Bioinformatics, 39(1), 2023, btac739
https://doi.org/10.1093/bioinformatics/btac739
Advance Access Publication Date: 16 November 2022
Original Paper

Data and text mining

scAWMV: an adaptively weighted multi-view learning framework for the integrative analysis of parallel scRNA-seq and scATAC-seq data

Pengcheng Zeng [1], Yuanyuan Ma [2] and Zhixiang Lin[3,*]

[1]Institute of Mathematical Sciences, ShanghaiTech University, Shanghai 201210, China [2]School of Computer and Information Engineering, Anyang Normal University, Henan 455000, China and [3]Department of Statistics, The Chinese University of Hong Kong, Hong Kong SAR, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on June 7, 2022; revised on October 16, 2022; editorial decision on November 10, 2022; accepted on November 15, 2022

Abstract

Motivation: Technological advances have enabled us to profile single-cell multi-omics data from the same cells, providing us with an unprecedented opportunity to understand the cellular phenotype and links to its genotype. The available protocols and multi-omics datasets [including parallel single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq) data profiled from the same cell] are growing increasingly. However, such data are highly sparse and tend to have high level of noise, making data analysis challenging. The methods that integrate the multi-omics data can potentially improve the capacity of revealing the cellular heterogeneity.

Results: We propose an adaptively weighted multi-view learning (scAWMV) method for the integrative analysis of parallel scRNA-seq and scATAC-seq data profiled from the same cell. scAWMV considers both the difference in importance across different modalities in multi-omics data and the biological connection of the features in the scRNA-seq and scATAC-seq data. It generates biologically meaningful low-dimensional representations for the transcriptomic and epigenomic profiles via unsupervised learning. Application to four real datasets demonstrates that our framework scAWMV is an efficient method to dissect cellular heterogeneity for single-cell multi-omics data.

Availability and implementation: The software and datasets are available at https://github.com/pengchengzeng/scAWMV.

Contact: zhixianglin@cuhk.edu.hk

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

1 Introduction

The recent advances in single-cell technologies have enabled us to probe multiple biological layers. These technologies include single-cell RNA sequencing (scRNA-seq) that profiles transcription, single-cell ATAC sequencing (scATAC-seq) that profiles accessible chromatin regions (Macaulay *et al.*, 2017; Mezger *et al.*, 2018) and other methods. There are increasing demands for computationally efficient methods for processing and analyzing the datasets (Rotem *et al.*, 2015; Rozenblatt-Rosen *et al.*, 2017) brought by these technologies. For example, clustering methods that group similar cells into sub-populations are often used as the first step in the analysis of single-cell genomic data. The clustering methods for scRNA-seq data include SIMLR (Wang *et al.*, 2017), SC3 (Kiselev *et al.*, 2017), SAFE-clustering (Yang *et al.*, 2018), SOUP (Zhu *et al.*, 2019), SHARP (Wan *et al.*, 2020) and other methods. The clustering methods for scATAC-seq data include Cusanovich 2018 (Cusanovich *et al.*, 2018), scABC (Zamanighomi *et al.*, 2018), SCALE (Xiong *et al.*, 2019), cisTopic (Gonzalez-Blas *et al.*, 2019) and other methods. However, the aforementioned methods are all designed for analyzing one type of genomic feature.

Recently, the technological advances have enabled us to profile high-dimensional multi-omics data at an unprecedented resolution, and the available protocols and multi-omics datasets are growing at an increasing pace. For example, scM&T-seq (Angermueller *et al.*, 2016) links transcriptional and epigenetic heterogeneity, scNMT-seq (Clark *et al.*, 2018) enables joint profiling of chromatin accessibility, DNA methylation and transcription in single cells, and more protocols have been developed for joint profiling of chromatin and transcriptome, including sci-CAR-seq (Cao *et al.*, 2018), scCAT-seq (Liu *et al.*, 2019), Paired-seq (Zhu *et al.*, 2019), SNARE-seq (Chen *et al.*, 2019), SHARE-seq (Ma *et al.*, 2020) and Paired-Tag (Zhu *et al.*, 2021). The resulting single-cell multi-omics datasets can provide insights into the cell's phenotype and links to its genotype (Macaulay *et al.*, 2017). However, such data tend to have high level

of noise and are highly sparse (Colomé-Tatché and Theis, 2018). These characteristics bring challenges for analyzing the multi-omics single-cell data. In order to analyze the complex biological process varying across cells, we need to integrate different types of genomic features via flexible but rigorous computational methods. The methods of data integration are growing recently, and they can be divided into three categories: (i) Non-negative matrix factorization (NMF)-based methods. coupleNMF (Duren *et al.*, 2018) is based on extensions of NMF, and the connection between chromatin accessibility and gene expression builds upon prediction models trained from bulk data with diverse cell types. LIGER (Welch *et al.*, 2019) implements integrative NMF to infer a shared low-dimensional space in multiple single-cell datasets. scAI (Jin *et al.*, 2020) aggregates epigenomic data in cell subpopulations that exhibit similar gene expression and epigenomic profiles through iterative learning in an unsupervised manner. JSNMF (Ma *et al.*, 2022) is based on jointly semi-orthogonal NMF and it enables effective and accurate integrative analysis of single-cell multiomics data. (ii) Methods based on probabilistic generative models. scACE (Lin *et al.*, 2019) is a model-based approach to jointly cluster single-cell chromatin accessibility and single-cell gene expression data. It does not rely on training data to connect the two data types and allows for statistical inference of the cluster assignment. MOFA+ (Argelaguet *et al.*, 2020) is based on extensions of factor analysis model and was designed to deal with increasingly large-scale multi-omics datasets. scAMACE (Wangwu *et al.*, 2021) is a model-based approach to the joint analysis of single-cell data on chromatin accessibility, gene expression and methylation, and it develops an efficient expectation–maximization algorithm to perform statistical inference. (iii) Other machine learning-based methods. Seurat (version 3) (Stuart *et al.*, 2019) anchors diverse datasets together with the capability of integrating single-cell measurements not only across scRNA-seq technologies, but also across different data modalities, that is, data types within single cells, for example, scRNA-seq data, scATAC-seq data and DNA methylation data. coupleCoC (Zeng *et al.*, 2020) and coupleCoC+ (Zeng and Lin, 2021) are transfer learning methods based on the information-theoretic co-clustering framework for the integrative analysis of single-cell genomics data. Seurat (version 4) (Hao *et al.*, 2021) introduces the 'weighted nearest neighbor' analysis to learn the relative utility of each genomic feature in each cell, enabling an integrative analysis of multi-omics data.

Two important issues should be taken into account while integrating single-cell multi-omics data. The first issue is how much weight should be assigned to each data modality. The information carried by different types of genomic features are complementary and it is desirable to integrate them. However, the data from different types of genomic features can have different levels of noise. As an example, consider the setting where scRNA-seq and scATAC-seq are profiled on the same cells. scATAC-seq data tend to have higher level of noise and higher degree of sparsity, and it will be intuitive to give smaller weight to scATAC-seq data compared with scRNA-seq data, while integrating the two data modalities. Another issue is how to link data from multiple omics in a way that is biologically meaningful. In the above setting, a subset of features in scATAC-seq data are linked with scRNA-seq data, because promoter accessibility/gene activity score are directly linked with gene expression. Effectively connecting the linked features is expected to be helpful in the integrative analysis of multi-omics data.

In this work, we present scAWMV—an adaptively weighted multi-view learning framework to integrate scRNA-seq data and scATAC-seq data measured from the same cells. Utilizing a unified matrix factorization model, our framework not only automatically assigns a weight to each modality of multi-omics data, but also connects the linked features across data types by adding a constraint (Fig. 1A). Unsupervised learning of this framework will result in biologically meaningful low-rank matrices that represent the transcriptomic and epigenomic profiles. These matrices allow for inferring low-dimensional representations (Fig. 1B), the identification of factor-specific marker genes (Fig. 1C) and the identification of cell types (Fig. 1D). In comparison with the recent multi-omics data

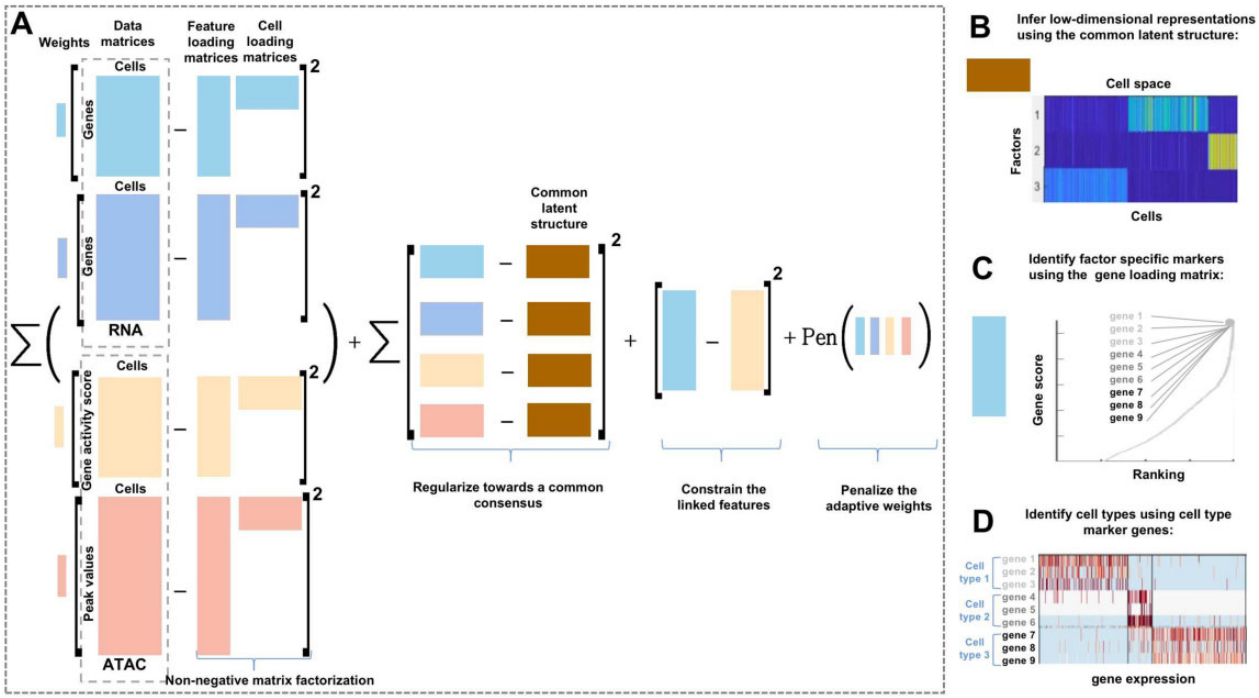

Fig. 1. Overview of scAWMV. (**A**) The sketch of the objective function in scAWMV, which is minimized via finding the optimal matrix factorization. It includes four components: (1) reconstruction errors by NMF for the data matrices from scRNA-seq and scATAC-seq, and each factorization is assigned an adaptive weight; (2) the regularization toward a common consensus for all the cell loading matrices; (3) the constraint on the gene loading matrices obtained from the NMF of the linked data, that is, gene expression data matrix in scRNA-seq and gene activity score matrix in scATAC-seq; (4) the penalty term for the adaptive weights. (**B**) Based on the common latent structure from (A), scAWMV uses Louvain clustering and groups the cells in the same clusters in the heatmap of the common latent structure. (**C**) scAWMV ranks genes based on the gene loading matrix for scRNA-seq data from (A). For example, genes 1–9 are labeled with the highest loadings. (**D**) scAWMV assigns cell type labels to cell clusters with known marker genes

integration methods in four published datasets, we demonstrate that our framework is efficient in revealing cellular heterogeneity.

2 The methodology

In this section, we first introduce the frameworks of single-view NMF (Wang and Zhang, 2013) and multi-view NMF (Liu *et al.*, 2013), and then extend them to our framework of adaptively weighted multi-view NMF for single-cell multi-omics data. We treat scATAC-seq data and scRNA-seq data that are measured from the same cells as two views of the single-cell genomic data. We assume that a subset of the features, that is, gene activity score in scATAC-seq data is linked with the gene expression in scRNA-seq data; and the other features, that is, accessibility of distal peaks in scATAC-seq data, are not directly linked with the genes in scRNA-seq data. We expect to improve the clustering performance of the cells by (i) integrating these views of data for uncovering the common shared structure; (ii) automatically assigning weights for adjusting for the importance of each view and (iii) adding a constraint to the linked features across data types for considering the biological dependency across different types of genomic features.

2.1 Single-view NMF

Let X be a p by n data matrix, representing the data on p features of n samples. A NMF $X = WH^T$ gives us a 'soft' clustering of the samples (Duren *et al.*, 2018): the ith column of basis matrix W gives the mean vectors of the ith cluster of samples and the jth row of coefficient matrix H gives the assignment weights of the jth sample to the clusters. The goal of NMF framework (Wang and Zhang, 2013) is to obtain the factorizations by solving the following optimization problem:

$$\underset{W\geq 0, H\geq 0}{\operatorname{argmin}} ||X - WH^T||_F^2, \tag{1}$$

where $||\cdot||_F$ is the Frobenius norm and $W \geq 0, H \geq 0$ stand for the constraints that all elements in the matrix are non-negative.

2.2 Multi-view NMF

Assume that the data have n views, and let $\{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$ denote the data of all the views. Here, for each view $X^{(v)}$, we have the single-view NMF: $X^{(v)} = W^{(v)}(H^{(v)})^T$. For different views, we have the same number of samples but allow for different number of features. Thus, $H^{(v)}$s are of the same shape but $W^{(v)}$s may differ along row dimension across multiple views. To integrate information from multiple views in clustering, Liu *et al.* (2013) developed multi-view NMF framework to cluster multiple views simultaneously to uncover the common latent structure shared by multiple views. A sample in different views would be assigned to the same cluster with high likelihood, thus the coefficient matrices $H^{(1)}, H^{(2)}, \ldots, H^{(n)}$ learnt from these views would be required to be softly regularized toward a common consensus H^*. The goal of multi-view NMF framework (Liu *et al.*, 2013) is to obtain the factorizations in multiple views by solving the following optimization problem:

$$\underset{\substack{W^{(v)}\geq 0, H^{(v)}\geq 0,\\ H^* \geq 0}}{\operatorname{argmin}} \sum_{v=1}^{n}(||X^{(v)} - W^{(v)}(H^{(v)})^T||_F^2 + \lambda^{(v)}||H^{(v)}Q^{(v)} - H^*||_F^2), \tag{2}$$

where $Q^{(v)} = \operatorname{Diag}(\sum_{i=1}^{p^{(v)}} W_{i,1}^{(v)}, \sum_{i=1}^{p^{(v)}} W_{i,2}^{(v)}, \ldots, \sum_{i=1}^{p^{(v)}} W_{i,K}^{(v)})$ is a diagonal matrix; K denotes the number of clusters of the samples; $W_{i,k}^{(v)}$ represents the element of the ith row and the kth column in the matrix $W^{(v)}$ and $p^{(v)}$ denotes the number of features for the vth view. Here, $Q^{(v)}$ is a normalization term for making $H^{(v)}$ in multiple views comparable at the same scale (Liu *et al.*, 2013). The hyperparameter $\lambda^{(v)}$ gives the relative weight between the standard NMF reconstruction error $||X^{(v)} - W^{(v)}(H^{(v)})^T||_F^2$ and the disagreement between the coefficient matrix and the consensus matrix $||H^{(v)}Q^{(v)} - H^*||_F^2$ for the vth view.

2.3 The proposed framework

We now extend the multi-view NMF (Liu *et al.*, 2013) to model single-cell multi-omics data. The gene activity score in scATAC-seq data is linked with gene expression in scRNA-seq data, and the accessibility of distal peaks in scATAC-seq data is not directly linked with the genes in scRNA-seq data. We use the index symbols $(v, 1)$ and $(v, 2)$ to represent the linked part and the unlinked part of data in the vth view, respectively. We write the vth view of data as $X^{(v)} = \begin{bmatrix} X^{(v,1)} \\ X^{(v,2)} \end{bmatrix}_{p^{(v)}\times n}$, where $p^{(v)} = p^{(v,1)} + p^{(v,2)}$, representing the sum of the number of features in the linked matrix $X^{(v,1)}$ and that in the unlinked matrix $X^{(v,2)}$. Here, we assume that $[X^{(1,1)}]_{p^{(1,1)}\times n}$ (i.e. gene expression) is directly linked with $[X^{(2,1)}]_{p^{(2,1)}\times n}$ (i.e. gene activity score), and $[X^{(1,2)}]_{p^{(1,2)}\times n}$ (i.e. gene expression) is not directly linked with $[X^{(2,2)}]_{p^{(2,2)}\times n}$ (i.e. accessibility of distal peaks). We consider these submatrices $X^{(1,1)}$, $X^{(1,2)}$, $X^{(2,1)}$ and $X^{(2,2)}$ as the four views of single-cell genomic datasets. We propose the following optimization problem (illustrated in Fig. 1A):

$$\begin{aligned}\underset{\substack{W^{(v,l)}\geq 0, H^{(v,l)}\geq 0,\\ H^* \geq 0, v,l=1,2}}{\operatorname{argmin}} & \sum_{v=1}^{2}\sum_{l=1}^{2}(\omega^{(v,l)}||X^{(v,l)} - W^{(v,l)}(H^{(v,l)})^T||_F^2 \\ & + \lambda^{(v,l)}||H^{(v,l)}Q^{(v,l)} - H^*||_F^2) + \beta||W^{(1,1)} - W^{(2,1)}||_F^2 \\ & + \gamma\sum_{v=1}^{2}\sum_{l=1}^{2}\omega^{(v,l)}\ln\omega^{(v,l)}, \\ \text{s.t.} & \sum_{v=1}^{2}\sum_{l=1}^{2}\omega^{(v,l)} = 1,\end{aligned} \tag{3}$$

where $Q^{(v,l)} = \operatorname{Diag}(\sum_{i=1}^{p^{(v,l)}} W_{i,1}^{(v,l)}, \sum_{i=1}^{p^{(v,l)}} W_{i,2}^{(v,l)}, \ldots, \sum_{i=1}^{p^{(v,l)}} W_{i,K}^{(v,l)})$, and $W_{i,k}^{(v,l)}$ represents the element of the ith row and the kth column in the matrix $W^{(v,l)}$. Note that $W^{(1,1)}$ and $W^{(2,1)}$ have the same number of rows, that is, $p^{(1,1)} = p^{(2,1)}$, under the assumption that $X^{(1,1)}$ is directly linked to $X^{(2,1)}$, and the number of columns of both $W^{(1,1)}$ and $W^{(2,1)}$ is equal to the number of factors K.

The naive integration of different views by multi-view NMF does not consider the difference in the relative importance across different views. Therefore, we assign each view $X^{(v,l)}$ with an adaptive weight $\omega^{(v,l)}$ in our framework (Equation 3). $\gamma\sum_{v=1}^{2}\sum_{l=1}^{2}\omega^{(v,l)}\ln\omega^{(v,l)}$ represents the penalty term for these weights. γ is the hyperparameter which controls the distribution of the adaptive weight: when γ is smaller, the view that has less reconstruction error NMF (which suggests that the factors carry more effective information) will be given higher weight. More details for the intuition of γ and the weights are provided in Section 2.4. The features in the linked data $X^{(1,1)}$ and $X^{(2,1)}$ are connected, because gene expression is positively associated with gene activity. We further encourage the difference between the basis matrices $W^{(1,1)}$ and $W^{(2,1)}$ to be small. It accounts for the dependence between gene expression in single-cell transcriptomic data and gene activity score in single-cell chromatin accessibility data. The hyperparameter β controls the strength of the constraint $||W^{(1,1)} - W^{(2,1)}||_F^2$. The rules for choosing the hyperparameters $\lambda^{(v,l)}(v, l = 1, 2)$, β, γ and the number of factors K will be discussed in Section 2.5. We call this adaptively weighted multi-view learning framework scAWMV for short.

2.4 Optimization algorithm

We optimize the objective function in Equation (3) by an iterative update procedure:

$$W_{i,k}^{(v,1)} \leftarrow W_{i,k}^{(v,1)}\frac{\omega^{(v,1)}X^{(v,1)}H^{(v,1)} + \lambda^{(v,1)}\sum_{j=1}^{n}H_{j,k}^{(v,1)}H_{j,k}^* + \beta W_{i,k}^{(v',1)}}{\omega^{(v,1)}(W^{(v,1)}(H^{(v,1)})^TH^{(v,1)})_{i,k} + \lambda^{(v,1)}\sum_{m=1}^{p^{(v,1)}}W_{m,k}^{(v,1)}\sum_{j=1}^{n}(H_{j,k}^{(v,1)})^2 + \beta W_{i,k}^{(v,1)}}, \quad v' = 3 - v, v = 1, 2;$$

$$W_{i,k}^{(v,2)} \leftarrow W_{i,k}^{(v,2)}\frac{\omega^{(v,2)}X^{(v,2)}H^{(v,2)} + \lambda^{(v,2)}\sum_{j=1}^{n}H_{j,k}^{(v,2)}H_{j,k}^*}{\omega^{(v,2)}(W^{(v,2)}(H^{(v,2)})^TH^{(v,2)})_{i,k} + \lambda^{(v,2)}\sum_{m=1}^{p^{(v,2)}}W_{m,k}^{(v,2)}\sum_{j=1}^{n}(H_{j,k}^{(v,2)})^2}, \quad v = 1, 2;$$

$$H_{j,k}^{(v,l)} \leftarrow H_{j,k}^{(v,l)}\frac{\omega^{(v,l)}((X^{(v,l)})^TW^{(v,l)})_{j,k} + \lambda^{(v,l)}H_{j,k}^*}{\omega^{(v,l)}(H^{(v,l)}(W^{(v,l)})^TW^{(v,l)})_{j,k} + \lambda^{(v,l)}H_{j,k}^{(v,l)}}, \quad v, l = 1, 2;$$

$$\omega^{(v,l)} = \frac{\exp\left\{-\frac{||X^{(v,l)} - W^{(v,l)}(H^{(v,l)})^T||_F^2}{\gamma}\right\}}{\sum_{v=1}^{2}\sum_{l=1}^{2}\exp\left\{-\frac{||X^{(v,l)} - W^{(v,l)}(H^{(v,l)})^T||_F^2}{\gamma}\right\}}, \quad v, l = 1, 2;$$

$$H^* = \frac{\sum_{v=1}^{2}\sum_{l=1}^{2}\lambda^{(v,l)}H^{(v,l)}Q^{(v,l)}}{\sum_{v=1}^{2}\sum_{l=1}^{2}\lambda^{(v,l)}};$$

where $(\cdot)_{i,k}$ represents the element of the ith row and the kth column in the matrix $(\cdot)$. Both $\omega^{(v,l)}$ and H^* have exact solutions in

each iteration. The intuition for $\omega^{(v,l)}$ and γ becomes clear by looking at the update for $\omega^{(v,l)}$: first, the views that have smaller reconstruction error in NMF (i.e. $||X^{(v,l)} - W^{(v,l)}(H^{(v,l)})^T||_F^2$) will be given higher weight; second, the hyperparameter γ acts as a bandwidth parameter. We stop the iterations when the difference of the loss function (Equation 3) is less than 10^{-4} between two adjacent iterations. The details for the derivation of these updating rules are given in the Supplementary Materials.

2.5 Selection of hyperparameters

We firstly obtain the initial solutions $\tilde{W}^{(v,l)}$ and $\tilde{H}^{(v,l)}(v,l=1,2)$ by solving the optimization problems:

$$\underset{W^{(v,l)},H^{(v,l)}}{\operatorname{argmin}} ||X^{(v,1)} - W^{(v,1)}(H^{(v,1)})^T||_F^2, \quad v,l=1,2,$$

using an approach similar to that in Liu *et al.* (2013), and use them as the initialization for our optimization problem (Equation 3). We initialize the adaptive weights $\tilde{\omega}^{(v,l)}$ as $\frac{1}{4}$ for $v,l=1,2$. We fix the value of the hyperparameter $\lambda^{(v,l)}$ as 0.01 for $v,l=1,2$, as suggested in multi-view NMF by Liu *et al.* (2013). We choose the hyperparameter

$$\beta = \frac{\sum_{v=1}^{2}\sum_{l=1}^{2}\tilde{\omega}^{(v,l)}||X^{(v,l)} - \tilde{W}^{(v,l)}(\tilde{H}^{(v,l)})^T||_F^2}{||\tilde{W}^{(1,1)} - \tilde{W}^{(2,1)}||_F^2}, \tag{4}$$

and choose the hyperparameter

$$\gamma = \max\left\{|\frac{||X^{(v,l)} - \tilde{W}^{(v,l)}(\tilde{H}^{(v,l)})^T||_F^2}{\ln\tilde{\omega}^{(v,l)}}|; v,l=1,2\right\}. \tag{5}$$

The number of factors K can be determined using an approach similar to Brunet *et al.* (2004). In this work, we set $K=20$ in all the real datasets. We will discuss in great detail the sensitivity of the choices of the hyperparameters in Section 3.4.

2.6 Evaluation of the clustering results

We first compute the cluster of cells by Louvain method (Blondel *et al.*, 2008) once we obtain the consensus matrix H^*, and then evaluate the clustering performance by three criteria: normalized mutual information (NMI), adjusted Rand index (ARI) (Christopher *et al.*, 2008) and the Residual Average Gini Index (RAGI) score (Chen *et al.*, 2019).

NMI and ARI require the known ground-truth labels of cells, and we use the cell type labels of gene expression data provided in the real datasets as the ground-truth labels. Assume that G is the known cell type labels and P is the predicted clustering assignments, we then calculate NMI as

$$\frac{I(G;P)}{\sqrt{E(G)\cdot E(P)}},$$

where $I(G;P)$ represents the mutual information over G and P, and $E(\cdot)$ represents the entropy. Assume that n is the total number of single cells, $n_{Q,i}$ is the number of cells assigned to the ith cluster in Q, $n_{G,j}$ is the number of cells belonging to the jth cell type in G and $n_{i,j}$ is the number of overlapping cells between the ith cluster in Q and the jth cell type in G. As a corrected-for-chance version of the Rand index, ARI is calculated as

$$\frac{\sum_{ij}\binom{n_{i,j}}{2} - \left[\sum_i\binom{n_{Q,i}}{2}\sum_j\binom{n_{G,j}}{2}/\binom{n}{2}\right]}{\frac{1}{2}\left[\sum_i\binom{n_{Q,i}}{2} + \sum_j\binom{n_{G,j}}{2}\right] - \left[\sum_i\binom{n_{Q,i}}{2}\sum_j\binom{n_{G,j}}{2}\right]/\binom{n}{2}}.$$

Higher values of NMI and ARI indicate better clustering performance.

RAGI score calculates the difference between the variability of marker gene expression across cell clusters and the variability of housekeeping gene expression across cell clusters. We note that no cell type labels are required when computing RAGI score on scRNA-seq data. First, we manually find the marker genes using the CellMarker database developed by Zhang *et al.* (2019) and find the housekeeping genes using the HRT Atlas v1.0 database developed by Hounkpe *et al.* (2021). Second, we compute Gini index (Gini, 1997) for each marker gene and each housekeeping gene. The Gini index measures how imbalanced the expression of a gene is across cell clusters. Third, based on the sets of Gini index values for marker genes and housekeeping genes, we calculate the difference between the mean Gini index for marker genes and the mean Gini index for housekeeping genes, that is, the RAGI score. Higher RAGI score represents better separation of the cell clusters.

2.7 Data preprocessing and feature selection

In order to get the linked data, we first compute the gene activity score using the gene scoring approach by Chen *et al.* (2019): the distance-weighted sum of reads (peak values in scATAC-seq data) within or near the region gives the accessibility at each transcription start site. Second, we extract the set of genes that have both gene expression and gene activity score, which can be considered as the linked features. Third, we use the scRNA-seq data to choose 2000 most highly variable genes from this set by R toolkit Seurat (Butler *et al.*, 2018; Stuart *et al.*, 2019). We then obtain the linked data $[X^{(1,1)}]_{2,000\times n}$ and $[X^{(2,1)}]_{2,000\times n}$. We also use the R toolkit Seurat to choose 2000 most highly variable genes from the set of genes that are not included in the linked data, and we have the unlinked scRNA-seq data $[X^{(1,2)}]_{2,000\times n}$. We choose 5000 unlinked features in scATAC-seq data, corresponding to the top 5000 largest summation of peak values over all cells, and we have the unlinked scATAC-seq data $[X^{(2,2)}]_{5,000\times n}$. We take log transformation for scRNA-seq data to alleviate the effect of extreme values in the data matrices. Before implementing our algorithm, we normalized each of these data matrices, including $[X^{(1,1)}]_{2,000\times n}$, $[X^{(2,1)}]_{2,000\times n}$, $[X^{(1,2)}]_{2,000\times n}$ and $[X^{(2,2)}]_{5,000\times n}$, to make different views of data comparable by dividing the sum of all the entries within that data matrix.

3 Results

In this section, we evaluated our methodology in four real datasets of single-cell multiome ATAC and gene expression, including peripheral blood mononuclear cell (PBMC) from healthy donors—granulocytes removed through cell sorting (one dataset has 2711 cells and another one has 11 898 cells), frozen healthy human brain tissue (3233 cells) and fresh frozen lymph node with B-cell lymphoma (14 566 sorted nuclei). All of these datasets are processed by Cell Ranger ARC 2.0.0 and are available at https://www.10xgenomics.com/. We compared our method scAWMV with scAI (Jin *et al.*, 2020), MOFA+ (Argelaguet *et al.*, 2020), Seurat V4 (Hao *et al.*, 2021) and multi-NMF (Liu *et al.*, 2013). To check whether the constraint on the linked data can improve the clustering performance, we compared a simplified version of scAWMV, where the objective function has no constraint on the linked data (i.e. $\beta||W^{(1,1)} - W^{(2,1)}||_F^2$) compared with that in scAWMV, and we denote it as scAWMV-no-link. To check whether the adaptive weights on different views can improve the clustering performance, we compared another simplified version of scAWMV, where the objective function has no adaptive weights compared with that in scAWMV, and we denote it as scAWMV-no-weights. To check whether the linked part in the datasets is helpful in clustering the cells, we also considered the case where the gene activity score of scATAC-seq data (i.e. $X^{(2,1)}$) is removed from the input, and we denote it as scAWMV-no-$X^{(2,1)}$. Except for Seurat V4 (Hao *et al.*, 2021) which can produce clusters of cells by itself, we implemented Louvain clustering on the low-dimensional representation of cells after dimension reduction by the other baseline methods: the cell loading matrix H given by scAI (Jin *et al.*, 2020), the latent factor Z given by MOFA+ (Argelaguet *et al.*, 2020) and the consensus V^* given by multi-NMF (Liu *et al.*, 2013). (Here, the notations H, Z and V^* are consistent with that in their original publications.) We used the NMI, ARI and RAGI to evaluate the clustering results (Table 1). When computing the cluster memberships of cells in four real

datasets by all methods except Seurat V4, we set the numbers of clusters the same as that given by the secondary analysis outputs from https://www.10xgenomics.com/.

3.1 Example 1: application to PBMCs from healthy donors

In the first example, we studied the paired single-cell multiome ATAC and gene expression of cryopreserved human PBMCs from two healthy female donors (Examples 1A and 1B). The numbers of cells in Examples 1A and 1B are 2711 and 11 898, respectively. We set the number of clusters in Example 1A as 9 and set the number of clusters in Example 1B as 17 when implementing Louvain method for clustering. These numbers of clusters are given by the analysis of gene expression data on the 10× Genomics website. Examples 1A and 1B in Table 1 show the results of clustering the cells. Our method performs the best in NMI, ARI and RAGI among the eight methods in Example 1A, and performs the best in NMI and RAGI among the eight methods in Example 1B. In both Examples 1A and 1B, the value of RAGI given by Seurat V4 ranks the joint first. On the one hand, both scAWMV-no-link and scAWMV-no-weight have slightly worse clustering results compared with scAWMV. This suggests that the removal of the constraint on the linked data or adaptive weights in the objective function of scAWMV has negative impacts on the clustering performance. On the other hand, when we excluded the data matrix $X^{(2,1)}$ from the input datasets of our method, the clustering performance worsens. It suggests that incorporating the gene activity score in scATAC-seq data and the connection between gene activity score and gene expression in scRNA-seq data are helpful in improving the result of clustering the cells.

We compared the heatmaps of the common latent structure $(H^*)^T$ given by scAWMV and the latent structure H given by scAI in Figure 2A and B for Example 1A, and in Figure 3A and B for Example 1B, respectively. They show that the consensus $(H^*)^T$ given by scAWMV can detect more clear patterns than that given by scAI. In Example s1A and 1B, we assigned the cell type labels for all the cell clusters based on marker genes (Supplementary Figs S1 and S2). We then assigned cell types to the factors in the common factor loading matrix $(H^*)^T$ by the following rule: we selected the top 200 entries in each row of $(H^*)^T$, calculated the proportions of the cell-type labels corresponding to these selected entries and chose the cell type with the largest proportion as the cell type for the factors. Factors 1 and 2 in Example 1A correspond to natural killer cells (proportion = 96%) and B cells (proportion = 97%), respectively. Factors 1 and 2 in Example 1B correspond to B cells (proportion = 91%) and natural killer cells (proportion = 89%), respectively. The marker genes for natural killer cells and B cells (Zhang *et al.*, 2019) also tend to have high scores in the first two factors in $W^{(1,1)}$ and $W^{(2,1)}$ (Fig. 2C and D for Example 1A and Fig. 3C and D for Example 1B), which is in agreement with the cell-type annotation for $(H^*)^T$, and demonstrates that the entries in the feature loading matrices provide biological insight on the factors.

We also studied the gene ontology (GO) enrichment analysis for the feature loading matrices $W^{(1,1)}$ and $W^{(2,1)}$, which correspond to gene expression data and gene activity score in chromatin accessibility data, respectively. We selected the top 200 linked genes with large values in each column of $W^{(1,1)}$ and $W^{(2,1)}$, and utilized Metascape (Zhou *et al.*, 2019) to implement GO enrichment analysis. The enriched biological processes and pathways for gene expression data and gene activity score tend to be consistent and they agree with the biological function of the underlying cell types that the factors represent (Supplementary Tables S1 and S2). Factor 1 in Example 1A corresponds to natural killer cells, which is a type of cytotoxic lymphocyte critical to the innate immune system that belong to the rapidly expanding family of innate lymphoid cells and represent 5–20% of all circulating lymphocytes in humans (Arachchige and Shavinda, 2021). As demonstrated in Supplementary Table S1, the first columns in $W^{(1,1)}$ and $W^{(2,1)}$ are both enriched for 'cell activation' ($\log(q\text{-value}) = -14.57$ for $W^{(1,1)}$ and $\log(q\text{-value}) = -10.75$ for $W^{(2,1)}$), 'regulation of cell activation' ($\log(q\text{-value}) = -10.84$ for $W^{(1,1)}$ and $\log(q\text{-value}) = -11.31$ for $W^{(2,1)}$) and 'inflammatory response' ($\log(q\text{-value}) = -6.24$ for $W^{(1,1)}$ and $\log(q\text{-value}) = -8.29$ for $W^{(2,1)}$). Factor 2 in Example 1A corresponds to B cells, which functions in the humoral

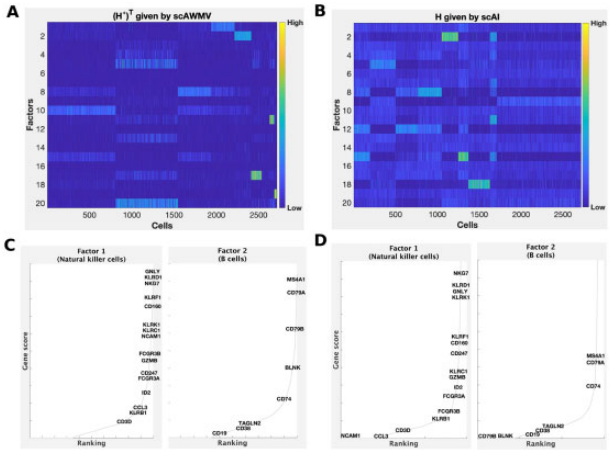

Fig. 2. Simultaneously analyzing the paired single-cell multiome ATAC and gene expression for PBMCs in Example 1A. (**A**) Heatmap of the common latent structure $(H^*)^T$ given by scAWMV. We grouped the cells in the same clusters based on Louvain clustering. (**B**) Heatmap of the latent structure H given by scAI. We grouped the cells in the same clusters based on Louvain clustering. (**C**) Ranking of known cell-type-specific marker genes in factors 1 and 2 from $W^{(1,1)}$ in scAWMV. (**D**) Ranking of known cell-type-specific marker genes in factors 1 and 2 from $W^{(2,1)}$ in scAWMV. Note that the marker genes for B cells and natural killer cells in PBMCs are collected from CellMarker database (Zhang *et al.*, 2019)

Table 1. The results of clustering the cells in four real datasets for three examples, evaluated by NMI, ARI and RAGI score

Clustering methods	Example 1A ($n = 2711$)			Example 1B ($n = 11\,898$)			Example 2 ($n = 3233$)			Example 3 ($n = 14\,566$)		
	NMI	ARI	RAGI	NMI	ARI	RAGI	NMI	ARI	RAGI	NMI	ARI	RAGI
scAI	0.65	0.47	0.42	0.62	0.37	0.39	0.64	0.44	0.29	0.47	0.18	0.25
MOFA+	0.60	0.43	0.45	0.58	0.39	0.39	0.54	0.33	0.28	0.42	0.22	0.25
Seurat V4	0.60	0.39	**0.48**	0.60	0.43	**0.40**	0.62	0.40	**0.38**	0.48	0.31	**0.28**
multi-NMF	0.68	0.48	0.44	0.61	0.43	0.36	0.64	0.44	0.32	0.42	0.24	0.24
scAWMV-no-link	0.68	0.53	0.47	0.60	0.34	0.36	0.63	0.43	0.31	0.52	0.35	0.26
scAWMV-no-weight	0.69	0.46	0.43	0.59	0.37	0.39	0.64	0.40	0.31	0.43	0.31	0.25
scAWMV-no-$X^{(2,1)}$	0.67	0.50	0.46	0.59	**0.45**	0.33	0.66	**0.47**	0.31	0.52	0.30	0.22
scAWMV	**0.71**	**0.54**	**0.48**	**0.64**	0.38	**0.40**	**0.69**	0.45	0.31	**0.54**	**0.35**	**0.28**

Notes: NMI and ARI were computed by the cell type labels provided by the analysis of gene expression data on the 10× Genomics website. RAGI was computed using marker genes and housekeeping genes. The bold numbers represent the best clustering results.

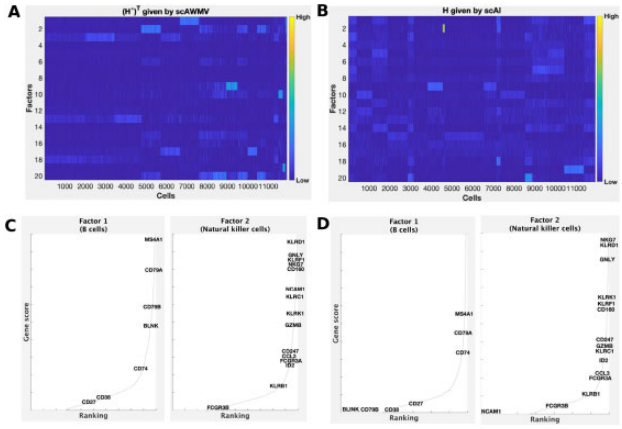

Fig. 3. Simultaneously analyzing the paired single-cell multiome ATAC and gene expression for PBMCs in Example 1B. (**A**) Heatmap of the common latent structure $(H^*)^T$ given by scAWMV. We grouped the cells in the same clusters based on Louvain clustering. (**B**) Heatmap of the latent structure H given by scAI. We grouped the cells in the same clusters based on Louvain clustering. (**C**) Ranking of known cell-type-specific marker genes in factors 1 and 2 from $W^{(1,1)}$ in scAWMV. (**D**) Ranking of known cell-type-specific marker genes in factors 1 and 2 from $W^{(2,1)}$ in scAWMV. Note that the marker genes for B cells and natural killer cells in PBMCs are collected from CellMarker database (Zhang *et al.*, 2019)

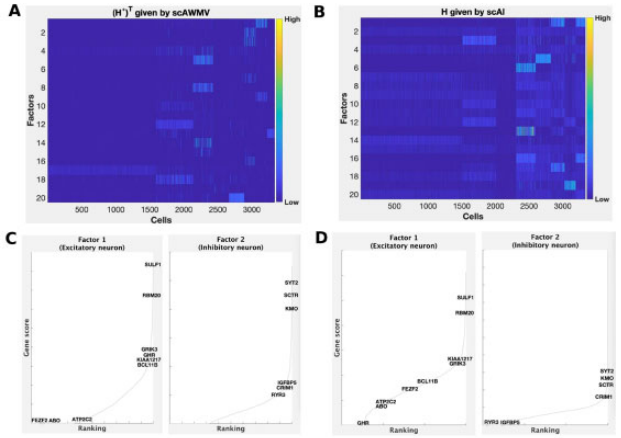

Fig. 4. Simultaneously analyzing the paired single-cell multiome ATAC and gene expression for frozen healthy human brain tissue in Example 2. (**A**) Heatmap of the common latent structure $(H^*)^T$ given by scAWMV. We grouped the cells in the same clusters based on Louvain clustering. (**B**) Heatmap of the latent structure H given by scAI. We grouped the cells in the same clusters based on Louvain clustering. (**C**) Ranking of known cell-type-specific marker genes in factors 1 and 2 from $W^{(1,1)}$ in scAWMV. (**D**) Ranking of known cell-type-specific marker genes in factors 1 and 2 from $W^{(2,1)}$ in scAWMV. Note that the marker genes for excitatory neuron and inhibitory neuron in human brain are collected from CellMarker database (Zhang *et al.*, 2019)

immunity component of the adaptive immune system (Murphy, 2012). The second columns in $W^{(1,1)}$ and $W^{(2,1)}$ are both enriched for 'regulation of immune effector process' ($\log(q\text{-value}) = -5.64$ for $W^{(1,1)}$ and $\log(q\text{-value}) = -9.07$ for $W^{(2,1)}$) and 'positive regulation of cytokine production' ($\log(q\text{-value}) = -4.73$ for $W^{(1,1)}$ and $\log(q\text{-value}) = -5.40$ for $W^{(2,1)}$). We observed the same trend in the enrichment analysis for the feature loading matrices $W^{(1,1)}$ and $W^{(2,1)}$ in Example 1B (Supplementary Table S2). These results suggest that the feature loading matrices $W^{(1,1)}$ and $W^{(2,1)}$ boost GO enrichment analysis and provide rich information on the biological interpretation of the factors. Also, the consistent trends in $W^{(1,1)}$ and $W^{(2,1)}$ indicate the effect of the constraint $||W^{(1,1)} - W^{(2,1)}||_F^2$ in scAWMV (Equation 3).

3.2 Example 2: application to frozen healthy human brain tissue (3k)

In the second example, we explored the paired single-cell multiome ATAC and gene expression of flash frozen healthy human brain tissue (cerebellum). The number of cells is 3233. We set the number of clusters as 8, given by the analysis of gene expression data on the 10× Genomics website. The clustering results in Example 2 from Table 1 show that scAWMV performs the best in NMI, scAWMV-no-$X^{(2,1)}$ performs the best in ARI and Seurat V4 performs the best in RAGI among the eight methods. The clustering performance by scAWMV is slightly better than that by scAWMV-no-link and scAWMV-no-weight. In Figure 4A and B, we compared the heatmaps of the common latent structure $(H^*)^T$ given by scAWMV and the latent structure H given by scAI, respectively. The patterns identified by $(H^*)^T$ in scAWMV are comparable and slightly clearer than that identified by H in scAI. The feature loading matrices $W^{(1,1)}$ and $W^{(2,1)}$ given by scAWMV have consistent trends and tend to be enriched for marker genes of the corresponding cell types (Fig. 4C and D). In Figure 4C and D, the enrichment for the modality of chromatin accessibility, $W^{(2,1)}$, tends to be weaker compared with the modality of gene expression, $W^{(1,1)}$, which is likely due to the higher level of noise in chromatin accessibility data.

3.3 Example 3: application to fresh frozen lymph node with B-cell lymphoma (14k)

In the last example, we studied the paired single-cell multiome ATAC and gene expression of flash frozen intra-abdominal lymph node tumor from a patient diagnosed with diffuse small lymphocytic lymphoma. The number of cells is 14 566. We set the number of clusters as 18, given by the analysis of gene expression data on the 10× Genomics website. Example 3 in Table 1 shows the results of clustering the cells. Our method scAWMV performs the best in NMI, ARI and RAGI among the eight methods, and the value of RAGI given by Seurat V4 ranks the joint first. Compared with scAWMV, both scAWMV-no-link and scAWMV-no-weight have slightly worse clustering results. The result by scAWMV-no-$X^{(2,1)}$ shows that the clustering performance of scAWMV becomes worse when we excluded the data matrix $X^{(2,1)}$ from the input datasets. We compared the heatmaps of the common latent structure $(H^*)^T$ given by scAWMV and H given by scAI in Figure 5A and B. They show that it is hard to detect very clear patterns from the heatmaps of $(H^*)^T$ and H, likely due to the high level of noise in this dataset. In Figure 5C and D, we chose factors 1 and 2 from coefficient matrices $W^{(1,1)}$ and $W^{(2,1)}$ given by scAWMV and demonstrate that known cell type markers tend to have higher rankings in the feature loading matrices. In Figure 5C and D, the enrichment for the modality of chromatin accessibility, $W^{(2,1)}$, tends to be weaker compared with the modality of gene expression, $W^{(1,1)}$, which is likely due to the higher level of noise in chromatin accessibility data.

3.4 Discussion

First, we note that the results from ARI, NMI and RAGI are not be all consistent with each other (Table 1), since the three evaluation criteria depend on different theories and use different inputs: (i) NMI, ARI and RAGI are based on Shannon information, pair-counting and Gini index, respectively, and (ii) both NMI and ARI need the cell type labels as input while RAGI does not. Instead, RAGI needs the list of marker genes and housekeeping genes as input.

Second, in order to study the baseline clustering performance when only one data type, that is, scRNA-seq or scATAC-seq data, is used in all these examples, we compared NMF-only-RNA and NMF-only-ATAC (Supplementary Table S3), which are NMF methods using only scRNA-seq or scATAC-seq data as input. We also summarized the adaptive weights obtained by scAWMV for all views (Supplementary Table S4). The results in Supplementary Table S3 show that the clustering performance of both NMF-only-RNA and NMF-only-ATAC is worse than scAWMV in all the above examples. It suggests that the integration of two data types is needed and adaptive weights need to be adjusted.

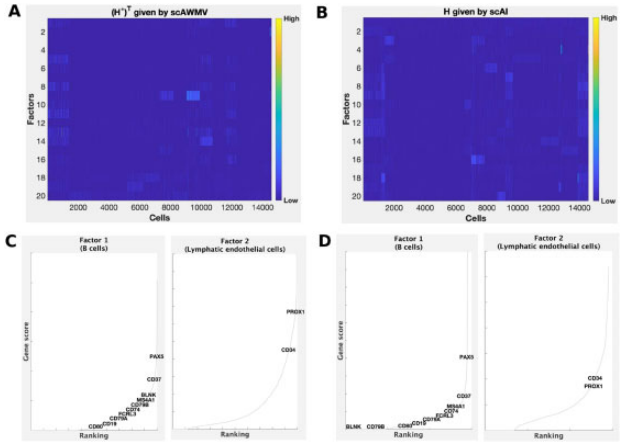

Fig. 5. Simultaneously analyzing the paired single-cell multiome ATAC and gene expression for fresh frozen lymph node with B-cell lymphoma in Example 3. (**A**) Heatmap of the common latent structure $(H^*)^T$ given by scAWMV. We grouped the cells in the same clusters based on Louvain clustering. (**B**) Heatmap of the latent structure H given by scAI. We grouped the cells in the same clusters based on Louvain clustering. (**C**) Ranking of known cell-type-specific marker genes in factors 1 and 2 from $W^{(1,1)}$ in scAWMV. (**D**) Ranking of known cell-type-specific marker genes in factors 1 and 2 from $W^{(2,1)}$ in scAWMV. Note that the marker genes for B cells and lymphatic endothelial cells in lymph node are collected from CellMarker database (Zhang *et al.*, 2019)

Third, we studied the sensitivity of the selection of hyperparameters by scAWMV (Supplementary Figs S3–S7) for the four real datasets in three examples. Supplementary Figure S3 presents the evaluation curves (NMI, ARI and RAGI) as functions of the number of factors K, ranging from 10 to 30, for the four real datasets in three examples. It shows that scAWMV performs more stable for Examples 1B and 3, and less stable for Examples 1A and 2. Supplementary Figure S4 presents the evaluation curves with respect to different initializations for the adaptive weights $\boldsymbol{v} = (\omega^{(1,1)}, \omega^{(1,2)}, \omega^{(2,1)}, \omega^{(2,2)})$, where these initialization values are listed in the figure caption. It shows that the clustering performance by scAWMV is stable for all examples except Example 2, and in all examples, scAWMV has relatively better performance when the adaptive weights are initialized as $(1/4, 1/4, 1/4, 1/4)$. The value $\lambda = 0.01$ was suggested in multi-view NMF by Liu *et al.* (2013). When we tried different values of λ, that is, $0.00001, 0.0001, 0.001, 0.01, 0.1, 1$, scAWMV performs the best when $\lambda = 0.01$, as shown in Supplementary Figure S5. Supplementary Figures S6 and S7 demonstrate the evaluations with respect to the tuning parameters β and γ, respectively, where β_0 and γ_0 are the default values calculated from formulas (4) and (5), respectively. They show that scAWMV performs the best when using the default setting $\beta = \beta_0$ and $\gamma = \gamma_0$. For Examples 1A, 1B and 3, the performance of scAWMV is more stable when β and γ vary; for Example 2, it is less stable. In summary, except for Example 2, the other three examples are stable to the initialization for the adaptive weights, and the values of β and γ.

4 Convergence and computational time

The algorithm scAWMV is guaranteed to converge as the objective function (Equation 3) is non-increasing in each iteration. It tends to converge in 30 iterations for the four real datasets. We summarized the running time (Supplementary Table S5) by five multi-view clustering methods, including multi-NMF, scAWMV, scAI, MOFA+ and Seurat V4 in each real dataset. The computational costs for Example 1B (∼10K cells) are 36.72 min (multi-NMF), 35.22 min (scAWMV), 600.70 min (scAI), 145.44 min (MOFA+) and 3.96 min (Seurat V4). The graph-based method Seurat V4 is the fastest and our scAWMV is faster or comparable to other methods based on matrix factorization (including multi-NMF, MOFA+ and scAI).

5 Conclusion

In this work, we proposed the framework scAWMV for the integrative analysis of parallel scRNA-seq data and scATAC-seq data from the same cells. scAWMV differs from other multi-view learning methods in two aspects: (i) It automatically assigns different weights for different views of data while the other methods tend to treat different views of data equally. (ii) It utilized the linked information between the parallel transcriptomic and epigenomic layers while the other methods tend to ignore this connection. The gene activity score in scATAC-seq data and the gene expression in scRNA-seq data are biologically linked, which may provide useful information in dissecting cellular heterogeneity. Application to four real-world datasets demonstrates that scAWMV is an efficient method to dissect cellular heterogeneity for single-cell multi-omics data.

Data and software availability

The raw datasets are available at 10× Genomics website https://www.10xgenomics.com/. The software is available at https://github.com/pengchengzeng/scAWMV.

Acknowledgements

We are grateful to the anonymous reviewers for helpful suggestions.

Funding

This work was supported by the HPC Platform of ShanghaiTech University, the Chinese University of Hong Kong startup grant [4930181], the Chinese University of Hong Kong's Project Impact Enhancement Fund (PIEF) and Science Faculty's Collaborative Research Impact Matching Scheme (CRIMS) and Hong Kong Research Grant Council [ECS 24301419, GRF 14301120].

Conflict of Interest: The authors declare that they have no conflict of interest.

References

Angermueller,C. *et al.* (2016) Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*, **13**, 229–232.
Arachchige,P.M. and Shavinda,A. (2021) Human NK cells: from development to effector functions. *Innate Immun.*, **27**, 212–229.
Argelaguet,R. *et al.* (2020) Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.*, **21**, 111.
Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech.*, **2008**, P10008.
Brunet,J.-P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA*, **101**, 4164–4169.
Butler,A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
Cao,J. *et al.* (2018) Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, **361**, 1380–1385.
Chen,S. *et al.* (2019) High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.*, **37**, 1452–1457.
Christopher,D.M. *et al.* (2008). *Introduction to Information Retrieval*. Cambridge University Press.
Clark,S.J. *et al.* (2018) scNMT-seq enables joint profiling of chromatin accessibility, DNA methylation and transcription in single cells. *Nat. Commun.*, **9**, 781.
Colomé-Tatché,M. and Theis,F.J. (2018) Statistical single cell multi-omics integration. *Curr. Opin. Syst. Biol.*, **7**, 54–59.
Cusanovich,D.A. *et al.* (2018) A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, **174**, 1309–1324.e18.
Duren,Z. *et al.* (2018) Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorization. *Proc. Natl. Acad. Sci. USA*, **115**, 7723–7728.
Gini,C. (1997) Concentration and dependency ratios. *Riv. Polit. Econ.*, **87**, 769–792.
Gonzalez-Blas,C.B. *et al.* (2019) Cistopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods*, **16**, 397–400.
Hao,Y. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.e29.

Hounkpe,B.W. *et al.* (2021) HRT atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res.*, **49**, D947–D955.

Jin,S. *et al.* (2020) scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol.*, **21**, 25.

Kiselev,V.Y. *et al.* (2017) Sc3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, **14**, 483–486.

Lin,Z.X. *et al.* (2019) Model-based approach to the joint analysis of single-cell data on chromatin accessibility and gene expression. *Stat. Sci.*, **35**, 2–13.

Liu,J. *et al.* (2013) Multi-view clustering via joint nonnegative matrix factorization. In: *2013 SIAM International Conference on Data Mining*, Society for Industrial and Applied mathematics, Philadelphia, PA, p. 252–260.

Liu,L. *et al.* (2019) Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat. Commnun.*, **10**, 470.

Ma,S. *et al.* (2020) Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*, **183**, 1103–1116.e20.

Ma,Y. *et al.* (2022) JSNMF enables effective and accurate integrative analysis of single-cell multiomics data. *Brief Bioinformatics*, **23**.

Macaulay,I.C. *et al.* (2017) Single-cell multiomics: multiple measurements from single cells. *Trends Genet.*, **33**, 155–168.

Mezger,A. *et al.* (2018) High-throughout chromatin accessibility profiling at single-cell resolution. *Nat. Commun.*, **9**, 34–67.

Murphy,K. (2012) *Janeway's Immunobiology*. 8th edn. Garland Science, New York.

Rotem,A. *et al.* (2015) Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, **33**, 1165–1172.

Rozenblatt-Rosen,O. *et al.* (2017) The human cell atlas: from vision to reality. *Nat. News*, **550**, 451–453.

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

Wan,S. *et al.* (2020) SHARP: hyperfast and accurate processing of single-cell RNA-seq data via ensemble random projection. *Genome Res.*, **30**, 205–213.

Wang,B. *et al.* (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods*, **14**, 414–416.

Wang,Y.X. and Zhang,Y.J. (2013) Nonnegative matrix factorization: a comprehensive review. *IEEE Trans. Knowl. Data Eng.*, **25**, 1336–1353.

Wangwu,J. *et al.* (2021) scAMACE: model-based approach to the joint analysis of single-cell data on chromatin accessibility, gene expression and methylation. *Bioinformatics*, **37**, 3874–3880.

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

Xiong,L. *et al.* (2019) SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.*, **10**, 4576.

Yang,Y. *et al.* (2018) Safe-clustering: single-cell aggregated (from ensemble) clustering for single-cell RNA-seq data. *Bioinformatics*, **35**, 1269–1277.

Zamanighomi,M. *et al.* (2018) Unsupervised clustering and epigenetic classification of single cells. *Nat. Commun.*, **9**, 2410.

Zeng,P. and Lin,Z. (2021) Couplecoc+: an information-theoretic co-clustering-based transfer learning framework for the integrative analysis of single-cell genomic data. *PLoS Comput. Biol.*, **17**, e1009064.

Zeng,P. *et al.* (2020) Coupled co-clustering-based unsupervised transfer learning for the integrative analysis of single-cell genomics data. *Brief Bioinformatics*, **22**.

Zhang,X. *et al.* (2019) Cellmarkers: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.*, **47**, 721–728.

Zhou,Y. *et al.* (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.*, **10**, 1–10.

Zhu,C. *et al.* (2019) An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.*, **26**, 1063–1070.

Zhu,C. *et al.* (2021) Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat. Methods*, **18**, 283–292.