Heredity **84** (2000) 525–537 *Received 22 June 1999, accepted 13 October 1999*

Maximum likelihood analysis of quantitative trait loci under selective genotyping

SHIZHONG XU*† & CLAUS VOGL‡§

†*Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, U.S.A. and* ‡*Department of Biology, University of Oulu, PO Box 3000, FIN-90401 Oulu, Finland*

Selective genotyping is a cost-saving strategy in mapping quantitative trait loci (QTLs). When the proportion of individuals selected for genotyping is low, the majority of the individuals are not genotyped, but their phenotypic values, if available, are still included in the data analysis to correct the bias in parameter estimation. These ungenotyped individuals do not contribute much information about linkage analysis and their inclusion can substantially increase the computational burden. For multiple trait analysis, ungenotyped individuals may not have a full array of phenotypic measurements. In this case, unbiased estimation of QTL effects using current methods seems to be impossible. In this study, we develop a maximum likelihood method of QTL mapping under selective genotyping using only the phenotypic values of genotyped individuals. Compared with the full data analysis (using all phenotypic values), the proposed method performs well. We derive an expectation–maximization (EM) algorithm that appears to be a simple modification of the existing EM algorithm for standard interval mapping. The new method can be readily incorporated into a standard QTL mapping software, e.g. MAPMAKER. A general recommendation is that whenever full data analysis is possible, the full maximum likelihood analysis should be performed. If it is impossible to analyse the full data, e.g. sample sizes are too large, phenotypic values of ungenotyped individuals are missing or composite interval mapping is to be performed, the proposed method can be applied.

Keywords: EM algorithm, QTL mapping, simplex algorithm, truncated selection.

Introduction

Statistical analysis of quantitative trait loci requires both the phenotypic data and marker genotypes of individuals sampled from a reference population. It is generally believed that a large sample size is required to map QTLs with small effects. However, obtaining a large sample can be very costly or even impossible. Usually, the cost of genotyping is higher than that of the phenotypic measurement. Lander & Botstein (1989) showed that one can selectively genotype individuals from the extremes of the phenotypic distribution, yet receive almost identical power as when the whole sample is genotyped. If the cost of the phenotypic measurement is low, selective genotyping can significantly reduce the cost. This selective genotyping technique has been widely utilized in QTL mapping experiments (e.g. Groover *et al.*, 1994).

Under selective genotyping, phenotypic values of ungenotyped individuals still have to be included in the analysis, with their marker genotypes treated as missing values, otherwise estimates of QTL effects will be biased (Lander & Botstein, 1989). A full likelihood function is given by Muranty & Goffinet (1997a). Exact maximum likelihood estimates (MLEs) can be achieved via an iterative approach. However, Muranty & Goffinet (1997a) derive approximate MLEs under the assumption that QTL effects are small relative to the residual standard deviation. Recently, Johnson *et al.* (1999) proposed an expectation–maximization (EM) algorithm implemented via Monte Carlo sampling for handling missing marker genotypes.

With phenotypic values of ungenotyped individuals excluded from the data analysis, Darvasi & Soller (1992) investigated an analysis of variance (ANOVA) approach

*Correspondence. E-mail: xu@genetics.ucr.edu
§Present address: Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, U.S.A.

to estimate QTL effects with a bias correction. Without providing a detailed implementation, they also proposed a maximum likelihood approach for such a truncated data analysis.

Muranty & Goffinet (1997b) extended their selective genotyping to multiple trait QTL mapping, showing that selection on one trait can increase the power of QTL detection for a correlated trait. They also proposed a selection index method for multiple trait selective genotyping. Instead of selecting the two tails of a single trait, they first established a selection index combining phenotypic values of all traits, and then selected the two extremes on the scale of the index. Again, phenotypic values of ungenotyped individuals must be included in the data analysis to remove the bias in the estimated QTL effect. In reality, different traits may be expressed in different stages. If selection is performed on an earlier displayed trait, individuals that fail to reach the criterion of selection in this stage may be removed, and thus do not have the opportunity to express their phenotype for a later trait. In this case, unbiased estimates of QTL effects for the later trait seem to be impossible based on the method of Muranty & Goffinet (1997a,b). Therefore, a new method is needed to handle missing values for both the genotypes and the phenotypes.

Such a method is now available as a result of work by Henshall & Goddard (1999). They adopted an entirely different approach by altering the roles of genotypes and phenotypes in the likelihood function. They treated phenotypes as independent variables and genotypes as dependent variables. Because genotypes are binary in a population with only two genotypes, they utilized a standard logistic regression approach. The advantages of this method are: a standard statistical package, such as SAS, is readily applied and estimates are not affected by selection of the phenotype. The second advantage is important in handling the problem of missing phenotypes. As recognized by the authors, the logistic regression method, however, has not been sufficiently generalized to handle populations with more than two segregating genotypes, e.g. the F_2 family. Furthermore, it is not clear how to implement the composite interval mapping (Jansen & Stam, 1994; Zeng, 1994) in the logistic regression framework.

The objectives of this study are to develop a maximum likelihood approach to QTL mapping using samples containing only the genotyped individuals and to compare its efficiency relative to that using samples of all individuals. The maximum likelihood solution will be achieved via an EM algorithm that is simple enough to be incorporated into any standard interval mapping software.

Theory and methods

Single-trait analysis

Consider an F_2 population of N individuals with phenotypic values measured for trait y. Among the N individuals, only n ($n \le N$) of them are selectively genotyped, with $n/2$ being selected from the upper extreme and $n/2$ being selected from the lower extreme in the scale of y. This selection regime can be viewed as 'disruptive selection' with two known artificial truncation points. Individuals are genotyped only if $y_j \le t_1$ or $y_j \ge t_2$, where t_1 and t_2 are, respectively, the $(n/2)$th and $(N - n/2)$th ascendingly ordered phenotypic values of y among the N individuals. In real data analyses, the two tails selected for genotyping may not be symmetrical. The two truncation points, t_1 and t_2, are not calculated from the distribution; rather, they take the largest phenotypic value in the lower tail and the smallest phenotypic value in the upper tail.

The phenotypic value of the jth individual is described by the following linear model:

$$y_j = \mu + z_j a + w_j d + e_j, \tag{1}$$

where a and d are the additive and dominance effects of a QTL, respectively, z_j and w_j are indicator variables for the genotype of the QTL, which are defined as:

$$z_j = \begin{cases} +1 \text{ for } Q_1Q_1 \\ 0 \text{ for } Q_1Q_2 \\ 1 \text{ for } Q_2Q_2 \end{cases} \text{ and } w_j = \begin{cases} +1 \text{ for } Q_1Q_2 \\ 1 \text{ for } Q_1Q_1 \text{ or } Q_2Q_2 \end{cases}$$

where Q_kQ_l for $k \le l = 1,2$ denotes the QTL genotype, and e_j is the residual effect distributed as $N(0, \sigma^2)$.

For notational simplicity, let us define $\mathbf{b} = [\mu, a, d]^T$ and $\mathbf{x}_j = [1, z_j, w_j]$, and rewrite model (1) by:

$$y_j = \mathbf{x}_j\mathbf{b} + e_j, \tag{2}$$

where $\mathbf{x}_j = \mathbf{u}_{kl}$ for genotype Q_kQ_l and

$$\begin{bmatrix} \mathbf{u}_{11} \\ \mathbf{u}_{12} \\ \mathbf{u}_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

This is a typical regression model. Because the QTL genotype is not observable but inferred from marker information, only the conditional distribution of $\mathbf{x}_j$ given the marker genotype is available. Define the conditional probabilities of the QTL genotype, and thus $\mathbf{x}_j$, by $p(\mathbf{x}_j)$ and the probability density of y_j given $\mathbf{x}_j$ by:

$$f(y_j|\mathbf{x}_j) = [1/\sqrt{(2\pi\sigma^2)}]\exp\{ \quad [1/2\sigma^2](y_j \quad \mathbf{x}_j\mathbf{b})^2\}.$$

The likelihood function is rewritten as:

$$\begin{aligned} L(\mathbf{b},\sigma^2|\mathbf{y}) &= \prod_{j=1}^{N}\left[\sum_{x_j\in S} p(\mathbf{x}_j) f(y_j|\mathbf{x}_j)\right] \\ &= \prod_{j=1}^{n}\left[\sum_{x_j\in S} p(\mathbf{x}_j) f(y_j|\mathbf{x}_j)\right] \\ &\quad \times \prod_{j=n+1}^{N}[(1/4)f(y_j|\mathbf{u}_{11}) + (1/2)f(y_j|\mathbf{u}_{12}) \\ &\quad + (1/4)f(y_j|\mathbf{u}_{22})], \end{aligned} \tag{3}$$

where $S = \{\mathbf{u}_{11}, \mathbf{u}_{12}, \mathbf{u}_{22}\}$. Note that the first n individuals are genotyped and the last $N - n$ individuals are ungenotyped, and $p(\mathbf{x}_j)$ takes its prior value for $j = n + 1, \ldots, N$.

The phenotypic values of ungenotyped individuals contribute very little information to linkage analysis. Their inclusion in the likelihood serves solely as a way to correct the bias in estimation of the QTL effects caused by selective genotyping. When the number of ungenotyed individuals becomes very large, the maximum likelihood method implemented this way is computationally inefficient. An alternative way to construct the likelihood function is to include only the phenotypic values of the genotyped individuals while still taking into account the selection process. The likelihood function then becomes:

$$L(\mathbf{b},\sigma^2|\mathbf{y}) = \prod_{j=1}^{n}\left[\sum_{x_j\in S} p(\mathbf{x}_j) g(y_j|\mathbf{x}_j)\right], \tag{4}$$

where:

$$g(y_j|\mathbf{x}_j) = f(y_j|\mathbf{x}_j)/\{\Phi(\tau_1|\mathbf{x}_j) + [1 \quad \Phi(\tau_2|\mathbf{x}_j)]\}. \tag{5}$$

The denominator of $g(y_j|\mathbf{x}_j)$ is $\Pr[(y_j \le t_1) \cup (y_j \ge t_2)|\mathbf{x}_j]$, the conditional probability that the jth individual is selected for genotyping given $\mathbf{x}_j$, where $\Phi(\tau_1|\mathbf{x}_j) = \Pr(y_j \le t_1|\mathbf{x}_j)$ and $\Phi(\tau_2|\mathbf{x}_j) = \Pr(y_j \le t_2|\mathbf{x}_j)$ are standardized normal functions, and $\tau_1|\mathbf{x}_j = (t_1 - \mathbf{x}_j\mathbf{b})/\sigma$ and $\tau_2|\mathbf{x}_j = (t_2 - \mathbf{x}_j\mathbf{b})/\sigma$ are the standardized truncation points.

The likelihood function can be searched via an EM algorithm that is described below. In the E step, the conditional posterior distribution of $\mathbf{x}_j$ is obtained using initial values of $\mathbf{b}$ and σ^2,

$$p^*(\mathbf{x}_j) = [p(\mathbf{x}_j)g(y_j|\mathbf{x}_j)] \Big/ \left[\sum_{x_j\in S} p(\mathbf{x}_j)g(y_j|\mathbf{x}_j)\right]. \tag{6}$$

The posterior distribution is then used to calculate the expectations of various quantities that involve $\mathbf{x}_j$. In the M step, we estimate the parameters based on the following equations:

$$\begin{aligned} \hat{\mathbf{b}} = &\left[\sum_{j=1}^{n} E_x\left(\mathbf{x}_j^T\mathbf{x}_j\right)\right]^{1} \\ &\times \left[\sum_{j=1}^{n} E_x\left\{\mathbf{x}_j^T\left[y_j + \sigma\frac{\phi\ \tau_1|\mathbf{x}_j) \quad \phi\ \tau_2|\mathbf{x}_j)}{1+\Phi\ \tau_1|\mathbf{x}_j) \quad \Phi\ \tau_2|\mathbf{x}_j)}\right]\right\}\right] \end{aligned} \tag{7}$$

and

$$\begin{aligned} \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n} E_x\Bigg[&(y_j \quad \mathbf{x}_j\mathbf{b})^2 \\ &+ \sigma^2\frac{\tau_1|\mathbf{x}_j)\phi\ \tau_1|\mathbf{x}_j) \quad \tau_2|\mathbf{x}_j)\phi\ \tau_2|\mathbf{x}_j)}{1+\Phi\ \tau_1|\mathbf{x}_j) \quad \Phi\ \tau_2|\mathbf{x}_j)}\Bigg], \end{aligned} \tag{8}$$

where $\phi(\tau_1|\mathbf{x}_j)$ and $\phi(\tau_2|\mathbf{x}_j)$ denote the standardized normal densities, different from $\Phi(\tau_1|\mathbf{x}_j)$ and $\Phi(\tau_2|\mathbf{x}_j)$. The notation E_x stands for expectation with respect to $\mathbf{x}_j$, the missing genotype. The initial values of parameters are then replaced by $\hat{\mathbf{b}}$ and $\hat{\sigma}^2$, forming a new cycle of iteration. After convergence, $\hat{\mathbf{b}}$ and $\hat{\sigma}^2$ will be the MLEs of $\mathbf{b}$ and σ^2. Note that the terms involving σ and σ^2 in the right hand sides of eqns (7) and (8) are because of selective genotyping. Without selection, these terms will vanish and the EM equations will reduce to the standard ones (Zeng, 1994). In the simple model described in this study, only one non-QTL effect, μ, is included in the model. If the model includes many covariates, as seen in composite interval mapping, the ECM approach should be adopted (Jiang & Zeng, 1995). The derivations of $\hat{\mathbf{b}}$ and $\hat{\sigma}^2$ are given in the Appendix.

Multiple-trait analysis

Let us define a $1 \times m$ vector for m traits measured in the jth individual by $\mathbf{y}_j = [y_{j1}, y_{j2}, \ldots, y_{jm}]$. The multivariate linear model is expressed by:

$$\mathbf{y}_j = \mathbf{x}_j\mathbf{B} + \mathbf{e}_j, \tag{9}$$

where $\mathbf{x}_j = [1, z_j, w_j]$ remains the same as in the single-trait model,

$$\mathbf{B} = [\mathbf{B}_1\ \mathbf{B}_2\ \cdots\ \mathbf{B}_m] = \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_m \\ a_1 & a_2 & \cdots & a_m \\ d_1 & d_2 & \cdots & d_m \end{bmatrix}$$

and

$$\mathbf{e}_j = [e_{j1}, e_{j2}, \ldots, e_{jm}]$$

is a $1 \times m$ vector for the residuals with a multivariate normal distribution, i.e. $\mathbf{e}_j \approx \mathrm{N}_m(\mathbf{0}, \mathbf{V})$, where:

$$\mathbf{V} = \mathrm{Var}(\mathbf{e}_j) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2 \end{bmatrix}.$$

Assume that the criterion of selection is a linear combination of all traits, called the selection index and denoted by $I_j = \sum_{k=1}^{m} c_k y_{kj} = \mathbf{y}_j \mathbf{c}$. The selection index is a generalized criterion of selection. If $c_1 = 1$ and $c_{j\neq 1} = 0$, then the index becomes the phenotypic value of the first trait. The score of the selection index can be similarly partitioned into a genetic and residual component:

$$I_j = \mathbf{x}_j \mathbf{b}_I + e_{Ij}, \tag{10}$$

where $\mathbf{b}_I = \sum_{k=1}^{m} c_k \mathbf{B}_k = \mathbf{Bc}$ and $e_{jI} = \sum_{k=1}^{m} c_k \mathrm{e}_{jk} = \mathbf{e}_j \mathbf{c}$. The expectation and variance of I_j are $\mathrm{E}(I_j|\mathbf{x}_j) = \mathbf{x}_j \mathbf{Bc}$ and $\mathrm{Var}(I_j|\mathbf{x}_j) = \sigma^2{}_I = \mathbf{c}^{\mathrm{T}} \mathrm{Var}(\mathbf{e}_j)\mathbf{c} = \mathbf{c}^{\mathrm{T}} \mathbf{V} \mathbf{c}$, respectively. The two truncation points of selection in the scale of the index are defined, again, by t_1 and t_2, respectively. The probability density of $\mathbf{y}_j$ without selection is:

$$f(\mathbf{y}_j|\mathbf{x}_j) = \{1/[(2\pi)^{m/2}|\mathbf{V}|^{1/2}]\} \times \exp[\ (1/2)(\mathbf{y}_j \quad \mathbf{x}_j\mathbf{B})\mathbf{V}^{\ 1}(\mathbf{y}_j \quad \mathbf{x}_j\mathbf{B})^{\mathrm{T}}]. \tag{11}$$

After truncation selection on index I_j, the joint density becomes $g(\mathbf{y}_j|\mathbf{x}_j) = [f(\mathbf{y}_j|\mathbf{x}_j)]/[1 + \Phi(\tau_2|\mathbf{x}_j) - \Phi(\tau_2|\mathbf{x}_j)$, where the denominator is the probability that the jth individual is selected for genotyping, i.e. $1 + \Phi(\tau_1|\mathbf{x}_j) - \Phi(\tau_2|\mathbf{x}_j) = 1 - \iint_{t_1 < y_{jc} < t_2} f(\mathbf{y}_j|\mathbf{x}_j)\mathrm{d}\mathbf{y}_j = 1 - \int_{t_1}^{t_2} f(I_j|\mathbf{x}_j)\mathrm{d}I_j$, where $\tau_1|\mathbf{x}_j = (t_1 - \mathbf{x}_j\mathbf{Bc})/\sigma_I$ and $\tau_2|\mathbf{x}_j = (t_1 - \mathbf{x}_j\mathbf{Bc})/\sigma_I$ are the standardized truncation points in the scale of the index.

The likelihood function appears the same as eqn (4). Again, the MLEs can be obtained by using an EM algorithm, which requires first calculating the posterior distribution of $\mathbf{x}_j$ and then maximizing the expectation of the log likelihood. The EM equations are given as follows:

$$\hat{\mathbf{B}} = \left[\sum_{j=1}^{n} \mathrm{E}_x\left(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j\right)\right]^{\ 1} \left\{\sum_{j=1}^{n} \mathrm{E}_x\left[\mathbf{x}_j^{\mathrm{T}}\left(\mathbf{y}_j + \boldsymbol{\beta}\sigma_I \frac{\phi\ \ \tau_1|\mathbf{x}_j) \quad \phi\ \ \tau_2|\mathbf{x}_j)}{1 + \Phi\ \ \tau_1|\mathbf{x}_j) \quad \Phi\ \ \tau_2|\mathbf{x}_j)}\right)\right]\right\} \tag{12}$$

and

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{E}_x\left[(\mathbf{y}_j \quad \mathbf{x}_j\mathbf{B})(\mathbf{y}_j \quad \mathbf{x}_j\mathbf{B})^{\mathrm{T}} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\sigma_I^2 \frac{\tau_1|\mathbf{x}_j)\phi\ \ \tau_1|\mathbf{x}_j) \quad \tau_2|\mathbf{x}_j)\phi\ \ \tau_2|\mathbf{x}_j)}{1 + \Phi\ \ \tau_1|\mathbf{x}_j) \quad \Phi\ \ \tau_2|\mathbf{x}_j)}\right], \tag{13}$$

where $\boldsymbol{\beta} = (1/\sigma_I^2)[\sigma_{1I}, \sigma_{2I}, \ldots, \sigma_{mI}]$, is a $1 \times m$ vector for the simple regression coefficients of the traits on the index. Note that the multivariate EM equations are simple extensions of the univariate EM by multiplying $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$ by the appropriate terms in eqns (7) and (8). Again, eqns (12) and (13) will reduce to the standard ones (Jiang & Zeng, 1995) under random selection.

Statistical power under selective genotyping

It is difficult to evaluate the power of QTL mapping when a genome-wide chromosomal scanning is performed because the distribution of the test statistic under either hypothesis (null or alternative) is unknown. The usual practice is to evaluate the power under the assumption that the position of the QTL is known so that only point-wise test statistics are considered (Muranty, 1996). The distribution of a point-wise test statistic is usually known, at least asymptotically. Although the power calculated this way cannot be applied to a whole genome-wide analysis, it may be used to compare relative efficiencies of different methods. It is certainly appropriate to use this power to evaluate mapping procedures under the candidate gene approach. Theoretical work has been conducted for systems with two contrasting genotypes in the segregating population, e.g. backcrosses or half-sibs (Darvasi & Soller, 1992). In this study, we evaluate the statistical power of QTL detection for a single-trait model in systems with three possible genotypes, e.g. F_2 families, under the assumption that the trait is controlled by a single QTL whose genotype is observable. Throughout the discussion, we will emphasize the difference in power between QTL detection with and without selective genotyping.

Power calculation without selective genotyping has been extensively investigated by researchers (e.g. Soller & Brody, 1976 and Muranty, 1996). Denote the general linear model in matrix notation by:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}. \tag{14}$$

The null hypothesis is $\mathrm{H}_0 : a = d = 0$, which is expressed in matrix notation by $\mathrm{H}_0 : \mathbf{Kb} = \mathbf{0}$, where:

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

The generalized likelihood ratio test statistic (Graybill, 1976) for testing this hypothesis is:

$$\lambda = \left(\frac{n \quad p}{q}\right)\left(\frac{\hat{\sigma}^2_{\Omega_0} \quad \hat{\sigma}^2_{\Omega_1}}{\hat{\sigma}^2_{\Omega_1}}\right), \tag{15}$$

where $q = 2$ is the rank of $\mathbf{K}$, $p = 3$ is the number of parameters in the full model, $\hat{\sigma}^2_{\Omega_1}$ and $\hat{\sigma}^2_{\Omega_0}$ are the residual variances estimated from the full model and the reduced model ($\mathbf{Kb} = \mathbf{0}$), respectively. Graybill (1976) showed that λ follows a noncentral F distribution denoted by $F(\lambda : q, n - p, \delta)$, where δ is the noncentrality parameter given by:

$$\delta = \{[\mathbf{Kb}]^{\mathrm{T}}[\mathbf{K}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{K}^{\mathrm{T}}]^{-1}[\mathbf{Kb}]\}/2\sigma^2. \tag{16}$$

Muranty (1996) called λ the F-test statistic because of the nature of F distribution. In genetic studies, a different sample will involve a different $\mathbf{X}$ because a completely different segregation process will occur for a different experiment. When the sample size is not too small, however, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ will be fairly constant from sample to sample. Therefore, we can substitute $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ by its expectation. Defining

$$\mathrm{E}\begin{bmatrix} z_j \\ w_j \end{bmatrix} = \begin{bmatrix} M_z \\ M_w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and

$$\mathrm{Var}\begin{bmatrix} z_j \\ w_j \end{bmatrix} = \begin{bmatrix} V_z & V_{zw} \\ V_{zw} & V_w \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1 \end{bmatrix},$$

we have

$$\mathrm{E}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = n\begin{bmatrix} 1 & M_z & M_w \\ M_z & V_z + M_z^2 & V_{zw} + M_z M_w \\ M_w & V_{zw} + M_z M_w & V_w + M_w^2 \end{bmatrix} = n\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Substituting $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ by $\mathrm{E}(\mathbf{X}^{\mathrm{T}}\mathbf{X})$ and after some algebraic manipulation, we get:

$$\delta = \frac{n(a^2/2 + d^2)}{2\sigma^2} = \frac{n\sigma^2_{\mathrm{G}}}{2\sigma^2}, \tag{17}$$

where $\sigma^2_{\mathrm{G}} = a^2/2 + d^2$ is the total genetic variance. The statistical power is then calculated as:

$$\Psi = \int_{F^{-1}(1-\alpha:q,n-p,0)}^{\infty} F(\lambda : q, n - p, \delta)\mathrm{d}\lambda, \tag{18}$$

where $1 - \Psi$ is the Type II error and $F^{-1}(1 - \alpha : q, n - p, 0)$ is the critical value for testing H_0 at a Type I error rate of α.

Under selective genotyping, the exact form of the distribution of the likelihood-ratio test statistic is unknown. To derive the power under selective genotyping, we must assume that the test statistic still follows a noncentral F distribution but with a different noncentrality parameter. This approximation is valid when the selection intensity is weak or the QTL in question has a small effect. In fact, Darvasi & Soller (1992) have already made this approximation when calculating the number of genotyped individuals required to achieve a given power under an additive effect model in a backcross design. Selective genotyping will change the conditional distribution of y_j given its genotype and the frequencies of the three genotypes in the mapping population. These changes will eventually modify a, d, σ^2 and $\mathrm{E}(\mathbf{X}^{\mathrm{T}}\mathbf{X})$, leading to an increase in the noncentrality parameter and thus an increase in the power.

Let us denote the phenotypic value in the selected population by y^*. Using the theory of truncated selection (Cohen, 1991), we found that the conditional expectation and variance of y^* given genotype $Q_k Q_l$ are:

$$\mathrm{E}(y^*|Q_kQ_l) = \mathbf{u}_{kl}\mathbf{b} + \sigma[\phi(\tau_2|\mathbf{u}_{kl}) \quad \phi(\tau_1|\mathbf{u}_{kl})]/[1 + \Phi(\tau_1|\mathbf{u}_{kl}) \quad \Phi(\tau_2|\mathbf{u}_{kl})]$$

and

$$\mathrm{Var}(y^*|Q_kQ_l) = \sigma^2\left[1 \quad \frac{(\tau_1|\mathbf{u}_{kl})\phi(\tau_1|\mathbf{u}_{kl}) \quad (\tau_2|\mathbf{u}_{kl})\phi(\tau_2|\mathbf{u}_{kl})}{1 + \Phi(\tau_1|\mathbf{u}_{kl}) \quad \Phi(\tau_2|\mathbf{u}_{kl})}\right] \sigma^2\left[\frac{\phi(\tau_1|\mathbf{u}_{kl}) \quad \phi(\tau_2|\mathbf{u}_{kl})}{1 + \Phi(\tau_1|\mathbf{u}_{kl}) \quad \Phi(\tau_2|\mathbf{u}_{kl})}\right]^2,$$

respectively, where $\tau_1|\mathbf{u}_{kl} = (t_1 - \mathbf{u}_{kl}\mathbf{b})/\sigma$ for $k \le l = 1,2$. Let us now define the probability that an individual with genotype $Q_k Q_l$ is selected for genotyping by $q_{kl} = 1 + \Phi(\tau_1|\mathbf{u}_{kl}) - \Phi(\tau_2|\mathbf{u}_{kl})$. According to Bayes' theorem, the frequency of genotype $Q_k Q_l$ in the selected population can be defined by $p_{kl} = [(1 + |k - l|)q_{kl}]/(q_{11} + 2q_{12} + q_{22})$ for $k \le l = 1, 2$.

The modified additive and dominance effects after selective genotyping become:

$$a^* = \mathrm{E}(y^*|Q_1Q_1) \quad [p_{11}\mathrm{E}(y^*|Q_1Q_1) + p_{22}\mathrm{E}(y^*|Q_2Q_2)]/[p_{11}+p_{22}] \tag{19}$$

and

$$d^* = \frac{1}{2}\left[\mathrm{E}(y^*|Q_1Q_2) \quad \frac{p_{11}\mathrm{E}(y^*|Q_1Q_1)+p_{22}\mathrm{E}(y^*|Q_2Q_2)}{p_{11}+p_{22}}\right], \tag{20}$$

respectively. The altered residual variance takes the weighted average of the within-genotype residual variances, i.e.

$$\sigma_*^2 = \sum_{k\le l}^{2} p_{kl}\mathrm{Var}(y^*|Q_kQ_l). \tag{21}$$

The modified $\mathrm{E}(\mathbf{X}^\mathrm{T}\mathbf{X})$ resulting from selective genotyping is:

$$\mathrm{E}^*(\mathbf{X}^\mathrm{T}\mathbf{X}) = n\begin{bmatrix} 1 & M_z^* & M_w^* \\ M_z^* & V_z^*+(M_z^*)^2 & V_{zw}^*+M_z^*M_w^* \\ M_w^* & V_{zw}^*+M_z^*M_w^* & V_w^*+(M_w^*)^2 \end{bmatrix}.$$

where

$$\begin{bmatrix} M_z^* \\ M_w^* \end{bmatrix} = \begin{bmatrix} p_{11} & p_{22} \\ p_{12} & (p_{11}+p_{22}) \end{bmatrix}$$

and

$$\begin{bmatrix} V_z^* & V_{zw}^* \\ V_{zw}^* & V_w^* \end{bmatrix} = \begin{bmatrix} (p_{11}+p_{22}) & (p_{11} & p_{22})^2 & p_{12}(p_{22} & p_{11}) \\ & p_{12}(p_{22} & p_{11}) & & 4p_{12}(1 & p_{12}) \end{bmatrix}.$$

Denote $\mathbf{b}^* = [\mu^* \; a^* \; d^*]^\mathrm{T}$ as the vector of parameters after selection, then the noncentrality parameter under selective genotyping is

$$\delta^* = \frac{(\mathbf{K}\mathbf{b}^*)^\mathrm{T}\{\mathbf{K}[\mathrm{E}^*(\mathbf{X}^\mathrm{T}\mathbf{X})]\ ^1\mathbf{K}^\mathrm{T}\}\ ^1(\mathbf{K}\mathbf{b}^*)}{2\sigma_*^2}. \tag{22}$$

Subsequently, the statistical power under selective genotyping is calculated using eqn (18) but with the noncentrality parameter replaced by δ^*.

Illustration

In this section we demonstrate the application of the method using simulated data and show the general behaviour of the method that one expects to observe in QTL mapping experiments.

Single-trait QTL mapping

In the first simulation study, we assumed that a single QTL is located at position 25 cM of a 100-cM chromosome segment covered by 11 evenly spaced codominant markers. The size of the QTL (measured by the percentage of phenotypic variance explained by the QTL) is 0.05. The actual genetic effects that generate such a QTL are $a = 0.229$ and $d = 0.162$. In an F_2 population, these genetic effects will make up a genetic variance of $\sigma_G^2 = a^2/2 + d^2 = 0.0525$. The residual variance was set at $\sigma^2 = 1.0$, leading to $h^2 = 0.0525/(0.0525 + 1.0) = 0.05$. The number of individuals genotyped was fixed at 100. We then varied the total number of individuals measured for the phenotype to control different levels of selection pressure. We set up four levels of proportion genotyped: 100%, 50%, 10% and 5%. The total numbers of phenotypically measured individuals corresponding to the four proportions were 100, 200, 1000 and 2000, respectively. Under selective genotyping, three methods of QTL mapping were compared: (i) full data analysis (FULL) where all phenotypic values, including ungenotyped individuals, were included in the data analysis with the marker genotypes of ungenotyped individuals treated as missing values; (ii) biased analysis (BIAS) where only phenotypic values of genotyped individuals were included in the analysis with the likelihood function constructed as if there were no selection; and (iii) the true method of selective genotyping (SELECT) proposed in this study where only genotyped individuals were included and the likelihood was constructed with correction for the bias. The QTL location was estimated as the mean chromosomal position that shows the highest value of the test statistic. Each simulation was repeated 100 times.

The mean and standard deviations of the estimates are given in Table 1. Under random selection with the low variance explained by the QTL and the small sample size ($n = 100$), estimation of the QTL position is not only severely biased towards the centre of the chromosome but is also subject to a large estimation error. Estimates of the QTL effects and the residual variance are quite close to the expected values, although with relatively large errors. With selective genotyping (SELECT), although the same numbers of individuals are included in the analysis, the bias in QTL position estimate has been progressively corrected as the selection intensity increases; for instance, when the proportion selected is 5%, the estimation is almost unbiased with the estimation error reduced to one-third of what is observed under random selection. Compared with the FULL method, the SELECT method has a slightly increased estimation error in the QTL position estimate. This indicates that inclusion of the large number of ungenotyped individuals

Table 1 Comparisons of quantitative trait loci (QTL) mapping procedures under different selection intensities, where cM_A is the estimated QTL position (true value is 25) and λ is the generalized likelihood ratio test statistic for the presence of QTL (including both the additive and dominance effects). The numbers listed in the table are the averages of 100 replicated simulations with the standard deviations in parentheses

Selection	Method†	cM_A	$\hat{\mu}$	$\hat{a}$	$\hat{d}$	$\hat{\sigma}^2$	λ
	True value	25	0.00	0.229	0.162	1.00	
100%		33.79 (23.41)	−0.01 (0.10)	0.243 (0.190)	0.164 (0.156)	0.998 (0.134)	4.45 (2.40)
50%	FULL	31.06 (18.29)	−0.02 (0.09)	0.271 (0.155)	0.193 (0.116)	0.992 (0.075)	7.70 (3.53)
	BIAS	30.35 (18.70)	0.00 (0.15)	0.426 (0.257)	0.349 (0.203)	1.686 (0.151)	6.91 (3.44)
	SELECT	30.34 (18.65)	−0.01 (0.09)	0.252 (0.166)	0.196 (0.118)	0.973 (0.103)	6.80 (3.40)
10%	FULL	27.07 (12.25)	−0.03 (0.04)	0.284 (0.110)	0.160 (0.075)	1.009 (0.041)	12.28 (5.55)
	BIAS	27.18 (13.03)	0.05 (0.20)	0.923 (0.369)	0.618 (0.338)	3.588 (0.539)	11.69 (4.30)
	SELECT	26.99 (13.10)	−0.01 (0.06)	0.255 (0.106)	0.155 (0.080)	0.993 (0.112)	10.57 (4.29)
5%	FULL	25.54 (6.10)	−0.04 (0.03)	0.280 (0.075)	0.170 (0.058)	0.996 (0.035)	15.31 (5.51)
	BIAS	25.36 (8.12)	0.04 (0.20)	1.116 (0.314)	0.811 (0.301)	4.272 (0.538)	13.73 (4.97)
	SELECT	25.25 (8.58)	−0.02 (0.06)	0.273 (0.095)	0.175 (0.076)	1.004 (0.121)	13.10 (4.29)

† See text for definitions of the three methods.

does provide a little information about linkage for a reason to be explained later. The BIAS method, using the same amount of phenotypic information as the SELECT method, has almost identical estimation error of the QTL position as the SELECT method. Both the FULL and the SELECT methods provide estimates of the QTL effects close to the expectations with similar estimation errors. The BIAS method, however, gives severely biased estimates of the QTL effects, because of the use of an incorrect likelihood function. The residual variance is estimated very closely to the expectation by both the FULL and SELECT methods. However, estimate of the FULL method has decreased the already small estimation error. This explains the slightly decreased estimation error of the QTL position by the FULL method. The BIAS method, again, gives a very biased estimate of the residual variance. Finally, selective genotyping has increased the score of the test statistic up to threefold (see the last column of Table 2).

Multiple trait QTL mapping

In the second simulation study, we investigated QTL mapping for two correlated traits under selective genotyping. The marker map remains the same as previously described. The first trait is controlled by one QTL at the same location (25 cM) with the same effect as described in the previous experiment, i.e. $a_1 = 0.229$ and $d_1 = 0.162$. The second trait is controlled by a QTL located at 85 cM with identical effects, i.e. $a_2 = 0.229$ and

Table 2 Comparison of multiple-trait mapping under selective genotyping and random selection from 50 replicated simulations. The estimated values are means of 50 replicates with standard deviations given in parentheses

			Estimate	
	QTL parameter	True value	Selective genotyping	Random selection
Trait one	Location	25 cM	25 cM	24 cM
	Additive effect	0.229	0.248 (0.065)	0.248 (0.096)
	Dominance effect	0.162	0.173 (0.042)	0.154 (0.081)
	Residual variance	1.00	1.00 (0.070)	0.966 (0.082)
Trait two	Location	85 cM	84 cM	85 cM
	Additive effect	0.229	0.225 (0.087)	0.206 (0.088)
	Dominance effect	0.162	0.156 (0.069)	0.152 (0.065)
	Residual variance	1.00	0.996 (0.040)	0.988 (0.070)
Joint	Residual covariance	0.50	0.500 (0.078)	0.501 (0.086)

$d_2 = 0.162$. The residual variances were set at $\sigma_1^2 = \sigma_2^2 = 1$ and the residual covariance set at $\sigma_{12} = 0.5$. The selection criterion was $I_j = c_1 y_{j1} + c_2 y_{j2}$ where $c_1 = 1$ and $c_2 = 0$, i.e. only the first trait was selected. A total of 2500 individuals were measured for phenotype, but 250 (10%) were selectively genotyped. The simulation was replicated 50 times. Figure 1(a) gives the average likelihood ratio test statistic profiles under selective genotyping (10%). The solid (T_1), dotted (T_2) and dashed (T_{12}) lines represent the likelihood-ratio test statistic profiles for the first trait, the second trait and both traits (joint test), respectively. Note that the likelihood-ratio test statistic profiles (functions of the F-test statistics) are used here. They are defined as:

$$T_1 = 2\left(\frac{n \quad 6}{2}\right)\left(\frac{\hat{\sigma}^2_{\Omega_1} \quad \hat{\sigma}^2_{\Omega_{12}}}{\hat{\sigma}^2_{\Omega_{12}}}\right),$$

$$T_2 = 2\left(\frac{n \quad 6}{2}\right)\left(\frac{\hat{\sigma}^2_{\Omega_2} \quad \hat{\sigma}^2_{\Omega_{12}}}{\hat{\sigma}^2_{\Omega_{12}}}\right)$$

and

$$T_{12} = 4\left(\frac{n \quad 6}{4}\right)\left(\frac{\hat{\sigma}^2_{\Omega_0} \quad \hat{\sigma}^2_{\Omega_{12}}}{\hat{\sigma}^2_{\Omega_{12}}}\right),$$

where $\hat{\sigma}^2_{\Omega}$ is the estimated residual variance under model Ω which defines the linear model by the set of parameters included in the model:

$$\begin{array}{llllll} \Omega_{12} \in \{\mu_1 & a_1 & d_1 & \mu_2 & a_2 & d_2\} \\ \Omega_1 \in \{\mu_1 & 0 & 0 & \mu_2 & a_2 & d_2\} \\ \Omega_2 \in \{\mu_1 & a_1 & d_1 & \mu_2 & 0 & 0\} \\ \Omega_0 \in \{\mu_1 & 0 & 0 & \mu_2 & 0 & 0\}. \end{array}$$

We used T instead of λ to depict the test statistic profiles because T approximates a χ^2_q distribution and thus bears the additive property, i.e. $T_{12} = T_1 + T_2$. Although the two traits have an identical genetic variance, the first trait has a substantially higher test statistic profile than the second one because the first trait is directly selected for genotyping. As a comparison, we repeated the simulation under random selection, i.e. we generated 250 individuals and genotyped all of them (100%) for mapping. The corresponding test statistic profiles are given in Fig. 1(b). Compared with random selection (Fig. 1b), the increase in the test statistic profile for the first trait (Fig. 1a) is obvious. A slight increase in the test statistic profile for the second trait is also observed because of its correlation to the first trait.

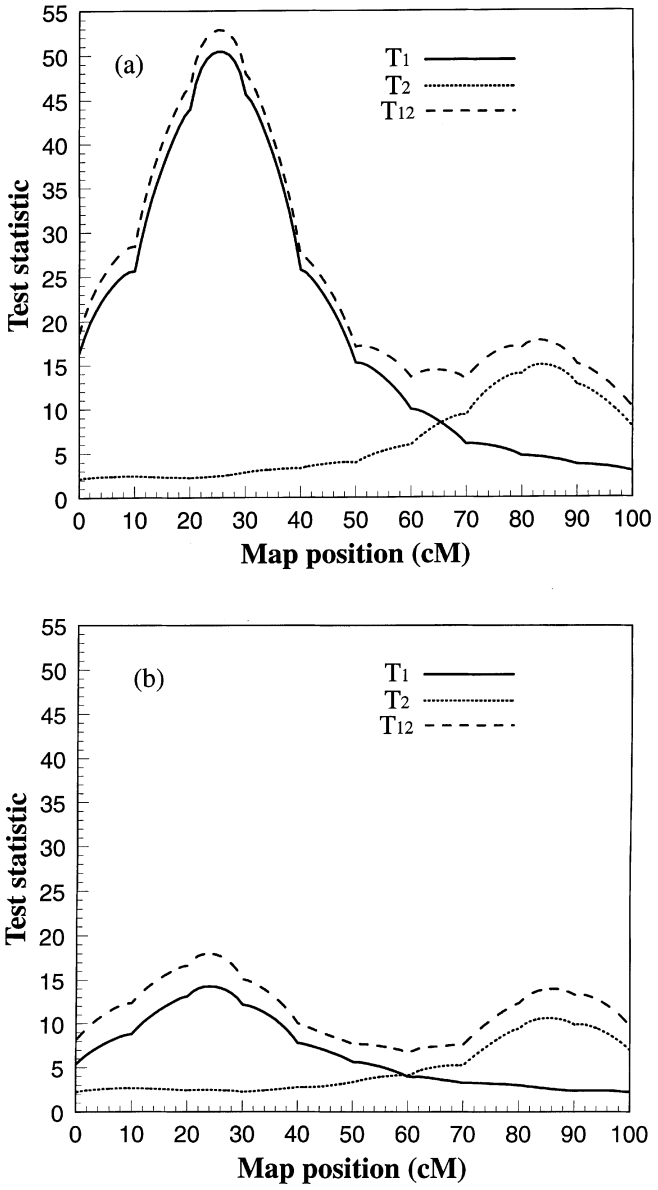

Fig. 1 (a) Test statistic profiles under selective genotyping for multivariate quantitative trait loci (QTL) mapping. (b) Test statistic profiles under random selection for multivariate QTL mapping. Two QTLs, each controlling one trait. The true QTL position of the first trait is 25 cM and that of the second trait is 85 cM. Selection is on the first trait. T_1, T_2 and T_{12} are test statistics for the presence of QTLs for the first trait, second trait and both traits, respectively (see text for definitions of the Ts).

Power under selective genotyping

As reported in this section, we first calculated the predicted powers under various proportions of genotyped individuals using the theoretical formula given in eqn (18). We then conducted simulation experiments to

Table 3 Statistical powers under various degrees of selective genotyping using the phenotypic values and marker data of $n = 100$ genotyped individuals. The case with 100% proportion genotyped represents random selection without selective genotyping. The trait is controlled by a single quantitative trait locus (QTL) explaining 5% of the phenotypic variance ($a = 0.229$ and $d = 0.162$)

Proportion genotyped	Number of individuals measured (N)	Lower† truncation (t_1)	Noncentrality (δ^*)	Power (ψ)	
				Theoretical	Simulated
100%	100	−0.00	2.623	0.511	0.526
75%	133	−0.33	3.555	0.648	0.635
50%	200	−0.70	5.190	0.818	0.793
25%	400	−1.17	8.424	0.960	0.935
10%	1000	−1.70	14.125	0.998	0.990
5%	2000	−2.00	18.599	1.000	0.999

† The upper truncation point is $t_2 = -t_1$.

verify our theoretical prediction. The effects of the QTL were again set at $a = 0.229$ and $d = 0.162$. The number of individuals genotyped remained at $n = 100$. We varied the total number of individuals measured (N) to control the proportions selected for genotyping (see column 2 of Table 3). Because the population mean was set at $\mu = 0$, the truncation points are symmetrical, and thus $t_2 = -t_1$, where the values of t_1 were found by trial and error so that the theoretical proportions selected equal the predetermined proportions (see column 3 of Table 3). The values of the noncentrality parameter are listed in column 4 of Table 3. The critical value for testing the hypothesis at a Type I error rate of $\alpha = 0.05$ is $F^{-1}(0.95 : 2, 97, 0) = 3.09$, which was used to calculate the theoretical powers (listed in column 5 of Table 3). We then simulated 1000 samples under each level of proportion and conducted QTL analysis for each sampled data set. The empirical power under each setting was calculated as the proportion of the samples that have the F-like test statistic greater than 3.09. These empirical powers, given in the last column of the table, are fairly close to the corresponding theoretical predictions.

Discussion

When the phenotypic values of ungenotyped individuals are included in the data analysis, standard methods with proper handling of missing markers are used (Lander & Botstein, 1989; Muranty & Goffinet, 1997a,b; Henshall & Goddard, 1999; Johnson *et al*., 1999). A problem occurs if the number of ungenotyped individuals is large because of the increased computational burden; for example, if 10% of the test population is genotyped, to genotype 250 individuals, one needs to measure an additional 2250 individuals for their phenotypes. The total sample size will be 2500. Because the 2250 ungenotyped individuals contribute very little to linkage analysis but serve as bias correctors, their phenotypic values do not have to be included in the analysis. These individuals, however, do contribute to the estimation of the residual variance. The estimate of the residual variance usually has very small estimation error. When the number of individuals genotyped is small, however, the residual variance estimate from only the genotyped individuals may not be sufficiently accurate. In this case, it is important to include the ungenotyped individuals. The methods described above (e.g. Muranty & Goffinet, 1997a,b; Johnson *et al*., 1999) are not the only ways to include the ungenotyped individuals. An alternative way is to ignore completely the genetic effects for the ungenotyped individuals, partition the residual variance of an ungenotyped individual into a genetic and a pure environmental component, and use a mixed-model approach. This can be accomplished via the following maximum likelihood analysis. Define the model for an ungenotyped individual by $y_j = \mu + r_j$ for $j = n + 1, \ldots, N$, where r_j is the residual effect with an $N(0, \sigma_G^2 + \sigma^2)$ distribution and $\sigma_G^2 = a^2/2 + d^2$, as defined previously. The probability density of y_j for the ungenotyped individual will be:

$$f(y_j) = \{1/\sqrt{[2\pi(\sigma_G^2 + \sigma^2)]}\} \times \exp\{ \quad [1/2(\sigma_G^2 + \sigma^2)][y_k \quad \mu]^2\}.$$

The likelihood function including all individuals will be:

$$L(\mathbf{b}, \sigma^2|\mathbf{y}) = \prod_{j=1}^{n} \left[\sum_{\mathbf{x}_j \in S} p(\mathbf{x}_j) f(y_j|\mathbf{x}_j) \right] \times \prod_{j=n+1}^{N} f(y_j).$$

Note that ungenotyped individuals do not contribute to the estimation of $\mathbf{b}$ except μ, but they are used to estimate $\sigma_G^2 + \sigma^2$. The MLE may be directly searched or obtained via an EM algorithm. In either way, the speed of convergence may be faster than the methods

that treat $\mathbf{x}_j$ as missing values because there is no need to update $p(\mathbf{x}_j)$ for an ungenotyped individual. Further investigation is required to explore the properties of this alternative approach.

It is not hard to imagine that ungenotyped individuals may not have a full measurement of phenotypic values. This may occur, for example, in QTL mapping for the trait of flowering time. An investigator may choose to visit the field for the first few days when the population of plants begins to flower and the last few days when the population approaches the end of the flowering season. In this case, plants that flower in the middle of the season may not have a record of phenotype. Another example comes from multiple trait analysis in forest trees. One may decide to select early growth rate for QTL mapping, but later the investigator may want to map QTLs for later growth rate as well. If selection is on the early trait, because of limited space, the investigator may not keep the culled individuals in the field. Then the final population would be a selected population with regard to the later growth rate. The maximum likelihood analysis proposed in this study is the proper tool for handling such centrally truncated data.

It is undesirable to use only one tail of the trait distribution to carry out QTL mapping because the total variance of the trait is artificially deflated. However, if the data happen to be single-tail truncated for some technical reasons, the proposed method can readily be applied for correcting the bias. A typical example of single-tail truncation can be seen in artificial selection of plant and animal breeding. Another example may come from longitudinal data analysis where the phenotypic value of an individual depends on it longevity. Only surviving individuals have a complete measurement of phenotype, whereas individuals not surviving only have partial information; for example, the yearly egg production of a chicken strongly depends on the viability of the chicken. If a chicken dies in the middle of the year, we do not know the phenotypic record of her yearly production, but we do know that her yearly egg production is greater than the current production in her record at the time when she dies. An unbiased analysis must be performed by taking into account these partial records.

For multiple-trait analysis, selective genotyping has been a problem because if all traits are deemed to be important to the researcher, which traits should be selected? The selection index approach of Muranty & Goffinet (1997b) is a compromise between the traits. Because the selection criterion now becomes a single 'trait', it is easy to apply in practice. Lin & Ritland (1997) suggested that an individual should be genotyped if at least one of m traits exhibits the extreme value. Under this selection regime, different individuals seem to have different criteria of selection; for example, if individual j is selected because its kth phenotypic value is first observed as being extreme, then the criterion for j is $(y_{jk} \le t_{1k}) \cup (y_{jk} \ge t_{2k})$. On the other hand, if individual i is selected because its lth phenotypic value is first observed as being extreme, then the criterion for i is $(y_{il} \le t_{1l}) \cup (y_{il} \ge t_{2l})$. In both the index selection and the method of Lin & Ritland (1997), the selection criterion of each individual is a single trait (one-dimensional selection), and thus the proposed method will apply. Another selection regime may be the so-called independent culling level selection where an individual will not be genotyped if any one of the m phenotypic values fails to reach the extreme. This selection regime is perhaps more rigorous than the previous two methods, but it is hard to programme because it is a multiple dimensional selection (requiring multiple integration). Further study may be necessary to compare different selection regimes. Nonetheless, when phenotypic values of ungenotyped individuals are included, methods of selection will be irrelevant to the statistical issue. Once selection is carried out on the phenotypic value of one trait (primary trait), QTL mapping for a highly correlated trait (secondary trait) will also benefit. However, if the two traits are not correlated, the effective sample size in terms of the secondary trait will be comparable to a *random* sample of n, where n is the number of genotyped individuals. Therefore, one should be cautious about the power of QTL mapping for traits less correlated to a highly selected primary trait.

An advantage of the logistic regression of Henshall & Goddard (1999) is that selection does not bias estimates of QTL effects, irrespective of whether phenotypic values of ungenotyped individuals are included in the data analysis. This is because the roles of marker genotypes and the phenotypes in the likelihood function have been altered, just like the discordant sib-pair mapping of Risch & Zhang (1995). Further investigation on the logistic regression, however, shows that selective genotyping can alter the estimation of the QTL effect. The equivalence between logistic regression and the maximum likelihood holds only approximately when the effect of a QTL is small. This can be shown by looking at the posterior probability of a QTL genotype given the phenotypic value of individual j:

$$p^*(\mathbf{x}_j) = [p(\mathbf{x}_j)g(y_j|\mathbf{x}_j)] \Big/ \left[\sum_{x_j \in S} p(\mathbf{x}_j)g(y_j|\mathbf{x}_j)\right],$$

where $p(\mathbf{x}_j)$ is the prior probability of the QTL genotype, independent of marker information, and $g(y_j|\mathbf{x}_j) = [f(y_j|\mathbf{x}_j)]/\{\Phi(\tau_1|\mathbf{x}_j) + [1 - \Phi(\tau_2|\mathbf{x}_j)]\}$. However, the logistic regression model uses $r(\mathbf{x}_j) = [p(\mathbf{x}_j)f(y_j|\mathbf{x}_j)]/[\sum_{x_j \in S} p(\mathbf{x}_j)f(y_j|\mathbf{x}_j)]$, i.e. the term $\Phi(\tau_1|\mathbf{x}_j) + [1 - \Phi(\tau_2|\mathbf{x}_j)]$ in the denominator of $g(y_j|\mathbf{x}_j)$ has vanished. The exact

maximum likelihood function should be built using $p^*(\mathbf{x}_j)$ instead of $r(\mathbf{x}_j)$. However, using $r(\mathbf{x}_j)$ may still be justifiable because: (i) when the size of the QTL is small, $\Phi(\tau_1|\mathbf{x}_j) + [1 - \Phi(\tau_2|\mathbf{x}_j)]$ can be considered as a constant across different genotypes so that the corresponding terms in the numerator and denominator cancel each other out, leading to $p^*(\mathbf{x}_j) \approx r(\mathbf{x}_j)$; and (ii) $r(\mathbf{x}_j)$ is much easier to handle than $p^*(\mathbf{x}_j)$ in the maximum likelihood analysis. In addition to the approximate nature of the logistic regression, there are two unsolved problems: (a) modification is required to map a QTL in an F_2 population; and (b) an exact interval mapping has not been available. An approximate interval mapping was accomplished via interpolation (Henshall & Goddard, 1999). Solving for the first problem requires a multicategorical response model, e.g. models for nominal or ordinal responses (Fahrmeir & Tutz, 1994). The second problem involves missing QTL genotypes and may be solved using the EM or MCMC algorithm of C. Vogl & S. Xu (unpublished results) for mapping viability loci.

Quantitative trait loci mapping is usually performed after a marker map is fully developed. If the trait under selection has a strong genetic component, selective genotyping can also cause distortion of the inferred marker map from the true one. The distortion is reflected by the change in both the marker order and the distances between markers. Because the severity of marker map distortion depends on the sizes and locations of QTLs, marker mapping and QTL mapping should be carried out concurrently under selective genotyping. One can apply the general idea of EM to concurrent mapping. To do this, one first maps QTLs under the assumption that the marker map is known without error, and then corrects the marker map by taking into account the distortion caused by selective genotyping under the assumption that the sizes and locations of QTLs are known. This completes one cycle of iteration, and the iteration should continue until a criterion of convergence is reached. The problem can be very complicated, especially when the marker order is allowed to change. Many theoretical and practical problems may exist in concurrent mapping, and further investigation is deemed necessary.

Maximization of the likelihood function is not an easy task. Special algorithms and computer programs are required. We developed an EM algorithm that appears to be a simple modification of the existing EM algorithm for standard interval mapping. As a result, it can be readily incorporated into a standard QTL mapping software, e.g. MAPMAKER (Lander & Botstein, 1989). One caveat about the EM algorithm is that when the selection intensity is too high, the EM algorithm may take a very large number of iterations to converge and sometime may not converge at all. This is not a problem of the EM algorithm itself, rather, it is caused by numerical overflow when the bias adjustment of the residual variance is conducted. Recall that we added an additional term in eqn (8), $\sigma^2[(\tau_1|\mathbf{x}_j)\phi(\tau_1|\mathbf{x}_j) - (\tau_2|\mathbf{x}_j)\phi(\tau_2|\mathbf{x}_j)]/[1 + \Phi(\tau_1|\mathbf{x}_j) - \Phi(\tau_2|\mathbf{x}_j)]$, to the standard EM estimation of the residual variance. When the selection intensity is high, this term can be numerically unstable. Proper handling of this numerical overflow is required which is, unfortunately, beyond our technical ability. We found that when the proportion selected is 40% or more, the problem rarely happens. In our simulation studies, when numerical overflow occurred, the EM algorithm was replaced by the simplex algorithm (Nelder & Mead, 1965) for direct search of MLEs. The simplex method is usually slower than the EM algorithm, but it can handle highly selected data. Another caveat is the sensitivity of the proposed method to departure from normality. Because the likelihood function involves $\Phi(\tau)$, in addition to $\phi(\tau)$, we anticipate that the method is more sensitive to deviation from normality than the methods using also the ungenotyped individuals.

In conclusion, we developed an exact maximum likelihood approach to map QTLs under selective genotyping using phenotypic values of genotyped individuals only. Compared with the full data analysis (using all phenotypic values), the proposed method performs well: the average test statistic is slightly lower; estimates of QTL parameters are almost identical; and the estimate of residual variance is subject to a relatively large error. The slightly lower test statistic value may be caused by the relatively large increase in the estimation error of the residual variance. A general recommendation is that whenever full data analysis is possible, the full maximum likelihood analysis should be performed. If it is impossible or difficult to analyse the full data, e.g. the sample size is too large, the phenotypic values of ungenotyped individuals are missing or composite interval mapping is to be performed, then the proposed method should be applied with the understanding that there is little to lose.

Acknowledgements

Many thanks are due to Drs W. M. Muir and O. Savolainen for their encouragement and support on the project. We thank Drs D. Gessler, J. Lin, R. Whitkus and M. Sillanpaa for their helpful comments on an earlier version of the manuscript. We would also like to thank two anonymous reviewers for their criticisms and comments which have greatly improved the presentation of the manuscript. This research was supported by the

National Institutes of Health Grant GM55321 and the USDA National Research Initiative Competitive Grants Program 97–35205–5075 to S.X.

References

COHEN, A. C. 1991. *Truncated and Censored Samples — Theory and Applications*. Marcel Dekker, New York.

DARVASI, A. AND SOLLER, M. 1992. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theor. Appl. Genet.*, **85**, 353–359.

FAHRMEIR, L. AND TUTZ, G. 1994. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York, NY.

GRAYBILL, F. A. 1976. *Theory and Application of the Linear Model*. Duxbury Press, North Scituate, MA.

GROOVER, A., DEVEY, M., FIDDLER, T., LEE, J., MEGRAW, R., MITCHELL-OLDS, T., *ET AL.* 1994. Identification of quantitative trait loci influencing wood specific gravity in an outbred pedigree of loblolly pine. *Genetics*, **138**, 1293–1300.

HENSHALL, J. M. AND GODDARD, M. E. 1999. Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression. *Genetics*, **151**, 885–894.

JANSEN, R. C. AND STAM, P. 1994. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136**, 1447–1455.

JIANG, C. AND ZENG, Z. B. 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, **140**, 1111–1127.

JOHNSON, D. L., VAN JANSEN, R. C. AND ARENDONK, A. M. 1999. Mapping quantitative trait loci in a selectively genotyped outbred population using a mixture model approach. *Genet. Res.*, **73**, 75–83.

LANDER, E. S. AND BOTSTEIN, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.

LIN, J. Z. AND RITLAND, K. 1997. Quantitative trait loci differentiating the outbreeding *Mimulus guttatus* from the inbreeding *M. platycalyx*. *Genetics*, **146**, 1115–1121.

MURANTY, H. 1996. Power of tests for quantitative trait loci detection using full-sib families in different schemes. *Heredity*, **76**, 156–165.

MURANTY, H. AND GOFFINET, B. 1997a. Selective genotyping for location and estimation of the effect of a quantitative trait locus. *Biometrics*, **53**, 629–643.

MURANTY, H. AND GOFFINET, B. 1997b. Multitrait and multi-population QTL search using selective genotyping. *Genet. Res.*, **70**, 259–265.

NELDER, J. A. AND MEAD, R. 1965. A simplex method for function minimization. *Comput. J.*, **7**, 308–313.

RISCH, N. AND ZHANG, H. 1995. Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science*, **268**, 1584–1589.

SOLLER, M. AND BRODY, T. 1976. On the power of experimental designs for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theor. Appl. Genet.*, **47**, 35–39.

ZENG, Z. B. 1994. Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457–1468.

Appendix

Derivation of the EM algorithm for single-trait analysis

First, let us define the log likelihood function by:

$$l(\mathbf{b}, \sigma^2|\mathbf{y}) = \sum_{j=1}^{n} \log\left[\sum_{x_j \in S} p(\mathbf{x}_j) g(y_j|\mathbf{x}_j)\right].$$

The MLEs of $\mathbf{b}$ and σ^2 are obtained by solving $\partial l/\partial \mathbf{b} = \mathbf{0}$ and $\partial l/\partial \sigma^2 = 0$ simultaneously.

Derivation of $\partial l/\partial \boldsymbol{b}$

$$\frac{\partial l}{\partial \mathbf{b}} = \sum_{j=1}^{n} \frac{\sum_{x_j \in S} p\ \mathbf{x}_j) \frac{\partial}{\partial \mathbf{b}} g\ y_j|\mathbf{x}_j)}{\sum_{x_j \in S} p\ \mathbf{x}_j) g\ y_j|\mathbf{x}_j)}$$

where

$$\sum_{x_j \in S} p\ \mathbf{x}_j) \frac{\partial}{\partial \mathbf{b}} g\ y_j|\mathbf{x}_j)$$
$$= \sum_{x_j \in S} p\ \mathbf{x}_j) g\ y_j|\mathbf{x}_j)$$
$$\times \left[\frac{1}{\sigma^2}\mathbf{x}_j^{\mathrm{T}}\ y_j \quad \mathbf{x}_j\mathbf{b})\right] \quad \frac{\frac{\partial}{\partial \mathbf{b}}\left[1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j)\right]}{1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j)}\Bigg].$$

In the above equation, the truncation points have been standardized, i.e. $\tau_1 = (t_1 - \mathbf{x}_j\mathbf{b})/\sigma$ and $\tau_2 = (t_2 - \mathbf{x}_j\mathbf{b})/\sigma$, where t_1 and t_2 are the truncation points in the original scale. Note that:

$$\frac{\partial}{\partial \mathbf{b}}\left[1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j)\right]$$
$$= \phi(\tau_1|\mathbf{x}_j)\frac{\partial \tau_1}{\partial \mathbf{b}} \quad \phi(\tau_2|\mathbf{x}_j)\frac{\partial \tau_1}{\partial \mathbf{b}}$$
$$= \frac{1}{\sigma}\left[\phi(\tau_1|\mathbf{x}_j) \quad \phi(\tau_2|\mathbf{x}_j)\right](\quad \mathbf{x}_j^{\mathrm{T}}).$$

Hence,

$$\sum_{x_j \in S} p\ \mathbf{x}_j) \frac{\partial}{\partial \mathbf{b}} g\ y_j|\mathbf{x}_j)$$
$$= \sum_{x_j \in S} p\ \mathbf{x}_j) g\ y_j|\mathbf{x}_j) \frac{1}{\sigma^2}\mathbf{x}_j^{\mathrm{T}}$$
$$\times \left[\ y_j \quad \mathbf{x}_j\mathbf{b}) + \sigma \frac{\phi(\tau_1|\mathbf{x}_j) \quad \phi(\tau_2|\mathbf{x}_j)}{1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j)}\right].$$

Define $p^*(\mathbf{x}_j) = [p(\mathbf{x}_j)g(y_j|\mathbf{x}_j)]/[\sum_{x_j \in S} p(\mathbf{x}_j)g(y_j|\mathbf{x}_j)]$ as the posterior distribution of $\mathbf{x}_j$. We now have:

$$\frac{\partial l}{\partial \mathbf{b}} = \sum_{j=1}^{n} \frac{\sum_{x_j \in S} p\ \mathbf{x}_j) \frac{\partial}{\partial \mathbf{b}} g\ y_j|\mathbf{x}_j)}{\sum_{x_j \in S} p\ \mathbf{x}_j) g\ y_j|\mathbf{x}_j)}$$

$$= \sum_{j=1}^{n} \sum_{x_j \in S} p^*(\mathbf{x}_j) \frac{1}{\sigma^2} \mathbf{x}_j^{\mathrm{T}} \times \left[\left(y_j + \sigma \frac{\phi(\tau_1|\mathbf{x}_j) \quad \phi(\tau_2|\mathbf{x}_j)}{1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j)} \right) \quad \mathbf{x}_j \mathbf{b} \right].$$

Solving for $\partial l/\partial \mathbf{b} = \mathbf{0}$, we have:

$$\hat{\mathbf{b}} = \left[\sum_{j=1}^{n} \mathrm{E}_x(\mathbf{x}_j^{\mathrm{T}} \mathbf{x}_j) \right]^{1} \times \left[\sum_{j=1}^{n} E_x \left\{ \mathbf{x}_j^{\mathrm{T}} \left[y_j + \sigma \frac{\phi(\tau_1|\mathbf{x}_j) \quad \phi(\tau_2|\mathbf{x}_j)}{1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j)} \right] \right\} \right].$$

Derivation of $\partial l/\partial \sigma^2$

$$\frac{\partial l}{\partial \sigma^2} = \sum_{j=1}^{n} \frac{\sum_{x_j \in S} p\ \mathbf{x}_j) \frac{\partial}{\partial \sigma^2} g\ y_j|\mathbf{x}_j)}{\sum_{x_j \in S} p\ \mathbf{x}_j) g\ y_j|\mathbf{x}_j)},$$

where

$$\sum_{x_j \in S} p\ \mathbf{x}_j) \frac{\partial}{\partial \sigma^2} g\ y_j|\mathbf{x}_j) = \sum_{x_j \in S} p\ \mathbf{x}_j) \left\{ \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\quad \frac{1}{2\sigma^2} (y_j \quad \mathbf{x}_j \mathbf{b})^2 \right]}{1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j)} \times \left[\frac{1}{2\sigma^4} (y_j \quad \mathbf{x}_j \mathbf{b})^2 \quad \frac{1}{2\sigma^2} \frac{\frac{\partial}{\partial \sigma^2} \left[1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j) \right]}{1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j)} \right] \right\}.$$

Note that

$$\frac{\partial}{\partial \sigma^2} \left[1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j) \right] = \phi(\tau_1|\mathbf{x}_j) \frac{\partial \tau_1}{\partial \sigma^2} \quad \phi(\tau_2|\mathbf{x}_j) \frac{\partial \tau_1}{\partial \sigma^2} = \left(\quad \frac{1}{2\sigma^2} \right) \left[\tau_1 \phi(\tau_1|\mathbf{x}_j) \quad \tau_2 \phi(\tau_2|\mathbf{x}_j) \right].$$

Therefore,

$$\sum_{x_j \in S} p\ \mathbf{x}_j) \frac{\partial}{\partial \sigma^2} g\ y_j|\mathbf{x}_j) = \sum_{x_j \in S} p\ \mathbf{x}_j) g\ y_j|\mathbf{x}_j) \times \left[\ y_j \quad \mathbf{x}_j \mathbf{b})^2 \quad \sigma^2 \left(1 \quad \frac{\tau_1 \phi(\tau_1|\mathbf{x}_j) \quad \tau_2 \phi(\tau_2|\mathbf{x}_j)}{1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j)} \right) \right].$$

Finally, we have

$$\frac{\partial L}{\partial \sigma^2} = \sum_{j=1}^{n} \frac{\sum_{x_j \in S} p\ \mathbf{x}_j) \frac{\partial}{\partial \sigma^2} g\ y_j|\mathbf{x}_j)}{\sum_{x_j \in S} p\ \mathbf{x}_j) g\ y_j|\mathbf{x}_j)} = \sum_{j=1}^{n} \left\{ \sum_{x_j \in S} p^* \ \mathbf{x}_j) \left[(y_j \quad \mathbf{x}_j \mathbf{b})^2 \quad \sigma^2 \left(1 \quad \frac{\tau_1 \phi(\tau_1|\mathbf{x}_j) \quad \tau_2 \phi(\tau_2|\mathbf{x}_j)}{1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j)} \right) \right] \right\}.$$

Solving for $\partial l/\partial \sigma^2 = 0$, we obtain

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} \mathrm{E}_x \left[\ y_j \quad \mathbf{x}_j \mathbf{b})^2 + \sigma^2 \frac{\tau_1 \phi(\tau_1|\mathbf{x}_j) \quad \tau_2 \phi(\tau_2|\mathbf{x}_j)}{1 + \Phi(\tau_1|\mathbf{x}_j) \quad \Phi(\tau_2|\mathbf{x}_j)} \right].$$