Bioinformatics Advance Access published August 4, 2015

Genetics and Population Analysis

NAM: Association Studies in Multiple Populations

Alencar Xavier[1,*], Shizhong Xu[3], William M. Muir[2], and Katy Martin Rainey[1]

[1]Department of Agronomy, Purdue University, 915 W. State St., Lilly Hall, West Lafayette, IN, 47907, USA.

[2]Department of Animal Science, Purdue University, 915 W. State St., Lilly Hall, West Lafayette, IN, 47907, USA.

[3]Department of Plant Science, University of California, 3134 Batchelor Hall, Riverside, CA, 92521, USA.

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Motivation: Mixed linear models (MLM) provide important techniques for performing genome-wide associations studies (GWAS). However, current models have pitfalls associated with their strong assumptions. Here, we propose a new implementation designed to overcome some of these pitfalls using an empirical Bayes algorithm.

Results: NAM is an R package that allows user to take into account prior information regarding population stratification to relax the linkage phase assumption of current methods. It allows markers to be treated as a random effect to increase the resolution, and uses a sliding-window strategy to increase power and avoid double fitting markers into the model.

Availability: NAM is an R package available in the CRAN repository. It can be installed in R by typing install.packages('NAM').

Contact: xaviera@purdue.edu

Supplementary Information: Supplementary information about the method and algorithms is available at *Bioinformatics* online.

1 INTRODUCTION

Since the advent of high-throughput genotyping technology, extensive efforts have focused on creating efficient mixed linear models (MLM) to address relatedness and computational issues in genome-wide association studies (GWAS) (Zhou and Stephens 2012, Kang et al. 2010). However, major pitfalls that still must be improved (Yang et al. 2014), including issues with resolution and detection power. Furthermore, MLM methods do not take into account the linkage phase associated with the multiple populations that comprise the association panel.

Association studies rely on persistent linkage disequilibrium (LD) between markers and quantitative trait loci (QTL). Such associations decay over time through recombination events, triggering LD that allows differentiation between populations (de Roos et al. 2008). Therefore, association panels containing multiple populations are more likely to display diverging linkage phases, what makes QTL undetectable (Wientjes et al. 2013).

Here we introduce "NAM," a statistical package for association studies that aims to overcome some limitations of the mixed model framework and supports users to work with multiple populations when a stratification factor is known.

2 STRUCTURE AND LINKAGE PHASE

Structure, crypto-relatedness (Yu et al. 2006) and unequal linkage phase across founders represent a major challenge for quantitative trait nucleotide (QTN) mapping (Lin et al. 2003). Association methods deal with multiple levels of relatedness through genomic kinship, eigenvectors and model-based approaches (Pritchard et al. 2000, Kang et al. 2010, Zhang et al. 2010) but are not able to handle linkage phase. Next-generation mapping populations such as nested association mapping (NAM) populations, it can address this issue by recoding the genotypic matrix to characterize haplotypes.

For example, in NAM populations alleles either come from the standard parent or from the founder. Thus, a given marker m can be represented as the number of alleles that come from each source: m=[a_s, a_1, a_2, … ,a_f], where a_s represents the number of alleles inherited from the standard parent and a_1 to a_f represent alleles inherited from founder parents. The haplotype representation of genotypes works as follows. A given locus in an individual that belongs to family 2: if homozygous to the standard parent, it is coded as m=[2,0,0,...,f]; if heterozygous, m=[1,0,1,...,f]; and m=[0,0,2,...,f] if homozygous to the founder. Similar approaches can work for a random population if structural factors are known. This makes possible to relax assumptions regarding the linkage phase between the molecular marker and the QTN across populations, allowing different populations to pursue distinct coefficients for the marker under evaluation.

If the family term (stratification) is specified, the NAM package initiates the association study by recoding alleles and building the genomic relationship matrix (GRM). After solving the MLM through the EMMA algorithm (Kang et al. 2008), NAM utilizes the P3D strategy (Zhang et al. 2010) to avoid updating the polygenic term for every marker. Using the empirical Bayes approach, each molecular marker is treated as a random effect and the model is refitted using Eigen decomposition (Zhou and Stephens 2012) and evaluated with the likelihood ratio test (LRT).

Datasets can still be analyzed by the empirical Bayes algorithm when no stratification factor is provided (Wang 2015), applicable to multi-parent advanced generation inter-cross (MAGIC), random or bi-parental populations.

3 MAJOR BACKGROUND EFFECT

Most association algorithms attempt to control the diffuse background effect and are unable to control genes of major effect (Se-

[*]To whom correspondence should be addressed.

gura et al. 2012) or use step-wise regression (Yu et al. 2008). To address this issue, our package implements a sliding-window algorithm (Xu and Atchley 1995). The approach consists of controlling the background by fitting a model with all markers outside a window, similar to whole-genome regression methods (Legarra et al. 2015). The use of a sliding window prevents the double-fitting of the markers in the model, once the marker under evaluation is included in the GRM (Yang et al. 2014). More details about the algorithm are available in the supplementary file.

4 METHODS COMPARISON

To demonstrate the increase in power and resolution of the NAM package, we compared to three standard algorithms of mixed linear models: the P3D/EMMAX algorithm with step-wise regression implemented in GAPIT (Lipka et al. 2012), the GRAMMAR-Gamma algorithm implemented in GenABEL (Svishcheva et al. 2012), and the GEMMA algorithm proposed and implemented by Zhou and Stephens (2012).

We used a simulated nested association panel with 840 individuals from six families, with 10 chromosomes of 100 cM and one marker by cM. A QTL was placed in the center of each chromosome (Figure 1). The NAM package was able to capture most QTL with few false positives and little background noise, while other packages provided lower resolution QTL.

5 ADDITIONAL TOOLS

The NAM package provides complimentary statistical tool, including the fixation indices (Weir and Cockerham 1984), estimator of gene content (Forneris et al. 2015), functions to deal with minor allele frequency and repeated markers, and the package performs imputation of missing loci through random forest (Stekhoven and Buhlmann 2012). Best linear unbiased predictors (BLUP) are often used to replace raw phenotypes (Robinson 1991) in association studies. Our package offers two algorithms to compute BLUP and variance components: REML (Kang et al. 2008) and Bayesian Gibbs Sampling (Sorensen and Gianola 2002). The latter allows users to perform Bayesian inferences.

6 CONCLUSIONS

The NAM package has implemented simple solutions to overcome pitfalls identified in association studies in mixed model frameworks, increasing the mapping power and resolution. The package includes an additional toolset for complimentary analysis of marker quality control, population stratification, and to calculate BLUPs.

ACKNOWLEDGEMENTS

We acknowledge William Beavis for providing the simulated data, and Tiago Pimenta and Quishan Wang for reviewing algorithms and optimization of the source code.

REFERENCES

de Roos, A. P. W. et al. (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics*, 179(3), 1503-1512.

Forneris, N. S. et al. (2015). Quality Control of Genotypes Using Heritability Estimates of Gene Content at the Marker. *Genetics*, genetics-114.

Kang, H. M. et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4), 348-354.

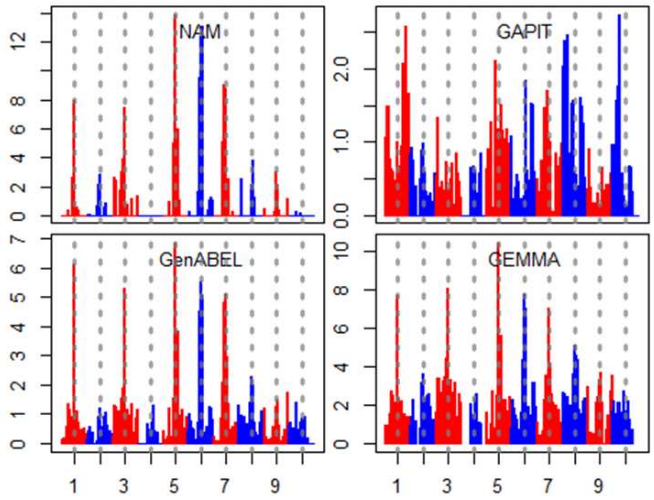

Fig. 1. Simulated data: Genome-wide association mapping performed with four different implementations. Vertical lines represent the position of the QTL.

Kang, H. M. et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4), 348-354.

Kang, H. M. et al. (2008). Efficient control of population structure in model organism association mapping. *Genetic*s, 178(3), 1709-1723.

Legarra, A. et al. (2015). A comparison of methods for whole-genome QTL mapping using dense markers in four livestock species. *Gen.Sel.Evol.*, 47(1), 6.

Lin, M. et al. (2003). A general statistical framework for mapping quantitative trait loci in nonmodel systems: issue for characterizing linkage phases. *Genetics,* 165(2), 901-913.

Lipka, A. E. et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, 28(18), 2397-2399.

Pritchard, J. K. et al. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945-959.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical science*, 15-32, 6(1).

Segura, V. et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics*, 44(7), 825-830.

Sorensen, D., and Gianola, D. (2002). *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. Springer Science & Business Media.

Stekhoven, D.J and Buhlmann, P. (2012). MissForest: non-parametric missing value imputation for mixed-tpe data. *Bioinformatic*s, 28(1), 112-118.

Svishcheva, G. R. et al. (2012). Rapid variance components-based method for whole-genome association analysis. *Nature genetics*, 44(10), 1166-1170.

Wang, Q. (2015). An Empirical Bayes Method for Genome-Wide Association Studies. *In Plant and Animal Genome XXXII*. W799/Statistical Genomics.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6), 1358-1370.

Wientjes, Y. C. et al. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetic*s, 193(2), 621-631.

Xu, S. and Atchley, W. R. (1995). A random model approach to interval mapping of quantitative trait loci. *Genetics*, 141(3), 1189-1197.

Yang, J. et al. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics*, 46(2), 100-106.

Yu, J. et al. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics*, 178(1), 539-551.

Yu, J. et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2), 203-208.

Zhang, Z. et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4), 355-360.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7), 821-824.

Empirical Bayes Genome-Wide Association Method

1. Polygenic model

We use a W-parent nested association mapping (NAM) population containing a standard parent and eight founders ($W = 7$) as an example to demonstrate the theory and methods. The method holds for any p-parents populations. Let y be an $n \times 1$ vector of phenotypic values for n individuals. Define Z_k as an $n \times (W + 1)$ matrix of founder allele inheritance for locus k. The jth row of matrix Z_k is defined as a $n \times (W + 1)$ vector. If this individual is a heterozygote carrying the first and second founder alleles, then

$$Z_{jk} = [\,1\;1\;0\;0\;0\;0\;0\;0\,]$$

If the individual is a homozygote inheriting both alleles from the fifth founder, then Z_{jk} is defined as

$$Z_{jk} = [\,0\;0\;0\;0\;2\;0\;0\;0\,]$$

The general rule for defining Z_{jk} is that there are at most two non-zero elements and the sum of all the eight elements equals two. Let

$$\gamma_k = \left[\gamma_{1k}\;\; \gamma_{2k}\;\; \gamma_{3k}\;\; \gamma_{4k}\;\; \gamma_{5k} \cdots \gamma_{(W+1)k}\,\right]^T$$

be an $(W + 1) \times 1$ vector of allelic effects for the eight founders. The phenotypic vector y is described by the following linear mixed model,

$$y = X\beta + \sum_{k=1}^{m} Z_k \gamma_k + \varepsilon \tag{1}$$

where X is a design matrix for fixed effects β, m is the number of (marker) loci available in the data and ε is an $n \times 1$ vector of residual errors. Assume that $\varepsilon \sim N(0, I_n\sigma^2)$ and $\gamma_k \sim N(0, I_8\phi_k^2)$, where σ^2 is the residual error variance and ϕ_k^2 is a common variance shared by all the eight founder alleles at locus k. Because γ_k are assumed to be a vector of random variables, the model is called the linear mixed model. The expectation of y is $\mathrm{E}(y) = X\beta$ and the variance-covariance matrix is

$$\mathrm{var}(y) = \sum_{k=1}^{m} Z_k Z_k^T \phi_k^2 + I\sigma^2 \tag{2}$$

When m is large, it is hard to estimate all m variance components in a simultaneous manner. Therefore, we make an assumption that all loci share the same variance component. This treatment implies that there are m polygenes in the model. This is a polygenic model and is treated as the null model for QTL detection. Under the polygenic model, we assume $\gamma_k \sim N(0, I_8\phi^2 / m)$ for all $k = 1,...,m$, where ϕ^2 is the polygenic variance (the sum of variances for all individual loci). Under the polygenic model, the variance-covariance matrix is

$$\text{var}(y) = \frac{1}{m}\sum_{k=1}^{m} Z_k Z_k^T \phi^2 + I\sigma^2 = K\phi^2 + I\sigma^2 = (K\lambda + I)\sigma^2 = H\sigma^2 \qquad (3)$$

where $\lambda = \phi^2 / \sigma^2$ is the variance ratio, $H = K\lambda + I$ is the covariance structure and

$$K = \frac{1}{m}\sum_{k=1}^{m} Z_k Z_k^T \qquad (4)$$

is a marker-generated kinship matrix.

2. Restricted maximum likelihood estimation

To estimate the variance components, we use the restricted maximum likelihood (REML) method to maximize the following likelihood function,

$$L(\theta) = -\frac{n-r}{2}\ln(\sigma^2) - \frac{1}{2}\ln|H| - \frac{1}{2\sigma^2}(y - X\beta)^T H^{-1}(y - X\beta) - \frac{1}{2}\ln|X^T H^{-1} X| \qquad (5)$$

where $\theta = \{\beta, \lambda, \sigma^2\}$ is the parameter vector and r is the rank of matrix *X*. Given λ, the restricted maximum likelihood estimates of β and σ^2 are

$$\begin{aligned} \hat{\beta} &= (X^T H^{-1} X)^{-1} X^T H^{-1} y \\ \hat{\sigma}^2 &= \frac{1}{n-r}(y - X\hat{\beta})^T H^{-1}(y - X\hat{\beta}) \end{aligned} \qquad (6)$$

The above estimated parameters are expressed as functions of λ. Substituting β and σ^2 in Equation (6) by $\hat{\beta}$ and $\hat{\sigma}^2$ in Equation (5) yields a profiled likelihood function that is only a function of λ, as shown below,

$$L(\lambda) = -\frac{1}{2}\ln|H| - \frac{1}{2}\ln|X^T H^{-1} X| - \frac{n-r}{2}\ln(y^T P y) \qquad (7)$$

where

$$P = H^{-1} - H^{-1} X (X^T H^{-1} X)^{-1} X^T H^{-1} \qquad (8)$$

A numeric solution of λ can be found iteratively using the Newton iteration algorithm,

$$\lambda^{(t+1)} = \lambda^{(t)} - \left[\frac{\partial^2 L(\lambda^{(t)})}{\partial \lambda^2}\right]^{-1} \left[\frac{\partial L(\lambda^{(t)})}{\partial \lambda}\right] \qquad (9)$$

Once the iteration process has converged, the solution is the REML estimate of λ, denoted by $\hat{\lambda}$. The log likelihood value of equation (7) evaluated at $\lambda = \hat{\lambda}$ is called $L_0 = L(\hat{\lambda})$ and it will be used in the likelihood ratio test (LRT) for individual QTL (to be discussed in a later section).

3. Eigenvalue decomposition

The likelihood function requires inverse and determinant of matrix *H*, an $n \times n$ matrix, and the computation can be demanding for large sample size. We used the eigenvalue decomposition to deal with the *K* matrix. Further investigation of Equation (7) shows that the profiled restricted log likelihood function only requires the log determinant of matrix *H* and various quadratic forms involving H^{-1}. Let us perform eigenvalue decomposition for K so that $K = UDU^T$, where $D = \text{diag}\{\delta_1, ..., \delta_n\}$ is a diagonal matrix for the eigenvalues and U is the eigenvectors, an $n \times n$ matrix. The eigenvectors have the property of $U^T = U^{-1}$ so that $UU^T = I$. Now, let us rewrite matrix *H* by

$$H = K\lambda + I = UDU^T\lambda + I = U(D\lambda + I)U^T \tag{10}$$

The determinant of *H* is

$$|H| = |U(D\lambda + I)U^T| = |D\lambda + I||UU^T| = |D\lambda + I| \tag{11}$$

where $D\lambda + I$ is a diagonal matrix. Therefore, the log determinant of matrix *H* is

$$\ln|H| = \sum_{j=1}^{n} \ln(\delta_j\lambda + 1) \tag{12}$$

The restricted log likelihood function also involves various quadratic terms in the form of $a^T H^{-1} b$, for example, $X^T H^{-1} X$, $X^T H^{-1} y$ and $y^T H^{-1} y$. Using eigenvalue decomposition, we can rewrite the quadratic form by

$$a^T H^{-1} b = a^T U(D\lambda + I)^{-1} U^T b = a^{*T}(D\lambda + I)^{-1} b^* = \sum_{j=1}^{n} a_j^{*T} b_j^* (\delta_j\lambda + 1)^{-1} \tag{13}$$

where $a^* = U^T a$ and $b^* = U^T b$. Note that a_j^* is the *j*th element (row) of vector (matrix) a^* and b_j^* is the *j*th element (row) of vector (matrix) b^*. Using eigenvalue decomposition, matrix inversion and determinant calculation have been simplified into simple summations, and thus, the computational speed can be substantially improved.

4. Genome scanning for quantitative trait loci

Once λ is estimated, we are able to scan the entire genome by controlling the polygenic covariance structure using the λ estimated from the null model. The genomic scanning model for the *k*th locus is

$$y = X\beta + Z_k\gamma_k + \xi + \varepsilon \tag{14}$$

where ξ is the polygene. The general error term $\xi + \varepsilon$ has $\text{E}(\xi + \varepsilon) = 0$ and $\text{var}(\xi + \varepsilon) = (K\hat{\lambda} + I)\sigma^2$, where the λ value is fixed at its estimated values under the polygenic model. This time, we assume $\gamma_k \sim N(0, I_8\phi_k^2)$ and perform a significance test for

$H_0 : \phi_k^2 = 0$. Under the null hypothesis, the *k*th locus is not linked to QTL. Because γ_k is assumed to be a random effect, the expectation of y in the above model remains $\mathrm{E}(y) = X\beta$, but the variance-covariance matrix is

$$\mathrm{var}(y) = Z_k Z_k^T \phi_k^2 + K\phi^2 + I\sigma^2 = (Z_k Z_k^T \lambda_k + K\hat{\lambda} + I)\sigma^2 \tag{15}$$

where $\lambda_k = \phi_k^2 / \sigma^2$ is the variance ratio. Let $y^* = U^T y$, $X^* = U^T X$ and $Z_k^* = U^T Z_k$ be transformed variables so that

$$y^* = X^*\beta + Z_k^*\gamma_k + U^T(\xi + \varepsilon) \tag{16}$$

The variance-covariance matrix of y^* is

$$\begin{aligned} \mathrm{var}(y^*) &= Z_k^* Z_k^{*T} \phi_k^2 + U^T(K\hat{\lambda} + I)U\sigma^2 \\ &= Z_k^* Z_k^{*T} \phi_k^2 + U^T U(D\hat{\lambda} + I)U^T U\sigma^2 \\ &= Z_k^* Z_k^{*T} \phi_k^2 + (D\hat{\lambda} + I)\sigma^2 \\ &= (Z_k^* Z_k^{*T} \lambda_k + R)\sigma^2 \end{aligned} \tag{17}$$

where $R = D\hat{\lambda} + I$ is a known diagonal matrix for the general covariance structure. Let $H_k = Z_k^* Z_k^{*T} \lambda_k + R$ and define the restricted log likelihood function for parameter vector $\theta = \{\beta, \lambda_k, \sigma^2\}$ by

$$L(\theta) = -\frac{n-r}{2}\ln(\sigma^2) - \frac{1}{2}\ln|H_k| - \frac{1}{2\sigma^2}(y^* - X^*\beta)^T H_k^{-1}(y^* - X^*\beta) - \frac{1}{2}\ln|X^{*T} H_k^{-1} X^*| \tag{18}$$

Given λ_k, the maximum likelihood estimates of β and σ^2 are

$$\begin{aligned} \hat{\beta} &= (X^{*T} H_k^{-1} X^*)^{-1} X^{*T} H_k^{-1} y^* \\ \hat{\sigma}^2 &= \frac{1}{n-r}(y^* - X^*\hat{\beta})^T H_k^{-1}(y^* - X^*\hat{\beta}) \end{aligned} \tag{19}$$

The above estimated parameters are expressed as functions of λ_k. Substituting β and σ^2 in Equation (18) by $\hat{\beta}$ and $\hat{\sigma}^2$ in Equation (19) yields a profiled likelihood function that is only a function of λ_k, as shown below,

$$L(\lambda_k) = -\frac{1}{2}\ln|H_k| - \frac{1}{2}\ln|X^{*T} H_k^{-1} X^*| - \frac{n-r}{2}\ln(y^{*T} P_k y^*) \tag{20}$$

where

$$P_k = H_k^{-1} - H_k^{-1} X^* (X^{*T} H_k^{-1} X^*)^{-1} X^{*T} H_k^{-1} \tag{21}$$

The Newton algorithm for the numeric solution of λ_k is

$$\lambda_k^{(t+1)} = \lambda_k^{(t)} - \left[\frac{\partial^2 L(\lambda_k^{(t)})}{\partial \lambda_k^2} \right]^{-1} \left[\frac{\partial L(\lambda_k^{(t)})}{\partial \lambda_k} \right] \tag{22}$$

Once the iteration process converges, the solution is the REML estimate of λ_k, denoted by $\hat{\lambda}_k$. The log likelihood value of Equation (20) evaluated at $\lambda_k = \hat{\lambda}_k$ is called $L_1 = L(\hat{\lambda}_k)$. The null hypothesis is $H_0 : \lambda_k = 0$. The likelihood ratio test (LRT) for the *k*th locus is defined by

$$\Gamma_k = -2(L_0 - L_1) \tag{23}$$

The entire genome is scanned one locus at a time. Locus *k* is declared as significant if $\Gamma_k > \Gamma_{1-0.05}$ where $\Gamma_{1-0.05}$ is the 95% percentile of the distribution of Γ_k under the null model. The 95% percentile threshold value is drawn from a permutation analysis (see the Result section of the manuscript).

5. Woodbury matrix identities

Efficient matrix inversion and determinant calculation is required to evaluate the log likelihood function shown in Equation (20). We use the Woodbury matrix identities to improve the computational speed. The Woodbury matrix identities are

$$\begin{aligned} H_k^{-1} &= (Z_k^* Z_k^{*T} \lambda_k + R)^{-1} \\ &= R^{-1} - R^{-1} Z_k^* (Z_k^{*T} R^{-1} Z_k^* + I_8 \lambda_k^{-1})^{-1} Z_k^{*T} R^{-1} \\ &= R^{-1} - \lambda_k R^{-1} Z_k^* (\lambda_k Z_k^{*T} R^{-1} Z_k^* + I_8)^{-1} Z_k^{*T} R^{-1} \end{aligned} \tag{24}$$

and

$$\begin{aligned} | H_k | &= | Z_k^* Z_k^{*T} \lambda_k + R | \\ &= | R || I_8 \lambda_k || Z_k^{*T} R^{-1} Z_k^* + I_8 \lambda_k^{-1} | \\ &= | R || \lambda_k Z_k^{*T} R^{-1} Z_k^* + I_8 | \end{aligned} \tag{25}$$

Because $R = D\hat{\lambda} + I$ is a diagonal matrix, the Woodbury identities convert the above calculations into inversion and determinant of matrices with dimension 8×8. The restricted likelihood function also involves various quadratic terms in the form of $a^T H_k^{-1} b$, which can be expressed as

$$a^T H_k^{-1} b = a^T R^{-1} b - \lambda_k a^T R^{-1} Z_k^* (\lambda_k Z_k^{*T} R^{-1} Z_k^* + I_8)^{-1} Z_k^{*T} R^{-1} b \tag{26}$$

Note that the above quadratic has been expressed as a function of various $a^T R^{-1} b$ terms. The simplified quadratic term is calculated using

$$a^T R^{-1} b = \sum_{j=1}^{n} a_j^T b_j (\delta_j \hat{\lambda} + 1)^{-1} \tag{27}$$

where a_j and b_j are the *j*th rows of matrices a and b, respectively, for $j = 1, ..., n$.

6. Best linear unbiased prediction of QTL effects

To derive the best linear unbiased prediction (BLUP) of γ_k, we need the following information,

$$\mathrm{E}\begin{bmatrix} y^* \\ \gamma_k \end{bmatrix} = \begin{bmatrix} X^* \beta \\ 0 \end{bmatrix} \tag{28}$$

and

$$\mathrm{var}\begin{bmatrix} y^* \\ \gamma_k \end{bmatrix} = \begin{bmatrix} (Z_k^* Z_k^{*T} \lambda_k + R) & Z_k^* \lambda_k \\ Z_k^{*T} \lambda_k & I_8 \lambda_k \end{bmatrix} \sigma^2 = \begin{bmatrix} H_k & Z_k^* \lambda_k \\ Z_k^{*T} \lambda_k & I_8 \lambda_k \end{bmatrix} \sigma^2 \tag{29}$$

The BLUP of γ_k can be derived as the conditional expectation of γ_k given y^*, assuming that β is known, which has the following expression,

$$\begin{aligned} \mathrm{E}(\gamma_k \mid y^*) &= \lambda_k Z_k^{*T} H_k^{-1} (y^* - X^* \beta) \\ &= \lambda_k Z_k^{*T} H_k^{-1} y^* - \lambda_k Z_k^{*T} H_k^{-1} X^* (X^{*T} H_k^{-1} X^*)^{-1} X^{*T} H_k^{-1} y^* \end{aligned} \tag{30}$$

The conditional variance is

$$\mathrm{var}(\gamma_k \mid y^*) = I_8 \lambda_k \sigma^2 - \lambda_k Z_k^{*T} H_k^{-1} Z_k^* \lambda_k \sigma^2 \tag{31}$$

Let $\hat{\gamma}_k = \mathrm{E}(\gamma_k \mid y^*)$ and $V_{\hat{\gamma}_k} = \mathrm{var}(\gamma_k \mid y^*)$, which provide an alternative test for the null hypothesis, $H_0 : \gamma_k = 0$. The test statistics is called the Wald test expressed by

$$\mathrm{Wald} = \hat{\gamma}_k^T V_{\hat{\gamma}_k}^{-1} \hat{\gamma}_k \tag{32}$$

7. Moving window scanning of the genome

The polygenic background control is similar to the composite interval mapping using co-factors to control the background effects. However, it does not eliminate the interference of the current locus from neighboring markers in the presence of linkage disequilibrium. Therefore, we extend the random model approach to addressing the problem of interference. We adopted the random model approach of Xu and Atchley (1995) by defining a window of fixed width that covers the locus of interest. Let Z_k be the allelic inheritance variables and γ_k be the QTL effects for locus *k*. Our target locus is *k* but we use Z_{k-1} and Z_{k+1} as the flanking markers to eliminate interference from effects of the left and right sides of the genome. The window size is fixed in *d* cM long with locus *k* right in the middle of the window. Note that Z_{k-1} and Z_{k+1} are not the genotype indicators for markers $k-1$ and $k+1$; rather, they are the genotype indicators for the left and right markers 0.5d cM deviating from marker *k*, respectively. These two markers define the moving window of d cM in width. The random model of this moving window scanning procedure is

$$y = X\beta + Z_{k-1}\gamma_{k-1} + Z_k\gamma_k + Z_{k+1}\gamma_{k+1} + \xi + \varepsilon \tag{33}$$

where only γ_k is the QTL effect under investigation but γ_{k-1} and γ_{k+1} appear also in the model to control potential interference. The QTL effects of the flanking markers of the window are also assumed to be random so that $\gamma_{k-1} \sim N(0, I_8\phi_{k-1}^2)$ and $\gamma_{k+1} \sim N(0, I_8\phi_{k+1}^2)$. The variance-covariance matrix of the model is

$$\begin{aligned} \mathrm{var}(y) &= Z_{k-1}Z_{k-1}^T\phi_{k-1}^2 + Z_kZ_k^T\phi_k^2 + Z_{k+1}Z_{k+1}^T\phi_{k+1}^2 + (K\hat{\lambda} + I)\sigma^2 \\ &= (Z_{k-1}Z_{k-1}^T\lambda_{k-1} + Z_kZ_k^T\lambda_k + Z_{k+1}Z_{k+1}^T\lambda_{k+1} + K\hat{\lambda} + I)\sigma^2 \end{aligned} \tag{34}$$

where $\lambda_{k-1} = \phi_{k-1}^2 / \sigma^2$ is the variance ratio. Let us define $W_k = Z_{k-1} \,||\, Z_k \,||\, Z_{k+1}$ as an $n \times 24$ matrix (column concatenation of the three *Z* matrices) and define

$$\psi_k = \begin{bmatrix} I_8\lambda_{k-1} & 0 & 0 \\ 0 & I_8\lambda_k & 0 \\ 0 & 0 & I_8\lambda_{k+1} \end{bmatrix}$$

as a 24×24 diagonal matrix. The variance-covariance matrix of *y* is rewritten as

$$\mathrm{var}(y) = (W_k\psi_kW_k^T + K\hat{\lambda} + I)\sigma^2 \tag{35}$$

Define $y^* = U^T y$ and $W_k^* = U^T W_k$, the variance-covariance matrix of the transformed *y* becomes

$$\mathrm{var}(y^*) = (W_k^* \psi_k W_k^{*T} + R)\sigma^2 = H_k \sigma^2 \tag{36}$$

where

$$H_k = W_k^* \psi_k W_k^{*T} + R \tag{37}$$

The profiled restricted log likelihood function for $\psi_k = f(\lambda_{k-1}, \lambda_k, \lambda_{k+1})$ is

$$L(\psi_k) = -\frac{1}{2}\ln|H_k| - \frac{1}{2}\ln|X^{*T} H_k^{-1} X^*| - \frac{n-r}{2}\ln(y^{*T} P_k y^*) \tag{38}$$

Evaluation of this likelihood function can be time consuming. However, we can use the Woodbury matrix identities to find the inverse and determinant of matrix H_k,

$$\begin{aligned} H_k^{-1} &= (W_k^* \psi_k W_k^{*T} + R)^{-1} \\ &= R^{-1} - R^{-1} W_k^* (W_k^{*T} R^{-1} W_k^* + \psi_k^{-1})^{-1} W_k^{*T} R^{-1} \\ &= R^{-1} - R^{-1} W_k^* \psi_k (W_k^{*T} R^{-1} W_k^* \psi_k + I_{24})^{-1} W_k^{*T} R^{-1} \\ &= R^{-1} - R^{-1} W_k^* (\psi_k W_k^{*T} R^{-1} W_k^* + I_{24})^{-1} \psi_k W_k^{*T} R^{-1} \end{aligned} \tag{39}$$

and

$$\begin{aligned} |H_k| &= |W_k^* \psi_k W_k^{*T} + R| \\ &= |R||\psi_k||W_k^{*T} R^{-1} W_k^* + \psi_k^{-1}| \\ &= |R||\psi_k W_k^{*T} R^{-1} W_k^* + I_{24}| \\ &= |R||W_k^{*T} R^{-1} W_k^* \psi_k + I_{24}| \end{aligned} \tag{40}$$

Dimension of the matrices required in the inversion and determinant calculation has increased from 8×8 (single marker analysis) to 24×24 in the moving window scanning method. Matrix *R* is diagonal and matrix $W_k^{*T} R^{-1} W_k^* \psi_k + I_{24}$ has low dimension. Therefore, the determinants of these two matrices are calculated with low cost. The quadratic term involved in the likelihood function is

$$a^T H_k^{-1} b = a^T R^{-1} b - a^T R^{-1} W_k^* \psi_k (W_k^{*T} R^{-1} W_k^* \psi_k + I_{24})^{-1} W_k^{*T} R^{-1} b \tag{41}$$

The Newton algorithm used before is now replaced by the Newton-Raphson algorithm because three parameters are estimated simultaneously. The likelihood value evaluated at $\psi_k = \hat{\psi}_k$ is denoted by $L_1 = L(\hat{\psi}_k)$.

Hypothesis test for $H_0 : \phi_k^2 = 0$ under the moving window scanning procedure is different from that introduced before because the null model is not the polygenic model but a model excluding ϕ_k^2 but keeping ϕ_{k-1}^2, ϕ_{k+1}^2 and the polygenic variance. This means that for every locus scanned, one must also calculate a locus specific L_0 in order to find the likelihood ratio test statistics. The likelihood ratio test statistic is again denoted by $\Gamma_k = -2(L_0 - L_1)$.

8. Moving window scanning with adjusted polygenic effect

The moving window scanning procedure will increase the resolution of QTL mapping, but may also reduce statistical power if the window is too narrow. In addition, the polygenic effect also contains QTL effects in the moving window. Essentially, QTL effects in a window are estimated twice, one by the polygenic effect and one by the moving window. The two estimates are competing with each other, leading to a lower power for QTL detection. We proposed the following remediation by releasing the effects in the moving window absorbed by the polygenic effect. The revised model is

$$y = X\beta + Z_{k-1}\gamma_{k-1} + Z_k\gamma_k + Z_{k+1}\gamma_{k+1} + \xi - \xi_k + \varepsilon \tag{42}$$

where ξ is still the polygenic effect and ξ_k is the polygenic effect linked to all markers covered by the current moving window, i.e., window *k*. This effect is estimated under the polygenic model (the null model). To minimize the revision of the model and maximize the computational speed, we rearranged the above model into

$$y_k = X\beta + Z_{k-1}\gamma_{k-1} + Z_k\gamma_k + Z_{k+1}\gamma_{k+1} + \xi + \varepsilon \tag{43}$$

where $y_k = y + \xi_k$ is a newly adjusted vector of phenotypic values. For each window, we used a window specific vector of phenotypic values and left all existing algorithm in the regular moving window scanning procedure intact. This is obvious because the right hand side of equation (43) is the same as that of equation (42). As a result, $\mathrm{E}(y_k) = X\beta$ and

$$\mathrm{var}(y_k) = (Z_{k-1}Z_{k-1}^T\lambda_{k-1} + Z_kZ_k^T\lambda_k + Z_{k+1}Z_{k+1}^T\lambda_{k+1} + K\hat{\lambda} + I)\sigma^2 \tag{44}$$

The only additional work is to find ξ_k for each window.

We now go back to the original polygenic model in equation (1). Under the polygenic model, all marker effects share the same variance, i.e., $\gamma_k \sim N(0, I_8\phi^2 / m)$, where $\phi^2 = \lambda\sigma^2$ is estimated from the data under the polygenic model. The BLUP estimate of γ_k is derived from the multivariate theorem. The joint distribution of y and γ_k are multivariate normal with expectation and variance given by

$$\mathrm{E}\begin{bmatrix} y \\ \gamma_k \end{bmatrix} = \begin{bmatrix} X\beta \\ 0 \end{bmatrix} \tag{45}$$

and

$$\operatorname{var}\begin{bmatrix} y \\ \gamma_k \end{bmatrix} = \begin{bmatrix} \operatorname{var}(y) & \operatorname{cov}(y, \gamma_k^T) \\ \operatorname{cov}(\gamma_k, y^T) & \operatorname{var}(\gamma_k) \end{bmatrix} = \begin{bmatrix} K\phi^2 + I\sigma^2 & Z_k\phi^2 / m \\ Z_k^T\phi^2 / m & I_8\phi^2 / m \end{bmatrix} \tag{46}$$

respectively. From the expectation and variance, we can find the conditional expectation of γ_k given y,

$$\mathrm{E}(\gamma_k \mid y) = Z_k\phi^2(mK\phi^2 + Im\sigma^2)^{-1}(y - X\beta) \tag{47}$$

which is the BLUP of γ_k if the parameters are known. The parameters are substituted by the estimated values under the polygenic model and thus the BLUP is in fact empirical Bayes estimates,

$$\hat{\gamma}_k = \mathrm{E}(\gamma_k \mid y) = Z_k\hat{\phi}^2(mK\hat{\phi}^2 + Im\hat{\sigma}^2)^{-1}(y - X\hat{\beta}) \tag{48}$$

We have a total of *m* markers and thus we will have *m* γ_k to estimate under the polygenic model (prior to the moving window scanning). When we scan the *k*th moving window, the polygenic effect covered by this window (*d* cM in width) is ξ_k, which is

$$\xi_k = \sum_{k'=1}^{m_k} Z_{k'}\hat{\gamma}_{k'} \tag{49}$$

where m_k is the number of markers covered by window k and $Z_{k'}$ is the k' marker genotype indicator variable. This polygenic adjusted moving window will avoid competing between the polygenic effect and the effect in the window. The method is computationally efficient because the polygenic effects are only estimated under the null model prior to the moving window scanning.

References

Xu, S., & Atchley, W. R. (1995). A random model approach to interval mapping of quantitative trait loci. *Genetics*, *141*(3), 1189-1197.