Statistical Science
2020, Vol. 35, No. 1, 2–13
https://doi.org/10.1214/19-STS714

Model-Based Approach to the Joint Analysis of Single-Cell Data on Chromatin Accessibility and Gene Expression

Zhixiang Lin, Mahdi Zamanighomi, Timothy Daley, Shining Ma and Wing Hung Wong

Abstract. Unsupervised methods, including clustering methods, are essential to the analysis of single-cell genomic data. Model-based clustering methods are under-explored in the area of single-cell genomics, and have the advantage of quantifying the uncertainty of the clustering result. Here we develop a model-based approach for the integrative analysis of single-cell chromatin accessibility and gene expression data. We show that combining these two types of data, we can achieve a better separation of the underlying cell types. An efficient Markov chain Monte Carlo algorithm is also developed.

Key words and phrases: Single-cell genomics, coupled clustering, Bayesian modeling, MCMC.

1. INTRODUCTION

Single-cell sequencing-based technologies have become the primary tool to profile genomic features for hundreds or even thousands of cells in parallel. The measurement of gene expression is an imperfect substitute for the quantification of protein abundance. Although the majority of single-cell genomic research published to date focuses on characterizing gene expression at the single cell level, other single-cell sequencing technologies that capture functional genomic features are emerging and the available datasets are growing (Rozenblatt-Rosen et al., 2017): including datasets from single-cell ChIP-seq (Rotem et al., 2015), single-cell methylation (Smallwood et al., 2014) and single-cell chromatin accessibility (Buenrostro et al., 2015b, Cusanovich et al., 2015). Different data types capture complementary information and together they provide a more complete view of the underlying biological process.

Despite the fact that the data structures (genomic features by samples/cells) are similar between single-cell genomic and bulk genomic data, the distinct characteristics of single-cell genomic data poses challenges for data analysis and opportunities for methodology development: (a) abundance of zeros: the zeros in the data matrix can be true biologically or false due to the failure to detect the biological signal. Technical failure to detect the signal is commonly observed in single-cell data and is referred to as *dropout* for single-cell gene expression experiments (Kharchenko, Silberstein and Scadden, 2014, Zhu et al., 2018). (b) batch effect/confounding variation: the standard balanced experimental designs are not possible for certain experimental protocols. These technical variabilities have been demonstrated to affect single-cell gene expression data analysis and a more comprehensive discussion is presented in Hicks et al. (2018).

The characterization of cell types based on their genomic signatures is one of the key computational challenges in single-cell genomics as the cell identity is unknown and needs to be inferred (Bacher and Kendziorski, 2016). The clustering methods developed so far have been mostly focused on single-cell gene expression data, and they can be classified as algorithm-based and probabilistic model-based methods.

Algorithm-based clustering methods usually build upon different similarity/distance metrics between the cells. SNN-Cliq (Xu and Su, 2015) uses shared nearest neighbor (SNN) graph based upon a subset of genes and clusters cells by identifying and merging sub-graphs; *pcaReduce* (Yau et al., 2016) integrates principal components analysis and hierarchical clustering; RaceID (Grün et al., 2016)

Zhixiang Lin is Assistant Professor, Department of Statistics, The Chinese University of Hong Kong, Sha Tin, Hong Kong SAR, China (e-mail: zhixianglin@cuhk.edu.hk). Mahdi Zamanighomi is Computational Biologist, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. Timothy Daley is Postdoctoral Fellow, Department of Statistics and Department of Bioengineering, Stanford University, Stanford, California, USA. Shining Ma is Postdoctoral Fellow, Department of Statistics, Stanford University, Stanford, California, USA. Wing Hung Wong is Professor, Department of Statistics and Department of Biomedical Data Science, Stanford University, Stanford, California, USA. (e-mail: whwong@stanford.edu).

uses an iterative k-means clustering algorithm based on a similarity matrix of Pearson's correlation coefficients; SC3 (Kiselev et al., 2017) is an ensemble clustering algorithm that combines the clustering outcomes of several other methods; CIDR (Lin, Troup and Ho, 2017) first imputes the gene expression profiles, calculate the dissimilarly matrix based on the imputed data matrix, perform dimension reduction by principal coordinate analysis and finally perform clustering on the first several principal coordinates; SIMLR (Wang et al., 2017) implements a kernel based similarity learning algorithm, where RBF kernel is utilized with Euclidean distance. The 'Corr' method proposes a new cell similarity measure based on cell-pair differentiability correlation and implements hierarchical clustering. SAFE-clustering (Yang et al., 2018) is another ensemble clustering algorithm that uses hypergraph-based partitioning algorithms. SOUP (Zhu et al., 2019) is a semi-soft clustering algorithm that first identifies the set of pure cells by exploiting the block structures in the cell-cell similarity matrix, use them to build the membership matrix, and then estimates the soft memberships for the other cells. For the analysis of single-cell epigenomic data, such as single-cell chromatin accessibility, *scABC* was proposed in Zamanighomi et al. (2018), where a weighted K-Medoids clustering algorithm was proposed, followed by aggregation of the reads within a cluster and cluster reassignment by the nearest neighbor.

Probabilistic model-based approaches for clustering single-cell data are still under-explored. DIMM-SC (Sun et al., 2017) builds upon a Dirichlet mixture model and is designed to cluster droplet-based single-cell transcriptomic data. The benefit for model-based approaches is that the clustering uncertainty can be quantified for each single cell, facilitating rigorous statistical inference and biological interpretations, which are typically unavailable from algorithm-based clustering methods (except for SOUP). Statistical inference of the clustering uncertainty can be particularly important when there are cells at the intermediate stage, which is expected in some biological processes, such as the stem cell differentiation process and carcinogenesis.

In this paper, we focus on the joint analysis of single-cell gene expression and single-cell chromatin accessibility data. Eukaryotic genomes are hierarchically packaged into chromatin, and the nature of this packaging plays a central role in gene regulation (Buenrostro et al., 2013). ATAC-seq maps transposase-accessible chromatin regions, and provides information for understanding this epigenetic structure of chromatin packaging and for understanding gene regulation (Buenrostro et al., 2015a). Single-cell chromatin accessibility (scATAC-Seq) maps chromatin accessibility at single-cell resolution and provides insight on the cell-to-cell variation of gene regulation (Buenrostro et al., 2015b).

Most current clustering methods are restricted to one data type and do not address the increasingly common situation where two or more types of single-cell genomic experiments are performed on different subsamples (i.e., cells) from the same cell population (i.e., tissue/biological sample) (Pollen et al., 2014, Buenrostro et al., 2015b, Lake et al., 2018, Zamanighomi et al., 2018, Duren et al., 2018). Given two data types obtained from the same cell population but from different cells, our goal is to cluster and match the cell types in these two data types (see Figure 1(a)). The benefits of solving this "coupled clustering" problem include the following: (1) the cell types may be better separated combining multiple data types. Different data types can have different power in separating the cell types (Corces et al., 2016) and combining all the information can help us separate the cell types. Moreover, the batch effect/confounding variation is expected to affect the data types differently and we may alleviate this technical artifact via the integrative analysis (Zang et al., 2016). (2) Matched clusters provide rich biological information. Different data types provide complementary biological information and it is beneficial to match the cell subpopulations at the cluster level (Duren et al., 2017). The "coupled clustering" problem was first introduced and tackled by the *coupleNMF* algorithm in Duren et al. (2018). The *coupleNMF* algorithm is based on extensions of nonnegative matrix factorization (NMF). The connection between chromatin accessibility and gene expression data builds upon prediction models trained from bulk data with diverse cell types.

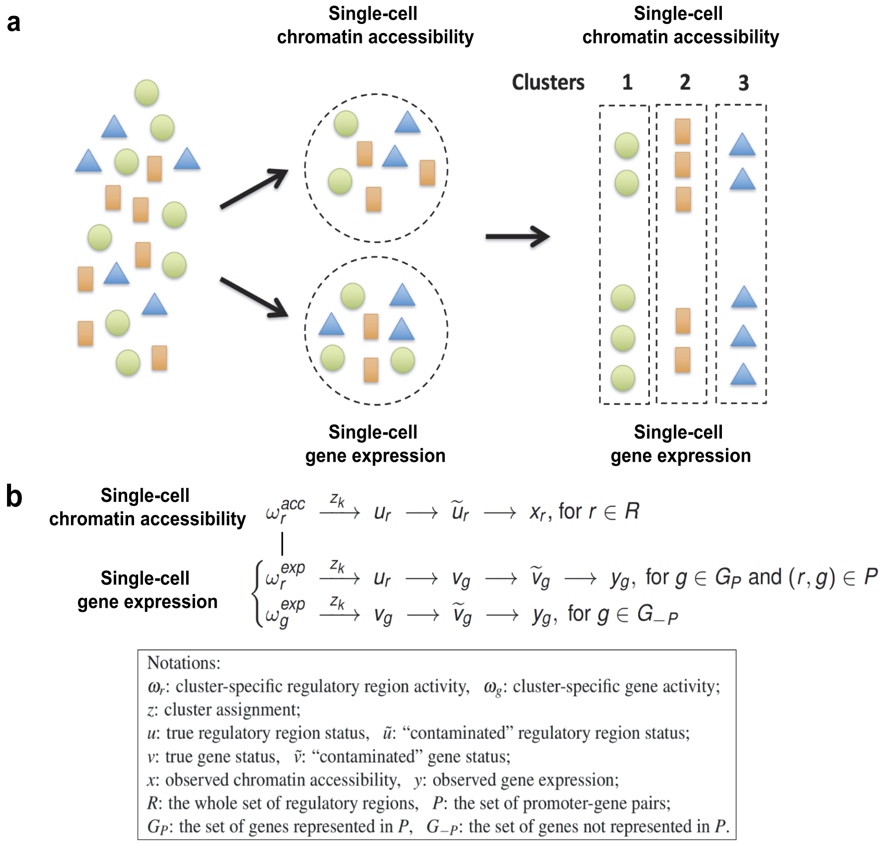
Here we propose a model-based approach to jointly cluster single-cell chromatin accessibility and single-cell gene expression data. Our approach has the following features: (1) our model does not rely on training data to connect the two data types; (2) The noisiness in single-cell experiments is taken into account by explicitly modeling the loss of biological signals; (3) How well the two data types are matched is adaptively inferred from the data; (4) Our model allows for statistical inference of the cluster assignment; (5) An efficient Markov chain Monte Carlo algorithm is developed that incorporates collapsing and auxiliary variables.

2. STATISITCAL MODEL AND METHODS

A graphical overview of the model is presented in Figure 1(b).

2.1 Modeling Single-Cell Chromatin Accessibility Data

Let u_{ir} denote the true status of a regulatory element in the single-cell experiment, where $u_{ir} = 1$ indicates that the element r is accessible in cell i and $u_{ir} = 0$ indicates that it is not accessible. Let $\tilde{u}_{ir}$ denote the "contaminated"


a
Single-cell chromatin accessibility
Single-cell chromatin accessibility
Clusters 1 2 3
Single-cell gene expression
Single-cell gene expression
b
Single-cell chromatin accessibility
$\omega_r^{acc} \xrightarrow{z_k} u_r \longrightarrow \tilde{u}_r \longrightarrow x_r$, for $r \in R$
Single-cell gene expression
$\omega_r^{exp} \xrightarrow{z_k} u_r \longrightarrow v_g \longrightarrow \tilde{v}_g \longrightarrow y_g$, for $g \in G_P$ and $(r,g) \in P$
$\omega_g^{exp} \xrightarrow{z_k} v_g \longrightarrow \tilde{v}_g \longrightarrow y_g$, for $g \in G_{-P}$

Notations:
ω_r: cluster-specific regulatory region activity, ω_g: cluster-specific gene activity;
z: cluster assignment;
u: true regulatory region status, $\tilde{u}$: "contaminated" regulatory region status;
v: true gene status, $\tilde{v}$: "contaminated" gene status;
x: observed chromatin accessibility, y: observed gene expression;
R: the whole set of regulatory regions, P: the set of promoter-gene pairs;
G_P: the set of genes represented in P, G_{-P}: the set of genes not represented in P.

FIG. 1. *(a) Schematic plot for the coupled clustering problem. (b) Graphical representation of the clustering model.*

status due to the loss of biological signal in the experimental process, which is commonly observed in single-cell experiments. Let x_{ir} denote the observed data for accessibility. We propose the following model for single-cell chromatin accessibility data:

$$\begin{aligned} \tilde{u}_{ir} \mid u_{ir} &\sim u_{ir}\,\mathrm{Bernoulli}(q_i) \\ &\quad + (1-u_{ir})\,\mathrm{Bernoulli}(0), \\ p(x_{ir} \mid \tilde{u}_{ir}) &= \tilde{u}_{ir} f_1(x_{ir}) + (1-\tilde{u}_{ir}) f_0(x_{ir}), \end{aligned} \tag{2.1}$$

where the loss of biological signal during the experimental process is modeled by a Bernoulli distribution with probability q_i, representing the capture rate, which is assumed to be cell-specific. Conditioning on the status $\tilde{u}_{ir}$, the observed accessibility x_{ir} follows a mixture distribution with components $f_1(\cdot)$ and $f_0(\cdot)$.

2.2 Modeling Single-Cell Gene Expression Data

Let v_{lg} denote the true status of a gene in the single-cell experiment, where $v_{lg} = 1$ indicates that the gene g is expressed in cell l and $v_{lg} = 0$ indicates that it is unexpressed. Here we use a different notation for the cell label to represent the fact that chromatin accessibility and gene expression are measured on different cells. Let $\tilde{v}_{lg}$ denote the "contaminated" gene status due to the loss of biological signal in the experimental process (i.e., the "dropout" event). Let y_{lg} denote the observed gene expression level. We propose the following model for single-cell gene expression data:

$$\begin{aligned} \tilde{v}_{lg} \mid v_{lg} &\sim v_{lg}\,\mathrm{Bernoulli}(q_l) \\ &\quad + (1-v_{lg})\,\mathrm{Bernoulli}(0), \\ p(y_{lg} \mid \tilde{v}_{lg}) &= \tilde{v}_{lg} g_1(y_{lg}) + (1-\tilde{v}_{lg}) g_0(y_{lg}), \end{aligned} \tag{2.2}$$

where the loss of biological signal is also modeled by Bernoulli distribution. Conditioning on the status $\tilde{v}_{lg}$, the observed gene expression y_{lg} follows a mixture distribution with components $g_1(\cdot)$ and $g_0(\cdot)$.

2.3 Coupling the Two Data Types via Promoter Activity

Let $P = \{(r, g)$, where promoter r regulates gene $g\}$ represent the set of promoter-gene pairs, and we have $P \subseteq R \times G$. Here we focus on the subset of genes where the gene is regulated by one promoter. Specifically, (r, g) is included in P only if gene g has a unique promoter r in RefSeq. This yields 5320 pairs of (r, g) for inclusion in P in human. To link the two data types, we introduce another layer of random variables in the hierarchical model:

$$\begin{aligned} v_{lg} \mid u_{lr} &\sim u_{lr}\,\mathrm{Bernoulli}(\pi_{l1}) \\ &\quad + (1-u_{lr})\,\mathrm{Bernoulli}(\pi_{l0}), \\ &\text{where } (r, g) \in P, \end{aligned} \tag{2.3}$$

where u_{lr} is the unobserved promoter status for cell l and gene g in the gene expression data. The additional layer brings the necessary stochasticity to connect the two data types: given that the promoter is accessible (i.e., $u_{lr} = 1$), the probability that the regulated gene being expressed is π_{l1}; On the other hand, the probability that the gene being expressed is π_{l0} when the promoter is not accessible. We assume that $\pi_{l1} \geq \pi_{l0}$. This is biologically meaningful as the gene is more likely to be expressed when the promoter is accessible.

2.4 The Model for Clustering

Let $z_{ik}, k = 1, \dots, K$ denote the cluster assignment for cell i in the accessibility data, that is, $z_{ik} = 1$ or 0 depending on whether cell i belongs to cluster k or not. Similarly, let $z_{lk}, k = 1, \dots, K$ denote the cluster assignment for cell l in the gene expression data.

Let R_P denote the set of promoters represented in P and R_{-P} denote the other regulatory regions. Let G_P denote the set of genes represented in P and G_{-P} denote the other genes.

Regulatory elements $r \in R$. We assume the following clustering model:

$$u_{ir} \mid z_{ik} = 1 \sim \text{Bernoulli}(\omega_{kr}), \tag{2.4}$$

where the random variable ω_{kr} denote the probability of regulatory region r to be accessible in cluster k.

Genes $g \in G_P$. Gene g is linked to promoter r and the distribution $v_{lg} \mid u_{lr}$ is specified in the previous section. We assume the following clustering model:

$$\begin{aligned} &u_{lr} \mid z_{lk} = 1 \sim \text{Bernoulli}(\omega_{kr}), \\ &\quad \text{for } g \in G_P \text{ and } (r, g) \in P. \end{aligned} \tag{2.5}$$

The accessibility data and the gene expression data are connected at the cluster level, and $\omega_{\cdot r}$ represent the cluster centers. The multiple-layer random variables incorporate the inherent stochasticity in the biological system. For $r \in R_P$, we use both the accessibility and the gene expression data to estimate $\omega_{\cdot r}$; And for $r \in R_{-P}$, we use the accessibility data alone to estimate $\omega_{\cdot r}$.

Genes $g \in G_{-P}$. These genes are not linked to the accessibility data and we assume the following clustering model:

$$v_{lg} \mid z_{lk} = 1 \sim \text{Bernoulli}(\omega_{kg}), \quad \text{for } g \in G_{-P}, \tag{2.6}$$

where ω_{kg} is the cluster-specific probability of gene g to be expressed, for $g \in G_{-P}$. We acknowledge the misuse of notation ω, which represents both the cluster-specific regulatory region activity and the cluster-specific gene activity. For $g \in G_{-P}$, we use the gene expression data alone to estimate $\omega_{\cdot g}$.

2.5 Priors for q_i, q_l, π_{l1} and π_{l0}

We assume the following flat priors for q_i and q_l:

$$q_i \sim \text{Beta}(\alpha = 1, \beta = 1),$$
$$q_l \sim \text{Beta}(\alpha = 1, \beta = 1).$$

We assume the following flat priors for π_{l1} and π_{l0}:

$$\pi_{l0} \sim \text{Beta}(\alpha = 1, \beta = 1),$$
$$\pi_{l1} \mid \pi_{l0} \sim \mathbb{1}_{\pi_{l1} > \pi_{l0}} \text{Beta}(\alpha = 1, \beta = 1).$$

Acknowledging the misuse of notation, we use $\mathbb{1}_{\pi_{l1} > \pi_{l0}} \text{Beta}(\cdot)$ to represent the truncated beta distribution. As discussed previously, this prior specification reflects the assumption that a gene is more likely to be expressed when the promoter is accessible.

2.6 Priors for ω and More Flexibility

It can be desirable to introduce additional flexibility in $\boldsymbol{\omega}_{\cdot r}$, $r \in R_P$. Instead of setting $\boldsymbol{\omega}_{\cdot r}$, $r \in R_P$ to be the same in the two data types, we may assume that $\boldsymbol{\omega}_{\cdot r}^{\text{acc}}$ and $\boldsymbol{\omega}_{\cdot r}^{\text{exp}}$ are different but related. In this case, the joint distribution for ω_{kr}^{acc} and ω_{kr}^{exp} is specified by the following conditional probability:

$$\left.\begin{aligned} &\omega_{kr}^{\text{acc}} \sim \text{Beta}(\mu = \mu_0, \nu = \nu_0), \\ &\omega_{kr}^{\text{exp}} \mid \omega_{kr}^{\text{acc}} \\ &\quad \sim \text{Beta}(\mu = \omega_{kr}^{\text{acc}}, \nu = \nu_1), \end{aligned}\right\} \quad \text{for } r \in R_P, \tag{2.7}$$
$$\nu_0 = 2, \qquad \nu_1 \sim \text{unif}[0, 50],$$

where the beta distributions are parametrized by the mean μ and precision ν. The key is the precision parameter ν_1, which represents how well the two data types are coupled and is learned adaptively from the data. When ν_1 is large, ω_{kr}^{exp} is expected to be close to ω_{kr}^{acc}, and the two data types are coupled well. An alternative specification for ω_{kr}^{acc} and ω_{kr}^{exp} is the multivariate beta distribution (Olkin and Rubin, 1964). However, we have found that the multivariate beta distribution does not seem to work well in practice for the joint model.

The priors for ω_{kr}^{acc}, $r \in R_{-P}$ and ω_{kg}^{exp}, $g \in G_{-P}$ are as follows:

$$\omega_{kr}^{\text{acc}} \sim \text{Beta}(\mu = \mu_0, \nu = \nu_0), \quad \text{for } r \in R_{-P},$$
$$\omega_{kg}^{\text{exp}} \sim \text{Beta}(\mu = \mu_1, \nu = \nu_0), \quad \text{for } g \in G_{-P}.$$

In practice, we set $\mu_0 = \mu_1 = 0.5$, and the priors for ω_{kr}^{acc} and ω_{kg}^{exp}(for $g \in G_{-P}$) are assumed to be flat.

2.7 The Mixture Components

For scRNA-Seq data, we fit a two-component gamma mixture model for the nonzero entries, through pooling log2(TPM+1) over all the samples, and then the spike at 0 is subsequently merged with the mixture component that has a smaller mean. For scATAC-Seq data, as the read count matrix is very sparse with the majority nonzero entries less than 5, we set $f_1(x) = 0$ if $x = 0$ and $f_0(x) = 0$ if $x > 0$. In the results section, we also implemented our model for bulk DNase-Seq and bulk RNA-seq data. For bulk DNase-Seq data, we fit a two component gamma mixture model for the nonzero entries, through pooling $\log 2(\text{accessibility fold change} + 1)$ over all samples, and the spike at 0 is subsequently merged with the mixture component that has a smaller mean. For bulk RNA-Seq data, we fit a two component gamma mixture model, through pooling $\log 2(\text{FPKM} + 1)$ over all samples, and the spike at 0 is subsequently merged with the mixture component that has a smaller mean. The gamma mixture models are estimated with the expectation-maximization (EM) algorithm as implemented in the R package *mixtools* (Benaglia et al., 2009).

2.8 Summary of the Clustering Model for the Joint Analysis

Here we summarize the model (Figure 1(b)) that couples scATAC-Seq and scRNA-Seq data.

We start from the coupled features in the two data types, where $r \in R_P$ is a promoter region and $g \in G_P$ is a gene that is uniquely mapped to r. ω_{kr} is the cluster-specific accessibility activity for region r in cluster k, and u_r is the true accessibility status for region r in a cell for scATAC-Seq and scRNA-Seq data. If the cell is assigned to cluster k ($z_k = 1$), we draw u_r from a Bernoulli distribution with probability ω_{kr} (equations (2.4) and (2.5)). The random variable $\boldsymbol{\omega_r}$ connects the two data types. We assume that $\boldsymbol{\omega_r^{\text{acc}}}$ (scATAC-Seq) and $\boldsymbol{\omega_r^{\text{exp}}}$ (scRNA-Seq) are different but dependent, to allow for more flexibility, and their distributions are specified by equation (2.7). The followings are the other components for the coupled features in the model:

- scATAC-Seq: $\tilde{u}_r$ is the "contaminated" accessibility status of region r due to the loss of biological signal in the experimental process, and $\tilde{u}_r \mid u_r$ follows a mixture of Bernoulli distributions, specified by equation (2.1); x_r is the observed scATAC-Seq data, $x_r \mid \tilde{u}_r$ follows a mixture distribution, specified by equation (2.1).
- scRNA-Seq: v_g is the true gene expression status for gene g regulated by region/promoter r, and $v_g \mid u_r$ follows a mixture of Bernoulli distributions, specified by equation (2.3). $\tilde{v}_g$ is the "contaminated" gene expression status of gene g due to the loss of biological signal in the experimental process, and $\tilde{v}_g \mid v_g$ follows a mixture of Bernoulli distributions, specified by equation (2.2). y_g is the observed scRNA-Seq data, and $y_g \mid \tilde{v}_g$ follows a mixture distribution, specified by equation (2.2).

Next we discuss the uncoupled features in scATAC-Seq data, where $r \in R_{-P}$. The data generating model is the same as that for $r \in R_P$, except that we use scATAC-Seq data alone to estimate $\boldsymbol{\omega_r^{\text{acc}}}$ for these features.

Finally, we discuss the uncoupled features in scRNA-Seq data, where $g \in G_{-P}$. ω_{kg} is the cluster-specific gene expression activity for gene g in cluster k, and v_g is the true gene expression status for gene g in a cell for scRNA-Seq data. If the cell is assigned to cluster k ($z_k = 1$), we draw v_g from a Bernoulli distribution with probability ω_{kg} (equation (2.2)). Specifications of $\tilde{v}_g \mid v_g$ and $x_g \mid \tilde{v}_g$ are the same as that for the coupled features. We use scRNA-Seq data alone to estimate ω_g^{exp} for these features.

3. STATISTICAL INFERENCE

We implement Markov chain Monte Carlo (MCMC) for statistical inference. To improve the mixing, we incorporate the collapsed Gibbs sampler and introduce auxiliary variable:

Collapsed Gibbs Sampler

In practice, we found that simple Gibbs sampling can get trapped at some local areas of the posterior probability. We implement the collapsed Gibbs sampler (Liu, Wong and Kong, 1994, Liu, 1994) by integrating out u_{ir}, u_{lr}, v_{lg}, and mixing is greatly improved:

$$\tilde{u}_{ir} \mid z_{ik} = 1 \sim \text{Bernoulli}(\omega_{kr}^{\text{acc}} q_i), \qquad \text{for } r \in R,$$

$$\begin{cases} \tilde{v}_{lg} \mid z_{lk} = 1 \sim \text{Bernoulli}(\omega_{kr}^{\text{exp}} q_l \pi_{l1} + (1 - \omega_{kr}^{\text{exp}}) q_l \pi_{l0}), \\ \quad \text{for } g \in G_P \text{ and } (r, g) \in P, \\ \tilde{v}_{lg} \mid z_{lk} = 1 \sim \text{Bernoulli}(\omega_{kg}^{\text{exp}} q_l), \\ \quad \text{for } g \in G_{-P}. \end{cases}$$

Auxiliary Variable

The clusters in the two data types can be mismatched, depending on the initialization. Although the posterior probability will be higher when the clusters are correctly matched, it is unlikely for Gibbs moves to escape from such mismatches, given the high dimension of $\boldsymbol{\omega}$. To achieve more efficient exploration of the alignment between clusters, we introduce an auxiliary variable $\boldsymbol{h}$, a permutation of $1, \ldots, K$, representing how gene expression clusters are matched to the accessibility clusters. We sample from h in MCMC.

The following are the details for our MCMC.

Update $\tilde{u}_{ir}$. Let $\eta_{kir} \equiv \omega_{kr}^{\text{acc}} q_i$. The variable can be sampled directly from

$$p(\tilde{u}_{ir} \mid \cdot) \propto f_1(x_{ir})^{\tilde{u}_{ir}} f_0(x_{ir})^{1-\tilde{u}_{ir}} \times \prod_{k=1}^{K} [(\eta_{kir})^{\tilde{u}_{ir}} (1 - \eta_{kir})^{1-\tilde{u}_{ir}}]^{z_{ik}}.$$

Update $\tilde{v}_{lg}$ *for* $g \in G_P$. Let $\lambda_{klg} \equiv \omega_{kr}^{\text{exp}} q_l \pi_{l1} + (1 - \omega_{kr}^{\text{exp}}) q_l \pi_{l0}$, where $(r, g) \in P$. The variable can be sampled directly from

$$p(\tilde{v}_{lg} \mid \cdot) \propto g_1(y_{lg})^{\tilde{v}_{lg}} g_0(y_{lg})^{1-\tilde{v}_{lg}} \times \prod_{k=1}^{K} \big[(\lambda_{klg})^{\tilde{v}_{lg}} (1-\lambda_{klg})^{1-\tilde{v}_{lg}}\big]^{z_{lk}}.$$

Update $\tilde{v}_{lg}$ *for* $g \in G_{-P}$. Let $\eta_{klg} \equiv \omega^{\exp}_{kg} q_l$. The variable can be sampled directly from

$$p(\tilde{v}_{lg} \mid \cdot) \propto g_1(y_{lg})^{\tilde{v}_{lg}} g_0(y_{lg})^{1-\tilde{v}_{lg}} \times \prod_{k=1}^{K} \big[(\eta_{klg})^{\tilde{v}_{lg}} (1-\eta_{klg})^{1-\tilde{v}_{lg}}\big]^{z_{lk}}.$$

Update $\omega^{\mathrm{acc}}_{kr}$ *for* $r \in R_P$. Let $\eta_{kir} \equiv \omega^{\mathrm{acc}}_{kr} q_i$. The variable can be updated by the Metropolis-Hastings (MH) algorithm:

$$\begin{aligned} p(\omega^{\mathrm{acc}}_{kr} \mid \cdot) &\propto (\omega^{\mathrm{acc}}_{kr})^{\nu_0\mu_0-1} (1-\omega^{\mathrm{acc}}_{kr})^{-\nu_0\mu_0+\nu_0-1} \\ &\times \prod_{i=1}^{n_{\mathrm{acc}}} \big[(\eta_{kir})^{\tilde{u}_{ir}} (1-\eta_{kir})^{1-\tilde{u}_{ir}}\big]^{z_{ik}} \\ &\times \mathrm{B}(\nu_1\omega^{\mathrm{acc}}_{kr}, -\nu_1\omega^{\mathrm{acc}}_{kr}+\nu_1) (\omega^{\exp}_{kr})^{\nu_1\omega^{\mathrm{acc}}_{kr}-1} \\ &\times (1-\omega^{\exp}_{kr})^{-\nu_1\omega^{\mathrm{acc}}_{kr}+\nu_1-1}, \end{aligned}$$

where $\mathrm{B}(\cdot)$ is the beta function, and n_{acc} is the number of cells in single-cell chromatin accessibility data.

Update $\omega^{\mathrm{acc}}_{kr}$ *for* $r \in R_{-P}$. The variable can be updated by the MH algorithm:

$$p(\omega^{\mathrm{acc}}_{kr} \mid \cdot) \propto (\omega^{\mathrm{acc}}_{kr})^{\nu_0\mu_0-1} (1-\omega^{\mathrm{acc}}_{kr})^{-\nu_0\mu_0+\nu_0-1} \times \prod_{i=1}^{n_{\mathrm{acc}}} \big[(\eta_{kir})^{\tilde{u}_{ir}} (1-\eta_{kir})^{1-\tilde{u}_{ir}}\big]^{z_{ik}}.$$

Update $\omega^{\exp}_{kr}$ *for* $g \in G_p$ *and* $(r, g) \in P$. The variable can be updated by the MH algorithm:

$$p(\omega^{\exp}_{kr} \mid \cdot) \propto (\omega^{\exp}_{kr})^{\nu_1\omega^{\mathrm{acc}}_{kr}-1} (1-\omega^{\exp}_{kr})^{-\nu_1\omega^{\mathrm{acc}}_{kr}+\nu_1-1} \times \prod_{l=1}^{n_{\exp}} \big[(\lambda_{klg})^{\tilde{v}_{lg}} (1-\lambda_{klg})^{1-\tilde{v}_{lg}}\big]^{z_{lk}},$$

where $n_{\exp}$ is the number of cells in single-cell gene expression data.

Update $\omega^{\exp}_{kg}$ *for* $g \in G_{-P}$. The variable can be updated by the MH algorithm:

$$p(\omega^{\exp}_{kg} \mid \cdot) \propto (\omega^{\exp}_{kg})^{\nu_0\mu_1-1} (1-\omega^{\exp}_{kg})^{-\nu_0\mu_1+\nu_0-1} \times \prod_{l=1}^{n_{\exp}} \big[(\eta_{klg})^{\tilde{u}_{lg}} (1-\eta_{klg})^{1-\tilde{u}_{lg}}\big]^{z_{lk}}.$$

Update q_i. The variable can be updated by the MH algorithm:

$$p(q_i \mid \cdot) \propto \prod_{k,r} \big[(\eta_{kir})^{\tilde{u}_{ir}} (1-\eta_{kir})^{1-\tilde{u}_{ir}}\big]^{z_{ik}}.$$

Update q_l. The variable can be updated by the MH algorithm:

$$p(q_l \mid \cdot) \propto \prod_{k} \Big\{ \prod_{g \in G_P} \big[(\lambda_{klg})^{\tilde{v}_{lg}} (1-\lambda_{klg})^{1-\tilde{v}_{lg}}\big] \times \prod_{g \in G_{-P}} \big[(\eta_{klg})^{\tilde{v}_{lg}} (1-\eta_{klg})^{1-\tilde{v}_{lg}}\big] \Big\}^{z_{lk}}.$$

Update π_{l0}. The variable can be updated by the MH algorithm:

$$p(\pi_{l0} \mid \cdot) \propto \mathbb{1}_{\pi_{l0} \le \pi_{l1}} \prod_{k} \Big\{ \prod_{g \in G_P} \big[(\lambda_{klg})^{\tilde{v}_{lg}} (1-\lambda_{klg})^{1-\tilde{v}_{lg}}\big] \Big\}^{z_{lk}}.$$

Update π_{l1}. The variable can be updated by the MH algorithm:

$$p(\pi_{l1} \mid \cdot) \propto \mathbb{1}_{\pi_{l1} \ge \pi_{l0}} \prod_{k} \Big\{ \prod_{g \in G_P} \big[(\lambda_{klg})^{\tilde{v}_{lg}} (1-\lambda_{klg})^{1-\tilde{v}_{lg}}\big] \Big\}^{z_{lk}}.$$

Update ν_1. The variable can be updated by the MH algorithm:

$$p(\nu_1 \mid \cdot) \propto \mathbb{1}_{0 \le \nu_1 \le 50} \prod_{k, r \in R_P} \big[\mathrm{B}(\nu_1\omega^{\mathrm{acc}}_{kr}, -\nu_1\omega^{\mathrm{acc}}_{kr}+\nu_1) \times (\omega^{\exp}_{kr})^{\nu_1\omega^{\mathrm{acc}}_{kr}-1} (1-\omega^{\exp}_{kr})^{-\nu_1\omega^{\mathrm{acc}}_{kr}+\nu_1-1}\big].$$

Update z_{ik}. For each sample i, $(z_{ik})_{k=1,\dots,K}$ follows multinomial distribution, and can be sampled directly:

$$p(z_{i\cdot} \mid \cdot) \propto \prod_{k,r} \big[(\eta_{kir})^{\tilde{u}_{ir}} (1-\eta_{kir})^{1-\tilde{u}_{ir}}\big]^{z_{ik}}.$$

Update z_{lk}. For each sample l, $(z_{lk})_{k=1,\dots,K}$ follows multinomial distribution, and can be sampled directly:

$$p(z_{l\cdot} \mid \cdot) \propto \prod_{k} \Big\{ \prod_{g \in G_P} \big[(\lambda_{klg})^{\tilde{v}_{lg}} (1-\lambda_{klg})^{1-\tilde{v}_{lg}}\big] \times \prod_{g \in G_{-P}} \big[(\eta_{klg})^{\tilde{v}_{lg}} (1-\eta_{klg})^{1-\tilde{v}_{lg}}\big] \Big\}^{z_{lk}}.$$

Update $\boldsymbol{h}$. The variable can be updated by the MH algorithm (a proposal h' is obtained by randomly switching two entries in h):

$$p(\boldsymbol{h} \mid \cdot) \propto \prod_{k} \prod_{r \in R_P} (\omega^{\exp}_{h_k r})^{\nu_1\omega^{\mathrm{acc}}_{kr}-1} (1-\omega^{\exp}_{h_k r})^{-\nu_1\omega^{\mathrm{acc}}_{kr}+\nu_1-1}.$$

After $\boldsymbol{h}$ is updated, we shuffle $\boldsymbol{\omega}^{\exp}$ and $\boldsymbol{z}$ according to $\boldsymbol{h}$, to match the clusters in the two data types.

Identifiability of the Cluster

The label switching problem arises when taking a Bayesian approach to clustering using mixture models

(Diebolt and Robert, 1994, Richardson and Green, 1997). The problems are mainly caused by the nonidentifiability of the components under symmetric priors, which leads to so-called label switching in the MCMC output (Stephens, 2000). We implement the Equivalence Classes Representatives algorithm (Iterative version 1) to relabel the cluster assignment matrix z in MCMC through the R package *label.switching* (Papastamoulis and Iliopoulos, 2010, Rodríguez and Walker, 2014, Papastamoulis, 2016). The algorithm only requires z as the input.

Variable Selection Before Clustering

It may be infeasible in practice to implement MCMC on the whole dataset with all the features. In Sections 5.1, 5.2 and 5.3, we performed variable selection before we implemented our clustering model. Although we may lose information in the clustering step, performing variable selection speeds up the computation and the number of variables becomes more balanced in the two data types after variable selection. The number of variables in chromatin accessibility data is much larger compared with that in gene expression data. Without variable selection, the clustering result may be dominated by the chromatin accessibility data. To select the relevant variables, we first apply simple clustering methods with just one data type (Zamanighomi et al., 2018), and then select cluster-specific genes/regions (Zamanighomi et al., 2018, Love, Huber and Anders, 2014). We select equal number of variables in each cluster to balance the number of variables across the clusters. In Sections 5.1, 5.2 and 5.3, we selected 1000 unlinked features (500 cluster-specific features in scATAC-Seq and 500 cluster-specific features in scRNA-Seq); And for the linked features, we selected 500 cluster-specific features in scATAC-Seq and 500 cluster-specific features in scRNA-Seq.

4. SIMULATION STUDIES

4.1 Simulation Setup

In the simulated data, we assume that all the features are linked. For simplicity of notation, we use j to represent the feature pair in the two data types. The number of samples $n^{\text{acc}} = 100$, and $n^{\text{exp}} = 100$. The number of features $p = 100$. The number of clusters $K = 2$. The followings are the simulation scheme:

Generate $\boldsymbol{\omega}^{\text{acc}}$. 20% of the features are differential across the clusters: $\omega_{kj}^{\text{acc}} = 0.8$ for $k = 1$ and $j = 1, \dots, 10$; $\omega_{kj}^{\text{acc}} = 0.2$ for $k = 2$ and $j = 1, \dots, 10$; $\omega_{kj}^{\text{acc}} = 0.8$ for $k = 2$ and $j = 11, \dots, 20$; $\omega_{kj}^{\text{acc}} = 0.2$ for $k = 1$ and $j = 11, \dots, 20$. For $j = 21, \dots, 100$, we generated $\omega_{1j}^{\text{acc}} = \omega_{2j}^{\text{acc}} \sim \text{Beta}(\mu = 0.5, \phi = 2)$.

Generate $\boldsymbol{\omega}^{\text{exp}}$. We generate $\omega_{kj}^{\text{exp}} \sim \text{Beta}(\mu = \omega_{kj}^{\text{acc}}, \phi = 10)$ for $j = 1, \dots, 20$. We generate $\omega_{1j}^{\text{exp}} = \omega_{2j}^{\text{exp}} \sim \text{Beta}(\mu = \omega_{1j}^{\text{acc}}, \phi = 10)$ for $j = 21, \dots, 100$.

Generate z^{acc} and z^{exp}. The cluster labels are generated with equal probability.

Generate u^{acc}. We generate u_{ij}^{acc} from Bernoulli$(\omega_{kj}^{\text{acc}})$ if $z_{ik}^{\text{acc}} = 1$.

Generate $\tilde{u}^{\text{acc}}$. We generate $\tilde{u}_{ij}^{\text{acc}}$ from Bernoulli(q_i) if $u_{ij}^{\text{acc}} = 1$ and set $\tilde{u}_{ij}^{\text{acc}} = 0$ otherwise. We set $q_i = 0.5$ for $i = 1, \dots, n^{\text{acc}}$.

Generate u^{exp}. We generate u_{lj}^{exp} from Bernoulli$(\omega_{kj}^{\text{exp}})$ if $z_{lk}^{\text{exp}} = 1$.

Generate $\tilde{v}^{\text{exp}}$. We generate $\tilde{v}_{lj}^{\text{exp}}$ from Bernoulli$(\theta_{l1} = q_l \pi_{l1})$ if $u_{lj}^{\text{exp}} = 1$, and from Bernoulli$(\theta_{l0} = q_l \pi_{l0})$ if $u_{lj}^{\text{exp}} = 0$.

Generate x. We generate x_{ij} from $\mathcal{N}(0, 0.8^2)$ if $\tilde{u}_{ij}^{\text{acc}} = 0$, and generate x_{ij} from $\mathcal{N}(2, 0.8^2)$ if $\tilde{u}_{ij}^{\text{acc}} = 1$.

Generate y. We generate y_{lj} from $\mathcal{N}(0, \sigma^2)$ if $\tilde{v}_{lj}^{\text{exp}} = 0$, and generate y_{lj} from $\mathcal{N}(2, \sigma^2)$ if $\tilde{v}_{lj}^{\text{exp}} = 1$.

4.2 Simulation Result

We cluster the samples by assignment to the cluster label with the maximum marginal posterior probability. The simulation results are present in Table 1. To quantify the clustering results, we used three criterions: purity, Rand index and normalized mutual information (NMI). We considered three simulation settings. The trends for the three criterions are very similar over different simulation settings. Compared with the first setting ($\sigma = 0.8$, $\theta_{l1} = 0.6$, $\theta_{l0} = 0.1$), the second setting has a larger variance (larger σ) in the mixture component for the gene expression data, and the third setting has a higher drop-out rate (smaller θ_{l1}). We implemented our clustering model either separately for the two data types ("Model, acc only" and "Model, exp only") or jointly ("Model, acc joint" and "Model, exp joint"). As expected, compared with k-means, the model-based approach achieves better clustering result. In the first simulation setting, combining the information from both the accessibility and gene expression data ("Model, acc joint" and "Model, exp joint") leads to better clustering, compared with implementing the model with the data types separately ("Model, acc only" and "Model, exp only"). In the second and third simulation settings, gene expression data is much noisier and has lower power to separate the cell types: the purity for clustering with gene expression data alone is close to 0.5, which is the expected purity for random assignment. Incorporating the information from the accessibility data, the clustering results for gene expression data improve significantly. Moreover, we did not observe a decrease in the clustering performance for the accessibility data. This is likely due to the fact that all the model parameters are learnt adaptively from the data in the model-based framework.

TABLE 1

Result of simulation study. NMI: normalized mutual information. The simulation results for 100 *independent runs were summarized*

Simulation	Method	Purity	Rand index	NMI
$\sigma = 0.8, \theta_{l1} = 0.6, \theta_{l0} = 0.1$	*k*-means, acc only	0.704 (0.009)	0.596 (0.008)	0.158 (0.013)
	k-means, exp only	0.711 (0.009)	0.599 (0.007)	0.162 (0.011)
	Model, acc only	0.807 (0.007)	0.695 (0.008)	0.318 (0.013)
	Model, acc joint	0.848 (0.005)	0.744 (0.006)	0.403 (0.011)
	Model, exp only	0.778 (0.008)	0.662 (0.008)	0.262 (0.013)
	Model, exp joint	0.836 (0.005)	0.727 (0.006)	0.375 (0.010)
$\sigma = 1.6, \theta_{l1} = 0.6, \theta_{l0} = 0.1$	*k*-means, acc only	0.704 (0.009)	0.596 (0.008)	0.158 (0.013)
	k-means, exp only	0.590 (0.005)	0.515 (0.003)	0.030 (0.004)
	Model, acc only	0.812 (0.006)	0.698 (0.007)	0.323 (0.012)
	Model, acc joint	0.821 (0.006)	0.709 (0.007)	0.342 (0.012)
	Model, exp only	0.596 (0.006)	0.519 (0.003)	0.035 (0.004)
	Model, exp joint	0.705 (0.006)	0.586 (0.004)	0.138 (0.007)
$\sigma = 0.8, \theta_{l1} = 0.4, \theta_{l0} = 0.1$	*k*-means, acc only	0.704 (0.009)	0.596 (0.008)	0.158 (0.013)
	k-means, exp only	0.574 (0.005)	0.508 (0.002)	0.020 (0.003)
	Model, acc only	0.807 (0.006)	0.693 (0.007)	0.314 (0.012)
	Model, acc joint	0.811 (0.006)	0.699 (0.007)	0.325 (0.013)
	Model, exp only	0.588 (0.006)	0.516 (0.003)	0.031 (0.004)
	Model, exp joint	0.696 (0.006)	0.580 (0.005)	0.131 (0.008)

5. APPLICATION TO REAL DATA

5.1 *In Silico* Mixture of Single Cells

Next, we evaluate our method by jointly clustering scATAC-Seq and scRNA-Seq data. We took 233 K562 and 91 HL60 scATAC-Seq samples (Buenrostro et al., 2015b), and 42 K562 and 54 HL-60 deeply sequenced scRNA-seq samples (Pollen et al., 2014). We perform *in silico* mixture of the single cells and use the true cell label as benchmark to evaluate the performance of the clustering methods. We first see if we can correctly match the two cell lines in scATAC-Seq and scRNA-Seq data. We implemented the model and use 0.5 as the threshold for posterior probability as there are two classes. The two cell types are perfectly separated and correctly matched in the two data types (Figure 2(a)).

We then downsampled the scRNA-Seq samples, mimicking single-cell experiments with shallow sequencing depth. We implemented the drop-out model (Pierson and Yau, 2015). Denote h_{ij} the number of reads for gene j in cell i, and let $\mu_{ij} = \log_2(h_{ij} + 1)$. The drop out probability (Pierson and Yau, 2015) $p_{ij} = \exp^{-\lambda\mu_{ij}^2}$. The downsampled read count $h'_{ij} = 0$ if gene j is dropped out in sample i, and otherwise $h'_{ij} = h_{ij}$. The parameter λ controls the drop-out rate. In the downsampling procedure, we chose different λs for the two cell types, so that the distributions for the number of expressed genes are more similar. We set $\lambda = 0.0025$ for K562 cells, and $\lambda = 0.00182$ for HL60 cells. For scATAC-Seq samples, we implemented a similar downsampling scheme with $\lambda = 0.025$ for both cell types as the number of zero entries are similar.

After downsampling, it is hard to separate K562 and HL60 using scRNA-Seq or scATAC-Seq alone (Figure 2(b)). Using our model-based approach, the two cell types are separated reasonably well and the cell types are correctly matched in the two data types (Figure 2(c)). The cluster purities for scATAC-Seq and scRNA-seq samples are 82.5% (259/314) and 85.4% (82/96). One advantage of our clustering method is that the model-based framework enables statistical inference of the cluster assignment. The distribution for the posterior probability of cluster assignment for scATAC-Seq data is shown in (Figure 2(d)). If we use a more stringent threshold (0.95) for the posterior probability of cluster assignment and do not cluster those cells with higher uncertainty, the clustering purity for scATAC-Seq samples improves to 86.1% (237/275).

5.2 Retinoic Acid Induction of Mouse Embryonic Stem Cells

This data set is from a recent experiment in our lab, where mouse embryonic stem cells (mESC) were treated with Retinoic Acid (RA) and the cells were harvest at induction day 4. Both scATAC-Seq (420 cells) and scRNA-Seq (464 cells) data are available under the GEO database with accession numbers GSE115968 and GSE115970. We compared the result of our model-based analysis to that of our previous analysis using coupleNMF (Duren et al., 2018). While coupleNMF connects the two data types based on prediction models trained from bulk data, our model-based approach does not rely on pre-trained prediction models. In general, the clustering results are a

scATAC-Seq. Model

	K562	HL60
Cluster 1	223	0
Cluster 2	0	91

scRNA-Seq. Model

	K562	HL60
Cluster 1	42	0
Cluster 2	0	54

b

scATAC-Seq downsample. *k*-means

	K562	HL60
Cluster 1	66	91
Cluster 2	25	132

scRNA-Seq downsample. *k*-means

	K562	HL60
Cluster 1	52	37
Cluster 2	2	5

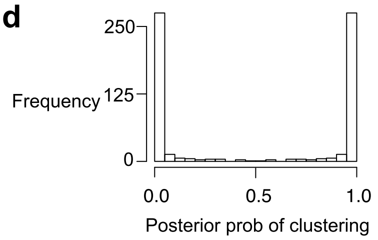
c

scATAC-Seq downsample. Model

	K562	HL60
Cluster 1	69	33
Cluster 2	22	190

scRNA-Seq downsample. Model

	K562	HL60
Cluster 1	41	1
Cluster 2	13	41



scATAC-Seq downsample. Model(0.95)

	K562	HL60
Cluster 1	61	22
Cluster 2	16	176

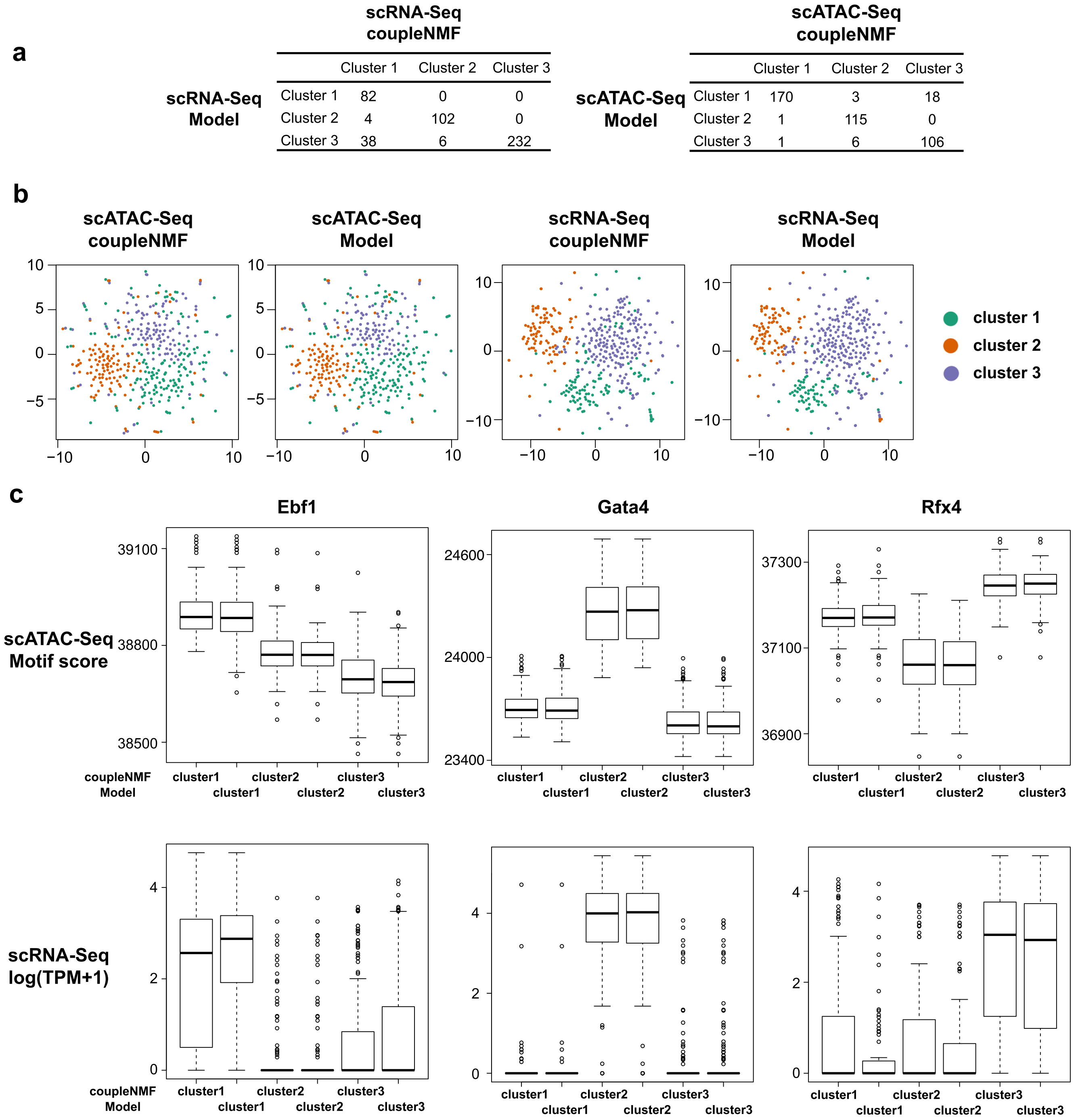
FIG. 2. *Results for in silico mixture of single cells*. (*a*) *Clustering table for model-based approach*. (*b*) *Clustering table for k-means clustering, downsampled data*. (*c*) *Clustering table for model-based clustering, here we cluster by maximum marginal posterior probability of clustering assignment*. (*d*) *Distribution for the posterior probability of clustering assignment* (*left*). *Clustering table for model-based approach* (*right*), *here we use* 0.95 *as the threshold*: *we only cluster cells that have maximum marginal posterior probability greater than* 0.95.

quite comparable between the two methods. For scATAC-Seq data, the clustering result is quite similar, with only 29 among 420 cells clustered differently. For scRNA-Seq data, the clustering result is more different, with 48 cells among 464 cells clustered differently. Visualization of the clustering result is presented in Figure 3(b). The visualization result for coupleNMF is slightly different from the plots in (Duren et al., 2018), due to the randomness in t-SNE algorithm. The major difference is the cluster assignment for clusters 1 and 3, where 18 scATAC-Seq cells assigned to cluster 3 by coupleNMF are assigned to cluster 1 by our model, and 38 scRNA-Seq cells assigned to cluster 1 by coupleNMF are assigned to cluster 3 by our model. For a more detailed comparison, we compared the clustering result via transcriptional factor (TF) motif score in scATAC-Seq data and TF gene expression in scRNA-Seq data. The genes Ebf1, Gata4 and Rfx4 are important TFs for the stem cell differentiation process and were identified as signatures for the clusters in Duren et al. (2018). Cluster 1 cells tend to have higher TF motif score and gene expression level for Ebf1, lower TF motif score and gene expression level for Rfx4, and cluster 3 cells are the opposite. The 18 cells that are clustered differently in scATAC-Seq data tend to have intermediate levels of motif score for Ebf1 and Rfx4. For scRNA-Seq data, separation of the clusters seems slightly better in our model, based on the expression levels of Ebf1 and Rfx4.

5.3 Demonstrating the Potential for Batch Effect Correction with a Bulk Data Example

The purpose of this example is to demonstrate that by combining chromatin accessibility and gene expression data, we may be able to correct for technical variation, which is commonly observed in single-cell data (Hicks et al., 2018). Although our modeling framework is designed for single-cell data, it can be applied to bulk data as well.

We tested our method with a collection of three cell types (fibroblasts, epithelial and endothelial cells) from the ENCODE Project Consortium (Dunham et al., 2012, Sloan et al., 2015) and the Roadmap Project (Kundaje et al., 2015). These are bulk DNase-Seq and bulk RNA-Seq samples. The three cell types are well separately using DNase-Seq data (Figure 4(a)). However, using RNA-Seq data alone, it seems hard to separate the three cell types (Figure 4(a) and (b)). The observation that RNA-Seq samples are hard to separate is likely due to technical reasons as the samples are processed on different batches from different laboratories. However, we did not observe the trend of the samples to cluster by laboratories (Supplementary Material Figure S2 (Lin et al., 2020)). Combining DNase-Seq data, we are able to achieve a good separation of the three cell types in RNA-Seq data (Figure 4(c)). To obtain the clustering table in Figure 4(c), we implemented our model in a stepwise manner: (1) we first cluster DNase-Seq samples with k-means (the clustering re-



scRNA-Seq Model \ scRNA-Seq coupleNMF	Cluster 1	Cluster 2	Cluster 3
Cluster 1	82	0	0
Cluster 2	4	102	0
Cluster 3	38	6	232

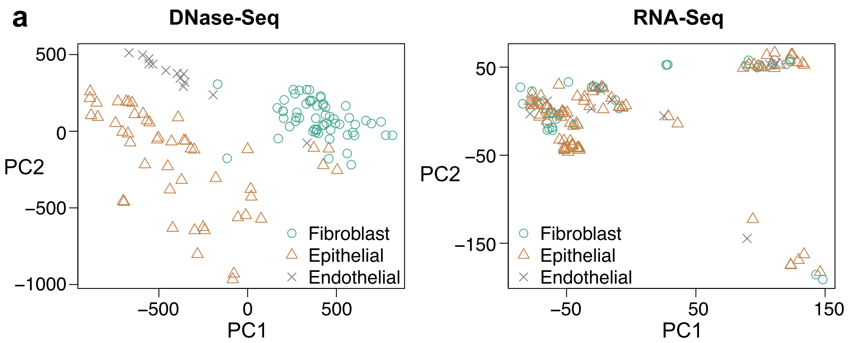
scATAC-Seq Model \ scATAC-Seq coupleNMF	Cluster 1	Cluster 2	Cluster 3
Cluster 1	170	3	18
Cluster 2	1	115	0
Cluster 3	1	6	106

FIG. 3. *Results for RA induction of mESC.* (*a*) *comparison of the clustering table, our model-based approach vs. coupleNMF.* (*b*) *t-SNE plot for visualization of the clustering result.* (*c*) *Boxplots for the motif score and expression level for the transcriptional factors Ebf*1, *Gata*4 *and Rfx*4.

sult is similar when we implement our model on DNase-Seq data alone), (2) then we estimate the model parameters using DNase-Seq data alone with the cluster labels fixed to the result in the first step, (3) and finally we cluster RNA-Seq samples via the model with the parameter ω_r fixed to the posterior mean in the second step. The benefit of the stepwise approach is mostly computational, as the chain converges much faster with some parameters fixed. When one data type is strong at separating the cell types, we do not lose much information by implementing the stepwise approach. The stepwise approach also suggest an option to combine our model-based approach with other clustering methods.

6. CONCLUSIONS

Unsupervised methods, such as dimension reduction and clustering are essential to the analysis of single-cell genomic data as the cell types need to be inferred. Model-based clustering methods have the advantage of quantifying the uncertainty in the cluster assignments and they are under-explored in the area of single-cell genomics. Combining the information across different types of genomic



b

RNA-Seq samples, *k*-means

	Fibroblast	Epithelial	Endothelial
Cluster 1	12	18	2
Cluster 2	26	58	10
Cluster 3	2	6	1

c

DNase-Seq samples, *k*-means

	Fibroblast	Epithelial	Endothelial
Cluster 1	57	4	1
Cluster 2	2	45	0
Cluster 3	0	0	12

RNA-Seq samples, model

	Fibroblast	Epithelial	Endothelial
Cluster 1	38	1	0
Cluster 2	2	79	1
Cluster 3	0	2	12

FIG. 4. *Results for the three cell types bulk data example*. (*a*) *Principal Component Analysis* (*PCA*) *for fibroblast*, *epithelial and endothelial samples*, *using* $\log 2(\textit{fold change} + 1)$ *of accessible regions in DNase-Seq data* (*left*) *and* $\log 2(\mathrm{FPKM} + 1)$ *of all genes in RNA-Seq data* (*right*). (*b*) *Clustering table using k-means clustering algorithm for the RNA-Seq samples*. (*c*) *Clustering table* (*left*) *using k-means clustering algorithm for the DNase-Seq* (*fibroblast*, *epithelial and endothelial*) *samples*. *Clustering table* (*right*) *using the model-based approach*, *where all RNA-Seq samples* (*fibroblast*, *epithelial and endothelial*) *are assigned to the DNase-Seq clusters with the highest posterior probability*.

features can provide rich biological insight and can lead to better separation of the cell types. We have developed a model-based approach that is specifically designed for single-cell genomic data and can jointly cluster single-cell chromatin accessibility and single-cell gene expression data. Our modeling framework is general and it can be extended to other types of single-cell genomic data, such as single-cell methylation data. The R package is available at https://github.com/cuhklinlab/scACE.

ACKNOWLEDGEMENTS

The authors would like to thank members in the Wong lab for their help in the course of this project, particularly Zhana Duren, Wanwen Zeng and Xi Chen for helpful discussions. This work has been supported by the Chinese University of Hong Kong direct grant 2018/2019 No. 4053360, and Hong Kong Research Grant Council Grant ECS 24301419. This work was partially supported by NIH Grant P50 HG007735 and NSF Grant DMS-1811920.

SUPPLEMENTARY MATERIAL

Supplement to "Model-Based Approach to the Joint Analysis of Single-Cell Data on Chromatin Accessibility and Gene Expression" (DOI: 10.1214/19-STS714SUPP; .pdf). Supplementary information.

REFERENCES

BACHER, R. and KENDZIORSKI, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol*. **17** 63. https://doi.org/10.1186/s13059-016-0927-y

BENAGLIA, T., CHAUVEAU, D., HUNTER, D. R. and YOUNG, D. (2009). mixtools: An R package for analyzing finite mixture models. *J. Stat. Softw.* **32** 1–29.

BUENROSTRO, J. D., GIRESI, P. G., ZABA, L. C., CHANG, H. Y. and GREENLEAF, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10** 1213.

BUENROSTRO, J. D., WU, B., CHANG, H. Y. and GREENLEAF, W. J. (2015a). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109** 21–29.

BUENROSTRO, J. D., WU, B., LITZENBURGER, U. M., RUFF, D., GONZALES, M. L., SNYDER, M. P., CHANG, H. Y. and GREENLEAF, W. J. (2015b). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523** 486–490.

CORCES, M. R., BUENROSTRO, J. D., WU, B., GREENSIDE, P. G., CHAN, S. M., KOENIG, J. L., SNYDER, M. P., PRITCHARD, J. K., KUNDAJE, A. et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48** 1193–1203.

CUSANOVICH, D. A., DAZA, R., ADEY, A., PLINER, H. A., CHRISTIANSEN, L., GUNDERSON, K. L., STEEMERS, F. J., TRAPNELL, C. and SHENDURE, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348** 910–914.

DIEBOLT, J. and ROBERT, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* **56** 363–375. MR1281940

DUNHAM, I., KUNDAJE, A., ALDRED, S. et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74. https://doi.org/10.1038/nature11247

DUREN, Z., CHEN, X., JIANG, R., WANG, Y. and WONG, W. H. (2017). Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl. Acad. Sci. USA* **114** E4914–E4923. https://doi.org/10.1073/pnas.1704553114

DUREN, Z., CHEN, X., ZAMANIGHOMI, M., ZENG, W., SATPATHY, A., CHANG, H., WANG, Y. and WONG, W. H. (2018). Integrative analysis of single cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl. Acad. Sci. USA* **115** 7723–7728. https://doi.org/10.1073/pnas.1805681115

GRÜN, D., MURARO, M. J., BOISSET, J.-C., WIEBRANDS, K., LYUBIMOVA, A., DHARMADHIKARI, G., VAN DEN BORN, M., VAN ES, J., JANSEN, E. et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19** 266–277. https://doi.org/10.1016/j.stem.2016.05.010

HICKS, S. C., TOWNES, F. W., TENG, M. and IRIZARRY, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19** 562–578. MR3867412 https://doi.org/10.1093/biostatistics/kxx053

KHARCHENKO, P. V., SILBERSTEIN, L. and SCADDEN, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11** 740–742. https://doi.org/10.1038/nmeth.2967

KISELEV, V. Y., KIRSCHNER, K., SCHAUB, M. T., ANDREWS, T., YIU, A., CHANDRA, T., NATARAJAN, K. N., REIK, W., BARAHONA, M. et al. (2017). SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14** 483.

KUNDAJE, A., MEULEMAN, W., ERNST, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* **518** 317–330. https://doi.org/10.1038/nature14248

LAKE, B. B., CHEN, S., SOS, B. C., FAN, J., KAESER, G. E., YUNG, Y. C., DUONG, T. E., GAO, D., CHUN, J. et al. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36** 70–80.

LIN, P., TROUP, M. and HO, J. W. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18** 59.

LIN, Z., ZAMANIGHOMI, M., DALEY, T., MA, S. and WONG, W. H. (2020). Supplement to "Model-Based Approach to the Joint Analysis of Single-Cell Data on Chromatin Accessibility and Gene Expression." https://doi.org/10.1214/19-STS714SUPP.

LIU, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.* **89** 958–966. MR1294740

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40. MR1279653 https://doi.org/10.1093/biomet/81.1.27

LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15** 550. https://doi.org/10.1186/s13059-014-0550-8

OLKIN, I. and RUBIN, H. (1964). Multivariate beta distributions and independence properties of the Wishart distribution. *Ann. Math. Stat.* **35** 261–269. MR0160297 https://doi.org/10.1214/aoms/1177703748

PAPASTAMOULIS, P. (2016). Label.switching: An R package for dealing with the label switching problem in MCMC outputs. *J. Stat. Softw.* **69** 1–24. https://doi.org/10.18637/jss.v069.c01.

PAPASTAMOULIS, P. and ILIOPOULOS, G. (2010). An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. *J. Comput. Graph. Statist.* **19** 313–331. MR2758306 https://doi.org/10.1198/jcgs.2010.09008

PIERSON, E. and YAU, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16** 241. https://doi.org/10.1186/s13059-015-0805-z

POLLEN, A. A., NOWAKOWSKI, T. J., SHUGA, J., WANG, X., LEYRAT, A. A. et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32** 1053–1058.

RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213 https://doi.org/10.1111/1467-9868.00095

RODRÍGUEZ, C. E. and WALKER, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *J. Comput. Graph. Statist.* **23** 25–45. MR3173759 https://doi.org/10.1080/10618600.2012.735624

ROTEM, A., RAM, O., SHORESH, N., SPERLING, R. A., GOREN, A., WEITZ, D. A. and BERNSTEIN, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33** 1165–1172. https://doi.org/10.1038/nbt.3383

ROZENBLATT-ROSEN, O., STUBBINGTON, M. J., REGEV, A. and TEICHMANN, S. A. (2017). The human cell atlas: From vision to reality. *Nat. News* **550** 451.

SLOAN, C. A., CHAN, E. T., DAVIDSON, J. M., MALLADI, V. S., STRATTAN, J. S., HITZ, B. C. and CHERRY, J. M. (2015). ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44** D726–D732. https://doi.org/10.1093/nar/gkv1160

SMALLWOOD, S. A., LEE, H. J., ANGERMUELLER, C., KRUEGER, F., SAADEH, H., PEAT, J., ANDREWS, S. R., STEGLE, O., REIK, W. et al. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11** 817.

STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 795–809. MR1796293 https://doi.org/10.1111/1467-9868.00265

SUN, Z., WANG, T., DENG, K., WANG, X.-F., LAFYATIS, R., DING, Y., HU, M. and CHEN, W. (2017). DIMM-SC: A Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics* **34** 139–146.

WANG, B., ZHU, J., PIERSON, E., RAMAZZOTTI, D. and BATZOGLOU, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14** 414–416. https://doi.org/10.1038/nmeth.4207

XU, C. and SU, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31** 1974–1980.

YANG, Y., HUH, R., CULPEPPER, H. W., LIN, Y., LOVE, M. I. and LI, Y. (2018). SAFE-clustering: Single-cell Aggregated (From Ensemble) clustering for single-cell RNA-seq data. *Bioinformatics*.

YAU, C. et al. (2016). PcaReduce: Hierarchical clustering of single cell transcriptional profiles. *BMC Bioinform.* **17** 140.

ZAMANIGHOMI, M., LIN, Z., DALEY, T., CHEN, X., DUREN, Z., SCHEP, A., GREENLEAF, W. J. and WONG, W. H. (2018). Unsupervised clustering and epigenetic classification of single cells. *Nat. Commun.* **9** 2410.

ZANG, C., WANG, T., DENG, K., LI, B., QIN, Q., XIAO, T., ZHANG, S., MEYER, C. A., HE, H. H. et al. (2016). High-dimensional genomic data bias correction and data integration using MANCIE. *Nat. Commun.* **7** 11305.

ZHU, L., LEI, J., DEVLIN, B. and ROEDER, K. (2018). A unified statistical framework for single cell and bulk RNA sequencing data. *Ann. Appl. Stat.* **12** 609–632. MR3773407 https://doi.org/10.1214/17-AOAS1110

ZHU, L., LEI, J., KLEI, L., DEVLIN, B. and ROEDER, K. (2019). Semisoft clustering of single-cell data. *Proc. Natl. Acad. Sci. USA* **116** 466–471. MR3904692 https://doi.org/10.1073/pnas.1817715116