DOI: 10.1534/genetics.103.019406

Joint Mapping of Quantitative Trait Loci for Multiple Binary Characters

Chenwu Xu,* Zhikang Li† and Shizhong Xu*,[1]

**Department of Botany and Plant Sciences, University of California, Riverside, California 92521 and*
†Chinese Academy of Agricultural Sciences, Beijing 100081, China

Manuscript received June 26, 2003
Accepted for publication October 11, 2004

ABSTRACT

Joint mapping for multiple quantitative traits has shed new light on genetic mapping by pinpointing pleiotropic effects and close linkage. Joint mapping also can improve statistical power of QTL detection. However, such a joint mapping procedure has not been available for discrete traits. Most disease resistance traits are measured as one or more discrete characters. These discrete characters are often correlated. Joint mapping for multiple binary disease traits may provide an opportunity to explore pleiotropic effects and increase the statistical power of detecting disease loci. We develop a maximum-likelihood method for mapping multiple binary traits. We postulate a set of multivariate normal disease liabilities, each contributing to the phenotypic variance of one disease trait. The underlying liabilities are linked to the binary phenotypes through some underlying thresholds. The new method actually maps loci for the variation of multivariate normal liabilities. As a result, we are able to take advantage of existing methods of joint mapping for quantitative traits. We treat the multivariate liabilities as missing values so that an expectation-maximization (EM) algorithm can be applied here. We also extend the method to joint mapping for both discrete and continuous traits. Efficiency of the method is demonstrated using simulated data. We also apply the new method to a set of real data and detect several loci responsible for blast resistance in rice.

MULTIPLE traits are measured virtually in all line-crossing experiments of QTL mapping. Yet, almost all data collected for multiple traits are analyzed separately for different traits. Joint analysis for multiple traits has shed new light in QTL mapping by improving the statistical power of QTL detection and increasing the accuracy of QTL localization when different traits segregating in the mapping population are genetically related. Joint analysis for multiple traits is defined as a method that includes all traits simultaneously in a single model, rather than analyzing one trait at a time and reporting the results in a format that appears to be multiple-trait analysis. In addition to the increased power and resolution of QTL detection, joint mapping can provide insights into fundamental genetic mechanisms underlying trait relationships such as pleiotropy *vs.* close linkage and genotype-by-environment ($G \times E$) interaction, which would otherwise be difficult to address if traits are analyzed separately.

Substantial work has been done in joint mapping for multiple quantitative traits (Jiang and Zeng 1995; Korol *et al.* 1995, 2001; Mangin *et al.* 1998; Henshall and Goddard 1999; Williams *et al.* 1999; Knott and Haley 2000; Hackett *et al.* 2001). In general, there are two ways to perform a joint mapping. One way is the true multivariate analysis in which a multivariate normal distribution is assumed for the multiple traits, and thus a multivariate Gaussian model is applied to construct the likelihood function. Parameter estimation is conducted via either the expectation-maximization (EM) algorithm (Dempster *et al.* 1977) or the multiple-traits least-squares method (Knott and Haley 2000). One problem with these multivariate analyses is that if the number of traits is large, there will be too many hypotheses to test and interpretation of the results will become cumbersome. The other way of multiple-trait analysis is to utilize a dimension reduction technique, *e.g.*, the principal component analysis, to transform the data into fewer variables, *i.e.*, "super traits," that explain the majority of the total variation of the entire set of traits. Analyzing the super traits requires little additional work (Korol *et al.* 1995, 2001; Mangin *et al.* 1998) compared to that for the single-trait genetic mapping statistics. However, as pointed out by Hackett *et al.* (2001), inferences based on the super traits might result in detection of spurious QTL. Furthermore, parameters of the super traits are often difficult to interpret biologically. Nevertheless, joint mapping provides a good opportunity to answer more questions about the genetic architecture of complex trait sets and deserves continued efforts from investigators in the QTL mapping community.

In contrast to that for multiple quantitative traits, relatively little work has been done in joint mapping for multiple discrete traits (Lange and Whittaker 2001).

[1] *Corresponding author:* Department of Botany and Plant Sciences, University of California, Riverside, CA 92521.
E-mail: xu@genetics.ucr.edu

In fact, multiple discrete traits, especially multiple binary disease traits, are frequently collected in plants and laboratory animals. Most disease resistance traits are measured as one or more dichotomous characters. For example, in experiments mapping disease resistance loci, multiple pathogen races or strains are commonly used to determine the number of race-specific resistance loci involved. In such cases, infection by each strain is measured as a binary trait and the overall infection spectrum is a vector of several binary measurements. In practice, scientists may be less interested in identifying resistance loci to individual strains, but more interested in loci with a wide spectrum of resistance, which, in principle, can be better addressed with the joint-mapping strategy. Unfortunately, there has been no report on such a joint-mapping analysis for multiple binary traits. Recently, Lange and Whittaker (2001) applied the generalized estimating equations (GEE; Liang and Zeger 1986) method to mapping multiple discrete trait loci. Results of GEE are hardly compared with those of single-trait mapping because there is no univariate version of the GEE.

Furthermore, it is not uncommon that investigators may collect both continuous and discrete traits in a single mapping experiment. For example, disease resistance characters may be measured in a QTL mapping experiment for yield traits, or vice versa. Combining the disease resistance traits (discrete) with the yield traits (continuous) may allow investigators to answer some important questions such as possible fitness penalty of resistance loci. Even if the associated quantitative traits are not directly responsible for the disease status but linked to the disease loci, joint analysis will still provide additional power in identifying the disease loci (Williams *et al.* 1999; Huang and Jiang 2003). So far, joint analysis of mixed types of traits has been attempted only in pedigree analysis of human populations under the random model methodology.

The goal of this study is to develop a formal multivariate version of the maximum-likelihood methodology for joint mapping of QTL underlying multiple binary traits and mixed types of traits in line-crossing experiments under the fixed-model framework. We analyzed both simulated data and data collected from field experiments.

METHODS

Joint mapping for multiple binary traits: *Statistical model:* Suppose that we have a sample of n individuals from an F_2 population derived from the cross of two inbred lines with observation on m binary traits. Assume that we also genotype a number of codominant molecular markers with known map positions for the species in question. Let w_{jk} denote the phenotype of the kth binary trait on the jth individual and $w_{jk} = 1$ if individual j is affected and $w_{jk} = 0$ if j is unaffected. Further define y_{jk} as the underlying latent variable, *i.e.*, the liability, for the kth binary trait on the jth individual. The relationship between the liability and the discrete phenotype is described by the following threshold model,

$$y_{jk} > 0 \Leftrightarrow w_{jk} = 1 \quad \text{and} \quad y_{jk} \leq 0 \Leftrightarrow w_{jk} = 0, \qquad (1)$$

where the threshold 0 is chosen in an arbitrary fashion. Each of the m liabilities is a continuous variable, similar to the phenotypic value of a quantitative trait. The difference between a liability and a quantitative trait is that the former is not observable but inferred from the discrete phenotype. The liabilities are described by the following linear models,

$$\begin{aligned} y_{j1} &= b_{01}x_{0j} + b_{11}x_{1j} + b_{21}x_{2j} + e_{j1},\\ y_{j2} &= b_{02}x_{0j} + b_{12}x_{1j} + b_{22}x_{2j} + e_{j2},\\ &\;\vdots\\ y_{jm} &= b_{0m}x_{0j} + b_{1m}x_{1j} + b_{2m}x_{2j} + e_{jm}, \end{aligned} \qquad (2)$$

where b_{0k} is the mean effect (intercept) for trait k in the scale of liability; b_{1k} and b_{2k} are, respectively, the additive and dominance effects of the putative QTL; x_{0j}, x_{1j}, and x_{2j} are the incidence variables for the mean effect and the additive and dominance effects, respectively; and e_{jk} is the residual error for trait k of individual j. We assume that the residual errors are independent among individuals but correlated among traits within individuals. In matrix notation, model (2) can be written as

$$\mathbf{y}_j = \mathbf{x}_j\mathbf{B} + \mathbf{e}_j, \qquad (3)$$

where $\mathbf{y}_j = [y_{j1} \ldots y_{jm}]$ is a $1 \times m$ vector for the liabilities; $\mathbf{x}_j = [x_{0j}\ x_{1j}\ x_{2j}]$, which is defined as $\mathbf{x}_j = \mathbf{h}_1 = [1\ 1\ 0]$ if individual j takes genotype QQ, $\mathbf{x}_j = \mathbf{h}_2 = [1\ 0\ 1]$ if j has a genotype Qq, and $\mathbf{x}_j = \mathbf{h}_3 = [1\ -1\ 0]$ if j is qq at the putative QTL position; $\mathbf{B}$ is a $3 \times m$ matrix defined as

$$\mathbf{B} = \begin{bmatrix} b_{01} & b_{02} & \ldots & b_{0m} \\ b_{11} & b_{12} & \ldots & b_{1m} \\ b_{21} & b_{22} & \ldots & b_{2m} \end{bmatrix}; \qquad (4)$$

and $\mathbf{e}_j$ is a $1 \times m$ vector for the residual errors, which has a covariance matrix

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \ldots & \sigma_m^2 \end{bmatrix}. \qquad (5)$$

Note that the variances are estimable from the latent variables but not from the observed data. Therefore, some restrictions are required when the binary data are taken into account (McCulloch 1994). The probability that individual j is affected by all the m binary diseases is

$$\Pr(w_{j1} = 1, \ldots, w_{jm} = 1|\mathbf{x}_j, \mathbf{B}, \mathbf{V}) = \int_0^\infty \ldots \int_0^\infty \phi_m(\mathbf{y}_j; \mathbf{x}_j\mathbf{B}, \mathbf{V})\, dy_{j1} \ldots dy_{jm}, \tag{6}$$

where

$$\phi_m(\mathbf{y}_j; \mathbf{x}_j\mathbf{B}, \mathbf{V}) = (2\pi)^{-m/2}|\mathbf{V}|^{-1/2}\exp\{-\tfrac{1}{2}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})\mathbf{V}^{-1}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})^{\mathrm{T}}\} \tag{7}$$

is the multivariate normal probability density. The probabilities for other joint binary phenotypes are derived similarly. For m dichotomous traits, there will be 2^m possible joint binary phenotypes. With the threshold model, mapping binary traits has been formulated as a problem of mapping quantitative trait loci. Therefore, methods of QTL mapping for multiple quantitative traits (Jiang and Zeng 1995) may be adopted here in multiple binary trait mapping.

Model (3) is a general multiple linear model (GMLM) with missing values in $\mathbf{x}_j$ because the QTL genotypes are not observable. The next step of the GMLM analysis with missing values is to infer the probabilities of QTL genotypes conditional on the marker information, denoted by $p_{jq} = \Pr(\mathbf{x}_j = \mathbf{h}_q|I_{\mathrm{M}})$ for $q = 1, 2, 3$, where I_{M} denotes marker information. We can compute p_{jq} using Table 1 of Jiang and Zeng (1995) if double recombinants are not considered or using Table 1 of Luo and Kearsey (1992) if no crossover interference is assumed. However, in general, we need to adopt the multipoint method (Jiang and Zeng 1997) to handle missing or dominance markers.

Maximum-likelihood estimation: Let us denote the parameters by a vector $\boldsymbol{\theta} = \{\mathbf{B}, \mathbf{V}\}$. The likelihood function for the jth individual conditional on $\mathbf{x}_j$ is

$$\Pr(\mathbf{w}_j|\mathbf{x}_j, \boldsymbol{\theta}) = \int_{g_1(w_{j1})}^{g_2(w_{j1})} \ldots \int_{g_1(w_{jm})}^{g_2(w_{jm})} \phi_m(\mathbf{y}_j; \mathbf{x}_j\mathbf{B}, \mathbf{V})\, dy_{j1} \ldots dy_{jm}$$
$$= \Phi_m(\mathbf{w}_j; \mathbf{x}_j\mathbf{B}, \mathbf{V}), \tag{8}$$

where $\mathbf{w}_j = [w_{j1}, \ldots, w_{jm}]$, $g_1(w_{jk}) = (w_{jk} - 1) \times \infty$, and $g_2(w_{jk}) = w_{jk} \times \infty$, for $k = 1, \ldots, m$. Note that $g_1(w_{jk}) = (w_{jk} - 1) \times \infty = -\infty$ and $g_2(w_{jk}) = w_{jk} \times \infty = 0$ if $w_{jk} = 0$, whereas $g_1(w_{jk}) = (w_{jk} - 1) \times \infty = 0$ and $g_2(w_{jk}) = w_{jk} \times \infty = \infty$ if $w_{jk} = 1$. The m-dimensional integral $\Phi_m(\mathbf{w}_j; \mathbf{x}_j\mathbf{B}, \mathbf{V})$ may be calculated with some special algorithms or by executing an intrinsic function from some software packages. A two-dimensional integral can be found in SAS (SAS Institute 1999). The probability $\Pr(\mathbf{w}_j|\mathbf{x}_j, \boldsymbol{\theta})$ is also called the penetrance of the QTL with genotype $\mathbf{x}_j$.

Since $\mathbf{x}_j$ is missing and only p_{jq} is calculated, the actual likelihood function for the jth individual is

$$\Pr(\mathbf{w}_j|\boldsymbol{\theta}) = \sum_{q=1}^{3} p_{jq}\Phi_m(\mathbf{w}_j; \mathbf{h}_q\mathbf{B}, \mathbf{V}). \tag{9}$$

The overall log-likelihood for the entire mapping population is

$$L(\boldsymbol{\theta}) = \sum_{j=1}^{n}\ln[\Pr(\mathbf{w}_j|\boldsymbol{\theta})]. \tag{10}$$

Solving the above log-likelihood function is tedious. We now introduce a simple EM algorithm to find the solution, which takes advantage of the simplicity of the original linear model with both $\mathbf{Y} = \{\mathbf{y}_j\}_{j=1}^{n}$ and $\mathbf{X} = \{\mathbf{x}_j\}_{j=1}^{n}$ being treated as missing values.

Instead of directly maximizing the log likelihood given in Equation 10, the EM algorithm deals with the complete-data log-likelihood function,

$$L(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Y}) = \text{const} - \frac{n}{2}\ln|\mathbf{V}|$$
$$- \frac{1}{2}\sum_{j=1}^{n}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})\mathbf{V}^{-1}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})^{\mathrm{T}}. \tag{11}$$

Because $\mathbf{X}$ and $\mathbf{Y}$ are treated as missing values, the EM algorithm starts with maximizing the expectation of the complete-data log-likelihood,

$$L(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E[L(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Y})] = \text{const} - \frac{n}{2}\ln|\mathbf{V}|$$
$$- \frac{1}{2}\sum_{j=1}^{n}E[(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})\mathbf{V}^{-1}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})^{\mathrm{T}}], \tag{12}$$

where the expectation is taken with respect to $\mathbf{X}$ and $\mathbf{Y}$, conditional on the data $\mathbf{W} = \{\mathbf{w}_j\}_{j=1}^{n}$ and the parameters at the current values $\boldsymbol{\theta}^{(t)}$. For brevity, we use $E[\xi(\mathbf{X}, \mathbf{Y})]$ to denote the conditional expectation throughout the entire text, but the full notation for the expectation should be

$$E[\xi(\mathbf{X}, \mathbf{Y})] = E_X\{E_{Y|X}[\xi(\mathbf{X}, \mathbf{Y})|\boldsymbol{\theta}^{(t)}, \mathbf{W}]\},$$

where $\xi(\mathbf{X}, \mathbf{Y})$ represents the term whose expectation is required in the EM algorithm. Note that Equation 12 differs from Equation 10 in two aspects: (i) Equation 10 takes the expectation before the log transformation whereas Equation 12 takes the expectation after the log transformation; and (ii) the expectations are taken using different probability distributions for the two equations. In Equation 12, the expectation is taken using the probability distribution conditional on the current parameter values and the phenotypic values. Maximizing Equation 12 with respect to the parameters, we get

$$\hat{\mathbf{B}} = \left[\sum_{j=1}^{n}E(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j)\right]^{-1}\left[\sum_{j=1}^{n}E(\mathbf{x}_j^{\mathrm{T}}\mathbf{y}_j)\right] \tag{13}$$

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{j=1}^{n}E\big[(\mathbf{y}_j - \mathbf{x}_j\hat{\mathbf{B}})^{\mathrm{T}}(\mathbf{y}_j - \mathbf{x}_j\hat{\mathbf{B}})\big]. \tag{14}$$

The MLE of $\mathbf{B}$ and $\mathbf{V}$ in the complete-data situation (both $\mathbf{X}$ and $\mathbf{Y}$ are observed) can be found in Anderson (1984, Sect. 8.2). Giri (1996, pp. 92–98) also provided the derivation in the simple case where $\mathbf{x}_j\mathbf{B} = \boldsymbol{\mu}$ is a $1 \times m$ vector of means. The results shown in Equations 13 and 14 are extensions of the results in the complete-data situation by adding the symbols of conditional expectation. appendix c gives a step-by-step derivation of Equations 13 and 14.

In binary data analysis under the liability model, the usual restriction is to let all the variances (diagonal elements of matrix **V**) equal unity (McCulloch 1994). If we had maximized the complete-data log likelihood function with such a restriction, the maximization step would be extremely complicated because there is no easy way to find a closed-form expression for $\hat{\mathbf{V}}$. Trying to find an explicit form for $\hat{\mathbf{V}}$, one would lose the advantage of the EM algorithm. Fortunately, Gueorguieva and Agresti (2001) showed that maximizing Equation 12 with **V** unrestricted, the EM algorithm converges to unique estimates of the fully identifiable ratios, $\mathbf{BS}^{-1}$, where

$$\mathbf{S} = \sqrt{\mathrm{diag}(\mathbf{V})} = \mathrm{diag}(\sigma_1 \ldots \sigma_m). \tag{15}$$

Therefore, Gueorguieva and Agresti (2001) maximized the log-likelihood function with **V** unrestricted and then standardized the model effects by taking $\mathbf{B}^* = \mathbf{BS}^{-1}$. The standardized covariance matrix became $\mathbf{R} = \mathbf{S}^{-1}\mathbf{VS}^{-1}$. At each EM iteration, Gueorguieva and Agresti (2001) estimated **B** and **V** and immediately replaced these two parameters by their standardized versions, $\mathbf{B}^*$ and **R**, before entering the next EM iteration to make sure that the EM converges. In our EM algorithm, we have already defined **B** as a standardized vector of genetic effects. We simply need to standardize **V** during each iteration of the EM algorithm to ensure the convergence of the iterations. The estimated correlation matrix $\hat{\mathbf{R}}$ is indeed the MLE of **R** based on the invariance property of the MLE (DeGroot 1986) because $\hat{\mathbf{V}}$ is the MLE of **V** and **R** is a function of **V**. Equations 13 and 14 represent the maximization step of the EM algorithm. We now investigate the expectation step of the EM algorithm. Recall that the probability of $\mathbf{x}_j$ conditional on marker information is denoted by p_{jq}. This probability may be called the prior probability. After incorporating the phenotypic value and the parameters, we obtain the posterior probability, denoted by

$$p^*_{jq} = \Pr(\mathbf{x}_j = \mathbf{h}_q | I_{\mathrm{M}}, \mathbf{w}_j) = \frac{p_{jq}\Phi_m(\mathbf{w}_j;\, \mathbf{h}_q\mathbf{B},\, \mathbf{R})}{\sum_{h=1}^{3} p_{jh}\Phi_m(\mathbf{w}_j;\, \mathbf{h}_h\mathbf{B},\, \mathbf{R})}. \tag{16}$$

Note that the **V** matrix has been replaced by the **R** matrix to reflect the standardization. In fact, the unrestricted covariance matrix **V** is used only once when we try to estimate it. Once **V** is estimated, it is immediately standardized into the form of **R**, which is then used in all steps of the EM iterations. The expectations are actually obtained using the posterior probabilities rather than the prior probabilities. Therefore, the conditional expectations given the data **W** are

$$\sum_{j=1}^{n} E(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j) = \sum_{j=1}^{n}\left[\sum_{q=1}^{3} p^*_{jq}\mathbf{h}_q^{\mathrm{T}}\mathbf{h}_q\right]$$

$$\sum_{j=1}^{n} E(\mathbf{x}_j^{\mathrm{T}}\mathbf{y}_j) = \sum_{j=1}^{n}\left[\sum_{q=1}^{3} p^*_{jq}\mathbf{h}_q^{\mathrm{T}}E(\mathbf{y}_j|\mathbf{w}_j,\, \mathbf{h}_q,\, \boldsymbol{\theta})\right], \tag{17}$$

where $E(\mathbf{y}_j|\mathbf{w}_j, \mathbf{h}_q, \boldsymbol{\theta})$ is the expectation of a truncated multivariate normal distribution. The residual error covariance matrix in Equation 14 becomes

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{j=1}^{n}\left[\sum_{q=1}^{3} p^*_{jq}(\mathbf{U}_{jq} + \boldsymbol{\mu}_{jq}^{\mathrm{T}}\boldsymbol{\mu}_{jq} - \hat{\mathbf{B}}^{\mathrm{T}}\mathbf{h}_q^{\mathrm{T}}\boldsymbol{\mu}_{jq} - \boldsymbol{\mu}_{jq}^{\mathrm{T}}\mathbf{h}_q\hat{\mathbf{B}} + \hat{\mathbf{B}}^{\mathrm{T}}\mathbf{h}_q^{\mathrm{T}}\mathbf{h}_q\hat{\mathbf{B}})\right], \tag{18}$$

where $\boldsymbol{\mu}_{jq} = E(\mathbf{y}_j|\mathbf{w}_j, \mathbf{h}_q, \boldsymbol{\theta})$ is a short notation for the conditional expectation and $\mathbf{U}_{jq} = \mathrm{Var}(\mathbf{y}_j|\mathbf{w}_j, \mathbf{h}_q, \boldsymbol{\theta})$ denotes the conditional covariance matrix of a truncated multivariate normal distribution. Neither $\boldsymbol{\mu}_{jq}$ nor $\mathbf{U}_{jq}$ has an explicit expression except in the special case when $m = 2$ and 3. These expectation and covariance matrices are calculated using the moment-generating function (Tallis 1963) or the Gibbs sampler (Chan and Kuk 1997). The moment-generating function approach needs multidimensional integrals and cannot be implemented easily in practice when m is large (see appendix a for the special case when $m = 2$). The Gibbs sampling approach requires Monte Carlo simulation, which is suitable for large m. Details of the Monte Carlo method are given in appendix b.

The EM algorithm may be summarized in the following steps:

Step 1. Choose the initial values for $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(0)} = \{\mathbf{B}^{(0)}, \mathbf{R}^{(0)}\}$.

Step 2. Calculate the posterior probabilities of the QTL genotype given the current values of all unknowns using Equation 16.

Step 3. Calculate the expectations using Equation 17, a process in the E-step.

Step 4. Calculate $\boldsymbol{\mu}_{jq} = E(\mathbf{y}_j|\mathbf{w}_j, \mathbf{h}_q, \boldsymbol{\theta})$ and $\mathbf{U}_{jq} = \mathrm{Var}(\mathbf{y}_j|\mathbf{w}_j, \mathbf{h}_q, \boldsymbol{\theta})$ using the moment-generating function or the Gibbs sampler (see appendix a and appendix b), another process of the E-step.

Step 5. Update parameter **B** using Equation 13, update parameter **V** using Equation 14, and convert **V** into **R**.

Step 6. Replace the initial parameters by the updated values and repeat steps 2–5 until convergence.

Likelihood-ratio test: Define the log-likelihood function evaluated at the maximum-likelihood estimate (MLE) of parameters as

$$L(\hat{\boldsymbol{\theta}}) = \sum_{j=1}^{n}\ln[\Pr(\mathbf{w}_j|\hat{\boldsymbol{\theta}}], \tag{19}$$

where $\Pr(\mathbf{w}_j|\hat{\boldsymbol{\theta}}) = \Sigma_{q=1}^{3} p_{jq}\Phi_m(\mathbf{w}_j;\, \mathbf{h}_q\hat{\mathbf{B}},\, \hat{\mathbf{R}})$ and $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$. This is also called the likelihood value under the full model. We need the likelihood values under various restricted models to test different hypotheses.

The overall null hypothesis is "no effect of QTL at the locus of interest," denoted by H_0: $b_{1k} = b_{2k} = 0$ for $k = 1, \ldots, m$ or H_0: $\mathbf{LB} = \mathbf{0}$, where

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

If we solve the MLE of the parameters under the restriction of $\mathbf{LB} = \mathbf{0}$ and evaluate the likelihood function evaluated at the solutions with this restriction, we have

$$L(\tilde{\boldsymbol{\theta}}|\mathbf{LB} = \mathbf{0}), \tag{20}$$

where $\tilde{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ under the restricted model. The likelihood-ratio test statistic is

$$\Lambda = -2[L(\tilde{\boldsymbol{\theta}}|\mathbf{LB} = \mathbf{0}) - L(\hat{\boldsymbol{\theta}})]. \tag{21}$$

Under the null hypothesis, this test statistic will approximately follow a chi-square distribution with $2(m - 1)$ d.f.

Various other test statistics may be defined by choosing different $\mathbf{L}$ matrices, as given by Jiang and Zeng (1995). For example, to test the additive effects for all traits, the $\mathbf{L}$ matrix is defined as $\mathbf{L} = [0\ 1\ 0]$. The degrees of freedom for this test are $m - 1$. To test the dominance effects for all traits, we use $\mathbf{L} = [0\ 0\ 1]$ with $m - 1$ d.f. for the test statistic.

To test trait-specific effects, we need to introduce another matrix, $\mathbf{T}$, which is used to postmultiply matrix $\mathbf{B}$. For example, to test QTL effects (both additive and dominance) for the kth trait, the null hypothesis is H_0: $\mathbf{LBT} = \mathbf{0}$, where

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and $\mathbf{T}$ is an $m \times 1$ vector with the kth element being one and all the remaining elements being zeros. The test statistic will be

$$\Lambda = -2[L(\tilde{\boldsymbol{\theta}}|\mathbf{LBT} = \mathbf{0}) - L(\hat{\boldsymbol{\theta}})] \tag{22}$$

with 2 d.f. In general, the $\mathbf{L}$ matrix controls the type of effects (population mean, additive, and dominance) being tested and the $\mathbf{T}$ matrix controls the traits (from 1 to m) being tested.

The position of the QTL is another parameter of interest. However, in the one-dimensional genome scan, the position is first treated as fixed and then the entire genome is tested for every putative position with a 1- or 2-cM increment. The likelihood-ratio test statistic forms a test statistic profile. The position corresponding to the highest peak is declared as the estimated QTL position if the peak surpasses a given critical value (Churchill and Doerge 1994; Diggle *et al.* 1996; Piepho 2001).

Joint mapping for mixed types of traits: We now describe a statistical model and likelihood analysis for joint mapping of loci that affect one binary trait and multiple quantitative traits. Let w_{j1} be the phenotype of the binary trait for the jth individual and defined as $w_{j1} = 1$ if individual j is affected and $w_{j1} = 0$ if it is not affected. Further define y_{jk} as the value of the kth observed quantitative trait, for $k = 2, \ldots, m$, on the jth individual. We also define y_{j1} as the liability for the binary trait,

$$y_{j1} > 0 \Leftrightarrow w_{j1} = 1 \quad \text{and} \quad y_{j1} \leq 0 \Leftrightarrow w_{j1} = 0.$$

The liability of the binary trait and the phenotypic values of the quantitative traits are arranged in a vector called

$$\mathbf{y}_j = [\underbrace{y_{j1}}_{\text{liability}}\ \underbrace{y_{j2}\ \ \ldots\ \ y_{jm}}_{\text{quantitative traits}}] = [y_{j1}\quad \mathbf{y}_{j\bar{1}}],$$

where $\mathbf{y}_{j\bar{1}} = [y_{j2} \ldots y_{jm}]$ is a special notation for a subvector of $\mathbf{y}_j$ that excludes the first element. The $\mathbf{y}_j$ vector is described by the same model as given in Equation 3. Let us further partition matrix $\mathbf{B}$ into $\mathbf{B} = [\mathbf{b}_1\ \mathbf{B}_{\bar{1}}]$, where $\mathbf{b}_1^{\mathrm{T}} = [b_{01}\ b_{11}\ b_{21}]$ is the first column of matrix $\mathbf{B}$ and $\mathbf{B}_{\bar{1}} = [\mathbf{b}_2 \ldots \mathbf{b}_m]$ is the remaining columns of $\mathbf{B}$. The residual errors are joint normal with mean zero and a covariance matrix $\mathbf{V}$, which can be partitioned into

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{bmatrix} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{1\bar{1}} \\ \mathbf{V}_{\bar{1}1} & \mathbf{V}_{\bar{1}\bar{1}} \end{bmatrix}, \tag{23}$$

where $\mathbf{V}_{11} = \sigma_1^2$, $\mathbf{V}_{1\bar{1}} = [\sigma_{12} \ldots \sigma_{1m}]$, $\mathbf{V}_{1\bar{1}} = \mathbf{V}_{\bar{1}1}^{\mathrm{T}}$, and

$$\mathbf{V}_{\bar{1}\bar{1}} = \begin{bmatrix} \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \ddots & \vdots \\ \sigma_{2m} & \cdots & \sigma_m^2 \end{bmatrix}$$

is the lower right block of matrix $\mathbf{V}$. The variance of the liability for the disease trait, however, is restricted to unity. Therefore, the standardized form of the $\mathbf{V}$ matrix is $\mathbf{R} = \mathbf{S}^{-1}\mathbf{V}\mathbf{S}^{-1}$, which is a function of the unrestricted covariance matrix $\mathbf{V}$, where $\mathbf{S} = \mathrm{diag}(\sigma_1\ 1 \ldots 1)$. Similar partitioning given in Equation 23 also applies to matrix $\mathbf{R}$. The joint distribution of the phenotype for individual j is

$$\Pr(w_{j1}, \mathbf{y}_{j\bar{1}}|\mathbf{x}_j, \boldsymbol{\theta}) = \phi_{m-1}(\mathbf{y}_{j\bar{1}}; \mathbf{x}_j\mathbf{B}_{\bar{1}}, \mathbf{R}_{\bar{1}\bar{1}})\Phi(w_{j1}; \mathbf{y}_{j\bar{1}}, \mathbf{x}_j\mathbf{b}_1, \mathbf{R}_{11}), \tag{24}$$

where

$$\phi_{m-1}(\mathbf{y}_{j\bar{1}}; \mathbf{x}_j\mathbf{B}_{\bar{1}}, \mathbf{R}_{\bar{1}\bar{1}}) = (2\pi)^{-(m-1)/2}|\mathbf{R}_{\bar{1}\bar{1}}|^{-1/2} \times \exp\{-\tfrac{1}{2}(\mathbf{y}_{j\bar{1}} - \mathbf{x}_j\mathbf{B}_{\bar{1}})\mathbf{R}_{\bar{1}\bar{1}}^{-1})(\mathbf{y}_{j\bar{1}} - \mathbf{x}_j\mathbf{B}_{\bar{1}})^{\mathrm{T}}\} \tag{25}$$

and

$$\Phi(w_{j1}; \mathbf{y}_{j\bar{1}}, \mathbf{x}_j\mathbf{b}_1, \mathbf{R}_{11}) = \int_{g_1(w_{j1})}^{g_2(w_{j1})} \phi(y_{j1}; \mathbf{y}_{j\bar{1}}, \mathbf{x}_j\mathbf{b}_1, \mathbf{R}_{11})\,dy_{j1}. \tag{26}$$

The probability density $\phi(y_{j1}; \mathbf{y}_{j\bar{1}}, \mathbf{x}_j\mathbf{b}_1, \mathbf{R}_{11})$ within the integral is a conditional density of y_{j1} given $\mathbf{y}_{j\bar{1}}$. It is a univariate normal with mean

$$E(y_{j1}|\mathbf{y}_{j\bar{1}}, \mathbf{x}_j, \boldsymbol{\theta}) = \mathbf{x}_j\mathbf{b}_1 + \mathbf{R}_{1\bar{1}}\mathbf{R}_{\bar{1}\bar{1}}^{-1}(\mathbf{y}_{j\bar{1}} - \mathbf{x}_j\mathbf{B}_{\bar{1}})^{\mathrm{T}} \tag{27}$$

and variance

$$\mathrm{Var}(y_{j1}|\mathbf{y}_{j\bar{1}}, \mathbf{x}_j, \boldsymbol{\theta}) = \mathbf{R}_{11} - \mathbf{R}_{1\bar{1}}\mathbf{R}_{\bar{1}\bar{1}}^{-1}\mathbf{R}_{\bar{1}1}. \tag{28}$$

Therefore, $\Phi(w_{j1}; \mathbf{y}_{j(-1)}, \mathbf{x}_j\mathbf{b}_1, \mathbf{R}_{11})$ is a truncated univariate normal probability.

The likelihood function for the jth individual is

$$\Pr(w_{j1}, \mathbf{y}_{j\bar{1}}|\boldsymbol{\theta}) = \sum_{q=1}^{3} p_{jq} \Pr(w_{j1}, \mathbf{y}_{j\bar{1}}|\mathbf{h}_q\mathbf{B}, \mathbf{R}). \tag{29}$$

The overall log-likelihood of the entire mapping population is

$$L(\boldsymbol{\theta}) = \sum_{j=1}^{n} \ln[\Pr(w_{j1}, \mathbf{y}_{j\bar{1}}|\boldsymbol{\theta})]. \tag{30}$$

Again, we adopt the EM algorithm to find the MLE of parameters. The maximization step is accomplished through Equations 13 and 14. The expectation step requires calculation of the posterior probabilities of QTL genotypes and then uses these probabilities to find various expectations. The posterior probability of a QTL genotype is

$$\begin{aligned} p^*_{jq} &= \Pr(\mathbf{x}_j = \mathbf{h}_q|I_{\mathrm{M}}, w_{j1}, \mathbf{y}_{j\bar{1}}) \\ &= \frac{p_{jq} \Pr(w_{j1}, \mathbf{y}_{j\bar{1}}|\mathbf{h}_q\mathbf{B}, \mathbf{R})}{\sum_{h=1}^{3} p_{jh} \Pr(w_{j1}, \mathbf{y}_{j\bar{1}}|\mathbf{h}_h\mathbf{B}, \mathbf{R})}, \end{aligned} \tag{31}$$

from which we get the expectations

$$\begin{aligned} \sum_{j=1}^{n} E(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j) &= \sum_{j=1}^{n} \left[\sum_{q=1}^{3} p^*_{jq} \mathbf{h}_q^{\mathrm{T}} \mathbf{h}_q \right] \\ \sum_{j=1}^{n} E(\mathbf{x}_j^{\mathrm{T}}\mathbf{y}_j) &= \sum_{j=1}^{n} \left[\sum_{q=1}^{3} p^*_{jq} \mathbf{h}_q^{\mathrm{T}} E(\mathbf{y}_j|w_{j1}, \mathbf{h}_q, \boldsymbol{\theta}) \right], \end{aligned} \tag{32}$$

where

$$E(\mathbf{y}_j|w_{j1}, \mathbf{h}_q, \boldsymbol{\theta}) = [E(y_{j1}|w_{j1}, \mathbf{h}_q, \boldsymbol{\theta}, \mathbf{y}_{j\bar{1}}) \quad \mathbf{y}_{j\bar{1}}] \tag{33}$$

is a 1 × m vector, which has been partitioned into a scalar $E(y_{j1}|w_{j1}, \mathbf{h}_q, \boldsymbol{\theta}, \mathbf{y}_{j\bar{1}})$ and a vector $\mathbf{y}_{j\bar{1}}$. The expectation is taken only for the liability of the binary trait. The remaining traits already take the observed values and thus no expectations are taken. The expectation for the liability, $E(y_{j1}|w_{j1}, \mathbf{h}_q, \boldsymbol{\theta}, \mathbf{y}_{j\bar{1}})$, is obtained from the truncated normal distribution with mean given by Equation B5 of APPENDIX B and variance given by Equation B6 of APPENDIX B. The residual error covariance matrix is given by Equation 18. However, the conditional expectation is replaced by

$$\boldsymbol{\mu}_{jq} = E(\mathbf{y}_j|w_{j1}, \mathbf{h}_q, \boldsymbol{\theta}) \tag{34}$$

and the conditional variance by

$$\mathbf{U}_{jq} = \mathrm{Var}(y_{j1}|w_{j1}, \mathbf{h}_q, \boldsymbol{\theta}, \mathbf{y}_{j\bar{1}}) \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{35}$$

where $\mathrm{Var}(y_{j1}|w_{j1}, \mathbf{h}_q, \boldsymbol{\theta}, \mathbf{y}_{j\bar{1}})$ is the variance of a truncated normal distribution. Both $E(y_{j1}|w_{j1}, \mathbf{h}_q, \boldsymbol{\theta}, \mathbf{y}_{j\bar{1}})$ and $\mathrm{Var}(y_{j1}|w_{j1}, \mathbf{h}_q, \boldsymbol{\theta}, \mathbf{y}_{j\bar{1}})$ can be found from the truncated normal distribution (COHEN 1991) and no Gibbs sampler is required.

TABLE 1

QTL parameters used in the simulation experiments

Heritability (%)	b_{01}	b_{11}	b_{21}	b_{02}	b_{12}	b_{22}
5	0	0.2	0.36	0	0.2	0.36
10	0	0.4	0.35	0	−0.4	0.35
15	0	0.5	0.45	0	0.5	−0.45

RESULTS

Simulation studies: To further evaluate the properties of the proposed method, we conducted two simulation experiments. For the sake of simplicity, we designed one experiment to evaluate the performance of joint mapping for two binary traits and another experiment for the mixture of one binary and one quantitative trait. In each experiment, one chromosome with 11 evenly distributed markers was simulated for an F_2 population. We simulated a single QTL located at 35 cM of the chromosome and the QTL effects of both traits under three different levels of heritability (proportion of variance in liability explained by the QTL). The effects of the QTL used in the simulation experiments are given in Table 1. The correlation coefficient between the residuals of the liabilities for the two traits was chosen at 0.25. Under these settings, both traits had the same heritability. The three levels of the QTL effects led to three different levels of the heritability: 5, 10, and 15%. The genetic correlation between the two traits was expected to be 1.0, −0.446, and 0.423, respectively, for the different chosen levels of the heritability. The sample size of the simulated F_2 population was $n = 200$. Each simulation experiment was replicated 100 times.

The first simulation experiment was to evaluate the efficiency of the joint binary trait mapping. We first simulated the liabilities of the two traits and then artificially truncated the continuous liabilities into two binary phenotypes using a threshold of zero. In the second simulation experiment, we truncated only the liability of the first trait to generate a binary phenotype but left the second trait intact so that we had one binary trait and one continuous trait.

Each data sample was analyzed using both the joint-mapping and single-trait-mapping statistics. For the single-trait analysis, we used LANDER and BOTSTEIN's (1989) method for the continuous trait and the method of XU *et al.* (2003) for the binary trait. Since we considered only two traits in the joint mapping, explicit formulas were used in each of the EM iteration steps (see APPENDIX A for the explicit formulas). The critical values for the chromosomewise type I error rate of 5% were determined by the approximate method of PIEPHO (2001). In real data analysis, one should use the permutation test (CHURCHILL and DOERGE 1994) to obtain the most appropriate critical values for significance tests. For the joint analysis, the empirical power was determined by the proportion of the replicated samples (out of 100)

TABLE 2

Comparison of joint mapping with single trait mapping from the first simulation experiment (both traits are binary)

Heritability (%)	Method	Power (%)	Position (cM)	b_{01}	b_{11}	b_{21}	b_{02}	b_{12}	b_{22}	ρ
5	J-12	42	33.02	−0.05	0.26	0.48	0.05	0.25	0.46	0.24
			(10.10)	(0.14)	(0.13)	(0.21)	(0.16)	(0.14)	(0.27)	(0.13)
	S-1	26	33.46	−0.10	0.28	0.59				
			(13.36)	(0.15)	(0.18)	(0.20)				
	S-2	29	36.13				−0.07	0.28	0.56	
			(12.68)				(0.15)	(0.16)	(0.17)	
10	J-12	90	34.02	0.01	0.41	0.36	−0.00	−0.42	0.38	0.24
			(7.47)	(0.14)	(0.13)	(0.24)	(0.13)	(0.14)	(0.21)	(0.10)
	S-1	78	35.42	−0.03	0.48	0.42				
			(12.44)	(0.14)	(0.13)	(0.27)				
	S-2	69	36.52				−0.01	−0.46	0.39	
			(11.94)				(0.15)	(0.11)	(0.24)	
15	J-12	100	35.08	−0.00	0.50	0.45	−0.00	0.49	−0.47	0.24
			(5.08)	(0.14)	(0.16)	(0.21)	(0.15)	(0.15)	(0.22)	(0.10)
	S-1	92	35.57	0.10	0.56	0.47				
			(9.23)	(0.17)	(0.14)	(0.26)				
	S-2	91	36.10				0.01	0.54	−0.48	
			(10.12)				(0.15)	(0.15)	(0.24)	

Entries for the QTL effect and location estimates are the average of 100 replicated simulations with the standard deviations among the 100 replicates given in parentheses. J-12, joint mapping; S-1, separate mapping for trait 1; S-2, separate mapping for trait 2.

whose highest test statistic values along the chromosome were greater than Piepho's (2001) critical value. The peak where the highest test statistic occurred was usually close to the true QTL position. However, a significant QTL was declared even if the peak was not exactly at the true position. For the separate analyses of individual traits, the statistical power was determined for the analysis of each trait as in the joint analysis. The critical value was recalculated for each trait in each replicate.

Tables 2 and 3 show the observed powers of QTL detection, the mean, and standard deviations (SD) of the estimated QTL locations and effects obtained from 100 replicated simulations. We compared the power of the joint analysis with that of a single-trait analysis for each trait separately. Joint analysis has a substantially higher power than either single-trait analysis. We understand that this may not be a fair comparison because joint analysis uses two traits while the single-trait analysis uses only one trait. However, this has been the standard way for comparison of joint mapping with separate mapping (Jiang and Zeng 1997). One may want to redefine the power for the separate analyses as the ability to detect at least one QTL effect (either additive or dominance) in at least one trait. Under this definition of the power, results of the two separate analyses may be combined so that the power is recalculated in the combined result. The combined power analysis requires either redefining the critical values by taking into account the multiple tests, which is difficult because the two separate analyses may be highly correlated, or simply using the sum of the powers of separate analyses (with an appropriate adjustment) as the combined power, which is $p_1 + p_2 - p_{12}$, where p_1 and p_2 are the powers for traits 1 and 2, respectively, and p_{12} is the proportion of the replicated simulations in which both traits are significant. For example, among the 100 replicates, if a significant QTL effect is detected in 50 samples for the first trait and a significant QTL effect is detected in 80 samples for the second trait, then $p_1 = 0.5$ and $p_2 = 0.8$. If QTL effects for both traits are detected in 40 samples, then $p_{12} = 0.4$. The combined power for the separate analyses will be $0.5 + 0.8 - 0.4 = 0.9$. Using this approach to calculating the power, the combined power of the separate trait analyses was almost identical to that of the joint analysis (data not shown). Therefore, power increase in joint mapping as opposed to separate mapping depends on how one defines the power in the separate analyses. From the traditional definition of statistical power for single-trait analysis (Jiang and Zeng 1997), joint analysis has higher power than single-trait analysis, *i.e.*, joint power greater than p_1 and joint power greater than p_2. But the joint analysis has an equivalent power to the combined power for separate analyses if the combined power is defined as $p_1 + p_2 - p_{12}$, *i.e.*, joint power $\approx p_1 + p_2 - p_{12}$.

The QTL effects and their standard deviations estimated from the joint mapping are comparable to those obtained from separate analyses. No obvious advantage of the joint mapping has been demonstrated from the simulation studies with respect to the estimates of QTL effects. The real advantage of the joint mapping over separate analyses has been demonstrated by the increased precision of the QTL position estimates in all situations examined (see Tables 2 and 3).

Overall, the parameter estimates are fairly close to the

TABLE 3

Comparison of joint mapping with single trait mapping from the second simulation experiment (one binary trait and one quantitative trait)

Heritability (%)	Method	Power (%)	Position (cM)	b_{01}	b_{11}	b_{21}	b_{02}	b_{12}	b_{22}	ρ	σ_2^2
5	J-12	65	35.94	−0.01	0.21	0.41	−0.03	0.24	0.42	0.28	0.98
			(9.40)	(0.15)	(0.17)	(0.26)	(0.11)	(0.10)	(0.18)	(0.10)	(0.10)
	S-1	20	37.20	−0.14	0.35	0.59					
			(13.68)	(0.15)	(0.14)	(0.24)					
	S-2	58	37.33				−0.04	0.28	0.43		0.97
			(12.11)				(0.10)	(0.11)	(0.14)		(0.11)
10	J-12	97	34.98	−0.01	0.39	0.36	−0.01	−0.40	0.37	0.27	0.98
			(6.50)	(0.15)	(0.13)	(0.20)	(0.09)	(0.11)	(0.15)	(0.09)	(0.10)
	S-1	44	36.27	−0.01	0.51	0.38					
			(12.46)	(0.19)	(0.11)	(0.29)					
	S-2	95	35.19				−0.02	−0.43	0.39		0.99
			(8.32)				(0.11)	(0.10)	(0.15)		(0.10)
15	J-12	100	35.26	0.02	0.49	0.42	0.02	0.52	−0.43	0.23	0.96
			(4.54)	(0.15)	(0.14)	(0.22)	(0.15)	(0.09)	(0.17)	(0.09)	(0.10)
	S-1	80	35.34	−0.10	0.56	0.48					
			(9.37)	(0.14)	(0.12)	(0.23)					
	S-2	99	35.15				0.01	0.49	−0.48		0.97
			(6.49)				(0.10)	(0.11)	(0.15)		(0.10)

Entries for the QTL effect and location estimates are the average of 100 replicated simulations with the standard deviations among the 100 replicates given in parentheses. J-12, joint mapping; S-1, separate mapping for trait 1; S-2, separate mapping for trait 2.

true parametric values. The general trend follows our expectation: high heritability tends to produce more accurate estimates than low heritability. If we compare the joint mapping of two binary traits with that of one binary and one continuous trait, we will note the power difference between the two experiments. Experiment 2 shows higher powers than experiment 1. This observation also follows our expectation because binary data are not as informative as continuously distributed data.

Mapping rice blast resistance loci: Developing blast resistance cultivars is one of the major objectives in rice (*Oryza sativa L.*) breeding in both tropical and temperate countries. The causal organism of the rice blast, *Pyricularia grisea*, is known for its high genetic variability, allowing it to overcome the resistance of the host plant. A framework linkage map was developed using 284 F_{10} recombinant inbred lines (RILs) from a "Lemont" × "Teqing" rice cultivar cross. A subset of 245 RILs innoculated with two rice blast races, IB54 and IG1, was used to map loci responsible for the hypersensitive reaction. Details of the experimental design, the measurements of phenotypes, and genotypes can be found in the original article by Tabien *et al.* (2000). The phenotypes were evaluated using a completely randomized design with three replicates. In other words, each line was evaluated three times for its reaction to each of the phathogen infections. The original scores of the plant response were measured from grade 0 to grade 5. The average score of the three replicates for each line was recorded as the raw data observation. The binary phenotype was defined as $w = 0$ if the average score was within the range 0–3 and $w = 1$ if the average score was 4–5. We were provided only with the binary data, not the original scores. The breeders were more interested in the genetic study of the qualitative dichotomous trait than in the genetic study of the numerical scores. This explains why we were approached by the breeders to analyze their data using the new methods.

Since the mapping population was a RIL population, a slight modification of our method for F_2 was required. We replaced the probability transition matrix of F_2 by that of F_{10} in calculating the conditional probability of QTL genotype (Jiang and Zeng 1997). There was still a 4% residual heterozygosity in the RIL lines (due to F_{10} instead of F_∞), which is sufficiently high to allow the dominance effects to be estimated. We treated the plant responses to blast pathogen races IB54 and IG1 as two separate binary traits. Therefore, joint mapping for both traits and separate mappings for individual traits were conducted for comparisons. The critical values of test statistics used to declare QTL were calculated using the method of Piepho (2001).

Table 4 shows the results of joint mapping and separate analyses. The joint mapping may have a greater power than separate analyses, as demonstrated by more detected QTL and higher test statistic values. A total of five resistance loci were identified by the joint mapping (qtl1–qtl5), but only four of them were detected with separate analyses (qtl4 was missed). Of these detected QTL, three of them (qtl1, qtl3, and qtl4) corresponded

TABLE 4

QTL mapping result for rice blast resistance in the "Lemont" × "Teqing" crossing experiment

Method	QTL	Chr.	Position (cM)	Λ	b_{01}	b_{11}	b_{21}	b_{02}	b_{12}	b_{22}	ρ
J-12	qtl1	2	7.8	33.92	−0.43	0.20	−0.78	−0.72	0.85	−0.44	0.73
	qtl2	3	137.4	25.22	−0.95	0.52	0.71	−1.00	0.37	0.56	0.65
	qtl3	11	3.0	37.55	−1.08	−0.84	0.13	−1.22	−0.71	0.42	0.63
	qtl4	12	54.2	25.01	−0.84	0.44	−1.81	−0.90	0.28	−1.75	0.69
	qtl5	12	87.2	46.57	−1.02	0.75	0.51	−0.98	0.30	0.46	0.66
S-1	qtl2	3	137.4	23.35	0.61	0.96	1.06				
	qtl3	11	2.0	29.75	−0.87	0.35	1.17				
	qtl5	12	84.5	48.57	0.78	0.35	1.08				
S-2	qtl1	2	7.8	28.95				1.11	0.10	1.06	

Λ is the likelihood-ratio test statistic and ρ is the residual correlation. J-12, joint mapping for IB54 and IG1; S-1, separate mapping for IB54; S-2, separate mapping for IG1. The critical values of the test statistic (Piepho 2001) used to declare QTL were 20.57 for the J-12 analysis and 18.15 for each of the separate analyses (S-1 and S-2). Chr., chromosome.

to *Pi-tq5*, *Pi-lm2*, and *Pi-tq6* detected previously on the basis of chi-square tests of individual marker-trait associations (Tabien *et al.* 2000). Two additional loci (qtl2 and qtl5) were detected on chromosomes 3 and 12, and they were not reported in the previous study (Tabien *et al.* 2000). For each of the two loci, the allele carried by the Lemont parent was responsible for the resistance. None of the genetic parameters, *e.g.*, the QTL effects and positions, were estimable in the previous chi-square tests conducted by the original authors (Tabien *et al.* 2000). The most striking result from the joint mapping was that all five resistance loci showed fairly consistent effects against both *P. grisea* races, while different resistance loci were detected separately by the single-trait analyses.

It is worth mentioning that results of joint mapping and separate mapping do not seem to be consistent in the real data analyses. This inconsistency, however, did not occur in the simulation studies. The reason for this is that we have taken a one-dimensional genome-scan approach, which uses a single-QTL model. In the simulation studies, we indeed simulated a single QTL. Therefore, the model adequately described the data. In the real data analysis, however, we used the single-QTL model to fit data controlled by apparently multiple QTL. The remaining QTL not fitted in the model may have caused all the inconsistencies observed between the joint and the separate analyses. In addition, the background QTL also have caused the observed high residual correlation. These problems can be solved by fitting a multiple-QTL model (see the discussion in a later section).

Table 5 shows the probabilities of the four possible phenotypic combinations under different genotypes of the identified QTL. For a single disease trait, penetrance is defined as the probability that a specific QTL genotype shows the affected phenotype. Penetrance has not been defined for multiple disease traits. Therefore, we listed the probabilities of all the four possible phenotype combinations for all genotypes of each detected QTL in Table 5. The penetrances of any particular genotypes for each QTL may be calculated from this table. For example, if we define the penetrance of a genotype as the probability that a plant with this genotype is affected by either of the two pathogens, the penetrance should be calculated using $1 - \Pr(\text{IB54} = \text{R and IG1} = \text{R})$. On the other hand, if we define the penetrance as the probability that the plant is affected by both pathogens, then we should use $\Pr(\text{IB54} = \text{S and IG1} = \text{S})$. The marginal penetrance for one pathogen, say pathogen IB54, should be defined as

$$\begin{aligned}\Pr(\text{IB54} = \text{S}) &= \Pr(\text{IBS} = \text{S and IG1} = \text{S})\\ &\quad + \Pr(\text{IBS} = \text{S and IG1} = \text{R}).\end{aligned}$$

Taking the first genotype of the first QTL, for example, we may be able to find penetrances defined in all possible ways, as shown below,

$$\begin{aligned}\Pr(\text{affected by either pathogen}|QQ) &= 1 - \Pr(\text{IB54} = \text{R and IG1} = \text{R})\\ &= 1 - 0.3863 = 0.6137,\\ \Pr(\text{affected by both pathogens}|QQ) &= \Pr(\text{IB54} = \text{S and IG1} = \text{S})\\ &= 0.3457,\\ \Pr(\text{affected by IB54}|QQ) &= \Pr(\text{IB54} = \text{S})\\ &= \Pr(\text{IB54} = \text{S and IG1} = \text{S})\\ &\quad + \Pr(\text{IB54} = \text{S and IG1} = \text{R})\\ &= 0.3457 + 0.0637 = 0.4094,\end{aligned}$$

and

$$\begin{aligned}\Pr(\text{affected by IG1}|QQ) &= \Pr(\text{IG1} = \text{S})\\ &= \Pr(\text{IB54} = \text{S and IG1} = \text{S})\\ &\quad + \Pr(\text{IB54} = \text{R and IG1} = \text{S})\\ &= 0.3457 + 0.2043 = 0.55.\end{aligned}$$

Interested rice geneticists and breeders may want to find out all kinds of penetrances of interest from Table 5. This table may also help rice breeders develop

TABLE 5

The penetrances of QTL genotypes for rice blast resistance in the "Lemont" × "Teqing" crossing experiment

QTL	QTL genotype	IB54 = S, IG1 = S	IB54 = S, IG1 = R	IB54 = R, IG1 = S	IB54 = R, IG1 = R
qtl1	*QQ*	0.3457	0.0637	0.2043	0.3863
	Qq	0.0602	0.0548	0.0628	0.8222
	qq	0.0504	0.2092	0.0084	0.7319
qtl2	*QQ*	0.1764	0.1582	0.0874	0.5779
	Qq	0.2294	0.1683	0.0975	0.5048
	qq	0.0304	0.0407	0.0551	0.8738
qtl3	*QQ*	0.0080	0.0207	0.0199	0.9514
	Qq	0.0961	0.0752	0.1148	0.7139
	qq	0.2190	0.1858	0.0874	0.5078
qtl4	*QQ*	0.1899	0.1533	0.0786	0.5782
	Qq	0.0009	0.0034	0.0034	0.9924
	qq	0.0516	0.0496	0.0496	0.8314
qtl5	*QQ*	0.1856	0.2048	0.0628	0.5468
	Qq	0.1822	0.1188	0.1184	0.5806
	qq	0.0224	0.0176	0.0784	0.8816

R, resistance; S, susceptibility. *Q* represents the allele from parent "Lemont" and *q* represents the allele from parent "Teqing."

an optimal marker-assisted seletion scheme to improve blast resistance in rice.

DISCUSSION

Joint mapping offers several advantages over single-trait analyses. First, joint mapping may increase statistical power of QTL detection compared to single-trait analyses. Second, joint analysis can improve the precision of parameter estimation. Third, joint mapping provides an opportunity to answer more questions related to the genetic architecture of complex traits. These have been discussed by many authors (Jiang and Zeng 1995; Korol *et al.* 1995; Mangin *et al.* 1998; Henshall and Goddard 1999; Knott and Haley 2000) in multiple quantitative traits QTL mapping. Similar advantages also have been demonstrated here in the joint mapping for multiple binary traits. In this study, we paid more attention to the development of the EM algorithm rather than to various hypotheses tests, because the latter have been fully addressed by Jiang and Zeng (1995). In addition, the method was derived in the context of interval mapping. Extension to composite interval mapping should be preferred in practice, but this is simply a matter of implementation. Furthermore, the proposed method for F_2 populations can be easily extended to other types of populations, *e.g.*, backcrosses or four-way crosses, as demonstrated by the extension from F_2 to RILs described in this study. The method differs from one mating design to another only by the possible different number of genotypes and different transition matrix from one locus to another.

In fact, there has been much work on single binary trait mapping (Hackett and Weller 1995; Xu and Atchley 1996; Yi and Xu 1999, 2000; Xu *et al.* 2003), using likelihood-based methods or Bayesian methods. However, the method of separate analyses of individual binary traits is, so far, the only approach currently available. For the first time, we developed the full probability model for joint mapping of multiple binary traits. The method requires numerical multiple integrals, as we know that high-dimensional numerical integration cannot be implemented easily in practice. Therefore, we presented the method using two traits as examples. In real data analysis, one may pay more attention to the information extracted from the data and thus may wish to perform joint mapping for more than two traits using the general algorithm developed here. Two factors may limit the number of traits included in the analysis. One is the computing time and the other is the difficulty in interpreting the results. For the rice blast data analysis with two binary traits, QTL search for the entire rice genome took ~10 min, which is quite reasonable. For more than two traits, computing time is a big factor of concern. We highly recommended using a different but fast numerical integration algorithm specially designed for high-dimensional integration, *e.g.*, Monte Carlo integration. The Bayesian method implemented via Markov chain Monte Carlo (MCMC) is an ideal tool to accomplish this. In addition, the Bayesian method can handle the multiple-QTL model with ease. To deal with the problem of interpretation, one must have some intuitive knowledge about the trait relationships and hypotheses underlying the traits. In the disease-resistance case, one would be interested not only in the number of loci involved, but also in the level of race specificity of individual resistance loci, since the hypersensitive response of rice to *P. grisea* is known to be controlled by the gene-for-gene system (Silué *et al.* 1992). However, this gene-for-gene system normally assumes that only two consequences, resistance or susceptibility, would result from interactions between alleles at a resistance locus of host plants and alleles at its corresponding avirulence locus in pathogens, which may not be always true, as is discussed in the following section; imperfect penetrance appears to be an important feature of resistance loci involved in the gene-for-gene interactions between host plants and their pathogens.

We took the maximum-likelihood approach and implemented the method via the EM algorithm. This is different from the GEE method described earlier. We favor the EM algorithm because it was developed on the basis of all existing theory and methods currently used for mapping loci of regular quantitative traits. In single binary trait mapping, one of the most frequently asked questions is "What is the advantage of using the probability model over the simple analysis that treats the binary traits as if they were continuous traits?" (Visscher *et al.* 1996). The same question also may be asked here for joint mapping of multiple binary traits. Although we did not try the joint analysis of binary traits by ignoring the binary nature of the traits, we predict that treating binary traits as if they were continuous traits may result in similar power in most situations. In some special cases, the probabilistic model may provide higher power than the simplified analysis, and we have not figured out the parameter range in which this will happen. The probabilistic model approach enables estimation of penetrance of a particular QTL genotype, which is an important property of genes involved in human diseases (Terwilliger and Weiss 1998) and plant disease resistance, as we demonstrated here. McIntyre *et al.* (2001) developed a different probabilistic model for QTL mapping of binary traits. The method also allows calculation of penetrance. Extension of their model to multiple binary trait analysis is another alternative approach. We did not choose this extension because the threshold model via the EM algorithm has a natural connection to existing methods of QTL mapping.

Finally, with the successful development of joint mapping of both multiple quantitative and qualitative traits, an important but largely unexploited area in genetic mapping begins to emerge, *i.e.*, the joint analysis of mixed types of traits. Although some authors (Williams *et al.* 1999; Huang and Jiang 2003) already exploited this idea in the context of human genetic mapping under the IBD-based random model framework, it has never been explored in QTL mapping of experimental populations. The method may be particularly useful in situations where mapping qualitative disease resistance of the gene-for-gene system is the primary objective while traits associated with quantitative resistance to the same or different pathogens are measured as by-products in the experiment. This coexistence of qualitative and quantitative resistance is widely present in many relationships between host plants and their pathogens in natural and agricultural systems (Leonard and Czochor 1980). Joint analyses of the correlated qualitative and quantitative phenotypes may substantially increase the power of detecting disease resistance loci and allow exploration of new features of loci involved.

We are grateful to three anonymous reviewers for their suggestive comments on early versions of the manuscript. This research was supported by the National Institutes of Health grants R01-GM55321 and the United States Department of Agriculture National Research Initiative competitive grants program 00-35300-9245 to S.X.

LITERATURE CITED

Anderson, T. W., 1984 *An Introduction to Multivariate Statistical Analysis*, Ed. 2. Wiley, New York.

Chan, J. S. K., and A. C. Kuk, 1997 Maximum likelihood estimation for probit-linear mixed models with correlated random effects. Biometrics **88:** 86–97.

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

Cohen, A. C., 1991 *Truncated and Censored Samples*. Marcel Dekker, New York.

DeGroot, M. H., 1986 *Probability and Statistics*. Addison-Wesley, Reading, MA.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM-algorithm. J. R. Stat. Soc. **39:** 1–38.

Devroye, T., 1986 *Non-Uniform Random Variable Generation*. Springer-Verlag, New York.

Diggle, A. P., K.-Y. Liang and S. L. Zeger, 1996 Permutation tests for multiple loci affecting a quantitative character. Genetics **142:** 285–294.

Giri, N. C., 1996 *Multivariate Statistical Analysis*. Marcel Dekker, New York.

Gueorguieva, R. V., and A. Agresti, 2001 A correlated probit model for joint modeling of clustered binary and continuous responses. J. Am. Stat. Assoc. **96:** 1102–1112.

Hackett, C. A., and J. I. Weller, 1995 Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics **51:** 1252–1263.

Hackett, C. A., R. C. Meyer and W. T. B. Thomas, 2001 Multi-trait QTL mapping in barley using multivariate regression. Genet. Res. **77:** 95–106.

Henshall, J. M., and M. E. Goddard, 1999 Multiple trait mapping of quantitative trait loci after selective genotyping using logistic regression. Genetics **151:** 885–894.

Huang, J., and Y. M. Jiang, 2003 Genetic linkage analysis of a dichotomous trait incorporating a tightly linked quantitative trait in affected sib pairs. Am. J. Hum. Genet. **72:** 946–960.

Jiang, C., and Z-B. Zeng, 1995 Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics **140:** 1111–1127.

Jiang, C., and Z-B. Zeng, 1997 Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica **101:** 47–58.

Knott, S. A., and C. S. Haley, 2000 Multitrait least squares for quantitative trait loci detection. Genetics **156:** 899–911.

Korol, A. B., Y. T. Ronin and V. M. Kirzhner, 1995 Interval mapping of quantitative trait loci employing correlated trait complexes. Genetics **140:** 1137–1147.

Korol, A. B., Y. T. Ronin, A. M. Itskovich, J. Peng and E. Nevo, 2001 Enhanced efficiency of quantitative trait loci mapping analysis based on multivariate complexs of quantitative traits. Genetics **157:** 1789–1803.

Lander, E. S., and S. D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

Lange, C., and J. C. Whittaker, 2001 Mapping quantitative trait loci using generalized estimating equations. Genetics **159:** 1325–1337.

Leonard, K. J., and R. J. Czochor, 1980 Theory of genetic interactions among populations of plants and their pathogens. Ann. Rev. Phytopathol. **18:** 237–258.

Liang, K. Y., and S. L. Zeger, 1986 Longitudinal data analysis using generalized linear models. Biometrika **73:** 13–22.

Luo, Z. W., and M. J. Kearsey, 1992 Interval mapping of quantitative trait loci in an F2 population. Heredity **69:** 236–242.

Mangin, B., P. Thoquet and N. Grimslev, 1998 Pleiotropic QTL analysis. Biometrics **54:** 88–99.

McCulloch, C. E., 1994 Maximum likelihood variance components estimation for binary data. J. Am. Stat. Assoc. **89:** 330–335.

McIntyre, L. M., C. Coffman and R. W. Doerge, 2001 Detection and location of a single binary trait locus in experimental populations. Genet. Res. **78:** 79–92.

Piepho, H. P., 2001 A quick method for computing approximate thresholds for quantitative trait loci detection. Genetics **157:** 425–432.

SAS Institute, 1999 *SAS/IML User's Guide Version 8*. SAS Institute, Cary, NC.

Silué, D., J. L. Notteghem and D. Tharreau, 1992 Evidence of a gene-for-gene relationships in the *Oryza sativa-Magnaporthe grisea* pathosystem. Phytopathology **82:** 577–580.

Tabien, R. E., Z. Li, A. H. Paterson, M. A. Marchetti, J. W. Stansel *et al.*, 2000 Mapping of four major rice blast resistance genes from 'Lemont' and 'Teqing', and evaluation of their combinatorial effect for field resistance. Theor. Appl. Genet. **101:** 1215–1225.

Tallis, G. M., 1963 The moment generating function of the truncated multi-normal distribution. J. R. Stat. Soc. Ser. B **23**: 223–229.

Terwilliger, J. D., and K. M. Weiss, 1998 Linkage disequilibrium mapping of complex disease: Fantasy or reality? Curr. Opin. Biotechnol. **9:** 578–594.

Visscher, P. M., C. S. Haley and S. A. Knott, 1996 Mapping QTLs for binary traits in backcross and F2 populations. Genet. Res. **68:** 55–63.

Williams, J. T., P. Van Eerdewegh, L. Almasy and J. Blangero, 1999 Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results. Am. J. Hum. Genet. **65:** 1134–1147.

Xu, S., and W. R. Atchley, 1996 Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics **143:** 1417–1424.

Xu, S., N. Yi, D. Burke, A. Galecki and R. A. Miller, 2003 An EM algorithm for mapping binary disease loci: application to fibrosarcoma in a four-way cross mouse family. Genet. Res. **82:** 127–138.

Yi, N., and S. Xu, 1999 Mapping quantitative trait loci for complex binary traits in outbred populations. Heredity **82:** 668–676.

Yi, N., and S. Xu, 2000 Bayesian mapping of quantitative trait loci for complex binary traits. Genetics **155:** 1391–1403.

Communicating editor: R. Doerge

APPENDIX A: MOMENTS OF TRUNCATED BIVARIATE NORMAL DISTRIBUTION

Let $\mathbf{z}^T = [z_1\ z_2]$ be a vector of two variables distributed as a standardized bivariate normal distribution with correlation ρ. Let $[a_1, c_1]$ and $[a_2, c_2]$ be the double truncation points on variables z_1 and z_2, respectively, and define $\alpha_1 = \Pr(z_1 > c_1, z_2 > c_2)$ as the area (integral) within the domain. Let us further define the first and second moments of the truncated standardized bivariate normal distribution at $z_1 > c_1$ and $z_2 > c_2$ as

$$E(z_1) = \frac{\phi(c_1)[1 - \Phi(d_1)] + \rho\phi(c_2)[1 - \Phi(d_2)]}{\alpha_1}$$

$$E(z_2) = \frac{\phi(c_2)[1 - \Phi(d_2)] + \rho\phi(c_1)[1 - \Phi(d_1)]}{\alpha_1}$$

$$E(z_1^2) = \frac{\alpha + c_1\phi(c_1)[1 - \Phi(d_1)] + \rho^2 c_2\phi(c_2)[1 - \Phi(d_2)] + \rho(1 - \rho^2)\phi_2(c_1, c_2; \rho)}{\alpha_1}$$

$$E(z_2^2) = \frac{\alpha + c_2\phi(c_2)[1 - \Phi(d_2)] + \rho^2 c_1\phi(c_1)[1 - \Phi(d_1)] + \rho(1 - \rho^2)\phi_2(c_1, c_2; \rho)}{\alpha_1}$$

$$E(z_1 z_2) = \frac{\alpha\rho + \rho c_1\phi(c_1)\Phi(d_1) + \rho c_2\phi(c_2)\Phi(d_2) + (1 - \rho^2)\phi_2(c_1, c_2; \rho)}{\alpha_1}, \tag{A1}$$

where

$$d_1 = \frac{c_1 - \rho c_2}{\sqrt{1 - \rho^2}} \tag{A2}$$

and

$$d_2 = \frac{c_2 - \rho c_1}{\sqrt{1 - \rho^2}}. \tag{A3}$$

Equations A1 can be found from Tallis (1963). Similarly, we can calculate the first and second moments of truncated standardized bivariate normal distribution at $z_1 > a_1$ and $z_2 > c_2$, $z_1 > a_2$ and $z_2 > c_1$, and $z_1 > a_1$ and $z_2 > a_2$, respectively. We further denote the above four truncated domains by 1, 2, 3, and 4, respectively. The following formula is used to calculate the moments under the double truncation with $[a_1, c_1]$ and $[a_2, c_2]$,

$$T = \frac{\alpha_1 T_1 + \alpha_4 T_4 - \alpha_2 T_2 - \alpha_3 T_3}{\alpha}, \tag{A4}$$

where T_i, $i = 1, 2, 3, 4$, represents the arbitrary first moment of (A1), α_i represents the probability under the corresponding truncated domain, and

$$\alpha = \int_{a_1}^{c_1}\int_{a_2}^{c_2} \phi_2(z_1, z_2; \rho)\, dz_1\, dz_2 = \alpha_1 + \alpha_4 - \alpha_2 - \alpha_3. \tag{A5}$$

APPENDIX B: CONDITIONAL EXPECTATION AND VARIANCE VIA GIBBS SAMPLER

The basic idea of the Gibbs sampler is to find the distribution of one element, say y_{jk}, conditional on the remaining components in vector $\mathbf{y}_j$ and sample y_{jk} from the conditional distribution. Under the assumption of multivariate normality for the liability vector, *i.e.*, $\mathbf{y}_j \sim N_m(\mathbf{x}_j\mathbf{B}, \mathbf{R})$, the conditional density of a single component is univariate normal with mean and variance described as follows. First, let us make the following matrix partitioning, $\mathbf{y}_j = [y_{jk}\ \mathbf{y}_{j\bar{k}}]$, where

$$\mathbf{y}_{j\bar{k}} = [y_{j1} \ldots y_{j(k-1)} \quad y_{j(k+1)} \ldots y_{jm}] \tag{B1}$$

is a special notation for a subset of vector $\mathbf{y}_j$ that excludes y_{jk}; *i.e.*, the subscript $\bar{k}$ indexes all elements except k. Using this special notation we can partition matrix $\mathbf{B}$ into $\mathbf{B} = [\mathbf{b}_k\ \mathbf{B}_{\bar{k}}]$, where $\mathbf{b}_k^{\mathrm{T}} = [b_{0k}\ b_{1k}\ b_{2k}]$ is the kth column of matrix $\mathbf{B}$ and

$$\mathbf{B}_{\bar{k}} = [\mathbf{b}_1 \ldots \mathbf{b}_{k-1} \quad \mathbf{b}_{k+1} \ldots \mathbf{b}_m] \tag{B2}$$

is a submatrix of $\mathbf{B}$ with the kth column left out. Let us further partition matrix $\mathbf{R}$ into

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{kk} & \boldsymbol{R}_{k\bar{k}} \\ \mathbf{R}_{\bar{k}k} & \mathbf{R}_{\bar{k}\bar{k}} \end{bmatrix}, \tag{B3}$$

where $\mathbf{R}_{kk} = 1$, $\mathbf{R}_{k\bar{k}} = [\rho_{1k} \ldots \rho_{(k-1)k} \quad \rho_{k(k+1)} \ldots \rho_{km}]$, $\mathbf{R}_{\bar{k}k} = \mathbf{R}_{k\bar{k}}^{\mathrm{T}}$, and

$$\mathbf{R}_{\bar{k}\bar{k}} = \begin{bmatrix} 1 & \cdots & \rho_{1(k-1)} & \rho_{1(k+1)} & \cdots & \rho_{1m} \\ \vdots & \ddots & \vdots & \vdots & ⋰ & \vdots \\ \rho_{1(k-1)} & \cdots & 1 & \rho_{(k-1)(k+1)} & \cdots & \rho_{(k-1)m} \\ \rho_{1(k+1)} & \cdots & \rho_{(k-1)(k+1)} & 1 & \cdots & \rho_{(k+1)m} \\ \vdots & ⋰ & \vdots & \vdots & \ddots & \vdots \\ \rho_{1m} & \cdots & \rho_{(k-1)m} & \rho_{(k+1)m} & \cdots & 1 \end{bmatrix}. \tag{B4}$$

Note that $\mathbf{R}_{\bar{k}\bar{k}}$ is the submatrix of $\mathbf{R}$ with the kth row and kth column removed. The above matrix partitionings allow us to define the conditional mean of y_{jk} as

$$E(y_{jk}|\mathbf{x}_j, \boldsymbol{\theta}, \mathbf{y}_{j\bar{k}}) = \mathbf{x}_j\mathbf{b}_k + \mathbf{R}_{k\bar{k}}\mathbf{R}_{\bar{k}\bar{k}}^{-1}(\mathbf{y}_{j\bar{k}} - \mathbf{x}_j\mathbf{B}_{\bar{k}})^{\mathrm{T}} \tag{B5}$$

and the conditional variance as

$$\mathrm{Var}(y_{jk}|\mathbf{x}_j, \boldsymbol{\theta}, \mathbf{y}_{j\bar{k}}) = \mathbf{R}_{kk} - \mathbf{R}_{k\bar{k}}\mathbf{R}_{\bar{k}\bar{k}}^{-1}\mathbf{R}_{k\bar{k}}. \tag{B6}$$

Having found the distribution of one component conditional on the remaining components, one can easily sample each element from its perspective univariate normal distribution. The binary phenotype for each trait has not played a role in the above sampling scheme. To incorporate this information, we need to sample each liability from a truncated normal distribution with the mean and variance given above. For example, if $w_{jk} = 0$, y_{jk} should be sampled only if $y_{jk} \leq 0$. If $w_{jk} = 1$, however, y_{jk} should be sampled only if $y_{jk} > 0$. In fact, we adopted the algorithm of Devroye (1986) to simulate a variable from a truncated normal distribution. This special algorithm has a 100% rate of acceptance. The Monte Carlo sampling process is repeated many times with the simulated $\mathbf{y}_j$ forming a large sample, $\mathbf{y}_j^{(1)}, \mathbf{y}_j^{(2)}, \ldots, \mathbf{y}_j^{(M)}$, where M is a large number. Discarding the observations during the burn-in period and thereafter saving one observation every few cycles, we get a sample containing roughly independent observations, from which the sampled mean vector and the covariance matrix are calculated. The sampled mean and covariance matrix are used in place of $\boldsymbol{\mu}_{jq} = E(\mathbf{y}_j|\mathbf{w}_j, \mathbf{h}_q, \boldsymbol{\theta})$ and $\mathbf{U}_{jq} = \mathrm{Var}(\mathbf{y}_j|\mathbf{w}_j, \mathbf{h}_q, \boldsymbol{\theta})$.

APPENDIX C: DERIVATION OF THE EM ALGORITHM

The expected likelihood function: The complete-data log likelihood is

$$L(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Y}) = \text{const} - \frac{n}{2}\ln|\mathbf{V}| - \frac{1}{2}\sum_{j=1}^{n}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})\mathbf{V}^{-1}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})^{\mathrm{T}}, \tag{C1}$$

where $\boldsymbol{\theta} = \{\mathbf{B}, \mathbf{V}\}$ is the vector of parameters and $\mathbf{X}$ and $\mathbf{Y}$ are the missing values. The data are the phenotypes of multiple binary traits, denoted by $\mathbf{W}$.

The expectation of the complete-data log-likelihood function conditional on the current parameter values and the data is

$$L(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_X\{E_{Y|X}[L(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Y})|\boldsymbol{\theta}^{(t)}, \mathbf{W}]\} = \text{const} - \frac{n}{2}\ln|\mathbf{V}| - \frac{1}{2}\sum_{j=1}^{n}E_X\{E_{Y|X}[(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})\mathbf{V}^{-1}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})^{\mathrm{T}}|\boldsymbol{\theta}^{(t)}, \mathbf{W}]\}, \tag{C2}$$

where the expectation is taken with respect to the missing values, $\mathbf{X}$ and $\mathbf{Y}$, conditional on the current parameters $\boldsymbol{\theta}^{(t)} = \{\mathbf{B}^{(t)}, \mathbf{V}^{(t)}\}$ and the data $\mathbf{W}$. Note that we use a special notation $E_{Y|X}$ to denote conditional expectation with respect to $\mathbf{Y}$ given $\mathbf{X}$. The expectation of the complete-data log-likelihood (C2) is the target function subject to maximization in the EM algorithm.

Maximization with respect to B: The expectation of the complete-data log-likelihood function relevant to $\mathbf{B}$ is

$$\begin{aligned}
L(\mathbf{B}|\boldsymbol{\theta}^{(t)}) &= \text{const} - \frac{1}{2}\sum_{j=1}^{n}E_X\{E_Y|_X[(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})(\mathbf{V}^{(t)})^{-1}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})^{\mathrm{T}}|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j]\} \\
&= \text{const} - \frac{1}{2}\sum_{j=1}^{n}\mathrm{tr}\,\{(\mathbf{V}^{(t)})^{-1}E_X\{E_{Y|X}[(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})^{\mathrm{T}}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j]\}\} \\
&= \text{const} - \frac{1}{2}\,\mathrm{tr}\,\{(\mathbf{V}^{(t)})^{-1}\sum_{j=1}^{n}E_X\{E_{Y|X}[(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})^{\mathrm{T}}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j]\}\}.
\end{aligned} \tag{C3}$$

The partial derivative of $L(\mathbf{B}|\boldsymbol{\theta}^{(t)})$ with respect to $\mathbf{B}$ is

$$\begin{aligned}
\frac{\partial}{\partial\mathbf{B}}L(\mathbf{B}|\boldsymbol{\theta}^{(t)}) &= -\frac{1}{2}\frac{\partial}{\partial\mathbf{B}}\,\mathrm{tr}\left\{(\mathbf{V}^{(t)})^{-1}\sum_{j=1}^{n}E_X\{E_{Y|X}[(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})^{\mathrm{T}}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j]\}\right\} \\
&= -\frac{1}{2}(\mathbf{V}^{(t)})^{-1}\sum_{j=1}^{n}E_X\left\{E_{Y|X}\left[\frac{\partial}{\partial\mathbf{B}}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})^{\mathrm{T}}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B})|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j\right]\right\} \\
&= (\mathbf{V}^{(t)})^{-1}\sum_{j=1}^{n}E_X\{E_{Y|X}[(\mathbf{y}_j^{\mathrm{T}}\mathbf{x}_j - \mathbf{B}^{\mathrm{T}}\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j)|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j]\} \\
&= (\mathbf{V}^{(t)})^{-1}\sum_{j=1}^{n}E_X[E_{Y|X}(\mathbf{y}_j^{\mathrm{T}}\mathbf{x}_j|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j)] - (\mathbf{V}^{(t)})^{-1}\mathbf{B}^{\mathrm{T}}\sum_{j=1}^{n}E_X[E_{Y|X}(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j)].
\end{aligned} \tag{C4}$$

Setting (C4) equal to zero and solving for $\mathbf{B}$, we obtain

$$\hat{\mathbf{B}} = \left\{\sum_{j=1}^{n}E_X[E_{Y|X}(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j)]\right\}^{-1}\left\{\sum_{j=1}^{n}E_X[E_{Y|X}(\mathbf{x}_j^{\mathrm{T}}\mathbf{y}_j|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j)]\right\}. \tag{C5}$$

In the main text, we used the following simple notation for the conditional expectations,

$$E(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j) = E_X[E_{Y|X}(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j)] \quad \text{and} \quad E(\mathbf{x}_j^{\mathrm{T}}\mathbf{y}_j) = E_X[E_{Y|X}(\mathbf{x}_j^{\mathrm{T}}\mathbf{y}_j|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j)].$$

With this short notation, the solution for $\mathbf{B}$ becomes

$$\hat{\mathbf{B}} = \left[\sum_{j=1}^{n}E(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j)\right]^{-1}\left[\sum_{j=1}^{n}E(\mathbf{x}_j^{\mathrm{T}}\mathbf{y}_j)\right], \tag{C6}$$

which concludes the proof of Equation 13 of the main text.

Maximization with respect to V: The expectation of the complete-data log-likelihood function relevant to $\mathbf{V}$ is

$$\begin{aligned}
L(\mathbf{V}|\boldsymbol{\theta}^{(t)}) &= \text{const} - \frac{n}{2}\ln|\mathbf{V}| - \frac{1}{2}\sum_{j=1}^{n}E_X\{E_{Y|X}[(\mathbf{y}_j - \mathbf{x}_j\mathbf{B}^{(t)})\mathbf{V}^{-1}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B}^{(t)})^{\mathrm{T}}|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j]\} \\
&= \text{const} + \frac{n}{2}\ln|\mathbf{V}^{-1}| - \frac{1}{2}\,\mathrm{tr}\left\{\mathbf{V}^{-1}\sum_{j=1}^{n}E_X\{E_{Y|X}[(\mathbf{y}_j - \mathbf{x}_j\mathbf{B}^{(t)})^{\mathrm{T}}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B}^{(t)})|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j]\}\right\}.
\end{aligned} \tag{C7}$$

The partial derivative of this likelihood function with respect to $\mathbf{V}$ is complicated, but the derivative of L with respect to $\mathbf{V}^{-1}$ is straightforward. On the basis of the invariance property of ML analysis, if $\widehat{\mathbf{V}^{-1}}$ is the MLE of $\mathbf{V}^{-1}$, then $(\widehat{\mathbf{V}^{-1}})^{-1} = \hat{\mathbf{V}}$ should be the MLE of $\mathbf{V}$. Therefore, we set the partial derivative of L with respect to $\mathbf{V}^{-1}$ equal to zero and solve for $\mathbf{V}$, as

$$\begin{aligned}\frac{\partial}{\partial \mathbf{V}^{-1}} L(\mathbf{V}|\boldsymbol{\theta}^{(t)}) &= \frac{n}{2}\frac{\partial}{\partial \mathbf{V}^{-1}} \ln|\mathbf{V}^{-1}| - \frac{1}{2}\frac{\partial}{\partial \mathbf{V}^{-1}} \operatorname{tr}\left\{\mathbf{V}^{-1}\sum_{j=1}^{n} E_X\left\{E_{Y|X}\left[(\mathbf{y}_j - \mathbf{x}_j\mathbf{B}^{(t)})^{\mathrm{T}}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B}^{(t)})|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j\right]\right\}\right\} \\ &= \frac{n}{2}\mathbf{V} - \frac{1}{2}\sum_{j=1}^{n} E_X\left\{E_{Y|X}\left[(\mathbf{y}_j - \mathbf{x}_j\mathbf{B}^{(t)})^{\mathrm{T}}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B}^{(t)})|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j\right]\right\}. \end{aligned} \tag{C8}$$

Setting (C8) equal to zero and solving for $\mathbf{V}$, we get

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{j=1}^{n} E_X\left\{E_{Y|X}\left[(\mathbf{y}_j - \mathbf{x}_j\mathbf{B}^{(t)})^{\mathrm{T}}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B}^{(t)})|\boldsymbol{\theta}^{(t)}, \mathbf{W}_j\right]\right\}. \tag{C9}$$

In the main text, we adopted a short notation for the expectation and denoted Equation C9 by

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{j=1}^{n} E\left[(\mathbf{y}_j - \mathbf{x}_j\mathbf{B}^{(t)})^{\mathrm{T}}(\mathbf{y}_j - \mathbf{x}_j\mathbf{B}^{(t)})\right]. \tag{C10}$$

This proves Equation 14 of the main text of the article.