Nonparametric Bayesian Two-Level Clustering for Subject-Level Single-Cell Expression Data

Qiuyu Wu, Xiangyu Luo

Institute of Statistics and Big Data, Renmin University of China

Supplementary Material

S1 Proof of proposition 1

S1.1 Results from Rodriguez et al. (2008)

In the hierarchical and nested nonparametric prior (3.1), given G_0, $G^{(j)}$'s come from the NDP, $\mathrm{DP}(\nu, \mathrm{DP}(\gamma, G_0))$. According to Rodriguez et al. (2008), we have the following results for any Borel set $A \in \mathcal{B}$.

1 $\mathbb{E}\left(G^{(j)}(A)|G_0\right) = G_0(A)$.

2 $\mathbb{E}\left(\left(G^{(j)}(A)\right)^2|G_0\right) = \frac{G_0(A)(\gamma G_0(A)+1)}{\gamma+1}$.

3 $\mathbb{E}\left(G^{(j)}(A)G^{(j')}(A)|G_0\right) = (G_0(A))^2 + \frac{1}{\nu+1}\frac{G_0(A)(1-G_0(A))}{\gamma+1}, j \neq j'$.

4 Assume the first and second moments of G_0 are a_1 and a_2, respectively.

When $D = 1$, let $\mu_i^{(j)}$ and $\mu_{i'}^{(j')}$ denote random variables from $G^{(j)}$ and $G^{(j')}$, respectively. Conditional on G_0, the expectation between $\mu_i^{(j)}$ and $\mu_{i'}^{(j')}$ for $i \neq i'$ is $\mathbb{E}\left(\mu_i^{(j)}\mu_{i'}^{(j)}|G_0\right) = \frac{1}{\gamma+1}a_2 + \frac{\gamma}{\gamma+1}a_1^2$.

5 $\mathbb{P}\left(G^{(j)}(A) = G^{(j')}(A)|G_0\right) = \frac{1}{\nu+1} > 0$.

S1.2 Proposition 1(1)

Proof.

$$\begin{aligned}\mathbb{E}\left(G^{(j)}(A)|H\right) &= \mathbb{E}\left(\mathbb{E}\left(G^{(j)}(A)|G_0\right)|H\right)\\ &= \mathbb{E}\left(G_0(A)|H\right)\\ &= H(A).\end{aligned}$$

□

S1.3 Proposition 1(2)

Proof. Since $G_0 \sim \mathrm{DP}(\alpha, H)$, $G_0(A) \sim \mathrm{Beta}\left(\alpha H(A), \alpha(1 - H(A))\right)$.

$$\begin{aligned}\mathbb{E}\left(\left(G^{(j)}(A)\right)^2|H\right) &= \mathbb{E}\left(\mathbb{E}\left(\left(G^{(j)}(A)\right)^2|G_0\right)|H\right)\\ &= \mathbb{E}\left(\frac{G_0(A)\left(\gamma G_0(A)+1\right)}{\gamma+1}|H\right)\\ &= \frac{\alpha\gamma\left(H(A)\right)^2 + (\alpha+\gamma+1)H(A)}{(\alpha+1)(\gamma+1)}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{V}\left(G^{(j)}(A)|H\right) &= \mathbb{E}\left(\left(G^{(j)}(A)\right)^2|H\right) - \left(\mathbb{E}\left(G^{(j)}(A)|H\right)\right)^2 \\ &= \frac{(\alpha+\gamma+1)\,H(A)\,(1-H(A))}{(\alpha+1)\,(\gamma+1)}.\end{aligned}$$

□

S1.4 Proposition 1(3)

Proof. Since $G_0 \sim \mathrm{DP}(\alpha, H)$, $G_0(A) \sim \mathrm{Beta}\left(\alpha H(A), \alpha(1-H(A))\right)$. For $j \neq j'$,

$$\begin{aligned}\mathbb{E}\left(G^{(j)}(A)G^{(j')}(A)|H\right) &= \mathbb{E}\left(\mathbb{E}\left(G^{(j)}(A)G^{(j')}(A)|G_0\right)|H\right) \\ &= \mathbb{E}\left(\left(G_0(A)\right)^2 + \frac{G_0(A)\,(1-G_0(A))}{(\gamma+1)(\nu+1)}|H\right) \\ &= \frac{\alpha\left(H(A)\right)^2 + H(A)}{\alpha+1} + \frac{\alpha H(A)\,(1-H(A))}{(\alpha+1)\,(\gamma+1)\,(\nu+1)}.\end{aligned}$$

Then,

$$\begin{aligned}\mathrm{Cov}\left(G^{(j)}(A), G^{(j')}(A)|H\right) &= \mathbb{E}\left(G^{(j)}(A)G^{(j')}(A)|H\right) - \mathbb{E}\left(G^{(j)}(A)|H\right)\mathbb{E}\left(G^{(j')}(A)|H\right) \\ &= \frac{\alpha\left(H(A)\right)^2 + H(A)}{\alpha+1} + \frac{\alpha H(A)\,(1-H(A))}{(\alpha+1)\,(\gamma+1)\,(\nu+1)} - \left(H(A)\right)^2 \\ &= \frac{(\alpha+(\gamma+1)\,(\nu+1))\,H(A)\,(1-H(A))}{(\alpha+1)\,(\gamma+1)\,(\nu+1)}.\end{aligned}$$

Therefore,

$$\begin{aligned}\operatorname{Cor}\left(G^{(j)}(A), G^{(j')}(A)|H\right) &= \frac{\operatorname{Cov}\left(G^{(j)}(A), G^{(j')}(A)|H\right)}{\left(\mathbb{V}\left(G^{(j)}(A)|H\right)\mathbb{V}\left(G^{(j')}(A)|H\right)\right)^{1/2}} \\ &= \frac{1}{1+\nu}\frac{\nu\gamma+\alpha+\gamma+\nu+1}{\alpha+\gamma+1}.\end{aligned}$$

□

S1.5 Proposition 1(4)

Proof. Since $G_0 \sim \mathrm{DP}(\alpha, H)$, we can write $G_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\mu_k}$, where $\mu_k \overset{i.i.d.}{\sim} H, \{\pi_k\}_{k=1}^{\infty} \sim \mathrm{GEM}(\alpha)$. Assume the first and second moments of G_0 are a_1 and a_2, respectively. Let $\mathbb{E}(\mu_k|H) = m_1$ and $\mathbb{E}(\mu_k^2|H) = m_2 < \infty$. Then $a_1 = \sum_{k=1}^{\infty} \pi_k \mu_k$ and $a_2 = \sum_{k=1}^{\infty} \pi_k \mu_k^2$ imply that $\mathbb{E}(a_1|H) = m_1$ and $\mathbb{E}(a_2|H) = m_2$. According to the construction of stick-breaking process, $\mathbb{E}(\pi_k^2) = \left(\frac{\alpha}{2+\alpha}\right)^{k-1} \frac{2}{(1+\alpha)(2+\alpha)}$.

First, we have

$$\begin{aligned}
\mathbb{E}\left(a_1^2|H\right) &= \mathbb{E}\left(\sum_{k=1}^{\infty}\pi_k\mu_k\left(\pi_k\mu_k+\sum_{\ell\neq k}\pi_\ell\mu_\ell\right)|H\right)\\
&= \sum_{k=1}^{\infty}\left[\mathbb{E}(\pi_k^2)\mathbb{E}(\mu_k^2|H)+\mathbb{E}(\pi_k\sum_{\ell\neq k}\pi_\ell)\mathbb{E}(\mu_k\mu_\ell|H)\right]\\
&= \sum_{k=1}^{\infty}\left[\mathbb{E}(\pi_k^2)m_2+\mathbb{E}(\pi_k\sum_{\ell\neq k}\pi_\ell)m_1^2\right]\\
&= \sum_{k=1}^{\infty}\left[\mathbb{E}(\pi_k^2)(m_2-m_1^2)\right]+m_1^2\\
&= \frac{1}{\alpha+1}(m_2-m_1^2)+m_1^2\\
&= \frac{1}{\alpha+1}m_2+\frac{\alpha}{\alpha+1}m_1^2.
\end{aligned}$$

In addition,

$$\begin{aligned}
\mathbb{E}\left(\mu_i^{(j)}|H\right) &= \mathbb{E}\left(\mathbb{E}\left(\mu_i^{(j)}|G_0\right)|H\right)\\
&= \mathbb{E}\left(a_1|H\right)=m_1,\\
\mathbb{V}\left(\mu_i^{(j)}|H\right) &= \mathbb{E}\left(\mathbb{E}\left(\left(\mu_i^{(j)}\right)^2|G_0\right)|H\right)-\mathbb{E}\left(\mu_i^{(j)}|H\right)^2\\
&= \mathbb{E}\left(a_2|H\right)-\mathbb{E}\left(\mu_i^{(j)}|H\right)^2=m_2-m_1^2.
\end{aligned}$$

For $i \neq i', j = j'$,

$$\begin{aligned}
\mathbb{E}\left(\mu_i^{(j)}\mu_{i'}^{(j')}|H\right) &= \mathbb{E}\left(\mathbb{E}\left(\mu_i^{(j)}\mu_{i'}^{(j)}|G_0\right)|H\right) \\
&= \mathbb{E}\left(\frac{1}{\gamma+1}a_2 + \frac{\gamma}{\gamma+1}a_1^2|H\right) \\
&= \frac{1}{\gamma+1}m_2 + \frac{\gamma}{\gamma+1}\left(\frac{1}{\alpha+1}m_2 + \frac{\alpha}{\alpha+1}m_1^2\right) \\
&= \frac{\alpha+\gamma+1}{(\alpha+1)(\gamma+1)}m_2 + \frac{\alpha\gamma}{(\alpha+1)(\gamma+1)}m_1^2.
\end{aligned}$$

Subsequently,

$$\begin{aligned}
\mathrm{Cov}\left(\mu_i^{(j)}, \mu_{i'}^{(j')}|H\right) &= \mathbb{E}\left(\mu_i^{(j)}\mu_{i'}^{(j)}|H\right) - \mathbb{E}\left(\mu_i^{(j)}|H\right)\mathbb{E}\left(\mu_{i'}^{(j)}|H\right) \\
&= \frac{\alpha+\gamma+1}{(\alpha+1)(\gamma+1)}m_2 + \frac{\alpha\gamma}{(\alpha+1)(\gamma+1)}m_1^2 - m_1^2 \\
&= \frac{\alpha+\gamma+1}{(\alpha+1)(\gamma+1)}(m_2 - m_1^2).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Cor}\left(\mu_i^{(j)}, \mu_{i'}^{(j')}|H\right) &= \frac{\mathrm{Cov}\left(\mu_i^{(j)}, \mu_{i'}^{(j)}|H\right)}{\left(\mathbb{V}\left(\mu_i^{(j)}|H\right)\mathbb{V}\left(\mu_{i'}^{(j)}|H\right)\right)^{1/2}} \\
&= \frac{\alpha+\gamma+1}{(\alpha+1)(\gamma+1)}.
\end{aligned}$$

For $j \neq j'$,

$$\begin{aligned}
\mathbb{E}\left(\mu_i^{(j)}\mu_{i'}^{(j')}|G_0\right) &= \mathbb{E}\left(\mu_i^{(j)}\mu_{i'}^{(j)}|G_0\right)\mathbb{P}\left(G^{(j)}(A) = G^{(j')}(A)|G_0\right) \\
&\quad + \mathbb{E}\left(\mu_i^{(j)}\mu_{i'}^{(j')}|G_0\right)\mathbb{P}\left(G^{(j)}(A) \neq G^{(j')}(A)|G_0\right) \\
&= \left(\frac{1}{\gamma+1}a_2 + \frac{\gamma}{\gamma+1}a_1^2\right)\frac{1}{\nu+1} + a_1^2\frac{\nu}{\nu+1} \\
&= \frac{1}{(\gamma+1)(\nu+1)}a_2 + \frac{\nu\gamma+\gamma+\nu}{(\gamma+1)(\nu+1)}a_1^2.
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}\left(\mu_i^{(j)}\mu_{i'}^{(j')}|H\right) &= \mathbb{E}\left(\mathbb{E}\left(\mu_i^{(j)}\mu_{i'}^{(j')}|G_0\right)|H\right) \\
&= \frac{\nu\gamma+\alpha+\gamma+\nu+1}{(\nu+1)(\alpha+1)(\gamma+1)}m_2 + \frac{\alpha(\nu\gamma+\gamma+\nu)}{(\nu+1)(\alpha+1)(\gamma+1)}m_1^2.
\end{aligned}$$

Subsequently,

$$\begin{aligned}
\mathrm{Cov}\left(\mu_i^{(j)}, \mu_{i'}^{(j')}|H\right) &= \mathbb{E}\left(\mu_i^{(j)}\mu_{i'}^{(j')}|H\right) - \mathbb{E}\left(\mu_i^{(j)}|H\right)\mathbb{E}\left(\mu_{i'}^{(j')}|H\right) \\
&= \frac{\nu\gamma+\alpha+\gamma+\nu+1}{(\nu+1)(\alpha+1)(\gamma+1)}(m_2 - m_1^2).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Cor}\left(\mu_i^{(j)}, \mu_{i'}^{(j')}|H\right) &= \frac{\mathrm{Cov}\left(\mu_i^{(j)}, \mu_{i'}^{(j')}|H\right)}{\left(\mathbb{V}\left(\mu_i^{(j)}|H\right)\mathbb{V}\left(\mu_{i'}^{(j')}|H\right)\right)^{1/2}} \\
&= \frac{\nu\gamma+\alpha+\gamma+\nu+1}{(\nu+1)(\alpha+1)(\gamma+1)}.
\end{aligned}$$

□

S1.6 Proposition 1(5)

Proof. Since $G_0 \sim \mathrm{DP}(\alpha, H)$, we can write $G_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\mu}_k}$, where $\boldsymbol{\mu}_k \overset{i.i.d.}{\sim} H, \{\pi_k\}_{k=1}^{\infty} \sim \mathrm{GEM}(\alpha)$. Assume the first and second moments of G_0 are $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, respectively. Let $\mathbb{E}(\boldsymbol{\mu}_k|H) = \boldsymbol{m}_1$, $\mathbb{E}(\boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top|H) = \boldsymbol{m}_2$ and $\mathrm{Cor}(\boldsymbol{\mu}_k|H) = \mathbf{R}_H$. Then $\boldsymbol{a}_1 = \sum_{k=1}^{\infty} \pi_k \boldsymbol{\mu}_k$ and $\boldsymbol{a}_2 = \sum_{k=1}^{\infty} \pi_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top$. Thus, $\mathbb{E}(\boldsymbol{a}_1|H) = \boldsymbol{m}_1$ and $\mathbb{E}(\boldsymbol{a}_2|H) = \boldsymbol{m}_2$. According to the construction of stick-breaking process, $\mathbb{E}(\pi_k^2) = \left(\frac{\alpha}{2+\alpha}\right)^{k-1} \frac{2}{(1+\alpha)(2+\alpha)}$.

First, we have

$$
\begin{aligned}
\mathbb{E}\left(\boldsymbol{a}_1\boldsymbol{a}_1^\top|H\right) &= \mathbb{E}\left(\sum_{k=1}^{\infty} \pi_k \boldsymbol{\mu}_k \left(\pi_k \boldsymbol{\mu}_k^\top + \sum_{\ell \neq k} \pi_\ell \boldsymbol{\mu}_\ell^\top\right)|H\right) \\
&= \sum_{k=1}^{\infty} \left[\mathbb{E}(\pi_k^2)\mathbb{E}(\boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top|H) + \mathbb{E}(\pi_k \sum_{\ell \neq k} \pi_\ell)\mathbb{E}(\boldsymbol{\mu}_k\boldsymbol{\mu}_\ell^\top|H)\right] \\
&= \sum_{k=1}^{\infty} \left[\mathbb{E}(\pi_k^2)\boldsymbol{m}_2 + \mathbb{E}(\pi_k \sum_{\ell \neq k} \pi_\ell)\boldsymbol{m}_1\boldsymbol{m}_1^\top\right] \\
&= \sum_{k=1}^{\infty} \left[\mathbb{E}(\pi_k^2)(\boldsymbol{m}_2 - \boldsymbol{m}_1\boldsymbol{m}_1^\top)\right] + \boldsymbol{m}_1\boldsymbol{m}_1^\top \\
&= \frac{1}{\alpha+1}(\boldsymbol{m}_2 - \boldsymbol{m}_1\boldsymbol{m}_1^\top) + \boldsymbol{m}_1\boldsymbol{m}_1^\top \\
&= \frac{1}{\alpha+1}\boldsymbol{m}_2 + \frac{\alpha}{\alpha+1}\boldsymbol{m}_1\boldsymbol{m}_1^\top.
\end{aligned}
$$

In addition,

$$
\begin{aligned}
\mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}|H\right) &= \mathbb{E}\left(\mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}|G_0\right)|H\right) \\
&= \mathbb{E}\left(\boldsymbol{a}_1|H\right) = \boldsymbol{m}_1,
\end{aligned}
$$

$$\begin{aligned}\mathbb{V}\left(\boldsymbol{\mu}_i^{(j)}|H\right) &= \mathbb{E}\left(\mathbb{E}\left(\left(\boldsymbol{\mu}_i^{(j)}\right)^2|G_0\right)|H\right) - \mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}|H\right)^2 \\ &= \mathbb{E}\left(\boldsymbol{a}_2|H\right) - \mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}|H\right)^2 = \boldsymbol{m}_2 - \boldsymbol{m}_1\boldsymbol{m}_1^\top.\end{aligned}$$

For $i \neq i', j = j'$,

$$\begin{aligned}\mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}\boldsymbol{\mu}_{i'}^{(j')}|H\right) &= \mathbb{E}\left(\mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}\boldsymbol{\mu}_{i'}^{(j)\top}|G_0\right)|H\right) \\ &= \mathbb{E}\left(\frac{1}{\gamma+1}\boldsymbol{a}_2 + \frac{\gamma}{\gamma+1}\boldsymbol{a}_1\boldsymbol{a}_1^\top|H\right) \\ &= \frac{1}{\gamma+1}\boldsymbol{m}_2 + \frac{\gamma}{\gamma+1}\left(\frac{1}{\alpha+1}\boldsymbol{m}_2 + \frac{\alpha}{\alpha+1}\boldsymbol{m}_1\boldsymbol{m}_1^\top\right) \\ &= \frac{\alpha+\gamma+1}{(\alpha+1)(\gamma+1)}\boldsymbol{m}_2 + \frac{\alpha\gamma}{(\alpha+1)(\gamma+1)}\boldsymbol{m}_1\boldsymbol{m}_1^\top.\end{aligned}$$

Subsequently,

$$\begin{aligned}\mathrm{Cov}\left(\boldsymbol{\mu}_i^{(j)}, \boldsymbol{\mu}_{i'}^{(j')}|H\right) &= \mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}\boldsymbol{\mu}_{i'}^{(j)\top}|H\right) - \mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}|H\right)\mathbb{E}\left(\boldsymbol{\mu}_{i'}^{(j)}|H\right)^\top \\ &= \frac{\alpha+\gamma+1}{(\alpha+1)(\gamma+1)}\boldsymbol{m}_2 + \frac{\alpha\gamma}{(\alpha+1)(\gamma+1)}\boldsymbol{m}_1\boldsymbol{m}_1^\top - \boldsymbol{m}_1\boldsymbol{m}_1^\top \\ &= \frac{\alpha+\gamma+1}{(\alpha+1)(\gamma+1)}(\boldsymbol{m}_2 - \boldsymbol{m}_1\boldsymbol{m}_1^\top).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{Cor}\left(\boldsymbol{\mu}_i^{(j)}, \boldsymbol{\mu}_{i'}^{(j')}|H\right) &= \mathrm{diag}\left(\mathbb{V}\left(\boldsymbol{\mu}_i^{(j)}|H\right)\right)^{-1/2}\mathrm{Cov}\left(\boldsymbol{\mu}_i^{(j)}, \boldsymbol{\mu}_{i'}^{(j)}|H\right)\mathrm{diag}\left(\mathbb{V}\left(\boldsymbol{\mu}_{i'}^{(j)}|H\right)\right)^{-1/2} \\ &= \frac{\alpha+\gamma+1}{(\alpha+1)(\gamma+1)}\mathbf{R}_H.\end{aligned}$$

For $j \neq j'$,

$$\begin{aligned}\mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}\boldsymbol{\mu}_{i'}^{(j')}|G_0\right) &= \mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}\boldsymbol{\mu}_{i'}^{(j)}|G_0\right)\mathbb{P}\left(G^{(j)}(A) = G^{(j')}(A)|G_0\right) \\ &\quad + \mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}\boldsymbol{\mu}_{i'}^{(j')}|G_0\right)\mathbb{P}\left(G^{(j)}(A) \neq G^{(j')}(A)|G_0\right) \\ &= \left(\frac{1}{\gamma+1}\boldsymbol{a}_2 + \frac{\gamma}{\gamma+1}\boldsymbol{a}_1\boldsymbol{a}_1^\top\right)\frac{1}{\nu+1} + \boldsymbol{a}_1\boldsymbol{a}_1^\top\frac{\nu}{\nu+1} \\ &= \frac{1}{(\gamma+1)(\nu+1)}\boldsymbol{a}_2 + \frac{\nu\gamma+\gamma+\nu}{(\gamma+1)(\nu+1)}\boldsymbol{a}_1\boldsymbol{a}_1^\top.\end{aligned}$$

Then

$$\begin{aligned}\mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}\boldsymbol{\mu}_{i'}^{(j')}|H\right) &= \mathbb{E}\left(\mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}\boldsymbol{\mu}_{i'}^{(j')}|G_0\right)|H\right) \\ &= \frac{\nu\gamma+\alpha+\gamma+\nu+1}{(\nu+1)(\alpha+1)(\gamma+1)}\boldsymbol{m}_2 + \frac{\alpha(\nu\gamma+\gamma+\nu)}{(\nu+1)(\alpha+1)(\gamma+1)}\boldsymbol{m}_1\boldsymbol{m}_1^\top.\end{aligned}$$

Subsequently,

$$\begin{aligned}\mathrm{Cov}\left(\boldsymbol{\mu}_i^{(j)}, \boldsymbol{\mu}_{i'}^{(j')}|H\right) &= \mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}\boldsymbol{\mu}_{i'}^{(j')}|H\right) - \mathbb{E}\left(\boldsymbol{\mu}_i^{(j)}|H\right)\mathbb{E}\left(\boldsymbol{\mu}_{i'}^{(j')}|H\right) \\ &= \frac{\nu\gamma+\alpha+\gamma+\nu+1}{(\nu+1)(\alpha+1)(\gamma+1)}(\boldsymbol{m}_2 - \boldsymbol{m}_1\boldsymbol{m}_1^\top).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{Cor}\left(\boldsymbol{\mu}_i^{(j)}, \boldsymbol{\mu}_{i'}^{(j')}|H\right) &= \mathrm{diag}\left(\mathbb{V}\left(\boldsymbol{\mu}_i^{(j)}|H\right)\right)^{-1/2}\mathrm{Cov}\left(\boldsymbol{\mu}_i^{(j)}, \boldsymbol{\mu}_{i'}^{(j')}|H\right)\mathrm{diag}\left(\mathbb{V}\left(\boldsymbol{\mu}_{i'}^{(j')}|H\right)\right)^{-1/2} \\ &= \frac{\nu\gamma+\alpha+\gamma+\nu+1}{(\nu+1)(\alpha+1)(\gamma+1)}\mathbf{R}_H.\end{aligned}$$

□

S2 The PLN distribution accounts for the over-dispersion

Let $Y \sim PLN(\eta, \sigma^2)$, which is equivalent to $\theta \sim N(\eta, \sigma^2)$ and $Y \sim Poi(e^\theta)$. The mean and variance of Y can be derived based on the law of the total expectation as follows.

$$\begin{aligned}
\mathbb{E}(Y) &= \mathbb{E}(\mathbb{E}(Y|\theta)) \\
&= e^{\eta+\frac{\sigma^2}{2}}, \\
\mathbb{V}(Y) &= \mathbb{E}(\mathbb{V}(Y|\theta)) + \mathbb{V}(\mathbb{E}(Y|\theta)) \\
&= e^{\eta+\frac{\sigma^2}{2}}(1 + e^{\eta+\frac{3\sigma^2}{2}} - e^{\eta+\frac{\sigma^2}{2}}).
\end{aligned}$$

The variance of PLN is larger than the mean, so PLN is able to model over-dispersed data.

S3 Library size

The cell library size is the total number of reads mapped to one single cell. For example, there are two genes and two cells. We assume the two genes in fact have the same expression in the two cells. However, two cells usually go through different polymerase chain reaction (PCR) amplification, resulting in different read counts on genes, e.g., gene 1 has 4 reads in cell 1 and 12 reads in cell 2, and gene 2 has 6 reads in cell 1 and 18 reads in cell 2. Then library sizes of two cells are 10 and 30, respectively. Without consideration

of cell library sizes, we will have the wrong conclusion that the two genes are differentially expressed between the two cells. If library sizes are accounted for, we divide the expressions in cell 2 by $30/10 = 3$ and thus obtain the correct results that the two genes have the same expression levels in the two cells.

S4 The proof of Theorem 1

We denote the set of all parameters in SCSC except $\boldsymbol{\mu}$ by $\boldsymbol{\Theta}_{-\boldsymbol{\mu}}$.

$$\begin{aligned}
\int \left|p^{KL}(\mathbf{x}) - p^{\infty\infty}(\mathbf{x})\right| d\mathbf{x} &\leq \iiint p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Theta}_{-\boldsymbol{\mu}}) \left|p^{KL}(d\boldsymbol{\mu}) - p^{\infty\infty}(d\boldsymbol{\mu})\right| p(d\boldsymbol{\Theta}_{-\boldsymbol{\mu}}) d\mathbf{x} \\
&= \int \left|p^{KL}(d\boldsymbol{\mu}) - p^{\infty\infty}(d\boldsymbol{\mu})\right| \\
&\leq \iint \left|p^{KL}(d\boldsymbol{\mu}|G_0) - p^{\infty\infty}(d\boldsymbol{\mu}|G_0)\right| p(dG_0) \\
&\leq \int \epsilon^{KL}(\nu, \gamma) p(dG_0) \\
&= \epsilon^{KL}(\nu, \gamma).
\end{aligned}$$

$\epsilon^{KL}(\nu, \gamma) = 4\left\{1 - \left[1 - \left(\frac{\nu}{1+\nu}\right)^{L-1}\right]^m \times \left[1 - \left(\frac{\gamma}{\gamma+1}\right)^{K-1}\right]^{\sum_{j=1}^m n_j}\right\}$. The last inequality in this theorem follows the proof of the Theorem 2 of Rodriguez et al. (2008).

S5 Proof of equivalence between Models (4.1) and (4.2)

We prove that Model (4.1) is equivalent to Model (4.2).

Proof. In Model (4.1), we first focus on the first three lines,

$$\begin{cases} G_0 \sim \mathrm{DP}(\alpha, H) \\ G^{(j)}|G_0 \sim \mathrm{DP}_L(\nu, \mathrm{DP}_K(\gamma, G_0)) \\ \boldsymbol{\mu}_i^{(j)}|G^{(j)} \sim G^{(j)} \end{cases} \iff (*) \begin{cases} G_0 \sim \mathrm{DP}(\alpha, H) \\ G_\ell^*|G_0 \sim \mathrm{DP}_K(\gamma, G_0) \text{ for } 1 \le \ell \le L \\ \boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_L) \sim \mathrm{GEM}_L(\nu) \\ G^{(j)}|G_\ell^*, \boldsymbol{\phi} \sim \sum_{\ell=1}^L \phi_\ell \delta_{G_\ell^*} \\ \boldsymbol{\mu}_i^{(j)}|G^{(j)} \sim G^{(j)}. \end{cases}$$

Note that $G_0 \sim \mathrm{DP}(\alpha, H)$ and $G_\ell^{*'}|G_0 \sim \mathrm{DP}(\gamma, G_0)$ are equivalent to $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots) \sim \mathrm{GEM}(\alpha)$, $\boldsymbol{\mu}_k \sim H$, $\boldsymbol{\pi}'_\ell = (\pi'_{1\ell}, \pi'_{2\ell}, \ldots)|\boldsymbol{\rho} \sim \mathrm{DP}(\gamma, \boldsymbol{\rho})$ and $G_\ell^{*'} = \sum_{k=1}^\infty \pi'_{k\ell}\delta_{\boldsymbol{\mu}_k}$ according to the results from the HDP paper (Teh et al., 2006). Subsequently, when there is a truncation K on the distribution $G_\ell^{*'} = \sum_{k=1}^\infty \pi'_{\ell k}\delta_{\boldsymbol{\mu}_k}$, we have $G_\ell^* = \sum_{k=1}^K \pi_{k\ell}\delta_{\boldsymbol{\mu}_k}$, where $\pi_{k\ell} = \pi'_{k\ell}$ for $1 \le k \le K-1$ and $\pi_{K\ell} = \sum_{i=K}^\infty \pi'_{i\ell}$. Therefore, the first two lines of the expression $(*)$ are

$$
\begin{cases} G_0 \sim \mathrm{DP}(\alpha, H) \\ G_\ell^* | G_0 \sim \mathrm{DP}_K(\gamma, G_0) \end{cases}
\iff
\begin{cases}
\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots) \sim \mathrm{GEM}(\alpha) \\
\boldsymbol{\mu}_k \sim H \\
\boldsymbol{\pi}_\ell = (\pi_{1,\ell}, \ldots, \pi_{K-1,\ell}, \pi_{K,\ell}) | \boldsymbol{\rho} \\
\qquad\qquad \sim \mathrm{Dir}(\gamma\rho_1, \ldots, \gamma\rho_{K-1}, \gamma \sum_{i=K}^{\infty} \rho_i) \\
G_\ell^* | \boldsymbol{\pi}_\ell, \boldsymbol{\mu}_k = \sum_{k=1}^{K} \pi_{k\ell} \delta_{\boldsymbol{\mu}_k}
\end{cases}
$$

$$
\iff
\begin{cases}
\boldsymbol{\xi} := (\xi_1 = \rho_1, \ldots, \xi_{K-1} = \rho_{K-1}, \xi_K = \sum_{i=K}^{\infty} \rho_i) \\
\qquad\qquad \sim \mathrm{GEM}_K(\alpha) \\
\boldsymbol{\mu}_k \sim H \\
\boldsymbol{\pi}_\ell | \boldsymbol{\xi} \sim \mathrm{Dir}(\gamma\xi_1, \ldots, \gamma\xi_K) \\
G_\ell^* | \boldsymbol{\pi}_\ell, \boldsymbol{\mu}_k \sim \sum_{k=1}^{K} \pi_{k\ell} \delta_{\boldsymbol{\mu}_k}.
\end{cases}
$$

The second equivalence holds because for any $\boldsymbol{\xi} \sim \mathrm{GEM}_K(\alpha)$ we can find a $\boldsymbol{\rho}$ following $\mathrm{GEM}(\alpha)$ by letting $\rho_k = \xi_k$ for $1 \le k \le K-1$, $\rho_K = (1 - \sum_{i=1}^{K-1} \rho_i) \cdot \rho'_K$, and $\rho_k = (1 - \sum_{i=1}^{K-1} \rho_i) \cdot \prod_{i=K}^{k-1} (1 - \rho'_i) \cdot \rho'_k$ for $k \ge K+1$, where $\rho'_i \sim \mathrm{Beta}(1, \alpha)$ $(i \ge K)$; and on the opposite direction, for any $\boldsymbol{\rho} \sim \mathrm{GEM}(\alpha)$, $\boldsymbol{\xi}$ follows $\mathrm{GEM}_K(\alpha)$ through the construction above. Next,

we plug the result above into the expression $(*)$, leading to

$$\begin{cases} G_0 \sim \mathrm{DP}(\alpha, H) \\ G^{(j)}|G_0 \sim \mathrm{DP}_L(\nu, \mathrm{DP}_K(\gamma, G_0)) \\ \boldsymbol{\mu}_i^{(j)}|G^{(j)} \sim G^{(j)} \end{cases} \iff \begin{cases} \boldsymbol{\xi} \sim \mathrm{GEM}_K(\alpha) \\ \boldsymbol{\mu}_k \sim H \\ \boldsymbol{\pi}_\ell|\boldsymbol{\xi} \sim \mathrm{Dir}(\gamma\xi_1, \ldots, \gamma\xi_K) \\ G_\ell^*|\boldsymbol{\pi}_\ell, \boldsymbol{\mu}_k = \sum_{k=1}^K \pi_{k\ell}\delta_{\boldsymbol{\mu}_k} \\ \boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_L) \sim \mathrm{GEM}_L(\nu) \\ G^{(j)}|G_\ell^*, \boldsymbol{\phi} \sim \sum_{\ell=1}^L \phi_\ell \delta_{G_\ell^*} \\ \boldsymbol{\mu}_i^{(j)}|G^{(j)} \sim G^{(j)}. \end{cases}$$

Considering $S^{(j)}$, $C_i^{(j)}$, and the distribution for scRNA-seq data, it follows that

$$
\begin{cases}
G_0 \sim \mathrm{DP}(\alpha, H) \\
G^{(j)}|G_0 \sim \mathrm{DP}_L(\nu, \mathrm{DP}_K(\gamma, G_0)) \\
\boldsymbol{\mu}_i^{(j)}|G^{(j)} \sim G^{(j)} \\
X_{gi}^{(j)}|\boldsymbol{\mu}_i^{(j)} \sim \\
\qquad \mathrm{ZIPLN}(\lambda_{g0}, \lambda_{g1}, s_i^{(j)}, \mu_{gi}^{(j)} + \beta_g^{(j)}, \sigma_g^2)
\end{cases}
\iff
\begin{cases}
\boldsymbol{\xi} \sim \mathrm{GEM}_K(\alpha) \\
\boldsymbol{\mu}_k \sim H \\
\boldsymbol{\pi}_\ell \sim \mathrm{Dir}(\gamma\xi_1, \ldots, \gamma\xi_K) \\
\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_L) \sim \mathrm{GEM}_L(\nu) \\
S^{(j)} \sim \mathrm{MN}(1; \phi_1, \phi_2, \ldots, \phi_L) \\
C_i^{(j)}|S^{(j)} = \ell \sim \mathrm{MN}(1; \pi_{1\ell}, \ldots, \pi_{K\ell}) \\
X_{gi}^{(j)}|S^{(j)} = \ell, C_i^{(j)} = k \sim \\
\qquad \mathrm{ZIPLN}(\lambda_{g0}, \lambda_{g1}, s_i^{(j)}, \mu_{gk} + \beta_{g\ell}, \sigma_g^2).
\end{cases}
$$

□

S6 SCSC-vs

Tadesse et al. (2005) proposed a Bayesian variable selection method to cluster high-dimensional samples and identify discriminating variables simultaneously, so we incorporated this idea into the proposed model SCSC, resulting in a variable selection version, which we termed SCSC-vs. SCSC-vs can identify genes that discriminate cell types to improve the clustering accuracy. Specifically, we introduced binary latent variables $\boldsymbol{Z} = \{z_1, \cdots, z_D\}$,

where $z_g = 1$ indicates that gene g is a marker across cell types ($1 \leq g \leq D$). We denoted the common cell-type effects by $\mu_{g,com}$ when $z_g = 0$, which means the gene g cannot distinguish cell types. Following Tadesse et al. (2005), the likelihood function of $\boldsymbol{Z}, \boldsymbol{\mu}$ given all other variables is

$$p(\boldsymbol{Z}, \boldsymbol{\mu}|-) = \prod_{i,j} \left(\prod_{g:z_g=0} \mathrm{N}\left(\theta_{gi}^{(j)}; \mu_{g,com} + \beta_{gS^{(j)}}, \sigma_g^2\right) \prod_{g:z_g=1} \mathrm{N}\left(\theta_{gi}^{(j)}; \mu_{gC_i^{(j)}} + \beta_{gS^{(j)}}, \sigma_g^2\right) \right),$$

where $\mathrm{N}(x; \mu, \sigma^2)$ is the normal density evaluated at x with mean μ and variance σ^2. We assigned a Bernoulli distribution $\mathrm{Ber}(q)$ to z_g and a normal distribution $\mathrm{N}(\eta_\mu, \tau_\mu^2)$ to $\mu_{g,com}$ and μ_{gk}.

By combining the conditional distribution above and the Model (4.2) in the manuscript, we implemented the MCMC algorithm to perform posterior inference. The following steps are different from those in the MCMC sampling scheme of SCSC in Section S7.

1. The augmented parameter $\theta_{gi}^{(j)}$ in the PLN distribution is generated from

$$p(\theta_{gi}^{(j)}|-) \propto \exp\left\{ -s_i^{(j)} e^{\theta_{gi}^{(j)}} + Y_{gi}^{(j)} \theta_{gi}^{(j)} - \frac{(\theta_{gi}^{(j)} - \mu_{gk} - \beta_{g\ell})^2}{2\sigma_g^2} \right\},$$

when $S^{(j)} = \ell$, $C_i^{(j)} = k$ and $z_g = 1$;

$$p(\theta_{gi}^{(j)}|-) \propto \exp\left\{ -s_i^{(j)} e^{\theta_{gi}^{(j)}} + Y_{gi}^{(j)} \theta_{gi}^{(j)} - \frac{(\theta_{gi}^{(j)} - \mu_{g,com} - \beta_{g\ell})^2}{2\sigma_g^2} \right\},$$

when $S^{(j)} = \ell$, $C_i^{(j)} = k$ and $z_g = 0$.

4. For each gene g, we first sample the marker indicator z_g from the Bernoulli distribution

$$\mathrm{Ber}(\frac{1}{1+d_g}),$$

where

$$\begin{aligned} d_g =& \frac{1-q}{q}(\tau_\mu^2)^{(K-1)/2} \frac{\prod_{k=1}^{K}\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mathbb{I}(C_i^{(j)}=k)/\sigma_g^2+\frac{1}{\tau_\mu^2}\right)^{1/2}}{\left(\sum_{j=1}^{m} n_j/\sigma_g^2+\frac{1}{\tau_\mu^2}\right)^{1/2}} \\ &\times \exp\left\{(K-1)\frac{\eta_\mu^2}{2\tau_\mu^2} + \frac{\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j}(\theta_{gi}^{(j)}-\beta_{gS^{(j)}})/\sigma_g^2+\frac{\eta_\mu}{\tau_\mu^2}\right)^2}{2\left(\sum_{j=1}^{m} n_j/\sigma_g^2+\frac{1}{\tau_\mu^2}\right)}\right\} \\ &\times \exp\left\{-\sum_{k=1}^{K}\frac{\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j}(\theta_{gi}^{(j)}-\beta_{gS^{(j)}})\mathbb{I}(C_i^{(j)}=k)/\sigma_g^2+\frac{\eta_\mu}{\tau_\mu^2}\right)^2}{2\left(\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mathbb{I}(C_i^{(j)}=k)/\sigma_g^2+\frac{1}{\tau_\mu^2}\right)}\right\}. \end{aligned}$$

If $z_g = 1$, then we sample the cell-type k effect on gene g, μ_{gk}, from the normal distribution

$$\mathrm{N}\left(\frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(\theta_{gi}^{(j)}-\beta_{gS^{(j)}})\mathbb{I}(C_i^{(j)}=k)/\sigma_g^2+\eta_\mu/\tau_\mu^2}{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mathbb{I}(C_i^{(j)}=k)/\sigma_g^2+1/\tau_\mu^2}, \frac{1}{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mathbb{I}(C_i^{(j)}=k)/\sigma_g^2+1/\tau_\mu^2}\right).$$

If $z_g = 0$, we sample the common cell effect on gene g, $\mu_{g,com}$, from

the normal distribution

$$\mathrm{N}\left(\frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(\theta_{gi}^{(j)}-\beta_{gS^{(j)}})/\sigma_g^2+\eta_\mu/\tau_\mu^2}{\sum_{j=1}^{m}n_j/\sigma_g^2+1/\tau_\mu^2},\frac{1}{\sum_{j=1}^{m}n_j/\sigma_g^2+1/\tau_\mu^2}\right).$$

5. Update the hyper-parameters in the cell-type effect prior,

$$\eta_\mu|-\sim \mathrm{N}\Bigg(\frac{\sum_{g=1}^{D}\Big(\sum_{k=1}^{K}\mu_{gk}\mathbb{I}(z_g=1)+K\mu_{g,com}\mathbb{I}(z_g=0)\Big)/\tau_\mu^2+u_\mu/\omega_\mu^2}{DK/\tau_\mu^2+1/\omega_\mu^2},\frac{1}{DK/\tau_\mu^2+1/\omega_\mu^2}\Bigg),$$

$$\tau_\mu^2|-\sim \mathrm{Inv}\Gamma\Bigg(b_{\mu1}+DK/2,b_{\mu2}+\sum_{g=1}^{D}\Big(\sum_{k=1}^{K}(\mu_{gk}-\eta_\mu)^2\mathbb{I}(z_g=1)+K(\mu_{g,com}-\eta_\mu)^2\mathbb{I}(z_g=0)\Big)/2\Bigg).$$

6. The subgroup ℓ effect on gene g for $\ell \geq 2$, $\beta_{g\ell}$, is sampled from the normal distribution

$$\mathrm{N}\left(\frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(\theta_{gi}^{(j)}-\mu_{gC_i^{(j)}})\mathbb{I}(S^{(j)}=\ell)/\sigma_g^2+\eta_\beta/\tau_\beta^2}{\sum_{j=1}^{m}\mathbb{I}(S^{(j)}=\ell)n_j/\sigma_g^2+1/\tau_\beta^2},\frac{1}{\sum_{j=1}^{m}\mathbb{I}(S^{(j)}=\ell)n_j/\sigma_g^2+1/\tau_\beta^2}\right),$$

when $z_g=1$;

$$\mathrm{N}\left(\frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(\theta_{gi}^{(j)}-\mu_{g,com})\mathbb{I}(S^{(j)}=\ell)/\sigma_g^2+\eta_\beta/\tau_\beta^2}{\sum_{j=1}^{m}\mathbb{I}(S^{(j)}=\ell)n_j/\sigma_g^2+1/\tau_\beta^2},\frac{1}{\sum_{j=1}^{m}\mathbb{I}(S^{(j)}=\ell)n_j/\sigma_g^2+1/\tau_\beta^2}\right),$$

when $z_g=0$.

We notice that subgroup 1 effect, β_{g1}, is restricted to zero across $1 \leq$

$g \leq D$ for identifying the subgroup and cell-type effects.

8. Update the variance σ_g^2 for gene g by sampling from the inverse-gamma distribution

$$\text{Inv}\Gamma(b_{\sigma 1} + \sum_{j=1}^{m} n_j/2, b_{\sigma 2} + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(\theta_{gi}^{(j)} - \mu_{gC_i^{(j)}} - \beta_{gS^{(j)}})^2/2),$$

when $z_g = 1$;

$$\text{Inv}\Gamma(b_{\sigma 1} + \sum_{j=1}^{m} n_j/2, b_{\sigma 2} + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(\theta_{gi}^{(j)} - \mu_{g,com} - \beta_{gS^{(j)}})^2/2),$$

when $z_g = 0$.

9. For each subject j, update the subtype indicator $S^{(j)}$ and the cell-type indicators $C_i^{(j)}$ for cell $i = 1, \ldots, n_j$ from multinomial distributions

$$P(S^{(j)} = \ell|-) \propto \phi_\ell \prod_{i=1}^{n_j}\Bigg[\sum_{k=1}^{K}\pi_{k\ell}\prod_{g:z_g=1}\text{N}(\theta_{gi}^{(j)}; \mu_{gk} + \beta_{g\ell}, \sigma_g^2)\times \prod_{g:z_g=0}\text{N}(\theta_{gi}^{(j)}; \mu_{g,com} + \beta_{g\ell}, \sigma_g^2)\Bigg],$$

$$P(C_i^{(j)} = k|S^{(j)} = \ell, -) \propto \pi_{k\ell}\prod_{g:z_g=1}\text{N}(\theta_{gi}^{(j)}; \mu_{gk} + \beta_{g\ell}, \sigma_g^2)\prod_{g:z_g=0}\text{N}(\theta_{gi}^{(j)}; \mu_{g,com} + \beta_{g\ell}, \sigma_g^2),$$

$\ell = 1, \ldots, L$ and $k = 1, \ldots, K$.

The other steps are the same as those in the MCMC sampling scheme in Section S7.

We acknowledge that the similar variable selection feature can be applied to subject-specific effects $\beta_g^{(j)}$ and even cell-type proportions parameters $\boldsymbol{\pi}_\ell$. However, the procedure would further complicate the derivation of

the sampling scheme and increase the computation burden of the MCMC algorithm especially for a large gene number and a large cell-type number. How to address the question efficiently and effectively is an interesting research direction in this project, and we thus leave it to the future work.

S7 Blocked Gibbs sampler

Given the priors and Model (4.2), we utilize the blocked Gibbs sampler to carry out the posterior sampling: ("$-$" means given all other variables)

1 The augmented parameter $\theta_{gi}^{(j)}$ in the PLN distribution is generated from

$$p(\theta_{gi}^{(j)}|-) \propto \exp\left\{-s_i^{(j)} e^{\theta_{gi}^{(j)}} + Y_{gi}^{(j)}\theta_{gi}^{(j)} - \frac{(\theta_{gi}^{(j)} - \mu_{gk} - \beta_{g\ell})^2}{2\sigma_g^2}\right\},$$

when $S^{(j)} = \ell$ and $C_i^{(j)} = k$.

2 Sample the missing variable $Y_{gi}^{(j)}$ for which its observation $X_{gi}^{(j)}$ equals zero from

$$p(Y_{gi}^{(j)}|-) \propto \begin{cases} (s_i^{(j)} e^{\theta_{gi}^{(j)}})^{Y_{gi}^{(j)}} / Y_{gi}^{(j)}! \cdot \Phi(\lambda_{g0} + \lambda_{g1}\log_2(Y_{gi}^{(j)} + 1)) & \text{if } Y_{gi}^{(j)} \geq 1 \\ 1 & \text{if } Y_{gi}^{(j)} = 0 \end{cases}.$$

3 Update the zero-inflation intensity parameters λ_{g0} and λ_{g1} by gener-

ating

$$p(\lambda_{g0}, \lambda_{g1}|-) \propto \prod_{(j,i):X_{gi}^{(j)}>0} \left(1 - \Phi(\lambda_{g0} + \lambda_{g1}\log_2(Y_{gi}^{(j)}+1))\right) \cdot \prod_{(j,i):X_{gi}^{(j)}=0, Y_{gi}^{(j)}>0} \Phi\left(\lambda_{g0} + \lambda_{g1}\log_2(Y_{gi}^{(j)}+1)\right) \cdot \mathrm{N}(\lambda_{g0}; \eta_{\lambda_{g0}}, \tau_{\lambda_{g0}}^2) \cdot \mathrm{N}(\lambda_{g1}; \eta_{\lambda_{g1}}, \tau_{\lambda_{g1}}^2)\mathbb{I}(\lambda_{g1} < 0),$$

where the $\mathbb{I}(A)$ is an indicator function, being one if A is true and zero otherwise; the $\mathrm{N}(x; a, b^2)$ represents the density value at x of a normal distribution with mean a and standard deviation b.

4 Sample the cell-type k effect on gene g, μ_{gk}, from the normal distribution

$$\mathrm{N}\left(\frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(\theta_{gi}^{(j)} - \beta_{gS^{(j)}})\mathbb{I}(C_i^{(j)}=k)/\sigma_g^2 + \eta_\mu/\tau_\mu^2}{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mathbb{I}(C_i^{(j)}=k)/\sigma_g^2 + 1/\tau_\mu^2}, \frac{1}{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mathbb{I}(C_i^{(j)}=k)/\sigma_g^2 + 1/\tau_\mu^2}\right).$$

5 Update the hyper-parameters in the cell-type effect prior,

$$\eta_\mu|- \sim \mathrm{N}\left(\frac{\sum_{g=1}^{D}\sum_{k=1}^{K}\mu_{gk}/\tau_\mu^2 + u_\mu/\omega_\mu^2}{DK/\tau_\mu^2 + 1/\omega_\mu^2}, \frac{1}{DK/\tau_\mu^2 + 1/\omega_\mu^2}\right),$$
$$\tau_\mu^2|- \sim \mathrm{Inv}\Gamma\left(b_{\mu 1} + DK/2, b_{\mu 2} + \sum_{g=1}^{D}\sum_{k=1}^{K}(\mu_{gk} - \eta_\mu)^2/2\right).$$

6 The subgroup ℓ effect on gene g for $\ell \geq 2$, $\beta_{g\ell}$, is sampled from the

normal distribution

$$\mathrm{N}\left(\frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(\theta_{gi}^{(j)}-\mu_{gC_i^{(j)}})\mathbb{I}(S^{(j)}=\ell)/\sigma_g^2+\eta_\beta/\tau_\beta^2}{\sum_{j=1}^{m}\mathbb{I}(S^{(j)}=\ell)n_j/\sigma_g^2+1/\tau_\beta^2},\frac{1}{\sum_{j=1}^{m}\mathbb{I}(S^{(j)}=\ell)n_j/\sigma_g^2+1/\tau_\beta^2}\right).$$

We notice that subgroup 1 effect, β_{g1}, is restricted to zero across $1 \le g \le D$ for identifying the subgroup and cell-type effects.

7 Update the hyper-parameters in the subgroup effect prior,

$$\eta_\beta|- \sim \mathrm{N}\left(\frac{\sum_{g=1}^{D}\sum_{\ell=2}^{L}\beta_{g\ell}/\tau_\beta^2+u_\beta/\omega_\beta^2}{D(L-1)/\tau_\beta^2+1/\omega_\beta^2},\frac{1}{D(L-1)/\tau_\beta^2+1/\omega_\beta^2}\right),$$

$$\tau_\beta^2|- \sim \mathrm{Inv}\Gamma\left(b_{\beta1}+D(L-1)/2, b_{\beta2}+\sum_{g=1}^{D}\sum_{\ell=2}^{L}(\beta_{g\ell}-\eta_\beta)^2/2\right).$$

8 Update the variance σ_g^2 for gene g by sampling from the inverse-gamma distribution

$$\mathrm{Inv}\Gamma(b_{\sigma1}+\sum_{j=1}^{m}n_j/2, b_{\sigma2}+\sum_{j=1}^{m}\sum_{i=1}^{n_j}(\theta_{gi}^{(j)}-\mu_{gC_i^{(j)}}-\beta_{gS^{(j)}})^2/2).$$

9 For each subject j, update the subtype indicator $S^{(j)}$ and the cell-type indicators $C_i^{(j)}$ for cell $i = 1, \ldots, n_j$ from multinomial distributions

$$P(S^{(j)}=\ell|-) \propto \phi_\ell \prod_{i=1}^{n_j}\left[\sum_{k=1}^{K}\pi_{k\ell}\prod_{g=1}^{D}\mathrm{N}(\theta_{gi}^{(j)};\mu_{gk}+\beta_{g\ell},\sigma_g^2)\right]$$

$$P(C_i^{(j)}=k|S^{(j)}=\ell,-) \propto \pi_{k\ell}\prod_{g=1}^{D}\mathrm{N}(\theta_{gi}^{(j)};\mu_{gk}+\beta_{g\ell},\sigma_g^2),$$

$\ell = 1, \ldots, L$ and $k = 1, \ldots, K$.

10 Update the subtype proportion vector $(\phi_1, \phi_2, \ldots, \phi_L)$. We first sample

$$\phi'_\ell \sim \text{Beta}(1 + m_\ell, \nu + \sum_{j=\ell+1}^{L} m_j)$$

for $\ell = 1, \ldots, L-1$ and $\phi'_L := 1$, where $m_\ell = \#\{j : S^{(j)} = \ell\}$ is the number of subjects allocated to subgroup ℓ. Subsequently, we let $\phi_1 = \phi'_1$ and $\phi_\ell = \phi'_\ell \prod_{i=1}^{\ell-1}(1-\phi'_i)$ for $\ell = 2, \ldots, L$.

11 Update the stick-breaking length vector $(\xi_1, \xi_2, \ldots, \xi_K)$. We first sample

$$p((\xi'_1, \ldots, \xi'_{K-1})|-) \propto \frac{\prod_{k=1}^{K-1}(1-\xi'_k)^{\alpha-1}}{\prod_{k=1}^{K}\Gamma^L(\gamma\prod_{i=1}^{k-1}(1-\xi'_i)\xi'_k)} \prod_{k=1}^{K}\left(\prod_{\ell=1}^{L}\pi_{k\ell}\right)^{\gamma\prod_{i=1}^{k-1}(1-\xi'_i)\xi'_k-1}.$$

and $\xi'_K := 1$. Subsequently, we let $\xi_1 = \xi'_1$ and $\xi_k = \prod_{i=1}^{k-1}(1-\xi'_i)\xi'_k$ for $k = 2, \ldots, K$.

12 The concentration parameter α is sampled from the gamma distribution

$$\Gamma\left(a_{\alpha 1} + K - 1, a_{\alpha 2} - \sum_{k=1}^{K-1}\log(1-\xi'_k)\right),$$

where and ξ'_ks are the variables generated in the previous iteration.

13 Sample the cell type proportions $(\pi_{1\ell}, \ldots, \pi_{K\ell})$ for each subtype ℓ from the Dirichlet distribution

$$\text{Dir}(\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mathbb{I}(S^{(j)} = \ell, C_i^{(j)} = 1) + \gamma\xi_1, \ldots, \sum_{j=1}^{m}\sum_{i=1}^{n_j}\mathbb{I}(S^{(j)} = \ell, C_i^{(j)} = K) + \gamma\xi_K).$$

S8 Proposal distributions and acceptance rates

S8.1 Details of the MH algorithm in steps 1, 2, 3 and 11

Updating of $\theta_{gi}^{(j)}$

The proposal density $q(\theta_{gi}^{(j)*}|\theta_{gi}^{(j)})$ for $\theta_{gi}^{(j)*}$ is set to $\mathrm{N}(\theta_{gi}^{(j)}, \tau_\theta^2)$. We set $\tau_\theta = 1$ in our implementation. The Metropolis-Hastings ratio is $r = \min(r^*, 1)$, where

$$
\begin{aligned}
r^* &= \frac{p(\theta_{gi}^{(j)*}|-)q(\theta_{gi}^{(j)}|\theta_{gi}^{(j)*})}{p(\theta_{gi}^{(j)}|-)q(\theta_{gi}^{(j)*}|\theta_{gi}^{(j)})} \\
&= \exp\left(-\left(e^{\theta_{gi}^{(j)*}} - e^{\theta_{gi}^{(j)}}\right) s_i^{(j)} + Y_{gi}^{(j)}\left(\theta_{gi}^{(j)*} - \theta_{gi}^{(j)}\right) - \frac{\left(\theta_{gi}^{(j)*} - \theta_{gi}^{(j)}\right)\left(\theta_{gi}^{(j)*} + \theta_{gi}^{(j)} - 2(\mu_{gk} + \beta_{g\ell})\right)}{2\sigma_g^2} \right).
\end{aligned}
$$

Updating of $Y_{gi}^{(j)}$

Proposal distribution 1: We generate a proposal $Y_{gi}^{(j)*}$ from a discrete uniform distribution

$$
q(Y_{gi}^{(j)*}|Y_{gi}^{(j)}) = Unif\{Y_{gi}^{(j)} - 5, Y_{gi}^{(j)} + 5\},
$$

where $Unif\{a, b\}$ samples an integer from $[a, b]$ uniformly.

The Metropolis-Hastings ratio is $r = \min(r^*, 1)$, where

$$r^* = \frac{p(Y_{gi}^{(j)*}|-)q(Y_{gi}^{(j)}|Y_{gi}^{(j)*})}{p(Y_{gi}^{(j)}|-)q(Y_{gi}^{(j)*}|Y_{gi}^{(j)})}$$

$$= \begin{cases} \exp\left(\left(\theta_{gi}^{(j)} + \log(s_i^{(j)})\right)\left(Y_{gi}^{(j)*} - Y_{gi}^{(j)}\right)\right)\frac{Y_{gi}^{(j)}!}{Y_{gi}^{(j)*}!} & \text{if } Y_{gi}^{(j)} = 0, Y_{gi}^{(j)*} = 0 \\ \exp\left(\left(\theta_{gi}^{(j)} + \log(s_i^{(j)})\right)\left(Y_{gi}^{(j)*} - Y_{gi}^{(j)}\right)\right)\frac{Y_{gi}^{(j)}!}{Y_{gi}^{(j)*}!}\frac{1}{\Phi(\lambda_{g0}+\lambda_{g1}\log_2(Y_{gi}^{(j)}+1))} & \text{if } Y_{gi}^{(j)} \geq 1, Y_{gi}^{(j)*} = 0 \\ \exp\left(\left(\theta_{gi}^{(j)} + \log(s_i^{(j)})\right)\left(Y_{gi}^{(j)*} - Y_{gi}^{(j)}\right)\right)\frac{Y_{gi}^{(j)}!}{Y_{gi}^{(j)*}!}\Phi(\lambda_{g0} + \lambda_{g1}\log_2(Y_{gi}^{(j)*}+1)) & \text{if } Y_{gi}^{(j)} = 0, Y_{gi}^{(j)*} \geq 1 \\ \exp\left(\left(\theta_{gi}^{(j)} + \log(s_i^{(j)})\right)\left(Y_{gi}^{(j)*} - Y_{gi}^{(j)}\right)\right)\frac{Y_{gi}^{(j)}!}{Y_{gi}^{(j)*}!}\frac{\Phi(\lambda_{g0}+\lambda_{g1}\log_2(Y_{gi}^{(j)*}+1))}{\Phi(\lambda_{g0}+\lambda_{g1}\log_2(Y_{gi}^{(j)}+1))} & \text{if } Y_{gi}^{(j)} \geq 1, Y_{gi}^{(j)*} \geq 1 \\ 0 & \text{else} \end{cases}$$

Proposal distribution 2: We generate a proposal $Y_{gi}^{(j)*}$ from a Poisson distribution which relies on the newly updated $\theta_{gi}^{(j)}$,

$$\text{Poi}(s_i^{(j)} \exp(\theta_{gi}^{(j)})),$$

where $\text{Poi}(a)$ represents a Poisson distribution with mean a.

The Metropolis-Hastings ratio is $r = \min(r^*, 1)$, where

$$r^* = \begin{cases} 1 & \text{if } Y_{gi}^{(j)} = 0, Y_{gi}^{(j)*} = 0 \\ \frac{1}{\Phi(\lambda_{g0}+\lambda_{g1}\log_2(Y_{gi}^{(j)}+1))} & \text{if } Y_{gi}^{(j)} \geq 1, Y_{gi}^{(j)*} = 0 \\ \Phi(\lambda_{g0} + \lambda_{g1}\log_2(Y_{gi}^{(j)*}+1)) & \text{if } Y_{gi}^{(j)} = 0, Y_{gi}^{(j)*} \geq 1 \\ \frac{\Phi(\lambda_{g0}+\lambda_{g1}\log_2(Y_{gi}^{(j)*}+1))}{\Phi(\lambda_{g0}+\lambda_{g1}\log_2(Y_{gi}^{(j)}+1))} & \text{if } Y_{gi}^{(j)} \geq 1, Y_{gi}^{(j)*} \geq 1 \\ 0 & \text{else} \end{cases}$$

Updating of λ_{g0} and λ_{g1}

The proposal densities $q(\lambda^*_{g0}|\lambda_{g0})$ and $q(\lambda^*_{g1}|\lambda_{g1})$ for λ_{g0} and λ_{g1} are $\mathrm{N}(\lambda_{g0}, \tau^2_{\lambda_0})$ and $\mathrm{N}(\lambda_{g1}, \tau^2_{\lambda_1})$, respectively. We set $\tau_{\lambda_0} = 0.15$ and $\tau_{\lambda_1} = 0.15$ in our implementation. The Metropolis-Hastings ratio is $r = \min(r^*, 1)$, where

$$\begin{aligned}
r^* =& \frac{p(\lambda^*_{g0}, \lambda^*_{g1}|-)q(\lambda_{g0}, \lambda_{g1}|\lambda^*_{g0}, \lambda^*_{g1})}{p(\lambda_{g0}, \lambda_{g1}|-)q(\lambda^*_{g0}, \lambda^*_{g1}|\lambda_{g0}, \lambda_{g1})} \\
=& \frac{p(\lambda^*_{g0}, \lambda^*_{g1}|-)q(\lambda_{g1}|\lambda^*_{g1})q(\lambda_{g0}|\lambda^*_{g0})}{p(\lambda_{g0}, \lambda_{g1}|-)q(\lambda^*_{g1}|\lambda_{g1})q(\lambda^*_{g0}|\lambda_{g0})} \\
=& \prod_{(j,i):X^{(j)}_{gi}=0, Y^{(j)}_{gi}>0} \frac{\Phi\left(\lambda^*_{g0} + \lambda^*_{g1}\log_2(Y^{(j)}_{gi}+1)\right)}{\Phi\left(\lambda_{g0} + \lambda_{g1}\log_2(Y^{(j)}_{gi}+1)\right)} \prod_{(j,i):X^{(j)}_{gi}>0} \frac{1-\Phi\left(\lambda^*_{g0} + \lambda^*_{g1}\log_2(Y^{(j)}_{gi}+1)\right)}{1-\Phi\left(\lambda_{g0} + \lambda_{g1}\log_2(Y^{(j)}_{gi}+1)\right)} \\
& \cdot \exp\left(-\frac{\left(\lambda^*_{g0} - \lambda_{g0}\right)\left(\lambda^*_{g0} + \lambda_{g0} - 2\eta_{\lambda_{g0}}\right)}{2\tau^2_{\lambda_{g0}}} - \frac{\left(\lambda^*_{g1} - \lambda_{g1}\right)\left(\lambda^*_{g1} + \lambda_{g1} - 2\eta_{\lambda_{g1}}\right)}{2\tau^2_{\lambda_{g1}}}\right) \mathbb{I}(\lambda^*_{g1} < 0).
\end{aligned}$$

Updating of $(\xi_1, \xi_2, \ldots, \xi_K)$

The proposal densities $q(\xi'^*_k|\xi'_k)$ for $\xi'_k (k = 1, 2, \cdots, K-1)$ is $\mathrm{N}(\xi'_k, \tau^2_{\xi'})$. We set $\tau_{\xi'} = 0.01$ in our implementation. The Metropolis-Hastings ratio is $r = \min(r^*, 1)$, where

$$\begin{aligned}
r^* &= \frac{p((\xi'^*_1, \ldots, \xi'^*_{K-1})|-)\prod_{k=1}^{K-1} q(\xi'_k|\xi'^*_k)}{p((\xi'_1, \ldots, \xi'_{K-1})|-)\prod_{k=1}^{K-1} q(\xi'^*_k|\xi'_k)} \\
&= \prod_{k=1}^{K-1}\left(\frac{1-\xi'^*_k}{1-\xi'_k}\right)^{\alpha-1} \prod_{k=1}^{K} \frac{\Gamma^L(\gamma\xi_k)}{\Gamma^L(\gamma\xi^*_k)} \prod_{k=1}^{K}\left(\prod_{\ell=1}^{L}\pi_{k\ell}\right)^{\gamma(\xi^*_k - \xi_k)},
\end{aligned}$$

where $\xi^*_1 = \xi'^*_1$ and $\xi^*_k = \prod_{i=1}^{k-1}(1-\xi'^*_i)\xi'^*_k$ for $k = 2, \ldots, K$.

S8.2 Acceptance rates

For the variance of proposal distributions in the MH steps, we actually tried multiple values and selected the one with the maximal effective sample size (Gelman et al., 2013). Using the current variance specification, the average acceptance rates for parameters $(\boldsymbol{\theta}, \boldsymbol{Y}, \boldsymbol{\xi}, \boldsymbol{\lambda})$ are $(0.155, 0.650, 0.263, 0.082)$ in the simulation study. We acknowledge that some adaptive schemes (e.g., Roberts and Rosenthal (2009)) can be adopted and could further improve the effective size.

For the missing variables $\boldsymbol{Y}$, if we used the uniform proposal distribution, the average acceptance rate was 0.650. If we employed the $\text{Pois}(s_i^{(j)} \exp(\theta_{gi}^{(j)}))$ proposal distribution, we can obtain a higher average acceptance rate 0.806. The trace plots for the two types of proposal distribution are shown in Figure S8.

S9 Data generation details in simulation

We generated data following Model (4.2) with three subject subgroups, four cell types. The subject size was selected as 50, and for each subject we sampled its corresponding cell number from a uniform distribution on integers between 15 and 35. The subject proportions for the three subgroups

were 40%, 30%, and 30%. For each subject subgroup, there were different cell type proportions $\boldsymbol{\pi}_\ell$ ($\ell = 1, 2, 3$). Subject subgroup 1 had 20%, 30%, 30%, and 20% for cell types one to four, respectively. If we denote those as $\boldsymbol{\pi}_1 = (0.2, 0.3, 0.3, 0.2)$, then we set $\boldsymbol{\pi}_2 = (0.4, 0.2, 0.3, 0.1)$ for subject subgroup 2, and $\boldsymbol{\pi}_3 = (0.3, 0.1, 0.3, 0.3)$ for subject subgroup 3. The number of genes for each cell was $1,000$, and the first 150 genes were treated as marker genes with differential expressions between at least two cell types, whereas the remaining genes had the same cell effects across all cell types. Regarding the subject subgroup effects, the subject subgroup 1 effects were fixed at zero. Using subgroup 1 as the reference, subgroup 2 had marker genes from 401 to 475, and subgroup 3 had marker genes from 501 to 575.

We generated cell type effects μ_{gk}s from normal distributions $N(\tilde{\mu}_{gk}, \sigma^2_{\mu_g})$. We set $\tilde{\mu}_{g1}$s of the first 75 genes to 2, and $\tilde{\mu}_{g1}$s of the genes from 76 to 150 to 1 in cell type 1. Subsequently, we set $\tilde{\mu}_{g2} = 5$ ($1 \leq g \leq 75$) and $\tilde{\mu}_{g2} = 1$ ($76 \leq g \leq 150$) in cell type 2; let $\tilde{\mu}_{g3}$ be 2 on the first 75 genes and $\tilde{\mu}_{g3}$ be 5 on the second 75 genes in cell type 3; and similarly in cell type 4 $\tilde{\mu}_{g4} = 5$ on $1 \leq g \leq 75$ and $\tilde{\mu}_{g4} = 4$ on $76 \leq g \leq 150$. The $\tilde{\mu}_{gk}$s for the genes from 151 to 1,000 in all cell types were set to 3. We finally let $\sigma^2_{\mu_g}(1 \leq g \leq 150) = 0.2^2$ and $\sigma^2_{\mu_g}(151 \leq g \leq 1,000) = 0.1^2$.

With regard to the subject subgroup effects, we used the subgroup 1 as

the reference subgroup. The subgroup 2 had intrinsic genes from 401 to 475, while the subgroup 3 had intrinsic genes from 501 to 575. We generated subject subgroup effects $\beta_{g\ell}$s of intrinsic genes from a normal distribution $N(1, 0.2^2)$. The rest of subject subgroup effects $\beta_{g\ell}$s were set to 0.

The scaling factors were fixed at one. The dropout coefficients λ_{g0} and λ_{g1} were sampled respectively from $\mathrm{N}(3, 0.1^2)$ and $\mathrm{N}(-1, 0.1^2)$, and σ_g's were fixed at 0.1.

In simulation, 10,000 iterations took approximately 3.49 hours using 24 CPU cores and retained the second half of the posterior samples for statistical inference. The trace plots of parameters are given in Figures S8 and S9, which demonstrate that the chain had attained convergence by the 5,000th iteration. The posterior mode of the number of the occupied subject clusters was three, and the posterior mode of the occupied cell types was four, both of which are the same as the truth. To measure the clustering accuracy, we used the adjusted Rand index (ARI), which is bounded above by one, and the larger the ARI, the more accurate the clustering results. The SCSC produced a perfect clustering for the subjects and cells with both ARIs being one. Hence, the SCSC model can automatically and accurately distinguish the underlying heterogeneity for subjects and cells.

With the available posterior samples of $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$, we further detected

the DE genes across cell types or subject subgroups using Bayesian credible intervals. For example, to test if gene g is DE between cell type 1 and cell type k ($k \geq 2$), we constructed the 99% credible interval for the difference $\mu_{gk} - \mu_{g1}$ using the posterior samples. If zero was not in this credible interval, we treated the gene as DE. Otherwise, the gene was non-DE. Although we conducted multiple hypothesis tests, it was unnecessary to implement multiple comparison adjustments as $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ were modeled in a hierarchical Bayesian fashion (Gelman et al., 2012). A similar procedure was applied to detecting DE genes across subject subgroups, and SCSC correctly detected most DE genes with high true positive rates (TPR) and small false positive rates (FPR) as shown in Supplementary Table S1.

We compared the SCSC against some popular cell and subject clustering approaches, respectively. For the cell clustering approaches k-means, SC3, DIMM-SC, and Seurat, we stacked the expression matrices for all subjects by row and used this large expression matrix as the input. Regarding the subject clustering approaches kmeans, SparseKmeans, and BCPlaid, we calculated the row means of its corresponding expression matrix (logarithm transformed) for each subject and combined all row means to form a gene-by-subject aggregated expression matrix. Overall, SCSC performed better in both cell clustering and subject clustering. When clustering cells,

SCSC borrows information across multiple subjects and considers the subject differences. When grouping subjects, the model exploits the cell information of each subject to discover the subtle difference. Owing to the two-way information-sharing strategy, SCSC outperforms competing methods in both cell clustering and subject grouping.

S10 Label-switching correction

We designed a strategy to deal with the label-switching by borrowing the idea from relabelling algorithms (Stephens, 2000). Specifically, we assume that T posterior samples were collected from MCMC, and for t^{th} posterior sample let ρ_t^* represent a permutation of $\{1, 2, \ldots, L\}$ and ρ_t^{**} denote a permutation of $\{1, 2, \ldots, K\}$. For each t from 2 to T, we minimize the squared loss, $\min_{\rho_t^*, \rho_t^{**}} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \sum_{g=1}^{D} (\theta_{gi,t-1}^{(j)} - \mu_{g\rho_t^{**}(C_{it}^{(j)}),t-1} - \beta_{g\rho_t^*(S_t^{(j)}),t-1})^2$. Subsequently, we reorder subject subgroup indicators $S_t^{(j)}$ and cell type indicators $C_{it}^{(j)}$ by $S_t^{(j)} \leftarrow \rho_t^*(S_t^{(j)})$ and $C_{it}^{(j)} \leftarrow \rho_t^{**}(C_{it}^{(j)})$. In addition, we have to modify the posterior samples of parameters which depend on clustering indicators. If we let $z_{gk\ell,t}$ be $\mu_{g\rho^{**-1}(k),t} + \beta_{g\rho^{*-1}(\ell),t}$, then we have $\mu_{gk,t} \leftarrow z_{gk1,t}$, $\beta_{g1,t} \leftarrow 0$ and $\beta_{g\ell,t} \leftarrow z_{g1\ell,t} - \mu_{g1,t}$.

In fact, we found that in the simulation and real application the identity permutations are usually the solutions for ρ_t^*, ρ_t^{**}, indicating the MCMC

chain usually explores only one mode out of $K! \times L!$ equivalent modes in the target distribution.

S11 Details of Implementing Competing Methods in the Simulation Study

S11.1 Cell clustering methods

We call $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(\ell)}, \ldots, \mathbf{X}^{(m)})$ the raw count data and call $\log_2(\mathbf{X}+1)$ the transformed data.

Kmeans We applied Kmeans to the transformed data. Regarding the number of clusters, we used the function "fviz_nbclust" to estimate the optimal number, where the argument "method" was set to "wss". We let the cluster function argument "FUNcluster" to "kmeans". Other arguments in the function "fviz_nbclust" were default. Subsequently, we used the estimate of cluster number as the input of "centers". Other parameters in the function "kmeans" were set to default. The R package is available on `https://github.com/cran/factoextra`.

DIMM-SC DIMMSC aims to cluster droplet-based single cell transcriptomic data. It uses the Dirichlet mixture prior to characterize the variations across different cell clusters. An EM algorithm is used for

the parameter estimation. We used DIMMSC on the raw count data. The number of desired clusters "K" was set to the true cell type number in the simulation. Other parameters in the function "DIMMSC" were default. The R package is available on `http://www.pitt.edu/~wec47/singlecell.html`.

SC3 Single-Cell Consensus Clustering (SC3) is a tool for unsupervised clustering of scRNA-seq data. SC3 achieves high accuracy and robustness by consistently integrating different clustering solutions through a consensus approach. We firstly created an object of SingleCellExperiment class with "counts" and "logcounts" being the raw and transformed data, respectively. Subsequently, we used the funtion "sc3_estimate_k" with the default arguments to find the optimal number of clusters, which was 5. Finally, we set the number of clusters "ks" to 5 in the function "sc3". The argument "biology", which defines whether to compute differentially expressed genes, marker genes and cell outliers, was set to "FALSE". The number of cores to be used on the machine "n_cores" was set to 20, and the seed of the random number generator "rand_seed" was set to 1. Other arguments in the function "sc3" were default. The R package is available on `http://bioconductor.org/packages/release/bioc/html/SC3.html`.

Seurat Seurat identified cell types through a shared nearest neighbor (SNN) modularity optimization based clustering algorithm. We used the raw count data in this method. The parameters in the function "FindClusters" were default. The R package is available on `https://github.com/satijalab/seurat`.

S11.2 Subject clustering methods

For each subject, we transformed the raw data matrix $\mathbf{X}^{(\ell)}$ to $rowMeans(log2(\mathbf{X}^{(\ell)}+1))$, the vector of row-means of $\log_2(\mathbf{X}^{(\ell)}+1)$.

Kmeans Regarding the number of clusters, we used the function "fviz_nbclust" to estimate the optimal number, where the argument "method" was set to "wss". We let the cluster function argument "FUNcluster" to "kmeans". Other arguments in the function "fviz_nbclust" were default. Subsequently, we used the estimate of cluster number as the input of "centers". Other parameters in the function "kmeans" were set to default. The R package is available on `https://github.com/cran/factoextra`.

SparseKmeans We used the same number of clusters in Kmeans. Firstly, we used the function "KMeansSparseCluster.permute" to find the tuning parameter which controls the L1 bound on the feature weights. The

range of tuning parameters "wbounds" was set to "seq(3,7,len=15)". The number of permutations "nperms" was 5. We then used the estimate of tunning parameter in the function "KMeansSparseCluster". Other parameters in the function "KMeansSparseCluster" were default. The R package is available on `https://github.com/cran/sparcl`.

BCPlaid This method performs Plaid Model Biclustering. This algorithm models data matrices via a sum of layers. We set the argument "method" in the function "biclust" to "BCPlaid". Other parameters in the function "biclust" were set to default. The package is available on `https://github.com/cran/biclust`

S12 Low signal scenarios and model misspecification cases

S12.1 Low signal scenarios

We further investigated the performances of SCSC and SCSC-vs in two scenarios with low signals. (1) We reduced the number of cell marker genes from 150 to 100 and 50. Supplementary Table S2 provides the ARI comparisons, showing that SCSC-vs has better clustering performance than

SCSC when cell marker gene number is small. When cell marker gene number is relatively high, the performances of SCSC and SCSC-vs are similar, but SCSC-vs has the advantage of automatically selecting important genes. Supplementary Table S3 displays the FPR and TPR of SCSC-vs in identifying the cell marker genes.

(2) The subject subgroup effects are set as zero in all the cell types $\beta_g^{(j)} = 0$. In this case, only the cell-type proportions contribute to the subject clustering. Supplementary Table S4 shows that compared to the ideal case where $\beta_g^{(j)}$ is nonzero on subject marker genes, the clustering result for cells are still good, but the mean ARI for subjects decreases. The clustering accuracy loss is reasonable as we lost the differential subgroup effect information and only resorted to the cellular composition information to separate subjects.

S12.2 Model misspecification cases

The performances of SCSC was also examined on three types of cases, where the model assumptions are violated. (1) Inspired by the paper (Shen-Orr et al., 2010), we let $\beta_g^{(j)}$ be present in only one cell type. Specifically, we let subject subgroup effects be 2 on genes from 400 to 600 and zero on other genes in one cell type and be zero across all genes in all other cell types.

Supplementary Table S4 displays the clustering results of SCSC and SCSC-vs in this scenario. We observed that the mean ARI values decreased for subject clustering compared to the case where assumptions hold (correct specification), while cell clustering is still satisfactory, and SCSC-vs has overall better performances than SCSC.

Moreover, we considered two more cases, where (2) there exist gene correlations with a large number of cell types and (3) the data distribution is from zero-inflated negative binomial distributions instead of zero-inflated Poisson log-normal distribution. Details are given as follows. Supplementary Table S5 provides the average and standard deviation of ARI after 10 replications, showing that SCSC is robust to cell number per subject, cell-type number, the violation of the expression independence assumption, and the misspecification of the data distribution.

(2) **Gene correlations with a large number of cell types.** The advantage of Poisson-log-normal distribution over negative binomial distribution is that we can easily incorporate gene expression correlations through the correlated multi-normal distribution. Therefore, in this setting, we considered three factors—number of cells per subject, cell-type number, and expression correlation—based on the zero-inflated Poisson-log-normal distribution. Specifically, we set two subject subgroups and 20 cell types. The

subject number was 20, and the cell number for each subject was sampled from a uniform distribution on integers between 100 and 110. The gene number for each cell was $1,000$, and K=30, L=10. To model the correlated expression, we generated observed count data $\boldsymbol{X}$ as follows.

$$\begin{aligned}
&\boldsymbol{\theta}_i^{(j)} \sim \mathrm{N}(\boldsymbol{\mu}_i^{(j)} + \boldsymbol{\beta}^{(j)}, \boldsymbol{\Sigma}),\\
&Y_{gi}^{(j)} \sim \mathrm{Poi}(s_i^{(j)} e^{\theta_{gi}^{(j)}}),\\
&X_{gi}^{(j)} = \begin{cases} 0 & \text{with probability } p(Y_{gi}^{(j)}) \\ Y_{gi}^{(j)} & \text{with probability } 1 - p(Y_{gi}^{(j)}) \end{cases},
\end{aligned}$$

where $\boldsymbol{\Sigma}$ was set as a blocked diagonal matrix, and each block was a 50×50 matrix with diagonal elements 0.3 and off-diagonal elements 0.1. We then applied SCSC and investigated its clustering performance. Table S5 provides the average and standard deviation of ARI after 10 replications, showing that SCSC is robust to cell number per subject, cell-type number, and the violation of the expression independence assumption.

(3) **Zero-inflated negative binomial distributions.** We considered a model-misspecified case where the observed count data were generated from a zero-inflated negative binomial distribution rather than zero-inflated Poisson-log-normal. In this case, we set 3 subgroups and 4 cell types. The subject number was 50, and cell number for each subject was sampled from

a uniform distribution on integers between 15 and 35. The gene number for each cell was $1,000$. Noting that negative binomial is equivalent to Poisson-gamma, we simulated observed data $\boldsymbol{X}$ as follows.

$$T_{gi}^{(j)} \sim \text{Gamma}\left(\frac{1}{e^{\sigma_g^2}-1}, \frac{1}{(e^{\sigma_g^2}-1)e^{\mu_{gi}^{(j)}+\beta_g^{(j)}+\sigma_g^2/2}}\right),$$

$$Y_{gi}^{(j)} \sim \text{Poi}(s_i^{(j)} T_{gi}^{(j)}),$$

$$X_{gi}^{(j)} = \begin{cases} 0 & \text{with probability } p(Y_{gi}^{(j)}) \\ Y_{gi}^{(j)} & \text{with probability } 1 - p(Y_{gi}^{(j)}) \end{cases}.$$

Table S5 provides the average and standard deviation of ARI after 10 replications. SCSC can achieve mean ARI=0.87 for cell clustering and mean ARI=0.97 for subject clustering, indicating that SCSC is not sensitive to the data distribution choice.

S13 Sensitivity analyses

In real application, we carried out sensitivity analyses for hyper-parameters $a_{\alpha_1}, a_{\alpha_2}, u_\mu, \omega_\mu^2, b_{\mu 1}, b_{\mu 2}, u_\beta, \omega_\beta^2, b_{\beta 1}, b_{\beta 2}, b_{\sigma 1}, b_{\sigma 2}, \eta_{\lambda_{g0}}, \tau_{\lambda_{g0}}^2, \eta_{\lambda_{g1}}, \tau_{\lambda_{g1}}^2, \nu, \gamma, K$, and L. For hyper-parameters $a_{\alpha_1}, a_{\alpha_2}, b_{\mu 1}, b_{\mu 2}, b_{\beta 1}, b_{\beta 2}, b_{\sigma 1}, b_{\sigma 2}$, we varied their values on 1, 2, 3, and 4; for hyper-parameters u_μ, u_β, we tried values $-1, 0, 1$, and 2; for hyper-parameters $\omega_\mu^2, \omega_\beta^2$, we adjusted values from 20^2 to $50^2, 100^2$, and 150^2; for hyper-parameter $\eta_{\lambda_{g0}}$, we set it as 1, 2, 3 and 4; for hyper-

parameter $\eta_{\lambda_{g1}}$, we changed its value from -1 to -2, -3, and -4; for hyper-parameters $\tau^2_{\lambda_{g0}}, \tau^2_{\lambda_{g1}}$, their values were set at 0.05^2, 0.1^2, 0.2^2, and 0.3^2; for hyper-parameters ν, γ, we tried values 0.1, 0.3, 0.5 and 0.7; for upper bounds K, L, we varied their values from 13 to 16. Figure S5, S6 and S7 displays how these hyper-parameters change the final results compared to the obtained clustering results in terms of ARI: a large ARI indicates that the clustering in the current hyper-parameter setting is similar to the obtained clustering result shown in the manuscript. We can see that SCSC is a little bit sensitive to the choice of K, L and is robust to hyper-parameters $a_{\alpha_1}, a_{\alpha_2}, u_\mu, \omega^2_\mu, b_{\mu 1}, b_{\mu 2}, u_\beta, \omega^2_\beta, b_{\beta 1}, b_{\beta 2}, b_{\sigma 1}, b_{\sigma 2}, \eta_{\lambda_{g0}}, \tau^2_{\lambda_{g0}}, \eta_{\lambda_{g1}}, \tau^2_{\lambda_{g1}}, \nu$, and γ.

S14 Validation of clustering results in real application

To validate the clustering results, we conducted the gene set enrichment analysis (Subramanian et al., 2005) for detected marker genes based on the KEGG database. The maker genes for Yoruba subgroups were called if they were DE in at least one subgroup. Similarly, the marker genes for iPSC types were called if they were DE in at least one cell type. We identified 2,932 DE genes between the Yoruba subgroups and found 10 significant pathways with q-value < 0.05 (Supplementary Table S6), where three pathways (KEGG_PARKINSONS_DISEASE, KEGG_HUNTINGTONS_DISEASE,

and KEGG_ALZHEIMERS_DISEASE) are all related to neurodegenerative disorders. Previous studies (Myers, 2004; Bertram and Tanzi, 2008; Shulman et al., 2011) have presented that the three diseases are likely caused by inheritable gene defects, and thus can be inherited across generations. These observations suggest that SCSC separated the Yoruba subjects possibly in terms of the lineage. In addition, we detected 2,698 DE genes across the cell types and identified 87 significant pathways (Supplementary Tables S7 and S8) including KEGG_P53_SIGNALING_PATHWAY, KEGG_WNT_SIGNALING_PATHWAY, KEGG_NOTCH_SIGNALING_PATHWAY, and KEGG_MTOR_SIGNALING_PATHWAY. The four signaling pathways may regulate the pluripotency of induced stem cells (Ye et al., 2012; Kate et al., 2014; Meng et al., 2018). These findings indicate the validity of the SCSC in discovering subject and cell heterogeneity.

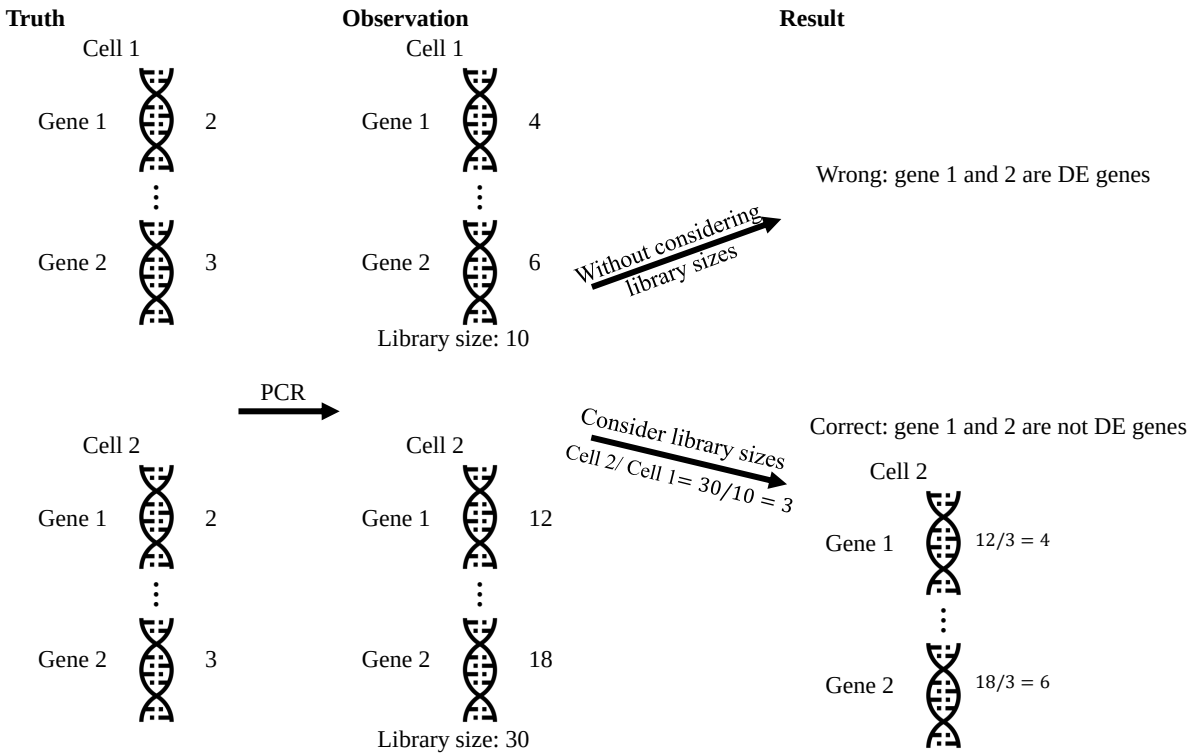

Figure S1: A simple demonstration of library sizes.

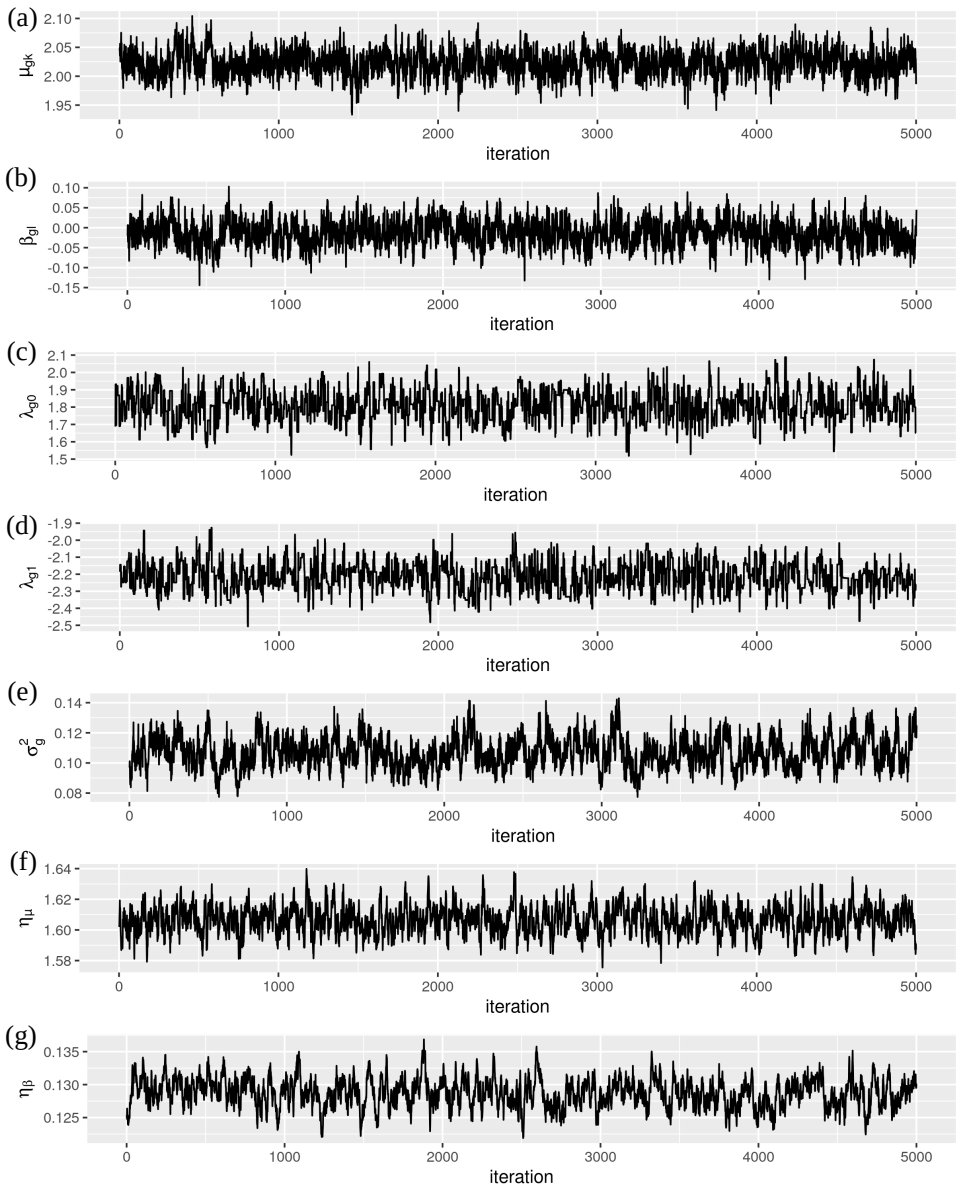

Figure S2: Trace plots for MCMC in the real application. (a)Trace plot of μ_{gk}. (b)Trace plot of β_{gl}. (c)Trace plot of λ_{g0}. (d)Trace plot of λ_{g1}. (e)Trace plot of σ_g^2. (f)Trace plot of η_μ. (g)Trace plot of η_β.

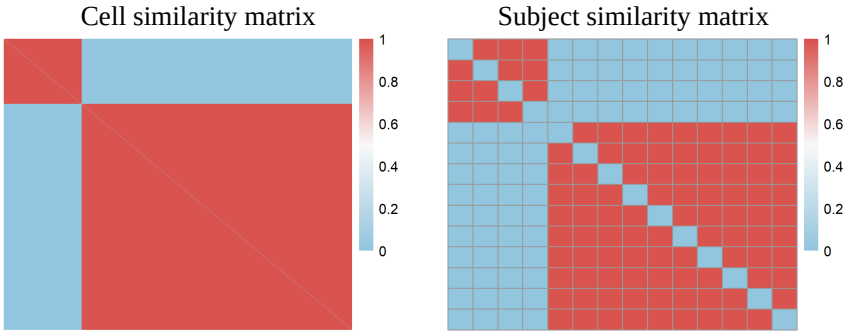

Figure S3: Posterior similarity matrix heatmaps for cells and subjects in the real application.

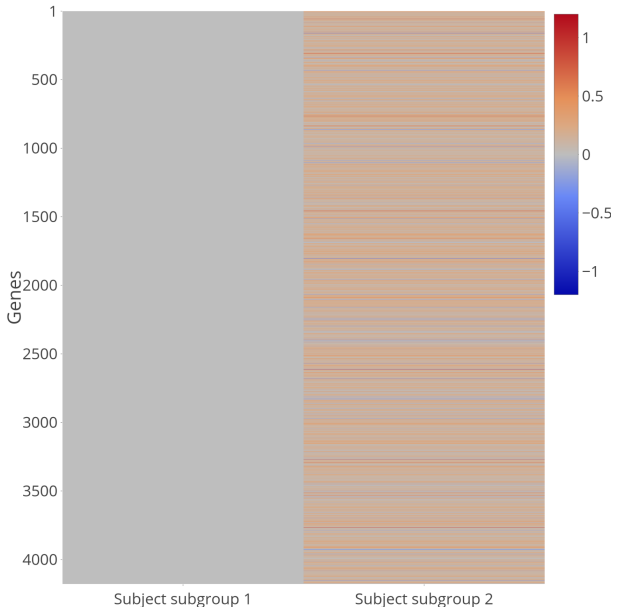

Figure S4: Estimated subject subgroup effects in the real application.

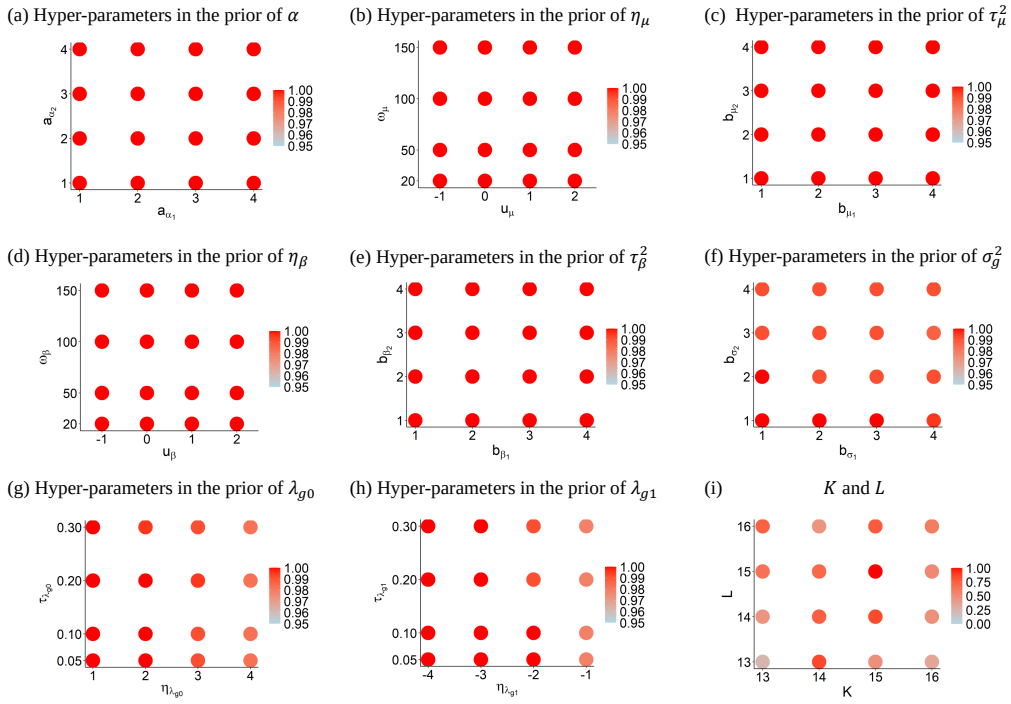

Figure S5: ARI values of cell clustering for different hyper-parameters, K and L. (a) ARI values for hyper-parameters in the prior of α. (b) ARI values for hyper-parameters in the prior of η_μ. (c) ARI values for hyper-parameters in the prior of τ_μ^2. (d) ARI values for hyper-parameters in the prior of η_β. (e) ARI values for hyper-parameters in the prior of τ_β^2. (f) ARI values for hyper-parameters in the prior of σ_g^2. (g) ARI values for hyper-parameters in the prior of λ_{g0}. (h) ARI values for hyper-parameters in the prior of λ_{g1}. (i) ARI values for K and L.

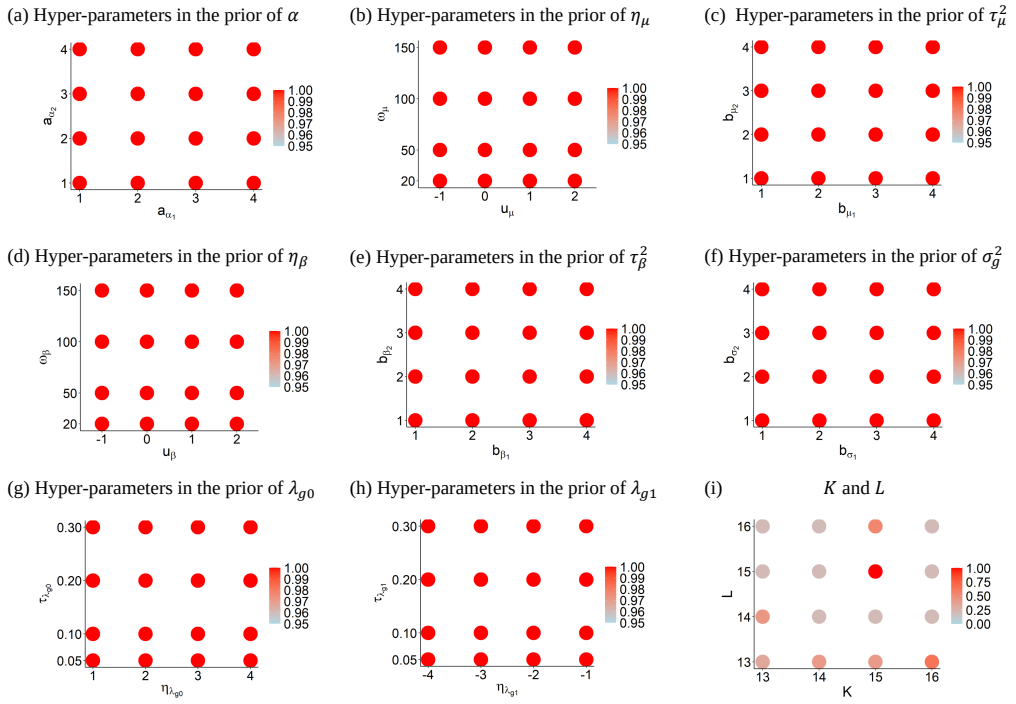

Figure S6: ARI values of subject clustering for different hyper-parameters, K and L. (a) ARI values for hyper-parameters in the prior of α. (b) ARI values for hyper-parameters in the prior of η_μ. (c) ARI values for hyper-parameters in the prior of τ_μ^2. (d) ARI values for hyper-parameters in the prior of η_β. (e) ARI values for hyper-parameters in the prior of τ_β^2. (f) ARI values for hyper-parameters in the prior of σ_g^2. (g) ARI values for hyper-parameters in the prior of λ_{g0}. (h) ARI values for hyper-parameters in the prior of λ_{g1}. (i) ARI values for K and L.

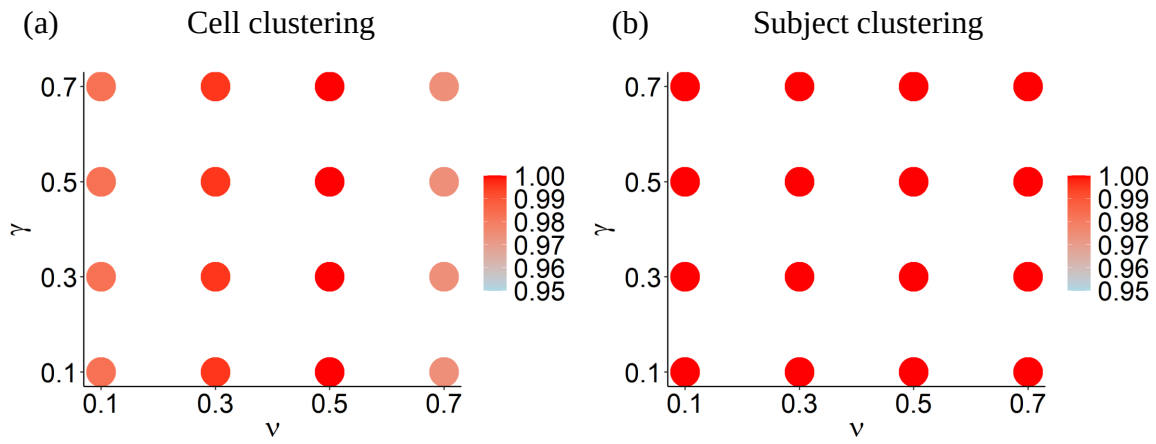

Figure S7: ARI values of the clustering for different pairs of γ and ν compared to the clustering result with $(\gamma, \nu) = (0.5, 0.5)$. (a) ARI values of cell clustering. (b) ARI values of subject clustering.

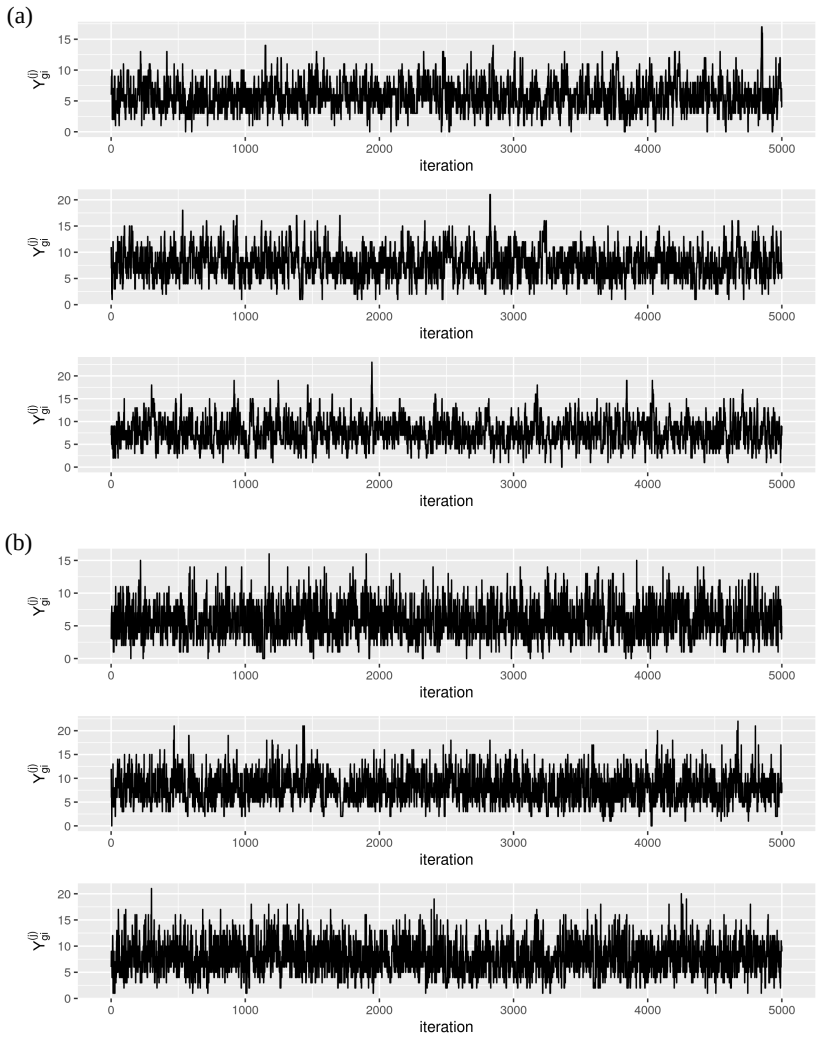

Figure S8: Trace plots of missing variables $\boldsymbol{Y}$ for 5,000 MCMC iterations after burn-in. (a) Uniform proposal distribution. (b) Pois($s_i^{(j)} \exp(\theta_{gi}^{(j)})$) proposal distribution.

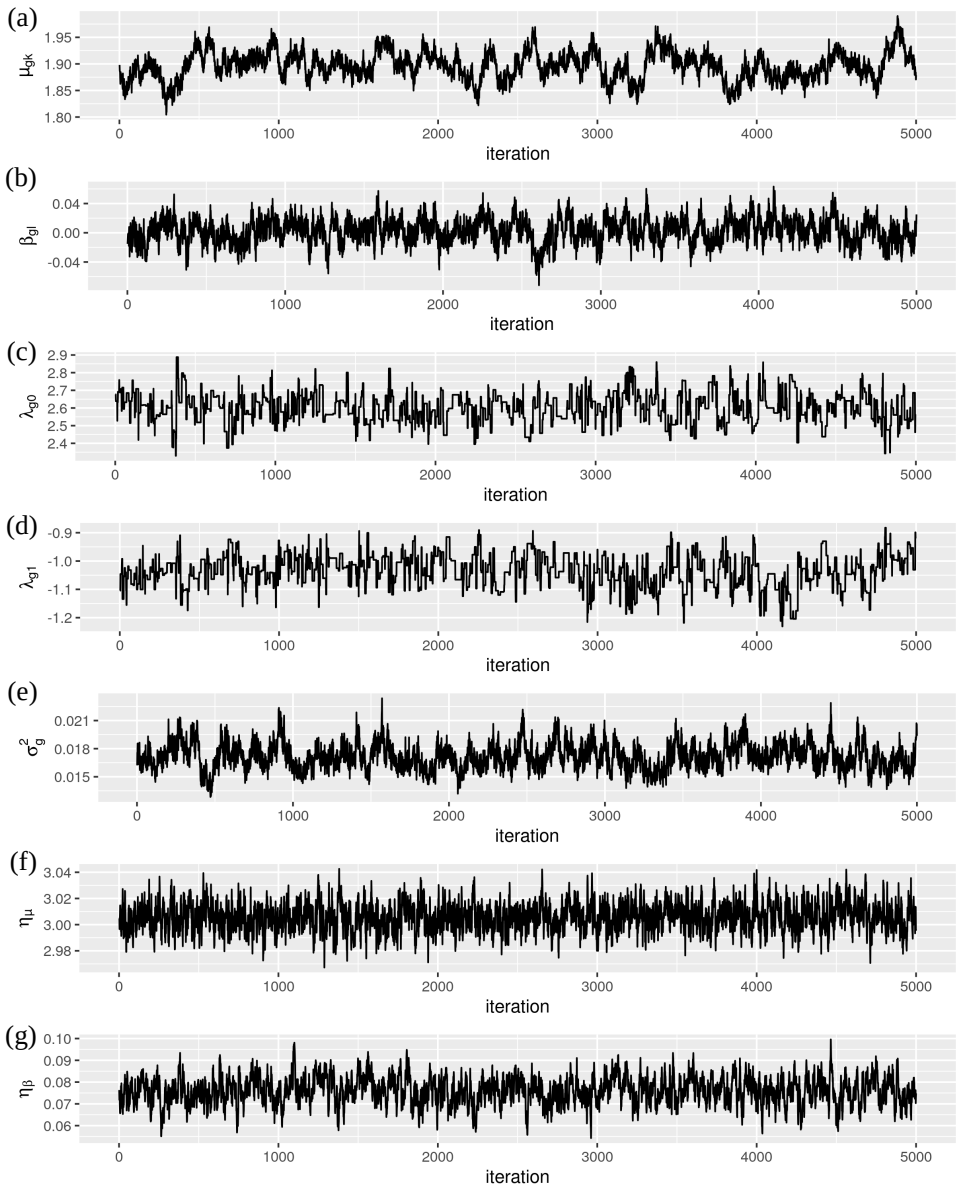

Figure S9: Trace plots for MCMC in the simulation. (a)Trace plot of μ_{gk}. (b)Trace plot of β_{gl}. (c)Trace plot of λ_{g0}. (d)Trace plot of λ_{g1}. (e)Trace plot of σ_g^2. (f)Trace plot of η_μ. (g)Trace plot of η_β.

Table S1: The TPR and FPR of the differentially expressed gene detection across subject subgroups and cell types in the simulation (three decimal places are kept). The subject subgroup 1 is chosen as the reference subject subgroup, and the cell type 1 is chosen as the reference cell type.

	Subject subgroup 2	Subject subgroup 3	cell type 2	cell type 3	cell type 4
FPR	0.006	0.003	0.011	0.008	0.006
TPR	1	1	1	1	1

Table S2: Clustering accuracy comparisons using ARI for SCSC and SCSC-vs based on 10 replicates. The number outside the parentheses is the mean, and the number in the parentheses represents the standard deviation.

	Cell marker gene number					
	50		100		150	
	Cell clustering	Subject clustering	Cell clustering	Subject clustering	Cell clustering	Subject clustering
SCSC	0.73(0.11)	0.91(0.10)	0.83(0.11)	0.81(0.11)	0.98(0.03)	0.96(0.04)
SCSC-vs	0.76(0.15)	0.94(0.07)	0.83(0.12)	0.95(0.05)	0.93(0.08)	0.95(0.08)

Table S3: TPR and FPR of SCSC-vs in identifying cell marker genes based on ten replicates. The number outside the parentheses is the mean, and the number in the parentheses represents the standard deviation.

		Cell marker gene number		
		50	100	150
SCSC-vs	TPR	1(0.00)	1(0.00)	1(0.00)
	FPR	0.00(0.00)	0.00(0.01)	0.00(0.00)

Table S4: Clustering accuracy using ARI for SCSC and SCSC-vs in different situations based on ten replicates. The number outside the parentheses is the mean, and the number in the parentheses represents the standard deviation.

	$\beta_g^{(j)}$ only on one cell type		All $\beta_g^{(j)} = 0$		Correct specification	
	Cell clustering	Subject clustering	Cell clustering	Subject clustering	Cell clustering	Subject clustering
SCSC	0.83(0.10)	0.46(0.14)	0.92(0.05)	0.33(0.14)	0.86(0.08)	0.87(0.08)
SCSC-vs	0.90(0.06)	0.57(0.25)	0.96(0.06)	0.36(0.15)	0.88(0.11)	0.89(0.09)

Table S5: Clustering accuracy using ARI for SCSC in different situations based on ten replicates. The number outside the parentheses is the mean, and the number in the parentheses represents the standard deviation.

	20 cell types with correlated genes	Zero-inflated negative binomial
Cell clustering	0.92(0.03)	0.87(0.12)
Subject clustering	0.77(0.19)	0.97(0.06)

Table S6: The gene set enrichment analysis for differentially expressed genes across subject subgroups. Since GSEA only allows up to 1,994 genes as input, we randomly selected 1,994 genes from 2,932 detected marker genes. Pathways with FDR q-value less than 0.05 are displayed. The pathways mentioned in Section S14 are colored in red.

Pathway names	Description	FDR q-value
KEGG_RIBOSOME	Ribosome	4.21E-41
KEGG_HUNTINGTONS_DISEASE	Huntington's disease	2.07E-40
KEGG_SPLICEOSOME	Spliceosome	2.57E-38
KEGG_PARKINSONS_DISEASE	Parkinson's disease	1.45E-36
KEGG_OXIDATIVE_PHOSPHORYLATION	Oxidative phosphorylation	6.38E-34
KEGG_ALZHEIMERS_DISEASE	Alzheimer's disease	3.12E-24
KEGG_PROTEASOME	Proteasome	7.72E-24
KEGG_CELL_CYCLE	Cell cycle	7.55E-17
KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	Ubiquitin mediated proteolysis	1.05E-15
KEGG_PROTEIN_EXPORT	Protein export	2.93E-10

Table S7: The gene set enrichment analysis for differentially expressed genes across cell types (Part I). Since GSEA only allows up to 1,994 genes as input, we randomly selected 1,994 genes from 2,698 detected marker genes. Pathways with FDR q-value less than 0.05 are displayed. The pathways mentioned in Section S14 are colored in red.

Pathway names	Description	FDR q-value
KEGG_RIBOSOME	Ribosome	1.56E-56
KEGG_HUNTINGTONS_DISEASE	Huntington's disease	4.90E-46
KEGG_SPLICEOSOME	Spliceosome	2.24E-38
KEGG_PARKINSONS_DISEASE	Parkinson's disease	8.35E-38
KEGG_OXIDATIVE_PHOSPHORYLATION	Oxidative phosphorylation	2.80E-36
KEGG_ALZHEIMERS_DISEASE	Alzheimer's disease	3.68E-33
KEGG_PROTEASOME	Proteasome	5.47E-21
KEGG_CELL_CYCLE	Cell cycle	4.13E-15
KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION	Pathogenic Escherichia coli infection	1.79E-12
KEGG_PURINE_METABOLISM	Purine metabolism	3.81E-11
KEGG_RNA_DEGRADATION	RNA degradation	5.16E-11
KEGG_PATHWAYS_IN_CANCER	Pathways in cancer	2.17E-10
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	Regulation of actin cytoskeleton	3.75E-10
KEGG_ADHERENS_JUNCTION	Adherens junction	4.13E-10
KEGG_OOCYTE_MEIOSIS	Oocyte meiosis	4.78E-10
KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	Ubiquitin mediated proteolysis	5.65E-09
KEGG_CARDIAC_MUSCLE_CONTRACTION	Cardiac muscle contraction	1.22E-08
KEGG_WNT_SIGNALING_PATHWAY	Wnt signaling pathway	6.58E-08
KEGG_PYRIMIDINE_METABOLISM	Pyrimidine metabolism	1.09E-07
KEGG_LYSOSOME	Lysosome	2.34E-07
KEGG_RNA_POLYMERASE	RNA polymerase	5.65E-07
KEGG_PROTEIN_EXPORT	Protein export	7.50E-07
KEGG_TIGHT_JUNCTION	Tight junction	1.09E-06
KEGG_FOCAL_ADHESION	Focal adhesion	1.49E-06
KEGG_GLUTATHIONE_METABOLISM	Glutathione metabolism	4.83E-06
KEGG_COLORECTAL_CANCER	Colorectal cancer	1.11E-05
KEGG_PROSTATE_CANCER	Prostate cancer	4.73E-05
KEGG_NUCLEOTIDE_EXCISION_REPAIR	Nucleotide excision repair	4.73E-05
KEGG_ONE_CARBON_POOL_BY_FOLATE	One carbon pool by folate	5.76E-05
KEGG_N_GLYCAN_BIOSYNTHESIS	N-Glycan biosynthesis	7.01E-05
KEGG_NOTCH_SIGNALING_PATHWAY	Notch signaling pathway	8.07E-05
KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	Vascular smooth muscle contraction	8.07E-05
KEGG_NEUROTROPHIN_SIGNALING_PATHWAY	Neurotrophin signaling pathway	8.07E-05
KEGG_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	Leukocyte transendothelial migration	8.75E-05
KEGG_VIBRIO_CHOLERAE_INFECTION	Vibrio cholerae infection	3.00E-04
KEGG_THYROID_CANCER	Thyroid cancer	3.08E-04
KEGG_PROGESTERONE_MEDIATED_OOCYTE_MATURATION	Progesterone-mediated oocyte maturation	3.52E-04
KEGG_P53_SIGNALING_PATHWAY	p53 signaling pathway	5.61E-04
KEGG_PYRUVATE_METABOLISM	Pyruvate metabolism	5.63E-04
KEGG_AMINOACYL_TRNA_BIOSYNTHESIS	Aminoacyl-tRNA biosynthesis	6.72E-04
KEGG_INSULIN_SIGNALING_PATHWAY	Insulin signaling pathway	6.77E-04
KEGG_LONG_TERM_POTENTIATION	Long-term potentiation	6.77E-04
KEGG_GLYCOLYSIS_GLUCONEOGENESIS	Glycolysis / Gluconeogenesis	9.01E-04
KEGG_ENDOCYTOSIS	Endocytosis	1.10E-03
KEGG_GLIOMA	Glioma	1.32E-03
KEGG_GAP_JUNCTION	Gap junction	1.93E-03
KEGG_RENAL_CELL_CARCINOMA	Renal cell carcinoma	2.44E-03
KEGG_CITRATE_CYCLE_TCA_CYCLE	Citrate cycle (TCA cycle)	2.56E-03

Table S8: The gene set enrichment analysis for differentially expressed genes across cell types (Part II). Since GSEA only allows up to 1,994 genes as input, we randomly selected 1,994 genes from 2,698 detected marker genes. Pathways with FDR q-value less than 0.05 are displayed. The pathways mentioned in Section S14 are colored in red.

Pathway names	Description	FDR q-value
KEGG_GLYOXYLATE_AND_DICARBOXYLATE_METABOLISM	Glyoxylate and dicarboxylate metabolism	3.09E-03
KEGG_CHRONIC_MYELOID_LEUKEMIA	Chronic myeloid leukemia	3.27E-03
KEGG_BLADDER_CANCER	Bladder cancer	3.27E-03
KEGG_FATTY_ACID_METABOLISM	Fatty acid metabolism	3.27E-03
KEGG_ENDOMETRIAL_CANCER	Endometrial cancer	3.27E-03
KEGG_TGF_BETA_SIGNALING_PATHWAY	TGF-beta signaling pathway	3.66E-03
KEGG_STEROID_BIOSYNTHESIS	Steroid biosynthesis	3.74E-03
KEGG_CYSTEINE_AND_METHIONINE_METABOLISM	Cysteine and methionine metabolism	3.94E-03
KEGG_MELANOGENESIS	Melanogenesis	4.68E-03
KEGG_PEROXISOME	Peroxisome	4.96E-03
KEGG_DNA_REPLICATION	DNA replication	5.33E-03
KEGG_PPAR_SIGNALING_PATHWAY	PPAR signaling pathway	6.28E-03
KEGG_MELANOMA	Melanoma	7.67E-03
KEGG_FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS	Fc gamma R-mediated phagocytosis	8.28E-03
KEGG_SMALL_CELL_LUNG_CANCER	Small cell lung cancer	8.34E-03
KEGG_BIOSYNTHESIS_OF_UNSATURATED_FATTY_ACIDS	Biosynthesis of unsaturated fatty acids	1.11E-02
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	Antigen processing and presentation	1.16E-02
KEGG_MISMATCH_REPAIR	Mismatch repair	1.32E-02
KEGG_LYSINE_DEGRADATION	Lysine degradation	1.48E-02
KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	Valine, leucine and isoleucine degradation	1.48E-02
KEGG_VASOPRESSIN_REGULATED_WATER_REABSORPTION	Vasopressin-regulated water reabsorption	1.48E-02
KEGG_PANCREATIC_CANCER	Pancreatic cancer	1.98E-02
KEGG_SELENOAMINO_ACID_METABOLISM	Selenoamino acid metabolism	2.13E-02
KEGG_ECM_RECEPTOR_INTERACTION	ECM-receptor interaction	2.18E-02
KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY_ARVC	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	2.71E-02
KEGG_SPHINGOLIPID_METABOLISM	Sphingolipid metabolism	2.90E-02
KEGG_AXON_GUIDANCE	Axon guidance	2.90E-02
KEGG_GLYCEROPHOSPHOLIPID_METABOLISM	Glycerophospholipid metabolism	3.30E-02
KEGG_MTOR_SIGNALING_PATHWAY	mTOR signaling pathway	3.30E-02
KEGG_CELL_ADHESION_MOLECULES_CAMS	Cell adhesion molecules (CAMs)	3.53E-02
KEGG_AMYOTROPHIC_LATERAL_SCLEROSIS_ALS	Amyotrophic lateral sclerosis (ALS)	3.55E-02
KEGG_ARGININE_AND_PROLINE_METABOLISM	Arginine and proline metabolism	3.81E-02
KEGG_NON_SMALL_CELL_LUNG_CANCER	Non-small cell lung cancer	3.81E-02
KEGG_GLYCINE_SERINE_AND_THREONINE_METABOLISM	Glycine, serine and threonine metabolism	3.88E-02
KEGG_BASAL_CELL_CARCINOMA	Basal cell carcinoma	4.09E-02
KEGG_EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLORI_INFECTION	Epithelial cell signaling in Helicobacter pylori infection	4.15E-02
KEGG_BETA_ALANINE_METABOLISM	beta-Alanine metabolism	4.68E-02
KEGG_LONG_TERM_DEPRESSION	Long-term depression	4.71E-02
KEGG_PROPANOATE_METABOLISM	Propanoate metabolism	4.71E-02

Bibliography

Bertram, L. and R. E. Tanzi (2008). Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses. *Nature Reviews Neuroscience 9*(10), 768.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. CRC press.

Gelman, A., J. Hill, and M. Yajima (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness 5*(2), 189–211.

Kate, Hawkins, and Tristan (2014). Cell signalling pathways underlying induced pluripotent stem cell reprogramming. *World Journal of Stem Cells 6*(5), 620.

Meng, D., A. R. Frank, and J. L. Jewell (2018). mTOR signaling in stem and progenitor cells. *Development 145*(1), dev152595.

Myers, R. H. (2004). Huntington's disease genetics. *NeuroRx 1*(2), 255–262.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics 18*(2), 349–367.

Rodriguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirich-

let process. *Journal of the American Statistical Association 103*(483), 1131–1154.

Shen-Orr, S. S., R. Tibshirani, P. Khatri, D. L. Bodian, F. Staedtler, N. M. Perry, T. Hastie, M. M. Sarwal, M. M. Davis, and A. J. Butte (2010). Cell type–specific gene expression differences in complex tissues. *Nature Methods 7*(4), 287–289.

Shulman, J. M., P. L. De Jager, and M. B. Feany (2011). Parkinson's disease: genetics and pathogenesis. *Annual Review of Pathology: Mechanisms of Disease 6*, 193–222.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 795–809.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences 102*(43), 15545–15550.

Tadesse, M. G., N. Sha, and M. Vannucci (2005). Bayesian variable selection

in clustering high-dimensional data. *Journal of the American Statistical Association 100*(470), 602–617.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*(476), 1566–1581.

Ye, D., G. Wang, Y. Liu, W. Huang, M. Wu, S. Zhu, W. Jia, A.-M. Deng, H. Liu, and J. Kang (2012). MiR-138 promotes induced pluripotent stem cell generation through the regulation of the p53 signaling. *Stem Cells 30*(8), 1645–1654.