*Bioinformatics*, 38(23), 2022, 5236–5244
https://doi.org/10.1093/bioinformatics/btac683
Advance Access Publication Date: 17 October 2022
Original Paper

Systems biology

# AlscEA: unsupervised integration of single-cell gene expression and chromatin accessibility via their biological consistency

**Elham Jafari [1], Travis Johnson [2], Yue Wang[3], Yunlong Liu[3], Kun Huang[2] and Yijie Wang [1,*]**

[1]Computer Science Department, Indiana University, Bloomington, IN 47408, USA, [2]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN 46202, USA and [3]Department of Medical & Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

*To whom correspondence should be addressed.

Associate Editor: Anthony Mathelier

Received on March 6, 2022; revised on October 7, 2022; editorial decision on October 10, 2022; accepted on October 14, 2022

## Abstract

**Motivation:** The integrative analysis of single-cell gene expression and chromatin accessibility measurements is essential for revealing gene regulation, but it is one of the key challenges in computational biology. Gene expression and chromatin accessibility are measurements from different modalities, and no common features can be directly used to guide integration. Current state-of-the-art methods lack practical solutions for finding heterogeneous clusters. However, previous methods might not generate reliable results when cluster heterogeneity exists. More importantly, current methods lack an effective way to select hyper-parameters under an unsupervised setting. Therefore, applying computational methods to integrate single-cell gene expression and chromatin accessibility measurements remains difficult.

**Results:** We introduce AlscEA—*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. AlscEA first defines a ranked similarity score to quantify the biological consistency between cell clusters across measurements. AlscEA then uses the ranked similarity score and a novel permutation test to identify cluster alignment across measurements. AlscEA further utilizes graph alignment for the aligned cell clusters to align the cells across measurements. We compared AlscEA with the competing methods on several benchmark datasets and demonstrated that AlscEA is highly robust to the choice of hyper-parameters and can better handle the cluster heterogeneity problem. Furthermore, AlscEA significantly outperforms the state-of-the-art methods when integrating real-world SNARE-seq and scMultiome-seq datasets in terms of integration accuracy.

**Availability and implementation:** AlscEA is available at https://figshare.com/articles/software/AlscEA_zip/21291135 on FigShare as well as {https://github.com/elhaam/AlscEA} onGitHub.

**Contact:** yijwang@iu.edu

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

## 1 Introduction

Advances in single-cell high-throughput technologies have enabled us to profile gene expression and chromatin accessibility at the single-cell resolution (Aparicio *et al.*, 2020; Chen *et al.*, 2019a; Eraslan *et al.*, 2019; Huang *et al.*, 2018; Luecken and Theis, 2019; Miao *et al.*, 2020; Risso *et al.*, 2018; Svensson *et al.*, 2020; van Dijk *et al.*, 2018; Vargo and Gilbert, 2020; Wang *et al.*, 2019). Several deep learning approaches have emerged to reveal more insights into cellular phenotypes (Cao and Gao, 2022; Chen *et al.*, 2021, 2022; Lin *et al.*, 2022; Liu *et al.*, 2021). Integration of the single-cell gene expression and chromatin accessibility measurements shed light on revealing gene regulation (Dong *et al.*, 2021; Efremova and Teichmann, 2020; Kelsey *et al.*, 2017; Lee *et al.*, 2020; Shema *et al.*, 2019). However, the heterogeneity among single cells presents challenges for integration (Efremova and Teichmann, 2020). Single-cell

gene expression and chromatin accessibility measure the cells at the transcriptomic and epigenomic layers, respectively. Identifying cluster–cluster or cell–cell correspondences is difficult across independently profiled measurements since they lack any shared features to integrate them. Single-cell dual-omics sequencing technologies (Chen *et al.*, 2019b; Liu *et al.*, 2019b) have been developed to tackle this problem by simultaneously profiling gene expression and chromatin accessibility in 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 single-cell measurements from different modalities.

Several unsupervised integrative methods have been developed to integrate the single-cell gene expression and chromatin accessibility measurements (Cao *et al.*, 2020; 2021; Cui *et al.*, 2014; Demetci *et al.*, 2022; Duren *et al.*, 2018; 2018; Liu *et al.*, 2019a; 2021; Wang *et al.*, 2020; Welch *et al.*, 2017). CoupleNMF (Duren *et al.*, 2018) utilizes the non-negative matrix factorization framework to integrate the single-cell gene expression and chromatin accessibility measurements at the cluster 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 the single-cell gene expression and chromatin accessibility profiles by minimizing the maximum mean discrepancy between them (Liu *et al.*, 2019a). 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 (Cao *et al.*, 2020). SCOT applies the Gromov–Wasserstein-based optimal transport to align the manifolds (Demetci *et al.*, 2022), but Pamona uses the partial Gromov–Wasserstein optimal transport (Cao *et al.*, 2021).

Another group of state-of-the-art utilizes biological feature relations across modalities. Seurat (Korsunsky *et al.*, 2019) embeds the data into a shared subspace using canonical correlation analysis and Harmony found the latent space using principal component analysis to project two domains in the shared space (Stuart *et al.*, 2019). GLUE (Cao and Gao, 2022) integrates the data using graph autoencoders by building a prior graph on regulatory inference and relation between different feature spaces. LIGER (Welch *et al.*, 2019) uses non-negative matrix factorization to map the common feature space into a shared latent space, and online iNMF (Gao *et al.*, 2021) extends the idea by utilizing online learning.

However, current methods suffer from two major problems. First, they are incapable of handling the cluster heterogeneity problem. When the clusters in the single-cell gene expression profile differ from those in the single-cell chromatin accessibility profile, they may generate poor alignment. CoupleNMF (Duren *et al.*, 2018) would fail because it requires the two datasets have the same number of clusters. Other methods assume that single-cell gene expression and chromatin accessibility share a similar manifold, which might not hold when the clusters across datasets are different. MMD-MA (Liu *et al.*, 2019a), UnionCom (Cao *et al.*, 2020) and SCOT (Demetci *et al.*, 2022) methods that rely on such assumptions would enforce the alignment between two different manifolds, which would lead to incorrect integration. Pamona (Cao *et al.*, 2021) attempts to resolve the cluster heterogeneity predicament by estimating the number of common cells across diverse measurements. However, the performance of the proposed estimation has not been comprehensively tested (Cao *et al.*, 2021). Second, all current methods' performance highly relies on hyper-parameter tuning, and they lack robustness to hyper-parameter selection. It is very challenging for current methods to find the optimal hyper-parameters under the unsupervised setting.

To overcome these limitations, we present AIscEA—*A*lignment-based *I*ntegration of single-cell gene *E*xpression and chromatin *A*ccessibility—a scalable and robust unsupervised computational method that explicitly uses biological consistency between gene expression and chromatin accessibility to guide the across-modality integration. AIscEA uses feature relations between two domains and is categorized in the same group as LIGER, iNMF and GLUE. First, AIscEA defines a rank-based similarity score to quantify the biological consistency between clusters across different domains. Then, based on the rank-based similarity and a novel designed permutation test, AIscEA identifies the domain-specific cell clusters and then finds corresponding cell clusters shared across single-cell gene expression and chromatin accessibility profiles. Furthermore, for these corresponding cell clusters across modalities, AIscEA applies a graph alignment method to elucidate the cell–cell correspondence (Zaslavskiy *et al.*, 2009).

We first validated the performance of AIscEA using SNARE-seq Human cell line mixtures data (Chen *et al.*, 2019b), which jointly captured accessible chromatin regions and gene expression profiles within the same cells, and therefore it provides cell–cell correspondence for validation. The benchmarking results demonstrate that AIscEA can resolve the cell-cluster heterogeneity problem and is robust to hyper-parameters while the state-of-the-arts were sensitive to hyper-parameter selection in experiments with existing heterogeneous cell cluster or another experiment when we slightly removed random cells. Furthermore, we show that AIscEA outperforms CoupleNMF (Duren *et al.*, 2018) in cell cluster alignment. In addition, we compared the performance of our method with state-of-the-art cell–cell integration methods MMD-MA (Liu *et al.*, 2019a), UnionCom (Cao *et al.*, 2020), SCOT (Demetci *et al.*, 2022), Pamona (Cao *et al.*, 2021), LIGER (Welch *et al.*, 2019), iNMF (Gao *et al.*, 2021) and GLUE (Cao and Gao, 2022) on real-world single-cell gene expression and chromatin accessibility profiles. We applied them to integrate SNARE-seq profilings of neonatal mouse cerebral cortex (Chen *et al.*, 2019b), adult mouse cerebral cortex (Chen *et al.*, 2019b) and two scMultiome-seq PBMC datasets from the healthy donors. We demonstrate that AIscEA significantly outperforms other methods in terms of the average FOSCTTM score (Liu *et al.*, 2019a), demonstrating its superiority in identifying the cell–cell correspondence. More importantly, AIscEA addresses hyper-parameter sensitivity problem in the state-of-the-arts as AIscEA is robust to hyper-parameter in an unsupervised setting to align cells between heterogeneous single-cell modalities.

## 2 Materials and methods

### 2.1 Method overview

AIscEA is an alignment-based method that can identify the cluster–cluster 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 (Cao *et al.*, 2021, 2020; Demetci *et al.*, 2022; Liu *et al.*, 2019a), AIscEA does not rely on the assumption of similarity between the manifolds of the entire single-cell gene expression and chromatin accessibility measurements. However, AIscEA 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 clusters and also between cells across the measurements (Halstead *et al.*, 2020; Quinlan and Hall, 2010; Rainer *et al.*, 2019; Silva *et al.*, 2016; Sun *et al.*, 2019). AIscEA quantifies such feature relations across domains using a rank-based similarity score and further unitizes the similarity score to direct the cluster and cell–cell alignments. As shown in Figure 1, AIscEA consists of three steps: (i) cluster identification, (ii) cluster–cluster alignment and (iii) cell–cell alignment. In the following, we will elaborate on the details of each step.

### 2.2 Cell cluster identification

We first identify cell clusters within single-cell gene expression and chromatin accessibility measurements (Fig. 1b) using state-of-the-art clustering methods. For single-cell gene expression, we used Scanpy package's classical graph-based clustering method (Becht *et al.*, 2019; Traag *et al.*, 2019; Troyanskaya *et al.*, 2002; Wolf *et al.*, 2018) to identify $n$ clusters in $\mathcal{C} = \{C_1, C_2, \ldots C_n\}$. For single-cell chromatin accessibility measurement, we first use cisTopic (Bravo González-Blas *et al.*, 2019) to extract regulatory topics and then use the extracted features to divide cells into $m$ clusters $\mathcal{D} = \{D_1, D_2, \ldots D_m\}$. More details are provided in Supplementary Section A.

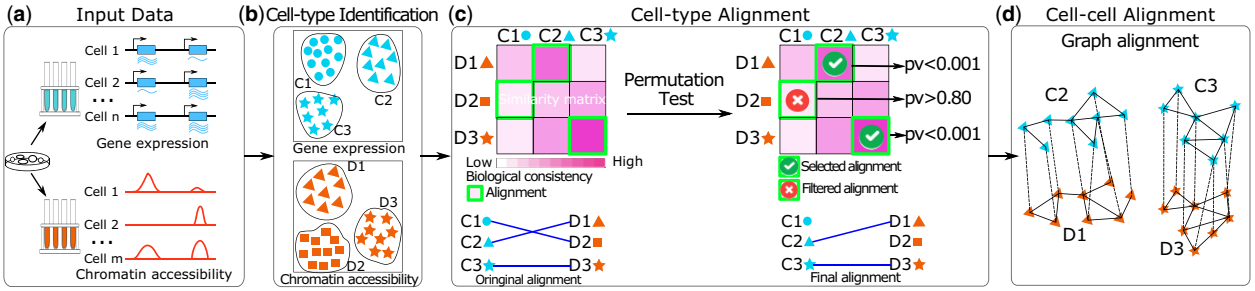


**Fig. 1.** Overview of AIscEA. (**a**) Input datasets of single-cell RNA-seq and single-cell ATAC-seq measurements. (**b**) Clustering and cell cluster identification. (**c**) Cluster alignment using biological consistency and calculating the *P*-values by a novel permutation test. (**d**) AIscEA finds cell–cell alignment using a graph alignment method for each pair of mapped clusters

Furthermore, for each cluster $C_i, \forall i$ identified in the single-cell gene expression measurement, AIscEA identifies the set of differential over-expressed genes $\mathcal{G}_{C_i} = \{g_1, g_2, \ldots\}$. AIscEA then ranks these genes by their expression log 2 fold changes with respect to their expression in the rest of the clusters 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 cluster $D_j, \forall j$ identified in the single-cell chromatin accessibility measurement, AIscEA identifies the significantly accessible locations using the predictive distribution calculated by cisTopic (Bravo González-Blas *et al.*, 2019). Next, we identify the overlap between these significantly accessible locations and the promoter regions of the expressed genes. AIscEA uses $\mathcal{H}_{D_j} = \{g_1, g_2, \ldots\}$ to present the set of genes, whose promoter regions overlap with the significantly accessible locations in cluster $D_j$. More details of this section is provided in Supplementary Section A.

## 2.3 Cell cluster alignment

After cell cluster (cluster for short) identification, $n$ and $m$ clusters 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 clusters $m$ and $n$ may differ. Furthermore, the correspondence is unknown between $n$ clusters in the single-cell gene expression measurements and $m$ clusters in the single-cell chromatin accessibility measurements.

We propose to use the biological consistency between gene expression and chromatin accessibility to align clusters across different modalities (as shown in Fig. 1). Furthermore, AIscEA adopts a novel permutation test to find statistically significant biological consistency between the aligned clusters. Sections 2.3.1 and 2.3.2 ensure we only map clusters with highly biological similarity score as we only keep aligned clusters with statistically significant similarity score and filter out heterogeneous clusters (Fig. 1c).

### 2.3.1 Cluster 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 (Halstead *et al.*, 2020; Quinlan and Hall, 2010; Rainer *et al.*, 2019; Silva *et al.*, 2016; Sun *et al.*, 2019). AIscEA defines a ranked similarity score $S$ to quantify such biological consistency between clusters. Mathematically, the ranked similarity score $S(C_i, D_j)$ between cluster $C_i$ in single-cell gene expression and cluster $D_j$ in single-cell chromatin accessibility data can be computed by:

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

where $\mathcal{G}_{C_i}$ is the set of differential over-expressed genes in cluster $C_i$ identified in the single-cell gene expression measurements. $\mathcal{H}_{D_j}$ is the set of genes whose promoter regions are significantly accessible in cluster $D_j$ 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 Eq. (1), we know that $S(C_i, D_j)$ is large when (i) $R_{C_i}(g)$ is small, meaning the top ranking genes' promoter regions should be significantly accessible; (ii) $|\mathcal{G}_{C_i} \cap \mathcal{H}_{D_j}|$ is large, meaning most of the highly over-expressed genes should have significantly accessible promoter regions. Figure 2a–c illustrates a toy example of how $S(C_i, D_j)$ is computed.

From Eq. (1), we compute the biological consistency between $n$ clusters in $\mathcal{C} = \{C_1, C_2, \ldots C_n\}$ and $m$ clusters in $\mathcal{D} = \{D_1, D_2, \ldots D_m\}$. Without loss of generality, we assume that $n \leq m$ (if $n > m$, we can add dummy clusters in $\mathcal{D}$ to make $n = m$). Then the cluster alignment across measurements can be obtained by maximizing the biological consistency between aligned clusters across measurements, which can be formulated as a linear assignment problem:

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

where $X$ is a binary assignment matrix, where $X_{ij} = 1$ denotes that cluster $C_i$ corresponds to cluster $D_j$. The constraint set $\Omega = \{X \in \{0,1\}^{n \times m} : X1_m = 1_n, X^\top 1_n \leq 1_m\}$ enforces each cluster in $\mathcal{C}$ is assigned to one and only one cluster in $\mathcal{D}$. The linear assignment problem can be efficiently solved by the Hungarian algorithm (Kuhn, 1955).

### 2.3.2 Resolving the cluster heterogeneity problem via a novel permutation test

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

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 $\Phi^{-1}_{\mathcal{S}_Z} : \mathbb{Z}^+ \rightarrow \mathcal{S}_Z$ 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 aligned clusters found by Eq. (2) are greater or equal to the ranked similarity scores of the randomly aligned clusters using all the similarity scores after permutation as a more comprehensive null. After solving Eq. (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

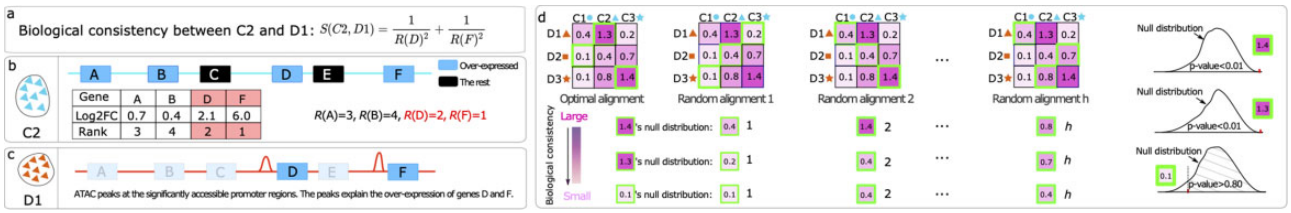


**Fig. 2.** (**a**) An illustration of computing the biological consistency between (**b**) a cell type in gene expression and (**c**) a cell type in chromatin accessibility. (**d**) An explanation of the proposed permutation test to calculate *P*-values for aligned cell clusters

can get the corresponding similarity score $S(C_i, D_j)$ and its ranking among all alignments by $I = \Phi_{S_{X^*}}(S(C_i, D_j))$. We then generate $h$ random cluster alignments by uniformly sampling $h = 1000$ assignment matrices $Z_1, \ldots, Z_h \in \Omega$. The null distribution of the $I$th ranking similarity score can be estimated by $\{\Phi^{-1}_{S_{Z_1}}(I), \ldots, \Phi^{-1}_{S_{Z_h}}(I)\}$ (where $\{\Phi^{-1}_{S_{Z_l}}(I)$ is the $I$th ranked similarity score in the random alignment $Z_l$). By comparing $S(C_i, D_j)$ with $\Phi^{-1}_{S_{Z_l}}(I), l = 1, \ldots, h$, we can calculate the *P*-value by $1 - \frac{1}{h}\left|\left\{l | \Phi^{-1}_{S_{Z_l}}(I) \leq S(C_i, D_j)\right\}\right|$, where $|\cdot|$ is the cardinality of a set. For the alignment $X^*_{ij} = 1$ whose corresponding *P*-value is significant ($\leq 0.01$), we consider it a true alignment. For the alignment $X^*_{ij} = 1$ whose corresponding *P*-value is not significant (*P*-value $> 0.01$), we consider the corresponding clusters in this alignment unique clusters within their measurement. Figure 2d illustrates how the permutation is calculated.

#### 2.3.3 Hyper-parameters selection scheme for the cluster–cluster alignment

Due to the heterogeneity across measurements the number of clusters $n$, $m$ typically differs ($n \neq m$). The selection of $n$ and $m$ would influence the performance of the cluster alignment in AIscEA. 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. AIscEA applies Leiden clustering (Traag *et al.*, 2019) to identify clusters using the resolution parameter. Therefore, we propose an effective and scalable scheme to select the resolution parameter rather than the number of clusters as following.

Our heuristic approach sets the range for the resolution parameter $r_e$ for single-cell gene expression measurement $r_e \in \{0.1, 0.15, 0.2, \ldots, 1.5\}$ and the resolution parameter $r_c$ for single-cell chromatin accessibility measurement $r_c \in \{0.1, 0.15, 0.2, \ldots, 1.5\}$. 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 clusters in gene expression measurement when the resolution parameter is set to $r_e$, $m_c$ is the number of identified clusters in chromatin accessibility measurement when the resolution parameter is set to $r_c$, and $o$ is the number of aligned clusters between $n_e$ and $m_c$ identified by the cluster alignment method in AIscEA (as explained in Section 2.3). In the end, after screening all resolution parameters, we select the ones yielding the largest *alignment ratio L*.

In the experiment Section 4.1.1, we empirically show that the proposed heuristic approach can select $r_e$ and $r_c$ that result in descent cluster alignment results for all datasets in a completely unsupervised manner.

### 2.4 Cell–cell alignment

Once we identify clusters $C_i$ and $D_j$ are aligned together, we can further find the cell–cell correspondence between the cells in $C_i$ and $D_j$. AIscEA assumes that $C_i$ and $D_j$ consist of a set of cells $C_i = \{c_i^1, c_i^2, \ldots\}$ and $D_j = \{d_j^1, d_j^2, \ldots\}$, respectively. Since the gene expression of cells in $C_i$ and the chromatin accessibility of cells in $D_j$ are different measurements for cells of the same cell type, we confidently assume that their low-dimensional manifold is similar. Hence, a graph alignment method is employed to find the cell–cell correspondence (Zaslavskiy *et al.*, 2009).

AIscEA 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 $V_1 = C_i = \{c_i^1, c_i^2, \ldots\}$ 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 $V_2 = D_j = \{d_j^1, d_j^2, \ldots\}$ are cells in $D_j$ (details in the Supplementary 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 Zaslavskiy *et al.* (2009). 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 AIscEA can be formulated (Zaslavskiy *et al.*, 2009) as:

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

$A_1$ 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 Eq. (1) (details in the Supplementary Materials). In the objective function in Eq. (3), the first term $\mathrm{Tr}(A_1^\top P A_2 P^\top)$ 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. $\lambda$ is a hyper-parameter that balances the trade-off in the objective function (Eq. 3). The optimization in Eq. (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 (Jaggi, 2013) algorithm and the path-relinking technique to solve (Eq. 3) (details in Supplementary Section B).

#### 2.4.1 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 $\lambda$ (the regularizer in Eq. 3). In the experiment Section 4.1.2, we show that the cell–cell alignment in AIscEA is robust to the selection of $k$ and $\lambda$. Therefore, we set $k$ and $\lambda$ to default values in practice.

## 3 Experimental setup

### 3.1 Competing methods

AIscEA can identify cluster alignment between scRNA-seq and scATAC-seq datasets, therefore, we compare AIscEA's performance on cluster alignment with CoupleNMF (Duren *et al.*, 2018), which is the state-of-the-art cluster alignment method. In addition, we compared AIscEA's cell–cell alignment with the current state-of-the-art cell–cell alignment methods MMD-MA (Liu *et al.*, 2019a), UnionCom (Cao *et al.*, 2020), SCOT (Demetci *et al.*, 2022), Pamona (Cao *et al.*, 2021), LIGER (Welch *et al.*, 2019), iNMF (Gao *et al.*, 2021) and GLUE (Cao and Gao, 2022).

### 3.2 Data

*SNARE-seq Human* (Chen *et al.*, 2019b) 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 cluster–cluster alignment and cell–cell alignment.

Moreover, we evaluated capability of handling the cluster heterogeneity problem for different methods by generating *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 toward hyper-parameter selection, we generated 10 *SNARE-seq Human_R5%* and 10 *SNARE-seq Human_R10%*, where 5% of cells and 10% cells are randomly removed from the original *SNARE-seq Human*. Additionally, 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 (10x Genomics, 2021a,b): *scMultiome PBMC 3k* (scMultiome-seq PBMC of a healthy donor with 3*k* cells) and *scMultiome PBMC 12k* (scMultiome-seq PBMC of a healthy donor with 12*k* 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 Section C.1.

### 3.3 Metrics

We first introduce the metric we use for evaluating the cluster alignment. When two clusters 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 clusters are expected to have a larger number of overlapping cells. Therefore, we use the overlap coefficient to measure the overlap between aligned clusters. Specifically, if cluster $C_i$ is aligned to cluster $D_j$, 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 clusters 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 clusters. Another metric we use to evaluate the cluster alignment is the average Silhouette score per cluster to measure cluster cohesion. We expect the cells in the same clusters to be similar to other cells in their own cluster type, but different from other cluster types.

For evaluating the cell–cell alignment, we use the average FOSCTTM score (Liu *et al.*, 2019a), which has been widely used for evaluating single-cell multi-omics integration methods (Cao *et al.*, 2021; Cao and Gao, 2022; Demetci *et al.*, 2022; Liu *et al.*, 2019a). 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 Section D.1. Another metric is the cell coverage, which is the number of cells with mapped 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 Sections 2.3.3 and 2.4.1. For CoupleNMF Duren *et al.* (2018), we set the number of clusters based on the ground truth and for the rest of the hyper-parameters, we use the suggested hyper-parameters. For MMD-MA (Liu *et al.*, 2019a), UnionCom (Cao *et al.*, 2020), SCOT (Demetci *et al.*, 2022), Pamona (Cao *et al.*, 2021), LIGER (Welch *et al.*, 2019), iNMF (Gao *et al.*, 2021) and GLUE (Cao and Gao, 2022) under the unsupervised setting, we use the following strategy to find the optimal hyper-parameters. We performed a grid search to find the optimal hyper-parameters using *SNARE-seq Human* dataset as golden standard. Then, we use the optimal hyper-parameters of *SNARE-seq Human* for real-world datasets (more details in Supplementary Sections SC.5 and SC.6).

### 3.5 Computational resource

All experiments are processed on an Intel(R) Core(TM) i7-6850K CPU @ 3.60 GHz CPU with 62 GB memory and GPU computations on a single GeForce GTX 1080 Ti with VRAM of 11 GB. If a method fails to run on a large-scale dataset due to memory shortage, we report a memory error as shown in Table 1.

## 4 Results

### 4.1 Benchmarking using SNARE-seq human

SNARE-seq human cell line mixtures provides the ground truth information for validating cluster alignment and cell–cell alignment. Therefore, we first use it to validate AIscEA's hyper-parameter selection scheme proposed in Section 2.3.3 for cluster 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 cluster heterogeneity problem, as in real-world datasets, the number of clusters may differ between two domains.

#### 4.1.1 Validation of the hyper-parameter selection scheme for the cluster alignment in AIscEA

In Section 2.3.3, we propose an approach to select the resolution hyper-parameters in the Leiden clustering in AIscEA, which determine the number of clusters in scRNA-seq *n*, and the number of clusters in scATAC-seq *m* for the cluster alignment in AIscEA. 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 cluster alignment.

We applied the proposed scheme in Section 2.3.3 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 data can be found in Section 3.2). We show the hyper-parameters screening results in Supplementary Figure S4 and the size of each dot corresponds to its *alignment ratios* defined in Section 2.3.3. 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 cluster alignment (computed as $U$ defined in Section 3.3). Darker blue means more number of

**Table 1.** Running time comparisons in minutes

| Sample size | 1047 | 2711 | 5081 | 10 309 | 11 898 |
|---|---|---|---|---|---|
| MMD-MA | 6.0 | 52.0 | 203.5 | 862.6 | 1276.9 |
| UnionCom | 19.6 | 26.3 | 216.5 | 1117.2 | E |
| SCOT | 2.2 | 2.8 | 8.0 | 10.8 | 19.9 |
| Pamona | 3.0 | 8.3 | 37.3 | 262.8 | 306.1 |
| LIGER | 7.0 | 3.2 | 5.3 | 99.3 | E |
| iNMF | 6.7 | 2.2 | 4.0 | 10.5 | E |
| GLUE | 14.4 | 20.2 | 75.6 | E | E |
| **AIscEA** | 10.6 | 16.5 | 131.0 | 223.2 | 123.0 |

*Note*: E means memory error as UnioinCom, LIGER, iNMF and GLUE require higher memory for a larger sample size.

overlapping cells are identified by the cluster alignment, which means the performance of the cluster alignment is more promising with higher cell coverage. As shown in Supplementary Figure S4, the large-size dots always appear in dark blue color, demonstrating that the *alignment ratio* and the performance of the cluster alignment method is positively correlated. Therefore, we can use the *alignment ratio* to guide the selection of the resolution hyper-parameters used in AIscEA 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 hyper-parameters may yield equivalently good cluster alignments. We have the same observation from the screening results for more real-world datasets in the Supplementary Figure S2.

#### 4.1.2 Benchmarking hyper-parameter robustness in cell–cell alignment

In this section, we compare AIscEA with the state-of-the-arts 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, prior knowledge to guide the hyper-parameter selection lacks. If a method is sensitive to hyper-parameters, its performance is unreliable for real-world applications.

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 2. Cell coverage for AIscEA and all other methods are 1047 cells. The grid search hyper-parameter tuning details are elaborated in Supplementary Section C.2. Although we tried using Seurat and Harmony in our experiments, we found a key function that Seurat and Harmony rely on cannot run through, and we are yet to resolve this. As shown in the Table 2, AIscEA is competitive with SCOT and GLUE on achieving the smallest FOSCTTM score, which is superior to the rest of the 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 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. We generated 10 *SNARE-seq Human_R5%* data and 10 *SNARE-seq Human_R10%* datasets (description of the data is in Section 3.2). Then we applied each method to them using their optimal set of hyper-parameters (in Supplementary Table S1). Cell coverage for AIscEA and competing methods in these experiments is 1047. Figure 3a and b exhibits 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. For best competitor GLUE, we also included their default set of hyper-parameters and we noticed a gap between this result and our best found set of hyper-parameters which showed GLUE's sensitivity to hyper-parameter selection. All these results demonstrate that AIscEA is more robust to the choice of hyper-parameters than all other competing methods.

#### 4.1.3 Benchmarking in solving the cluster heterogeneity problem

Next, we benchmark all methods on their ability to resolve the cluster 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 clusters. If we cannot distinguish the clusters that have correspondence and other clusters that have not, the alignment between the two measurements would be misleading.

To simulate the cluster heterogeneity problem, we generated the *SNARE-seq Human_Heterogeneity* data (description of the data is in Section 3.2). We first ran each competing method over an extensive grid of suggested set of hyper-parameters and showed the results

**Table 2.** 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.

| | *SNARE-seq Human* | | | *Human_Heterogeneity* | | |
|---|---|---|---|---|---|---|
| Method | Minimum | Mean | SD | Minimum | Mean | SD |
| 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 |
| LIGER | 0.425 | 0.444 | 0.009 | – | – | – |
| iNMF | 0.474 | 0.488 | 0.002 | – | – | – |
| GLUE | **0.143** | 0.299 | 0.105 | 0.198 | 0.340 | 0.068 |

*Note*: LIGER and iNMF exitted with an error for *Human_Heterogeneity*. Bold entries in each column show the best FOSCTTM values.

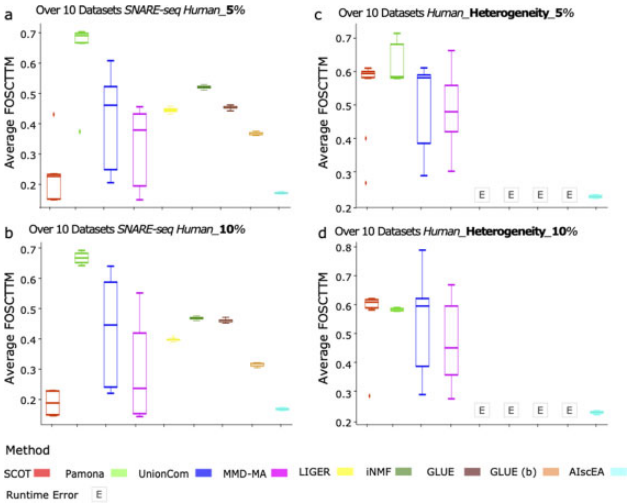


**Fig. 3.** Box plots of the average FOSCTTM over 10 (**a**) *SNARE-seq Human_R5%*, (**b**) *SNARE-seq Human_R10%*, (**c**) *SNARE-seq Human_Heterogeneity_R5%* and (**d**) *SNARE-seq Human_Heterogeneity_R10%* datasets (40 downsampled datasets in total). UMAP of the RNA-seq and ATAC-seq data are shown in Supplementary Figure S3. 5% and 10% of cells are removed from SNARE-seq_Human (a, b) and Human_Heterogeneity (c, d). We used their best set of hyper-parameters from grid search analysis SNARE-seq_Human and Human_Heterogeneity (Supplementary Table S1 in Supplementary Materials) for SCOT, Pamona, UnionCom, MMD-MA, LIGER, iNMF and GLUE. We also included default recommended hyper-parameters in our further experiments only for the highest competitor, GLUE. In further analyses, GLUE's performance using default hyper-parameters is shown as GLUE (b). LIGER, iNMF and GLUE faced an issue during alignment for SNAREseq_Heterogeneous data, and we marked them under 'Runtime Error' in (c, d)

for *SNARE-seq Human_Heterogeneity* in Table 2. AIscEA can identify the heterogeneous cell type, exclude it, and map the shared clusters 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 cluster heterogeneity problem.

Furthermore, we applied each method using its optimal hyper-parameters found on *SNARE-seq Human_Heterogeneity* data (shown in Supplementary Table S1) to 10 *Human_Heterogeneity_R5%* data and 10 *SNARE-seq Human_Heterogeneity_R10%* data (description in 3.2). Figure 3c and d shows the comparison results. Apparently, AIscEA achieved the smallest average FOSCTTM score and was more robust to its hyper-parameters. When one cluster was missing, GLUE's FOSCTTM score is the second best after our method, as shown in

Table 2. But in the experiment over 10 *Human_Heterogeneity_R5%* data and 10 *SNARE-seq Human_Heterogeneity_R10%* data, GLUE faced a Run time Error that no meta cells are found so it could not perform the alignment, as shown in Figure 3. All above experiments demonstrate that AIscEA is the most robust, having the lowest FOSCTTM score and on-par coverage among all methods to resolve the cluster heterogeneity problem.

## 4.2 Comparison of the cluster alignment

In this section, we compare AIscEA with CoupleNMF (Duren *et al.*, 2018) in terms of cluster alignment. We applied both methods to *SNARE-seq Human*, *SNARE-seq Mouse 5k*, *SNARE-seq Mouse 10k*, *scMultiome-seq PBMC 3k* 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 clusters. 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 Figure 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 Figure S1e–h. As shown in Figure 4, for AIscEA, the cells in the aligned clusters in both scRNA-seq and scATAC-seq are well isolated as expected for a good cluster alignment. However, for CoupleNMF, the cells in the aligned clusters are mixed together. We further evaluated the performance of both methods in terms of overlapping coefficient [defined in Eq. (4)] and the Silhouette score shown in Table 3. Clearly, AIscEA achieves much higher overlapping coefficients and Silhouette scores, which demonstrates that AIscEA outperform CoupleNMF in terms of cluster alignment.

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

We compared AIscEA with competing cell–cell alignment methods MMD-MA (Liu *et al.*, 2019a), UnionCom (Cao *et al.*, 2020), SCOT (Demetci *et al.*, 2022), Pamona (Cao *et al.*, 2021), LIGER (Welch *et al.*, 2019), iNMF (Gao *et al.*, 2021) and GLUE (Cao and Gao, 2022) 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.

### 4.3.1 AIscEA outperforms current methods on the real-world SNARE-seq and shows high robustness to hyper-parameters

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

Figure 5a and b illustrates the cluster alignment identified by AIscEA for *SNARE-seq Mouse 5k*. And Figure 5c shows the comparison between different methods in terms of the average FOSCTTM score and cell coverage. As shown, GLUE achieves the lowest average FOSCTTM score, and AIscEA is the second best for this dataset. 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 average FOSCTTM score and robustness, it is obvious that AIscEA significantly outperforms all the current methods.

Figure 5d and e illustrate the cluster alignment identified by AIscEA for *SNARE-seq Mouse 10k*. Figure 5f 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 (9373 cells out of 10 309 cells). LIGER and iNMF could not run on our system for this data due to high memory they required. Overall, considering average FOSCTTM score, cell coverage and robustness using default hyper-parameters in Supplementary Table S1, AIscEA significantly outperforms all the current methods.

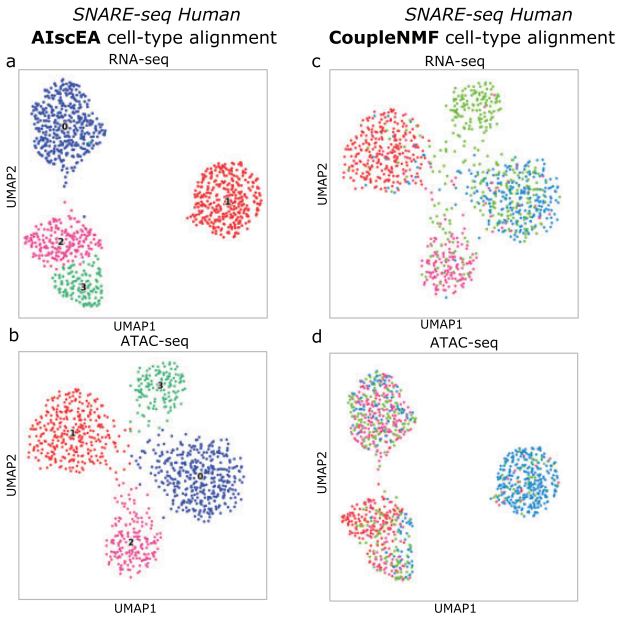


**Fig. 4.** (**a** and **b**) Aligned cell clusters identified by AIscEA compared to (**c**, **d**) aligned clusters identified by CoupleNMF using in *SNARE-seq Human*. See also Supplementary Figure S1

**Table 3.** Cell cluster alignment comparison.

| Method | *SNARE-seq Human* | *scMultiome-seq PBMC 3k* |
|---|---|---|
| **AIscEA** | Overlap coef: **0.911** | Overlap coef: **0.884** |
| | Silhouette score: **0.618** | Silhouette score: **0.463** |
| **CoupleNMF** | Overlap coef: 0.202 | Overlap coef: 0.398 |
| | Silhouette score: 0.146 | Silhouette score: 0.032 |

*Note*: See also Figure 4 and Supplementary Figure S3. CoupleNMF failed to run on other datasets due to memory error.

Bold entries show the best overlap coefficient and silhouette score achieved.

### 4.3.2 AIscEA outperforms 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).

Figure 5g and h illustrate the cluster alignment identified by AIscEA for *scMultiome-seq PBMC 3k*. Figure 5i compares 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.

Figure 5j and k illustrate the cluster alignment identified by AIscEA for *scMultiome-seq PBMC 12k*. Figure 5l 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 minimal average FOSCTTM score, AIscEA outperforms all the current methods. Although cell coverage is lower for this dataset due to higher noise in clustering, we filtered out several noisy clusters from our downstream analysis. This ensures having high-quality cell–cell alignment and not sacrificing quality for quantity of aligned cells.

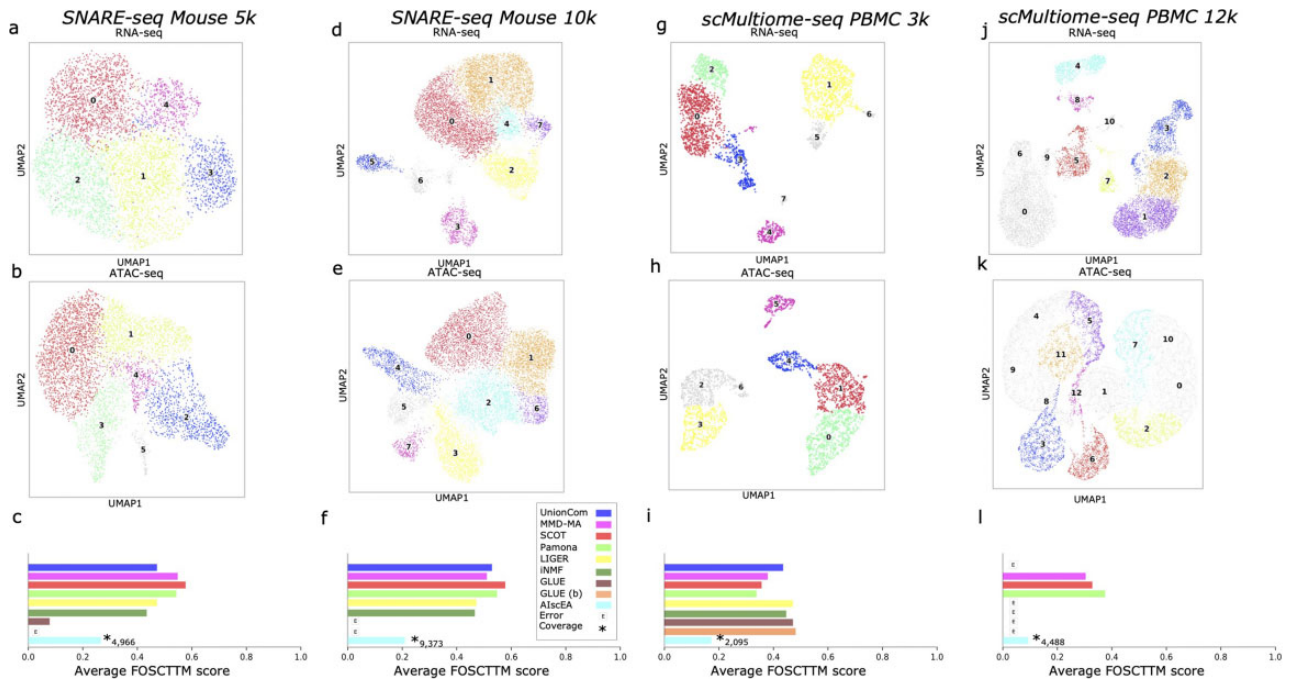


**Fig. 5.** (**a** and **b**) The UMAPs of the RNA-seq and ATAC-seq of *SNARE-seq Mouse 5k*. Clusters 0, 1, 2, 3, 4 in RNA-seq data are aligned to clusters 0, 1, 3, 2, 4 in ATAC-seq data, respectively. (**c**) The bar plots of the average FOSCTTM scores for all methods. The shorter the bar the better the method performs. The number after '*' for AIscEA shows the cell coverage in each data. The cell coverage for state-of-the-art (methods except AIscEA) is the sample size of each data. Cell coverage for AIscEA is 4966, and cell coverage of the other state-of-the-art methods is 5081. (**d** and **e**) The UMAPs of the RNA-seq and ATAC-seq of *SNARE-seq Mouse 10k*. Clusters 0, 1, 2, 3, 4, 5, 7 in RNA-seq data are aligned to clusters 0, 1, 3, 7, 2, 4, 6 in ATAC-seq data, respectively. (**f**) The bar plots of the average FOSCTTM scores for all methods for *SNARE-seq Mouse 10k* with 10 309 cells. (**g** and **h**) The UMAPs of the RNA-seq and ATAC-seq of *scMultiome PBMC 3k*. Clusters 0, 1, 2, 3, 4 in RNA-seq data are aligned to clusters 1, 3, 0, 4, 5 in ATAC-seq data, respectively. (**i**) The bar plots of the average FOSCTTM scores for all methods for *scMultiome PBMC 3k*. Cell coverage for AIscEA is 2095, and cell coverage of the other state-of-the-art methods is 2711. (**j** and **k**) The UMAPs of the RNA-seq and ATAC-seq of *scMultiome PBMC 12k*. Clusters 1, 2, 3, 4, 5, 7, 8 in RNA-seq data are aligned to clusters 5, 11, 3, 7, 6, 2, 12 in ATAC-seq data, respectively. (**l**) The bar plots of the average FOSCTTM scores for all methods for *scMultiome PBMC 12k*. Cell coverage for AIscEA is 4488, and cell coverage of the other state-of-the-art methods is 11 898

On the other hand, AIscEA showed scalability in larger datasets, as shown in Figure 5 UnionCom, LIGER, iNMF and GLUE are not scalable and need high memory for large datasets (See Table 1 for runtime analysis).

## 5 Conclusion

In this study, we proposed AIscEA, a robust unsupervised computational method for integrating single-cell gene expression and chromatin accessibility measurements. Unlike other approaches, AIscEA relies on the biological consistency between feature spaces of two measurements to guide the integration. We compared AIscEA with the state-of-the-art on *SNARE-seq* human cell line mixtures datasets and demonstrated that AIscEA can effectively select hyper-parameters. Moreover, AIscEA handles the cluster heterogeneity problem. Furthermore, we showed that AIscEA significantly outperforms previous methods in the real-world mouse SNARE-seq and scMultiome-seq datasets. Considering the trade-off between average FOSCTTM score and cell coverage, AIscEA outperforms the competing methods yielding the lowest FOSCTTM score as it strength. We addressed the gap that state-of-the-art approaches showed high coverage but poor FOSCTTM score.

Several innovations developed in this work contributed to the performance of AIscEA in cell–cell alignment. First, the ranked similarity score enables us to compare the clusters across measurements. The ranked similarity score is the key to estimating the similarity between clusters from different modalities. Second, the novel permutation test can distinguish the true cluster 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 *k*-nearest neighbor graphs to characterize the low-dimensional manifold. It is a notable advantage that AIscEA can identify heterogeneous clusters and exclude them from the cell–cell alignment in further analysis. Our future direction is to recruit more cells in the integration. However, it is important to remember that AIscEA assumes the input single-cell data to consist of separable cell clusters and it is not designed to perform an integrative analysis of single-cell trajectories datasets that do not fulfill this constraint.

AIscEA is computationally efficient (shown in Table 1) and practical in real-world data due to its robustness to hyper-parameter selection and superior and biologically meaningful cell alignment. We are highly interested in increasing the cell coverage and sharing more insights with the research community in our future work. We believe AIscEA is the milestone for integrating single-cell gene expression and chromatin accessibility measurements. Furthermore, it also provides a stepping stone for integrating other single-cell measurements.

## Funding

This research was supported by National Institutes of Health [R35GM147241-01 to Y.W.] and precision health intuitive at Indiana University for Y.W.

*Conflict of Interest*: none declared.

## Data availability

No new data were generated or analysed in support of this research.

## References

10x Genomics. (2021a) 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. https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0.
10x Genomics. (2021b) 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. https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0.
Aparicio,L. *et al.* (2020) A random matrix theory approach to denoise single-cell data. *Patterns (NY)*, **1**, 100035.
Bravo González-Blas,C. *et al.* (2019) cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods*, **16**, 397–400.
Becht,E. *et al.* (2019) Dimensionality reduction for visualizing single-cell data using Umap. *Nat. Biotechnol.,* **37**, 38–44.
Cao,K. *et al.* (2020) Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, **36**, i48–i56.
Cao,K. *et al.* (2021) Manifold alignment for heterogeneous singlecell multiomics data integration using Pamona. *Bioinformatics*, **38**, 211–219.
Cao,Z.J. and Gao,G. (2022) Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.*, **40**, 1458–1466.
Chen,H. *et al.* (2019a) Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.*, **20**, 241.
Chen,S. *et al.* (2019b) High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.*, **37**, 1452–1457.
Chen,S. *et al.* (2021) RA3 is a reference-guided approach for epigenetic characterization of single cells. *Nat. Commun.*, **12**, 2177.
Chen,X. *et al.* (2022) Cell type annotation of single-cell chromatin accessibility data via supervised bayesian embedding. *Nat. Mach. Intell.*, **4**, 116–126.
Cui,Z. *et al.* (2014) Generalized unsupervised manifold alignment. *Adv. Neural Inf. Process. Syst.*, **27**, 2429–2437.
Demetci,P. *et al.* (2022) SCOT: Single-cell multi-omics alignment with optimal transport. *J. Comput. Biol.,* **29**, 3–18.
Dong,X. *et al.* (2021) Review of multi-omics data resources and integrative analysis for human brain disorders. *Brief. Funct. Genomics*, **20**, 223–234.
Duren,Z. *et al.* (2018) Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl. Acad. Sci. USA*, **115**, 7723–7728.
Efremova,M. and Teichmann,S.A. (2020) Computational methods for single-cell omics across modalities. *Nat. Methods*, **17**, 14–17.
Eraslan,G. *et al.* (2019) Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.*, **10**, 390.
Gao,C. *et al.* (2021) Iterative single-cell multi-omic integration using online learning. *Nat. Biotechnol.*, **39**, 1000–1007.
Halstead,M.M. *et al.* (2020) Systematic alteration of ATAC-seq for profiling open chromatin in cryopreserved nuclei preparations from livestock tissues. *Sci. Rep.*, **10**, 5230.
Huang,M. *et al.* (2018) SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods*, **15**, 539–542.
Jaggi,M. (2013) Revisiting Frank-Wolfe: projection-free sparse convex optimization. In: *International Conference on Machine Learning*. PMLR, Atlanta, GA, USA, pp. 427–435.
Kelsey,G. *et al.* (2017) Single-cell epigenomics: recording the past and predicting the future. *Science*, **358**, 69–75.
Korsunsky,I. *et al.* (2019) Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, **16**, 1289–1296.
Kuhn,H.W. (1955) The Hungarian method for the assignment problem. *Naval Res. Logistics*, **2**, 83–97.
Lee,J. *et al.* (2020) Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med.*, **52**, 1428–1442.
Lin,Y. *et al.* (2022) scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nat. Biotechnol.*, **40**, 703–710.
Liu,J. *et al.* (2019a) Jointly embedding multiple single-cell omics measurements. *Algorithms Bioinform.*, **143**, 10.
Liu,L. *et al.* (2019b) Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat. Commun.*, **10**, 470.
Liu,Q. *et al.* (2021) Simultaneous deep generative modeling and clustering of single cell genomic data. *Nat. Mach. Intell.*, **3**, 536–544.
Luecken,M.D. and Theis,F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, **15**, e8746.
Miao,Z. *et al.* (2020) Putative cell type discovery from single-cell gene expression data. *Nat. Methods*, **17**, 621–628.
Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
Rainer,J. *et al.* (2019) ensembldb: an R package to create and use ensembl-based annotation resources. *Bioinformatics*, **35**, 3151–3153.
Risso,D. *et al.* (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.*, **9**, 284.
Shema,E. *et al.* (2019) Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nat. Genet.*, **51**, 19–25.
Silva,T.C. *et al.* (2016) Analyze cancer genomics and epigenomics data using bioconductor packages. *F1000Res*, **5**, 1542.
Stuart,T. *et al.* (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.e21.
Sun,Y. *et al.* (2019) Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas*, **156**, 29.
Svensson,V. *et al.* (2020) A curated database reveals trends in single-cell transcriptomics. *Database (Oxford)*, **2020**, baaa073.
Traag,V.A. *et al.* (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.*, **9**, 5233.
Troyanskaya,O.G. *et al.* (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454–1461.
van Dijk,D. *et al.* (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell*, **174**, 716–729.e27.
Vargo,A.H.S. and Gilbert,A.C. (2020) A rank-based marker selection method for high throughput scRNA-seq data. *BMC Bioinformatics*, **21**, 477.
Wang,C. *et al.* (2020) Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol.*, **21**, 198.
Wang,J. *et al.* (2019) Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods*, **16**, 875–878.
Welch,J.D. *et al.* (2017) MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol.*, **18**, 138.
Welch,J.D. *et al.* (2019) Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, **177**, 1873–1887.e17.
Wolf,F.A. *et al.* (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
Zaslavskiy,M. *et al.* (2009) Global alignment of protein–protein interaction networks by graph matching methods. *Bioinformatics*, **25**, i259–267.