Integrative Analyses of Single-Cell Multi-Omics Data: A Review from a Statistical Perspective

Zhixiang Lin

Abstract Advances in technology during the past few years have enabled us to profile various types of genomic features at single-cell resolution. Different types of genomic features capture different aspects of the cells, and together they more accurately depict the biology of the cells. This emerging research area will significantly advance our understanding of complex biological systems and human diseases. Here we review computational methods for analysis and integration of single-cell data across different molecular modalities, and we will emphasize on the statistical aspects of these methods.

The advances in technology in the past few years have enabled us to profile various types of genome-wide molecular features at single-cell resolution, including DNA, gene expression, protein-binding, histone modifications, and chromatin accessibility. Previously, such genomic approaches could only be applied to bulk tissue samples comprised of an ensemble of many cells, providing average genomic measures, but masking the cellular difference [1]. Different types of genomic features capture complementary information and together they provide a more complete biological picture.

The high technical variation and high level of noise present in single-cell datasets, especially in single-cell epigenomic datasets, pose challenges for the extraction of biological variation. The large-scale of single-cell datasets also necessitates efficient algorithms to analyze the datasets. In this review, we focus on computational methods developed for single-cell multi-omics data. Depending on the data structure, the computational methods fall into two broad categories: methods that integrate data from multiple molecular modalities profiled in different cells but similar biological tissue, and methods that integrate data from multiple molecular

Z. Lin (✉)
Department of Statistics, the Chinese University of Hong Kong, Hong Kong, Hong Kong
e-mail: zhixianglin@cuhk.edu.hk

H. H.-S. Lu et al. (eds.), *Handbook of Statistical Bioinformatics*,
Springer Handbooks of Computational Statistics,
https://doi.org/10.1007/978-3-662-65902-1_3

modalities profiled simultaneously in the same cells. In this review, we use "scRNA-Seq data" to represent single-cell gene expression data, and "scATAC-Seq data" to represent single-cell chromatin accessibility data, acknowledging the fact that there are multiple platforms with different names that can profile gene expression and chromatin accessibility at single-cell resolution.

Methods that integrates multiple scRNA-Seq datasets have also been developed [2–6], and they will not be discussed in detail in this review. Methods that integrate multi-omics data obtained from bulk tissues have also been developed. These methods have been summarized and reviewed in [7].

1 Multi-Omics Data Profiled on Different Cells

Cells are sacrificed in single-cell experiments, and it is experimentally more challenging to obtain multiple types of genomic data from the same cell, compared with the relative ease of obtaining such genomic data from the same sample in bulk genomic experiments. Computational methods are developed for the setting where multiple types of genomic data are obtained from different subsets of cells from similar cell population (i.e., tissue).

Based on the goal, the methods can be classified in the following categories: (a) methods that learn low-dimensional embeddings where different types of genomic features are aligned to the same latent space. After cells from different modalities are aligned, cell type identification can be achieved by a separate clustering step using the low-dimension representation of the cells. The methods that fall into this category include coupled NMF [8], DC3 [9], Seurat V3 [10], LIGER [11], online iNMF [12], UINMF [13], and MAESTRO [14]; (b) Methods that directly perform joint clustering on the original data space, where the shared and unshared cell types across data modalities are identified through joint clustering. The methods that fall into this category include scACE [15] and scAMACE [16]; (c) Transfer learning-based methods where one dataset (typically scRNA-Seq data) facilitates the analysis of another noisier dataset (typically single-cell epigenomic data). The methods that fall into this category include coupleCoC [17], coupleCoC+ [18], and scJoint [19].

Because different types of genomic features are profiled in different cells, these methods require that at least a subset of features are connected across the multi-omics data: To connect scATAC-Seq data with scRNA-Seq data, gene activity score[20] that summarizes the peak accessibility near the gene body was used in online iNMF [12], UINMF[21], MAESTRO [14], scAMACE[16], coupleCoC [17], coupleCoC+ [18], and scJoint [19], promoter accessibility was used in scACE [15], prediction model trained from reference data was used in coupled NMF [8], and external chromatin conformation data that links regulatory regions to genes was used in DC3 [9]; To connect single-cell methylation data with scRNA-Seq data, gene body mCH methylation was used in LIGER [11], online iNMF [12], scAMACE [16], coupleCoC [17], and coupleCoC+ [18], promoter methylation was also used in scAMACE[16].

coupled NMF [8] was designed for integrative analysis of scRNA-Seq and scATAC-Seq data obtained from different set of cells. Let $\mathbf{O}$ be a p_1 by n_1 data matrix for scATAC-Seq data, where p_1 is the number of regions and n_1 is the number of cells. Let $\mathbf{E}$ be a p_2 by n_2 data matrix for scRNA-Seq data, where p_2 is the number of genes and n_2 is the number of cells. The following optimization problem was proposed in coupled NMF:

$$\underset{\mathbf{W}_1,\mathbf{H}_1,\mathbf{W}_2,\mathbf{H}_2\geq 0}{\arg\min} \frac{1}{2}\|\mathbf{O}-\mathbf{W}_1\mathbf{H}_1\|_F^2+\frac{\lambda_1}{2}\|\mathbf{E}-\mathbf{W}_2\mathbf{H}_2\|_F^2-\lambda_2 tr(\mathbf{W}_2^T\mathbf{A}\mathbf{W}_1)+\mu(\|\mathbf{W}_1\|_F^2+\|\mathbf{W}_2\|_F^2), \tag{1}$$

where $\mathbf{W}_1$ is the $p_1 \times K$ region-factor matrix for scATAC-Seq data, $\mathbf{W}_2$ is the $p_2 \times K$ gene-factor matrix for scRNA-Seq data, $\mathbf{H}_1$ and $\mathbf{H}_2$ are matrices of dimensions $K \times n_1$ and $K \times n_2$, representing the low-dimensional embeddings for the cells in scATAC-Seq and scRNA-Seq data, respectively. coupled NMF is based on non-negative matrix factorization [22], and the entries in $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{H}_1$, and $\mathbf{H}_2$ are non-negative. In coupled NMF, the regions in scATAC-Seq data and the genes in scRNA-Seq data are connected through the term $tr(\mathbf{W}_2^T\mathbf{A}\mathbf{W}_1)$, where $\mathbf{A}$ is a known $p_1 \times p_2$ matrix obtained from training non-negative least squares regression models on bulk gene expression and chromatin accessibility datasets. The matrix $\mathbf{A}$ is set to the regression coefficients, where bulk gene expression data was used as the outcome, and bulk chromatin accessibility data was used as the predictor. The term $\|\mathbf{W}_1\|_F^2 + \|\mathbf{W}_2\|_F^2$ penalizes the scales of $\mathbf{W}_1$ and $\mathbf{W}_2$. λ_1, λ_2, and μ are tuning parameters.

DC3 [9] is a follow-up work based on the general framework of coupled NMF. Instead of a pre-trained regression model, DC3 connects the regions in scATAC-Seq data and genes in scRNA-Seq data through bulk HiChIP data obtained from similar tissues as that in scRNA-Seq and scATAC-Seq data. Simultaneous to clustering, DC3 also performs deconvolution of bulk HiChIP data to different cell subpopulations. The following is the objective function proposed in DC3:

$$\begin{aligned}&\min_{\mathbf{W}_1,\mathbf{H}_1,\mathbf{W}_2,\mathbf{H}_2,\alpha,\mathbf{\Lambda}\geq 0} \frac{\mu_1}{2}\|\mathbf{O}-\mathbf{W}_1\mathbf{H}_1\|_F^2+\frac{\mu_2}{2}\|\mathbf{E}-\mathbf{W}_2\mathbf{H}_2\|_F^2+\frac{1}{2}\|\mathbf{C}-\alpha\mathbf{D}\odot(\mathbf{W}_2\Lambda\mathbf{W}_1^T)\|_F^2\\ &\text{subject to } \sum_{k=1}^{K} h_{1,kj}=1 \text{ for } j=1,2,\cdots,n_1;\ \sum_{k=1}^{K} h_{2,kj}=1, \text{ for } j=1,2,\cdots,n_2;\ \sum_{k=1}^{K}\lambda_k=1,\end{aligned} \tag{2}$$

where $\mathbf{O}$, $\mathbf{E}$, $\mathbf{W}_1$, $\mathbf{H}_1$, $\mathbf{W}_2$, $\mathbf{H}_2$ are the same as the corresponding matrices in coupled NMF; $h_{1,kj}$ and $h_{2,kj}$ are the kjth entry in $\mathbf{H}_1$ and $\mathbf{H}_2$, respectively; $\mathbf{C}$ is a $p_1 \times p_2$ bulk HiChIP data matrix, representing the enhancer-promoter interaction strength. The bulk HiChIP data is obtained from similar tissues as that in scRNA-Seq and scATAC-Seq data, and it is considered as a mixture of different cell subpopulations. The term $\|\mathbf{C}-\alpha\mathbf{D}\odot(\mathbf{W}_2\Lambda\mathbf{W}_1^T)\|_F^2$ is the key for deconvolution of bulk HiChIP to cell subpopulation-specific enhancer-promoter interaction, and for linking genes (promoters) in scRNA-Seq data and regions (enhancers) in scATAC-Seq data. α is

a scaling factor and $\mathbf{D}$ is a masking matrix that extracts the entries in $\mathbf{C}$ that are larger than 1: $d_{ij} = 1$ if $c_{ij} \geq 1$ and $d_{ij} = 0$ if $c_{ij} < 1$. $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_K)$ is a diagonal matrix. $\mathbf{W}_2 \Lambda \mathbf{W}_1^T = \sum_{k=1}^K \lambda_k \mathbf{w}_{2,\cdot k} \mathbf{w}_{1,\cdot k}^T$, where $\mathbf{w}_{1,\cdot k}$ and $\mathbf{w}_{2,\cdot k}$ denote the kth columns in $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively. $\mathbf{w}_{2,\cdot k} \mathbf{w}_{1,\cdot k}^T$ can be interpreted as the enhancer-promoter interaction strength in the kth cell subpopulation (represented by the kth factor in NMF), which provides deconvolution for $\mathbf{C}$, and λ_k can be interpreted as the proportion of cells that belong to the kth cell subpopulation.

Seurat V3 [10] integrates multiple single-cell datasets. Examples were demonstrated which integrate multiple scRNA-Seq datasets, scRNA-Seq with scATAC-Seq data, and *in situ* gene expression and scRNA-Seq datasets. The features are assumed to be the same across datasets: gene activity score was used for scATAC-Seq data. Seurat V3 implements the following four steps to integrate two datasets, $\mathbf{Y}$ and $\mathbf{X}$. The correction procedure in Seurat V3 can also be extended to multiple datasets.

Step 1: Data preprocessing and feature selection with highly variable genes.

Step 2: Dimension reduction and identify "anchor" correspondences between datasets. $\mathbf{X}$ is a $p \times n_X$ single-cell dataset, and $\mathbf{Y}$ is a $p \times n_Y$ single-cell dataset, where p is the number of features (i.e., genes), n_X and n_Y are the number of cells. Seurat V3 performs canonical correlation analysis (CCA) for dimension reduction of $\mathbf{X}$ and $\mathbf{Y}$. The first pair of canonical vectors $\mathbf{u} \in \mathbb{R}^{n_X}$ and $\mathbf{v} \in \mathbb{R}^{n_Y}$ are obtained by solving the following problem:

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T \mathbf{X}^T \mathbf{Y} \mathbf{v}, \\ &\text{subject to } \|\mathbf{u}\|_2^2 \leq 1, \|\mathbf{v}\|_2^2 \leq 1. \end{aligned} \tag{3}$$

Note that the implementation of CCA in 3 is different from its usual implementation in statistics, where the projection vectors are implemented in the feature spaces. Seurat V3 obtains the first k pairs of canonical vectors, and then normalizes the canonical vectors so the ℓ_2-norm of the vector for each cell equals to 1. The normalized canonical vectors are used as the low-dimensional representation of the cells. Mutual nearest neighbors (MNN; pairs of cells, with one from each dataset, that are contained within each other's neighborhoods) are obtained from the low-dimensional representations. These pairwise correspondences are referred as "anchors."

Step 3: Filtering, scoring, and weighting of anchor correspondences. The initial anchor pairs obtained in step 2 are filtered, so they are also supported by the original high-dimensional space. The anchors are then scored based on their strength using an approach that is similar to the shared nearest neighbor graphs. Suppose the matrix $\mathbf{X}$ is used to correct the matrix $\mathbf{Y}$. $\mathbf{W}$ is a weight matrix for the cells in $\mathbf{Y}$, and it has dimension $n_Y \times$ number of anchor cells. w_{ij} represents the weighted similarity between cell i in $\mathbf{Y}$ and anchor cell j in $\mathbf{Y}$, which not only considers the distance between cells i and j but also considers the anchor score of cell j: if cell j has higher anchor score, w_{ij} will tend to be larger.

Step 4: Data matrix correction. Let $\mathbf{a}_X$ and $\mathbf{a}_Y$ denote the sets of anchor cell pairs in $\mathbf{X}$ and $\mathbf{Y}$. Seurat V3 first computes the differences between the pairs of anchor cells in the two data matrices:

$$\mathbf{B} = \mathbf{Y}[, \mathbf{a}_Y] - \mathbf{X}[, \mathbf{a}_X]. \tag{4}$$

The corrected data matrix $\hat{\mathbf{Y}}$ is obtained as the following:

$$\hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{B}\mathbf{W}^T. \tag{5}$$

LIGER [11] employs integrative non-negative matrix factorization (iNMF) [23]. LIGER uses gene as the feature to connect different datasets. It uses one minus non-CpG (mCH) gene body methylation for single-cell methylation data, because non-CpG gene body methylation is generally negatively correlated with gene expression in neurons. Though the integrative analysis of scRNA-Seq and scATAC-Seq was not presented in [11], scATAC-Seq data can be incorporated in principle using gene activity score. The objective function in LIGER is as the following:

$$\underset{\mathbf{W},\mathbf{H}_i,\mathbf{V}_i \geq 0}{\arg\min} \sum_i \|\mathbf{E}_i - (\mathbf{W} + \mathbf{V}_i)\mathbf{H}_i\|_F^2 + \lambda \sum_i \|\mathbf{V}_i\mathbf{H}_i\|_F^2, \tag{6}$$

where $\mathbf{E}_i$ denotes dataset i, which is of dimension $n_i \times m$, where n_i denotes the number of cells in dataset i, and m denotes the number of genes. $\mathbf{W}$ is of dimension $K \times m$, and it is the shared factor loadings across datasets. $\mathbf{V}_i$ is of dimension $K \times m$, and it is the factor loading that is unique to dataset i. $\mathbf{H}_i$ is of dimension $n_i \times K$, and it denotes the low-dimensional embedding for the cells in dataset i. In the objective function 6, $\mathbf{E}_i$ is approximated by $\mathbf{W}\mathbf{H}_i + \mathbf{V}_i\mathbf{H}_i$, where $\mathbf{W}\mathbf{H}_i$ represents the shared variation across datasets, and $\mathbf{V}_i\mathbf{H}_i$ denotes the dataset-specific effect. The regularization term $\lambda \sum_i \|\mathbf{V}_i\mathbf{H}_i\|_F^2$ controls the strength of the dataset-specific variation, and λ is a tuning parameter. After obtaining the low-dimensional embedding $\mathbf{H}$ for the cells across datasets, LIGER further builds a shared factor neighborhood graph in which cells are connected based on their similarity in the low-dimensional embeddings, and joint clusters are identified by performing community detection on this graph.

Other than integrative analysis of single-cell multi-omics data, examples that integrate multiple scRNA-Seq datasets from different individuals, time points, species, and spatial gene expression data were also presented in LIGER. Methods have been developed based on extensions of LIGER, including online iNMF [12] and UINMF [21].

Online iNMF [12] has the same objective function as formula 6 in LIGER. The major advantage of online iNMF is its computational efficiency and fixed memory usage for large datasets. It enables integration of large, multi-modal datasets by cycling through the data multiple times in small mini-batches and integration

of continually arriving datasets, where the entire dataset is not available at any point during training. Online iNMF builds upon the online non-negative matrix factorization approach in [24].

UINMF [21]. One limitation of LIGER is that the features that are not linked across datasets are not utilized. For example, peaks in the intergenic regions in scATAC-Seq data are not directly linked to the genes in scRNA-Seq data, so the peaks were not included in the objective function of LIGER. To address this limitation, UINMF was developed to include these unlinked features. The objective function of UINMF is as the following:

$$\underset{\mathbf{W},\mathbf{H}_i,\mathbf{U}_i,\mathbf{V}_i\geq 0}{\arg\min} \sum_i \left\{ \|\mathbf{E}_i - (\mathbf{W}+\mathbf{V}_i)\mathbf{H}_i\|_F^2 + \|\mathbf{P}_i - \mathbf{U}_i\mathbf{H}_i)\|_F^2 \right\} + \lambda_i \sum_i \left\{ \|\mathbf{V}_i\mathbf{H}_i\|_F^2 + \|\mathbf{U}_i\mathbf{H}_i\|_F^2 \right\}. \tag{7}$$

In formula 7, the terms $\mathbf{E}_i$, $\mathbf{H}_i$, $\mathbf{V}_i$, and $\mathbf{W}$ are the same as those in formula 6. $\mathbf{P}_i$ is a matrix of dimension $n_i \times z_i$, where n_i is the number of cells and z_i is the number of unlinked features in the ith dataset. The matrices for the linked features $\mathbf{E}_i$ and the unlinked features $\mathbf{P}_i$ share the same $\mathbf{H}_i$, which is the low-dimensional embedding for the cells. Note that the tuning parameter λ_i is different across datasets, and the variation of the unlinked features $\mathbf{U}_i\mathbf{H}_i$ is included in the penalization term.

MAESTRO [14] provides a comprehensive open-source computational workflow for the integrative analyses of scRNA-Seq and scATAC-Seq data from multiple platforms. MAESTRO provides functions for preprocessing, alignment, quality control, expression and chromatin accessibility quantification, clustering, differential analysis, and annotation. Most other methods in this review start from the processed datasets, while MAESTRO supports input from fastq files for a wide variety of single-cell sequencing-based platforms. To integrate the cells from scRNA-Seq and scATAC-Seq, MAESTRO first calculates the regulatory potential for each gene in each cell, which measures the scATAC-Seq reads near the gene weighted by an exponential decay of the read distance to the transcriptional start site of the gene. Note that regulatory potential is computed similarly as the gene activity score. MAESTRO then performs a canonical correlation analysis between gene expression from scRNA-Seq and regulatory potential from scATAC-Seq. A pair of cells, one from scRNA-Seq and the other from scATAC-Seq, can be anchored using mutual nearest neighbors after dimension reduction. Then, MAESTRO transfers the cell type labels from scRNA-Seq (cell type labels in scRNA-Seq data are obtained from clustering by Seurat) to scATAC-Seq using the anchored cell pairs.

After integrating scRNA-Seq and scATAC-Seq cells, MAESTRO combines the transcriptional regulators predicted from scRNA-Seq data using LISA [25] and scATAC-Seq data using GIGGLE [26], and uses the rank product to combine the two. The final candidate regulators are further filtered based on the regulator expression from scRNA-Seq.

scACE [15] and **scAMACE** [16] are clustering methods built upon Bayesian hierarchical models. scACE integrates scRNA-Seq and scATAC-Seq data profiled on different set of single cells. scAMACE builds upon scACE and extends it to

model scRNA-Seq, scATAC-Seq, and sc-methylation data. The goal in scACE and scAMACE is to cluster similar cell types within and across different molecular modalities. The followings are details for the model in scAMACE. The model for scRNA-Seq data:

$$
\begin{aligned}
&\boldsymbol{\omega}_{\cdot g}^{rna} \xrightarrow{z_{l\cdot}} u_{lg} \longrightarrow v_{lg} \longrightarrow y_{lg} \\
&\boldsymbol{z}_l \sim \text{Categorical}(\boldsymbol{\psi}^{rna}), \\
&u_{lg} \mid z_{lk} = 1 \sim \text{Bernoulli}(\omega_{kg}^{rna}), \\
&v_{lg} \mid u_{lg} = 1 \sim \text{Bernoulli}(\pi_{l1});\ v_{lg} \mid u_{lg} = 0 \sim \text{Bernoulli}(\pi_{l0}), \\
&p(y_{lg} \mid v_{lg}) = v_{lg} g_1(y_{lg}) + (1 - v_{lg}) g_0(y_{lg}).
\end{aligned}
$$

Assume that there are K cell clusters in total, the random variable z_{lk} denotes whether cell l belongs to cluster $k \in \{1, \ldots, K\}$, and $\boldsymbol{z}_{l\cdot}$ follows categorical distribution with probability ψ_k^{rna} for cluster k. ω_{kg}^{rna} denotes the probability that gene g is active in cluster k. u_{lg} is a binary latent variable representing whether gene g is active in cell l and $u_{lg} = 1$ represents that it is active. v_{lg} denotes whether gene g is expressed in cell l and $v_{lg} = 1$ represents that it is expressed. When gene g is active in cell l ($u_{lg} = 1$), the probability that gene g is expressed in cell l ($v_{lg} = 1$) is π_{l1}, while the probability that gene g is expressed is π_{l0} if the gene is not active ($u_{lg} = 0$). Since genes are more likely to be expressed when they are active, it was assumed that $\pi_{l1} \geq \pi_{l0}$. y_{lg} denotes the observed gene expression for gene g in cell l (after normalization to account for sequencing depth and gene length), and it was assumed that $y_{lg} \mid v_{lg}$ follows a mixture distribution, where $g_1(.)$ and $g_0(.)$ are density functions of the expression level conditional on v_{lg}.

The model for scATAC-Seq data:

$$
\begin{aligned}
&\boldsymbol{\omega}_{\cdot g}^{acc} \xrightarrow{z_{i\cdot}} u_{ig} \longrightarrow o_{ig} \longrightarrow x_{ig} \\
&\boldsymbol{z}_{i\cdot} \sim \text{Categorical}(\boldsymbol{\psi}^{acc}), \\
&u_{ig} \mid z_{ik} = 1 \sim \text{Bernoulli}(\omega_{kg}^{acc}), \\
&o_{ig} \mid u_{ig} = 1 \sim \text{Bernoulli}(\pi_{i1});\ o_{ig} \mid u_{ig} = 0 \sim \text{Bernoulli}(\pi_{i0}), \\
&p(x_{ig} \mid o_{ig}) = o_{ig} f_1(x_{ig}) + (1 - o_{ig}) f_0(x_{ig}).
\end{aligned}
$$

The random variables ω_{kg}^{acc}, z_{ik}, ψ_k^{acc}, and u_{ig} have similar interpretations to their corresponding variables in the model for scRNA-Seq data. The cells in the scATAC-Seq data are different from the cells in the scRNA-Seq data, as indicated by the different notation i which represent the cells. x_{ig} denotes the observed gene activity score for gene g in cell i. It was modeled by a mixture distribution with density functions $f_1(.)$, $f_0(.)$, and binary latent variable o_{ig}. $o_{ig} = 1$, and 0 represent the mixture components with high (f_1) and low (f_0) gene scores, respectively. Accessibility tends to be positively associated with activity of the gene. This positive

relationship was modeled by the distribution $o_{ig} \mid u_{ig}$. When gene g is active in cell i ($u_{ig} = 1$), the probability that it has high gene score ($o_{ig} = 1$) is π_{i1}; When gene g is inactive in cell i ($u_{ig} = 0$), the probability that it has high gene score ($o_{ig} = 1$) is π_{i0}. π_{i1} was assumed to be larger than π_{i0} to represent the positive relationship.

The model for sc-methylation data:

$$\boldsymbol{\omega}_{\cdot g}^{met} \xrightarrow{z_{d\cdot}} u_{dg} \longrightarrow m_{dg} \longrightarrow t_{dg}$$

$$z_{d\cdot} \sim \text{Categorical}(\boldsymbol{\psi}^{met}),$$

$$u_{dg} \mid z_{dk} = 1 \sim \text{Bernoulli}(\omega_{kg}^{met}),$$

$$m_{dg} \mid u_{dg} = 1 \sim \text{Bernoulli}(\pi_{d1}); m_{dg} \mid u_{dg} = 0 \sim \text{Bernoulli}(\pi_{d0}),$$

$$p(t_{dg} \mid m_{dg}) = m_{dg} h_1(t_{dg}) + (1 - m_{dg}) h_0(t_{dg}).$$

The random variables ω_{kg}^{met}, z_{dk}, ψ_k^{met} and u_{dg} have similar interpretations to their corresponding variables in the model for scRNA-Seq data. The cells in the sc-methylation data are different from the cells in the scRNA-Seq data, as indicated by the different notation d which represents the cells. The binary random variable m_{dg} denotes whether gene g is methylated in cell d, and $m_{dg} = 1$ represents that it is methylated. Methylation of a gene (promoter methylation/gene body methylation) tends to be negatively associated with activity of the gene, and this negative relationship is modeled with $m_{dg} \mid u_{dg}$: when the gene g is active in cell d ($u_{dg} = 1$), it is less likely to be methylated ($m_{dg} = 1$), as $\pi_{d1} \leq \pi_{d0}$. t_{dg} denotes the observed methylation level for gene g in cell d, and $t_{dg} \mid m_{dg}$ was assumed to follow a mixture distribution, where $h_1(.)$ and $h_0(.)$ are density functions conditional on m_{dg}.

The model that connects the three molecular modalities: For scATAC-Seq data, ω_{kg}^{acc} was assumed to follow beta distribution with mean μ_{kg}^{acc} and precision ϕ^{acc}. The variable μ_{kg}^{acc} is connected with ω_{kg}^{rna} in scRNA-Seq data through the logit function: $logit(\mu_{kg}^{acc}) = \eta + \gamma \omega_{kg}^{rna} + \tau(\omega_{kg}^{rna})^2$. For sc-methylation data, the mean of ω_{kg}^{met}, μ_{kg}^{met}, is connected with ω_{kg}^{rna} through the logit function: $logit(\mu_{kg}^{met}) = \delta + \theta \omega_{kg}^{rna}$. Methylation and chromatin accessibility regulate gene expression biologically. The model was specified in the reverse order, so gene expression plays a central role. This is because scRNA-Seq data is usually less noisy compared with single-cell epigenomic data, the model specified this way will improve the clustering performance of single-cell epigenomic data, without sacrificing much the clustering performance of scRNA-Seq data.

scJoint [19] is a transfer learning method that integrates atlas-scale, heterogeneous collections of scRNA-Seq and scATAC-Seq data. It is a semi-supervised approach where the cell type labels for scRNA-Seq data are assumed to be known. The goal of scJoint is to transfer knowledge from massive scRNA-Seq data to scATAC-Seq through joint embedding in a low-dimensional space, and it also transfers the cell type labels from scRNA-Seq to scATAC-Seq data. scJoint uses gene activity score for scATAC-Seq data.

The neural network in scJoint consists of one input layer and two fully connected layers. Linear activation functions were used. Let $\{\mathbf{x}_i^{(s)}\}_{i=1}^{N_s}$ be the expression profiles for the cells in batch s in scRNA-Seq data, and $\mathbf{y}_i^{(s)} \in \{1, \cdots, K\}$ is the cell type label for cell i. Let $\{\mathbf{x}_i^{(t)}\}_{i=1}^{N_t}$ denote the gene activity scores for the cells in batch t in scATAC-Seq data. $f_{\theta,i}^{(s)} = f(\mathbf{x}_i^{(s)}; \theta)$ and $f_{\theta,i}^{(t)} = f(\mathbf{x}_i^{(t)}; \theta) \in \mathbb{R}^D$, $D = 64$, are the outputs of the joint embedding layer for scRNA-Seq and scATAC-Seq data, where θ denotes the parameters in the neural network and it is shared in the two datasets. Note that although the same notation i is used to represent cells in scRNA-Seq and scATAC-Seq data, the two types of data are obtained on different sets of cells. $h(f(\mathbf{x}_i^{(s)}; \theta))$ and $h(f(\mathbf{x}_i^{(t)}; \theta))$ are the outputs from the prediction layer for scRNA-Seq and scATAC-Seq data, respectively. $g_{\theta,i}^{(s)} = \text{softmax}(h(f(\mathbf{x}_i^{(s)}; \theta)))$ and $g_{\theta,i}^{(t)} = \text{softmax}(h(f(\mathbf{x}_i^{(t)}; \theta)))$ are vectors of length K, representing the probabilities of the assignment of cells to the K cell types.

There are three steps in scJoint. The first step is to train the neural network with the following loss function:

$$\mathcal{L}_1(\mathcal{B}_0, \theta) = \sum_{s=1}^{S} \left(\mathcal{L}_{\text{NNDR}}(\mathcal{B}^{(s)}, \theta) + \mathcal{L}_{\text{entropy}}(\mathcal{B}^{(s)}, \theta)\right) + \sum_{t=1}^{T} \left(\mathcal{L}_{\text{NNDR}}(\mathcal{B}^{(t)}, \theta) + \mathcal{L}_{\text{COS}}(\mathcal{B}^{(t)}, \mathcal{B}_R, \theta)\right), \tag{8}$$

where $\mathcal{B}^{(s)}$ denotes the data for batch s in scRNA-Seq data, $\mathcal{B}^{(t)}$ denotes the data for batch t in scATAC-Seq data, and $\mathcal{B}_0 = \{\mathcal{B}^{(s)}\}_{s=1}^{S} \cup \{\mathcal{B}^{(t)}\}_{t=1}^{T}$. In a spirit similar to PCA, the NNDR loss $\mathcal{L}_{\text{NNDR}}(\cdot)$ aims to capture low-dimensional orthogonal features in the joint embedding layer represented by the function $f(\cdot)$. The cosine similarity loss $\mathcal{L}_{\text{COS}}(\cdot)$ aims to align a subset of scRNA-Seq and scATAC-Seq cells in the joint embedding space. The cross entropy loss $\mathcal{L}_{\text{entropy}}(\cdot)$ represents the supervised component, where it penalizes the disagreement between the predicted cell type probabilities given by the function $g(\cdot)$ and the known cell types labels in scRNA-Seq data. The second step in scJoint transfers cell type labels from scRNA-Seq data to scATAC-Seq through k-nearest neighbor in the joint embedding space. The third step in scJoint refines the joint embedding space and improves mixing of cells from the same cell type in scRNA-Seq and scATAC-Seq data. The neural network is trained with the following loss function:

$$\mathcal{L}_{\text{scJoint}}(\mathcal{B}_0, \theta) = \mathcal{L}_1(\mathcal{B}_0, \theta) + \mathcal{L}_{\text{entropy}}(\mathcal{B}^{(t)}, \theta) + \mathcal{L}_{\text{center}}(\mathcal{B}_0, \theta), \tag{9}$$

where $\mathcal{L}_1(\mathcal{B}_0, \theta)$ is the same as the loss in step 1; $\mathcal{L}_{\text{entropy}}(\mathcal{B}^{(t)}, \theta)$ is the cross entropy loss using the transferred cell type labels for scATAC-Seq data, which are obtained in step 2; The term $\mathcal{L}_{\text{center}}(\mathcal{B}_0, \theta)$ encourages cells with the same cell type label to form clusters in the joint embedding space (determined by the function $f(\cdot)$), and it is similar to the loss function in k-means clustering, which encourages cells with the same cell type label to be close to the center of the cell type.

coupleCoC [17] and **coupleCoC+** [18] are based on the information-theoretic co-clustering [42] transfer learning framework, where the features and observations are clustered simultaneously and the co-clustering result achieves minimal loss in mutual information.

The goal of coupleCoC+ is to utilize one dataset, the source data (S), to facilitate the analysis of another dataset, the target data. Depending on whether the features are linked with the source data, the target data can be partitioned into two parts, data T that contains the linked features, and data U that contains the unlinked features. As an example, consider the setting where scRNA-Seq and scATAC-Seq are profiled on similar cell subpopulations but different cells. It is desirable to utilize the information in scRNA-Seq data to help cluster scATAC-Seq data, which is typically sparser and noisier. So scRNA-Seq data can be used as the source data S, and scATAC-Seq data can be used as the target data. In scATAC-Seq data, the data matrix of gene activity score are directly linked with gene expression in scRNA-Seq data, so it can be regarded as data T; the data matrix of peak accessibility can be regarded as data U, because the peaks that are distal to the genes are not directly linked with gene expression.

In coupleCoC+, both the genomic features and the cells are clustered. C_Y, C_X, C_Z, C_U denote the clustering functions for the cells in target data, the cells in source data, the linked features in the two datasets, and the unlined features that are unique in the target data. The following objective function was proposed in coupleCoC+:

$$\begin{aligned} \underset{\substack{C_Y, C_X, C_Z, C_U \\ h_{\mathrm{T}}, N_{\mathrm{sub}}, h_{\mathrm{S}, N_{\mathrm{sub}}}}}{\operatorname{argmin}} \quad & \ell_{\mathrm{T}}(C_Y, C_Z) + \lambda \ell_{\mathrm{S}}(C_X, C_Z) + \beta \ell_{\mathrm{U}}(C_Y, C_U) \\ & + \gamma D_{\mathrm{KL}}(\hat{p}_{\mathrm{T}}(\tilde{Y}_{h_{\mathrm{T}}, N_{\mathrm{sub}}}, \tilde{Z}_{\mathrm{T}}) \| \hat{p}_{\mathrm{S}}(\tilde{X}_{h_{\mathrm{S}}, N_{\mathrm{sub}}}, \tilde{Z}_{\mathrm{S}})). \end{aligned} \tag{10}$$

The first two terms $\ell_{\mathrm{T}}(C_Y, C_Z)$ and $\ell_{\mathrm{S}}(C_X, C_Z)$ are the losses in mutual information for co-clustering the cells and the shared features in the target data and source data, respectively. The shared features Z have the same cluster C_Z in both the target data and the source data. C_Z can be viewed as a bridge to transfer knowledge between the source data and the target data. The dimension of the feature space shared by the source data S and the data T is reduced by clustering and aggregating similar features. Aggregating similar features guided by the source data S enables knowledge transfer between the source data S and the data T, which reduces the noise in the single-cell data and can generally improve the clustering performance of the cells in target data. The term $\ell_{\mathrm{U}}(C_Y, C_U)$ corresponds to the loss in mutual information for co-clustering the cells and the features that are unique in the target data. The clustering of the cells in target data, C_Y, is the same in terms $\ell_{\mathrm{U}}(C_Y, C_U)$ and $\ell_{\mathrm{T}}(C_Y, C_Z)$. The term $D_{\mathrm{KL}}(\hat{p}_{\mathrm{T}}(\tilde{Y}_{h_{\mathrm{T}}, N_{\mathrm{sub}}}, \tilde{Z}_{\mathrm{T}}) \| \hat{p}_{\mathrm{S}}(\tilde{X}_{h_{\mathrm{S}}, N_{\mathrm{sub}}}, \tilde{Z}_{\mathrm{S}}))$ aims to match a subset of the cell clusters in the two datasets. λ, β, and γ are tuning parameters.

The objective function in coupleCoC is similar to coupleCoC+. Its differences from coupleCoC+ include that coupleCoC does not consider the unlinked features across datasets, and the matching of cell types across datasets is implemented in a separate step, instead of being integrated in the objective function. Apart

from the integrative analysis of scRNA-Seq and scATAC-Seq data, coupleCoC and coupleCoC+ were utilized for the integrative analysis of sc-methylation and scRNA-Seq data, and scRNA-Seq data from mouse and human.

2 Multi-Omics Data Profiled on the Same Single Cells

Technologies that can profile multiple types of genomic features simultaneously in the same cells are beginning to emerge, and have the potential to reveal causal regulatory relations [27–33]. Methods have been developed for integrative analysis of these datasets, where one major goal is to integrate multiple molecular modalities profiled on the same cells to obtain better dimension reduction and clustering results compared with using single modalities alone [34–39]. These methods do not require the features to be linked across different molecular modalities.

MOFA+ [34] was designed to capture a common latent space, which integrates multi-omics data obtained from the same set of cells. It also considers sample structure (batches, donors, etc.) in the factor analysis model. Let M denote the number of data modalities, MOFA+ assumes the following factor analysis model for the mth data modality:

$$\mathbf{Y}_{gm} = \mathbf{Z}_g \mathbf{W}_m^T + \boldsymbol{\epsilon}_{gm}. \tag{11}$$

$\mathbf{Y}_{gm}$ is a $N_g \times D_m$ matrix, where N_g is the number of cells in group/batch g, and D_m is the number of features in modality m; $\mathbf{Z}_g$ is a $N_g \times K$ matrix, which represents the matrix of K factors in group g; $\mathbf{W}_m$ is a $D_m \times K$ matrix, which is the weight matrix for the mth modality. $\boldsymbol{\epsilon}_{gm}$ is the residual noise matrix. In MOFA+, the factor matrix $\mathbf{Z}_g$ is shared across different modalities within group g, and the weight matrix $\mathbf{W}_m$ is shared for the same modality across different groups. Element-wise spike-and-slab prior was assumed for the entries in $\mathbf{Z}_g$ and $\mathbf{W}_m$ for regularization. MOFA+ also extends model 11 and supports non-Gaussian likelihoods, including a Poisson model for count data and a Bernoulli model for binary data. Inference of the models was achieved using stochastic variational inference, which can scale up to large datasets.

WNN (Seurat V4) [35] was primarily designed for the analysis of CITE-Seq data (RNA + surface protein abundance), and it was also applied to paired measurement of RNA and chromatin accessibility for the same cells. The key is to construct a weighted nearest neighbor (WNN) graph, defined as a K-nearest neighbor (KNN) graph constructed using a weighted similarity metric, which combines the information in the two modalities. The WNN graph can then be used for downstream analysis, including data visualization, clustering, and trajectory analysis. The weighted similarity between cell i and cell j is defined as

$$\theta_{weighted}(i, j) = w_{rna}(i)\theta_{rna}(\mathbf{r}_i, \mathbf{r}_j) + w_{protein}(i)\theta_{protein}(\mathbf{p}_i, \mathbf{p}_j), \tag{12}$$

where $\mathbf{r}$ represents the observed RNA profile for a cell, $\mathbf{p}$ represents the observed surface protein level for a cell. $w_{rna}(i)$ and $w_{protein}(i)$ are the weights for RNA and protein profiles, respectively, and the weights depend on the cell label i. $\theta_{rna}(\mathbf{r}_i, \mathbf{r}_j)$ and $\theta_{protein}(\mathbf{p}_i, \mathbf{p}_j)$ denotes the affinities between cell i and cell j computed from RNA levels and protein levels, respectively, and they are defined as the following:

$$\begin{aligned}\theta_{rna}(\mathbf{r}_i, \mathbf{r}_j) &= \exp\left(\frac{-\max(d(\mathbf{r}_i, \mathbf{r}_j) - d(\mathbf{r}_i, \mathbf{r}_{knn_{r,i,1}}), 0)}{\sigma_{r,i} - d(\mathbf{r}_i, \mathbf{r}_{knn_{r,i,1}})}\right), \\ \theta_{protein}(\mathbf{p}_i, \mathbf{p}_j) &= \exp\left(\frac{-\max(d(\mathbf{p}_i, \mathbf{p}_j) - d(\mathbf{p}_i, \mathbf{p}_{knn_{p,i,1}}), 0)}{\sigma_{p,i} - d(\mathbf{p}_i, \mathbf{p}_{knn_{p,i,1}})}\right),\end{aligned} \tag{13}$$

where $\mathbf{r}_{knn_{r,i,1}}$ denotes the RNA profile for the cell that is closest to cell i, using RNA data to calculate the distance; $\mathbf{p}_{knn_{p,i,1}}$ denotes the protein profile for the cell that is closest to cell i, using protein data to calculate the distance. So the affinities $\theta_{rna}(\mathbf{r}_i, \mathbf{r}_j)$ and $\theta_{protein}(\mathbf{p}_i, \mathbf{p}_j)$ represent the similarities between cells i and j using RNA and protein profiles, while considering the distance between cell i and its nearest neighbor.

The weights $w_{rna}(i)$ and $w_{protein}(i)$ are chosen as the following:

$$\begin{aligned}s_{rna}(i) &= \frac{\theta_{rna}(\mathbf{r}_i, \hat{\mathbf{r}}_{i,knn_r})}{\theta_{rna}(\mathbf{r}_i, \hat{\mathbf{r}}_{i,knn_p}) + \epsilon}, s_{protein}(i) = \frac{\theta_{protein}(\mathbf{p}_i, \hat{\mathbf{p}}_{i,knn_p})}{\theta_{protein}(\mathbf{p}_i, \hat{\mathbf{p}}_{i,knn_r}) + \epsilon}, \\ w_{rna}(i) &= \frac{e^{s_{rna}(i)}}{e^{s_{rna}(i)} + e^{s_{protein}(i)}}, w_{protein}(i) = \frac{e^{s_{protein}(i)}}{e^{s_{rna}(i)} + e^{s_{protein}(i)}},\end{aligned} \tag{14}$$

where $\hat{\mathbf{r}}_{i,knn_r}$ and $\hat{\mathbf{r}}_{i,knn_p}$ are the average RNA profiles among the neighbors of cell i: in $\hat{\mathbf{r}}_{i,knn_r}$, the neighborhood is obtained by the closest distances in RNA profiles; While in $\hat{\mathbf{r}}_{i,knn_p}$, the neighborhood is obtained by the closest distances in protein profiles. $\hat{\mathbf{p}}_{i,knn_p}$ and $\hat{\mathbf{p}}_{i,knn_r}$ are the average protein profiles among the neighbors of cell i, where the neighborhoods are obtained by the closest distances in protein profiles and RNA profiles, respectively. The intuition for choosing the weight is that when the neighborhood obtained from RNA profiles better predicts the RNA profiles and protein profiles for cell i, compared with the neighborhood obtained from the protein profiles, the weight $w_{rna}(i)$ will tend to be larger.

TotalVI [36] was developed for CITE-Seq data and it is based on variational autoencoder [40]. Suppose that there are B batches. $\boldsymbol{s}_n$ is a vector of length B, which represents the known one-hot batch index for cell n. The batch index $\boldsymbol{s}_n$ is the same for RNA data and protein data. TotalVI learns a shared latent representation for RNA and protein data. z_n is the latent representation for cell n, the prior on z_n is specified as

$$z_n \sim \text{LogisticNormal}(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \boldsymbol{I}). \tag{15}$$

The following hierarchical model was assumed for the RNA levels in CITE-Seq data.

Similar to the specification in scVI [41], the size factor for RNA data for cell n, represented as $\ell_n \in \mathbb{R}_+$, was assumed to be latent and it depends on the batch index $\boldsymbol{s}_n$:

$$\ell_n \mid \boldsymbol{s}_n \sim \text{LogNormal}(\mu = \boldsymbol{\ell}_{\mu}^T \boldsymbol{s}_n, \sigma^2 = \boldsymbol{\ell}_{\sigma^2}^T \boldsymbol{s}_n), \tag{16}$$

where $\boldsymbol{\ell}_{\mu}$ and $\boldsymbol{\ell}_{\sigma^2}$ are vectors of length B, and their entries are set to the empirical mean and variance of the log(RNA library size) calculated from the cells within individual batches. Let x_{ng} denote the observed RNA count for gene g in cell n, and it was modeled as the following:

$$\begin{aligned} x_{ng} \mid l_n, \boldsymbol{\rho}_n, \theta_g &\sim \text{NB}(\text{mean} = l_n \rho_{ng}, \text{dispersion} = 1/\theta_g), \\ \boldsymbol{\rho}_n &= f_{\boldsymbol{\rho}}(\boldsymbol{z}_n, \boldsymbol{s}_n), \end{aligned} \tag{17}$$

where NB stands for negative binomial distribution. The function $f_{\rho}(\boldsymbol{z}_n, \boldsymbol{s}_n)$ is a neural network: its inputs are $\boldsymbol{z}_n$ and $\boldsymbol{s}_n$, and the output is a vector $\boldsymbol{\rho}_n$, which represents the abundance of the genes in cell n. The model specification for RNA data is very similar to that in scVI, and the major difference is that zero inflation was not considered in TotalVI.

The following hierarchical model was assumed for the protein levels in CITE-Seq data.

Let y_{nt} denote the observed count for protein t in cell n. It was assumed to follow a negative binomial mixture distribution:

$$\begin{aligned} y_{nt} \mid v_{nt}, \beta_{nt}, \alpha_{nt} \sim\ & v_{nt}\text{NB}(\text{mean} = \beta_{nt}, \text{dispersion} = 1/\phi_t) + \\ & (1 - v_{nt})\text{NB}(\text{mean} = \beta_{nt}\alpha_{nt}, \text{dispersion} = 1/\phi_t), \end{aligned} \tag{18}$$

where v_{nt} is a binary latent variable representing the mixture component. β_{nt} represents the background intensity, and $\alpha_{nt} > 1$ represents the fold change in mean for the mixture component with the larger mean. So $v_{nt} = 0$ represents the mixture component with a larger mean. The distribution for v_{nt} was specified as the following:

$$\begin{aligned} v_{nt} \mid \boldsymbol{\pi}_n &\sim \text{Bernoulli}(\pi_{nt}), \\ \boldsymbol{\pi}_n &= h_{\boldsymbol{\pi}}(\boldsymbol{z}_n, \boldsymbol{s}_n), \end{aligned} \tag{19}$$

where $h_{\boldsymbol{\pi}}(\boldsymbol{z}_n, \boldsymbol{s}_n)$ is a neural network: its inputs are $\boldsymbol{z}_n, \boldsymbol{s}_n$, and its output is a vector of probabilities $\boldsymbol{\pi}_n$ for cell n.

The distribution of β_{nt} is specified as

$$\beta_{nt} \mid \boldsymbol{s}_n \sim \text{LogNormal}(\mu = \boldsymbol{c}_t^T \boldsymbol{s}_n, \sigma^2 = \boldsymbol{d}_t^T \boldsymbol{s}_n), \tag{20}$$

where $\boldsymbol{c}_t$ and $\boldsymbol{d}_t$ are parameters to be estimated from the data. The variable α_{nt} is specified as $\boldsymbol{\alpha}_n = g_{\boldsymbol{\alpha}}(\boldsymbol{z}_n, \boldsymbol{s}_n)$, where $g_{\boldsymbol{\alpha}}(\boldsymbol{z}_n, \boldsymbol{s}_n)$ is a neural network. Inference of TotalVI was performed under the variational autoencoder framework.

scAI [37] was developed for the integrative analysis of single-cell transcriptome and epigenome profiled in the same single cells, and it is based on non-negative matrix factorization. The following optimization problem was proposed in scAI:

$$\underset{\mathbf{W}_1,\mathbf{W}_2,\mathbf{H},\mathbf{Z}\geq 0}{\arg\min}\ \alpha\|\mathbf{X}_1-\mathbf{W}_1\mathbf{H}\|_F^2+\|\mathbf{X}_2(\mathbf{Z}\circ\mathbf{R})-\mathbf{W}_2\mathbf{H}\|_F^2+\lambda\|\mathbf{Z}-\mathbf{H}^T\mathbf{H}\|_F^2+\gamma\sum_j\|\mathbf{H}_{\cdot j}\|_1^2. \tag{21}$$

$\mathbf{X}_1$ is the normalized $p \times n$ (p genes in n cells) data matrix for single-cell transcriptomic data, and $\mathbf{X}_2$ is the normalized $q \times n$ (q regions in n cells) data matrix for single-cell epigenomic data. $\mathbf{W}_1$ and $\mathbf{W}_2$ are the gene loading and region loading matrices with dimensions $p \times K$ and $q \times K$, respectively. $\mathbf{H}$ is the $K \times n$ cell loading matrix shared by the transcriptomic and epigenomic data. $\mathbf{Z}$ is the $n \times n$ cell-cell similarity matrix. $\mathbf{R}$ is a binary matrix generated by a binomial distribution with probability s. The symbol $\circ$ represents element-wise multiplication. The term $\mathbf{X}_2(\mathbf{Z} \circ \mathbf{R})$ has a smoothing effect on the single-cell epigenomic data matrix, where the epigenomic profiles from similar cells are being aggregated based on the cell-cell similarity matrix $\mathbf{Z}$, and this term is helpful to deal with the sparsity and high level of noise in single-cell epigenomic data.

JSNMF [38] was also developed for the integrative analysis of single-cell transcriptome and epigenome profiled in the same single cells. The following optimization problem was proposed in JSNMF:

$$\begin{aligned}\min_{\mathbf{W}_i,\mathbf{H}_i,\mathbf{Z},\lambda_i}&\sum_{i=1}^{2}\|\mathbf{X}_i-\mathbf{W}_i\mathbf{H}_i\|_F^2+\frac{\alpha}{2}\sum_{i=1}^{2}\|\mathbf{Z}-\mathbf{H}_i^T\mathbf{H}_i\|_F^2+\sum_{i=1}^{2}\frac{\varphi_i}{2}\|\mathbf{H}_i\mathbf{H}_i^T\\&-\mathbf{I}\|_F^2+\eta\|\mathbf{1}^T\mathbf{Z}-\mathbf{1}^T\|_F^2+\gamma\sum_{i=1}^{2}\lambda_i^2\,tr(\mathbf{H}_i\mathbf{L}_i\mathbf{H}_i^T)\\&\text{s.t. }\mathbf{W}_i,\mathbf{H}_i,\mathbf{Z},\lambda_i\geq 0,\text{ for } i\in\{1,2\};\ \sum_{i=1}^{2}\lambda_i^2=1.\end{aligned} \tag{22}$$

Similar to scAI, JSNMF is also based on non-negative matrix factorization. $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{W}_1$, and $\mathbf{W}_2$ have similar interpretation as those in scAI. One key difference between JSNMF and scAI is that JSNMF assumes different cell loading matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ for the two data modalities and integrate the information in $\mathbf{H}_1$ and $\mathbf{H}_2$ through consensus graph fusion, $\sum_{i=1}^{2}\|\mathbf{Z}-\mathbf{H}_i^T\mathbf{H}_i\|_F^2$. This integration strategy was shown to be beneficial when the data from different types of genomic features have different levels of noise. The term $\sum_{i=1}^{2}\frac{\varphi_i}{2}\|\mathbf{H}_i\mathbf{H}_i^T-\mathbf{I}\|_F^2$ improves interpretability of the factors. $\|\mathbf{1}^T\mathbf{Z}-\mathbf{1}^T\|_F^2$ is a normalization term that encourages the columns in $\mathbf{Z}$ to have summations close to 1. $\mathbf{L}_i \in R^{n\times n}$ is the Laplacian graph for

the ith data modality, and it captures the high-dimensional geometrical structure in the original data space. The term $\sum_{i=1}^{2} \lambda_i^2 \ tr(\mathbf{H}_i \mathbf{L}_i \mathbf{H}_i^T)$ encourages the low-dimensional embeddings $\mathbf{H}_i$ to preserve the high-dimensional geometrical structure. In JSNMF, formula 22 was also extended to the integration of more than two molecular modalities profiled on the same cells and the integration of multiple single-cell multi-omics experiments. JSNMF also includes a module that infers cell type-specific region-gene associations.

3 Challenges and Future Perspectives

Different molecular modalities capture different aspects of the cell. Most methods in this review focus on exploratory analysis, including dimension reduction and clustering. The natural next step is methodology development for downstream analysis, including estimating the transcriptional regulatory network, data integration with the summary statistics in genome-wide association analysis to unravel the mechanism of human diseases, and relating single-cell multi-omics with the clinical outcome of the patients.

Multi-omics data obtained from the same single cells tend to be noisier compared with single-omic data. It will be interesting to integrate these data with other existing reference data, especially atlas-scale data, to help deal with the high noise level. Computational burden will be another challenge following technology developments that increase the throughput of cells. Single-cell epigenomic data tend to have much more features compared with scRNA-Seq data, and the analysis of these datasets will be more demanding computationally.

Acknowledgments We thank Prof. Wing Hung Wong for the constructive comments in preparing this review. This work has been supported by the Chinese University of Hong Kong startup grant (4930181), Hong Kong Research Grant Council (ECS 24301419, GRF 14301120).

References

1. The Human Cell Atlas Participants (2017) Science forum: the human cell atlas. Elife 6:e27041
2. Haghverdi L, Lun AT, Morgan MD, Marioni JC (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36(5):421–427
3. Hie B, Bryson B, Berger B (2019) Efficient integration of heterogeneous single-cell transcriptomes using scanorama. Nat Biotechnol 37(6):685–691
4. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park JE (2020) Bbknn: fast batch alignment of single cell transcriptomes. Bioinformatics 36(3):964–965
5. Song F, Chan GMA, Wei Y (2020) Flexible experimental designs for valid single-cell rna-sequencing experiments allowing batch effects correction. Nat Commun 11(1):1–15
6. Peng M, Li Y, Wamsley B, Wei Y, Roeder K (2021) Integration and transfer learning of single-cell transcriptomes via cFIT. Proc Natl Acad Sci 118(10):e2024383118

7. Richardson S, Tseng GC, Sun W (2016) Statistical methods in integrative genomics. Ann Rev Stat Appl 3:181–209
8. Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, Wang Y, Wong WH (2018) Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. Proc Natl Acad Sci 115(30):7723–7728
9. Zeng W, Chen X, Duren Z, Wang Y, Jiang R, Wong WH (2019) DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. Nat Commun 10(1):1–11
10. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck III WM, Hao Y, Stoeckius M, Smibert P, Satija R (2019) Comprehensive integration of single-cell data. Cell 177(7):1888–1902
11. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ (2019) Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell 177(7):1873–1887
12. Gao C, Liu J, Kriebel AR, Preissl S, Luo C, Castanon R, Sandoval J, Rivkin A, Nery JR, Behrens MM, et al. (2021) Iterative single-cell multi-omic integration using online learning. Nat Biotechnol 39(8):1000–1007
13. Kriebel AR, Welch JD (2021) Nonnegative matrix factorization integrates single-cell multi-omic datasets with partially overlapping features. bioRxiv
14. Wang C, Sun D, Huang X, Wan C, Li Z, Han Y, Qin Q, Fan J, Qiu X, Xie Y et al. (2020) Integrative analyses of single-cell transcriptome and regulome using MAESTRO. Genome Biol 21(1):1–28
15. Lin Z, Zamanighomi M, Daley T, Ma S, Wong WH (2020) Model-based approach to the joint analysis of single-cell data on chromatin accessibility and gene expression. Stat Sci 35(1):2–13
16. Wangwu J, Sun Z, Lin Z (2021) scAMACE: model-based approach to the joint analysis of single-cell data on chromatin accessibility, gene expression and methylation. Bioinformatics 37(21):3874–380
17. Zeng P, Wangwu J, Lin Z (2020) Coupled co-clustering-based unsupervised transfer learning for the integrative analysis of single-cell genomic data. Briefings Bioinform 22(4):bbaa347
18. Zeng P, 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(6):e1009064
19. Lin Y, Wu TY, Wan S, Yang JY, Wong WH, Wang Y (2022) scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. Nat Biotechnol 40(5):703–710
20. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, et al. (2018) A single-cell atlas of in vivo mammalian chromatin accessibility. Cell 174(5):1309–1324
21. Kriebel AR, Welch JD (2022) UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. Nat Commun 13(1):1–17
22. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791
23. Yang Z, Michailidis G (2016) A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics 32(1):1–8
24. Mairal J, Bach F, Ponce J, Sapiro G (2010) Online learning for matrix factorization and sparse coding. J Mach Learn Res 11(1)
25. Qin Q, Fan J, Zheng R, Wan C, Mei S, Wu Q, Sun H, Brown M, Zhang J, Meyer CA et al. (2020) Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. Genome Biol 21(1):1–14
26. Layer RM, Pedersen BS, DiSera T, Marth GT, Gertz J, Quinlan AR (2018) GIGGLE: a search engine for large-scale integrated genome analysis. Nat Methods 15(2):123–126
27. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, Daza RM, McFaline-Figueroa JL, Packer JS, Christiansen L, Steemers FJ, Adey AC, Trapnell C, Shendure J (2018) Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science 361(6409):1380–1385

28. Chen S, Lake BB, Zhang K (2019) High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol 37(12):1452–1457
29. Zhu C, Yu M, Huang H, Juric I, Abnousi A, Hu R, Lucero J, Behrens MM, Hu M, Ren B (2019) An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. Nat Struct Mol Biol 26:1063–1070
30. Argelaguet R, Clark SJ, Mohammed H, Stapel LC, Krueger C, Kapourani C, Imaz-Rosshandler I, Lohoff T, Xiang Y, Hanna CW, Smallwood S, Ibarra XS, Buettner F, Sanguinetti G, Xie W, Krueger F, Gottgens B, Rugg PJG, Kelsey G, Dean W, Nicholas J, Stegle O, Marioni JC, Reik W (2019) Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature 576(7787):487–491
31. Ma S, Zhang B, LaFave L, Chiang Z, Hu Y, Ding J, Brack A, Kartha VK, Law T, Lareau C, Hsu YC, Regev A, Buenrostro JD (2020) Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell 183(4):1103–1116
32. Zhu C, Zhang Y, Li YE, Lucero J, Behrens MM, Ren B (2021) Joint profiling of histone modifications and transcriptome in single cells from mouse brain. Nat Methods 18(3):283–292
33. Xiong H, Luo Y, Wang Q, Yu X, He A (2021) Single-cell joint detection of chromatin occupancy and transcriptome enables higher-dimensional epigenomic reconstructions. Nat Methods 18(6):652–660
34. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O (2020) MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol 21(1):1–17
35. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R (2021) Integrated analysis of multimodal single-cell data. Cell 184(13):3573–3587.e29
36. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, Yosef N (2021) Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat Methods 18(3):272–282
37. Jin S, Zhang L, Nie Q (2020) scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. Genome Biol 21(1):1–19
38. Ma Y, Sun Z, Zeng P, Zhang W, Lin Z (2022) JSNMF enables effective and accurate integrative analysis of single-cell multiomics data. Briefings Bioinform 23(3):p.bbac105
39. Liu Q, Chen S, Jiang R, Wong WH (2021) Simultaneous deep generative modelling and clustering of single-cell genomic data. Nat Mach Intell 3(6):536–544
40. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:13126114
41. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N (2018) Deep generative modeling for single-cell transcriptomics. Nature Methods 15(12):1053
42. Dhillon IS, Mallela S, Modha DS (2003) Information-theoretic co-clustering. Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, pp 89–98