Mapping Quantitative Trait Loci by Controlling Polygenic Background Effects

Shizhong Xu[1]
Department of Botany and Plant Sciences, University of California, Riverside, California 92521

ABSTRACT A new mixed-model method was developed for mapping quantitative trait loci (QTL) by incorporating multiple polygenic covariance structures. First, we used genome-wide markers to calculate six different kinship matrices. We then partitioned the total genetic variance into six variance components, one corresponding to each kinship matrix, including the additive, dominance, additive × additive, dominance × dominance, additive × dominance, and dominance × additive variances. The six different kinship matrices along with the six estimated polygenic variances were used to control the genetic background of a QTL mapping model. Simulation studies showed that incorporating epistatic polygenic covariance structure can improve QTL mapping resolution. The method was applied to yield component traits of rice. We analyzed four traits (yield, tiller number, grain number, and grain weight) using 278 immortal F2 crosses (crosses between recombinant inbred lines) and 1619 markers. We found that the relative importance of each type of genetic variance varies across different traits. The total genetic variance of yield is contributed by additive × additive (18%), dominance × dominance (14%), additive × dominance (48%), and dominance × additive (15%) variances. Tiller number is contributed by additive (17%), additive × additive (22%), and dominance × additive (43%) variances. Grain number is mainly contributed by additive (42%), additive × additive (19%), and additive × dominance (31%) variances. Grain weight is almost exclusively contributed by the additive (73%) variance plus a small contribution from the additive × additive (10%) variance. Using the estimated genetic variance components to capture the polygenic covariance structure, we detected 39 effects for yield, 39 effects for tiller number, 24 for grain number, and 15 for grain weight. The new method can be directly applied to polygenic-effect-adjusted genome-wide association studies (GWAS) in human and other species.

EPISTATIC effects refer to interactions of allelic effects between different loci. Depending on the natures of populations, the variance of epistatic effects for a quantitative trait may be partitioned into several different types of variance components, *e.g.*, additive × additive, additive × dominance, dominance × additive, and dominance × dominance (Cockerham 1954). The relative importance of each variance component usually varies across different traits. Accurate partitioning of these epistatic variance components can help understand the genetic mechanisms of complex traits and develop more efficient breeding programs. Prior to the genome era, special designs of experiments were required to partition the epistatic variance into different variance components, *e.g.*, the NC design III (Cockerham and Zeng 1996). The well-known Cockerham's epistatic model (Cockerham 1954) is the principle basis for data analysis. With molecular data, marker genotypes can be observed and thus complicated designs are no longer crucial. For example, the recombinant inbred line (RIL) design allows us to estimate the additive × additive effects between pairs of markers. The F2 design facilitates estimation of all the four types of epistatic effects. These simple designs of experiments are very popular in crops and laboratory animals.

Quantitative trait locus (QTL) mapping for epistatic effects has attracted much attention from geneticists (Cockerham and Zeng 1996; Kao and Zeng 2002; Yi and Xu 2002; Xu 2007; Xu and Jia 2007; Garcia *et al.* 2008). The main hurdle in epistatic effect QTL mapping is the large number of interaction effects to estimate. For low-density marker maps, all pairwise interactions can be fit simultaneously into a single model (Yi and Xu 2002; Xu 2007). With high-density marker maps, simultaneous fit of all genetic effects to a single model is prohibited. An *ad hoc* method is to fit one pair of loci at a time and the entire genome is then scanned in

doi: 10.1534/genetics.113.157032
Manuscript received July 7, 2013; accepted for publication September 16, 2013; published Early Online September 27, 2013.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.157032/-/DC1
[1]Address for correspondence: Department of Botany and Plant Sciences, University of California, Riverside, CA 92521. E-mail: shizhong.xu@ucr.edu

a two-dimensional approach (Bell *et al.* 2011). It is well known in genome-wide association studies (GWAS) that such a genome-wide scanning has ignored the polygenic effect and thus will inflate the residual error variance (Yu *et al.* 2006; Zhang *et al.* 2010; Zhou and Stephens 2012). As a consequence, the statistical power may be low. When the pedigree relationship of individuals is known, one can fit a polygenic effect to the model using the standard linear mixed-model approach (Henderson 1975), as done for the inbred lines of maize (Zhang *et al.* 2005). However, the pedigree relationship is rarely known in a randomly sampled population. Yu *et al.* (2006) proposed using marker-inferred kinship matrix to capture the polygenic variance. This approach has now become the standard method for GWAS. Efficient algorithms have been proposed afterward to improve the computational efficiency (Kang *et al.* 2008; Zhang *et al.* 2010; Lippert *et al.* 2011; Zhou and Stephens 2012). However, such a polygenic mixed linear model has not been investigated in GWAS for epistatic effects.

QTL mapping (genome-wide linkage studies) is slightly different from GWAS because the target populations for the two approaches of genetic mapping are often different. However, if analyses are focused on markers only (ignoring pseudo-markers between observed markers), the statistical methods are much the same for the two approaches except that the random population GWAS requires an extra term in the model to control for population structures. The linear mixed-model GWAS incorporating marker-inferred kinship matrix as the covariance structure can be naturally adopted here for QTL mapping to capture the background genetic information. For a well-controlled mapping population, *e.g.*, the F2 population, the marker inferred kinship matrix is equivalent to the identical-by-descent (IBD) matrix of polygenes. Using this matrix to model the covariance structure plays a similar role to the cofactors added to the interval mapping, the so-called composite interval mapping as proposed by Jansen (1994) and Zeng (1994). A more appealing feature of this kinship-corrected QTL mapping approach is that it is easy to control the genetic background information caused by epistatic effects. The original composite interval mapping, in principle, can take into account pairwise interaction terms as cofactors, but no attempt has been made because of the difficulty of choosing the pairwise cofactors. If the covariance structure includes kinship matrices for various epistatic effects, the genetic background due to epistatic effects can be controlled.

Motivated by the above arguments, we propose a mixed-model GWAS for epistatsis using ultra-high-density SNP markers. Xie *et al.* (2010) and Yu *et al.* (2011) released a high-density SNP map containing about 270,000 SNPs. From the SNPs, they inferred recombination breakpoints for 240 RILs of a rice population derived from two inbred lines. The authors further combined consecutive SNP markers of cosegregation into bins. Within a bin, there are no breakpoints and thus all SNPs within the same bin have exactly the same genotypes. The >270,000 SNPs were eventually converted into 1619 bins. Each bin was then treated as a synthetic marker. Genetic analysis was then performed on the bin genotypes. Using the binned genotypes, Yu *et al.* (2011) conducted QTL mapping for seven yield component traits in rice and identified numerous QTL. Using the 240 RILs, Hua *et al.* (2002) and Hua *et al.* (2003) created 360 crosses to dissect epistatic effects and to investigate the genetic basis of heterosis for yield component traits. Most recently, Zhou *et al.* (2012) reinvestigated the genetic basis of heterosis using 278 of the 360 crosses. They called the RIL-generated crosses immortal F2 (IMF2) because the genotypes of the crosses mimic the genotypes of the regular F2 population. The IMF2 crosses are also called recombinant inbred intercrosses (RIX) in laboratory mice derived from crosses of multiple RILs (Zou *et al.* 2005). Zhou *et al.* (2012) used the 1619 bins as synthetic markers to test all pairwise interaction effects using a two-dimensional scanning approach by fitting one pair of bins at a time. The purpose of their study was to evaluate the relative importance of dominance and epistasis to heterosis.

In this study, we analyzed the same IMF2 population using genotypes of the 1619 bins. We first evaluated the relative importance of each type of epistatic effects on four yield component traits using variance component analysis and then mapped epistatic effects using marker-generated kinship matrices to control the epistatic genetic background.

Material and Methods

Plant materials and SNP bin map

An IMF2 population of rice was investigated in this study. The IMF2 population was generated by Dr. Qifa Zhang's laboratory and previously described in detail by Hua *et al.* (2002) and Hua *et al.* (2003). It consisted of 360 crosses made by random matches of 240 RILs derived by single-seed descent from a cross between Zhenshan 97 and Minghui 63. Field data of yield (YIELD), number of tillers per plant (TILLER), number of grains per panicle (GRAIN), and thousand grain weight (KGW) were collected in the 1998 and 1999 rice growing seasons from replicated field trials on the experimental farm of Huazhong Agricultural University, Wuhan, China. Over a quarter million high-density SNP markers were used to infer recombination breakpoints (crossovers), which were then used to construct bins. The bins were treated as "new markers" for association studies. The number of bins in this rice population was 1619. The bin map was constructed by genotyping the RILs with population sequencing (Xie *et al.* 2010; Yu *et al.* 2011). Among the 360 crosses, only 278 of them were available in both phenotypes and bin genotypes. Therefore, the bin genotype data were stored in a $n \times m = 278 \times 1619$ matrix. Each genotype was coded by A for the Zhenshan 97 genotype, B for the Minghui 63 genotype, and H for the heterozygote. The genotypes and phenotypes were downloaded from the journal website of the study by Zhou *et al.* (2012). For each trait, there were two temporal replications (years

Table 1 Formulas used to calculate marker generated kinship matrices

Type of effect	Original kinship matrix	Normalization factor[a]	Kinship matrix
Additive (*a*)	$K_a^* = \sum_{k=1}^{m} Z_k Z_k^{\mathsf{T}}$	$c_a = \text{mean}[\text{diag}(K_a^*)]$	$K_a = (1/c_a)K_a^*$
Dominance (*d*)	$K_d^* = \sum_{k=1}^{m} W_k W_k^{\mathsf{T}}$	$c_d = \text{mean}[\text{diag}(K_d^*)]$	$K_d = (1/c_d)K_d^*$
Additive × additive (*aa*)	$K_{aa}^* = \sum_{k=1}^{m-1}\sum_{k'=k+1}^{m}(Z_k \# Z_{k'})(Z_k \# Z_{k'})^{\mathsf{T}}$	$c_{aa} = \text{mean}[\text{diag}(K_{aa}^*)]$	$K_{aa} = (1/c_{aa})K_{aa}^*$
Dominance × dominance (*dd*)	$K_{dd}^* = \sum_{k=1}^{m-1}\sum_{k'=k+1}^{m}(W_k \# W_{k'})(W_k \# W_{k'})^{\mathsf{T}}$	$c_{dd} = \text{mean}[\text{diag}(K_{dd}^*)]$	$K_{dd} = (1/c_{dd})K_{dd}^*$
Additive × dominance (*ad*)	$K_{ad}^* = \sum_{k=1}^{m-1}\sum_{k'=k+1}^{m}(Z_k \# W_{k'})(Z_k \# W_{k'})^{\mathsf{T}}$	$c_{ad} = \text{mean}[\text{diag}(K_{ad}^*)]$	$K_{ad} = (1/c_{ad})K_{ad}^*$
Dominance × additive (*da*)	$K_{da}^* = \sum_{k=1}^{m-1}\sum_{k'=k+1}^{m}(W_k \# Z_{k'})(W_k \# Z_{k'})^{\mathsf{T}}$	$c_{da} = \text{mean}[\text{diag}(K_{da}^*)]$	$K_{da} = (1/c_{da})K_{da}^*$

[a] Each marker-generated kinship matrix is normalized so that the diagonal elements are roughly equal to one. This is achieved by dividing the original kinship matrix by the mean value of all the diagonal elements of the original matrix. Using the normalized kinship matrix will bring the estimated genetic variance in the same scale as the residual error variance.

1998 and 1999). The phenotypic values of the two replicates were pooled for each cross after removing the year effects using

$$y_j = \frac{1}{2}\Big[\big(y_{j1} - \bar{y}_1\big) + \big(y_{j2} - \bar{y}_2\big)\Big], \tag{1}$$

where $\bar{y}_1$ and $\bar{y}_2$ are the mean values of the trait measured in 1998 and 1999, respectively. This pooled trait value was treated as the actual phenotypic value for analysis. Apparently, we ignored the genotype by year interaction (G × E) effects, if there is any.

Polygenic variance component analysis

We first numerically coded the genotype of individual j at bin k into two variables,

$$Z_{jk} = \begin{cases} +1 & \text{for } A \\ 0 & \text{for } H \\ -1 & \text{for } B, \end{cases} \qquad W_{jk} = \begin{cases} 0 & \text{for } A \\ 1 & \text{for } H \\ 0 & \text{for } B, \end{cases} \tag{2}$$

where Z_{jk} and W_{jk} represent the codes for additive and dominance effects, respectively. Let y be an $n \times 1 = 278 \times 1$ column vector for the phenotypic values of all the IMF2 crosses. The full epistatic model for $m = 1619$ bins is described by

$$\begin{aligned} y = X\beta &+ \sum_{k=1}^{m} Z_k a_k + \sum_{k=1}^{m} W_k d_k + \sum_{k=1}^{m-1}\sum_{k'=k+1}^{m}(Z_k \# Z_{k'})(aa)_{kk'} \\ &+ \sum_{k=1}^{m-1}\sum_{k'=k+1}^{m}(W_k \# W_{k'})(dd)_{kk'} + \sum_{k=1}^{m-1}\sum_{k'=k+1}^{m}(Z_k \# W_{k'})(ad)_{kk'} \\ &+ \sum_{k=1}^{m-1}\sum_{k'=k+1}^{m}(W_k \# Z_{k'})(da)_{kk'} + \varepsilon, \end{aligned} \tag{3}$$

where $X\beta$ represents some nongenetic effects, *e.g.*, intercept in this case, and $\varepsilon \sim N(0, I\sigma^2)$ is a vector of residual errors with a normal distribution of unknown variance σ^2. In addition, $Z_k \# W_k$ represents element-wise vector multiplication, and a_k and d_k are the additive and dominance effects, respectively, for bin k. The four terms, $(aa)_{kk'}$, $(dd)_{kk'}$, $(ad)_{kk'}$ and $(da)_{kk'}$ are the additive × additive, dominance × dominance, additive × dominance and dominance × additive effects, respectively, between bins k and k'. These four terms are called the epistatic effects. The total number of genetic effects in the above model is

$$2m + 4m(m-1)/2 = 2m^2 = 2 \times 1619^2 = \ 5,242,322, \tag{4}$$

which is beyond the capability of any models currently available in genome-wide association studies in a simultaneous manner. By treating each genetic effect as a randomly distributed normal variable with mean zero and a common variance across all bins or pairs of bins, the model becomes a mixed model. Let $a_k \sim N(0, \sigma_a^2)$, $d_k \sim N(0, \sigma_d^2)$,$(aa)_{kk'} \sim N(0, \sigma_{aa}^2)$, $(dd)_{kk'} \sim N(0, \sigma_{dd}^2)$, $(ad)_{kk'} \sim N(0, \sigma_{ad}^2)$, and $(da)_{kk'} \sim N(0, \sigma_{da}^2)$ be the distributions. The fact that the variance of each type of genetic effects does not depend on the bins or bin pairs implies that all effects of the same kind share a common genetic variance. These common variances are called polygenic variances. The expectation of y is $\text{E}(y) = X\beta$ and the variance–covariance matrix of y is

$$\begin{aligned} \text{var}(y) = &K_a\sigma_a^2 + K_d\sigma_d^2 + K_{aa}\sigma_{aa}^2 + K_{dd}\sigma_{dd}^2 + K_{ad}\sigma_{ad}^2 \\ &+ K_{da}\sigma_{da}^2 + I\sigma^2, \end{aligned} \tag{5}$$

where the K's are marker-generated kinship matrices; *i.e.*, they depend on Z_k and W_k. Formulas to calculate these kinship matrices are given in Table 1. Note that the kinship matrices for epistatic variances presented here are different from the one given by Ober *et al.* (2012), who used the Hadamard square of the additive kinship matrix as the additive × additive kinship matrix. They did not detect any epistatic variance in their *Drosophila melanogaster* population, likely due to the use of wrong additive × additive kinship matrix. Given these marker-generated kinship matrices, the variance components were estimated using standard mixed-model procedures (Henderson 1975).

One may rewrite the variance–covariance matrix in Equation 5 by

$$\begin{aligned} \text{var}(y) = &(K_a\lambda_a + K_d\lambda_d + K_{aa}\lambda_{aa} + K_{dd}\lambda_{dd} + K_{ad}\lambda_{ad} \\ &+ K_{da}\lambda_{da} + I)\sigma^2, \end{aligned} \tag{6}$$

where $\lambda_a = \sigma_a^2/\sigma^2$ is the ratio of the additive genetic variance to the residual variance and other λ's are defined similarly. A more useful measurement of the relative importance of a type of genetic effect, *e.g.*, the additive effect, is

$$h_a^2 = \frac{\sigma_a^2}{\sigma_P^2} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_d^2 + \sigma_{aa}^2 + \sigma_{dd}^2 + \sigma_{ad}^2 + \sigma_{da}^2 + \sigma^2}, \quad (7)$$

where the denominator (the sum of all variance components) is called the phenotypic variance. The above ratio is called the narrow-sense heritability. The broad-sense heritability is define by

$$H = \frac{\sigma_a^2 + \sigma_d^2 + \sigma_{aa}^2 + \sigma_{dd}^2 + \sigma_{ad}^2 + \sigma_{da}^2}{\sigma_a^2 + \sigma_d^2 + \sigma_{aa}^2 + \sigma_{dd}^2 + \sigma_{ad}^2 + \sigma_{da}^2 + \sigma^2}. \quad (8)$$

Note that the broad-sense heritability using marker-generated kinship is often close to unity when the marker density is sufficiently high. Therefore, it is not an informative piece of information. However, the proportion of each variance component relative to the total phenotypic variance is informative and is the focus of discussion in this study.

Model for genome scanning

Individual bin effects and bin pair interaction (epistatic) effects can be estimated after fitting the polygenic effects. The idea of the mixed-model GWAS (Yu *et al.* 2006) can be adopted here. The polygenic term remains in the model except that we used six kinship matrices to describe the covariance structure. The genetic effects of each bin and the epistatic effects of each pair of bins are modeled like the usual GWAS (Yu *et al.* 2006) with the polygenic structure included. The epistatic model appears like

$$\begin{aligned} y = {} & X\beta + \xi + \varepsilon \\ & + Z_k a_k + W_k d_k + Z_{k'} a_{k'} + W_k d_{k'} + (Z_k \# Z_{k'})(aa)_{kk'} \\ & + (W_k \# W_{k'})(dd)_{kk'} + (Z_k \# W_{k'})(ad)_{kk'} \\ & + (W_k \# Z_{k'})(da)_{kk'} \end{aligned} \quad (9)$$

for loci k and k', where ξ is the polygenic effect (the sum of all genetic effects across the entire genome). There are two additive effects, two dominance effects, and four epistatic effects (denoted by double symbols with double subscripts) for each pair of bins. A complete scan under the epistatic model requires a two-dimensional search for all pairs of bins. We noted that the additive and dominance effects would be estimated multiple times. The inclusion of polygenic effects allows us to search for bin effects and epistatic effects separately. The model that includes the additive and dominance effects is

$$y = X\beta + Z_k a_k + W_k d_k + \xi + \varepsilon. \quad (10)$$

A complete scan for this model (10) requires m calls of the above model and two genetic effects are estimated under each call. The epistatic effect model excluding the additive and dominance effects is

$$\begin{aligned} y = {} & X\beta + \xi + \varepsilon + (Z_k \# Z_{k'})(aa)_{kk'} + (W_k \# W_{k'})(dd)_{kk'} \\ & + (Z_k \# W_{k'})(ad)_{kk'} + (W_k \# Z_{k'})(da)_{kk'}. \end{aligned} \quad (11)$$

A complete scan for this model (11) requires $m(m-1)/2$ calls of the above model and four epistatic effects are estimated under each call. The mixed model can be written generically as

$$y = X\beta + Z\gamma + e, \quad (12)$$

where $e = \xi + \varepsilon$ is a general error term with $E(e) = 0$ and

$$\mathrm{var}(e) = \mathrm{var}(\xi) + I\sigma^2 = (K + I)\sigma^2. \quad (13)$$

Matrix K is a weighted sum of all effect-specific kinship matrices, it is not the kinship matrix given by Yu *et al.* (2006) in the original GWAS study,

$$K = K_a\lambda_a + K_d\lambda_d + K_{aa}\lambda_{aa} + K_{dd}\lambda_{dd} + K_{ad}\lambda_{ad} + K_{da}\lambda_{da}, \quad (14)$$

and the weights are the ratios of the variance components to the residual variance. For the additive and dominance effect model, $Z = Z_k||W_k$ and $\gamma = [\,a_k \quad d_k\,]^{\mathrm{T}}$. For the epistatic effect model,

$$\begin{aligned} & Z = (Z_k \# Z_{k'})||(W_k \# W_{k'})||(Z_k \# W_{k'})||(W_k \# Z_{k'}) \\ & \gamma = [\,(aa)_{kk'}(dd)_{kk'}(ad)_{kk'}(da)_{kk'}]^{\mathrm{T}}, \end{aligned} \quad (15)$$

where the special notation $a||b$ indicates column concatenation of matrices a and b. The vector of genetic effects γ has been treated as fixed effects in the classical GWAS (Yu *et al.* 2006; Zhou and Stephens 2012). In this study, we treat γ as random effects with a multivariate $N(0, I\sigma_\gamma^2)$ distribution. The shared common variance σ_γ^2 is estimated from the data. When σ_γ^2 is set to ∞, the genetic effects γ become fixed effects.

A two-step approach to genome scanning

Polygenic analysis: The mixed-model analysis would include seven genetic variance components (six polygenic variances and one genetic variance for each bin or bin pair). The six polygenic variances change very little across bins or bin pairs. Therefore, we adopted the two-step approach for parameter estimation to save computational time (Zhang *et al.* 2010; Lippert *et al.* 2011). In the first step, we estimated only the polygenic variances using the model

$$y = X\beta + \xi + \varepsilon, \quad (16)$$

where

$$\mathrm{var}(y) = \mathrm{var}(\xi) + \mathrm{var}(\varepsilon) = (K + I)\sigma^2. \quad (17)$$

This model allows us to estimate the six variance ratios, λ_X, which are needed to construct the K matrix. The K matrix estimated from this model is then treated as a constant (known value) in the second step. The polygenic model was treated as the null model for the next step genome scan. The likelihood function evaluated under the polygenic model is denoted by L_0 in the subsequent likelihood ratio tests.

Genome scanning: We proposed the following eigen decomposition algorithm that does not require repeated inverse of matrix $K+I$. Eigen decomposition for GWAS was originally proposed by Lippert *et al.* (2011) and later extended by Zhou and Stephens (2012). The eigen decomposition for the kinship matrix is $K = UDU^{\mathrm{T}}$, where U is the eigenvector (an $n \times n$ matrix) and $D = \mathrm{diag}(\delta_1, ..., \delta_n)$ is the eigenvalue (a diagonal matrix). One property of the eigen decomposition is $K^{-1} = UD^{-1}U^{\mathrm{T}}$, which has avoided matrix inversion because D^{-1} is simply the inverse of a diagonal matrix. The error variance is rewritten by

$$\begin{aligned}\mathrm{var}(e) &= \mathrm{var}(\xi + \varepsilon) = (K+I)\sigma^2 = (UDU^T + I)\sigma^2 \\ &= U(D+I)U^T\sigma^2. \end{aligned} \tag{18}$$

Before calling the mixed-model procedure, we performed data transformations so that the mixed model is rewritten by

$$U^{\mathrm{T}}y = U^{\mathrm{T}}(X\beta + Z\gamma + e) = U^{\mathrm{T}}X\beta + U^{\mathrm{T}}Z\gamma + U^{\mathrm{T}}e. \tag{19}$$

Let $y^* = U^{\mathrm{T}}y$ be the transformed phenotypic values. The model for the transformed data is

$$y^* = X^*\beta + Z^*\gamma + e^*. \tag{20}$$

The expectation and variance matrix of the transformed data are $E(y^*) = X^*\beta$ and

$$\begin{aligned}\mathrm{var}(e^*) &= \mathrm{var}(U^{\mathrm{T}}e) = U^{\mathrm{T}}(K+I)U\sigma^2 = U^{\mathrm{T}}(UDU^T + I)U\sigma^2 \\ &= (D+I)\sigma^2, \end{aligned} \tag{21}$$

respectively. Let

$$W = (D+I)^{-1} = \mathrm{diag}\left[(\delta_1 + 1)^{-1}, ..., (\delta_n + 1)^{-1}\right] \tag{22}$$

be a weight matrix. The residual variance structure for the transformed data are

$$\mathrm{var}(e^*) = R = W^{-1}\sigma^2. \tag{23}$$

The mixed-model equations for BLUE of β and BLUP of γ are

$$\begin{bmatrix} X^{*T}WX^* & X^{*T}WZ^* \\ Z^{*T}WX^* & Z^{*T}WZ^* + I/\sigma_\gamma^2 \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} X^{*T}Wy^* \\ Z^{*T}Wy^* \end{bmatrix}. \tag{24}$$

The null hypothesis may be tested in one of two ways. One way is the likelihood ratio test under the null model $H_0 : \sigma_\gamma^2 = 0$. The other way is the Wald test under the null model $H_0 : \gamma = 0$. The mixed-model equations provide both the BLUP of γ and the variance–covariance matrix of the BLUP estimate. The Wald test statistic is

$$\mathrm{Wald} = \hat{\gamma}^{\mathrm{T}}[\mathrm{var}(\hat{\gamma})]^{-1}\hat{\gamma} \tag{25}$$

When $\gamma = 0$ is true, the Wald test statistic follows a chi-square distribution with 2 degrees of freedom for the additive-dominance model and 4 degrees of freedom for the epistatic effect model.

Software implementation

All data analyses were implemented using the restricted maximum-likelihood (REML) method (Patterson and Thompson 1971) implemented via the MIXED and HPMIXED procedures in SAS (SAS Institute 2009b). PROC IML (SAS Institute 2009a) was also used to calculate eigenvalues and eigenvectors and to perform data transformations. The polygenic variance component analysis was conducted using PROC MIXED. Genome scans were implemented using PROC HPMIXED, which is a specialized high-performance version of the MIXED procedure. The SAS codes are provided in Supporting Information, File S11 for PROC MIXED and File S12 for PROC HPMIXED. The six types of kinship matrices are given in File S1. The phenotypes of four quantitative traits are given in File S2. Genotypes of two bins (bins 729 and 1064) are given in File S3.

Results

Results of a simulation experiment

Before analyzing the rice data, we wanted to show the difference of genomic scans between the model including the polygenic variances and the model without the polygenic variances using simulated data. Similar comparison has been performed in the traditional mixed-model GWAS that ignored the epistatic polygenic variance (Yu *et al.* 2006). We used the genotypes of 217 bins on the first chromosome for the 278 IMF2 crosses of rice as the simulated genotypes and added some effects on bins and bin pairs to generate genetic values of all crosses. We then added a normally distributed residual error to the genetic value of each cross to form a phenotypic value for the cross. To simplify the model, we simulated only the additive effects and the additive × additive effects. For 217 bins, the model included 217 additive effects and $217(217-1)/2 = 23437$ additive × additive effects. Therefore, the polygenic model contained two genetic variance components.

In the first simulation experiment, we added only one additive effect on bin 62 and one additive × additive effect on bin pair (112, 182). The true effects and estimated effects are illustrated in Figure 1, red and blue, respectively. Figure 1, A and B, shows the main effects and the epistatic effects, respectively, under the model that ignored the polygenic covariance structure. Figure 1, C and D, shows the main effects and the epistatic effects, respectively, under the model that incorporated only the additive covariance structure. The improvement can be seen for the main effects but not the epistatic effects. Figure 1, E and F, shows the main effects and the epistatic effects, respectively, under the model that incorporated both the additive and the additive × additive covariance structures. Further improvement can be seen, especially for the epistatic effects where the peak of the estimated effects is much

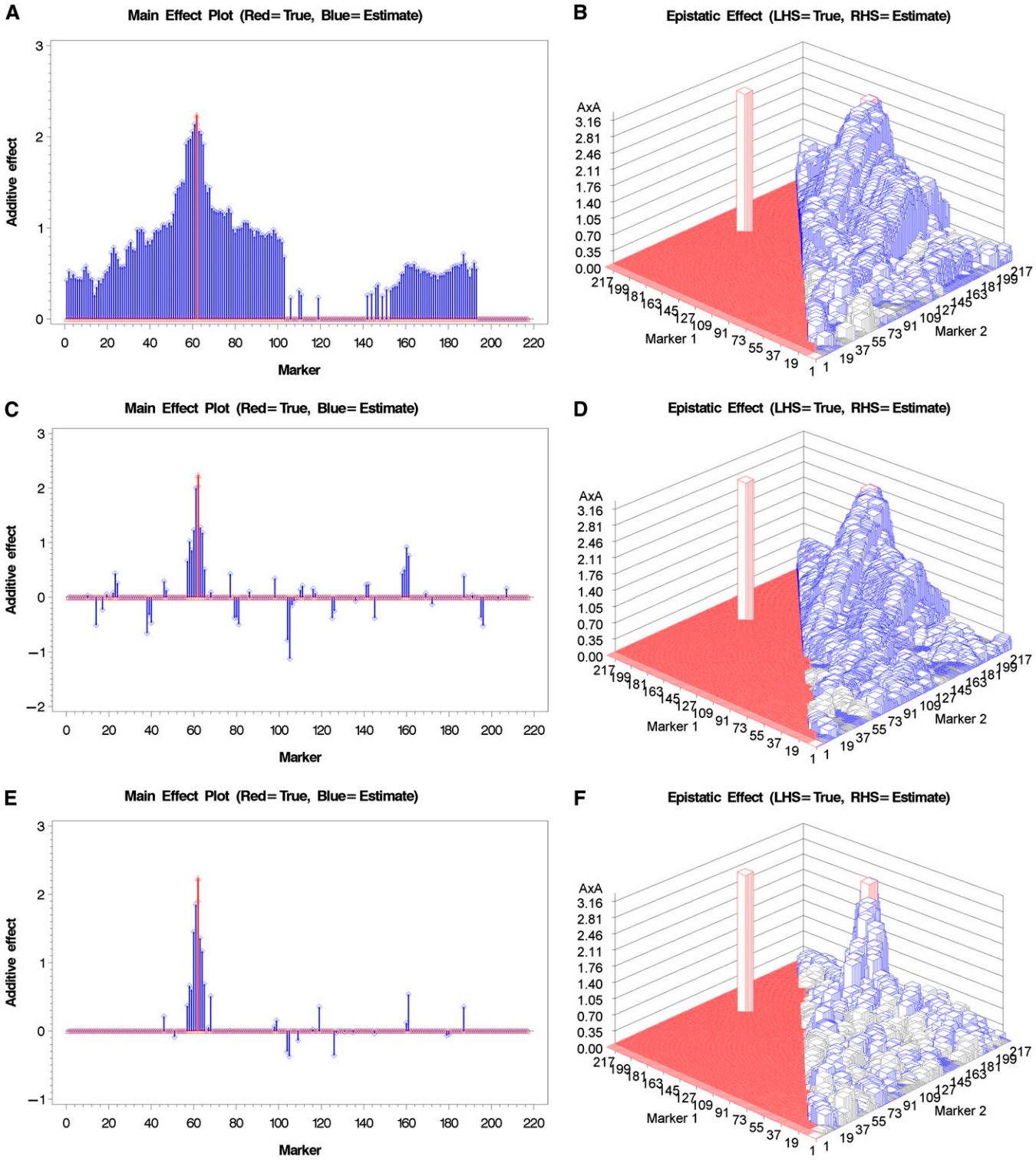

Figure 1 Comparison of models with and without fitting the polygenic covariance structures in the first simulation experiment. (A and B) The main and epistatic effects, respectively, under the model without the polygenic covariance structure. (C and D) The main and epistatic effects, respectively, under the model with only the additive covariance structure. (E and F) The main and epistatic effects, respectively, under the model with both the additive and the additive × additive covariance structures. True and estimated effects are in red and blue, respectively.

sharper than the previous two models. For the simple experiment, the model ignoring polygenic covariance structure already performed well.

In the second simulation experiment, we added 9 additive effects and 13 additive × additive interaction effects. The true effects and the estimated effects are illustrated in

Figure 2, red and blue, respectively. The improvement of the models including the polygenic effects over the model ignoring them is more obvious in the complicated experiment. For the model ignoring the polygenic covariance structure (Figure 2, A and B), peaks of the estimated main effects do not correspond to the true values at all. The estimated epistatic effects correspond to the true effects when the true effects are large. When the additive covariance structure is incorporated in the model, improvement for the main effects is obvious (Figure 2C) but the epistatic effects have very little improvement (Figure 2D). Figure 2, E and F, show the main and epistatic effects, respectively, for the model incorporated both the additive and additive × additive covariance structures. The improvement is much clearer compared with the previous two models. The simulation experiments by no means were exhausted but they provided visual evidence that the polygenic-effect-adjusted model performed better than the model without the polygenic effects. Incorporating epistatic covariance structure led to further improvement.

Polygenic variance analysis for yield component traits in rice

Results of the variance component analyses are summarized in Table 2. The variance ratios ($\lambda_X = \sigma_X^2/\sigma^2$) are also listed in the table and they were used to calculate the weighted sum of different kinship matrices. The most important information from the table is the proportion of the phenotypic variance contributed by each type of genetic variance ($h_X^2 = \sigma_X^2/\sigma_P^2$). Several general conclusions were drawn from the observations: (1) None of the four traits were contributed by the dominance variance; (2) the additive variances contributed the most in two of the four traits (GRAIN and KGW); (3) the additive × dominance variance is important for YIELD and dominance × additive variance is important for TILLER; (4) the KGW trait was almost exclusively controlled by the additive variance; and (5) all the four traits had high broad sense heritability with KGW being the highest ($H = 0.9636$). The goodness of fit (squared correlation between predicted and observed phenotypic values) was almost perfect for all traits. The likelihood ratio test (LRT) statistics for the polygenic effect model (including all six genetic variance components) were all high (37.9, 58.2, 82.2, and 248.9) with extremely small *P*-values (1.17*E*–06, 1.04 *E*–10, 1.25 *E*–15, and 7.04 *E*–51) for the four traits. The significant polygenic variance for each of the four traits implied that polygenic-effect-adjusted model should perform better than the model ignoring the polygenic effects. Note that the goodness of fit is not exactly the same as the broad sense heritability because the two are calculated using different formulas, in which var(y) was calculated from the data and σ_P^2 was calculated from the sum of all variance components. The polygenic analysis provides only a general picture of the relative importance of each type of variance components. When a variance component is zero, it means that this type of effect is not as important as some other types of effects in general. It does not exclude the possibility that some bins may have significant effects of this type, which have been "diluted" in the polygenic analysis (see more discussion in the last section). File S11 gives the SAS code of PROC MIXED for the polygenic analysis.

Genome scans for yield component traits in rice

Main effect analysis: We used two models to perform the main effect analysis. One model contains the additive and dominance effects of each bin plus the polygenic effect captured by the polygenic covariance structure (all six types of polygenic variances were considered). The method is called weighted mixed-model analysis. The second model contains the additive and dominance effects for each bin but ignores the polygenic covariance structure. This method is called unweighted mixed-model analysis. The comparison provides a real life demonstration of the differences between the two models. The LRT statistic was used to declare statistical significance for each bin. The critical values for the LRT test statistics were determined via multiple permutations to adjust for multiple tests (Churchill and Doerge 1994). The polygenic covariance structures were reserved. The number of reshuffled samples in the permutation tests was 1000 per trait within each method. The maximum LRT was recorded for each permuted sample. For a total of 1000 permuted samples, we obtained 1000 such maximum LRT values. These maximum LRT values form an empirical null distribution. The LRT value of each bin from the original data analysis (not from the reshuffled data) was then compared to the null distribution. The *P*-value is the proportion of the 1000 permuted samples whose LRT values are greater than the LRT of the original data analysis. First, let us evaluate the critical values of the test statistics from the null distributions. The 90, 95, 99, and 100 percentiles of the null distributions are listed in Table 3 (the top and middle portions of the table). The critical values for the unweighted method are always higher than the weighted method. For example, the critical value for YIELD at type I error of 0.05 is 3.6711 for the weighted method, but the corresponding value for the unweighted method is 10.0605. The critical values also vary across different traits, indicating that permutation tests are necessary.

Figure 3 gives the LRT profiles for each trait under the two models. The weighted method usually has low LRT but with sharper peaks than the unweighted method. Due to lower empirical critical values for the weighted method, the powers are not necessarily low. In fact, many peaks co-occur between the two methods. For example, the peaks on chromosomes 6, 7, and 8 for YIELD are very similar between the two methods. Differences between the two methods are also obvious. Taking trait YIELD, for example, the weighted mixed model shows a sharp peak in the beginning of chromosome 12 ($P < 0.05$) but the unweighted method shows a much wider peak in the middle of the chromosome and this wide peak is not significant ($P > 0.05$). Because the peaks for the unweighted method are usually wider, more bins are significant but these significant bins are clustered

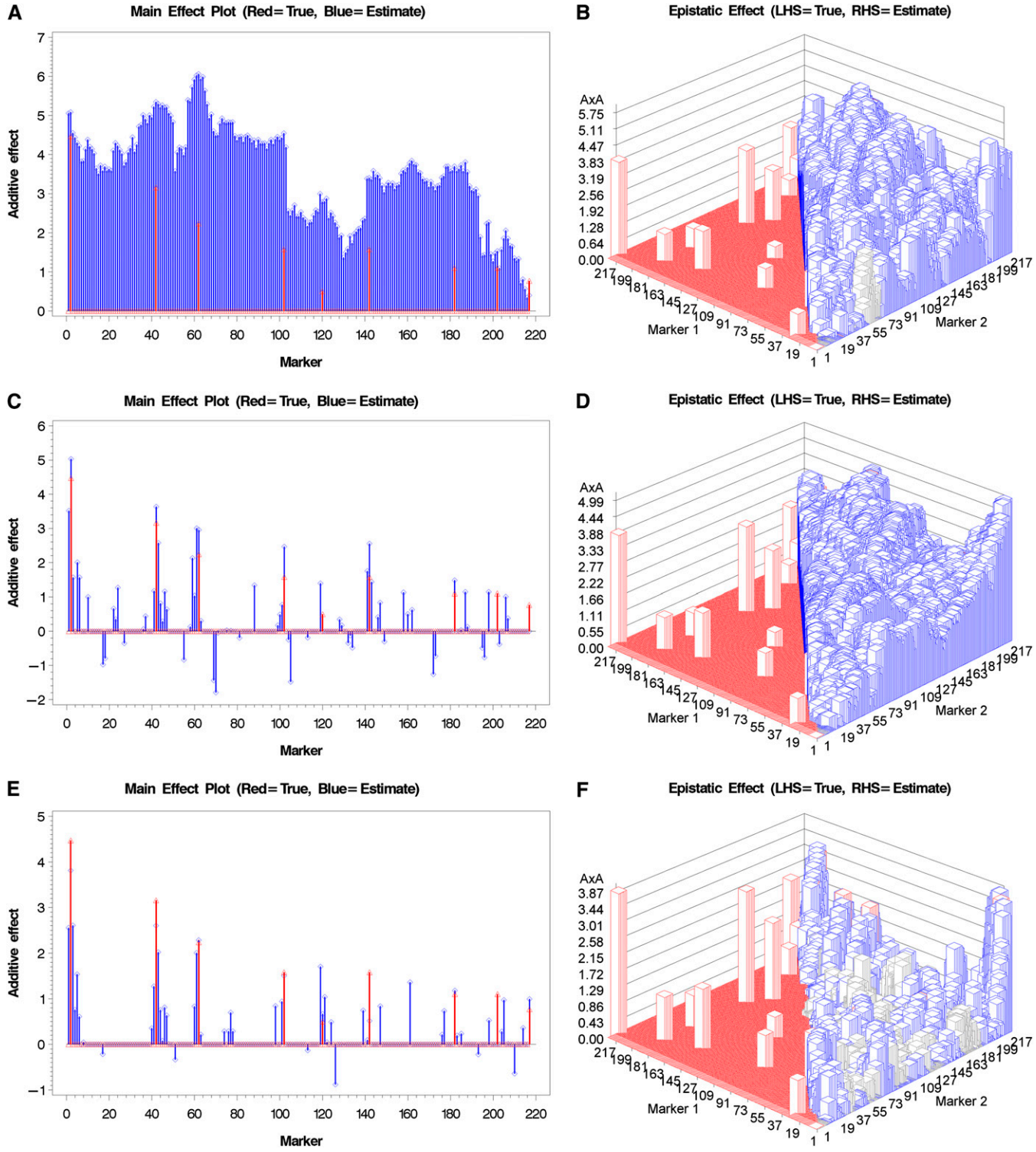

Figure 2 Comparison of models with and without fitting the polygenic covariance structures of the second simulation experiment. (A and B) The main and epistatic effects, respectively, under the model without the polygenic covariance structure. (C and D) The main and epistatic effects, respectively, under the model with only the additive covariance structure. (E and F) The main and epistatic effects, respectively, under the model with both the additive and the additive × additive covariance structures. True and estimated effects are in red and blue, respectively.

due to linkage disequilibrium. For example, about 50 bins covering one-third of chromosome 3 for trait GRAIN are significant for the unweighted method due to high linkage disequilibrium, but the weighted method has narrowed down the peak into a small region consisting of only 10 significant bins. The estimated effects and the test statistics

Table 2 Estimated variance components and proportions of variances over the phenotypic variance

Trait	σ_a^2	σ_d^2	σ_{aa}^2	σ_{dd}^2	σ_{ad}^2	σ_{da}^2	σ^2	σ_P^2
YIELD	0	0	7.0096	5.2713	18.2771	5.5485	1.963	38.0695
TILLER	0.4452	0	0.5851	0	0.1204	1.1305	0.3738	2.655
GRAIN	150.9134	0	66.8374	6.5794	110.1801	0	21.5148	356.0251
KGW	2.2656	0	0.312	0.2273	0	0.188	0.1131	3.106
	λ_a	λ_d	λ_{aa}	λ_{dd}	λ_{ad}	λ_{da}		
YIELD	0	0	3.5708	2.6853	9.3108	2.8265		
TILLER	1.1910	0	1.5652	0	0.3220	3.0243		
GRAIN	7.0144	0	3.1065	0.3058	5.1211	0		
KGW	20.0318	0	2.7586	2.0097	0	1.6622		
	h_a^2	h_d^2	h_{aa}^2	h_{dd}^2	h_{ad}^2	h_{da}^2	e^2	H
YIELD	0	0	0.1841	0.1385	0.4801	0.1457	0.0516	0.9484
TILLER	0.1677	0	0.2204	0	0.0454	0.4258	0.1408	0.8592
GRAIN	0.4239	0	0.1877	0.0185	0.3095	0	0.0604	0.9396
KGW	0.7294	0	0.1005	0.0732	0	0.0605	0.0364	0.9636

$\lambda_X = \sigma_X^2/\sigma^2$: variance ratio. $h_X^2 = \sigma_X^2/\sigma_P^2$: proportion of phenotypic variance contributed by each type of variance. H: the broad sense heritability defined by $H = \sigma_G^2/\sigma_P^2 = (\sigma_P^2 - \sigma^2)/\sigma_P^2$. e^2: the proportion of the environmental contribution defined by $e^2 = \sigma^2/\sigma_P^2 = 1 - H$.

along with the *P*-values of the 1619 bins for all the four traits are provided in File S4 and File S5 for the weighted and unweighted methods, respectively.

We now focus only on the weighted mixed-model analysis and report all bins that passed a preliminary screen to eliminate all bins with LRT less than 4.61, equivalent to LOD score less than one. All bins with *P*-values <0.05 were declared as significant at the genome-wide type I error of 0.05. Using the *P*-value <0.05 criterion, the number of significant bins are 25, 4, 32, and 23 for traits YIELD, TILLER, GRAIN, and KGW, respectively. These numbers are not the numbers of detected QTL because consecutive bins tended to show cosegregation. Therefore, the actual numbers of QTL are less than the number of significant bins. For example, among the four significant bins (bins 534, 1004, 1005, and 1262) for trait TILLER, only three QTL can be declared because bins 1004 and 1005 are counted as one due to their high linkage. File S6 lists all the bins passing the preselection, including the significant bins drawn from the permutation tests for the four traits. The estimated numbers of QTL are presented later using another round of screening to remove redundant bins caused by close linkage.

Epistatic effect analysis: Due to limitation of computing power, we used only the polygenic-effect-adjusted model (weighted method) to analyze the yield component traits. The LRT statistic was used to declare statistical significance for bin × bin interactions. The critical value for the LRT statistics was determined via a permutation analysis (Churchill and Doerge 1994) to adjust for multiple tests. For the epistatic effect model, we first eliminated all bin pairs with LRT <9.22, equivalent to LOD score <2. We then used the survived bin pairs to perform permutation analysis. The maximum LRT test statistics of 1000 permuted samples form an empirical null distribution. The 90, 95, 99, and 100 percentiles of the permutation-drawn null distributions for the four traits are given in Table 3 (bottom). These critical values also vary across traits. Trait KGW had the highest critical values and trait TILLER had the lowest critical values. The percentiles for the epistatic effect model are substantially higher than the percentiles for the main effect model. For example, the LRT critical value of KGW at 0.05 type I error is 5.47 for the main effect model but it is 15.77 for the epistatic effect model. The *P*-values were drawn for each pair of bins and the significant bin pairs along with their estimated effects are listed in File S7. The numbers of significant bin pairs are 1071, 205, 573, and 66 for traits YIELD, TILLER, GRAIN, and KGW, respectively. Again, these numbers are not the estimated numbers of epistatic effects due to tight linkage of consecutive bins. The estimated number of QTL interactions is addressed below using another cycle of screening.

For trait KGW in the epistatic effect model analysis, bin pair $(k, l) = (729, 1064)$ had an LRT of 18.23049 with a *P*-value of 0.016. We now provide a detailed analysis for this pair of bins. We used a model that includes the additive and dominance effects for both bins and all the four epistatic effects. The six polygenic covariance structures were also included in the model. This revised model now contains eight genetic effects and an intercept. Reanalysis of this bin pair produced eight estimated genetic effects, which are further converted into nine genotypic values using the transformation

$$\begin{bmatrix} 0.913736 \\ 0.857518 \\ 0.163836 \\ 0.004339 \\ -0.36711 \\ -0.34794 \\ -0.72723 \\ -1.01407 \\ -0.35034 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 & 0 & 0 & -1 & 0 \\ -1 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.538786 \\ 0.093253 \\ -0.1718 \\ -0.07827 \\ 0.281697 \\ 0.397007 \\ 0.082886 \\ -0.11703 \end{bmatrix}, \tag{26}$$

where the vector in the right-hand side stores the estimated genetic effects in the following order: $\{\hat{a}_k, \hat{a}_l, \hat{d}_k, \hat{d}_l, (\widehat{aa})_{kl}, (\widehat{ad})_{kl}, (\widehat{da})_{kl}, (\widehat{dd})_{kl}\}$. The vector in the left-hand side

Table 3 Critical values for the LRT test statistics drawn from 1000 permuted samples

		Trait			
	Percentile	YIELD	TILLER	GRAIN	KGW
Main (weighted)	100	9.7700	9.6460	14.5973	15.0849
	99	6.3322	5.1701	8.5974	8.3822
	95	3.6711	2.7129	5.5586	5.4752
	90	2.9402	1.8485	4.3127	4.2637
Main (unweighted)	100	17.0252	29.8881	19.8806	18.9779
	99	13.5334	12.2396	11.2226	14.8465
	95	10.0605	9.5286	8.3632	10.3167
	90	8.3845	8.1208	6.9026	8.4707
Epistatic	100	15.3727	21.8627	23.1942	26.2208
	99	14.4295	14.1096	18.0734	18.9523
	95	11.7136	11.0299	14.3440	15.7716
	90	10.4775	9.1725	12.6711	14.0365

gives the nine genotypic values. These genotypic values along with the marginal values are arranged in the following 4×4 matrix:

For bin $l = 1064$, the marginal effects are not significantly different among the three genotypes. However, the

	A_l	H_l	B_l	Marginal
A_k	0.913736	0.857518	0.163836	0.64503
H_k	0.004339	−0.36711	−0.34794	−0.2369
B_k	−0.72723	−1.01407	−0.35034	−0.69721
Marginal	0.063615	−0.17455	−0.17815	

differences are very significant under the A_k background. This detailed analysis shows the importance of the epistatic effects between the two bins for trait KGW. The two bins jointly contribute a genetic variance of 0.2991, which explains $h^2_{kl} = 0.2991/3.1092 \approx 10\%$ of the total trait variance. This is a significant contribution from only two loci for a quantitative trait. File S12 gives the SAS code of PROC HPMIXED for the analysis of the two bins. The genotypes of the two bins are given in File S3.

Numbers of bins and bin × bin interactions

This step provides the last screen of the bins and bin × bin interactions that have survived all previous selections. We noted that due to close linkages between consecutive bins, multiple bins are detected in a wide range of chromosome regions. In the original study of the IMF2 population by Zhou *et al.* (2012), the authors set up a subjective criterion to eliminate the extra significant bins. For example, they pooled several bins in the neighborhood of a few centimorgans of the genome into a single cluster. Each cluster was treated as one QTL. The authors realized that this approach is subjective and may eliminate some true effects. Here, we adopted a different and more objective strategy to eliminate those superfluous bins. Since the numbers of significant bins and bin × bin interactions are substantially smaller than the total number of bins and bin × bin pairs available in the data, we can fit all these significant bins and bin pairs simultaneously in a single model. For example, there are 23 significant bins (*a* and *d*) and 66 significant bin pairs (*aa*, *dd*, *ad*, and *da*) for trait KGW, and simultaneous analysis of all $23 \times 2 + 66 \times 4 = 310$ effects with a population size of 278 is possible, provided that a penalized regression is used to overcome the overfitting problem. For trait YIELD, the total number of significant bins and bin × bin pairs is 25 + 1071 = 1096, leading to $25 \times 2 + 1071 \times 4 = 4334$ effects, which is even harder to deal with than the KGW trait. Therefore, we used a penalized regression approach to perform parameter estimation, the Lasso method (Tibshirani 1996) implemented via the GlmNet/R package (Friedman *et al.* 2010). The method selects effects with nonzero effects (variable selection). After a minor modification of the Lasso method, the program also provides a test for each effect included in the model (Xu 2013).

The selected (nonzero) effects and their *P*-values obtained from the Lasso method are given in File S8 for all four traits. In this data set, if bin1 and bin2 have the same number, the effect represents a main effect; otherwise, the effect is an epistatic effect. The variable named "type" provides information about the type of effect. For example, type = "*aa*" indicates the additive × additive effect. Significant effects are labeled 1 in the last column named "significant." The numbers of significant effects sorted by effect type for the four traits are given in Table 4, where the top shows the number of (nonzero) effects selected by the Lasso method and the bottom gives the corresponding numbers that are significant at *P*-value <0.05. There are 39 significant effects apiece for YIELD and TILLER. Traits GRAIN and KGW have 24 and 15 significant effects, respectively. The number of significant epistatic effects is often higher than the number of main effects due to the substantially larger number of bin × bin pairs than the number of bins. The only exception occurs for trait KGW, which has 8 significant main effects and 7 significant epistatic effects. Table 5 shows the relative contributions of different types of effects to the total phenotypic variance. The significant effects collectively contributed 0.83, 0.69, 0.68, and 0.61 of the phenotypic variances for traits YIELD, TILLER, GRAIN, and KGW, respectively. These proportions are different from the broad sense heritability in the polygenic analysis (see Table 2). File S9 stores only the significant effects for the four traits (a subset of File S8). File S10 gives the independent variables converted from the genotypes for all the significant effects. For example, trait KGW has 15 significant effects, and thus the data set (sheet KWG in File S10) is stored in a 15×278 matrix plus the bin and effect type information.

Detailed examination of Table 5 shows that the relative contributions of different types of effects vary across traits. Epistatic effects play more important roles than main effects for traits YIELD, TILLER, and GRAIN while the additive effect is more important for trait KGW. This general conclusion is consistent with that of the polygenic analysis.

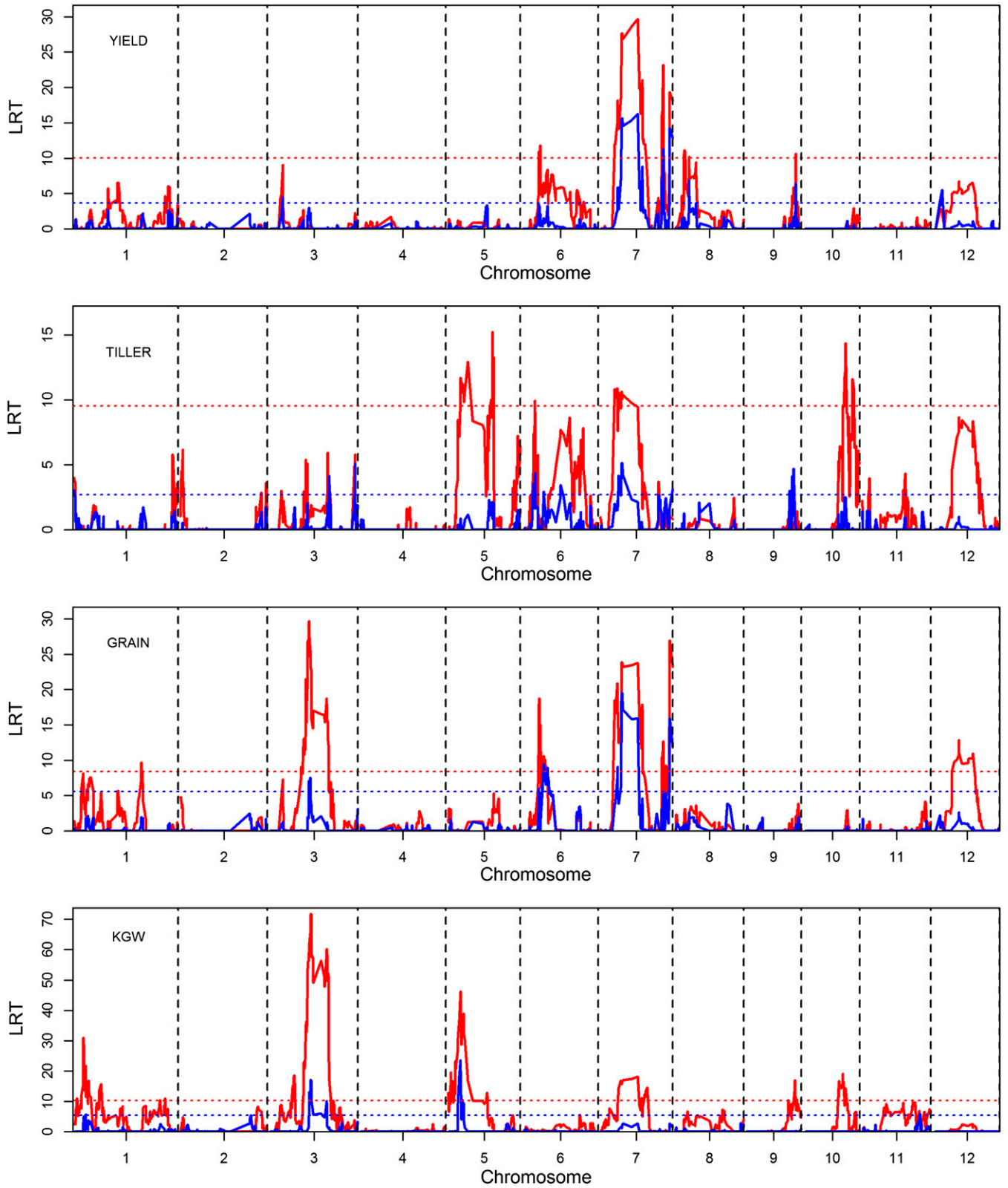

Figure 3 Comparison of the likelihood ratio test (LRT) statistics between the main effect model fitted with the polygenic covariance structure (weighted linear mixed model whose curve is blue) and the main effect model without the polygenic covariance structure (unweighted mixed model whose curve is red). Chromosomes are separated by the vertical dashed lines. The two horizontal dotted lines are permutation generated critical values of the LRT statistics under the 0.05 genome-wide type I error rate.

Table 4 Numbers of non-zero effects and numbers of significant effects ($P < 0.05$) obtained from the Lasso analysis

		Trait			
Catagory	Effect type	YIELD	TILLER	GRAIN	KGW
No. of nonzero effects	*a*	1	3	2	8
	d	1	2	0	7
	aa	40	22	27	3
	dd	35	23	25	4
	ad	48	29	32	10
	da	34	32	27	2
	Total	159	111	113	34
No. of significant effects	*a*	0	2	1	5
	d	1	2	0	3
	aa	6	10	6	3
	dd	13	5	7	1
	ad	11	13	6	3
	da	8	7	4	0
	Total	39	39	24	15

Discussion

Epistasis is often considered important to the variation of complex traits (Cockerham 1954). However, dissection of epistasis is difficult prior to the genome era because epistatic variances are often confounded with the additive and residual variances. To separate epistatic variances from additive and dominance variances, one needs pedigree information and the pedigrees must be very complicated to make the IBD matrices of epistasis sufficiently different from the kinship matrix for additive effects. In human populations, pedigree information is often incomplete and shallow (traced back to just one or two generations) and thus does not give us sufficient power to dissect epistasis. In the genome era, genome-wide marker information is available, which gives us an ample opportunity to model epistatic effects. Pedigree information is no longer crucial and a simple line cross experiment may be sufficient. In this study, we used binned genotypes of an IMF2 population to derive all the four types of kinship matrices to facilitate estimation of all types of epistatic variance components.

The result of the polygenic variance component analysis showed that adding epistatic variances can increase the model goodness of fit. We treated the additive model as the basis and progressively added more variance components to the model in the following sequence: a, $a + d$, $a + d + aa$, $a + d + aa + dd$, $a + d + aa + dd + ad$, and $a + d + aa + dd + ad + da$. Table 6 shows the increase of goodness of fit expressed as the squared correlation coefficient between the observed and predicted phenotypic values for the 278 IMF2 crosses. Depending on the characteristics of different traits, the goodness of fit has increased from 51.5 to 99.8% for YIELD and from 89.8 to 99.6% for KGW. The high goodness of fit implies an overfitting issue for the full model. Increasing sample size will partially correct the overfitting problem. Interestingly, the additive model alone already fits well for all the traits, especially for trait KGW. The reason may be explained for the high correlation between different estimated variance components, which may be caused by high similarities of various types of kinship matrices. The estimated variance components under different model sizes are listed in Table S1. Under the full model (all six genetic variance components are fit), the estimated dominance variances are zero for all the four traits. However, when the model size decreases, this estimated variance component can be nonzero, indicating that the actual dominance variance may not be zero but it is captured by other variance components due to the high correlation. The high correlation of estimated genetic variance components and high similarity between different kinship matrices suggest that benefit from fitting multiple polygenic covariance structures may not be significant for small sample sizes. Large sample sizes are necessary to separate the different genetic variances. How large of a sample size is sufficiently large? Theoretical investigation along with more simulation studies are required to have a clear answer for this question. The idea of incorporating multiple epistatic variance components in GWAS may be extended to incorporating multiple chromosome specific variance components in GWAS. How much improvement can we get by fitting multiple chromosome variances? Again, there is no clear answer at this moment. We may need a very large sample size to show the benefit.

A zero-estimated variance component for a particular type of effect does not mean that there is no genetic effect of this type for a particular bin or bin pair. The polygenic variance component of each type is the average of all bin- or bin pair-specific variances. A small number of bins or bin pairs may have large effects, but their effects are "diluted" under the polygenic model, which causes a zero estimation of the variance component. Our data analysis shows a lot of significant dominance effects but the estimated dominance variances obtained from the full model are zero for all traits.

In the original study of the IMF2 population, Zhou *et al.* (2012) used a different model to identify epistatic effects. They fit a two-locus model using a 3×3 factorial design (two loci and three genotypes per locus) and used a two-dimensional genome scan. They did not fit the polygenic variances. How much improvement have we achieved by fitting the polygenic variances? We do not have the answer at the moment because that would involve reanalysis of all the entire genome using the unweighted mixed model. Analysis of one trait, including permutation tests, took about 10 days of computing time because the large number of model effects fit to the model under the epistatic effect model, which is $4 \times 1619(1619 - 1)/2 = 5,239,084$, where the 4 in the expression represents the four types of epistatic effects. Advantages of fitting the polygenic variances (weighted mixed model) have been demonstrated under the main effect ($a + d$) model using both simulated data as well as the rice data. Our results are also not comparable with the results of Zhou *et al.* (2012) because we pooled data of the 2 years together and ignored the potential G $\times$ E interactions.

Table 5 Genetic variances contributed by each type of genetic effects for the yield component traits in rice

Trait		*a*	*d*	*aa*	*dd*	*ad*	*da*	Genotype[a]	Phenotype[b]
YIELD	No. effects	0	1	6	13	11	8	39	
	Proportion	0	0.017948	0.097888	0.144975	0.14334	0.096371	0.830223	
	Variance	0	0.612323	3.339706	4.946172	4.890387	3.287921	28.3251	34.117502
TILLER	No. effects	2	2	10	5	13	7	39	
	Proportion	0.02379	0.00998	0.15154	0.04523	0.19865	0.12923	0.6855	
	Variance	0.05834	0.02447	0.37166	0.11092	0.4872	0.31694	1.68123	2.4525721
GRAIN	No. effects	1	0	6	7	6	4	24	
	Proportion	0.0243	0	0.12329	0.09619	0.12649	0.09752	0.67615	
	Variance	6.99571	0	35.4907	27.6897	36.413	28.0731	194.643	287.87109
KGW	No. effects	5	3	3	1	3	0	15	
	Proportion	0.35732	0.03345	0.03845	0.04702	0.087	0	0.61046	
	Variance	1.11098	0.10402	0.11955	0.14618	0.27049	0	1.89803	3.1091762

[a] The number of effects in this column is the sum of all the six effect-specific numbers. The variance of this column is not the sum of the six effect-specific variances because it includes the covariance terms in addition to the effect-specific variances. The proportion of phenotypic variance contributed by genotype is also not the sum of the effect-specific proportions; rather it is the ratio of the genotypic variance to the phenotypic variance.

[b] The phenotypic variance in this column is directly calculated from the trait value, denoted by var(y), and thus it is different from the phenotypic variance shown in Table 1 that is denoted by σ_P^2.

Our study is different from the conventional GWAS (Yu *et al.* 2006) because we fit epistatic effects. However, the statistical model in our study shares similar properties with the improved eigenvalue decomposition methods (Lippert *et al.* 2011; Zhou and Stephens 2012) due to the fit of polygenic variances. We fit six variance components while the improved methods fit only one. Another difference is that we treated the bin effects and bin × bin interactions as random effects while Lippert *et al.* (2011) and Zhou and Stephens (2012) treated them as fixed effects. The advantages of fitting the random effects come from the shared variances and the tests of variance components. For instance, all the four epistatic effects share the same epistatic variance and are tested as one unit using a single LRT statistics. This approach may share some of the natures in rare variant detection (Wu *et al.* 2011), in which a group of rare variants are detected together using a single score test. Compared with the additive effect, an epistatic effect is a rare variant because the coefficient, *e.g.*, $Z_k \# Z_l$, has a smaller variance across individuals than the variance of Z_k or Z_l. Testing all four epistatic effects together will increase the statistical power for the same reason as the rare variant score test.

Finally, existing SAS programs were used here and the SAS codes are extremely simple (see File S11 and File S12). Theoretically, for each bin or bin pair, the six genetic variance components and the bin or bin pair specific variance can be estimated simultaneously using the HPMIXED procedure in SAS. This would be called the exact method (Zhou and Stephens 2012), as opposed to the approximate method (Lippert *et al.* 2011), where the polygenic variance is estimated only once before the genome is scanned. Our method is considered the approximate method because we estimated the polygenic variances prior to the two-dimensional scan. We did not adopt the exact method for two reasons: (1) there is a computational concern because we would have to estimate seven genetic variances per bin pair (high computational cost), and (2) there is no empirical evidence that the exact method outperforms the approximate method, although the former is theoretically superior over the latter. Using simulated data for the main effect model, we did not see much difference between the exact and approximate methods (data not shown). Zhou and Stephens (2012) proposed the exact method not because the exact method is better than the approximate method, but because it will avoid any "potential risk" of the approximate method and the exact method does not increase the computing time due to a special algorithm (eigen decomposition) for likelihood function evaluation. Once the polygenic variances are treated as constants in the approximate method, we provided a weight variable (see Equation 22 for the weight) to the HPMIXED procedure without modifying the code. In addition, the input variables were transformed using the eigenvector matrix prior to calling the SAS procedure. For readers without easy access to the SAS software packages, an R code is available from the author on request.

Acknowledgments

The author is grateful to two anonymous reviewers for their comments on early versions of the manuscript. The author also appreciates Dr. Qifa Zhang for sharing some additional data for the IMF2 population of rice beyond the data posted on the journal website. The project was supported by the

Table 6 Change of the goodness of fit[a] as the model size changes

Model	Model size	Yield	Tiller	Grain	KGW
a	1	0.5148	0.6052	0.7280	0.8980
a+d	2	0.7001	0.6052	0.7798	0.9350
a+d+aa	3	0.8769	0.8509	0.9227	0.9783
a+d+aa+dd	4	0.9172	0.8509	0.9342	0.9890
a+d+aa+dd+ad	5	0.9900	0.8500	0.9950	0.9890
a+d+aa+dd+ad+da	6	0.9980	0.9860	0.9950	0.9960

[a] The goodness of fit is defined as the squared correlation coefficient between observed and predicted phenotypic values.

U.S. Department of Agriculture National Institute of Food and Agriculture Grant 2007-02784.

Literature Cited

Bell, J. T., N. J. Timpson, N. W. Rayner, E. Zeggini, T. M. Frayling *et al.*, 2011 Genome-wide association scan allowing for epistasis in type 2 diabetes. Ann. Hum. Genet. 75: 10–19.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics 138: 963–971.

Cockerham, C. C., 1954 An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 39: 859–882.

Cockerham, C. C., and Z. B. Zeng, 1996 Design III with marker loci. Genetics 143: 1437–1456.

Friedman, J., T. Hastie, and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. J. Stat. Software 33: 1–22.

Garcia, A. A., S. Wang, A. E. Melchinger, and Z. B. Zeng, 2008 Quantitative trait loci mapping and the genetic basis of heterosis in maize and rice. Genetics 180: 1707–1724.

Henderson, C. R., 1975 Best linear unbiased estimation and prediction under a selection model. Biometrics 31: 423–447.

Hua, J. P., Y. Z. Xing, C. G. Xu, X. L. Sun, S. B. Yu *et al.*, 2002 Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance. Genetics 162: 1885–1895.

Hua, J., Y. Xing, W. Wu, C. Xu, X. Sun *et al.*, 2003 Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. Proc. Natl. Acad. Sci. USA 100: 2574–2579.

Jansen, R. C., 1994 Controlling the type I and type II errors in mapping quantitative trait loci. Genetics 138: 871–881.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman *et al.*, 2008 Efficient control of population structure in model organism association mapping. Genetics 178: 1709–1723.

Kao, C. H., and Z. B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics 160: 1243–1261.

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

Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu *et al.*, 2012 Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. PLoS Genet. 8: e1002685.

Patterson, H. D., and R. Thompson, 1971 Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545–554.

SAS Institute, 2009a *SAS/IML User's Guide, Version 9.3*. SAS Institute Inc., Cary, NC.

SAS Institute, 2009b *SAS/STAT: Users' Guide, Version 9.3*. SAS Institute Inc., Cary, NC.

Tibshirani, R., 1996 Regression shrinkage and selection via the Lasso. J. R. Stat. Soc., B 58: 267–288.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke *et al.*, 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89: 82–93.

Xie, W., Q. Feng, H. Yu, X. Huang, Q. Zhao *et al.*, 2010 Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc. Natl. Acad. Sci. USA 107: 10578–10583.

Xu, S., 2007 An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 63: 513–521.

Xu, S., 2013 Genetic mapping and genomic selection using recombination breakpoint data. Genetics 195:1103–1115.

Xu, S., and Z. Jia, 2007 Genomewide analysis of epistatic effects for quantitative traits in barley. Genetics 157: 1955–1963.

Yi, N., and S. Xu, 2002 Mapping quantitative trait loci with epistatic effects. Genet. Res. 79: 185–198.

Yu, H., W. Xie, J. Wang, Y. Xing, C. Xu *et al.*, 2011 Gains in QTL detection using an ultra-high-density SNP map based on population sequencing relative to traditional RFLP/SSR markers. PLoS One 6: e17595.

Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203–208.

Zeng, Z.-B., 1994 Precision mapping of quantitative trait loci. Genetics 136: 1457–1468.

Zhang, Y.-M., Y. Mao, C. Xie, H. Smith, L. Luo *et al.*, 2005 Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). Genetics 169: 2267–2275.

Zhang, Z., E. Ersoz, C. Q. Lai, R. J. Todhunter, H. K. Tiwari *et al.*, 2010 Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42: 355–360.

Zhou, G., Y. Chen, W. Yao, C. Zhang, W. Xie *et al.*, 2012 Genetic composition of yield heterosis in an elite rice hybrid. Proc. Natl. Acad. Sci. USA 109: 15847–15852.

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

Zou, F., J. A. L. Gelfond, D. C. Airey, L. Lu, K. F. Manly *et al.*, 2005 Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. Genetics 170: 1299–1311.

Communicating editor: N. Yi

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.157032/-/DC1

Mapping Quantitative Trait Loci by Controlling Polygenic Background Effects

Shizhong Xu

DOI: 10.1534/genetics.113.157032

Files S1-S10

All files are available for download at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.157032/-/DC1.

File S1: This dataset contains all six types of kinship matrices, each with a dimension 278×278. The type of kinship matrix is indicated by the first column (variable named parm) with 1, 2, 3, 4, 5, 6 indicating a, d, aa, dd, ad and da, respectively. The second column of this dataset gives the row numbers of each type of the kinship matrix. This special format is required by the PROC MIXED program. The data must contain $n+2$ variables with variable names of parm, row, col1, …, coln, where n is the number of lines. Do not mess up the variable names! The type=lin(6) option in the random statement of PROC MIXED means the program is looking for 6 kinship matrices.

File S2: This dataset stores the fixed-effect-adjusted phenotypic values of the four traits, yield, tiller, grain and kgw. The first two columns give the line id and line names. The last column gives the fold id that is used in the five-fold cross validation analysis by the Lasso method implemented via the GlmNet/R program.

File S3: This dataset gives the genotypes of two selected bins, bin1 = 729 and bin2 = 1064. The first column gives the names of the 278 IMF2 crosses. This dataset is used by the PROC HPMIXED program shown in Script S2.

File S4: This dataset list the estimated additive (a) and dominance (d) effects for all the 1619 bins obtained from the weighted mixed model (model incorporating the polygenic covariance structure). The standard errors corresponding to the additive and dominance effects are denoted by a_err and d_err, respectively. Effect specific tests are denoted by f_a and f_d (F test or Wald test). The LRT and p-value are also provided in the dataset. The last column gives the significant status with value 1 for $p < 0.05$ and 0 for $p > 0.05$.

File S5: This dataset list the estimated additive (a) and dominance (d) effects for all the 1619 bins obtained from the unweighted mixed model (model ignoring the polygenic covariance structure). The standard errors corresponding to the additive and dominance effects are denoted by a_err and d_err, respectively. Effect specific tests are denoted by f_a and f_d (F test or Wald test). The LRT and p-value are also provided in the dataset. The last column gives the significant status with value 1 for $p < 0.05$ and 0 for $p > 0.05$.

File S6: This dataset stores all bins selected with LRT > 4.61 (LOD >1) for the main effects (additive and dominance) from the main effect model analysis. The estimated additive and dominance effects along with the standard errors are given in columns headed by a, d, a_err and d_err respectively. The p-values drawn from permutation tests (1000 permuted samples) are also given. The last column indicates the significance status with 1 for $p < 0.05$ and 0 for $p > 0.05$. There are four sheets in the file, one for each trait.

File S7: This excel spread dataset stores all bin pairs selected with LRT > 9.22 (LOD > 2) for the epistatic effects (aa, dd, ad and da) from the epistatic model analysis. The estimated additive × additive, dominance × dominance, additive × dominance and dominance × additive effects along with the standard errors are given in columns headed by aa, dd, ad, da, aa_err, dd_err, ad_err and da_err, respectively. The p-values drawn from permutation tests are also given. The last column indicates the significance status with 1 for $p < 0.05$ and 0 for $p > 0.05$. There are four sheets in the file, one for each trait.

File S8: This dataset contains the estimated non-zero effects from the Lasso analysis for all bins and bin pairs that have survived the preliminary screen. Note that in the preliminary screen, the criteria were LOD > 1 for bins and LOD > 2 for bin pairs. When bin_1 equals bin_2, the effect represents a main effect (a for additive and d for dominance). When bin_1 does not equal bin_2, the effect represents an epistatic effect whose type is indicated by the column headed by type, i.e., aa, dd, ad and da, for the four types of epistatic effects, respectively. The estimated effect and the standard error of the estimate are given in columns headed by estimate and stderr. Wald test and LOD score along with the p-value are also given in the file. The last column shows the significance status from the Lasso analysis with 1 indicating $p < 0.05$ and 0 indicating $p > 0.05$. There are four sheets in the file, one for each trait.

File S9: This dataset contains the significant effects from the Lasso analysis for all bins and bin pairs that have survived the preliminary screen (LOD > 1 for bins and LOD > 2 for bin pairs). This dataset is a subset of Data S8 that excludes all effects with $p > 0.05$. All effects listed in this table are deemed to be significant. When bin_1 equals bin_2, the effect represents a main effect (a for additive and d for dominance). When bin_1 does not equal bin_2, the effect represents an epistatic effect whose type is indicated by the column headed by type, i.e., aa, dd, ad and da, for the four types of epistatic effects, respectively. The

estimated effect and the standard error of the estimate are given in columns headed by estimate and stderr. Wald test and LOD score along with the p-value are also given in the file. There are four sheets in the file, one for each trait.

File S10: This dataset contains the design matrix for all the significant effects listed in Data S9. The first four columns give the bin and bin pair information along with the effect type information. Variable named typeA shows the type of effect. A numerical version of the type is given in column headed by typeN, with 1, 2, 3, 4, 5 and 6 represent a, d, aa, dd, ad and da, respectively. The numerical type allows sorting by type in the natural order. The remaining $n = 278$ columns store the elements in the design matrix for multiple regression analysis using only the significant effects. This matrix needs to be transposed before fitting a multiple regression model. When bin_1 equals bin_2, the effect represents a main effect (a for additive and d for dominance). When bin_1 does not equal bin_2, the effect represents an epistatic effect whose type is indicated by the column headed by type, i.e., aa, dd, ad and da, for the four types of epistatic effects, respectively. There are four sheets in the file, one for each trait.

File S11

Script S1: SAS PROC MIXED for Polygenic Variance Components Analysis

This is the program code for PROC MIXED in SAS. The code also includes PROC IMPORT used to read the input data. The outputs are directly printed out in the window. In addition, estimated parameters and predicted genomic values are also written in SAS datasets. These SAS datasets can be exported later into physical files using PROC EXPORT (not provided). To make sure that PROC MIXED produces legal estimates of variance components, a lowerb= option is given in the parms statement. The lower bound option of 1e-5 means that each variance component is bounded at 1e-5, i.e., $\sigma_X^2 \geq 1e-5$. There are seven estimated variance components (six polygenic variances and one residual variance). All initial values of the variances are set to 1.0. Users can choose different initial values, depending on the properties of the data. The initial value of one (1) is the default initial value for the variance parameters in PROC MIXED. The trait shown in the code is KGW.

```
/*begin code*/

%let dir=C:\Users\SHXU\Programs;
filename kk "&dir\Data S1.csv" lrecl=20000;
filename phe "&dir\Data S2.csv";

proc import datafile=kk out=kk dbms=csv replace;
proc import datafile=phe out=phe dbms=csv replace;
run;

proc mixed data=phe method=reml;
class line;
model kgw=/solution;
random line/type=lin(6) ldata=kk solution;
parms (1) (1) (1) (1) (1) (1) (1)/lowerb=1e-5 1e-5 1e-5 1e-5 1e-5 1e-5 1e-5;
ods output SolutionR=blup SolutionF=fixed CovParms=covar;
run;

data pred;
   merge phe blup;
run;

proc corr data=pred;
   var kgw estimate;
run;

/*end code*/
```

Comments: The program takes two input files stored in a user defined folder (c:\users\shxu\programs in this example), one file for the kinship matrices (named Data S1.csv in this example) and one for the phenotypic values (named Data S2.csv in this example). The Data S2.csv file must contain a variable for the id number of lines (named line in this example) and a variable for the phenotypic values of the trait in question (named kgw in this example). The program will generate three SAS datasets. One SAS dataset is called blup, which gives the predicted polygenic value for each line, the second SAS dataset is named fixed, which gives the estimated fixed effects and the third SAS dataset named covar gives all the seven estimated variance components, including six polygenic variances and one residual variance. The two input files are provided in Supplemental Data S1 for the kinship matrix and Data S2 for fixed-effect adjusted phenotypic values of 278 lines for four quantitative traits.

File S12

Script S2: SAS PROC HPMIXED for Association Studies

This is the program code for PROC HPMIXED in SAS. The code also includes PROC IML for eigenvalue/eigenvector calculation and data manipulation. This program only shows the mixed model association study for two bins, bin1 = 729 and bin2 = 1064. The trait analyzed is KGW because this trait has shown that the two bins have strong interactions in the whole genome analysis. The model contains eight genetic effects (a1, a2, d1, d2, aa, ad, da, and dd). The program requires polygenic variance ratios (lambda values) calculated from the PROC MIXED program. The SAS dataset named lambda contains pre-calculated lambda values for all the four traits and thus the data matrix dimension is 6×4 (six observations and four variables). The program will print all outputs on the screen and also write various output tables into SAS datasets. The most important output is the estimated genetic effects in an output dataset called blup1. In the script, the PROC HPMIXED program is called twice, one for the polygenic model (null model) without fitting the two bins. This call produces a likelihood value ($-2L_0$) under the null model. The second call of this procedure produces a likelihood value ($-2L_1$) under the full model (fitting the two bins). A dataset called lrt is generated by taking the difference between the two likelihood values, lrt = ($-2L_0$) – ($-2L_1$) = $-2(L_0-L_1)$. This likelihood ratio test statistic is testing the significance of the two bins jointly.

```
/*begin code*/

%let dir=C:\Users\SHXU\Programs;
filename kk "&dir\Data S1.csv" lrecl=20000;
filename phe "&dir\Data S2.csv";
filename gen "&dir\Data S3.csv" lrecl=20000;

proc import datafile=kk out=kk dbms=csv replace;
proc import datafile=phe out=phe dbms=csv replace;
proc import datafile=gen out=gen dbms=csv replace;
run;

data lambda;
   input yield tiller grain kgw;
cards;
5.09424E-06   1.190743713   7.00747898    20.04424779
5.09424E-06   2.67523E-05   4.64533E-07   8.84956E-05
3.570809985   1.563937935   3.10387885    2.761061947
2.685277636   2.67523E-05   0.305281739   2.010619469
9.310901681   0.322632424   5.118688159   8.84956E-05
2.826439124   3.024879615   4.64533E-07   1.666371681
;

proc iml;
use lambda;
   read all var{kgw} into lambda;
close;
use phe;
   read all var{kgw} into y;
close;
p=nrow(lambda);
n=nrow(y);
kk=j(n,n,0);
do k=1 to p;
   range=((k-1)*n+1):(k*n);
   use kk;
      read point range into kk0;
   close;
   kk0=kk0[,3:(n+2)];
   kk=kk+kk0*lambda[k];
end;
call eigen(delta,uu,kk);
create delta from delta;
   append from delta;
```

```
close;
create uu from uu;
   append from uu;
close;
x=j(n,1,1);
xu=uu`*x;
yu=uu`*y;
w=1/(delta+1);
yxw=yu||xu||w;
varname={"y" "x" "w"};
create yxw from yxw[colname=varname];
   append from yxw;
close;
quit;

proc hpmixed data=yxw;
   model y = x/noint solution;
   weight w;
   ods output CovParms=parm0 FitStatistics=fit0 ParameterEstimates=fixed0;
   nloptions  maxiter=10000 gconv=1e-8;
run;

proc iml;
use gen;
  read all var{bin729 bin1064} into zz;
close;
k=1;
l=2;
zk=(zz[,k]='A')-(zz[,k]='B');
wk=(zz[,k]='H');
zl=(zz[,l]='A')-(zz[,l]='B');
wl=(zz[,l]='H');
create zz from zz;
append from zz;
close;

use yxw;
  read all into yxw;
close;
use uu;
   read all into uu;
close;
z=zk||zl||wk||wl||(zk#zl)||(zk#wl)||(wk#zl)||(wk#wl);
zu=uu`*z;
yxwz=yxw||zu;
varname={"y" "x" "w" "a1" "a2" "d1" "d2" "aa" "ad" "da" "dd"};
create yxwz from yxwz[colname=varname];
    append from yxwz;
close;
quit;

proc hpmixed data=yxwz;
   effect z=collection(a1 a2 d1 d2 aa ad da dd);
   model y=x/noint solution;
   weight w;
   random z / solution;
   ods output CovParms=parm1 FitStatistics=fit1
              ParameterEstimates=fixed1
              SolutionR=blup1 ConvergenceStatus=conv1;
   nloptions  maxiter=10000 gconv=1e-8;
run;

data lrt;
```

```
   merge fit0(rename=(value=l0)) fit1(rename=(value=l1));
   lrt=l0-l1;
   if _n_=1;
run;

/*end code*/
```

One of the outputs generated from PROC HPMIXED is the estimated genetic effects for the two bins (bin1 = 729 and bin2 = 1064).

Solution for Random Effects

Effect	z	Estimate	Std Err Pred	DF	t Value	Pr > |t|
z	a1	0.5388	0.1688	277	3.19	0.0016
z	a2	0.09325	0.1619	277	0.58	0.5651
z	d1	-0.1718	0.1315	277	-1.31	0.1924
z	d2	-0.07827	0.1302	277	-0.60	0.5482
z	aa	0.2817	0.1164	277	2.42	0.0162
z	ad	0.3970	0.1296	277	3.06	0.0024
z	da	0.08289	0.1327	277	0.62	0.5329
z	dd	-0.1170	0.1595	277	-0.73	0.4636

Comments: The program takes three input files stored in a user defined folder (c:\users\shxu\programs in this example), one file for the kinship matrices (named Data S1.csv in this example), one for the phenotypic values (named Data S2.csv in this example) and the third file for the bin genotypes (named Data S3 in this example). The Data S2.csv file must contain a variable for the id number of lines (named line in this example) and a variable for the phenotypic values of the trait in question (named kgw in this example). The program also requires a SAS dataset named lambda to store the six variance ratios obtained from the polygenic analysis via PROC MIXED described in Script S1. The program will generate several SAS datasets. One SAS dataset is called blup1, which gives the estimated genetic effects in the following order: a1, a2, d1, d2, aa, ad, da and dd. The three input files are provided in Supplemental Datas S1, S2 and S3, respectively.

Table S1 Estimated genetic and residual variance components under different models sorted by model size for the yield component traits in rice. This table gives the estimated genetic variance components under different models sorted by model size for the yield component traits obtained from the IMF2 population of rice. Six models were compared and the sizes of the six models range from one to six. The model labeled a is the additive model with only one additive variance component (model size is one). The model labeled a + d is the main effect model with additive and dominance variance components (model size is two). Other models are defined in the same way. The estimated genetic variance components are arranged in lower triangular of a square table for each trait because a particular type of variance component is only available from a model that includes this type of genetic variance.

Trait	Model	a	d	aa	dd	ad	da	Residual
Yield	a	14.4911						23.3308
	a+d	13.1129	9.0443					19.2088
	a+d+aa	11.2609	7.73516	7.38906				14.0566
	a+d+aa+dd	10.7365	0.00001	7.80710	7.23373			12.2930
	a+d+aa+dd+ad	4.32190	0.00001	7.27218	5.442035	16.7565		4.90684
	a+d+aa+dd+ad+da	0.00001	0.00001	7.00964	5.271305	18.2771	5.54845	1.96303
Tiller	a	1.38792						1.39981
	a+d	1.38792	0.00001					1.39981
	a+d+aa	1.05890	0.00001	0.61650				0.97550
	a+d+aa+dd	1.05895	0.00001	0.61645	0.00001			0.97549
	a+d+aa+dd+ad	1.05896	0.00001	0.61650	0.00001	0.00001		0.97545
	a+d+aa+dd+ad+da	0.44516	0.00001	0.58511	0.00001	0.12044	1.13048	0.37376
Grain	a	254.636						124.165
	a+d	245.599	24.3225					113.847
	a+d+aa	193.179	14.9761	69.4831				74.4618
	a+d+aa+dd	192.990	0.00001	68.5104	17.41830			69.7084
	a+d+aa+dd+ad	150.877	0.00001	66.8473	6.55935	110.184		21.5337
	a+d+aa+dd+ad+da	150.913	0.00001	66.8373	6.579376	110.180	0.00001	21.5147
KGW	a	2.82000						0.54720
	a+d	2.69618	0.26283					0.43694
	a+d+aa	2.38064	0.21624	0.32088				0.25818
	a+d+aa+dd	2.34714	0.00001	0.32697	0.245812			0.18642
	a+d+aa+dd+ad	2.34717	0.00001	0.32699	0.245813	0.00001		0.18641
	a+d+aa+dd+ad+da	2.26561	0.00001	0.31201	0.227324	0.00001	0.18803	0.11309