ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

scDA: Single cell discriminant analysis for single-cell RNA sequencing data

Qianqian Shi [a], Xinxing Li [a], Qirui Peng [a], Chuanchao Zhang [b,*], Luonan Chen [b,c,d,*]

[a] Agricultural Bioinformatics Key Laboratory of Hubei Province, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[b] Key Laboratory of Systems Health Science of Zhejiang Province, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou 310024, China
[c] State Key Laboratory of Cell Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China
[d] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

ARTICLE INFO

Article history:
Received 29 January 2021
Received in revised form 25 May 2021
Accepted 25 May 2021
Available online 29 May 2021

Keywords:
Single-cell RNA-sequencing
Discriminant analysis
Discriminant features
Cell-by-cell representation graph
Cell annotation

ABSTRACT

Single-cell RNA-sequencing (scRNA-seq) techniques provide unprecedented opportunities to investigate phenotypic and molecular heterogeneity in complex biological systems. However, profiling massive amounts of cells brings great computational challenges to accurately and efficiently characterize diverse cell populations. Single cell discriminant analysis (scDA) solves this problem by simultaneously identifying cell groups and discriminant metagenes based on the construction of cell-by-cell representation graph, and then using them to annotate unlabeled cells in data. We demonstrate scDA is effective to determine cell types, revealing the overall variabilities between cells from eleven data sets. scDA also outperforms several state-of-the-art methods when inferring the labels of new samples. In particular, we found scDA less sensitive to drop-out events and capable to label a mass of cells within or across data-sets after learning even from a small set of data. The scDA approach offers a new way to efficiently analyze scRNA-seq profiles of large size or from different batches. scDA was implemented and freely available at https://github.com/ZCCQQWork/scDA.

1. Introduction

Single-cell RNA sequencing profiles gene expression of individual cells, allowing to detailly characterize multicellular organisms than bulk RNA-seq [1–3]. While, given a scRNA-seq transcriptomic data set, one challenge is to unsupervisedly find out distinguishing features (or genes) for different cell populations [4,5]. In particular, reliable features would greatly improve efficiency to annotate large data sets or newly profiled cells [5,6].

Recently, a popular way is a two-step schema [7], i.e., firstly identifying cell groups based on clustering approaches, then obtaining discriminant genes between these cell groups. These genes are usually referred to as marker genes [5]. For instance, the pipeline of SC3 [8] is to first infer cell clusters, and then conduct binary comparisons (i.e., one cluster vs the remaining cells) for selecting cluster-specific markers. SparseDC [9] model can extract diverse types of marker genes with known cell labels based on two-sample statistical theory. Unfortunately, this schema roughly defines a stepwise solution, which ignores the inherent correlations between genes and cells. Both results may be overfull dependent on cell clustering and easily lead to biased analysis. Meanwhile, group-level comparisons construct a separating line or hyper-plane in the original data but are difficult to capture the crucial information underlying heterogeneous structures. This could largely depress their performances to discriminate cells from new data sets.

Such issues thus turn the attention to those feature extraction methods, which enable to discover a set of informative genes (or metagenes) by preserving the supportive structure inherent in data. Actually, many models are ever developed and widely used for analogous problems in bulk RNA-seq [10,11]. For example, principal component analysis (PCA), is directly borrowed to analyze scRNA-seq data sets [12], or modified to handle noisy data (e.g., dropouts) in experiments [13]. Considering high heterogeneity among cells, some nonlinear methods, such as *t*-SNE [14] and

[*] Corresponding authors at: Key Laboratory of Systems Health Science of Zhejiang Province, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou 310024, China (L. Chen).

E-mail addresses: chaozhangchuan@163.com (C. Zhang), lnchen@sibs.ac.cn (L. Chen).

https://doi.org/10.1016/j.csbj.2021.05.046

SNN-Cliq [15], were also proposed. They aim to preserve local variations of data in a lower dimensional feature space and help improve the visualization of natural cell groups, appearing more viable for scRNA-seq profiles. However, there are two main shortcomings in these involved approaches: (i) they only retain part of the data structures in new feature space, e.g., global linear structure (e.g., PCA) or neighborhood relationship (e.g., *t*-SNE), which cannot take advantage of the full population heterogeneity; and (ii) consequently the obtained features are usually ineffective for annotating new samples. In most cases, these limitations further require costly re-analysis when handling merged data sets, or may take more time and memories with large amount of cells (e.g., *t*-SNE).

To address such challenges in real applications, we propose a novel model, which can unsupervisedly capture the overall relationship of cells and meanwhile obtain powerful discriminants (metagenes) from scRNA-seq data sets. In particular, the cell-by-cell relationship can be expressed as representations of multiple neighboring cells [16,17], which provides information on the gross (global and local) variabilities among the inherent cell types (Supplementary note 1). The representation graph in our method is more robust to noise and data heterogeneity in profiles [16], rather than the pair-wise distances. Moreover, the metagenes are also identified to best fit in with the full representation configuration. These features carry optimal discriminative information, and can be used to discriminate unlabeled cells from new dataset.

With the model, we made the analytic pipeline of single cell discriminant analysis (scDA) for scRNA-seq data (Fig. 1). We separate the whole data set into two parts: discovery (or old) set and validation (or new) set. On this basis, scDA implements two main steps: identify data structures (cell quantitative affinities and cell clusters) and discriminant features (metagenes) with the discovery set, and then use them to label the cells in the validation set. Thus, scDA can avoid unnecessary re-clustering, and is actually a combinational approach simultaneously performing both clustering and classification. We demonstrate the effectiveness and efficiency of our scDA on eleven scRNA-seq data sets involving different biological systems and technologies.

2. Materials and methods

2.1. Processing of scRNA-seq expression profiles

Ten scRNA-seq data sets were selected for method evaluation in total (Table 1); we kept the originally normalized data from corresponding researches. All these datasets were log-transformed after adding a pseudo count of 1, and further preprocessed by different approaches for different feature measurements.

With the four datasets from Camp to Pollen, we considered the effect of dropouts on clustering and classification, and generated series of corresponding data subsets with different numbers of genes. Since the sparsity rate is empirical and uncertain, we simply filtered out those genes when a gene has more than a certain proportion of zeros (i.e., $0.1 \sim 0.9$ by 0.1) across cells. Thus, these benchmarks were expanded to 4×9 data sets.

With pancreas datasets, we tested clustering and classification performances of scDA within and across datasets. For the first purpose, we generated two large datasets, namely Large1 and Large2. Large1 is a mix of 4 profiles including Lawlor, Segerstolpe, Muraro and Enge. Large2 is Baron. The raw data sets were firstly preprocessed, including quality control and normalization. Then Large1 sets are further batch corrected. All the process steps are similar as Haghverdi's work[18] except that we use all the genes rather than highly variable genes when removing batch effects of Large1. We abandoned these 'Unknown' cells and retained the other six types of cells in all. We applied scDA to Large1 under sparsity rate of 0.3 and Large2 under 0.1 to make a similar number of genes $\sim$ 10,000. The specialized experiments can be used to test the robustness of scDA as well as to evaluate intra-dataset classification performances. For the evaluation of inter-dataset classification performances, we keep all the common (i.e., 15,446) genes across the five pancreas datasets for fair comparison instead.

We also used the Galen data for classification evaluation across different subjects. The whole amount of genes, i.e., 27,672, passing the filtering thresholds in the original paper are used in our work. 1 M dataset (obtained from https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons) is mainly

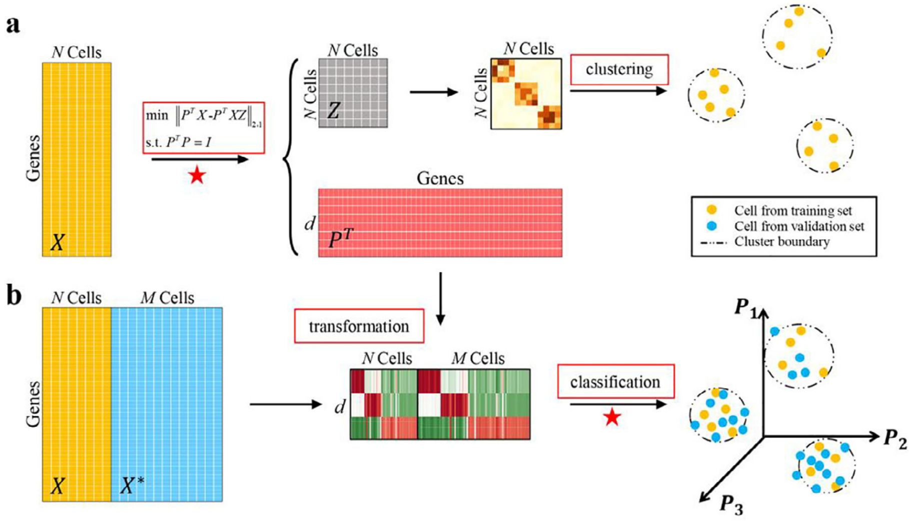

Fig. 1. Illustrative scheme of scDA. (a) Identify cell affinities and discriminants from profiles of discovery cohort. Matrix *Z* and *P* respectively denote representation matrix and discrimination matrix, which are solved through the formula using expression matrix of *X*. The representation matrix can be converted to similarity matrix, then used to define cell clusters in discovery data set. (b) Classify new samples from validation cohort (i.e., X^*) based on the obtained cell clusters and discriminants. Red stars highlight scDA implemented in two available ways: (i) unsupervised mode only with scRNA-seq dataset; (ii) supervised mode with data and given labels of cells in training set. In supervised mode, the prior cell annotations can help constrain scDA model either to obtain features (i.e., representation and discrimination matrices) or build classifiers (i.e., with the discriminant vectors and cell annotations). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Datasets for scDA evaluation.

Datasets	*k*	*N*	Unit	Description	Protocol	Reference
Camp	7	777	FPKM	Human liver bud	SMARTer	Ref. [33]
Goolam	5	124	CPM	Mouse embryo	Smart-Seq2	Ref. [34]
Li	9	561	FPKM	Human cell line	SMARTer	Ref. [35]
Pollen	11	301	TPM	Human cell line	SMARTer	Ref. [36]
Lawlor	6	597	RSEM	Human pancreas	Smart-Seq2	Ref. [37]
Segerstolpe	6	1,812	RPKM	Human pancreas	Smart-Seq2	Ref. [38]
Muraro	6	1,940	Count	Human pancreas	CEL-Seq2	Ref. [39]
Enge	5	2,174	Count	Human pancreas	Smart-Seq2	Ref. [40]
Baron	6	7,742	Count	Human pancreas	inDrop	Ref. [41]
Galen	15	4,677	CPM	Human bone marrow aspirate	Seq-Well	Ref. [42]
1 M	20	1,306,127	Count	Mouse brain	10X Genomics Chromium	–

k: number of groups provided by the authors.
N: number of cells.

used to test time efficiency and computing capacities of scDA, where top 2000 highly variable genes are remained for simplicity.

2.2. Method overview

Given the processed scRNA-seq data (e.g., from discovery set), scDA could construct two matrices, defined as the ‘representation matrix’ and ‘discrimination matrix’ respectively, to accomplish two tasks: characterizing cell types, i.e. clustering in discovery set (Fig. 1a), and predicting new cells in unannotated set (Fig. 1b) based on the model-determined or prior-given labels, i.e., identified in unsupervised and supervised modes (details in Sections 2.3–2.5).

The ‘representation matrix’ describes the inherent relationship of cells, available for cell clustering, and the ‘discrimination matrix’ contains representation-learning-based features, qualified for classification. In fact, the key issue of scDA approach is to solve the two matrices that involves three basic assumptions: (i) different cell populations belong to distinct biological subspaces or processes, (ii) each cell can be mathematically represented by other cells with similar biological properties, and (iii) the intrinsic heterogeneity as described in (i)-(ii) can be approximately reconstructed on a common feature subspace.

Under assumptions (i) and (ii), which cover global differences and local similarities of underlying data, the cell-by-cell representation graph can be built based on expression profiles by subspace segmentation theory [17,19]. It presents as a sparsely filled matrix, where the coefficients are equal to zero if cells are faraway or in different clusters (Supplementary note 1), and reveals implicit subspaces of various cell populations. Note that representation-based measurement not only describes cell affinities as well as pairwise metrics, but also has robust performances against data biases or noises from scRNA-seq experiments (Fig. 2 and Supplementary Figs. S1–S4). Broadly, the graph quantitatively defines intrinsic heterogeneity from cells, which is of great importance to obtain discriminant features.

While, it is usually considered that the number of features inherent in data is much smaller than the total number of profiled genes [5,18]. As assumed in (iii), our dimensionality reduction is constrained by the obtained representation structure (Supplementary note 2). Based on metric-learning theory [20], this approach aims to project the original data onto a subspace which best fits in with the detailed configuration of within- and between- groups [21]. Therefore, the projection matrix contains the information of underlying cell differences and has more discriminating power on new objects than PCA or *t*-SNE.

However, it is generally impossible to simultaneously obtain the optimal solutions of the two matrices. Inspired by the phenotypic and feature (metagene) correlations, here we design an alternately iterative optimization algorithm, which is guaranteed to converge theoretically [17,19]. Our method is robust to a variety of the hyperparameter settings (Supplementary Figs. 5 and 6). We emphasize that the whole solving process forms a data-driven closed loop to alternately compute the cell-by-cell and feature-by-gene matrices until it reaches convergence. The loop ensures scDA could unbiasedly derive inherent information from (discovery) data sets.

Then, with the optimal representation matrix, scDA is capable to estimate the involved cell types through a graph-based clustering method, e.g., spectral clustering [22]; and classify the unlabeled cells to the acquired assignments based on discriminant vectors, e.g., *K*NN classifier. The hybrid approach dispenses with re-analysis, appearing more feasible to deal with large amount of cells within or across datasets. We also provide supervised mode to accommodate valuable prior-knowledge from experiencers to obtain more reliable annotations. Regardless of running modes, the performance of scDA is considered to be closely related to the characteristics of these two matrices.

2.3. Optimization problem of scDA

Suppose we describe a scRNA-seq data with g biological measurements and N samples as a matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]$. scDA could learn the representation matrix $\boldsymbol{Z}$ and the discrimination matrix $\boldsymbol{P}$ from the input data matrix by solving the following objective function:

$$\min_{Z,\,P} \parallel Z \parallel_* + \lambda \parallel P^T X \text{ - } P^T XZ \parallel_{2,1}$$
$$\text{s.t.} \begin{cases} P^T P = I \\ Z^T \mathbf{1} = \mathbf{1} \\ Z_{ij} = 0, \qquad (i,j) \in \bar{\Omega} \end{cases} \tag{1}$$

where $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N]$ is an $N \times N$ matrix containing all the coefficient vectors of $\boldsymbol{x}_i (1 \le i \le N)$. $\|\cdot\|_*$ represents the nuclear norm of the matrix, which is the sum of singular values. The projection matrix $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_d] \in R^{g \times d}$ denotes the d discrimination vectors transformed from all the profiled genes and $l_{2,1}$-norm is used to measure the reconstructive error matrix due to its robustness [19]. In the constraint conditions, $\boldsymbol{I}$ denotes the identity matrix, indicating $\boldsymbol{P}$ is an orthogonal subspace without redundancy. Here $\mathbf{1}$ is an all-one vector for normalization. And $\overline{\Omega}$ is the complement of Ω, where Ω is a set of edges between the samples in a predefined adjacency graph. For example, if $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ are not neighbors, then we have$(i,j) \in \overline{\Omega}$. Thus, the third constraint can guarantee the preservation of local structure underlying data. In unsupervised mode, we use K nearest neighbor (*K*NN) algorithm with pair-wise Euclidean distances to determine the sample adjacency graph; in supervised

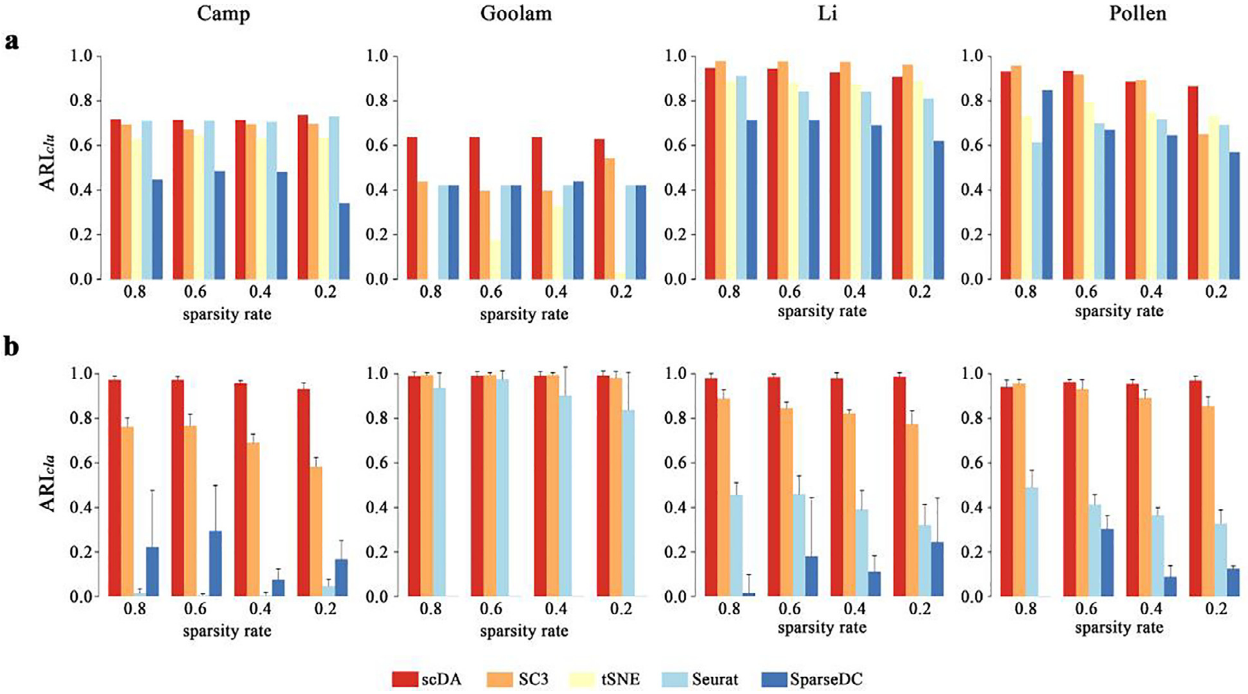

Fig. 2. Comparative performances of scDA approach to existing methods. (a, b) Adjusted Rand index (ARI) values of clustering (a) and classification (b) for small data subsets in scDA supervised mode at different sparsity rates, e.g., from 0.8 to 0.2 by 0.2. Performances at other sparsity rates are seen in Supplementary Figs. 1 and 2. Clustering and classification are respectively abbreviated to *clu* and *cla*, which are used as footnotes in Figs. 2, 4 and Supplementary Figs. 1–3, 5, 9, 10, 13. Bars with zero height indicate NA values. Error bars in (b) indicate upper 95% confidence interval of cross-validation results.

mode, the local constraint is fully or partly determined by prior-knowledge, e.g., known cell labels. While, parameter K is actually free to set and besides it, the equation (1) have a tuning parameterλ to balance the two terms. They both can be selected according to data properties, or settled empirically.

Next, we focus on the solution of the objective function. Considering either of the objective terms can be solved separately, we design an alternately iterative algorithm to compute one matrix by fixing the other one. The splitting problems can then be seen as follows.

(1) Solving the optimal matrix ***Z*** by fixing ***P***

Let's convert the Eq. (1) to an equivalent form when ***P*** is fixed:

$$\min_{Z,\,E} \quad \| Z \|_* + \lambda \| E \|_{2,1}$$

$$\text{s.t.} \begin{cases} P^TX = P^TXZ + E \\ Z^T\mathbf{1} = \mathbf{1} \\ Z_{ij} = 0, \qquad (i,j) \in \bar{\Omega} \end{cases} \tag{2}$$

where ***E*** presents the reconstructive error matrix. It can be solved via modified low-rank representation algorithm [17] (Supplementary Note 1).

(2) Solving the optimal matrix ***P*** by fixing ***Z***

When ***Z*** is given, the substantial problem becomes as:

$$P* = \underset{P}{\operatorname{argmin}} \quad \| P^TX - P^TXZ \|_{2,1}, \ \text{s.t.}\ P^TP = I \tag{3}$$

The problem (3) can be solved using $l_{2,1}$-norm minimization technique [23] (Supplementary Note 2).

2.4. scDA clustering for discovery set

Given representation matrix ***Z***, the corresponding affinity matrix ***W*** is obtained by $\boldsymbol{W} = \left(|\boldsymbol{Z}| + |\boldsymbol{Z}^T|\right)/2$. With the affinity graph, we used spectral clustering algorithm (e.g., RatioCut [24]) to identify the underlying groups of cells. And the number of groups, i.e., k, is estimated in this work by the eigengap heuristic [22], or determined by other approaches, e.g., silhouette coefficients [25]. If ***Z*** is constrained fully by known cell labels, clustering step can be omitted, e.g. study with Galen.

2.5. scDA classification for unannotated set

For 4 small data sets, we adopted five-fold cross-validation experiments. All the cells in data are randomly separated into five folds, while each fold is treated as the testing set and the remaining folds as the training set. We use a set of discriminant vectors obtained from training set and (i) inferred labels with representation matrix or (ii) the gold-standard cell labels to build KNN classifiers, i.e., classifications in unsupervised /supervised way, for annotating testing cells. Additionally, the prediction accuracy is impacted by the number of discriminant vectors. We pick the first d eigen vectors of the discrimination matrix ***P***, as they are ordered ascendingly by the eigen values.

For pancreas data sets in Results section, we have small discovery subsets, of which (i) the cell number ranges from 3% to 10% of all the cells from large datasets for classification intra-datasets, (ii) less than 10% size of Baron for classification cross-datasets, and the remaining cells as validation subsets. Under the same sizes, the paired subsets were randomly generated for 100 times. At every time, we randomly selected the same proportion of cells from all the six cell types to guarantee a full learning with training sets. Then, the classification experiments in unsupervised /supervised way are similar to the above descriptions in small data sets.

For 1 M dataset, we firstly run PCA and get the 2,000 medoids with 50 PCs as discovery landmarks for clustering, and then completed classification step.

2.6. Methods comparison

SC3 (version 1.14.0), t-SNE (package Rtsne version 0.15), Seurat (version 2.3.4) and SparseDC (version 0.1.17) were also

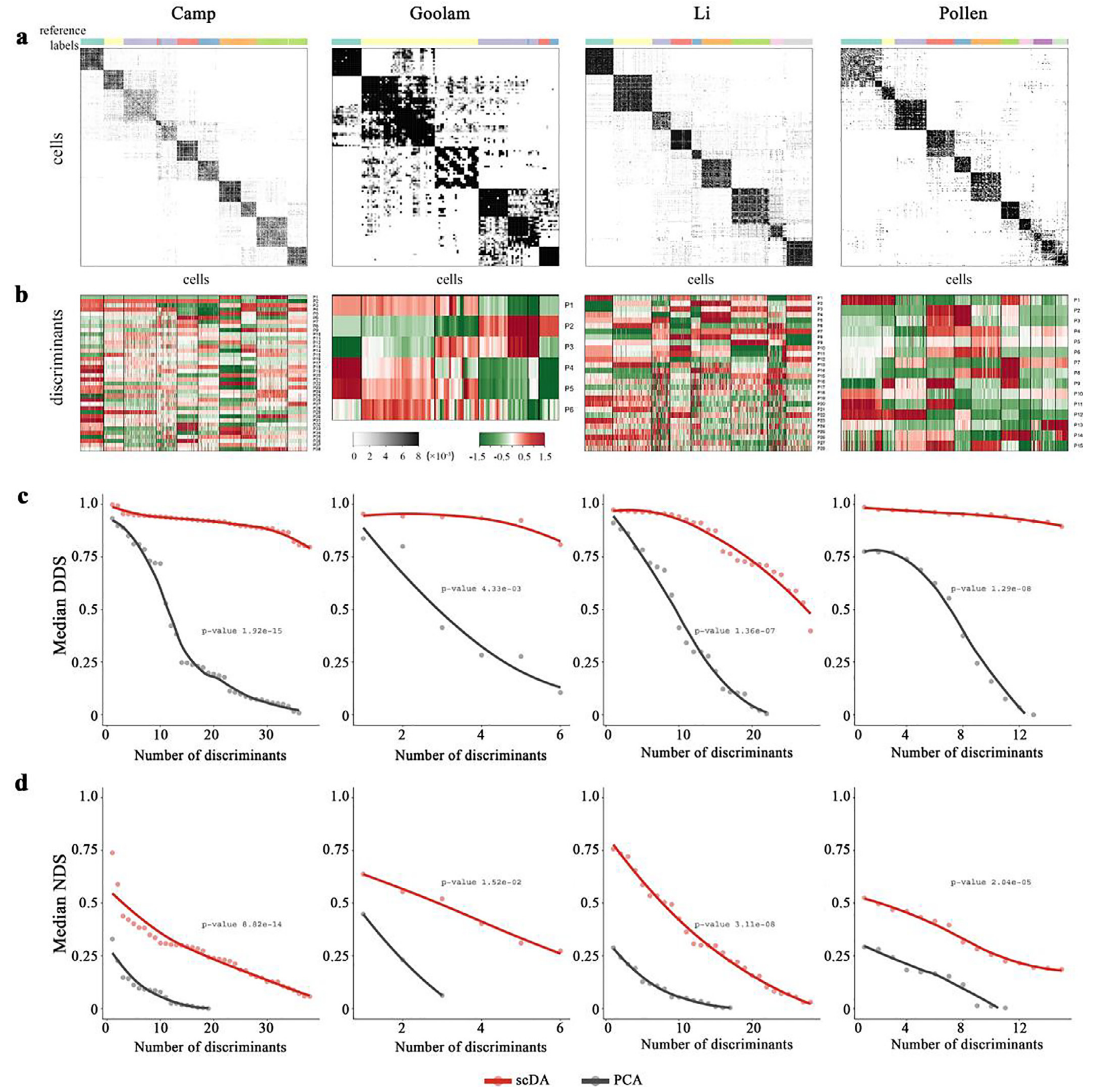

Fig. 3. Characteristics of representation and discrimination matrices from scDA with small data sets. (a, b) Illustration of representation matrices (a) and projected data (b) from Fig. 2a. Column-side colors indicate the reference cell types provided by the original authors. Sample orders are the same in (a) and (b). The optimal groups estimated by the largest eigen gap are separated by black vertical lines in (b). The top discriminants (eigen vectors) occupying 5% of the total number of involved cells, were selected and marked according to eigen values. (c, d) Discriminative abilities of the example discriminants to separate model-identified cell populations between the most distant pairs (c) or neighboring groups (d). DDS, discrimination score for distant groups; NDS, discrimination score for neighboring groups (see Methods). Discrimination scores smaller than 0 are not shown. Fitness curves were created under 'loess' regression. *P*-value was calculated using Wilcoxon test.

applied to the small data sets, where all four are to benchmark cell clustering, and three for classification except *t*-SNE. For SC3, the default parameters provided by the author are used for clustering and marker-genes identification. For *t*-SNE, given the existing cell types in each data set, we took the expected smallest amount of cells as its perplexity parameter, and the projected data are clustered following *k*-means. For Seurat, we chose the improved PCA-based clustering pipeline as representative method from several packaged dimensionality reduction models. For SparseDC, we randomly divided the cells into two groups as it requires two conditions before identifying cell types and the related marker genes. Note that for SparseDC we only selected the cell type-specific genes for classification. The number of cell clusters is inferred by the default functions for SC3 and Seurat, or directly set as original group number for SparseDC and *t*-SNE. The whole sets of the identified marker genes are used when built corresponding *K*NN classifiers for SC3, Seurat and SparseDC.

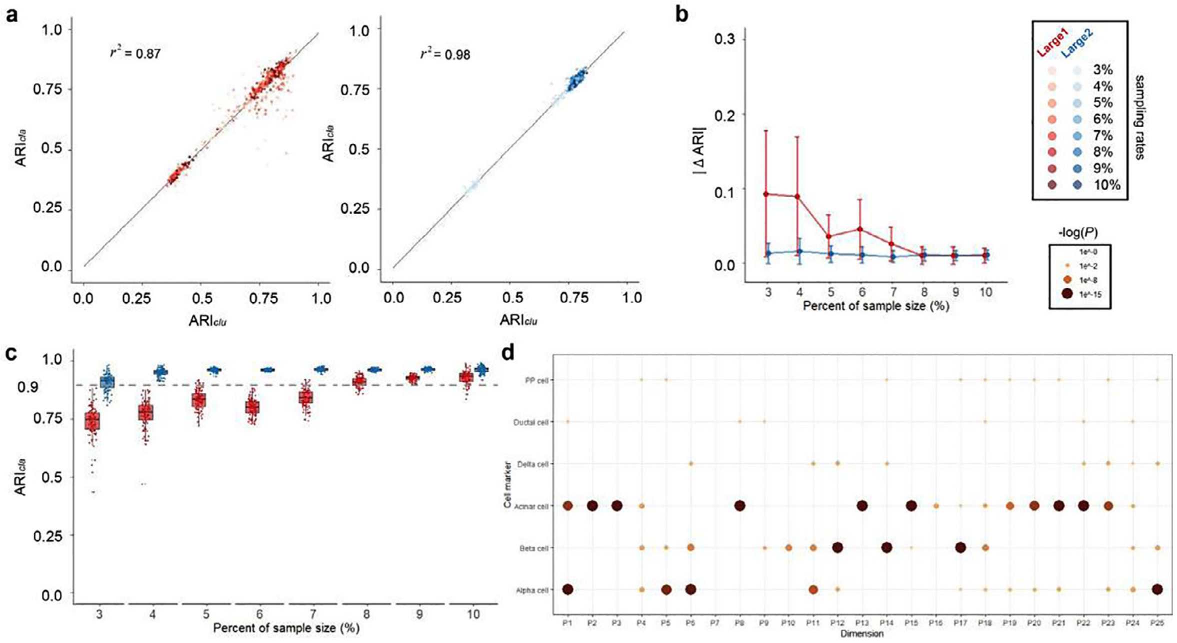

Fig. 4. Application of scDA on two large data sets. (a) Clustering (for discovery cohort) and classification (for validation cohort) performances in unsupervised manner with trained datasets of different sizes. For each data set, random sampling was repeated 100 times to construct the discovery sets at certain sampling rates, depicting the percentage of cell amount in total. The validation set is the rest dataset with removal of discovery set. (b) ARI difference (i.e., mean $|\Delta ARI|\pm$ *s.e.*) between clustering and classification performances from Fig. 4a. (c) Classification performance with construction of classifiers using reference labels. The dashed gray line indicates ARI of 0.9. (d) Enrichment results of marker genes with discriminant vectors from Large1 pancreas dataset. Marker genes in Fi. 4 and Supplementary Fig. S12 are obtained from CellMarker database [28]. The top 5% dimensions (or discriminant vectors) of discrimination matrix is used for enrichment analysis. Discriminant vectors were ascendingly sorted according to the eigen values. The visual example is selected owing to the highest clustering accuracy at sampling rate of 8% with Large1 dataset (k = 40). Each dot indicates enrichment significance.

2.7. Evaluation metrics

We used two metrics to evaluate the performances of scDA, one is adjusted Rand index (ARI) and the other is discrimination score (DS).

ARI is used to quantify the similarity between the reference labels and the clusters/predictions obtained by scDA and other involved methods, as the cell labels were known or ever inferred by the authors in all the applied data sets. For n objects, there are two groups denoted as a and b, and the relationship of the two groups can be summarized by a contingency table. Each entry of the table stores the number of common observations between a and b. Given the table, the ARI is calculated as

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2} + \sum_j\binom{b_j}{2}\right] - \left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/\binom{n}{2}}$$

where n_{ij} is the entry at the ith row and jth column in the contingency table, a_i, b_j are the marginal sums for the corresponding groups, and () represents a binomial coefficient.

The definition of DS is inspired by the fisher's discriminant rule [26] and a measure of group coherence, i.e., silhouette coefficient [25]. We use it to assess how effective each projection vector (feature) is to discriminate a given pair of groups, for instance, denoted as a and b. With the cell i in a, let $a(i)$ be the median distance of cell i to all other cells within a, and $b(i)$ denotes the median distance of cell i to all cells in b. The DS for cell i is defined as:

$$ds(i) = \begin{cases} 1 - \frac{a(i)}{b(i)},\ \text{if}\ a(i) < b(i) \\ 0,\ \text{if}\ a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1,\ \text{if}\ a(i) > b(i) \end{cases}$$

For simplicity, we further define distant DS (DDS) to measure the separations between the most distant groups, and neighboring DS (NDS) for neighboring groups. Both the scores range from + 1 to −1. And its median value over all the cells can then be used to measure the discriminative power of the examined feature to separate different groups. With the value bigger than 0, the feature could make positive contributions to distinguish groups. While if the value is closed to 1, it suggests that on the projected axis (or direction), cells are grouped tightly within respective clusters and far from the other clusters (similar to the Fisher's ratio of between- to within-group distance), then indicates that the projection feature brilliantly discriminates cells from the two groups.

2.8. Biological insights

We adopted GSEA approach [27] to explore the biological significance of discriminant features (metagenes). Features are sorted in a descending order according to the absolute scores of $\boldsymbol{P}$ (i.e., discrimination matrix), and gene sets of certain biological functions or sources are then tested. After FDR correction, the significant gene sets, e.g., P-value $\leq$ 0.05, can reveal the true biological information contained in the features. For example, when we applied scDA to the large data sets, we used the marker genes of known pancreatic cell types from CellMarker database [28] to determine whether the features are interpretable by the prior knowledge. It is implemented by phenoTest package (version 1.34.0).

3. Results

3.1. Study on small benchmark datasets

To exploit the characteristics of representation and discrimination matrices identified by scDA, we considered four small scRNA-seq datasets with well-defined cell types and different sequencing protocols (Table 1). As the representation graph is used to identify cell groups, and with discriminant vectors together to predict labels of new cells, we simply evaluate the two types of matrices by clustering/classification accuracies. We also conducted robustness tests using series of data subsets with different sparsity rates (see Methods). Note that both matrices and clustering/classification performances are assessed by adjusted Rand index (see Methods), which scores closer to 1 when the tested labels are similar to gold-standard ones. Here, we considered four established methods: SC3, *t*-SNE, Seurat[29] and SparseDC for fair comparisons.

In respect of clustering, scDA performed generally better than other involved methods, or pretty close to the best results (Fig. 2a and Supplementary Fig. S1). It is remarkable that scDA was the most stable method with respect to the percent of dropouts. Overall, the clustering achievements reveal that scDA can recognize the inherent data structures, and this feature is crucially important as the discriminative projection learns from such representation configuration (Fig. 1a). To classify unlabeled cells with discriminant vectors, scDA outperformed the three tested models (*t*-SNE not included): stable and accurate across the spectrum of input gene sets (Fig. 2b and Supplementary Figs. S2, S3). While, the second best method is SC3, which uses marker genes from binary classifiers. But the individual markers are less informative to discriminate different cell clusters with the increase of sample size; scDA showed more obvious superiority, such as on Li and Camp data sets, in both unsupervised and supervised modes. Furthermore, we compared clustering time cost of the best two algorithms and found scDA (implemented in python) also computes faster than SC3 (Supplementary Fig. S7).

In addition to these simple assessments, the cell-by-cell representation matrices could provide more affinity details than plain separations (Fig. 3a). Their illustrations appear nearly block-diagonal, indicating cells are tightly grouped in the identified clusters. Such "block" structures even can be seen in individual dimensions of the discrimination space (Fig. 3b). Here we use discrimination score (DS) to evaluate how powerful each single feature can discriminate different cell types (see Methods). We also compared scDA to principal component analysis (PCA), which also yields orthogonal rotations but without restrictions of cell relations. According to the scores, these projection vectors derived from scDA are overwhelmingly more favorable for distinct cell populations (Fig. 3c, d). Together with the minor differences between clustering and classifications (Supplementary Fig. S4), all the results showed that the scDA projection space with the preservation of overall-relationship is indeed spanned by remarkable discriminant vectors.

Above all, scDA model can unsupervisedly recognize the intrinsic sample patterns, and also obtain a number of powerful discriminants. These characteristics allow us to apply scDA to large scRNA-seq profiles or across datasets.

3.2. Performance within large benchmark datasets

Inspired by the performance of scDA on small data sets, we evaluated the applicability of scDA with two large scRNA-seq profiles denoted as Large1 and Large2 (see Methods). Both data sets are generated to study human pancreas, a highly heterogeneous tissue with several determined cell types (Supplementary Table S1). The difference between them is that some cell groups inherent in the mixed data set (after batch correction), i.e., Large1, present less coherent than those in the single-sourced Large2 (Supplementary Fig. S8), thereby enabling to make comparative and contrasted evaluations with datasets in different qualities. Furthermore, we expected capacity of scDA to use small subsets of cells to discriminate the majority of cells, and thus conducted random experiments varying different sample sizes of discovery set (see Methods). Here, we use clustering ARI, classification ARI and their agreement to assess the utility of scDA.

Firstly, we did unsupervised clustering with training data sets, and used the inferred labels to classify the other unlabeled cells. The results show that classification performance has strong correlation with clustering qualities in both large datasets (Fig. 4a and Supplementary Figs. S9, S10). However, more misclassifications were made when model learnt data patterns from very few samples, e.g., less than 5% of the total cells. This is probably caused due to the unbalanced distribution of given cell types (Supplementary Table S1), and small sampling ratio is more likely to obtain invalid training sets. Moreover, such mistakes still may occur even if sampling ratio increased in the dataset of Large1. Using this mixed dataset, we can also see more variances of ARI difference (Fig. 4b). The unsatisfactory performances may be subject to data qualities, for Large1 is multi-sourced and some confounding effects may still exist even after batch correction. While such situations get greatly improved to obtain accuracies and consistencies very close to those outcomes in Large2 when we use more training cells, e.g., 8% but still less than 10% of the total cell amount for Large1 data (Fig. 4b). All the results suggest the robustness of scDA to data coherences.

Furthermore, we then constructed the classifier with the reference cell labels instead, and also observed the classification can achieve comparably reliable outcomes between Large1 and Large2 when training the classifier with cells occupying 8% $\sim$ 10% of total sample size (Fig. 4c). Our discriminant vectors show fair discrimination capacities to separate the farthest or closest cell groups (Supplementary Fig. S11) and these metagenes are found highly enriched with the known cell markers (Fig. 4d and Supplementary Fig. S12). All the results indicate the discrimination matrix can capture true biological variances inherent in scRNA-seq datasets; thus, with the matrix, scDA can well predict the unlabeled cells in unsupervised (Supplementary Fig. S10) or supervised (Fig. 4c) manners. Besides, we were able to analyze a very large 10X dataset with more than 1.3 million cells and 20 clusters, generating results in good agreement with its original annotations (Supplementary Table S2). Taken together, the above analyses show excellent generalization of scDA, enabling the model to use a handful of cells to well predict the labels of abundant cells within data sets.

3.3. Classification evaluation across datasets

We next tested the prediction ability of scDA across datasets as batches or batch effects are really common between different scRNA-seq experiments [30], e.g., from different labs, sequencing protocols or subjects. An approach, which could handle cross-dataset classification, is more preferred and practical. We used the human pancreas and bone marrow aspirate datasets (Table 1) for inter-dataset evaluation, and made directional classifications where to use small datasets to predict large datasets.

We firstly tested classification performance of scDA across labs/protocols with five pancreas data sets. Specially, we made extreme tests to have scDA build classifiers from cells no more than 10% of Baron, the largest of pancreas profiles, and randomly divided all other datasets into different subsets, i.e., 2 folds for Lawlor, 3 folds for Segerstolpe, Muraro and Enge. Applied to the subsets, scDA obtained good cell annotations within each dataset in unsupervised way (Supplementary Fig. S13), consistent with our findings in previous sub-section. While, we then built classifiers with identified discrimination matrix and cell gold labels, and summarized their performances in Fig. 5a. All the classifiers performed well on predicting Lawlor's dataset due to its small amount of cells. However, the classifiers originated from Enge subsets predicted

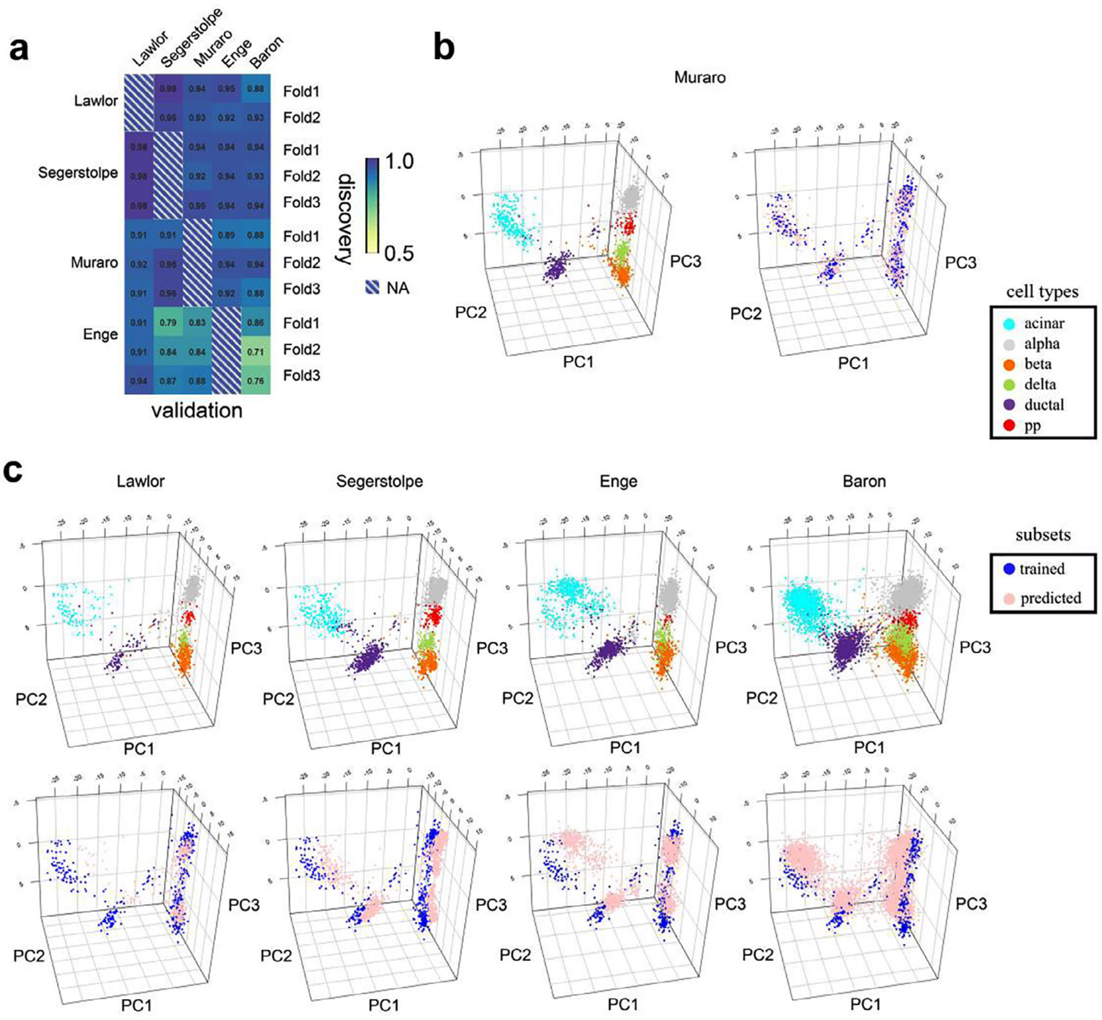

Fig. 5. Classification performance across pancreatic datasets. (a) Heatmap of cross-dataset classification ARI values. Each dataset in discovery set is randomly divided into 2 or 3 folds to unsupervisedly get representation and discrimination matrices. Classifiers with prior-provided cell annotations are to classify the whole cells in each prediction dataset. Classification ARI values are shown explicitly. Blue striped rectangle means NA value. (b, c) Visualization of trained and predicted cells in 3-dimenstional discrimination space. Fold2 of Muraro is used to train classifier for classification of other two Muraro folds in (b); then also to predict cell labels across Lawlor, Segerstolpe, Enge and Baron data sets in (c) Each point in graph represents a cell and colored by two annotation categories: data source and cell type. Cells are colored by given labels in training fold and predicted labels in testing datasets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the other datasets not as well, but still acceptable (ARI$\geq$ 0.79), as the rest classifiers (Fig. 5a). This may be because Enge is the only dataset that doesn't contain PP (pancreatic polypeptide) cells (Supplementary Table S2) and produces incomplete classifiers. While, from Lawlor, Segerstolpe, Muraro subsets, the classifiers were theoretically unbiasedly trained, and performed much better on across-dataset classifications, even on predicting the largest dataset, i.e., Baron, with unsupervised cell annotations (Supplementary Fig. S14). We then used principal component analysis to project our discrimination space onto three-dimensional visualization space (Fig. 5b and c), which clearly shows the "shifted gaps" across the involved datasets. These gaps explicitly reveal external differences between batches. The results indicate that the scDA discrimination space preserves biological variances underlying trained dataset, therefore it allows to label unannotated dataset of the same tissue or similar condition in regardless of batch effects.

We next tested classification performance of scDA across subjects with bone marrow (BM) aspirate dataset, which contains 4 human donors named as BM1 to BM4 according to the ascending order of available cell amounts. These sets have a total of 15 cell types belonging to 4 developmental lineages (Supplementary Fig. S15). To guarantee that the training set has at least 10 cells for one cell type, we trained with BMs1–3, i.e., about 25% the size of BM4, to classify cells in BM4. We first obtained corresponding discrimination matrix to the representation matrix, which was initially constrained by *K*NN using given cell annotations of BMs1-3. The top 14 discriminant vectors contribute most to different types of cells (Fig. 6a), thus are then used to build the classifier for BM4. Subsequently, we compared the classified assignments with the original annotations (Fig. 6b). Around 83.0% of cells are classified correctly according to their cell types. Among the mis-classified cells (17.0%), 13.0% are assigned to their related cell types within the same lineages, or predicted between HSC/Prog and other committed progenitor cells (2.4%), early Ery and proB, GMP (1.0%). It means that our mis-classifications mainly come from predicting the intermediated states or cell types with differentiation potency along the continuum of cells. Besides, we can also observe the putative differentiation relationships from undifferentiated cells to other lineages with our discrimination space (Fig. 6c and Supplementary Fig. S16). And the expression of marker signatures also supports our classification performance across different donors (Fig. 6d).

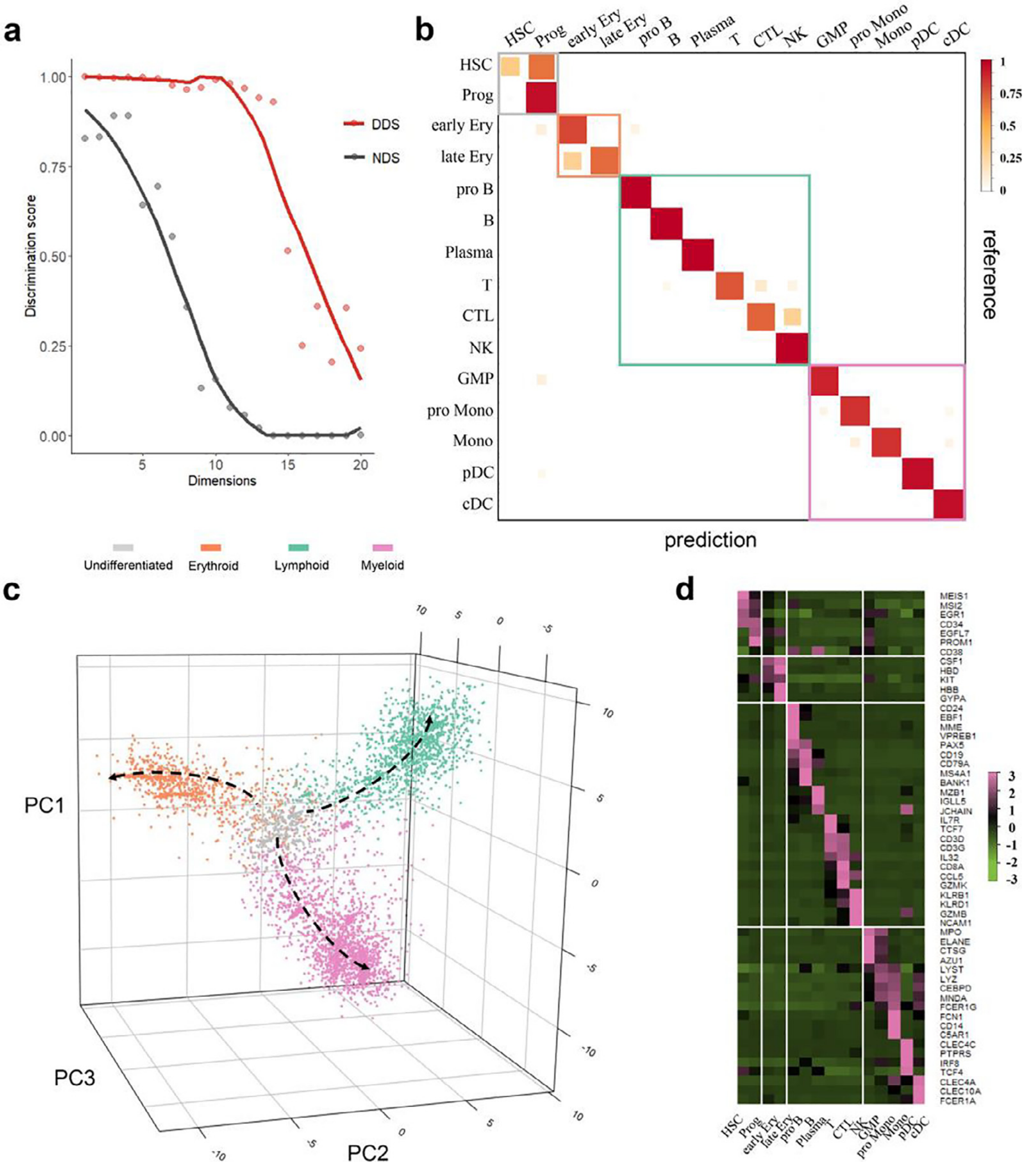

Fig. 6. Classification performance using subjects of BMs1-3 to annotate BM4. (a) Scatter chart of discrimination scores of top 20 discriminant vectors. We select the top discriminants by: (i) before the steep drop of median DDS, and (ii) with median NDS bigger than 0. (b) Prediction comparison for 3,738 hematopoietic cells in BM4. The square of heatmap represents the ratio of cells in prediction class over all cells of the same annotation. (c) Visualization of BMs1-4 in 3-dimenstional discrimination space. Dashed lines with arrows imply committed lineages from undifferentiated to mature states, which are inferred by cell labels (Supplementary Fig. 16). (d) Expression heatmap of 55 selected cell-type-specific genes (rows) across our annotated cell classes (columns). The median expression was calculated for each cell group. HSC: Hematopoietic stem cell; Prog: Progenitor; early Ery: Early erythroid progenitor; late Ery: Late erythroid progenitor; pro B: Progenitor B cell; B: Mature B cell; Plasma: Plasma cell; T: Naïve T cell; CTL: Cytotoxic T Lymphocyte; NK: Natural Killer cell; GMP: Granulocyte-macrophage progenitor; pro Mono: Promonocyte; Mono: Monocyte; pDC: Plasmacytoid dendritic cell; cDC: Conventional dendritic cell.

Overall, the evaluation of cross-dataset classification performance revalidates the effectiveness of our discrimination matrix to predict large amount of cells with small dataset. The discrimination space embodies inherent information from discovery dataset, that is, it is capable to overcome batches or batch effects for new dataset. Meanwhile, our approach allows to accommodate more prior knowledge, making the results of great biological interpretation.

4. Discussion

Identifying genes (or metagenes) to discriminate different cell populations is effective to annotate large amount of cells within or across scRNA-seq data sets. The challenge is that cell states are temporally and spatially heterogeneous but always unknown in biological systems. While, existing methods usually adopt (i) label-based discrimination approaches, e.g., scID [31], which needs a reference annotation to identify cluster or cell type specific genes; or (ii) unsupervised dimensionality reductions, e.g., PCA, *t*-SNE, to learn hidden structures from unlabeled cells. Supervised discrimination methods can be used to infer unlabeled cells in validation datasets, but limited to the biological variances and cell compositions underlying the discovery dataset. While, unsupervised methods transform genes to metagenes, generally without the ability to transfer cell labels between different datasets. In fact, neither of the two types accounts for overall affinities inherent in multiple cell populations, thereby easily resulting in misclassification on new data. We demonstrated that our scDA method is able to successfully achieve powerful discriminant features in both unsupervised and supervised manners using eleven scRNA-seq data sets.

The notable advantage of scDA is that the model finds out discriminant features based on cell-by-cell representation graph. The graph quantitatively reveals both local (within cell groups) and global (across different groups) variations in data sets, and is robust to data heterogeneity. scDA features learnt from full representation configuration, accommodate affinity differences between any pairs of cells, and present more specific for cell discrimination. Moreover, scDA uses an alternately iterative solution to take advantage of the inherent correlations between samples and features. Such approach makes our model data driven and can correct the obtained information from data sets. The optimal results become reliable when considering the underlying heterogeneity across profiled cells.

We demonstrated that scDA accurately predicts plentiful unlabeled cells after obtaining discriminant metagenes and estimating potential cell types from a small amount of cells, such that unnecessary re-analysis can be avoided for large studies. The differences between estimation and prediction are proven quite small, suggesting that scDA is flexible to only implement the objective of recognizing cell populations for those small or finished studies. Moreover, the separable pipeline of scDA adopts spectral clustering and *K*NN classification in this work, but it also allows integrations of other useful approaches for scRNA-seq data, for example, hierarchical clustering or support vector machine algorithm. Theoretically, even the representation matrix of core model can be alternatively initialized (rather than *K*NN graph), e.g., with prior knowledge of cell groups, which helps to obtain more interpretable outcomes. Its potential scalability would deal with different practical issues and thus be well in conjunction with other information or analytical procedures [32].

CRediT authorship contribution statement

Qianqian Shi: Conceptualization, Methodology, Investigation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Project administration, Funding acquisition. **Xinxing Li:** Software, Validation. **Qirui Peng:** Data curation. **Chuanchao Zhang:** Conceptualization, Methodology, Writing - review & editing. **Luonan Chen:** Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant Nos. 61802141, 31930022 and 31771476), Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB38040400), Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01) and Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (Grant No. 2662017QD043).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.05.046.

References

[1] Han X, Wang R, Zhou Y, Fei L, Sun H, et al. Mapping the mouse cell atlas by microwell-seq. Cell 2018;172:1091.
[2] Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, et al. The human cell atlas. Elife 2017;6.
[3] Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 2014;344:1396–401.
[4] Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res 2014;42:8845–60.
[5] Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 2019;20:273–82.
[6] Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol 2019;20:194.
[7] Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 2018;50:96.
[8] Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods 2017;14:483–6.
[9] Barron M, Zhang S, Li J. A sparse differential clustering algorithm for tracing cell type changes via single-cell RNA-sequencing data. Nucleic Acids Res 2018;46:e14.
[10] Jean-Philippe B, Pablo T, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 2004;101:4164–9.
[11] Markus R. What is principal component analysis? Nat Biotechnol 2008;26:303.
[12] Barbara T, Brownfield DG, Wu AR, Neff NF, Mantalas GL, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 2014;509:371–5.
[13] Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol 2015;16:241.
[14] Laurens VDM, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9:2579–605.
[15] Chen X, Zhengchang S. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics 2015;31:1974–80.
[16] Chen J, Yang J. Robust subspace segmentation via low-rank representation. IEEE Trans Cybern 2014;44:1432–45.
[17] Zhuang L, Wang J, Lin Z, Yang AY, Ma Y, et al. Locality-preserving low-rank representation for graph construction from nonlinear manifolds. Neurocomputing 2015;175:715–22.
[18] Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 2018;36:421–7.
[19] Wai Keung W, Zhihui L, Jiajun W, Xiaozhao F, Yuwu L. Low-rank embedding for robust image feature extraction. IEEE Trans Image Process 2017;26:2905–17.
[20] Yang L. An overview of distance metric learning. Proc Computer Vision & Pattern Recognition 2007.
[21] Cox TF, Ferry G. Discriminant analysis using non-metric multidimensional scaling. Pattern Recogn 1993;26:145–53.
[22] Von Luxburg U. A tutorial on spectral clustering. Statist Comput 2007;17:395–416.
[23] Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning. Adv Neural Inform Process Syst 2006;19:41–8.
[24] Hagen L, Kahng AB. New spectral methods for ratio cut partitioning and clustering. IEEE Trans Comput Aided Des Integr Circuits Syst 1992;11:1074–85.
[25] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1999;20:53–65.
[26] Lachenbruch PA, Goldstein M. Discriminant analysis. Biometrics 1979:69–85.
[27] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50.
[28] Zhang X, Lan Y, Xu J, Quan F, Zhao E, et al. Cell marker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res 2019;47:D721–8.
[29] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–20.
[30] Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol 2020;21:12.
[31] Boufea K, Seth S, Batada NN. scID uses discriminant analysis to identify transcriptionally equivalent cell types across single-cell RNA-Seq data with batch effect. iScience 2020;23:100914.
[32] Shi Q, Zhang C, Peng M, Yu X, Zeng T, et al. Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. Bioinformatics 2017;33:2706–14.
[33] Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, et al. Multilineage communication regulates human liver bud development from pluripotency. Nature 2017;546:533–8.
[34] Goolam M, Scialdone A, Graham SJL, Macaulay IC, Jedrusik A, et al. Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. Cell 2016;165:61–74.

[35] Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet 2017;49:708–18.

[36] Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol 2014;32:1053–8.

[37] Lawlor N, George J, Bolisetty M, Kursawe R, Sun L, et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. Genome Res 2017;27:208–22.

[38] Segerstolpe A, Palasantza A, Eliasson P, Andersson EM, Andreasson AC, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metab 2016;24:593–607.

[39] Muraro MJ, Dharmadhikari G, Grun D, Groen N, Dielen T, et al. A single-cell transcriptome atlas of the human Pancreas. Cell Syst 2016;3:385–394.e383.

[40] Enge M, Arda HE, Mignardi M, Beausang J, Bottino R, et al. Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. Cell 2017;171:321–330.e314.

[41] Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst 2016;3:346–360.e344.

[42] van Galen P, Hovestadt V, Wadsworth Ii MH, Hughes TK, Griffin GK, et al. Single-Cell RNA-Seq reveals AML hierarchies relevant to disease progression and immunity. Cell 2019;176:1265–1281.e1224.