BMC Genetics

Research article

Open Access

Multiple-trait quantitative trait locus mapping with incomplete phenotypic data

Zhigang Guo[1,2] and James C Nelson*[1]

Address: [1]Department of Plant Pathology, Kansas State University, Manhattan, Kansas, USA 66506 and [2]Syngenta Seeds, Inc., Clinton, Illinois, 61727, USA

Email: Zhigang Guo - zhigang_guo@syngenta.com; James C Nelson* - jcn@ksu.edu

* Corresponding author

Published: 5 December 2008

BMC Genetics 2008, **9**:82 doi:10.1186/1471-2156-9-82

Received: 28 March 2008

Accepted: 5 December 2008

This article is available from: http://www.biomedcentral.com/1471-2156/9/82

Abstract

Background: Conventional multiple-trait quantitative trait locus (QTL) mapping methods must discard cases (individuals) with incomplete phenotypic data, thereby sacrificing other phenotypic and genotypic information contained in the discarded cases. Under standard assumptions about the missing-data mechanism, it is possible to exploit these cases.

Results: We present an expectation-maximization (EM) algorithm, derived for recombinant inbred and F_2 genetic models but extensible to any mating design, that supports conventional hypothesis tests for QTL main effect, pleiotropy, and QTL-by-environment interaction in multiple-trait analyses with missing phenotypic data. We evaluate its performance by simulations and illustrate with a real-data example.

Conclusion: The EM method affords improved QTL detection power and precision of QTL location and effect estimation in comparison with case deletion or imputation methods. It may be incorporated into any least-squares or likelihood-maximization QTL-mapping approach.

Background

Statistical methods for identifying and mapping genes controlling complex traits, commonly known as quantitative trait loci or QTL, have been developed to a high degree. The primary focus has been on methods for single traits ([1-8] and many others). It was proposed [9,10] that multi-trait QTL mapping methods that consider simultaneously several correlated phenotypic traits, or a single trait measured in several environments, offer increased detection power and precision of location and effect estimation over single-trait QTL mapping. This is because trait-by-trait QTL-searching neglects information contained in the data about the common influence of a QTL on more than one trait, in more than one environment, or [11] at more than one developmental stage. Multi-trait (MT) QTL mapping allows a formal test of pleiotropy of a QTL for multiple traits or QTL-by-environment interaction for a single trait measured across multiple environments. The enhancement by MT of QTL-detection power is greatest when the QTL induces covariation between the tested traits in the direction opposite to that from "background" sources [12,13]. These advantages have been exploited in animal [14-16] and plant [17] studies.

With the promise of increased power from a multivariate approach comes an interesting problem: what to do when some of the multivariate data are missing.

Two main statistical approaches have been elaborated for multi-trait QTL analysis: regression [10,18-21] and maximum likelihood or ML [9]. Regression QTL-mapping methods, though easier to implement and faster to compute, give biased parameter estimates with sparse markers [22] or when QTLs interact or are closely linked [23], while ML methods are free of these defects [23]. It has also been proposed to transform multiple traits into canonical variates so that conventional univariate interval QTL mapping can be applied [18,24,25], but interpretation of the results is difficult.

Though QTL-mapping data are often incomplete, information-recovery methods are at present applied only to genotypic data. For incompletely informative marker-genotype data, posterior distributions are readily estimated from flanking markers in the same individual [26]. For unknown QTL genotypes at tested positions in map intervals, ML methods estimate posterior distributions simultaneously with the parameters of a phenotypic mixture distribution [4], while regression methods [1] replace missing QTL genotypes with their expectations given flanking markers. Variations based on sampling include multiple imputation (MI) as described by [27] and [22] and Bayesian approaches (*e.g.* [5-7,28,29]).

In contrast to genotypic data, missing phenotypic data for any trait results in discarding all cases (individuals) lacking even one value, sacrificing all other phenotypic and genotypic information available for these cases. The problem was recognized by [20], but they provided no solution, nor does conventional QTL-mapping software offer an alternative to this "casewise" [30] deletion. Is there one?

Completion of incomplete multivariate data may be done by imputation (single or multiple), by EM algorithm, or by Bayesian approaches. Single imputation typically replaces missing data with three kinds of values: a value drawn from a model-based distribution, a mean of other observations of the same variable, or a conditional mean calculated by least-squares regression on predictors. MI [31,32] fills in missing data by imputing multiple (*e.g.* 3–5) times to produce several complete datasets, with parameter estimates calculated as the average over the results from these datasets. The defect of imputation methods, in analyses such as QTL mapping where we want ML estimates of statistics, is that bias is introduced by maximization of the likelihood over both original and imputed data. In contrast, the EM algorithm as described by [33] focuses not on replacing a missing value with its expectation, but on using the information available in the original dataset. In the framework of EM, missing data imputed are in effect integrated out of the complete-data log likelihood by iterative refinement of their expectation. An EM method described by [34] addressed the problem of missing genotype or phenotype data in single-marker QTL analyses by the use of flanking-marker genotypes. While free of the dependence on recombination estimates to which interval-mapping methods are subject, the method accommodated only single traits. [35] provided an EM algorithm for incomplete multivariate data and extended it to accommodate multiple regression with missing responses. A Bayesian approach developed by [36] for joint mapping with multiple traits in outbred populations employed an identity-by-descent (IBD)-based variance-components model and reversible-jump Markov-chain Monte Carlo (MCMC) estimation, but did not consider missing phenotypic and genotypic data. It would be possible to derive an MCMC algorithm to sample missing phenotypic entries from their posterior distribution, though Bayesian approaches are computationally intensive and often criticized for lacking a test statistic.

Here we describe an adaptation of the EM method of [35] to the case of multi-trait QTL mapping with incomplete phenotypic data. For simplicity we have limited our scope of mating design to biparental crosses between inbreds. We show that the tests for QTL main effects may be constructed as in [9], and we describe the properties and behavior of the test statistics and QTL effect and position estimates based on simulation studies and a real example.

Results

Power

As expected, power was highest when data were complete (Figure 1). When data were missing, EM, MS and CMS gave power superior to CaD in all cases. MS and CMS gave similar power, equal to or lower than that of EM. The gain in power for EM over CaD increased with the proportion of missing data. This trend was also seen for gain in power of EM over MS or CMS, but to a lower degree.

Figure 1 also shows that EM gave QTL detection power about equal to that supplied by CaD with half the proportion of missing data. Simple probability calculations yield the numbers to which this power relationship corresponds. As an example, in a population of size 300 with 0.4 of the data missing from each of two traits, the EM method was operating on only 108 lines carrying complete data and another 144 lines with partial data, but achieved power corresponding to 192 lines with complete data. The increase in effective (equivalent-power) number of complete records achieved by the EM method can be estimated graphically from Figure 2. Here the effective complete-data sample sizes achieved by EM were about 271, 255, 230, and 190, representing gains of 1, 12, 38 and 82 over the number of complete records available for CaD at missing levels of 0.05, 0.1, 0.2 and 0.4.

Specificity

All the methods gave similar QTL-detection specificity of 0.98 to 1.00, except with sample size 100 and missing

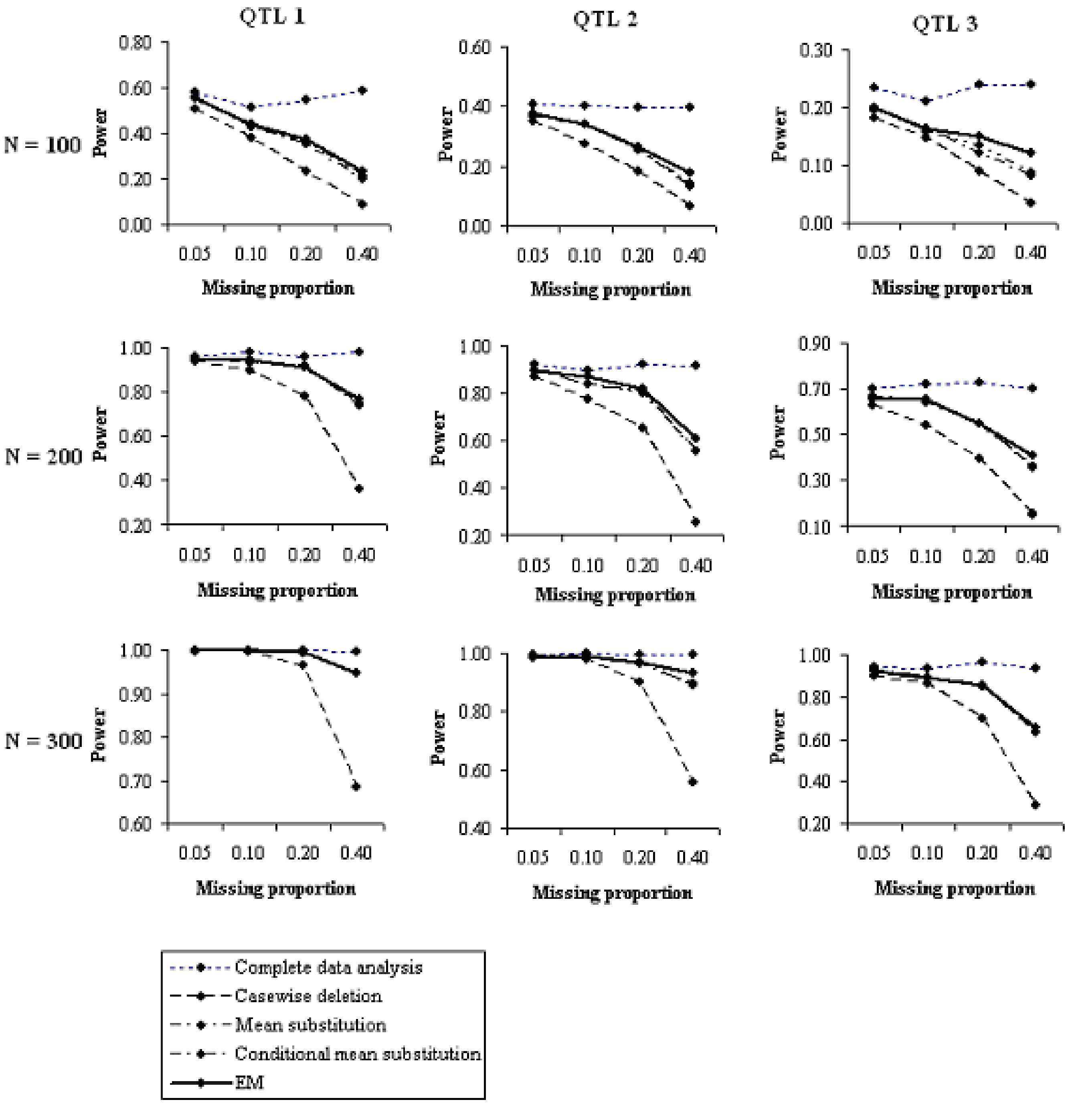

Figure 1
Statistical power of five multiple-trait QTL-mapping methods in simulated populations with four levels of missing data.

proportion 0.40, where CaD gave specificity as low as 0.93.

Accuracy and precision of QTL effect estimation

All methods gave reasonable estimates of QTL positions. CoD and CaD provided the highest and lowest precisions for QTL position estimation (Figure 3), while those of MS, MS, and EM were very similar and intermediate. For QTL effects (Figure 4), CoD, CaD and EM provided unbiased estimates, while both MS and CMS underestimated these parameters, CMS by slightly less. The extent of underesti-

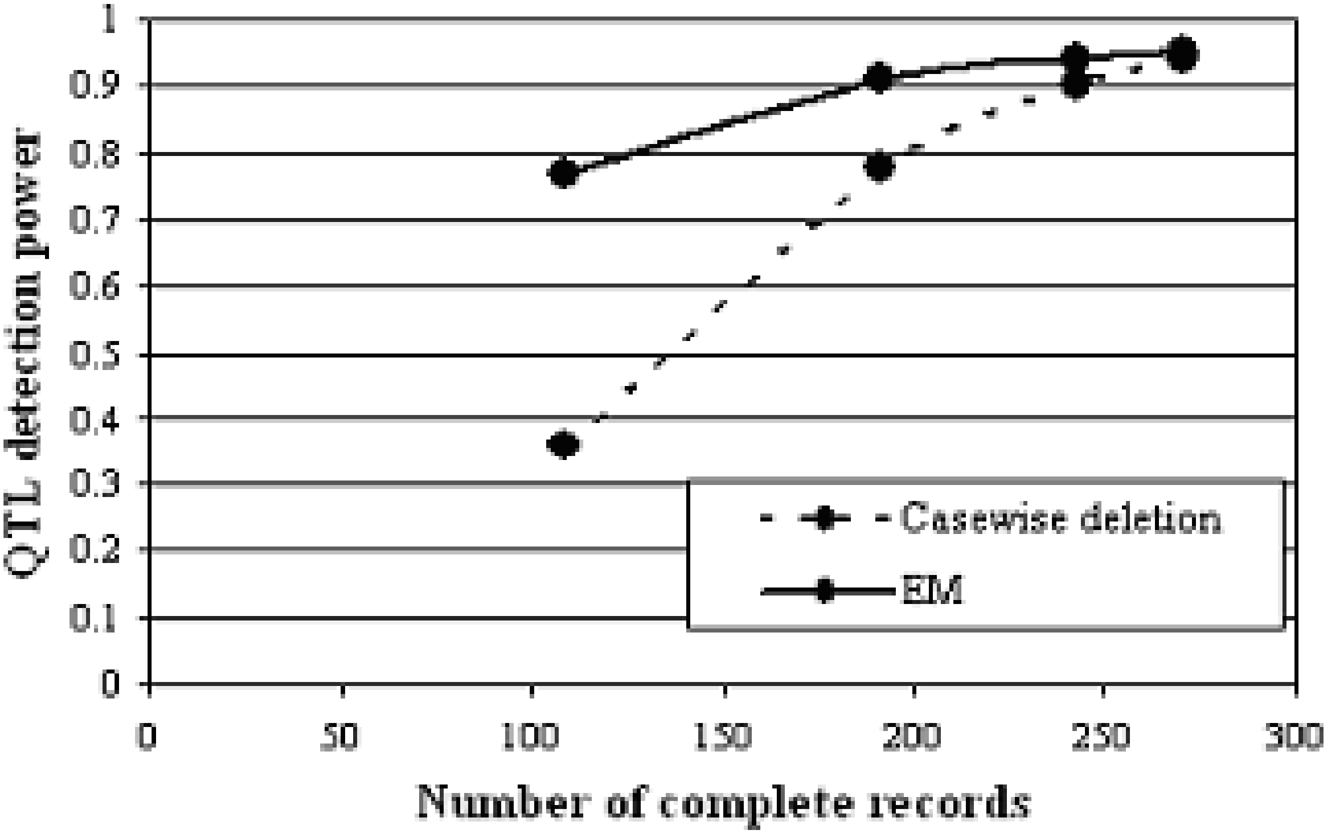

Figure 2
Power of QTL 1 detection after casewise deletion and by the EM method as a function of the number of complete trait records for 300 simulated RILs.

mation tended to increase with missing percentage and decrease with sample size (not shown here).

Real data analyses

In the selected example from a real dataset, EM provided higher power for the detection of a QTL on rice chromosome 12 than CaD and MS (Figure 5). While CMS also identified the QTL, its position estimate was somewhat biased, as were those of CaD and MS (Table 1). In general, EM provided estimates of QTL position and effect closer to those from CoD than were the estimates from CaD, MS and CMS.

Discussion

While any substitution of unavailable data by their expectations based on trends in available data will increase the precision of parameter estimates over those obtained by CaD, the EM-based multi-trait QTL mapping method we propose here is superior to MS and CMS for several reasons. MS underestimates phenotypic variation and QTL effect due to fill-in of missing data with a single value, resulting in decreased power compared with our method especially when amounts of missing data are relatively large. The same trend can be observed for CMS, which, as a precursor of the EM algorithm, is closely related to a single EM iteration [35]. Although CMS gave better estimates of QTL effect than MS, it still underestimates variance [35].

While we did not include MI [31,37] in the simulation study, we doubt its potential utility for multi-trait QTL mapping with missing trait data. We investigated MI by filling in missing trait data with values sampled from their conditional distributions under the null and alternative hypotheses given the observed trait values. Resulting logarithm-of-odds (LOD) profiles were sawtoothed (not shown here) due to random sampling, and a different profile could be obtained with each analysis even with many imputations (*e.g.* 100 compared with 3–5 in regular MI) performed at each QTL test position. For these reasons, apart from the high computational cost, we did not pursue this method further.

For MS, CMS, and even MI, the effects on QTL mapping of introducing imputed data need further study. Although simulation results showed specificities close to those of

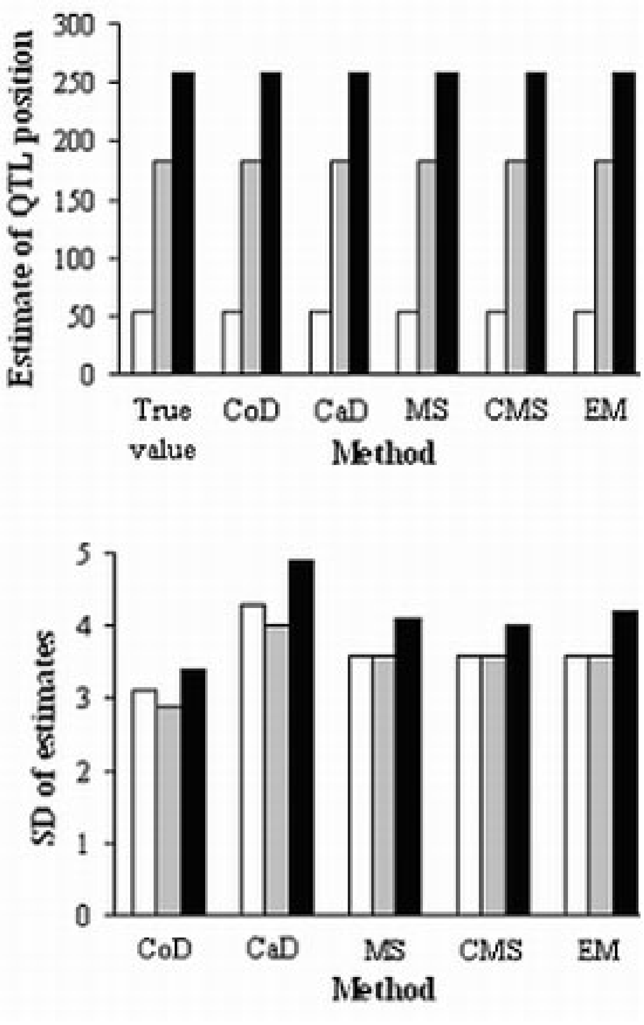

Figure 3
QTL position estimates and standard deviations with 200 simulated RILs and 0.40 missing proportion for each trait. White, gray, and black bars represent QTLs 1, 2 and 3. CoD: complete data analysis; CaD: casewise deletion; MS: mean substitution; CMS: conditional mean substitution; EM: EM algorithm; SD: standard deviation

our method, complete-data analysis, and CaD, the bias imposed on the LOD test statistic by introduction of these "artificial" data remains unknown. Interestingly, in the real-data example we chose for illustration, CMS, besides apparently biasing the location of the QTL, gave higher background LOD scores than CoD in regions away from the QTL, while the EM and other approaches did not (Fig. 5). In fact, imputation of missing data is also performed in the E step of our EM algorithm. But this kind of imputation only furnishes a pivot to facilitate parameter estimation and is actually not involved in the likelihood calculation. Thus, theoretically, the EM-based method does not bias QTL detection and parameter estimation as may imputation methods.

The information gain of our method over CaD, MS, and CMS depends on the amount of missing trait data. The reason is readily explained by the following example for CaD. Consider a sample of 200 individuals with missing proportion 0.1 for each of two traits independently. The average number of individuals available for CaD is 162 and that for EM 198, and the difference is 36. This difference expands to 96 with a missing proportion of 0.4. In other words, power is lost more slowly with data loss when the information-recovering EM method is applied.

Cofactor markers too may lack genotype data. In our model, these are replaced by expectations given flanking markers, computed by the method of [26]. Prediction of missing trait data employs these imputed cofactor genotypes, with a resulting potential for error in parameter estimation. While such error could be minimized by extension of our method to include missing cofactor data, we suspect that the improvement would hardly justify the computation, and remark that the issue is shared by all composite interval-mapping methods.

While the EM method should give more power than the information-discarding alternatives regardless of the chosen QTL acceptance threshold, we do not know how to find the optimum threshold for multiple-trait mapping when our algorithm is applied. Indeed, the question has not to our knowledge been satisfactorily answered to date even for the case of complete data. Thresholds in our simulation experiment were based on analyses of simulated populations lacking QTLs but with complete trait data conforming to the same variance structure as the QTL population – an option not available to an analyst in practice. A working method might be to adopt a threshold lying between the overly conservative one calculated from permutation (of individual records for all traits) applied to the complete data remaining after CaD and the insufficiently conservative one that would be obtained from permutation applied under our reconstruction algorithm. For nonpathological data sets our simulation results suggest that this range will be relatively narrow. The quick and approximate method of [38] might serve as well as any for establishing thresholds.

Some extensions of the EM method are promising. First, we have derived the EM calculation of the hypothesis test for QTL main effect. By following the procedure of [9], one may derive specific EM implementations for other hypothesis tests including for QTL-by-environment interaction, pleiotropy, and pleiotropy vs. close linkage. Second, the EM method may be extended to multiple interval mapping [3] with multiple traits and incomplete phenotypic data. Third, mixed-model QTL mapping as recommended by [9] can now be applied to incomplete trait data as an alternative method for multi-trait QTL mapping. When multiple traits are actually different expres-

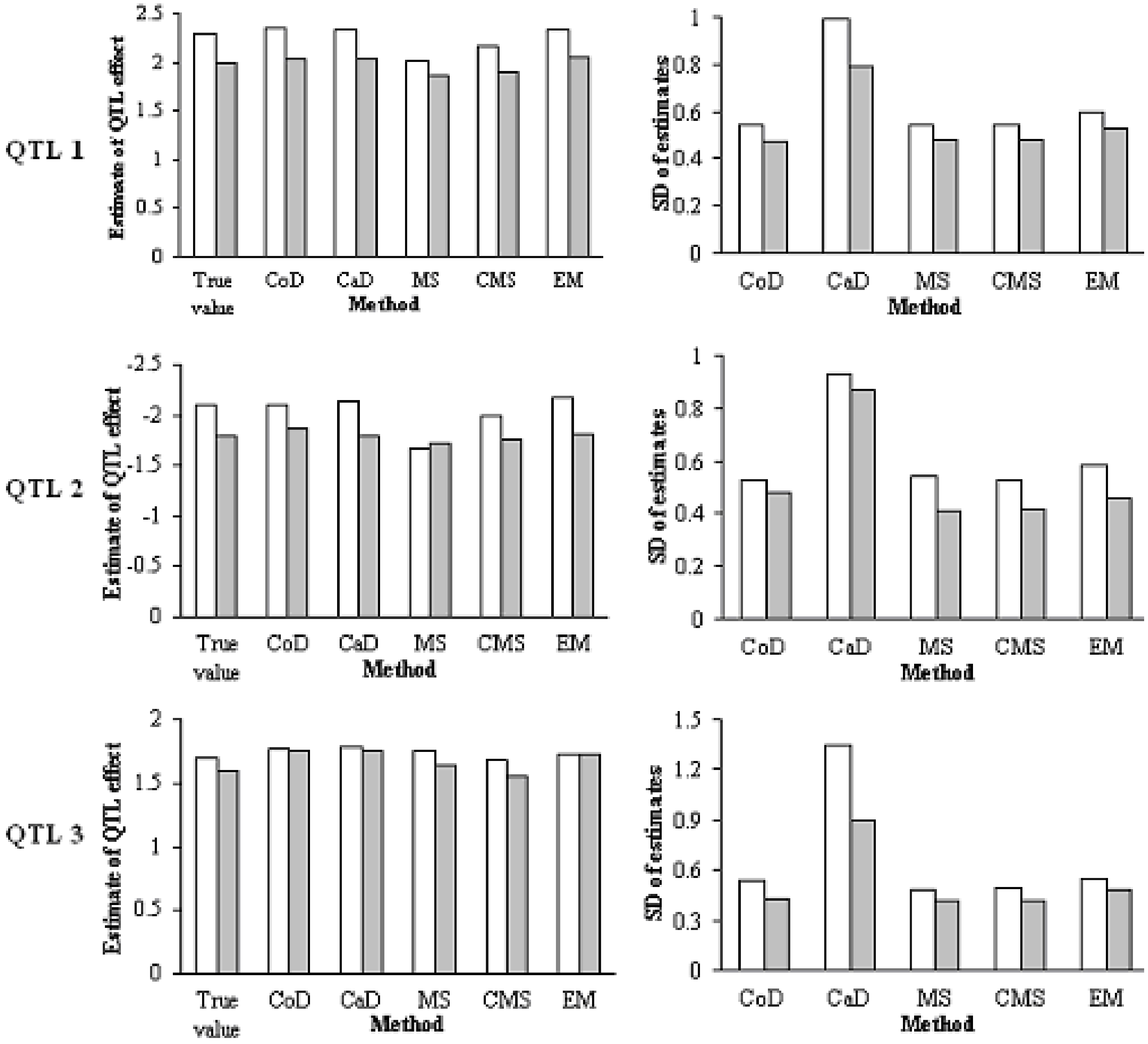

Figure 4
QTL effect estimates and their standard deviations with 200 simulated RILs and 0.40 missing percentage for each trait. White bars represent trait 1 and gray trait 2. CoD: complete data analysis; CaD: casewise deletion; MS: mean substitution; CMS: conditional mean substitution; EM: EM algorithm; SD: standard deviation

sions of a single trait in different environments (locations or years), a mixed model allows treating environmental effect as a random and QTL effect as a fixed factor [39,40]. One of the advantages of the mixed model is in accommodating both balanced and unbalanced data structure.

In principle, our EM approach may be used to handle multi-trait QTL mapping with any proportion of missing phenotypic data. But if the data may not be MAR, as is especially likely when the missing proportion is large, a more prudent course of analysis is to find out why not, and work out an appropriate statistical method for QTL mapping.

The method we have presented requires more computing time than the conventional EM or ECM interval-mapping algorithm. There are two reasons for this. First, to obtain parameter estimates, the EM algorithm must be applied

Table 1: QTL statistics for analysis of two traits in a 325-line doubled-haploid rice population with 10% simulated missing data, estimated by several approaches.

Method	Position (cM)	LOD	Additive effect	
			Sheath blight	Heading days
EM	21	4.47	-0.38	0.80
CaD	24	3.07	-0.33	0.70
MS	22	3.76	-0.32	0.66
CMS	16	4.41	-0.55	0.59
CoD	21	6.58	-0.44	1.06

Key EM: EM algorithm; CaD: casewise deletion; MS: mean substitution; CMS: conditional mean substitution; CoD: complete data analysis.

under both null and alternative hypotheses, because the trait data are missing in both cases. In contrast, conventional methods require EM iteration only under the alternative hypothesis. Second, our EM algorithm is used to complete both QTL genotype and phenotype in the case of ML-based QTL mapping, while the conventional method must complete only QTL genotype. The computing load increases with the proportion of missing data, but the extreme amounts of missing data we have simulated are unusual in real experiments.

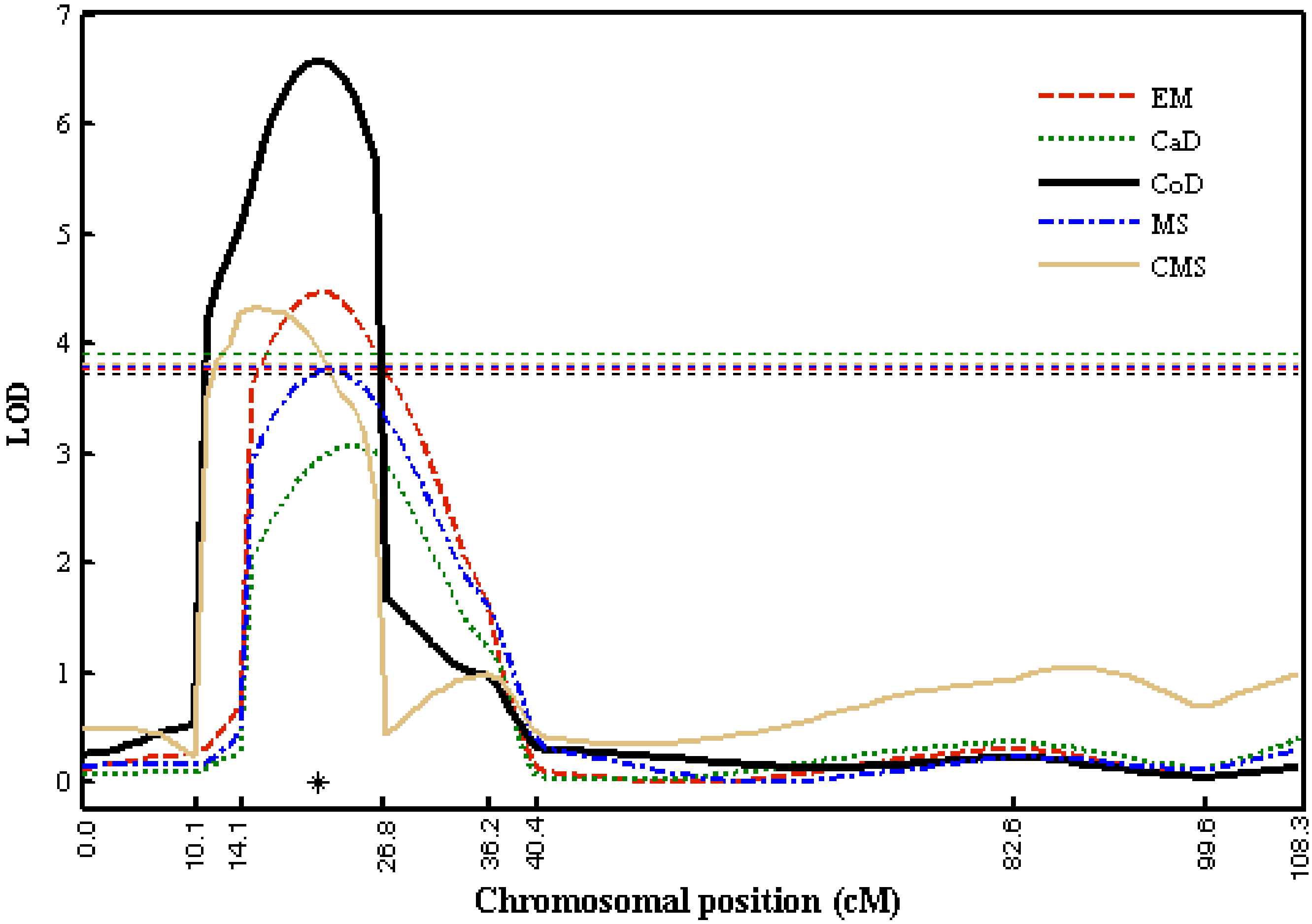

Figure 5
LOD profiles of several approaches to analysis of a putative QTL for two traits in a 325-line doubled-haploid rice population with 10% simulated missing data. CoD: complete data analysis; CaD: casewise deletion; MS: mean substitution; CMS: conditional mean substitution; EM: EM algorithm. Horizontal dotted lines in the corresponding colors show the empirical LOD thresholds for these methods. The asterisk marks the position (at 21 cM) of the QTL identified by CoD.

Methods

Missing-data mechanism is ignored

Several kinds of "missingness" have been defined [37]. Under MAR, "missing at random", the probability of missing phenotypic data within any genotype class is unrelated to the phenotypic value. Either for MAR or the stronger assumption, MCAR or "missing completely at random" (missingness also independent of genotype), estimation methods need not model a missing-data mechanism. Either assumption seems reasonable in conventional mapping practice and is accommodated by the method described here.

Multivariate regression with incomplete data

Consider the linear model

$$\mathbf{Y}_{n\times m} = \mathbf{X}_{n\times p}\,\mathbf{B}_{p\times m} + \mathbf{E}_{n\times m}, \tag{1}$$

where **Y** is a $(n \times m)$ response matrix with n the number of individuals and m the number of traits (or environments); **X** is a $(n \times p)$ design matrix with p predictors; **B** is a matrix of regression coefficients associated with **X**, **E** is an error matrix; and E_i $(i = 1, 2, ..., n)$ follows a multivariate normal distribution with means zero and variance-covariance matrix

$$\mathbf{V} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1m}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2m}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1}^2 & \sigma_{m2}^2 & \cdots & \sigma_{mm}^2 \end{pmatrix} \tag{2}$$

For a random sample, the log likelihood of observations is given by

$$\ell(\mathbf{B},\mathbf{V};\mathbf{Y}) = -\frac{nm}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\ln|\mathbf{V}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}_i\mathbf{B})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}_i - \mathbf{x}_i\mathbf{B}) \tag{3}$$

where $\mathbf{y}_i$ is the response and $\mathbf{x}_i$ the predictor vector of the i^{th} individual, $\mathbf{Y}^{\mathrm{T}} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]$, and $\mathbf{X}^{\mathrm{T}} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$. Parameters **B** and **V** are estimated by maximization of (3).

Suppose there are some missing entries in $\mathbf{y}_i$. The log likelihood of observations can be written as

$$\ell(\mathbf{B},\mathbf{V};\mathbf{Y}) = -\frac{nm}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\ln\left|\mathbf{V}_i^{obs,obs}\right| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i^{obs} - \mathbf{x}_i^{obs}\mathbf{b}^{obs})^{\mathrm{T}}(\mathbf{V}_i^{obs,obs})^{-1}(\mathbf{y}_i^{obs} - \mathbf{x}_i^{obs}\mathbf{b}^{obs}) \tag{4}$$

where $\mathbf{y}_i^{obs}$ is the observed part of $\mathbf{y}_i$, $\mathbf{x}_i^{obs}$ is the part of the predictor vector associated with $\mathbf{y}_i^{obs}$, $\mathbf{b}^{obs}$ contains the regression coefficients associated with $\mathbf{x}_i^{obs}$, and $\mathbf{V}_i^{obs,obs}$ is the submatrix of **V** representing the variance-covariance matrix of the traits for which the i^{th} individual has complete data. Since missing entries vary among individuals, the log likelihood is a logarithm sum of multivariate normal probabilities of varying dimensions.

Estimates of parameters **B** and **V** cannot be obtained by direct maximization of (4) with respect to the individual parameters. To estimate parameters in the presence of missing data we may apply the EM algorithm of [35]. For the i^{th} individual with some missing trait entries, we partition its trait $\mathbf{y}_i$, its mean $\mu_i = \mathbf{x}_i\mathbf{B}$, and the variance-covariance matrix **V** as

$$\mathbf{y}_i = [\mathbf{y}_i^{obs}, \mathbf{y}_i^{miss}], \tag{5}$$

where $\mathbf{y}_i^{miss}$ is a vector composed of the missing trait data of individual i,

$$\boldsymbol{\mu}_i = [\boldsymbol{\mu}_i^{obs}, \boldsymbol{\mu}_i^{miss}], \tag{6}$$

and

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_i^{obs,obs} & \mathbf{V}_i^{obs,miss} \\ \mathbf{V}_i^{miss,obs} & \mathbf{V}_i^{miss,miss} \end{bmatrix}. \tag{7}$$

Given these partitions, the MLEs of parameters in model (1) are obtained as follows.

ALGORITHM 1: Starting with random initial values $\hat{\boldsymbol{\theta}}^{(0)} = [\hat{\mathbf{B}}^{(0)}, \hat{\boldsymbol{\mu}}^{(0)}, \hat{\mathbf{V}}^{(0)}]$, iterate the following two steps until convergence.

E step: use the following equation to predict missing trait data conditional on the observed trait data and variance-covariance matrix.

$$E(\mathbf{y}_i^{miss(k+1)} \,\big|\, \mathbf{y}_i^{obs}, \boldsymbol{\theta}^{(k)}) = \boldsymbol{\mu}_i^{miss(k+1)} + (\mathbf{y}_i^{obs} - \boldsymbol{\mu}_i^{obs(k+1)})(\hat{\mathbf{V}}_i^{obs,miss})^{(k)}(\hat{\mathbf{V}}_i^{obs,obs})^{-1(k)}. \tag{8}$$

Then reconstruct complete phenotypic matrix $\mathbf{y}_i$ with the observed and predicted trait data (k indexes iterations):

$$\mathbf{y}_i^{(k+1)} = (\mathbf{y}_i^{obs}, \mathbf{y}_i^{miss(k+1)}). \tag{9}$$

M step:

$$\hat{\mathbf{B}}^{(k+1)} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}^{(k+1)}, \tag{10}$$

$$\hat{\boldsymbol{\mu}}^{(k+1)} = \mathbf{X}\hat{\mathbf{B}}^{(k+1)}, \tag{11}$$

$$\hat{\mathbf{V}}^{(k+1)} = \frac{(\mathbf{Y}^{(k+1)} - \hat{\boldsymbol{\mu}}^{(k+1)})^{\mathrm{T}}(\mathbf{Y}^{(k+1)} - \hat{\boldsymbol{\mu}}^{(k+1)})}{n}. \tag{12}$$

Multi-trait QTL mapping with incomplete phenotypic data by regression

We now describe our multi-trait QTL mapping method with incomplete data. Though the method given is based on a recombinant inbred line (RIL) population, it is readily extended to other mating designs, as we show for the F_2. According to the statistical model for multiple-trait analysis [9,10,19] based on complete phenotypic data, the model for incomplete phenotypic data is written as

$$\mathbf{Y}_{n \times m} = \mathbf{z}_{n \times 1}\mathbf{a}_{(1 \times m)} + \mathbf{x}_{n \times (p+1)}\mathbf{b}_{(p+1) \times m} + \mathbf{E}_{n \times m} \tag{13}$$

where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2,..., \mathbf{y}_n]'$ is a matrix of phenotypic data for n lines and m traits, and $\mathbf{y}_1, \mathbf{y}_2,...,\mathbf{y}_n$ are $1 \times m$ vectors composed of observations and missing trait data; $\mathbf{z}$ is a matrix of QTL genotypes represented as 2 for *QQ* and 0 for *qq*; $\mathbf{a}$ is a matrix of additive effects of a putative QTL at a tested position; $\mathbf{x}$ is a matrix of genotypes of p cofactor markers with the first column ones; $\mathbf{b}$ is a matrix of cofactor marker effects; and $\mathbf{E}$ is a matrix of residual errors e_{ij} ($i = 1, 2, ..., n$; $j = 1, 2,..., m$), which are assumed to be correlated between traits and to follow a multivariate normal distribution with means zero and covariance matrix as in (2). Equation (13) is readily seen to be a variant of (1).

In this model, QTL genotype is replaced with its conditional expectation given flanking-marker genotypes [1,26]. Least-squares estimates of the parameters can then be obtained by multiple regression based on ALGORITHM 1. If considering a F_2 population, we may instead use the model

$$\mathbf{Y}_{n \times m} = \mathbf{z}_{n \times 1}\mathbf{a}_{1 \times m} + \mathbf{w}_{n \times 1}\mathbf{d}_{1 \times m} + \mathbf{x}_{n \times (p+1)}\mathbf{b}_{(p+1) \times m} + \mathbf{E}_{n \times m}, \tag{14}$$

where $\mathbf{w}$ is a matrix of QTL genotypes represented as 1 for *Qq* and 0 for *qq* and QQ, and $\mathbf{d}$ is a matrix of dominance effects of a putative QTL at a tested position. Since w is unobservable, it is also replaced by its probability conditional on flanking markers.

Multitrait QTL mapping with incomplete phenotypic data by ECM

Instead of replacing a missing QTL genotype with its expectation given flanking markers, ECM (expectation/conditional maximization) treats QTL genotype as missing data included in model (14) and estimates parameters at a QTL position by repeatedly updating the posterior probability of QTL genotype given both flanking-marker genotypes and phenotypes. Since we now have two types of missing data in model (14), QTL genotype and phenotype, we may extend the ECM method of [9] for multi-trait QTL mapping as follows:

ALGORITHM 2: Starting with initial values of parameters

$$\boldsymbol{\theta}^{(0)} = [\mathbf{a}^{(0)}, \mathbf{b}^{(0)}, \boldsymbol{\mu}^{(0)}, \mathbf{V}^{(0)}], \tag{15}$$

iterate the following two steps until convergence (< 10^{-7} change in log likelihood between two iterations).

E step:

$$q_{1i}^{(k+1)} = \frac{p_{1i}f_1^{(k)}(\mathbf{y}_i^{obs} \,|\, \boldsymbol{\mu}_{i,QQ}^{(k)}, \mathbf{V}_i^{obs,obs(k)})}{p_{1i}f_1^{(k)}(\mathbf{y}_i^{obs} \,|\, \hat{\boldsymbol{\mu}}_{i,QQ}^{(k)}, \hat{\mathbf{V}}_i^{obs,obs(k)}) + p_{2i}f_2^{(k)}(\mathbf{y}_i^{obs} \,|\, \hat{\boldsymbol{\mu}}_{i,qq}^{(k)}, \hat{\mathbf{V}}_i^{obs,obs(k)})}, \tag{16}$$

$$q_{2i}^{(k+1)} = \frac{p_{1i}f_2^{(k)}(\mathbf{y}_i^{obs} \,|\, \boldsymbol{\mu}_{i,qq}^{(k)}, \mathbf{V}_i^{obs,obs(k)})}{p_{1i}f_1^{(k)}(\mathbf{y}_i^{obs} \,|\, \hat{\boldsymbol{\mu}}_{i,QQ}^{(k)}, \hat{\mathbf{V}}_i^{obs,obs(k)}) + p_{2i}f_2^{(k)}(\mathbf{y}_i^{obs} \,|\, \hat{\boldsymbol{\mu}}_{i,qq}^{(k)}, \hat{\mathbf{V}}_i^{obs,obs(k)})}, \tag{17}$$

where p_{1i} and p_{2i} are the conditional probabilities of QTL genotypes *QQ* and *qq* given flanking markers and recombination distances [26], f the multivariate normal probability density function, and q_{1i} and q_{2i} the posterior probability of QTL genotypes given flanking markers and phenotypes [9].

$$E(\mathbf{y}_i^{miss(k+1)} \,|\, \mathbf{y}_i^{obs}, \boldsymbol{\theta}^{(k)}) = \boldsymbol{\mu}_{i,E}^{miss(k+1)} + (\mathbf{y}_i^{obs} - \boldsymbol{\mu}_{i,E}^{obs(k+1)})\mathbf{V}_i^{obs,miss(k)}(\mathbf{V}_i^{obs,obs})^{-1(k)}, \tag{18}$$

$$\mathbf{y}_i^{(k+1)} = (\mathbf{y}_i^{obs}, \mathbf{y}_i^{miss(k+1)}). \tag{19}$$

M step:

$$\hat{\mathbf{a}}^{(k+1)} = \frac{0.5\mathbf{q}_2^{(k+1)\mathrm{T}}(\mathbf{Y}^{(k+1)} - \mathbf{x}\hat{\mathbf{b}}^{(k+1)})}{\mathbf{q}_2^{(k+1)\mathrm{T}})\mathbf{l}}, \tag{20}$$

where $\mathbf{l}$ is a ($n \times 1$) matrix of ones.

$$\hat{\mathbf{b}}^{(k+1)} = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}[\mathbf{Y}^{(k+1)} - 2\mathbf{q}_2^{(k+1)}\hat{\mathbf{a}}^{(k+1)}], \tag{21}$$

$$\hat{\boldsymbol{\mu}}_E^{(k+1)} = \mathbf{x}\hat{\mathbf{b}}^{(k+1)} + 2\mathbf{q}_2^{(k+1)}\hat{\mathbf{a}}^{(k+1)}, \tag{22}$$

where $\hat{\boldsymbol{\mu}}_E^{(k+1)}$ is the predicted phenotypic mean given QTL and cofactor markers,

$$\hat{\boldsymbol{\mu}}_{QQ}^{(k+1)} = \hat{\boldsymbol{\mu}}_{E}^{(k+1)}, \tag{23}$$

$$\hat{\boldsymbol{\mu}}_{qq}^{(k+1)} = \mathbf{x}\hat{\mathbf{b}}^{(k+1)} \tag{24}$$

where $\hat{\boldsymbol{\mu}}_{QQ}^{(k+1)}$ and $\hat{\boldsymbol{\mu}}_{qq}^{(k+1)}$ are the predicted phenotypic means of QTL genotypes *QQ* and *qq* given cofactor markers, and

$$\hat{\mathbf{V}}^{(k+1)} = \frac{(\mathbf{Y}^{(k+1)} - \hat{\boldsymbol{\mu}}_{E}^{(k+1)})^{\mathrm{T}}(\mathbf{Y}^{(k+1)} - \hat{\boldsymbol{\mu}}_{E}^{(k+1)})}{n}. \tag{25}$$

For a F_2 population, we need to consider both additive **a** and dominance effects **d** in terms of model (14). In this case, the E-step is as in equations 16–19, and the M-step is as follows:

$$\hat{\mathbf{a}}^{(k+1)} = \frac{0.5\mathbf{q}_2^{(k+1)\mathrm{T}}}{\mathbf{q}_2^{(k+1)\mathrm{T}}\mathbf{l}}(\mathbf{Y}^{(k+1)} - \mathbf{x}\hat{\mathbf{b}}^{(k+1)}), \tag{26}$$

$$\hat{\mathbf{d}}^{(k+1)} = \left(\frac{\mathbf{q}_1^{(k+1)\mathrm{T}}}{\mathbf{q}_1^{(k+1)\mathrm{T}}\mathbf{l}} - \frac{0.5\mathbf{q}_2^{(k+1)\mathrm{T}}}{\mathbf{q}_2^{(k+1)\mathrm{T}}\mathbf{l}}\right)(\mathbf{Y}^{(k+1)} - \mathbf{x}\hat{\mathbf{b}}^{(k+1)}) \tag{27}$$

where **l** is a (n × 1) matrix of ones,

$$\hat{\mathbf{b}}^{(k+1)} = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}[\mathbf{Y}^{(k+1)} - 2\mathbf{q}_2^{(k+1)}\hat{\mathbf{a}}^{(k+1)} - \mathbf{q}_1^{(k+1)}(\hat{\mathbf{a}}^{(k+1)} + \hat{\mathbf{d}}^{(k+1)})], \tag{28}$$

$$\hat{\boldsymbol{\mu}}_{E}^{(k+1)} = \mathbf{x}\hat{\mathbf{b}}^{(k+1)} + 2\mathbf{q}_2^{(k+1)}\hat{\mathbf{a}}^{(k+1)} + \mathbf{q}_1^{(k+1)}(\hat{\mathbf{a}}^{(k+1)} + \hat{\mathbf{d}}^{(k+1)}), \tag{29}$$

where $\hat{\boldsymbol{\mu}}_{E}^{(k+1)}$ is the predicted phenotypic mean given QTL and cofactor markers,

$$\hat{\boldsymbol{\mu}}_{QQ}^{(k+1)} = \mathbf{x}\hat{\mathbf{b}}^{(k+1)} + 2\mathbf{q}_2^{(k+1)}\hat{\mathbf{a}}^{(k+1)}, \tag{30}$$

$$\hat{\boldsymbol{\mu}}_{Qq}^{(k+1)} = \mathbf{x}\hat{\mathbf{b}}^{(k+1)} + \mathbf{q}_1^{(k+1)}(\hat{\mathbf{a}}^{(k+1)} + \hat{\mathbf{d}}^{(k+1)}), \tag{31}$$

$$\hat{\boldsymbol{\mu}}_{qq}^{(k+1)} = \mathbf{x}\hat{\mathbf{b}}^{(k+1)}, \tag{32}$$

where $\hat{\boldsymbol{\mu}}_{QQ}^{(k+1)}$, $\hat{\boldsymbol{\mu}}_{Qq}^{(k+1)}$ and $\hat{\boldsymbol{\mu}}_{qq}^{(k+1)}$ are the predicted phenotypic means of QTL genotypes *QQ*, *Qq*, and *qq* given cofactor markers, and **V** is updated by (25).

In the ECM algorithm, incautious selection of initial parameter values may lead to convergence on local maxima. We used additive and dominance effects of 0, and the estimates of μ and **V** under the null hypothesis H_0: $a = 0$, $d = 0$.

Hypothesis tests for QTL effects with missing phenotypic data

Hypothesis tests for QTL main effects, pleiotropy effects and close linkage vs. pleiotropy are constructed according to [9] and can be tested by ALGORITHM 1 if regression is chosen or ALGORITHM 2 if the ECM method is used. As test statistic the likelihood ratio (LR) or its transformation to a logarithm-of-odds (LOD) are commonly used. For example, to test main QTL effects in a two-trait example, the hypotheses can be formulated as H_0: $a_1 = 0$, $a_2 = 0$ and H_1: at least one a 0. For the regression method, parameters under H_0 or H_1 are estimated by ALGORITHM 1 (Equations 8–12) depending on whether or not QTL effects are included in model (13). If the ECM method is used, first these quantities are estimated under H_0 by ALGORITHM 1 without inclusion of QTL effect and then those of the full model under H_1 are obtained by ALGORITHM 2 (Equations 16–25 for RIL or 16–19, 26–32, and 25 for F_2). Then the LR is obtained as $LR = -2(<_{reduced} - <_{full})$, where $<_{reduced}$ is the log likelihood of the reduced model under H_0, and $<_{full}$ is that of the full model under H_1 [4]. Both are calculated from (4) and a LOD score is calculated as $LR/(2 \ln 10)$.

Simulations

To compare the properties of the EM method with those of casewise deletion (CaD), mean substitution (MS), conditional mean substitution (CMS) and complete data (CoD), we performed simulation experiments. RIL populations of size 100, 200 and 300 were generated based on a 300-cM chromosome with 31 evenly spaced markers. For CMS, missing data were replaced with their conditional expectations calculated by regression of each trait on the other(s). Three pleiotropic QTLs controlling two traits were simulated at cM positions 53, 182, and 258 with effects listed in Table 2.

Trait values of each line were calculated as the sum of QTL effects plus a random vector of environmental effects with means zero and variance given in Table 2. Then a specified proportion (0.05, 0.10, 0.20, or 0.40) of values for each trait independently was set to missing. Lines lacking data for both traits were dropped. Analyses were performed on 500 replicates.

In the QTL analyses, the calculation interval (step size) used was 1 cM. Cofactor markers for each trait were selected by forward stepwise regression at a significance level of 0.01 and combined for multi-trait analysis. Cofac-

Table 2: QTL effects and variances for two traits used for simulation of multi-trait QTL mapping.

Parameter	QTL	Trait	
		1	2
QTL effect	**1**	2.3	2
	2	-2.1	-1.8
	3	1.7	1.6
QTL variance	**1**	5.3 (8.4%)	4.0 (8.2%)
	2	4.4 (7.0%)	3.2 (6.6%)
	3	2. 9 (4.6%)	2.6 (5.2%)
Total genetic variance		12.6 (20%)	9.8 (20%)
Environmental variance		50.0 (80%)	39.2 (80%)
Phenotypic variance		62.6 (100%)	49.0 (100%)

tors lying within 10 cM of a QTL testing position were dropped from the model.

Though the LR at each test position asymptotically follows a chi-square distribution with degrees of freedom determined by the corresponding hypothesis test, an acceptance threshold applying over all test positions must be found. Genomewide LOD thresholds of 3.71, 3.54 and 3.43 for n = 100, 200, and 300 at significance level 0.05 were calculated from 5000 simulations with no missing data under the null hypothesis of no QTL [20]. When sample size or heritability is relatively small, the effect of a QTL may extend to adjacent intervals due to sampling error. Rather than including heritability as an experimental parameter for investigation, we chose the cautious expedient of declaring a QTL if a LOD peak higher than threshold was found within the interval containing the simulated QTL and the intervals on either side of the QTL interval. Power of QTL detection was calculated as the number of correctly declared ("true positive") QTLs divided by the number of actual QTLs simulated, while specificity was calculated as the number of true positive QTLs divided by the total number declared.

Real data analysis

We applied the EM method to a population of 325 doubled-haploid lines (unpublished data) tested for rice sheath-blight disease in field and greenhouse studies and genotyped with 114 codominant markers. The traits we chose for analysis were sheath-blight score and heading date measured in Stuttgart, Arkansas in 2006 and correlated at r = -0.57. The preliminary SIM and CIM analyses indicated that a QTL on chromosome 12 influenced both traits, and suggested a multi-trait analysis.

Since there were no missing trait data, we generated a new sample with 10% randomly missing phenotype scores. Genomewide LOD thresholds 3.73, 3.77, 3.90, 3.80, and 3.83 for EM, CoD, CaD, MS, and CMS at significance level 0.05 were calculated from 1000 permutations, based on shuffling the phenotypic records for both traits at once in order to preserve their correlation structure.

Authors' contributions

ZG conceived the problem, developed the statistical approach, and drafted the manuscript. JCN guided the design and conduct of the experimental study and co-wrote the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

We thank the reviewers for close attention leading to improvement of the article. Support for this work has been provided by National Science Foundation NSF-DBI grant 0109879 and U. S. Department of Agriculture Cooperative State Research, Education and Extension Service – National Research Initiative – Applied Plant Genomics Program grant 2004-35317-14867 entitled "RiceCAP: A coordinated research, education, and extension project for the application of genomic discoveries to improve rice in the United States". This is contribution KAES 08-378-J from the Kansas Agriculture Experiment Station.

References

1. Haley CS, Knott SA: **A simple regression method for mapping quantitative trait loci in line crosses using flanking markers.** *Heredity* 1992, **69:**315-324.
2. Jansen RC: **Interval mapping of multiple quantitative trait loci.** *Genetics* 1993, **135:**205-211.
3. Kao C-H, Zeng Z-B, Teasdale RD: **Multiple interval mapping for quantitative trait loci.** *Genetics* 1999, **152:**1203-1216.
4. Lander ES, Botstein D: **Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps.** *Genetics* 1989, **121:**185-199.
5. Satagopan JM, Yandell BS, Newton MA, Osborn TC: **A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo.** *Genetics* 1996, **144:**805-816.
6. Wang H, Zhang Y-M, Li X, Masinde GL, Mohan S, Baylink DJ, Xu S: **Bayesian shrinkage estimation of quantitative trait loci parameters.** *Genetics* 2005, **170:**465-480.
7. Yi N, Xu S: **Bayesian mapping of quantitative trait loci under complicated mating designs.** *Genetics* 2001, **157:**1759-1771.
8. Zeng Z-B: **Precision mapping of quantitative trait loci.** *Genetics* 1994, **136:**1457-1468.
9. Jiang C, Zeng Z-B: **Multiple trait analysis of genetic mapping for quantitative trait loci.** *Genetics* 1995, **140:**1111-1127.
10. Korol AB, Ronin YI, Kirzhner VM: **Interval mapping of quantitative trait loci employing correlated trait complexes.** *Genetics* 1995, **140:**1137-1147.
11. Ma CX, Casella G, Wu RL: **Functional mapping of quantitative trait loci underlying the character process: A theoretical framework.** *Genetics* 2002, **161:**1751-1762.
12. Allison DB, Thiel B, St Jean P, Elston RC, Infante MC, Schork NJ: **Multiple phenotype modeling in gene-mapping studies of quantitative traits: Power advantages.** *Am J Hum Genet* 1998, **63:**1190-1201.
13. Evans DM: **The power of multivariate quantitative-trait loci linkage analysis is influenced by the correlation between variables.** *Am J Hum Genet* 2002, **70(6):**1599-1602.
14. Kučerová J, Lund MS, Sorensen P, Sahana G, Guldbrandtsen B, Nielsen VH, Thomsen B, Bendixen C: **Multitrait quantitative trait loci mapping for milk production traits in Danish Holstein cattle.** *J Dairy Sci* 2006, **89(6):**2245-2256.
15. Neuschl C, Brockmann GA, Knott SA: **Multiple-trait QTL mapping for body and organ weights in a cross between NMRI8 and DBA/2 mice.** *Genet Res* 2007, **89(1):**47-59.
16. Thomasen J: **Quantitative trait loci affecting calving traits in Danish Holstein cattle.** *J Dairy Sci* 2008, **91(5):**2098-2105.

17. Malosetti M: **A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (*Zea mays* L.).** *Euphytica* 2008, **161(1–2):**241-257.
18. Calinski T, Kaczmarek Z, Krajewski P, Frova C, Sari-Gorla M: **A multivariate approach to the problem of QTL localization.** *Heredity* 2000, **84(3):**303-310.
19. Hackett CA, Meyer RC, Thomas WTB: **Multi-trait QTL mapping in barley using multivariate regression.** *Genet Res* 2001, **77:**95-106.
20. Knott SA, Haley CS: **Multitrait least squares for quantitative trait loci detection.** *Genetics* 2000, **156:**899-911.
21. Korol AB, Ronin YI, Nevo E, Hayes PM: **Multi-interval mapping of correlated trait complexes: simulation analysis and evidence from barley.** *Heredity* 1998, **80:**273-284.
22. Xu S: **A comment on the simple regression method for interval mapping.** *Genetics* 1995, **141:**1657-1659.
23. Kao C-H: **On the differences between maximum likelihood and regression interval mapping in the analyis of quantitative trait loci.** *Genetics* 2000, **156:**855-865.
24. Mangin B, Thoquet P, Grimsley N: **Pleiotropic QTL analysis.** *Biometrics* 1998, **54:**88-99.
25. Weller JI, Wiggans GR, VanRaden PM, Ron M: **Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment.** *Theor Appl Genet* 1996, **92:**998-1002.
26. Jiang C, Zeng Z-B: **Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines.** *Genetica* 1997, **101:**47-58.
27. Sen S, Churchill GA: **A statistical framework for quantitative trait mapping.** *Genetics* 2001, **159:**371-387.
28. Sillanpaa MJ, Arjas E: **Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data.** *Genetics* 1999, **151(4):**1605-1619.
29. Sillanpää MJ, Arjas E: **Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data.** *Genetics* 1998, **148:**1373-1388.
30. Allison PD: **Missing Data.** Thousand Oaks, Calif.: Sage Publications; 2002.
31. Rubin DB: **Multiple Imputation for Nonresponse in Surveys.** New York: Wiley; 1987.
32. Rubin DB: **Multiple imputation after 18+ years.** *Journal of the American Statistical Association* 1996, **91:**473-489.
33. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm.** *J Royal Stat Soc B* 1977, **39(1):**1-38.
34. Niu T, Ding AA, Kreutz R, Lindpaintner K: **An expectation-maximization-likelihood-ratio test for handling missing data: application in experimental crosses.** *Genetics* 2005, **169(2):**1021-1031.
35. Little RJA, Rubin DB: **Statistical Analysis with Missing Data.** Hoboken, New Jersey: John Wiley & Sons; 2001.
36. Liu JF, Liu YJ, Liu XG, Deng HW: **Bayesian mapping of quantitative trait loci for multiple complex traits with the use of variance components.** *Am J Hum Genet* 2007, **81(2):**304-320.
37. Rubin DB: **Inference and missing data.** *Biometrika* 1976, **63:**581-592.
38. Piepho HP: **A quick method for computing approximate thresholds for quantitative trait loci detection.** *Genetics* 2001, **157:**425-432.
39. Piepho H-P: **A mixed-model approach to mapping quantitative trait loci in barley on the basis of multiple environment data.** *Genetics* 2000, **156:**2043-2050.
40. Wang DL, Zhu J, Li ZK, Paterson AH: **Mapping QTLs with epistatic effects and QTL × environment interactions by mixed linear model approaches.** *Theor Appl Genet* 1999, **99:**1255-1264.