Bioinformatics, 37(19), 2021, 3263–3269
doi: 10.1093/bioinformatics/btab250
Advance Access Publication Date: 11 May 2021
Original Paper

Gene expression

MAT^2: manifold alignment of single-cell transcriptomes with cell triplets

Jinglong Zhang [1,2], Xu Zhang[2], Ying Wang [2,3], Feng Zeng[2,3,4,*] and Xing-Ming Zhao [1,5,6,*]

[1]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China, [2]Department of Automation, Xiamen University, Xiamen 361005, China, [3]Xiamen Key Laboratory of Big Data Intelligent Analysis and Decision, Xiamen University, Xiamen 361005, China, [4]Department of Neuroscience, School of Medicine, Xiamen University, Xiamen 361005, China, [5]MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, and MOE Frontiers Center for Brain Science, Shanghai 200433, China and [6]Research Institute of Intelligent Complex Systems, Fudan University, Shanghai 200433, China

*To whom correspondence should be addressed.
Associate Editor: Valentina Boeva

Received on January 25, 2021; revised on March 30, 2021; editorial decision on April 14, 2021; accepted on April 20, 2021

Abstract

Motivation: Aligning single-cell transcriptomes is important for the joint analysis of multiple single-cell RNA sequencing datasets, which in turn is vital to establishing a holistic cellular landscape of certain biological processes. Although numbers of approaches have been proposed for this problem, most of which only consider mutual neighbors when aligning the cells without taking into account known cell type annotations.

Results: In this work, we present MAT^2 that aligns cells in the manifold space with a deep neural network employing contrastive learning strategy. Compared with other manifold-based approaches, MAT^2 has two-fold advantages. Firstly, with cell triplets defined based on known cell type annotations, the consensus manifold yielded by the alignment procedure is more robust especially for datasets with limited common cell types. Secondly, the batch-effect-free gene expression reconstructed by MAT^2 can better help annotate cell types. Benchmarking results on real scRNA-seq datasets demonstrate that MAT^2 outperforms existing popular methods. Moreover, with MAT^2, the hematopoietic stem cells are found to differentiate at different paces between human and mouse.

Availability and implementation: MAT^2 is publicly available at https://github.com/Zhang-Jinglong/MAT2.

Contact: zengfeng@xmu.edu.cn or xmzhao@fudan.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Over the last decade, single-cell RNA sequencing (scRNA-seq) has attracted more and more attention for unbiased exploration of transcriptome variation at single-cell level (Tanay and Regev, 2017). Accordingly, scRNA-seq has been widely utilized for investigating cellular heterogeneity and transcriptional dynamics under various conditions, e.g. the development lineage of certain cells or tissues (Park *et al.*, 2020; Treutlein *et al.*, 2014; Zhong *et al.*, 2020). Generally, multiple scRNA-seq datasets may be generated by different labs for a same biological problem of interest, where each individual dataset may cover only a limited number of cell types. Therefore, the integration of multiple scRNA-seq datasets is a promising strategy for uncovering the heterogeneity of cellular compositions and the holistic biological process. However, it is a big challenge for jointly analyzing multiple scRNA-seq datasets that are generated by different labs due to technical variations or batch effects, etc.

The integration of scRNA-seq datasets can be regarded as an alignment problem (Cao *et al.*, 2020) where the individual datasets will be aligned against each other. When aligning multiple scRNA-seq datasets, identifying cellular correspondences across datasets is one of the most important steps. Accordingly, the existing approaches for aligning multiple scRNA-seq datasets can be grouped into two classes, i.e. cell alignment and cluster alignment approaches.

As typical cell alignment approaches, MNNCorrect (Haghverdi *et al.*, 2018), Seurat (Stuart *et al.*, 2019), Scanorama (Hie *et al.*, 2019) and BBKNN (Polański *et al.*, 2020) select mutual nearest neighbors (Haghverdi *et al.*, 2018) from different datasets as anchors based on their gene expression profiles or latent vectors. Recently, cell cluster approaches become popular with robust results of alignment, where cell clusters instead of cells are used for alignment at the population level. Among them, scMerge (Lin *et al.*, 2019) looks for mutual nearest neighbors between cell clusters, and Harmony (Korsunsky *et al.*, 2019) maximizes the mixing of cells within clusters through soft clustering. Although the above methods

perform well when aligning multiple datasets composed of common cell types, they may fail to work on those datasets with few shared cell types. In addition, most of existing approaches work in unsupervised way by looking for cells with similar gene expression profiles as same cell types, whereas the annotation of known cell types has not been fully utilized.

In this work, we present a novel approach, namely manifold alignment of single-cell transcriptomes with cell triplets (MAT^2), to align multiple scRNA-seq datasets in their latent manifold space. Compared with existing methods that only consider cells with similar gene expression profiles as positive anchors across datasets, MAT^2 also takes into account pairs of cells with functional difference as negative anchors by employing contrastive learning (Chen *et al.*, 2020; Hoffer and Ailon, 2015), and composes the cells and the cells forming the positive and negative anchors into cell triplets to guide the alignment in a discriminative way. With contrastive learning, MAT^2 is able to take into account of the prior knowledge of cell types and define a more reasonable manifold space for cellular transcriptomes, where the cells of the same type will stay closer. Moreover, by reconstructing both consensus and batch-specific matrices from the latent manifold space, MAT^2 can be used to recover the batch-effect-free gene expression that can be used for downstream analysis. When benchmarking on real scRNA-seq datasets, MAT^2 outperforms other popular approaches with robust alignment results for cell type annotation and performs especially well over datasets with few shared cell types. We further integrated the human and mouse embryo datasets, verifying that HSC-primed hemogenic endothelial cells (HECs) develop at different paces in human and mouse.

2 Materials and methods

2.1 Datasets and pre-processing

As shown in Table 1, we collected nine datasets and grouped them into four datasets according to tissues, each of which contains only the genes shared by all batchs within every dataset. All datasets were downloaded from NCBI's Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) and EBI's ArrayExpress database (https://www.ebi.ac.uk/arrayexpress/). References for the datasets are available in Supplementary Table S1.

For each dataset, we will first select the top 2000 highly variable genes as features through variance stabilize transformation (vst) in Seurat (Stuart *et al.*, 2019), then normalize the gene expression profile of each cell as shown in equation (1).

$$\boldsymbol{X} = \boldsymbol{X}_0 \cdot diag(\boldsymbol{s})^{-1} \tag{1}$$

where $\boldsymbol{X}_0 = [\boldsymbol{x}_{01}, \boldsymbol{x}_{02}, \ldots, \boldsymbol{x}_{0N}]$ denotes the original dataset with N cells, G genes and $\boldsymbol{x}_{0i}$ denotes the original gene expression profile of cell i. And $\boldsymbol{s} = [s_1, s_2, \ldots, s_N]^T$ denotes the size factors for every cell with $s_i = ||\boldsymbol{x}_{0i}||^1/G$. The normalized dataset $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$ is used as the input of MAT^2, where $\boldsymbol{x}_i \in \mathbb{R}^G$ represents the normalized expression profile of cell i.

2.2 Manifold alignment of single-cell transcriptomes with cell triplets

As shown in Figure 1, MAT^2 aligns multiple scRNA-seq datasets in a manifold space. Briefly, MAT^2 consists of the following steps: (a) Each gene expression matrix of a scRNA-seq dataset will be transformed into a manifold space with the utility of encoder, and the cells will be aligned against each other based on their gene expression with the help of contrastive learning. In contrastive learning, for a cell of interest denoted as C, a cell from the same type but different datasets denoted as C_p and a cell from different types denoted as C_n will be considered, and a triplet (C, C_p, C_n) will be formed. A consensus manifold will be achieved for all datasets, where C and C_p will stay close while C and C_n will be separated from each other (Fig. 1a); (b) For each cell type with its consensus manifold, two decoders ($\mathcal{D}$ and $\mathcal{R}$) will be built to recover the deviation of cellular transcriptomes from the consensus manifold and the consensus transcriptome across multi-datasets, respectively (Fig. 1b). The former ($\mathcal{D}$) reflects the technology- or platform-specific effect on gene expression, while the latter ($\mathcal{R}$) reveals the cell-type-specific regulation of transcriptome. Therefore, MAT^2 can be regarded as a decomposition of the original cellular transcriptome data into a consensus gene expression matrix and a batch deviation matrix; and (c) With the cell-type-specific gene expression matrix, the downstream analysis, e.g. trajectory analysis and differential expression analysis can be performed (Fig. 1c).

2.2.1 Alignment of single-cell manifolds

MAT^2 trains an encoder neural network to learn a nonlinear mapping function to map normalized dataset $\boldsymbol{X}$ into a low dimensional latent space $\boldsymbol{Z} = [z_1, z_2, \ldots, z_N]$ with $z_i \in \mathbb{R}^K$ ($K \ll G$). We define a positive anchor as a pair of cells of the same type but from different datasets, while a negative anchor as a pair of cells of different types. The goal of alignment is to minimize the distance between cells in a positive anchor and maximize that in a negative anchor in the latent space $\boldsymbol{Z}$. For this, the contrastive learning is employed with the following objective function.

$$\mathcal{L}(\boldsymbol{Z}) = \sum_{i=1}^{N} \sum_{j \in C_i^p} \sum_{k \in C_i^n} T(z_i, z_j, z_k, \alpha_{ijk}) \tag{2}$$

where $T(\cdot)$ denotes the triplet loss as shown in equation (3).

Table 1. Statistics of the four datasets

	Platform and number of cells	Number of cell types	Cell state	Shared cell types	Difference between datasets	Accession number
Retina	All by Drop-seq • Batch 1: 13660 • Batch 2: 12825	19, 19 (19 in total)	Mature	100%	Different batches	GSE81904
Pancreas	• CelSeq: 1004 • CeqSeq2: 2285 • Fluidigm C1: 638 • SMART-seq2: 2394	13, 13, 13, 13 (13 in total)	Mature	100%	Different platforms	• GSE81076 • GSE85241 • GSE86469 • E-MTAB-5061
Hematopoietic	• SMART-seq2: 1494 • MARS-seq: 2699	6, 3 (6 in total)	Intermediate	50%	Different platforms and developmental stages	• GSE81682 • GSE72857
Embryo	All by STRT-seq • human: 528 • mouse: 597	5, 6 (6 in total)	Intermediate	83.3%	Different species	• GSE135202 • GSE139389

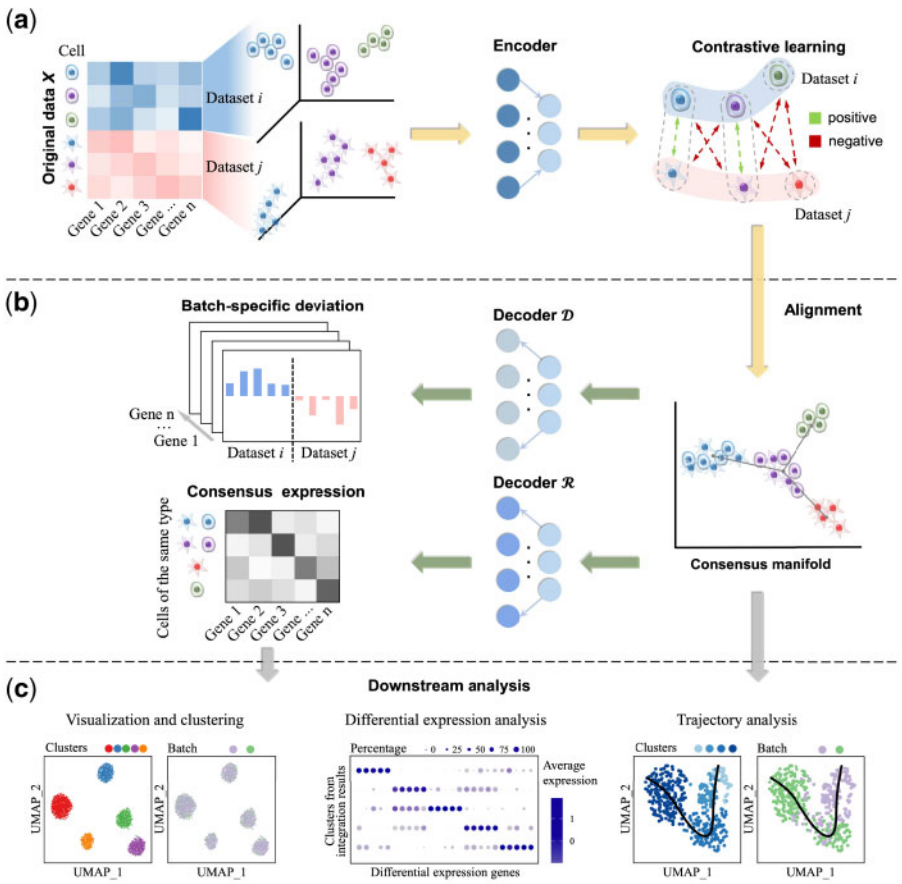

Fig. 1. The schematic view of MAT^2. (**a**) The gene expression matrix of each dataset is transformed into a latent consensus manifold, and the manifold alignment of multiple datasets is achieved at the same time. (**b**) With the manifold, two decoders are employed to divide the original gene expression matrix into a consensus expression matrix and a batch-specific deviation matrix. (**c**) The manifold and consensus gene expression can be used for downstream analysis, such as trajectory analysis and differential expression analysis

$$T(\boldsymbol{a},\boldsymbol{p},\boldsymbol{n},\alpha)=\begin{cases}||\boldsymbol{a}-\boldsymbol{p}||-||\boldsymbol{a}-\boldsymbol{n}||+\alpha, & \text{if } T>0\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

and C_i^p denotes a set of cells that can form a positive anchor with cell i, and C_i^n denotes a set of cells that can form a negative anchor with cell i. For any cell triplet (i, j, k) with $j \in C_i^p$ and $k \in C_i^n$, the distance between cells i and k is expected to be significantly larger than that between cells i and j. In supervised settings, MAT^2 utilizes cell type annotations to construct positive and negative anchors, where α_{ijk} is set to 1.0 for any triplet (i, j, k). In unsupervised settings, for an anchor (i, j) generated by Seurat with a score s_{ij}, MAT^2 randomly selects a cell k to form a triplet (i, j, k) with $\alpha_{ijk} = (s_{ij} + 1)/2$. By combining the triplets formed with cells of known types and those formed with anchors generated by Seurat based on unlabeled cells, MAT^2 can be easily extended to work in semi-supervised mode as shown in Section 3.2.

With the above objective function, an encoder neural network composed of fully connected layers is trained with Pytorch and optimized with Adam (Kingma and Ba, 2014), where the input size is equal to the number of genes and the output size is set to 20 as default. The neural network uses rectified linear unit (ReLU) (Glorot *et al.*, 2011) as the activation function in the two hidden layers and linear activation function for the output layer. Two more operations were adopted during learning, including L2 regularization with a ratio of 0.01 and dropout (Srivastava *et al.*, 2014) with a ratio of no more than 0.3. The mini-batch stochastic gradient descent (Cotter *et al.*, 2011) is used for optimization by randomly selecting triplets for training, where a mini-batch contains 256 cell triplets.

2.2.2 Reconstruction of gene expression

After alignment in the manifold space, MAT^2 constructs two decoders f^R and f^D to learn the mapping functions from optimized latent space $\hat{\boldsymbol{Z}} = [\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_N]$ to gene expression space. The consensus gene expression $\boldsymbol{X}^r = [\boldsymbol{x}_1^r, \boldsymbol{x}_2^r, \ldots, \boldsymbol{x}_N^r]$ with $\boldsymbol{x}_i^r = f^R(\hat{z}_i)$ was built through the decoders f^R, and the batch-specific deviation $\boldsymbol{X}^d = [\boldsymbol{x}_1^d, \boldsymbol{x}_2^d, \ldots, \boldsymbol{x}_N^d]$ with $\boldsymbol{x}_i^d = f^D(\hat{z}_i, b_i)$ was built through decoders f^D, where $\boldsymbol{b} = [b_1, b_2, \ldots, b_N]^T$ represents the batch to which the cells belong. Here, the reconstruction error for cell i is defined as mean square error as shown in equation (4).

$$E_i = ||\boldsymbol{x}_i - \boldsymbol{x}_i^r - \boldsymbol{x}_i^d||^2 \tag{4}$$

When training the decoders, contrastive learning is utilized to force f^D to describe the batch-specific deviation. We choose the same batch of cells as cell i to form B_i^+, and the different batch of cell i to form B_i^-. Then the objective function for two decoders is shown in equation (5).

$$\mathcal{L}(\boldsymbol{X}^r, \boldsymbol{X}^d) = \sum_{i=1}^{N} \sum_{p \in B_i^+} \sum_{q \in B_i^-} \{T(\boldsymbol{x}_i^d, \boldsymbol{x}_p^d, \boldsymbol{x}_q^d, \beta) + E_{i,p,q}\} \tag{5}$$

where the value of β is 1.0, and $E_{i,p,q}$ denotes the average value of reconstruction error for cells i, p and q.

The two decoders used for reconstruction adopt the same training strategy as the encoder for alignment. The difference between the two decoders is that f^R and f^D are single-layer and two-layer fully connected networks, respectively, where ReLU is used as activation function of the output layer in f^R to obtain non-negative gene expression.

3 Results

3.1 Benchmark results on pancreas and retina scRNA-seq datasets

We evaluated the performance of MAT^2 and compared it with the state-of-the-art methods on retina datasets (Shekhar *et al.*, 2016) and pancreas datasets (Grün *et al.*, 2016; Lawlor *et al.*, 2017; Muraro *et al.*, 2016; Segerstolpe *et al.*, 2016) (see Table 1). Here, we selected four methods for comparison, including two state-of-the-art methods, namely Seurat (version 3) (Stuart *et al.*, 2019) and Harmony (Korsunsky *et al.*, 2019), as reported recently (Tran *et al.*, 2020), scMerge (Lin *et al.*, 2019) that can work in both supervised and unsupervised modes, and a neural network model named scANVI (Xu *et al.*, 2021) that can work in supervised mode with provided cell type annotations. We noticed that a similar approach named INSCT (Simon *et al.*, 2020) was proposed when this manuscript was drafted, which also uses cell triplets for alignment, and we also compared MAT^2 against INSCT here. Since MAT^2, INSCT and scMerge can work in supervised and unsupervised settings, we used MAT^2-s and MAT^2-u to respectively denote MAT^2 working in these two settings, and the same for INSCT-s, INSCT-u, scMerge-s and scMerge-u. The same inputs to MAT^2 were also given to scANVI so that it can work in supervised setting. Once the multiple datasets were aligned into a single dataset in the gene expression space or latent manifold space by a certain method, Louvain (Blondel *et al.*, 2008) was employed at different resolutions from 0 to 0.5 at interval of 0.01 to cluster cells to determine the cell types, where each cluster represents a cell type. The adjusted rand index (ARI) (Hubert and Arabie, 1985) was utilized to check the accuracy of cell types assigned by clustering. For each method, the maximum ARI calculated based on the clustering results of different resolutions was used as the ARI of this method. Furthermore, local inverse Simpson's index (LISI) (Korsunsky *et al.*, 2019) was adopted to assess the degree of dataset mixing for each aligned dataset.

Figure 2 shows the results by the nine approaches over the benchmark datasets. Overall, MAT^2, especially MAT^2-s, achieved the best results when assigning cells to corresponding cell types with the highest ARI over the two datasets compared with the other methods (Fig. 2a, Supplementary Tables S2 and S3). For dataset mixing, the LISIs of MAT^2 were among the best on these two datasets, proving its ability to effectively integrate datasets from distinct batches and platforms. Although INSCT-s, scMerge-s and scANVI also work in supervised way, MAT^2-s significantly outperformed scMerge-s by 47.1%, INSCT-s by 21.0% and scANVI by 19.4% with respect to LISI on average, indicating the effectiveness of MAT^2 when integrating multiple datasets.

To better demonstrate the performance of MAT^2, Figure 2b and c show the visualization results of pancreas datasets by UMAP (Becht *et al.*, 2019) without or with integration, respectively (the results on retina datasets can be found in Supplementary Fig. S1). From the results, it can be seen that all the methods except for Harmony performed very well when separating distinct cell types (Fig. 2a, right and Fig. 2c, top), where MAT^2-s outperformed other approaches while MAT^2-u performed comparably well with Seurat. We also noticed that MAT^2, Seurat and Harmony perform well with respect to LISI (Fig. 2c, bottom), indicating the effectiveness of these approaches to integrate multiple datasets. Surprisingly, scMerge showed poor performance with respect to LISI, implying insufficient dataset mixing of the same cell type, which is consistent with previous reports that scMerge may fail to work in some cases (Tran *et al.*, 2020). Although we found a structure similar to MAT^2 in INSCT, the actual performance of the two is significantly different in the above dataset. In order to further make a fair comparison, we also tested MAT^2 and INSCT with the datasets used in Simon *et al.* (2020) (Supplementary Fig. S2). Whether using ARI or LISI for metrics, MAT^2 had obvious superiority.

Since MAT^2, INSCT and scANVI can work in supervised setting and show best performance in above results, we investigated the effect of the number of training cells on the performance of MAT^2-s, INSCT-s and scANVI. Figure 2d shows the ARI results of MAT^2-s, INSCT-s and scANVI with the percentage of cells used for training from 100% to 5%. With the number of training cells decreasing, MAT^2-s showed robust performance with only 3.8% and 4.0% reductions of ARIs on retina and pancreas, respectively (Supplementary Table S4). INSCT-s and scANVI had a significant decrease in ARI over the pancreas and retina datasets, respectively. It can be seen that MAT^2-s trained with few (hundreds) samples still worked efficiently, which is especially important for large datasets with very few known cell types.

When integrating multiple scRNA-seq datasets, computation consumption is also the issue to be concerned. We compared the nine approaches with respect to their computation time and memory usage with a Linux server equipped with 12-core AMD Ryzen Threadripper 1920X, 125 GB RAM and GeForce RTX 2080 Ti. The running time and memory usage of the nine approaches on the cell sets composed of various numbers of cells sampled from pancreas datasets were respectively shown in Figure 2e and f, where the time of reading data was not recorded. The CPU time for all the approaches considered here with or without GPU can be found in Supplementary Table S5, and the detailed memory usage can be found in Supplementary Table S6. Note that it takes more time for the additional step of reconstructing gene expression in MAT^2. Even so, from the results, it can be seen that MAT^2 performed comparatively well with popular approaches. The results also indicate that MAT^2 has the potential to integrate millions of cells on personal computers.

3.2 MAT^2 aligns cellular states along developmental lineage

The scRNA-seq technology has been widely used for tracing cell lineage during developmental procedure, where various datasets may be generated for the same procedure. Here, different approaches were applied to align two hematopoietic datasets (Nestorowa *et al.*, 2016; Paul *et al.*, 2015) with obvious batch effect as shown in Supplementary Fig. S3a. The whole differentiation procedure contains six cell types, i.e. long-term HSC (LT-HSC), megakaryocyte/erythroid progenitor (MEP), granulocyte/monocyte progenitor (GMP), common myeloid progenitor (CMP), multipotent progenitor (MPP) and lymphoid multipotent progenitor (LMPP).

The UMAP visualization results based on the aligned datasets generated by MAT^2, scMerge, INSCT, Seurat, Harmony and scANVI were shown in Supplementary Fig. S3. With cell type annotations, MAT^2-s could successfully identify most cell types with the help of contrastive learning, and performed especially well to prevent LT-HSC, MPP and LMPP that only occur in a single dataset from being confused with the other three types. scMerge-s could successfully identify most of the cells belonging to CMP, GMP and MEP shared between datasets, but failed to identify the other three cell types. INSCT-s was difficult to integrate these two hematopoietic datasets. When cell type annotations were not used, the unsupervised approach including MAT^2-u could identify MEP and GMP cells while it failed to separate the other four cell types. In particular, a large number of anchors describing the correspondence between cells (e.g. 39.5% for Seurat) link LT-HSC, MPP and LMPP to CMP, which made it difficult to separate the four cell types. When looking at the results of cell annotations with respect to ARI, we noticed that MAT^2-s performed best on assigning cell types with ARI of 0.922, surpassing the second ranked scMerge-s by 18.1%.

To evaluate whether gene expression reconstructed by MAT^2 conserves the original biological signals, we clustered the cells in the reconstructed gene expression space and performed the differential expression analysis (P-value = 0.01) with Seurat R package (Stuart *et al.*, 2019) (Supplementary Fig. S4). In the results of MAT^2-s, the differentially expressed genes (DEGs) with the same cell types had similar expression profiles between datasets. scMerge-s also produced an accurate integration of the cells of CMP, GMP and MEP in two datasets and generated the corrected gene expression. However, the expression profiles of the DEGs derived from the scMerge-s' gene expression were significantly different between datasets. Therefore, MAT^2 not only placed cells with the same cell types together but also could get rid of batch effect to recover the

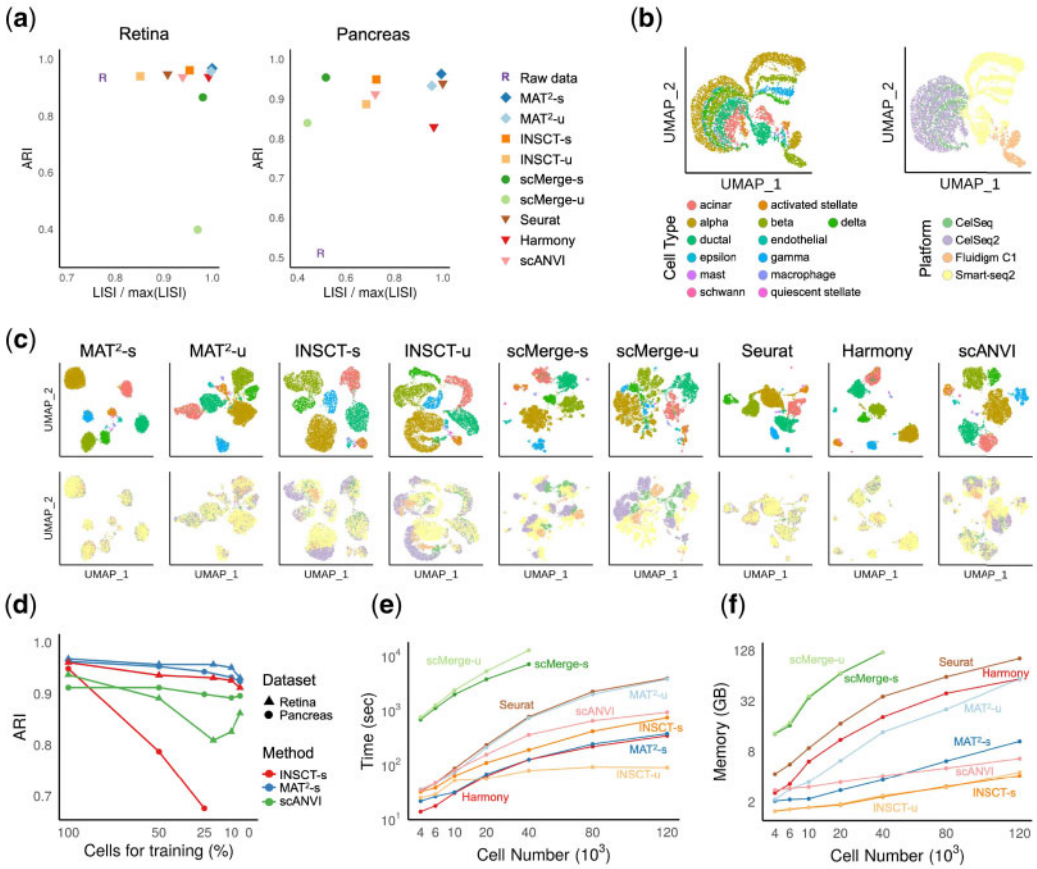

Fig. 2. Benchmarking of MAT^2 and other the state-of-the-art methods on retina and pancreas datasets. (**a**) The results of cell type assignment (ARI) and dataset mixing (LISI) on retina and pancreas datasets by the nine methods. (**b**) Visualization of the pancreas dataset with respect to both cell types (left) and platforms (right), where the pancreas dataset has significant platform difference. (**c**) Visualization of the results of the nine methods using the same color in **b** to mark cell types (upper) and platforms (lower). (**d**) The ARI results on the integrated retina and pancreas datasets by MAT^2-s, INSCT-s and scANVI with the proportion of training cells ranging from 100% to 5%. (**e** and **f**) The computation time and memory usage of the nine methods on pancreas datasets with increased sampling sizes

true biological signals. Considering the transition between cell types is a continuous procedure, we extended MAT^2 to the semi-supervised mode, denoted as MAT^2-semi hereafter. MAT^2-semi first forms the cell triplets based on labeled cells in the same way as that of MAT^2-s, and then mixes them with the cell triplets that have been used in MAT^2-u for training. In this way, MAT^2-semi can utilize known cell types and anchor scores for cells without labels. We evaluated MAT^2-semi on hematopoietic datasets by keeping 10% of the cell type annotations as labeled. MAT^2-semi was able to show the transition of gene expression patterns from CMPs to MEPs more obviously (see Supplementary Figs S3 and S4) with only few cell type annotations. At the same time, the DEGs found based on the results of MAT^2 were consistent with the marker genes reported in the literature. For example, both *Car1* and *Car2* were significantly upregulated in MEP, which is consistent with previous reports that these two genes are marker genes of erythroid progenitors (Paul *et al.*, 2015). In addition, *Dntt* was upregulated in MPP and LMPP, which was confirmed by a previous work that regarded it as a marker gene in both MPP and lymphoid lineages (Herman *et al.*, 2018). The above results demonstrate that the consensus gene expression of MAT^2 is effective for aligning cellular states along developmental lineage in datasets that are interfered by batch effects.

3.3 MAT^2 infers species-specific cell lineage during HSC-primed HEC development

In this section, we investigated the performance of MAT^2 when aligning two scRNA-seq datasets of HSC-primed HEC development from human (Zeng *et al.*, 2019) and mouse (Hou *et al.*, 2020) embryos (Table 1). The human and mouse datasets contain cells generated from the aorta-gonad-mesonephros (AGM) tissues of embryos during the equal period. According to previous annotations (Hou *et al.*, 2020), the mouse cells annotated with the five cell types involved in the mouse HEC development were used as training samples, where the cell types consist of early and late arterial endothelial cell (aEC), hematopoietic cell (HC), HEC and venous endothelial cell (vEC). Furthermore, the non-EC negative control cells (Neg) from mouse dataset were used as control. With the clustering results over the aligned dataset by MAT^2, all the cells will be grouped into six clusters with each cluster annotated with one of the six above cell types. Therefore, the human AGM-associated endothelial cells will be annotated as one of the five cell types, i.e. the early aECs, late aECs, HCs, vECs and HECs.

We first looked at the gene expression reconstructed from the results of aligned human and mouse data. As shown in Figure 3a, the reconstructed gene expression was significantly correlated with the original one (Pearson correlation, human, $R = 0.69$, $P < 2.2e{-}16$; mouse, $R = 0.72$, $P < 2.2e{-}6$), indicating the effectiveness of MAT^2. Next, we performed the tSNE (Maaten and Hinton, 2008) analysis on the reconstructed gene expressions as shown in Figure 3b. From the results, we can see that the five cell types can be well separated from each other and are far from the negative controls (Neg). In addition, we looked at the marker genes specifying distinct subpopulations of endothelial cells in both human and mouse and wanted to see whether these marker genes have conserved expression patterns between mouse and human. We noticed

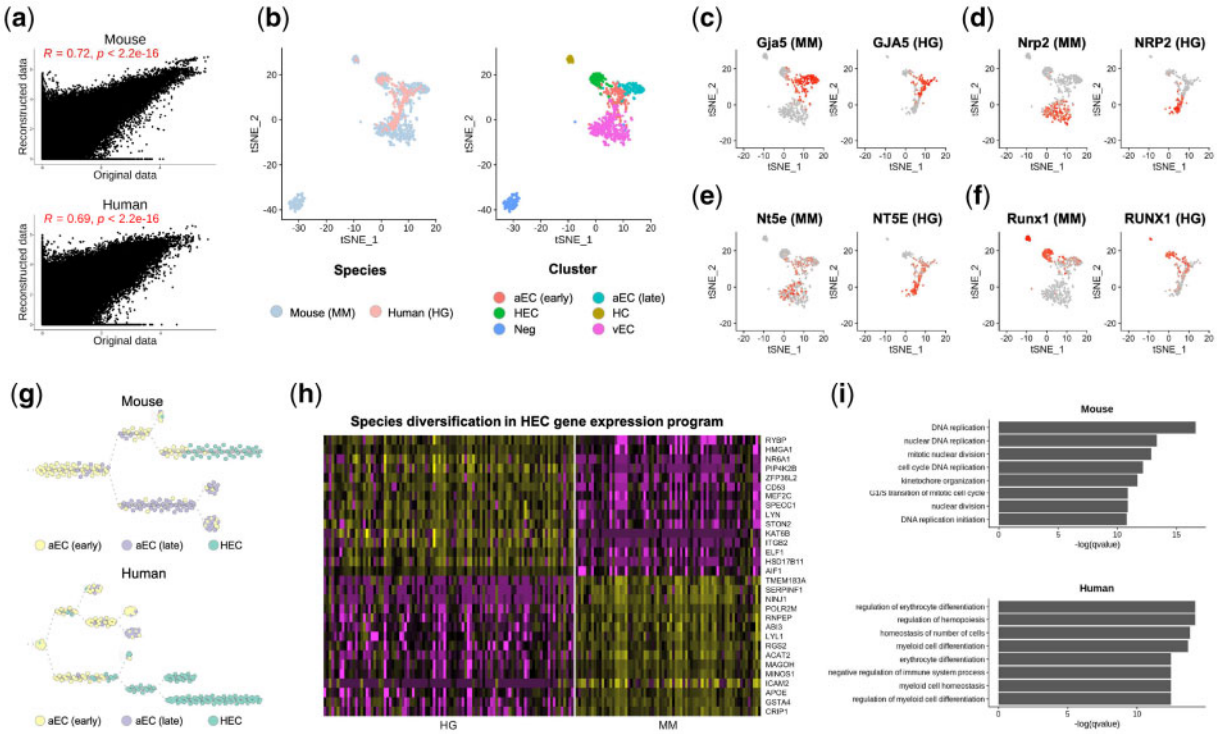

Fig. 3. Alignment of cells during human and mouse HSC-primed HEC development with MAT^2. (**a**) Pearson correlation between original gene expression data and reconstructed one. (**b**) tSNE visualization of mouse and human cells from AGM area, which were colored according to species (left) or cell clusters (right). (**c–f**) The expression of marker genes in human (HG) and mouse (MM). (**g**) The trajectory analysis results from aECs to HECs in mouse and human. (**h**) The heatmap of genes differentially expressed between human and mouse HSC-primed HECs. (**i**) The functional terms enriched in those genes from biological process

that four genes have similar expression patters in human and mouse cells, where *GJA5* was highly expressed in aECs, *NRP2* was highly expressed in vECs, *NT5E* was highly expressed in vECs, and *RUNX1* was highly expressed in HCs and HECs (Figure 3c–f). For example, the transcription factor *RUNX1* was highly expressed in both human and mouse HECs (Fig. 3f), which is coincident with the observation that the upregulation of *RUNX1* marks the initiation of the endothelial-to-hemogenic transition (EHT) and priming of endothelial cells toward HECs (Hou *et al.*, 2020).

Considering species specificity, we further inferred the trajectories from aECs to HECs with ti_slingshot (Street *et al.*, 2018) in Dyno (Saelens *et al.*, 2019) based on the reconstructed transcriptome for human and mouse, respectively (Fig. 3g). It was a consensus that the early aECs differentiated into the late aECs and HECs in human and mouse data. However, the developmental paces of HECs were different in two species. In human, the early aECs differentiated to HECs ahead of the differentiation toward the late aECs (Fig. 3g bottom). In contrast, in mouse, the early aECs first differentiated to the late aECs and then HECs (Fig. 3g, top). With genes differentially expressed between human and mouse HECs identified with MAST (Finak *et al.*, 2015) ($P < 0.01$), the discrepancy of gene expression programs between human and mouse HECs was revealed as shown in Figure 3h. The upregulated genes in human HECs included *HMGA1*, *CD53* and *ITGB2*, which marked the development of HCs. Furthermore, the gene ontology (GO) analysis of biological processes performed by clusterProfiler (Yu *et al.*, 2012) R package confirmed that the human HEC-specific genes were enriched in hematopoietic lineage differentiation (Fig. 3i). On the other hand, the upregulated genes in mouse HECs were intensively involved in cell division and disjunction. Both the trajectory and functional analysis indicated that two commitment events of the early aECs happened at different temporal orders in human and mouse embryos respectively, which is consistent with earlier speculation of Hou *et al.* (2020). The above results indicate that MAT^2 can help researchers align scRNA-seq datasets of different species to study species-specific gene expression.

4 Discussion

The joint analysis of multiple scRNA-seq datasets can help elucidate the comprehensive landscape of cellular compositions of certain tissues or processes. In this work, we present a novel approach, MAT^2, for aligning cells from multiple datasets by taking into account negative anchors with contrastive learning, which outperforms other popular approaches over several real datasets. Except for cell alignment in the manifold space, MAT^2 is able to reconstruct the consensus gene expression profile for a certain cell type, which in turn can be used for downstream analyses, e.g. trajectory analysis and differential gene expression analysis. We showcased that the transcriptome reconstructed by MAT^2 can help reveal the gene expression conservation and discrepancy between human and mouse AGM-associated HSC-primed HECs.

The good performance of MAT^2 is attributed to contrastive learning, which utilizes both positive and negative anchors when integrating multiple scRNA-seq datasets. In contrast, both existing cell alignment and cluster alignment approaches generally rely on positive anchors when aligning single-cell transcriptomes, whereas reliable positive anchors may be not always available especially for datasets with unshared cell types. The superiority of using negative anchors is that it can prevent distinct cell types from being mapped to the same ones, thereby improving the stability of scRNA-seq data integration, which is especially helpful for identifying rare cell types or inferring cell lineages.

At present, the rapid development of multi-modal data makes it a practical need to integrate these data for analysis. Compared with using a single modality to identify cell states, multi-modal data can provide more complete information to reveal the nuances of cell

types, thus enhancing the understanding of regulatory mechanisms. Existing methods have made some attempts in this direction. For example, Seurat (Stuart *et al.*, 2019) uses scRNA-seq data to transfer cell types to scATAC-seq data, and LIGER (Welch *et al.*, 2019) uses scRNA-seq and DNA methylation data to jointly identify cell types. Next, we envisage extending MAT^2 to single-cell multi-modal data to gain a deeper understanding of the cell state and its transcriptional regulation mechanism.

Funding

This work was partly supported by the National Key R&D Program of China (2020YFA0712403, 2018YFC0910500), National Natural Science Foundation of China (61932008, 61772368, 61503314), Shanghai Science and Technology Innovation Fund (19511101404), Shanghai Municipal Science and Technology Major Project (2018SHZDZX01) and Natural Science Foundation of Fujian Province, China (2019J01041).

Conflict of Interest: none declared.

References

Becht,E. *et al.* (2019) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*, **37**, 38–44.

Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **2008**, P10008.

Cao,Z.-J. *et al.* (2020) Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. *Nat. Commun.*, **11**, 3458.

Chen,T. *et al.* (2020) A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv: 2002.05709*.

Cotter,A. *et al.* (2011) Better mini-batch algorithms via accelerated gradient methods. In Proceedings of the 24th International Conference on Neural Information Processing Systems, 1647–1655.

Finak,G. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.

Glorot,X. *et al.* (2011) Deep sparse rectifier neural networks. *J. Mach. Learn. Res.*, **15**, 315–323.

Grün,D. *et al.* (2016) De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*, **19**, 266–277.

Haghverdi,L. *et al.* (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, **36**, 421–427.

Herman,J.S. *et al.* (2018) FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods*, **15**, 379–386.

Hie,B. *et al.* (2019) Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.*, **37**, 685–691.

Hoffer,E. and Ailon,N. (2015) Deep metric learning using triplet network. In: Feragen A., Pelillo M., Loog M. (eds) *Similarity-Based Pattern Recognition*. Cham. Springer International Publishing, https://doi.org/10.1007/978-3-319-24261-3_7. pp. 84–92.

Hou,S. *et al.* (2020) Embryonic endothelial evolution towards first hematopoietic stem cells revealed by single-cell transcriptomic and functional analyses. *Cell Res.*, **30**, 376–392.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

Kingma,D.P. and Ba,J. (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv: 1412.6980*.

Korsunsky,I. *et al.* (2019) Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods*, **16**, 1289–1296.

Lawlor,N. *et al.* (2017) Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome Res.*, **27**, 208–222.

Lin,Y. *et al.* (2019) scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc. Natl. Acad. Sci. USA*, **116**, 9775–9784.

Maaten,L. v d. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

Muraro,M.J. *et al.* (2016) A single-cell transcriptome atlas of the human pancreas. *Cell Syst.*, **3**, 385–394.e3.

Nestorowa,S. *et al.* (2016) A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, **128**, e20–e31.

Park,J.-E. *et al.* (2020) A cell atlas of human thymic development defines T cell repertoire formation. *Science*, **367**, eaay3224.

Paul,F. *et al.* (2015) Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, **163**, 1663–1677.

Polański,K. *et al.* (2020) BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics*, **36**, 964–965.

Saelens,W. *et al.* (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, **37**, 547–554.

Segerstolpe,Å. *et al.* (2016) Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.*, **24**, 593–607.

Shekhar,K. *et al.* (2016) Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, **166**, 1308–1323.e30.

Simon,L.M. *et al.* (2020) INSCT: Integrating millions of single cells using batch-aware triplet neural networks. *bioRxiv*, 2020.05.16.100024. doi: 10.1101/2020.05.16.100024.

Srivastava,N. *et al.* (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.

Street,K. *et al.* (2018) Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, **19**, 477.

Stuart,T. *et al.* (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.e21.

Tanay,A. and Regev,A. (2017) Scaling single-cell genomics from phenomenology to mechanism. *Nature*, **541**, 331–338.

Tran,H.T.N. *et al.* (2020) A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.*, **21**, 12.

Treutlein,B. *et al.* (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, **509**, 371–375.

Welch,J.D. *et al.* (2019) Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, **177**, 1873–1887.e17.

Xu,C. *et al.* (2021) Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.*, **17**, e9620.

Yu,G. *et al.* (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS J. Integr. Biol.*, **16**, 284–287.

Zeng,Y. *et al.* (2019) Tracing the first hematopoietic stem cell generation in human embryo by single-cell RNA sequencing. *Cell Res.*, **29**, 881–894.

Zhong,S. *et al.* (2020) Decoding the development of the human hippocampus. *Nature*, **577**, 531–536.