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, Hong Kong SAR, China

SUPPLEMENTARY MATERIALS

1 Supplementary Text

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

$$\boldsymbol{P}(\boldsymbol{Y}, \boldsymbol{U}_{l.}, \boldsymbol{V}_{l.}, \boldsymbol{Z}_{l.} | \boldsymbol{\psi}^{rna}, \boldsymbol{\omega}^{rna}, \boldsymbol{\pi}_l) = \prod_l \prod_k [\psi_k^{rna} * \boldsymbol{A}]^{z_{lk}},$$

$$\boldsymbol{A} = \prod_g \left\{ \omega_{kg}^{rna} \left[\pi_{l1} g_1 \right]^{v_{lg}} * \left[(1-\pi_{l1}) g_0 \right]^{1-v_{lg}} \right\}^{u_{lg}} * \left\{ (1-\omega_{kg}^{rna}) \left[\pi_{l0} g_1 \right]^{v_{lg}} * \left[(1-\pi_{l0}) g_0 \right]^{1-v_{lg}} \right\}^{1-u_{lg}}.$$

- sc-methylation data

$$\boldsymbol{P}(\boldsymbol{T}, \boldsymbol{U}_{d.}, \boldsymbol{M}_{d.}, \boldsymbol{Z}_{d.} | \boldsymbol{\psi}^{met}, \boldsymbol{\omega}^{met}, \boldsymbol{\pi}_d) = \prod_d \prod_k [\psi_k^{met} * \boldsymbol{B}]^{z_{dk}},$$

$$\boldsymbol{B} = \prod_g \left\{ \omega_{kg}^{met} (\pi_{d1} h_1)^{m_{dg}} * \left[(1-\pi_{d1}) h_0 \right]^{1-m_{dg}} \right\}^{u_{dg}} * \left\{ (1-\omega_{kg}^{met}) (\pi_{d0} h_1)^{m_{dg}} * \left[(1-\pi_{d0}) h_0 \right]^{1-m_{dg}} \right\}^{1-u_{dg}}.$$

*Corresponding author: zhixianglin@cuhk.edu.hk

1.2 Q-function

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

$$
\begin{aligned}
&ln(\boldsymbol{P}(\mathbf{\Phi}, \mathbf{\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 param-

eters.

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

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

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

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

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

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

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

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 numbers of clusters $K_x = K_y = K_t = 3$. $f(\omega_{kg}^y) = \eta + \gamma\omega_{kg}^y + \tau(\omega_{kg}^y)^2 = -1 + 7\omega_{kg}^y - 2(\omega_{kg}^y)^2$, $g(\omega_{kg}^y) = \delta + \theta\omega_{kg}^y = -2 + 5\omega_{kg}^y$, $\phi^x = 10$ 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 $z_{ik} = 1$.

- 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 $m_{dg} = 0$.

We set different parameter values for the four additional simulation settings mentioned in Section 4 as following:

(1) Simulation setting 1: imbalanced dataset, where the numbers of cells, n_x, n_y, and n_t are different across modalities (Table S.1). Data is generated as described above, but we set $n_x = 1000$, $n_y = 2000$, and $n_t = 500$.

(2) Simulation setting 2: unequal number of clusters, where the numbers of clusters in the three modalities, K_x, K_y, K_t are different (Table S.2). Data is generated as described above, but we applied following scheme to generate $\boldsymbol{\omega}^y$ and $\boldsymbol{\psi}_x, \boldsymbol{\psi}_y, \boldsymbol{\psi}_t$ so that $K_x = 3, K_y = 7, K_t = 5$.

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

For $g = 1, ..., 350$:

$$\omega^y_{kg} \sim \begin{array}{c} \begin{matrix} (1,\ldots,50) & (51,\ldots,100) & (101,\ldots,150) & (151,\ldots,200) & (201,\ldots,250) & (251,\ldots,300) & (301,\ldots,350) \end{matrix} \\ \begin{bmatrix} 0.800 & 0.733 & 0.533 & 0.500 & 0.467 & 0.267 & 0.200 \\ 0.733 & 0.533 & 0.500 & 0.467 & 0.267 & 0.200 & 0.800 \\ 0.533 & 0.500 & 0.467 & 0.267 & 0.200 & 0.800 & 0.733 \\ 0.500 & 0.467 & 0.267 & 0.200 & 0.800 & 0.733 & 0.533 \\ 0.467 & 0.267 & 0.200 & 0.800 & 0.733 & 0.533 & 0.500 \\ 0.267 & 0.200 & 0.800 & 0.733 & 0.533 & 0.500 & 0.567 \\ 0.200 & 0.800 & 0.733 & 0.533 & 0.500 & 0.467 & 0.267 \end{bmatrix} \end{array}$$

For $g = 351, ..., 1000$:

$$\omega_{kg}^{y} \sim \begin{cases} Beta\ (\alpha = 2, \beta = 2), & \text{for } k = 1 \\ \omega_{1g}^{y}, & \text{for } k = 2, \ldots, 7 \end{cases}$$

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

For $g = 1, ..., 350$:

$\omega_{kg}^{x} \sim Beta\ (\mu_{kg}^{x} = \frac{1}{1+e^{-f(\omega_{kg}^{y})}}, \phi^x)$, for $k = 1, \ldots, 7$

For $g = 351, ..., 1000$:

$$\omega_{kg}^{x} \sim \begin{cases} Beta\ (\mu_{kg}^{x} = \frac{1}{1+e^{-f(\omega_{1g}^{y})}}, \phi^x), & \text{for } k = 1 \\ \omega_{1g}^{x}, & \text{for } k = 2, \ldots, 7 \end{cases}$$

For $g = 1, ..., 350$:

$\omega_{kg}^{t} \sim Beta\ (\mu_{kg}^{t} = \frac{1}{1+e^{-g(\omega_{kg}^{y})}}, \phi^t)$, for $k = 1, \ldots, 7$

For $g = 351, ..., 1000$:

$$\omega_{kg}^{t} \sim \begin{cases} Beta\ (\mu_{kg}^{t} = \frac{1}{1+e^{-g(\omega_{1g}^{y})}}, \phi^t), & \text{for } k = 1 \\ \omega_{1g}^{t}, & \text{for } k = 2, \ldots, 7 \end{cases}$$

C. Generate $\boldsymbol{z}^x$, $\boldsymbol{z}^y$ and $\boldsymbol{z}^t$. The cluster labels are generated by $\boldsymbol{\psi}_x = (0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0)$, $\boldsymbol{\psi}_y = (\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7})$, and $\boldsymbol{\psi}_t = (0, 0, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$ so that $K_x = 3, K_y = 7, K_t = 5$.

The remaining steps are the same as described above.

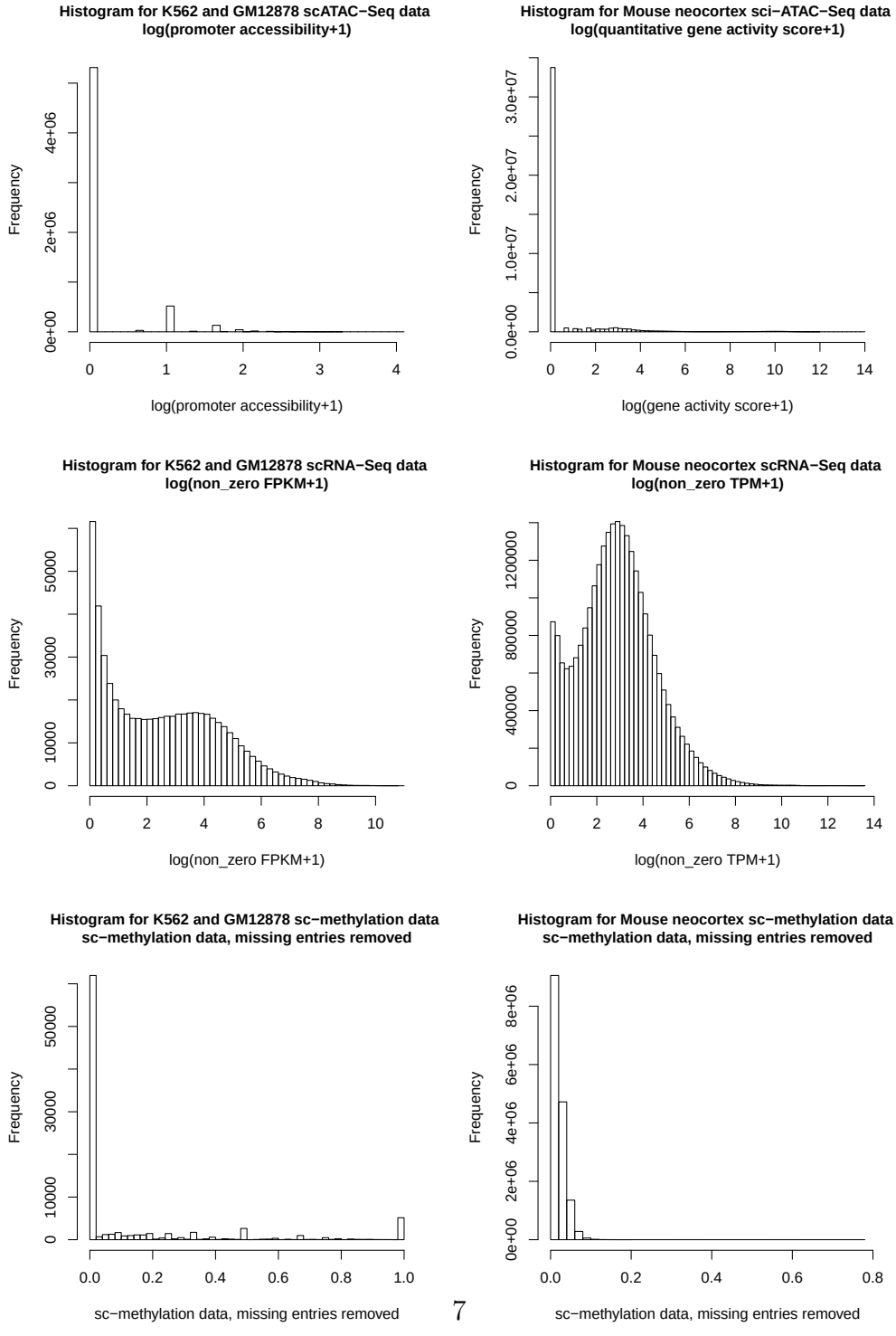
(3) Simulation setting 3: imbalanced cluster sizes, where the proportions of different cell types in the three modalities, $\boldsymbol{\psi}_x$, $\boldsymbol{\psi}_y$, $\boldsymbol{\psi}_t$ are different (Table S.3). Data is generated as described above, but we set $\boldsymbol{\psi}_x = (0.3, 0.1, 0.6), \boldsymbol{\psi}_y = (0.6, 0.3, 0.1), \boldsymbol{\psi}_t = (0.6, 0.1, 0.3)$. There are rare cell types (10% of the cells) in the three modalities.

(4) Simulation setting 4: smaller number of features (Table S.4). Data is generated as described above, but we set $p = 500$.

2 Supplementary Figures

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).



2.2 Determination of number of clusters K for real data applications

We applied the Silhouette method [1] 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 [2], including astrocytes, oligodendrocytes, and three subtypes of neurons. There are three cell types in sci-ATAC-Seq data [3], including astrocytes, oligodendrocytes, and excitatory neurons CPN. And there are three cell types in sc-methylation dataset [4], 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).

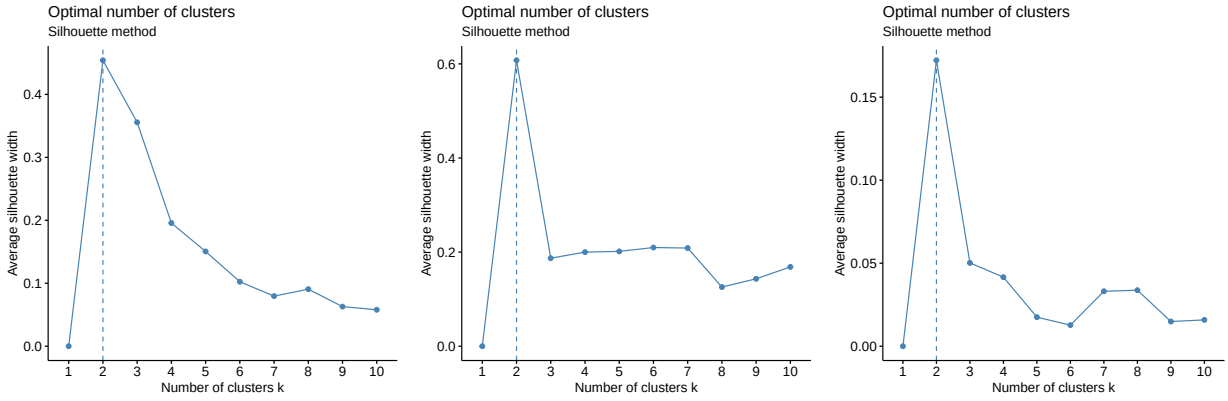
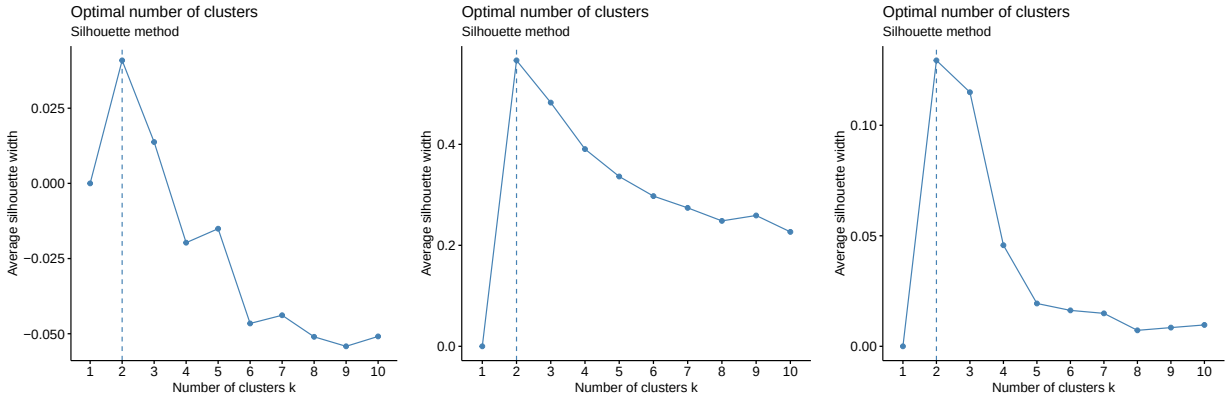


Figure S.3: Average Silhouette width v.s. K for Application 2, mouse neocortex data: sci-ATAC-Seq (Left), scRNA-Seq (Middle), and sc-methylation data (Right).



2.3 Empirical distribution of $\boldsymbol{\omega}$

To assess whether the linear and quadratic models work well, we performed the following steps. We first obtain $\hat{\omega}_{kg}^{acc}$, $\hat{\omega}_{kg}^{rna}$, $\hat{\omega}_{kg}^{met}$ as described in Section 2.6 by setting $K = 1$ and fitting the model on the three modalities separately. We plotted the distributions of $\hat{\omega}_{kg}^{rna}$ v.s. $\hat{\omega}_{kg}^{acc}$ and $\hat{\omega}_{kg}^{rna}$ v.s. $\hat{\omega}_{kg}^{met}$ for the two real data applications in the left panels in Figures S.4 and S.5, respectively. For better visualization on how $\hat{\omega}_{kg}^{acc}$ and $\hat{\omega}_{kg}^{met}$ change with $\hat{\omega}_{kg}^{rna}$, we plotted the boxplots of $\hat{\omega}_{kg}^{acc}$ and $\hat{\omega}_{kg}^{met}$ in different ranges of $\hat{\omega}_{kg}^{rna}$. We estimated $\{\eta, \gamma, \tau, \delta, \theta, \phi^{acc}, \phi^{met}\}$ by beta regression [5] using $\hat{\omega}_{kg}^{acc}$, $\hat{\omega}_{kg}^{rna}$, $\hat{\omega}_{kg}^{met}$. Using $\hat{\omega}_{kg}^{rna}$ and $\{\hat{\eta}, \hat{\gamma}, \hat{\tau}, \hat{\delta}, \hat{\theta}, \hat{\phi}^{acc}, \hat{\phi}^{met}\}$, we then generated ω_{kg}^{acc} and ω_{kg}^{met} by random sampling following the quadratic and linear models. The distributions of the generated ω_{kg}^{acc} and ω_{kg}^{met} are plotted in the right panels in Figures S.4 and S.5, respectively. In comparison of the generated values (left panels in Figures S.4 and S.5) with the estimated $\hat{\omega}_{kg}^{acc}$ and $\hat{\omega}_{kg}^{met}$ (right panels in Figures S.4 and S.5), we can see that the linear and quadratic models capture the trends on how $\hat{\omega}_{kg}^{acc}$ and $\hat{\omega}_{kg}^{met}$ changes with $\hat{\omega}_{kg}^{rna}$.

Figure S.4: Application 1: Distribution of $\hat{\omega}_{kg}^{rna}$ v.s. $\hat{\omega}_{kg}^{acc}$ where $\hat{\omega}_{kg}^{rna}$ and $\hat{\omega}_{kg}^{acc}$ are obtained by setting $K = 1$ and fitting the model on scRNA-Seq data and scCAS data seperately (Top left); Distribution of $\hat{\omega}_{kg}^{rna}$ v.s. generated ω_{kg}^{acc} by random sampling from the quadratic model with $\{\hat{\eta}, \hat{\gamma}, \hat{\tau}, \hat{\phi}^{acc}\}$ and $\hat{\omega}_{kg}^{rna}$ (Top right). Distribution of $\hat{\omega}_{kg}^{rna}$ v.s. $\hat{\omega}_{kg}^{met}$ where $\hat{\omega}_{kg}^{rna}$ and $\hat{\omega}_{kg}^{met}$ are obtained by setting $K = 1$ and fitting the model on scRNA-Seq data and sc-methylation data seperately (Bottom left); Distribution of $\hat{\omega}_{kg}^{rna}$ v.s. generated ω_{kg}^{met} by random sampling from the linear model with $\{\hat{\delta}, \hat{\theta}, \hat{\phi}^{met}\}$ and $\hat{\omega}_{kg}^{rna}$ (Bottom right). The estimated values $\hat{\eta} = -1.190, \hat{\gamma} = 4.376, \hat{\tau} = -3.036, \hat{\phi}^{acc} = 2.684, \hat{\delta} = 0.117, \hat{\theta} = 0.731, \hat{\phi}^{met} = 3.186$.

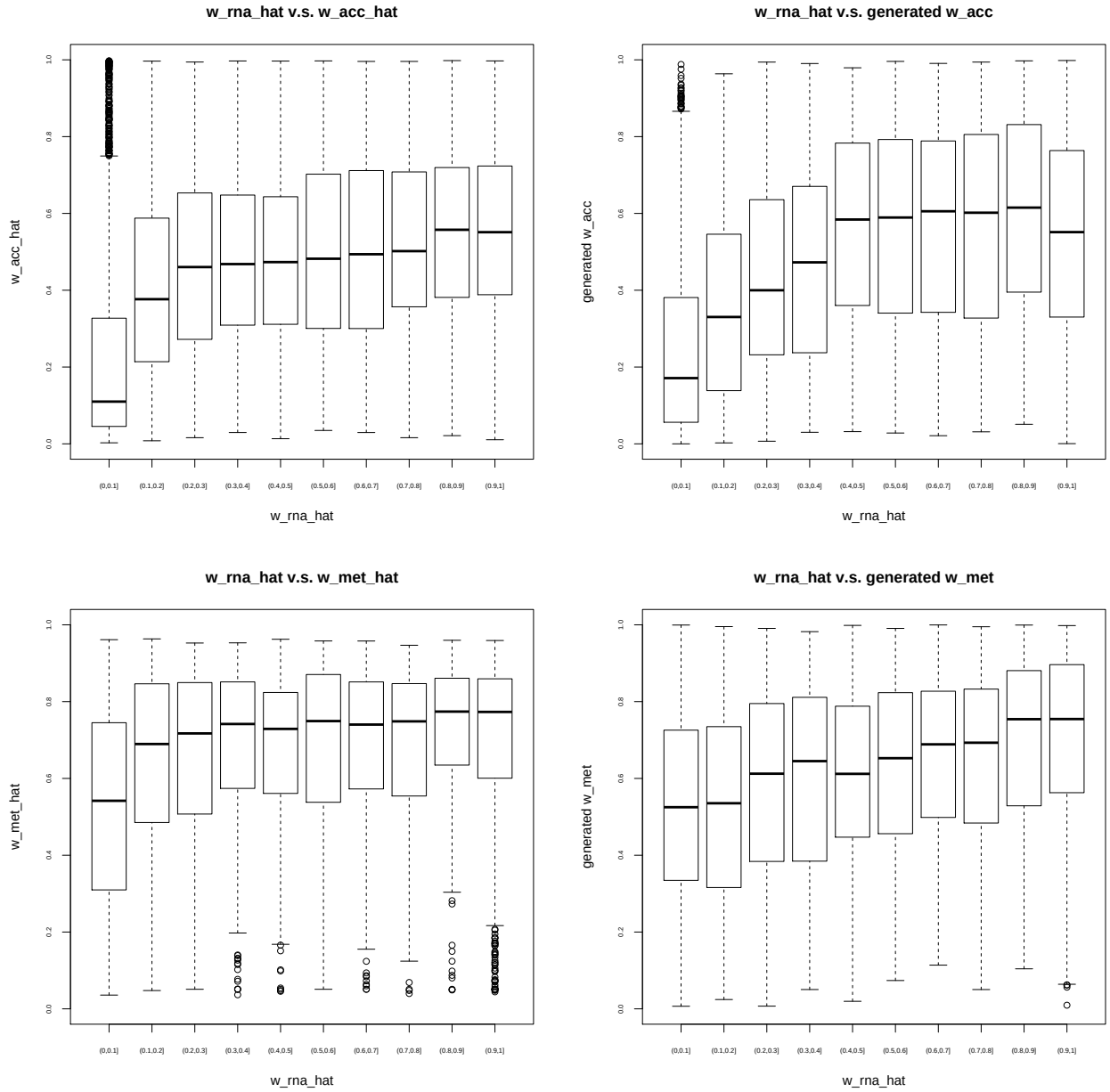
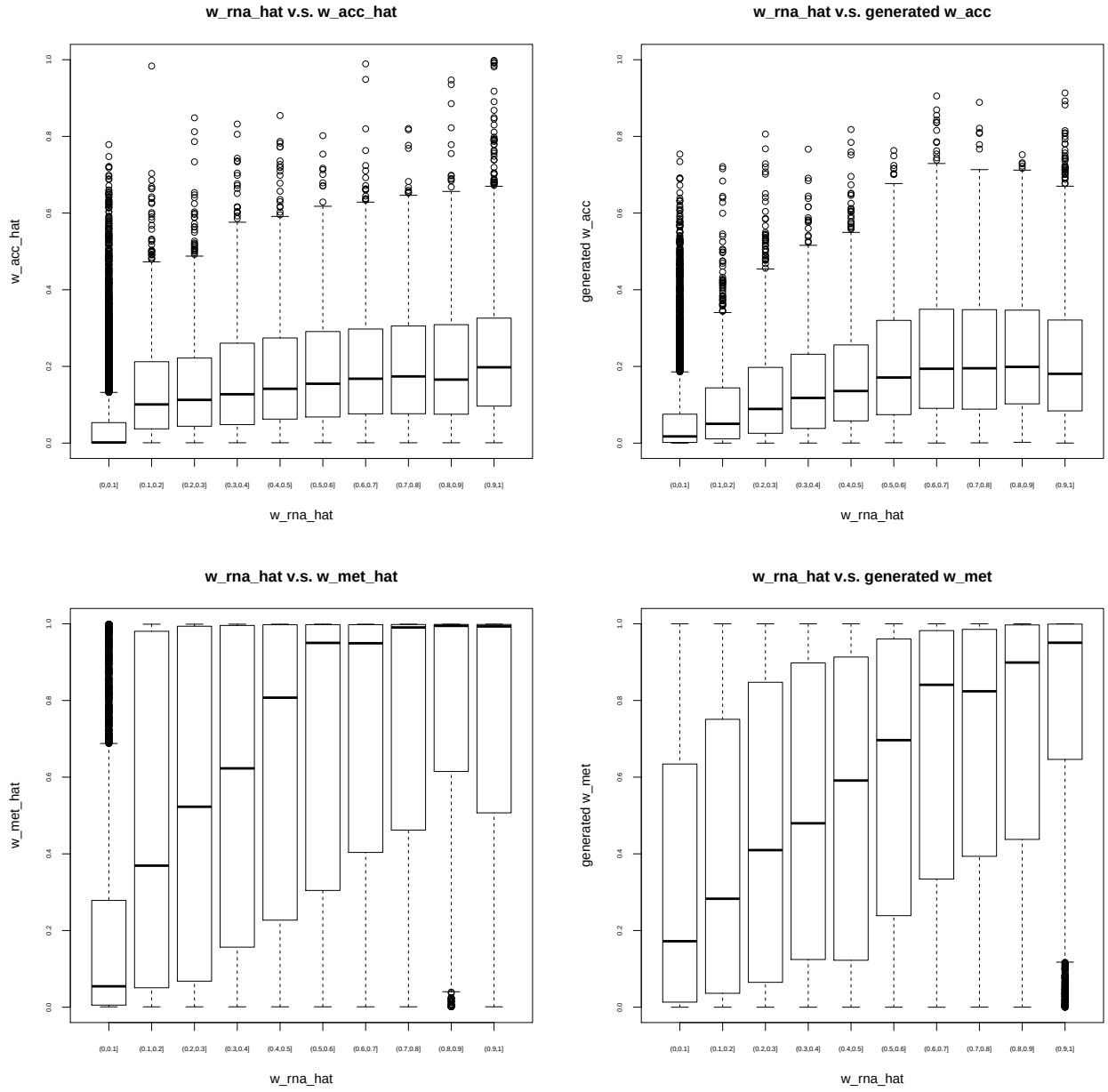


Figure S.5: Application 2: Distribution of $\hat{\omega}_{kg}^{rna}$ v.s. $\hat{\omega}_{kg}^{acc}$ where $\hat{\omega}_{kg}^{rna}$ and $\hat{\omega}_{kg}^{acc}$ are obtained by setting $K = 1$ and fitting the model on scRNA-Seq data and scCAS data seperately (Top left); Distribution of $\hat{\omega}_{kg}^{rna}$ v.s. generated ω_{kg}^{acc} by random sampling from the quadratic model with $\{\hat{\eta}, \hat{\gamma}, \hat{\tau}, \hat{\phi}^{acc}\}$ and $\hat{\omega}_{kg}^{rna}$ (Top right). Distribution of $\hat{\omega}_{kg}^{rna}$ v.s. $\hat{\omega}_{kg}^{met}$ where $\hat{\omega}_{kg}^{rna}$ and $\hat{\omega}_{kg}^{met}$ are obtained by setting $K = 1$ and fitting the model on scRNA-Seq data and sc-methylation data seperately (Bottom left); Distribution of $\hat{\omega}_{kg}^{rna}$ v.s. generated ω_{kg}^{met} by random sampling from the linear model with $\{\hat{\delta}, \hat{\theta}, \hat{\phi}^{met}\}$ and $\hat{\omega}_{kg}^{rna}$ (Bottom right). The estimated values $\hat{\eta} = -2.713, \hat{\gamma} = 4.334, \hat{\tau} = -2.826, \hat{\phi}^{acc} = 4.799, \hat{\delta} = -0.735, \hat{\theta} = 2.008, \hat{\phi}^{met} = 0.804$.



3 Supplementary Tables

Table S.1: Simulation setting 1: imbalanced datasets across the three modalities. Data is generated as described in Section 4, but we set the numbers of cells in the three modalities, n_x, n_y, and n_t to be different. Mean and sd (in parentheses) of purity, rand index, adjusted rand index (ARI) and normalized mutual information (NMI) for 50 independent runs are shown.

		Data type	Purity	Rand Index	ARI	NMI
$n_x = 1000$ $n_y = 2000$ $n_t = 500$	scAMACE (joint)	$\boldsymbol{x}$	0.697(0.022)	0.688(0.017)	0.299(0.037)	0.254(0.031)
		$\boldsymbol{y}$	0.911(0.006)	0.890(0.007)	0.752(0.016)	0.673(0.017)
		$\boldsymbol{t}$	0.705(0.025)	0.693(0.019)	0.311(0.043)	0.268(0.035)
	scAMACE (seperate)	$\boldsymbol{x}$	0.658(0.023)	0.661(0.015)	0.239(0.034)	0.203(0.029)
		$\boldsymbol{y}$	0.872(0.008)	0.846(0.008)	0.652(0.018)	0.572(0.018)
		$\boldsymbol{t}$	0.650(0.031)	0.655(0.019)	0.227(0.043)	0.196(0.036)
	K-Means	$\boldsymbol{x}$	0.382(0.019)	0.558(0.003)	0.007(0.007)	0.008(0.006)
		$\boldsymbol{y}$	0.811(0.011)	0.784(0.010)	0.514(0.023)	0.442(0.021)
		$\boldsymbol{t}$	0.394(0.024)	0.559(0.005)	0.009(0.009)	0.011(0.009)
	Hierarchical Clustering	$\boldsymbol{x}$	0.358(0.008)	0.493(0.045)	0.001(0.001)	0.002(0.002)
		$\boldsymbol{y}$	0.356(0.007)	0.526(0.023)	0.001(0.001)	0.002(0.001)
		$\boldsymbol{t}$	0.372(0.014)	0.524(0.033)	0.002(0.002)	0.005(0.004)
	Spectral Clustering	$\boldsymbol{x}$	0.390(0.022)	0.560(0.004)	0.011(0.009)	0.012(0.009)
		$\boldsymbol{y}$	0.806(0.010)	0.779(0.010)	0.503(0.021)	0.431(0.019)
		$\boldsymbol{t}$	0.399(0.025)	0.560(0.005)	0.010(0.010)	0.013(0.009)

Table S.2: Simulation setting 2: unequal number of clusters across the three modalities. Data is generated as described in Section 4, but we set the numbers of clusters in the three modalities, K_x, K_y, K_t to be different. Mean and sd (in parentheses) of purity, rand index, adjusted rand index (ARI) and normalized mutual information (NMI) for 50 independent runs are shown.

		Data type	Purity	Rand Index	ARI	NMI
$K_x = 3$ $K_y = 7$ $K_t = 5$	scAMACE (joint)	$\boldsymbol{x}$	0.727(0.023)	0.714(0.017)	0.353(0.039)	0.291(0.032)
		$\boldsymbol{y}$	0.775(0.015)	0.889(0.006)	0.549(0.024)	0.559(0.022)
		$\boldsymbol{t}$	0.628(0.019)	0.773(0.008)	0.325(0.076)	0.291(0.020)
	scAMACE (seperate)	$\boldsymbol{x}$	0.717(0.020)	0.705(0.014)	0.322(0.033)	0.231(0.024)
		$\boldsymbol{y}$	0.710(0.018)	0.864(0.007)	0.446(0.027)	0.466(0.026)
		$\boldsymbol{t}$	0.591(0.017)	0.762(0.006)	0.241(0.020)	0.236(0.018)
	K-Means	$\boldsymbol{x}$	0.416(0.027)	0.566(0.008)	0.285(0.016)	0.024(0.016)
		$\boldsymbol{y}$	0.282(0.024)	0.767(0.004)	0.050(0.016)	0.079(0.020)
		$\boldsymbol{t}$	0.276(0.016)	0.684(0.002)	0.190(0.008)	0.023(0.008)
	Hierarchical Clustering	$\boldsymbol{x}$	0.363(0.012)	0.485(0.050)	0.372(0.062)	0.003(0.002)
		$\boldsymbol{y}$	0.188(0.006)	0.704(0.038)	0.002(0.001)	0.012(0.003)
		$\boldsymbol{t}$	0.240(0.009)	0.637(0.029)	0.215(0.024)	0.007(0.002)
	Spectral Clustering	$\boldsymbol{x}$	0.434(0.031)	0.573(0.010)	0.292(0.017)	0.039(0.022)
		$\boldsymbol{y}$	0.323(0.022)	0.774(0.004)	0.079(0.014)	0.121(0.018)
		$\boldsymbol{t}$	0.295(0.019)	0.688(0.003)	0.197(0.009)	0.036(0.010)

Table S.3: Simulation setting 3: imbalanced cluster sizes across the three modalities. Data is generated as described in Section 4, but we set the proportions different cell types across the three modalities, $\boldsymbol{\psi}_x$, $\boldsymbol{\psi}_y$, and $\boldsymbol{\psi}_t$ to be different. Mean and sd (in parentheses) of purity, rand index, adjusted rand index (ARI) and normalized mutual information (NMI) for 50 independent runs are shown.

		Data type	Purity	Rand Index	ARI	NMI
$\boldsymbol{\psi}_x = (0.3, 0.1, 0.6)$ $\boldsymbol{\psi}_y = (0.6, 0.3, 0.1)$ $\boldsymbol{\psi}_t = (0.6, 0.1, 0.3)$	scAMACE (joint)	$\boldsymbol{x}$	0.732(0.023)	0.680(0.023)	0.350(0.047)	0.238(0.034)
		$\boldsymbol{y}$	0.908(0.010)	0.872(0.013)	0.743(0.026)	0.603(0.030)
		$\boldsymbol{t}$	0.756(0.017)	0.689(0.017)	0.376(0.034)	0.243(0.030)
	scAMACE (seperate)	$\boldsymbol{x}$	0.729(0.025)	0.673(0.024)	0.336(0.050)	0.222(0.038)
		$\boldsymbol{y}$	0.870(0.012)	0.823(0.015)	0.644(0.030)	0.499(0.029)
		$\boldsymbol{t}$	0.729(0.017)	0.665(0.015)	0.326(0.031)	0.208(0.026)
	K-Means	$\boldsymbol{x}$	0.601(0.016)	0.516(0.006)	0.007(0.006)	0.008(0.006)
		$\boldsymbol{y}$	0.769(0.021)	0.625(0.016)	0.230(0.031)	0.198(0.028)
		$\boldsymbol{t}$	0.598(0.015)	0.519(0.007)	0.011(0.001)	0.011(0.008)
	Hierarchical Clustering	$\boldsymbol{x}$	0.601(0.016)	0.500(0.012)	0.007(0.007)	0.002(0.002)
		$\boldsymbol{y}$	0.600(0.014)	0.505(0.008)	0.006(0.005)	0.003(0.002)
		$\boldsymbol{t}$	0.598(0.015)	0.506(0.010)	0.008(0.008)	0.003(0.002)
	Spectral Clustering	$\boldsymbol{x}$	0.601(0.016)	0.519(0.007)	0.011(0.010)	0.011(0.008)
		$\boldsymbol{y}$	0.754(0.034)	0.620(0.017)	0.220(0.031)	0.189(0.026)
		$\boldsymbol{t}$	0.601(0.017)	0.525(0.009)	0.022(0.015)	0.021(0.013)

Table S.4: Simulation setting 4: smaller number of features. Data is generated as described in Section 4, but we set the number of features p to be smaller. Mean and sd (in parentheses) of purity, rand index, adjusted rand index (ARI) and normalized mutual information (NMI) for 50 independent runs are shown.

		Data type	Purity	Rand Index	ARI	NMI
$p = 500$	scAMACE (joint)	$\boldsymbol{x}$	0.557(0.029)	0.605(0.013)	0.115(0.029)	0.101(0.025)
		$\boldsymbol{y}$	0.770(0.012)	0.746(0.010)	0.429(0.023)	0.368(0.020)
		$\boldsymbol{t}$	0.564(0.024)	0.603(0.014)	0.120(0.025)	0.107(0.020)
	scAMACE (seperate)	$\boldsymbol{x}$	0.507(0.027)	0.585(0.010)	0.070(0.022)	0.063(0.0198)
		$\boldsymbol{y}$	0.658(0.021)	0.660(0.014)	0.238(0.030)	0.205(0.025)
		$\boldsymbol{t}$	0.511(0.023)	0.581(0.011)	0.072(0.019)	0.065(0.016)
	K-Means	$\boldsymbol{x}$	0.373(0.013)	0.557(0.002)	0.003(0.004)	0.005(0.004)
		$\boldsymbol{y}$	0.461(0.045)	0.577(0.013)	0.049(0.028)	0.045(0.024)
		$\boldsymbol{t}$	0.377(0.017)	0.557(0.002)	0.005(0.005)	0.006(0.004)
	Hierarchical Clustering	$\boldsymbol{x}$	0.359(0.010)	0.472(0.045)	0.001(0.001)	0.003(0.002)
		$\boldsymbol{y}$	0.359(0.009)	0.501(0.038)	0.001(0.001)	0.002(0.002)
		$\boldsymbol{t}$	0.360(0.010)	0.527(0.023)	0.001(0.001)	0.002(0.001)
	Spectral Clustering	$\boldsymbol{x}$	0.374(0.015)	0.557(0.002)	0.004(0.004)	0.005(0.004)
		$\boldsymbol{y}$	0.501(0.052)	0.591(0.015)	0.079(0.034)	0.072(0.028)
		$\boldsymbol{t}$	0.379(0.002)	0.558(0.002)	0.004(0.004)	0.006(0.004)

Table S.5: Clustering tables for K562, GM12878 scRNA-Seq, scATAC-Seq and sc-methylation data before and after the transformation on sc-methylation data.

		scAMACE (joint)						scAMACE (seperate)					
		Before transformation			After transformation			Before transformation			After transformation		
		1	2	ARI	1	2	ARI	1	2	ARI	1	2	ARI
scATAC-Seq	GM12878	368	5	0.958	368	5	0.958	254	119	0.192	254	119	0.192
	K562	6	660		6	660		171	495		171	495	
scRNA-Seq	GM12878	128	0	1.000	128	0	1.000	128	0	1.000	128	0	1.000
	K562	0	73		0	73		0	73		0	73	
sc-methyl	GM12878		19	0.000	16	3	0.628		19	0.000	7	12	0.260
	K562		11		0	11			11		11	0	

Table S.6: Clustering tables for the mouse neocortex scRNA-Seq, sci-ATAC-Seq, and sc-methylation data before and after the transformation on sc-methylation data.

		scAMACE (joint)											
		Before transformation						After transformation					
		1	2	3	4	5	ARI	1	2	3	4	5	ARI
sci-ATAC-Seq	Astro	550	0		1		0.998	550	0		1		0.998
	Ex. neurons CPN	0	1391		0			0	1391		0		
	Oligo	0	1		457			0	1		457		
scRNA-Seq	Astro	368	0	0	0	0	0.997	368	0	0	0	0	0.997
	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		25	0	0	66	0	
	Pvalb	0	0	0	0	1337		0	0	0	0	1337	
sc-methyl	L4					412	0.000		411	1	0	0	0.932
	L6-2					729			20	703	6	0	
	Pvalb					154			0	0	1	153	

		scAMACE (seperate)											
		Before transformation						After transformation					
		1	2	3	4	5	ARI	1	2	3	4	5	ARI
sci-ATAC-Seq	Astro	550	0	1			0.998	550	0	1			0.998
	Ex. neurons CPN	1	1390	0				1	1390	0			
	Oligo	0	0	458				0	0	458			
scRNA-Seq	Astro	368	0	0	0	0	0.997	368	0	0	0	0	0.997
	L4	0	1401	0	0	0		0	1401	0	0	0	
	L6 CT	0	0	960	0	0		0	0	960	0	0	
	Oligo	27	0	0	64	0		27	0	0	64	0	
	Pvalb	0	0	0	0	1337		0	0	0	0	1337	
sc-methyl	L4	412					0.000	412					0.000
	L6-2	729						729					
	Pvalb	154						154					

Table S.7: 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.8: 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

Table S.9: 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.10: 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

Table S.11: 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.12: 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.13: 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

References

[1] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley, 1990.

[2] Bosiljka Tasic, Zizhen Yao, Lucas T. Graybuck, Kimberly A. Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N. Economo, Sarada Viswanathan, Osnat Penn, Trygve Bakken, Vilas Menon, Jeremy Miller, Olivia Fong, Karla E. Hirokawa, Kanan Lathia, Christine Rimorin, Michael Tieu, Rachael Larsen, Tamara Casper, Eliza Barkan, Matthew Kroll, Sheana Parry, Nadiya V. Shapovalova, Daniel Hirschstein, Julie Pendergraft, Heather A. Sullivan, Tae Kyung Kim, Aaron Szafer, Nick Dee, Peter Groblewski, Ian Wickersham, Ali Cetin, Julie A. Harris, Boaz P. Levi, Susan M. Sunkin, Linda Madisen, Tanya L. Daigle, Loren Looger, Amy Bernard, John Phillips, Ed Lein, Michael Hawrylycz, Karel Svoboda, Allan R. Jones, Christof Koch, and Hongkui Zeng. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, Nov 2018.

[3] Darren A. Cusanovich, Andrew J. Hill, Delasa Aghamirzaie, Riza M. Daza, Hannah A. Pliner, Joel B. Berletch, Galina N. Filippova, Xingfan Huang, Lena Christiansen, William S. DeWitt, Choli Lee, Samuel G. Regalado, David F. Read, Frank J. Steemers, Christine M. Disteche, Cole Trapnell, and Jay Shendure. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174(5):1309 – 1324.e18, 2018.

[4] Chongyuan Luo, Christopher L. Keown, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R. Nery, Justin P. Sandoval, Brian Bui, Terrence J. Sejnowski, Timothy T. Harkins, Eran A. Mukamel, M. Margarita Behrens, and Joseph R. Ecker. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, 357(6351):600–604, 2017.

[5] Ferrari Silvia and Cribari-Neto Francisco. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.