Briefings in Bioinformatics, 2020, pp. 1–11

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Cell-type annotation with accurate unseen cell-type identification using multiple references

Yi-Xuan Xiong,[1,2,†] Meng-Guo Wang,[1,2,†] Luonan Chen[3,4,5,6,*] and Xiao-Fei Zhang[1,2,*]

[1]School of Mathematics and Statistics, Central China Normal University, Wuhan, 430079, China, [2]Hubei Key Laboratory of Mathematical Sciences, Central China Normal University, Wuhan, 430079, China, [3]State Key Laboratory of Cell Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai, 200031, China, [4]School of Life Science and Technology, ShanghaiTech University, Shanghai, 201210, China, [5]Key Laboratory of Systems Health Science of Zhejiang Province, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou, 310024, China and [6]Guangdong Institute of Intelligence Science and Technology, Hengqin, Zhuhai, Guangdong, 519031, China
[*]To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Automated cell-type annotation using a well-annotated single-cell RNA-sequencing (scRNA-seq) reference relies on the diversity of cell types in the reference. However, for technical and biological reasons, new query data of interest may contain unseen cell types that are missing from the reference. When annotating new query data, identifying unseen cell types is fundamental not only to improve annotation accuracy but also to new biological discoveries. Here, we propose mtANN (**m**ul**t**iple-reference-based scRNA-seq data **ann**otation), a new method to automatically annotate query data while accurately identifying unseen cell types with the help of multiple references. Key innovations of mtANN include the integration of deep learning and ensemble learning to improve prediction accuracy, and the introduction of a new metric defined from three complementary aspects to identify unseen cell types. We demonstrate the advantages of mtANN over state-of-the-art methods for cell-type annotation and unseen cell-type identification on two benchmark dataset collections, as well as its predictive power on a collection of COVID-19 datasets.

Key words: single-cell RNA sequencing, cell-type annotation, unseen cell-type identification, ensemble learning

Introduction

Single-cell RNA sequencing (scRNA-seq) techniques outline the expression profile of a sample in single-cell resolution. Their recent advances have stimulated efforts to identify and characterize the cellular composition of tissues, revolutionizing the understanding of the heterogeneity of complex tissues. With the various sequencing technologies, like 10x Genomics Chromium, Drop-seq, and Smart-seq2, having emerged, understanding the complex tissues has turned into a problem of cell-type annotation for new sequencing data [1].

There are two typical solutions for the cell-type annotation tasks. One of the solutions is to unsupervised cluster cells into groups based on the similarity of their gene expression profiles, and annotate cell populations by assigning labels to each cluster according to cluster-specific marker genes [2, 3]. However, the appropriate marker genes selection needs to conduct extensive literature reviews and manually test various combinations of marker genes, which is not only time-consuming but also unrepeatable across different experiments within and across research groups [4]. Another strategy annotates new scRNA-seq data with well-annotated data as a reference atlas [5, 6, 7]. These methods make predictions on query datasets by training classifiers on the reference atlas [8, 9, 10], or transfer cell type labels based on the similarity between the reference atlas and query dataset [11, 12]. The reference-based methods can alleviate the problems involved with clustering-based methods. However, previous reference-based methods barely consider the new issue that the query data may contain cell types not included in the reference.

In reality, limited by various technical and biological factors, it is often difficult to collect all cell types in the reference that

may be present in the query data. For example, in scRNA-seq, cells that are not easily dissociated in certain tissues are easily lost, and some sensitive cells may be impaired by excessive dissociation [13]. These issues can be prevented in single-nucleus RNA sequencing (snRNA-seq) because snRNA-seq only requires the isolation of single nuclei [14]. When the reference is from scRNA-seq, there may be unseen cell types in the query data from snRNA-seq. Another example is that the reference is from healthy samples (e.g., the Human Cell Atlas Project [15]), but we need to annotate query data from disease (e.g., tumor) samples. In this context, identifying unseen cell types in query data may lead to new biological discoveries. Therefore, reference-based methods need to simultaneously consider two key factors: (i) accurately distinguishing cells belonging to unseen cell types from cells belonging to known cell types; (ii) annotating cells belonging to known cell types with the correct type.

In order to handle the above two tasks, we propose mtANN (**m**ul**t**iple-reference-based scRNA-seq data **ann**otation), a novel method that automatically identifies unseen cell types while accurately annotating query datasets by integrating multiple well-annotated scRNA-seq datasets as references. The main idea of mtANN is first to learn multiple deep classification models from multiple reference datasets, and the multiple prediction results are used to calculate the metric for unseen cell-type identification and to vote for the final annotation. mtANN has the following characteristics: (i) it integrates multiple reference datasets to enrich cell types in the reference atlas to alleviate the unseen cell-type problem; (ii) it combines the ideas of deep learning and ensemble learning to improve prediction accuracy; (iii) it proposes a new metric from three complementary aspects to measure whether a cell belongs to an unseen cell-type; and (iv) it introduces a new data-driven approach to automatically determining the threshold for unseen cell-type identification. We benchmark mtANN using two collections of benchmark datasets, each from different tissues, sequencing technologies, and containing different cell types. We prepare a total of 75 benchmark tests, including annotations across different technologies when different cell types are the unseen cell types. We also use a COVID-19 dataset and prepare a total of 249 tests to assess the performance of mtANN. Experimental results demonstrate that mtANN outperforms state-of-the-art methods in both unseen cell-type identification and cell-type annotation.

Methods

Notations and problem statement

For convenience, we first introduce some notations (Supplementary Table S1). We assume that M well-labeled reference datasets with the same tissue type as the query dataset are collected. Let $\{(X^{r,i}, Y^{r,i})\}_{i=1}^{M}$ denote the references, where $X^{r,i}$ is an $n^{r,i} \times p^{r,i}$ matrix that denotes the gene expression matrix after library size normalization of the i-th reference dataset with rows representing cells and columns representing genes, and $Y^{r,i}$ denotes the corresponding cell type labels. The number of cells and genes of the i-th reference dataset are denoted by $n^{r,i}$ and $p^{r,i}$ separately. Let $K^{r,i}$ denote the set of cell types observed in $Y^{r,i}$ and $m^{r,i} = |K^{r,i}|$ denote the cardinality of $K^{r,i}$, i.e., the number of cell types observed in $Y^{r,i}$. Let $K = union\left(\{K^{r,i}\}_{i=1}^{M}\right)$ denote all cell types present in all reference datasets. Let X^q be an $n^q \times p^q$ matrix that denotes the gene expression matrix after library size normalization of the query dataset. The number of cells and genes of the query dataset are denoted by n^q and p^q separately. Let Y^q denote the corresponding cell type labels which is unknown. Let $\mathbf{1}_{[\cdot]}$ denote the indicative function, i.e., $\mathbf{1}_{[\cdot]} = 1$ when $[\cdot]$ is true, otherwise, it is equal to 0. Let $\|\cdot\|_F$ denote the Frobenius norm of a matrix.

In this study, we focus on annotating cells in a new query dataset with multiple well-labeled references. Mathematically, our goal is to estimate Y^q based on observed data, $\{(X^{r,i}, Y^{r,i})\}_{i=1}^{M}$ and X^q. Note that the cell type labels of these reference datasets should be provided with consistent terminology. In addition, although multiple reference datasets are integrated, there may still be cell types in the query dataset that are not observed in any reference dataset. We call such cell types "unseen cell types". It is necessary to identify cells belonging to unseen cell types to avoid misclassification. To achieve our goal, we propose a novel multiple-reference-based scRNA-seq data annotation method (Figure 1, Supplementary section 1). Our method consists of four steps. First, we adopt eight gene selection methods to generate a series of subsets that retain distinct genes for each reference dataset. This step facilitates the detection of biologically important genes and increases data diversity for effective ensemble learning. Second, we train a series of neural network-based deep classification models based on all subsets of all reference datasets. Third, we annotate the query dataset through integrating the results output by all base classification models. Finally, we identify cells that may belong to unseen cell types and mark them as "unassigned".

Gene selection

For each reference dataset $(X^{r,i}, Y^{r,i})$, we adopt eight gene selection methods to pick genes from different perspectives, including Limma, Bartlett's test, Kolmogorov-Smirnov test, Chi-squared test, Bimodality index, Gini index, Dispersion and Variance-stabilizing transformation. Details can be found in Supplementary section 2. We index these gene selection methods using $j = 1, \ldots, 8$. Let $G^{r,ij}$ denote the genes selected by the j-th gene selection method for the i-th reference dataset and G^q denote all genes in the query dataset. Based on the j-th gene selection method, we construct the j-th subset of $X^{r,i}$ and the corresponding (i,j)-th subset of X^q as follows:

$$X^{r,ij} = X^{r,i}[:, G^{r,ij} \cap G^q], \quad X^{q,ij} = X^q[:, G^{r,ij} \cap G^q]. \quad (1)$$

After obtaining $X^{r,ij}$ and $X^{q,ij}$, logarithmic transformation, z-score standardization, and Min-Max scaling are applied to preprocess them. For convenience, we still denote the preprocessed results by $X^{r,ij}$ and $X^{q,ij}$. Based on $X^{r,ij}$, $Y^{r,i}$ and $X^{q,ij}$, we construct a dataset pair $(X^{r,ij}, Y^{r,i}, X^{q,ij})$, as the training set for the next step to train a base classification model.

Deep classification model training

Based on each dataset pair $(X^{r,ij}, Y^{r,i}, X^{q,ij})$, we train a classification model based on deep learning. The classification model involves two components: the embedding component for extracting cell type-related features and the linear classifier layer for distinguishing cell types. Let E^{ij} and C^{ij} denote the embedding component and the linear classifier layer separately. We take the reference subset as input, and the forward propagation result of the classification model after softmax transformation can be defined as $\hat{P}^{r,ij} =$

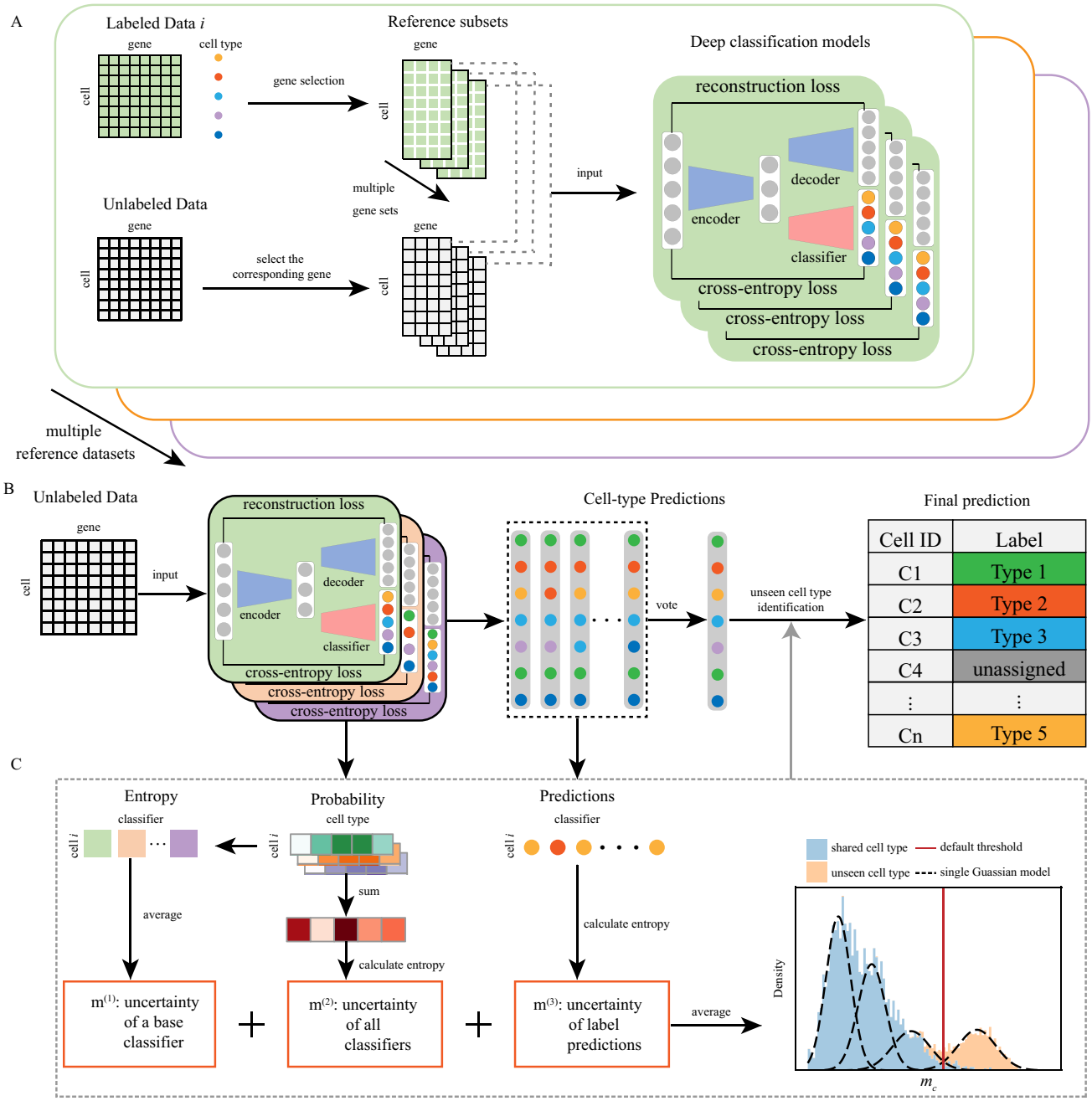

Fig. 1. Overview of mtANN. **(A)** The training process of mtANN using multiple datasets as reference datasets. The labeled data i is used as an example. Eight gene selection methods are applied on data i, obtaining multiple reference subsets. The gene sets selected by the eight gene selection methods intersect with all the genes in the query dataset, determining the input genes of multiple deep classification models. We conduct the same steps for every labeled data, thus getting multiple deep classification models. **(B)** The testing process of mtANN on the query dataset. The multiple deep classification models obtained on each reference subset make predictions on the query dataset. We perform majority voting and unseen cell type identification to obtain the final prediction result. **(C)** Unseen cell identification process. Cell i is used as an example. The unseen cell type identification metric is defined by averaging three entropy indexes which are calculated from the prediction probability of multiple base classification models and vote probability. The Gaussian mixture model is applied to the metric to select the threshold.

$softmax\left(C^{ij}\left(E^{ij}\left(X^{r,ij}\right)\right)\right)$, where $\hat{P}^{r,ij}$ is an $n^{r,i} \times m^{r,i}$ matrix and the (c,k)-th element of $\hat{P}^{r,ij}$ can be regarded as the predicted probability of cell c in the reference subset belonging to cell type k. The cross-entropy loss is used to train the classification model, and the loss function can be formulated as

$$\mathcal{L}_{ce} = -\frac{1}{n^{r,i}} \sum_{c=1}^{n^{r,i}} \sum_{k \in K^{r,i}} \mathbf{1}_{[k=Y_c^{r,i}]} \log \hat{P}_{ck}^{r,ij}. \quad (2)$$

In order to enable the embedding component E^{ij} to fully capture the characteristics of cells and make the classification model better fit the query dataset, we employ

the embedding component as an encoder and use a mirror image of the embedding component as a decoder to construct an autoencoder [16, 17]. The reconstruction loss of cells both from the reference subset and the query subset is taken into consideration when training the classification model. Let D^{ij} denote the decoder component. The forward propagation results of the autoencoder can be defined as $\hat{X}^{r,ij} = D^{ij}\left(E^{ij}\left(X^{r,ij}\right)\right)$ and $\hat{X}^{q,ij} = D^{ij}\left(E^{ij}\left(X^{q,ij}\right)\right)$, where $\hat{X}^{r,ij}$ and $\hat{X}^{q,ij}$ denote the reconstruction of $X^{r,ij}$ and $X^{q,ij}$ separately. The reconstruction loss is measured by the mean squared error, and the loss function can be formulated as

$$\mathcal{L}_{re} = \frac{1}{n^{r,i}p^{ij}}\left\|\hat{X}^{r,ij} - X^{r,ij}\right\|_F^2 + \frac{1}{n^{q}p^{ij}}\left\|\hat{X}^{q,ij} - X^{q,ij}\right\|_F^2, \tag{3}$$

where p^{ij} represents the number of genes in this dataset pair.

Therefore, the final optimization problem for training the classification model for dataset pair $\left(X^{r,ij}, Y^{r,i}, X^{q,ij}\right)$ can be written as

$$\min_{E^{ij},D^{ij},C^{ij}} \quad \mathcal{L}_{ce} + \lambda\mathcal{L}_{re}, \tag{4}$$

where λ is the tuning parameter and the default value is 1. Details of the neural network architecture, hyperparameter settings, and initialization can be found in Supplementary section 3. Let $\left\{\left(\hat{E}^{ij}, \hat{C}^{ij}\right)\right\}_{i=1,\ldots,M,j=1,\ldots,8}$ denote all trained base classification models.

Query dataset annotation

Based on one base classification model $\left(\hat{E}^{ij}, \hat{C}^{ij}\right)$, we take the corresponding query subset $X^{q,ij}$ as input. The forward propagation result along the model after softmax transformation can be formulated as $\hat{P}^{q,ij} = softmax\left(\hat{C}^{ij}\left(\hat{E}^{ij}\left(X^{q,ij}\right)\right)\right)$. The (c,k)-th element of $\hat{P}^{q,ij}$ can be regarded as the predicted probability of cell c in the query dataset belonging to cell type k. For each cell in the query dataset, we assign the cell type label with the highest probability to it according to $\hat{P}^{q,ij}$. Let $\hat{Y}^{q,ij}$ and $\tilde{P}^{q,ij}$ denote the predicted specific cell type labels and the corresponding probabilities, separately. For cell c,

$$\hat{Y}_c^{q,ij} = \underset{k\in K^{r,i}}{\arg\max}\, \hat{P}_{ck}^{q,ij}, \quad \tilde{P}_c^{q,ij} = \max_{k\in K^{r,i}} \hat{P}_{ck}^{q,ij}. \tag{5}$$

Based on all base classification models $\left\{\left(\hat{E}^{ij}, \hat{C}^{ij}\right)\right\}$, we obtain a series of predictions $\left\{\left(\hat{Y}^{q,ij}, \tilde{P}^{q,ij}\right)\right\}$ of the query dataset. Then, we integrate all these predictions for consensus annotation, denoted by $\hat{Y}^q$. For cell c, we calculate

$$\hat{Y}_c^q = \underset{k\in K}{\arg\max} \frac{\sum_{i=1}^{M}\sum_{j=1}^{8}\mathbf{1}_{[k=\hat{Y}_c^{q,ij}]}}{\sum_{i=1}^{M}\sum_{j=1}^{8}\mathbf{1}_{[k\in K^{r,i}]}}, \tag{6}$$

where the numerator represents the number of times that cell c is predicted to belong to cell type k across all predictions and the denominator represents the number of dataset pairs containing cell type k. The role of the denominator is to handle the situation where a cell is predicted to belong to a cell type only a small number of times, even if the cell does belong to that cell type since the cell type is only present in part of the dataset pairs. In addition, if the maximum value corresponds to more than one cell type, we determine the final label according to the maximum sum of their corresponding probabilities obtained from $\left\{\tilde{P}^{q,ij}\right\}$.

Uncertain cell identification

Since there is no training data in the reference datasets for the unseen cell types, the predictions for the cells belonging to these cell types can be more uncertain. The uncertainty can be detected from two perspectives based on all predicted results of the classification models, including the intra-model perspective and inter-model perspective. For the former, no one cell type dominates the probability when making predictions based on a single classification model. For the latter, there is a large inconsistency among the predictions obtained by different classification models. Therefore, we design three entropy-based measures, denoted by $m^{(1)}$, $m^{(2)}$ and $m^{(3)}$, to quantitatively characterize the uncertainty, where $m^{(1)}$ is from the intra-model perspective, and $m^{(2)}$ and $m^{(3)}$ are from the inter-model perspective.

The first measure $m^{(1)}$ characterizes uncertainty from the intra-model perspective by calculating the entropy of the probabilities of belonging to different cell types within each classification model and then averaging all entropy values for each cell. Higher entropy indicates higher uncertainty in prediction. For cell c,

$$m_c^{(1)} = \frac{1}{8M}\sum_{i,j} H\left(\hat{P}_c^{q,ij}\right), \tag{7}$$

where $H(\cdot)$ represents entropy and is defined as $H\left(\hat{P}_c^{q,ij}\right) = -\sum_{k\in K^{r,i}} \hat{P}_{ck}^{q,ij}\log_2\left(\hat{P}_{ck}^{q,ij}\right)$.

The second measure $m^{(2)}$ characterizes uncertainty from the inter-model perspective by first integrating the probabilities of all classification models and then calculating the entropy. Let $Q^{(2)}$ be an $n^q \times K$ matrix that denotes the integrated result. The (c,k)-th element of $Q^{(2)}$ is defined as

$$Q_{ck}^{(2)} = \frac{\sum_{i,j}\mathbf{1}_{[k\in K^{r,i}]}\hat{P}_{ck}^{q,ij}}{\sum_{i,j}\mathbf{1}_{[k\in K^{r,i}]}}. \tag{8}$$

The value of $Q_{ck}^{(2)}$ represents the average of the predicted probabilities that cell c is assigned to cell type k across all classification models. Then, $Q^{(2)}$ is transformed into a probability matrix $\tilde{Q}^{(2)}$ by dividing each value by the row sum. If the predictions of different classification models for cell c are inconsistent, no one cell type dominates the probability in $\tilde{Q}_c^{(2)}$. $m^{(2)}$ is the entropy of each cell calculated on the basis of the probability matrix to characterize the inconsistency. For cell c,

$$m_c^{(2)} = H\left(\tilde{Q}_c^{(2)}\right). \tag{9}$$

The larger $m_c^{(2)}$ is, the more inconsistent the predictions are, and thus the more uncertain the prediction of cell c is.

The third measure $m^{(3)}$ is similar to $m^{(2)}$. The difference is that it integrates the predicted specific cell type labels of all classification models, just as it does when estimating $\hat{Y}^q$. Let $Q^{(3)}$ be an $n^q \times K$ matrix that denotes the integrated result for this measure. The (c,k)-th element of $Q^{(3)}$ is defined as

$$Q_{ck}^{(3)} = \frac{\sum_{i,j}\mathbf{1}_{[k=\hat{Y}_c^{q,ij}]}}{\sum_{i,j}\mathbf{1}_{[k\in K^{r,i}]}}. \tag{10}$$

Then, as before, we transform $Q^{(3)}$ into a probability matrix $\tilde{Q}^{(3)}$. If the predicted specific cell type labels for cell c are inconsistent, no one cell type dominates the probability in $\tilde{Q}_c^{(3)}$

as in $\tilde{Q}_c^{(2)}$. We calculate the entropy to get $m^{(3)}$, i.e., for cell c,

$$m_c^{(3)} = H\left(\tilde{Q}_c^{(3)}\right). \tag{11}$$

After obtaining $m^{(1)}$, $m^{(2)}$ and $m^{(3)}$, the values are scaled to $[0, 1]$ linearly through Min-Max scaling within each measure, and the results are denoted by $\bar{m}^{(1)}$, $\bar{m}^{(2)}$ and $\bar{m}^{(3)}$ separately. The ensemble uncertainty measure m is defined as the mean of these three measures for each cell, i.e., for cell c,

$$m_c = \frac{\bar{m}_c^{(1)} + \bar{m}_c^{(2)} + \bar{m}_c^{(3)}}{3}. \tag{12}$$

A larger value of m_c indicates a higher probability that cell c belongs to an unseen cell type. However, how determining the threshold to distinguish cells belonging to unseen cell types from all cells remains challenging. To this end, we provide a new method for automatically recognizing the cells with higher uncertainty than others. The identification method is based on Gaussian mixture models, which can be written as

$$p(m) = \sum_{s=1}^{S} \pi_s N(\mu_s, \sigma_s^2), \tag{13}$$

where μ_s, σ_s^2 represent the mean and variance of the s-th component and π_s is the weight of the s-th component. We determine the number of mixture components S by trying different numbers from 1 to 5 and selecting the number according to the Akaike information criterion (AIC). If the most suitable value of S determined by AIC is 1, we consider that there is no cell belonging to unseen cell types. Otherwise, all the cells are divided into different groups according to the ensemble uncertainty measure m through the Gaussian mixture model, and then the uncertain groups are distinguished based on the mean of the ensemble uncertainty of cells within each group. If there are groups with a mean greater than or equal to 0.6, these groups are considered to be uncertain groups. Otherwise, the group with the largest mean is considered to be the uncertain group. All the cells in the uncertain groups are annotated as "unassigned".

Experiments and results

Benchmark mtANN for unseen cell-type identification

To demonstrate the ability of mtANN in identifying unseen cell types, we use two collections of datasets from two tissues: peripheral blood mononuclear cells (PBMCs) collection [18] and pancreas collection [19, 20, 21, 22] (Supplementary section 4 and Supplementary Tables S3-4). In each collection, each dataset is used as a query dataset alternatively and the remaining datasets are reference datasets. We remove one cell type shared by all the reference datasets and the query dataset for one test experiment (for details, please see Supplementary Figure S1, Tables S5-S6). We compare mtANN with existing popular methods, including scmap-clust, scmap-cell, Seurat v3, ItClust, scGCN (entropy), scGCN (enrichment), and scANVI (Supplementary section 5) in terms of the auprc scores (Supplementary section 6).

The results are presented in Figure 2 and Supplementary Figure S2. The boxplots show that mtANN is superior to the competing methods on all the datasets in PBMCs collection (Figure 2A). mtANN exceeds the competing methods when "Baron human", "Segerstolpe" and "Xin" are used as the query datasets in pancreas collection (Supplementary Figure S2A). We also count the number of times each method ranks first in auprc scores across all tests, and we find that the number of mtANN ranking first is much higher than the competing methods in both PBMCs and pancreas collections (Supplymentary Figures S2B-C).

To further investigate which unseen cell type is better identified, we compare the auprc scores when each cell type is missing from references on PBMCs collection. mtANN outperforms all the competing methods when B cell, CD14+ monocyte, Megakaryocyte, CD4+ T cell, and Plasmacytoid dendritic cell are treated as unseen cell types in the query data (Supplementary Figure S3). When the unseen cell type is similar to a known cell type, it is difficult to identify cells belonging to the truly unseen cell type as "unassigned". For example, when we remove B cells in all the references, the competing methods confuse unseen cell types and shared cell types while mtANN can clearly distinguish these two types of cells (Figure 2B and Supplementary Figure S4). We also remove CD14+ monocyte and Megakaryocyte from all the references, and the distribution of metrics calculated from mtANN also validates its superiority (Supplementary Figures S5-S6).

To validate the performance of the threshold selection, mtANN is compared with Seurat v3, scmap-clust, and scmap-cell in terms of F1 score (Supplementary section 6). The results are displayed in Figure 3 and Supplementary Figure S7. The barplots represent that mtANN achieves the best performance when "Celseq", "inDrop", "Smart-seq2", "10X v2", and "10X v3" are used as the query datasets (Figure 3A), and mtANN outperforms other methods when "Baron human" and "Xin" datasets are the query datasets (Supplementary Figure S7). We also find that the number of mtANN ranking first has large margins than other methods in PBMCs and pancreas collections (Figures 3B-C). These results demonstrate that, compared to scmap using a fixed threshold, and Seurat v3 selecting the 20-th percentile of scores as the threshold (the proportion of unseen cell types is substantially fixed), the data-driven approach proposed by mtANN is more flexible in choosing an appropriate threshold.

Benchmark mtANN for cell-type annotation when there are unseen cell types in the query dataset

To evaluate the performance of mtANN to annotate unlabeled sequencing data when there are unseen cell types, we also use the PBMCs and pancreas collections. Based on previous benchmarks, We compared the accuracy of mtANN with other methods on the entire dataset (Supplementary section 6). As threshold selection will affect the entire annotation accuracy of the query dataset, we conduct two ways to determine the thresholds: the actual proportion of unseen cells and the default threshold provided by each method.

When using the actual proportion of unseen cell types in the query dataset as the threshold to assign unseen cells, the performances of mtANN and other methods are presented in Figure 4 and Supplementary Figure S8. In Figure 4A, we find that mtANN outperforms other methods in all the datasets from PBMCs collection, and mtANN has the best performance when "Baron human", "Segerstolpe" and "Xin" in the pancreas collection are used as the query datasets (Supplementary Figure S8A). We count the number of times each method ranks first in accuracy across all tests, and the results show that the number of ranking first of mtANN is much higher than other methods

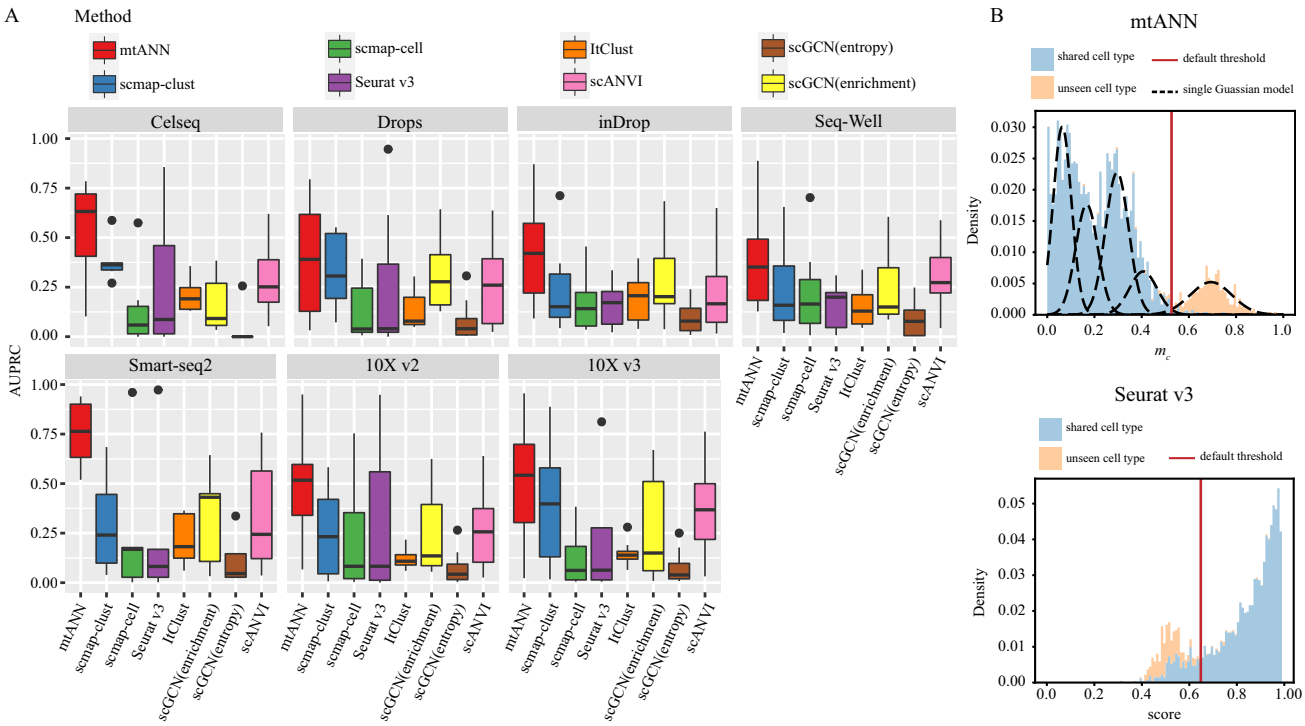

Fig. 2. Performances of mtANN and other compared methods in identifying unseen cell types. **(A)** Boxplots of the auprc scores of different methods on the PBMCs collection. The results with different query datasets are displayed in different panels. **(B)** The distributions of mtANN's metric of uncertainty and the score of Seurat v3. The test when "10X v3" is the query dataset and the B cell is the unseen cell type is taken as an example. The color of the histogram distinguishes unseen cell types from shared cell types. The black dotted line represents the subpopulations of the Gaussian mixture model fitted by mtANN. The red solid line represents the default threshold selected by each method. Cells with a metric greater than the threshold are identified as "unassigned" in mtANN and cells with a score less than the threshold are identified as "unassigned" in Seurat v3.

in PBMCs and pancreas collections (Supplementary Figures S8B-C).

To further illustrate the specific annotation of each cell type when there are unseen cell types, we use the heatmap of the confusion matrix between the real cell type labels and the predicted cell type labels. We obtain the hierarchical relationship of cell types by performing hierarchical clustering on the average expression profiles of the cell types (Supplementary Figure S9). As an example, we remove the B cells from all the reference datasets and use mtANN and other methods to annotate the query dataset. The confusion matrices of mtANN and other methods are shown in Figure 4B. From the heatmaps, we can find that mtANN identifies most B cells as "unassigned" while all the other compared methods annotate most B cells as similar cell types (CD4+ T cell). For shared cell-type annotation, mtANN is better at distinguishing two similar cell types in the query dataset (such as CD14+ monocyte and CD16+ monocyte, CD4+ T cell and Cytotoxic T cell), while scmap-clust, scmap-cell, and, ItClust annotate a part of Cytotoxic T cells as CD4+ T cell, and Seurat v3 annotates most CD16+ monocyte cells as CD14+ monocyte. We also remove CD14+ monocyte cells from all reference datasets and find that mtANN identifies most CD14+ monocyte as "unassigned" while scmap-clust, scmap-cell, and, Seurat v3 identify most CD14+ monocyte cells as CD16+ monocyte (Supplementary Figure S10). This illustrates that mtANN can better distinguish two types of cells with small biological differences when an unseen cell type is present.

In reality, we cannot know the actual proportion of unseen cell types, so the default threshold provided by each method is more practical and essential. When using the default method to select threshold, the prediction accuracy of mtANN, scmap-clust, scmap-cell, and, Seurat v3 in all the tests are exhibited in Supplementary Figure S11. We can see that the accuracy of mtANN is higher than those of the compared methods when "Celseq", "Drops", "inDrop", "Smart-seq2", "10X v2", and "10X v3" are evaluated as the query datasets (Supplementary Figure S11A). Supplementary Figure S11B shows that mtANN also has the best performance when "Baron human" and "Xin" are used as the query datasets. We find that mtANN has the best performance in almost all tests (Supplementary Figures S11C-D) when we count the number of times each method ranks first in accuracy across all tests. In particular, the result of mtANN under the default threshold is similar to the result under the actual proportion (Supplementary Figures S12-S13), which shows that the threshold selected by mtANN is comparable to the actual proportion of unseen cells.

Cell-type annotation of COVID-19 patients with different symptoms

Coronavirus disease 2019 (COVID-19) has caused more than 536 million infections and more than 6.3 million deaths, according to World Health Organization (WHO) statistics as of June 19, 2022. It is thus important to annotate the cell type of the sequencing data from patients for understanding the disease mechanism. With many scRNA-seq data from COVID-19 patients available, we select the study of COVID-19 [23] which offers a comprehensive immune landscape, including 284 samples from 196 COVID-19 patients and controls to assess the

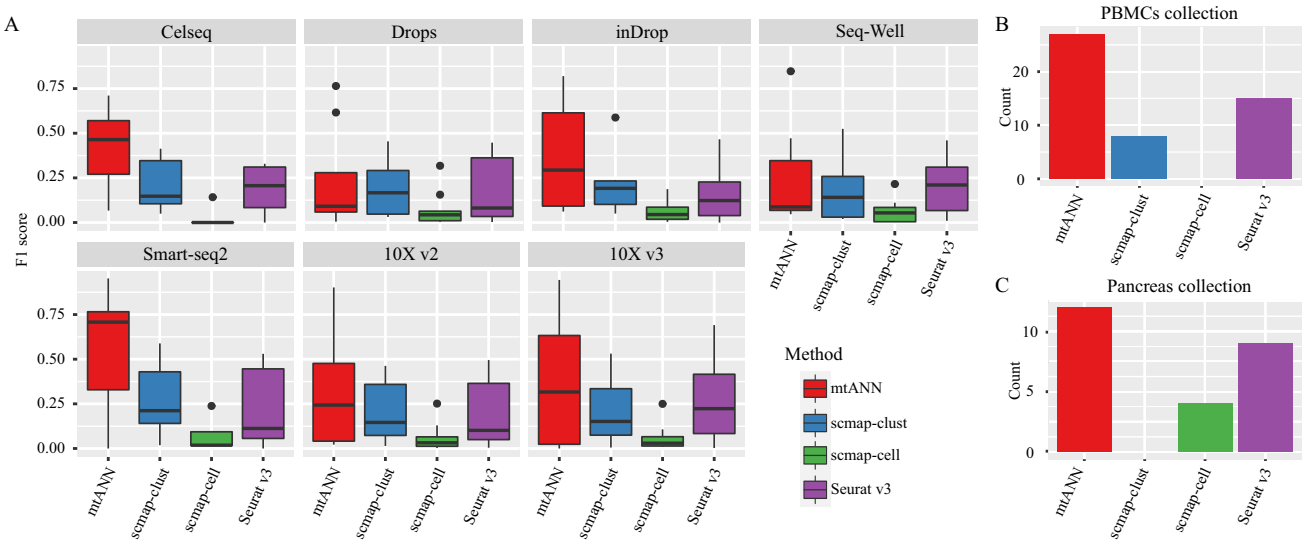

Fig. 3. The performance of each method for identifying unseen cell types under the default threshold. **(A)** Boxplots of the F1 scores of different methods on the PBMCs collection. The results with different query datasets are displayed in different panels. **(B)** and **(C)** Barplots of the statistics of the number of times that the F1 score ranks first for each method in all tests on the PBMCs **(B)** and pancreas **(C)** collections. The x-axis represents the method, and the y-axis represents the counts.

performance of mtANN on real data. We use the dataset from PBMC cells in the COVID-19 dataset as the query datasets and the PBMCs collection we used above as reference [18] to evaluate the performance of mtANN and other methods.

We group the cells according to samples' id, resulting in 249 query datasets. mtANN is compared with scmap-clust, scmap-cell and, Seurat v3 under the default threshold parameters of identifying unseen cells. The accuracies of mtANN and other methods on the 249 query datasets are presented in Figure 5A. It can be seen that the prediction accuracy of mtANN for patients with different symptoms is higher than other methods, and scmap-cell suffers a decrease. We further conduct a one-to-one comparison and find that mtANN significantly (two-sided paired Wilcoxon test, p-value < 0.01) outperforms the compared methods (Figure 5B). We compare the composition of cell types between patients with different symptoms and find that the proportion of B cells increases in patients with severe symptoms, and the percentage of dendritic cells and T cells decreases, particularly in patients with severe symptoms (Figure 5C), which is consistent with the lymphopenia phenomenon previously reported [24]. We get the conclusion that the percentage of megakaryocyte and CD14+ monocyte elevates in patients with severe symptoms, which is also the same phenomenon as observed in the original datasets [23].

Discussion

With the development of single-cell sequencing technology, traditional unsupervised clustering-based cell-type annotation methods are difficult to adapt to rapidly generated datasets since they are time-consuming [25, 26]. Another method for automatic cell-type annotation based on a reference atlas has been widely studied, but these methods are barely able to discover unseen cell types [27]. The identification of an unseen cell type may lead to new biological discoveries, while the erroneous identification may lead to missing new biological discoveries or lead improper biological conclusions. Only some of the previous methods for automatically annotating cell types address the problem of identifying unseen cell types [11, 12], and all of them only set a default threshold instead of proposing a methodology to automatically select a threshold. The choice of threshold determines the exactitude and useability of the method.

In this study, we propose a novel ensemble learning-based cell-type annotation method, mtANN, to annotate cell type labels for a query dataset automatically. Firstly, mtANN integrates datasets containing different cell types, enriching the cell types of the reference atlas to reduce the presence of unseen cell types in the query dataset. Secondly, mtANN proposes a metric to efficiently discriminate unseen cell types in query datasets. Finally, mtANN provides a data-driven methodology to adaptively select thresholds, enabling mtANN to automatically identify unseen cell types and simultaneously annotate shared cell types. Recently, several methods have emerged that integrate multiple reference datasets to annotate query datasets [28, 29], but they do not focus on the presence of unseen cell types. Our comprehensive benchmark and application on an extensive set of publicly available benchmark datasets indicate that mtANN has achieved state-of-the-art performance for unseen cell-type identification and cell-type annotation in the meantime.

There may be two challenges in integrating multiple reference datasets that we have not considered in this work. The first one is the inconsistent terminology of cell types across different reference datasets. In this work, we avoid this problem by collecting reference datasets with as consistent terminology as possible. However, if we want to integrate more datasets, this problem will be inevitable. Several approaches can be attempted in the future to match cell types between datasets, such as matching based on marker genes of cell types or matching by mutual prediction between datasets [30]. Another more challenging problem is that the annotation resolutions of the reference datasets may be inconsistent [31]. Two directions can be taken into consideration in the future. On the one hand, based on the existing labels of the reference datasets, we

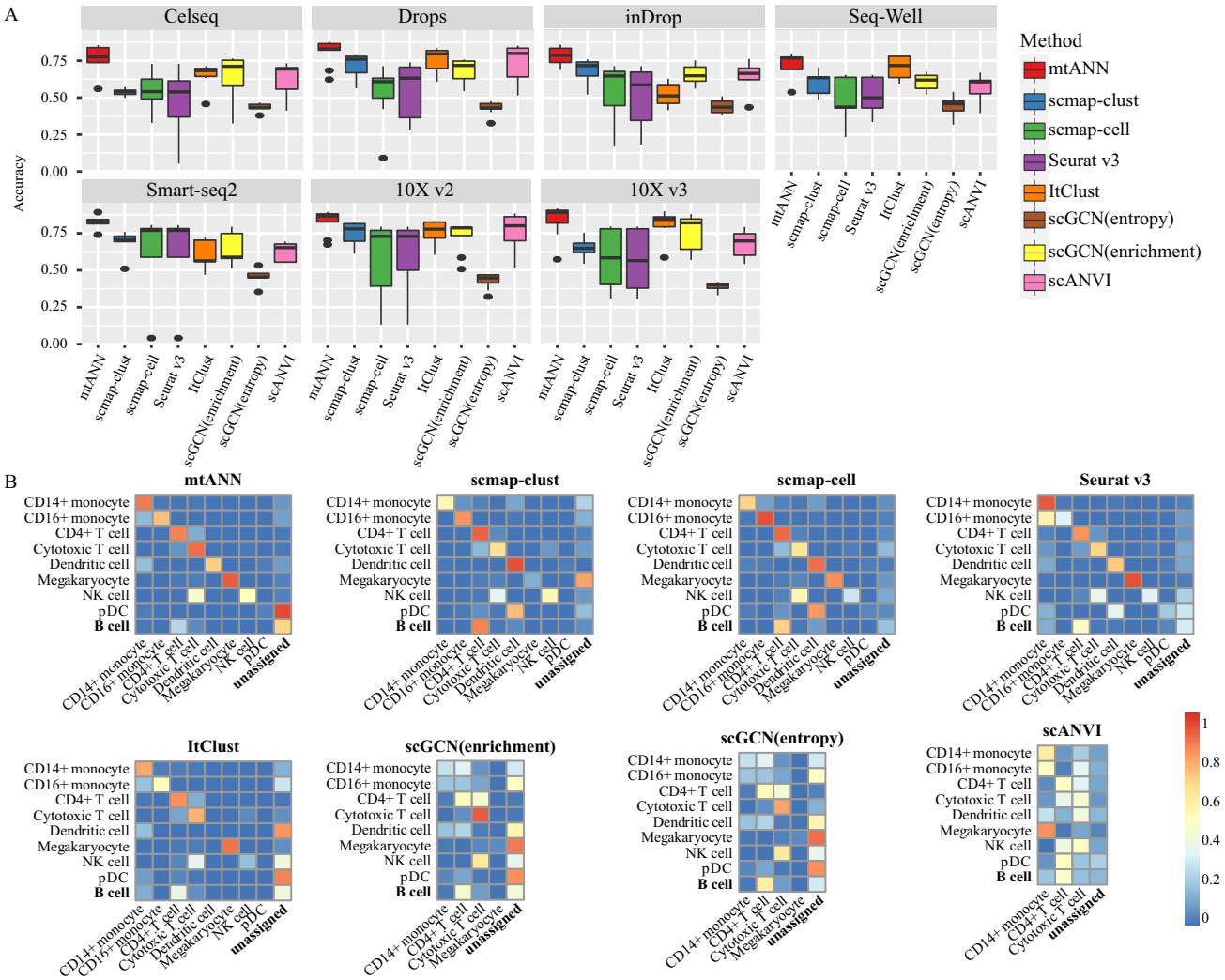

Fig. 4. Performances of mtANN and other compared methods in cell-type annotation when using the actual proportion of unseen cells as thresholds to identify unseen cells. **(A)** Boxplots of the accuracies of different methods on the PBMCs collection. The results with different query datasets are displayed in different panels. **(B)** Heatmaps of the confusion matrices of mtANN and other methods when B cells in the query dataset are the real unseen cell type. In a confusion matrix, the row and column names correspond to the true cell labels and the predicted cell labels of the query dataset, while the element represents the proportion of cells belonging to one cell type that is predicted to be of other cell types. Note that NK cell and pDC are abbreviations for the Natural killer cell and the Plasmacytoid dendritic cell.

can further subcluster these datasets separately to a uniform resolution level. On the other hand, incorporating hierarchical relationships among cell types which can be constructed according to some prior knowledge (e.g., Cell Ontology [32]) into the development of cell-type annotation methods can not only help to improve the validity of a method but also provide clues for exploring the identities of the uncertain cells [7, 33].

So far, we have marked the cells that are considered to belong to unseen cell types as "unassigned". One limitation of our method is that we do not provide a further biological interpretation of these cells. A straightforward way to explore the identities of these cells is to use unsupervised annotation methods. In addition, as mentioned earlier, integrating Cell Ontology into the method may enable automatic annotation of "unassigned" cells, such as assigning them to supertypes of an observed cell type [34]. In the future, we will extend our method to implement this functionality.

Supplementary data and code

Supplementary data is available online. The source code is available at `https://github.com/Zhangxf-ccnu/mtANN`.

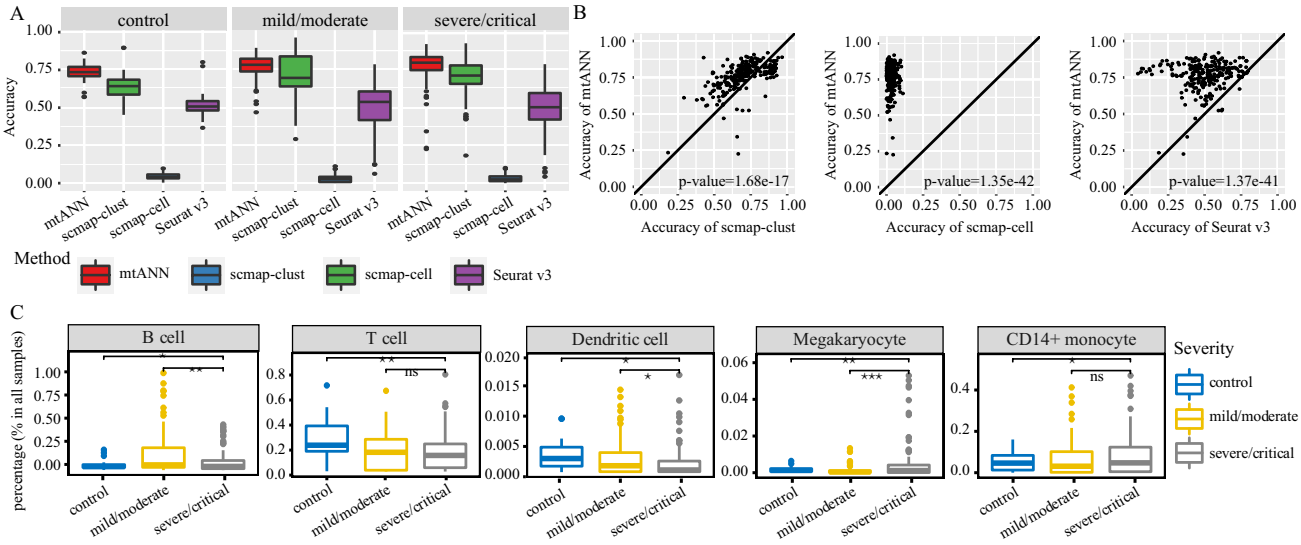

Fig. 5. Performances of mtANN and other compared methods in cell-type annotation on the COVID-19 dataset. **(A)** Boxplots of the accuracies of different methods on samples in different symptoms. **(B)** One-to-one comparison between mtANN and scmap-clust, scmap-cell and Seurat v3. Each point represents a query dataset. P-values of two-sided paired Wilcoxon signed-rank tests used to test the performance difference are reported. **(C)** Boxplots of the compositions of B cells, T cells, dendritic cells, megakaryocyte cells and, CD14+ monocyte between samples with different symptoms. The significance of the two-sided T-test is represented by stars where one star, two stars and, three stars mean the corresponding p-value less than 0.05, 0.01 and, 0.001, respectively and *ns* means the corresponding p-value greater than 0.05.

Key Points

- Supervised cell-type annotation relies on the diversity of cell types in the reference. For technical and biological reasons, new query data of interest may contain unseen cell types that are missing from the reference. Identifying unseen cell types is critical for new biological discoveries.
- We propose a new method to automatically annotate query data while accurately identifying unseen cell types. It improves predictive power by combining the ideas of deep learning and ensemble learning. It also introduces a new metric to measure whether a cell belongs to an unseen cell type and a new data-driven approach to automatically determining the corresponding threshold.
- Using two collections of datasets, we conduct a total of 75 benchmark experiments to show that our method outperforms state-of-the-art methods in both unseen cell-type identification and cell-type annotation. We also demonstrate the predictive power of mtANN on a total of 249 tests using a COVID-19 dataset.
- A Python package mtANN is developed to implement our proposed cell-type annotation procedure.

Funding

This work was supported by the National Natural Science Foundation of China [12271198,11871026], and Hubei Provincial Science and Technology Innovation Base (Platform) Special Project [2020DFH002].

References

1. Lähnemann D, Köster J, Szczurek E, et al. Eleven grand challenges in single-cell data science. *Genome Biol.*, 21(1):1–35, 2020.
2. Kiselev VY, Kirschner K, Schaub MT, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nat. Med.*, 14(5):483–486, 2017.
3. Wang B, Zhu J, Pierson E, et al. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nat. Methods*, 14(4):414–416, 2017.
4. Brbić M, Zitnik M, Wang S, et al. Mars: discovering novel cell types across heterogeneous single-cell experiments. *Nat. Methods*, 17(12):1200–1206, 2020.
5. Pasquini G, Rojo Arias JE, Schäfer P, et al. Automated methods for cell type annotation on scrna-seq data. *Comput. Struct. Biotechnol. J.*, 19:961–969, 2021.
6. Zhao X, Wu S, Fang N, et al. Evaluation of single-cell classifiers for single-cell rna sequencing data sets. *Brief. Bioinformatics*, 21(5):1581–1595, 2020.
7. Lin Y, Cao Y, Kim HJ, et al. scclassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Mol. Syst. Biol*, 16(6):e9389, 2020.
8. Hu J, Li X, Hu G, et al. Iterative transfer learning with neural network for clustering and cell type classification in single-cell rna-seq analysis. *Nat. Mach. Intell.*, 2(10):607–618, 2020.
9. Song Q, Su J, and Zhang W. scgcn is a graph convolutional networks algorithm for knowledge transfer in single cell omics. *Nat. Commun.*, 12(1):1–11, 2021.
10. Xu C, Lopez R, Mehlman E, et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.*, 17(1):e9620, 2021.
11. Kiselev VY, Yiu A, and Hemberg M. scmap: projection of single-cell rna-seq data across data sets. *Nat. Methods*, 15(5):359–362, 2018.
12. Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

13. Denisenko E, Guo BB, Jones M, et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus rna-seq workflows. *Genome Biol.*, 21(1):1–25, 2020.
14. Slyper M, Porter CBM, Ashenberg O, et al. A single-cell and single-nucleus rna-seq toolbox for fresh and frozen human tumors. *Nat. Med.*, 26(5):792–802, 2020.
15. Rozenblatt-Rosen O, Stubbington MJT, Regev A, et al. The human cell atlas: from vision to reality. *Nature*, 550(7677):451–453, 2017.
16. Chen L, He Q, Zhai Y, et al. Single-cell rna-seq data semi-supervised clustering and annotation via structural regularized domain adaptation. *Bioinformatics*, 37(6):775–784, 2021.
17. Li C, Ma H, Yuan Y, et al. Improved deep embedded clustering with local structure preservation. In *IJCAI*, pages 1753–1759, 2017.
18. Ding J, Adiconis X, Simmons SK, et al. Systematic comparison of single-cell and single-nucleus rna-sequencing methods. *Nat. Biotechnol.*, 38(6):737–746, 2020.
19. Baron M, Veres A, Wolock SL, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Syst.*, 3(4):346–360, 2016.
20. Muraro MJ, Dharmadhikari G, Grün D, et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst.*, 3(4):385–394, 2016.
21. Segerstolpe Å, Palasantza A, Eliasson P, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.*, 24(4):593–607, 2016.
22. Xin Y, Kim J, Okamoto H, et al. Rna sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.*, 24(4):608–615, 2016.
23. Ren X, Wen W, Fan X, et al. Covid-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell*, 184(7):1895–1913, 2021.
24. Chen Z and John Wherry E. T cell responses in patients with covid-19. *Nat. Rev. Immunol.*, 20(9):529–536, 2020.
25. Guo C, Li B, Ma H, et al. Single-cell analysis of two severe covid-19 patients reveals a monocyte-associated and tocilizumab-responding cytokine storm. *Nat. Commun.*, 11(1):1–11, 2020.
26. Lee JTH and Hemberg M. Supervised clustering for single-cell analysis. *Nat. Methods*, 16(10):965–966, 2019.
27. Zhang Y, Zhang F, Wang Z, et al. scmagic: accurately annotating single cells using two rounds of reference-based classification. *Nucleic Acids Res.*, 50(8):e43–e43, 2022.
28. Duan B, Chen S, Chen X, et al. Integrating multiple references for single-cell assignment. *Nucleic Acids Res.*, 49(14):e80–e80, 2021.
29. Yuan M, Chen L, and Deng M. scmra: a robust deep learning method to annotate scrna-seq data with multiple reference datasets. *Bioinformatics*, 38(3):738–745, 2022.
30. Michielsen L, Reinders MJT, and Mahfouz A. Hierarchical progressive learning of cell identities in single-cell data. *Nat. Commun.*, 12(1):1–12, 2021.
31. Li J, Sheng Q, Shyr Y, et al. scmrma: single cell multiresolution marker-based annotation. *Nucleic Acids Res.*, 50(2):e7–e7, 2022.
32. Diehl AD, Meehan TF, Bradford YM, et al. The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semant.*, 7(1):1–10, 2016.
33. Bernstein MN, Ma Z, Gleicher M, et al. Cello: Comprehensive and hierarchical cell type classification of human cells with the cell ontology. *Iscience*, 24(1):101913, 2021.
34. Wang S, Pisco AO, McGeever A, et al. Leveraging the cell ontology to classify unseen cell types. *Nat. Commun.*, 12(1):1–11, 2021.

Yi-Xuan Xiong is a PhD student at School of Mathematics and Statistics, Central China Normal University, Wuhan, China. Her research interests include bioinformatics and machine learning.

Meng-Guo Wang is a PhD student at School of Mathematics and Statistics, Central China Normal University, Wuhan, China. His research interests include bioinformatics and machine learning.

Luonan Chen is a professor and the executive director with the Key Laboratory of Systems Biology, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China. His current research interests include systems biology, bioinformatics and nonlinear dynamics.

Xiao-Fei Zhang is a professor at School of Mathematics and Statistics, Central China Normal University, Wuhan, China. His current research interests include data mining, machine learning and bioinformatics.