

AIsceEA: Unsupervised Integration of Single-cell Gene Expression and Chromatin Accessibility via Their Biological Consistency

Elham Jafari¹, Travis Johnson², Yue Wang³, Yunlong Liu³, Kun Huang², and Yijie Wang¹

¹ Luddy School of Informatics, Computing, and Engineering, Indiana University, 700 N. Woodlawn Avenue, Bloomington, IN 47408, United States, USA

² Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN 46202, USA

³ Department of Medical & Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA
yijwang@iu.edu

Abstract. Since the integrative analysis of single-cell gene expression and chromatin accessibility measurements is essential for revealing gene regulation at the single-cell resolution, integrating these two measurements becomes one of the key challenges in computational biology. Because gene expression and chromatin accessibility are measurements from different modalities, no common features can be directly used to guide their integration. Current state-of-the-art methods assume that the number of cell types across the measurements is the same. However, when cell-type heterogeneity exists, they might not generate reliable results. Furthermore, current methods do not have an effective way to select the hyper-parameter under the unsupervised setting. Therefore, applying computational methods to integrate single-cell gene expression and chromatin accessibility measurements remains difficult.

We introduce AIsceEA – **A**lignment-based **I**ntegration of single-cell gene **E**xpression and chromatin **A**ccessibility – a computational method that integrates single-cell gene expression and chromatin accessibility measurements using their biological consistency. AIsceEA first defines a ranked similarity score to quantify the biological consistency between cell types across measurements. AIsceEA then uses the ranked similarity score and a novel permutation test to identify the cell-type alignment across measurements. For the aligned cell types, AIsceEA further utilizes graph alignment to align the cells across measurements. We compared AIsceEA with the competing methods on several benchmark datasets and demonstrated that AIsceEA is more robust to hyper-parameters and can better handle the cell-type heterogeneity problem. Furthermore, we demonstrate that AIsceEA significantly outperforms the state-of-the-art methods when integrating real-world SNARE-seq and scMultiome-seq datasets in terms of integration accuracy.

1 Introduction

Advances in single-cell high-throughput technologies have enabled us to profile gene expression and chromatin accessibility at the single-cell resolution [1–11]. Integration of the single-cell gene expression and chromatin accessibility measurements shed light on revealing gene regulation for specific cells [12–16]. However, the heterogeneity among single cells presents challenges for such integration [15]. Single-cell gene expression and single-cell chromatin accessibility measure the cells at the transcriptomic and epigenomic layers, respectively. When both measurements are profiled independently, it is difficult to identify the cell-type or cell-cell correspondences across the measurements since they exist in heterogeneous cellular modalities and lack any shared features to integrate them. Single-cell dual-omics sequencing technologies [17, 18] have been developed to tackle this problem by simultaneously profiling gene expression and chromatin accessibility for the same cells. However, most available single-cell gene expression and chromatin accessibility datasets are still profiled independently. Therefore, a reliable computational method is needed to integrate these two kinds of single-cell measurements from different modalities.

Several unsupervised integrative methods have been developed to integrate the single-cell gene expression and chromatin accessibility measurements [19–27]. CoupleNMF [19] unitizes the non-negative matrix factorization framework to integrate the single-cell gene expression and chromatin accessibility measurements at the cell type level. Other state-of-the-art methods focus on the integration at the cell-cell level. They assume that single-cell gene expression and chromatin accessibility measurements share similar low-dimensional manifolds and apply different computational methods to align the corresponding manifolds. MMD-MA aligns the manifold of single-cell gene expression profile and the manifold of single-cell chromatin accessibility profile by minimizing the maximum mean discrepancy between them [23]. UnionCom relies on the generalized unsupervised manifold alignment and uses local and global properties of the cells to align the single-cell gene expression and chromatin accessibility measurements [24]. SCOT applies the Gromov-Wasserstein-based optimal transport to align the manifolds [25], but Pamona unitizes the partial Gromov-Wasserstein optimal transport to align the manifolds [26].

However, all current methods mentioned above suffer from two major problems. First, they are incapable of handling the cell-type heterogeneity problem. Specifically, when the cell types in the single-cell gene expression profile differ from those in the single-cell chromatin accessibility profile, they may generate poor alignment. CoupleNMF [19] would fail because it requires the two datasets have the same number of cell types. Other methods assume that single-cell gene expression and chromatin accessibility share a similar manifold, which might not hold when the cell types across datasets are different. MMD-MA [23], UnionCom [24], and SCOT [25] methods that rely on such assumptions would enforce the alignment between two different manifolds, which would lead to incorrect integration. Pamona [26] attempts to resolve the cell-type heterogeneity predicament by estimating the number of the common cells across diverse measurements. However, the performance of the proposed estimation has not been comprehensively tested [26]. Second, all current methods’ performance highly relies on hyper-parameter selection. Under the unsupervised setting, it is very challenging to find the optimal hyper-parameter for different datasets.

To overcome these limitations, we present AIsceEA – **A**lignment-based **I**ntegration of single-cell gene **E**xpression and chromatin **A**ccessibility – an unsupervised computational method that explicitly uses biological consistency between gene expression and chromatin accessibility to guide the across-modality integration. First, AIsceEA defines a rank-based similarity score to quantify the biological consistency between cell types in gene expression and chromatin accessibility profiles. Then, based on the rank-based similarity and a novel designed permutation test, AIsceEA identifies the domain-specific cell types and then finds corresponding cell types shared across single-cell gene expression and chromatin accessibility profiles. Furthermore, for these corresponding cell types across modalities, AIsceEA applies a graph alignment method to elucidate the cell-cell correspondence [28].

We first validated the performance of AIsceEA using SNARE-seq Human cell line mixtures data [17], which jointly captured accessible chromatin regions and gene expression profiles within the same cells, and therefore it provides cell-cell correspondence for benchmarking. The benchmarking results demonstrate that AIsceEA can resolve the cell-type heterogeneity problem and is more robust to hyper-parameters. Furthermore, we show that our AIsceEA outperforms CoupleNMF [19] in terms of cell-type alignment. In addition, we compared the performance of our method with state-of-the-art cell-cell integration methods MMD-MA [23], UnionCom [24], SCOT [25], and Pamona [26] on real-world single-cell gene expression and chromatin ac-

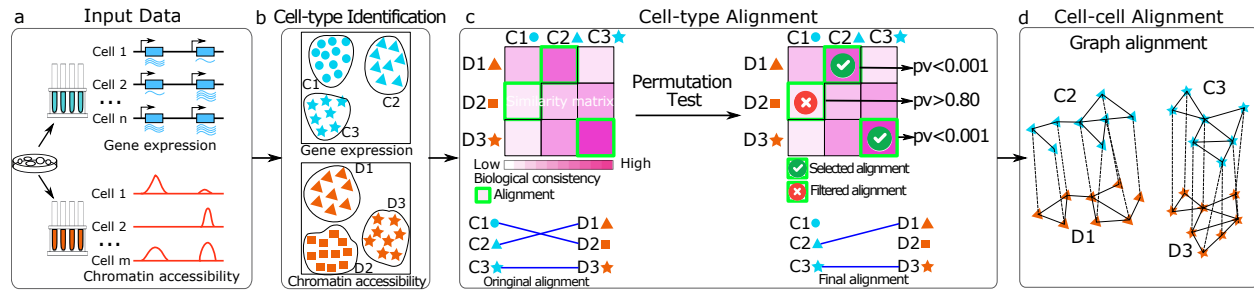


Fig. 1: Overview of AIsceEA. (a) Input datasets of single-cell RNA-seq and single-cell ATAC-seq measurements. (b) Clustering and cell-type identification. (c) Cell-type alignment using biological consistency and calculating the p-values by a novel permutation test. (d) AIsceEA finds cell-cell alignment using a graph alignment method for each pair of mapped clusters.

cessibility profiles. We applied them to integrate SNARE-seq profilings of neonatal mouse cerebral cortex [17], adult mouse cerebral cortex [17], and two scMultiome-seq PBMC datasets from the healthy donors.

We demonstrate that AIsceEA significantly outperforms other methods in terms of the average FOSCTTM score [23], demonstrating its outperformance in identifying the cell-cell correspondence.

2 Methods

2.1 Method Overview

AIsceEA is an alignment-based method that can identify the cell-type and cell-cell correspondence between single-cell gene expression and chromatin accessibility measurements profiled from the same tissue. In contrast to the current state-of-the-art methods [23–26], AIsceEA does not rely on the assumption of similarity between the manifolds of the entire single-cell gene expression and chromatin accessibility measurements. However, AIsceEA relies on biological consistency, which is the fact that the promoter regions of over-expressed genes should be significantly accessible to guide the alignment between cell types and also between cells across the measurements [29–33]. AIsceEA quantifies such biological consistency using a rank-based similarity score and further unitizes the similarity score to direct the cell-type and cell-cell alignments. As shown in Fig. 1, AIsceEA consists of three steps: (i) cell-type identification, (ii) cell-type alignment, and (iii) cell-cell alignment. In the following, we will elaborate on the details of each step.

2.2 Cell-type Identification

We first identify cell types within single-cell gene expression and chromatin accessibility measurements via commonly used clustering methods. For single-cell gene expression, we use the classical graph-based clustering method [34–37] to identify n cell-types in $\mathcal{C} = \{C_1, C_2, \dots, C_n\}$. For single-cell chromatin accessibility measurement, we first use cisTopic [38] to extract regulatory topics and then use the extracted features to cluster cells into m cell-types $\mathcal{D} = \{D_1, D_2, \dots, D_m\}$. In practice, we use the Leiden clustering as the default clustering method to identify cell types in both single-cell gene expression and chromatin accessibility measurements.

Furthermore, for each cell-type $C_i, \forall i$ identified in the single-cell gene expression measurement, AIsceEA identifies the set of differential over-expressed genes $\mathcal{G}_{C_i} = \{g_1, g_2, \dots\}$. AIsceEA then ranks these genes by their expression log 2 fold changes with respect to their expression in the rest of the cells in descending order. We further use a function $R_{C_i} : \mathcal{G}_{C_i} \rightarrow \mathbb{Z}^+$ to retrieve the ranking of a gene in \mathcal{G}_{C_i} . Similarly, for each cell-type $D_j, \forall j$ identified in the single-cell chromatin accessibility measurement, AIsceEA identifies the significantly accessible locations using the predictive distribution calculated by cisTopic [38]. Next, we identify the overlap between these significantly accessible locations and the promoter regions of the expressed genes. AIsceEA uses $\mathcal{H}_{D_j} = \{g_1, g_2, \dots\}$ to present the set of genes, whose promoter regions overlap with the significantly accessible locations in cell-type D_j .

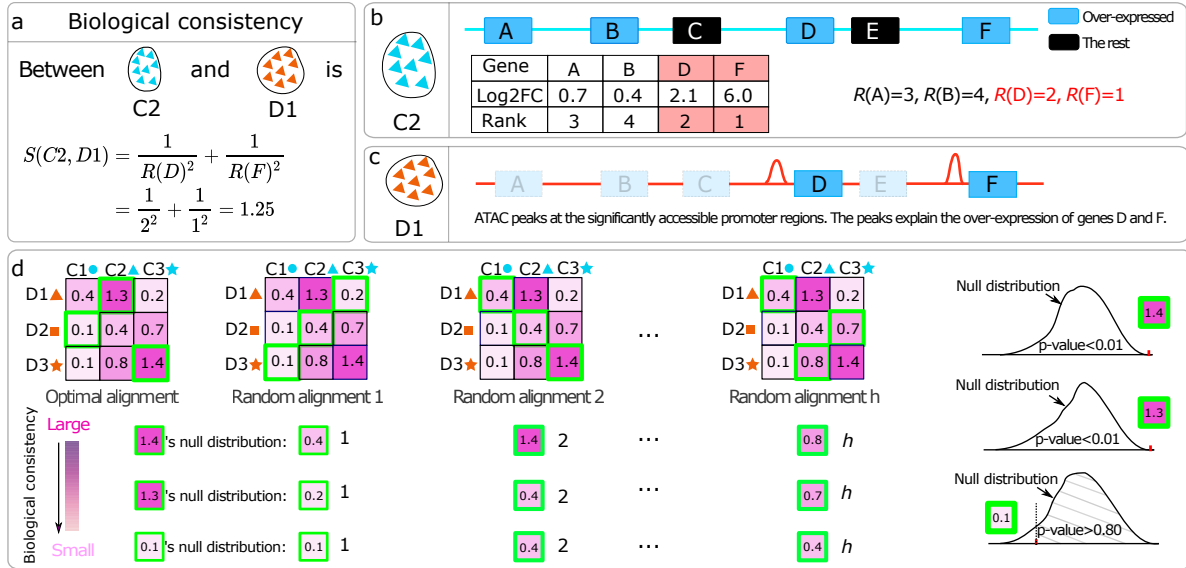


Fig. 2: (a) An illustration of computing the biological consistency between (b) a cell type in gene expression, (c) and a cell-type in chromatin accessibility measurement. (d) An explanation of the proposed permutation test to calculate p-values for aligned cell types.

2.3 Cell-type Alignment

After cell-type identification, as explained in section 2.3, n and m cell types are obtained in the single-cell gene expression and chromatin accessibility measurements, respectively. Although both measurements are profiled from the same tissue, due to cellular heterogeneity, in general, the number of cell types m and n may differ. Furthermore, the correspondence is unknown between n cell-types in the single-cell gene expression measurements and m cell-types in the single-cell chromatin accessibility measurements.

We propose to use the biological consistency between gene expression and chromatin accessibility to find the alignment of cell types across different modalities (as shown in Fig. 1). Furthermore, AIsCEA adopts a novel permutation test to find statistically significant biological consistency between the aligned cell types. We will keep the aligned cell types that are statistically significant and filter out the rest, as shown in Fig. 1c.

Cell-type alignment by biological consistency The biological consistency AIsCEA anchored on is the fact that the promoter regions of over-expressed genes should be significantly accessible [29–33]. AIsCEA defines a ranked similarity score S to quantify such biological consistency between cell types. Mathematically, the ranked similarity score $S(C_i, D_j)$ between cell-type C_i in single-cell gene expression and cell-type D_j in single-cell chromatin accessibility data can be computed by:

$$S(C_i, D_j) = \sum_{g \in \mathcal{G}_{C_i} \cap \mathcal{H}_{D_j}} \frac{1}{R_{C_i}(g)^2}, \quad (1)$$

where \mathcal{G}_{C_i} is the set of differential over-expressed genes in cell-type C_i identified in the single-cell gene expression measurements. \mathcal{H}_{D_j} is the set of genes whose promoter regions are significantly accessible in cell-type D_j in single-cell chromatin accessibility measurements. $\mathcal{G}_{C_i} \cap \mathcal{H}_{D_j}$ extracts all differentially over-expressed genes whose promoter regions are significantly accessible. $R_{C_i} : \mathcal{G}_{C_i} \rightarrow \mathbb{Z}^+$ is the function that takes a gene and returns the ranking of the gene in terms of its expression log 2 fold change. The larger the log2 fold change of a gene expression is, the higher rank it has (the rank of the top gene is 1). Based on the definition of $S(C_i, D_j)$ in (1), we know that $S(C_i, D_j)$ is large when 1) $R_{C_i}(g)$ is small, meaning the top ranking genes' promoter regions should be significantly accessible; 2) $|\mathcal{G}_{C_i} \cap \mathcal{H}_{D_j}|$ is large, meaning most of the highly over-expressed genes should have significantly accessible promoter regions. Fig. 2(a-c) illustrates a toy example of how $S(C_i, D_j)$ is computed.

From (1), we compute the biological consistency between n cell-types in $\mathcal{C} = \{C_1, C_2, \dots, C_n\}$ and m cell-types in $\mathcal{D} = \{D_1, D_2, \dots, D_m\}$. Without loss of generality, we assume that $n \leq m$ (if $n \geq m$, we can add dummy cell-types in \mathcal{D} to make $n = m$). Then the cell-type alignment across measurements can be

obtained by maximizing the biological consistency between aligned cell types across measurements, which can be formulated as a linear assignment problem:

$$\begin{aligned} \max_X : & \sum_{i=1}^n \sum_{j=1}^m S(C_i, D_j) X_{ij} \\ \text{s.t. } & X \in \Omega, \end{aligned} \quad (2)$$

where X is a binary assignment matrix, where $X_{ij} = 1$ denotes that cell-type C_i corresponds to cell-type D_j . The constraint set $\Omega = \{X \in \{0, 1\}^{n \times m} : X \mathbf{1}_m = \mathbf{1}_n, X^\top \mathbf{1}_n \leq \mathbf{1}_m\}$ enforces each cell-type in \mathcal{C} is assigned to one and only one cell-type in \mathcal{D} . The linear assignment problem can be efficiently solved by the Hungarian algorithm [39].

Resolving the cell-type heterogeneity problem via a novel permutation test. The set of cell types \mathcal{C} in single-cell gene expression could be different from the set of cell types \mathcal{D} in the single-cell chromatin accessibility data, which results in the cell-type heterogeneity problem. To elucidate the cell-type heterogeneity across measurements, we develop a novel permutation test to distinguish statistically significant corresponding cell types across modalities and find the unique cell types within the measurements.

Before introducing the permutation test, let us first introduce some notations. Given an assignment matrix $Z \in \Omega$, we can obtain the corresponding ranked similarity scores for each alignment and collect them in the set $\mathcal{S}_Z = \{S(C_i, D_j) | Z_{ij} = 1, \forall i, j\}$. We further sort the ranked similarity scores in \mathcal{S}_Z in descending order and define a function $\Phi_{\mathcal{S}_Z} : \mathcal{S}_Z \rightarrow \mathbb{Z}^+$ that returns the ranking of a similarity score in \mathcal{S}_Z . We also define $\Phi_{\mathcal{S}_Z}$'s inverse function $\Phi_{\mathcal{S}_Z}^{-1} : \mathbb{Z}^+ \rightarrow \mathcal{S}_Z$ that applies to a given ranking and returns the corresponding similarity score.

The null hypothesis of our novel permutation test is that the ranked similarity score between the aligned cell-types found by (2) are greater or equal to the ranked similarity scores of the randomly aligned cell types. After solving (2), we obtain an optimal assignment X^* and the corresponding similarity scores $\mathcal{S}_{X^*} = \{S(C_i, D_j) | X_{ij}^* = 1, \forall i, j\}$. For a specific alignment $X_{ij}^* = 1$, we can get the corresponding similarity score $S(C_i, D_j)$ and its ranking among all alignments by $I = \Phi_{\mathcal{S}_{X^*}}(S(C_i, D_j))$. We then generate k random cell-type alignments by uniformly sampling $h = 1,000$ assignment matrices $Z_1, \dots, Z_h \in \Omega$ from Ω . The null distribution of the I th ranking similarity score can be estimated by $\{\Phi_{\mathcal{S}_{Z_1}}^{-1}(I), \dots, \Phi_{\mathcal{S}_{Z_h}}^{-1}(I)\}$ (where $\{\Phi_{\mathcal{S}_{Z_l}}^{-1}(I)$ is the I th ranked similarity score in the random alignment Z_l). By comparing $S(C_i, D_j)$ with $\Phi_{\mathcal{S}_{Z_l}}^{-1}(I), l = 1, \dots, h$, we can calculate the p -value by $1 - \frac{1}{h} |\{I | \Phi_{\mathcal{S}_{Z_l}}^{-1}(I) \leq S(C_i, D_j)\}|$, where $|\cdot|$ is the cardinality of a set. For the alignment $X_{ij}^* = 1$ whose corresponding p -value is significant (≤ 0.01), we consider it a true alignment. For the alignment $X_{ij}^* = 1$ whose corresponding p -value is not significant ($p\text{-value} > 0.01$), we consider the corresponding cell-types in this alignment may be considered unique cell-types within their measurement. Fig. 2 (d) illustrates how the permutation is calculated.

Hyper-parameters selection scheme for the cell-type alignment Due to the heterogeneity between the single-cell gene expression and chromatin accessibility measurements, the number of cell types n, m found in both measurements typically differs ($n \neq m$). The selection of n and m would influence the performance of the cell-type alignment in AIsceEA. Currently, under the unsupervised setting, there is no effective way to select n and m . To fill the gap, we propose a heuristic approach to select them. AIsceEA applies Leiden clustering [35] to identify cell types using the resolution parameter. Therefore, we propose an effective scheme to select the resolution parameter rather than the number of cell types as following.

Our heuristic approach sets the range for the resolution parameter r_e for single-cell gene expression measurement $r_e \in \{0.1, 0.15, 0.2, \dots, 1.5\}$ and the resolution parameter r_c for single-cell chromatin accessibility measurement $r_c \in \{0.1, 0.15, 0.2, \dots, 1.5\}$. Then we screen different combinations of r_e and r_c to compute the *alignment ratio* defined as $L = \frac{o}{n_e} + \frac{o}{m_c}$, where n_e is the number of identified cell types in gene expression measurement when the resolution parameter is set to r_e , m_c is the number of identified cell types in chromatin accessibility measurement when the resolution parameter is set to r_c , and o is the number of aligned cell types between n_e and m_c identified by the cell-type alignment method in AIsceEA (as explained in Section 2.3). In the end, after screening all resolution parameters, we select the ones yielding the largest *alignment ratio* L .

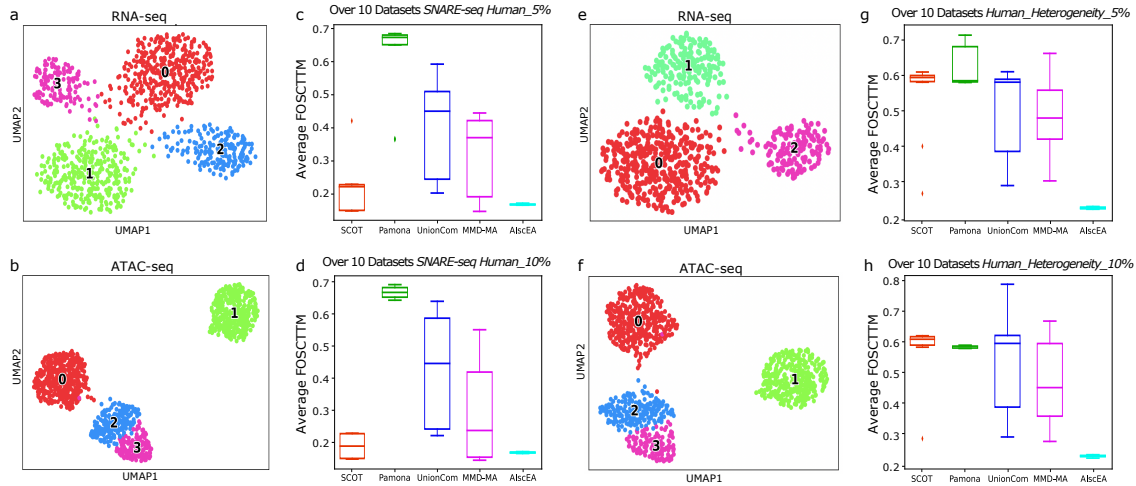


Fig. 3: (a) The UMAP of the RNA-seq in *SNARE-seq Human* data. (b) The UMAP of the ATAC-seq in *SNARE-seq Human* data. (c) The box plot of the average FOSCTTM over 10 *SNARE-seq Human_R5%* data. (d) The box plot of the average FOSCTTM over 10 *SNARE-seq Human_R10%* data. (e-f) The UMAPs of *SNARE-seq Human_Heterogeneity* data that ATAC-seq has one more cell type than RNA-seq. (g) The box plot of the average FOSCTTM over 10 *SNARE-seq Human_R5%* data. (h) The box plot of the average FOSCTTM over 10 *SNARE-seq Human_R10%* data.

In the experiment Section 4.1, we empirically show that the proposed heuristic approach can select r_e and r_c that result in descent cell-type alignment results for all datasets in a completely unsupervised manner.

2.4 Cell-cell Alignment

Once we identify cell-types C_i and D_j are aligned together, we can further find the cell-cell correspondence between the cells in C_i and D_j . AlScEA assumes that C_i and D_j consist of a set of cells $C_i = \{c_i^1, c_i^2, \dots\}$ and $D_j = \{d_j^1, d_j^2, \dots\}$, respectively. Since the gene expression of cells in C_i and the chromatin accessibility of cells in D_j are considered as different measurements for cells of the same kind, we confidently assume that their low-dimensional manifold is similar. Hence, a graph alignment method is employed to find the cell-cell correspondence [28].

AlScEA constructs a symmetric k -nearest neighbor graph $G_1 = (V_1, E_1)$ to present the manifold of the cells in the cluster C_i , where vertices in $V_1 = C_i = \{c_i^1, c_i^2, \dots\}$ are cells in C_i . Similarly, we construct a symmetric k -nearest neighbor graph $G_2 = (V_2, E_2)$ to present the manifold of the cells in D_j , where vertices in $V_2 = D_j = \{d_j^1, d_j^2, \dots\}$ are cells in D_j (details in the Supplementary Materials Section B). Therefore, in the following, we can safely assume $|V_1| = |V_2| = N$. If $|V_1| \neq |V_2|$, we can add dummy node to make them equal as done in [28]. The manifold matching between cells in C_i and the cells in D_j can be achieved by the graph alignment between G_1 and G_2 . Mathematically, the graph alignment step in AlScEA can be formulated [28] as:

$$\begin{aligned} \max_P : & \text{Tr}(A_1^\top P A_2 P^\top) + \lambda \text{Tr}(P L) \\ \text{s.t. } & P \in \{P \in \{0, 1\}^{N \times N}, P^\top \mathbf{1}_N = \mathbf{1}_N, P \mathbf{1}_N = \mathbf{1}_N\}. \end{aligned} \quad (3)$$

A_1 and A_2 are the adjacency matrices for G_1 and G_2 , respectively. P is constrained to be a permutation matrix that enforces one-to-one mapping between cells in G_1 and G_2 . L is the similarity matrix between cells in V_1 and V_2 and L_{kl} estimates the biological consistency between cells c_i^k and d_j^l . L_{kl} can be computed by a ranked similarity score, which is similar to (1) (details in the Supplementary Materials). In the objective function in (3), the first term $\text{Tr}(A_1^\top P A_2 P^\top)$ computes the number of overlapping edges between G_1 and G_2 (more overlapping edges imply that the manifolds represented by G_1 and G_2 are similar) and the second term computes total similarity between the aligned cells. λ is a hyper-parameter that balances the trade-off in the objective function (3). The optimization in (3) finds a one-to-one cell-cell alignment such that the number of overlapping edges between G_1 as well as G_2 and the total similarity between the aligned cells are maximized simultaneously. We propose applying the Frank-Wolfe [40] algorithm and the path-relinking technique to solve (3) (details in Supplementary Materials Section B).

Hyper-parameters for the cell-cell alignment There are two hyper-parameters that needs to be selected for the cell-cell alignment used in AIsCEA: k (the number of nearest neighbors when constructing the symmetric k -nearest neighbor graph) and λ (the regularizer in Eq. 3). In the experiment section 4.1, we show that the cell-cell alignment in AIsCEA is robust to the selection of k and λ . Therefore, we set k and λ to default values in practice.

3 Experimental Setup

3.1 Competing methods

AIsCEA can identify cell-type alignment between scRNA-seq and scATAC-seq datasets, therefore, we compare AIsCEA’s performance on cell-type alignment with CoupleNMF [19], which is the state-of-the-art cell-type alignment method. In addition, AIsCEA is able to find cell-cell alignment, therefore, we compare AIsCEA with the current state-of-the-art cell-cell alignment methods MMD-MA [23], UnionCom [24], SCOT [25], and Pamona [26].

3.2 Data

SNARE-seq Human [17] is a joint profiling of accessible chromatin and RNA of the mixture of human cell lines BJ, H1, K562, and GM12878. We use *SNARE-seq Human* to benchmark the competing methods because it provides the ground truth for both cell-type alignment and cell-cell alignment. Moreover, to evaluate different methods on handling the cell-type heterogeneity problem, we generate *SNARE-seq Human.Heterogeneity* data by manually removing cells of BJ cell type from the scRNA-seq data in *SNARE-seq Human*. Furthermore, To compare different methods’ robustness of hyper-parameter selection, we generate *SNARE-seq Human.R5%* and *SNARE-seq Human.R10%*, where 5% of cells and 10% cells are randomly removed from the original *SNARE-seq Human*. Furthermore, we generate *SNARE-seq Human.Heterogeneity.R5%* and *SNARE-seq Human.Heterogeneity.R10%*, where 5% of cells and 10% cells were randomly removed from the *SNARE-seq Human.Heterogeneity* data.

Additionally, we compare all the competing methods on real-world datasets. We first benchmark our method against all competing methods on two SNARE-seq real-world datasets: *SNARE-seq Mouse 5k* (SNARE-seq of neonatal mouse cerebral cortex that contains 5k cells) and *SNARE-seq Mouse 10k* (SNARE-seq of adult mouse cerebral cortex that has 10k cells). Then we apply all competing methods on two scMultiome datasets [41, 42]: *scMultiome PBMC 3k* (scMultiome-seq PBMC of a healthy donor with 3k cells) and *scMultiome PBMC 12k* (scMultiome-seq PBMC of a healthy donor with 12k cells). All these datasets provide cell-cell correspondence information, which is used to evaluate the competing methods. More details of these datasets can be found in Supplementary Materials Section C.1.

3.3 Metrics

We first introduce the metric we use for evaluating the cell-type alignment. When two cell types are aligned between scRNA-seq and scATAC-seq, we expect the cells in scRNA-seq cell type to appear in the aligned scATAC-seq cell type (for the existing cells). In other words, the two aligned cell types are expected to have a larger number of overlapping cells. Therefore, we use the overlap coefficient to measure the overlap between aligned cell types. Specifically, if cell-type C_i is aligned to cell-type D_j , the overlap coefficient between cells in C_i and cells in D_j can be computed as:

$$O(C_i, D_j) = \frac{|C_i \cap D_j|}{\min(|C_i|, |D_j|)}. \quad (4)$$

Furthermore, we can compute the total number of the overlapped cells over all aligned cell types as:

$$U = \sum_{(C_i, D_j) \in \mathcal{A}} |C_i \cap D_j|, \quad (5)$$

where \mathcal{A} is the collection of all aligned cell types. Another metric we use to evaluate the cell-type alignment is the average Silhouette score per cluster to measure cluster cohesion. We expect the cells in the same cell types to be similar to other cells in their own cell type, but different from other cell types.

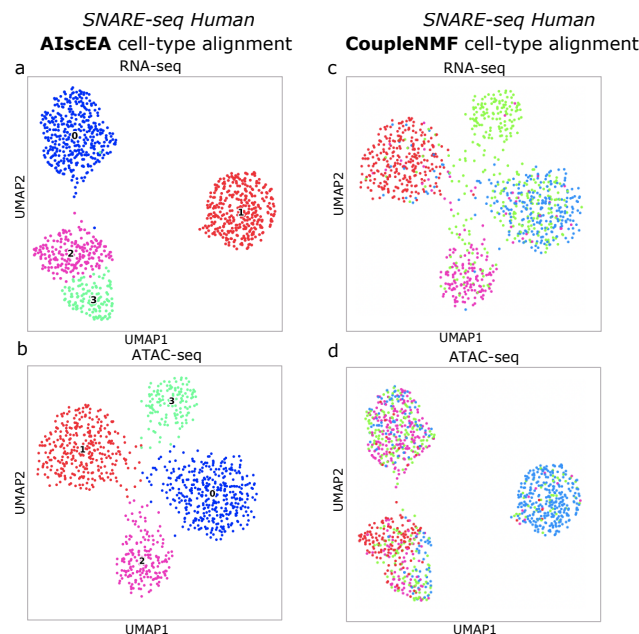


Fig. 4: (a-b) The aligned cell types in the RNA-seq and ATAC-seq in *SNARE-seq Human* data identified by AIscEA. (c-d) The aligned cell types in the RNA-seq and ATAC-seq in *SNARE-seq Human* data identified by CoupleNMF.

For evaluating the cell-cell alignment, we use the average FOSCTTM score [23], which has been widely used for evaluating single-cell multi-omics integration methods [23, 25, 26]. FOSCTTM stands for “fraction of samples closer than the true match”, therefore, the lower the better. The details of how FOSCTTM is computed are elaborated in Supplementary Materials Section D.1. Another metric is the cell coverage, which is the number of cells that have been found correspondence across the scRNA-seq and scATAC-seq datasets.

3.4 Hyper-parameter selection

We select the hyper-parameter for AIscEA using the approaches described in Section 2.3 and Section 2.4. For CoupleNMF [19], we set the number of cell-types based on the ground truth and for the rest of the hyper-parameters, we use the suggested hyper-parameters. For MMD-MA [23], UnionCom [24], SCOT [25], and Pamona [26], under the unsupervised setting, they do not have an effective way to find the optimal hyper-parameters. In this paper, we use the following strategy to find the optimal hyper-parameters for them. We performed a grid search to find the optimal hyper-parameters using *SNARE-seq Human* dataset. Then, we use the optimal hyper-parameters found in *SNARE-seq Human* for real-world datasets (more details in Supplementary Materials Section C.5 and C.6).

3.5 Computational resource

All experiments are processed on an Intel(R) Core(TM) i7-6850K CPU @ 3.60GHz CPU with 62GB memory and GPU computations on a single GeForce GTX 1080 Ti with VRAM of 11GB. If a method fails to run on a large-scale dataset due to memory shortage, we report a memory error.

4 Results

4.1 Benchmarking using SNARE-seq human cell line mixtures

SNARE-seq human cell line mixtures provide the ground truth information for validating cell-type alignment and cell-cell alignment. Therefore, we first use it to validate AIscEA’s hyper-parameter selection scheme proposed in Section 2.3 for cell-type alignment. Furthermore, we use it to evaluate all methods’ robustness to the choice of hyper-parameters for the cell-cell alignment. Last but not least, we use it to benchmark the performance of the competing methods on handling the cell-type heterogeneity problem, as in real-world datasets, the number of cell types may differ between two domains.

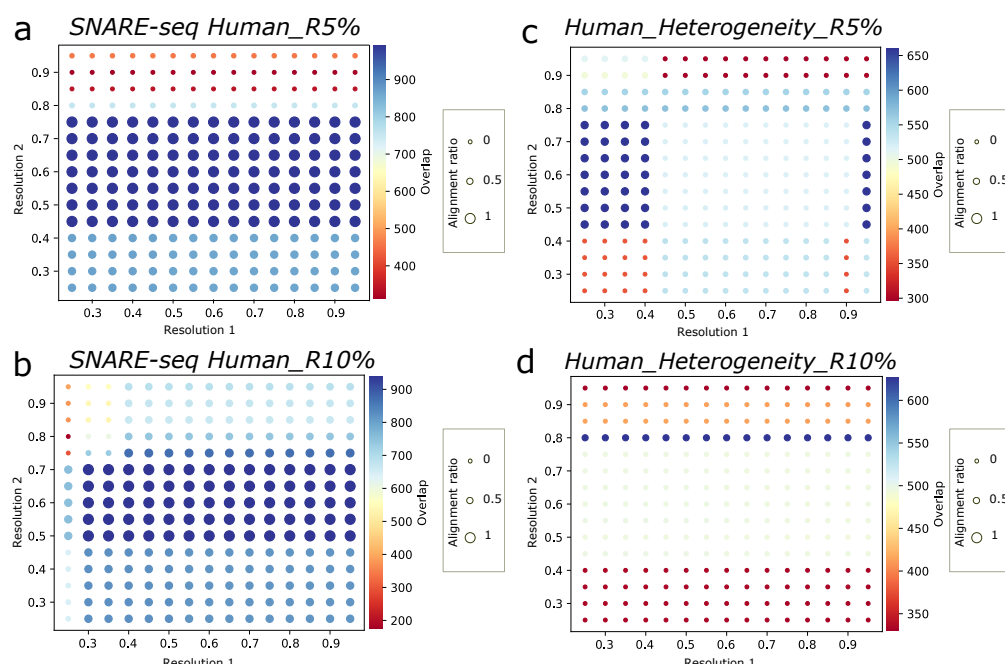


Fig. 5: Screening the resolution parameters. Color bar denotes the number of overlapped cells identified by the cell-type alignment. Dark blue implies the cell-type alignment has good performance. The size of the dots corresponds to the *alignment ratio* defined in 2.3. Large size means the the *alignment ratio* is large. (a-b) Screening results for *SNARE-seq Human_R5%* and *SNARE-seq Human_R10%*. (c-d) Screening results for *Human_Heterogeneity_R5%* and *SNARE-seq Human_Heterogeneity_R10%*.

Validation of the hyper-parameter selection scheme for the cell-type alignment in AIsceEA In Section 2.3, we propose an approach to select the resolution hyper-parameters in the Leiden clustering in AIsceEA, which determine the number of cell types in scRNA-seq n , and the number of cell types in scATAC-seq m for the cell-type alignment in AIsceEA. This section uses the *SNARE-seq Human* cell line mixtures to demonstrate that our unsupervised parameter selection scheme can select the resolution hyper-parameters that result in promising cell-type alignment.

We applied the proposed scheme in 2.3 to *SNARE-seq Human_R5%* data, *SNARE-seq Human_R10%* data, *Human_Heterogeneity_R5%* data, and *SNARE-seq Human_Heterogeneity_R10%* data (description of these four data can be found in Section 3.2) and show the hyper-parameters screening results in Fig. 5. In the Fig. 5, the size of each dot corresponds to its *alignment ratios* defined in 2.3. Larger size of the dots means the corresponding *alignment ratio* is higher. The color of each dot for each pair of resolution values indicates the number of overlapping cells identified by the cell-type alignment (computed as U defined in Section 3.3). Darker blue means more number of overlapping cells are identified by the cell-type alignment, which means the performance of the cell-type alignment is more promising. As shown in Fig. 5, the large-size dots always appear in dark blue color, demonstrating that the *alignment ratio* and the performance of the cell-type alignment method is positively correlated. Therefore, we can use the *alignment ratio* to guide the selection of the resolution hyper-parameters used in AIsceEA in an unsupervised setting. In addition, we noticed that many dots have the same size and color. Such observation implies that different combinations of resolution parameters may yield equivalently good cell-type alignments. We have the same observation from the screening results for more real-world datasets in the Supplementary Materials Fig. S2.

Benchmarking hyper-parameter robustness in the cell-cell alignment In this section, we compare AIsceEA with all competing cell-cell alignment methods in terms of their robustness to the choice of hyper-parameters. Such robustness is of practical importance since the real-world application is completely unsupervised; therefore, we do not have any prior knowledge to guide the hyper-parameter selection. If a method is sensitive to hyper-parameters, its performance is unreliable for real-world applications.

Table 1: The statistics of average FOSCTTM scores over the grid search of the hyper-parameter for each method using *SNARE-seq Human* and *Human.Heterogeneity* data.

Method	<i>SNARE-seq Human</i>			<i>Human.Heterogeneity</i>		
	Minimum	Mean	Std.	Minimum	Mean	Std.
AIscEA	0.150	0.162	0.001	0.152	0.156	0.007
SCOT	0.149	0.383	0.157	0.267	0.463	0.118
Pamona	0.227	0.402	0.130	0.159	0.463	0.170
MMD-MA	0.157	0.335	0.132	0.210	0.511	0.124
UnionCom	0.243	0.514	0.166	0.395	0.467	0.035

We applied all methods to the *SNARE-seq Human* data. We ran each method over an extensive grid search of suggested hyper-parameters and showed the results for *SNARE-seq Human* in Table 1. Cell coverage for AIscEA and all other methods are exactly 1,047 cells. The grid search hyper-parameter tuning details are elaborated in Supplementary Materials Section C.2. As shown in the table 1, AIscEA is competitive with SCOT on achieving the smallest FOSCTTM score, which is superior to the rest of methods. However, AIscEA has the smallest standard deviation, implying that AIscEA is more robust to the choice of hyper-parameters.

To further confirm the robustness of hyper-parameters for each method, we applied the optimal hyper-parameters found on *SNARE-seq Human* to the datasets that are slightly different from *SNARE-seq Human*. The goal is to check whether the optimal parameters on one dataset would still yield good results on a slightly different dataset.

To evaluate different methods in this scenario, we generated 10 *SNARE-seq Human.R5%* data and 10 *SNARE-seq Human.R10%* datasets (description of the data is in 3.2). Then we applied each method to them using their optimal set of hyper-parameters (in Supplementary Materials Table S1). Cell coverage for AIscEA and competing methods in these experiments is the number of shared cells between two domains for all methods. Fig. 3 c and d exhibit the box plots of the average FOSCTTM scores obtained by each method over 10 *SNARE-seq Human.R5%* data and 10 *SNARE-seq Human.R10%*, respectively. Clearly, our method shows the smallest variance on both figures. We further found that the mean of the average FOSCTTM scores achieved by AIscEA is significantly smaller than the rest of the methods. All these results demonstrate that AIscEA is more robust to the choice of hyper-parameters than all other competing methods.

Benchmarking in solving the cell-type heterogeneity problem Next, we benchmark all methods on their ability to resolve the cell-type heterogeneity problem. In a real-world application, we may not have any prior knowledge of whether the single-cell RNA-seq measurement and the single-cell ATAC-seq measurement have the same cell types. If we cannot distinguish the cell types that have correspondence and other cell types that have not, the alignment between the two measurements would be misleading.

To simulate the cell-type heterogeneity problem, we generated the *SNARE-seq Human.Heterogeneity* data (description of the data is in 3.2). We first ran each competing method over an extensive grid of suggested set of hyper-parameters and showed the results for *SNARE-seq Human.Heterogeneity* in Table 1. AIscEA can identify the heterogeneous cell type, exclude it, and map the shared cell types between two domains. Cell coverage for AIscEA in this experiment consists of the number of cells in all three shared clusters. As shown, AIscEA achieved the smallest average FOSCTTM score with the smallest standard deviation, indicating AIscEA is the best method to handle the cell-type heterogeneity problem.

Furthermore, we applied each method using its optimal hyper-parameters found on *SNARE-seq Human.Heterogeneity* data (shown in Supplementary Materials Table S.1) to 10 *Human.Heterogeneity.R5%* data and 10 *SNARE-seq Human.Heterogeneity.R10%* data (description in 3.2). Fig. 3e and f show the comparison results. Apparently, AIscEA achieved the smallest average FOSCTTM score and was more robust to its hyper-parameters. All above experiments demonstrate that AIscEA is the best method to resolve the cell-type heterogeneity problem.

4.2 Comparison of the cell-type alignment

In this section, we compare AIscEA with CoupleNMF [19] in terms of cell-type alignment. We applied both methods to *SNARE-seq Human*, *SNARE-seq Mouse 5k*, *SNARE-seq Mouse 10k*, *scMultiome-seq PBMC 3k*

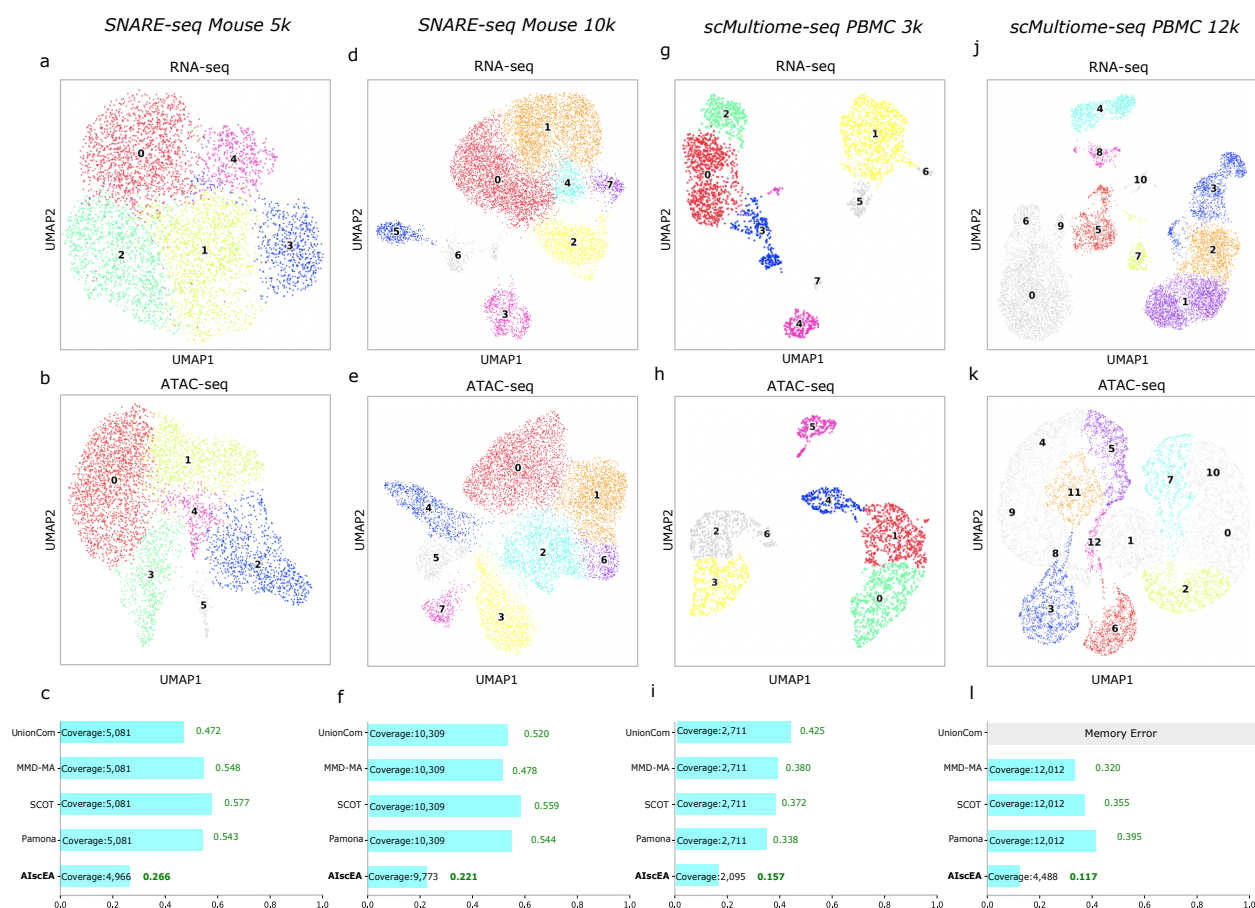


Fig. 6: (a-b) The UMAPs of the RNA-seq and ATAC-seq of *SNARE-seq Mouse 5k*. Aligned cell types identified by AIsCEA are shown in the same color, but grey color shows filtered out cell types by AIsCEA. Each color indicates a specific aligned cell type across the measurements. (c) The bar plots of the average FOSCTTM scores for all methods. The shorter the bar the better the method performs. The number shows on the bar is the cell coverage for the method. (d-e) The UMAPs of the RNA-seq and ATAC-seq of *SNARE-seq Mouse 10k*. The color coding scheme is the same as (a-b). (f) The bar plots of the average FOSCTTM scores for all methods along with cell coverage for *SNARE-seq Mouse 10k*. (g-h) The UMAPs of the RNA-seq and ATAC-seq of *scMultiome-seq PBMC 3k*. The color coding scheme is the same as (a-b). (i) The bar plots of the average FOSCTTM scores for all methods along with cell coverage for *scMultiome-seq PBMC 3k*. (j-k) The UMAPs of the RNA-seq and ATAC-seq of *scMultiome-seq PBMC 12k*. The color coding scheme is the same as (a-b). (l) The bar plots of the average FOSCTTM scores for all methods along with cell coverage for *scMultiome-seq PBMC 12k*.

and *scMultiome-seq PBMC 12k*, except *SNARE-seq Human Heterogeneity* because CoupleNMF requires the scRNA-seq and scATAC-seq data share the same number of cell types. CoupleNMF only generated results for two datasets with small number of cells, which are *SNARE-seq Human* and *scMultiome-seq PBMC 3k*. For the rest of the datasets, CoupleNMF failed and ran out of memory (memory error).

In Fig. 4, we illustrate the comparison between AIsCEA and CoupleNMF on *SNARE-seq Human*. The same comparison for *scMultiome-seq PBMC 3k* is shown in Supplementary Materials Fig. S.1(e-h). As shown in Fig. 4, for AIsCEA, the cells in the aligned cell types in both scRNA-seq and scATAC-seq are well isolated. However, for CoupleNMF, the cells in the aligned cell types are mixed together. We further evaluated the performance of both methods in terms of overlapping coefficient (defined in (4)) and the Silhouette score shown in Table 2. Clearly, AIsCEA achieves much higher overlapping coefficients and Silhouette scores, which demonstrates that AIsCEA outperform CoupleNMF in terms of cell-type alignment.

4.3 Comparison of cell-cell alignment using real-world data

We compared AIsCEA with competing cell-cell alignment methods MMD-MA [23], UnionCom [24], SCOT [25], and Pamona [26] on both real-world SNARE-seq data and real-world scMultiome-seq data. We selected hyper-parameters for each method following the strategy we described in. Section 3.4.

Table 2: Cell-type alignment comparison.

Method	<i>SNARE-seq Human</i>	<i>scMultiome-seq PBMC 3k</i>
AIsCEA	Overlap coef: 0.911 Silhouette score: 0.618	Overlap coef: 0.884 Silhouette score: 0.463
CoupleNMF	Overlap coef: 0.202 Silhouette score: 0.146	Overlap coef: 0.398 Silhouette score: 0.032
CoupleNMF failed to run on other datasets due to memory error.		

AIsCEA outperforms current methods on the real-world SNARE-seq data. We applied all competing cell-cell alignment methods on *SNARE-seq Mouse 5k* and *SNARE-seq Mouse 10k* (description in 3.2). We compared their performance in terms of the average FOSCTTM score and cell coverage (description in Section 3.3).

Fig. 6 a and b illustrate the cell-type alignment identified by AIsCEA for *SNARE-seq Mouse 5k*. And Fig. 6c shows the comparison between different methods in terms of the average FOSCTTM score and cell coverage. As shown, AIsCEA achieves the lowest average FOSCTTM score, which is much smaller than the rest of the methods. The cell coverage of AIsCEA is slightly smaller than the other methods (4966 cells out of 5081 cells, only around 2% of cells are missed by AIsCEA). But considering both average FOSCTTM score and cell coverage, it is obvious that AIsCEA significantly outperforms all the current methods.

Fig. 6 d and e illustrate the cell-type alignment identified by AIsCEA for *SNARE-seq Mouse 10k*. Fig. 6f shows the comparison between different methods in terms of the average FOSCTTM score and cell coverage. As shown, AIsCEA achieves the lowest average FOSCTTM score, which is much smaller than the rest of the methods. The cell coverage of AIsCEA is slightly smaller than the other methods (9773 cells out of 10,309 cells, only around 5% of cells are missed by AIsCEA). But considering both average FOSCTTM score and cell coverage, AIsCEA significantly outperforms all the current methods.

AIsCEA outperforms the current methods on the real-world scMultiome-seq data. We applied all competing cell-cell alignment methods on *scMultiome-seq PBMC 3k* and *scMultiome-seq PBMC 12k* (description in Section 3.2). We compared their performance in terms of the average FOSCTTM score and cell coverage (description in Section 3.3).

Fig. 6 g and h illustrate the cell-type alignment identified by AIsCEA for *scMultiome-seq PBMC 3k*. Fig. 6i shows the comparison between different methods in terms of the average FOSCTTM score and cell coverage. As illustrated, AIsCEA attained the lowest average FOSCTTM score, which is much smaller than the rest of the methods. Considering both the average FOSCTTM score and cell coverage, AIsCEA significantly outperforms all the current methods.

Fig. 6 j and k illustrate the cell-type alignment identified by AIsCEA for *scMultiome-seq PBMC 12k*. Fig. 6l shows the comparison between different methods in terms of the average FOSCTTM score and cell coverage. AIsCEA yielded the lowest average FOSCTTM score with a large margin. Although The cell coverage of AIsCEA can be smaller than the other methods, considering both average FOSCTTM score and cell coverage, AIsCEA outperforms all the current methods.

5 Conclusion

In this study, we proposed AIsCEA, an unsupervised computational method for integrating single-cell gene expression and chromatin accessibility measurements. Unlike other approaches, AIsCEA relies on the biological consistency between the two measurements to guide the integration. We compared AIsCEA with the state-of-the-art methods on the *SNARE-seq* human cell line mixtures benchmark datasets [17] and demonstrated that AIsCEA can effectively select hyper-parameters as well as better handle the cell-type heterogeneity problem. Furthermore, we showed that AIsCEA significantly outperforms previous methods when applying to the real-world mouse SNARE-seq and scMultiome-seq datasets.

Several innovations developed in this work contributed to the performance of AIsCEA. First, the ranked similarity score enables us to compare the cell types across measurements. The ranked similarity score is the key to estimating the similarity between cell types from different modalities. Second, the novel permutation

test can distinguish the true cell-type alignment if the corresponding ranked similarity score is significantly larger than the random ranked similarity score in the background. Last but not least, the graph alignment method uses the symmetric k nearest neighbor graph to characterize the low-dimensional manifold. It is a notable advantage that AIsCEA can identify cell types that appear only in one domain and exclude them from the cell-cell alignment in further analysis.

Our future direction is to recruit more cells in the integration. We believe AIsCEA is the milestone for the integration of single-cell gene expression and chromatin accessibility measurements. Furthermore, it also provides a stepping stone for integrating other single-cell measurements.

References

- [1] Z. Miao et al. “Putative cell type discovery from single-cell gene expression data”. In: *Nat Methods* 17.6 (June 2020), pp. 621–628.
- [2] V. Svensson, E. da Veiga Beltrame, and L. Pachter. “A curated database reveals trends in single-cell transcriptomics”. In: *Database (Oxford)* 2020 (Nov. 2020).
- [3] H. Chen et al. “Assessment of computational methods for the analysis of single-cell ATAC-seq data”. In: *Genome Biol* 20.1 (Nov. 2019), p. 241.
- [4] L. Aparicio, M. Bordyuh, A. J. Blumberg, and R. Rabadan. “A Random Matrix Theory Approach to Denoise Single-Cell Data”. In: *Patterns (N Y)* 1.3 (June 2020), p. 100035.
- [5] J. Wang et al. “Data denoising with transfer learning in single-cell transcriptomics”. In: *Nat Methods* 16.9 (Sept. 2019), pp. 875–878.
- [6] G. Eraslan et al. “Single-cell RNA-seq denoising using a deep count autoencoder”. In: *Nat Commun* 10.1 (Jan. 2019), p. 390.
- [7] M. Huang et al. “SAVER: gene expression recovery for single-cell RNA sequencing”. In: *Nat Methods* 15.7 (July 2018), pp. 539–542.
- [8] D. van Dijk et al. “Recovering Gene Interactions from Single-Cell Data Using Data Diffusion”. In: *Cell* 174.3 (July 2018), pp. 716–729.
- [9] M. D. Luecken and F. J. Theis. “Current best practices in single-cell RNA-seq analysis: a tutorial”. In: *Mol Syst Biol* 15.6 (June 2019), e8746.
- [10] D. Risso et al. “A general and flexible method for signal extraction from single-cell RNA-seq data”. In: *Nat Commun* 9.1 (Jan. 2018), p. 284.
- [11] A. H. S. Vargo and A. C. Gilbert. “A rank-based marker selection method for high throughput scRNA-seq data”. In: *BMC Bioinformatics* 21.1 (Oct. 2020), p. 477.
- [12] J. Lee, D. Y. Hyeon, and D. Hwang. “Single-cell multiomics: technologies and data analysis methods”. In: *Exp Mol Med* 52.9 (Sept. 2020), pp. 1428–1442.
- [13] G. Kelsey, O. Stegle, and W. Reik. “Single-cell epigenomics: Recording the past and predicting the future”. In: *Science* 358.6359 (Oct. 2017), pp. 69–75.
- [14] E. Shema, B. E. Bernstein, and J. D. Buenrostro. “Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution”. In: *Nat Genet* 51.1 (Jan. 2019), pp. 19–25.
- [15] M. Efremova and S. A. Teichmann. “Computational methods for single-cell omics across modalities”. In: *Nat Methods* 17.1 (Jan. 2020), pp. 14–17.
- [16] X. Dong, C. Liu, and M. Dozmorov. “Review of multi-omics data resources and integrative analysis for human brain disorders”. In: *Brief Funct Genomics* 20.4 (July 2021), pp. 223–234.
- [17] S. Chen, B. B. Lake, and K. Zhang. “High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell”. In: *Nat Biotechnol* 37.12 (Dec. 2019), pp. 1452–1457.
- [18] L. Liu et al. “Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity”. In: *Nat Commun* 10.1 (Jan. 2019), p. 470.
- [19] Z. Duren et al. “Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations”. In: *Proc Natl Acad Sci U S A* 115.30 (July 2018), pp. 7723–7728.
- [20] Zhen Cui, Hong Chang, Shiguang Shan, and Xilin Chen. “Generalized unsupervised manifold alignment”. In: *Advances in Neural Information Processing Systems* 27 (2014), pp. 2429–2437.
- [21] J. D. Welch, A. J. Hartemink, and J. F. Prins. “MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics”. In: *Genome Biol* 18.1 (July 2017), p. 138.
- [22] T. Stuart et al. “Comprehensive Integration of Single-Cell Data”. In: *Cell* 177.7 (June 2019), pp. 1888–1902.
- [23] R. Singh et al. “Unsupervised manifold alignment for single-cell multi-omics data”. In: *ACM BCB* 2020 (Sept. 2020), pp. 1–10.
- [24] K. Cao, X. Bai, Y. Hong, and L. Wan. “Unsupervised topological alignment for single-cell multi-omics integration”. In: *Bioinformatics* 36.Suppl.1 (July 2020), pp. i48–i56.
- [25] Pinar Demetci et al. “Gromov-Wasserstein optimal transport to align single-cell multi-omics data”. In: *bioRxiv* (2020).

- [26] K. Cao, Y. Hong, and L. Wan. “Manifold alignment for heterogeneous singlecell multiomics data integration using Pamona”. In: *Bioinformatics* (Aug. 2021).
- [27] C. Wang et al. “Integrative analyses of single-cell transcriptome and regulome using MAESTRO”. In: *Genome Biol* 21.1 (Aug. 2020), p. 198.
- [28] M. Zaslavskiy, F. Bach, and J. P. Vert. “Global alignment of protein-protein interaction networks by graph matching methods”. In: *Bioinformatics* 25.12 (June 2009), pp. i259–267.
- [29] Y. Sun, N. Miao, and T. Sun. “Detect accessible chromatin using ATAC-sequencing, from principle to applications”. In: *Hereditas* 156 (2019), p. 29.
- [30] A. R. Quinlan and I. M. Hall. “BEDTools: a flexible suite of utilities for comparing genomic features”. In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842.
- [31] T. C. Silva et al. “Analyze cancer genomics and epigenomics data using Bioconductor packages”. In: *F1000Res* 5 (2016), p. 1542.
- [32] J. Rainer, L. Gatto, and C. X. Weichenberger. “ensemldb: an R package to create and use Ensembl-based annotation resources”. In: *Bioinformatics* 35.17 (Sept. 2019), pp. 3151–3153.
- [33] M. M. Halstead et al. “Systematic alteration of ATAC-seq for profiling open chromatin in cryopreserved nuclei preparations from livestock tissues”. In: *Sci Rep* 10.1 (Mar. 2020), p. 5230.
- [34] F. A. Wolf, P. Angerer, and F. J. Theis. “SCANPY: large-scale single-cell gene expression data analysis”. In: *Genome Biol* 19.1 (Feb. 2018), p. 15.
- [35] V. A. Traag, L. Waltman, and N. J. van Eck. “From Louvain to Leiden: guaranteeing well-connected communities”. In: *Sci Rep* 9.1 (Mar. 2019), p. 5233.
- [36] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).
- [37] O. G. Troyanskaya et al. “Nonparametric methods for identifying differentially expressed genes in microarray data”. In: *Bioinformatics* 18.11 (Nov. 2002), pp. 1454–1461.
- [38] C. Bravo González-Blas et al. “cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data”. In: *Nat Methods* 16.5 (May 2019), pp. 397–400.
- [39] Harold W Kuhn. “The Hungarian method for the assignment problem”. In: *Naval research logistics quarterly* 2.1-2 (1955), pp. 83–97.
- [40] Martin Jaggi. “Revisiting Frank-Wolfe: Projection-free sparse convex optimization”. In: *International Conference on Machine Learning*. PMLR. 2013, pp. 427–435.
- [41] 10x Genomics. “PBMC from a healthy donor - granulocytes removed through cell sorting (3k), Single Cell Multiome ATAC + Gene Expression Dataset by Cell Ranger ARC 2.0.0”. In: (2021).
- [42] 10x Genomics. “PBMC from a healthy donor - granulocytes removed through cell sorting (10k), Single Cell Multiome ATAC + Gene Expression Dataset by Cell Ranger ARC 2.0.0”. In: (2021).