Supplementary Materials for Dirichlet Process Mixture Model for Correcting Technical Variation in Single-Cell Gene Expression Data

A. Stochastic generative process

The generative process is as follows:

1. Compute the empirical mean $\boldsymbol{\mu}''$ and empirical covariance Σ'' of $X_{d\times n}$.
2. Sample the hyperparameters: $\boldsymbol{\mu}'$, Σ', H' and σ.
3. Sample parameters for each cluster $\boldsymbol{\mu}_k, \Sigma_k$, for $k = \{1, ..., K\}$.
4. Select one of the K clusters with probabilities $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_K\}$ where $\sum_{k=1}^{K} \pi_k = 1$.
5. Sample cell-specific parameters for scaling mean and covariance matrix as α_j, β_j respectively.
6. Sample an observation $\boldsymbol{x}_j$ from the probability distribution of the selected cluster by accounting for the scaling: $\boldsymbol{x}_j \sim \mathcal{N}(\alpha_j \boldsymbol{\mu_k}, \beta_j \Sigma_k)$.
7. Repeat steps 4 to 6, n times to sample n i.i.d. cells.

B. Algorithm

Algorithm 1 Inference Algorithm

for each Gibbs iteration t **do**
  $k =$ active clusters, $\zeta =$ auxiliary classes
  $\varphi^{-1} \sim Gamma(1, 1)$
  Update mixture component parameters μ_k and Σ_k based on z and hyperparameters:
  for every class k **do**
    $\Sigma_k'^{-1} = (\Sigma'^{-1} + n_k \Sigma_k^{-1})$
    $\mu_k' = \Sigma_k'^{-1}(\mu' \Sigma' + n_k \Sigma_k^{-1} \bar{x}_k)$
    $\Sigma_{\mu_k}^{-1} = \sum_{i \in (z_i = k)} (x_i - \mu_k)(x_i - \mu_k)^T$
    $\mu_k \sim \mathcal{N}(\mu_k', \Sigma_k')$
    $\Sigma_k^{-1} \sim Wish(\sigma' + n_k, (\sigma' H' + \Sigma_{\mu_k}^{-1})^{-1})$
  end for
  Update hyper parameters based on μ_k and Σ_k:
  $\mu' \sim \mathcal{N}((\Sigma''^{-1} + K\Sigma'^{-1})^{-1}(K\Sigma'^{-1}\bar{\mu} + \mu''\Sigma''), (\Sigma'' + K\Sigma'^{-1})^{-1})$
  $\Sigma_{\mu'}^{-1} = \sum_{i \in (1, \cdots, K)} (\mu_i - \mu')(\mu_i - \mu')^T$
  $\Sigma' \sim Wish(d + K, (d\Sigma''^{-1} + \Sigma_{\mu'}^{-1})^{-1})$
  $H' \sim Wish(d + K\sigma', d\Sigma'' + \sum_{k=1}^{K} \Sigma_k^{-1})$
  $\sigma' \sim InvGamma(1, \frac{1}{d})$
  for each cell i from 1 to n **do**
    if z_i is a singleton class **then**
      Make z_i an auxiliary class
    end if
    Construct $\boldsymbol{\rho}$
    $\forall$ k: $\boldsymbol{\rho}(k) = \frac{n_k - 1}{n - 1 + \varphi} * \mathcal{N}(\alpha_j \boldsymbol{\mu}_k, \beta_j \Sigma_k)$
    $\forall \zeta : \boldsymbol{\mu}_\zeta \sim \mathcal{N}(\mu', \Sigma'), \Sigma_\zeta^{-1} \sim Wish((\Sigma'' * d)^{-1}, d)$
    $\boldsymbol{\rho}(l) = \frac{\varphi / \zeta}{n - 1 + \varphi} * \mathcal{N}(\alpha_j \boldsymbol{\mu}_\zeta, \beta_j \Sigma_\zeta)$
    $\boldsymbol{\rho} = \frac{\boldsymbol{\rho}}{\sum \boldsymbol{\rho}}$
    $z_i \sim Mult(\boldsymbol{\rho})$
    Remove empty classes.
  end for
  for each cell j from 1 to n **do**
    $\alpha_j \sim \mathcal{N}(\nu^p, \delta^{p^2})$
    $\beta_j \sim InvGamma(\omega^p, \theta^p)$
    Impose identifiability constraints
  end for
end for

C. Derivation of posterior distributions

The conditional posterior distributions used in Gibbs sampling based on CRP have analytical forms and can be written as:

- For the component proportions π_k, using Bayes' Theorem and conjugacy of the Dirichlet prior.

$$\begin{aligned} f(\pi_k|\boldsymbol{z}, \{x\}_{\forall j}^{(1,\cdots,d)}, \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}, \varphi) &= f(\pi_k|\boldsymbol{z}, \boldsymbol{\varphi}) \\ &\propto f(\pi_k|\boldsymbol{\varphi}) f(\boldsymbol{z}|\boldsymbol{\pi}) \\ &\sim Dir(\tfrac{\varphi}{K} + \sum_{\forall i} \delta_1(z_i), \cdots, \tfrac{\varphi}{K} + \sum_{\forall i} \delta_K(z_j)) \end{aligned} \tag{4}$$

  where $\delta_i(z_j)$ is the Kronecker delta.

- For the latent class assignment variable, z_j, one integrates out $\boldsymbol{\pi}$ to get:

$$\begin{aligned} f(z_j|\boldsymbol{z}_{-j}, x_j, \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}, \alpha, \beta, \varphi) &= f(z_j|\boldsymbol{z}_{-j}, \boldsymbol{\pi}, \varphi) f(x_j|\boldsymbol{\mu}, \Sigma, \alpha, \beta) \\ f(z_j|\boldsymbol{z}_{-j}, \boldsymbol{\pi}, \varphi) &\propto \int_{\boldsymbol{\pi}} f(z_j|\boldsymbol{z}_{-j}, \boldsymbol{\pi}) f(\boldsymbol{\pi}|\varphi) d(\boldsymbol{\pi}) \\ &\propto f(z_j|\varphi) \coloneqq CRP(z_j|\varphi) \end{aligned}$$

  The $CRP(z_j|\varphi)$ can be written as:

$$\begin{aligned} f(z_j = k|\boldsymbol{z}_{-j}, \varphi) &\propto \frac{n_k - 1}{n - 1 + \varphi} \quad \text{for an existing class} \\ f(z_j = k|\boldsymbol{z}_{-j}, \varphi) &\propto \frac{\varphi/\zeta}{n - 1 + \varphi} \quad \text{for an auxiliary class} \\ \therefore f(z_j|\boldsymbol{z}_{-j}, x_j, \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}, \alpha, \beta, \varphi) &= CRP(z_j|\varphi) f(x_j|\boldsymbol{\mu}, \Sigma, \alpha, \beta) \\ &= \frac{n_k - 1}{n - 1 + \varphi} \mathcal{N}(x_j|\alpha_j \boldsymbol{\mu}_k, \beta_j \Sigma_k) \\ &\quad \text{(for an existing k)} \\ &= \frac{\varphi/\zeta}{n - 1 + \varphi} \mathcal{N}(x_j|\alpha_j \boldsymbol{\mu}_\zeta, \beta_j \Sigma_\zeta) \\ &\quad \text{(for an auxiliary class } \zeta) \end{aligned} \tag{5}$$

where $\boldsymbol{\mu}_\zeta$ and Σ_ζ are sampled from their base distributions; $\boldsymbol{\mu}_\zeta \sim \mathcal{N}(\mu', \Sigma')$ and $\Sigma_\zeta^{-1} \sim Wish(d, \frac{1}{d\Sigma''})$ (Görür & Rasmussen, 2010; Neal, 2000).

- For the mixing component parameters, sample $\boldsymbol{\mu}_k$ and Σ_k conditioned on $\boldsymbol{y}_k$ (Rasmussen, 1999; Neal, 2000):

$$
\begin{aligned}
f(\mu_k|\mu',\Sigma',\Sigma_k,x_j,\alpha_j,\beta_j,z_j=k) &\propto f(x_j|\mu_k,\Sigma_k,\alpha_j,\beta_j,z_j=k)f(\mu_k|\mu',\Sigma') \\
f(x_j|\mu_k,\Sigma_k,z_j=k,\alpha_j,\beta_j) &\sim \mathcal{N}(\alpha_j\mu_k,\beta_j\Sigma_k) \\
&\propto \frac{1}{2}(x_j-\alpha_j\mu_k)^T(\beta_j\Sigma_k)^{-1}(x_j-\alpha_j\mu_k) \\
&\propto (\mu_k-\frac{x_j}{\alpha_j})^T(\frac{\beta_j}{\alpha_j^2}\Sigma_k)^{-1}(\mu_k-\frac{x_j}{\alpha_j}) \\
f(\mu_k|\mu',\Sigma') &\sim \mathcal{N}(\mu',\Sigma') \\
&= \frac{1}{2}(\mu_k-\mu')^T\Sigma'^{-1}(\mu_k-\mu') \\
f(\mu_k|\mu',\Sigma',\Sigma_k,x_j,\alpha_j,\beta_j,z_j=k) &\sim \mathcal{N}(\mu_k^p,\Sigma_k^p) \\
\mu_k^p &= (\Sigma_k^{-1}\frac{\alpha_j^2}{\beta_j}+\Sigma'^{-1})^{-1}(\Sigma_k^{-1}\frac{\alpha_j x_j}{\beta_j}+\Sigma'^{-1}\mu') \\
\Sigma_k^p &= (\Sigma_k^{-1}\frac{\alpha_j^2}{\beta_j}+\Sigma'^{-1})^{-1}
\end{aligned}
\tag{6}
$$

However, we need to sum on all cells in the same cluster, hence:

$$
\begin{aligned}
f(\mu_k|\mu',\Sigma',\Sigma_k,x_k,z_j=k,\alpha_j,\beta_j) &\propto f(x_k|\mu_k,\Sigma_k,z_j=k,\alpha_j,\beta_j)f(\mu_k|\mu',\Sigma') \\
f(x_k|\mu_k,\Sigma_k,z_j=k,\alpha_j,\beta_j) &= \sum_{\forall j;z_j=k}\frac{1}{2}(\mu_k-\frac{x_j}{\alpha_j})^T(\frac{\beta_j}{\alpha_j^2}\Sigma_k)^{-1}(\mu_k-\frac{x_j}{\alpha_j}) \\
&\sim \mathcal{N}((\sum_j\{\frac{\alpha_j^2}{\beta_j}\Sigma_k^{-1})^{-1}\}\sum_j\{\frac{x_j}{\alpha_j^2}\frac{\alpha_j^2}{\beta_j}\Sigma_k^{-1}\},(\sum_j\{\frac{\alpha_j^2}{\beta_j}\Sigma_k^{-1}\})^{-1}) \\
&\sim \mathcal{N}((\sum_j\frac{\alpha_j^2}{\beta_j})^{-1}(\sum_j\frac{x_j}{\beta_j}),\Sigma_k(\sum_j\frac{\alpha_j^2}{\beta_j})^{-1}) \\
f(\mu_k|\mu',\Sigma') &= \frac{1}{2}(\mu_k-\mu')^T\Sigma'^{-1}(\mu_k-\mu') \\
f(\mu_k|\mu',\Sigma',\Sigma_k,x_k,\alpha_j,\beta_j,z_j=k) &\sim \mathcal{N}(\mu_k^p,\Sigma_k^p) \\
\mu_k^p &= \Sigma_k^p(\Sigma'^{-1}\mu'+\Sigma_k^{-1}(\sum_j\frac{x_j}{\beta_j})) \\
\Sigma_k^p &= (\Sigma'^{-1}+\Sigma_k^{-1}\sum_j\frac{\alpha_j^2}{\beta_j})^{-1}
\end{aligned}
\tag{7}
$$

Now for Σ_k, if x_k is the set of all x_js assigned to cluster k such that $z_j = k$:

$$
\begin{aligned}
f(\Sigma_k^{-1}|\sigma', H', \mu_k, x_k, \alpha_j, \beta_j, z_j = k) &\propto f(x_k|\Sigma_k, \mu_k, \alpha_j, \beta_j, z_j = k) f(\Sigma_k^{-1}|\sigma', H') \\
f(x_k|\Sigma_k^{-1}, \mu_k, \alpha_j, \beta_j, z_j = k) &\sim \mathcal{N}(\alpha_j \mu_k, \beta_j \Sigma_k) \\
&\propto \prod_{\forall j; z_j = k} |\beta_j \Sigma_k|^{-1/2} exp\{-\frac{1}{2}(x_j - \alpha_j \mu_k)^T (\beta_j \Sigma_k)^{-1} (x_j - \alpha_j \mu_k)\} \\
f(\Sigma_k^{-1}|\sigma', H') &\sim Wish({H'}^{-1}, \sigma') \\
&= |\Sigma_k^{-1}|^{\frac{\sigma' - d - 1}{2}} exp\{-\frac{tr(H'\Sigma_k^{-1})}{2}\} / (2^{\sigma' d/2} |H'|^{-\sigma'/2} \Gamma_d(\sigma'/2)) \\
f(\Sigma_k^{-1}|\sigma', H', \mu_k, x_j, \alpha_j, \beta_j, z_j = k) &\propto |\Sigma_k|^{\frac{1 - \sigma' + d}{2}} exp\{-\frac{tr(H'\Sigma_k^{-1})}{2}\} / (2^{\sigma' d/2} |H'|^{-\sigma'/2} \Gamma_d(\sigma'/2)) \times \\
&\quad \prod_{\forall j; z_j = k} |\beta_j \Sigma_k|^{-1/2} exp\{-\frac{1}{2}(x_j - \alpha_j \mu_k)^T (\beta_j \Sigma_k)^{-1} (x_j - \alpha_j \mu_k)\} \\
&= |\Sigma_k|^{\frac{1 - \sigma' + d}{2}} (\prod_{\forall j; z_j = k} |\Sigma_k|^{-1/2}) (\prod_j \beta_j^{-d/2}) / (2^{\sigma' d/2} |H'|^{-\sigma'/2} \Gamma_d(\sigma'/2)) \times \\
&\quad exp\{-\frac{tr(H'\Sigma_k^{-1})}{2} - \sum_j \frac{1}{2}(x_j - \alpha_j \mu_k)^T (\beta_j \Sigma_k)^{-1} (x_j - \alpha_j \mu_k)\} \\
&= |\Sigma_k|^{(d - \sigma' - n_k + 1)/2} (\prod_j \beta_j^{-d/2}) / (2^{\sigma' d/2} |H'|^{-\sigma'/2} \Gamma_d(\sigma'/2)) \times \\
&\quad exp\{-\frac{tr(H'\Sigma_k^{-1})}{2} - \frac{1}{2} \sum_j tr((\beta_j \Sigma_k)^{-1} (x_j - \alpha_j \mu_k)(x_j - \alpha_j \mu_k)^T)\} \\
&= |\Sigma_k|^{(d - \sigma' - n_k + 1)/2} (\prod_j \beta_j^{-d/2}) / (2^{\sigma' d/2} |H'|^{-\sigma'/2} \Gamma_d(\sigma'/2)) \times \\
&\quad exp\Big(-\frac{tr(H'\Sigma_k^{-1})}{2} - \frac{1}{2} tr\Big(\Sigma_k^{-1} \sum_j ((x_j - \alpha_j \mu_k)(x_j - \alpha_j \mu_k)^T)/\beta_j\Big)\Big) \\
&= |\Sigma_k^{-1}|^{(\sigma' - d + n_k - 1)/2} exp\{-\frac{1}{2} tr\Big((H' + S_x)\Sigma_k^{-1}\Big)\} \times \\
&\quad (\prod_j \beta_j^{-d/2}) / (2^{\sigma' d/2} |H'|^{-\sigma'/2} \Gamma_d(\sigma'/2)) \\
S_x &= \sum_j ((x_j - \alpha_j \mu_k)(x_j - \alpha_j \mu_k)^T)/\beta_j) \\
f(\Sigma_k^{-1}|\sigma', H', \mu_k, x_j, \alpha_j, \beta_j, z_j = k) &\sim Wish(H, \sigma) \\
H &= (H' + S_x)^{-1} \\
\sigma &= \sigma' + n_k
\end{aligned}
\tag{8}
$$

- For scaling parameters α_j, if we assume Normal distributions with positive mean and narrow variance (as approxima-

tion to the right skewed distribution of positive values), we can also derive posteriors:

$$
\begin{aligned}
f(\alpha_j|\nu,\delta,\mu_k,\Sigma_k,x_j,\beta_j,z_j=k) &\propto f(x_j|\alpha_j,\beta_j,\mu_k,\Sigma_k,z_j=k)f(\alpha_j|\nu,\delta)\\
f(x_j|\alpha_j,\beta_j,\mu_k,\Sigma_k,z_j=k) &\sim \mathcal{N}(\alpha_j\mu_k,\beta_j\Sigma_k)\\
&\propto \frac{1}{2}(x_j-\alpha_j\mu_k)^T(\beta_j\Sigma_k)^{-1}(x_j-\alpha_j\mu_k)\\
&= \frac{1}{2}(x_j-\alpha_j\mu_k)^T(\beta_j\Sigma_k)^{-T/2}(\beta_j\Sigma_k)^{-1/2}(x_j-\alpha_j\mu_k)\\
&= \frac{1}{2}(Ax_j-A\mu_k\alpha_j)^T I(Ax_j-A\mu_k\alpha_j)\\
&= \frac{1}{2}\sum_{i=1}^{d}\{((Ax_j)^i-(A\mu_k)^i\alpha_j)((Ax_j)^i-(A\mu_k)^i\alpha_j)\}\\
&= \frac{1}{2}\sum_{i=1}^{d}\{((Ax_j)^i./(A\mu_k)^i-\alpha_j)(A\mu_k)^{i^2}((Ax_j)^i./(A\mu_k)^2-\alpha_j)\} \qquad (9)\\
\alpha_j &\sim \mathcal{N}(\nu^x,\delta^{x2})\\
\nu^x &= \delta^{x2}\sum_i\{(Ax_j)^i(A\mu_k)^i\}\\
\delta^{x2} &= (\sum_i(A\mu_k)^i)^{-1}\\
f(\alpha_j|\nu,\delta) &\sim \mathcal{N}(\nu,\delta^2)\\
f(\alpha_j|\nu,\delta,\mu_k,\Sigma_k,x_j,\beta_j,z_j=k) &\sim \mathcal{N}(\nu^p,\delta^{p^2})\\
\nu^p &= \delta^{p2}(\nu^x/\delta^{x2}+\nu/\delta^2)\\
\delta^{p2} &= (1/\delta^{x2}+1/\delta^2)^{-1}
\end{aligned}
$$

where $A=(\beta_j\Sigma_k)^{-1/2}$, and $(\cdot)^i$ denotes the ith elements, and $./$ is element-wise division.
For β_j we assume an Inverse-gamma distribution.

$$
\begin{aligned}
f(\beta_j|\omega,\theta,\mu_k,\Sigma_k,x_j,\alpha_j,z_j=k) &\propto f(x_j|\alpha_j,\beta_j,\mu_k,\Sigma_k,z_j=k)f(\beta_j|\omega,\theta)\\
f(x_j|\alpha_j,\beta_j,\mu_k,\Sigma_k,z_j=k) &\sim \mathcal{N}(\alpha_j\mu_k,\beta_j\Sigma_k)\\
&\propto |\beta_j\Sigma_k|^{-1/2}exp\{-\frac{1}{2}(x_j-\alpha_j\mu_k)^T(\beta_j\Sigma_k)^{-1}(x_j-\alpha_j\mu_k)\}\\
&\propto |\beta_j\Sigma_k|^{-1/2}exp\{-\frac{1}{2\beta_j}(x_j-\alpha_j\mu_k)^T\Sigma_k^{-1}(x_j-\alpha_j\mu_k)\}\\
f(\beta_j|\omega,\theta) &\sim InvGamma(\omega,\theta)\\
f(\beta_j|\omega,\theta) &= \frac{\theta^\omega}{\Gamma(\omega)}\beta_j^{-\omega-1}exp(\frac{-\theta}{\beta_j})\\
f(\beta_j|\omega,\theta,\mu_k,\Sigma_k,x_j,\alpha_j,z_j=k) &\propto \frac{\theta^\omega}{\Gamma(\omega)}\beta_j^{-\omega-1-d/2}|\Sigma_k|^{-1/2}exp\{\frac{-1}{\beta_j}(\theta+\frac{1}{2}(x_j-\alpha_j\mu_k)^T\Sigma_k^{-1}(x_j-\alpha_j\mu_k))\}\\
&\sim InvGamma(\omega^p,\theta^p)\\
\omega^p &= \omega+d/2\\
\theta^p &= \theta+\frac{1}{2}(x_j-\alpha_j\mu_k)^T\Sigma_k^{-1}(x_j-\alpha_j\mu_k)
\end{aligned}
\qquad (10)
$$

- For the hyperparameters viz. μ', Σ' and H', sample them conditioned on the new component parameters $\boldsymbol{\mu}_k$ and Σ_k

(Rasmussen, 1999),(Neal, 2000):

$$
\begin{aligned}
f(\mu'|\mu'', \Sigma'', \{\mu_k\}_{k=1}^K, \{x_j\}_{j=1}^n) &\propto \Big(\prod_{k=1}^K f(\mu_k|\mu', \Sigma')\Big) f(\mu'|\mu'', \Sigma'') \\
f(\mu'|\mu'', \Sigma'') &\sim \mathcal{N}(\mu'|\mu'', \Sigma'') \\
&\propto |\Sigma''|^{-\frac{1}{2}} exp(-\frac{1}{2}(\mu' - \mu'')^T \Sigma''^{-1}(\mu' - \mu'')) \\
\Big(\prod_{k=1}^K f(\mu_k|\mu', \Sigma')\Big) &\sim (\text{unnormalised})\, \mathcal{N}(f(\mu_k|\mu', \Sigma')) \\
&\propto \prod_{k=1}^K |\Sigma'|^{-\frac{1}{2}} exp(-\frac{1}{2}(\mu_k - \mu')^T \Sigma'^{-1}(\mu_k - \mu')) \\
\therefore f(\mu'|\mu'', \Sigma'', \{\mu_k\}_{k=1}^K, \{x_j\}_{j=1}^n) &\propto |\Sigma''|^{-\frac{1}{2}} exp(-\frac{1}{2}(\mu' - \mu'')^T \Sigma''^{-1}(\mu' - \mu'')) \times \\
\prod_{k=1}^K |\Sigma'|^{-\frac{1}{2}} exp(-\frac{1}{2}(\mu_k - \mu')^T \Sigma'^{-1}(\mu_k - \mu')) & \\
&\propto |\Sigma''|^{-\frac{1}{2}} exp(-\frac{1}{2}(\mu' - \mu'')^T \Sigma''^{-1}(\mu' - \mu'')) \times \\
|\Sigma'|^{-\frac{K}{2}} exp(\sum_{k=1}^K -\frac{1}{2}(\mu_k - \mu')^T \Sigma'^{-1}(\mu_k - \mu')) & \\
&\propto |\Sigma''|^{-\frac{1}{2}} exp(-\frac{1}{2}(\mu' - \mu'')^T \Sigma''^{-1}(\mu' - \mu'')) \times \\
|K\Sigma'|^{-\frac{1}{2}} exp(\sum_{k=1}^K -\frac{1}{2}(\mu_k - \mu')^T \Sigma'^{-1}(\mu_k - \mu')) & \\
&\sim \mathcal{N}(\mu_{\mu'}, \Sigma_{\mu'}) \\
\mu_{\mu'} &= \Sigma_{\mu'}(\Sigma''^{-1}\mu'' + K^2\Sigma'^{-1}\bar{\mu'}) \\
\Sigma_{\mu'} &= (\Sigma''^{-1} + K\Sigma'^{-1})^{-1}
\end{aligned}
\tag{11}
$$

where K is the number of currently populated clusters, d the data dimensionality, $\bar{\mu'}$ is the mean over μ_ks and μ'' and

Σ'' are the empirical mean and covariance of the data.

$$
\begin{aligned}
f(\Sigma'^{-1}|\mu', \Sigma''^{-1}, \{\mu_k\}_{k=1}^K, \{x_j\}_{j=1}^n) &\propto \Big(\prod_{k=1}^K f(\mu_k|\mu', \Sigma')\Big) f(\Sigma'^{-1}|\Sigma''^{-1}) \\
f(\Sigma'^{-1}|\Sigma''^{-1}) &\sim \mathit{Wish}(\Sigma'^{-1}|\frac{\Sigma''^{-1}}{d}, d) \\
&\propto |\Sigma'^{-1}|^{\frac{d-d-1}{2}} \exp\big(-\frac{tr(d\Sigma''\Sigma'^{-1})}{2}\big) \\
\Big(\prod_{k=1}^K f(\mu_k|\mu', \Sigma')\Big) &\sim (\text{unnormalised})\, \mathcal{N}(f(\mu_k|\mu', \Sigma')) \\
&\propto \prod_{k=1}^K |\Sigma'|^{-\frac{1}{2}} exp\big(-\frac{1}{2}(\mu_k-\mu')^T\Sigma'^{-1}(\mu_k-\mu')\big) \\
&\propto |\Sigma'|^{-\frac{K}{2}} exp\big(\sum_{k=1}^K -\frac{1}{2}(\mu_k-\mu')^T\Sigma'^{-1}(\mu_k-\mu')\big) \\
\therefore f(\Sigma'^{-1}|\mu', \Sigma''^{-1}, \{\mu_k\}_{k=1}^K, \{x_j\}_{j=1}^n) &\propto |\Sigma'^{-1}|^{\frac{d-d-1}{2}} \exp\big(-\frac{tr(d\Sigma''\Sigma'^{-1})}{2}\big)\times \\
|\Sigma'^{-1}|^{\frac{K}{2}} exp\big(\sum_{k=1}^K -\frac{1}{2}(\mu_k-\mu')^T\Sigma'^{-1}(\mu_k-\mu')\big) & \\
&\propto |\Sigma'^{-1}|^{\frac{d-d-1+K}{2}} \exp\big(-\frac{tr(d\Sigma''\Sigma'^{-1})}{2}\big)\times \\
exp\big(-\Sigma'^{-1}\sum_{k=1}^K \frac{1}{2}(\mu_k-\mu')^T(\mu_k-\mu')\big) & \\
\text{Substituting } \Sigma_{rss} &= \sum_{k=1}^K \frac{1}{2}(\mu_k-\mu')^T(\mu_k-\mu') \text{ above, we get:} \\
f(\Sigma'^{-1}|\mu', \Sigma''^{-1}, \{\mu_k\}_{k=1}^K, \{x_j\}_{j=1}^n) &\propto |\Sigma'^{-1}|^{\frac{d-d-1+K}{2}} \\
\exp\big(-\frac{tr(d\Sigma''\Sigma'^{-1})}{2}\big) \times \exp\big(-\Sigma'^{-1}\Sigma_{rss}\big) & \\
&\propto |\Sigma'^{-1}|^{\frac{d-d-1+K}{2}} \exp\big(-\frac{tr(d\Sigma''\Sigma'^{-1})}{2} - \frac{2tr(\Sigma'^{-1}\Sigma_{rss})}{2}\big) \\
&\propto |\Sigma'^{-1}|^{\frac{d-d-1+K}{2}} \exp\big(-\frac{tr(d\Sigma''\Sigma'^{-1} + 2\Sigma'^{-1}\Sigma_{rss})}{2}\big) \\
&\propto |\Sigma'^{-1}|^{\frac{d-d-1+K}{2}} \exp\Big(-\frac{tr(\Sigma'^{-1}(d\Sigma'' + 2\Sigma_{rss}))}{2}\Big) \\
&\sim \mathcal{W}(V_{\Sigma'^{-1}}, d_{\Sigma'^{-1}}) \\
V_{\Sigma'^{-1}} &= (d\Sigma'' + 2\Sigma_{rss})^{-1} \\
d_{\Sigma'^{-1}} &= d + K
\end{aligned}
\tag{12}
$$

where K is the number of currently populated clusters, d the data dimensionality and μ'' and Σ'' are the empirical

mean and covariance of the data.

$$
\begin{aligned}
f(H'|\sigma', \Sigma'', \{\Sigma_k\}_{k=1}^K, \{x_j\}_{j=1}^n) &\propto \Big(\prod_{k=1}^K f(\Sigma_k^{-1}|H'^{-1}, \sigma')\Big) f(H'|\Sigma'') \\
f(H'|\Sigma'') &\sim Wish(H'|\frac{\Sigma''}{d}, d) \\
&\propto |H'|^{\frac{d-n-1}{2}} |d\Sigma''^{-1}|^{-\frac{d}{2}} \exp\big(-\frac{tr(d\Sigma''^{-1}H')}{2}\big) \\
\Big(\prod_{k=1}^K f(\Sigma_k^{-1}|H'^{-1}, \sigma')\Big) &\sim Wish(\Sigma_k^{-1}|H'^{-1}, \sigma') \\
&\propto \prod_{k=1}^K \Big(|\Sigma_k'^{-1}|^{\frac{\sigma'-d-1}{2}} |H'|^{\frac{\sigma'}{2}} \exp\big(-\frac{tr(H'\Sigma_k^{-1})}{2}\big)\Big) \\
&\propto |\Sigma_k'^{-1}|^{\frac{K(\sigma'-d-1)}{2}} |H'|^{\frac{K\sigma'}{2}} \exp\Big(\sum_{k=1}^K \big(-\frac{tr(H'\Sigma_k^{-1})}{2}\big)\Big) \\
&\propto |\Sigma_k'^{-1}|^{\frac{K(\sigma'-d-1)}{2}} |H'|^{\frac{K\sigma'}{2}} \exp\Big(-\frac{tr\sum_{k=1}^K (H'\Sigma_k^{-1})}{2}\Big) \\
&\propto |\Sigma_k'^{-1}|^{\frac{K(\sigma'-d-1)}{2}} |H'|^{\frac{K\sigma'}{2}} \exp\Big(-\frac{trH'\sum_{k=1}^K (\Sigma_k^{-1})}{2}\Big) \\
\therefore f(H'|\sigma', \Sigma'', \{\Sigma_k\}_{k=1}^K, \{x_j\}_{j=1}^n) &\propto |H'|^{\frac{d-n-1}{2}} |d\Sigma''^{-1}|^{-\frac{d}{2}} \exp\big(-\frac{tr(d\Sigma''^{-1}H')}{2}\big) \times \\
|\Sigma_k'^{-1}|^{\frac{K(\sigma'-d-1)}{2}} |H'|^{\frac{K\sigma'}{2}} \exp\Big(-\frac{trH'\sum_{k=1}^K (\Sigma_k^{-1})}{2}\Big) & \\
&\propto |H'|^{\frac{d-n-1+K\sigma'}{2}} \\
\exp\Big(-\frac{tr(H'(d\Sigma''^{-1} + \sum_{k=1}^K (\Sigma_k^{-1})))}{2}\Big) & \\
&\sim Wish(V_{H'}, d_{H'}) \\
V_{H'} &= (d\Sigma''^{-1} + \sum_{k=1}^K (\Sigma_k^{-1}))^{-1} \\
d_{H'} &= d + K\sigma'
\end{aligned}
\tag{13}
$$

where K is the number of currently populated clusters, d the data dimensionality and μ'' and Σ'' are the empirical mean and covariance of the data.

D. Theorems

D.1. Model Identifiability

As we intend to learn interpretable and consistent structures (rather than building a solely predictive model), we need to insure model identifiability. Specifically, we need to set constraints on parameters $\alpha_j, \beta_j, \boldsymbol{\mu}_k$ such that the parameter estimates are valid.

Lemma S1. *A finite mixture of multivariate Gaussian distributions $f(X|\boldsymbol{m}_k, S_k)$ with means $\boldsymbol{m}_k$ and covariance S_k for component k, is identifiable with permutations in components, i.e. $\Sigma_{k=1}^K \pi_k f(X|\boldsymbol{m}_k, S_k) = \Sigma_{l=1}^{K^\star} \pi_l^\star f(X|\boldsymbol{m}_l^\star, S_l^\star)$ implies that $K = K^\star$ and mixtures are equivalent with permutations in components.*

Proof. Follows results from Yakowitz & Spragins (1968) and Titterington (1985). □

Suppose we define the parameter set $\Theta = \{\forall j, k : (\alpha_j \boldsymbol{\mu}_k, \beta_j \Sigma_k)\} \cup \{\boldsymbol{\pi}\}$, using Lemma 1, we have the identifiability of $f(X|\Theta)$ with $\boldsymbol{m}_k = \alpha_j \boldsymbol{\mu}_k$ and $S_k = \beta_j \Sigma_k$ with permutation in the components. We would like to extend this result to identifiability of $f(X|\Phi)$ where $\Phi = \{\forall j, k : (\alpha_j, \boldsymbol{\mu}_k, \beta_j, \Sigma_k)\} \cup \{\boldsymbol{\pi}\}$.

Theorem S2. *Suppose that* $\Theta = \Theta^\star$ *and for the prior distributions we have* $\forall j, k : f(\alpha_j, \boldsymbol{\mu}_k, \beta_j, \Sigma_k) = f(\alpha_j^\star, \boldsymbol{\mu}_k{}^\star, \beta_j^\star, \Sigma_k^\star)$. *If the following condition holds we have* $\Phi = \Phi^\star$.

- $\forall j : \boldsymbol{\mu}_k \geq \boldsymbol{\mu}' + diag(\Sigma')(\alpha_j - \nu)/\delta$
- $\forall j : \beta_j \geq \frac{\theta}{\omega+1}$

Proof. We present the proof sketch. According to the result from Lemma 1, $\alpha_j \boldsymbol{\mu}_k$ and $\beta_j \Sigma_k$ are identified. Given the priors for $\alpha_j \sim \mathcal{N}(\nu, \delta^2)$ and $\boldsymbol{\mu}_k \sim \mathcal{N}(\mu', \Sigma')$, and the identifiability of $\alpha_j \boldsymbol{\mu}_k$, to insure one solution set for the intersection of the Normal probability distributions and $\alpha_j \boldsymbol{\mu}_k = const.$, we limit the solution space such that $\forall j : \boldsymbol{\mu}_k - \boldsymbol{\mu}' \geq diag(\Sigma')(\alpha_j - \nu)/\delta$. Similarly, given the identifiability of $\beta_j \Sigma_k$, we confine the space of β_js such that they are always larger than the mode of its prior: $\beta_j \geq \theta/(\omega + 1)$. These conditions impose truncated priors instead of the full distribution for $\boldsymbol{\mu}_k, \boldsymbol{\beta}$. Given the structure of the prior $f(\alpha_j, \boldsymbol{\mu}_k, \beta_j, \Sigma_k)$ and above conditions, any change in one of the parameters (e.g. α_j) with respect to their mean leads to either increasing the probabilities given priors for all parameters or decreasing the probabilities given priors for all parameters. For example, an increase in α_j will have the following effect given the conditions: $\alpha_j \uparrow \Rightarrow \boldsymbol{\mu}_k \uparrow$ which given the form of priors leads to a decrease in both $f(\alpha_j)$, and $f(\boldsymbol{\mu}_k)$ and hence a decrease in $f(\alpha_j, \boldsymbol{\mu}_k)$. Therefore, in order to guarantee $\forall j, k : f(\alpha_j, \boldsymbol{\mu}_k, \beta_j, \Sigma_k) = f(\alpha_j^\star, \boldsymbol{\mu}_k{}^\star, \beta_j^\star, \Sigma_k^\star)$ we need to have $\Phi = \Phi^\star$.

□

D.2. Weak Posterior Consistency

Let $f_0(\boldsymbol{x}) := \mathcal{N}(\alpha\boldsymbol{\mu}, \beta\Sigma) \in \mathbb{R}^d$ be the *true* Gaussian density of $\boldsymbol{x}$ with identifiability constraints imposed on $\boldsymbol{\mu}, \alpha, \beta$ as given in Theorem S2. Let P be the mixing distribution and $\mathcal{F}$ be the space of all density functions in $\mathbb{R}^d$ with respect to the Lebesque measure. Let Π be the prior over $\mathcal{F}$ induced by BISCUIT i.e. $\mathcal{F} \sim \Pi$. Each $\boldsymbol{x}_j \sim \mathcal{N}(\alpha_j \boldsymbol{\mu}_k, \beta_j \Sigma_k)$, $(\boldsymbol{\mu}_k, \Sigma_k) \sim \mathcal{G}, \mathcal{G} \sim DP(\varphi, \mathcal{G}_0)$ and $\alpha_j, \beta_j \in \mathbb{R}$. The base distribution $\mathcal{G}_0$ is the prior density over the distribution of the model parameters. Weak consistency of a distribution relates to how close the posterior distribution, $\Pi(f \in \mathcal{F}|\boldsymbol{x}_{i=1}^n)$ concentrates around arbitrarily small neighborhoods of $f_0(\boldsymbol{x})$ as $n \to \infty$. For a given radius ϵ, the KL-neighborhood $KL_\epsilon(f_0) := \{f \in \mathcal{F} : KL(f_0, f) < \epsilon\}$.

Theorem S3. *If* $f_0(\boldsymbol{x})$ *is compactly supported and* $\mathcal{G}_0$ *has support* $\mathrm{R}^d \times \mathrm{R}_+^{d \times d}$, *then for weak consistency we show that* $f_0(\boldsymbol{x}) \in KL(\Pi)$ *every* $\epsilon > 0$.

Proof outline: The proof closely follows Wu & Ghosal (2010). We restate Theorem S3 in terms of Schwartz's theorem (Schwartz, 1965) based on the Kullback-Leibler (KL) property. It states that weak consistency conditions are equivalent to showing that $f_0 \in KL(\Pi)$ i.e. for every $\epsilon > 0, \Pi(KL_\epsilon(f_0)) > 0$ where the KL-neighborhood, $KL_\epsilon(f_0) := \{f \in \mathcal{F} : KL(f_0, f) < \epsilon\} = \{f \in \mathcal{F} : \int f_0(\boldsymbol{x}) log \frac{f_0(\boldsymbol{x})}{f(\boldsymbol{x})} d\boldsymbol{x} < \epsilon\}$. In other words, if Π puts positive probability on all $KL_\epsilon(f_0)$ for every $\epsilon > 0$, then Π is weakly consistent at f_0.

Let $\phi_d(\boldsymbol{x}, \Sigma)$ be the multivariate Gaussian density for $\boldsymbol{x}$. Following the assumptions in Theorem 2 in (Wu & Ghosal, 2010), we bound Σ to $\beta h I_d$ where I_d is the identity matrix of order d, $h \in H$ and $H \subset \mathrm{R}^+$ to show:

$$\begin{aligned}
\int f_0(\boldsymbol{x}) log \frac{f_0(\boldsymbol{x})}{f(\boldsymbol{x})} &= \int f_0(\boldsymbol{x}) log \frac{f_0(\boldsymbol{x})}{\int \phi_d(\boldsymbol{x} - \alpha\theta, \Sigma) dP(\theta)} d\boldsymbol{x} \\
&\leq \int f_0(\boldsymbol{x}) log \frac{f_0(\boldsymbol{x})}{\int \phi_d(\boldsymbol{x} - \alpha\theta, \beta h I_d) dP(\theta)} d\boldsymbol{x} + constant \\
&\leq \epsilon
\end{aligned}$$

Thus for all weak neighborhoods $KL_\epsilon(f_0) < \epsilon$, meaning Π puts positive probability on all weak neighborhoods of f_0. This concludes Theorem S3.

Proposition S4. *One sweep of the Gibbs sampler in BISCUIT can be computed in* $O(n)$ *time.*

Proof. For every Gibbs sweep, there are principally four blocks viz. block 1 that computes the mixture component parameters for k components, block 2 that updates the hyperparameters, block 3 that assigns the class membership probability to each of the n cells and block 4 that samples the cell-specific α and β. Blocks 1 and 2 run in $O(1)$ time. The run time for Block 3 can be computed as follows: We need to assign a class to the j^{th} cell, given the assignments of the remaining objects and all other parameter values. The j^{th} cell's assignment probabilities for all existing clusters and auxiliary clusters are calculated based on equations in Section 5. From the resulting categorical distribution we sample a new assignment,say z_{new}. This procedure repeats for all cells $j = 1, \cdots, n$. The computational cost of sampling the class assignment z_{new} is proportional to the number of classes, $k \therefore$ as $n \to \infty$, $k \to \varphi log(n)$ almost surely (Korwar & Hollander, 1973; Antoniak, 1974). The time complexity for block 4 to update the scale parameters α_j, β_j is also constant and done for n observations. Therefore the per-sweep time complexity for clustering n observations is $O(nk+n) = O(\varphi nlog(n)+n) \approx O(\varphi nlog(n))$. Given that $n \gg k$ most of the time, the time complexity approaches $O(n)$.

□

E. Comparison experiments.

We compared the performance of BISCUIT with a number of alternative methods including the naive HDPMM (Görür & Rasmussen, 2010) along with two normalization methods typically used for single-cell data. a) a Generalized Linear Model-based normalization (GLMnorm) where counts are regressed against the library size to get a residual count matrix and b) a Mean-normalized method (MeanNorm) each cell is scaled by the average library size. Both the residual and mean normalized matrices are log-transformed and used as input to the naive HDPMM. Additionally we compare with non-MCMC methods such as Spectral clustering (Ng et al., 2002) and Phenograph (Levine et al., 2015).

We use a confusion matrix C to assess the quality of inferred clusters. For the MCMC methods, zs are taken from Gibbs samples after a certain burn-in period. For graph-based methods, a series of z estimates are created by varying the nearest neighbor parameter for Phenograph and by varying the number of clusters in Spectral clustering. The confusion matrices for the different MCMC methods are shown in Figure 5. Next the upper triangular matrix of C^{true} (left-most in Figure 5) is used to create a binary vector $\mathbf{h}$ encoding the 'true' z of each cell which is then compared with $\mathbf{h}^*_{met}$ that contains inferred zs with *met* referring to BISCUIT, HDPMM, GLMnorm, MeanNorm, Phenograph and Spectral clustering. When $C^{true}_{(i,j)} = 0$ and $C^{met}_{(i,j)} = 0$, $\mathbf{h}^*_{met}$ is set to the number of valid iterations. When $C^{true}_{(i,j)} = 1$ and $C^{met}_{(i,j)} \geq 1$, $\mathbf{h}^*_{met}$ is assigned the value in $C^{met}_{(i,j)}$. The agreement of $\mathbf{h}$ and $\mathbf{h}^*_{met}$ is measured using the F-measure. The top panel in Figure S2 shows boxplots of F-scores obtained in 15 experiments with randomly generated X for the different methods. The bottom panel shows the outcome of a Friedman test with post-hoc analysis for assessing the significance of the pairwise differences (Hollander & Wolfe, 1999). The better performance of BISCUIT is due to its ability to account for cell-specific scalings.

Model convergence diagnostics. Figure S4 depicts the trace of number of active clusters during Gibbs sampling in BISCUIT for $X_{50\times100}$ with 3 clusters. The sampler stabilizes after roughly 150 sweeps and when initialized with one cluster, the traceplot (Figure S4) shows an 'almost monotone' increase during burn-in. Table S2 compares runtimes and memory usage between the methods.

F. Supplementary Figures & Tables

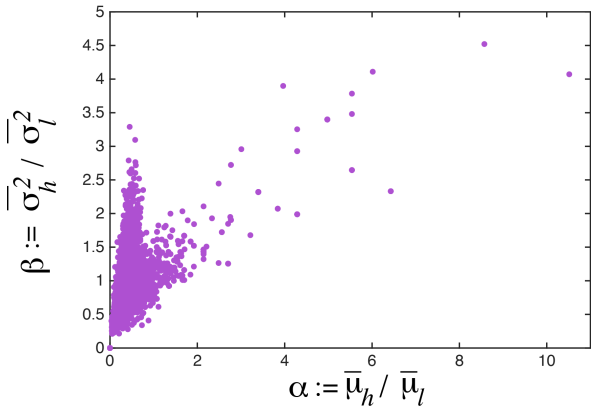

Figure S1: No clear dependence between the ratios of variances vs ratios of means motivates modeling moment-scalings as separate cell-specific parameters.

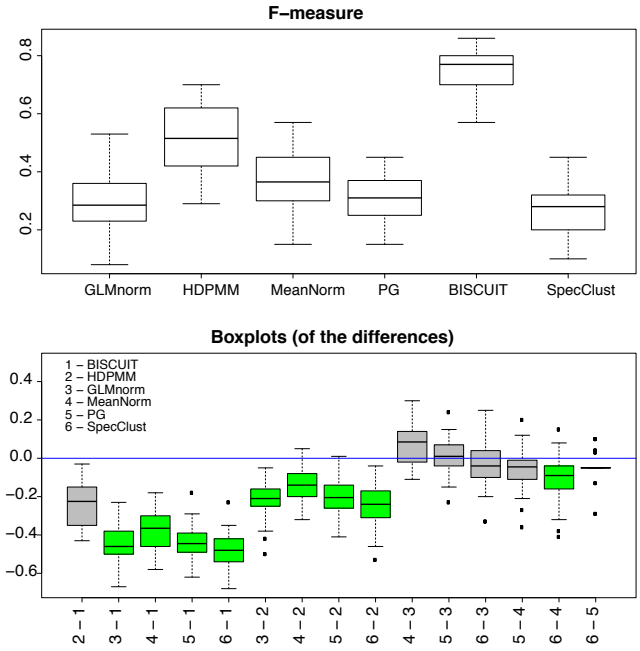

Figure S2: **Top**: Boxplots of F-scores obtained in 15 experiments with randomly-generated X for various methods. **Bottom**: Boxplot of pairwise differences with color-coded significance (green, if multiple-testing-corrected $p < 0.05$) computed by a Friedman test with post-hoc analysis. (Wilcoxon-Nemenyi-McDonald-Thompson test (Hollander & Wolfe, 1999)).

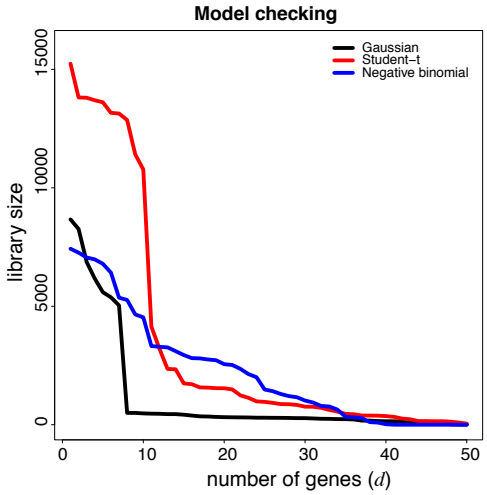

Figure S3: Generative distributions used to check model mismatch assumptions. The hypothesized multivariate Gaussian versus two heavy-tailed distributions: a non-central Student's t and a negative binomial.

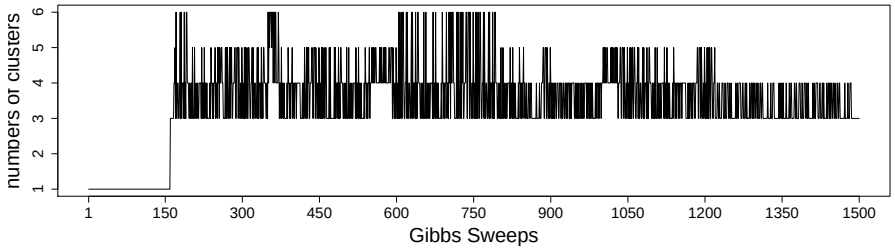

Figure S4: Traceplot of the number of active clusters.

	BISCUIT	HDPMM	GLMnorm	MeanNorm	PG	SpecClust
$\sim$ Runtime (min)	1.53 ($\pm$) 0.15	2.10 ($\pm$) 0.02	2.16 ($\pm$) 0.04	2.16 ($\pm$) 0.12	0.01 ($\pm$) 0.03	0.06 ($\pm$) 0.01
$\sim$ Memory (MB)	23.4	26.2	30.3	31.1	7.6	10.3

Table S1: For a randomly-generated $X_{50\times100}$, appproximate runtime comparisons and memory usages for different methods. All simulations are carried out on a computational cluster with a single core of 2.30 GHz processor and 32 GB memory.

	BISCUIT	HDPMM	PG	SpecClust	DBScan	BASiCS+HDPMM	BASiCS+PG	BASiCS+SpecClust
F-score	0.9127	0.7913	0.7417	0.3205	0.2486	0.7125	0.6173	0.1425

Table S2: F-scores for BISCUIT versus other competing clustering techniques for 3005 cells in the Zeisel et al. (2015) dataset.

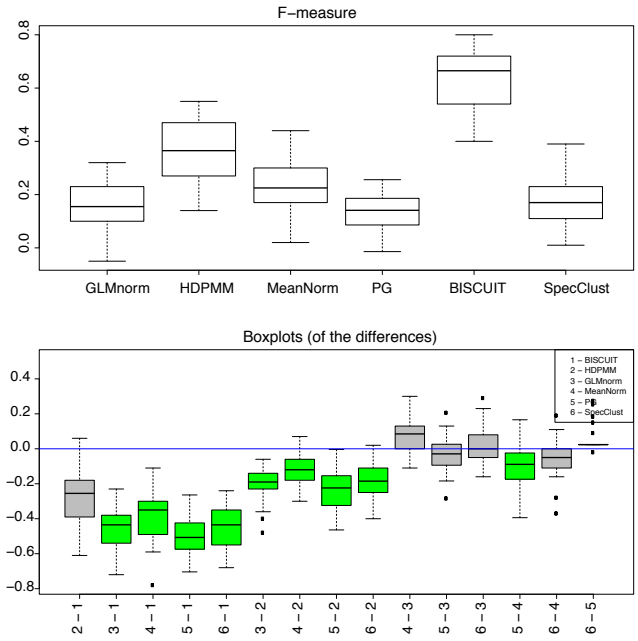

Figure S5: **Top**: Boxplots of F-scores obtained in 15 experiments with randomly-generated X from a negative binomial distribution for BISCUIT, HDPMM, GLMnorm, MeanNorm, Phenograph (PG) and Spectral Clustering (SpecClust). **Bottom**: Boxplot of pairwise differences together with color-coded significance (green, if multiple-testing-corrected $p < 0.05$) computed by a non-parametric Friedman test with post-hoc analysis (Wilcoxon-Nemenyi-McDonald-Thompson test (Hollander & Wolfe, 1999)).

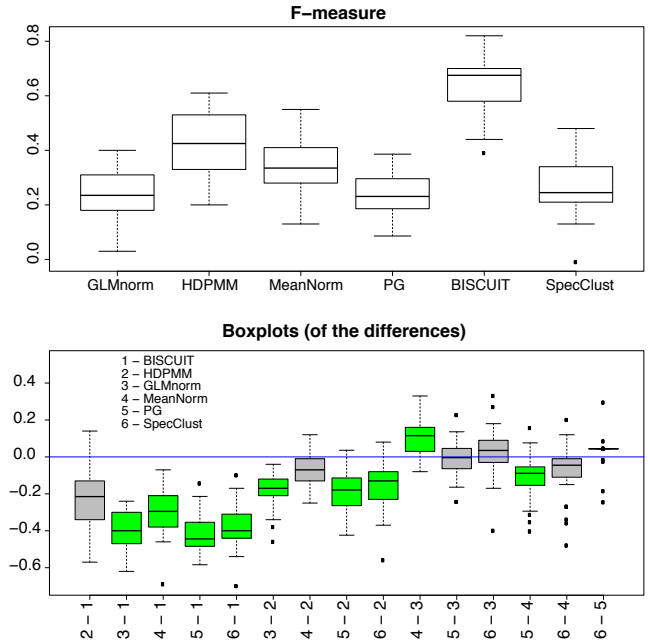

Figure S6: **Top**: Boxplots of F-scores obtained in 15 experiments with randomly-generated X from a non-central Student's t for BISCUIT, HDPMM, GLMnorm, MeanNorm, Phenograph (PG) and Spectral Clustering (SpecClust). **Bottom**: Boxplot of pairwise differences together with color-coded significance (green, if multiple-testing-corrected $p < 0.05$) computed by a non-parametric Friedman test with post-hoc analysis (Wilcoxon-Nemenyi-McDonald-Thompson test (Hollander & Wolfe, 1999)).

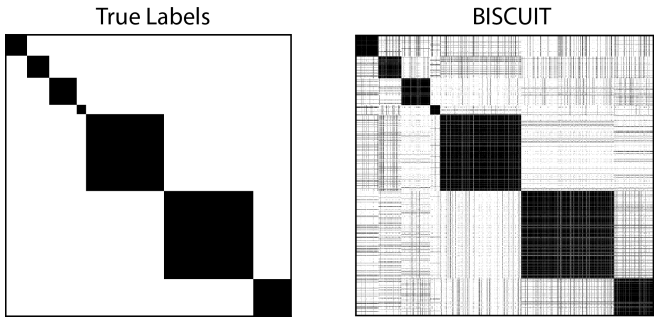

Figure S7: Confusion matrix for inferred cluster assignments using BISCUIT for 3005 cells in the Zeisel et al. (2015) dataset (**right**), compared to actual cell types (**left**).

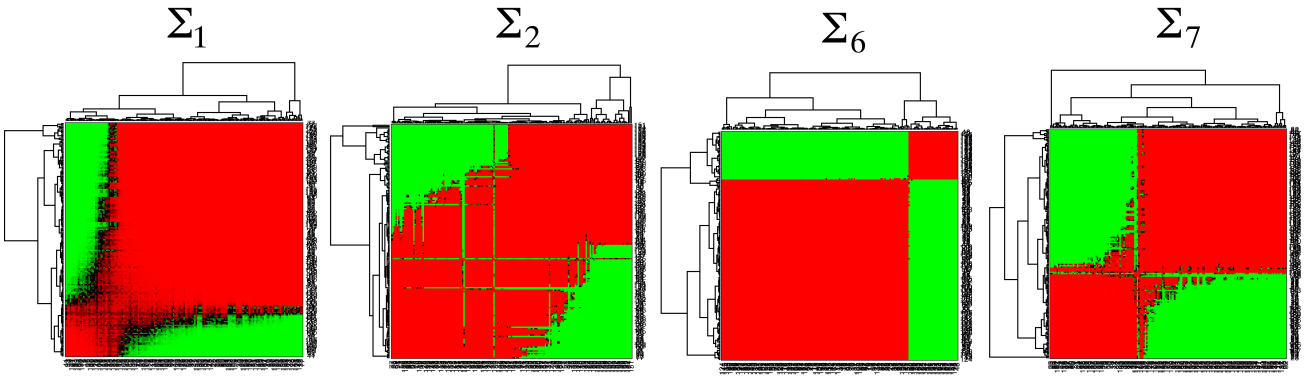

Figure S8: Inferred Σ_k showing different patterns of co-expression for genes in different inferred clusters of cells. In these heatmaps, green shows negative and red shows positive covariance.

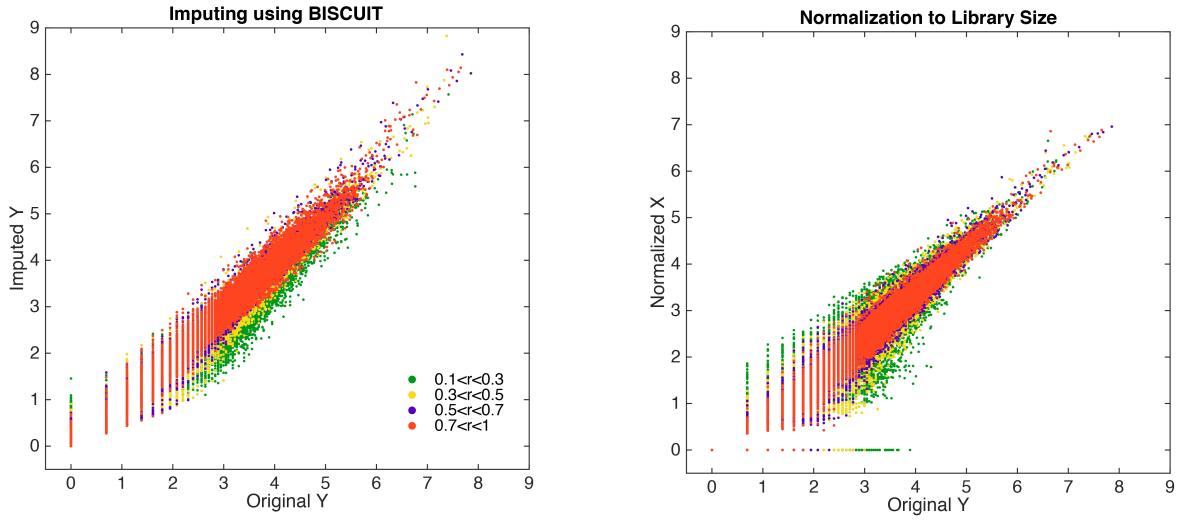

Figure S9: **Left**: Comparison of imputed values per cell per gene in a down-sampled dataset using BISCUIT to original values. **Right**: Values corrected using commonly used normalization approach viz. scaling by mean library size, compared to original values. Each point is a cell colored by its DS rate.