bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

scAMACE: Model-based approach to the joint analysis of single-cell data on chromatin accessibility, gene expression and methylation

Jiaxuan Wangwu[1], Zexuan Sun[1], and Zhixiang Lin [*1]

[1]Department of Statistics, The Chinese University of Hong Kong, HK, SAR China

March 30, 2021

Abstract

The advancement in technologies and the growth of available single-cell datasets motivate integrative analysis of multiple single-cell genomic datasets. Integrative analysis of multimodal single-cell datasets combines complementary information offered by single-omic datasets and can offer deeper insights on complex biological process. Clustering methods that identify the unknown cell types are among the first few steps in the analysis of single-cell datasets, and they are important for downstream analysis built upon the identified cell types. We propose scAMACE for the integrative analysis and clustering of single-cell data on chromatin accessibility, gene expression and methylation. We demonstrate that cell types are better identified and characterized through analyzing the three data types jointly. We develop an efficient expectation-maximization (EM) algorithm to perform statistical inference, and evaluate our methods on both simulation study and real data applications. We also provide the GPU implementation of scAMACE, making it scalable to large datasets. The software and datasets are available at https://github.com/cuhklinlab/scAMACE_py (pythom implementation) and https://github.com/cuhklinlab/scAMACE (R implementation).

Keywords: Bayesian statistics, Clustering, Data integration, Multi-omics, Single-cell genomic

1 Introduction

Recent developments in single-cell technologies enable multiple measurements of different genomic features (Lahnemann *et al.*, 2020). Sequencing technologies include single-cell RNA sequencing (scRNA-seq)

[*]Corresponding author: zhixianglin@cuhk.edu.hk

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which measures transcription, single-cell ATAC sequencing (scATAC-seq) and the assay based on combinatorial indexing (sci-ATAC-seq) (Cusanovich *et al.*, 2018b) that measure chromatin accessibility, and single-nucleus methylcytosine sequencing (snmC-seq) (Luo *et al.*, 2017) which meansures methylome at the single cell resolution. High technical variation is presented in single-cell datasets due to the limited amount of genomic materials and the experimental procedures to amplify the signals (Lahnemann *et al.*, 2020).

Because cell types are usually unknown beforehand, clustering methods are needed to identify the cell types. Majority of existing clustering algorithms only take one single dataset as input. Beside the widely used K-Means clustering algorithm, hierarchical clustering (Ward, 1963) forms hierarchical groups of mutually exclusive subsets on the basis of their similarity with respect to specified characteristics by considering the union of all possible $\frac{k(k-1)}{2}$ pairs and accepting the union with which an optimal value of the objective function is associated. Spectral Clustering (Ng *et al.*, 2001) uses the top k eigenvectors of a matrix derived from the distance between points simultaneously for clustering. Several algorithms are developed specifically for scRNA-seq data. SC3 (Kiselev *et al.*, 2017) combines multiple clustering outcomes through a consensus approach. SIMLR (Wang *et al.*, 2017) learns a distance metric by multiple kernels and clusters with affinity propagation. CIDR (Lin *et al.*, 2017) imputes the gene expression profiles, calculates the dissimilarity based on the imputed gene expression profiles for every pair of single cells, performs principal coordinate analysis using the dissimilarity matrix, and finally performs clustering using the first few principal coordinates. SOUP (Zhu *et al.*, 2019) semi-softly classifies both pure and intermediate cell types: it first identifies the set of pure cells by special block structure and estimates a membership matrix, then estimates soft membership for the other cells. For the analysis of single-cell chromatin accessibility data, scABC (Zamanighomi *et al.*, 2018) first weights cells and applies weighted K-medoids clustering, then calculate landmarks for each cluster, and finally clusters the cells by assignment to the closest landmark based on Spearman correlation. Cusanovich (Cusanovich *et al.*, 2018a) makes use of singular value decomposition on TF-IDF transformed matrix and density peak clustering algorithm. cisTopic (Bravo González-Blas *et al.*, 2019) uses latent Dirichlet allocation with a collapsed Gibbs sampler to iteratively optimize the region-topic distribution and the topic-cell distribution. SCALE (Xiong *et al.*, 2019) combines the variational autoencoder framework with the Gaussian Mixture Model which extracts latent features that characterize the distributions of input scATAC-seq data, and then uses the latent features to cluster cell mixtures into subpopulations. Clustering methods are also developed for single-cell methylation data. BPRMeth (Kapourani and Sanguinetti, 2016) uses probabilistic machine learning to extract higher order features across a defined region and to cluster promoter-proximal regions by Binomial distributed probit regression (BPR) and mixture modeling. PDclust (Hui *et al.*, 2018) leverages the methylation state of individual CpGs to obtain pairwise dissimilarity (PD) values, and calculates Euclidean distances between each pair of cells using their PD values and performed hierarchical clustering. Melissa (Kapourani and Sanguinetti, 2019) implements a Bayesian hierarchical model that jointly learns the methylation profiles of genomic regions of interest and clusters cells based on their genome-wide methylation patterns. pCSM (Yin *et al.*, 2019) implements a semi-reference-free procedure to perform virtual methylome dissection using the nonnegative matrix factorization algorithm. It first determines putative cell-type-specific methylated loci and then clusters the loci into groups based on their correlations in methylation profiles.

Studies based on single-omic data provide only a partial landscape of the entire cellular heterogeneity (Ma *et al.*, 2020). High technical noise and the growth of available datasets measuring different genomic

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

features encourage integrative analysis (Lahnemann *et al.*, 2020). By combining complementary information from multiple datasets, the cell types may be better seperated and characterized (Corces *et al.*, 2016; Duren *et al.*, 2017). The integrative analysis of gene expression and chromatin activity may better define cell types and lineages, especially in complex tissues (Duren *et al.*, 2018). Seurat V3 (Stuart *et al.*, 2019) uses Canonical Correlation Analysis (CCA) to reduce the dimension of the datasets. It identifies the pairwise correspondences of single cells across datasets, termed 'anchors', and then transfers labels from a reference dataset onto a query dataset. coupleNMF (Duren *et al.*, 2018) is based on the coupling of two non-negative matrix factorizations, where a 'soft' clustering can be obtained following the matrix factorizations. It enables integrative analysis of scRNA-seq and scATAC-seq data. LIGER (Welch *et al.*, 2019) integrates multimodal datasets via integrative non-negative matrix factorization (iNMF) to learn a low-dimensional space defined by dataset-specific factors and shared factors across datasets, and then build a neighborhood graph based on the shared factors to identify joint clusters by performing community detection on this graph. scACE (Lin *et al.*, 2020) is a model-based approach that jointly analyzes single-cell chromatin accessibility and scRNA-Seq data, and it quantifies the uncertainty of cluster assignments. MAESTRO (Wang *et al.*, 2020) integrates scRNA-seq and scATAC-seq data from multiple platforms. It also provides comprehensive functions for pre-processing, alignment, quality control, and quantification of expression and accessibility. coupleCoC (Zeng *et al.*, 2020) performs co-clustering of the cells and the features simultaneously in the source data and the target data, and it also matches the cell clusters between the source data and the target data through minimizing the distribution divergence. scMC (Zhang and Nie, 2021) integrates multiple scRNA-Seq datasets or multiple scATAC-Seq datasets, where it learns biological variation via variance analysis to subtract technical variation inferred in an unsupervised manner. The three data types, including gene expression, chromatin accessibility and methylation, have distinct characteristics and complex relationships with each other. The aforementioned methods for integrative analysis are not designed to integrate all three data types. Moreover, these methods (except scACE) do not provide statistical inference on the cluster assignments, which may be important when there are cells at the intermediate stages during development.

In this work, we extend scACE (Lin *et al.*, 2020) to scAMACE (integrative Analysis of single-cell Methylation, chromatin ACcessibility, and gene Expression). scAMACE considers the biological and technical variabilities when integrating multiple data types, and it can provide statistical inference on the assignment of clusters. We reason that by combining complementary biological information from multiple data types, better cell type seperation can be achieved. We present our model in Section 2, and statistical inference using the Expectation-Maximization (EM) algorithm in Section 3. Simulation study and real data applications are presented in Sections 4 and 5, respectively. The conclusion is presented in Section 6.

2 Methods

An overview of scAMACE is presented in Fig. 1.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

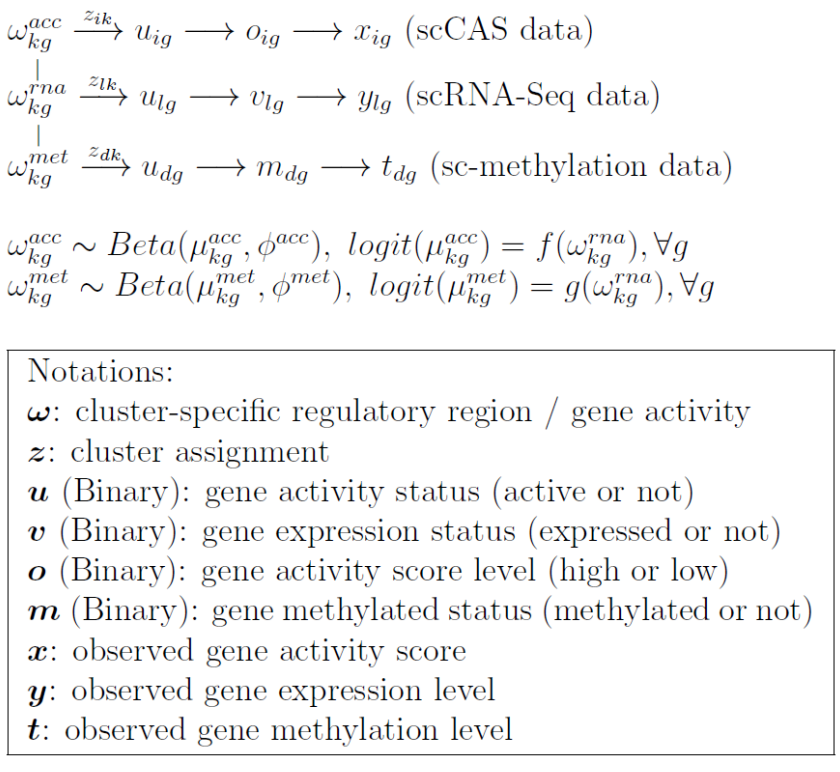


Figure 1: Graphical representation of scAMACE.

2.1 Model for scRNA-Seq data

The model specification for scRNA-Seq data is as the following.

$$
\begin{aligned}
&\omega_{kg}^{rna} \xrightarrow{z_{lk}} u_{lg} \longrightarrow v_{lg} \longrightarrow y_{lg}\ \forall g,\\
&P(z_{lk} = 1) = \psi_k^{rna},\\
&u_{lg} \mid z_{lk} = 1 \sim Bernoulli(\omega_{kg}^{rna}),\\
&v_{lg} \mid u_{lg} \sim u_{lg} Bernoulli(\pi_{l1}) + (1 - u_{lg}) Bernoulli(\pi_{l0}),\\
&\pi_{l0} \sim Beta(\alpha = 1, \beta = 1), \pi_{l1} \sim \mathbb{1}(\pi_{l1} \geq \pi_{l0}) Beta(\alpha = 1, \beta = 1),\\
&p(y_{lg} \mid v_{lg}) = v_{lg} g_1(y_{lg}) + (1 - v_{lg}) g_0(y_{lg}).
\end{aligned}
$$

We assume that there are K cell clusters in total, the random variable z_{lk} denotes whether cell l belongs to cluster $k \in \{1, ..., K\}$, 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, we assume that $\pi_{l1} \geq \pi_{l0}$ and the prior distributions of π_{l1} and π_{l0} are assumed to be flat.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Let y_{lg} denote the observed gene expression for gene g in cell l (after normalization to account for sequencing depth and gene length), and we assume 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}.

2.2 Model for single-cell chromatin accessibility (scCAS) data

The model specification for scCAS data is as the following.

$$
\begin{aligned}
&\omega_{kg}^{acc} \xrightarrow{z_{ik}} u_{ig} \longrightarrow o_{ig} \longrightarrow x_{ig} \ \forall g,\\
&P(z_{ik}=1) = \psi_k^{acc},\\
&u_{ig} \mid z_{ik}=1 \sim Bernoulli(\omega_{kg}^{acc}),\\
&o_{ig} \mid u_{ig} \sim u_{ig} Bernoulli(\pi_{i1}) + (1-u_{ig}) Bernoulli(\pi_{i0}),\\
&\pi_{i1} \sim Beta(\alpha_{acc}=1, \beta_{acc}=1), \text{set } \pi_{i0}=0,\\
&p(x_{ig} \mid o_{ig}) = o_{ig} f_1(x_{ig}) + (1-o_{ig}) f_0(x_{ig}),\\
&\omega_{kg}^{acc} | \omega_{kg}^{rna} \sim Beta(\mu_{kg}^{acc}, \phi^{acc}),\ logit(\mu_{kg}^{acc}) = f(\omega_{kg}^{rna}).
\end{aligned}
$$

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. We use a different notation i to represent that the cells in the scCAS data are different from the cells in the scRNA-Seq data.

x_{ig} denotes the observed gene score for gene g in cell i. The gene score summarizes the accessibility of the regions around the gene body (Cusanovich *et al.*, 2018a). We model it 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. We model this positive relationship by the distribution $o_{ig}|u_{ig}$. 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}. We assume that $\pi_{i1} \geq \pi_{i0}$ to represent the positive relationship. In practice, we found that fixing $\pi_{i0}=0$ leads to good real data performance, and we set $\pi_{i0}=0$ by default. The prior distribution $\pi_{i1} \sim Beta(\alpha=1, \beta=1)$. In real data example 1, the observed data is promoter accessibility and we use the same model as that for gene score.

We assume that ω_{kg}^{acc} follows 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: $logit(\mu_{kg}^{acc}) = f(\omega_{kg}^{rna})$. Details on the specification of $f(\cdot)$ are presented in Section 2.6.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.3 Model for single-cell methylation data

The model specification for sc-methylation data is as the following.

$$
\begin{aligned}
&\omega_{kg}^{met} \xrightarrow{z_{dk}} u_{dg} \longrightarrow m_{dg} \longrightarrow t_{dg} \ \forall g,\\
&P(z_{dk}=1)=\psi_k^{met},\\
&u_{dg} \mid z_{dk}=1 \sim Bernoulli(\omega_{kg}^{met}),\\
&m_{dg} \mid u_{dg} \sim u_{dg} Bernoulli(\pi_{d1}) + (1-u_{dg}) Bernoulli(\pi_{d0}),\\
&\pi_{d0} \sim Beta(\alpha=1,\beta=1), \pi_{d1} \sim \mathbb{1}(\pi_{d1} \leq \pi_{d0}) Beta(\alpha=1,\beta=1),\\
&p(t_{dg} \mid m_{dg}) = m_{dg} h_1(t_{dg}) + (1-m_{dg}) h_0(t_{dg}),\\
&\omega_{kg}^{met} | \omega_{kg}^{rna} \sim Beta(\mu_{kg}^{met}, \phi^{met}),\ logit(\mu_{kg}^{met}) = g(\omega_{kg}^{rna}).
\end{aligned}
$$

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. We use a different notation d to represent that the cells in the sc-methylation data are different from the cells in the scRNA-Seq data.

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 we model this negative relationship with the model $m_{dg} \mid u_{dg}$: when the gene g is active in cell d ($u_{dg}=1$), it is less likely to be methylated ($m_{dg}=1$), as we assume that $\pi_{d1} \leq \pi_{d0}$.

t_{dg} denotes the observed methylation level for gene g in cell d, and we assume that $t_{dg} \mid m_{dg}$ follows a mixture distribution, where $h_1(.)$ and $h_0(.)$ are density functions conditional on m_{dg}. The technologies/features differ for the two real data applications to be presented: promoter methylation for the gene (Pott, 2017), and gene body methylation at non-CG sites (Luo *et al.*, 2017).

Similar to scCAS data, we connect μ_{kg}^{met}, which is the mean of ω_{kg}^{met}, and ω_{kg}^{rna} through the logit function: $logit(\mu_{kg}^{met}) = g(\omega_{kg}^{rna})$. Details on specification of $g(\cdot)$ are presented in Section 2.6.

2.4 More on model specification

Methylation and chromatin accessibility regulate gene expression biologically. Our model is specified in the reverse order, so gene expression plays a central role. This is because scRNA-Seq data is usually less noisy compared with scCAS data and sc-methylation data, the model specified this way will improve the clustering performance of scCAS data and sc-methylation data, without sacrificing much the clustering performance of scRNA-Seq data.

2.5 Prior specifications

We assume the following priors for $\psi^{acc}, \psi^{rna}, \psi^{met}, \omega_{kg}^{rna}$.

$$
\begin{aligned}
&\boldsymbol{\psi}^{acc} \sim Dir(2,...,2), \boldsymbol{\psi}^{rna} \sim Dir(2,...,2), \boldsymbol{\psi}^{met} \sim Dir(2,...,2),\\
&\omega_{kg}^{rna} \sim Beta\ (\alpha_1=2,\beta_1=2)
\end{aligned}
$$

The prior specification $Beta\ (\alpha=2,\beta=2)$ improves the stability of the EM algorithm in Section 3 over uniform distritbution.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.6 Determination of $f(\omega_{kg}^{rna})$, $g(\omega_{kg}^{rna})$, ϕ^{acc} and ϕ^{met}

We assume that $f(\omega_{kg}^{rna}) = \eta + \gamma\omega_{kg}^{rna} + \tau(\omega_{kg}^{rna})^2$ and $g(\omega_{kg}^{rna}) = \delta + \theta\omega_{kg}^{rna}$. The parameters $\{\eta, \gamma, \tau, \delta, \theta, \phi^{acc}, \phi^{met}\}$ are estimated empirically from the datasets. We first set $K = 1$ and use the model to estimate ω_{kg}^{rna}, ω_{kg}^{acc} and ω_{kg}^{met} separately without considering the links on $\boldsymbol{\omega}$ across the three datasets, and then fix ω_{kg}^{rna}, ω_{kg}^{acc} and ω_{kg}^{met} to estimate $\{\eta, \gamma, \tau, \delta, \theta, \phi^{acc}, \phi^{met}\}$ by beta regression (Silvia and Francisco, 2004). The rationale for fixing $K = 1$ to estimating the parameters in $f(.)$ and $g(.)$ is that the majority of the features may not change much across the cell types. We fix $\{\hat{\eta}, \hat{\gamma}, \hat{\tau}, \hat{\delta}, \hat{\theta}, \hat{\phi}^{acc}, \hat{\phi}^{met}\}$ when implementing the EM algorithm in Section 3. Estimating $\{\eta, \gamma, \tau, \delta, \theta, \phi^{acc}, \phi^{met}\}$ separately from the EM algorithm improves computational efficiency and avoids problematic local modes. Empirical distributions of ω_{kg}^{acc} v.s. ω_{kg}^{rna} and ω_{kg}^{met} v.s. ω_{kg}^{rna} for the two real data applications are presented in Supplementary Materials Figures S.4 and S.5.

2.7 The mixture components

For scCAS data, we apply $f_1(x) = 0, f_0(x) = 1$ if $x = 0$ and $f_1(x) = 1, f_0(x) = 0$ if $x > 0$, due to the sparsity of the data matrix.

For scRNA-Seq data, we first normalize read counts to TPM (transcripts per million) or FPKM (fragments per kilobase of exon model per million reads mapped), then fit a two-component gamma mixture model for the nonzero entries, through pooling ln(TPM+1) or ln(FPKM+1) over all the samples, and then the remaining zero entries are merged with the mixture component that has a smaller mean.

sc-methylation data represents the proportion of methylated sites within a given genomic interval, where the entries in the data matrix take values between 0 and 1. For the methylation data in each cell, we first divide the entries by (1 − entries) to map them into $[0, \infty)$. We then normalize the entries by dividing the median of non-zero entries in each cell, and then take square of the entries to boost the signals. Because the transformed entries represent the relative evidence of the methylation status, we input the transformed entries directly as the ratio $\frac{h_1(\cdot)}{h_0(\cdot)}$ in the EM algorithm. Histograms for the distributions of the sc-methylation data are presented in the Supplementary Materials Figure S.1.

2.8 Feature selection

Since scRNA-Seq data is usually the least noisy data type, we use scRNA-Seq data for feature selection. We first cluster scRNA-Seq data with SC3 and then use the cluster assignments to select top 1,000 features with large mean shift across different clusters. More specifically, denote the data matrix as $\boldsymbol{X}_{n\times p}$ (x_{ij} denotes the observation for the i-th cell and j-th feature), the cluster assignments as $\boldsymbol{L}_{n\times 1}$ ($l_i = k$ denotes that the i-th cell belongs to the k-th cluster) and total number of clusters as K. For feature j, we first calculate the difference between the mean of the cells within one cell type and the mean of cells in other cell types; the differences are represented as $\boldsymbol{D}(j) = (d_{1j}, ..., d_{Kj})$, where $d_{kj} = \underset{i:l_i=k}{mean}(x_{ij}) - \underset{i:l_i\neq k}{mean}(x_{ij})$. We take the maximum entry in $D(j)$: $m(j) = \max_k \boldsymbol{D}(j)$. When $m(j)$ is large, it represents that feature j has high expression in one cluster, compared with all other clusters. Finally, we select the top 1,000 features with highest values in $m(j)$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.9 Determination of the number of clusters K

We determine the number of clusters K for the three single-cell datasets seperately before we apply scAMACE. We first run K-Means for each K and calculate the average silhouette width of observations (Kaufman and Rousseeuw, 1990). Silhouette width measures how well an observation has been classified. For each observation i, the silhouette value $s(i)$ is calculated as follows. First denote by A the cluster to which observation i has been assigned and then calculate

$$a(i) = \text{average Euclidean distance of } i \text{ to all other objects of } A.$$

Now consider any cluster C different from A and define

$$d(i, C) = \text{average Euclidean distance of } i \text{ to all objects of } C.$$

$$b(i) = \min_{C \neq A} d(i, C).$$

Then $s(i) = \frac{b(i)-a(i)}{max(b(i),a(i))}$. When cluster A contains only a single observation, we simply set $s(i) = 0$. The average of $s(i)$ for $i = 1, 2, ..., n$ is denoted by $\bar{s}(k)$, and it is called the average silhouette width for the entire data set. $\bar{s}(k)$ is used for the selection of K. Higher value in $\bar{s}(k)$ indicates better clustering outcome. We select K that has the maximum average silhouette width. Details for selecting K in the two real data applications are presented in Supplementary Materials Figures S.2 and S.3. When the similarity of the cell types is high, the Silhoutte method may choose a smaller K than the number of cell types (Figure S.3), and we may choose a larger K instead.

3 Statistical inference: EM algorithm

Given the observed scCAS data $\boldsymbol{X}$, scRNA-Seq data $\boldsymbol{Y}$, and sc-methylation data $\boldsymbol{T}$, we treat the latent variables $\boldsymbol{\Gamma} = \{\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{O}, \boldsymbol{V}, \boldsymbol{M}\}$ as missing data, and use the Expectation-Maximization (EM) algorithm to estimate the parameters $\boldsymbol{\Phi}$=$\{\boldsymbol{\psi}^{acc}$,$\boldsymbol{\omega}^{acc}$,$\boldsymbol{\pi}_i$,$\boldsymbol{\psi}^{rna}$,$\boldsymbol{\omega}^{rna}$,$\boldsymbol{\pi}_l$, $\boldsymbol{\psi}^{met}$,$\boldsymbol{\omega}^{met}$,$\boldsymbol{\pi}_d\}$. The Q-function is $\boldsymbol{Q}(\boldsymbol{\Phi}|\boldsymbol{\Phi}_{old}) = \mathbb{E}_{old}(ln(\boldsymbol{P}(\boldsymbol{\Phi}, \boldsymbol{\Gamma}|obs.)))$, where the expectation is over $\boldsymbol{\Gamma}$ under distribution $\boldsymbol{P}(\boldsymbol{\Gamma}|\boldsymbol{\Phi}^{old}, obs.)$.

In the M-step, we maximize $\boldsymbol{Q}(\boldsymbol{\Phi}|\boldsymbol{\Phi}_{old})$ with respect to $\boldsymbol{\Phi}$ and update parameters as follows.

$$\psi_k^{acc} = \frac{1 + \sum_i \mathbb{E}_{old}(z_{ik})}{K + n_{acc}},$$

$$\psi_k^{rna} = \frac{1 + \sum_l \mathbb{E}_{old}(z_{lk})}{K + n_{rna}},$$

$$\psi_k^{met} = \frac{1 + \sum_d \mathbb{E}_{old}(z_{dk})}{K + n_{met}},$$

$$\pi_{i1} = \frac{\sum_k \sum_g \mathbb{E}_{old}(z_{ik} u_{ig} o_{ig}) + \alpha_{acc} - 1}{\sum_k \sum_g \mathbb{E}_{old}(z_{ik} u_{ig}) + \alpha_{acc} + \beta_{acc} - 2},$$

$$\pi_{l1} = \frac{\sum_k \sum_g \mathbb{E}_{old}(z_{lk} u_{lg} v_{lg})}{\sum_k \sum_g \mathbb{E}_{old}(z_{lk} u_{lg})},$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\pi_{l0} = \frac{\sum_k \sum_g \mathbb{E}_{old}[z_{lk}(1-u_{lg})v_{lg}]}{\sum_k \sum_g \mathbb{E}_{old}[z_{lk}(1-u_{lg})] - 1},$$

$$\pi_{d1} = \frac{\sum_k \sum_g \mathbb{E}_{old}(z_{dk}u_{dg}m_{dg})}{\sum_k \sum_g \mathbb{E}_{old}(z_{dk}u_{dg})},$$

$$\pi_{d0} = \frac{\sum_k \sum_g \mathbb{E}_{old}[z_{dk}(1-u_{dg})m_{dg}] - 1}{\sum_k \sum_g \mathbb{E}_{old}[z_{dk}(1-u_{dg})] - 1},$$

$$\omega_{kg}^{acc} = \frac{\sum_i \mathbb{E}_{old}(z_{ik}u_{ig}) + \mu_{kg}^{acc}\phi^{acc} - 1}{\sum_i \mathbb{E}_{old}(z_{ik}) + \phi^{acc} - 2},$$

$$\omega_{kg}^{met} = \frac{\sum_d \mathbb{E}_{old}(z_{dk}u_{dg}) + \mu_{kg}^{met}\phi^{met} - 1}{\sum_d \mathbb{E}_{old}(z_{dk}) + \phi^{met} - 2}.$$

We use grid search to update ω_{kg}^{rna} because its optimal value does not have an explicit form.

We iterate between E-step and M-step until converge. $\mathbb{E}(\boldsymbol{Z}_{i.})$, $\mathbb{E}(\boldsymbol{Z}_{l.})$ and $\mathbb{E}(\boldsymbol{Z}_{d.})$ in the last iteration are used for clustering. Details for the derivations are presented in the Supplementary Materials.

4 Simulation studies

To validate scAMACE, we generated three different types of simulated data $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{t}$ following the model assumption. In the simulated data, the sample sizes $n_x = 900$, $n_y = 1100$, and $n_t = 1000$. The number of features $p = 1000$. The number of clusters $K = 3$. $f(\omega_{kg}^y) = \eta + \gamma\omega_{kg}^y + \tau(\omega_{kg}^y)^2 = -1 + 7\omega_{kg}^y - 2(\omega_{kg}^y)^2$, $g(\omega_{kg}^y) = \delta + \theta\omega_{kg}^y = -2 + 5\omega_{kg}^y$, $\phi^x = 10$ and $\phi^t = 10$. The detailed simulation scheme is presented in the Supplementary Materials.

For the first data type, x, we set $f_1(x) = 0$ if $x = 0$, and $f_0(x) = 0$ if $x = 1$. We fit a two-component gamma mixture model for $\boldsymbol{y}$ using 'gammamixEM' in R (Young *et al.*, 2019) and beta mixture model for $\boldsymbol{t}$ using 'betamix' in R (Cribari-Neto and Zeileis, 2010; Grun *et al.*, 2012) to estimate the mixture densities. We apply the method in Section 2.6 to estimate parameters in $f(\omega_{kg}^y)$ and $g(\omega_{kg}^y)$. We then implement scAMACE using the estimated densities and $\hat{\eta}, \hat{\gamma}, \hat{\tau}, \hat{\phi}^x, \hat{\delta}, \hat{\theta}, \hat{\phi}^t$.

We use purity, rand index, adjusted rand index and normalized mutual information to evaluate the clustering results. We implement scAMACE either on the three data types seperately ('scAMACE (seperate)') without borrowing information or jointly ('scAMACE (joint)'). Table 1 presents the simulation results. As expected, scAMACE performs the best compared with the other methods. This is likely due to integration of information from all three data sets.

In the following two real data applications, we apply methods mentioned in Section 2.7 instead of fitting a beta mixture model to sc-methylation data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1: Mean and sd (in parentheses) of purity, rand index, adjusted rand index (ARI) and normalized mutual information (NMI) for 50 independent runs.

	Data type	Purity	Rand Index	ARI	NMI
scAMACE (joint)	$\boldsymbol{x}$	0.690(0.025)	0.683(0.018)	0.288(0.041)	0.245(0.035)
	$\boldsymbol{y}$	0.897(0.009)	0.874(0.010)	0.716(0.022)	0.637(0.022)
	$\boldsymbol{t}$	0.704(0.021)	0.693(0.016)	0.310(0.034)	0.265(0.030)
scAMACE (seperate)	$\boldsymbol{x}$	0.659(0.028)	0.662(0.018)	0.241(0.041)	0.205(0.034)
	$\boldsymbol{y}$	0.838(0.012)	0.810(0.012)	0.573(0.028)	0.498(0.026)
	$\boldsymbol{t}$	0.643(0.020)	0.651(0.012)	0.216(0.028)	0.185(0.024)
K-Means	$\boldsymbol{x}$	0.383(0.020)	0.558(0.004)	0.007(0.008)	0.008(0.007)
	$\boldsymbol{y}$	0.714(0.036)	0.702(0.026)	0.331(0.058)	0.283(0.049)
	$\boldsymbol{t}$	0.388(0.021)	0.560(0.004)	0.010(0.008)	0.106(0.008)
Hierarchical Clustering	$\boldsymbol{x}$	0.360(0.011)	0.488(0.047)	0.001(0.001)	0.003(0.002)
	$\boldsymbol{y}$	0.366(0.011)	0.520(0.026)	0.002(0.002)	0.003(0.002)
	$\boldsymbol{t}$	0.360(0.010)	0.532(0.022)	0.001(0.001)	0.002(0.001)
Spectral Clustering	$\boldsymbol{x}$	0.395(0.020)	0.561(0.004)	0.012(0.009)	0.013(0.008)
	$\boldsymbol{y}$	0.722(0.018)	0.708(0.018)	0.344(0.041)	0.295(0.034)
	$\boldsymbol{t}$	0.400(0.025)	0.562(0.005)	0.014(0.012)	0.015(0.011)

5 Application to real data

5.1 Application 1: K562 and GM12878 scRNA-Seq, scATAC-Seq and sc-methylation data

We evaluate scAMACE by jointly clustering scRNA-Seq, scATAC-Seq and sc-methylation data generated from two cell types, K562 and GM12878 (Buenrostro *et al.*, 2015; Li *et al.*, 2017; Pott, 2017). We set $K = 2$, and use the true cell labels as a benchmark to evaluate the performance of the clustering methods. Table 2 presents the clustering results. scAMACE performs well in seperating the cell types. scRNA-Seq is perfectly seperated, while there are only three cells that are not classified correctly in the sc-methylation dataset and eleven misclassifications in the scATAC-Seq dataset. In addition, the two cell types are correctly matched across the three datasets. Compared with the clustering results given by implementing scAMACE seperately on the three datasets, jointly clustering the three datasets improves the overall clustering performance, especially for scATAC-Seq data, which is likely due to the integration of information across the three datasets.

We compared scAMACE with Seurat V3 (Stuart *et al.*, 2019), LIGER (Welch *et al.*, 2019) and scMC (Zhang and Nie, 2021), which are methods for integrative analysis of single-cell data. Examples were presented in Seurat V3 (Stuart *et al.*, 2019) where scRNA-Seq and scATAC-Seq data were integrated. So we implemented Seurat V3 to integrate these two data types. Seurat V3 did not perform well for scATAC-Seq data (Table 2). Seurat V3 is not applicable to integrate sc-methylation data with the other two datasets. Examples were presented in LIGER (Welch *et al.*, 2019) where scRNA-Seq data and sc-methylation data were integrated. So we implemented LIGER to integrate these two data types. LIGER did not perform well on sc-methylation data (Table 2). We also implemented LIGER to integrate all three datasets, and LIGER still did not perform well on sc-methylation data (Supplementary Materials

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2: Clustering tables for K562, GM12878 scRNA-Seq, scATAC-Seq and sc-methylation data.

		scAMACE (joint)			scAMACE (seperate)		
		1	2		1	2	
scATAC-Seq	GM12878	368	5		254	119	
	K562	6	660		171	495	
scRNA-Seq	GM12878	128	0		128	0	
	K562	0	73		0	73	
sc-methyl	GM12878	16	3		7	12	
	K562	0	11		11	0	
		Seurat V3			LIGER		
		1	2	3	1	2	3
scATAC-Seq	GM12878	346	27				
	K562	499	167				
scRNA-Seq	GM12878	101	2	25	127	0	1
	K562	0	73	0	10	63	0
sc-methyl	GM12878						19
	K562						11

Table 3: Comparison of the performance of different methods on the K562, GM12878 dataset by adjusted rand index (ARI).

	scAMACE (joint)	scAMACE (seperate)	Seurat V3	LIGER	scMC
scATAC-Seq	0.958	0.192	0.033		0.000
scRNA-Seq	1.000	1.000	0.713	0.800	0.771
sc-methyl	0.628	0.260		0.000	0.000

Table S.1), this may be due to the small sample size in sc-methylation data. scMC (Zhang and Nie, 2021) was developed for the integrative analysis of multiple single-cell datasets with the same data type. Since the features in scATAC-Seq data, scRNA-Seq data and sc-methylation data are linked, scMC can be implemented in principle. scMC did not perform well on scATAC-Seq data and sc-methylation data (Supplementary Materials Table S.1). This may be due to the fact that the characteristics of different data types are very different, and ignoring the difference leads to suboptimal performance.

5.2 Application 2: Mouse neocortex scRNA-Seq, sci-ATAC-Seq and sc-methylation data

In this example, we evaluate scAMACE for the joint analysis of single-cell datasets where the cell types are different across the datasets.

We collected single-cell datasets generated from mouse neocortex. There are five cell types in scRNA-Seq data (Tasic *et al.*, 2018), including astrocytes (Astro), glutamatergic neurons in layer 4 (L4) , corticothalamic glutamatergic neurons in layer 6 (L6 CT), oligodendrocytes (Oligo) and Pvalb+ GABAergic neurons (Pvalb). There are three cell types in sci-ATAC-Seq data (Cusanovich *et al.*, 2018b), including

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

astrocytes (Astro), excitatory neurons CPN (Ex. neurons CPN), and oligodendrocytes (Oligo). There are three cell types in sc-methylation dataset (Luo *et al.*, 2017), including excitatory neurons in layer 4 (L4), excitatory neurons in layer 6 (labeled as L6-2 in (Luo *et al.*, 2017)), and Pvalb+ GABAergic neurons (Pvalb). In the three datasets, the optimal numbers of clusters chosen by the Silhoutte method, $\hat{K} = 2$, tend to be smaller than the numbers of cell types, which is likely due to the similarity of the neuronal subtypes. We set K=5 when we implement scAMACE, instead of the value given by the Silhoutte method. The true cell labels are used as a benchmark for evaluating the performance of the clustering methods.

The clustering results are presented in Table 4. Even though K is larger than the number of cell types in sci-ATAC-Seq data and sc-methylation data, scAMACE still determines the correct number of cell types in sci-ATAC-Seq data. Although the cells in sc-methylation data fall into four clusters, there are only seven cells in cluster 4. Cell types in all three datasets are well seperated. Astrocytes and oligodendrocytes are matched across scRNA-Seq data and sci-ATAC-Seq data. Excitatory neurons CPN in sci-ATAC-Seq data are matched with glutamatergic neurons in layer 4 in the scRNA-Seq data. We note that most excitatory neurons are glutamatergic neurons. Excitatory neurons in layers 4 and 6, and Pvalb+ GABAergic neurons are matched between scRNA-Seq data and sc-methylation data.

Compared with implementing scAMACE on the three datasets separately, the joint analysis leads to improvement in clustering, especially for sc-methylation dataset. This is likely because the joint model borrows information across the three datasets. Similar to application 1, we implemented Seurat V3 to integrate scRNA-Seq and sci-ATAC-Seq data. Seurat V3 (Stuart *et al.*, 2019) does not perform well on sci-ATAC-Seq data (Table 4). We implemented LIGER (Welch *et al.*, 2019) to integrate scRNA-Seq and sc-methylation data. LIGER does not seperate excitatory neurons in layer 4 and layer 6 in sc-methylation data (Table 4). We also integrated all three datasets by LIGER (Welch *et al.*, 2019) and scMC (Zhang and Nie, 2021). LIGER and scMC did not perform well (Supplementary Materials Table S.3). Overall, scAMACE performed the best compared with the other methods.

5.3 Computational cost

LIGER, Seurat V3 and scMC only provide the versions that are implemented on CPU, while scAMACE can be implemented on both CPU and GPU. We summarized the computational time for scAMACE (CPU version and GPU version in python), LIGER (Welch *et al.*, 2019), Seurat V3 (Stuart *et al.*, 2019) and scMC (Zhang and Nie, 2021). We implemented scAMACE, LIGER and scMC to cluster the three types of data simultaneously, and we implemented Seurat V3 to cluster scCAS data and scRNA-Seq data.

On real data application 2 (∼8,000 cells), the computational time for scAMACE are 418.858 seconds on one 3.4GHz Intel Xeon Gold CPU and 69.652 seconds on one 3.1GHz Dual Intel Xeon Gold GPU. Compared with LIGER (80.389 seconds on one 3.4GHz Intel Xeon Gold CPU), scMC (372.323 seconds on one 3.4GHz Intel Xeon Gold CPU) and Seurat V3 (116.688 seconds for scRNA-Seq and sci-ATAC-Seq data on one 3.4GHz Intel Xeon Gold CPU), scAMACE has competitive computational speed, especially the GPU version.

Next, we generated a dataset with sample size=30,000 ($n_{acc} = n_{rna} = n_{met} = 10,000$) by sampling the cells with replacement from real data application 2. The computational time for scAMACE are 1534.631 seconds on one 3.4GHz Intel Xeon Gold CPU and 250.089 seconds on one 3.1GHz Dual Intel Xeon Gold GPU. Compared with LIGER (555.574 seconds on one 3.4GHz Intel Xeon Gold CPU), scMC (3667.878

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 4: Clustering tables for the mouse neocortex scRNA-Seq, sci-ATAC-Seq, and sc-methylation data.

		scAMACE (joint)					scAMACE (seperate)				
		1	2	3	4	5	1	2	3	4	5
sci-ATAC-Seq	Astro	550	0		1		550	0	1		
	Ex. neurons CPN	0	1391		0		1	1390	0		
	Oligo	0	1		457		0	0	458		
scRNA-Seq	Astro	368	0	0	0	0	368	0	0	0	0
	L4	0	1401	0	0	0	0	1401	0	0	0
	L6 CT	0	0	960	0	0	0	0	960	0	0
	Oligo	25	0	0	66	0	27	0	0	64	0
	Pvalb	0	0	0	0	1337	0	0	0	0	1337
sc-methyl	L4		411	1	0	0	412				
	L6-2		20	703	6	0	729				
	Pvalb		0	0	1	153	154				

		Seurat V3									LIGER						
		1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7
sci-ATAC-Seq	Astro	296	29	1	2	223											
	Ex. neurons CPN	110	461	243	0	577											
	Oligo	64	90	1	0	303											
scRNA-Seq	Astro	0	0	0	0	0	368	0	0	0	6	0		0	362		0
	L4	1028	0	0	0	373	0	0	0	0	0	1401		0	0		0
	L6 CT	0	960	0	0	0	0	0	0	0	2	0		958	0		0
	Oligo	0	0	0	0	0	0	0	60	31	31	0		0	0		60
	Pvalb	0	0	647	498	0	0	192	0	0	1337	0		0	0		0
sc-methyl	L4										0	11	679			0	
	L6-2										0	13	399			0	
	Pvalb										68	0	0			86	

Table 5: Comparison of the performance of different methods on the mouse neocortex dataset by adjusted rand index (ARI).

	scAMACE (joint)	scAMACE (seperate)	Seurat V3	LIGER	scMC
sci-ATAC-Seq	0.998	0.998	0.058		0.019
scRNA-Seq	0.997	0.997	0.697	0.983	0.145
sc-methyl	0.932	0.000		0.316	0.001

seconds on one 3.4GHz Intel Xeon Gold CPU) and Seurat V3 (290.640 seconds for scRNA-Seq and sci-ATAC-Seq data on one 3.4GHz Intel Xeon Gold CPU), scAMACE has competitive computational speed on datasets with larger scale.

6 Conclusion

Unsupervised methods including dimension reduction and clustering are essential to the analysis of single-cell genomic data as the cell types are usually unknown. We have developed scAMACE, a model-based approach for integratively clustering single-cell data on chromatin accessibility, gene expression and methylation. scAMACE provides statistical inference of cluster assignments and achieves better cell type seperation combining biological information across different types of genomic features. The cells in our real data examples are differentiated and mature cells. In the future, we will investigate the

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

performance of scAMACE on immature cells undergoing differentiation.

Acknowledgements

We would like to thank Jinwen Yang, Wenyu Zhang and Pengcheng Zeng for the helpful discussions.

Funding

This work has been supported by the Chinese University of Hong Kong direct grants (4053360, 4053423), the Chinese University of Hong Kong startup grant (4930181), the Chinese University of Hong Kong's Project Impact Enhancement Fund (PIEF) and Science Faculty's Collaborative Research Impact Matching Scheme (CRIMS), and Hong Kong Research Grant Council (ECS 24301419, GRF 14301120).

References

Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., and Aerts, S. (2019). cistopic: cis-regulatory topic modeling on single-cell atac-seq data. *Nature Methods*, **16**(5), 397–400.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**(7561), 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., Greenleaf, W. J., Majeti, R., and Chang, H. Y. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, **48**(10), 1193–1203.

Cribari-Neto, F. and Zeileis, A. (2010). Beta regression in r. *Journal of Statistical Software, Articles*, **34**(2), 1–24.

Cusanovich, D. A., Reddington, J. P., Garfield, D. A., Daza, R. M., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H. A., Christiansen, L., Qiu, X., Steemers, F. J., Trapnell, C., Shendure, J., and Furlong, E. E. M. (2018a). The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*, **555**(7697), 538–542.

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C., and Shendure, J. (2018b). A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, **174**(5), 1309 – 1324.e18.

Duren, Z., Chen, X., Jiang, R., Wang, Y., and Wong, W. H. (2017). Modeling gene regulation from paired expression and chromatin accessibility data. *Proceedings of the National Academy of Sciences*, **114**(25), E4914–E4923.

Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A. T., Chang, H. Y., Wang, Y., and Wong, W. H. (2018). Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proceedings of the National Academy of Sciences*, **115**(30), 7723–7728.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Grun, B., Kosmidis, I., and Zeileis, A. (2012). Extended beta regression in r: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software, Articles*, **48**(11), 1–25.

Hui, T., Cao, Q., Wegrzyn-Woltosz, J., O'Neill, K., Hammond, C. A., Knapp, D. J., Laks, E., Moksa, M., Aparicio, S., Eaves, C. J., Karsan, A., and Hirst, M. (2018). High-resolution single-cell dna methylation measurements reveal epigenetically distinct hematopoietic stem cell subpopulations. *Stem Cell Reports*, **11**(2), 578 – 592.

Kapourani, C.-A. and Sanguinetti, G. (2016). Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics*, **32**(17), i405–i412.

Kapourani, C.-A. and Sanguinetti, G. (2019). Melissa: Bayesian clustering and imputation of single-cell methylomes. *Genome Biology*, **20**(1), 61.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., and Hemberg, M. (2017). Sc3: consensus clustering of single-cell rna-seq data. *Nature Methods*, **14**(5), 483–486.

Lahnemann, D., Koster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S. O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. d., Cappuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korbel, J. O., Kozlov, A. M., Kuo, T.-H., Lelieveldt, B. P., Mandoiu, I. I., Marioni, J. C., Marschall, T., Molder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J. d., Saliba, A.-E., Somarakis, A., Stegle, O., Theis, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P., and Schonhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, **21**(1), 31.

Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., Kong, S. L., Chua, C., Hon, L. K., Tan, W. S., Wong, M., Choi, P. J., Wee, L. J. K., Hillmer, A. M., Tan, I. B., Robson, P., and Prabhakar, S. (2017). Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature Genetics*, **49**(5), 708–718.

Lin, P., Troup, M., and Ho, J. W. K. (2017). Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome Biology*, **18**(1), 59.

Lin, Z., Zamanighomi, M., Daley, T., Ma, S., and Wong, W. H. (2020). Model-based approach to the joint analysis of single-cell data on chromatin accessibility and gene expression. *Statist. Sci.*, **35**(1), 2–13.

Luo, C., Keown, C. L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J. R., Sandoval, J. P., Bui, B., Sejnowski, T. J., Harkins, T. T., Mukamel, E. A., Behrens, M. M., and Ecker, J. R. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, **357**(6351), 600–604.

Ma, A., McDermaid, A., Xu, J., Chang, Y., and Ma, Q. (2020). Integrative methods and practical challenges for single-cell multi-omics. *Trends in Biotechnology*, **38**(9), 1007 – 1022.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 849856, Cambridge, MA, USA. MIT Press.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pott, S. (2017). Simultaneous measurement of chromatin accessibility, dna methylation, and nucleosome phasing in single cells. *eLife*, **6**, e23203.

Silvia, F. and Francisco, C.-N. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, **177**(7), 1888 – 1902.e21.

Tasic, B., Yao, Z., Graybuck, L. T., Smith, K. A., Nguyen, T. N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M. N., Viswanathan, S., Penn, O., Bakken, T., Menon, V., Miller, J., Fong, O., Hirokawa, K. E., Lathia, K., Rimorin, C., Tieu, M., Larsen, R., Casper, T., Barkan, E., Kroll, M., Parry, S., Shapovalova, N. V., Hirschstein, D., Pendergraft, J., Sullivan, H. A., Kim, T. K., Szafer, A., Dee, N., Groblewski, P., Wickersham, I., Cetin, A., Harris, J. A., Levi, B. P., Sunkin, S. M., Madisen, L., Daigle, T. L., Looger, L., Bernard, A., Phillips, J., Lein, E., Hawrylycz, M., Svoboda, K., Jones, A. R., Koch, C., and Zeng, H. (2018). Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, **563**(7729), 72–78.

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. *Nature Methods*, **14**(4), 414–416.

Wang, C., Sun, D., Huang, X., Wan, C., Li, Z., Han, Y., Qin, Q., Fan, J., Qiu, X., Xie, Y., Meyer, C. A., Brown, M., Tang, M., Long, H., Liu, T., and Liu, X. S. (2020). Integrative analyses of single-cell transcriptome and regulome using maestro. *Genome Biology*, **21**(1), 198.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**(301), 236–244.

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, **177**(7), 1873 – 1887.e17.

Xiong, L., Xu, K., Tian, K., Shao, Y., Tang, L., Gao, G., Zhang, M., Jiang, T., and Zhang, Q. C. (2019). Scale method for single-cell atac-seq analysis via latent feature extraction. *Nature Communications*, **10**(1), 4576.

Yin, L., Luo, Y., Xu, X., Wen, S., Wu, X., Lu, X., and Xie, H. (2019). Virtual methylome dissection facilitated by single-cell analyses. *Epigenetics & Chromatin*, **12**(1), 66.

Young, D. S., Chen, X., Hewage, D. C., and Nilo-Poyanco, R. (2019). Finite mixture-of-gamma distributions: estimation, inference, and model-based clustering. *Advances in Data Analysis and Classification*, **13**(4), 1053–1082.

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. *Nature Communications*, **9**(1), 2410.

Zeng, P., Wangwu, J., and Lin, Z. (2020). Coupled co-clustering-based unsupervised transfer learning for the integrative analysis of single-cell genomic data. *Briefings in Bioinformatics*. bbaa347.

Zhang, L. and Nie, Q. (2021). scmc learns biological variation through the alignment of multiple single-cell genomics datasets. *Genome Biology*, **22**(1), 10.

Zhu, L., Lei, J., Klei, L., Devlin, B., and Roeder, K. (2019). Semisoft clustering of single-cell data. *Proceedings of the National Academy of Sciences*, **116**(2), 466–471.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SUPPLEMENTARY MATERIALS

S.1 Supplementary Text

S.1.1 Joint likelihood

- scCAS data

$$\begin{aligned}&\boldsymbol{P}(\boldsymbol{X}, \boldsymbol{U}_{i.}, \boldsymbol{O}_{i.}, \boldsymbol{Z}_{i.} | \boldsymbol{\psi}^{acc}, \boldsymbol{\omega}^{acc}, \boldsymbol{\pi}_i) \\ &= \prod_i \prod_k \left\{ \psi_k^{acc} * \prod_g \left[\omega_{kg}^{acc} (\pi_{i1} f_1)^{o_{ig}} ((1-\pi_{i1}) f_0)^{1-o_{ig}} \right]^{u_{ig}} * \left[(1-\omega_{kg}^{acc}) f_0 \right]^{1-u_{ig}} \right\}^{z_{ik}}.\end{aligned}$$

- scRNA-Seq data

$$\begin{aligned}&\boldsymbol{P}(\boldsymbol{Y}, \boldsymbol{U}_{l.}, \boldsymbol{V}_{l.}, \boldsymbol{Z}_{l.} | \boldsymbol{\psi}^{rna}, \boldsymbol{\omega}^{rna}, \boldsymbol{\pi}_l) = \prod_l \prod_k [\psi_k^{rna} * \boldsymbol{A}]^{z_{lk}}, \\ &\boldsymbol{A} = \prod_g \left\{ \omega_{kg}^{rna} \left[\pi_{l1} g_1 \right]^{v_{lg}} * \left[(1-\pi_{l1}) g_0 \right]^{1-v_{lg}} \right\}^{u_{lg}} * \left\{ (1-\omega_{kg}^{rna}) \left[\pi_{l0} g_1 \right]^{v_{lg}} * \left[(1-\pi_{l0}) g_0 \right]^{1-v_{lg}} \right\}^{1-u_{lg}}.\end{aligned}$$

- sc-methylation data

$$\begin{aligned}&\boldsymbol{P}(\boldsymbol{T}, \boldsymbol{U}_{d.}, \boldsymbol{M}_{d.}, \boldsymbol{Z}_{d.} | \boldsymbol{\psi}^{met}, \boldsymbol{\omega}^{met}, \boldsymbol{\pi}_d) = \prod_d \prod_k [\psi_k^{met} * \boldsymbol{B}]^{z_{dk}}, \\ &\boldsymbol{B} = \prod_g \left\{ \omega_{kg}^{met} (\pi_{d1} h_1)^{m_{dg}} * \left[(1-\pi_{d1}) h_0 \right]^{1-m_{dg}} \right\}^{u_{dg}} * \left\{ (1-\omega_{kg}^{met}) (\pi_{d0} h_1)^{m_{dg}} * \left[(1-\pi_{d0}) h_0 \right]^{1-m_{dg}} \right\}^{1-u_{dg}}.\end{aligned}$$

S.1.2 Q-function

Let $\boldsymbol{\Gamma}$ denote the missing data, and let $\boldsymbol{\Phi}$ denote the parameters. the Q-function is $\boldsymbol{Q}(\boldsymbol{\Phi}|\boldsymbol{\Phi}_{old}) = \mathbb{E}_{old}(ln(\boldsymbol{P}(\boldsymbol{\Phi}, \boldsymbol{\Gamma}|obs.)))$, where the expectation is over $\boldsymbol{\Gamma}$ under distribution $\boldsymbol{P}(\boldsymbol{\Gamma}|\boldsymbol{\Phi}^{old}, obs.) := \boldsymbol{P}_{old}(\boldsymbol{\Gamma})$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$
\begin{aligned}
&ln(\boldsymbol{P}(\boldsymbol{\Phi},\boldsymbol{\Gamma}|obs.)) \\
&= \sum_i \sum_k z_{ik} ln(\psi_k^{acc}) + \sum_i \sum_k z_{ik} \sum_g \left[u_{ig} ln(\omega_{kg}^{acc}) + (1-u_{ig}) ln(1-\omega_{kg}^{acc})\right] \\
&+ \sum_i \sum_k z_{ik} \sum_g \left[u_{ig} o_{ig} ln(\pi_{i1}) + u_{ig}(1-o_{ig}) ln(1-\pi_{i1}) + u_{ig} o_{ig} ln(f_1) + (1-u_{ig} o_{ig}) ln(f_0)\right] \\
&+ \sum_l \sum_k z_{lk} ln(\psi_k^{rna}) + \sum_l \sum_k z_{lk} \sum_g \left[u_{lg} ln(\omega_{kg}^{rna}) + (1-u_{lg}) ln(1-\omega_{kg}^{rna})\right] \\
&+ \sum_l \sum_k z_{lk} \sum_g \left[u_{lg} v_{lg} ln(\pi_{l1}) + u_{lg}(1-v_{lg}) ln(1-\pi_{l1})\right] \\
&+ \sum_l \sum_k z_{lk} \sum_g \left[(1-u_{lg}) v_{lg} ln(\pi_{l0}) + (1-u_{lg})(1-v_{lg}) ln(1-\pi_{l0})\right] \\
&+ \sum_l \sum_k z_{lk} \sum_g \left[v_{lg} ln(g_1) + (1-v_{lg}) ln(g_0)\right] \\
&+ \sum_d \sum_k z_{dk} ln(\psi_k^{met}) + \sum_d \sum_k z_{dk} \sum_g \left[u_{dg} ln(\omega_{kg}^{met}) + (1-u_{dg}) ln(1-\omega_{kg}^{met})\right] \\
&+ \sum_d \sum_k z_{dk} \sum_g \left[u_{dg} m_{dg} ln(\pi_{d1}) + u_{dg}(1-m_{dg}) ln(1-\pi_{d1})\right] \\
&+ \sum_d \sum_k z_{dk} \sum_g \left[(1-u_{dg}) m_{dg} ln(\pi_{d0}) + (1-u_{dg})(1-m_{dg}) ln(1-\pi_{d0})\right] \\
&+ \sum_d \sum_k z_{dk} \sum_g \left[m_{dg} ln(h_1) + (1-m_{dg}) ln(h_0)\right] \\
&+ \sum_k ln(\psi_k^{acc}) + \sum_k ln(\psi_k^{rna}) + \sum_k ln(\psi_k^{met}) \\
&+ \sum_i \left[(\alpha_{acc}-1) ln(\pi_{i1}) + (\beta_{acc}-1) ln(1-\pi_{i1})\right] + \sum_l \left[-ln(1-\pi_{l0})\right] + \sum_d \left[-ln(\pi_{d0})\right] \\
&+ \sum_k \sum_g \left[(\alpha_1-1) ln(\omega_{kg}^{rna}) + (\beta_1-1) ln(1-\omega_{kg}^{rna})\right] \\
&+ \sum_k \sum_g \left\{(\mu_{kg}^{acc}\phi^{acc}-1) ln(\omega_{kg}^{acc}) + (\phi^{acc}-\mu_{kg}^{acc}\phi^{acc}-1) ln(1-\omega_{kg}^{acc}) - ln\left[Beta(\mu_{kg}^{acc}\phi^{acc}, \phi^{acc}-\mu_{kg}^{acc}\phi^{acc})\right]\right\} \\
&+ \sum_k \sum_g \left\{(\mu_{kg}^{met}\phi^{met}-1) ln(\omega_{kg}^{met}) + (\phi^{met}-\mu_{kg}^{met}\phi^{met}-1) ln(1-\omega_{kg}^{met}) - ln\left[Beta(\mu_{kg}^{met}\phi^{met}, \phi^{met}-\mu_{kg}^{met}\phi^{met})\right]\right\} \\
&+ \boldsymbol{C},
\end{aligned}
$$

where $\mu_{kg}^{acc} = \frac{1}{1+e^{-f(\omega_{kg}^{rna})}}$, $\mu_{kg}^{met} = \frac{1}{1+e^{-g(\omega_{kg}^{rna})}}$, and $\boldsymbol{C}$ is a constant that does not depend on the parameters.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S.1.3 Expectations in E-Step

- scCAS data

$$\mathbb{E}_{old}(z_{ik}) \propto \psi_k^{acc} \prod_g \left\{ \left[\omega_{kg}^{acc}(\pi_{i1}f_1 + (1-\pi_{i1})f_0)\right] + \left[(1-\omega_{kg}^{acc})f_0\right] \right\},$$

$$\mathbb{E}_{old}(z_{ik}u_{ig}) = \frac{\omega_{kg}^{acc}(\pi_{i1}f_1 + (1-\pi_{i1})f_0)}{\omega_{kg}^{acc}(\pi_{i1}f_1 + (1-\pi_{i1})f_0) + (1-\omega_{kg}^{acc})f_0} * P_{old}(z_{ik}=1),$$

$$\mathbb{E}_{old}(z_{ik}u_{ig}o_{ig}) = \frac{\pi_{i1}f_1}{\pi_{i1}f_1 + (1-\pi_{i1})f_0} * P_{old}(z_{ik}=1, u_{ig}=1).$$

- scRNA-Seq data

$$\mathbb{E}_{old}(z_{lk}) \propto \psi_k^{rna} \prod_g \left\{ \omega_{kg}^{rna}\left[\pi_{l1}g_1 + (1-\pi_{l1})g_0\right] + (1-\omega_{kg}^{rna})\left[\pi_{l0}g_1 + (1-\pi_{l0})g_0\right] \right\},$$

$$\mathbb{E}_{old}(z_{lk}u_{lg}) = \frac{\omega_{kg}^{rna}\left[\pi_{l1}g_1 + (1-\pi_{l1})g_0\right]}{\omega_{kg}^{rna}\left[\pi_{l1}g_1 + (1-\pi_{l1})g_0\right] + (1-\omega_{kg}^{rna})\left[\pi_{l0}g_1 + (1-\pi_{l0})g_0\right]} * P_{old}(z_{lk}=1),$$

$$\mathbb{E}_{old}(z_{lk}(1-u_{lg})) = \mathbb{E}_{old}(z_{lk}) - \mathbb{E}_{old}(z_{lk}u_{lg}),$$

$$\mathbb{E}_{old}(z_{lk}u_{lg}v_{lg}) = \frac{\pi_{l1}g_1}{\pi_{l1}g_1 + (1-\pi_{l1})g_0} * P_{old}(z_{lk}=1, u_{lg}=1),$$

$$\mathbb{E}_{old}(z_{lk}(1-u_{lg})v_{lg}) = \frac{\pi_{l0}g_1}{\pi_{l0}g_1 + (1-\pi_{l0})g_0} * P_{old}(z_{lk}=1, u_{lg}=0),$$

$$\mathbb{E}_{old}(z_{lk}v_{lg}) = \mathbb{E}_{old}(z_{lk}u_{lg}v_{lg}) + \mathbb{E}_{old}(z_{lk}(1-u_{lg})v_{lg}).$$

- sc-methylation data

$$\mathbb{E}_{old}(z_{dk}) \propto \psi_k^{met} \prod_g \left\{ \omega_{kg}^{met}\left[\pi_{d1}h_1 + (1-\pi_{d1})h_0\right] + (1-\omega_{kg}^{met})\left[\pi_{d0}h_1 + (1-\pi_{d0})h_0\right] \right\},$$

$$\mathbb{E}_{old}(z_{dk}u_{dg}) = \frac{\omega_{kg}^{met}\left[\pi_{d1}h_1 + (1-\pi_{d1})h_0\right]}{\omega_{kg}^{met}\left[\pi_{d1}h_1 + (1-\pi_{d1})h_0\right] + (1-\omega_{kg}^{met})\left[\pi_{d0}h_1 + (1-\pi_{d0})h_0\right]} * P_{old}(z_{dk}=1),$$

$$\mathbb{E}_{old}(z_{dk}(1-u_{dg})) = \mathbb{E}_{old}(z_{dk}) - \mathbb{E}_{old}(z_{dk}u_{dg}),$$

$$\mathbb{E}_{old}(z_{dk}u_{dg}m_{dg}) = \frac{\pi_{d1}h_1}{\pi_{d1}h_1 + (1-\pi_{d1})h_0} * P_{old}(z_{dk}=1, u_{dg}=1),$$

$$\mathbb{E}_{old}(z_{dk}(1-u_{dg})m_{dg}) = \frac{\pi_{d0}h_1}{\pi_{d0}h_1 + (1-\pi_{d0})h_0} * P_{old}(z_{dk}=1, u_{dg}=0),$$

$$\mathbb{E}_{old}(z_{dk}m_{dg}) = \mathbb{E}_{old}(z_{dk}u_{dg}m_{dg}) + \mathbb{E}_{old}(z_{dk}(1-u_{dg})m_{dg}).$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S.1.4 Simulation scheme

We generated three different types of simulated data $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{t}$ following the model assumption. In the simulated data, the sample sizes $n_x = 900$, $n_y = 1100$, and $n_t = 1000$. The number of features $p = 1000$. The number of clusters $K = 3$. $f(\omega_{kg}^y) = \eta + \gamma\omega_{kg}^y + \tau(\omega_{kg}^y)^2 = -1 + 7\omega_{kg}^y - 2(\omega_{kg}^y)^2$, $g(\omega_{kg}^y) = \delta + \theta\omega_{kg}^y = -2 + 5\omega_{kg}^y$, $\phi^x = 10$ and $\phi^t = 10$. The followings are the simulation scheme:

A. Generate $\boldsymbol{\omega}^y$

For $g = 1, ..., 150$:

$$\omega_{kg}^y \sim \begin{array}{c} \begin{matrix} (1,\ldots,50) & (51,\ldots,100) & (101,\ldots,150) \end{matrix} \\ \begin{bmatrix} \omega & 0.5 & 1-\omega \\ 0.5 & 1-\omega & \omega \\ 1-\omega & \omega & 0.5 \end{bmatrix} \end{array}$$

We set $\omega = 0.8$.

For $g = 151, ..., 1000$:

$$\omega_{kg}^y \sim \begin{cases} Beta\ (\alpha = 2, \beta = 2), & \text{for } k = 1 \\ \omega_{1g}^y, & \text{for } k = 2, 3 \end{cases}$$

To summarize, we set the first 150 features to be differential and for the remaining $151, \ldots, p$ features, we set ω_{kg}^y to be the same across different clusters k.

B. Generate $\boldsymbol{\omega}^x$ and $\boldsymbol{\omega}^t$.

For $g = 1, ..., 150$:

$\omega_{kg}^x \sim Beta\ (\mu_{kg}^x = \frac{1}{1+e^{-f(\omega_{kg}^y)}}, \phi^x)$, for $k = 1, 2, 3$

For $g = 151, ..., 1000$:

$$\omega_{kg}^x \sim \begin{cases} Beta\ (\mu_{kg}^x = \frac{1}{1+e^{-f(\omega_{1g}^y)}}, \phi^x), & \text{for } k = 1 \\ \omega_{1g}^x, & \text{for } k = 2, 3 \end{cases}$$

For $g = 1, ..., 150$:

$\omega_{kg}^t \sim Beta\ (\mu_{kg}^t = \frac{1}{1+e^{-g(\omega_{kg}^y)}}, \phi^t)$, for $k = 1, 2, 3$

For $g = 151, ..., 1000$:

$$\omega_{kg}^t \sim \begin{cases} Beta\ (\mu_{kg}^t = \frac{1}{1+e^{-g(\omega_{1g}^y)}}, \phi^t), & \text{for } k = 1 \\ \omega_{1g}^t, & \text{for } k = 2, 3 \end{cases}$$

C. Generate $\boldsymbol{z}^x$, $\boldsymbol{z}^y$ and $\boldsymbol{z}^t$. The cluster labels are generated with equal probability, $P(\boldsymbol{z}_\cdot = 1) = P(\boldsymbol{z}_\cdot = 2) = P(\boldsymbol{z}_\cdot = 3) = \frac{1}{3}$.

D. Data type 1: $\boldsymbol{x}$

- Generate $\boldsymbol{u}^x$. We generate u_{ig} from $Bernoulli(\omega_{kg}^x)$ if $z_{ik} = 1$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- Generate $\boldsymbol{o}^x$. We generate o_{ig} from $Bernoulli(\pi_{i1})$ if $u_{ig} = 1$, and set $o_{ig} = 0$ if $u_{ig} = 0$. We set $\pi_{i1} = 0.2$ for $i = 1, \cdots, n_x$.
- Generate $\boldsymbol{x}$. We generate $x_{ig} = 1$ if $o_{ig} = 1$, and generate $x_{ig} = 0$ if $o_{ig} = 0$.

E. Data type 2: $\boldsymbol{y}$

- Generate $\boldsymbol{u}^y$. We generate u_{lg} from $Bernoulli(\omega^y_{kg})$ if $z_{lk} = 1$.
- Generate $\boldsymbol{v}^y$. We generate v_{lg} from $Bernoulli(\pi_{l1})$ if $u_{lg} = 1$, and from $Bernoulli(\pi_{l0})$ if $u_{lg} = 0$. We set $\pi_{l1} = 0.7, \pi_{l0} = 0.3$ for $l = 1, \cdots, n_y$.
- Generate $\boldsymbol{y}$. We generate y_{lg} from $Gamma(shape = 7, scale = 0.5)$ if $v_{lg} = 1$, and generate y_{lg} from $Gamma(shape = 1, scale = 1)$ if $v_{lg} = 0$.

F. Data type 3: $\boldsymbol{t}$

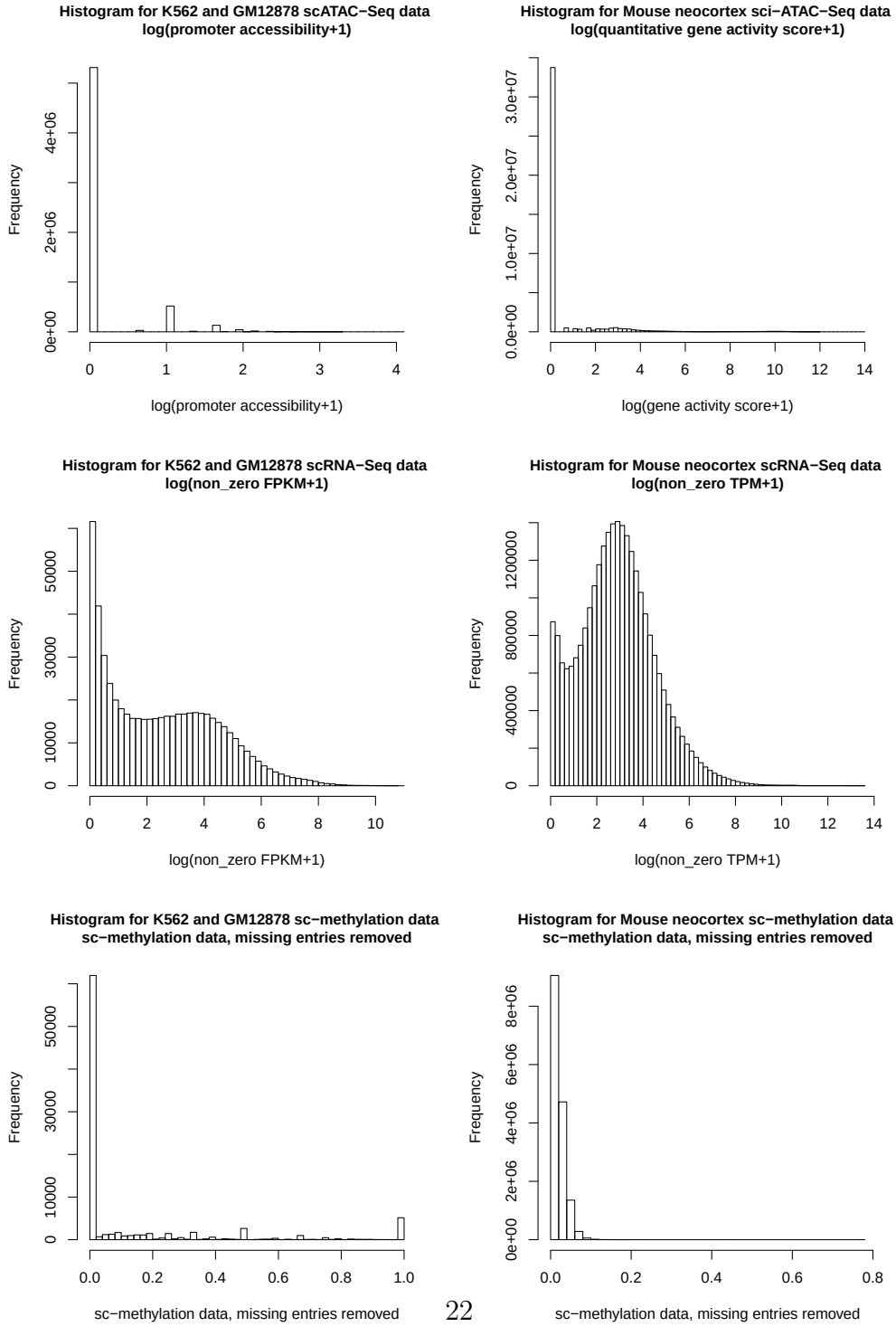
- Generate $\boldsymbol{u}^t$. We generate u_{dg} from $Bernoulli(\omega^t_{kg})$ if $z_{dk} = 1$.
- Generate $\boldsymbol{m}^t$. We generate m_{dg} from $Bernoulli(\pi_{d1})$ if $u_{dg} = 1$, and from $Bernoulli(\pi_{d0})$ if $u_{dg} = 0$. We set $\pi_{d1} = 0.4, \pi_{d0} = 0.7$ for $d = 1, \cdots, n_t$.
- Generate $\boldsymbol{t}$. We generate t_{dg} from $Beta(\alpha = 0.5, \beta = 0.5)$ if $m_{dg} = 1$, and generate t_{dg} from $Beta(\alpha = 1, \beta = 10)$ if $m_{dg} = 0$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S.2 Supplementary Figures

S.2.1 Histograms for two real data applications

Figure S.1: Histograms for Application 1 (Left) and Appliation 2 (Right) scCAS data (Upper), scRNA-Seq data (Middle) and sc-methylation data(Lower).



bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S.2.2 Determination of number of clusters K for real data applications

We applied the Silhouette method (Kaufman and Rousseeuw, 1990) mentioned in Section 2.9 on the two real data applications to determine K before we apply scAMACE. The result for real application 1 is presented in Figure S.2: $\hat{K} = 2$ is chosen for the three single-cell datasets, where the true number of cell types is 2. The results for real data application 2 are presented in Figure S.3. There are five cell types in scRNA-Seq data (Tasic *et al.*, 2018), including astrocytes, oligodendrocytes, and three subtypes of neurons. There are three cell types in sci-ATAC-Seq data (Cusanovich *et al.*, 2018b), including astrocytes, oligodendrocytes, and excitatory neurons CPN. And there are three cell types in sc-methylation dataset (Luo *et al.*, 2017), including three subtypes of neurons. In the three datasets, the optimal numbers of clusters chosen by the Silhoutte method ($\hat{K} = 2$) tend to be smaller than the numbers of cell types, which is likely due to the similarity of the neuronal subtypes. We chose $K = 5$ when we implement scAMACE, instead of the suggested $\hat{K} = 2$ by the Silhoutte method.

Figure S.2: Average Silhouette width v.s. K for Application 1, K562 and GM12878 cells: scATAC-Seq (Left), scRNA-Seq (Middle), and sc-methylation data (Right).

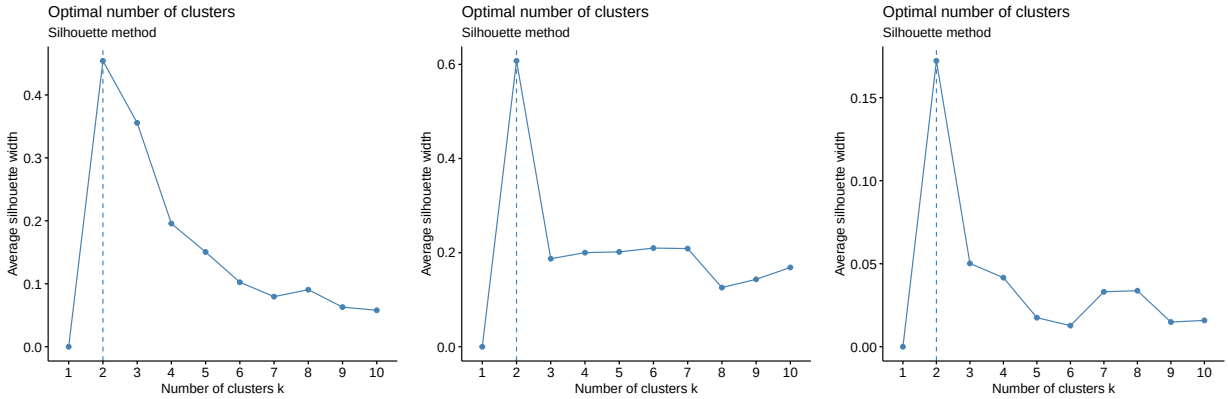
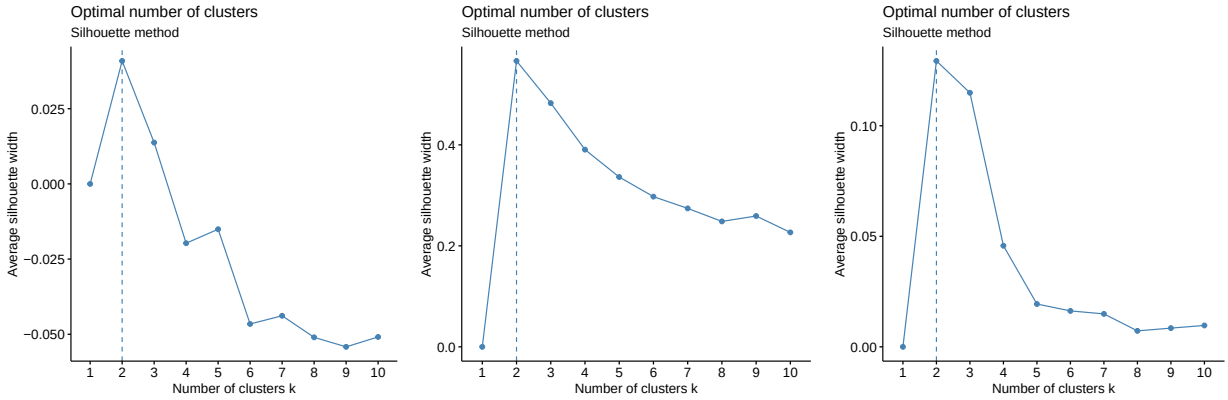


Figure S.3: Average Silhouette width v.s. K for Application 2, mouse neocortex data: sci-ATAC-Seq (Left), scRNA-Seq (Middle), and sc-methylation data (Right).

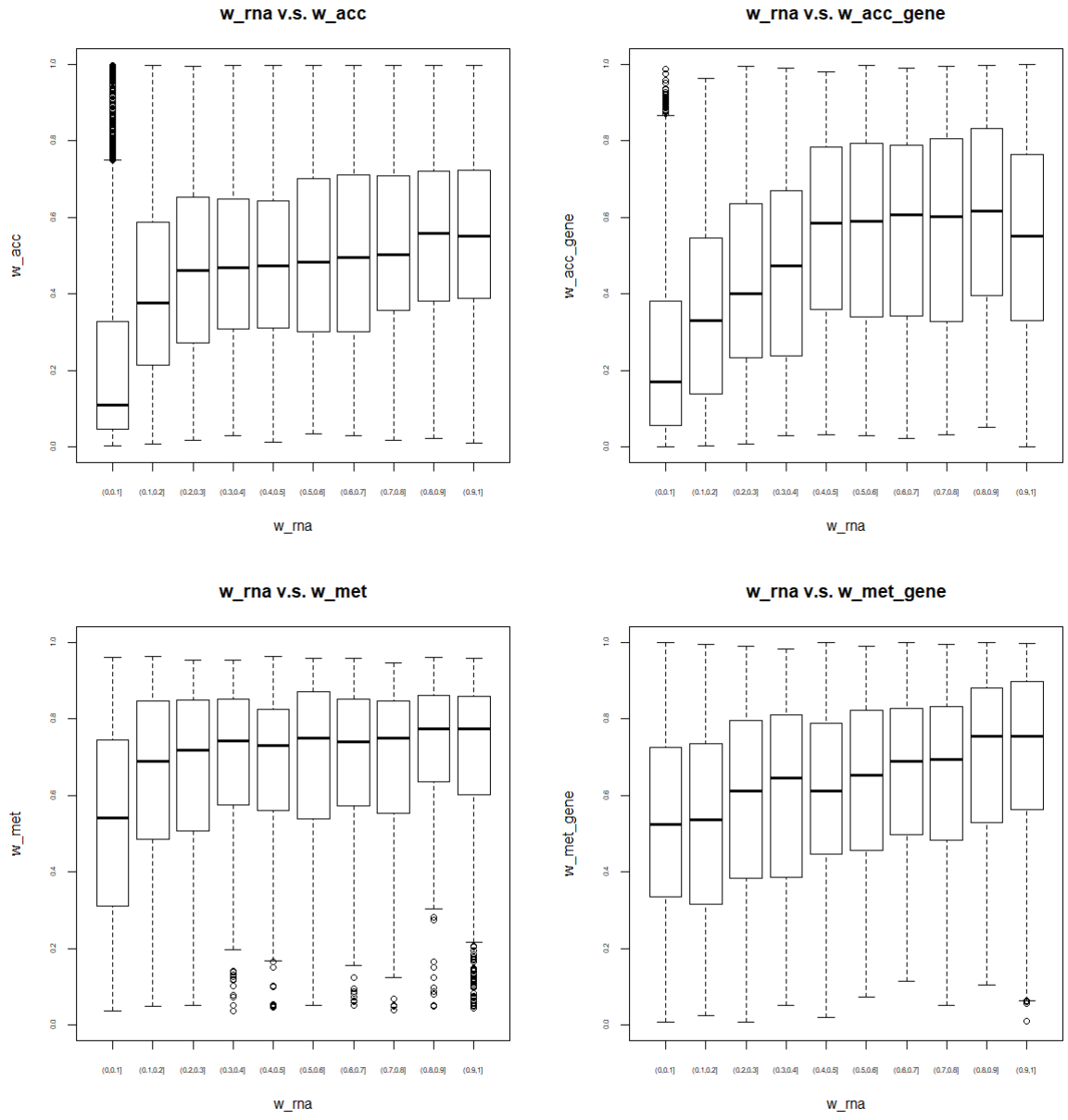


bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S.2.3 Empirical distribution of ω

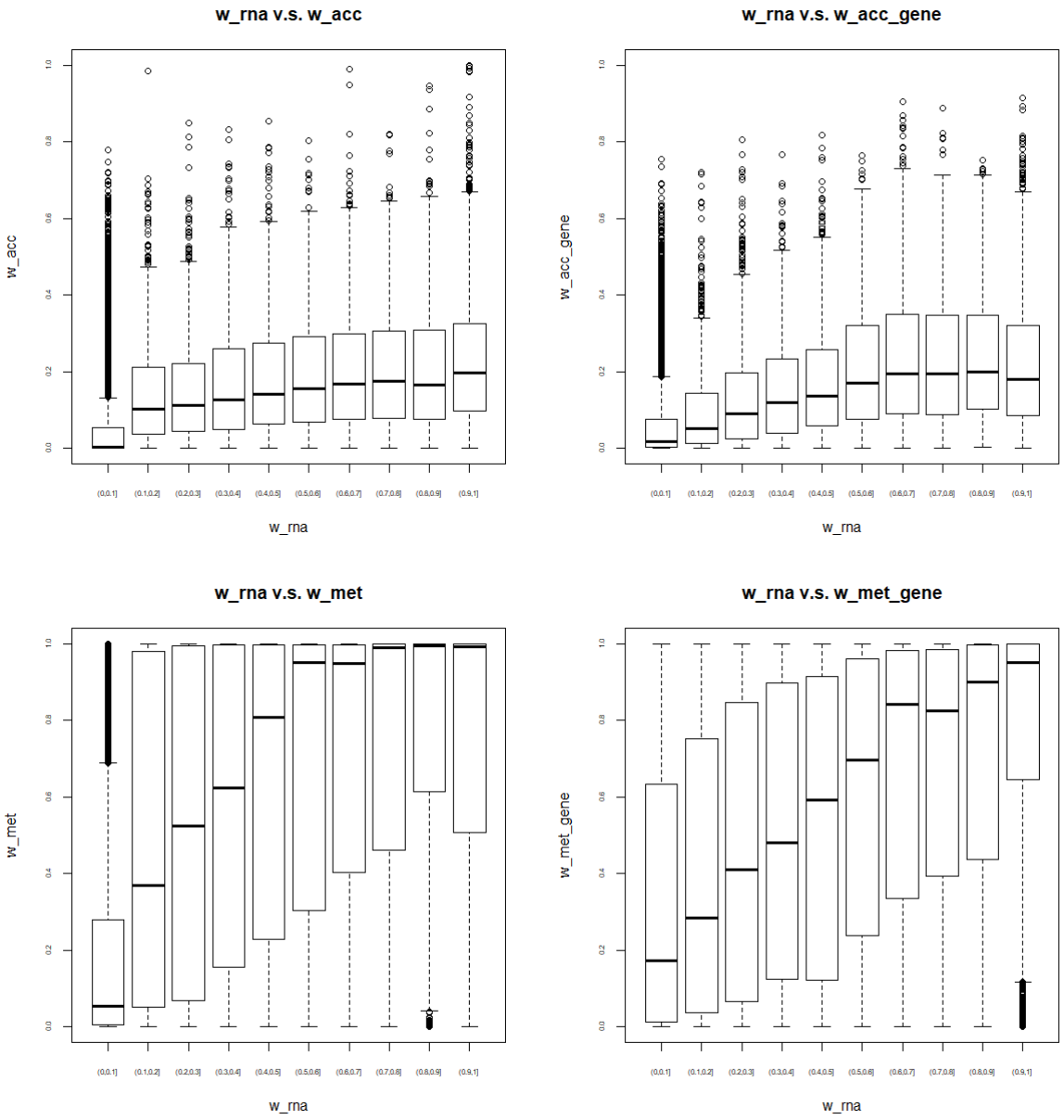
We apply the method mentioned in Section 2.6 to estimate $\eta, \gamma, \tau, \phi^{acc}, \delta, \theta, \phi^{met}$ by beta regression (Silvia and Francisco, 2004) before we run scAMACE. The empirical distributions of ω_{kg}^{acc} v.s. ω_{kg}^{rna} and ω_{kg}^{met} v.s. ω_{kg}^{rna} for two real data applications are shown in the Figures S.4 and S.5.

Figure S.4: Application 1: Empirical distribution of ω_{kg}^{acc} v.s. ω_{kg}^{rna} (Top left) and distribution generated by the model (Top right); Empirical distribution of ω_{kg}^{met} v.s. ω_{kg}^{rna} (Bottom left) and distribution generated by the model (Bottom right). $\hat{\eta} = -1.190, \hat{\gamma} = 4.376, \hat{\tau} = -3.036, \hat{\phi}^{acc} = 2.684, \hat{\delta} = 0.117, \hat{\theta} = 0.731, \hat{\phi}^{met} = 3.186$.



bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S.5: Application 2: Empirical distribution of ω_{kg}^{acc} v.s. ω_{kg}^{rna} (Top left) and distribution generated by the model (Top right); Empirical distribution of ω_{kg}^{met} v.s. ω_{kg}^{rna} (Bottom left) and distribution generated by the model (Bottom right). $\hat{\eta} = -2.713, \hat{\gamma} = 4.334, \hat{\tau} = -2.826, \hat{\phi}^{acc} = 4.799, \hat{\delta} = -0.735, \hat{\theta} = 2.008, \hat{\phi}^{met} = 0.804$.



bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S.3 Supplementary Tables

Table S.1: Supplementary clustering tables for K562, GM12878 scRNA-Seq, scATAC-Seq and sc-methylation data.

		LIGER					scMC			
		1	2	3	4	5	1	2	3	4
scATAC-Seq	GM12878	2	346	24	1	0	353			
	K562	649	0	15	0	2	611			
scRNA-Seq	GM12878	2	1	30	95	0	13		115	0
	K562	0	0	9	0	64	11		0	62
sc-methyl	GM12878			19				19		
	K562			11				11		

Table S.2: Supplementary comparison of the performance of different methods on the K562, GM12878 dataset by adjusted rand index (ARI).

	LIGER	scMC
scATAC-Seq	0.918	0.000
scRNA-Seq	0.603	0.771
sc-methyl	0.000	0.000

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S.3: Supplementary clustering tables for the mouse neocortex scRNA-Seq, sci-ATAC-Seq, and sc-methylation data.

		LIGER											
		1	2	3	4	5	6	7	8	9	10	11	12
sci-ATAC-Seq	Astro	145	184	31		58	12	34	16	10	37	6	0
	Ex. neurons CPN	582	112	194		135	51	58	87	36	49	3	7
	Oligo	156	62	51		64	45	25	14	3	13	17	0
scRNA-Seq	Astro	49	250	5		27	5	5	0	8	16	3	0
	L4	1153	82	66		1	13	16	13	14	5	38	0
	L6 CT	791	17	52		3	40	5	6	27	0	12	7
	Oligo	41	25	3		3	0	0	2	1	7	9	0
	Pvalb	1078	24	78		3	8	6	9	36	1	22	72
sc-methyl	L4	137	328	23	174	3	3	6	1	3	6	3	3
	L6-2	68	223	18	95	1	1	1	0	3	2	0	0
	Pvalb	79	27	2	41	2	0	0	0	0	0	1	2

		scMC						
		1	2	3	4	5	6	7
sci-ATAC-Seq	Astro	8	1	7	326	52	23	133
	Ex. neurons CPN	103	18	27	836	164	29	212
	Oligo	13	2	10	226	86	16	105
scRNA-Seq	Astro	0			362	4	1	1
	L4	1			1	1362	0	37
	L6 CT	0			0	959	0	1
	Oligo	0			14	58	1	18
	Pvalb	1			0	1331	0	5
sc-methyl	L4					1	689	
	L6-2					0	412	
	Pvalb					0	154	

Table S.4: Supplementary comparison of the performance of different methods on the mouse neocorex dataset by adjusted rand index (ARI).

	LIGER	scMC
sci-ATAC-Seq	0.052	0.019
scRNA-Seq	0.099	0.145
sc-methyl	0.031	0.001

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.29.437485; this version posted March 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S.5: Summary of the computation time by scAMACE and other clustering methods for Application 1. The unit of measurement is second. We implemented scAMACE (jointly on the three datastes, and seperately on the three datasets) by R and Python, run in 200 iterations. Seurat V3, scMC and LIGER are run in R by the downloaded R packages. Unless specified, all methods are implemented on one 3.4GHz Intel Xeon Gold CPU.

scAMACE (joint) by R 153.348	scAMACE (seperate) by R 91.200(scATAC-Seq)+21.619(scRNA-Seq)+3.282(sc-methyl)=116.101
scAMACE (joint) by Python 19.731	scAMACE (seperate) by Python 7.519(scATAC-Seq)+1.957(scRNA-Seq)+0.881(sc-methyl)=10.357
scAMACE (joint) by Python (using GPU) 7.640	scAMACE (seperate) by Python (using GPU) 1.971(scATAC-Seq)+0.671(scRNA-Seq)+0.364(sc-methyl)=3.006
Seurat V3 (scATAC-Seq+scRNA-Seq) 30.689	scMC(scATAC-Seq+scRNA-Seq+sc-methyl) 52.391
LIGER(scRNA-Seq+sc-methyl) 23.100	LIGER(scATAC-Seq+scRNA-Seq+sc-methyl) 73.788

Table S.6: Summary of the computation time by scAMACE and other clustering methods for Application 2. The unit of measurement is second. We implemented scAMACE (jointly on the three datastes, and seperately on the three datasets) by R and Python, run in 200 iterations. Seurat V3, scMC and LIGER are run in R by the downloaded R packages. Unless specified, all methods are implemented on one 3.4GHz Intel Xeon Gold CPU.

scAMACE (joint) by R 2317.787	scAMACE (seperate) by R 249.891(sci-ATAC-Seq)+950.388(scRNA-Seq)+171.823(sc-methyl)=1372.102
scAMACE (joint) by Python 418.858	scAMACE (seperate) by Python 51.878(sci-ATAC-Seq)+217.728(scRNA-Seq)+25.994(sc-methyl)=295.6
scAMACE (joint) by Python (using GPU) 69.652	scAMACE (seperate) by Python (using GPU) 7.651(sci-ATAC-Seq)+33.435(scRNA-Seq)+4.161(sc-methyl)=45.247
Seurat V3 (sci-ATAC-Seq+scRNA-Seq) 116.688	scMC(sci-ATAC-Seq+scRNA-Seq+sc-methyl) 372.323
LIGER(scRNA-Seq+sc-methyl) 1618.965	LIGER(sci-ATAC-Seq+scRNA-Seq+sc-methyl) 80.389

Table S.7: Summary of the computation time by scAMACE and other clustering methods using 30,000 bootstrap samples ($n_{acc} = n_{rna} = n_{met} = 10,000$) from Application 2. The unit of measurement is second. We implemented scAMACE (jointly on the three datastes, and seperately on the three datasets) by R and Python, run in 200 iterations. Seurat V3, scMC and LIGER are run in R by the downloaded R packages. Unless specified, all methods are implemented on one 3.4GHz Intel Xeon Gold CPU.

scAMACE (joint) by R 7663.651	scAMACE (seperate) by R 1366.109(sci-ATAC-Seq)+2570.669(scRNA-Seq)+1391.464(sc-methyl)=5328.242
scAMACE (joint) by Python 1534.631	scAMACE (seperate) by Python 307.066(sci-ATAC-Seq)+548.065(scRNA-Seq)+373.092(sc-methyl)=1228.223
scAMACE (joint) by Python (using GPU) 250.089	scAMACE (seperate) by Python (using GPU) 64.549(sci-ATAC-Seq)+82.53(scRNA-Seq)+49.898(sc-methyl)=196.977
Seurat V3 (sci-ATAC-Seq+scRNA-Seq) 290.640	scMC(sci-ATAC-Seq+scRNA-Seq+sc-methyl) 3667.878
LIGER(scRNA-Seq+sc-methyl) 5319.259	LIGER(sci-ATAC-Seq+scRNA-Seq+sc-methyl) 555.574