scDeepInsight: a supervised cell-type identification method for scRNA-seq data with deep learning

Shangru Jia[1], Artem Lysenko[2,3], Keith A Boroevich[3], Alok Sharma[2,3,4,5,*], Tatsuhiko Tsunoda[1,2,3,5,*]

[1] Laboratory for Medical Science Mathematics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Japan.

[2] Laboratory for Medical Science Mathematics, Department of Biological Sciences, School of Science, The University of Tokyo, Japan.

[3] Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Japan.

[4] Institute for Integrated and Intelligent Systems, Griffith University, Australia.

[5] Co-last authors

* Corresponding authors: alok.fj@gmail.com, tsunoda@bs.s.u-tokyo.ac.jp

Abstract

Annotation of cell-types is a critical step in the analysis of single-cell RNA sequencing (scRNA-seq) data that allows to study heterogeneity across multiple cell populations. Currently this is most commonly done using unsupervised clustering algorithms, which project single-cell expression data into a lower dimensional space and then cluster cells based on their distances from each other. However, as these methods do not use reference datasets, they can only achieve a rough classification of cell-types, and it is difficult to improve the recognition accuracy further. To effectively solve this issue we propose a novel supervised annotation method, scDeepInsight. The scDeepInsight method is capable of performing manifold assignments. It is competent in executing data integration through batch normalization, performing supervised training on the reference dataset, doing outlier detection and annotating cell-types on query datasets. Moreover, it can help identify active genes or marker genes related to cell-types. The training of the scDeepInsight model is performed in a unique way. The tabular scRNA-seq data are first converted to corresponding images through the DeepInsight methodology. DeepInsight can provide a trainable image transformer to convert non-image RNA data to images by comprehensively comparing interrelationships among multiple genes. Subsequently, the converted images are fed into convolutional neural networks (CNNs) such as EfficientNet. This enables automatic feature extraction to identify the cell-types of scRNA-seq samples. We benchmarked scDeepInsight with five other mainstream cell annotation methods. The average accuracy rate of scDeepInsight reached 88.2%, which is more than 7% higher compared with the state-of-the-art methods.

Introduction

Rapidly developing single-cell RNA sequencing (scRNA-seq) technologies have made it possible to observe gene expression at the single-cell level and enhance our understanding of complex biological systems and diseases, such as cancer and chronic diseases. These methods have significant implications for studying a wide variety of tissues and the different types of cells within them. Accurate cell annotation is a prerequisite for downstream analysis of single-cell data. However, cell annotation is a time-consuming and expert-dependent step in single-cell analysis. The annotation process can be divided into three steps [1], automatic annotation, expert manual annotation and verification. Due to emergence of a large number of single-cell datasets, manual annotation by experts based on empirical analysis is practically impossible to meet research needs. Furthermore, manual annotation is not only time-consuming and laborious, but the annotation results can be subjective.

Owning to these factors, improving the accuracy of automatic annotation has become an essential area of research and in response to these practical needs the researchers have proposed a variety of relevant annotation methods. These can be roughly divided into three different types according to the theoretical basis [2]: 1) annotation based on marker genes, 2) annotation using correlation analysis with reference datasets, and 3) annotation by supervised classifiers trained on reference datasets.

Most single-cell annotation methods start with unsupervised cell clustering analysis. Cells are firstly clustered through clustering methods such as k-means [3], Single-Cell Consensus Clustering (SC3) [4], and shared nearest neighbor (SNN) [5]. The clusters are then mapped to different cell-types by analyzing the abundance of

marker genes within each cluster. However, these marker gene-based methods have many limitations. The first issue is the accuracy of the marker gene database. Even though there are already databases such as PanglaoDB [6], ScType [7], and CellMarker [8], the selection of some marker genes still depends on the prior research knowledge. The second is that information on marker genes is often insufficient for many cell subtypes and, in particular, newly discovered cell-types.

Annotation methods based on correlation analysis with reference datasets tend to be more accurate than marker gene-based methods [9]. This is because gene-gene correlations are generally ignored when analyzing marker gene lists. Annotation can be done more comprehensively by correlating target annotated datasets with reference datasets of similar biological tissues. However, batch effects between the reference dataset and the target dataset can hinder the correct annotation of cell-types. Technical differences, such as sequencing methods and experimental batches, will affect the results of single-cell sequencing. While current annotation methods by correlation analysis, such as SingleR [10], often do not offer batch effect processing methods as part of their pipeline. When annotating datasets obtained in different experiments, it is very difficult to eliminate the influence of the batch effects.

Single-cell datasets are often high-dimensional and sparse. In this case, machine learning (ML) is a good choice for processing complex sequencing data. After learning the expression patterns of multiple genes in different cell-types on the reference dataset, ML methods can transfer labels from the reference dataset to the target dataset. Furthermore, as a new field of machine learning research, deep learning is capable of learning highly abstracted representations from data such as images, sounds, and texts. Given the robustness of deep learning methods and the availability of finely annotated reference datasets, supervised learning models have gradually become widely used in

the analysis of reference datasets. Currently, representative machine learning methods for processing single-cell omics data mainly include Bidirectional Encoder Representations from Transformers (BERT) [2], Autoencoders (AEs) [11] and Recurrent Neural Networks (RNNs) [12]. Single-cell omics data are often treated as texts or sequence data since they are not images and do not have graph network structures.

In this paper, we propose scDeepInsight, an original method integrating the whole cell-type identification process (Figure 1). scDeepInsight can directly annotate the query dataset based on the model trained on the reference dataset. In the first step, scDeepInsight does preprocessing of scRNA-seq data, including quality control and integration through batch normalization. By integrating DeepInsight [13], our method is capable of converting the scRNA-seq samples to corresponding images. Images generated from the reference dataset are used to train a CNN, which can then be used to predict cell types found in the query dataset(s). As a machine learning model under deep supervised learning, the absolute superiority in image classification and feature extraction of CNN has been widely recognized [14]. Furthermore, extracted features in the training process are helpful in investigating marker genes through DeepFeature [15]. By uniquely converting scRNA-seq to image data, our method will help to fully exploit the advantages of CNNs in cell-type classification. scDeepInsight enables accurate and efficient annotation of multiple cell subtypes and can perform outlier detection. Cell-type prediction results are validated using reliable pre-annotated cell labels. Also, some rare/unknown cell-types which are not included in the reference datasets can be detected during the annotation process. Further details are explained in the Materials and Methods section.

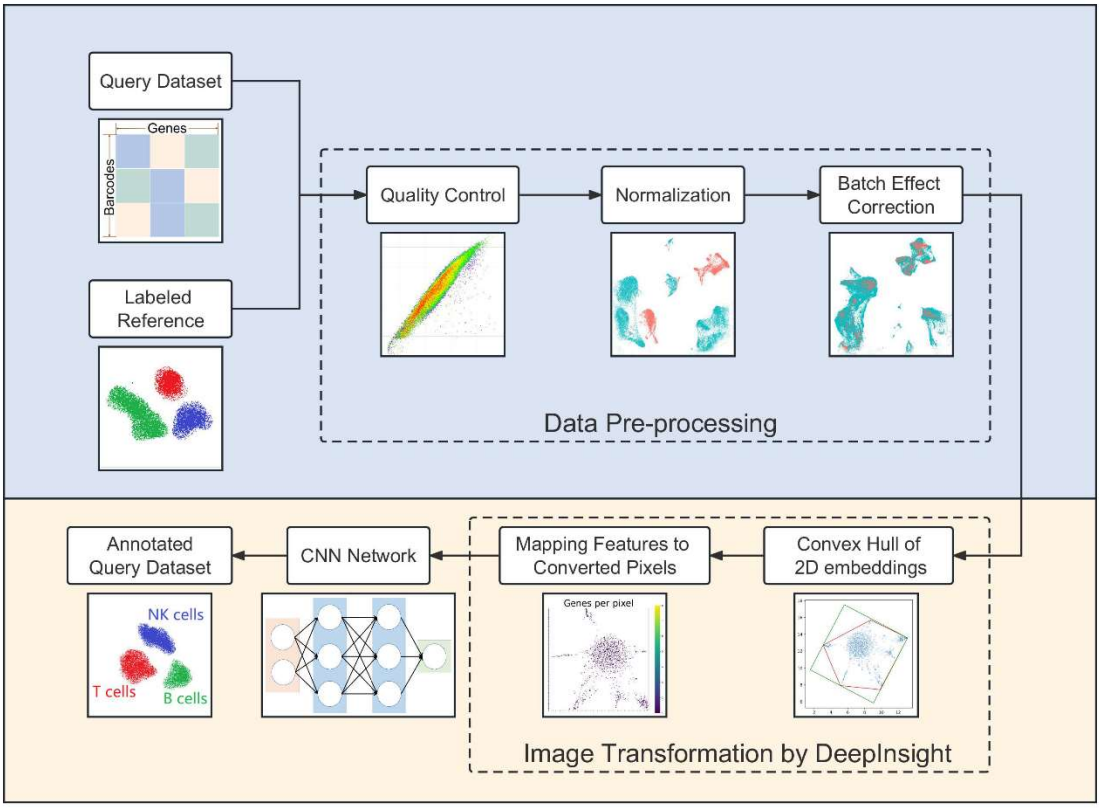

Figure 1. The scDeepInsight pipeline**:** the key steps performed by scDeepInsight from inputting single-unique molecular identifier count matrix to outputting cell annotation prediction. A reference dataset with a query data is processed via quality control, normalization and correction of batch effects. Then processed tabular data are converted into 2D embeddings. After framing and feature mapping, single-cell expression data are transformed into corresponding images. After this step, the reference dataset is used in training the CNN model. In the training step, no query dataset is used. Once the CNN model is trained, it is used to cluster single-cell samples from query dataset into cell-types. For subsequent query datasets, no further training of the reference dataset is performed, and therefore, the previously trained model can be directly used for clustering and annotation.

Materials and Methods

Overview of scDeepInsight

The workflow of scDeepInsight integrates the whole process of single-cell annotation, including data preprocessing, image conversion, neural network training

and cell-type prediction (Figure 1). It takes single unique molecular identifier (UMI) count matrices generated after sequencing as input. Each row of the matrix represents one cell with a unique barcode, and each column represents a gene. scDeepInsight is a single-cell labeling model based on supervised learning, so a reference dataset is also required. After preparing reference and test datasets, data preprocessing is performed. This step includes quality control, normalization, and correction of batch effect between the query dataset and reference dataset. Afterwards, DeepInsight is utilized to convert the processed non-image data into images. First, the processed data are transformed into two-dimensional embeddings by a visualization method like t-SNE. By mapping genes to pixels, the expression of different genes in a sample is transformed into a unique image. The lighter the pixel in the image, the higher the expression level of the gene in this cell. Next, the processed images converted from the reference dataset are used as the input for the CNN. After training, images transformed from the query dataset can be input into the trained model to complete the annotation of cell-types. The output of the model in the fully connected layer is the predicted probability of a specific cell-type. The label corresponding to the maximum probability is the predicted cell-type of this single-cell sample.

Sample quality control

Quality control directly impacts the reliability of the downstream analysis. By controlling the number of specific genes detected in a single-cell (nFeature_RNA), the total number of UMI detected (nCount_RNA) and the proportion of mitochondrial genes (percent.mt) in each cell, cells with little effective information can be filtered out. Low-quality cells or empty droplets typically have very few genes detected. However, if there is an overlap where two or more cells are captured simultaneously, UMI

detected in such cells will also become abnormally large. Both of these two cases should be removed. In this paper, we generally limit nFeature_RNA to between 300 and 4000 for single-cell samples. In addition, dying cells often exhibit extensive mitochondrial contamination. In this step, we set the threshold of percent.mt to 15 and cells exceeding this value are filtered out to avoid excessive influence of mitochondrial genes.

Gene expression normalization

Commonly used normalization methods such as Scanpy::zheng17 [16] include selecting highly expressed genes, normalization, and scaling. In scDeepInsight, we use SCTransform [17] to normalize the expression data. This method performs regularized negative binomial regression on total UMI counts per cell to eliminate the variance due to sequencing depth. UMI reads are commonly positively correlated with the sequencing depth of the cells. Traditional logarithmic normalization process cannot remove this correlation among different cell samples. When extracting variable genes, genes selected by SCTransform also demonstrate more biologically meaningful variations than traditional normalization methods [18]. In addition, by performing regression on the percentage of mitochondrial genes or cell cycle of samples, SCTransform can also remove the influence of these factors.

Batch Effect Correction

Batch effects have a quantitative impact on single-cell gene expression values. As a consequence, cells that should have been clustered together may end up divided into different clusters due to batch effects. We use Canonical Correlation Analysis (CCA) to eliminate the batch effect. CCA is a multivariate statistical method to study the correlation between two groups of variables, and it can reveal internal relationships

between variables. By calling Seurat::IntegrateData [19], the CCA method can be used to find anchors between datasets and integrate multi-sample datasets accordingly. This step can avoid the impact of biological heterogeneity caused by different experiment batches or sequencing technology on the accuracy of subsequent analysis.

Data Scaling

The processed data needs to be scaled to the range [0, 1] before it is passed on to the image converter. For scRNA-seq data, two scaling methods are commonly used. The first is according to the maximum and minimum values of the expression data on this gene. The second is to treat the dataset as a whole and scale it according to the global maximum and minimum values. In actual experiments, it will appear that the data of some highly expressed genes in cells is much higher than that of other genes. For this reason, the data processing function normalize_total of Scanpy also provides the option to ignore some particularly highly expressed genes when calculating regularization. In this context, differences in gene expression across cells are attenuated if scaled using global maximum and minimum. Therefore, each gene was scaled separately according to corresponding maximum and minimum values of expression.

Generation of images from tabular data

DeepInsight [13] is a widely recognized algorithm for converting non-image data into images. This algorithm can convert different types of non-image data, such as RNA-seq, artificial data, speech, and texts, into image data, which can be easily input into CNN for classification. In brief, DeepInsight can convert the input feature/element vector into Cartesian coordinates through the built-in visualization technique such as t-SNE, PCA, Kernal PCA and UMAP. However, it does not use these visualization

techniques in the usual manner. Instead, it positions the features (not samples) in the Cartesian plane, and therefore, avoids the dimensionality reduction of features [20]. The position of features in this two-dimensional plane depends on the similarity between features. After the positioning of features, DeepInsight uses the convex hull algorithm to find the smallest rectangle containing all feature points and rotate it to the horizontal or vertical direction applicable to CNN architecture. Next, Cartesian coordinates are converted to pixels according to the specified output image size. Finally, the feature/gene values are mapped to these locations in the pixel frame. The intensity of the pixel indicates the degree of corresponding gene expression in a single-cell. In this paper, pyDeepInsight [https://github.com/alok-ai-lab/pyDeepInsight] was used to perform image conversion. To avoid overlapping of genes in the same pixel, a proper resolution of the converted image should be set. The parameters of visualization techniques also have an impact on the experimental results. For instance, when selecting t-SNE, parameters for perplexity and distance function should be manually set. Since the number of cells in the sample used in the reference data set is relatively large, it is recommended to use a larger perplexity value [21].

CNN model training and validation

To take advantage shorter training time achievable by using transfer learning, a pre-trained version of EfficientNet-b3 [22] model was used for all analysis in this paper. Additionally, label smoothing and early stopping optimization techniques were used to reduce overfitting. All datasets used in the experiments were first subdivided into training and validation parts at 85:15 ratio. One of the training accuracy plots is shown in Supplement figure 1.

Results

Datasets and preprocessing

Database of peripheral blood mononuclear cells (PBMC) offers readily accessible heterogeneous cell samples that contain several similar but distinct cell-types. PBMC datasets are usually dominated by several cell-types and are very unbalanced datasets. Furthermore, there are many similar cell subtypes, such as CD4+ Central Memory T (CD4 TCM) and CD4+ Central Effector T (CD4 TEM). Correctly annotating PBMC cells has always been a challenging task for researchers [7, 11]. In this paper, we used a PBMC dataset that had been labeled by experts in prior experiments [23] as a reference dataset for scDeepInsight training. The reference dataset contained over 160,000 single-cell samples and 31 different cell-types.

Independent test (or query) datasets used in this paper for benchmarking scDeepInsight are described in Table 1. The number of cell samples and cell-types contained in test datasets were relatively large. Test datasets were obtained by two different sequencing technologies, 10x Multiome 3' v2 and v3.

Table 1. Summary of datasets

Dataset	Genes	Cells	Protocol	Cell types	
Hao	20729	161764	10x Multiome 3' v3	31	Reference
Yazar	36571	255734	10x Multiome 3' v2	25	Test
Schulte-Schrepping	33228	45787	10x Multiome 3' v2,v3	20	Test
Arunachalam	22878	25954	10x Multiome 3' v3	25	Test
Lee	22021	43512	10x Multiome 3' v3	24	Test
10x-Multiome-Pbmc10k	29095	9631	10x Multiome	19	Test

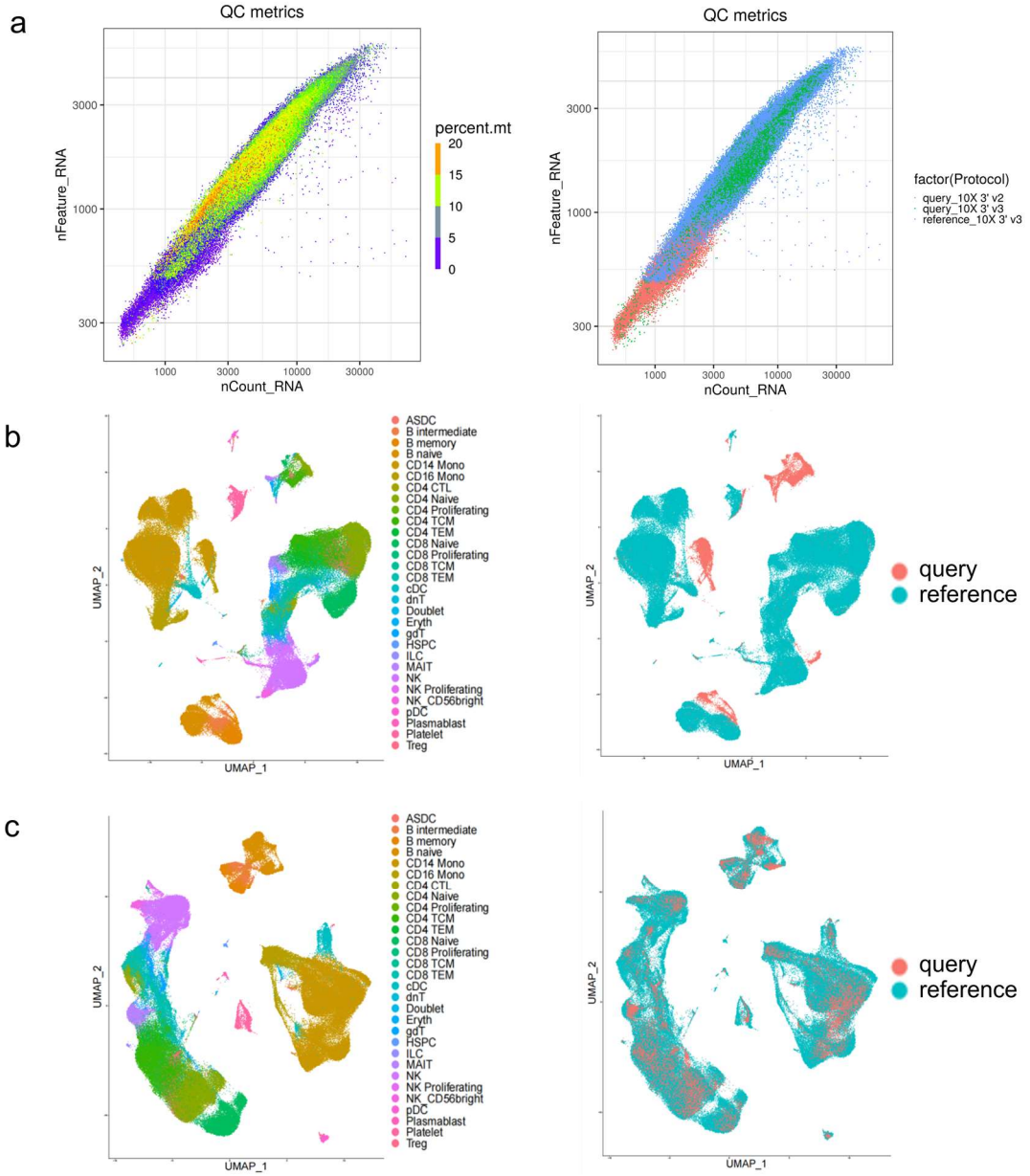

Figure 2. Preprocessing results of the test dataset Schulte-Schrepping. (a) The left section is the quality control plot of the test dataset. The cells whose nFeature_RNA was less than 300 or more than 4000 were filtered out. Also, cells with percent.mt larger than 15 were also excluded. The right plot is labeled by the sequencing technology of data. (b) The Uniform Manifold Approximation and Projection (UMAP) representation of the reference before batch effect correction labeled by cell-types and data sources (c) The UMAP representation of datasets after batch effect correction.

The test dataset needs to be preprocessed before using DeepInsight for image conversion. After quality control, cell samples with poor sequencing quality were filtered out (Figure 2a; see Materials and Methods section). Then we used SCTransform to complete the screening for highly expressed genes and do the normalization. Dealing with batch effects between reference and test datasets is an important step in preprocessing. Samples sequenced by 10x Multiome 3' v3 in the test dataset used the same sequencing technology as the reference dataset, so the sequencing depth (total UMI counts per cell) was approximately the same. However, in the part of the test set sequenced by 10x Multiome v2, fewer genes and UMI counts were detected, and the sequencing depth was shallower than that of the reference dataset. Such differences brought by sequencing platforms should be addressed when dealing with batch effects. Experiments showed that CCA could successfully eliminate the batch effect (Figure 2b). Prior to batch effect correction distribution patterns displayed high degree of variation between the samples from reference and test datasets, even though these cell samples were of the same cell-type (such as CD4 TCM and CD4 TEM cells). After eliminating the batch effect, the distribution difference between expression data in UMAP results was mainly determined by cell-types, rather than the batch effect brought by the sequencing experiment between the test dataset and the reference dataset (Figure 2c).

Performance of scDeepInsight and comparison with other methods

Accuracy was verified on five independent PBMC test datasets with high-quality pre-annotated labels. We used accuracy and adjusted Rand index (ARI) to measure the performance of cell-type prediction. Accuracy represents the percentage of predicted cell-types that perfectly fit with the cell-types labeled in previous reliable experiments.

ARI is a measure of similarity between real cell labels and predicted cell-type clusters. The higher the value of ARI, the more similar the predicted results are to the original labels of cell samples.

To benchmark the accuracy of our method, we implemented different kinds of mainstream cell annotation methods. The first type was annotation based on marker genes, including SCINA, SC3 and Seurat::FindClusters. As a graph-based clustering algorithm provided by Seurat, FindClusters is able to identify clusters through shared nearest neighbor (SNN) modularity optimization. Different from SC3 and SNN clustering, SCINA does not implement unsupervised clustering. SCINA directly performs enrichment analysis on the specific marker gene list to assign type annotation. Furthermore, we chose ScType as a marker gene database. Integrating two marker gene databases, CellMarker and PanglaoDB, it provides a comprehensive database of specific markers covering many cell-types. In this paper, we refer to marker genes of the immune system in the ScType database to complete the annotation of clustering results. In order to ensure the fairness of the comparison, the marker genes were not manually selected when implementing these methods. For the reference-based annotation method, we chose SingleR, which achieves cell-type annotation by calculating and comparing the Spearman correlation coefficient between single-cell samples in query and reference datasets. To guarantee the validity of benchmarking, the same reference dataset used in scDeepInsight was utilized for learning. We also tested CellTypist to make the comparison more comprehensive. Being an automatic labeling method, CellTypist does not need an additional reference data set. In contrast, it directly offers built-in, pre-trained models for different tissues by integrating multiple single-cell data sets. The model we chose was Healthy_COVID19_PBMC, which best fitted the test datasets.

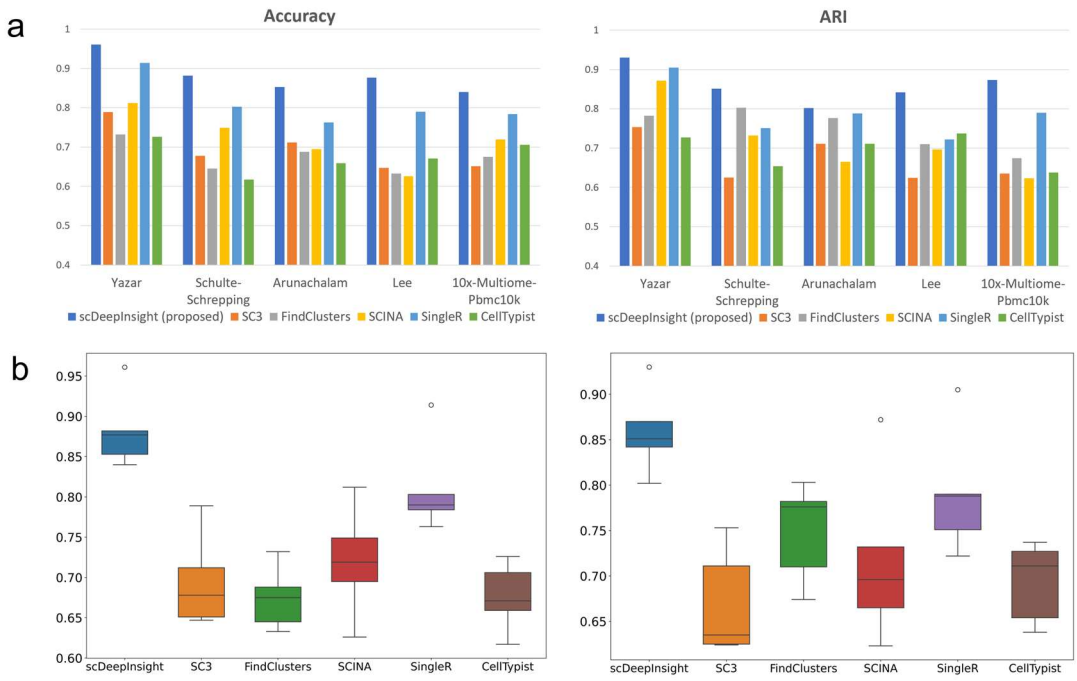

Figure 3. The performance of scDeepInsight. (a) Accuracy and ARI of scDeepInsight compared to other methods: SC3, FindClusters, SCINA, SingleR, and CellTypist, across the five datasets: Yaza, Schulte-Schrepping, Arunachalam, Lee, and 10x-Multiome-Pbmc10k. (b) The accuracy and ARI box plots of scDeepInsight and the other five methods used in benchmarking are depicted.

On five test datasets, our method outperformed the other five methods (Figure 3). The average accuracy of scDeepInsight (88.2%) was more than 7% greater than the next best performing method, singleR. Specifically, on dataset Yazar, the accuracy was the highest (96.1%). Also, the average ARI of scDeepInsight (0.859) on these five datasets was about 0.07 higher than the other five methods, which shows that the prediction results of scDeepInsight could better reflect real clustering patterns of cell labels. In addition, these query datasets included both 10X Multiome 3'v2 and v3 sequencing technologies, which also highlights the robustness of scDeepInsight.

Furthermore, scDeepInsight had a high recognition accuracy for main cell-types in the test dataset (Figure 4). As shown in Figure 4a, 99.4% of CD14 Monocytes could be correctly labeled as CD14 Mono by scDeepInsight. CD14 Monocytes accounted for

more than 35% of the test dataset (Figure 4b). The heatmap of confusion matrix could clearly demonstrate the high recognition accuracy on most cell-types by scDeepInsight (Figure 4c; corresponding detailed confusion matrix is shown in Supplement Figure 2). The cell-type prediction results by scDeepInsight were primarily consistent with the original labels and further subdivided into some cell-types (Figure 4d). The other unsupervised clustering methods rarely found the correct number of clusters. Some of the cell subtypes, including CD4 Proliferating and CD8 Proliferating, could not be recognized using other annotation methods. Cell-type annotation results of the Schulte-Schrepping dataset using SC3 with ScType are shown in Figure 4e. Samples were first clustered by SC3 and then annotated by referring to the marker gene database ScType. Compared with prediction results by scDeepInsight, this annotation method recognized fewer clusters and cell subtypes.

The performance results of scDeepInsight in the benchmarking were more reliable than some methods that directly divide the original data set into the training set and test set and were more in line with real annotation requirements. For example, when using 90% of the Zheng68K dataset as the training set and the remaining 10% as the test set, the prediction accuracy of scBERT [2] was 0.759. Under this premise, the accuracy rate of scDeepInsight's cell-type annotation on this Zheng68K dataset could still reach 0.837. However, in this benchmarking approach, the test set was not completely independent from the reference set, which was not very reliable, so this comparison was not included in the results table.

In conclusion, compared with other mainstream sequencing methods, our method was not only more accurate but also more capable of detecting similar cell subtypes.

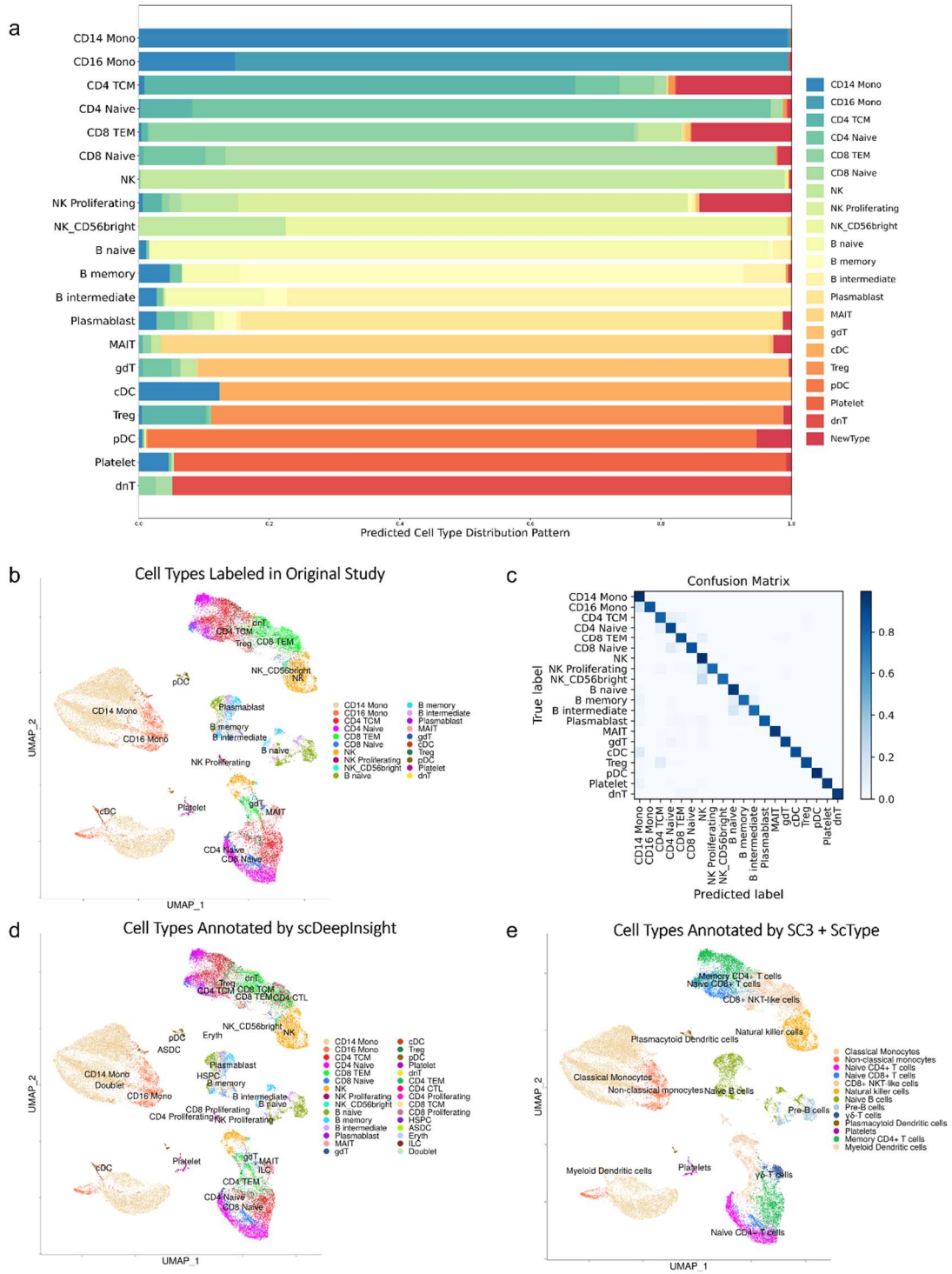

Figure 4. Cell-type labels of the reference dataset and prediction results on dataset Schulte-Schrepping. (a) The stacked percentage column chart of the prediction results on the Schulte-Schrepping dataset. (b) UMAP representation colored by cell-types in the original study (c) Heatmap of the confusion matrix. (d) UMAP representation colored by cell-types predicted by

scDeepInsight. (e) UMAP representation colored by annotation results using SC3 clustering with ScType.

Identification of marker genes

By analyzing the fully connected layer in a trained CNN, DeepFeature [15] can construct Class Activation Mapping (CAM) to extract features of different classes. In this study of the single-cell classification, we also introduced CAM to help analyze the features deferentially expressed among multiple cell-types. After applying DeepFeature to single-cell datasets, it was found that the genes extracted in different cell-types contained marker genes corresponding to these types. Generated CAM graphs are shown in Supplement Figure 5. Through the analysis of the extracted marker genes, the accuracy of the cell-type annotation model established by scDeepInsight at the biological level could also be proven. We selected three marker genes for Monocytes extracted by scDeepInsight from the reference dataset and colored the UMAP representation of the reference according to the expression values (Figure 5a). By referring to the cell-type distribution of the reference dataset shown in Figure 5b, the differential expression of these three genes in different regions also confirmed their reliability as marker genes.

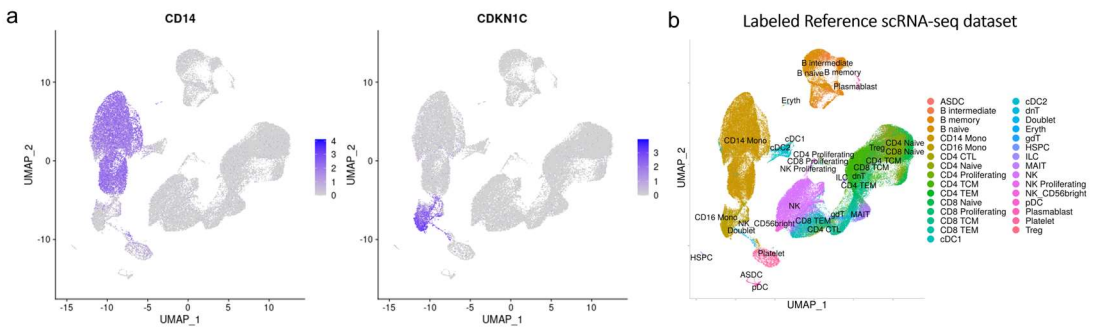

Figure 5. (a) CD14 [24] and CDKN1C [25] were proven to be marker genes for monocytes, CD14 Monocytes and CD16 Monocytes correspondingly in the previous study. (b) The UMAP 2D embedding after performing normalization and principal component analysis (PCA) dimensionality

reduction on the reference dataset. Cells are grouped and colored by known labels from previous studies.

Summary of overall performance

Compared with unsupervised clustering algorithms, supervised labeling methods with reference datasets can often achieve more accurate identification. However, supervised annotation methods require a correctly labeled reference dataset as the prerequisite, which is a limitation. In addition, training on reference datasets also brings additional time costs. Nonetheless, as an original supervised cell-type annotation method, scDeepInsight converts preprocessed single-cell data into images and utilizes CNN's strengths in feature extraction and classification, enabling accurate identification of cells in a few epochs. On large-scale PBMC test datasets, ARI and annotation accuracy have improved compared with SC3, SCINA, SingleR and other mainstream methods.

Discussion

The ability to discover new cell-types

In actual experiments, reference datasets or marker gene databases may not cover all cell-types contained in the test dataset. Unsupervised methods can reveal new cell-types when detecting clusters, which cannot be annotated to known cell-types. However, some supervised cell-type annotation methods simply label all samples as known cell-types contained in the reference dataset. These methods ignore rare/unknown cell-types, thus affecting the labeling accuracy. What our method outputs in the fully connected layer is the probability that each single-cell should be labeled as respective cell-types. The higher the prediction probability, the more likely the trained model is to label the

sample as this cell-type. Therefore, by setting a threshold on the predicted probability, we can filter out the cases where the predicted probability of the target sample is very low for all cell-types. In this situation, a sample which is not similar to any known cell-types in the reference dataset should be predicted as an unknown type.

The reference set and the five test sets used for benchmarking in the paper are all derived from healthy donors, and the cell-types do not contain neutrophils. While one dataset contains cells obtained from donors infected with COVID-19, it can be found that such datasets contain few neutrophils. However, neutrophils do not exist in the reference PBMC dataset, and these cells would be filtered out. When using data obtained from infected donors in the Lee dataset as the test set, the predicted probability returned by the fully connected layer is shown in Supplement Figure 3. When the 1% of cells with the smallest predicted probability are screened out, 72 out of a total of 89 neutrophils contained in this query dataset can be correctly detected and annotated as unknown cell-types.

Different sequencing methods can bring biological difference

Current batch effect correction methods, such as IntegrateData, Combat and Harmony, can correct the technical differences between datasets due to different platforms and batches of experiments. However, different sequencing experiments may also reveal biological differences that cannot be eliminated by batch effect correction (Supplement Figure 4). Unlike the reference dataset sequenced by Chromium 10X 3' v3, the query dataset Wilk [26] was sequenced by Seq-well [27]. nFeature_RNA and nCount_RNA of this dataset were much smaller than those of the reference dataset. Expression levels vary greatly between reference and test datasets. In this case, the distinct batch effect affected the integration of data between the reference and test

datasets. Moreover, since too few genes were detected in cell samples, the number of genes that could be analyzed by the CNN model was limited, which would affect the accuracy of the prediction results. Even so, the annotation accuracy of scDeepInsight on this test dataset could still reach 80.2%, achieving an accuracy improvement of more than 9% relative to the other five methods used in the benchmarking process.

The choice of CNN models

CNN completes the extraction of image features through convolution and pooling operations, and processes extracted feature information in the fully connected layer to achieve image classification. According to the difference in network structure and connection mode, different CNN models have been developed. scDeepInsight has good support for multiple CNN models such as ResNet [28] and DenseNet [29]. Considering the comparison between EfficientNet-b3 [22] and other image networks, the efficiency and accuracy are relatively high.

Conclusions and future direction

In this paper, we proposed scDeepInsight, which integrates the whole operation of data preprocessing, image conversion and constructing the CNN model for cell-type prediction. The prediction results correctly reflected the number of clusters and further enabled outlier detection for unknown types. First, we selected a PBMC dataset containing trusted cell-type labels as a reference. By performing standard tests on five different independent test sets, it was shown that scDeepInsight had significantly higher accuracy and could identify more cell subtypes than the other five mainstream cell labeling methods. In addition, we addressed classical problems in single-cell annotation, such as batch effect correction methods and the detection of rare/unknown cell-types.

Finally, by applying DeepFeature to extract marker genes of cell-types, the accuracy of scDeepInsight in feature extraction was further proved in biological significance.

In the future, we will integrate more reference datasets obtained from different tissues to construct a more comprehensive cell-type classification model. Also, we will try to integrate protein and spatial chromatin accessibility information to further improve the accuracy of cell-type identification.

Data Availability

All datasets used in this paper are publicly available. The reference PBMC dataset [23] (GSE164378) can be obtained from [https://atlas.fredhutch.org/nygc/multimodal-pbmc/]. Dataset 10x-Multiome-Pbmc10k is provided by scglue [11] and can be downloaded directly from [https://scglue.readthedocs.io/en/latest/data.html]. Query datasets Yazar [30] (GSE196830), Schulte-Schrepping [31] (EGAS00001004571), Arunachalam (GSE155673) and Lee [32] (GSE149689) are all available from the CELLxGENE database: [https://cellxgene.cziscience.com/datasets].

Code Availability

We also provide pretrained annotation models to save users time. The entire code base, including the implementation of the proposed scDeepInsight pipeline, training and testing processes in benchmarking, and the pretrained models, is available at: [https://github.com/shangruJia/scDeepInsight]. Gene IDs and barcodes of the cell samples used in this paper are recorded in Supplement File 2, which can also be downloaded from this repository.

Acknowledgements

The results shown in this paper are in part based upon PBMC dataset generated by New York Genome Center (NYGC). We appreciate Gene Expression Omnibus (GEO) and European Genome-phenome Archive (EGA) providing access to the query datasets.

Author contributions

SJ implemented the whole pipeline, evaluated the performance, and wrote the first draft and contributed in the subsequent versions of the manuscript. AL advised for the model and the evaluation, contributed in the manuscript writeups. KAB checked the model, and helped in the manuscript writeup. AS perceived, supervised, and contributed in the manuscript writeups. TT perceived, supervised, and contributed in the manuscript writeups. All authors read and approved the manuscript.

Funding

This work was funded by JSPS KAKENHI Grant Number JP20H03240, Japan and JST CREST Grant Number JPMJCR2231, Japan.

References

1. Clarke ZA, Andrews TS, Atif J et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods, Nat Protoc 2021;16:2749-2764.
2. Yang F, Wang W, Wang F et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data, Nature Machine Intelligence 2022;4:852-866.

3. Hicks SC, Liu R, Ni Y et al. mbkmeans: Fast clustering for single cell data using mini-batch k-means, PLoS Comput Biol 2021;17:e1008625.

4. Kiselev VY, Kirschner K, Schaub MT et al. SC3: consensus clustering of single-cell RNA-seq data, Nat Methods 2017;14:483-486.

5. Waltman L, van Eck NJ. A smart local moving algorithm for large-scale modularity-based community detection, The European Physical Journal B 2013;86:471.

6. Franzen O, Gan LM, Bjorkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data, Database (Oxford) 2019;2019.

7. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data, Nat Commun 2022;13:1246.

8. Hu C, Li T, Xu Y et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data, Nucleic Acids Res 2023;51:D870-D876.

9. Ranjan B, Sun W, Park J et al. DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data, Nature Communications 2021;12:5849.

10. Aran D, Looney AP, Liu L et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage, Nat Immunol 2019;20:163-172.

11. Cao ZJ, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding, Nat Biotechnol 2022;40:1458-1466.

12. Song T, Dai H, Wang S et al. TransCluster: A Cell-Type Identification Method for single-cell RNA-Seq data using deep learning based on transformer, Front Genet 2022;13:1038919.

13. Sharma A, Vans E, Shigemizu D et al. DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture, Sci Rep 2019;9:11399.

14. Sultana F, Sufian A, Dutta P. Advancements in Image Classification using Convolutional Neural Network. 2019.

15. Sharma A, Lysenko A, Boroevich KA et al. DeepFeature: feature selection in nonimage data using convolutional neural network, Briefings in Bioinformatics 2021;22.

16. Zheng GX, Terry JM, Belgrader P et al. Massively parallel digital transcriptional profiling of single cells, Nat Commun 2017;8:14049.

17. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression, Genome Biol 2019;20:296.

18. Lause J, Berens P, Kobak D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data, Genome Biol 2021;22:258.

19. Butler A, Hoffman P, Smibert P et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species, Nat Biotechnol 2018;36:411-420.

20. Ye, Andre and Wang, Zian, Modern deep learning for tabular data: novel approaches to common modelling problems, Apress 1st ed. Edition, ISBN-13 978-1484286913, Dec-30, 2022.

21. Martin W et al. How to Use t-SNE Effectively. Distill. Vol. 1(10). DOI: 10.23915/distill.00002.

22. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. 2019, p. 6105-6114. PMLR.

23. Hao Y, Hao S, Andersen-Nissen E et al. Integrated analysis of multimodal single-cell data, Cell 2021;184:3573-3587 e3529.

24. Kapellos TS, Bonaguro L, Gemund I et al. Human Monocyte Subsets and Phenotypes in Major Chronic Inflammatory Diseases, Front Immunol 2019;10:2035.

25. Ancuta P, Liu KY, Misra V et al. Transcriptional profiling reveals developmental relationship and distinct biological functions of CD16+ and CD16- monocyte subsets, BMC Genomics 2009;10:403.

26. Wilk AJ, Rustagi A, Zhao NQ et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19, Nat Med 2020;26:1070-1076.

27. Gierahn TM, Wadsworth MH, 2nd, Hughes TK et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput, Nat Methods 2017;14:395-398.

28. He K, Zhang X, Ren S et al. Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015:770-778.

29. Huang G, Liu Z, Weinberger KQ. Densely Connected Convolutional Networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016:2261-2269.

30. Yazar S, Alquicira-Hernandez J, Wing K et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease, Science 2022;376:eabf3041.

31. Arunachalam PS, Wimmers F, Mok CKP et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans, Science 2020;369:1210-1220.

32. Lee JS, Park S, Jeong HW et al. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19, Sci Immunol 2020;5.