PNAS nexus

PNAS Nexus, 2022, **1**, 1–14

https://doi.org/10.1093/pnasnexus/pgac165

Advance access publication date: 19 August 2022

Research Report

SECANT: a biology-guided semi-supervised method for clustering, classification, and annotation of single-cell multi-omics

Xinjun Wang[a,b], Zhongli Xu[c,d], Haoran Hu[a], Xueping Zhou[a], Yanfu Zhang[e], Robert Lafyatis[f], Kong Chen[f], Heng Huang[e], Ying Ding[a], Richard H. Duerr[f,1] and Wei Chen[a,c,*,1]

[a]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15213, USA
[b]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[c]Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA 15224, USA
[d]School of Medicine, Tsinghua University, Beijing 100084, China
[e]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15261, USA
[f]Department of Medicine, University of Pittsburgh, Pittsburgh, PA 15261, USA
*To whom correspondence should be addressed: Email: wec47@pitt.edu
Edited By: Shibu Yooseph

Abstract

The recent advance of single cell sequencing (scRNA-seq) technology such as Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) allows researchers to quantify cell surface protein abundance and RNA expression simultaneously at single cell resolution. Although CITE-seq and other similar technologies have gained enormous popularity, novel methods for analyzing this type of single cell multi-omics data are in urgent need. A limited number of available tools utilize data-driven approach, which may undermine the biological importance of surface protein data. In this study, we developed SECANT, a biology-guided SEmi-supervised method for Clustering, classification, and ANnoTation of single-cell multi-omics. SECANT is used to analyze CITE-seq data, or jointly analyze CITE-seq and scRNA-seq data. The novelties of SECANT include (1) using confident cell type label identified from surface protein data as guidance for cell clustering, (2) providing general annotation of confident cell types for each cell cluster, (3) utilizing cells with uncertain or missing cell type label to increase performance, and (4) accurate prediction of confident cell types for scRNA-seq data. Besides, as a model-based approach, SECANT can quantify the uncertainty of the results through easily interpretable posterior probability, and our framework can be potentially extended to handle other types of multi-omics data. We successfully demonstrated the validity and advantages of SECANT via simulation studies and analysis of public and in-house datasets from multiple tissues. We believe this new method will be complementary to existing tools for characterizing novel cell types and make new biological discoveries using single-cell multi-omics data.

Keywords: scRNA-Seq, CITE-Seq, single-cell multi-omics, semi-supervised learning

Significance Statement:

The recent advance of single-cell sequencing technology such as CITE-seq, which quantifies cell surface protein abundance and RNA expression, simultaneously at single-cell resolution, has quickly gained enormous popularity. Motivated by the fact that novel statistical methods or bioinformatical tools are in urgent need for analyzing such new data type, we developed SECANT to analyze CITE-seq data or jointly analyze CITE-seq and scRNA-seq data. As a biology-driven method, SECANT utilizes biological guidance derived from cell gating, which is considered as gold standard for cell type classification and will be of great usefulness to immunologists who are used to working with flow or mass cytometry data. Besides, SECANT quantifies the uncertainty of result with direct interpretation through a model-based approach.

Introduction

Single-cell RNA-sequencing (scRNA-seq) technologies have advanced rapidly for understanding cell heterogeneity and discovering rare cell types from normal and disease tissues (1, 2, 3, 4, 5). Embedded in the popular scRNA-seq platform such as the 10x Genomics Chromium System (6), the recently developed CITE-seq (Cellular Indexing of Transcriptomes and Epitopes by sequencing) (7) [or similar REAP-seq (RNA expression and protein sequencing) (8)], and cell hashing technologies (9) allow for immunophenotyping of single cells based on cell surface expression of specific proteins together with simultaneous transcriptome profiling and sample origin detection within a cell. More omics types of single

Competing Interests: The authors declare no competing interests.
[1]R.D. and W.C. contributed equally to this work.
Received: February 16, 2022. **Accepted:** August 16, 2022

cell data are emerging (10, 11, 12). In these single-cell multi-omics experiments, the abundance of different kinds of features such as mRNA or cell surface protein is converted into a quantitative and sequenceable readout through the use of DNA-barcoded antibodies and can be measured by the count of Unique Molecular Index (UMI) and Antibody-Derived Tags (ADT), respectively, simultaneously at single-cell resolution.

Although there are a large number of existing tools for analyzing droplet-based scRNA-seq data (13, 14, 15, 16, 17, 18, 19), model-based statistical methods for analyzing single-cell multi-omics data are still in urgent need. We will focus on analyzing CITE-seq data, one of the most informative multi-omics types, in this paper, although the method can be generalized to other types of bimodal multi-omics data where one data modality focuses on generating confident cell type labels in a supervised manner while the other data modality focuses on clustering cells in an unsupervised manner. For convenience, we use ADT data to denote surface protein data in this paper. CITE-seq refers to single-cell multi-omics data with both scRNA-seq and ADT measurements for each cell. There are a number of cutting-edge methods for surface protein imputation with scRNA-seq data (20, 21) and for joint clustering of both protein and RNA features (20, 22, 23). These methods, although having demonstrated their outstanding performance, all tend to utilize a data-driven approach but do not use much of existing biological knowledge. For example, it is a common approach in joint clustering methods to integrate protein and RNA data by transforming both features onto a similar space. However, by doing so, the important underlying biological information from surface protein marker could be undermined. On the contrary, biological researchers often consider cell surface markers as the gold standard to define cell types in molecular biology, where researchers identify distinguished cell types through cell gating such as flow cytometry with a list of classic differentiation (CD) markers, such as CD3, CD4, CD8, and CD19 (24, 25, 26, 27, 28). Thus, a more biological knowledge-driven approach should consider putting more weight on ADT data for the purpose of cell clustering and cell type identification. For example, ADT data are used to first label the well-defined (confident) cell types, such as B cells, Monocytes, CD4 + T cells, CD8 + T cells, and natural killer (NK) cells, which are then employed as guidance for clustering with RNA data. This approach utilizes a great amount of biological knowledge to avoid a common issue that some cell clusters identified with RNA data are in fact mixtures of multiple general cell types with respect to protein data (29, 30, 7). The recently developed scDCC proposed to integrate prior information into the modeling process to guide a deep learning model for latent representation and clustering, which can be applied to CITE-seq data (31). Besides cell clustering, ADT data can also play a role in cluster annotation. Current methods for cluster annotation rely on post-hoc differential expression (DE) analysis on RNA data, and researchers need to select a couple of plausible DE gene markers from a long list. However, it is often challenging since gene markers are not always correlated with their corresponding surface markers as observed from the data (32). Thus, we expect it vastly beneficial to provide researchers some confident cell type annotation from ADT data to help figure out the identities of cell subtypes. Another popular research topic is to jointly analyze data from CITE-seq and scRNA-seq, by which we can assume the cell compositions are similar though batch effect may exist. There are many advantages of joint analyzing CITE-seq and scRNA-seq data, e.g. the addition of an extra RNA data could help increase clustering performance due to a larger sample size, and we can also provide the additional confident cell type annotation identified with ADT data to scRNA-seq data.

Motivated by the above demands, in this study, we propose a novel framework, namely SECANT, for protein-guided cell clustering and general cluster annotation with CITE-seq data. If additional scRNA-seq datasets from similar cell populations are available, SECANT can be used to jointly analyze data from CITE-seq and scRNA-seq to predict confident cell types for scRNA-seq data, and enhance the performance of cell clustering and general annotation of confident cell types for each cell cluster. Our method utilizes a model-based approach motivated by classic statistical models in semi-supervised learning (33). As a biological knowledge-driven approach, the input of our method from ADT data is the confident cell type label, which can be obtained through cell gating or other existing methods (24, 25, 26, 27, 28). To overcome a common issue in cell gating that there are always cells distributed near or on the gating boundaries (e.g. the vertical or horizontal cut-off lines in a two-parameter scatter plot), which makes it hard to correctly identify their cell type with protein data, instead of excluding those cells from the analysis, which will cause the loss of sample size and potential the drop of some novel cell subtypes, our method is specifically designed to accommodate those cells with "uncertain" labels from protein data into our model so that we can fully utilize their transcriptomic information. In addition, as a model-based approach, our method can provide clustering and prediction uncertainty through posterior probability, which can be useful in downstream analysis, and readily adapted to rigorous statistical inference in a confirmatory study. We use extensive simulation studies to demonstrate the validity of our proposed method, and we illustrate the usefulness and easy interpretation of our method with CITE-seq datasets from human peripheral blood mononuclear cells (PBMC), bone marrow, and upper lobe lung tissues.

Results

General workflow of SECANT

The general workflow of SECANT is shown in Fig. 1. Our method can work with CITE-seq data (scRNA + ADT) only or jointly analyze CITE-seq and scRNA-seq data. When analyzing CITE-seq data only, the raw data matrices need to first undergo some data preprocessing steps, and the inputs of SECANT include the confident cell type labels built from ADT data and the latent space of RNA data after dimension reduction. SECANT considers ADT cell type labels as general guidance for cell clustering with RNA data by introducing certain constraints through a probabilistic concordance matrix. We establish a statistical model and maximize the log-likelihood of the observed data to estimate the concordance matrix and ADT-guided cell clustering results. Through the estimated concordance matrix, SECANT enables confident cell type annotation for each cluster (e.g. cluster 1 and cluster 2 are potential subclusters of B cells). For joint analysis of CITE-seq and scRNA-seq data, the latent spaces of pooled RNA data are required to be similar in distribution so that the clustering parameters are commonly shared by both data. The inclusion of the RNA data from scRNA-seq into the model could increase the precision of estimating concordance matrix as well as ADT-guided cell clustering result. Also, our model can predict the ADT confident cell type labels for cells from scRNA-seq experiment, where the latter does not have information regarding protein abundance. Other important benefits of SECANT include utilizing cells with uncertain cell type label from ADT data, and providing uncertainty of the results

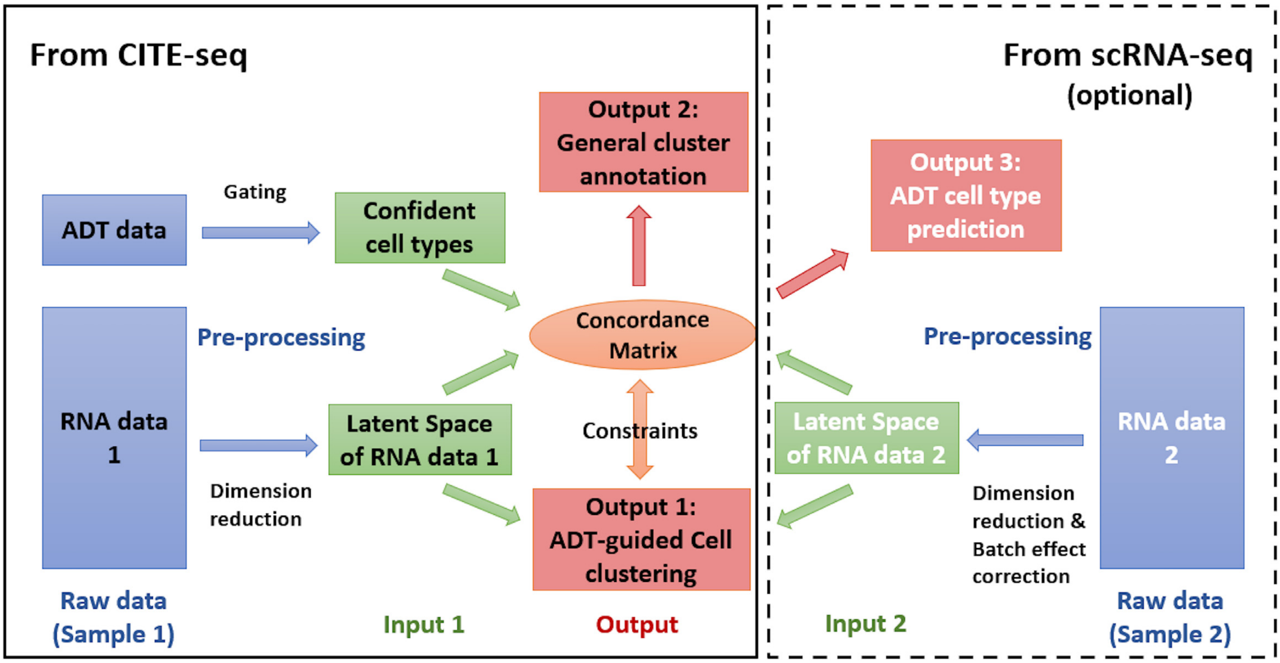
From CITE-seq
From scRNA-seq (optional)
ADT data
Gating
Confident cell types
Output 2: General cluster annotation
Output 3: ADT cell type prediction
RNA data 1
Pre-processing
Concordance Matrix
Pre-processing
RNA data 2
Latent Space of RNA data 1
Constraints
Latent Space of RNA data 2
Dimension reduction
Output 1: ADT-guided Cell clustering
Dimension reduction & Batch effect correction
Raw data (Sample 1)
Input 1
Output
Input 2
Raw data (Sample 2)

Fig. 1. General workflow of SECANT. In this study, we used manual gating or Gaussian mixture model (GMM) to classify confident cell types with ADT data, and used scVI for dimensional reduction and batch effect correction with RNA data for data preprocessing.

("soft labels") in terms of posterior probability. Details about our statistical models can be found in the "Methods" section.

Identifying confident cell types with ADT data

In the first step, we obtain the confident cell type label directly through manual gating with ADT data, motivated by the fact that it has been a widely used tool for cell type identification with flow cytometry and mass cytometry data, and the corresponding pipelines proposed by the biologists are quite mature (26, 28). In addition, the bi-modal or multimodal mixture Gaussian distribution structure of log-transformed ADT count fits gating pipeline naturally. In Fig. S1, we summarize a workflow to illustrate how to gate some confident cell types for PBMC from ADT data. However, one of the major challenges of manual gating is its subjective choice of gating boundary. In general, a less stringent boundary will introduce mixture of target cells with other cell types, while a more stringent boundary will lead to less target cells to be identified (Fig. S2). To overcome this challenge, our method is designed to utilize cells with uncertain cell type label. It is worth noting that tools other than manual gating can also be used for identifying confident cell types with ADT data (24, 25, 27).

Preprocessing of scRNA-seq data

RNA data need to undergo preprocessing for dimension reduction and batch effect correction. In this paper, we applied scVI to process real data for both purposes. scVI is a popular Python-based tool that utilizes variational autoencoder for nonlinear dimensional reduction and batch effect correction (15). We also tested the performance of batch effect correction on two public PBMC datasets (Fig. S3). To be specific, we first processed RNA data with scVI, and the resulting latent space, which follows a low-dimensional multivariate Gaussian distribution, is then used as the input of SECANT. In general, other tools for dimension reduction and batch effect correction such as Seurat (34) can also be used for data preprocessing in our proposed framework, although the distribution assumed in the statistical model is subject to change for better data fitting.

Simulation results

Clustering performance of SECANT with CITE-seq data

We first assessed the performance of SECANT with simulation studies. For convenience, we simulated data from mixture multivariate Gaussian distribution to mimic the latent space of RNA data through scVI, where the distribution parameters including cluster-specific mean vector, covariance matrix, and cluster weight are obtained from fitting SECANT on a public PBMC dataset processed by scVI. We fixed the number of confident cell types to be 5, assigned confident cell type labels according to the estimated best configuration (e.g. cluster 1 and 2 belong to confident cell type 1, etc.) and generated random noise on the simulated label to mimic the manual gating result from ADT data. To be specific, we randomly sampled a subset of cells, each with probability p^U, from the pool and changed their label to "uncertain." To assess the performance of our method under different scenarios, we varied the total number of cells (or sample size, $N = 500, 1000, 2000$), number of clusters ($K = 8, 10, 12$), dimensionality ($D = 5, 10, 20$), and proportion of uncertain labels ($p^U = 0, 0.2, 0.4, 0.6$) under different settings. We simulated 100 datasets under each setting.

In Fig. 2, we use Uniform Manifold Approximation and Projection (UMAP) plot (35), a popular nonlinear dimension reduction tool used in single cell analysis, to visualize an example of our simulated data ($N = 1000$, $K = 10$, $D = 10$, $p^U = 0.2$). Figure 2A is colored by the five confident cell types (type 1 to type 5), which refer to B cells, CD14 + Monocytes, CD4 + T cells, and CD8 + T cells, and NK cells, respectively, as in a PBMC dataset. Figure 2B is colored by the 10 clusters (cluster 1 to cluster 10), where clusters 1 and 2 belong to cell type 1, clusters 3 and 4 belong to cell type 2, clusters 5, 6, and 7 belong to cell type 3, clusters 8 and 9 belong to

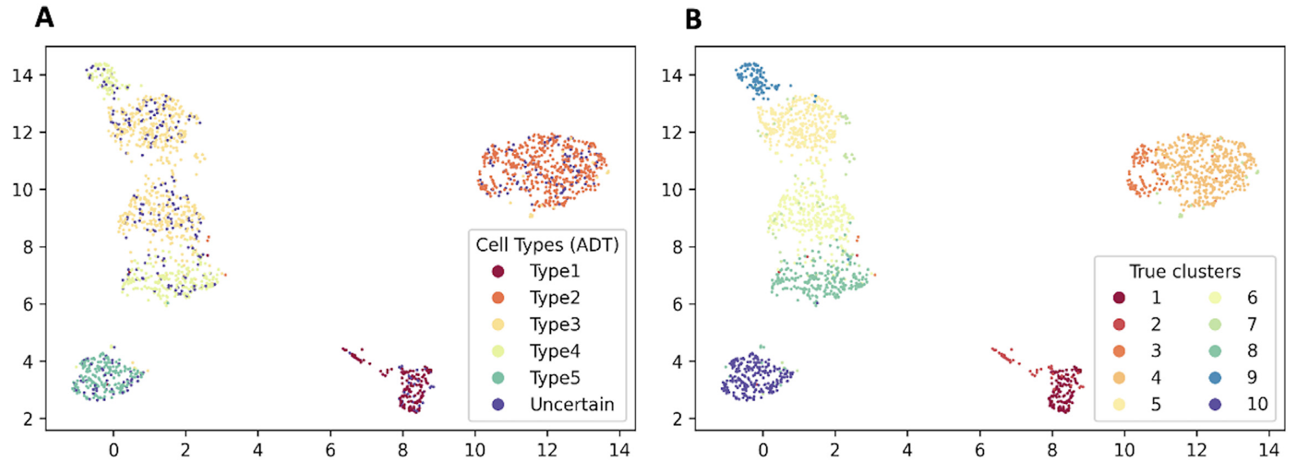

Fig. 2. UMAP visualization of an example simulated data (mimic latent space of RNA data) under our simulation setting. 2A: cells are colored by simulated true cell types (mimic input from ADT data). 2B: cells are colored by the simulated true cluster assignments.

cell type 4, and cluster 10 belongs to cell type 5. This simulation setting largely reflects the real-world scenario that CD4 + T cells and CD8 + T cells are often clustered together with scRNA-seq data.

We applied K-means and multivariate GMM to the simulated data (without information from ADT data), set the number of clusters as the true value as in the simulation design, and summarized their results for reference. We alter the total number of cells, denoted by N, as well as the proportion of cells randomly assigned with the uncertain label, p^U. In Fig. 3A and B, we show the boxplot of adjusted random index (ARI) (36) from SECANT under different p^U settings with varying N and K. We observe that the clustering performance of SECANT increases with a larger sample size, but decreases with a larger number of clusters. Indeed, as a model-based approach, a larger sample size will lead to better parameter estimation, which will increase the clustering performance. However, with a fixed total number of cells ($N = 1000$), a larger number of clusters leads to a smaller cluster-specific sample size, which diminishes the performance of a model-based approach. We also observe that the clustering performance decreases with larger p^U, which is as expected since we get less information from ADT confident cell type with increased p^U. In Fig. 3C, we assess the effect of feature dimension on clustering performance of SECANT when $N = 1000$, and observe that SECANT performs best when $D = 10$. Although data with higher dimensionality contain more information, which explains why $D = 10$ leads to a better result than $D = 5$, for a model-based approach, the parameters to be estimated also increase, which requires a larger sample size for a good estimation result. To further explore on this, in Fig. 3D, we increase the sample size to $N = 2000$, and observe that the performance of $D = 10$ and $D = 20$ are very similar, both outperform $D = 5$ notably. Based on this result, we generally recommend setting $D = 10$, which is also the default setting in scVI for dimension reduction. In practice, users may increase the value of D with larger sample size (Fig. S4). In addition, compared with the performance of multivariate GMM and K-means, SECANT performs the best among the three across all scenarios, even when 60% of cells are labeled as “uncertain” in the input ADT cell type label. The result of adjusted mutual information (AMI) (37) shows consistent pattern as ARI (Fig. S5).

Performance of joint analysis of SECANT with CITE-seq and scRNA-seq data

To assess the performance of SECANT for joint analysis of CITE-seq and scRNA-seq data, in each simulation we generated a pair of datasets with the same distribution parameters (i.e. cluster-specific mean, covariance matrix, and cluster weight), and the aforementioned data generation method is used. Therefore, both datasets are composed of ADT label and RNA data after dimension reduction. We set $D = 10$, $K = 10$, with varying N and p^U. Next, we masked the ADT label from one dataset, which is pretended as scRNA-seq data. Under this setting, we can evaluate if the additional “unlabeled” scRNA-seq data could help increase the clustering performance, and also assess the prediction accuracy of ADT confident cell type label for the dataset whose ADT labels are masked. Here, we claim the predicted confident cell type for a cell is accurate if it is the same as the simulated true confident cell type, and compute the proportion of accurate predictions among all cells. In Fig. S6A and B, we compare the clustering performance of SECANT with CITE-seq data only and with paired CITE-seq and scRNA-seq data. In addition to the similar patterns in Fig. 3A, we conclude that the inclusion of additional scRNA-seq data into our model can help increase the clustering performance with regard to ARI and AMI, especially when sample size is relatively small.

In practice, it could happen when the paired CITE-seq and scRNA-seq data have different cluster weights. For example, one PBMC data have relatively more CD4 + T cells while the other data have relatively more CD8 + T cells. To evaluate the performance of SECANT under such circumstance, we designed several additional simulation scenarios (in addition to Scenario 1) as follows:

- Scenario 1: scRNA-seq data have the same set of cluster weights as CITE-seq data.
- Scenario 2: The cluster weights of scRNA-seq data are evenly distributed (different from cluster weights of CITE-seq data).
- Scenario 3: The cluster weights of scRNA-seq data are evenly distributed conditional on missing two small-size clusters as in CITE-seq data.
- Scenario 4: The cluster weights of scRNA-seq data are evenly distributed conditional on missing two big-size clusters as in CITE-seq data.

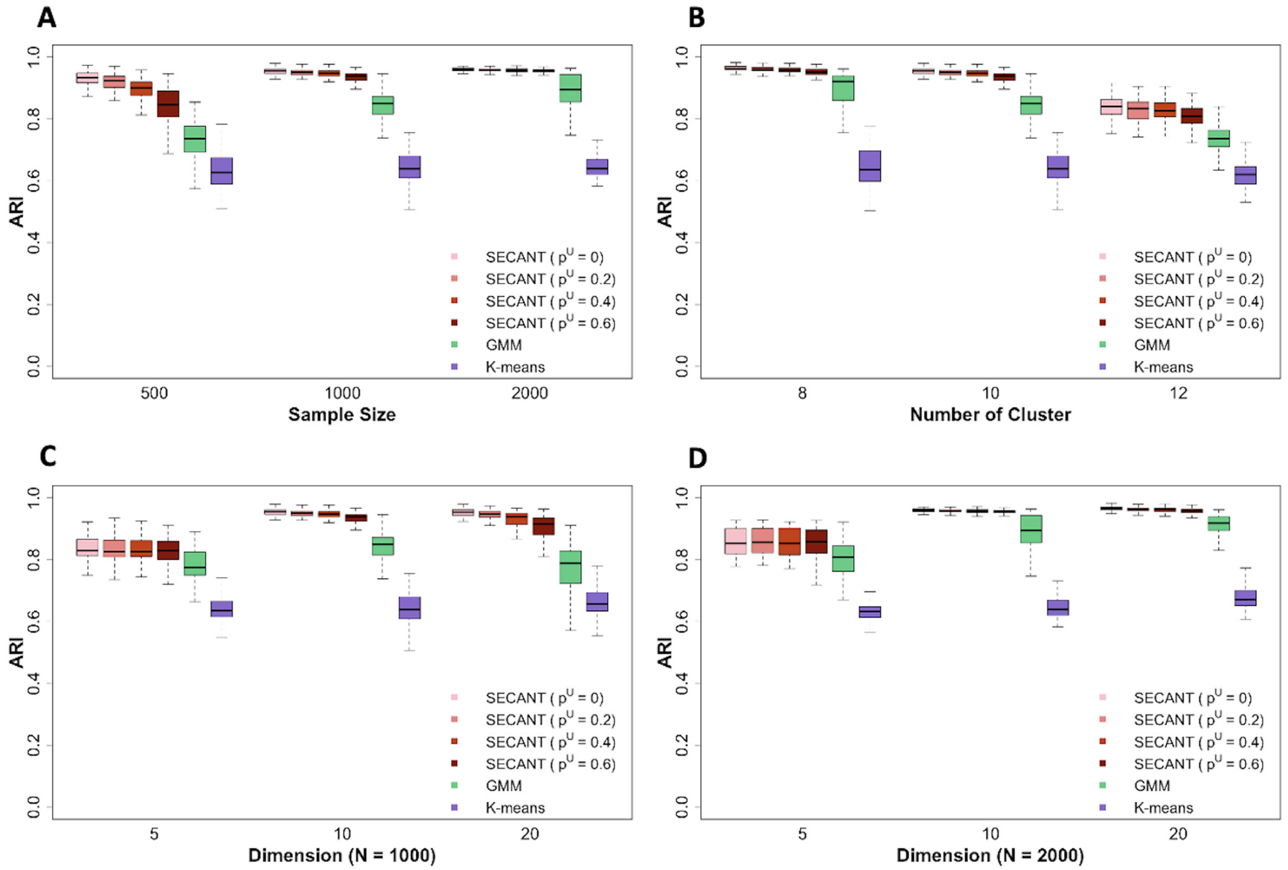

Fig. 3. The distribution of ARI of clustering result from SECANT (with different p^U setting), K-means, and GMM compared with the true cluster label generated in simulation study. 3A: varying by sample size N. 3B: varying by the number of simulated clusters K. 3C: varying by feature dimension D when $N = 1000$. 3D: varying by feature dimension D when $N = 2000$.

The specific values of cluster weights under each scenario are summarized in Table S1. For each scenario, we computed the mean and standard deviation (SD) of the bias of our estimated cluster weight, compared with the prespecified values, and the results are summarized in Table S2. In general, the absolute values of bias as well as SD decrease as sample size increases, and we conclude our method estimates the cluster weight very well across all scenarios even with a low sample size of 500.

We further assess the accuracy of predicting ADT confident cell types for scRNA-seq data under each of the four scenarios described above with varying N and p^U (Fig. 4A to D). To be specific, if the predicted confidence cell type is uncertain, we classify this outcome as inaccurate. Although it is observed that the mean predication accuracy decreases with increasing p^U, the prediction accuracy actually remains very high if the uncertain rate of input label is less than 40%. The performance breaks down drastically when p^U is greater than 50%, which is as expected since more than half of the input label provides no information. Similar results are found in robust mixture discriminant analysis (RMDA) (38). Also, it is interesting to observe that the prediction accuracy increases with larger sample size when p^U is smaller than 50%, but such a trend is reversed when p^U is greater than 50%. We further evaluated the cause of the observed low prediction accuracy when p^U is greater than 50%, and found that most cells are predicted as "uncertain" by SECANT when p^U is high (Fig. S7). Overall, we demonstrate the validity and the outstanding performance of SECANT for jointly analyzing CITE-seq data and scRNA-seq data for clustering and ADT confident cell type prediction through simulation studies.

Detecting cell clusters when there is no prior biology knowledge from surface protein data

In practice, there could exist some cell clusters that we fail to identify their confident cell type label with ADT data due to the lack of biology knowledge or unavailable on-shelf markers (e.g. the residual cells from cell gating). Different from other cell clusters which only have a subset of cells labeled as uncertain cell types, cells in such clusters are entirely labeled as uncertain cell type. Since SECANT is biology-driven, it could lose power when there lacks prior biology knowledge. To mimic this real-world situation, we first simulated data with five confident cell types and 10 underlying clusters similar to the setting in previous *"Clustering performance of SECANT with CITE-seq data"* section. Then, in addition to the random noise, we changed the confident cell type label of entire cells in a specific cluster into "uncertain" to pretend that we do not have prior knowledge to identify those cells. We further designed two scenarios: (1) a large-size cluster has been "converted" (cluster 6 in Fig. 2B, 15.3%); and (2) a small-size cluster has been "converted" (cluster 2 in Fig. 2B, 2.9%), which are shown in Fig. S8A and C. For scenario (1), SECANT can successfully identify those converted cells as a unique cell cluster (Table S3A; Fig. S8B) and achieves high clustering performance (ARI = 0.944; AMI = 0.934). For scenario (2), SECANT fails to identify those converted cells as a unique cell cluster but merge them with a small number of other

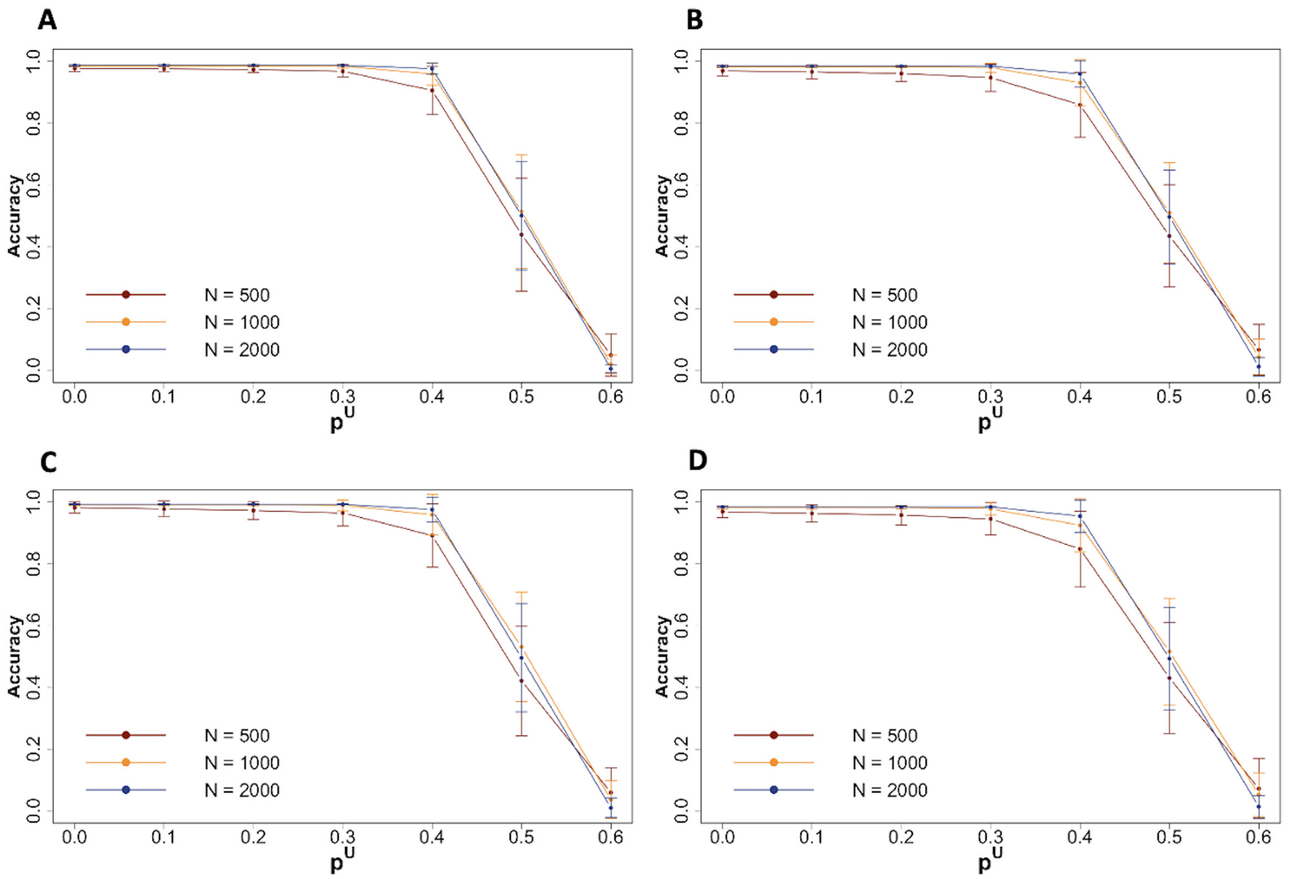

Fig. 4. Prediction accuracy of ADT confident cell types for simulated scRNA-seq data (on latent space) with various p^U and N settings. A to D correspond to simulation scenario 1 to 4 as described in the section "Performance of joint analysis of SECANT with CITE-seq and scRNA-seq data," with different simulated cluster weights.

cells (Table S3B; Fig. S8D), although the overall clustering performance is still outstanding (ARI = 0.933; AMI = 0.917) since the majority of cells are still clustered appropriately. In summary, as a model-based and biology-driven approach, our method is powerful to detect cell clusters either with large size or with prior knowledge, but not sensitive to identify small clusters especially with limited biology knowledge.

Real data applications

ADT-guided cell clustering with CITE-seq dataset

To illustrate the usefulness of ADT-guided cell clustering proposed in SECANT, we applied SECANT to three CITE-seq datasets from different human tissues, including human PBMC, bone marrow, and upper lobe lung. For PBMC and bone marrow data, we classified ADT confident cell types through manual gating (Fig. S1), and identified five and six major cell types, respectively (Figs. 5A and 6A). For upper lobe lung data, we failed to manually gate confident cell types due to lack of existing pipeline, so we applied GMM as an alternative approach and identify three major cell types. We applied scVI to RNA count matrix in each CITE-seq data for dimension reduction, and a 10D latent space was extracted as the input of SECANT.

The public PBMC CITE-seq dataset, denoted by 10x10k_PBMC (7,865 cells), is from a healthy donor and provided by 10x Genomics. Cells are gated into five confident cell types with ADT data, including B cells, CD14 + Monocytes, CD4 + T cells, CD8 + T cells, and NK cells (Fig. 5A). Through a relatively conservative gating, the proportion of cells labeled as uncertain cell type is 16.3%. We failed to identify CD16 + Monocytes with ADT data due to its low amount. We set the number of clusters to be 11. The clustering result is shown in Fig. 5B, from which we observe that none of the clusters is obviously a mixture of multiple ADT confident cell types. Further, the estimated concordance matrix from SECANT provides the correspondence between ADT confident cell types and RNA clusters (Table 1, Fig. S9), which can help guide the following annotation step. Based on DE genes through post-hoc analysis and existing literature, we successfully annotated each cluster (Table 1) (39). We also applied Seurat and totalVI to this PBMC dataset. Both Seurat and totalVI are data-driven methods and utilize graphic-based algorithm for cell clustering with CITE-seq data as the input. We controlled the number of clusters to be 11 when applying Seurat and totalVI, the same value we set in SECANT. In general, we observe that most of the identified clusters are consistent among three methods (Fig. 5C and D). Although there is no ground truth to compare with, the pairwise ARI among those three methods are 0.75 (SECANT versus Seurat), 0.82 (SECANT versus totalVI), and 0.78 (Seurat versus totalVI), which indicates the concordances among three clustering results are at similar level. With a more detailed comparison (Table S4), we find that the major differences of clustering results between SECANT and the other two methods are the identification of marginal zone B cells (cluster 2 in SECANT), dendritic cells (cluster 3 in SECANT), and Gamma Delta T cells (cluster 7 in SECANT), possibly due to their low amount. In addition, as demonstrated in Fig. 5 and Table S4, SECANT identifies three subtypes of

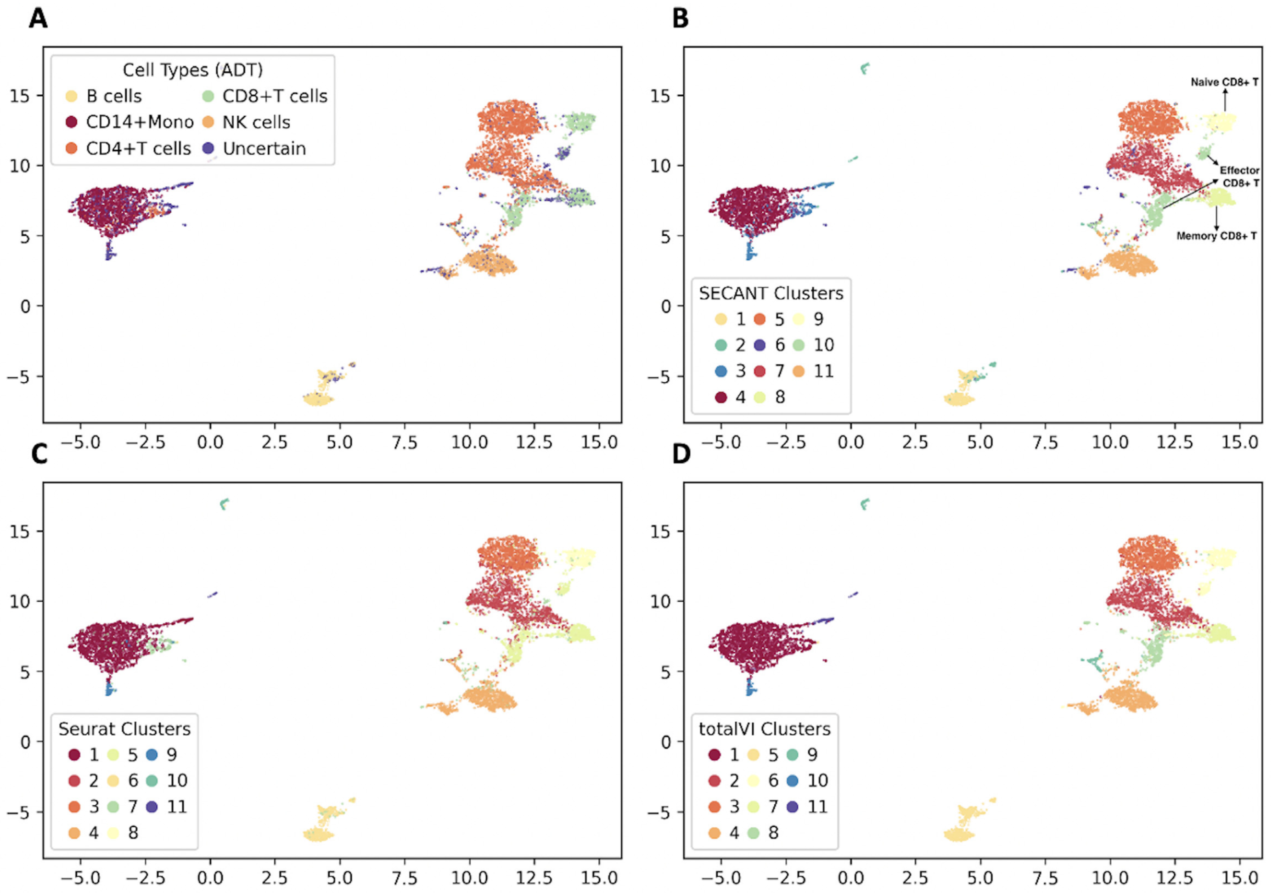

Fig. 5. UMAP visualization of the latent space (dimension reduction with scVI) of RNA data from 10x10k_PBMC. 5A: cells are colored by ADT confident cell types through manual gating. 5B: cells are colored by SECANT result. 5C: cells are colored by Seurat result. 5D: cells are colored by totalVI result.

Table 1. Estimated concordance matrix and post-hoc subtype identification from SECANT for 10x10k_PBMC dataset.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11
B cells	0.969	0.474	0	0	0	0	0	0	0	0	0
CD14 + Monocytes	0	0	0.202	0.877	0	0	0	0	0	0	0
CD4 + T cells	0	0	0	0	0.986	0.628	0.910	0	0	0	0
CD8 + T cells	0	0	0	0	0	0	0	0.745	0.934	0.744	0
NK cells	0	0	0	0	0	0	0	0	0	0	0.884
Uncertain	0.031	0.526	0.798	0.123	0.014	0.372	0.090	0.255	0.066	0.256	0.116
Cluster weight	0.064	0.031	0.054	0.219	0.151	0.030	0.154	0.053	0.049	0.079	0.117
SECANT annotation	**Follicular B cells**	**Marginal zone B cells**	**Dendritic cells**	**CD14 + Monocytes**	**Naïve CD4 + T cells**	**Gamma delta T cells**	**Memory CD4 + T cells**	**Memory CD8 + T cells**	**Naïve CD8 + T cells**	**Effector CD8 + T cells**	**NK cells**
Selected DE genes	IGHM CD79A MS4A1 IGHD CD22	MZB1 TNFRSF 17 CD1 CD27	CSF1R CST3	S100A9 S100A8 LYZ	CCR7 SELL CD3D	GZMK KLRB1	TRAC IL7R LTB	GZMK IL7R CD8A CD69 KLRB1	SELL CCR7	CD8B CD8A GZMK NKG7 GZMM	

CD8 + T cells, including naïve CD8 + T cells (cluster 9), effector CD8 + T cells (cluster 10), and memory CD8 + T cells (cluster 8), but Seurat only identifies two clusters where memory CD8 + T cells and effector CD8 + T cells are combined in one cluster (cluster 5).

The human bone marrow CITE-seq dataset (30,672 cells) is public in literature and also available in Seurat package ("bmcite") (34). Although Seurat identifies 27 cell clusters in this dataset, many of which are small-size clusters (e.g. less than 1%). Thus, we set a smaller number of clusters ($K = 13$) when running SECANT. Through manual gating with ADT data, in addition to the five confident cell types identified in the aforementioned PBMC dataset, we also detected CD16 + Monocytes (Fig. 6A). The proportion of cells labeled as uncertain cell type is 19.7%. Compared to clusters identified with Seurat (Fig. 6B), we fail to classify some cell types, such as progenitor B cells, progenitor dendritic cells and hematopoietic stem cells, into one of the confident cell types according to our cell gating pipeline (Fig. S1). As a result, SECANT

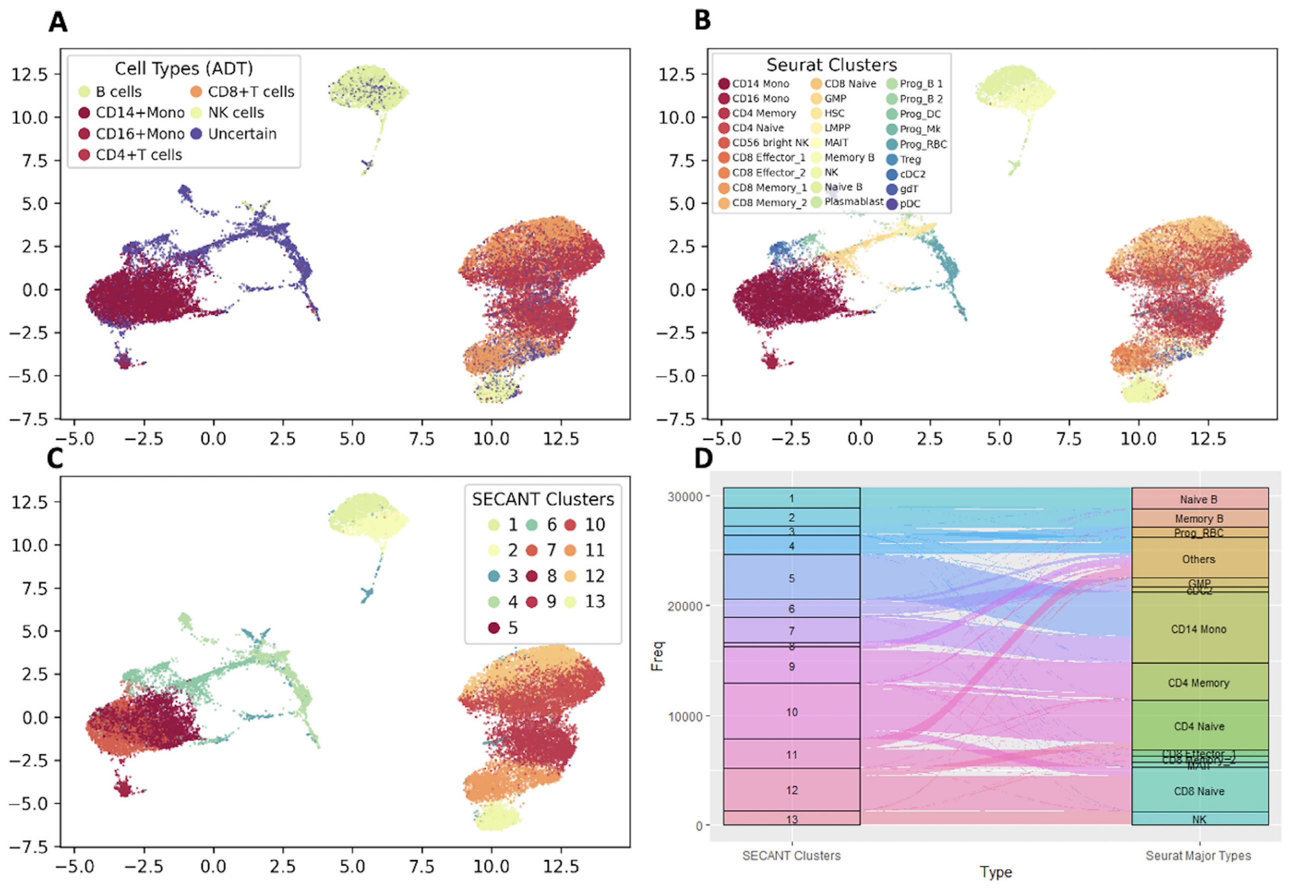

Fig. 6. UMAP visualization of the latent space of RNA data from human bone marrow dataset. 6A: cells are colored by ADT confident cell types through manual gating. 6B: cells are colored by annotated clusters provided in Seurat package. 6C: cells are colored by SECANT result. 6D: alluvial plot showing the correspondence between SECANT clustering result and Seurat annotated result. *For better visualization, we grouped small-size clusters (<1.5%, 14 clusters) in Seurat into one cell type, named "Others," in alluvial plot.

does not perform well to identify those cell types due to lack of biological guidance. On the other hand, the clustering result from SECANT is in general consistent with Seurat result especially for medium or large-size clusters (Fig. 6C). We provide cluster annotation based on our estimated concordance matrix as well as post-hoc DE analysis in Table S5A. We further compare the clustering result from SECANT with Seurat cluster annotation. As expected, Seurat performs better at detecting small-size clusters, such as subtypes of different progenitor cells or CD8 + T cells (Fig. 6D, Table S5B).

The human upper lobe lung CITE-seq dataset was publicly available on Gene Expression Omnibus (GEO) under GSE128169 (sample SC277) (40), which contains unfiltered raw feature barcodes. Thus, as a data processing step, we applied *DropletUtils* to remove background noise (41, 42). The filtered data contain 5,756 cells. Due to lack of existing pipeline for manual gating, we applied GMM as an alternative approach to log transformed ADT data (17 surface markers) and set the number of clusters to be 7. Based on post-hoc DE analysis (Fig. S10), we identified three confident cell types from six clusters, including epithelium cells, endothelium cells and immune cells, and selected those cells for clustering analysis (5,451 cells). Since GMM is a model-based approach, we can estimate the clustering uncertainty for each cell. Thus, we labeled cells with more than 10% uncertainty about their cluster assignments as uncertain cell type (13.2%), and the obtained ADT confident cell types are relatively conservative (Fig. S11A). We set the number of clusters to be 6, and the clustering result from SECANT can be visualized through UMAP (Fig. S11B). Based on post-hoc analysis, we not only confirmed the identity of epithelium cells and endothelium cells, but also detected three specific types of immune cells, including monocytes, T cells, and Macrophage, although one of the clusters appear to be a mixture of multiple cell types and could undergo further subclustering (Table S6).

To assess the "soft-clustering" property of SECANT, we computed the posterior probability of clustering assignment for each cell in the aforementioned three CITE-seq datasets, and utilized UMAP plot for visualization (Fig. S12). In general, we observe that cells with relatively low confidence concentrate on the boundary of different cell clusters on UMAP plot. Therefore, compared to "hard-clustering" algorithms, SECANT can provide probabilistic uncertainty measurement for each cell, which can be used to enhance precision in downstream analysis, or to assess the robustness of the result in a sensitivity analysis.

Joint analysis of CITE-seq and scRNA-seq data for confident cell type prediction

In addition to ADT-guided clustering and cluster annotation, SECANT can also jointly analyze CITE-seq and scRNA-seq data to predict ADT confident cell types for scRNA-seq data. Based on the result from simulation study, a general assumption is that the two RNA data, from CITE-seq and scRNA-seq, should have similar cell type compositions (e.g. from same tissue), although the cluster weights could differ. In this study, we prepared a pair of public

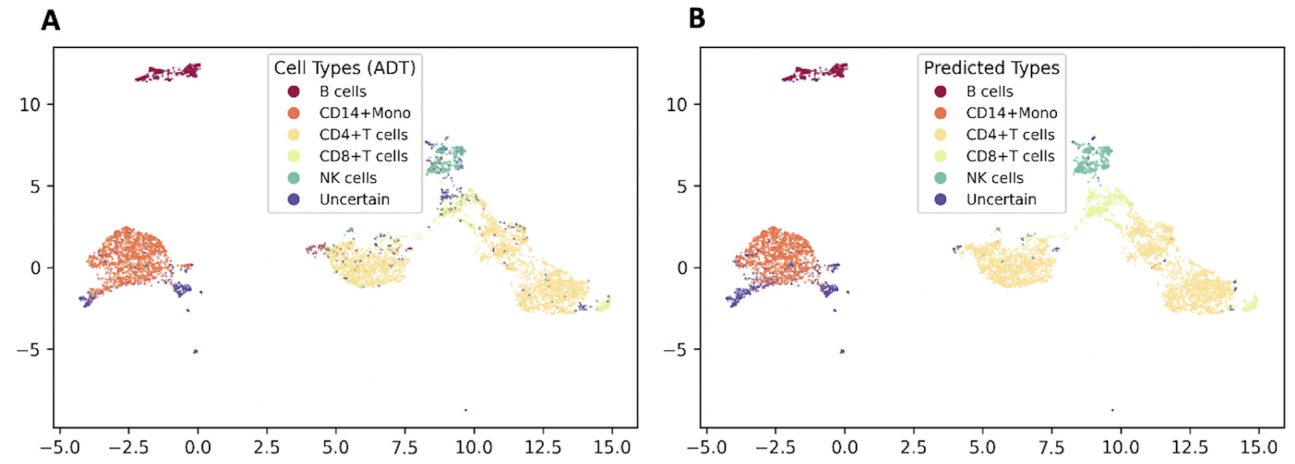

Fig. 7. UMAP visualization of the latent space of RNA data from 10x5k_PBMC. 7A: cells are colored by ADT confident cell types through manual gating. 7B: cells are colored by predicted confident cell types from SECANT.

PBMC CITE-seq datasets as well as a pair of in-house PBMC CITE-seq datasets to illustrate the performance of SECANT for confident cell type prediction. For each CITE-seq dataset, we first used manual gating to get confident cell types with ADT data. Then, for each pair of datasets, we removed the ADT confident cell type label from one of the datasets and pretended that dataset was generated from scRNA-seq. Next, we applied SECANT to jointly analyze each pair of datasets, and predicted the ADT confident cell type for the dataset without ADT labels, which is then compared to the true label we previously removed. The major differences between public and in-house PBMC datasets are (1) there is batch effect in public datasets but not in in-house datasets; and (2) the paired in-house PBMC datasets are more homogenous since they are aliquots of the same sample. Again, we applied scVI to RNA count matrices in paired datasets for both batch effect correction and dimension reduction, and a 10D latent space from each dataset was extracted as the input of SECANT.

The two public PBMC CITE-seq datasets are available on 10x Genomics website, denoted by 10x10k_PBMC (7,865 cells) and 10x5k_PBMC (5,527 cells). In previous section, we identified five confident cell types through gating with ADT data for 10x10k_PBMC dataset. Similarly, we identified the same five confident cell types in 10x5k_PBMC datasets (11.5% uncertain labels), but we temporarily removed these labels to pretend 10x5k_PBMC is a scRNA-seq dataset. As shown in Fig. S3A and B, although scVI has largely reduced the batch effect between two datasets, we still observe a cluster of cells that are dominated by 10x5k, which could be due to the different cluster weights between two samples. We set the number of clusters to be 11, and then applied SECANT for ADT-guided clustering for both datasets and predicting ADT confident cell types for 10x5k_PBMC dataset. The clustering results are shown in Fig. S13, and the estimated concordance matrix is summarized in Table S7, from which we observe the estimated cluster weights are obviously different for some clusters, possibly because the two samples are not homogenous. To assess the performance of ADT confident cell types prediction, we compared the predicted labels with the observed labels, the latter of which were not used in SECANT. The UMAP visualization of the result is shown in Fig. 7A and B, and we observe that the major difference is among cells that are either classified as "uncertain" or predicted as "uncertain." In general, the predicted labels are close to the observed labels. We also computed the confusion matrix between the predicted and the observed cell types (Table 2). Excluding cells with observed "uncertain" cell type, the overall prediction accuracy achieves 89.1%. This result is consistent with our simulation result that different cluster weights do not influence much on the prediction accuracy.

To further investigate the performance of confident cell type prediction using SECANT, we generated two in-house CITE-seq datasets of PBMCs from a healthy donor. The antibody concentration used in each sample is different to mimic different quality of ADT data, denoted as concen_high and concen_low. Similar to the joint analysis of aforementioned paired public PBMC datasets, we first generated ADT confident cell type label (i.e. the same 5 types as in public PBMC data) through manual gating. The proportion of uncertain cell type in concen_high (1,587 cells) is 17.5%, whereas in concen_low (2,112 cells) the proportion increases to 39.4% due to a lower concentration of antibodies. We temporally removed the ADT confident cell type labels in the concen_low dataset and pretended this dataset was generated from scRNA-seq. Since there is no batch effect between two RNA data (Fig. S14), we only applied scVI for dimension reduction but not correcting for batch effect. Due to small sample size, we set the number of clusters to be 8, and then applied SECANT for ADT-guided clustering for both datasets and predicting ADT confident cell types for concen_low dataset. The clustering results are shown in Fig. S15, and the estimated concordance matrix is summarized in Table S8, from which we observe the estimated cluster weights are close between two data for all clusters, because the two samples are homogenous in nature. Similarly, we find the predicted labels are quite close to the observed labels (Fig. 8A and B). We also computed the confusion matrix between the predicted and the observed cell types (Table 3). Excluding cells that are classified as "uncertain", the overall prediction accuracy achieves 95.2%.

Discussion

In this study, we have developed SECANT, a biology-guided semi-supervised method for cell clustering, cell type classification, and annotation for analyzing CITE-seq data alone or jointly with scRNA-seq data. Different from other existing tools for single-cell multi-omics, SECANT utilizes a biology-driven approach and considers that cell surface protein data can provide confident cell type labels, which are assumed to be the gold standard in single cell proteomics experiments such as flow cytometry and mass cytometry, and thus should be used to guide cell clustering with

Table 2. Confusion matrix of predicted confident cell types versus observed cell types built with ADT data from 10x5k_PBMC dataset.

	Observed					
Predicted	**B cells**	**CD14 + Monocytes**	**CD4 + T cells**	**CD8 + T cells**	**NK cells**	**Uncertain**
B cells	305	0	0	0	0	*26*
CD14 + Monocytes	1	1,086	19	0	0	*29*
CD4 + T cells	0	33	2,350	124	4	*116*
CD8 + T cells	0	0	65	336	1	*109*
NK cells	1	0	11	7	280	*60*
Uncertain	22	187	42	4	12	*297*

Numbers italicized are excluded when computing accuracy.

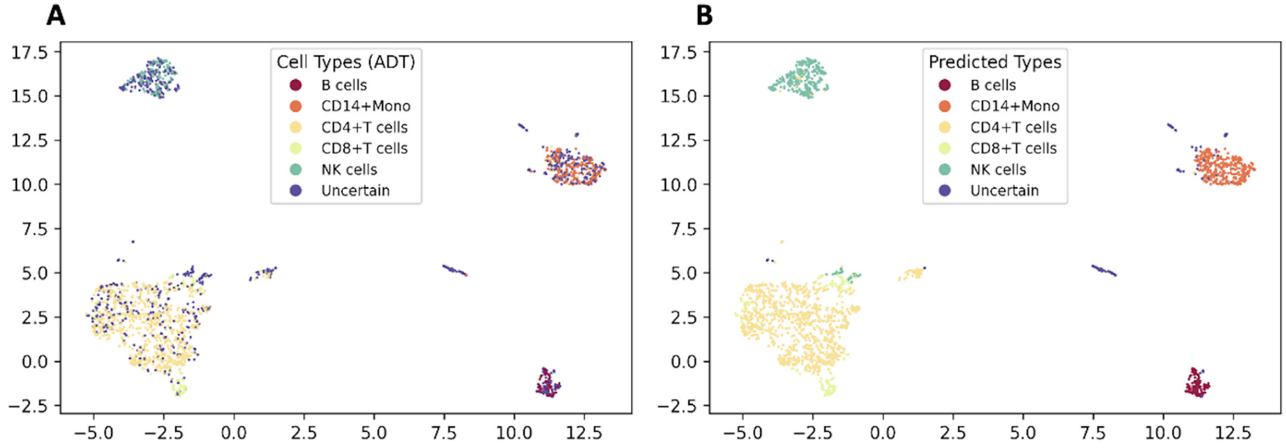

Fig. 8. UMAP visualization of the latent space of RNA data from concen_low. 8A: cells are colored by ADT confident cell types through manual gating. 8B: cells are colored by predicted confident cell types from SECANT.

Table 3. Confusion matrix of predicted confident cell types versus observed cell types built with ADT data from concen_low dataset.

	Observed					
Predicted	**B cells**	**CD14 + Monocytes**	**CD4 + T cells**	**CD8 + T cells**	**NK cells**	**Uncertain**
B cells	70	0	0	0	0	*65*
CD14 + Monocytes	0	173	10	0	0	*139*
CD4 + T cells	0	0	735	16	1	*256*
CD8 + T cells	0	1	11	108	0	*60*
NK cells	0	0	1	1	132	*183*
Uncertain	3	14	3	0	0	*130*

Numbers italicized are excluded when computing accuracy.

RNA data. Our proposed method is developed based on model-based semi-supervised learning, and we introduce a probabilistic concordance matrix to implement ADT constraints as well as for cluster annotation. When several related scRNA-seq data are available, jointly analyzing CITE-seq and scRNA-seq data with SECANT can provide annotation of confident cell types, which are constructed with ADT data from CITE-seq, for cells from scRNA-seq, and the ADT-guided clustering performance is expected to enhance.

Still, several limitations exist for SECANT. First, in this study, the input of SECANT from ADT data is the confident cell type label built through manual gating, which undergoes a relatively subjective process. For example, a less conservative gating approach will introduce noise to cell label, while a more conservative approach will lead to the loss of sample size. In practice, as a preliminary step of SECANT, we suggest that researchers gate cells more conservatively, and leave cells on the boundary as “uncertain” cell type (e.g. Fig. S2). SECANT is designed to fully utilize cells with uncertain cell type identified with ADT data, thus a conservative gating approach would not lead to the loss of sample size, but could sufficiently reduce the labeling noise. In addition, other methods, e.g. auto gating, can also be used to build confident cell type label as the input for SECANT. Second, SECANT employs stochastic gradient descent (SGD) for optimization, which is a computationally expensive approach. To speed up, we implement our algorithm in PyTorch (a Python library from Facebook) and utilizes tensor broadcasting. Although PyTorch is well-known for deep learning, it can be used to optimize a target function through SGD without building neural networks. Also, it is extremely convenient with PyTorch to use graphics processing units (GPUs) for strong acceleration, and we have implemented our algorithm with GPU setting (e.g. freely available on Google Colab). We further benchmarked the computational speed and memory consumption of SECANT in real data applications in Table S9 under GPU setting. Third, to estimate the configuration of the matrix form of the concordance matrix C without any prior knowledge, currently we need to run all possible configurations, and then select the one with the maximum log-likelihood. To speed up, one can run SECANT in parallel, each thread running a different configuration. In addition, we are developing an alternative approach for optimization, which utilizes the alternating direction method of multipliers (ADMM), in a separate study. The alternative approach is expected to be more efficient than the current approach. Lastly, SECANT is a model-based approach, which requires sufficient sample size for proper parameter estimation and valid statistical inference. Therefore, SECANT may not be powerful to detect small cell clusters (e.g. rare cell types) when the total number of cells in the dataset is relatively small. However, given the rapid advance in technology, a typical CITE-seq experiment can now measure the expression profile of more than 5,000 cells, which can be further increased to over 10,000 if coupled with sample multiplexing methods such as cell hashing, which no longer limits the usage of SECANT in real applications.

In summary, we propose a novel statistical method, SECANT, which utilizes model-based semi-supervised learning for surface protein guided cell clustering, classification and annotation with CITE-seq data or joint analysis with CITE-seq and scRNA-seq data. Our model framework can be extended to accommodate single cell data from other two data sources (e.g. the recently developed ASAP-seq (12)), or to analyze data from other fields. Additionally, our well-designed in-house CITE-seq datasets will be valuable for researchers to develop novel methods. We believe SECANT would quickly gain popularity among medical researchers, particularly in immunology filed.

Methods

Statistical models

We first denote L_i the confident cell type label for cell i obtained from ADT data. We assume there are in total M confident cell types identified with ADT data, and the support of L is $\{1, 2, \ldots, M, M+1\}$, where $L = 1, 2, \ldots, M$ corresponds to each of the confident cell types, and $L = M+1$ refers to the additional "uncertain" group. We then denote Z_i the cell cluster label for cell i estimated from RNA data. Assuming the total number of clusters is K, the support of Z is $\{1, 2, \ldots, K\}$. The core assumption of our approach is that cells should not fall into the same cluster identified with RNA data if they are classified as different confident cell types (not including the "uncertain" group) with ADT data. For example, if one cell is identified as a CD4 + T cell and another as a CD8 + T cell confidently from ADT data, then we should avoid these two cells being clustered together with RNA data. As an exception, this constraint does not apply to cells falling into the "uncertain" group, which makes our assumption biologically plausible. This constraint can be described mathematically as follows:

$$\text{For cell } i \text{ and cell } j,\ \text{where } i \neq j,\ \text{if } L_i \neq M+1,\ L_j \neq M+1,$$
$$\text{and } L_i \neq L_j,\ \text{then } Z_i \neq Z_j, \tag{1}$$

Equivalently, we can state our constraint [1] in a statistical way by introducing a concordance matrix $C_{(M+1)\times K}$, as shown in Table 4. We denote p_{mk} in the matrix the conditional probability $P(L_i = m|Z_i = k)$ for cell i. Under constraint [1] and general assumptions, we have the following constraints on the first M rows of C, denoted by $C^*_{M\times K}$, where each row corresponds to a confident cell type:

1) Each column of C^* contains exactly one nonzero parameter.
2) Each row of C^* contains at least one nonzero parameter.

The last row of C, referring to the "uncertain" group, can then be decided with parameters implemented in C^*. In general, there are multiple configurations of the matrix form of C that fulfill the constraints described above, and there are K parameters to be estimated for each configuration. We will discuss more about the concordance matrix C in the next section. After introducing the concordance matrix C, we can then write out the likelihood function.

Scenario 1: CITE-seq data only

We denote Y the latent space of RNA data, and each element Y_{ij} represents the value for feature j in cell i, where i runs from 1 to the total number of cells N and j runs from 1 to the number of dimensions in latent space D. We further assume the total number of clusters identified from RNA data is K. The likelihood can be written as

$$P(Y,\ L) = \prod_{i=1}^{N}\{\sum_{k=1}^{K} P(L_i|Z_i = k)P(Z_i = k|Y_i)\} \prod_{i\,=\,1}^{N} \left\{ \prod_{k\,=\,1}^{K} f(Y_i|\theta_k)^{1(Z_i=k)} \right\}, \tag{2}$$

where $P(Z_i = k|Y_i) = \frac{\tau_k f(Y_i|\theta_k)}{\sum_{k\,=\,1}^{K} \tau_k f(Y_i|\theta_k)}$ refers to the posterior probability of cell i belonging to cluster k, $\tau_k = P(Z_i = k)$ refers to the proportion of cluster k, and $f(Y_i|\theta_k)$ refers to the cluster-specific distribution of RNA data on latent space. In this study, we assume $f(Y_i|\theta_k) = \frac{\exp(-\frac{1}{2}(Y_i-\mu_k)^T \Sigma_k^{-1}(Y_i-\mu_k))}{\sqrt{(2\pi)^D|\Sigma_k|}}$, the probability density function (pdf) of multivariate Gaussian distribution with $\theta_k = \{\mu_k, \Sigma_k\}$, where μ_k stands for a D-dimensional cluster-specific mean vector and Σ_k stands for a D by D cluster-specific covariance matrix. In addition, $P(L_i|Z_i = k)$ are elements of the aforementioned concordance matrix C.

Scenario 2: CITE-seq data and scRNA-seq data

We denote $Y^{(1)}$ the latent space of RNA data (of size N_1 by D) from CITE-seq, and each element $Y^{(1)}_{ij}$ represents the value for feature j in cell i, where i runs from 1 to N_1, and j runs from 1 to D. We then denote $Y^{(2)}$ the latent space of RNA data (of size N_2 by D) from scRNA-seq, and each element $Y^{(2)}_{ij}$ represents the value for feature j in cell i, where i runs from $N_1 + 1$ to $N_1 + N_2$, and j runs from 1 to D. We assume $Y^{(1)}$ and $Y^{(2)}$ have batch effect corrected, and the features are exactly matched. Similarly, we also denote $Z^{(1)}_i$ and $Z^{(2)}_i$ the cell cluster label for cell i estimated from CITE-seq and scRNA-seq data, respectively. We further assume the common total number of clusters identified from RNA data is K. Similar to Scenario 1, the likelihood can be written as

$$\begin{aligned} P\left(Y^{(1)}, Y^{(2)}, L\right) &= P\left(L|Y^{(1)}\right) P\left(Y^{(1)}\right) P\left(Y^{(2)}\right) \\ &= \prod_{i=1}^{N_1} \left\{\sum_{k=1}^{K} P\left(L^{(1)}|Z_i^{(1)} = k\right) P\left(Z_i^{(1)} = k|Y_i^{(1)}\right)\right\} \prod_{i=1}^{N_1} \\ &\quad \times \left\{\prod_{k=1}^{K} f(Y_i^{(1)}|\theta_k)^{1\left(Z_i^{(1)}=k\right)}\right\} \prod_{i=N_1+1}^{N_1+N_2} \left\{\prod_{k=1}^{K} f(Y_i^{(2)}|\theta_k)^{1\left(Z_i^{(2)}=k\right)}\right\}, \end{aligned} \tag{3}$$

where $P(Z_i^{(1)} = k|Y_i^{(1)}) = \frac{\tau_k^{(1)} f(Y_i^{(1)}|\theta_k)}{\sum_{k\,=\,1}^{K} \tau_k^{(1)} f(Y_i^{(1)}|\theta_k)}$ refers to the posterior probability of cell i belonging to cluster k, and $\tau_k^{(1)} = P(Z_i^{(1)} = k)$ refers to the proportion of cluster k in RNA data from CITE-seq. Similarly, $\tau_k^{(2)} = P(Z_i^{(2)} = k)$ refers to the proportion of cluster k in RNA data from scRNA-seq. $f(Y_i^{(1)}|\theta_k)$ and $f(Y_i^{(2)}|\theta_k)$ refer to the cluster-specific distribution of RNA data from CITE-seq and scRNA-seq on latent space, respectively. Similar to Scenario 1, we assume

Table 4. An example of concordance matrix with M confident cell types identified with ADT data and K clusters identified with RNA data under ADT guidance.

		Clusters from RNA data						
		Cluster 1	Cluster 2	. . .	Cluster k	Cluster $k+1$	. . .	Cluster K
Confident cell types from ADT data	Cell type 1	p_{11}	0		0	0		0
	Cell type 2	0	p_{22}		0	0		0
	. . .							
	Cell type m	0	0		p_{mk}	$p_{m,\ k+1}$		0
	. . .							
	Cell type M	0	0		0	0		p_{MK}
	Uncertain $*$	$1-p_{11}$	$1-p_{22}$		$1-p_{mk}$	$1-p_{m,\ k+1}$		$1-p_{MK}$

Each entry in the matrix represents the conditional probability of a cell belong to a certain cell type given its cluster category.

$f(Y_i^{(1)}|\theta_k)$ and $f(Y_i^{(2)}|\theta_k)$ are the pdf of multivariate Gaussian distribution with $\theta_k = \{\mu_k, \Sigma_k\}$. For model flexibility, data-specific cluster proportions, $\tau_k^{(1)}$ and $\tau_k^{(2)}$, are allowed to differ. Note that the two data sources (after batch effect correction) share the common cluster-specific parameters. Again, $P(L_i|Z_i = k)$ are elements of the concordance matrix C. For prediction of confident cell types for scRNA-seq data, we can compute the posterior probability of cell i belonging to confident cell type m, $P(L_i^{(2)} = m|Y_i^{(2)}) = \sum_{k=1}^{K} P(L_i = m|Z_i^{(2)} = k)P(Z_i^{(2)} = k|Y_i^{(2)})$, where $P(Z_i^{(2)} = k|Y_i^{(2)}) = \frac{\tau_k^{(2)} f(Y_i^{(2)}|\theta_k)}{\sum_{k=1}^{K} \tau_k^{(2)} f(Y_i^{(2)}|\theta_k)}$ refers to the posterior probability of cell i belonging to cluster k for scRNA-seq data.

Modeling and space reduction of concordance matrix C

The concordance matrix C has two major functions: (1) to associate ADT data and RNA data by considering the ADT confident cell type label as guidance for cell clustering with RNA data; and (2) to provide the confident cell type annotation for each cluster. In general, a concordance matrix with M confident cell types and K clusters, as shown in Table 4 has $\binom{K-1}{M-1}$ different configurations of matrix form that fulfill the aforementioned constraints. Here, the notation $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}$ refers to the number of k-combinations of a set S with n elements. The number $\binom{K-1}{M-1}$ is derived from an analogy of listing all possible configurations for arranging K balls into M different boxes providing that each box has at least one ball. For example, when $K = 11$ and $M = 5$, the total number of different configurations is 210. In practice, one can either specify one or several plausible configurations with prior knowledge or to run our algorithm with all possible configurations in parallel, and then select the best configuration with the largest log-likelihood. Although the number $\binom{K-1}{M-1}$ can be large, a great number of matrix forms are actually not practical, (e.g. when one confident cell type is assumed to have six subtypes while the other five confident cell types each has only one subtype). As a result, the number can be largely reduced by restricting the maximum number of clusters a confident cell type corresponds to, denoted by K_{Sub}^{Max}. For example, when we set $K_{Sub}^{Max} = 3$, the total number of matrix forms is reduced to 45 from 210 for the situation when $K = 11$ and $M = 5$. K_{Sub}^{Max} can be selected based on prior biological knowledge, or information from a UMAP plot (e.g. Fig. 5A).

Optimization method

We use SGD method to directly optimize the log-likelihood (by minimizing the negative log-likelihood) of complete data, where the likelihood function for each scenario (1. CITE-seq data only; 2. CITE-seq data and scRNA-seq data) is defined above. The parameters to be estimated through SGD include cluster-specific parameters $\{\mu_k, \Sigma_k\}$ from the clustering part, and the nonzero p'_{mk}s in the concordance matrix C. As described in previous section, each configuration of the matrix form of C is also a parameter in the likelihood function, which can be maximized through parallel computing (each thread with a different configuration).

Initialization of SECANT and selection of the number of clusters

For the initialization of clustering-related parameters $\{\tau_k,\ \mu_k, \Sigma_k\}$, we exclude cells with uncertain ADT label and run separate multivariate GMM with cells from each ADT confident cell type. The number of mixtures for each multivariate GMM is determined based on the concordance matrix C. For probability parameters in C, by default we set $p_{mk} = 0.5$ as the initial value. In general, SECANT is robust to different initializations (Table S10). For the number of clusters to be specified in SECANT, we suggest users choose this value based on prior biological knowledge, especially when the tissue type is well-studied (e.g. PBMC) or the study is confirmatory. On the other hand, since SECANT is a likelihood-based method, people can utilize the gradient of Bayesian information criterion (BIC) scores curve to help decide the number of clusters as a data-driven approach.

Determining the best configuration of concordance matrix in this study

To determine the best configuration of concordance matrix in both simulation study and real data applications, we first search the entire configuration space in each scenario (in parallel) and run each configuration with 10 different initializations, and then determine the best configuration based on largest log-likelihood. Under each setting, we summarize the boxplot of log-likelihood from 10 initializations across all configurations (Figs. S16 to S22). In most settings, all the top 10 log-likelihoods are from the correct configuration, which infers running SECANT with one initialization is sufficient. The only exceptions occur in simulation study, when sample size decreases or the number of parameters to estimate increases (e.g. increasing number of clusters or feature dimension), which can be remedied by an increasing sample size (e.g. from N = 1000 to N = 2000).

In practice, there are many situations where people can directly specify the configuration without searching for the entire space. For instance, if the study is confirmatory (the cell type compositions are prespecified) or the tissue type is well-studied (e.g. PBMC). On the other hand, for complex tissues like bone marrow, one may consider running SECANT with multiple initializations (e.g. five) for each configuration if the number of cells is low, and then select the one with the largest log-likelihood among all combinations. However, to avoid a large number of configurations to search, which will largely increase computational burden, we generally suggest that users specify the total number of clusters less than 15.

Evaluation metrics for clustering performance

We assessed the performance of ADT-guided clustering of SECANT by computing ARI (36) and AMI (37) with the simulated clustering truth. Both ARI and AMI are commonly used metrics for the concordance of two clustering results. Comparing to the truth, an ARI or AMI of value 1 indicates the clustering result is identical to the truth, while value 0 indicates the clustering result is a random assignment. A previous study suggests using ARI for balanced clustering situation, while using AMI for unbalanced clustering situation (43).

In-house CITE-seq datasets

We generated two in-house CITE-seq datasets of PBMCs from a healthy donor. Cells from both datasets are from the same aliquot of the sample, and thus are homogenous in nature. Cells were stained with the newly released TotalSeq-A panel (Human Universal Cocktail, V1.0) with a total of 154 unique cell surface antigens from BioLegend and are prepared using the 10x Genomics platform with Gel Bead Kit V2. Different antibody concentrations were used in each sample. In addition, we used cell hashing for sample multiplexing to eliminate batch effect between RNA data from each dataset. The prepared assay is subsequently sequenced on an Illumina HiSeq with a depth of 50 K reads per cell.

Competing methods for comparisons

Seurat's weighted-nearest neighbor (WNN) is a novel analytical framework to integrate single-cell multi-omics data to jointly define cellular state (22). Seurat (WNN) utilizes an unsupervised approach (specifically through constructing k-nearest neighbor graphs for each modality and performing within and cross-modality predictions) to estimate cell-specific modality "weights," which reflects the relative importance of each data modality for each cell in downstream analyses. For Seurat (WNN), the raw RNA and ADT count data matrices were used as input. In general, data preprocessing followed general suggestion in the Seurat tutorial (https://satijalab.org/seurat/articles/weighted_nearest_neighbor_analysis.html). The number of important principal components (PCs) after dimension reduction was determined based on Elbow plot. The resolution parameter in clustering algorithm was adjusted according to the prespecified number of clusters. All the other parameters (such as the number of multimodal neighbors to compute) were set to default settings.

totalVI is a deep generative model that can jointly analyze paired protein and RNA data in CITE-seq (20). Essentially, it utilizes a probabilistic latent variable model to learn a joint probabilistic representation of the observed paired data, which accounts for the technical biases and noise from each data modality. For totalVI, the raw RNA and ADT count data matrices were used as input. The dataset was first filtered and the top 4,000 highly variable genes were selected, as suggested in the totalVI tutorial (https://docs.scvi-tools.org/en/0.6.5/tutorials/totalvi.html). The resolution parameter in clustering algorithm was adjusted according to the prespecified number of clusters. All the other parameters (such as hyperparameters, learning rate, and number of epochs) were set to default settings.

Acknowledgments

This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. Specifically, this work used the HTC cluster, which is supported by NIH award number S10OD028483, and the H2P cluster, which is supported by NSF award number OAC-2117681.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

This study was funded by the National Institutes of Health grants [P01AI106684 to W.C., U01DK062420 to W.C. and R.D., P50AR060780 to R.L. and W.C., and R01HL137709 to K.C.]. This study was also supported in part by a National Cancer Institute grant [P30 CA008748 to X.W.] and the Children's Hospital of Pittsburgh of the UPMC Health System.

Authors' Contributions

X.W., R.D., and W.C. designed research; X.W., Z.X., H. Hu, and X.Z. performed research; Y.Z. and H. Huang provided knowledge for speeding-up algorithm; R.L., K.C., and R.D. conducted biological experiments and provided guidance from biological perspective; and X.W., Y.D., and W.C. wrote the paper.

Software Availability

Our algorithm is implemented in Python based on PyTorch, and is available on GitHub at https://github.com/tarot0410/SECANT.

Data Availability

The two in-house human PBMC CITE-seq datasets are available on GEO under GSE168264. The public 10x10k_PBMC and 10x5k_PBMC CITE-seq datasets can be downloaded from 10× Genomic website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_protein_v3 and https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.2/5k_pbmc_v3_nextgem). The public human bone marrow CITE-seq dataset is available in Seurat package ("bmcite"). The public human upper lobe lung CITE-seq dataset is available on GEO under GSE128169 (sample SC277).

References

1. Gawad C, Koh W, Quake SR. 2016. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 17: 175–188.
2. Grun D, *et al.* 2015. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 525: 251–255.
3. Treutlein B, *et al.* 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 509: 371–375.

4. Tsoucas D, Yuan GC. 2017. Recent progress in single-cell cancer genomics. Curr Opin Genet Dev. 42: 22–32.
5. Yuan GC, *et al.* 2017. Challenges and emerging directions in single-cell analysis. Genome Biol. 18: 84.
6. Zheng GX, *et al.* 2017. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 8: 1–12.
7. Stoeckius M, *et al.* 2017. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 14: 865.
8. Peterson VM, *et al.* 2017. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 35: 936–939.
9. Stoeckius M, *et al.* 2018. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 19: 1–12.
10. Buenrostro JD, *et al.* 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 523: 486–490.
11. Cusanovich DA, *et al.* 2015. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 348: 910–914.
12. Mimitou EP, *et al.* 2021. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat Biotechnol. 39: 1246–1258.
13. Ji Z, Ji H. 2016. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res. 44: e117–e117.
14. Kiselev VY, *et al.* 2017. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 14: 483–486.
15. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. Nat Methods. 15: 1053–1058.
16. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 33: 495–502.
17. Sun Z, *et al.* 2019. A Bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. Nat Commun. 10: 1–10.
18. Sun Z, *et al.* 2018. DIMM-SC: a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. Bioinformatics. 34: 139–146.
19. Wang B, *et al.* 2018. SIMLR: a tool for large-scale genomic analyses by multi-kernel learning. Proteomics. 18: 1700232.
20. Gayoso A, *et al.* 2021. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat Methods. 18: 272–282.
21. Zhou Z, Ye C, Wang J, Zhang NR. 2020. Surface protein imputation from single cell transcriptomes by deep neural networks. Nat Commun. 11: 1–10.
22. Hao Y, *et al.* 2021. Integrated analysis of multimodal single-cell data. Cell. 184, 3573–3587.e29,
23. Wang X, *et al.* 2020. BREM-SC: a Bayesian random effects mixture model for joint clustering single cell multi-omics data. Nucleic Acids Res. 48: 5814–5824.
24. Aghaeepour N, Nikolic R, Hoos HH, Brinkman RR. 2011. Rapid cell population identification in flow cytometry data. Cytometry A. 79A: 6–13.
25. Lian Q, *et al.* 2020. Artificial-cell-type aware cell-type classification in CITE-seq. Bioinformatics. 36: i542–i550.
26. Maecker HT, McCoy JP, Nussenblatt R. 2012. Standardizing immunophenotyping for the human immunology project. Nat Rev Immunol. 12: 191–200.
27. Qian Y, *et al.* 2010. 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. 78B: S69–S82.
28. Verschoor CP, Lelic A, Bramson JL, Bowdish DM. 2015. An introduction to automated flow cytometry gating tools and their implementation. Front Immunol. 6: 380.
29. Chen G, *et al.* 2002. Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics. 1: 304–313.
30. Haider S, Pal R. 2013. Integrated analysis of transcriptomic and proteomic data. Curr Genomics. 14: 91–110.
31. Tian T, Zhang J, Lin X, Wei Z, Hakonarson H. 2021. Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data. Nat Commun. 12: 1–12.
32. Li J, Zhang Y, Yang C, Rong R. 2020. Discrepant mRNA and protein expression in immune cells. Curr Genomics. 21: 560–563.
33. Bouveyron C, Celeux G, Murphy TB, Raftery AE. 2019. Model-based clustering and classification for data science: with applications in R, Vol. 50. Cambridge University Press.
34. Stuart T, *et al.* 2019. Comprehensive integration of single-cell data. Cell. 177: 1888–1902.e21.
35. McInnes L, Healy J, Melville J. 2018. Umap: uniform manifold approximation and projection for dimension reduction. arXiv:180203426.
36. Rand WM. 1971. Objective criteria for the evaluation of clustering methods. J Am Statist Assoc. 66: 846–850.
37. Nguyen XV, Epps J, Bailey J. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009; Montreal, Quebec, Canada.
38. Bouveyron C, Girard S. 2009. Robust supervised classification with mixture models: learning from data with uncertain labels. Pattern Recognit. 42: 2649–2658.
39. Nguyen HH, *et al.* 2016. Naïve CD8+ T cell derived tumor-specific cytotoxic effectors as a potential remedy for overcoming TGF-β immunosuppression in the tumor microenvironment. Sci Rep. 6: 1–10.
40. Valenzi E, *et al.* 2019. Single-cell analysis reveals fibroblast heterogeneity and myofibroblasts in systemic sclerosis-associated interstitial lung disease. Ann Rheum Dis. 78: 1379–1387.
41. Griffiths JA, Richard AC, Bach K, Lun AT, Marioni JC. 2018. Detection and removal of barcode swapping in single-cell RNA-seq data. Nat Commun. 9: 1–6.
42. Lun AT, Riesenfeld S, Andrews T, Gomes T, Marioni JC. 2019. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. 20: 1–9.
43. Romano S, Vinh NX, Bailey J, Verspoor K. 2016. Adjusting for chance clustering comparison measures. J Mach Learn Res. 17: 4635–4666.