Chapter 7
Genome-wide Association Studies (GWAS)

1. Introduction

There are many different technologies, study designs and analytical tools for identifying genetic risk factors in human and quantitative trait loci in agriculture species. Genome-wide association studies (GWAS) are statistical approaches that involve rapidly scanning markers across the complete sets of DNA, or genomes, of a randomly selected population to find genetic variations associated with a particular trait. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses. For animal and plant geneticists and breeders, genetic associations can help identify quantitative trait loci, clone genes and facilitate marker assisted selection. GWAS typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major human diseases and grain yield in plants, but can equally be applied to any other organisms. Genome-wide association studies have evolved over the last ten years into a powerful tool for investigating the genetic architecture of human diseases and complex traits in agricultural species. In this chapter, we review the key concepts underlying GWAS and introduce various statistical methods used to perform GWAS. We will start with the simple method and then introduce the most advanced methods. The target traits for GWAS can be case-control (binary) outcomes or continuous traits. We will only focus on continuous traits.

Note that GWAS in random populations and QTL mapping in line crosses can use exactly the same set of statistical methods. The only difference is that there is no need to fit population structures (covariates) in QTL mapping because a family of line crosses is often homogenous. Therefore, the statistical methods introduced in this chapter also apply to QTL mapping (Xu 2013).

2. Terminology in genome-wide association studies

2.1 Population structure

A GWAS population is often heterogeneous in background, e.g., Black, White and Asian subpopulations in human, different breeds in large animals and various geographical locations in trees. Some alleles may be associated with a subpopulation for a historical reason, but not necessarily associated with any traits. However, if a model ignores the population structure, these subpopulation-associated alleles will be detected as associated alleles with the trait of interest. A population structure can be diagnosed via a cluster analysis, e.g., the K-mean cluster analysis, or via a mixture analysis using the “STRUCTURE” software package. However, most recently, people try to use principal component analysis (PCA) to identify population structure and use the first few, say 3 or 4, principals as independent variables (covariates) to be included in a GWAS model.

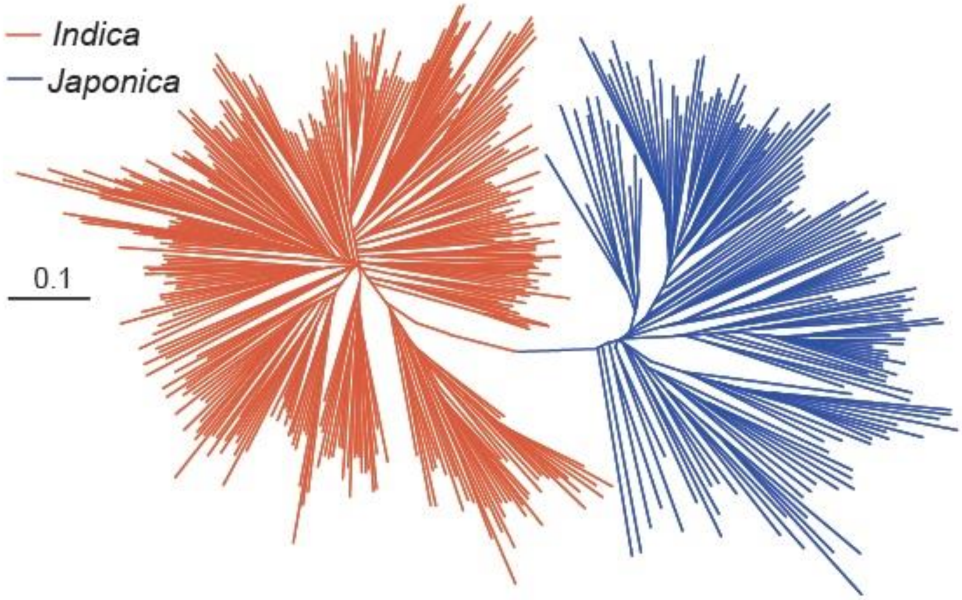


Figure 1. Hierarchical clusters of rice using SNP markers

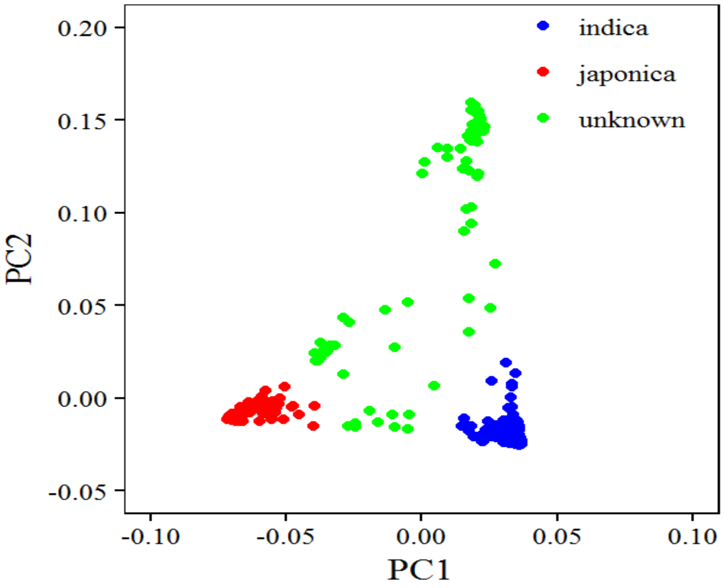


Figure 2. Principal component plots of rice using SNP markers

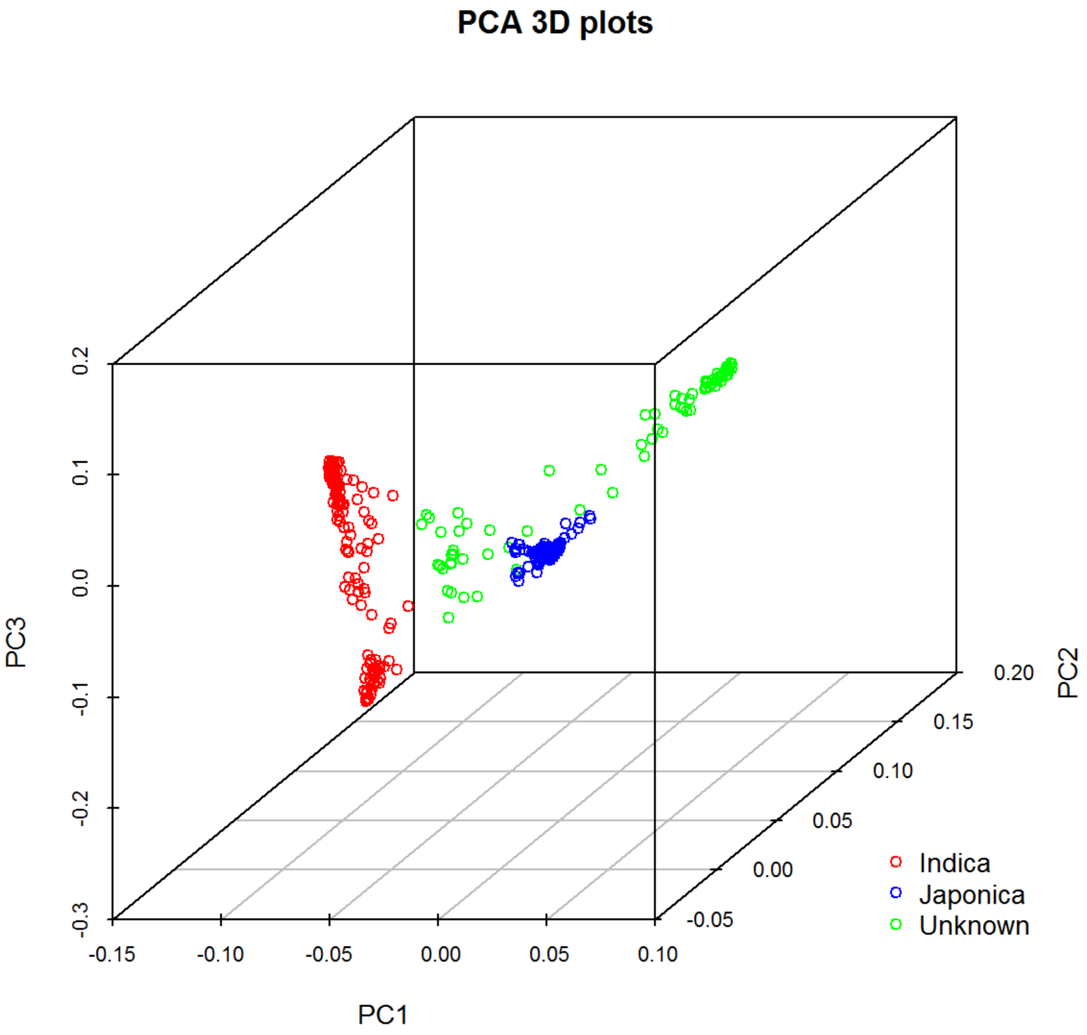


Figure 3. Principal component plots (3D) of rice using SNP markers

The principal components are calculated from the linear combinations of all markers weighted by the eigenvectors of the marker relationship matrix. For example, the first PCA is the linear combination of markers (numerical codes) weighted by the first eigenvector. These PCAs are treated as covariates and included in the linear mixed model of GWAS as fixed effects to control the population structure effects.

2.2. Kinship matrix

Individuals in a GWAS population may be genetically related with various degrees. If the pedigree relationship is known, the pair-wise relatedness of all individuals form an $n \times n$ additive relationship matrix, denoted by *A*. The matrix is the coancetry matrix multiplied by two. This matrix is incorporated in the linear mixed model to capture the genetic covariance structure. If the genetic relatedness of individuals is unknown, we can calculate an empirical additive relationship matrix using markers, called marker inferred kinship matrix denoted by *G*. Matrix *G* no longer represents the expected IBD (identity-by-descent); instead, it is an IBS (identity-by-state) matrix.

2.3. Manhattan plots

GWAS is a genome scanning method to test each marker of the genome for its association with the target trait. The test statistic is then converted to a p-value from the distribution of the test statistic under the model. The p-value is then converted into $-\log_{10}(p)$. This new test statistic is plotted against the marker location on the genome. Such as plot is called the Manhattan plot (Figure 4), which is very similar to the real Manhattan image (Figure 5)

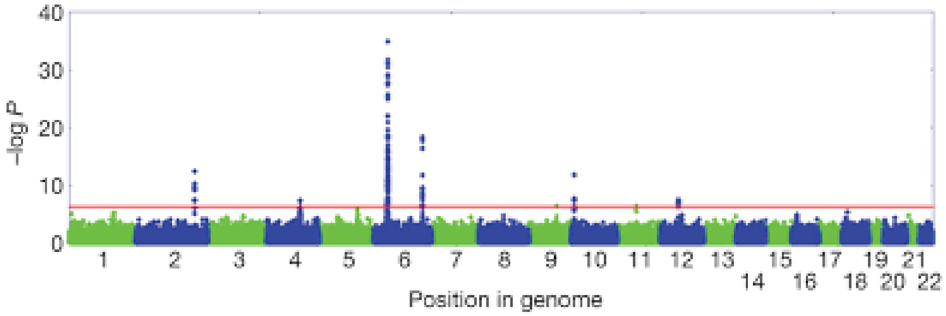


Figure 4. Manhattan plot of a genome-wide association study.

Figure 5. An image of Manhattan in New York City, US

2.4. Quantite-Quantile (QQ) plot

QQ plot is a graphical method to examine the validity of the test statistic, whether the population structure is properly controlled and whether the polygenic background has been removed from the test. If all these things are controlled properly and the test statistic is valid, we should expect that the majority of the loci are neutral and only a few loci are expected to stand out. In this situation, the p-values of the majority loci (neutral loci) will follow a uniform distribution between 0 and 1. The QQ plot is a plot of the observed $-\log_{10}(p)$ (from the data) against the expected $-\log 10(p)$ (extracted from the (0,1) uniform distribution). In the idea situation, the plot will show that the majority of the points are on the diagonal and only a few significant loci show deviations from the diagonal (Figure 6).

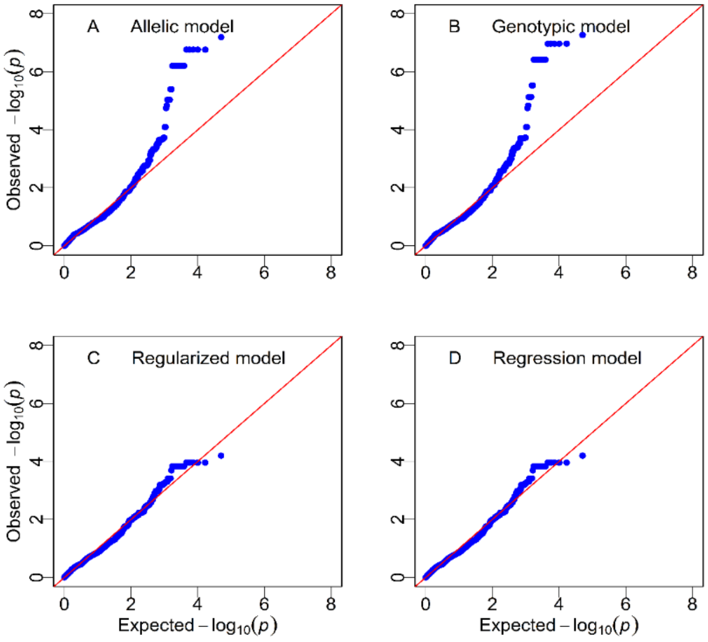


Figure 6. QQ plot drawn from a genome-wide associate study.

The upper panels of Figure 6 are expected to see when the methods are valid and all things are under control. The lower panels show that the methods are not powerful to detect any loci. Figure 7 below shows QQ plots for several models from the same dataset. The gray plot shows the test with correct control of population structure and polygenic background.

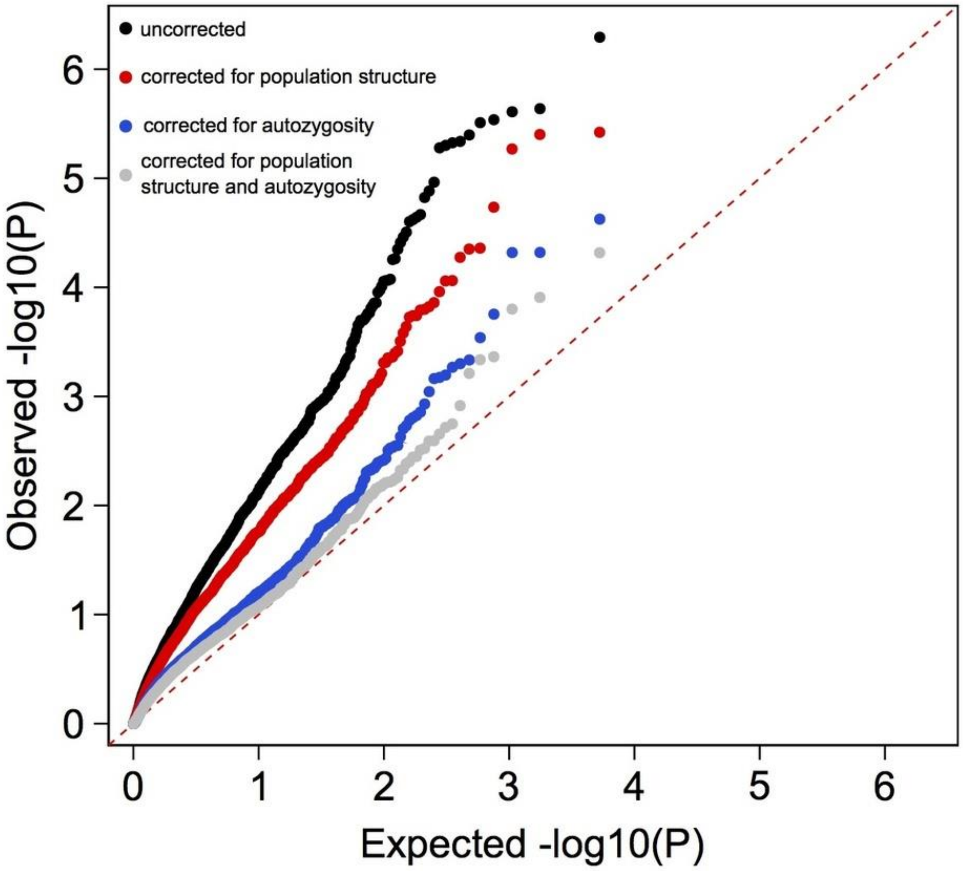


Figure 7. QQ-plots of several different models from the same GWAS data.

3. Simple regression analysis

Simple regression analysis is an approach of genome scanning for testing one marker at a time. The model may contain other fixed effects such as covariates drawn from population structures, but there is no polygenic control. As a result, simple regression method cannot control false positives properly. Let y be the phenotypic values of *n* individuals in the population of interest. The simple linear model is

$$y = X\beta + Z_k\gamma_k + e \tag{1}$$

where *X* is a design matrix of covariates, β is a vector of effects for the covariates to reduce interference from other non-genetic factors, γ_k is the effect of marker *k* for

$k = 1, ..., m$ where *m* is the total number of markers and *e* is a vector of residual errors with an assumed $N(0, \sigma^2)$. For SNP data, Z_k is an $n \times 1$ vector of genotype indicator variables and the *j*th element (individual) is defined as

$$Z_{jk} = \begin{cases} +1 & \text{for } A_1A_1 \\ 0 & \text{for } A_1A_2 \\ -1 & \text{for } A_2A_2 \end{cases} \tag{2}$$

The size of this locus is measured by the total phenotypic variance explained by this locus, which is

$$h_k^2 = \frac{\sigma_{G_k}^2}{\sigma_{G_k}^2 + \sigma^2} = \frac{\text{var}(Z_k)\gamma_k^2}{\text{var}(Z_k)\gamma_k^2 + \sigma^2} = \frac{\sigma_{Z_k}^2 \gamma_k^2}{\sigma_{Z_k}^2 \gamma_k^2 + \sigma^2} \tag{3}$$

where $\text{var}(Z_k) = \sigma_{Z_k}^2$ represents the variance of the *k*th marker and $\sigma_{G_k}^2 = \sigma_{Z_k}^2 \gamma_k^2$ is the genetic variance of the *k*th marker. Estimation and hypothesis test for the simple regression analysis is straightforward. In matrix notation, the estimates are

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma}_k \end{bmatrix} = \begin{bmatrix} X^TX & X^TZ_k \\ Z_k^TX & Z_k^TZ_k \end{bmatrix}^{-1} \begin{bmatrix} X^Ty \\ Z_k^Ty \end{bmatrix} \tag{4}$$

The residual error variance is estimated using

$$\hat{\sigma}^2 = \frac{1}{n-q-1}(y - X\hat{\beta} - Z_k\hat{\gamma}_k)^T(y - X\hat{\beta} - Z_k\hat{\gamma}_k) \tag{5}$$

where *q* is the number of columns (covariates including the intercept) of matrix *X*. When the sample size is sufficiently large, the numerator of equation (5) can be replaced by *n*. The variance-covariance matrix of the estimates is

$$\text{var}\begin{bmatrix} \hat{\beta} \\ \hat{\gamma}_k \end{bmatrix} = \begin{bmatrix} \text{var}(\hat{\beta}) & \text{cov}(\hat{\beta}, \hat{\gamma}_k) \\ \text{cov}(\hat{\gamma}_k, \hat{\beta}) & \text{var}(\hat{\gamma}_k) \end{bmatrix} = \begin{bmatrix} X^TX & X^TZ_k \\ Z_k^TX & Z_k^TZ_k \end{bmatrix}^{-1} \hat{\sigma}^2 = \begin{bmatrix} C_{XX} & C_{XZ} \\ C_{ZX} & C_{ZZ} \end{bmatrix} \hat{\sigma}^2 \tag{6}$$

Note that $\text{var}(\hat{\gamma}_k) = C_{ZZ}\hat{\sigma}^2$ and C_{ZZ} is the lower left diagonal element of matrix *C* and $C_{ZZ} \neq (Z_k^TZ_k)^{-1}$. The Wald test statistic for $H_0 : \gamma_k = 0$ is

$$W_k = \frac{\hat{\gamma}_k^2}{\text{var}(\hat{\gamma}_k)} \tag{7}$$

Under the null hypothesis, W_k follows a Chi-square distribution with 1 degree of freedom (exact distribution is *F* with 1 and $n - q - 1$ degrees of freedom). The p value is calculated using

$$p_k = 1 - \Pr(\chi_1^2 < W_k) \tag{8}$$

where χ_1^2 is a Chi-square variable with 1 degree of freedom. In GWAS, people often convert the *p* value into

$$\eta_k = -\log_{10}(p_k) \tag{9}$$

and report η_k as the test statistic. The entire genome is then scanned. Any markers with η_k larger than a predetermined critical value are declared as being associated with the trait of interest. The SAS function to calculate the p value is

```
data one;
  w=6.5;
  p=1-cdf('chisquare',w,1);
run;
proc print;
run;
```

Obs	w	p
1	6.5	0.010787

The major flaw of the simple regression analysis is the inability to control false positive. Therefore, this method is often used as a “control” to demonstrate the advantages of more advanced methods of GWAS.

3. Mixed model methodology incorporating pedigree information

Zhang et al. (2005) developed a mixed model methodology for mapping QTL using multiple inbred lines of maize. They collected over 400 varieties (inbred lines) of maize derived from the past 70 years in the US. The pedigree relationship of the lines was known exactly. Their mixed model included covariates (fixed effects), a marker (treated as a random effect) and a polygenic effect (random effect) whose covariance structure is determined by the additive relationship matrix inferred from the pedigree. The model is described as

$$y = X\beta + Z_k\gamma_k + W\alpha + e \tag{10}$$

The followings define entries of the above model. $X\beta$ remains the same as defined earlier in the simple regression analysis. Z_k is a design matrix with *n* rows and *s* columns, where *s* is the number of founders in the pedigree. This matrix is inferred from marker information. It contains the probabilities of lines inheriting from the founder alleles given the marker genotypes. Therefore, γ_k is a $s\times 1$ vector of founder effects. W has the same dimension as Z_k but it is the expectation of Z_k across all loci. α is an $s\times 1$ vector of polygenic effects of all founders. *e* is a vector of residual errors with an assumed $N(0, I\sigma^2)$ distribution (multivariate normal distribution). Both the marker specific effects and polygenic effects are assumed to be random with $\gamma_k \sim N(0, I\sigma_k^2)$ and $\alpha \sim N(0, I\sigma_a^2)$, respectively. Under these assumptions, the expectation of *y* is $E(y) = X\beta$ and the variance is

$$\text{var}(y) = V = Z_k Z_k^T \sigma_k^2 + WW^T\sigma_\alpha^2 + I\sigma^2 \tag{11}$$

Zhang et al. (2005) defined $\Pi_k = Z_k Z_k^T$ and called it the marker specific IBD matrix, $\Pi = WW^T$ and called it the additive relationship matrix, and rewrote the variance as

$$\text{var}(y) = V = \Pi_k\sigma_k^2 + \Pi\sigma_\alpha^2 + I\sigma^2 \tag{12}$$

They used the restricted maximum likelihood (REML) method to estimate the variance components and performed a likelihood ratio test for $H_0 : \sigma_k^2 = 0$,

$$\lambda_k = -2\left[L(0) - L(\hat{\sigma}_k^2) \right] \tag{13}$$

where $L(0)$ is the log likelihood function evaluated under the null model (the pure polygenic model) and $L(\hat{\sigma}_k^2)$ is the log likelihood function evaluated under the alternative model. Under the null model, λ_k follows a mixture of two chi-square distributions with equal mixing proportion, $\lambda_k \sim 0.5\chi_0^2 + 0.5\chi_1^2$, where χ_0^2 represents a point mass at 0. The p-value is calculated as follows,

$$p_k = \begin{cases} 1 & \text{for} \quad \lambda_k = 0 \\ 0.5\Pr(\chi_1^2 > \lambda_k) & \text{for} \quad \lambda_k > 0 \end{cases} \tag{14}$$

In practice, the calculated λ_k may be negative due to floating point error. Therefore, we often truncate λ_k at zero when it has a very small value, say $\lambda_k = 0$ if $\lambda_k < 10^{-8}$. Alternatively, we can modify the above equation using

$$p_k = \begin{cases} 1 & \text{for} \quad \lambda_k \le 10^{-8} \\ 0.5\Pr(\chi_1^2 > \lambda_k) & \text{for} \quad \lambda_k > 10^{-8} \end{cases} \tag{15}$$

The method was implemented using the mixed procedure in SAS (PROC MIXED). Since the title of the paper did not include the key word “GWAS”, the paper was published in Genetics (not Nature Genetics) and Zhang et al. (2005) did not provide a user friendly software package, nobody knows its exist. The paper has not been cited after 12 years of its publication, although it was published a year earlier than the Q+K mixed model of Yu et al. (2006).

4. Mixed model methodology using marker-inferred kinship

A year later after Zhang et al. (2005), Yu et al. (2006) published a similar mixed model procedure for GWAS, but they used a marker inferred kinship matrix *K* in place of Zhang’s Π matrix. They replaced the *X* matrix by a *Q* matrix (population structure) and treated γ_k as a fixed effect and directly estimated and tested this marker effect. They also implemented the GWAS in SAS using PROC MIXED. The model was written in the following form

$$y = Q\alpha + X\beta + Zu + e \tag{16}$$

where $Q\alpha$ is the fixed effect part for controlling population structure, $X\beta$ is the marker genotype indicator multiplied by marker effect (replacing our $Z_k\gamma_k$), Zu captures the polygene (replacing our $W\alpha$). The expectation of *y* is $E(y) = Q\alpha + X\beta$ and the variance is

$$\text{var}(y) = V = ZKZ^T\sigma_u^2 + I\sigma^2 \tag{17}$$

where *K* is the kinship matrix and it is the covariance structure $\text{var}(u) = K\sigma_u^2$. The polygenic variance is σ_u^2. The *Z* matrix depends on the type of populations and often take the form of identity if *u* is an $n \times 1$ polygenic effects. The “kinship” matrix is calculated from markers of the whole genome. The genotype indicator for marker *k* is

X and the marker effect is β (marker index is ignored). The test statistics is again the Wald test defined as

$$W = \frac{\hat{\beta}^2}{\mathrm{var}(\hat{\beta})} \tag{18}$$

Such a test is done for every marker and the entire genome is scanned to complete the GWAS. This mixed model is called the Q + K model and is considered the first appropriate method of GWAS. Jianming Yu was eventually honored the “father” of the modern GWAS procedure when he was still a postdoc. The linear mixed model procedure (they call it mixed linear model) is implemented in a software package called TASSEL (Bradbury et al. 2007).

5. Iteratively reweighted least squares

The mixed model in the original form is computationally very expensive because the inclusion of the kinship matrix. Kang et al. (2008) proposed an eigen-decomposition method to improve the computational speed. They first decomposed the *K* matrix to simplify the computation and then estimated the polygenic and residual variance components for each marker screened by treating the variance ratio $\lambda = \sigma_u^2 / \sigma^2$ as the parameter. The method is called efficient mixed model association (EMMA). Zhou and Stevens (2012) further investigated the method and gave a new name called genome-wide efficient mixed model association (GEMMA). Other than the recursive algorithm to evaluate the likelihood function and the beautiful presentation, I do not see much value of GEMMA compared with EMMA. I originally thought that EMMA treated the estimated δ from the pure polygenic model (no marker effect is included) as a constant and used it to scan the genome without further estimation of δ. I thought that GEMMA re-estimated δ for each marker and this was the improvement of GEMMA over EMMA. In fact, EMMA is also re-estimating δ for each marker. This puzzled me how GEMMA was published in Nature Genetics.

I now introduce these methods using my notation with some modification to ease the understanding from readers. The linear mixed model for the *k*th marker ($k = 1, ..., m$) is

$$y = X\beta + Z_k\gamma_k + \xi + \varepsilon \tag{19}$$

where X is a design matrix for *q* fixed effects (including population structure) and β are the fixed effects themselves. The fixed effects are included in the model to capture effects of any covariates that are not relevant to marker effects, e.g., population structure, age effect etc. The remaining variables in the model are defined as follows. The Z_k vector stores the genotype indicator variable for all individuals and the *j*th element of it is defined as $Z_{jk} = 0$ for the homozygote of the major allele, $Z_{jk} = 1$ for the heterozygote and $Z_{jk} = 2$ for the homozygote of the minor allele. The coding system is arbitrary and you can use the $Z_{jk} = \{-1, 0, 1\}$ coding system. The effect of the *k*th marker is denoted by γ_k and treated as fixed effect. The polygenic term ξ is assumed to be random with variance

$$\mathrm{var}(\xi) = K\sigma_\xi^2 \tag{20}$$

where K is the covariance structure (kinship matrix) and σ_ξ^2 is the polygenic variance. You may wonder how the *K* matrix is derived from the markers. We can write ξ explicitly using

$$\xi = \sum_{k=1}^{m} Z_k \delta_k \tag{21}$$

The Z_k's in equations (19) and (21) are the same, but the γ_k and δ_k in the two equations are different because they are from different distributions; the former is a fixed effect (no distribution) and the latter is sampled from a normal distribution, i.e., $\delta_k \sim N(0, \sigma_\xi^2)$ for all $k = 1, ..., m$. So, each marker is assumed to have two effects, one as a major effect and one contributes as a component of the polygene. The major effect determines the association but the polygenic component is small and assume to be sampled from the same normal distribution for all markers. The variance of ξ is

$$\mathrm{var}(\xi) = \sum_{k=1}^{m} Z_k \,\mathrm{var}(\delta_k) Z_k^T = \sum_{k=1}^{m} Z_k Z_k^T \sigma_\xi^2 = K\sigma_\xi^2 \tag{22}$$

Therefore,

$$K = \sum_{k=1}^{m} Z_k Z_k^T \tag{23}$$

If *m* is small, we can directly use matrix multiplication to calculate *K* without using the summation. In other words, $K = ZZ^T$, where Z is an $n \times m$ matrix. There are more than a dozen different forms of *K* in the literature, but this is the only one directly derived from the properties of variance. The residual errors are assumed to be $\varepsilon \sim N(0, I\sigma^2)$. Going back to model (19), the expectation is

$$E(y) = X\beta + Z_k \gamma_k \tag{24}$$

and the variance is

$$\mathrm{var}(y) = K\sigma_\xi^2 + I\sigma^2 = (K\lambda_k + I)\sigma^2 \tag{25}$$

where $\lambda_k = \sigma_\xi^2 / \sigma^2$ is the variance ratio. This ratio is marker specific because it depends on the fixed effects, which include the *k*th marker effect. We use the restricted maximum likelihood (REML) to estimate parameters. Evaluation of the restricted log likelihood function can be very costly when the sample size is very large. Let $T_k = (X \,\|\, Z_k)$ be the horizontal concatenation of the two matrices, equivalent to the cbind(X,Zk) function in R. The log likelihood function for REML is

$$L(\lambda_k) = -\frac{1}{2}\ln|H_k| - \frac{1}{2}\ln|T_k^T H_k^{-1} T_k| - \frac{n-r}{2}\ln(y^T P_k y) \tag{26}$$

where

$$H_k = K\lambda_k + I \tag{27}$$

and

$$P_k = H_k^{-1} - H_k^{-1} T_k (T_k^T H_k^{-1} T_k)^{-1} T_k^T H_k^{-1} \tag{28}$$

The fixed effect and residual variance have been absorbed (profiled likelihood function). Therefore, this likelihood function only involves one parameter λ_k. Any numerical algorithm can be used to search for the REML estimate, denoted by $\hat{\lambda}$. Once λ is replaced by $\hat{\lambda}$, we can find β and γ_k using

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma}_k \end{bmatrix} = \begin{bmatrix} X^T H_k^{-1} X & X^T H_k^{-1} Z_k \\ Z_k^T H_k^{-1} X & Z_k^T H_k^{-1} Z_k \end{bmatrix}^{-1} \begin{bmatrix} X^T H_k^{-1} y \\ Z_k^T H_k^{-1} y \end{bmatrix} \tag{29}$$

The residual error variance is estimated using

$$\hat{\sigma}^2 = \frac{1}{n-q-1}(y - X\hat{\beta} - Z_k\hat{\gamma}_k)^T H_k^{-1}(y - X\hat{\beta} - Z_k\hat{\gamma}_k) \tag{30}$$

The variance matrix of the estimated fixed effects is

$$\mathrm{var}\begin{bmatrix} \hat{\beta} \\ \hat{\gamma}_k \end{bmatrix} = \begin{bmatrix} \mathrm{var}(\hat{\beta}) & \mathrm{cov}(\hat{\beta}, \hat{\gamma}_k) \\ \mathrm{cov}(\hat{\gamma}_k, \hat{\beta}) & \mathrm{var}(\hat{\gamma}_k) \end{bmatrix} = \begin{bmatrix} X^T H_k^{-1} X & X^T H_k^{-1} Z_k \\ Z_k^T H_k^{-1} X & Z_k^T H_k^{-1} Z_k \end{bmatrix}^{-1} \hat{\sigma}^2 \tag{31}$$

The Wald test statistic for $H_0 : \gamma_k = 0$ is

$$W_k = \frac{\hat{\gamma}_k^2}{\mathrm{var}(\hat{\gamma}_k)} \tag{32}$$

Under the null hypothesis, W_k follows a Chi-square distribution with 1 degree of freedom (exact distribution is *F* with 1 and $n-q-1$ degrees of freedom). The analysis is identical to the simple regression analysis except that there is a weight matrix H_k^{-1} involved in the linear mixed model. The mixed model requires multiple iterations for estimating λ_k, which is marker specific, and the entire genome has to be scanned. This method is called the exact method by Zhou and Stevens (2012). It is equivalent to the iteratively reweighted least squares (IRWLS) method where the weights keep changing over the iterations.

Corresponding to the exact method, there is an approximate method, in which $\lambda_k = \hat{\lambda}$ where $\hat{\lambda}$ is the estimated variance ratio from the pure polygenic model. In other words, the variance ratio is fixed for all loci. The restricted log likelihood function of the pure polygenic model is

$$L(\lambda) = -\frac{1}{2}\ln|H| - \frac{1}{2}\ln|X^T H^{-1} X| - \frac{n-r}{2}\ln(y^T P y) \tag{33}$$

where

$$H = K\lambda + I \tag{34}$$

and

$$P = H^{-1} - H^{-1}X(X^T H^{-1} X)^{-1} X^T H^{-1} \tag{35}$$

Such an approximated method is a typical weighted least square method. The weight matrix H^{-1} is only calculated once from the pure polygenic model and will be kept as constant when the genome is scanned. This method is the same as the classical weighted least squares (WLS) method.

The eigen-decomposition algorithm has been implemented in both the EMMA and GEMMA methods to ease the computational load. Eigen decomposition is performed on the kinship matrix *K* as

$$K = UDU^T \tag{36}$$

where U is an $n \times n$ matrix of eigenvectors and D is a diagonal matrix holding the eigenvalues, $D = \mathrm{diag}\{\delta_1, ..., \delta_n\}$. The eigenvector matrix is orthogonal in the sense that $UU^T = I$ (i.e., $U^{-1} = U^T$). The *H* matrix is

$$H = K\lambda + I = UDU^T + I = U(D\lambda + I)U^T \tag{37}$$

Therefore, the log determinant of matrix *H* is

$$\ln|H| = \ln|D\lambda + I| = \sum_{j=1}^{n} \ln(\delta_j \lambda + 1) \tag{38}$$

because *D* is a diagonal matrix and $UU^T = I$. Various quadratic forms are involve in the likelihood function 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

$$a^T H^{-1} b = a^T U (D\lambda + I)^{-1} U^T b = a^{*T} (D\lambda + I)^{-1} b^* = \sum_{j=1}^{n} a_j^{*T} b_j^* (\delta_j \lambda + 1)^{-1} \tag{39}$$

where $a^* = U^T a$ and $b^* = U^T b$. With such a decomposition, matrix inversion has been replaced by the above summations and the computational speed has been improved significantly. As a result, the exact method by updating λ_k for each locus is also possible (Zhou and Stevens 2012), although the improvement of the exact method over the approximate method has never been shown to be sufficiently significant to change the conclusions of a GWAS study.

6. Further improvement of statistical power

The polygenic background control is similar to the composite interval mapping using co-factors to control the background effects in QTL mapping. However, if the number of markers used to calculate the kinship matrix is small, the effect from the polygene (δ_k) will compete with the effect of the *k*th marker (γ_k). This will lead to a decreased power. The decreased false positive will be traded off by a decreased power, which is not a desired property of the mixed model method. Can we avoid or reduce the competition? Wang et al. (2016) proposed a method that can boost the power but still keep the low false positive rate. The method first calculates the BLUP value for each δ_k using

$$\hat{\delta}_k = \mathrm{E}(\delta_k | y) = Z_k \hat{\lambda} (K\hat{\lambda} + I)^{-1} (y - X\hat{\beta}) \tag{40}$$

From this BLUP value, we define $\hat{\xi}_k = Z_k \hat{\delta}_k$ as the polygenic component contributed by marker *k*. Using this term, we can adjust the original model (19) by

$$y = X\beta + Z_k \gamma_k + \xi - \hat{\xi}_k + \varepsilon \tag{41}$$

A rearrangement of this equation lead to

$$y + \hat{\xi}_k = X\beta + Z_k \gamma_k + \xi + \varepsilon \tag{42}$$

Let $y_k = y + \hat{\xi}_k$, we have

$$y_k = X\beta + Z_k \gamma_k + \xi + \varepsilon \tag{43}$$

The right hand side of the equation is the original model (19), but the left hand side is an adjusted phenotype. Such an adjustment is not necessary when the number of markers used to calculate the kinship matrix is large, say > 100,000. Since the kinship matrix defined this way, $K = ZZ^T$, depends on the number of loci, a very large number of loci will produce a very large value for the *K* matrix, which eventually leads to an extremely small $\hat{\delta}_k$, as shown in equation (40). A small $\hat{\delta}_k$ will not be able to compete with γ_k. However, if the number of loci is small, the competition can be serious and the adjustment is necessary.

7. A random model approach to genome-wide association studies

The above adjustment has been implemented by Wang et al (2016) and Wei et al. (2016). It works very well in boosting the power. However, the method is still *ad hoc* in the sense that the kinship matrix is not adjusted. Yao et al. (2017, unpublished result) developed an exact method that adjust the kinship matrix every time a new marker is scanned. The method was developed for QTL mapping but applies to GWAS equally well. This section is beyond the scope of this course and will not be taught in the class. However, students with very strong statistical background are encouraged to read this section as supplementary material (the result has not been published yet).

Let *y* be the phenotypic values of a trait (an $n \times 1$ vector where *n* is the sample size) and define Z_k as a design matrix (dummy variables) for the genotype of locus *k*. We now introduce the following mixed model for *y*,

$$y = X\beta + Z_k\gamma_k + \sum_{k' \neq \kappa}^{m} Z_{k'}\gamma_{k'} + \varepsilon \tag{44}$$

where $\kappa = \{k-1, k, k+1\}$ is a set of three markers (a triplet) surrounding the *k*th marker, β represents some fixed effects, $\gamma_k \sim N(0, \phi_k^2)$ is a random effect of marker *k* with variance ϕ_k^2 and $\gamma_{k'} \sim N(0, \phi^2 / (m-3))$ is a random effect of marker $k' \neq \kappa$ with variance $\phi^2 / (m-3)$. We call $\gamma_{k'}$ a polygenic effect. The number 3 in $m-3$ defines a triplet (three markers around marker *k* that are excluded from the polygene to avoid competition between the *k*th marker and its counterpart in the polygene). The dummy variables are defined as an $n \times 2$ matrix if the population is BC or RIL population (with two possible genotypes per marker) and an $n \times 3$ matrix if it is an F$_2$ population (with three different genotypes per marker). In general, if there are q genotypes per marker, the design matrix should have a dimension of $n \times q$. For individual *j*, we define the *j*th row of matrix Z_k by $Z_{jk} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$ if the individual's genotype for this marker is A_1A_1, by $Z_{jk} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$ if the individual's genotype is A_1A_2 and by $Z_{jk} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$ the genotype is A_2A_2. The expectation of *y* is $\mathrm{E}(y) = X\beta$ and the variance matrix is $\mathrm{var}(y) = V$, where

$$
\begin{aligned}
V &= Z_k Z_k^T \phi_k^2 + \sum_{k'\neq\kappa}^{m} Z_{k'} Z_{k'}^T \phi^2 / (m-3) + I\sigma^2 \\
&= Z_k Z_k^T \phi_k^2 + \frac{1}{m-3} \sum_{k'\neq\kappa}^{m} Z_{k'} Z_{k'}^T \phi^2 + I\sigma^2 \\
&= K_{-\kappa}\phi^2 + Z_k Z_k^T \phi_k^2 + I\sigma^2
\end{aligned}
\tag{45}
$$

where

$$
K_{-\kappa} = \frac{1}{m-3} \sum_{k'\neq\kappa}^{m} Z_{k'} Z_{k'}^T
\tag{46}
$$

is a kinship matrix inferred from $m-3$ markers that excludes the markers in set $\kappa = \{k-1, k, k+1\}$. This marker specific kinship matrix will change for every marker scanned. This will present a tremendous increase in computational burden. Therefore, we will replace this marker specific kinship matrix by a kinship matrix inferred from all markers of the entire genome. Let us define the kinship matrix inferred from all *m* markers by

$$
K = \frac{1}{m} \sum_{k'=1}^{m} Z_{k'} Z_{k'}^T
\tag{47}
$$

After some algebraic manipulation (see Appendix A), we show that,

$$
\begin{aligned}
V &= \frac{m\phi^2}{m-3} K - \frac{\phi^2}{m-3} Z_{k-1} Z_{k-1}^T + \left(\phi_k^2 - \frac{\phi^2}{m-3} \right) Z_k Z_k^T - \frac{\phi^2}{m-3} Z_{k+1} Z_{k+1}^T + I\sigma^2 \\
&= \left\{ \frac{m\lambda}{m-3} K + I - \frac{\lambda}{m-3} Z_{k-1} Z_{k-1}^T + \left(\lambda_k - \frac{\lambda}{m-3} \right) Z_k Z_k^T - \frac{\lambda}{m-3} Z_{k+1} Z_{k+1}^T \right\} \sigma^2
\end{aligned}
\tag{48}
$$

where $\lambda = \phi^2 / \sigma^2$ and $\lambda_k = \phi_k^2 / \sigma^2$ are the variance ratios. Define

$$
c = \lambda m / (m-3)
\tag{49}
$$

and

$$
C_k = \begin{bmatrix} -\lambda/(m-3) I_{q\times q} & 0 & 0 \\ 0 & [\lambda_k - \lambda/(m-3)] I_{q\times q} & 0 \\ 0 & 0 & -\lambda/(m-3) I_{q\times q} \end{bmatrix}
\tag{50}
$$

Let $Z_\kappa = \begin{bmatrix} Z_{k-1} & Z_k & Z_{k+1} \end{bmatrix}$ be the design matrix for the triplet. The *V* matrix can be rewritten by

$$
V = (cK + I + Z_\kappa C_k Z_\kappa^T)\sigma^2 = (H + Z_\kappa C_k Z_\kappa^T)\sigma^2 = H_k \sigma^2
\tag{51}
$$

where $H = cK + I$ and $H_k = H + Z_\kappa C_k Z_\kappa^T$. If there are *q* genotypes per locus, the number of columns of matrix Z_κ is $3q$ and the dimension of matrix C_k is $(3q)\times(3q)$. For example, in QTL mapping for F2 population, the number of genotypes per locus is $q = 3$. Therefore, Z_κ is a $n\times 9$ matrix and C_k is a 9×9 matrix. In summary, the expectation and variance of *y* given in model (44) are

$$
\begin{aligned}
\mathrm{E}(y) &= \mu = X\beta \\
\mathrm{var}(y) &= V = H_k \sigma^2
\end{aligned}
\tag{52}
$$

The above terms are used to construct the following restricted likelihood function for parameters $\theta = \{\theta_1, \theta_2\}$, where $\theta_1 = \{\beta, \sigma^2\}$ and $\theta_2 = \{\lambda, \lambda_k\}$. The log likelihood function is

$$\begin{aligned} L(\theta) &= -\frac{1}{2}\ln|V| - \frac{1}{2}\ln|X^T V^{-1} X| - \frac{1}{2}(y-\mu)V^{-1}(y-\mu) \\ &= -\frac{1}{2}\ln|\sigma^2 H_k| - \frac{1}{2}\ln|\sigma^{-2} X^T H_k^{-1} X| - \frac{1}{2\sigma^2}(y - X\beta)H_k^{-1}(y - X\beta) \\ &= -\frac{1}{2}\ln|H_k| - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\ln|X^T H_k^{-1} X| + \frac{r}{2}\ln|\sigma^2| - \frac{1}{2\sigma^2}(y - X\beta)H_k^{-1}(y - X\beta) \\ &= -\frac{1}{2}\ln|H_k| - \frac{n-r}{2}\ln(\sigma^2) - \frac{1}{2}\ln|X^T H_k^{-1} X| - \frac{1}{2\sigma^2}(y - X\beta)H_k^{-1}(y - X\beta) \end{aligned} \quad (53)$$

Note that H_k is only a function of $\theta_2 = \{\lambda, \lambda_k\}$ and contains no information about $\theta_1 = \{\beta, \sigma^2\}$. Therefore, given H_k the likelihood function for $\theta_1 = \{\beta, \sigma^2\}$ is

$$L(\theta_1) = -\frac{n-r}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)H_k^{-1}(y - X\beta) \quad (54)$$

Taking the partial derivatives of this likelihood function with respect θ_1 and letting them equal zeros, we get

$$\hat{\beta} = (X^T H_k^{-1} X)^{-1} X^T H_k^{-1} y \quad (55)$$

and

$$\hat{\sigma}^2 = \frac{1}{n-r}(y - X\hat{\beta})^T H_k^{-1}(y - X\hat{\beta}) = \frac{1}{n-r} y^T P_k y \quad (56)$$

where

$$P_k = H_k^{-1} - H_k^{-1} X_k (X_k^T H_k^{-1} X_k)^{-1} X_k^T H_k^{-1} \quad (57)$$

Substituting θ_1 in equation (53) by $\hat{\theta}_1 = \{\hat{\beta}, \hat{\sigma}^2\}$, we have

$$\begin{aligned} L(\theta_2) &= -\frac{1}{2}\ln|H_k| - \frac{1}{2}(n-r)\ln(\hat{\sigma}^2) - \frac{1}{2}\ln|X^T H_k^{-1} X| - \frac{1}{2\hat{\sigma}^2}(y - X\hat{\beta})H_k^{-1}(y - X\hat{\beta}) \\ &= -\frac{1}{2}\ln|H_k| - \frac{1}{2}\ln|X^T H_k^{-1} X| - \frac{1}{2}(n-r)\ln(y^T P_k y) + \frac{1}{2}(n-r)\ln(n-r) - \frac{1}{2}(n-r) \end{aligned}$$

(58)

This likelihood function contains only parameters $\theta_2 = \{\lambda, \lambda_k\}$ because $\theta_1 = \{\beta, \sigma^2\}$ have been absorbed and it is called the profiled likelihood function. After deleting the last two terms (constants), we have

$$L(\lambda, \lambda_k) = -\frac{1}{2}\ln|H_k| - \frac{1}{2}\ln|X^T H_k^{-1} X| - \frac{1}{2}(n-r)\ln(y^T P_k y) \quad (59)$$

A numeric solution of $\theta_2 = \{\lambda, \lambda_k\}$ can be found iteratively using the Newton-Raphson algorithm,

$$\begin{bmatrix} \lambda^{(t+1)} \\ \lambda_k^{(t+1)} \end{bmatrix} = \begin{bmatrix} \lambda^{(t)} \\ \lambda_k^{(t)} \end{bmatrix} - \left[\frac{\partial^2 L(\theta_2^{(t)})}{\partial \theta_2^2}\right]^{-1} \left[\frac{\partial L(\theta_2^{(t)})}{\partial \theta_2}\right] \quad (60)$$

Once the iteration process converges, the solution is the REML estimate of $\theta_2 = \{\lambda, \lambda_k\}$.

Looking back at the REML estimation of $\theta_2 = \{\lambda, \lambda_k\}$, the best linear unbiased estimate (BLUE) of the fixed effects and the estimated residual error variance, we realized that the H_k matrix never occurs alone. It always appears in the form of $A^T H_k^{-1} B$ or $|H_k|$, where *A* and *B* are low ranking matrices or vectors. This makes the computation of the quadratic forms very efficient by taking advantage of eigen-decomposition of the kinship matrix *K*, which is $K = UDU^T$ where $D = \text{diag}\{\delta_1, \delta_2, ..., \delta_n\}$ is a diagonal matrix of eigenvalues and U is the eigenvector matrix, where $UU^T = UU^{-1} = I$. We now use eigen-decomposition to calculate the inverse and determinant of matrix $H = cK + I$ and thus the corresponding inverse and determinant of matrix $H_k = K + Z_\kappa C_k Z_\kappa^T$.

$$H = cK + I = cUDU^T + I = U(cD + I)U^T \tag{61}$$

The inverse of H is then

$$H^{-1} = U(cD + I)^{-1} U^T \tag{62}$$

Since $cD + I$ is diagonal, its inverse is very easy to calculate. The determinant of *H* is

$$|H| = |U(cD + I)U^T| = |cD + I||UU^T| = |cD + I| \tag{63}$$

where $cD + I$ is a diagonal matrix and the log determinant of matrix *H* is

$$\ln|H| = \ln|cD + I| = \sum_{j=1}^{n} \ln(c\delta_j + 1) = \sum_{j=1}^{n} \ln\left[m(m-3)^{-1}\lambda\delta_j + 1\right] \tag{64}$$

where the very last expression is due to $c = m(m-1)^{-1}\lambda$. The restricted log likelihood function also involves various quadratic terms in the form of $A^T H^{-1} B$, where *A* and *B* are low ranking matrices or vectors, for example, $X^T H^{-1} X$, $X^T H^{-1} y$ and $y^T H^{-1} y$. Using eigen-decomposition, we can rewrite the quadratic form by

$$A^T H^{-1} B = A^T U (cD + I)^{-1} U^T B = a^T (cD + I)^{-1} b = \sum_{j=1}^{n} a_j^T b_j (c\delta_j + 1)^{-1} \tag{65}$$

where $a = U^T A$ and $b = U^T B$ are computed just once using eigenvectors of the original kinship matrix *K*. Now, let us look at the quadratic form $A^T H_k^{-1} B$ and the log determinant $|H_k|$. Recall that

$$H_k = H + Z_\kappa C_k Z_\kappa^T \tag{66}$$

is a highly structured matrix and a special algorithm can be used to find the inverse and determinant. Based on the Sherman-Morris-Woodbury matrix identity (citation), the determinant of H_k can be expressed as

$$\begin{aligned} |H_k| &= |H||Z_\kappa^T H^{-1} Z_\kappa + C_k^{-1}||C_k| \\ &= |H||Z_\kappa^T H^{-1} Z_\kappa C_k + I||C_k^{-1}||C_k| \\ &= |H||Z_\kappa^T H^{-1} Z_\kappa C_k + I| \end{aligned} \tag{67}$$

and

$$\ln|H_k| = \ln|H| + \ln|Z_k^T H^{-1} Z_k C_k + I| \tag{68}$$

Therefore, we only need to obtain the determinant by updating the determinant of matrix *H* with a factor $|Z_\kappa^T H^{-1} Z_\kappa C_k + I|$, which is the determinant of a $(3q) \times (3q)$ matrix where $3q$ is the number of columns of matrix Z_κ (3 times the number of genotypes per marker). The Sherman-Morris-Woodbury matrix identity (citation) also shows

$$\begin{aligned} H_k^{-1} &= (H + Z_\kappa C_k Z_\kappa^T)^{-1} \\ &= H^{-1} - H^{-1} Z_\kappa (Z_\kappa^T H^{-1} Z_\kappa + C_k^{-1})^{-1} Z_\kappa^T H^{-1} \\ &= H^{-1} - H^{-1} Z_\kappa C_k (Z_\kappa^T H^{-1} Z_\kappa C_k + I)^{-1} Z_\kappa^T H^{-1} \end{aligned} \qquad (69)$$

This identity means that we do not need to calculate H_k^{-1} anew but simply update from H^{-1}. The inverse matrix involved in the updating, $(Z_\kappa^T H^{-1} Z_\kappa C_k + I)^{-1}$, is a $(3q)\times(3q)$ matrix and is easy to calculate. Having defined H_k^{-1}, we now need to find the quadratic forms of this inverse matrix,

$$A^T H_k^{-1} B = A^T H^{-1} B - A^T H^{-1} Z_\kappa C_k (Z_\kappa^T H^{-1} Z_\kappa C_k + I)^{-1} Z_\kappa^T H^{-1} B \qquad (70)$$

Let $z_\kappa = U^T Z_\kappa$, $x = U^T X$ and so on, i.e., the lower case variable is the eigen-transformed upper case variable. The above quadratic form can be calculated as

$$A^T H_k^{-1} B = a^T (cD+I)^{-1} b - a^T (cD+I)^{-1} z_\kappa C_k \left[z_\kappa^T (cD+I)^{-1} z_\kappa C_k + I \right]^{-1} z_\kappa^T (cD+I)^{-1} b \qquad (71)$$

With eigen-decomposition, all the quadratic forms and determinants are expressed as sum of terms across all individual observations. The cost saving can be enormous.

Once the variance components are estimated from REML, we can calculate the best linear unbiased prediction (BLUP) of γ_k using,

$$\begin{aligned} \hat{\gamma}_k &= \phi_k Z_k^T V^{-1}(y - X\beta) \\ &= \lambda_k Z_k^T H_k^{-1}(y - X\beta) \\ &= \lambda_k Z_k^T H_k^{-1} y - \lambda_k Z_k^T H_k^{-1} X\beta \end{aligned} \qquad (72)$$

Note that Z_k is used here, not Z_κ, and the two matrices have different dimensions, the former is $n\times q$ and the latter is $n\times(3q)$. The variance matrix of the BLUP of marker effects is

$$\mathrm{var}(\hat{\gamma}_k) = (\lambda_k I - \lambda_k Z_k^T H_k^{-1} Z_k \lambda_k)\sigma^2 \qquad (73)$$

Let us define the additive and dominance effects of locus *k* by linear combinations of γ_k

$$\begin{bmatrix} a_k \\ d_k \end{bmatrix} = \begin{bmatrix} -0.5 & 0 & 0.5 \\ -0.5 & 1 & -0.5 \end{bmatrix} \begin{bmatrix} \gamma_{1k} \\ \gamma_{2k} \\ \gamma_{3k} \end{bmatrix} = \begin{bmatrix} -0.5\gamma_{1k} + 0.5\gamma_{3k} \\ -0.5\gamma_{1k} + \gamma_{2k} - 0.5\gamma_{3k} \end{bmatrix} \qquad (74)$$

Define $g_k = \begin{bmatrix} a_k & d_k \end{bmatrix}^T$ and

$$L = \begin{bmatrix} -0.5 & 0 & 0.5 \\ -0.5 & 1 & -0.5 \end{bmatrix} \qquad (75)$$

We have $g_k = L\gamma_k$ and thus $\mathrm{var}(\hat{g}_k) = L\,\mathrm{var}(\hat{\gamma}_k)L^T$. The Wald test for $H_0 : g_k = 0$ is

$$W_k = \hat{g}_k^T \left[\mathrm{var}(\hat{g}_k)\right]^{-1} \hat{g}_k = \hat{\gamma}_k^T L^T \left[L\,\mathrm{var}(\hat{\gamma}_k) L^T \right]^{-1} L\hat{\gamma}_k \qquad (76)$$

Under the null hypothesis, this test statistic will follow approximately a Chi-square distribution with degree of freedom

$$d_k = \mathrm{rank}(L) - \frac{1}{\phi_k^2}\mathrm{tr}\left[(LL^T)^{-1}L\,\mathrm{var}(\hat{\gamma}_k)L^T\right] \tag{77}$$

where $\mathrm{rank}(L) = 2$ is the number of rows of matrix L. Therefore, the p-value for marker *k* is

$$p_k = 1 - \Pr(\chi^2_{d_k} < W_k) \tag{78}$$

where $\chi^2_{d_k}$ is a Chi-square variable with d_k degree of freedom.

If the population is an RIL or BC population, there are only two genotypes. The additive effect is defined as

$$\hat{a}_k = \begin{bmatrix} -0.5 & 0.5 \end{bmatrix}\begin{bmatrix} \hat{\gamma}_{1k} \\ \hat{\gamma}_{2k} \end{bmatrix} = L\hat{\gamma}_k \tag{79}$$

where $L = \begin{bmatrix} -0.5 & 0.5 \end{bmatrix}$ has only one row and thus $\mathrm{rank}(L) = 1$. The test for the F_2 population applies to the RIL population.

Alternatively, we can perform likelihood ratio test (LRT) under the null hypothesis of $H_0 : \lambda_k = 0$, which is defined as

$$\Gamma_k = -2\left[L_0(\hat{\lambda}, 0) - L_1(\hat{\lambda}, \hat{\lambda}_k)\right] \tag{80}$$

where $L_1(\hat{\lambda}, \hat{\lambda}_k)$ is the log likelihood function evaluated at the REML estimated parameters under the full model and $L_0(\hat{\lambda}, 0)$ is the log likelihood function evaluated at the REML estimated parameter under the null model. Note that the estimated λ under the two models are different because they are estimated from different models. Under the null hypothesis, this likelihood ratio test follows approximately a mixture of χ^2_0 and χ^2_1 distribution with equal proportion. Therefore, the p-value is calculated using

$$p_k = \begin{cases} 1 & \text{if} \quad \Gamma_k = 0 \\ \frac{1}{2}\left[1 - \Pr(\chi^2_1 < \Gamma_k)\right] & \text{otherwise} \, \Gamma_k > 0 \end{cases} \tag{81}$$

8. Multiple marker models

The mixed model GWAS methods described above are all “single QTL model”, which is never correct if a trait is controlled by more than one locus. In fact, a multiple locus model is always better than the single locus model because it may represent the correct model. We scan the genome, one marker at a time, not because the method is great but because we have no other alternative to replace it WHEN the number of markers is extremely large. For example, for human data, the number of SNPs can easily reach 500,000. A multiple marker model with such a high dimension may fail to generate meaningful result or take forever to complete the analysis. Therefore, the single QTL model is the only choice for genome scanning with saturated markers. In many studies, the number of markers may not be that high. Number of markers ranging from 1,000 to 10,000 is very common in species other than human, major crops and laboratory animals. When the number of markers is not that high, a multiple marker model is more preferable. Here, we introduce two multiple marker models: LASSO and EBAYES.

8.1. LASSO

Least absolute shrinkage and selection operator **(**LASSO) (Tibshirani 1996) is perhaps the most popular statistical method for model selection. Ever since it was published in 1996, there are huge number of studies trying to apply the method to more broad area or modify the method to handle special type of data. However, the original LASSO is still the best in general. Some modified versions may perform better for a special dataset, but in general, the plain version of the LASSO is still the most robust method of model selection.

LASSO is a penalized regression method. To apply the method to GWAS, the population structure effects must be adjusted prior to the LASSO analysis. In our notation, the model to estimate population structure effects is

$$y = X\beta + e \tag{82}$$

The residual $\hat{e} = y - X\hat{\beta}$ is treated as the response variable *y* and then subject to analysis. Therefore, the response variable (*y*) in the LASSO analysis is actually the residual after adjustment for any other covariates, including population structures. The LASSO linear model in our notation is

$$y = Z\gamma + \varepsilon = \sum_{k=1}^{m} Z_k\gamma_k + \varepsilon \tag{83}$$

where Z is an $n \times m$ matrix of marker genotype indicators and γ is an $m \times 1$ vector. The marker effects are estimated by minimizing the penalized sum of squares,

$$Q(\lambda,\gamma) = (y - Z\gamma)^T (y - Z\gamma) + \lambda \sum\nolimits_{k=1}^{m} |\gamma_k| \tag{84}$$

where λ is a Lagrange multiplier for the penalty term. The interpretation of the above penalized least squares is that the first part can be minimized by including as many non-zero regression coefficients as possible but the second part will penalize it and favor more zero coefficient. The balance between the two term will generate the “best result”. The Lagrange multiplier is also a parameter, but is it obtained by cross validation (10-fold cross validation by default). Although there is a SAS procedure called PROC GLMSELECT that can perform LASSO variable selection, it does not produce the same result as the R package called GLMNET (Friedman et al. 2010) that was written by the original contributors of the LASSO method. The result of GLMSELECT is often not satisfactory. This is the first time I noticed that SAS is inferior to R in terms of correctness of result (not computational speed).

The problem with LASSO for GWAS is that the package only reports the markers selected and does not provide a test for a marker. Xu (2013) proposed an approximated Wald test for marker *k*,

$$W_k \approx \frac{\hat{\gamma}_k^2}{\hat{\sigma}^2} Z_k^T Z_k + 1 \tag{85}$$

This test statistic is often too liberal because the estimated σ^2 is often too small. We used the generalized cross validation (GCV) approach to estimate the residual error variance, which is defined as

$$\hat{\sigma}^2 = \frac{1}{n\left[n - \mathrm{tr}(H)\right]^2}(y - \sum_{k=1}^{m} Z_k \hat{\gamma}_k)^T (y - \sum_{k=1}^{m} Z_k \hat{\gamma}_k) \tag{86}$$

where

$$H = Z_{\mathrm{Select}}(Z_{\mathrm{Select}}^T Z_{\mathrm{Select}})^- Z_{\mathrm{Select}}^T \tag{87}$$

is the hat matrix and Z_{Select} is the *Z* matrix containing columns corresponding the non-zero $\hat{\gamma}_k$. The *p* value is calculated using the usual Chi-square distribution. Because the model is a multiple locus model, the critical value to declare significance can be set at 0.05, rather than from the Bonferroni correction.

8.2. Empirical Bayes (EBAYES)

The method was developed by Xu (2007) to estimate and test gene by gene interaction (epistatic) effects. It is a Bayesian method by assigning each marker effect a normal prior distribution. The prior variance is unknown but estimated from the data. Therefore, it is called empirical Bayes rather than Bayes. Since the prior variance of each marker is different across the genome, the degrees of shrinkage of marker effects are different. Large effects will have less shrinkage and small effects will have large shrinkage. Therefore, the method is also called selective shrinkage.

8.2.1. Hierarchical linear mixed model

Let y be a vector of phenotypic values of a quantitative trait collected from *n* individuals. Define Z_{jk} as a genotyped indicator variable of individual *j* with three values, 1, 0 and -1, representing the three possible genotypes of locus *k*, A_1A_1, A_1A_2 and A_2A_2, respectively. The linear mixed model for *y* is

$$y = \sum_{l=1}^{q} X_l \beta_l + \sum_{k=1}^{m} Z_k \gamma_k + \varepsilon \tag{88}$$

where X_l and β_l represent the design matrix and effect for the *l*th fixed effect (non-genetic), Z_k is a genotype indicator vector for marker *k* and γ_k is the effect of this marker and ε is the residual error with an assumed $\varepsilon \sim N(0, I\sigma^2)$ distribution. The marker effect γ_k is treated as a random variable with an assumed $N(0, \phi_k^2)$ distribution with a prior variance ϕ_k^2. To control the sparseness of the model, we further assign a hierarchical prior distribution to ϕ_k^2. We choose a scaled inverse chi-square distribution as the hierarchical prior,

$$p(\phi_k^2) \propto (\phi_k^2)^{(\tau+2)/2} \exp\left(\tfrac{\omega}{2\phi_k^2}\right) \tag{89}$$

where τ (degree of freedom) and ω (scale) are hyper parameters in the hierarchical prior. Note that each of the fixed effects β_l and the residual variance σ^2 has a default uniform prior.

8.2.2. Conditional posterior mode estimation of marker effects

We propose to estimate one parameter as a time conditional on values of other parameters. Let us define

$$\zeta = \sum_{l=1}^{q} X_l \beta_l \tag{90}$$

as the sum of all fixed effects and

$$\xi = \sum_{k=1}^{m} Z_k \gamma_k \tag{91}$$

as the sum of all marker effects (polygene). Let us further define

$$\zeta_{-l} = \sum_{l' \neq l}^{q} X_{l'} \beta_{l'} = \zeta - X_l \beta_l \tag{92}$$

and

$$\xi_{-k} = \sum_{k' \neq k}^{m} Z_{k'} \gamma_{k'} = \xi - Z_k \gamma_k \tag{93}$$

The conditional posterior mode estimate of β_l is obtained using the following linear fixed model

$$y_l = X_l \beta_l + \varepsilon \tag{94}$$

where $y_l = y - \zeta_{-l} - \xi$ is the phenotypic value adjusted by all other effects except $X_l \beta_l$, the effect to be estimated. The simple least square estimate of β_l conditional on all other effects is

$$\hat{\beta}_l = (X_l^T X_l)^{-1} (X_l^T y_l) \tag{95}$$

for $l = 1, ..., q$. Note that matrix inversion is replaced by the inverse of a scalar. We now discuss the marker effects. Since γ_k is a random effect, its estimate is called the best linear unbiased prediction (BLUP). The conditional BLUP of γ_k given all other effects is obtained using the following linear random model

$$y_k = Z_k \gamma_k + \varepsilon \tag{96}$$

where $y_k = y - \zeta - \xi_{-k}$. The BLUP of γ_k is

$$\hat{\gamma}_k = Z_k^T \phi_k^2 \left(Z_k Z_k^T \phi_k^2 + I\sigma^2 \right)^{-1} y_k \tag{97}$$

The inverse matrix can be obtained using the Woodbury matrix identity,

$$\begin{aligned} \left(Z_k Z_k^T \phi_k^2 + I\sigma^2 \right)^{-1} &= \frac{1}{\sigma^2} I - \frac{1}{\sigma^2} Z_k \left(Z_k^T Z_k / \sigma^2 + 1/\phi_k^2 \right)^{-1} Z_k^T \frac{1}{\sigma^2} \\ &= \frac{1}{\sigma^2} I - \frac{1}{\sigma^2} \lambda_k Z_k \left(Z_k^T Z_k \lambda_k + 1 \right)^{-1} Z_k^T \\ &= \frac{1}{\sigma^2} \left(I - \frac{\lambda_k Z_k Z_k^T}{\lambda_k Z_k^T Z_k + 1} \right) \end{aligned} \tag{98}$$

where $\lambda_k = \phi_k^2 / \sigma^2$ is the variance ratio. Substituting this inverse matrix into the BLUP equation in equation (97) yields

$$\hat{\gamma}_k = \lambda_k Z_k^T y_k - \frac{\lambda_k^2 (Z_k^T Z_k)(Z_k^T y_k)}{\lambda_k Z_k^T Z_k + 1} \tag{99}$$

for $k = 1, ..., m$, where matrix calculation has been avoided. The variance of the BLUP estimate is

$$\text{var}(\hat{\gamma}_k) = \left[\lambda_k - \lambda_k^2 \left(Z_k^T Z_k - \lambda_k \frac{(Z_k^T Z_k)^2}{\lambda_k Z_k^T Z_k + 1} \right) \right] \sigma^2 \tag{100}$$

The estimated residual variance conditional on all model effects is

$$\sigma^2 = \frac{(y - \hat{\zeta} - \hat{\xi})^T (y - \hat{\zeta} - \hat{\xi})}{n - q - p_0} \tag{101}$$

where

$$m_0 = \sum_{k=1}^{m} \lambda_k \left(Z_k^T Z_k - \frac{\lambda_k Z_k^T Z_k Z_k^T Z_k}{\lambda_k Z_k^T Z_k + 1} \right) \tag{102}$$

is the effective number of markers (Tipping 2001; Xu 2013).

The BLUP estimates of marker effects depend on $\lambda_k = \phi_k^2 / \sigma^2$. When ϕ_k^2 is replaced by the estimated value, the estimate is no longer BLUP; it is called the empirical Bayes estimate (Xu 2007). Therefore, we need to estimate ϕ_k^2 also from the data, which will be discussed in the next section. Since the estimate of any parameter is conditioned on values of all other parameters, the estimation process is iterative. We must iterate the process repeatedly until each parameter converges to a constant.

8.2.3. Conditional posterior mode estimation of marker variances

We now derive a simple method to estimate ϕ_k^2. Using the random model given in equation (96), we see that $\text{E}(y_k) = 0$ and the variance is

$$\text{var}(y_k) = Z_k Z_k^T \phi_k^2 + I\sigma^2 \tag{103}$$

The log posterior probability of ϕ_k^2 after incorporate the hyper parameters is

$$L(\phi_k^2) = -\frac{1}{2} \ln | Z_k Z_k^T \phi_k^2 + I\sigma^2 | - \frac{1}{2} y_k^T (Z_k Z_k^T \phi_k^2 + I\sigma^2)^{-1} y_k - \frac{\tau + 2}{2} \ln(\phi_k^2) - \frac{\omega}{2\phi_k^2} \tag{104}$$

Using Woodbury matrix identity, we can rewrite the above log posterior as

$$L(\phi_k^2) = -\frac{1}{2} \ln(Z_k^T Z_k \phi_k^2 / \sigma^2 + 1) + \frac{\phi_k^2 y_k^T Z_k Z_k^T y_k}{2\sigma^4 (Z_k^T Z_k \phi_k^2 / \sigma^2 + 1)} - \frac{\tau + 2}{2} \ln(\phi_k^2) - \frac{\omega}{2\phi_k^2} \tag{105}$$

where terms irrelevant to ϕ_k^2 have been ignored. Let $s_k = Z_k^T Z_k / \sigma^2$ and $h_k = Z_k^T y_k / \sigma^2$, we get

$$L(\phi_k^2) = -\frac{1}{2} \ln(s_k \phi_k^2 + 1) + \frac{\phi_k^2 h_k^2}{2(s_k \phi_k^2 + 1)} - \frac{\tau + 2}{2} \ln(\phi_k^2) - \frac{\omega}{2\phi_k^2} \tag{106}$$

When we set $\partial L(\phi_k^2)/\partial\phi_k^2 = 0$, many terms will be cancelled out and leaves a cubic function of ϕ_k^2 (see Xu 2012). This means that there are three possible solutions for ϕ_k^2. To simplify the problem, we now set $\omega = 0$ so that

$$L(\phi_k^2) = -\frac{1}{2}\ln(s_k\phi_k^2 + 1) + \frac{\phi_k^2 h_k^2}{2(s_k\phi_k^2 + 1)} - \frac{\tau + 2}{2}\ln(\phi_k^2) \tag{107}$$

It is obvious that the global solution is $\phi_k^2 = 0$. In this case, however, we need a local solution (the largest positive solution). The derivative of the above log posterior is

$$\frac{\partial}{\partial\phi_k^2} L(\phi_k^2) = -\frac{s_k}{2(s_k\phi_k^2 + 1)} + \frac{h_k^2(s_k\phi_k^2 + 1) - h_k^2\phi_k^2 s_k}{2(s_k\phi_k^2 + 1)^2} - \frac{\tau + 2}{2\phi_k^2} \tag{108}$$

Setting $\partial L(\phi_k^2)/\partial\phi_k^2 = 0$ leads to

$$-(\tau + 3)s_k^2\phi_k^4 - \left[(2\tau + 5)s_k + h_k^2)\right]\phi_k^2 - (\tau + 2) = 0 \tag{109}$$

which is a quadratic function with the largest positive solution equal to

$$\phi_k^2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \tag{110}$$

where

$$\begin{aligned} a &= -(\tau + 3)s_k^2 \\ b &= -\left[(2\tau + 5)s_k + h_k^2)\right] \\ c &= -(\tau + 2) \end{aligned} \tag{111}$$

Since we know the global solution of ϕ_k^2 is 0, whenever a solution is negative or illegal, we should set $\phi_k^2 = 0$. Note that we can use τ to control the model sparseness. When we set $\tau = -2$, the solution is

$$\phi_k^2 = \frac{h_k^2 - s_k}{s_k^2} \tag{112}$$

which represents the least sparseness. The sparseness will increase as τ increases. Setting $\tau = 0$ is equivalent to using the Jeffreys' prior $p(\phi_k^2) \propto 1/\phi_k^2$.

8.2.4. Summary of the computational algorithm

Before the algorithm starts, we need to calculate all the following terms, $X_l^T X_l$, $X_l^T y$, $Z_k^T Z_k$, $Z_k^T y$ and $y^T y$, because they do not involve parameters and are only calculated once. The iteration process is summarized as follows.

Step (1): Initialize the following variables, $\zeta = \xi = 0$, $\beta_l = \gamma_k = 0$ and $\sigma^2 = 1$.
Step (2): Update β_l one at a time until all β_l's have been updated.
Step (3): Estimate ϕ_k^2 and thus obtain $\lambda_k = \phi_k^2/\sigma^2$.
Step (4): Update γ_k using BLUP given in equation (99).
Step (5): Loop steps (3)-(4) for all $k = 1,...,m$.

Step (6): Update σ^2 based on updated ζ and ξ.
Step (7): Repeat from steps (2) to (6) until iteration converges.

One important property of the algorithm is that ζ and ξ are updated instantly when an effect (β_l or γ_k) is estimated (instead of waiting until all effects are estimated). The computation cost is $O\left[n(m+q)t\right]$ where *t* is the number of iterations required for the program to converge.

8.2.5. Hypothesis test

The Wald test statistic is used to test null hypothesis $H_0:\gamma_k=0$, which is defined as

$$W_k=\frac{\hat{\gamma}_k^2}{\text{var}(\hat{\gamma}_k)} \tag{113}$$

where the variance in the denominator is simplified from equation (100) to the following form

$$\text{var}(\hat{\gamma}_k)=\phi_k^2-\phi_k^4\left(s_k-\frac{\phi_k^2 s_k^2}{\phi_k^2 s_k+1}\right) \tag{114}$$

Assume that W_k follows a Chi-square distribution with one degree of freedom, the p-value is

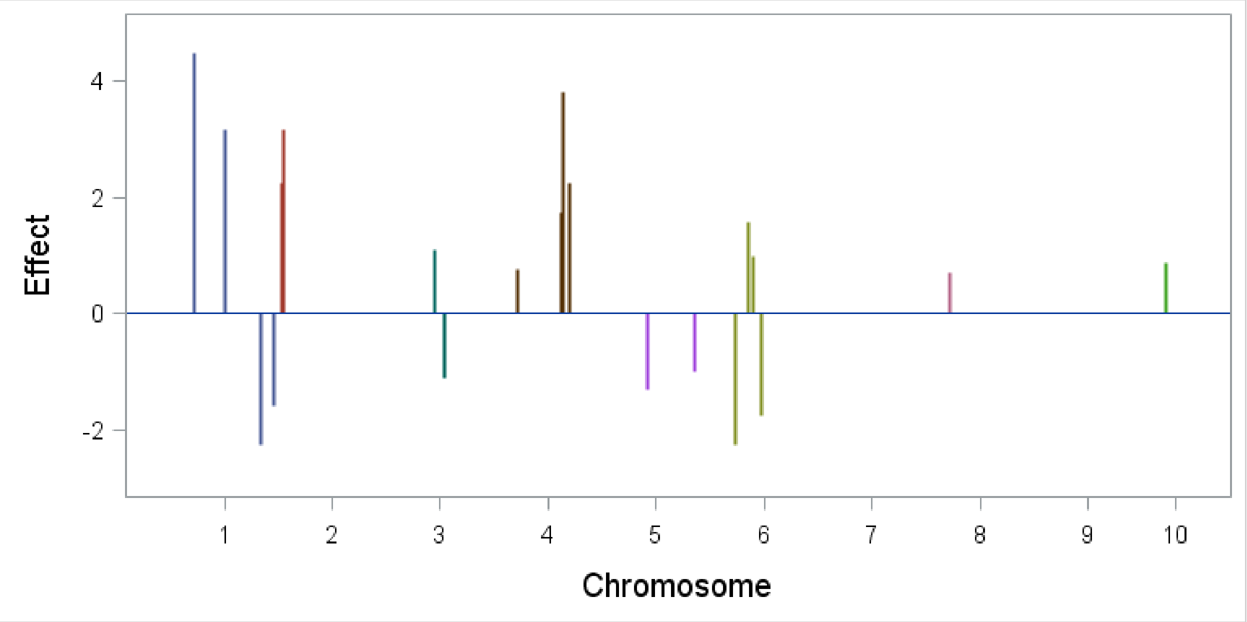
$$p_k=\Pr(\chi_1^2>W_k) \tag{115}$$

The nominal probability of 0.05 may be used. Bonferroni correction for multiple tests may not be necessary because of the shrinkage nature of the sparse Bayesian estimates of marker effects.

9. Simulated data analysis

We simulated an F2 population with a sample size 1000. There is no population structure effect in the model, different from a random population. A total of 10 chromosomes with a total length of 2400 cM were simulated. A total of 961 markers were evenly placed along the genome with 2.5 cM distance per marker interval. We simulated 20 QTL with sizes and locations depicted in Figure 8. The residual variance was set at 10. The largest QTL contributes 14% of the phenotypic variance and the smallest QTL contributes 0.36% of the phenotypic variance. The 20 QTL collectively contribute 85% of the total phenotypic variance. The mean of the simulated trait was 10 and the residual error variance was 10. No polygene was simulated but when a marker is scanned, the QTL not overlapping with the scanned marker will go to the polygene and be captured by the polygene in the model. The dimension of the *Z* matrix is 1000×961.

Figure 8
QTL effects used in the simulation experiment, where effects from 10 chromosomes are color coded.

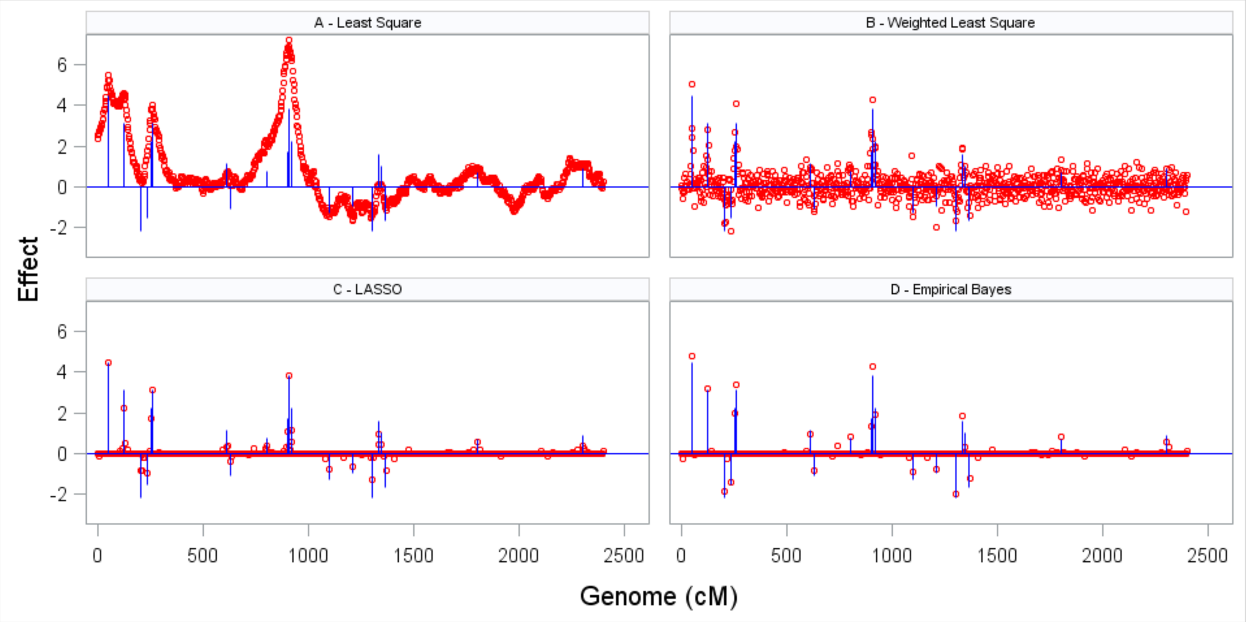


We used four methods to estimate marker effects and perform significance tests. The four methods are: A – the simple least squares method; B – the weighted least squares method; C – the LASSO method; D – the empirical Bayes method. The estimated effects from the four methods along with the true effects are depicted in Figure 9. The simple least squares method (panel A) does not have the resolution to separate linked QTL. Markers near the simulated QTL also show large estimated effects due to linkage disequilibrium. These markers are false positive. Effects of linked QTL with opposite effects are cancelled out each other. The weighted least squares method (panel B) increases the resolution significantly and is able to separate closely liked QTL even with effects in opposite directions. The estimated effects for the null markers are all around zero but not exactly at zero. The LASSO (panel C) and empirical Bayes (panel D) are all generate results that are unbelievably good. The two methods can even pick the smallest QTL. It appears that the empirical Bayes method is slightly better than the LASSO method for this particular data.

The most important part of GWAS is the Manhattan plot: plot of the test statistic (either the Wald test or the –Log10(p)) against the genome location of the marker. The Manhattan plots of the simulated data from the four methods are shown in Figure 10.

Figure 9

Estimated QTL effects of the simulated data from four methods: A - the simple least squares method (without polygenic control); B – the weighted least squares method (with polygenic control); C – LASSO; D – Empirical Bayes. Red circles are estimated effects and blue needles are the true effects.



The Manhattan plots show that there are too many QTL detected by the simple least squares method and too few QTL detected by the weighted least squares method. The LASSO and empirical Bayes methods detected the corrected number of QTL. The smallest simulated QTL on chromosome 8 is also detected by the empirical Bayes method, but missed by the LASSO method.

In GWAS analysis, people often monitor the QQ-plot to see how the test statistic behave. A QQ-plot is the plot of the observed –Log10(p) against the expected –Log10(p). The latter is just a uniform distribution within the range of the observed test statistic. The QQ-plot only applies to the genome scanning approach, e.g., least squares and weighted least squares. Figure 11 shows the QQ-plots of the two methods for the simulated data. The observed test statistics are way above the expected test statistics for the least squares method (without polygenic control). The weighted least squares method behaves well as most of the observed test statistics are around the diagonal line while only a small proportion of the data points are away from the diagonal.

Figure 10

Manhattan plots of the simulated data from four methods: A - the simple least squares method (without polygenic control); B – the weighted least squares method (with polygenic control); C – LASSO; D – Empirical Bayes. The horizontal reference line is the Bonferroni corrected critical value for the test statistics (4.28).

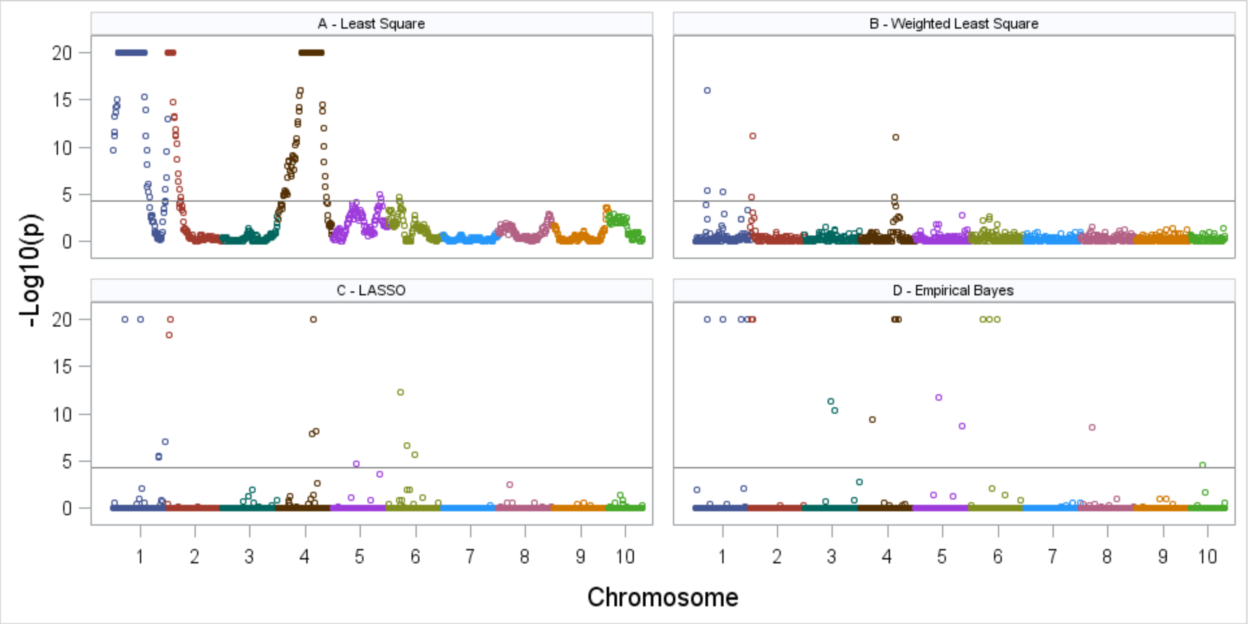
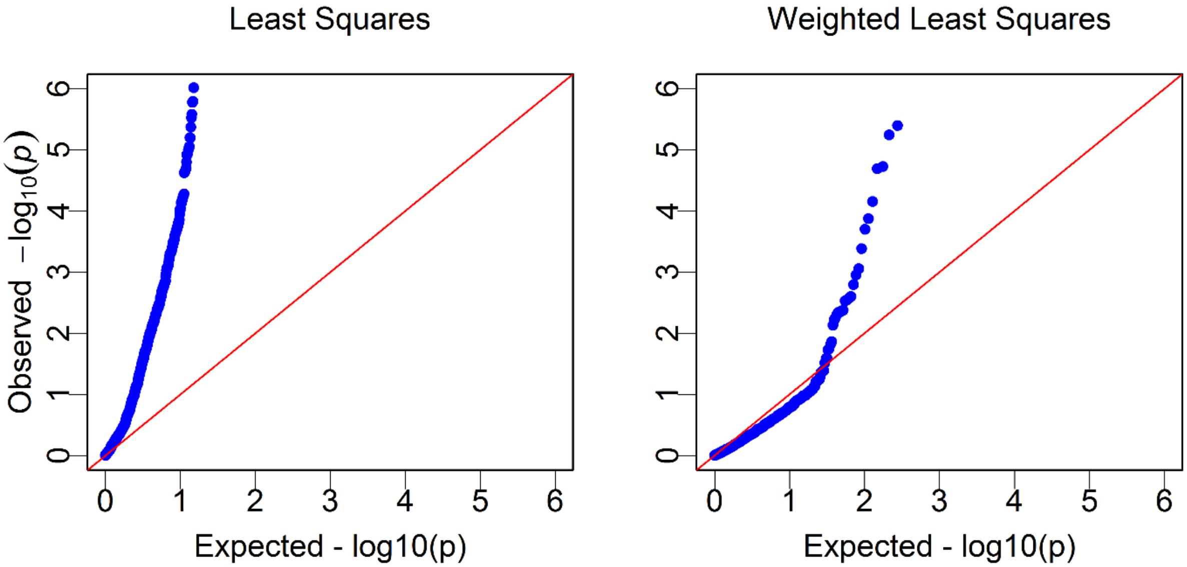


Figure 11

QQ-plots of the simulated data from the least squares and weighted least squares methods.



The data and SAS codes to perform the above GWAS are describe here. There are three datasets generated from the simulation experiment. The “sim-phe-961.csv” file holds the simulated phenotypes of 1000 individuals. The “sim-map-961.csv” file is the map file to store the marker IDs and the true effects of the simulated markers. The file “sim-gen-961.csv” holds the genotypes of 961 markers for the 1000 individuals (matrix Z). Two other files were generated from these original simulated data, one is the kinship matrix generated from $K = ZZ^T$ (sim-kk-961.csv) and the other is the eigenvector transformed *X*, *y* and *Z* matrices (sim-eig-961.csv). The SAS codes are stored in several files (to be introduced in the Practice Section). The LASSO method was implemented using an R package called GLMNET (Friedman et al. 2010). The empirical Bayes method was implemented using an R program written by Xu (2007).

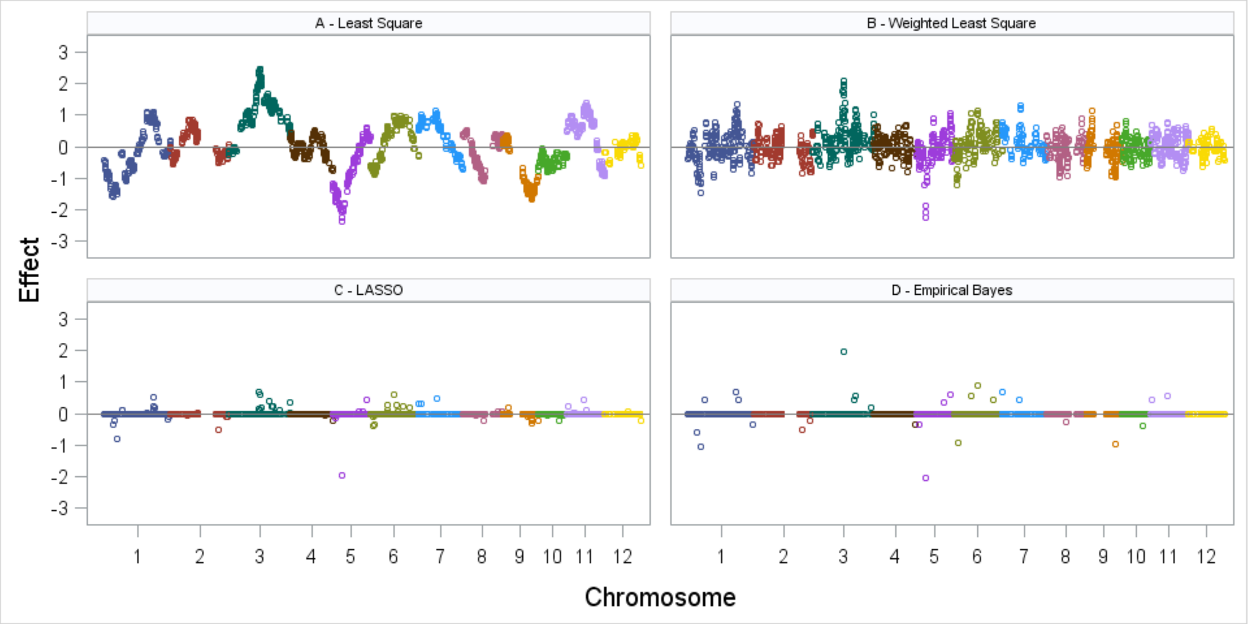
10. Rice data analysis

We analyzed a population of 210 recombinant inbred lines of rice (Oryza sativa) derived from the cross between Zhenshan 97 and Minghui 63 (Hua et al., 2002, 2003; Xing et al., 2002). This hybrid (Shanyou 63) is the most widely grown hybrid in China. A total of 210 RILs were derived by single-seed descent from this hybrid. We analyzed the 1000-grain weight (KGW) trait as an example to illustrate the GWAS procedures. The trait was measured from four replicated experiments (1997 and 1998 from one location, 1998 and 1999 from another location). In each replicate, eight plants from each line were sampled and the average phenotype represented the phenotypic value of the line in that environment (Xing et al., 2002; Yu et al., 2011). To mimic GWAS, we treated the four replications (environments) as “population structures” and included them in the model as fixed effects. The number markers (bins) is 1619. Therefore, the *Z* matrix has 210 rows and 1619 columns for each replicate. After combining data of all four replicates, the *y* vector has a dimension of 840 x 1, where 840 = 210 x 4. The overall Z matrix has a dimension of 840 x 1619. The kinship matrix for the overall analysis is of 840 x 840.

We used the same four methods to estimate marker effects and perform significance tests. The estimated effects from the four methods are depicted in Figure 12. All methods detected a marker (bin729) with a large effect on chromosome 5. This bin contains a cloned gene for KGW. The Manhattan plots of the simulated data from the four methods are shown in Figure 13.

Figure 12

Estimated QTL effects of KGW of the rice data from four methods: A - the simple least squares method (without polygenic control); B – the weighted least squares method (with polygenic control); C – LASSO; D – Empirical Bayes.



The Manhattan plots show that there are too many QTL detected by the simple least squares method and too few QTL detected by the weighted least squares method. The LASSO and empirical Bayes methods detected more QTL than the weighted least squares method. These additional QTL deserve further investigation. Figure 14 shows the QQ-plots of the two methods for the simulated data. The observed test statistics are way above the expected test statistics for the least squares method (without polygenic control). The weighted least squares method behaves well as most of the observed test statistics are around the diagonal line while only a small proportion of the data points are away from the diagonal.

Figure 13

Manhattan plots of the KGW of the rice data from four methods: A - the simple least squares method (without polygenic control); B – the weighted least squares method (with polygenic control); C – LASSO; D – Empirical Bayes. The horizontal reference line is the Bonferroni corrected critical value for the test statistics (4.51).

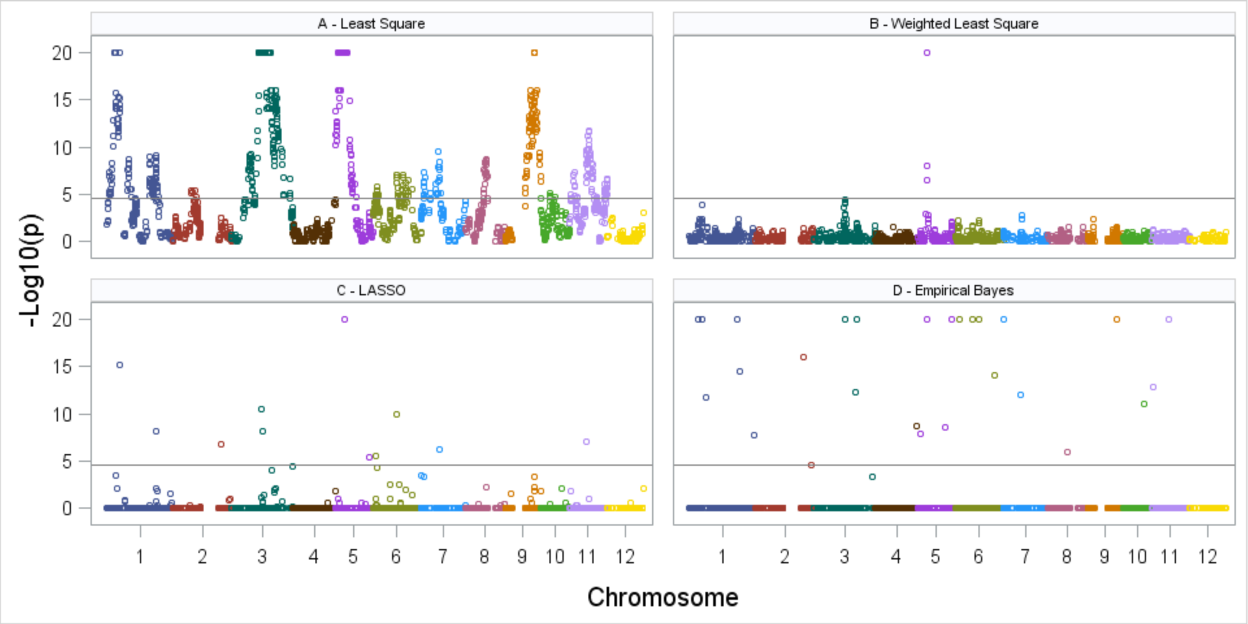
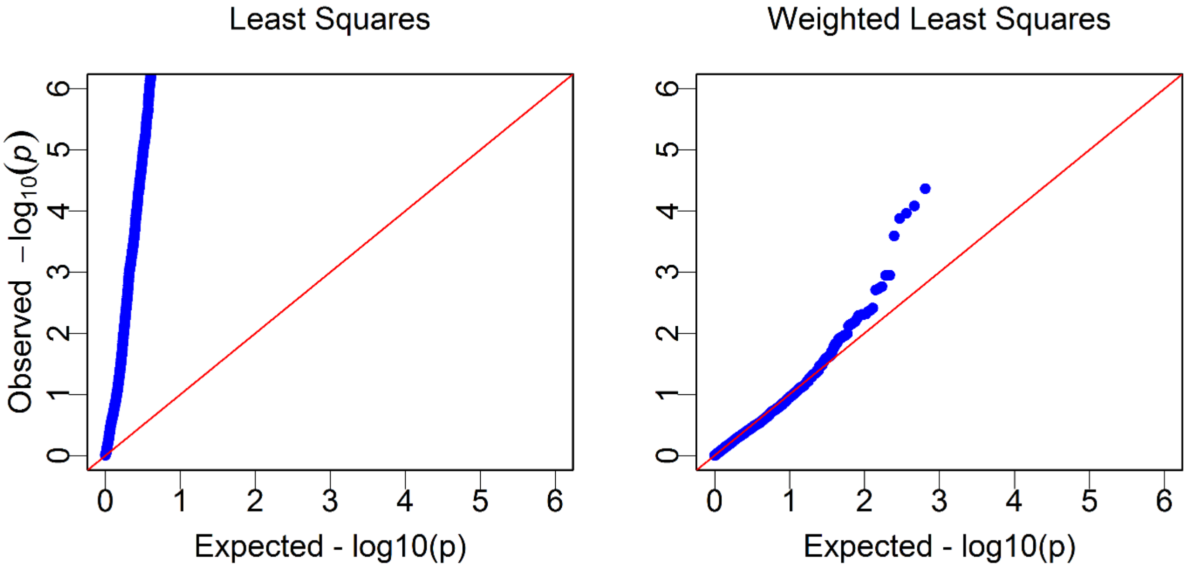


Figure 7

QQ-plots of KGW of the rice data from the least squares and weighted least squares methods.



Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

An efficient empirical Bayes method for genomewide association studies

Q. Wang[1,2], J. Wei[2,3], Y. Pan[1] & S. Xu[2]

1 Department of Animal Science, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China
2 Department of Botany and Plant Sciences, University of California, Riverside, CA, USA
3 College of Animal Science and Technology, China Agricultural University, Beijing, China

Keywords

Effective number of tests; empirical Bayes; genomewide association studies; linear mixed model.

Correspondence

S. Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA.
Tel: (951)-827-5898;
Fax: (951) 827-4437;
E-mail: shizhong.xu@ucr.edu

Received: 24 June 2015;
accepted: 2 October 2015

Summary

Linear mixed model (LMM) is one of the most popular methods for genomewide association studies (GWAS). Numerous forms of LMM have been developed; however, there are two major issues in GWAS that have not been fully addressed before. The two issues are (i) the genomic background noise and (ii) low statistical power after Bonferroni correction. We proposed an empirical Bayes (EB) method by assigning each marker effect a normal prior distribution, resulting in shrinkage estimates of marker effects. We found that such a shrinkage approach can selectively shrink marker effects and reduce the noise level to zero for majority of non-associated markers. In the meantime, the EB method allows us to use an 'effective number of tests' to perform Bonferroni correction for multiple tests. Simulation studies for both human and pig data showed that EB method can significantly increase statistical power compared with the widely used exact GWAS methods, such as GEMMA and FaST-LMM-Select. Real data analyses in human breast cancer identified improved detection signals for markers previously known to be associated with breast cancer. We therefore believe that EB method is a valuable tool for identifying the genetic basis of complex traits.

Introduction

Genomewide association studies (GWAS) is an important tool for identifying genes underlying human diseases and agriculturally important traits. One of the most efficient model of GWAS is the linear mixed model (LMM) where population structure and marker effects are assumed as fixed and a polygene is treated as a random effect whose variance is captured using a marker-inferred additive relationship matrix (Yu *et al.* 2006). Since the original publication of the linear mixed model GWAS, numerous improved methods have been developed under the LMM framework. Majority of the methods focused on improving computational speed. For example, the efficient mixed model association (EMMA) utilized eigendecomposition to evaluate the likelihood function (Kang *et al.* 2008) and the new method was proved to be more efficient than the original method (Yu *et al.* 2006). The EMMA method is called the exact method because the polygenic variance is re-estimated along with each marker scanned. Recently, Zhou & Stephens (2012) developed an improved method called genomewide efficient mixed model association (GEMMA). This new algorithm can improve the computational speed orders of magnitude faster than the original EMMA. Other improved methods include factored spectrally transformed linear mixed models (FaST-LMM) developed by Lippert *et al.* (2011) and later improved by Listgarten *et al.* (2012). The FaST-LMM method achieves the fast speed by choosing a selected subset of markers to capture the polygenic effect. In contrast to these exact methods, there are several approximate methods available to improve the

doi:10.1111/jbg.12191

computational speed by fixing the polygenic variance at the estimated value under the pure polygenic model. These methods include EMMAX (Kang *et al.* 2010) and P3D (Zhang *et al.* 2010). The latter can also reduce computing time and improve statistical power by clustering individuals into groups. The elements of the kinship matrix are used as similarity measures in the clustering analysis. Summary statistics of the kinship between and within groups are then used as the elements of a reduced kinship matrix.

Recently, Segura *et al.* (2012) presented a multilocus mixed model approach for GWAS that fits multiple markers into the same model via a forward selection algorithm. They claimed that this method can increase power and decrease false discovery rate. Zhou *et al.* (2013) developed a hybrid method between GWAS and genomic prediction, called Bayesian sparse linear mixed model (BSLMM), and hope to perform both in a single data analysis. The BSLMM was implemented using the Markov chain Monte Carlo (MCMC) algorithm. All the multiple-marker analysis algorithms gain the accuracy at the cost of computational speed. Unless the gain is substantial compared with the single marker scanning approach, people may still prefer the latter for its simplicity and the ability to handle extremely large data sets.

These single marker scanning methods mentioned above mainly focus on algorithm development for improving computational speed. However, there are two major issues in GWAS that have not been fully addressed before. The two issues are (i) the genomic background noise and (ii) low statistical power after Bonferroni correction. For the first issue, the genomic background noise may blur the true signals and lead to markers with high LD to a causal gene to be falsely claimed as associated with the trait of interest. For the second issue, Bonferroni correction for multiple tests may cause a true associated marker to be missed because the criterion is too stringent for p-value 0.05 dividing by the number of markers (~1.0 million). Current statistical methods are incapable of dealing with these two issues.

Motivated by these concerns, we present a new GWAS approach by treating a scanned marker effect as random and assumed to be normally distributed. The variance of the distribution is estimated using the restricted maximum-likelihood (REML) method. The estimated variance is then used to facilitate the best linear unbiased prediction (BLUP) of the marker effect. If we consider the marker effect as the parameter of interest and the marker variance as a prior variance, the BLUP of marker effect is the posterior mean in terms of Bayesian analysis. As the variance is estimated from the same data set, the BLUP is called the empirical Bayes (EB) estimate of the marker effect. There are several advantages of the EB method for GWAS over existing LMM procedures: (i) The shrinkage nature of EB estimate can substantially reduce the type I error by the shrinkage nature of the method; (ii) It uses a new method for multiple test correction to improve statistical power; (iii) The new EB method still takes both the advantage of eigendecomposition and FaST-LMM to improve computational speed.

Materials and methods

Theory

Random model approach to single marker scan

We developed the following linear mixed model to scan a marker for its association with a quantitative trait. The entire genome is then scanned one marker at a time until all markers are evaluated to complete the GWAS. Let y be an $n \times 1$ vector of phenotypic values of a quantitative trait for n individuals. The linear mixed model for the kth marker ($k = 1,\ldots,m$) is

$$y = X\beta + Z_k\gamma_k + \xi + \varepsilon \tag{1}$$

where X is an $n \times c$ matrix for fixed effects including a column vector of 1 and β is a $c \times 1$ vector of fixed effects including the intercept. The fixed effects are included in the model to capture effects of any covariates that are not relevant to marker effects, for example population structure and age effect, etc. The Z_k vector ($n \times 1$) stores the numerical codes of genotypes for all individuals for marker k and the jth element of it is defined as $Z_{jk} = 0$ for the homozygote of major allele, $Z_{jk} = 1$ for the heterozygote and $Z_{jk} = 2$ for the homozygote of the minor allele. The effect of the kth marker is denoted by γ_{k} with an assumed $N(0, \phi_k^2)$ prior distribution, and the marker-specific prior variance is ϕ_k^2. This particular treatment of γ_{k} is called the random model approach. Under the Bayesian framework, γ_k is the parameter of interest and ϕ_k^2 is simply a prior variance. When an estimated ϕ_k^2 is used to perform the Bayesian analysis, the estimated γ_k is called the empirical Bayes (EB) estimate. The term ξ is an $n \times 1$ vector of polygenic effects, which is expressed by

$$\xi = \sum_{k=1}^{m} Z_k\eta_k \tag{2}$$

The Z_k's in Equations (1) and (2) are the same, but the γ_k and η_k in the two equations follow different distributions, $\gamma_k \sim N(0, \phi_k^2)$, $\eta_k \sim N(0, \phi^2/m)$, where ϕ^2 is called the polygenic variance. The additional term ε is an $n \times 1$ vector of errors, and the term $\xi + \varepsilon$ has $E\ (\xi + \varepsilon) = 0$ and $\mathrm{var}(\xi + \varepsilon) = (K\lambda + I)\sigma^2$, where $K = \frac{1}{m}\sum_{k=1}^{m} Z_k Z_k^T$ is an $n \times n$ matrix, I is an $n \times n$ identity matrix and $\lambda = \phi^2/\sigma^2$ is fixed at its estimated value under pure polygenic model (Data S1 in Supplementary Material for more details).

Because γ_k is assumed to be a random effect, the expectation of y in Equation (1) remains $E(y) = X\beta$, but the variance–covariance matrix is

$$\mathrm{var}(y) = Z_k Z_k^T \phi_k^2 + K\phi^2 + I\sigma^2 = (Z_k Z_k^T \lambda_k + K\lambda + I)\sigma^2 \tag{3}$$

where λ_k is called the variance ratio for the kth marker, defined as $\lambda_k = \phi_k^2/\sigma^2$. We use the REML method to estimate the variance ratio λ_k with the fast computation method described as follows.

Fast computation for single marker scan

Evaluation of the restricted log likelihood function can be very costly when the sample size is large. Two special algorithms were adopted to evaluate the likelihood function, which are eigendecomposition and Woodbury matrix identity (Kang *et al.* 2008; Zhou & Stephens 2012).

We used eigendecomposition to deal with the K matrix so that $K = \mathrm{UDU^T}$, where $D = \mathrm{diag}\{\delta_1, \ldots, \delta_n\}$ is a diagonal matrix for the eigenvalues and U is an $n \times n$ matrix for the eigenvectors. Let $y^* = U^T y$, $X^* = U^T X$ and $Z_k^* = U^T Z_k$ be transformed variables so that

$$y^* = X^*\beta + Z_k^*\gamma_k + U^T(\xi + \varepsilon) \tag{4}$$

The variance–covariance matrix of y^* is

$$\mathrm{var}(y^*) = (Z_k^* Z_k^{*T}\lambda_k + R)\sigma^2 = H_k\sigma^2 \tag{5}$$

where $R = \mathrm{D}\lambda + I$ is a known diagonal matrix $(n \times n)$ and $H_k = Z_k^* Z_k^{*T}\lambda_k + R$ is a general covariance structure $(n \times n)$. After replacing β and σ^2 with λ_k, we have the following profiled restricted log likelihood function,

$$\begin{aligned} L(\lambda_k) = &-\frac{1}{2}\ln|H_k| - \frac{1}{2}\ln|X^{*T}H_k^{-1}X^*| \\ &- \frac{n-r}{2}\ln(y^{*T}P_k y^*), \end{aligned} \tag{6}$$

where r is the total number of fixed effects and

$$P_k = H_k^{-1} - H_k^{-1}X^*(X^{*T}H_k^{-1}X^*)^{-1}X^{*T}H_k^{-1} \tag{7}$$

This likelihood function contains only one parameter λ_k The Newton algorithm can be used for finding the numeric solution of λ_k. We then used the Woodbury matrix identities to improve the computational speed. The Woodbury matrix identities are

$$\begin{aligned} H_k^{-1} &= (Z_k^* Z_k^{*T}\lambda_k + R)^{-1} \\ &= R^{-1} - \lambda_k R^{-1} Z_k^* (\lambda_k Z_k^{*T} R^{-1} Z_k^* + 1)^{-1} Z_k^{*T} R^{-1} \end{aligned} \tag{8}$$

and

$$|H_k| = |Z_k^* Z_k^{*T}\lambda_k + R| = |R||\lambda_k Z_k^{*T} R^{-1} Z_k^* + 1| \tag{9}$$

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

$$\begin{aligned} a^T H_k^{-1} b = &\ a^T R^{-1} b \\ &- \lambda_k a^T R^{-1} Z_k^* (\lambda_k Z_k^{*T} R^{-1} Z_k^* + 1)^{-1} Z_k^{*T} R^{-1} b \end{aligned} \tag{10}$$

where both a and b can be any matrix or vector corresponding to H_k^{-1}, for example, $X^{*T}H_k^{-1}X^*$, $X^{*T}H_k^{-1}y^*$ and $y^{*T}H_k^{-1}y^*$.

Note that the above quadratic has been expressed as a function of various $a^T R^{-1} b$ terms,

$$a^T R^{-1} b = \sum_{j=1}^{n} a_j^T b_j (\delta_j \hat{\lambda} + 1)^{-1} \tag{11}$$

where a_j is the jth element (row) of vector (matrix) a and b_j is the jth element (row) of vector (matrix) b, respectively, for $j = 1, \ldots, n$.

EB estimate of a marker effect

Recalling the model for y in Equation (1), the EB estimate of γ_k can be formulated using the joint distribution of y and γ_k, which has a multivariate normal distribution with expectation

$$\mathrm{E}\begin{bmatrix} y \\ \gamma_k \end{bmatrix} = \begin{bmatrix} X\beta \\ 0 \end{bmatrix} \tag{12}$$

and variance

$$\mathrm{var}\begin{bmatrix} y \\ \gamma_k \end{bmatrix} = \begin{bmatrix} Z_k Z_k^T \phi_k^2 + K\phi^2 + I\sigma^2 & Z_k\phi_k^2 \\ Z_k^T\phi_k^2 & \phi_k^2 \end{bmatrix} \tag{13}$$

The EB estimate of γ_k is the conditional mean of γ_k, given y. From the multivariate normal theorem (Giri 2003), the conditional mean $(\hat{\gamma}_k)$ is

$$\begin{aligned} \mathrm{E}(\gamma_k|y) &= \mathrm{E}(\gamma_k) + \mathrm{cov}(\gamma_k, y)[\mathrm{var}(y)]^{-1}[y - \mathrm{E}(y)] \\ &= Z_k^T \hat{\phi}_k^2 \left(Z_k Z_k^T \hat{\phi}_k^2 + K\hat{\phi}^2 + I\hat{\sigma}^2 \right)^{-1} (y - X\hat{\beta}) \end{aligned} \tag{14}$$

where all parameters are substituted by the REML estimates. The conditional variance $\mathrm{var}(\hat{\gamma}_k)$ is

$$\begin{aligned}\mathrm{var}(\gamma_k|y) =& \mathrm{var}(\gamma_k) - \mathrm{cov}(\gamma_k, y)[\mathrm{var}(y)]^{-1}\mathrm{cov}(y, \gamma_k)\\ =& \hat{\phi}_k^2 - \hat{\phi}_k^2 Z_k^T\left(Z_k Z_k^T \hat{\phi}_k^2 + K\hat{\phi}^2 + I\hat{\sigma}^2\right)^{-1} Z_k \hat{\phi}_k^2\end{aligned} \tag{15}$$

Hypothesis test

We proposed the following Wald test statistics to scan the genome. The Wald test statistic for H_0:$\gamma_k = 0$ is

$$W_k = \frac{\hat{\gamma}_k^2}{\mathrm{var}(\hat{\gamma}_k)} \tag{16}$$

where $\hat{\gamma}_k$ is the EB estimate of the marker effect and $\mathrm{var}(\hat{\gamma}_k)$ is the variance of the EB estimate. This test is under the Bayesian framework because we essentially treated γ_k as the parameter and $\hat{\phi}_k^2 = \hat{\lambda}_k \hat{\sigma}^2$ as an estimated prior variance. Assume that W_k follows a chi-square distribution with one degree of freedom. The p-value is calculated using $p_k = \Pr(\chi_1^2 > W_k)$.

Bonferroni correction using the effective number of tests

The 'effective number of tests' was adopted from the 'effective number of parameters' in Bayesian analysis (MacKay 1992; Tipping 2001; Gelman *et al.* 2004). MacKay (1992) called it the number of 'well-measured parameters'. For a single random effect, Tipping (2001) interpreted it as a measure of how 'well-determined' a model effect is by the data. Gelman *et al.* (2004) used it to measure the complexity of a Bayesian hierarchical model and called it the 'effective number of parameters'. Xu (2013) adopted the effective number of tests for Bonferroni correction in QTL mapping. The EB method also allows us to calculate the 'effective number of tests' and use this number to perform Bonferroni correction.

For each marker, a degree of confidence is calculated using

$$d_k = 1 - \mathrm{var}(\hat{\gamma}_k)/\hat{\phi}_k^2 \tag{17}$$

where $\mathrm{var}(\hat{\gamma}_k)$ is the posterior variance of γ_k and $\hat{\phi}_k^2$ is the (estimated) prior variance. By definition, the posterior variance is always smaller than the prior variance. Therefore, $0 \le d_k \le 1$ and it represents the degree of confidence for marker k. The EB method often shrinks majority of the marker effects to zero (because the estimated variance component is bounded at zero) and these markers will have zero confidence. The effective number of tests is then defined as

$$m_e = \sum_{k=1}^{m} d_k \tag{18}$$

Here, we derived d_k using the expectation and maximization (EM) algorithm. The well-known formula to derive the EM algorithm for estimating $\hat{\phi}_k^2$ is

$$\begin{aligned}\phi_k^2 &= E(\gamma_k^2) = [E(\gamma_k|\mathrm{data})]^2 + \mathrm{var}(\gamma_k|\mathrm{data})\\ &= \hat{\gamma}_k^2 + \mathrm{var}(\hat{\gamma}_k)\end{aligned} \tag{19}$$

where $\hat{\gamma}_k = E(\gamma_k|\mathrm{data})$ is the posterior mean and $\mathrm{var}(\hat{\gamma}_k) = \mathrm{var}(\gamma_k|\mathrm{data})$ is the posterior variance of γ_k Rearranging this equation leads to

$$\hat{\gamma}_k^2 = E(\gamma_k^2) - \mathrm{var}(\hat{\gamma}_k) \tag{20}$$

Dividing both sides of the above equation by ϕ_k^2 yields

$$\begin{aligned}\frac{1}{\phi_k^2}\hat{\gamma}_k^2 &= \frac{E(\gamma_k^2) - \mathrm{var}(\hat{\gamma}_k)}{\phi_k^2} = \frac{\phi_k^2 - \mathrm{var}(\hat{\gamma}_k)}{\phi_k^2} = 1 - \frac{\mathrm{var}(\hat{\gamma}_k)}{\phi_k^2}\\ &= d_k\end{aligned} \tag{21}$$

Under a modified Bonferroni correction, a marker is declared as significant when the p-value is smaller than $p_{mBonf} = 0.05/m_e$, instead of $p_{Bonf} = 0.05/m$ in the original Bonferroni correction. So the EB method can be called EB-mBonf using modified Bonferroni correction and called EB-Bonf using the original Bonferroni correction.

Real human and pig data

The human breast cancer data (Hunter *et al.* 2007), as part of the Cancer Genetic Markers of Susceptibility (CGEMS) Project, were downloaded from the dbGAP databases (phs000147.v2.p1). The data include 2287 subjects with 1145 breast cancer cases and 1142 controls. The genotype data consist of 509 979 SNPs on 22 autosomes after exclusion of rare SNPs (Minor allele frequency <1%). The missing values of genotypes were imputed by the mean values. The phenotypic value of a breast cancer patient was assigned a value 1 and a control a value 0. We first performed a principal component analysis (PCA) using all markers and selected the first five principal components as covariates (fixed effects) to control the population structure. The fixed effects also included six different age groups (<55, 55–59, 60–64, 65–69, 70–74 and >74). The 509 979 SNPs were also used to calculate the kinship relationship matrix, which was further used to estimate the polygenic variance for genomic background control. The pig data set was collected from 820 commercial female pigs (Fan *et al.* 2011). The SNPs were filtered with minor allele frequency <5% and p value of chi squared test for Hardy–Weinberg equilibrium ≥1E−5. After filtering, 43 537 SNPs located on 18 autosomes were used for further analysis. Population stratification analysis showed that no significant genetic difference existed among this pig population (Fan *et al.* 2011), so PCA was not used as covariates for the pig data.

Simulation studies

To test the statistical power and false-positive rates, we performed two simulation experiments using real human breast cancer and commercial female pig genotype data.

Experiment 1

To examine the empirical power and false-positive rates of our EB methods under a fixed heritability value, we assigned genetic effects to 15 randomly selected SNPs sampled from the first chromosome of the human data. The effects of the 15 QTN were sampled from a normal distribution with mean 0 and variance 1. We then added a normal residual error drawn from $N(0,\sigma^2)$ to the genetic value of each individual to generate a phenotypic value. This is the situation where no polygenic effect was simulated. In another situation, we added a simulated polygenic effect drawn from a normal distribution $N(0,\sigma_\xi^2)$, where σ_ξ^2 was equal to σ^2. The residual variance was selected so that the proportions of phenotypic variance contributed by the simulated QTN (also called heritability) were 0.5. The empirical statistical power and false-positive rates were calculated as the ratios of true QTN and false QTN that passed the Bonferroni-corrected thresholds at the genomewide type I error of 0.05. The simulation was replicated 1000 times, and the average power and false-positive rate over the 1000 replicates were reported. For the pig data, the simulation procedure was the same as the human data except 15 SNPs were randomly selected from the whole genome of the pig data.

Experiment 2

To examine the impact of heritability on the performance of our EB methods, we further generated phenotypic values by varying residual variance or the number of QTN. The simulation was conducted under two scenarios: (i) The simulated procedure and QTN number (15 QTN) were the same as Experiment 1, but two different values of the residual variance were selected so that the heritability contributed by the simulated QTN were 0.25 and 0.75, respectively. (ii) The simulated procedure and SNP contribution to the phenotypic variance (0.5) were the same as Experiment 1, but three different levels of SNP number (167, 36 and 10 for both human and pig data) were chosen and they represent three proportions of the phenotypic variance explained by each QTN (0.3%, 1.4% and 5.0%). Again, the simulation was replicated 1000 times. The empirical power was calculated following the approaches used in Experiment 1.

Results

Simulation studies

We compared our EB method with two existing exact GWAS methods (GEMMA and FaST-LMM-Select) through a series of simulation experiments. In the first simulation experiment, we simulated multiple QTNs with randomly selected effect sizes and genome locations (see *Materials and Methods*). We first examined the empirical power and type I error of the EB method under two criteria of Bonferroni correction (EB-Bonf and EB-mBonf) and compared the EB method with GEMMA and FaST-LMM-Select under 0.5 heritability. Without polygene, EB-mBonf performed the highest statistical powers (0.63), followed by GEMMA (0.57) and EB-Bonf (0.57) for the human data (see Table 1). The statistical powers of EB-mBonf and EB-Bonf were 8% and 2% higher than FaST-LMM-Select, respectively. The same rankings can also be found in the pig data (Table 1). When a polygene was added to the model, the corresponding powers slightly increased but the rankings of the four methods remain the same for both human and pig data.

In all situations, the empirical false-positive rates were well controlled for both human and pig data. The results are shown in Table 2, where EB-Bonf provided the best control of false positives, followed by FaST-LMM-Select. EB-mBonf controlled the false-positive rates at a similar level to the GEMMA method. As described in Equation (17), d_k ranges between 0 and 1, so the effective number of tests is always smaller than m. The p-value threshold under the modified Bonferroni correction (p_{mBonf}) should

Table 1 Empirical powers of four methods drawn from 1000 replicated simulation experiments under 0.5 heritability for the real human and pig genotype data

	Human data		Pig data	
Method	QTN model	QTN + Polygene	QTN model	QTN + Polygene
EB-mBonf	0.63	0.65	0.59	0.62
GEMMA	0.57	0.60	0.54	0.57
EB-Bonf	0.57	0.59	0.54	0.57
Fast-LMM-Select	0.55	0.56	0.51	0.53

Table 2 Empirical false-positive rates of four methods drawn from 1000 replicated simulation experiments under 0.5 heritability for the real human and pig genotype data at the genomewide type I error of 0.05

	Human data		Pig data	
Method	QTN model	QTN + Polygene	QTN model	QTN + Polygene
EB-mBonf	0.0030	0.0035	0.0017	0.0018
GEMMA	0.0031	0.0035	0.0018	0.0018
EB-Bonf	0.0014	0.0014	0.0006	0.0007
Fast-LMM-Select	0.0021	0.0022	0.0012	0.0012

always be higher than the corresponding threshold under the classical Bonferroni correction (p_{Bonf}). As the EB method already selectively shrunken marker effects and reduced the noise level to zero for majority of the non-associated markers, we expected the EB method to be over conservative using the p_{Bonf} criterion. The p_{mBonf} criterion for the EB method has corrected this over conservative behaviour to a level comparable to the GEMMA method.

We then examined the impact of heritability on the performance of the methods compared under three levels of total heritability (0.25, 0.5 and 0.75) and four levels of per QTN heritability (0.3%, 1.4%, 3.3% and 5.0%) for both the human and pig data. The results are depicted in Figure 1 and Figure S1, where the upper panels show the power comparison under the three levels of heritability and the lower panels show the power comparison under four levels of per QTN heritability. The overall conclusions are that EB-mBonf has the highest power, followed by GEMMA and EB-Bonf, and FaSTt-LMM-Select has the lowest power, regardless whether a polygene was added or not.

Application to human breast cancer

The first GWAS study using the CGEMS breast cancer data identified several SNPs within a gene called FGFR2 (Hunter *et al.* 2007), which has been detected by many independent studies (Easton *et al.* 2007; Liang *et al.* 2008). We compared our EB method with two exact methods (GEMMA and FaST-LMM-Select), one approximate method (EMMAX) and a logistic regression method implemented in PLINK (Purcell *et al.* 2007). The FGFR2 gene is known to be located on chromosome 10. Therefore, we first analysed this chromosome only in detail and then followed by a whole-genome scanning. The Manhattan plots of chromosome 10 for methods GEMMA, FaST-LMM-Select and EB-mBonf are shown in Figure 2 and plots of the remaining two methods (EMMAX and PLINK) are illustrated in Figure S2. Although all methods show clear peaks for SNPs of gene FGFR2, only peaks of the EB-mBonf passed the significance threshold. The gene was detected and has been confirmed, but GEMMA and FaST-LMM-Select failed to detect it, indicating that the classical Bonferroni correction is over conservative. The modified Bonferroni correction with the effective number of tests has corrected this over conservation and declared significance of this gene. The two exact methods are fixed model approaches, and they cannot use the modified Bonferroni correction to lower the criterion.

The genomewide SNPs identified by our EB-mBonf method were listed in Table 3. The EB-mBonf identified three SNPs on chromosome 10, two SNPs located within gene FGFR2 and one in the proximity of gene BUB3 (77 kb). None of the SNPs reached the significance level using the other two methods (see Table 3). Gene BUB3 was recently discovered to be related to breast cancer (Briollais *et al.* 2014). The Manhattan plot of genomewide SNPs for EB-mBonf method along with the result from GEMMA and FaST-LMM-Select is shown in Figure S3.

Analysis of the CGEMS breast cancer data provided a good example to show the selective shrinkage nature of the EB method. As illustrated in Figure 2, majority of the markers are noises and their test statistics have been shrunken to zero by the new method. In addition, the EB method allows us to use the 'effective number of tests' to correct for multiple tests. Table 4 lists the number of SNPs and the effective number of tests for each chromosome. The total number of markers is 509 979, but the effective number of tests using the EB method is only 7766. The Bonferroni-corrected p-value for the genomewide significance threshold is $0.05/509\ 979 = 9.8 \times 10^{-8}$ using the total number of SNPs. This threshold has been reduced to $0.05/7766 = 6.44 \times 10^{-6}$ using the effective number of tests. Clearly, the Bonferroni correction using the total number of SNPs is too conservative. The SNPs in gene FGFR2 are already confirmed to be associated with breast cancer, but none of the SNPs were significant using GEMMA and FaST-LMM-Select with this conservative criterion. These two SNPs are significant using the EB method with the modified Bonferroni correction. However, the effective number of tests only applies to the empirical Bayes method (the random model approach), not the fixed model approach.

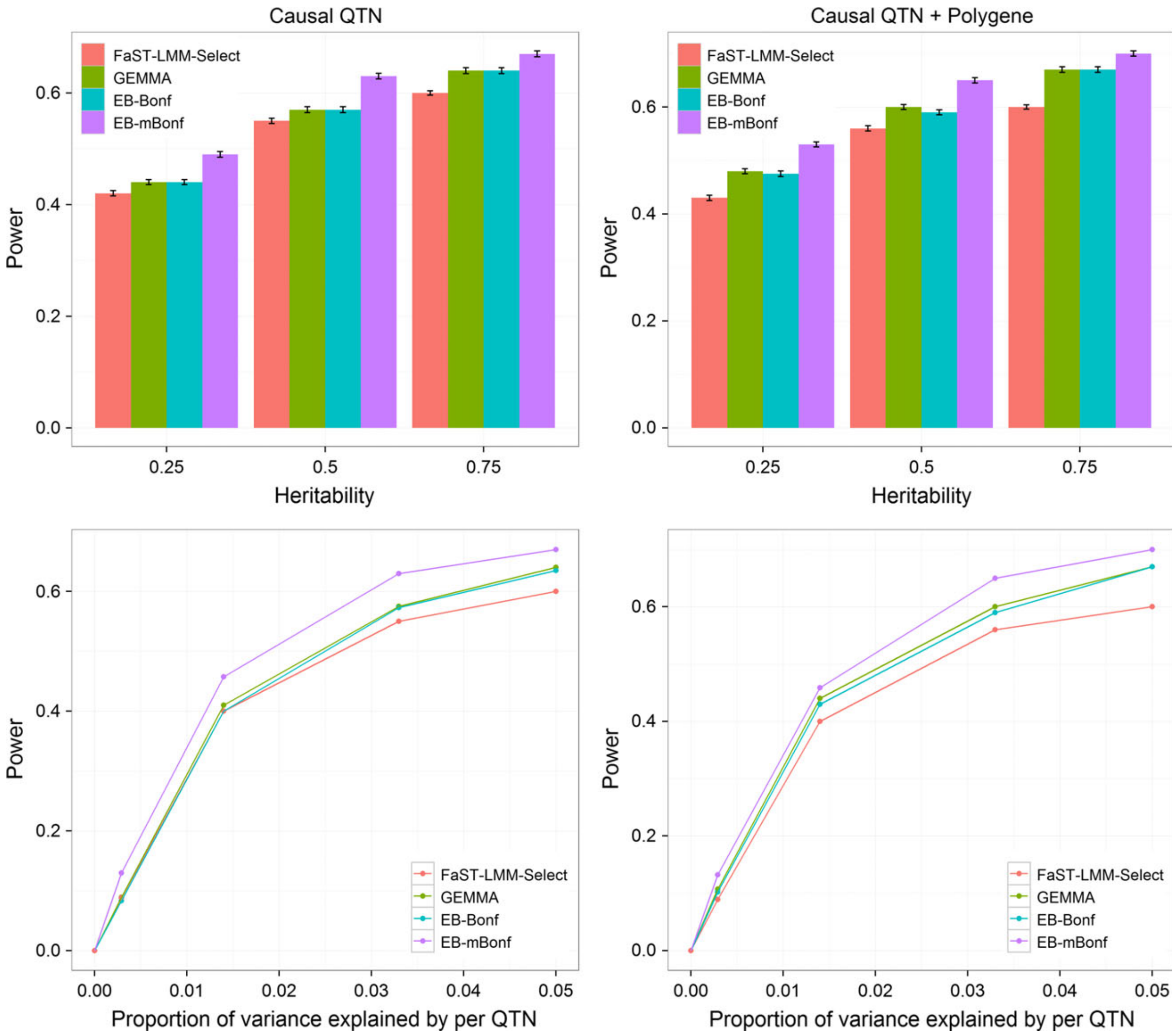


Figure 1 Empirical statistical powers of four methods (GEMMA, FaST-LMM-Select, EB-Bonf and EB-mBonf) drawn from 1000 replicated simulations from human data. The upper panels show the powers under three levels of heritability (0.25, 0.5 and 0.75). The lower panels show the powers of the four methods under four levels of per QTN heritability (0.3%, 1.4%, 3.3% and 5.0%). The left panels show the results of simulation without polygene. The right panels show the results of simulation with polygene.

Discussion

In summary, we developed a random model approach to GWAS in which the effect of a scanned marker is treated as a random variable and assumed to be normally distributed. When we consider the marker effect as the parameter of interest, the normal distribution can be interpreted as a prior distribution under the Bayesian framework. As the variance of the prior is estimated from the data, the estimated marker effect is called the empirical Bayes estimate. Unlike many other mixed model-based GWAS methods, the EB method was not designed for improving computational speed; rather, it was developed based on a new model and represented a methodology improvement. Properties of the EB method have been put forward in the introduction and demonstrated in both simulation studies and real-life data analyses. Improving marker detection signal, and thus increasing mapping resolution, was the original motivation of developing such a method (taking advantage of the shrinkage nature of the method). However, the method ended up with an increased power with an ‘effective number of tests’. The classical mixed model GWAS methods have been

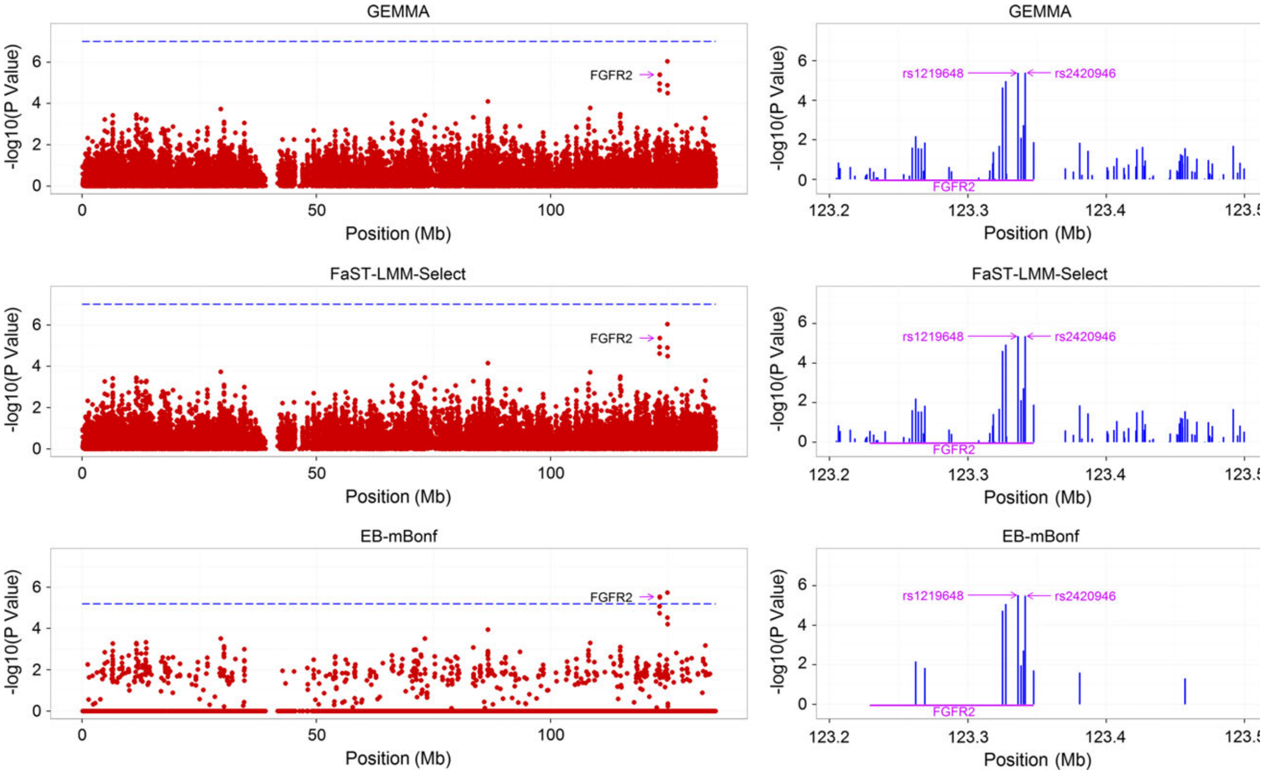


Figure 2 Comparison of GWAS results for chromosome 10 of human breast cancer in the CGEMS data set (left panels) and for gene FGFR2 (right panels) within chromosome 10 using three methods (GEMMA, FaST-LMM-Select and EB-mBonf). The blue dashed horizontal line of each panel of the Manhattan plots represents the Bonferroni-corrected or modified Bonferroni-corrected threshold of genomewide type I error of 0.05.

Table 3 Three SNPs identified by the EB-mBonf method for human breast cancer

	rs1219648	rs2420946	rs10510126
Chr.	10	10	10
Position (bp)	123346190	123351324	125002485
Closest gene	*FGFR2* (0 kb)	*FGFR2* (0 kb)	*BUB3* (77 kb)
EB-mBonf[a]	2.97E-06*	3.18E-06*	1.87E-06*
GEMMA[b]	4.08E-06	4.03E-06	9.26E-07
FaST-LMM-Selectb	4.30E-06	4.30E-06	9.20E-07
Validation	Hunter *et al.* (2007)	Hunter *et al.* (2007)	Briollais *et al.* (2014)

*The star indicates significance at <0.05 genomewide type I error.

[a]The modified Bonferroni-corrected threshold of the p value is 6.44E-06.

[b]The classical Bonferroni-corrected threshold of p value is 9.8E-08.

well developed, which are effective in preventing false-positive associations and increasing in statistical power (Yang *et al.* 2014). Any slight improvement over existing methods is proved to be difficult. However, the new EB method has shown a dramatic improvement in the statistical power and controlled the false-positive rate equally well as the widely used exact mixed model method GEMMA.

The EB method is to some degree similar to the Bayesian method in genomic prediction (Bayes B) proposed by Meuwissen *et al.* (2001) because both methods treated marker effects as random. The two methods differ in two aspects: (i) Bayes B is a multiple-marker model while EB proposed here is a single marker model and (ii) Bayes B is implemented via the MCMC algorithm but EB is a posterior mean-based method. Theoretically, Bayes B is preferred because it deals with the correct model while the single marker EB method deals with a simplified model that contains only one marker at a time. However, Bayes B cannot handle the large number of markers commonly seen in a GWAS data set (~1 million) within a reasonable computational time while the EB method has no limitation on the number of markers. Therefore, the single marker EB method will remain the first choice for GWAS in the foreseeable future.

The EB method of GWAS is computationally more demanding than the classical mixed model GWAS in which the scanned marker effect is treated as a fixed effect. Without resorting to special computational algorithms, it is unrealistic to use the EB method for GWAS. We adopted the eigendecomposition algo-

Table 4 Effective number of tests drawn from the EB analysis for the human breast cancer data

Chromosome	Number of SNPs	Effective number of tests
1	39 054	597
2	42 110	755
3	35 136	410
4	31 070	484
5	32 300	556
6	34 220	508
7	28 097	442
8	29 604	416
9	25 115	344
10	27 128	501
11	25 398	391
12	25 309	338
13	19 316	260
14	17 255	265
15	15 525	214
16	15 856	250
17	13 606	211
18	15 717	259
19	9057	124
20	13 356	198
21	7788	108
22	7962	135
Total	509 979	7766

rithm to ease the computation for the polygenic model and the Woodbury matrix identity to handle the full model. Combining the two special computational algorithms, the EB method can handle data set with large sample size and large number of markers that the classical mixed model GWAS can. The Woodbury matrix identity algorithm for calculating matrix inverse and determinant has been implemented by Lippert *et al.* (2011) in the classical mixed model GWAS, but they gave the algorithm a different name – factored spectrally transformed linear mixed models (FaST-LMM). In addition, they adopted the FaST algorithm for the polygenic model using a selected subset of genomewide markers to calculate the kinship matrix. Neither the eigendecomposition nor the FaST algorithm alone will allow us to estimate more than one genetic variance components. Combining the two algorithms, we have been able to estimate multiple genetic variance components under the random model approach for GWAS.

Excluding eigendecomposition of the kinship matrix, each evaluation of the log restricted likelihood has a run time that is linear in n. Therefore, for testing m markers, the time complexity is $O(n^3)$ for calculating the eigenvalues and eigenvectors of the kinship matrix, $O(n^2m)$ for calculating transformed variables U^Ty and U^TX, and $O(tnm)$ for performing t iterations of the log restricted likelihood for the optimization over λ, where t is the number of iterations of the Newton–Raphson or the Brent's algorithm used for finding the numeric solution. Consequently, the total time complexity of EB, given the kinship, is $O(n^3 + n^2m + tnm)$, which is the same as the full-rank form of FaST-LMM (Lippert *et al.* 2011) and GEMMA (Zhou & Stephens 2012). For the pure polygenic model, this complexity reduces to $O(n^3 + n^2m + tn)$.

One of the important properties of the EB method of GWAS is the availability of the prior variance for each marker. Combining the prior variance and the posterior variance, we are able to calculate the 'confidence' of each marker. The sum of the 'confidences' of all markers leads to an 'effective number of tests'. This effective number of tests is then used to perform the Bonferroni correction. It has been well known that Bonferroni correction using the total number of markers is over conservative, but there has been no easy way to fix the problem. For example, the human breast data GWAS showed high peaks in the test statistic profile for SNPs in gene FGFR2 on chromosome 10 and these SNPs have been confirmed in independent studies to be associated with breast cancer. However, they did not pass the criterion of Bonferroni correction using the total number of markers. The EB method, however, showed that these markers have passed the criterion of Bonferroni correction when the 'effective number of tests' is used for the correction. One may argue that if the effective number of tests based on Bonferroni correction was used for the classical mixed model GWAS, these SNPs would also be significant. However, an effective number of tests does not exist in the mixed model GWAS when the marker effects are treated as fixed.

Zhou & Stephens (2012) performed an exact method for the mixed model GWAS. Their mixed model treats marker effects as fixed and the polygenic effect as random. There is only one genetic variance (the polygenic variance). Therefore, they used the exact method to estimate the marker effect and the polygenic variance simultaneously for every marker. They claimed that the exact method may not be easily extended to multiple variance component estimation due to one singular value decomposition of the kinship matrix. Here, we show that eigendecomposition and Woodbury matrix identity can be used jointly to estimate multiple genetic variance components (Data S2 in Supplementary Material for more details). Theoretically, the exact method may prevent any undesirable properties of the approximate method where the polygenic variance is only estimated once under the pure polygenic model. In practice, the difference between the exact and approximate methods can be safely ignored. We did use the exact method to analyse the human data. The result is indeed much the same as the approximate method. The Manhattan plot comparison between the approximate and exact EB methods is shown in Figure S4.

Finally, the random model approach for GWAS may be easily extended to genetic association studies in populations initiated from multiple parents, for example the multiparent advanced generation intercross (MAGIC) population, where each marker involves multiple alleles. All allelic effects within a locus can be assigned the same normal distribution with a common variance. Regardless of how many parents, each locus only has one variance component to estimate and test. Similar extension can be made to handle genotype by environment (G × E) interaction. If the effect of a locus varies across multiple environments, a G × E is present. The variance of the effects across these environments represents exactly this G × E interaction. Therefore, it is natural to use the random model approach to performing G × E interaction analysis.

Conclusions

There are several advantages of the EB method for GWAS over existing procedures: (i) The shrinkage nature of EB estimate will selectively shrink marker effects and reduce the noise level to zero for majority of non-associated markers; (ii) It provides a new method for multiple test correction using an 'effective number of tests;' (iii) The new EB method still takes advantage of eigendecomposition and FaST-LMM to improve computational speed. The EB method may be easily extended to genetic association studies in populations initiated from multiple parents and G × E interaction. We therefore believe that EB method is a valuable tool for identifying the genetic mechanism of complex traits.

Software availability

We implemented the EB method in an R program that is available on our personal website (http://klab.sjtu.edu.cn/eb/) or can be downloaded from Baidu Cloud (http://pan.baidu.com/s/1mgBVAoK).

Acknowledgements

The project was supported by the United States Department of Agriculture National Institute of Food and Agriculture Grant 2007-02784 to SZX and the National Natural Science Foundation (grant no. 31370043, 31472069), National Key Technology Support Program (2012BAD12B08) and Shanghai Jiao Tong University Agri-X fund to QSW and YCP.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SZX designed the study. SZX and YCP supervised the study. QSW performed simulation and analysed the real data. SZX and QSW wrote the manuscript. QSW and JLW implemented the EB method in an R program. All authors read and approved the manuscript.

References

Briollais L., Dobra A., Liu J.N., Ozcelik. H., Massam H. (2014) A bayesian graphical model for genome-wide association studies(GWAS). *Submitted to J. Stat. Softw.*, http://research.lunenfeld.ca/mprime_briollais/?page=Publications

Easton D.F., Pooley K.A., Dunning A.M., Pharoah P.D., Thompson D., Ballinger D.G., Struewing J.P., Morrison J., Field H., Luben R. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087–1093.

Fan B., Onteru S.K., Du Z.Q., Garrick D.J., Stalder K.J., Rothschild M.F. (2011) Genome-wide association study identifies Loci for body composition and structural soundness traits in pigs. *PLoS ONE*, **6**, e14726.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. (2004) Bayesian Data Analysis. Chapman & Hall/CRC, New York.

Giri N.C. (2003) Multivariate statistical analysis: revised and expanded. *CRC Press*, **171**, 131–151.

Hunter D.J., Kraft P., Jacobs K.B., Cox D.G., Yeager M., Hankinson S.E., Wacholder S., Wang Z., Welch R., Hutchinson A., Wang J., Yu K., Chatterjee N., Orr N., Willett W.C., Colditz G.A., Ziegler R.G., Berg C.D., Buys S.S., McCarty C.A., Feigelson H.S., Calle E.E., Thun M.J., Hayes R.B., Tucker M., Gerhard D.S., Fraumeni J.F. Jr, Hoover R.N., Thomas G., Chanock S.J. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.*, **39**, 870–874.

Kang H.M., Zaitlen N.A., Wade C.M., Kirby A., Heckerman D., Daly M.J., Eskin E. (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.

Kang H.M., Sul J.H., Service S.K., Zaitlen N.A., Kong S.Y., Freimer N.B., Sabatti C., Eskin E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–54.

Liang J., Chen P., Hu Z., Zhou X., Chen L., Li M., Wang Y., Tang J., Wang H., Shen H. (2008) Genetic variants in fibroblast growth factor receptor 2 (FGFR2) contribute to susceptibility of breast cancer in Chinese women. *Carcinogenesis*, **29**, 2341–2346.

Lippert C., Listgarten J., Liu Y., Kadie C.M., Davidson R.I., Heckerman D. (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.

Listgarten J., Lippert C., Kadie C.M., Davidson R.I., Eskin E., Heckerman D. (2012) Improved linear mixed models for genome-wide association studies. *Nat. Methods*, **9**, 525–526.

MacKay D.J.C. (1992) Bayesian interpolation. *Neural Comput.*, **4**, 415–447.

Meuwissen T.H., Hayes B.J., Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J., Sham P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Segura V., Vilhjalmsson B.J., Platt A., Korte A., Seren U., Long Q., Nordborg M. (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.*, **44**, 825–830.

Tipping M.E. (2001) Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, **1**, 211–244.

Xu S.Z. (2013) Genetic mapping and genomic selection using recombination breakpoint data. *Genetics*, **195**, 1103–1105.

Yang J., Zaitlen N.A., Goddard M.E., Visscher P.M., Price A.L. (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, **46**, 100–106.

Yu J., Pressoir G., Briggs W.H., Vroh Bi.I., Yamasaki M., Doebley J.F., McMullen M.D., Gaut B.S., Nielsen D.M., Holland J.B., Kresovich S., Buckler E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**, 203–208.

Zhang Z., Ersoz E., Lai C.Q., Todhunter R.J., Tiwari H.K., Gore M.A., Bradbury P.J., Yu J., Arnett D.K., Ordovas J.M., Buckler E.S. (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.*, **42**, 355–360.

Zhou X., Stephens M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821–824.

Zhou X., Carbonetto P., Stephens M. (2013) Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.*, **9**, e1003264.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Empirical statistical powers of four methods (GEMMA, FaST-LMM-Select, EB-Bonf and EB-mBonf) drawn from 1000 replicated simulations from pig data.

Figure S2. Comparison of GWAS results for chromosome 10 of human breast cancer in the CGEMS data set (left panels) and for gene FGFR2 (right panels) within chromosome 10 for two additional methods not included in the main test (PLINK and EMMAX).

Figure S3. Manhattan plots of genome-wide SNPs for the human breast cancer data. The blue dashed horizontal line in each panel represents the Bonferroni corrected threshold at genome-wide type I error of 0.05.

Figure S4. Comparison of GWAS results for chromosome 10 of human breast cancer in the CGEMS data set (left panels) and for gene FGFR2 (right panels) within chromosome 10 for the exact and approximate EB methods (Data S2 in Supplementary Material for more details).

Data S1. Fast computation for the pure polygenic model.

Data S2. Exact EB method of GWAS.