

Supplemental Material for ‘FLASH-MM: fast and scalable single-cell differential expression analysis using linear mixed-effects models’

Contents

1	Linear mixed-effects models	1
1.1	LMM parameter estimation	2
1.2	Hypothesis testing	6
1.3	Prediction of random effects	7
1.4	Numerical algorithms	7
2	Summary statistics based algorithm	8
2.1	Computational complexity	10
2.2	LMM with dimension reduction	10
3	Simulation methods	10
3.1	Performance of simuRNAseq simulator	12
3.2	Differential expression analysis of simulated scRNA-seq data	12
References		12

1 Linear mixed-effects models

In this section, we first describe the linear mixed-effects model (LMM) and then introduce the methods for the LMM parameter estimation, hypothesis testing, prediction of random effects, and gradient algorithms for computing the LMM variance components, see Searle, Casella, and McCulloch (2006), Jiang (2007) and references therein for details.

A linear mixed-effects model (LMM) is an extension of general linear model, which contains both fixed effects and random effects as expressed below (Searle, Casella, and McCulloch 2006)

$$y = X\beta + Zb + \epsilon, \quad (1)$$

where y is an $n \times 1$ vector of observations, X is an $n \times p$ design matrix for fixed effects β , Z is an $n \times q$ design matrix for random effects b , and ϵ is an $n \times 1$ vector of residual errors. The term of random effects may be a combination of various random-effect components:

$$Zb = Z_1b_1 + \cdots + Z_Kb_K,$$

where $Z = [Z_1, \dots, Z_K]$, $b = [b_1^T, \dots, b_K^T]^T$, K is the number of random-effect components, Z_k is an $n \times q_k$ design matrix for the k -th random-effect component, and $\sum_{k=1}^K q_k = q$. The superscript T denotes a transpose of vector or matrix. The basic assumptions are as follows:

- (1) The design matrix X is of full rank, satisfying conditions of estimability for the parameters.
- (2) The random vectors b_k and ϵ are independent and follow a normal distribution, $b_k \sim N(\mathbf{0}, \sigma_k^2 I_{q_k})$ and $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$, where σ_k^2 and σ^2 are unknown parameters, called variance components, $\mathbf{0}$ is a vector or matrix of zero elements, and I_n is an $n \times n$ identity matrix.

Assumption (1) implies $p < n$. We also assume $q_k < n$. If $q_k \geq n$, we can use principal component analysis (PCA) to obtain an equivalent LMM with the number of random effects less than n , as described late by (45).

Hartley and Rao (1967) specified the LMM (1) as mixed analysis of variance (ANOVA) models. Harville (1977) introduced the general linear mixed-effects models with covariance matrices $Cov(b) = D_\phi$ and $Cov(\epsilon) = D_\psi$, specified by the unobservable parameter vectors ϕ and ψ . Laird and Ware (1982) described the linear mixed-effects models for longitudinal data, repeated measures data, or grouped data.

The random effects reflect variations between groups (subjects) and correlations within groups (subjects). Suppose $Cov(b) = \sigma_b^2 I_q$. Then the variance of the i th observation and the correlation between two observations i and j are

$$\begin{aligned} Var(y_i) &= \sigma^2 + \sigma_b^2 z_i^T z_i, \\ Cor(y_i, y_j) &= \frac{\sigma_b^2 z_i^T z_j}{(\sigma^2 + \sigma_b^2 z_i^T z_i)^{1/2} (\sigma^2 + \sigma_b^2 z_j^T z_j)^{1/2}}, \end{aligned}$$

where z_i is the i th row of Z corresponding to the i th observation. If two observations i and j come from different subjects, usually $z_i^T z_i \neq z_j^T z_j$ and then $Var(y_i) \neq Var(y_j)$, reflecting the variation between subjects (inter-subject variability). If the two observations come from same subject, usually $z_i = z_j$ and then $Cor(y_i, y_j) > 0$, that is, the two observations within a subject are correlated, reflecting the intra-subject correlation.

1.1 LMM parameter estimation

Hartley and Rao (1967) developed maximum likelihood (ML) method for estimating the LMM parameters, that is, the fixed effects and variance components. The ML method estimates all parameters of fixed effects and variance components together. Patterson and Thompson (1971) proposed a modified maximum likelihood procedure which partitions the data into two mutually uncorrelated parts, one being free of the fixed effects used for estimating variance components, called restricted maximum likelihood (REML) estimators. The REML estimator is unbiased, and the ML estimator (MLE) of variance components is biased in general. For given variance components, both ML and REML methods provide with the same estimates of fixed effects. In this section, we briefly describe both ML and REML methods, see Hartley and Rao (1967), Patterson and Thompson (1971), Harville (1977), and Jiang (2007) for the details.

1.1.1 Maximum likelihood estimation

Under the assumptions that random vectors b_i and ϵ are independent and have a normal distribution, $b_i \sim N(\mathbf{0}, \sigma_i^2 I_{q_i})$ and $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$, we have $y \sim N(X\beta, V_\theta)$, with a probability density function (pdf)

$$f(y|\beta, \theta) = \frac{1}{\sqrt{(2\pi)^n \det(V_\theta)}} \exp\left[-\frac{1}{2}(y - X\beta)^T V_\theta^{-1} (y - X\beta)\right], \quad (2)$$

where

$$\begin{aligned} \theta &= [\theta_0, \theta_1, \dots, \theta_K]^T = [\sigma^2, \sigma_1^2, \dots, \sigma_K^2]^T, \\ V_\theta &= \sigma^2 I_n + \sigma_1^2 Z_1 Z_1^T + \dots + \sigma_K^2 Z_K Z_K^T = \sigma^2 I_n + Z D_\theta Z^T, \\ D_\theta &= \begin{bmatrix} \theta_1 I_{q_1} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \theta_K I_{q_K} \end{bmatrix}. \end{aligned}$$

The maximum likelihood estimation (MLE) is obtained by maximizing the log-likelihood

$$l(\beta, \theta) = \log f(y|\beta, \theta) = -\frac{1}{2}n \log 2\pi - \frac{1}{2} \log \det(V_\theta) - \frac{1}{2}(y - X\beta)^T V_\theta^{-1} (y - X\beta). \quad (3)$$

The first derivatives of the log-likelihood are

$$\begin{cases} \frac{\partial l}{\partial \beta} = X^T V_\theta^{-1} y - X^T V_\theta^{-1} X\beta, \\ \frac{\partial l}{\partial \theta_i} = -\frac{1}{2} \text{tr}(V_\theta^{-1} V_i) + \frac{1}{2}(y - X\beta)^T V_\theta^{-1} V_i V_\theta^{-1} (y - X\beta), \quad i = 0, \dots, K, \end{cases} \quad (4)$$

where $V_i = Z_i Z_i^T$ and $Z_0 = I_n$. By equating the first derivatives to zero, we have the MLE equations:

$$\begin{cases} X^T V_\theta^{-1} X \beta = X^T V_\theta^{-1} y, \\ \text{tr}(V_\theta^{-1} V_i) = y^T R_\theta V_i R_\theta y, \quad i = 0, \dots, K, \end{cases} \quad (5)$$

where

$$R_\theta = V_\theta^{-1} - V_\theta^{-1} X (X^T V_\theta^{-1} X)^{-1} X^T V_\theta^{-1}. \quad (6)$$

With θ given, from (5), we have the MLE of β and the covariance matrix:

$$\begin{aligned} \hat{\beta} &= (X^T V_\theta^{-1} X)^{-1} X^T V_\theta^{-1} y, \\ \text{var}(\hat{\beta}) &= (X^T V_\theta^{-1} X)^{-1}. \end{aligned} \quad (7)$$

Note

$$P_{V_\theta^{-1/2} X} = V_\theta^{-1/2} X (X^T V_\theta^{-1} X)^{-1} X^T V_\theta^{-1/2}$$

is a projection matrix, and $R_\theta = V_\theta^{-1/2} (I_n - P_{V_\theta^{-1/2} X}) V_\theta^{-1/2}$ is a residual (maker) matrix. The following properties can be readily verified:

$$R_\theta X = \mathbf{0}, \quad R_\theta V_\theta R_\theta = R_\theta. \quad (8)$$

The information matrix is given by the second derivatives of $l(\beta, \theta)$

$$\begin{cases} \frac{\partial^2 l}{\partial \beta \partial \beta^T} = -X^T V_\theta^{-1} X \\ \frac{\partial^2 l}{\partial \beta \partial \theta_i} = -X^T V_\theta^{-1} V_i V_\theta^{-1} (y - X\beta) \\ \frac{\partial^2 l}{\partial \theta_i \partial \theta_j} = \frac{1}{2} \text{tr}(V_\theta^{-1} V_i V_\theta^{-1} V_j) - (y - X\beta)^T V_\theta^{-1} V_i V_\theta^{-1} V_j V_\theta^{-1} (y - X\beta) \end{cases} \quad (9)$$

Using $E(y - X\beta) = \mathbf{0}$ and $E[(y - X\beta)^T A (y - X\beta)] = \text{tr}(AV_\theta)$ for any $n \times n$ matrix A , we have Fisher information matrix:

$$I(\beta, \theta) = - \begin{bmatrix} E\left(\frac{\partial^2 l}{\partial \beta \partial \beta^T}\right) & E\left(\frac{\partial^2 l}{\partial \beta \partial \theta^T}\right) \\ E\left(\frac{\partial^2 l}{\partial \theta \partial \beta^T}\right) & E\left(\frac{\partial^2 l}{\partial \theta \partial \theta^T}\right) \end{bmatrix} = \begin{bmatrix} \text{var}(\hat{\beta})^{-1} & \mathbf{0} \\ \mathbf{0} & I(\theta) \end{bmatrix}, \quad (10)$$

where

$$\begin{aligned} \text{var}(\hat{\beta}) &= (X^T V_\theta^{-1} X)^{-1}, \\ I(\theta) &= -E\left(\frac{\partial^2 l}{\partial \theta \partial \theta^T}\right) = \left\{ \frac{1}{2} \text{tr}(V_\theta^{-1} V_i V_\theta^{-1} V_j) \right\}_{0 \leq i, j \leq K}. \end{aligned}$$

Recall that I_n denotes an $n \times n$ identity matrix. Without confusion, $I(\cdot)$ denotes a function of the Fisher information matrix with corresponding parameters as variables. The Fisher information matrix may also be written as

$$I(\beta, \theta) = E \left[\left(\begin{pmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \theta} \end{pmatrix} \right) \left(\begin{pmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \theta} \end{pmatrix} \right)^T \right],$$

which is a positive semidefinite matrix.

Let $\gamma_i = \sigma_i^2 / \sigma^2$, $i = 1, \dots, K$. Then $V_\theta = \sigma^2 V_\gamma$, where

$$V_\gamma = I_n + \gamma_1 Z_1 Z_1^T + \dots + \gamma_K Z_K Z_K^T = I_n + Z D_\gamma Z^T. \quad (11)$$

From Weinstein–Aronszajn identity:

$$\det(I_n + AB^T) = \det(I_q + B^T A) \quad (12)$$

for any $n \times q$ matrices A and B , we have

$$\det(V_\theta) = \sigma^{2n} \det(I_n + Z D_\gamma Z^T) = \sigma^{2n} \det(I_q + D_\gamma Z^T Z).$$

Note $V_\theta^{-1}(y - X\hat{\beta}) = R_\theta y$. Given the MLEs, $\hat{\beta}$ and $\hat{\theta}$, the log-likelihood (3) can be estimated by

$$l(\hat{\beta}, \hat{\theta}) = -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2} \log \det(I_q + D_{\hat{\gamma}} Z^T Z) - \frac{1}{2} y^T R_\theta V_\theta R_\theta y, \quad (13)$$

where $\hat{\gamma}_i = \hat{\sigma}_i^2 / \hat{\sigma}^2$. From the MLE equations (5),

$$y^T R_\theta V_\theta R_\theta y = \sum_{i=0}^K \hat{\sigma}_i^2 y^T R_\theta V_i R_\theta y = \sum_{i=0}^K \hat{\sigma}_i^2 \text{tr}(V_\theta^{-1} V_i) = n.$$

Then the log-likelihood (13) is reduced as

$$l(\hat{\beta}, \hat{\theta}) = -\frac{n}{2} - \frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2} \log \det(I_q + D_{\gamma} Z^T Z). \quad (14)$$

From $y^T R_\theta V_\theta R_\theta y = \sigma^{-2} y^T R_\gamma V_\gamma R_\gamma y$ and $R_\gamma y = V_\gamma^{-1}(y - X\hat{\beta})$, we also have

$$\hat{\sigma}^2 = \frac{1}{n} y^T R_{\hat{\gamma}} V_{\hat{\gamma}} R_{\hat{\gamma}} y = \frac{1}{n} (y - X\hat{\beta})^T V_{\hat{\gamma}}^{-1} (y - X\hat{\beta}). \quad (15)$$

Remark: The Weinstein–Aronszajn identity (12) can be derived from the equality

$$\begin{bmatrix} I_n & \mathbf{0} \\ B^T & I_q \end{bmatrix} \begin{bmatrix} I_n + AB^T & A \\ \mathbf{0} & I_q \end{bmatrix} \begin{bmatrix} I_n & \mathbf{0} \\ -B^T & I_q \end{bmatrix} = \begin{bmatrix} I_n & A \\ \mathbf{0} & I_q + B^T A \end{bmatrix}.$$

1.1.2 Restricted maximum likelihood

Let Q be an $n \times (n - p)$ full column rank matrix such that $Q^T X = \mathbf{0}$, and $L = V_\theta^{-1} X$, an $n \times p$ full column rank matrix. Then the data y can be partitioned into two parts: $z = Q^T y$ and $u = L^T y$. The data z and u are uncorrelated since $\text{Cov}(z, u) = Q^T V_\theta L = Q^T X = \mathbf{0}$. From (1),

$$z = Q^T Z b + Q^T \epsilon \sim N(\mathbf{0}, Q^T V_\theta Q).$$

The transformed data, z , does not contain the fixed effects β . Based on the log-likelihood function of z ,

$$l_R(\theta) = -\frac{n-p}{2} \log 2\pi - \frac{1}{2} \log \det(Q^T V_\theta Q) - \frac{1}{2} y^T Q (Q^T V_\theta Q)^{-1} Q^T y,$$

similar to the MLE equations (5), we have the restricted maximum likelihood (REML) equations:

$$\text{tr}[(Q^T V_\theta Q)^{-1} Q^T V_i Q] = y^T Q (Q^T V_\theta Q)^{-1} Q^T V_i Q (Q^T V_\theta Q)^{-1} Q^T y, \quad i = 0, \dots, K. \quad (16)$$

The REML equations can be further simplified. Let

$$P_A = A(A^T A)^{-1} A^T$$

be the projection matrix for a full column rank matrix A . Based on $Q^T X = \mathbf{0}$ and both X and Q are of full rank, for an $n \times n$ positive definite matrix V , we have the following equation (Jiang 2007)

$$P_{V^{1/2} Q} = I_n - P_{V^{-1/2} X},$$

that is,

$$Q(Q^T V Q)^{-1} Q^T = V^{-1} - V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1}. \quad (17)$$

With (17), the REML equations (16) reduces to

$$\text{tr}(R_\theta V_i) = y^T R_\theta V_i R_\theta y, \quad i = 0, \dots, K, \quad (18)$$

where R_θ is defined as (6),

$$R_\theta = V_\theta^{-1} - V_\theta^{-1} X (X^T V_\theta^{-1} X)^{-1} X^T V_\theta^{-1}.$$

The REML equations (18) do not contain Q . Thus the REML estimator does not depend on Q .

Let $Q = (I_n - P_X)C$, where C is an $n \times (n - p)$ matrix. Thus

$$z = Q^T y = C^T (I_n - P_X) y = C^T (y - X \hat{\beta}_{LS}),$$

where $\hat{\beta}_{LS} = (X^T X)^{-1} X^T y$ is the least square estimator of β , is a linear combination of residuals obtained after fitting the fixed effects. So the z is also called error contrasts (Harville 1977).

The REML equations do not include the fixed effects. The fixed effects can be estimated based on the second part of the data: $u = L^T y$, where $L = V_\theta^{-1} X$. From (1),

$$u = L^T X \beta + L^T Z b + L^T \epsilon \sim N(L^T X \beta, L^T V_\theta L).$$

With θ fixed, based on the data u , the MLE of β is given by

$$\hat{\beta} = [X^T L (L^T V_\theta L)^{-1} L^T X]^{-1} [X^T L (L^T V_\theta L)^{-1} L^T y] = (X^T V_\theta^{-1} X)^{-1} X^T V_\theta^{-1} y. \quad (19)$$

For a given θ , the MLE of β based on u is exactly the same with that based on y in (7).

The first and second derivatives of l_R are given as follows:

$$\frac{\partial l_R}{\partial \theta_i} = -\frac{1}{2} \text{tr}(R_\theta V_i) + \frac{1}{2} y^T R_\theta V_i R_\theta y, \quad (20)$$

$$\frac{\partial^2 l_R}{\partial \theta_i \partial \theta_j} = \frac{1}{2} \text{tr}(R_\theta V_i R_\theta V_j) - y^T R_\theta V_i R_\theta V_j R_\theta y. \quad (21)$$

From (21), using $E(y^T A y) = \text{tr}(A V_\theta) + \beta^T X^T A X \beta$, $R_\theta X = \mathbf{0}$ and $R_\theta V_\theta R_\theta = R_\theta$, we have Fisher information matrix:

$$I(\theta) = -E\left(\frac{\partial^2 l_R}{\partial \theta \partial \theta^T}\right) = \left\{ \frac{1}{2} \text{tr}(R_\theta V_i R_\theta V_j) \right\}_{0 \leq i, j \leq K}. \quad (22)$$

Recall $V_\theta = \sigma^2 (I_n + Z D_\gamma Z^T)$. From (17) and the Weinstein–Aronszajn identity (12),

$$\begin{aligned} \det(Q^T V_\theta Q) &= \det[\sigma^2 (Q^T Q + Q^T Z D_\gamma Z^T Q)] \\ &= \sigma^{2(n-p)} \det(Q^T Q) \det(I_{n-p} + (Q^T Q)^{-1} Q^T Z D_\gamma Z^T Q) \\ &= \sigma^{2(n-p)} \det(Q^T Q) \det(I_q + D_\gamma Z^T Q (Q^T Q)^{-1} Q^T Z) \\ &= \sigma^{2(n-p)} \det(Q^T Q) \det(I_q + D_\gamma Z^T R Z), \end{aligned} \quad (23)$$

where

$$R = Q(Q^T Q)^{-1} Q^T = I_n - X(X^T X)^{-1} X^T = I_n - P_X.$$

From (17), $Q(Q^T V_\theta Q)^{-1} Q^T = R_\theta$. Then the REML log-likelihood function can also be expressed as

$$l_R(\theta) = -\frac{1}{2} \log \det(Q^T Q) - \frac{n-p}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log \det(I_q + D_\gamma Z^T R Z) - \frac{1}{2} y^T R_\theta y. \quad (24)$$

This also shows that the REML estimator, obtained by maximizing $l_R(\theta)$, doesn't depend on the choice of Q . We can choose Q such that $Q^T Q = I_{n-p}$. From the REML equations (18),

$$y^T R_\theta V_\theta R_\theta y = \text{tr}(R_\theta V_\theta) = \text{tr}(I_n - V_{\hat{\theta}}^{-1} X (X^T V_{\hat{\theta}}^{-1} X)^{-1} X^T) = n-p.$$

From (8), $y^T R_{\hat{\theta}} y = y^T R_{\hat{\theta}} V_{\hat{\theta}} R_{\hat{\theta}} y = n-p$, and then the REML log-likelihood function can be estimated by

$$l_R(\hat{\theta}) = -\frac{n-p}{2} - \frac{n-p}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2} \log \det(I_q + D_{\hat{\gamma}} Z^T R Z). \quad (25)$$

From $y^T R_\theta V_\theta R_\theta y = \sigma^{-2} y^T R_\gamma V_\gamma R_\gamma y$ and $V_\theta^{-1} (y - X \hat{\beta}) = R_\theta y$, we also have

$$\hat{\sigma}^2 = \frac{1}{n-p} y^T R_{\hat{\gamma}} V_{\hat{\gamma}} R_{\hat{\gamma}} y = \frac{1}{n-p} (y - X \hat{\beta})^T V_{\hat{\gamma}}^{-1} (y - X \hat{\beta}). \quad (26)$$

1.2 Hypothesis testing

The hypotheses for testing fixed effects and variance components can be respectively defined as

$$H_{0,i} : \beta_i = 0 \text{ versus } H_{1,i} : \beta_i \neq 0,$$

$$H_{0,k} : \sigma_k^2 = 0 \text{ versus } H_{1,k} : \sigma_k^2 > 0.$$

Under regularity conditions, the MLE is consistent and asymptotically normal with asymptotic covariance matrix equal to the inverse of Fisher information matrix (Jiang 2007), that is, asymptotically

$$\hat{\beta} - \beta \sim N(\mathbf{0}, \text{var}(\hat{\beta})), \quad (27)$$

$$\hat{\theta} - \theta \sim N(\mathbf{0}, I(\theta)^{-1}). \quad (28)$$

The asymptotic normality is inappropriate for the variance components under null hypothesis, $\sigma_k^2 = 0$, because the zero variance is on the boundary of the parameter space. To avoid the problem of zero-boundary in hypothesis testing, we reparameterize the variance components by $\theta_k = \sigma^2 \gamma_k$, $k = 1, \dots, K$, and allow the parameters, γ_k or θ_k , to take negative values. The lower boundary of the parameters can be a negative value such that the variance-covariance matrix

$$V_\theta = \sigma^2 (I_n + \gamma_1 Z_1 Z_1^T + \dots + \gamma_K Z_K Z_K^T),$$

is well defined (positive-definite). Note that V_θ is positive-definite when $\gamma_k > -1/\lambda_{\max}$, where $\lambda_{\max} > 0$, is the largest singular value of ZZ^T . That is, V_θ can be well-defined with negative parameters θ_k . Then the hypotheses for the variance components are extended as

$$H_{0,k} : \theta_k \leq 0 \text{ versus } H_{1,k} : \theta_k > 0,$$

in which the zero components, $\theta_k = 0$, are no longer on the boundary of the parameter space, and the the MLE asymptotic properties hold under regularity conditions, which enables the use of z-statistics or t-statistics for testing fixed effects and variance components. If $\theta_k > 0$, $\sigma_k^2 = \theta_k$ is definable and the mixed-effects model is well-specified. Otherwise, $\theta_k \leq 0$, implies that the term of random effects is not needed in the model design.

The test statistics for fixed effects are given by

$$T_i = \frac{\hat{\beta}_i}{\sqrt{\text{var}(\hat{\beta}_i)}} = \frac{\hat{\beta}_i}{\sqrt{\text{var}(\hat{\beta})_{ii}}} \sim t(n-p). \quad (29)$$

The test statistic for a contrast, a linear combination of the estimated fixed effects, $c^T \hat{\beta}$, is

$$T_c = \frac{c^T \hat{\beta}}{\sqrt{c^T \text{var}(\hat{\beta}) c}} \sim t(n-p). \quad (30)$$

The test statistics for the parameters of variance components are given by

$$T_{\theta_k} = \frac{\hat{\theta}_k}{\sqrt{[I(\hat{\theta})^{-1}]_{kk}}} \sim N(0, 1). \quad (31)$$

If $T_{\theta_k} > 0$, then $\sigma_k^2 = \theta_k$ is definable. Otherwise, $T_{\theta_k} \leq 0$, means that the random effects are not required or can be ignored.

1.3 Prediction of random effects

With the assumption of normality: $b \sim N(\mathbf{0}, D_\theta)$ and $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$, from (1), we have

$$\begin{bmatrix} b \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ X\beta \end{bmatrix}, \begin{bmatrix} D_\theta & D_\theta Z^T \\ ZD_\theta & V_\theta \end{bmatrix}\right),$$

Note $V_\theta = \sigma^2 I_n + ZD_\theta Z^T$. The prediction of random effects is given by a conditional mean:

$$\hat{b} = E(b|y) = D_\theta Z^T V_\theta^{-1} (y - X\beta). \quad (32)$$

Since

$$E_{b,y}(\|\hat{b} - b\|^2) = E_y(\|\hat{b} - E(b|y)\|^2) + E_{b,y}(\|b - E(b|y)\|^2),$$

$\hat{b} = E(b|y)$, is the best predictor in the sense of minimum mean squared error (MSE) of prediction. It is seen from $E(\hat{b}) = E_y[E(b|y)] = E(b)$, that the best predictor is unbiased. Without any assumption of normality, the prediction (32) can also be derived from the best linear predictor of the form

$$b_{BLP} = a + A[y - E(y)],$$

by minimizing $E(\|b_{BLP} - b\|^2)$ which yields

$$a = E(b) = \mathbf{0}, \quad A = \text{Cov}(b, y)[\text{Cov}(y)]^{-1} = D_\theta Z^T V_\theta^{-1}.$$

Therefore the \hat{b} is the best linear unbiased prediction (BLUP) (Searle, Casella, and McCulloch 2006). It is also known that the MLE of β ,

$$\hat{\beta} = (X^T V_\theta^{-1} X)^{-1} X^T V_\theta^{-1} y,$$

is the best linear unbiased estimator (BLUE), which does not require the normality assumption. This can be readily verified. Let $\beta_{LUE} = \hat{\beta} + Cy$ be any linear unbiased estimator of β . Then we have $CX = \mathbf{0}$ and

$$E(\|\beta_{LUE} - \beta\|^2) = E(\|\hat{\beta} - \beta\|^2) + \text{tr}(CV_\theta C^T) \geq E(\|\hat{\beta} - \beta\|^2).$$

Thus $\hat{\beta}$ minimizing the MSE is the BLUE.

The BLUP of $E(y|b) = X\beta + Zb$ is give by

$$\begin{aligned} \hat{y} &= X\beta + Z\hat{b} \\ &= X\beta + ZD_\theta Z^T V_\theta^{-1} (y - X\beta) \\ &= X\beta + (I_n - \sigma^2 V_\theta^{-1})(y - X\beta) \\ &= y - \sigma^2 V_\theta^{-1} (y - X\beta). \end{aligned} \quad (33)$$

Substituting $\hat{\beta}$ and $\hat{\theta}$ for the unknown parameters in (32) and (33), we have the empirical BLUPs (Harville 1991),

$$\hat{b} = E(b|y) = D_{\hat{\theta}} Z^T V_{\hat{\theta}}^{-1} (y - X\hat{\beta}) = D_{\hat{\theta}} Z^T R_{\hat{\theta}} y, \quad (34)$$

$$\hat{y} = E(y|b) = y - \hat{\sigma}^2 R_{\hat{\theta}} y, \quad (35)$$

where R_θ is defined by (6), that is,

$$R_\theta = V_\theta^{-1} - V_\theta^{-1} X (X^T V_\theta^{-1} X)^{-1} X^T V_\theta^{-1}.$$

1.4 Numerical algorithms

With variance components estimated, the fixed effects estimated by ML and REML methods are given by (7) or (19) as follows

$$\hat{\beta} = (X^T V_\theta^{-1} X)^{-1} X^T V_\theta^{-1} y.$$

Estimating variance components by either ML or REML is a numerical optimization problem. Various iterative methods based on the log likelihood, called gradient methods, have been proposed to compute the ML and REML estimates (Searle, Casella, and McCulloch 2006). The gradient methods are represented by the iteration equation

$$\theta^{(i+1)} = \theta^{(i)} + \Gamma(\theta^{(i)}) \frac{\partial l(\theta^{(i)})}{\partial \theta}, \quad (36)$$

where $\theta^{(i)}$ represents the value of the estimate of the parameter vector θ at the i th iteration, $\partial l(\theta)/\partial \theta$ is the gradient of the log likelihood function, and $\Gamma(\theta)$ is a modifier matrix of the gradient direction. Let $H(\theta)$ and $I(\theta)$ be the Hessian matrix and Fisher information matrix of the log likelihood function with respect to θ . Let $I_A(\theta) = [I(\theta) - H(\theta)]/2$ be the average information matrix. The modifier matrix can be specified by

- 1) Newton–Raphson: $\Gamma(\theta) = -H(\theta)^{-1}$;
- 2) Fisher scoring: $\Gamma(\theta) = I(\theta)^{-1}$;
- 3) Average information: $\Gamma(\theta) = I_A(\theta)^{-1}$.

For MLE, from (4), (5), (9) and (10), we have

$$\begin{aligned} \frac{\partial l}{\partial \theta_i} &= -\frac{1}{2} [tr(V_\theta^{-1} V_i) - y^T R_\theta V_i R_\theta y], \\ H_{ij} &= \frac{\partial^2 l}{\partial \theta_i \partial \theta_j} = \frac{1}{2} tr(V_\theta^{-1} V_i V_\theta^{-1} V_j) - y^T R_\theta V_i V_\theta^{-1} V_j R_\theta y, \\ I_{ij} &= -E(H_{ij}) = \frac{1}{2} tr(V_\theta^{-1} V_i V_\theta^{-1} V_j), \\ I_{A,ij} &= \frac{1}{2}(I_{ij} - H_{ij}) = \frac{1}{2} y^T R_\theta V_i V_\theta^{-1} V_j R_\theta y. \end{aligned} \quad (37)$$

For REML, $l = l_R(\theta)$. From (20), (21) and (22), we have

$$\begin{aligned} \frac{\partial l_R}{\partial \theta_i} &= -\frac{1}{2} [tr(R_\theta V_i) - y^T R_\theta V_i R_\theta y], \\ H_{ij} &= \frac{\partial^2 l_R}{\partial \theta_i \partial \theta_j} = \frac{1}{2} tr(R_\theta V_i R_\theta V_j) - y^T R_\theta V_i R_\theta V_j R_\theta y, \\ I_{ij} &= -E(H_{ij}) = \frac{1}{2} tr(R_\theta V_i R_\theta V_j), \\ I_{A,ij} &= \frac{1}{2}(I_{ij} - H_{ij}) = \frac{1}{2} y^T R_\theta V_i R_\theta V_j R_\theta y. \end{aligned} \quad (38)$$

Here H_{ij} , I_{ij} and $I_{A,ij}$ represent the (i, j) entry of $H(\theta)$, $I(\theta)$ and $I_A(\theta)$, respectively. Recall $V_i = Z_i Z_i^T$, defined in (4), and

$$\begin{aligned} V_\theta &= \sigma^2 I_n + Z D_\theta Z^T, \\ R_\theta &= V_\theta^{-1} - V_\theta^{-1} X (X^T V_\theta^{-1} X)^{-1} X^T V_\theta^{-1}, \end{aligned}$$

defined in (2) and by (6), respectively.

2 Summary statistics based algorithm

This section describes our work to improve LMM solver scalability and memory usage efficiency in the context of single cell transcriptomics data. The gradient methods given by (36), (37) and (38) have a computational complexity of $O(n^3)$. The computational burden arises from the $n \times n$ matrix inverse, $V^{-1}(\theta)$, and the matrix-matrix products, $V_\theta^{-1} V_i V_\theta^{-1} V_j$ and $R_\theta V_i R_\theta V_j$. Now we derive a summary statistics based algorithm to implement the gradient methods for the LMM estimation. The $n \times n$ matrix inverse and matrix-matrix products are computed through the low dimension $p \times p$ and $q \times q$ matrices. The algorithm achieves a computational complexity of $O(n(p^2 + q^2))$, which is fast and linearly scalable with the sample size n . In addition, using the summary statistics requires less computer memory usage to estimate the LMM parameters, which enables the LMM application to the large-scale data analysis. By precomputing and directly using the summary statistics as inputs, the algorithm has a computational complexity of $O(p^3 + q^3)$, which makes computations independent of the sample size n and achieves both speed and memory efficiency.

Recall that $\gamma_k = \sigma_k^2 / \sigma^2$, $k = 1, \dots, K$, and

$$\begin{aligned} V_\gamma &= I_n + \gamma_1 Z_1 Z_1^T + \dots + \gamma_K Z_K Z_K^T = I_n + Z D Z^T, \\ R_\gamma &= V_\gamma^{-1} - V_\gamma^{-1} X (X^T V_\gamma^{-1} X)^{-1} X^T V_\gamma^{-1}, \end{aligned}$$

where $D = D_\gamma$. Then $V_\theta = \sigma^2 V_\gamma$, $R_\theta = \sigma^{-2} R_\gamma$, and

$$\begin{aligned}
\text{tr}(V_\theta^{-1} V_i) &= \sigma^{-2} \text{tr}(Z_i^T V_\gamma^{-1} Z_i), \\
\text{tr}(V_\theta^{-1} V_i V_\theta^{-1} V_j) &= \sigma^{-4} \text{tr}[(Z_i^T V_\gamma^{-1} Z_j)^T (Z_i^T V_\gamma^{-1} Z_j)], \\
y^T R_\theta V_i V_\theta^{-1} V_j R_\theta y &= \sigma^{-6} \text{tr}[(Z_i^T R_\gamma y)^T (Z_i^T V_\gamma^{-1} Z_j) (Z_j^T R_\gamma y)], \\
\text{tr}(R_\theta V_i) &= \sigma^{-2} \text{tr}(Z_i^T R_\gamma Z_i), \\
\text{tr}(R_\theta V_i R_\theta V_j) &= \sigma^{-4} \text{tr}[(Z_i^T R_\gamma Z_j)^T (Z_i^T R_\gamma Z_j)], \\
y^T R_\theta V_i R_\theta y &= \sigma^{-4} \text{tr}[(Z_i^T R_\gamma y)^T (Z_i^T R_\gamma y)], \\
y^T R_\theta V_i R_\theta V_j R_\theta y &= \sigma^{-6} \text{tr}[(Z_i^T R_\gamma y)^T (Z_i^T R_\gamma Z_j) (Z_j^T R_\gamma y)]. \tag{39}
\end{aligned}$$

Using the formulas of matrix identities given by Harville (1977) and Rao (1979), we have

$$V_\gamma^{-1} = I_n - Z(I_q + DZ^T Z)^{-1} DZ^T,$$

$$R_\gamma = R - RZ(I_q + DZ^T RZ)^{-1} DZ^T R,$$

where

$$R = I_n - X(X^T X)^{-1} X^T.$$

Let

$$\begin{aligned}
M_0 &= (I_q + DZ^T Z)^{-1}, \\
M &= (I_q + DZ^T RZ)^{-1}. \tag{40}
\end{aligned}$$

Then $V_\gamma^{-1} = I_n - ZM_0 DZ^T$ and $R_\gamma = R - RZMDZ^T R$. By the identity equation

$$DZ^T Z M_0 = M_0 DZ^T Z = I_q - M_0,$$

$$DZ^T RZ M = MDZ^T RZ = I_q - M,$$

and $R^2 = R$, we have $V_\gamma^{-2} = V_\gamma^{-1} - ZM_0^2 DZ^T$, $R_\gamma^2 = R_\gamma - RZM^2 DZ^T R$, and

$$\begin{aligned}
Z_i^T V_\gamma^{-1} Z_j &= Z_i^T Z_j - (Z_i^T Z) M_0 D(Z^T Z_j), \\
Z_i^T V_\gamma^{-2} Z_i &= Z_i^T V_\gamma^{-1} Z_i - (Z_i^T Z) M_0^2 D(Z^T Z_i), \\
\text{tr}(V_\gamma^{-1}) &= n - q + \text{tr}(M_0), \\
\text{tr}(V_\gamma^{-2}) &= n - q + \text{tr}(M_0^2), \tag{41}
\end{aligned}$$

$$\begin{aligned}
Z_i^T R_\gamma Z_j &= Z_i^T RZ_j - (Z_i^T RZ) M D(Z^T RZ_j), \\
Z_i^T R_\gamma^2 Z_i &= Z_i^T R_\gamma Z_i - (Z_i^T RZ) M^2 D(Z^T RZ_i), \\
\text{tr}(R_\gamma) &= \text{tr}(R) - q + \text{tr}(M), \\
\text{tr}(R_\gamma^2) &= \text{tr}(R) - q + \text{tr}(M^2), \tag{42}
\end{aligned}$$

$$\begin{aligned}
Z_i^T R_\gamma y &= Z_i^T Ry - (Z_i^T RZ) M D(Z^T Ry), \\
y^T R_\gamma^2 y &= y^T R_\gamma y - (y^T RZ) M^2 D(Z^T Ry), \\
y^T R_\gamma y &= y^T Ry - (y^T RZ) M D(Z^T Ry), \tag{43}
\end{aligned}$$

where

$$\begin{aligned}
Z_i^T RZ_j &= Z_i^T Z_j - (Z_i^T X)(X^T X)^{-1}(X^T Z_j), \\
Z^T Ry &= Z^T y - (Z^T X)(X^T X)^{-1}(X^T y), \\
y^T Ry &= y^T y - (y^T X)(X^T X)^{-1}(X^T y). \tag{44}
\end{aligned}$$

Thus the gradient methods given by (36), (37) and (38) can be implemented by using the summary statistics: $X^T X$, $X^T Z$, $X^T y$, $Z^T y$ and $Z^T Z$, based on the formulas (39)~(44).

2.1 Computational complexity

The complexity for computing the summary statistics: $X^T X$, $X^T y$, $Z^T X$, $Z^T y$ and $Z^T Z$, is $O(n(p^2 + q^2))$. The complexity for computing the $q \times q$ matrix inverse, M_0 and M , by (40) is $O(q^3)$ or even $O(q^2 \log(q))$. Using the summary statistics as the inputs, the algorithms given by (42), (43) and (44) have a complexity of $O(p^3 + q^3)$. Then the summary statistics based algorithm given by equations (39)~(44) has the complexity

$$O(n(p^2 + q^2) + p^3 + q^3) = O(n(p^2 + q^2)),$$

since $p < n$, $q_k < n$ and $q < Kn$. The summary statistics can be computed in advance and stored in a computer with less storage. Once the summary statistics are computed and directly used as the inputs, the algorithm complexity, $O(p^3 + q^3)$, doesn't depend on the sample size n .

2.2 LMM with dimension reduction

The LMM estimation algorithm can be further sped up by reducing the number of random effects. For a large number of random effects, we may combine the correlated random effects by cluster analysis or principal component analysis (PCA). By PCA, we have

$$Z_k = U_k V_k^T,$$

where U_k is an $n \times r_k$ matrix of principal components (PCs), V_k is an $q_k \times r_k$ matrix, $r_k \leq \min(n, q_k)$ is the rank of Z_k , $U_k^T U_k$ is diagonal, and $V_k^T V_k = I_{r_k}$, $k = 1, \dots, K$. Thus the LMM (1) can be rewritten as

$$y = X\beta + U_1 v_1 + \dots + U_K v_K + \epsilon = X\beta + Uv + \epsilon, \quad (45)$$

where $v_k = V_k^T b_k \sim N(\mathbf{0}, \sigma_k^2 I_{r_k})$, is a r_k -vector of random effects. The LMM (45) is equivalent to the LMM (1). The random effect design matrices U_k in the LMM (45) have a lower dimension r_k . We may even approximate the design matrix Z_k using a smaller number of PCs such that $r_k \ll \min(n, q_k)$.

3 Simulation methods

To simulate the multi-subject multi-cell-type single-cell RNA-seq (scRNA-seq) dataset, we developed a scRNA-seq simulator, named simuRNaseq, by using a reference dataset based on a negative binomial (NB) distribution. The reference dataset contains genes-by-cells counts matrix and meta data. The simulated genes are randomly selected from the reference data. The simulated cells are randomly selected from the meta data that specifies subjects, cell-types and treatment conditions for the cells. If the reference data is not available, it will be generated randomly. The simuRNaseq workflow is illustrated in the Supplemental Figure S1, which consists of following main steps:

- (1) Estimate the dispersion and mean of the NB distribution for each gene using the reference count data.
- (2) Select the differentially expressed (DE) genes and non-DE genes randomly from the genes in the reference count data and the cells from the meta data, with given numbers.
- (3) Generate counts by sampling from the NB distributions for the non-DE genes without log-fold-change (logFC) between treatments and for the DE genes with assigned logFC between treatments.

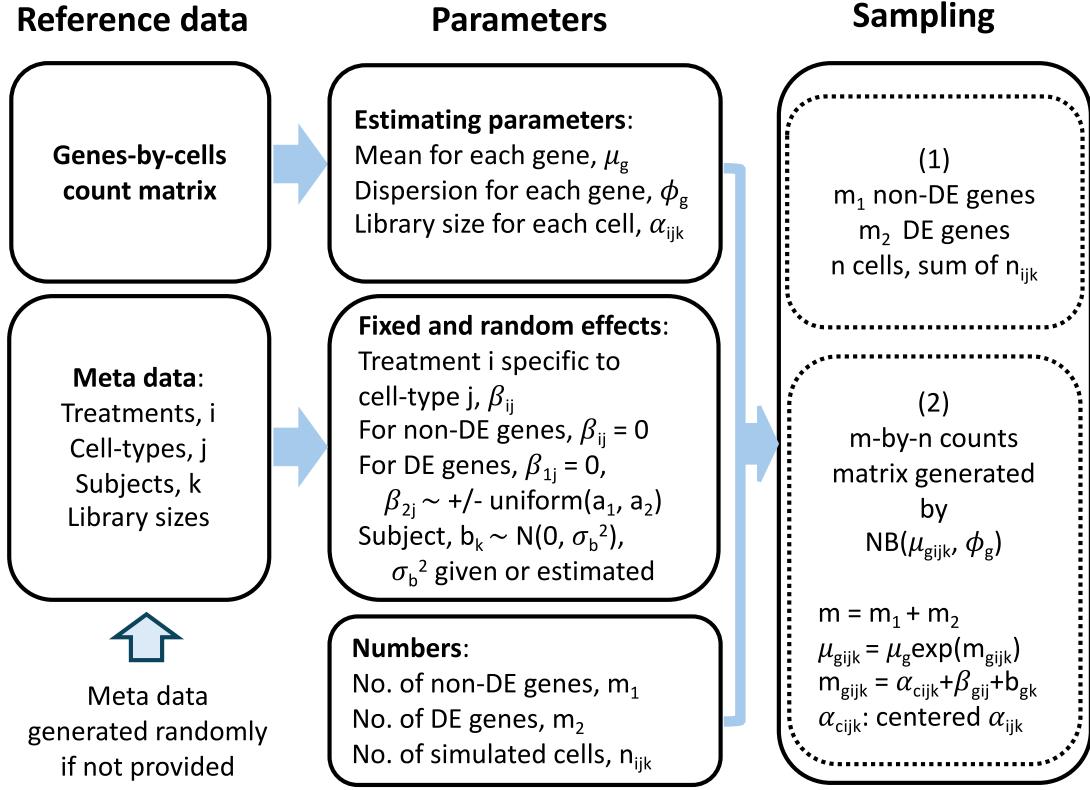


Figure S1: Workflow of the scRNA-seq simulator, simuRNAseq.

Let $y_{g,ijk}$ be the count for gene g and the cell from subject $k = 1, \dots, n_s$ and cell-type $j = 1, \dots, n_c$ with treatment $i = 1, 2$. The count is generated by the NB distribution with dispersion ϕ_g and mean $\mu_{g,ijk}$,

$$y_{g,ijk} \sim NB(\mu_{g,ijk}, \phi_g), \quad (46)$$

where

$$\log(\mu_{g,ijk}) = \log(\mu_g) + \alpha_{ijk} + \beta_{g,ij} + b_{g,k},$$

ϕ_g and μ_g are the dispersion and mean estimated from the reference data for gene g , α_{ijk} is a centered log-library-size for the cell c_{ijk} , $\beta_{g,ij}$ represents the effect of treatment i specific to cell-type j , $b_{g,k} \sim N(0, \sigma_b^2)$ is a random effect of subject variation. For the non-DE genes, $\beta_{g,1j} = \beta_{g,2j} = 0$. For the DE genes, $\beta_{g,1j} = 0$ and $\beta_{g,2j} \sim \pm Uniform(a_1, a_2)$, a uniform distribution with $a_1 a_2 > 0$, where a_1 and a_2 are the lower and upper bounds of the DE gene effect sizes.

The mean of NB distribution is taken as the sample mean for each gene. The dispersion of NB distribution is computed by the method-of-moments estimate (MME) (Clark and Perry 1989), which is computationally more simple and performs reasonably well compared to the maximum likelihood estimate (MLE). The simuRNAseq simulator shares similarities with muscat (Crowell et al. 2020) and GLMsim (Wang et al. 2024). However, both muscat and GLMsim have limitations in scRNA-seq simulations. Muscat estimates the dispersion of the negative binomial (NB) distribution using the edgeR package based on a subset of the reference data. Muscat cannot be applied to large-scale scRNA-seq data and only reflects partial information from the reference dataset. GLMsim estimates the coefficients and dispersion parameters of the NB models for each gene using glm.nb from the MASS package, which is time-consuming and only generates data of a fixed size matching the reference data. In contrast, The simuRNAseq simulator is fast and flexible, and utilizes the whole information from the reference dataset.

We used the PBMC scRNA-seq data as a reference to simulate the scRNA-seq data. The PBMC data from eight lupus patients (Kang et al. 2018), available through Bioconductor's ExperimentHub package, contains

35,635 genes and 29,065 cells (14,619 control cells and 14,446 stimulated cells) consisting of eight identified cell types. After quality control filtering (filtering the cells with few or many detected genes and the genes lowly expressed), the data contains 7,040 genes and 26,820 cells.

3.1 Performance of simuRNAseq simulator

We examined the performance of simuRNAseq simulator by comparing dispersions estimated by MME and MLE, means and variances across cells and library sizes of real reference data and simulated data using scatterplots. The simulated scRNA-seq data is generated by simuRNAseq with the same genes and cells as the reference data.

It is shown that the dispersions estimated by MME and MLE methods are coincident, see Supplemental Figure S2 (a). The MME method took only 0.48 seconds while the MLE method needed about 51.5 minutes to estimate the dispersions. This verified that the MME method is fast and performs very well. The means and variances of counts across cells and library sizes of the real and simulated data are shown in Supplemental Figure S2 (b), (c) and (d), respectively. The scatterplots display the similarity between the real and simulated data.

3.2 Differential expression analysis of simulated scRNA-seq data

We generated scRNA-seq datasets consisting of 6,000 genes with 6 different numbers of cells (sample sizes) from 20,000 to 120,000 using the PBMC data as a reference, as described above. The genes to be simulated were randomly selected from the reference data. We randomly generated the meta data comprising 25 subjects and 12 cell-types which were treated by one of two treatments. The treatments, cell-types and subjects are assigned randomly with equal probability. There are 480 DE genes specific to a cell-type. For the DE genes $b_{g,2j} \sim \pm \text{Uniform}(0.25, 1)$. The variance component $\sigma_b^2 = 0.16$.

We did differential gene expression analysis for the 6 simulated scRNA-seq datasets using LMM with model formula:

$$\sim \log(\text{library.size}) + \text{cell.type} + \text{cell.type : treatment} + (1|\text{subject}). \quad (47)$$

The interaction term $\text{cell.type : treatment}$ in the model formula (47) represents treatment effects in a specific cell-type. The last term represent random effects of subject variation. We fit the LMM to the log-transformed counts, $\log_2(1 + y)$, by FLASH-MM method. Note that FLASH-MM doesn't directly use model formula as an argument and need design matrices of fixed and random effects that can be created by, respectively,

$$X = \text{model.matrix}(\sim \log(\text{library.size}) + \text{cell.type} + \text{cell.type : treatment}),$$

$$Z = \text{model.matrix}(\sim 0 + \text{subject}).$$

Table S1: Computation time in minutes for running FLASH-MM, lmer and nebula in differential expression analysis of the simulated data across various sample sizes n .

	n=20000	n=40000	n=60000	n=80000	n=100000	n=120000
FLASH-MM	0.46	0.51	0.67	0.67	0.88	1.04
lmer	25.15	49.70	74.90	94.02	122.44	149.12
nebula	39.71	86.45	140.95	205.98	274.87	331.89

References

- Clark, Suzanne J., and Joe N. Perry. 1989. “Estimation of the Negative Binomial Parameter k by Maximum Quasi-Likelihood.” *Biometrics* 45 (1). <https://doi.org/10.2307/2532055>.
- Crowell, Helena L., Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, Dheeraj Malhotra, and Mark D. Robinson. 2020. “Muscat Detects Subpopulation-Specific State Transitions from Multi-Sample Multi-Condition Single-Cell Transcriptomics Data.” *Nature Communications* 11. <https://doi.org/10.1038/s41467-020-19894-4>.

- Hartley, H. O., and J. N. K. Rao. 1967. "Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model." *Biometrika* 54 (1/2): 93–108. <http://www.jstor.org/stable/2333854>.
- Harville, David A. 1977. "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems." *Journal of the American Statistical Association* 72 (358): 320–38. <https://doi.org/10.1080/01621459.1977.10480998>.
- . 1991. "That BLUP Is a Good Thing: The Estimation of Random Effects: Comment." *Statistical Science* 6 (1): 35–39. <http://www.jstor.org/stable/2245697>.
- Jiang, Jiming. 2007. *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer. <https://doi.org/10.1007/978-0-387-47946-0>.
- Kang, Hyun Min, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, et al. 2018. "Multiplexed Droplet Single-Cell RNA-Sequencing Using Natural Genetic Variation." *Nature Biotechnology* 36 (89–94). <https://doi.org/10.1038/nbt.4042>.
- Laird, Nan M., and James H. Ware. 1982. "Random-Effects Models for Longitudinal Data." *Biometrics* 38 (4): 963–74. <http://www.jstor.org/stable/2529876>.
- Patterson, H. D., and R. Thompson. 1971. "Recovery of Inter-Block Information When Block Sizes Are Unequal." *Biometrika* 58 (3): 545–54. <https://doi.org/10.2307/2334389>.
- Rao, C. Radhakrishna. 1979. "Minqe Theory and Its Relation to ML and MML Estimation of Variance Components." *The Indian Journal of Statistics, Series B (1960-2002)* 41: 138–53.
- Searle, Shayle R., George Casella, and Charles E. McCulloch. 2006. *Variance Components*. New Jersey: John Wiley & Sons, Inc.
- Wang, Jianan, Lizhong Chen, Rachel Thijssen, Belinda Phipson, and Terence P. Speed. 2024. "GLMsim: A GLM-Based Single Cell RNA-Seq Simulator Incorporating Batch and Biological Effects." *bioRxiv*. <https://doi.org/10.1101/2024.03.20.586030>.

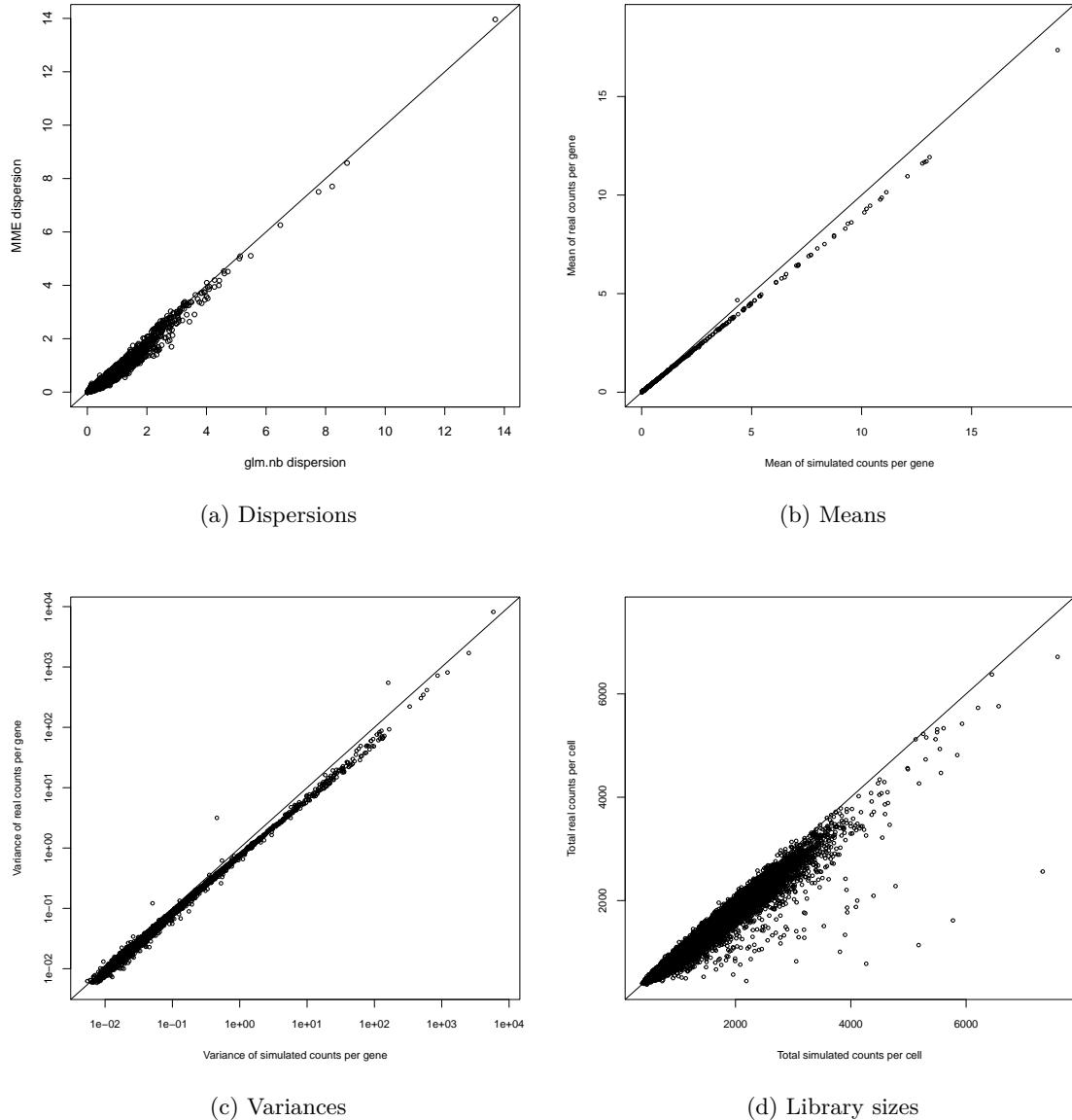


Figure S2: Performance of the scRNA-seq simulator, simuRNAseq: (a) Scatterplot of dispersions estimated by MME and MLE (`glm.nb`). (b) Scatterplot of means of real and simulated counts across cells. (c) Scatterplot of variances of real and simulated counts across cells. (d) Scatterplot of library sizes of real and simulated counts.