

Efficient and accurate mixed model association tool for single-cell eQTL analysis Supplementary Note

1 Notations of the SAIGE-QTL method

Assuming the data contain read counts from RNA sequencing mapped to a gene for N cells across n individuals, c_i cells belong to the i th individual and $N = \sum_{i=1}^n c_i$. The sequencing read count for the gene of interest, y_{ij} , of the j th cell from the i th individual is assumed to follow $\text{Poisson}(\mu_{ij})$, where $\mu_{ij} = \exp(\eta_{ij})$ and $\eta_{ij} = X_{cij}\alpha_c + X_{di}\alpha_d + G_i\beta + b_i$. Let G_i represent the allele count (0,1, or 2) of the i th individual for the genetic variant to test, X_{cij} represent p_c cell-level covariates, such as cell states, including the intercept for the j th cell of the i th individual, and X_{di} represent p_d individual (donor)-level covariates, such as age, gender, and ancestry PCs, for the i th individual. The random effects \mathbf{b}_d is a length n vector of b_i and is assumed to follow multivariate Normal distribution $b_i \sim N(0, \sum_{k=1}^K \tau_k \Psi_k)$, where τ_k are the variance component parameters and Ψ_k are known $n \times n$ variance-covariance matrices, which can include the identity matrix to account for intra-individual variance of read counts across multiple cells and the genetic relationship matrix to account for sample relatedness, if any.

The read counts y_{ij} are assumed to be independent conditional on the random effect $(X_{cij}, X_{di}, G_i, b_i)$ and follows the Poisson distribution with mean an variance $E(y_{ij}|X_{cij}, X_{di}, G_i, b_i) = \mu_{ij}$ and $\text{Var}(y_{ij}|X_{cij}, X_{di}, G_i, b_i) = \phi v(\mu_{ij})$, where $v(\mu_{ij}) = \mu_{ij}$ is the variance function and the dispersion parameter $\phi = 1$.

Let \mathbf{Z} be the $N \times n$ design matrix. Each row in \mathbf{Z} contains a 1 and all other elements are 0s to indicate which individual has the cell. The Poisson mixed model can be written as

$$\boldsymbol{\eta} = \mathbf{X}_c \boldsymbol{\alpha}_c + \mathbf{Z} \mathbf{X}_d \boldsymbol{\alpha}_d + \mathbf{Z} \mathbf{G} \beta + \mathbf{b}$$

, where $\boldsymbol{\eta}$ is a length of N vector of η_{ij} , \mathbf{X}_c is a $N \times p_c + 1$ matrix containing cell-level covariates including the intercept, \mathbf{X}_d is a $n \times p_d$ matrix containing individual-level covariates, and the random effects for all cells $\mathbf{b} = \mathbf{Z} \mathbf{b}_d$ is a vector with length of N . $\mathbf{b} \sim N(0, \sum_{k=1}^K \tau_k \mathbf{Z} \Psi_k \mathbf{Z}^T)$ accounts for all sources of correlation in the data.

1.1 Poisson generalized linear mixed model and penalized quasi-likelihood

Similar algorithms that were used in previous GWAS methods [1, 2, 3] are used in SAIGE-QTL to fit the Poisson generalized linear mixed model under the null hypothesis $H_0 : \beta = 0$. To estimate

$(\boldsymbol{\alpha}, \phi, \boldsymbol{\tau})$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_c, \boldsymbol{\alpha}_d)$, the integrated quasi-likelihood function can be written as

$$\begin{aligned} L(\boldsymbol{\alpha}, \beta = 0, \phi, \boldsymbol{\tau}) &= \int \exp\left\{\sum_{i=1}^n \sum_{j=1}^{c_i} ql_{ij}(\boldsymbol{\alpha}, \beta = 0|\mathbf{b})\right\} \\ &\quad \times (2\pi)^{-\frac{N}{2}} \frac{1}{\left|\sum_{k=1}^K \tau_k \mathbf{V}_K\right|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{b}^\top \left(\sum_{k=1}^K \tau_k \mathbf{V}_K\right)^{-1} \mathbf{b}\right) d\mathbf{b} \\ &= (2\pi)^{-\frac{N}{2}} \frac{1}{\left|\sum_{k=1}^K \tau_k \mathbf{V}_K\right|^{1/2}} \\ &\quad \times \int \exp\left(\sum_{i=1}^n \sum_{j=1}^{c_i} ql_{ij}(\boldsymbol{\alpha}, \beta = 0|\mathbf{b}) - \left(\frac{1}{2} \mathbf{b}^\top \left(\sum_{k=1}^K \tau_k \mathbf{V}_K\right)^{-1} \mathbf{b}\right)\right) d\mathbf{b} \end{aligned}$$

, where $\mathbf{V}_k = \mathbf{Z}\boldsymbol{\Psi}_k\mathbf{Z}^\top$

$$ql_{ij}(\boldsymbol{\alpha}, \beta = 0|\mathbf{b}) = \int_{y_{ij}}^{\mu_{ij}} \frac{(y_{ij} - \mu)}{\phi v(\mu)} d(\mu) = \int_{y_{ij}}^{\mu_{ij}} \frac{y_{ij} - \mu}{\mu} d(\mu)$$

is the quasi-likelihood for the j th cell from the i th individual given the random effect \mathbf{b} .

Let $f(\mathbf{b}) = \sum_{i=1}^n \sum_{j=1}^{c_i} ql_{ij}(\boldsymbol{\alpha}, \beta = 0|\mathbf{b}) - \frac{1}{2} \mathbf{b}^\top \left(\sum_{k=1}^K \tau_k \mathbf{V}_K\right)^{-1} \mathbf{b}$, we can then approximate the integral using the Laplace approximation,

$$\int \exp f(\mathbf{b}) d\mathbf{b} \approx (2\pi)^{N/2} \left| -f''(\tilde{\mathbf{b}}) \right|^{-1/2} \exp(f(\tilde{\mathbf{b}})),$$

, where $\tilde{\mathbf{b}} = \arg_b \max f(\mathbf{b})$ is the solution to $f'(\mathbf{b}) = 0$. Therefore, the log integrated quasi-likelihood function can be approximated by

$$\begin{aligned} f'(\mathbf{b}) &= \frac{\partial \sum_{i=1}^n \sum_{j=1}^{c_i} ql_{ij}(\boldsymbol{\alpha}, \beta = 0|\mathbf{b})}{\partial \mathbf{b}} = \sum_{i=1}^n \sum_{j=1}^{c_i} \frac{\partial ql_{ij}(\boldsymbol{\alpha}, \beta = 0|\mathbf{b})}{\partial \mathbf{b}} \\ \frac{\partial^2 ql_{ij}(\boldsymbol{\alpha}, \beta = 0|\mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}^\top} &= -\frac{\frac{\partial \mu_{ij}}{\partial \eta_{ij}} \frac{\partial \eta_{ij}}{\partial \mathbf{b}_{ij}} \mathbf{Z}^\top}{-\mu_{ij} \mathbf{Z}_{ij} \mathbf{Z}_{ij}^\top} \\ f''(\tilde{\mathbf{b}}) &= -\left(\sum_{i=1}^n \sum_{j=1}^{c_i} \mu_{ij} \mathbf{Z}_{ij} \mathbf{Z}_{ij}^\top\right) - \left(\sum_{k=1}^K \tau_k \mathbf{V}_K\right)^{-1} \\ ql(\boldsymbol{\alpha}, \beta = 0, \phi, \boldsymbol{\tau}) &\approx -\frac{1}{2} \log \left| \sum_{k=1}^K \tau_k \mathbf{V}_K \right| - \frac{1}{2} \log \left| \left(\sum_{i=1}^n \sum_{j=1}^{c_i} \mu_{ij} \mathbf{Z}_{ij} \mathbf{Z}_{ij}^\top\right) + \left(\sum_{k=1}^K \tau_k \mathbf{V}_K\right)^{-1} \right| \\ &\quad + \sum_{i=1}^n \sum_{j=1}^{c_i} ql_{ij}(\boldsymbol{\alpha}, \beta = 0|\tilde{\mathbf{b}}) - \frac{1}{2} \tilde{\mathbf{b}}^\top \left(\sum_{k=1}^K \tau_k \mathbf{V}_K\right)^{-1} \tilde{\mathbf{b}} \\ &= -\frac{1}{2} \log \left| \sum_{k=1}^K \mathbf{Z}(\tau_k \boldsymbol{\Psi}_k \mathbf{W} + I) \mathbf{Z}^\top \right| + \sum_{i=1}^n \sum_{j=1}^{c_i} ql_{ij}(\boldsymbol{\alpha}, \beta = 0|\tilde{\mathbf{b}}) - \frac{1}{2} \tilde{\mathbf{b}}^\top \left(\sum_{k=1}^K \tau_k \mathbf{V}_K\right)^{-1} \tilde{\mathbf{b}} \end{aligned} \tag{1}$$

, where $\mathbf{W} = \text{diag}(\mu)$

1.2 Estimation of the fixed effects and random effects given the variance components

We can obtain the score functions by taking derivatives of $ql(\alpha, \beta = 0, \phi, \tau)$ with respect to β, α , and \mathbf{b} .

$$\begin{aligned}\frac{\partial ql(\alpha, \beta = 0, \phi, \tau)}{\partial \alpha} &= \mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}) \\ \frac{\partial ql(\alpha, \beta = 0, \phi, \tau)}{\partial \beta} &= \mathbf{G}^T(\mathbf{y} - \boldsymbol{\mu}) \\ \frac{\partial ql(\alpha, \beta = 0, \phi, \tau)}{\partial \mathbf{b}} &= (\mathbf{y} - \boldsymbol{\mu}) - \left(\sum_{k=1}^K \tau_k \mathbf{V}_K\right)^{-1} \mathbf{b}\end{aligned}$$

, where $\mathbf{X} = [\mathbf{X}_c \ \mathbf{Z}\mathbf{X}_d]$. For the score test of $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$, we only estimate the $(\hat{\alpha}, \hat{\beta}, \hat{\mathbf{b}})$ that maximize $ql(\alpha, \beta = 0, \phi, \tau)$ under H_0 , setting $\beta = 0$. We denote the working outcome vector $\mathbf{Y} = \boldsymbol{\eta} + \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\mu})$, where $\mathbf{W} = \text{diag}(\mu_i)$, $\boldsymbol{\eta} = (\eta_1, \dots, \eta_N)$, $\boldsymbol{\delta} = (\delta_1, \dots, \delta_N)$, and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$. Then, $\mathbf{y} - \boldsymbol{\mu} = \mathbf{W}(\mathbf{Y} - \boldsymbol{\eta}) = \mathbf{W}(\mathbf{Y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b})$, and the score equations can be written as,

$$\begin{bmatrix} \mathbf{X}^T \mathbf{W} \mathbf{X} & \mathbf{X}^T \mathbf{W} \\ \mathbf{W} \mathbf{X} & \mathbf{W} + \left(\sum_{k=1}^K \tau_k \mathbf{V}_K\right)^{-1} \end{bmatrix} \begin{bmatrix} \alpha \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \mathbf{W} \mathbf{Y} \\ \mathbf{W} \mathbf{Y} \end{bmatrix}$$

Let $\boldsymbol{\Sigma} = \mathbf{W}^{-1} + \sum_{k=1}^K \tau_k \mathbf{V}_K$, then

$$\begin{aligned}\hat{\alpha} &= (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y} \\ \hat{\mathbf{b}} &= \left(\sum_{k=1}^K \tau_k \mathbf{V}_K\right) \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{X}\hat{\alpha}).\end{aligned}$$

1.3 Estimation of the variance component

Given $\hat{\alpha} = \hat{\alpha}(\tau)$, $\hat{\mathbf{b}} = \hat{\mathbf{b}}(\tau)$ estimated, from (1) the log-likelihood of the variance component can be derived as,

$$ql(\hat{\alpha}(\tau, \phi), \beta = 0, \phi, \tau) = c - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \mathbf{Y}^T \mathbf{P} \mathbf{Y},$$

where $\mathbf{P} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \mathbf{X} (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma} = \mathbf{W}^{-1} + \sum_{k=1}^K \tau_k \mathbf{V}_k$. We maximize the corresponding restricted maximum-likelihood (REML),

$$ql_R(\hat{\alpha}(\tau, \phi), \beta = 0, \phi, \tau) = c_R - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \log |\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}| - \frac{1}{2} \mathbf{Y}^T \mathbf{P} \mathbf{Y}.$$

The score function with respect to τ are given by,

$$\mathbf{U}_\tau = \frac{\partial ql_R(\hat{\alpha}(\tau, \phi), \beta = 0, \phi, \tau)}{\partial \tau_k} = \frac{1}{2} [\mathbf{Y}^T \mathbf{P} \mathbf{Z} \boldsymbol{\Psi}_k \mathbf{Z}^T \mathbf{P} \mathbf{Y} - \text{tr}(\mathbf{P} \mathbf{Z} \boldsymbol{\Psi}_k \mathbf{Z}^T)].$$

The corresponding observed information function, and the expected information function are given by

$$\begin{aligned}\mathbf{J}_\tau &= -\frac{\partial^2 ql_R(\hat{\alpha}(\tau, \phi), \beta = 0, \phi, \tau)}{\partial \tau_k^2} = -\frac{1}{2} \text{tr}(\mathbf{P} \mathbf{V}_k \mathbf{P} \mathbf{V}_k) + \mathbf{Y}^T \mathbf{P} \mathbf{V}_k \mathbf{P} \mathbf{V}_k \mathbf{P} \mathbf{Y}, \\ E(\mathbf{J}_\tau) &= E \left[-\frac{\partial^2 ql_R(\hat{\alpha}(\tau, \phi), \beta = 0, \phi, \tau)}{\partial \tau_k^2} \right] = \frac{1}{2} \text{tr}(\mathbf{P} \mathbf{V}_k \mathbf{P} \mathbf{V}_k),\end{aligned}$$

respectively. Evaluating both observed and expected information functions involves computationally expensive trace computations. To avoid the trace computations, the average information is used in the AI-REML[1, 2, 4] algorithm. The average information is expressed as the average of \mathbf{J}_τ and $E(\mathbf{J}_\tau)$,

$$\mathbf{AI}_{\tau_k \tau_q} = \frac{1}{2} \mathbf{Y}^T \mathbf{P} \mathbf{V}_k \mathbf{P} \mathbf{V}_q \mathbf{P} \mathbf{Y}.$$

, where \mathbf{AI} is an $K \times K$ matrix with the (k, q) -th element $\mathbf{AI}_{\tau_k \tau_q}$.

1.3.1 Algorithm to fit the null mixed model

The null model fitting algorithm can be summarized as,

1. Fit a Poisson linear model with $\tau = 0$ to get initial estimates $\hat{\boldsymbol{\alpha}}^{(0)}$ and working outcome vector $\mathbf{Y}^{(0)}$.
2. At the i -th step, update $\hat{\boldsymbol{\tau}}$ using $\hat{\boldsymbol{\tau}}^{(i)} = \hat{\boldsymbol{\tau}}^{(i-1)} + \left\{ \mathbf{A}\mathbf{I} \Big|_{\tau=\hat{\boldsymbol{\tau}}^{(i-1)}} \right\}^{-1} \left\{ \mathbf{U} \Big|_{\tau=\hat{\boldsymbol{\tau}}^{(i-1)}} \right\}$.
3. Update $\hat{\boldsymbol{\alpha}}, \hat{\mathbf{b}}$ using \mathbf{Y} and $\hat{\boldsymbol{\tau}}^{(i)}$.
4. Update \mathbf{Y} using $\hat{\mathbf{b}}^{(i)}, \hat{\boldsymbol{\alpha}}^{(i)}, \hat{\boldsymbol{\tau}}^{(i)}$.
5. Repeat steps 2–4, until $\max \left\{ \frac{|\hat{\beta}^{(i)} - \hat{\beta}^{(i-1)}|}{|\hat{\beta}^{(i)}| + |\hat{\beta}^{(i-1)}|}, \frac{|\hat{\tau}^{(i)} - \hat{\tau}^{(i-1)}|}{|\hat{\tau}^{(i)}| + |\hat{\tau}^{(i-1)}|} \right\} \leq \textit{tolerance}$.

1.4 Score test

The score test statistic under the null hypothesis is given by,

$$T = \frac{\partial q_l}{\partial \beta} \Big|_{(\hat{b}, \beta=0, \hat{\alpha}, \hat{\tau})} = \mathbf{G}^\top \mathbf{Z}^\top (\mathbf{y} - \hat{\boldsymbol{\mu}}) = \tilde{\mathbf{G}}^\top (\mathbf{y} - \hat{\boldsymbol{\mu}}), \quad (2)$$

where $\tilde{\mathbf{G}} = \mathbf{Z}\mathbf{G} - \mathbf{X} \left(\mathbf{X}^\top \hat{\mathbf{W}}\mathbf{X} \right)^{-1} \mathbf{X}^\top \hat{\mathbf{W}}\mathbf{Z}\mathbf{G}$ is the covariate adjusted genotype vector, and $\mathbf{X} = [\mathbf{X}_c \ \mathbf{Z}\mathbf{X}_d]$ is the covariate matrix with intercept included. The information matrix corresponding to the score equations in (2) is given by,

$$\mathbf{I}(\alpha, \beta, b) = \begin{bmatrix} \mathbf{X}^\top \mathbf{W}\mathbf{X} & \mathbf{X}^\top \mathbf{W}\mathbf{G} & \mathbf{X}^\top \mathbf{W} \\ \mathbf{G}^\top \mathbf{W}\mathbf{X} & \mathbf{G}^\top \mathbf{W}\mathbf{G} & \mathbf{G}^\top \mathbf{W} \\ \mathbf{W}\mathbf{X} & \mathbf{Z}^\top \mathbf{W}\mathbf{G} & \mathbf{W} + \left(\sum_{k=1}^K \tau_k \mathbf{V}_k \right)^{-1} \end{bmatrix},$$

A_{ij} and $\mathbf{A}_{\cdot i}$ denote the (i, j) -th element and i -th column of a matrix A , respectively. Then, the variance of the score statistic under H_0 is given by,

$$\text{Var}_{H_0}(T) = \left(\mathbf{I}(\hat{\alpha}, \hat{\beta}, \hat{b}) \right)^{-1}_{22} = \mathbf{G}^\top \hat{\mathbf{P}}\mathbf{G} = \tilde{\mathbf{G}}^\top \hat{\mathbf{P}}\tilde{\mathbf{G}},$$

where $\hat{\mathbf{P}} = \hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X} \left(\mathbf{X}^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X} \right)^{-1} \mathbf{X}^\top \hat{\boldsymbol{\Sigma}}^{-1}$, $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{W}}^{-1} + \sum_{k=1}^K \hat{\tau}_k \mathbf{V}_k$, and $\mathbf{V}_k = \mathbf{Z}\boldsymbol{\Psi}_k\mathbf{Z}^\top$.

2 Variance ratio approximation

Computation of the variance of the score statistic $\text{Var}_{H_0}(T) = \tilde{\mathbf{G}}^\top \hat{\mathbf{P}}\tilde{\mathbf{G}}$ requires calculating $\hat{\mathbf{P}}\tilde{\mathbf{G}}$ repeatedly for all markers, which is computationally expensive. To avoid calculating $\hat{\mathbf{P}}\tilde{\mathbf{G}}$ for all the markers, previous linear mixed model [5, 6, 7] and logistic mixed model methods [2] first estimate the variance ratio $\hat{r} = \tilde{\mathbf{G}}^\top \hat{\mathbf{P}}\tilde{\mathbf{G}} / \tilde{\mathbf{G}}^\top \hat{\mathbf{W}}\tilde{\mathbf{G}}$ using a small set of markers, and then approximate the variance of the score statistic for all markers by $\hat{r}\tilde{\mathbf{G}}^\top \hat{\mathbf{W}}\tilde{\mathbf{G}}$ in step 2. It has been shown in the studies that \hat{r} is approximately constant for all genetic variants. This saves substantial computation time since $\hat{\mathbf{W}}$ is a diagonal matrix. Following the Computation of the variance of the score statistic $\text{Var}_{H_0}(T) = \tilde{\mathbf{G}}^\top \hat{\mathbf{P}}\tilde{\mathbf{G}}$ requires calculating $\hat{\mathbf{P}}\tilde{\mathbf{G}}$ repeatedly for all markers, which is computationally expensive. To avoid calculating $\hat{\mathbf{P}}\tilde{\mathbf{G}}$ for all the markers, we estimate the variance ratio $\hat{r} = \tilde{\mathbf{G}}^\top \hat{\mathbf{P}}\tilde{\mathbf{G}} / \tilde{\mathbf{G}}^\top \hat{\mathbf{W}}\tilde{\mathbf{G}}$ using a small set of markers, and then approximate the variance of the score statistic for all markers by $\hat{r}\tilde{\mathbf{G}}^\top \hat{\mathbf{W}}\tilde{\mathbf{G}}$ in step 2. This saves substantial computation time since $\hat{\mathbf{W}}$ is a diagonal matrix. Such variance ratio approximation approaches were previously used in various linear[5, 6, 7] and logistic mixed effects models[2] to speed up computation. Following the theoretical justification by Dey et al.[3] to show the approximation also works in the frailty models, we can show that it also works in the Poisson mixed model for eQTL mapping with scRNA-seq data.

Let $E(G_i) = \mu_g$ and the covariance matrix of \mathbf{G} is given by $\sigma_g^2 \Psi$, where Ψ is the correlation matrix of \mathbf{G} and represents the $n \times n$ kinship matrix. The exact characterization of Ψ is not needed for the proof. Then the elements of \mathbf{G} follows the $Bin(2, p_g)$ distribution, then $\mu_g = 2p_g$, and $\sigma_g^2 = 2p_g(1-p_g)$. Let \mathbf{G}_c be the $n \times 1$ unadjusted (but mean-centered) genotype vector. Then $E(\mathbf{G}_c) = 0$ and $Cov(\mathbf{G}_c) = \sigma_g^2 \Psi$. Let $\mathbf{Q}_{\tilde{\mathbf{X}}} = \mathbf{X}(\mathbf{X}^\top \hat{\mathbf{W}}\mathbf{X})^{-1} \mathbf{X}^\top \hat{\mathbf{W}}$ be the weighted projection matrix. Then, $E(\tilde{\mathbf{G}}) = (\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}})\mathbf{Z}\mu_g\mathbf{1} = 0$, and $Cov(\tilde{\mathbf{G}}) = \sigma_g^2(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}})\mathbf{Z}\Psi\mathbf{Z}^\top(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}})^\top$, where $\mathbf{1}$ is the $n \times 1$ vector of all element equal to unity. We scale both the numerator and denominator of the variance ratio by N^{-1} so that they don't blow to infinity when looked at individually. Then, for the numerator,

$$E(N^{-1}\tilde{\mathbf{G}}^\top \hat{\mathbf{P}}\tilde{\mathbf{G}}) = \frac{\sigma_g^2}{N} tr \left[\hat{\mathbf{P}}(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}})\mathbf{Z}\Psi\mathbf{Z}^\top(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}})^\top \right] = \frac{\sigma_g^2}{N} tr(\hat{\mathbf{P}}\mathbf{Z}\Psi\mathbf{Z}^\top),$$

since $(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}})^\top \hat{\mathbf{P}}(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}}) = \hat{\mathbf{P}}$. Similarly, for the denominator,

$$E(N^{-1}\mathbf{G}_c^\top \mathbf{Z}^\top \hat{\mathbf{W}}\mathbf{Z}\mathbf{G}_c) = \frac{\sigma_g^2}{N} tr(\mathbf{Z}^\top \hat{\mathbf{W}}\mathbf{Z}\Psi),$$

As the eigenvalues of $\hat{\mathbf{P}}, \hat{\mathbf{W}}, \Psi$ are bounded, and the distribution of \mathbf{G} has bounded support, the variances of the numerator and the denominator terms are both $O(N^{-1})$, and the variance ratio converges to,

$$\hat{r} = \frac{\tilde{\mathbf{G}}^\top \hat{\mathbf{P}}\tilde{\mathbf{G}}}{\mathbf{G}_c^\top \mathbf{Z}^\top \hat{\mathbf{W}}\mathbf{Z}\mathbf{G}_c} \xrightarrow{p} \frac{\lim_{N \rightarrow \infty} \left\{ N^{-1} tr(\hat{\mathbf{P}}\mathbf{Z}\Psi\mathbf{Z}^\top) \right\}}{\lim_{N \rightarrow \infty} \left\{ N^{-1} tr(\mathbf{Z}^\top \hat{\mathbf{W}}\mathbf{Z}\Psi) \right\}}.$$

The ratio on the right-hand side is constant across all markers as the individual limits in the numerator and denominator exist and are bounded away from zero.

In addition, as we have previously shown in SAIGE-GENE[8], the variation of the estimated variance ratio can be smaller when $\hat{\Sigma}^{-1}$ is used in the denominator than when only $\hat{\mathbf{W}}$ is used. Therefore, in SAIGE-QTL, we incorporate $\hat{\Sigma}^{-1}$ and the covariate adjusted genotype vector into the denominator to approximate the $Var_{H_0}(T)$ with a smaller variation. More specifically, the variance ratio $\hat{r}_s = \tilde{\mathbf{G}}^\top \hat{\mathbf{P}}\tilde{\mathbf{G}} / \tilde{\mathbf{G}}_d^\top \hat{\Sigma}^{-1} \tilde{\mathbf{G}}_d$ is estimated. Let $\tilde{\mathbf{X}}_d = [\mathbf{1} \quad \mathbf{X}_d]$ represent the donor-level covariate matrix with intercept included with the size $n \times (p_d + 1)$, then $\tilde{\mathbf{G}}_d$ is the donor-level covariate adjusted genotype vector with size $N \times 1$ and

$$\begin{aligned} \tilde{\mathbf{G}}_d &= \mathbf{Z}\mathbf{G} - \mathbf{Z}\tilde{\mathbf{X}}_d \left(\tilde{\mathbf{X}}_d^\top \mathbf{Z}^\top \hat{\mathbf{W}}\mathbf{Z}\tilde{\mathbf{X}}_d \right)^{-1} \tilde{\mathbf{X}}_d^\top \mathbf{Z}^\top \hat{\mathbf{W}}\mathbf{Z}\mathbf{G} \\ &= \mathbf{Z}(\mathbf{G} - \tilde{\mathbf{X}}_d \left(\tilde{\mathbf{X}}_d^\top \mathbf{Z}^\top \hat{\mathbf{W}}\mathbf{Z}\tilde{\mathbf{X}}_d \right)^{-1} \tilde{\mathbf{X}}_d^\top \mathbf{Z}^\top \hat{\mathbf{W}}\mathbf{Z}\mathbf{G}) \end{aligned}$$

. Similarly, Let $\mathbf{Q}_{\tilde{\mathbf{X}}_d} = \tilde{\mathbf{X}}_d \left(\tilde{\mathbf{X}}_d^\top \mathbf{Z}^\top \hat{\mathbf{W}}\mathbf{Z}\tilde{\mathbf{X}}_d \right)^{-1} \tilde{\mathbf{X}}_d^\top \mathbf{Z}^\top \hat{\mathbf{W}}\mathbf{Z}$ be the weighted projection matrix to project out the donor-level covariates with intercept, then $\tilde{\mathbf{G}}_d = \mathbf{Z}(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}_d})\mathbf{G}$, $E(\tilde{\mathbf{G}}_d) = \mathbf{Z}(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}_d})\mu_g\mathbf{1} = 0$, and $Cov(\tilde{\mathbf{G}}_d) = \sigma_g^2 \mathbf{Z}(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}_d})\Psi(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}_d})^\top \mathbf{Z}^\top$.

For the denominator,

$$E(N^{-1}\tilde{\mathbf{G}}_d^\top \hat{\Sigma}^{-1} \tilde{\mathbf{G}}_d) = \frac{\sigma_g^2}{N} tr \left(\hat{\Sigma}^{-1} \mathbf{Z}(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}_d})\Psi(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}_d})^\top \mathbf{Z}^\top \right),$$

Similarly, as the eigenvalues of $\hat{\mathbf{P}}, \hat{\mathbf{W}}, \Psi, \hat{\Sigma}^{-1}$ are bounded, and the distribution of \mathbf{G} has bounded support, the variances of the numerator and the denominator terms are both $O(N^{-1})$, and the variance ratio converges to,

$$\hat{r}_s = \frac{\tilde{\mathbf{G}}^\top \hat{\mathbf{P}}\tilde{\mathbf{G}}}{\tilde{\mathbf{G}}_d^\top \hat{\Sigma}^{-1} \tilde{\mathbf{G}}_d} \xrightarrow{p} \frac{\lim_{N \rightarrow \infty} \left\{ N^{-1} tr(\hat{\mathbf{P}}\mathbf{Z}\Psi\mathbf{Z}^\top) \right\}}{\lim_{N \rightarrow \infty} \left\{ N^{-1} tr(\hat{\Sigma}^{-1} \mathbf{Z}(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}_d})\Psi(\mathbf{I} - \mathbf{Q}_{\tilde{\mathbf{X}}_d})^\top \mathbf{Z}^\top) \right\}}.$$

, which is constant across all markers as the individual limits in the numerator and denominator exist and are bounded away from zero.

For each genetic variant, \hat{r} is first used to approximate the variance $Var_{H_0}(T)$, if the p-value ≤ 0.05 , \hat{r}_s is used to obtain a more accurate $Var_{H_0}(T)$.

3 Using saddlepoint approximation[9] (SPA) for the null distribution of the score statistic

In traditional score tests, the distribution of the score statistic under H_0 is approximated by a Normal distribution, which uses the first two moments, mean and variance. This approach can perform poorly in the tail regions, especially if the underlying distribution is highly skewed when the studying event is very rare or the testing genetic variant has a very low minor allele count (MAC). Here, similar to what has been applied in the logistic mixed models [10, 11, 2, 12] previously, we use the SPA to approximate the null distribution of the score statistic to obtain accurate p-values. For this, we utilize the fact that given the random effects \mathbf{b} , the phenotype Y , which are read counts from RNA sequencing, independently follow Poisson distribution. $T_{adj} = \tilde{\mathbf{G}}^\top (\mathbf{y} - \hat{\boldsymbol{\mu}}) / \sqrt{\hat{r} \tilde{\mathbf{G}}^\top \hat{\mathbf{W}} \tilde{\mathbf{G}}}$ is a weighted sum of independent Poisson random variable. The approximated cumulant generating function (CGF) of T_{adj} is

$$K(\xi; \hat{\boldsymbol{\mu}}, c) = \sum_{i=1}^N \hat{\mu}_i (e^{\tilde{G}_i c \xi} - \tilde{G}_i c \xi - 1)$$

, where $c = (\tilde{\mathbf{G}}^\top \mathbf{W} \tilde{\mathbf{G}})^{-1/2}$. $K'(\xi; \hat{\boldsymbol{\mu}}, c)$ and $K''(\xi; \hat{\boldsymbol{\mu}}, c)$ are first and second derivatives of K with respect to ξ .

$$K'(\xi; \hat{\boldsymbol{\mu}}, c) = \sum_{i=1}^N \hat{\mu}_i \tilde{G}_i c (e^{\tilde{G}_i c \xi} - 1),$$

$$K''(\xi; \hat{\boldsymbol{\mu}}, c) = \sum_{i=1}^N \hat{\mu}_i \tilde{G}_i^2 c^2 e^{\tilde{G}_i c \xi}$$

To calculate the probability that $T_{adj} < s$, where s is the observed test statistic, we use the following formula

$$Pr(T < s) = \Phi \left\{ w + \frac{1}{w} \log \left(\frac{v}{w} \right) \right\},$$

where $w = \text{sign}(\hat{\xi}) [2\{\hat{\xi}s - K(\hat{\xi})\}]^{1/2}$, $v = \hat{\xi} [K''(\hat{\xi})]^{1/2}$, $\hat{\xi}$ is the solution of $K'(\hat{\xi}) = s$, and Φ is the standard normal distribution function.

4 Approaches to reduce computation and memory cost

We developed SAIGE-QTL based on the SAIGE framework, which utilizes several state-of-the-art approaches to reduce computation and memory cost to fit the null GLMM and to test for genetic associations. Given that in the single-cell RNA sequencing data, each individual may have hundreds and thousands of cells ($N \gg n$), which N is the total number of cells and n is the total number of individuals, we use several approaches to make SAIGE-QTL conduct matrix operations on the scale of number of individuals to reduce a computation cost from $O(N)$ to $O(n)$.

First, to compute the test statistics $\mathbf{T} = \mathbf{G}^\top \mathbf{Z}^\top (\mathbf{y} - \hat{\boldsymbol{\mu}})$ for each genetic variant, we pre-compute the statistics $\mathbf{Z}^\top (\mathbf{y} - \hat{\boldsymbol{\mu}})$, so the computation cost of obtaining \mathbf{T} is $O(n)$.

Second, to calculate $Var_{H_0}(T) = \tilde{\mathbf{G}}^\top \hat{\mathbf{P}} \tilde{\mathbf{G}}$ when estimating the variance ratios using randomly selected genetic markers, we need to compute quantities of the form $\hat{\boldsymbol{\Sigma}}^{-1} \mathbf{a}$, where \mathbf{a} is a vector of length N . The standard computation technique of inverting $\hat{\boldsymbol{\Sigma}}$ (computation cost $O(N^3)$) and multiplying $\hat{\boldsymbol{\Sigma}}^{-1}$ with \mathbf{a} can be extremely time consuming when N is large. To compute quantities of the form $\hat{\boldsymbol{\Sigma}}^{-1} \mathbf{a}$, we implemented the pre-conditioned conjugate gradient[13] (PCG) method, which computes $\hat{\boldsymbol{\Sigma}}^{-1} \mathbf{a} = \mathbf{x}$ by iteratively solving the linear system of equations $\hat{\boldsymbol{\Sigma}} \mathbf{x} = \mathbf{a}$. Only $\hat{\boldsymbol{\Sigma}} \mathbf{a}$ needs to be computed when the PCG method is used. Instead of storing the $N \times N$ variance-covariance matrix \mathbf{V} (memory cost $O(N^2)$), we only store the K matrices $\boldsymbol{\Psi}_{\mathbf{k}}$ with the size $n \times n$ (memory cost $O(kn^2)$) and the $N \times n$ design matrix \mathbf{Z} , which only cost $4N$ bytes if stored using a sparse matrix with 1s. Then, $\hat{\boldsymbol{\Sigma}} \mathbf{a}$ involves the computation for $\mathbf{Z} \boldsymbol{\Psi}_{\mathbf{k}} \mathbf{Z}^\top \mathbf{a}$, which is computed consecutively, with the time cost $O(Nnk)$. As implemented in SAIGE, we use the Hutchinson's randomized trace estimation[14, 15] for calculating $tr(\mathbf{P}\mathbf{V})$.

We also implemented multithreaded parallel computation for the matrix-vector multiplications in the PCG steps using Intel Threading Building Block (TBB) from the RcppParallel[16] package.

Third, as described in , in step 2, the variance of the score statistic for all markers can be approximated by $\hat{r}\mathbf{G}_c^\top\mathbf{Z}^\top\hat{\mathbf{W}}\mathbf{Z}\mathbf{G}_c$ or $\hat{r}_s\tilde{\mathbf{G}}_d^\top\hat{\Sigma}^{-1}\tilde{\mathbf{G}}_d$. We pre-compute both $\mathbf{Z}^\top\hat{\mathbf{W}}\mathbf{Z}$ and $\mathbf{Z}^\top\hat{\Sigma}^{-1}\mathbf{Z}$, so the computation to approximate the variance of the score statistic becomes $O(n)$ and for $\sim 5\%$ genetic variants, it costs $O(n(p_d + 1))$.

Fourth, to reduce the redundancies of reading genotypes for each genetic variant when mapping eQTLs for all 20,000 genes, SAIGE-QTL allows for analyzing multiple genes. We observed that computation time has been dropped dramatically compared to analyzing each gene separately. This is particularly useful for conducting genome-wide trans-eQTL mapping.

References

- [1] Han Chen, Chaolong Wang, Matthew p Conomos, Adrienne m Stilp, Zilin Li, Tamar Sofer, Adam a Szpiro, Wei Chen, John m Brehm, Juan c Celedón, Susan Redline, George j Papanicolaou, and Timothy a Thornton. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4):653–666, 2016.
- [2] Wei Zhou, Jonas B Nielsen, Lars G Fritsche, Rounak Dey, Maiken E Gabrielsen, Brooke N Wolford, Jonathon Lefaive, Peter Vandehaar, Sarah A Gagliano, Aliya Gifford, Lisa A Bastarache, Wei-Qi Wei, Joshua C Denny, Maoxuan Lin, Kristian Hveem, Hyun Min Kang, Goncalo R Abecasis, Cristen J Willer, and Seunggeun Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics*, 50(9):1335–1341, 2018.
- [3] Rounak Dey, Wei Zhou, Tuomo Kiiskinen, Aki Havulinna, Amanda Elliott, Juha Karjalainen, Mitja Kurki, Ashley Qin, FinnGen, Seunggeun Lee, et al. An efficient and accurate frailty model approach for genome-wide survival association analysis controlling for population structure and relatedness in large-scale biobanks. *bioRxiv*, pages 2020–10, 2020.
- [4] Arthur R. Gilmour, Robin Thompson, and Brian R. Cullis. Average information reml: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 51(4):1440–1450, 1995.
- [5] Gulnara R Svishcheva, Tatiana I Axenovich, Nadezhda M Belonogova, Cornelia M Van Duijn, and Yurii S Aulchenko. Rapid variance components-based method for whole-genome association analysis. *Nature Genetics*, 44(10):1166, 2012.
- [6] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, Nick Patterson, and Alkes L Price. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 47(3):284, 2015.
- [7] Longda Jiang, Zhili Zheng, Ting Qi, Kathryn Kemper, Naomi Wray, Peter Visscher, and Jian Yang. A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, 51(12):1749–2, 2019.
- [8] Wei Zhou, Zhangchen Zhao, Jonas B Nielsen, Lars G Fritsche, Jonathon LeFaive, Sarah A Gagliano Taliun, Wenjian Bi, Maiken E Gabrielsen, Mark J Daly, Benjamin M Neale, et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nature genetics*, 52(6):634–639, 2020.
- [9] H. E. Daniels. Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, 25(4):631–650, 1954.
- [10] Rounak Dey, Ellen M Schmidt, Goncalo R Abecasis, and Seunggeun Lee. A fast and accurate algorithm to test for binary phenotypes and its application to phewas. *The American Journal of Human Genetics*, 101(1):37–49, 2017.

- [11] Rounak Dey, Jonas B. Nielsen, Lars G. Fritsche, Wei Zhou, Huanhuan Zhu, Cristen J. Willer, and Seunggeun Lee. Robust meta-analysis of biobank-based genome-wide association studies with unbalanced binary phenotypes. *Genetic Epidemiology*, 43(5):462–476, 2019.
- [12] Wenjian Bi, Zhangchen Zhao, Rounak Dey, Lars G Fritsche, Bhramar Mukherjee, and Seunggeun Lee. A fast and accurate method for genome-wide scale phenome-wide g x e analysis and its application to uk biobank. *The American Journal of Human Genetics*, 105(6):1182–1192, 2019.
- [13] S Tsuruta, I Misztal, and I Strandén. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. *Journal of animal science*, 79(5):1166–1172, 2001.
- [14] M.F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 19(2):433–450, 1990.
- [15] Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2):1–34, 2011.
- [16] JJ Allaire, Romain Francois, Kevin Ushey, Gregory Vandenbrouck, Marcus Geelnard, and Intel. *RcppParallel: Parallel Programming Tools for 'Rcpp'*, 2019. R package version 4.4.4.