Briefings in Bioinformatics, 00(00), 2021, 1–13

https://doi.org/10.1093/bib/bbab281
Problem Solving Protocol

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021

scAdapt: virtual adversarial domain adaptation network for single cell RNA-seq data classification across platforms and species

Xiang Zhou, Hua Chai, Yuansong Zeng, Huiying Zhao and Yuedong Yang

Corresponding author: Yuedong Yang, School of Computer Science and Engineering, Key Laboratory of Machine Intelligence and Advanced Computing (MOE), Sun Yat-sen University, Guangzhou 510000, China. Tel.: +86 020-37106020; Fax.: +86 020-37106020. E-mail: yangyd25@mail.sysu.edu.cn

Abstract

In single cell analyses, cell types are conventionally identified based on expressions of known marker genes, whose identifications are time-consuming and irreproducible. To solve this issue, many supervised approaches have been developed to identify cell types based on the rapid accumulation of public datasets. However, these approaches are sensitive to batch effects or biological variations since the data distributions are different in cross-platforms or species predictions. In this study, we developed scAdapt, a virtual adversarial domain adaptation network, to transfer cell labels between datasets with batch effects. scAdapt used both the labeled source and unlabeled target data to train an enhanced classifier and aligned the labeled source centroids and pseudo-labeled target centroids to generate a joint embedding. The scAdapt was demonstrated to outperform existing methods for classification in simulated, cross-platforms, cross-species, spatial transcriptomic and COVID-19 immune datasets. Further quantitative evaluations and visualizations for the aligned embeddings confirm the superiority in cell mixing and the ability to preserve discriminative cluster structure present in the original datasets.

Key words: single cell classification; batch correction; batch effects; virtual adversarial training; single cell RNA-seq; spatial transcriptomic

Introduction

Single-cell RNA-seq technologies have been successfully employed to generate high-resolution cell atlas and to improve our understanding of cellular heterogeneity in human diseases, and one major step of single-cell RNA sequencing (scRNA-seq) analyses is cell-type identification [1]. Typically, cells are first grouped into different clusters, and each cell cluster will be manually assigned to one label based on the uniquely high expression levels of canonical makers. Nevertheless, visual inspection of cluster-specific gene is labor intensive in practice and irreproducible, and the assignments of cell types require expert knowledge of canonical makers [2]. Thus, it is necessary to develop automated computational methods for cell annotations.

A growing list of classification methods have been developed to annotate cells based on public data of known cell types [3]. The most typical methods are similarity-based methods that assign cell labels through scanning reference cell databases for similar cells. For example, SingleR [4] and CHETAH [5] used Spearman correlation for similarity measurement. The scmap [6] combined three metrics, cosine distance, Pearson correlation and Spearman correlation, to quantify the closeness between query cell and the centroid of each reference cell cluster. Though

Xiang Zhou is a PhD student in the School of Computer Science and Engineering at the Sun Yat-sen University.
Hua Chai is a postdoctoral fellow in the School of Computer Science and Engineering at the Sun Yat-sen University.
Yuansong Zeng is a PhD student in the School of Computer Science and Engineering at the Sun Yat-sen University.
Huiying Zhao is an Associate Research Fellow in the Sun Yat-sen Memorial Hospital at the Sun Yat-sen University.
Yuedong Yang is a Professor in the School of Computer Science and Engineering and the National Super Computer Center at Guangzhou, Sun Yat-sen University.
Submitted: 16 April 2021; **Received (in revised form):** 29 June 2021

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

these methods are robust, their linear metrics cannot reflect the complex nonlinear relations between genes. To overcome this issue, machine-learning-based methods were proposed to train models on the reference dataset and to use the trained models for predicting cell labels in target datasets. For example, scPred [7] took advantage of singular value decomposition to obtain small number of informative features for training a support vector machine (SVM) model. In singleCellNet [8], the reference data were pair-transformed into binary matrix that was then used to train a Random Forest classifier. Seurat [9] identified anchoring cell pairs by projecting query cells onto precomputed reference principal component analysis (PCA) structure and used these anchors to train a weighted vote classifier for cell annotation. However, these methods were applicable to test sets following the same distribution as the training set [10] and thus did not always work well due to batch effects or biological factors (e.g. treatments, individuals, species) difference between datasets.

To solve the distribution mismatch between samples, many methods have been proposed to align cell distribution, such as fastMNN [11], Harmony [12] and LIGER [13]. However, most of these methods did not support label predictions, while other cell annotation tools such as singleCellNet and SingleR running on the aligned data did not show much improvement due to the transformation of gene expressions according to previous benchmark analyses [3, 14]. The only method to support a joint batch effect removal and cell annotation was Seurat [9]; when the batch difference is obvious, Seurat provides the option to learn an aligned subspace across datasets using canonical correlation analysis (CCA) instead of PCA and identify anchors for cell classification. Although Seurat took batch difference into consideration with improvement in practice, the method suffers from two limitations. First, the integration is unsupervised without effectively using the cell-type information in reference data for aligning cell clusters, and thus, it may mismatch cells of different cell types across datasets. Second, since it optimizes distribution-alignment and label projection independently, the features for sample alignments are not necessarily optimal for cell classification. Therefore, it should be beneficial to combine sample alignment and cell annotation in one step.

In fact, the alignment of different scRNA-seq data can be considered as a typical domain adaption task in the computer vision community, with each batch as a domain [15]. The recently developed domain adaptation network could perform domain alignment and classifier training jointly, which has been shown to enhance the generalization of the classification model [16]. A typical domain adaptation framework is to reduce the distribution mismatch of the latent feature via domain adversarial learning [10]. On the other hand, unlabeled data (target single cell dataset) could be used to better estimate the decision boundary between the different classes, and thus improves the classifier's accuracy through the semi-supervised learning (SSL) [17]. More importantly, SSL can also be used to alleviate the adverse impact of domain discrepancy by jointly training the classifier on labeled source and unlabeled target data [18]. As a mainstream SSL method, the virtual adversarial training (VAT) technique used adversarial examples generated from labeled and unlabeled data to make the classifier robust against local perturbations or noises, which has been proven to be effective in many SSL tasks [19].

In this study, we have developed a new method (scAdapt) to make use of both labeled source and unlabeled target data for accurate cell classification by combining the domain adaptation and VAT-based semi-supervised learning. The domain adaptation network includes not only the adversary-based global distribution alignment but also category-level alignment [20] to preserve the discriminative structures of cell clusters in low dimensional feature (i.e. embedding) space. We demonstrate that scAdapt consistently outperforms existing methods for classification and batch correction in simulated, cross-platforms, cross-species, spatial transcriptomic and COVID-19 immune datasets. Further quantitative evaluations and visualizations for the aligned embeddings confirm the superiority in cell mixing and the ability to preserve discriminative cluster structure present in the original datasets.

Materials and Methods

Datasets and preprocessing

Simulated data

We used the *R* package Splatter [21] to generate simulated scRNA-seq counts data of different batches with similar cell-type compositions. We simulated two batches with 2000 and 1000 cells considered as source and target dataset, respectively, and each cell has 10 000 genes. Each batch was uniformly split into four cell groups with cell proportion set to 0.25 by the parameter *group.prob*. To simulate datasets with different magnitudes of batch effects, we adjusted the batch parameter *batch.facLoc* and *batch.facScale* with increasing values {0.2, 0.4, 0.6, 0.8, 1.0} where larger values corresponding to stronger batch effects. For brevity, we set *batch.facLoc* = *batch.facScale*. To simulate datasets with different magnitudes of clustering difficulty, we set the parameter de.fracScale to 0.2 for simulated datasets with weak clustering signal and 0.3 for simulated datasets with strong clustering signal. Simulation was run five times with different random seeds and average results were reported. For other parameters, default values were used unless otherwise specified.

Cross-platforms datasets

The human Peripheral Blood Mononuclear Cells (PBMC) scRNA-seq data were retrieved directly from the SeuratData package with dataset name 'pbmcsca' [22]. The data consist of seven batches from seven different sequencing platforms. We removed the cells annotated as 'Unknown' and the resulting datasets contains a total of 30 975 cells and each cell has 33 694 genes. We combined the data from the 10× Chromium (v2) and 10× Chromium (v3) platform as source data and the rest five platforms: CEL-Seq2 (CL), Drop-seq (DR), inDrop (iD), Smart-seq2 (SM2) and Seq-Well (SW) as target data. As a result, we have five pairs of cross-platform datasets: 10×-CL, 10×-DR, 10×-iD, 10×-SM2 and 10×-SW. For all the datasets, raw counts were extracted from the Seurat object for further processing.

Cross-species datasets

The human and mouse pancreas data were downloaded from SingleCellNet GitHub page where five ready-to-use datasets are provided. For data batch generated by Baron, Segerstolpe and Tabula Muris (TM) cell atlas, raw counts are provided for further processing. For datasets from Murano and Xin, normcounts are provided. Following the filtering step in previous benchmark study [23], we removed the cells labeled as 'unclear', 'co-expression', 'unclassified', 'unclassified endocrine', 'alpha.contaminated', 'beta.contaminated', 'delta.contaminated' or 'gamma.contaminated', and merged 'activated_stellate', 'PSC' and 'quiescent_stellate' cells into 'stellate'. The resulting datasets contain a total of 17 574 cells.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021

To obtain compatible gene names for cross-species analysis, we used the homologous genes provided by SingleCellNet to convert gene names and only the intersection gene set between the human data and mouse data were kept. To construct a large source with enough training samples and cover more cell types in the source, we combined the mouse data from the Baron and TM as source data.

Spatial transcriptomic datasets

We downloaded two mouse brain (hypothalamic preoptic region) datasets from Gene Expression Omnibus (GSE113576) and Dryad repositories, respectively [24]. The spatial transcriptomic dataset has 64 373 cells measured with spatially resolved multiplexed error robust fluorescence in situ hybridization (MERFISH) and the scRNA-seq dataset has 30 370 cells measured by 10× Chromium; 10× data have full transcriptome with 22 067 genes, while MERFISH data have only 154 targeted genes. We combined the two datasets with the 154 intersecting genes.

COVID-19 datasets

The COVID-19 immune atlas was downloaded from GEO (GSE158055), which was generated by Single Cell Consortium for COVID-19 in China [25]. It has 1 462 702 single cells in the lung and peripheral blood. We used this dataset as source dataset. The COVID-19 Immunodeficiency PBMC dataset was download from https://www.covid19cellatlas.org/ and has 56 840 cells. We used it as target dataset.

Preprocessing

Seurat R package (version 3.2.0) was used for preprocessing. For both the simulated and real datasets (except for Murano, Xin and MERFISH where counts are already normalized), the counts matrix were normalized by the NormalizeData function in Seurat with default 'LogNormalize' normalization method and a scale factor of 10 000. Top 2000 highly variable genes were selected based on the log-normalized counts using the FindVariableFeatures function with default 'vst' method. For real datasets, the cell-type annotations from the corresponding publications were considered as the ground truth for evaluations. Because the brain data have preselected markers, we did not select variable gene but used all the 154 intersecting genes.

The datasets analyzed in this study are summarized in Table S1, see Supplementary Data available online at http://bib.oxfordjournals.org/.

The architecture of scAdapt

Our scAdapt model includes two modules. The classification module, based on cross-entropy loss and VAT loss, aims to improve the accuracy of cell annotation using both labeled source and unlabeled target data. The batch correction module contains two loss. The adversarial domain adaptation loss aims to reduce distribution discrepancy at embedding space of source and target, while the semantic alignment loss can make the embeddings better clustered and more separable. We optimized these two modules jointly in order to improve domain alignment and final classification simultaneously.

The overall structure of scAdapt is illustrated in Figure 1. It consists of a feature extractor G with two hidden layers, a domain classifier D with two hidden layers and a label predictor F with a linear output layer followed by a softmax operation. The input includes source gene expression matrix $X_s = [x_1^s, \ldots, x_{m_s}^s] \in R^{m_s \times n}$ of m_s labeled cells with $Y_s = \{y_i^s\}_{i=1}^{m_s}$ being the corresponding labels and target gene expression matrix $X_t = [x_1^t, \ldots, x_{m_t}^t] \in R^{m_t \times n}$ of m_t unlabeled cells, where n is the number of common genes shared by the source and target data. In domain adaptation setting, X_s and X_t are assumed to be different but related [16].

To minimize the source sample classification error with known labels, standard cross-entropy loss is used as below

$$L_{CE} = -\frac{1}{m_s}\sum_{i=1}^{m_s}\sum_{c=1}^{K} y_{i,c}^s \log\left(p_{i,c}\right), \tag{1}$$

where $y_{i,c}^s$ is a binary indicator (0 or 1) if the cell label c is the correct label for cell i, $p_{i,c}$ is the predicted probability of the cell i belonging to cth cell label and K is the number of class.

We used VAT to incorporate the information of data distribution from unlabeled data, which can better estimate the decision boundary between different classes [17]. VAT is an effective data augmentation technique which does not need prior label information and is hence applicable to semi-supervised learning. It assigns similar labels to each input data and its neighbors in the adversarial direction where the perturbation will alter the model's output distribution the most. Then, the model is robust to small perturbations or noises in the inputs. The loss function of VAT is given by

$$\begin{aligned} L_{VAT}\left(X_t, \theta\right) &= D_{KL}\left[p\left(Y_t|X_t, \theta\right), p\left(Y_t|X_t + r_{vat}, \theta\right)\right] - 2\|F\left(G\left(X_t\right)\right)\|_*, \\ \text{where} r_{vat} &= \underset{r,\|r\|_2 \le \epsilon}{\arg\max} D_{KL}[p(Y_t|X_t, \theta), p(Y_t|X_t + r)], \end{aligned} \tag{2}$$

where r_{vat} denotes the virtual adversarial perturbation maximizing the difference between the model output of perturbed input and nonperturbed input, and θ is the model parameter to train. The output distribution is parameterized as $p(Y_t|X_t, \theta)$, and $D_{KL}[\cdot, \cdot]$ is KullbackLeibler divergence that measures the difference between two probability distributions. The last penalty term $\|F(G(X_t))\|_*$ in (2) is designed to improve both the prediction discriminability and diversity, and $\|\cdot\|_*$ is the nuclear-norm [18].

To learn the domain-invariant features, adversarial adaptation loss is adopted, where the feature extractor G and domain classifier D are trained by playing a two-player minimax game: the first player is domain classifier which distinguishes whether the feature is from the source domain or target domain, and the second player is feature extractor which aims to output domain-invariant features to confuse the domain classifier. Domain alignment is expected when the game reaches an equilibrium. Formally, the domain classifier D is trained by minimizing the binary cross-entropy loss

$$L_{DA} = -\frac{1}{m_s}\sum_{i=1}^{m_s}\log\left(D\left(G\left(x_i^s\right)\right)\right) - \frac{1}{m_t}\sum_{j=1}^{m_t}\log\left(1 - D\left(G\left(x_j^t\right)\right)\right), \tag{3}$$

while the feature extractor G is trained to maximize the L_{DA} loss (fool the domain classifier D). In order to update the parameters of D and G simultaneously, gradient reverse is used to flip the sign of the gradient between D and G during backpropagation [10].

Besides the global domain-invariance, discriminability must also be preserved, which ensures the embeddings of same class, but different domains are mapped nearby. An intuitive solution is to perform semantic alignment for samples of each class

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021

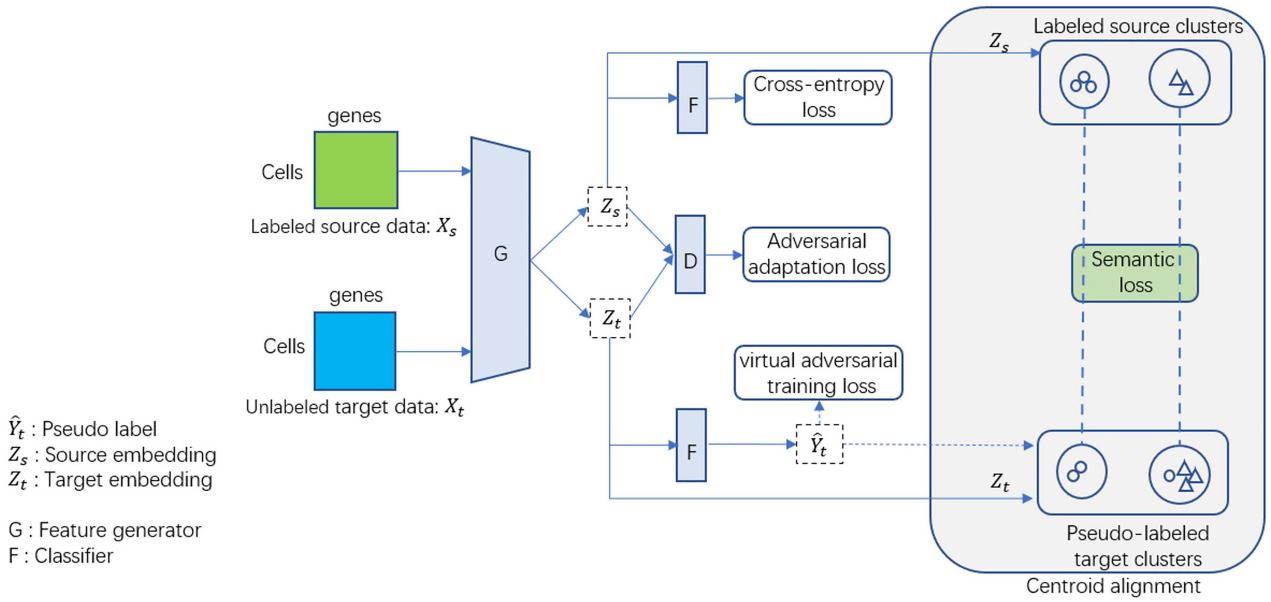

Figure 1. The overall structure of scAdapt for cell-type classification and batch correction on source and target dataset. With source and target data as input, the feature extractor G learns to capture low-dimensional embedding Z_s and Z_t which are then used to train the label predictor F with the cross-entropy loss and VAT loss, respectively. At the embedding space, batch correction is achieved at global- and class-level: adversarial domain adaptation loss is employed to perform global distribution alignment and semantic alignment loss minimizes the distance between the labeled source centroid and pseudo-labeled target centroid. The target pseudo label is estimated by label predictor F.

directly. However, the explicit alignment for each class is impossible since no label information provided for target domain. We approached the problem by assigning pseudo labels to target samples with the classifier F and then explicitly align the centroid for each class in source and target domain [20]. The centroid is defined as the mean embedding of each class. For each target class, all samples with correct or wrong pseudo labels are used for centroid calculation, and thus the noise or bias brought by partial false pseudo labels are expected to be suppressed by correct pseudo labels with a dominating portion. We formulate the following semantic alignment loss to minimize the distance between the target centroids and their corresponding source centroids:

$$L_{\mathrm{SM}}\left(\mathrm{X_s},\mathrm{X_t}\right)=\frac{1}{K}\sum_{k=1}^{K}\left\|C_s^k-C_t^k\right\|_2^2+\frac{1}{m_s}\sum_{i=1}^{m_s}\left\|G\left(\mathrm{x}_i^s\right)-C_s^k\right\|_2^2, \quad (4)$$

where C_s^k and C_t^k denote the source and target centroids, respectively. The second term of (4) is designed to penalize big intra-class distances and enforce better cluster compactness [26].

The overall loss function can be formulated as

$$L_{\mathrm{overall}}=L_{\mathrm{CE}}+\lambda_0 L_{\mathrm{VAT}}+\lambda_1 L_{\mathrm{DA}}+\lambda_2 L_{\mathrm{SM}}, \quad (5)$$

where λ_0, λ_1 and λ_2 are the regularization coefficients controlling the contribution of VAT, global domain alignment and semantic alignment to the total loss function, respectively.

Identifying cell-type important genes

Since neural networks are often considered as black box models with no clear interpretation, examining the importance of each gene relative to classification output is favorable for understanding the reason behind classification decisions. We identified key genes for each cell type by activation maximization method [27]. Formally, let θ be the fixed model parameters after training the network, and $h_i(\theta, x)$ be the activation of ith neuron in the last layer of neural network with input x, i.e. the classification score for cell type i. Activation maximization looks for input patterns which maximize the classification score

$$x^*=\arg\max_x h_i(\theta,x). \quad (6)$$

A locally optimal solution of (6) can be found through gradient ascent in the input space, where the gradient of $h_i(\theta, x)$ with respect to x was computed to iteratively update the inputx. It should be noted that the optimization was performed with respect to the input x, which is different from the training procedure of neural network for optimizing the model parameters θ. The inputx was initialized with a zero vector and updated for 100 iterations with learning rate set to 1. The changes of resulting x^* compared with the initialization values were calculated as the gene importance score. To evaluate whether the identified top-important genes are reliable, we selected the top genes with the largest importance score for each cell type and compared them with cell-type markers in the PanglaoDB database [28] and the marker gene reported in original publication [24]. We also performed Gene Ontology (GO) enrichment analysis on these selected genes, using the R package clusterProfiler [29].

Hyper-parameters setting

All the neural network layers are fully connected. The two hidden layers of feature extractor G have 512 and 256 nodes, respectively. For spatial transcriptomic data with only 154 genes, we set the nodes in each hidden layer as 128 and 128. The size of hidden layers in domain classifier D is set to 1024. Rectified linear unit function is used as activation function for the hidden layers while softmax activation function and sigmoid function applied to the last layer of F and D, respectively. The network is trained

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021

by mini-batch stochastic gradient descent with a momentum of 0.9 and a weight decay of 5×10^{-4}. We follow the same annealing strategy of learning rate as described in [10], i.e. the learning rate η_p is adjusted following $\eta_p = \eta_0/(1+ap)^b$, where p is the training progress linearly increasing from 0 to 1, η_0 is the initial learning rate set to 0.001, $a = 10$ and $b = 0.75$. The batch size is set to 256. Throughout all experiments, we set regularization parameter $\lambda_0 = 0.1\lambda$, $\lambda_1 = \lambda$ and $\lambda_2 = \lambda$, where the penalty parameter λ is updated from 0 to 1 by a progressive schedule according to Ganin *et al.* [10]

$$\lambda = \frac{2}{1+e^{-10t}} - 1, \tag{7}$$

where t is the training progress linearly increasing from 0 to 1. With this schedule, the model can first focus on training the model with labeled source data and then focus on the optimization of VAT, global domain alignment and semantic alignment whose information are noisy and inaccurate at early training stages. The method was implemented based on PyTorch [30].

Benchmarking classification methods

To evaluate the performance of scAdapt, we benchmarked it against other cell-type annotation tools, including Seurat V3, scmap, scPred, CHETAH, SingleR, singleCellNet and SVM. For Seurat V3, we used both the CCA-based and PCA-based label propagation to evaluate whether the classification can benefit from aligned data. The default hyperparameters recommended in these annotation tools and accompanying tutorials were used for performance evaluation.

Evaluation metric

We evaluated the classification performance of each method using the accuracy score, which is defined as the proportion of correctly annotated cells. We computed the accuracy for each class in the test data and reported averaged accuracy across all the classes. Throughout the evaluation, the previously published cell type annotations provided by original datasets were considered as ground truth.

Benchmarking batch correction methods

Seurat V3, fastMNN, Harmony and LIGER were used as competing methods. We evaluated the performance by quantitative measure and visual inspection. Silhouette score and divergence score were used to measure the quality of batch correction [31]. An accurate batch correction method should result in a high silhouette score (preserving the original structure of the data) and low divergence score (keeping the same-type cells across batches well mixed). Uniform manifold approximation and projection (UMAP) was used for visualizing cells in a two-dimensional space [32]. During the benchmark, all competing methods were run with their default hyperparameters, or the hyperparameters provided in the accompanying tutorials.

Evaluation metric

We used divergence score to quantify how well the same population between different batches is mixed after batch correction. A smaller divergence score means better mixing of the same cell population. The quality of mixing is estimated by the universal k-nearest-neighbor (kNN) divergence [33]. The kNN divergence between UMAP embeddings $Z_{s,l}$ and $Z_{t,l}$of class l can be formulated as

$$D_{\mathrm{kNN}}\left(Z_{s,l} \| Z_{t,l}\right) = \frac{d}{n}\sum_{i=1}^{n} \log \frac{v(i)}{\rho(i)} + \log \frac{m}{n-1}, \tag{8}$$

where d is dimension size of embeddings, m and n are the number of source and target samples in class l, respectively, $\rho(i)$ is the Euclidean distance between sample i and its kNNs in the same batch, and $v(i)$ is the distance from sample i to its kNNs in the other batch. The average kNN divergence over all classes is calculated as divergence score. In all experiments for batch-correction evaluation, k was set to 30 for kNN computation.

Evaluation only by divergence score is not sufficient, since we can obtain a perfect score by randomly mixing the data regardless of the cell type. Thus, we used silhouette score to quantify how well different cell types are separated after batch correction and ensure that datasets integration can conserve true biological signals in original datasets. Let $a(i)$denote the average distance between cell i to all other cells in the same cluster and $b(i)$ be the average distance between the cell i and cells in the next closest cluster. The distance is calculated using Euclidean distance based on the UMAP embeddings of the batch-corrected data. The silhouette coefficient of cell i can be formulated as

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(j)\}} \in [-1, 1], i = 1, \ldots, n. \tag{9}$$

The average silhouette coefficient across all cells can be calculated as silhouette score. A higher silhouette score indicates better cell-type assignment.

Results

To showcase the strength of scAdapt, we analyzed simulated dataset generated by '*Splatter*' and real scRNA-seq datasets with different sequencing platforms and species, and of spatial transcriptomics. The scAdapt was compared with seven cell-type classification methods and four batch correction methods, and scAdapt was shown to consistently outperform these existing methods in both cell-type annotation and batch correction.

Performance on the simulated dataset

We first evaluated the classification accuracy under different degrees of batch effects with weak clustering signal strength (*de.facScale* = 0.2). As shown in Figure 2a, the accuracies of all classification methods decrease with increasing *batch.fracScale*, confirming the classification challenges brought by the batch effects. The scAdapt always outperforms the competing methods across all batch effects settings: the accuracy of scAdapt is ∼0.98 until the *batch.fracScale* value increases to 0.6 and remains >0.94 at *batch.fracScale* = 1.0. The superior performances show that scAdapt can effectively reduce performance degradation brought by batch difference, even under high batch effects. By comparison, the accuracy of competing methods drop dramatically as the *batch.fracScale* increases. Although SingleR achieves the second highest performance when *batch.fracScale* ≤ 0.6, its accuracy demonstrates a sharp drop from 0.85 to 0.61 when *batch.fracScale* changes from 0.6 to 1.0. At the largest batch.fracScale, Seurat-PCA has a relatively low accuracy of 0.39 which is improved to 0.69 by Seurat-CCA

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021

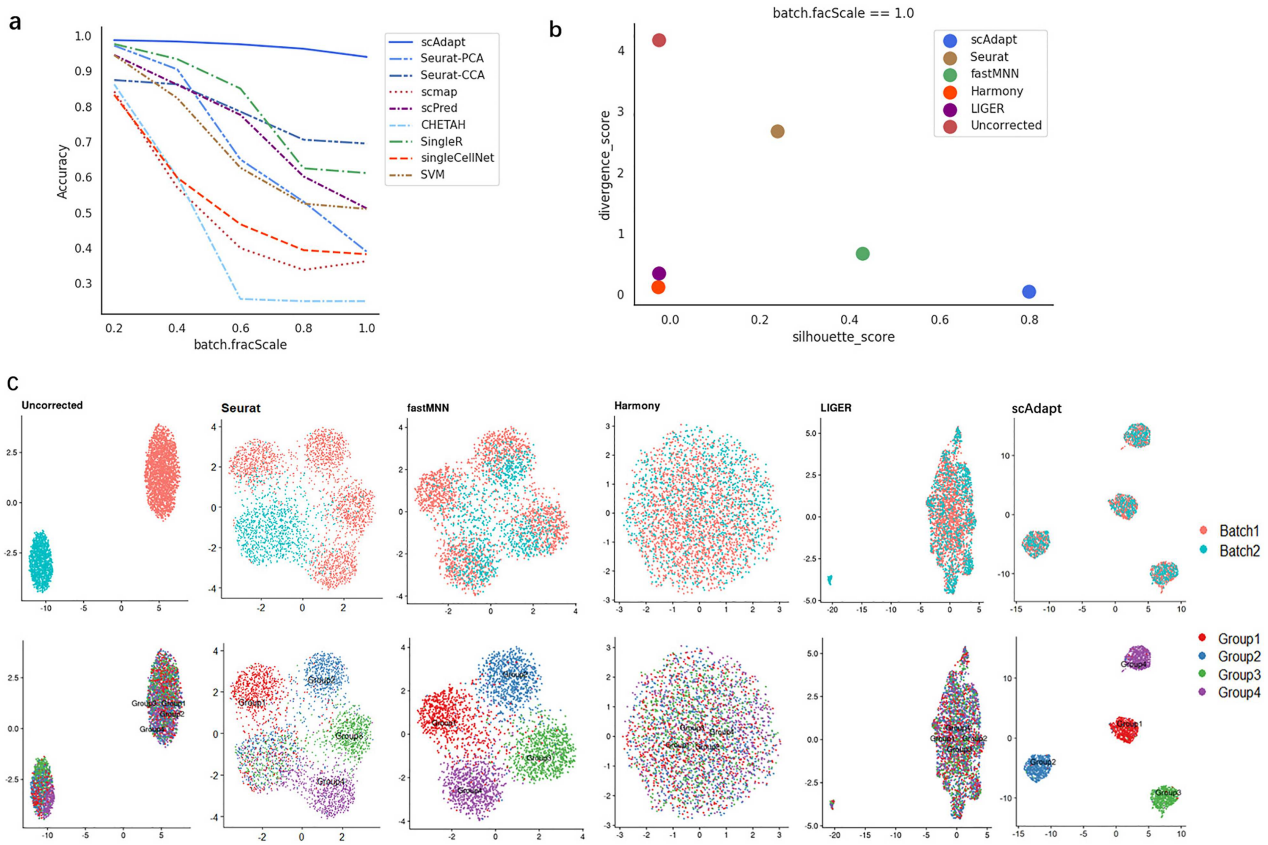

Figure 2. Benchmarking of scAdapt against seven classification methods and four batch correction methods on simulated data with two batches and four different cell types. (**a**) Average accuracy under increasing *batch.fracScale* values where larger values corresponding to stronger batch effects. (**b**) The integration quality measured by divergence score versus silhouette score at *batch.fracScale* = 1.0. Specifically, a lower divergence score means better cell mixing across datasets and a higher silhouette score indicates better cell type assignment. (**c**) UMAP plots colored by batch and cell type at *batch.fracScale* = 1.0.

(78% improvement), confirming that batch-correction in Seurat-CCA can enhance the classification model under a high batch difference.

Next, we evaluated the performance of scAdapt and other four batch correction methods using the most challenging simulation setting with *batch.fracScale* = 1.0 (Figure 2b and c). Ideally, there should be four distinct cell groups (each representing a cell type) in the UMAP visualization after the batch correction, and the cells from both batches are well mixed in each group. Visualization of the uncorrected data shows that the cells are distinctly grouped by batch, resulting in the lowest silhouette score (−0.026) and the highest divergence score (4.18). After removing batch effects by scAdapt, the batch distinctions are effectively removed with the same-type cells across batches uniformly mixed while maintaining the cell-type structure in original batches. scAdapt achieves not only the highest silhouette score (0.80) but also presents the lowest divergence score (0.05) over others. Although Harmony and LIGER also have low divergence score (≈ 0.10), their silhouette scores are much lower (≈ 0) due to over-correction problem with all cell types mixed together. The fastMNN method, although produces a proper balance of batch mixing and cell type mixing with a divergence score of 0.75 and a silhouette score of 0.43, suffers from under-correction where the cell types across batches are not well aligned despite relatively clear separations between cell types. Seurat produces the highest divergence score (2.5) and a low silhouette score (0.24) since it fails to perform cell-type alignment and the cell types in the 2nd batch are far less discernable. These results suggest that scAdapt improves batch correction relative to unsupervised methods that ignore label information of source data.

To demonstrate the separate contributions of different components in scAdapt, we performed ablation studies with *batch.fracScale* = 1.0 for three variants of scAdapt: Baseline (removing both VAT and DA module from scAdapt) or removing either module. As shown in Figure S1a, see Supplementary Data available online at http://bib.oxfordjournals.org/, the addition of VAT (Baseline+VAT) can notably improve the classification accuracy relative to the baseline, while the addition of DA (Baseline+DA) is beneficial for batch correction. The full model by combing both components can further enhance the performance of batch correction through guiding cell-type alignment with more accurate pseudo labels. We also visualized the embeddings of scAdapt and its three variants through UMAP in Figure S1b,see Supplementary Data available online at http://bib.oxfordjournals.org/. For the Baseline model, two batches are completely mismatched and the four clusters in the target batch essentially overlap with each other, making it hard to separate the cells. By using the auxiliary information from the target sample, Baseline+VAT can separate the cell types well but ignores alignment across batches. On the other hand, Baseline+DA aligns the cell types correctly, but the group 1 and group 2 are not well separated. By comparison, scAdapt can well align the same cell types across batches and discriminate

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021

different cell types, confirming the necessity of combining VAT and DA for batch correction.

We also evaluated the performance on datasets with strong clustering signal strength (*de.facScale* = 0.3) where the uncorrected data have low cell type noise and distinguishable cluster structure in each batch (Figure S2, see Supplementary Data available online at http://bib.oxfordjournals.org/). As expected, all methods (except LIGER) show improved performance in the easy dataset, while scAdapt still outperforms the competing methods.

Performance on the cross-platform datasets

In realistic scenario, the source and target datasets are often generated from different experimental platforms in different labs. To evaluate the performance of scAdapt on this realistic setting, we conducted cross-platform test on five paired source-target PBMC datasets where the cell types from the source dataset were mapped to those in the target dataset. In this setting, each dataset is profiled by different sequencing platforms.

As shown in Figure 3a, scAdapt consistently outperforms other methods by the accuracy on the five test pairs, indicating that integrating source and target dataset can make the classification method resilient against batch effects. The detailed confusion matrices between cell classification show that scAdapt has a more balanced performance on each cell type with minimum accuracy >0.75, compared with other methods with minimum accuracy ranging from 0.31 to 0.62 (Figure S3a, see Supplementary Data available online at http://bib.oxfordjournals.org/). The performance drop of competing methods mainly come from the misclassification of closely related cell types. For example, the second ranked Seurat-CCA method incorrectly assigned 21% of Cytotoxic T cell as CD4+ T cell, and 10% of them to Natural killer cell. For completeness, we also ran CHETAH, SingleR, singleCellNet and SVM with gene expression data corrected by Seurat V3 (other methods failed to run over corrected datasets) and observed that using Seurat V3 correction can improve the performance of CHETAH, SingleR, singleCellNet and SVM by 15, 4, 3 and 4%, respectively (Figure S3b, see Supplementary Data available online at http://bib.oxfordjournals.org/). It is worth noting that the improved accuracies of CHETAH (0.83), SingleR (0.84), singleCellNet (0.8) and SVM (0.8) are still lower than that of scAdapt (0.86).

In many studies, the target data may contain novel cell types absent in the collected source datasets. Good classifier should provide low classification confidence scores for these unknown cells and high scores for known cells. To test whether scAdapt can identify these novel cell types, we trained it on the source dataset with 'Cytotoxic T cell' cells removed and tested it on the five target datasets. We performed false positive rate (FPR) control experiments to evaluate the accuracy for known and unknown cells. Here, FPR is calculated as the ratio between the number of unknown cells falsely assigned as known and the total number of unknown cells that should be rejected. We chose SingleR and Seurat-CCA for comparison since they demonstrated overall top performance in the benchmark. The evaluation of Seurat and SingleR with 'unassigned' function was following the previous study [9], which can examine their prediction/confidence scores by assigning the cells with the lowest values to be 'unassigned'. In Figure S4a, see Supplementary Data available online at http://bib.oxfordjournals.org/, at the FPR of 0.05 and 0.1, we showed that scAdapt consistently achieves the highest accuracy of 0.55 and 0.61 for known cells, respectively. The accuracies of SingleR (0.29 and 0.33) and Seurat-CCA (0.40 and 0.48) are much lower with the same FPR cutoffs, indicating the advantage of scAdapt in classification accuracy of known and unknown cell type. When the 'Cytotoxic T cell' are not removed, scAdapt achieves similar high accuracy at the FPR of 0.05 and 0.1 (Figure S4b, see Supplementary Data available online at http://bib.oxfordjournals.org/).

The quantitative performance of the five integration methods is summarized in Figure 3b. As expected, scAdapt is the top method with lowest divergence score and highest silhouette score across all dataset pairs, which is congruent with visualizations plots in Figures 3c and S5, see Supplementary Data available online at http://bib.oxfordjournals.org/. Specifically, divergence score is reduced by 17–61% on top of LIGER, fastMNN, Harmony and Seurat V3, respectively. The silhouette score is also improved substantially by 65–110% when compared with the four competing methods. Although the competing methods can separate the distinct cell types such as B cell and Megakaryocyte well, the highly similar cell types (CD4+ T cells, Cytotoxic T cells and Natural killer cells) [34] are tightly connected in their visualization plots. Since the target data are unlabeled in practice, it would be difficult to visually distinguish them as different cell types. In contrast, scAdapt is able to visibly separate three clusters, highlighting the contribution of the proposed semantic alignment loss and accurate pseudo label.

Compared with batch-corrected embedding space, batch correction in the gene expression space is more useful for downstream analysis like the identification of differentially expressed (DE) gene. To address this issue, we added a decoder layer to reconstruct batch-corrected gene expression from batch-corrected embeddings with mean squared error loss. We used the DE gene detection as a performance measure to evaluate the quality of corrected gene expression. Seurat and fastMNN that produce batch-corrected gene expression matrix were chosen for comparison. We identified DE genes between 'B cell' and all other cells using Wilcoxon Rank Sum test. The DE gene identified in uncorrected source and target data individually were considered as 'working truth' (logFC > 0.25 and adjusted *P*-value < 0.01). We compared the overlap of equal number of top-scoring batch-corrected DE genes and uncorrected (source and target) DE genes and computed true positive rate (TPR) for each comparison. From Figure S6a, see Supplementary Data available online at http://bib.oxfordjournals.org/, we found that, in both the source and target comparisons, scAdapt achieves higher TPR (0.74 and 0.68) than those in Seurat (0.71 and 0.55) and fastMNN (0.67 and 0.55) results. These results indicate that the gene expression reconstructed by scAdapt benefits from batch-corrected embeddings and accurately retain original gene expression information. Furthermore, we investigate the impact of different normalization methods on the performance (Figure S6a–c, see Supplementary Data available online at http://bib.oxfordjournals.org/). Besides LogNormalize, we included the centered log ratio (CLR) normalization from the Seurat package and the multiBatchNorm from the batchelor package [11] for comparison. The multiBatchNorm will rescale the size factors between batches to make them comparable and then perform LogNormalize. From Figure S6a–c, see Supplementary Data available online at http://bib.oxfordjournals.org/, we found that the advantage of scAdapt is consistent across the considered three normalization methods. CLR reduces the average TPR by 10.3% on top of LogNormalize and thus is not suitable for batch correction processing. multiBatchNorm does not demonstrate the performance improvement, and the potential gain may be neutralized by the batch correction methods. Thus, we recommend LogNormalize for preprocessing.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021

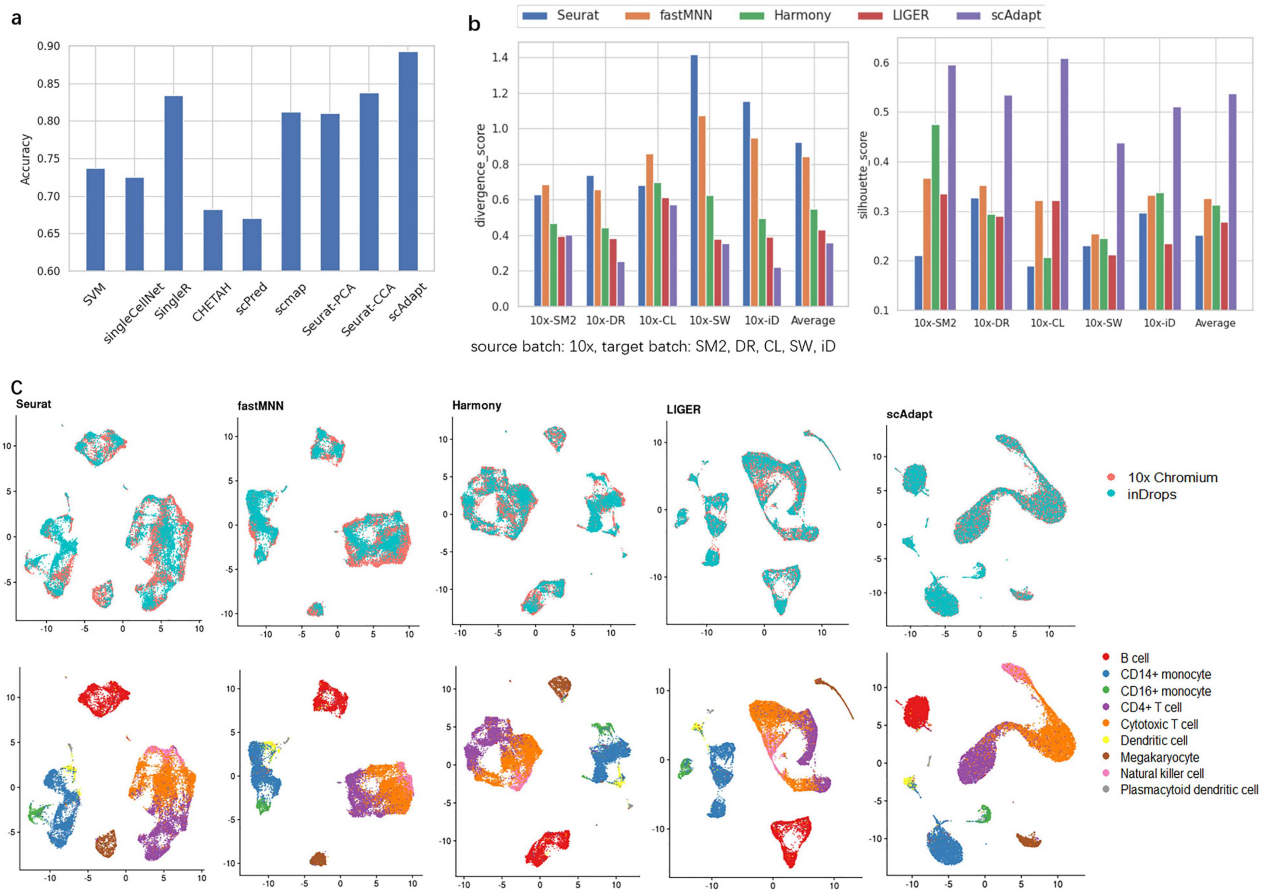

Figure 3. Comparison of classification methods and integration methods for five pairs of cross-platform human PBMC dataset. (**a**) The overall accuracy of different classification methods across the five test pairs. (**b**) Divergence score and silhouette score of different integration methods. (**c**) UMAP plots of the 10×-iD test pair colored by batch and cell type.

Performance on the cross-species datasets

Similar to the test in singleCellNet and Garnett [35], we evaluated scAdapt on cross-species datasets, which is a more challenging scenario with large distribution difference than cross-platform. The differences mainly come from two sources of variations: species and platforms. We used mouse pancreas data from the Baron and TM cell atlas as source data, and human pancreas data from the Baron, Segerstolpe, Murano and Xin as target data. Different from the slight superiority in cross-platform experiments, scAdapt achieves much higher average accuracy (0.93) than the second-ranked scmap (0.81), confirming the ability of scAdapt to deal with large difference between species (Figure 4a). Seurat-CCA, which achieves competitive performance in cross-platform test, suffers an accuracy drop of 22.5% compared with scAdapt. Other six methods are low-performing with accuracy <0.7. Further inspection of the classification by Sankey plots reveals that the competing methods cannot effectively differentiate several major cell types (Figure S7a, see Supplementary Data available online at http://bib.oxfordjournals.org/). For the alpha cell type that accounts for the highest proportion (35%) in target data, Seurat-CCA, Seurat-PCA, scPred and SVM only correctly classify 55, 26, 31 and 65% cells, respectively. For beta cell type with the second highest proportion (26%), 33–89% cells are correctly categorized by competing methods. In contrast, the accuracy of scAdapt on these two cell types is both larger than 0.95. We also performed batch corrected classification evaluations as done in the cross-platform test and found that the four competing methods can benefit from correction with an accuracy of 0.81 (CHETAH), 0.86 (SingleR), 0.78 (singleCellNet) and 0.81 (SVM) (Figure S7b, see Supplementary Data available online at http://bib.oxfordjournals.org/) but are still lower than that of scAdapt (0.93). These results suggest the effectiveness of scAdapt to overcome the variations from both species and platform.

The divergence score and silhouette score show that scAdapt is again the leading method for batch correction (Figure 4b). Similar to simulation and cross-platform scenario, fastMNN ranks the second for silhouette score, despite the relatively poor data mixing. Harmony and LIGER produce 69 and 83% lower silhouette scores than scAdapt, respectively. Visual inspection reveals that the performance degradation of Harmony and LIGER mainly come from under-correction (Figures 4c and S8, see Supplementary Data available online at http://bib.oxfordjournals.org/). For example, part of the alpha and beta cells are separated as human-specific cell types in the xin dataset, which is not consistent with the original assignments. Additionally, scAdapt can clearly separate acinar and ductal cells which come from the same progenitors and are closely associated [36], while none of the competing methods can separate them in all of the cross-species integration tests. These results suggest that scAdapt is able to maintain biological heterogeneity while effectively reducing unwanted species-specific noise.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021

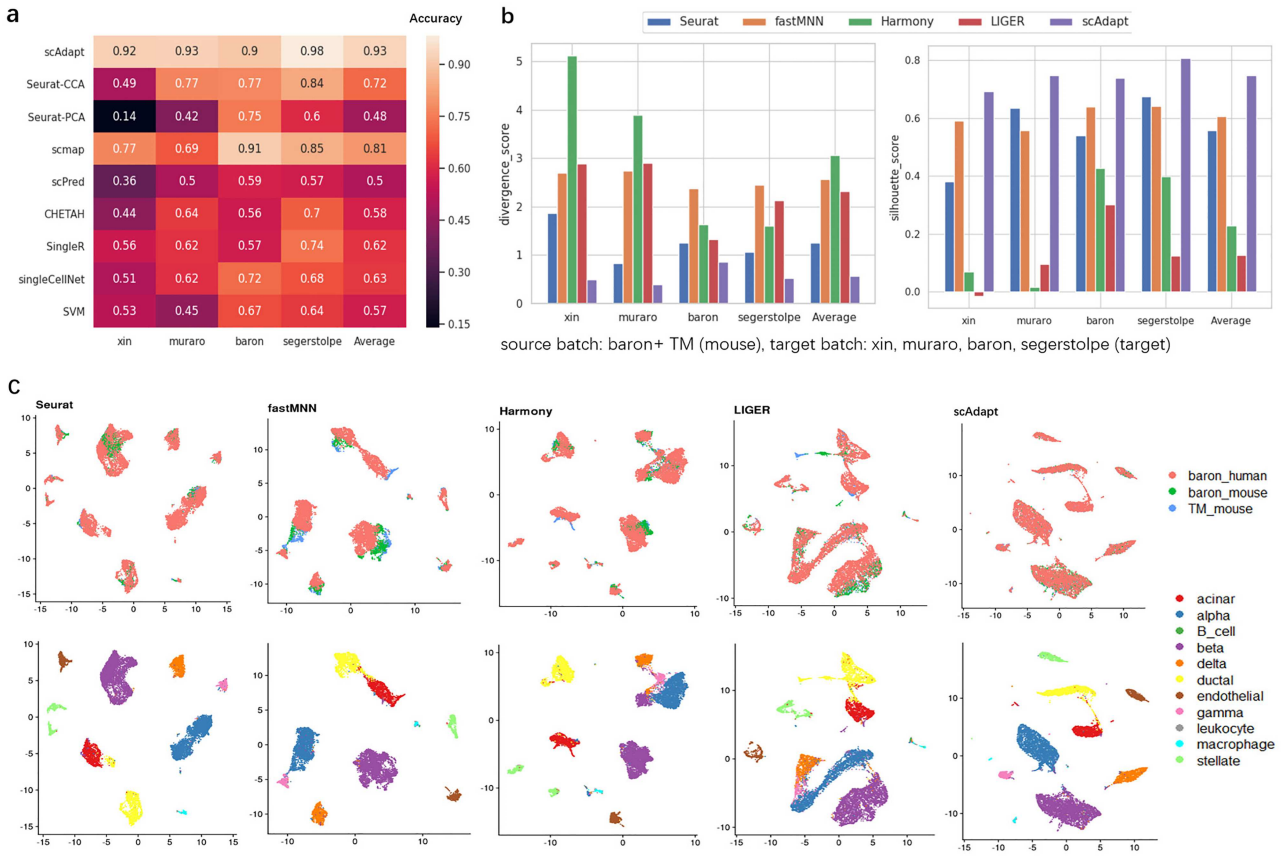

Figure 4. Comparison of classification methods and integration methods for four cross-species pancreas dataset pairs. **(a)** Heatmap showing the accuracy of different classification methods. **(b)** Divergence score and silhouette score of different integration methods. **(c)** UMAP plots of the Baron human dataset mapped to mouse source datasets (Baron and TM) colored by batch and cell type.

Due to the difference between species, the reference data may not contain all cell types in the query data. In order to assess how well scAdapt discover new or unknown cell types, we trained it on the mouse dataset with 'alpha' cells removed and tested it on the human Baron dataset. We labeled the cell 'unassigned' if its highest output probability was smaller than 0.5. We found that scAdapt achieved a high accuracy for known cell types (accuracy = 0.9), while 95% of 'alpha' cells are correctly recognized as 'unassigned'. Removing 'beta' cells obtains similar results with an accuracy of 0.95 (known) and 0.89 ('beta'). The potentially new cell types in the UAMP visualizations also occupy a distinct region and are clearly distinguishable (Figure S9, see Supplementary Data available online at http://bib.oxfordjournals.org/). These results show that scAdapt is able to identify cell types that are not in the reference dataset.

Application to the spatial transcriptomic dataset

Current scRNA-seq technologies require cell dissociation, resulting in loss of the spatial localization. Novel spatial transcriptomics methods, such as MERFISH [24], can retain spatial cell information but capture only a small number of genes that can be simultaneously measured per cell. With the limited shared information, it is challenging to find and merge similar cell types across these two data types. To assess how well scAdapt performs in this setting, we obtained two datasets profiled from the hypothalamic preoptic region of mouse brain, where we used the dissociated scRNA-seq dataset sequenced by the 10× Chromium as source and the spatial transcriptomics dataset measured with MERFISH as target. These two datasets have only 154 overlapped genes.

As shown in Figures 5a and S10, see Supplementary Data available online at http://bib.oxfordjournals.org/, scAdapt achieves an average accuracy of 0.87 over the nine cell types in target dataset, higher than the accuracies of competing classification methods ranging from 0.35 (scPred) to 0.78 (Seurat-CCA). The cell types predicted by scAdapt also demonstrate more consistent patterns in spatial distribution than competing methods in the previous report [24] (Figure S11, see Supplementary Data available online at http://bib.oxfordjournals.org/). For example, the predicted ependymal cells by scAdapt are enriched in a single layer lining the third ventricle, while this pattern is missed by Seurat-CCA since it misclassifies most of the ependymal cells as the inhibitory neurons. The batch correction results in Figure 5b and S12, see Supplementary Data available online at http://bib.oxfordjournals.org/, suggest that scAdapt successfully maps cells of the same cell types between the two datasets into a shared embedding, with lower divergence score (0.30) and higher silhouette score (0.47) than alternative approaches (divergence score: 0.33–2.50, silhouette score: 0.20–0.43).

To make our neural network model more interpretable, we used the activation maximization method to identify the most important genes for predicting certain cell type. We listed the top 10 genes with the largest importance scores by order in Table S2, see Supplementary Data available online at http://bib.oxfordjou

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021

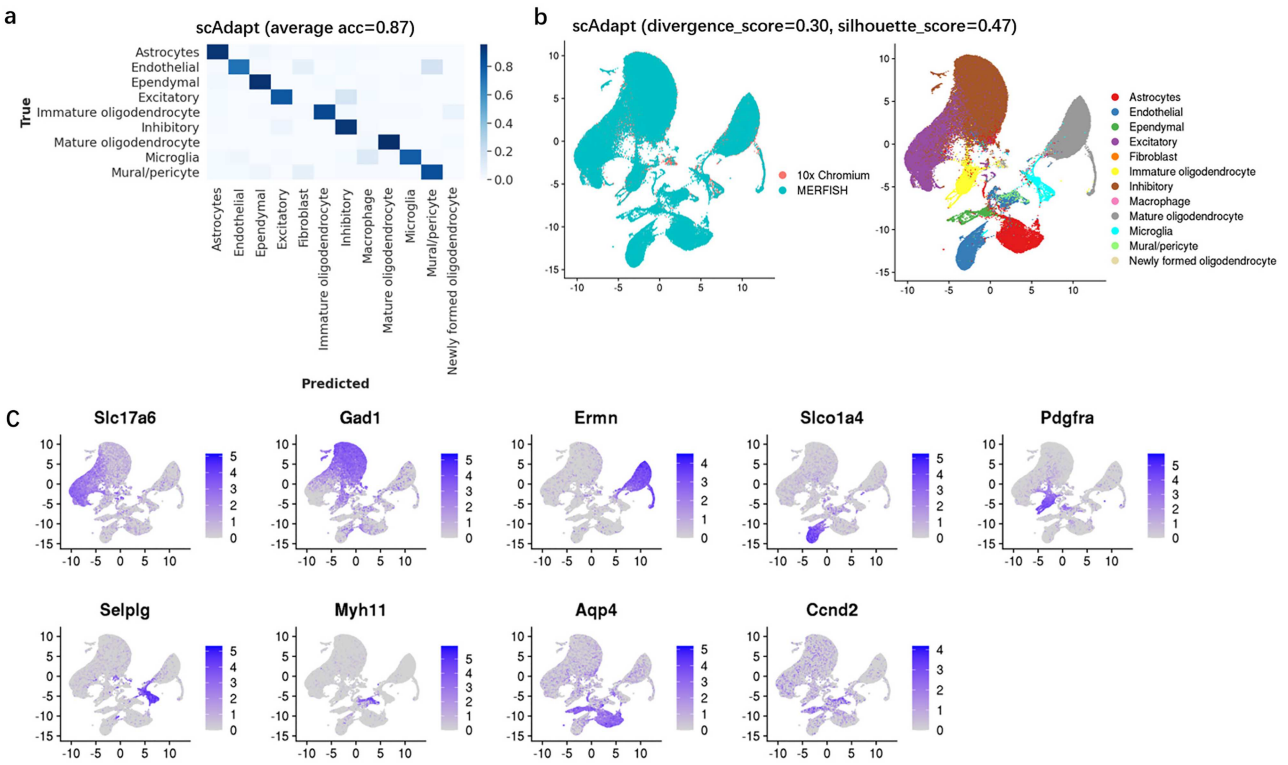

Figure 5. Predicting cell types in spatial transcriptomic dataset (MERFISH) with dissociated scRNA-seq (10× Chromium) dataset from the hypothalamic preoptic region of mouse brain. (**a**) Heatmap for the confusion matrix of our method with average accuracy in the bracket. (**b**) UMAP plots of the MERFISH dataset mapped to 10× Chromium reference by our method with divergence score and silhouette score in the bracket. Cells are colored by batch (left) and cell type (right). (**c**) Expression patterns of the top cell-type-specific marker genes identified by activation maximization in MERFISH dataset. Cells are colored based on the log-normalized expression of marker gene. The gene names are listed in the title of the panel.

rnals.org/. We found that among the nine top-1 marker genes identified by our method for each cell type, six genes (Excitatory: *Slc17a6*, Inhibitory: *Gad1*, Immature_oligodendrocyte: *Pdgfra*, Microglia: *Selplg*, Mural_pericyte: *Myh11*, Astrocytes: *Aqp4*) are the same as the ones reported in original publication [24], and the other three genes (Mature_oligodendrocyte: *Ermn*, Endothelial: *Slco1a4*, Ependymal: *Ccnd2*) are in the marker gene lists of corresponding cell type from the PanglaoDB database [28]. In Figure 5c, these genes also exhibited clear expression patterns correlated with corresponding cell types. The results of GO enrichment analysis on the top 10 genes of each cell type are presented in Figure S13, see Supplementary Data available online at http://bib.oxfordjournals.org/. We can see that the selected genes are significantly enriched on GO terms relevant to the biological processes of nervous system, such as glutamate secretion term (GO: 0014047) for excitatory cell, myelination (GO: 0042552) for immature oligodendrocyte cell. These results suggest that the identified genes are consistent with prior biology knowledge and verify the reliability and interpretability of our scAdapt model.

We also identified DE genes from the batch-corrected expression matrix to evaluate whether the batch correction can preserve the results of DE analysis run on the original datasets. We selected DE genes by performing DE analysis between inhibitory cells and all other cells using the Wilcoxon rank sum test with $\log FC > 0.25$ and adjusted P-value < 0.01. We compared the intersection of DE genes from the batch-corrected dataset and uncorrected dataset to evaluate whether the batch correction method can preserve the results of DE analysis on the original datasets. From Figure S14, see Supplementary Data available online at http://bib.oxfordjournals.org/, we found that gene expression corrected by scAdapt retain more raw DE genes than those by Seurat and fastMNN (76 versus 67 and 69). Further GO enrichment analysis shows that the DE genes detected by scAdapt are significantly enriched for GO terms relevant to the process of neural communication and development, such as neuropeptide signaling pathway and positive regulation of neuron projection development (Figure S15, see Supplementary Data available online at http://bib.oxfordjournals.org/). These results suggest that scAdapt can effectively preserve original biological features after batch correction.

Running time and memory evaluation

We tested the scalability of scAdapt on large-scale datasets. Recently, a single-cell transcriptome atlas was generated by Single Cell Consortium for COVID-19 in China, which has 1.46 million single cells in the lung and peripheral blood. We used this dataset as source dataset, and a COVID-19 Immunodeficiency PBMC dataset as target dataset. The target dataset has 56 840 cells. We downsampled the source dataset per cell type to generate different sample sizes (14 k, 140 k, 700 k and 1.46 million) and profiled the memory usage and the computing time of classification method and batch correction method. All analyses were performed on a server with Intel(R) Xeon(R) CPU E5-2650 v4 (2.20GHz), 256 GB RAM and GeForce GTX 1080Ti GPU. As shown in Table S3, see Supplementary Data available online at http://bib.oxfordjournals.org/, the computing time of scAdapt increases with respect to the increase of sample size, ranging from about 1 min for 14 k cells to 26 min for 1.46 million cells,

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021

which are faster than Seurat and SingleR that need 51 min and 142 h for 1.46 million cells, respectively. Under the pretrain-finetune mode, the runtime of scAdapt can be further reduced to 21 (pretrain) and 1 min (finetune) for 1.46 million cells. In terms of memory usage, all methods consumed similar memory with increasing cells and required memory less than 100 GB for 1.46 million cells. Besides, scAdapt also achieves higher accuracy (0.91) than Seurat (0.81) and SingleR (0.87). As for batch correction methods (Table S4, see Supplementary Data available online at http://bib.oxfordjournals.org/), only fastMNN was able to complete runs on all sample sizes (from 7 min to 2.4 h), while the remainder did not complete due to memory allocation error. We also find that the average memory consumption of batch correction methods is larger than those of classification methods. fastMNN requires 175 GB memory to run the biggest dataset, which is higher than the memory requirement of classification methods at the same sample size (SingleR: 90 GB, Seurat: 82 GB, scAdapt: 75 GB). By comparison, scAdapt performs classification and batch correction simultaneously, which can save more computational resources. scAdapt also provides the option for CPU implementation. Without GPU acceleration, the runtime of scAdapt increases to 44 min for 1.46 million cells, slightly faster than that of Seurat. We also used the Covid-19 dataset with 1.46 million cells as large source and large target dataset. We found that SingleR and all batch correction methods cannot finish this task within 24 h or hit an out of memory error (Table S5, see Supplementary Data available online at http://bib.oxfordjournals.org/). Seurat is faster but also took 21 h. scAdapt requires much shorter time with 160 min and the finetune process only consumes 20 min. These results reflect that scAdapt is computationally efficient when analyzing large-scale dataset.

Discussion

In this work, we developed a novel virtual adversarial domain adaptation framework, scAdapt, to perform cell-type classification for datasets with batch effects. The virtual adversarial-based semi-supervised learning in scAdapt improves classification accuracy using both labeled source dataset and unlabeled target dataset, and domain alignment removes batch effects in the embedding space by making use of label information in the source. For quantitative benchmarks, we used simulated scRNA-seq datasets that vary in the intensity of batch effects, real cross-platform, cross-species, spatial transcriptomic and COVID-19 immune datasets. Experiments with quantitative measure validated the superiority of scAdapt. Visual inspection also demonstrated that the method preserved discriminative cluster structure present in the original datasets with the same types of cells well mixed. To gain the biological interpretability behind model decisions, we also identified cell-type specific marker genes and one portion of them were validated by the PanglaoDB database.

We demonstrated that our method could also overcome strong batch effects, while other classification methods did not perform well when there is large batch difference between source and target datasets. Additionally, our method is robust to remove batch effects in the combined dataset even if we combine datasets from different platforms or species as the source. It should be noted that the source dataset for model training should contain a reasonable number of cells per cell type for reliable cell-type annotation. We recommended including at least 10 cells per cell type to adequately represent the transcriptional program as well as variance.

Unlike the cross-dataset mapping approaches, classifiers based on cell-type marker genes such as Garnett [35] and CellAssign [37] can also overcome the batch effect issues. We compared scAdapt with these two classifiers on the human PBMC dataset which has comprehensive resource of marker genes for various cell types. We used the marker genes from three sources: the original author, CellMarker database [38] and top 10 differentially expressed markers of the respective cell types in the training data. From Figure S16, see Supplementary Data available online at http://bib.oxfordjournals.org/, we found that the selection of marker genes has large impact on the performance of Garnett (accuracy: 0.53–0.71) and CellAssign (accuracy: 0.51–0.62). CellAssign shows lower performance compared with other classifiers and Garnett only surpasses the accuracy of CHETAH and scPred. These results are in line with the previous benchmark conclusions that marker gene knowledge is not beneficial and the performance strongly depends on the selected markers [3, 14].

For constructing large reference datasets and pretrained models, we also test scAdapt in a pretrain-finetune manner. Specifically, we first pretrain the model only with the cross-entropy loss on the labeled source data, and then fine-tune the model with VAT loss on the unlabeled target data. This manner does not require rerunning the full training pipeline on source data and thus save computational resources. We found that our fine-tuned models only experience a slight accuracy loss on the cross-platform (2%), cross-species (1%) and Covid-19 (2%) datasets. These results validate the effectiveness of the proposed pretrain-finetune approach. We have uploaded the pretrained classifiers to GitHub for further use.

As for guideline to use scAdapt, we first need to choose a well-annotated reference dataset which contains similar cell types with the query data. Cell atlases, such as the Human Cell Atlas [39] with comprehensive tissue and cell-type information, are excellent for building reference. Once such standard references are constructed, scAdapt can be used to characterize cells from different tissues, diseases and states. Another typical scenario to use scAdapt is that users can share their reference data as pretrained network without any sharing of private data, such as human data due to legal restrictions. This makes the learned models easily transferable, shareable and reproducible.

It is important to note that the semi-supervised classification module of scAdapt can deal with different degrees of batch effects and does not need to perform batch correction explicitly. During the training process, the classification module is trained in the first step and then the batch correction module is trained with the pseudo labels produced by the classification module. If only performing classification task, users can close the batch effect removal step by zeroing out the regularization coefficients of batch correction module in the loss function.

While scAdapt performed the best in the experiments, there is still room for improvement. scAdapt needs source cell labels for supervised batch correction. For the scenario without source data label information, we can perform a preliminary clustering in the source data and use the inferred cluster labels as input labels for scAdapt training. Integrating the preclustering step and scAdapt into a unified deep learning framework may accommodate such unlabeled scenario more effectively, which will be left for our future work. Another future direction is to enhance the classifier by distinguishing similar subtypes at deeper annotation level since the subtle biological difference between subtypes is often masked by noise from experimental batches and sequencing platforms, and hard to be recognized. This problem might be solved by the recently developed fine-grained image classification that can learn a more discriminative feature representation [40].

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021

Key Points

- With the rapid accumulation of labeled single cell datasets, such as Human Cell Atlas, many supervised approaches have been developed to identify cell types for cells in the new unannotated data. However, existing approaches are sensitive to batch effects or biological variations since the data distributions are different in cross-platforms or species predictions.
- To address this critical issue, we developed scAdapt, a virtual adversarial domain adaptation network, to transfer cell labels between datasets with batch effects. scAdapt used both the labeled source and unlabeled target data to train an enhanced classifier and aligned the labeled source centroids and pseudo-labeled target centroids to generate a batch-corrected embedding.
- Our scAdapt method shows clear advantages over seven state-of-the-art single cell classification methods and four batch effect correction methods in simulated, cross-platforms, cross-species, spatial transcriptomic and COVID-19 immune datasets. Visualizations for the batch-corrected embeddings confirm the superiority of scAdapt in mixing cells of the same cell type across batches and the ability to preserve discriminative cluster structure present in the original datasets.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data and software availability

The datasets and code for scAdapt are publicly available at GitHub: https://github.com/biomed-ai/scAdapt.

Funding

National Key R&D Program of China (2020YFB020003); National Natural Science Foundation of China (61772566, 81801132); Guangdong Frontier & Key Tech Innovation Program (2018B010109006, 2019B020228001); Natural Science Foundation of Guangdong, China (2019A1515012207); Introducing Innovative and Entrepreneurial Teams (2016ZT06D211).

References

1. Lahnemann D, Koster J, Szczurek E, *et al*. Eleven grand challenges in single-cell data science. *Genome Biol* 2020;**21**(1):31.
2. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 2019;**15**(6):e8746.
3. Abdelaal T, Michielsen L, Cats D, *et al*. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* 2019;**20**(1):194.
4. 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**(2):163–72.
5. de Kanter JK, Lijnzaad P, Candelli T, *et al*. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res* 2019;**47**:e95–5.
6. Kiselev VY, Yiu A, Hemberg M. Scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods* 2018;**15**(5):359–62.
7. Alquicira-Hernandez J, Sathe A, Ji HP, *et al*. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol* 2019;**20**:264.
8. Tan Y, Cahan P. SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. *Cell Syst* 2019;**9**:207–213 e202.
9. Stuart T, Butler A, Hoffman P, *et al*. Comprehensive integration of single-cell data. *Cell* 2019;**177**(7):1888–1902.e21.
10. Ganin Y, Lempitsky V. *Unsupervised domain adaptation by backpropagation*, *International Conference on Machine Learning*, Lille, France, 2015, 1180–9.
11. Haghverdi L, Lun ATL, Morgan MD, *et al*. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 2018;**36**(5):421–7.
12. Korsunsky I, Millard N, Fan J, *et al*. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**(12):1289–96.
13. Welch JD, Kozareva V, Ferreira A, *et al*. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 2019;**177**(7):1873–1887.e17.
14. Huang Q, Liu Y, Du Y, *et al*. Evaluation of cell type annotation R packages on single cell RNA-seq data. *Genomics, Proteomics & Bioinformatics* 2020. doi:https://doi.org/10.1016/j.gpb.2020.07.004.
15. Ge S, Wang H, Alavi A, *et al*. Supervised adversarial alignment of single-cell RNA-seq data. In: *International Conference on Research in Computational Molecular Biology*. New York City: Springer, 2020, 72–87.
16. Wang M, Deng W. Deep visual domain adaptation: a survey. *Neurocomputing* 2018;**312**:135–53.
17. Ouali Y, Hudelot C, Tami M. An overview of deep semi-supervised learning. arXiv preprint arXiv:.05278 2020.
18. Cui S, Wang S, Zhuo J, *et al*. Towards discriminability and diversity: batch nuclear-norm maximization under label insufficient situations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, Washington, USA 2020, 3941–50.
19. Miyato T, Maeda S-I, Koyama M, *et al*. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2018;**41**:1979–93.
20. Xie S, Zheng Z, Chen L, *et al*. *Learning semantic representations for unsupervised domain adaptation*. Stockholm, Sweden, International Conference on Machine Learning, 2018, 5423–32.
21. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* 2017;**18**(1):174.
22. Ding J, Adiconis X, Simmons SK, *et al*. Systematic comparative analysis of single cell RNA-sequencing methods. *BioRxiv* 2019;632216.
23. Tran HTN, Ang KS, Chevrier M, *et al*. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 2020;**21**:1–32.
24. Moffitt JR, Bambah-Mukku D, Eichhorn SW, *et al*. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* 2018;**362**(6416):eaau5324.
25. Ren X, Wen W, Fan X, *et al*. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* 2021;**184**(7):1895–1913.e19.
26. Wen Y, Zhang K, Li Z, *et al*. A discriminative feature learning approach for deep face recognition. In: *European Con-*

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021

ference on Computer Vision. New York City: Springer, 2016, 499–515.
27. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:.05278 2013. 2013.
28. Franzén O, Gan L-M, Björkegren JL. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* 2019;**2019**:baz046. https://doi.org/10.1093/database/baz046.
29. Yu G, Wang L-G, Han Y, *et al*. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology* 2012;**16**(5):284–7.
30. Paszke A, Gross S, Massa F, *et al*. Pytorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems*, Vancouver, BC, Canada 2019, 8026–37.
31. Wang T, Johnson TS, Shao W, *et al*. BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome Biol* 2019;**20**(1):165.
32. Becht E, McInnes L, Healy J, *et al*. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2019;**37**(1):38–44.
33. Wang Q, Kulkarni SR, Verdú S. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Trans Inf Theory* 2009;**55**(5):2392–405.
34. Bezman NA, Kim CC, Sun JC, *et al*. Molecular definition of the identity and activation of natural killer cells. 2012;**13**: 1000–9.
35. Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid annotation of cell atlases. *Nat Methods* 2019;**16**(10):983–6.
36. Reichert M, Rustgi AK. Pancreatic ductal cells in development, regeneration, and neoplasia. *J Clin Invest* 2011;**121**(12):4572–8.
37. Zhang AW, O'Flanagan C, Chavez EA, *et al*. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* 2019;**16**(10):1007–15.
38. Zhang X, Lan Y, Xu J, *et al*. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res* 2019;**47**(D1):D721–8.
39. Regev A, Teichmann SA, Lander ES, *et al*. Science forum: the human cell atlas. *Elife* 2017;**6**:e27041.
40. Lin T-Y, RoyChowdhury A, Maji S. Bilinear cnn models for fine-grained visual recognition. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449–57.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbab281/6326525 by Sun Yat-Sen University user on 15 August 2021