Mapping Quantitative Trait Loci for Longitudinal Traits

in Line Crosses

Runqing Yang*, Quan Tian* and Shizhong Xu†

*School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai, 201101, P. R. China
†Department of Botany and Plant Science, University of California, Riverside, CA 92521, USA
Running head: Mapping QTL for Longitudinal Traits

Key words: EM algorithm, Longitudinal trait, Mixed model, Polynomial, QTL.

Corresponding author:
Dr. Shizhong Xu
Department of Botany and Plant Sciences
University of California
Riverside, CA 92521
Tel: (909) 787 5898
Fax: (909) 787 4437
E-mail: xu@genetics.ucr.edu
ABSTRACT

Quantitative traits whose phenotypic values change over time are called longitudinal traits. Genetic analyses of longitudinal traits can be conducted using any of the following approaches: (1) treating the phenotypic values at different time points as repeated measurements of the same trait and analyzing the trait under the repeated measurements framework; (2) treating the phenotypes measured from different time points as different traits and analyzing the traits jointly based on the theory of multivariate analysis; and (3) fitting a growth curve to the phenotypic values across time points and analyzing the fitted parameters of the growth trajectory under the theory of multivariate analysis. The third approach has been used in QTL mapping for longitudinal traits by fitting the data to a logistic growth trajectory. This approach only applies to the particular s-shaped growth process. In practice, a longitudinal trait may show a trajectory of any shape. We demonstrate that one can describe a longitudinal trait with orthogonal polynomials, which are sufficiently general for fitting any shaped curve. We develop a mixed model methodology for QTL mapping of longitudinal traits and a maximum likelihood method for parameter estimation and statistical tests. The expectation-maximization (EM) algorithm is applied to search for the maximum likelihood estimates of parameters. The method is verified with simulated data and demonstrated with experimental data from a pseudo backcross family of Populus (poplar) trees.
INTRODUCTION

The genetic variance of a quantitative trait is controlled by the segregation of many genes, each with a small effect (FALCONER AND MACKAY 1996). Small genetic effects are collectively called the polygenic effects. In the infinitesimal model, only polygenic effects exist. An alternative to this classical definition of genetic variance of a quantitative trait is the theory that the genetic variances of many quantitative traits are actually controlled by the segregation of one or more major genes plus numerous small genetic effects. Using the latter definition, the phenotype of interest still shows a continuous distribution because of the joint contribution from the polygenic effects and random environmental variant. Such a model is called the oligogenic model, which can be tested using segregation analysis (ELSTON and STEWART 1971; MORTON and MACLEAN 1974). In fact, the oligogenic model is the basis for QTL mapping because QTL analysis with the current statistical methods and sample sizes ($n < 1000$) can only detect genes with major or moderate effects (e.g., HALEY and KNOTT 1992; ZENG 1994; KAO et al. 1999; XU and YI 2000).

When a trait is measured repeatedly over time, it is called a time dependent trait or longitudinal trait. Three approaches are currently available for analyzing longitudinal traits. The first approach is to treat the phenotypic values at different time points as repeated measurements of the same trait and analyze the trait under the repeated measurements framework. The second approach is to treat the phenotypes measured from different time points as different traits and analyze the traits jointly based on the theory of multivariate analysis (Wu et al. 1999). The third approach is to fit a growth curve to the phenotypic
values across different time points and analyze the fitted parameters of the growth trajectory under the theory of multivariate analysis in a much reduced dimension (Wu et al. 2002). The method of fitting a growth trajectory is considered to be the optimal one because it treats phenotypes measured over time as different traits, but takes into account the correlation generated by the ordered time points.

Abundant molecular markers are now available for QTL mapping. Many QTL mapping studies have focused on traits measured at a single time point (e.g., Hill and Knott 1990; Zeng 1994; Xu and Yi 2000). Methods for single trait QTL mapping cannot be directly adopted for longitudinal traits because phenotypes measured at different time points may be controlled by different sets of genes (Cheverud et al. 1996; Nuzhdin et al. 1997; Verhaegen et al. 1997; Emebiri et al. 1998; Wu et al. 1999, Yan et al. 1998a, b; Wu et al. 2002a). Several methods have been proposed for multiple trait analysis in the context of QTL mapping (e.g., Jiang and Zeng 1995; Korol et al. 1995; Ronin et al. 1995; Eaves et al. 1996; Wu et al. 1999; Knott and Haley 2000). The method of Jiang and Zeng (1995) is one of the first methods for multivariate QTL mapping. In practice, the method is suitable only for a few traits. As the number of traits increases, computational time will be a prohibitive factor. To reduce the dimensionality of the multivariate analysis, a principle component analysis has been proposed (Mangin et al. 1998; Korol et al. 2001), in which the first principle component, called a "super trait", is used as a representative of the multiple trait complex. QTL mapping can be performed on the "super trait," which, methodologically, is not different from the univariate QTL mapping. However, the interpretation of the "super trait" in a biological setting is challenging.
It is in general agreement that the phenotypic values measured over time are correlated. The correlation may be generated by some underlying biological process. In fact, if the trait were measured at an infinite number of time points, the phenotypic value would form a smooth curve as a function of time. This phenotypic profile is called the growth trajectory, which may be governed by a few growth parameters. Genetic analysis of traits with time-course repeated measurements may be conducted on the few growth parameters. The work of Wu et al. (2002) was among the early studies on QTL mapping based on the growth models. The authors first fit the phenotypes of a longitudinal trait to a growth model (equations 3 and 4 of Wu et al. 2002) and then treated the estimated growth parameters as multiple traits. QTL mapping was then conducted on the growth parameters using a standard multivariate QTL mapping procedure (Jiang and Zeng 1995). The two-step approach may not be optimal because errors in the estimated growth parameters (the first step) have not been taken into account. Nevertheless, the idea presented by Wu et al. (2002) is fundamentally important in the study of longitudinal traits. Other scientists have been attracted to this important area of QTL mapping. Rongling Wu and his colleagues (Wu et al. 2002b, 2003, 2004; Ma et al. 2002) presented substantial improvement on the method of Weiren Wu and his group (Wu et al. 2002). The improvement was reflected by the change from two steps to a single step. Ma et al. (2002) used a logistic curve to model the growth trajectory, in which the genetic value is described by $g(t) = a/(1 + be^{-rt})$, where $t$ is the time point, $a$ is the asymptotic value of $g$, $r$ is the relative rate of growth and $b$ (combined with $a$) is a parameter describing the starting point of $g$. The combination of the three parameters, $\{a, b, r\}$, determines the actual shape of the growth trajectory. Each of the three
parameters is further partitioned into an additive and a dominance effect. If the trait is measured at \( m + 1 \) time points where \( m \geq 2 \), analysis of the \( m + 1 \) measurements is always performed in three dimensions, regardless of the dimension of \( m \). The time-course functional trait analysis replaces a continuous function with a finite number of collected data points to facilitate estimation of the functional parameters.

A growth trajectory is governed by the biological law of growth (Gould 1977; Alberch et al. 1979; West et al. 2001). The shape, in general, is sigmoid or \( s \)-like, a monotonic increasing function of time. Many biological curves, however, cannot be described by the monotonic logistic function. For example, the yearly gain of growth in trees is not sigmoid with time; rather, it is a negative exponential. Daily milk production of cows may be better described by a quadratic function of time. A reaction norm (a trait measured in a continuously varying environment) may have an arbitrary shape. The logistic function cannot be applied to these curves as they are not sigmoid.

Even if the biological process has an \( s \)-shaped trajectory, the lack of linearity in the logistic regression limits the use of extensively developed inference methods available for linear models. As a result, the EM algorithm developed for the logistic function cannot be generalized to other types of functions. If we follow the approach of logistic analysis and develop methods using a different specialized function for a different curve, a different EM algorithm would be required for each of the functions, because each function involves a different set of parameters with different meanings. This is another area that needs further investigation.

We propose to use orthogonal polynomials to fit the biological curve and perform QTL
mapping for longitudinal traits with the following justifications. First, the polynomial analysis is sufficiently flexible for fitting a biological curve in an arbitrary shape. This can be achieved by choosing different orders of the polynomials. Secondly, the polynomial function is linear in the parameters, although it is non-linear in time. As a consequence, we can take advantage of the extensively developed inference methods available for linear models. Thirdly, the same EM algorithm developed for a linear mixed effects model can be applied to all types of biological curves, requiring no modification, which is clearly in contrast to the aforementioned logistic approach.

Orthogonal polynomials have been applied to genetic analysis for longitudinal traits in pedigree data (Henderson 1982; Schaeffer and Dekkers 1994; Swalve 2000; Jensen 2001). These analyses utilized the random regression model (RRM) because the regression (polynomial) coefficients of individual animals were treated as random effects. Meyer and Hill (1997) showed the equivalence of the covariance function of Kirpatrick et al. (1990) to the RRM. A key element in the RRM analysis is the function sub-models for the random regression (Misztal et al. 2000). One such orthogonal polynomial is the normalized Legendre polynomial (Kirpatrick et al. 1990; Meyer 2000; Guo and Schaeffer 2002), which has been extensively used to evaluate breeding values for various longitudinal traits of large farm animals (Schaeffer 2004). The publication of the first QTL mapping study for longitudinal traits in pedigrees has been conducted using Legendre orthogonal polynomials (Macgregor et al. 2005), demonstrating the applied importance of this approach. Notably, dominance and epistatic effects of QTL were assumed to be absent in that study. Here we extend this idea to a much more general framework, which will include dominance as well as
traditional QTL populations such as the F2 and backcross.

In this study, we take the polynomial approach to mapping QTL for longitudinal traits, but address the problem in the context of line crossing experiments. Line crosses are major designs of the experiment for genetic analysis. An F2 population derived from the cross of two inbred lines is commonly used genetic material for QTL mapping. We first introduce the theory and methodology based on the F2 mating design. We then conduct a series of simulation experiments to validate the method. Finally, we apply the method to QTL mapping for the growth trajectory in a pseudo backcross family of Populus (poplar) trees and compare the result with that of MA et al. (2002).

THEORY AND METHODS

Genetic model for longitudinal traits

Based on Mendel's law of inheritance, there are three possible genotypes in an F2 population for a segregating locus, denoted by AA, Aa and aa, respectively, for the three genotypes. Let \( y_i(\tau) \) be the phenotypic value of individual \( i \) measured at time \( \tau \) (a standardized time point ranging from \(-1\) to \(+1\), see APPENDIX B for the definition of standardized time point). This phenotypic value \( y_i(\tau) \) can be described by the following linear model,

\[
y_i(\tau) = \mu(\tau) + x_i \alpha(\tau) + z_i \delta(\tau) + \xi_i(\tau) + \epsilon_i,
\]

for \( i = 1, \cdots, n \), where \( n \) is the number of individuals, \( \mu(\tau) \) is the population mean, \( x_i \) is a genotype indicator variable (defined as 1 for AA, 0 for Aa and -1 for aa) for the \( i \)th individual at the genetic locus of interest, \( z_i \) is a dominance indicator variable defined as
1 for the heterozygote and -1 for the homozygote, \( \alpha(\tau) \) is the additive effect, \( \delta(\tau) \) is the dominance effect, and \( \xi(\tau) \) is an individual-specific time-dependent residual error with an i.i.d. \( N(0, \sigma^2(\tau)) \) distribution and \( \epsilon_i \) is an individual-specific time-independent experimental error with an i.i.d. \( N(0, \sigma^2) \) distribution. This is a mixed effects model with \( \mu(\tau), \alpha(\tau) \) and \( \delta(\tau) \) being treated as the fixed effects and \( \xi(\tau) \) as the random effect. The purpose of the analysis is to estimate and test \( \alpha(\tau) \) and \( \delta(\tau) \), the time-dependent functional genetic effects of a hypothesized QTL at a putative position of the genome.

The genetic variance contributed by the genetic locus of interest at time \( \tau \) is expressed as

\[
\sigma^2_g(\tau) = \sigma^2_x(\tau) + \sigma^2(\tau) = \frac{1}{2} \alpha^2(\tau) + \delta^2(\tau),
\]

where \( \sigma^2_x = 1/2 \) and \( \sigma^2_z = 1 \) under the current definitions of \( x \) and \( z \) and the assumption of no segregation distortion (see APPENDIX A for the derivations). The phenotypic variance should be written as

\[
\sigma^2_p(\tau) = \sigma^2_g(\tau) + \sigma^2(\tau) = \frac{1}{2} \alpha^2(\tau) + \delta^2(\tau) + \sigma^2_z(\tau) + \sigma^2.
\]

The proportion of the phenotypic variance contributed by the QTL is

\[
h^2(\tau) = \frac{\sigma^2_z(\tau)}{\sigma^2_p(\tau)}.
\]

**Orthogonal polynomial for longitudinal traits**

All the model parameters, except \( \sigma^2 \), are functions of time. The functional relationships between genetic parameters and the time variable are described with an orthogonal polynomial. Define \( \psi(\tau) = [\psi_0(\tau) \ \psi_1(\tau) \ \cdots \ \psi_r(\tau)] \) as a basis of the polynomial with order \( r \). A basis is also called a sub-model. Method to construct the basis
of an orthogonal polynomial is given in APPENDIX B. Let us define
\[ \mathbf{\mu} = [\mu_0 \ \mu_1 \ \cdots \ \mu_r]^\top \]
as a vector of population means, which is time-independent. The
time-dependent population mean \( \mu(\tau) \) can be described as a linear combination of \( \mathbf{\mu} \)
weighted by the basis of the polynomials, i.e., \( \mu(\tau) = \mathbf{\psi}(\tau)\mathbf{\mu} \). Similarly, we can describe
other parameters with the same basis, e.g., \( \alpha(\tau) = \mathbf{\psi}(\tau)\mathbf{\alpha} \) and \( \delta(\tau) = \mathbf{\psi}(\tau)\mathbf{\delta} \), where
\[ \mathbf{\alpha} = [\alpha_0 \ \alpha_1 \ \cdots \ \alpha_r]^\top \]
and \( \mathbf{\delta} = [\delta_0 \ \delta_1 \ \cdots \ \delta_r]^\top \). The random effect at time \( \tau \) can
also be expressed as a linear function of the time-independent effects, \( \xi_i(\tau) = \mathbf{\psi}(\tau)\xi_i \),
where \( \xi_i = [\xi_{i0} \ \xi_{i1} \ \cdots \ \xi_{ir}]^\top \) for \( i = 1,\ldots,n \). Since \( \xi_i \) is treated as a vector of
random effects, we assume that \( \xi_i \) is i.i.d. \( N(\mathbf{0}, \mathbf{\Sigma}) \), where \( \mathbf{\Sigma} \) is an \( (r+1) \times (r+1) \)
positive definite covariance matrix.

Recall that we are dealing with a longitudinal trait (a trait measured repeatedly over
some time interval on the same individual). If the phenotypic value measured at any given
time point were treated as an individual trait, for \( m+1 \) time points, we would be dealing
with \( m+1 \) different traits. The vector of genetic effects, say \( \mathbf{\alpha} \), would have a dimension
of \( m+1 \) because each element of vector \( \mathbf{\alpha} \) represents an effect for a trait. In the
polynomial analysis of order \( r \), however, the dimension of vector \( \mathbf{\alpha} \) is reduced to \( r+1 \)
for \( r < m \). Therefore, the polynomial analysis for longitudinal traits is a special dimension
reduction technique.

With the polynomial reparameterization, model (1) is now rewritten as a linear
function of the time-independent parameters,
\[ y_i(\tau) = \mathbf{\psi}(\tau)\mathbf{\mu} + x_i\mathbf{\psi}(\tau)\mathbf{\alpha} + z_i\mathbf{\psi}(\tau)\mathbf{\delta} + \mathbf{\psi}(\tau)\xi_i + \epsilon_i. \]
(5)
The expectation function of $y_i(\tau)$ conditional on the fixed effects is

$$E\left(y_i(\tau)\right) = \psi(\tau)\mu + x_i \psi(\tau)\alpha + z_i \psi(\tau)\delta.$$  
(6)

The covariance function between $y_i(\tau)$ and $y_i(s)$ conditional on the fixed effects is

$$\text{cov}\left(y_i(\tau), y_i(s)\right) = \psi(\tau)\Sigma \psi^\top(s) + \eta(\tau, s)\sigma^2,$$
(7)

where $s$ is another standardized time point, $\eta(\tau, s) = 1$ for $\tau = s$ and $\eta(\tau, s) = 0$ for $\tau \neq s$. The first term in the right hand side of equation (7), $\psi(\tau)\Sigma \psi^\top(s)$, is actually the covariance between $\xi_i(\tau)$ and $\xi_i(s)$ for $\tau \neq s$. When $\tau = s$, it becomes the variance of $\xi_i(\tau)$, i.e., $\sigma^2_{\xi}(\tau) = \psi(\tau)\Sigma \psi^\top(\tau)$.

Model (5) is now linear in parameters, which can be estimated under the framework of the mixed model analysis. We first estimate $\mu$, $\alpha$, $\delta$, $\Sigma$ and $\sigma^2$, and then use the basis of the polynomial to convert $\mu$, $\alpha$, $\delta$ into their time-dependent functional counterparts.

The polynomial transformation allows us to express the total genetic variance as a function of the quadratic terms of the time-independent genetic parameters, as shown below,

$$\sigma^2_g(\tau) = \frac{1}{2} \alpha^2(\tau) + \delta^2(\tau) = \psi(\tau)\left[\frac{1}{2} \alpha \alpha^\top + \delta \delta^\top\right] \psi^\top(\tau).$$
(8)

The phenotypic variance may be rewritten as

$$\sigma^2_p(\tau) = \sigma^2_g(\tau) + \sigma^2_{\xi}(\tau) + \sigma^2 = \psi(\tau)\left[\frac{1}{2} \alpha \alpha^\top + \delta \delta^\top + \Sigma\right] \psi^\top(\tau) + \sigma^2.$$  
(9)

The proportion of the phenotypic variance contributed by the major gene is
One can examine the behavior of each genotype by evaluating the change of the genotypic value across time. Let $g_{aa}(\tau)$, $g_{Aa}(\tau)$ and $g_{AA}(\tau)$ be the values of the three genotypes at time $\tau$. They are expressed as

$$
\begin{align*}
g_{aa}(\tau) &= \mu(\tau) - \alpha(\tau) - \delta(\tau) = \psi(\tau)(\mu - \alpha - \delta) \\
g_{Aa}(\tau) &= \mu(\tau) + \delta(\tau) = \psi(\tau)(\mu + \delta) \\
g_{AA}(\tau) &= \mu(\tau) + \alpha(\tau) - \delta(\tau) = \psi(\tau)(\mu + \alpha - \delta)
\end{align*}
$$

Likelihood function

The phenotypic values for each individual are collected at $m+1$ fixed time points, denoted by $\tau_0, \tau_1, \ldots, \tau_m$. We add a subscript to variable $\tau$ to indicate that $\tau$ now has taken a particular value from one of the $m+1$ time points. We use an $(m+1) \times 1$ vector to denote an array of the phenotypic values for the $i$th individual,

$$
y_i = [y_i(\tau_0) \quad y_i(\tau_1) \quad \cdots \quad y_i(\tau_m)]^T, \text{ for } i = 1, \ldots, n.
$$

Define

$$
\psi = 
\begin{bmatrix}
\psi_0(\tau_0) & \psi_0(\tau_1) & \cdots & \psi_0(\tau_m) \\
\psi_1(\tau_0) & \psi_1(\tau_1) & \cdots & \psi_1(\tau_m) \\
\vdots & \vdots & \ddots & \vdots \\
\psi_r(\tau_0) & \psi_r(\tau_1) & \cdots & \psi_r(\tau_m)
\end{bmatrix}
$$

as an $(r+1) \times (m+1)$ matrix. In matrix notation, the linear model for vector $y_i$ is

$$
y_i = \psi^T \mu + x_i \psi^T \alpha + z_i \psi^T \delta + \psi^T \xi_i + \epsilon_i, \quad (12)
$$

where $\epsilon_i = [\epsilon_{i\tau_0} \quad \cdots \quad \epsilon_{i\tau_m}]^T$ is an $(m+1) \times 1$ vector for the experimental errors with $\epsilon_i \sim N(0, I \sigma^2)$, where $I$ is an $(m+1) \times (m+1)$ identity matrix. The conditional

$$
h^2(\tau) = \frac{\sigma^2_{g}(\tau)}{\sigma^2_{p}(\tau)} = \frac{\psi(\tau) [\frac{1}{2} \alpha \alpha^T + \delta \delta^T] \psi^T(\tau)}{\psi(\tau) [\frac{1}{2} \alpha \alpha^T + \delta \delta^T + \Sigma] \psi^T(\tau) + \sigma^2}.
$$
expectation of model (12) given the fixed effects is

\[ E(y_i) = M_i = \psi^T \mu + x_i \psi^T \alpha + z_i \psi^T \delta \]  \hspace{1cm} (13)

and the variance-covariance matrix is

\[ \text{var}(y_i) = V = \psi^T \Sigma \psi + I \sigma^2 \]  \hspace{1cm} (14)

which applies to all \( i = 1, \ldots, n \).

The models discussed so far are based on known genotypes of the QTL. In practice, the genotypes are not observable and the likelihood function must be constructed by taking into account all three genotypes for all individuals. Let us order the three genotypes, \( aa, Aa \) and \( AA \), as genotypes 1, 2 and 3 (indexed by \( k \)), respectively, and use a numerical variable \( G \) to denote the three ordered genotypes. Let \( G_i \) be the genotype of the major gene for individual \( i \). Before we observe the phenotypic values, we should use markers to infer the probability of \( G_i = k \). This conditional probability is denoted by \( \pi_{ik} = \Pr(G_i = k \mid I_M) \) where \( I_M \) denotes marker information. The conditional probability \( \pi_{ik} \) varies across individuals because different individuals are supposed to have different marker genotypes. Methods to calculate \( \pi_{ik} \) can be found in LANDER and BOTSTEIN (1989) for interval mapping and JIANG and ZENG (1997) for multipoint mapping. We take the multipoint method because it facilitates an automatic mechanism to handle missing marker genotypes.

Recall that \( E(y_i \mid G_i) = M_i \) is the conditional expectation of \( y_i \) given the genotype of individual \( i \) and the population parameters. Since each individual can take one of three genotypes, there are only three distinguishable conditional expectations, denoted by
The probability density of phenotypes for the \(i\)th individual with the \(k\)th genotype \((k = 1, 2, 3)\) is

\[
p(y_i \mid G_i = k) \propto |V|^{-1/2} \exp\left[ -\frac{1}{2}(y_i - H_k)^T V^{-1} (y_i - H_k) \right].
\]  

Let \(\theta = \{\mu, \alpha, \delta, \Sigma, \sigma^2\}\) be the parameters. The likelihood function is

\[
L(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{3} \pi_{ik} p(y_i \mid G_i = k).
\]  

Several numerical methods can be used to search for the maximum likelihood estimates (MLE) of the parameters, e.g., the simplex method of NELDER and MEAD (1965). We decided to use the EM algorithm (DEMPSTER et al. 1977; LANDER and BOTSTEIN 1989; JANSEN 1993) because explicit equations for the maximization step are available in the complete data situation.

**EM algorithm for parameter estimation**

With the EM algorithm, we classify variables into data, missing values and parameters. In QTL mapping, \(y = \{y_i\}\) and \(I_M\) are data, \(\theta\) are parameters, and \(\{x_i, z_i, \xi_i\}\) are variables with missing values. The genotype indicator variables \(\{x_i, z_i\}\), however, are determined by \(G_i\). So, the missing values can be rewritten as \(\{G_i, \xi_i\}\). The likelihood function given in (17) with all the missing values integrated out is called the observed likelihood function. Instead of directly maximizing the observed likelihood function, the EM algorithm deals with the following complete-data likelihood function,
\[
L(\theta, G, \xi) = \prod_{i=1}^{n} p(y_i \mid G_i, \xi_i) p(G_i \mid I_M) p(\xi_i),
\]

where

\[
p(y_i \mid G_i, \xi_i) = (\sigma^2)^{-\frac{1}{2} (m+1)} \exp \left[ -\frac{1}{2\sigma^2} (y_i - M_i - \psi^T \xi)^T (y_i - M_i - \psi^T \xi) \right],
\]

\[
p(\xi_i) = |\Sigma|^{-\frac{1}{2}} \exp(-\frac{1}{2} \xi_i^T \Sigma^{-1} \xi_i)
\]

and

\[
p(G_i \mid I_M) = \prod_{k=1}^{3} p_{G_k}^{(G_i, k)} .
\]

Note that we introduce a new variable \( \eta \) whose value is determined by \( G_i \) as

\[\eta(G_i, k) = 1 \quad \text{for} \quad G_i = k \quad \text{and} \quad \eta(G_i, k) = 0 \quad \text{for} \quad G_i \neq k .\]

Because the complete-data likelihood function involves missing values, integration (or expectation) is taken with respect to the missing values, not for the complete-data likelihood function, but for the following log transformed complete-data likelihood function,

\[
l(\theta, G, \xi) = \ln[L(\theta, G, \xi)] = \sum_{i=1}^{n} \{ \ln[p(y_i \mid G_i, \xi_i)] + \ln[p(G_i \mid I_M)] + \ln(p(\xi_i)) \}.
\]

The expectation of equation (22) with respect to the missing values is

\[
l(\theta) = E[l(\theta, G, \xi)] = T_1(\theta) + T_2(\theta) + T_3(\theta),
\]

where

\[
T_1(\theta) = -\frac{1}{2} n(m+1) \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} E(y_i - M_i - \psi^T \xi_i)^T (y_i - M_i - \psi^T \xi_i),
\]

\[
T_2(\theta) = \sum_{i=1}^{n} \{ E[\eta(G_i, 1)] \ln(p_{\pi_i}) + E[\eta(G_i, 2)] \ln(p_{\pi_i}) + E[\eta(G_i, 3)] \ln(p_{\pi_i}) \}
\]

and

\[
T_3(\theta) = -\frac{1}{2} n \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} E(\xi_i^T \Sigma^{-1} \xi_i).
\]
The EM algorithm requires: (1) finding the partial derivatives of \(l(\theta)\) with respect to the parameters, (2) setting these partial derivatives equal to zero and (3) searching for explicit solutions for the parameters as a function of the missing values. This completes the maximization step. The expectation step requires taking the expectations for all terms that involve the missing values. Since the complete-data likelihood function is just a likelihood function for a typical mixed model, we can simply utilize the existing mixed model EM algorithm to find the MLE of parameters (HENDERSON 1986). The following are the EM steps we implement for the mixed model analysis.

**Step 0**: Set \(\zeta = 0\) and initialize all parameters with values in their legal domain, denoted by \(\theta^{(\zeta)}\).

**Step 1 (E1)**: Compute the posterior probabilities of the three genotypes for each individual

\[
p(G_i = k \mid I_M, y_i) = E[\eta(G_i, k)] = \frac{\pi_{ik} p(y_i \mid G_i = k)}{\sum_{k'} \pi_{ik'} p(y_i \mid G_i = k')} \text{ for } k = 1, 2, 3.
\]

**Step 2 (E2)**: Find the posterior distribution of the random effect \(\xi_i\). This posterior distribution turns out to be a mixture of three normal distributions with a mean

\[
E(\xi_i) = \Sigma \psi V^{-1} \left( y_i - \psi^T \mu - E(x_i) \psi^T \alpha - E(z_i) \psi^T \delta \right)
\]

and a covariance matrix

\[
\text{var}(\xi_i) = \Sigma - \Sigma \psi V^{-1} \psi^T \Sigma.
\]

All parameters appearing in the right hand side of the equations should be substituted by their most current values of the parameters at iteration \(t\). This statement also holds for subsequent steps.
Step 3 (E3): Compute all the expectations involved in the following maximization steps (see the next paragraph for detailed expressions of the expectations).

Step 4 (M1): Update the population mean

\[
\mu^{(\zeta+1)} = \left( n \psi V^{-1} \psi^\top \right)^{-1} \psi V^{-1} \sum_{i=1}^n \left( y_i - E(x_i) \mu - E(x_i, z_i) \psi^\top \delta \right).
\] (30)

Step 5 (M2): Update the additive effect

\[
\alpha^{(\zeta+1)} = \left( \sum_{i=1}^n E(x_i^2) \psi V^{-1} \psi^\top \right)^{-1} \psi V^{-1} \sum_{i=1}^n \left( E(x_i) y_i - E(x_i) \mu - E(x_i, z_i) \psi^\top \delta \right).
\] (31)

Step 6 (M3): Update the dominance effect

\[
\delta^{(\zeta+1)} = \left( \sum_{i=1}^n E(z_i^2) \psi V^{-1} \psi^\top \right)^{-1} \psi V^{-1} \sum_{i=1}^n \left( E(z_i) y_i - E(z_i) \mu - E(x_i, z_i) \psi^\top \alpha \right).
\] (32)

Step 7 (M4): Update the covariance matrix of the random effect

\[
\Sigma^{(\zeta+1)} = \frac{1}{n} \sum_{i=1}^n E(\xi_i \xi_i^\top) = \frac{1}{n} \sum_{i=1}^n \left[ \text{var}(\xi_i) + E(\xi_i) E(\xi_i^\top) \right].
\] (33)

Step 8 (M5): Update the residual variance

\[
\sigma^2(\zeta+1) = \frac{1}{n(m+1)} \sum_{i=1}^n y_i^\top \left( y_i - \psi^\top \mu - E(x_i) \psi^\top \alpha - E(z_i) \psi^\top \delta - \psi^\top E(\xi_i) \right).
\] (34)

Step 9: Increment \( \zeta \) by one, i.e., let \( \zeta = \zeta + 1 \), and repeat Steps 1 to 8 (three E-steps and five M-steps) until \( \| \theta^{(\zeta+1)} - \theta^{(\zeta)} \| < 10^{-8} \) holds.

In this paragraph, we provide explicit expressions for the expectations of various terms involved in the EM steps. The expectation of the quadratic term of the random effects is

\[
E(\xi_i, \xi_i^\top) = \text{var}(\xi_i) + E(\xi_i) E(\xi_i^\top),
\] (35)

where \( E(\xi_i) \) and \( \text{var}(\xi_i) \) are already given in Step 2 (eqs. 28 and 29). We only need to provide the expectations related to the genotype indicator variables. These expectations are
\[
\begin{align*}
E(x_i) &= E[\eta(G_i,3)] - E[\eta(G_i,1)] \\
E(z_i) &= E[\eta(G_i,2)] - [E[\eta(G_i,1)] + E[\eta(G_i,3)]] \\
E(x_i z_i) &= E[\eta(G_i,3)] - E[\eta(G_i,1)] \\
E(x_i^2) &= E[\eta(G_i,1)] + E[\eta(G_i,3)] \\
E(z_i^2) &= E[\eta(G_i,1)] - E[\eta(G_i,2)] + E[\eta(G_i,3)] = 1 
\end{align*}
\]

(36)

The EM algorithm was implemented via a FORTRAN90 program, which is available from the authors on request.

**Hypothesis tests**

Several important hypothesis tests are interesting and can be performed via the likelihood ratio test statistics. The first hypothesis is “no segregation of a QTL at the current position,” which is denoted by \( H_0 : \alpha = \delta = 0 \). The test statistic for this hypothesis is

\[
\hat{\lambda}_0 = -2 \left[ \ln L(\hat{\theta}_0) - \ln L(\hat{\theta}) \right],
\]

(37)

where \( \ln L(\hat{\theta}) \) is the log likelihood function evaluated at \( \hat{\theta} = \{\hat{\mu}, \hat{\alpha}, \hat{\delta}, \hat{\Sigma}, \hat{\sigma}^2\} \) and \( \ln L(\hat{\theta}_0) \) is the log likelihood under the null model evaluated at \( \hat{\theta}_0 = \{\hat{\mu}, 0, 0, \hat{\Sigma}, \hat{\sigma}^2\} \). Note that \( \hat{\theta}_0 \) differs from \( \hat{\theta} \) in two aspects: (1) \( \hat{\theta}_0 \) has a lower dimension because \( \alpha = \delta = 0 \) has been enforced, (2) the three parameter sets in \( \hat{\theta}_0 \) are estimated from the null model and they are different from the counterparts estimated under the full model. The EM algorithm for estimating \( \theta_0 \) is a simple special case of the EM algorithm for the full model. Under the null hypothesis, \( \lambda_0 \) will follow approximately a chi-square distribution with \( \text{df} = \dim(\theta) - \dim(\theta_0) = 2(r+1) \), where \( \dim(\theta) \) and \( \dim(\theta_0) \) are the dimensions of \( \theta \) and \( \theta_0 \), respectively. These two vectors differ by \( \alpha \) and \( \delta \), each having a dimension of \( r+1 \), which explains why \( \text{df} = 2(r+1) \).
When \( H_0 \) is rejected, we can further test whether the QTL effects are due to additive or dominance effects. First, we can test the departure from additivity, i.e., the dominance effect. The null hypothesis of “no dominance effect” is \( H_1: \delta = 0 \) and the test statistic is

\[
\lambda_1 = -2 \left[ \ln L(\hat{\theta}_1) - \ln L(\hat{\theta}) \right],
\]

where \( \hat{\theta}_1 = \{\hat{\mu}, \hat{\alpha}, 0, \hat{\Sigma}, \hat{\sigma}^2\} \) is the parameter vector estimated under \( H_1 \).

To test the additive effect, our null hypothesis becomes “no additive effect,” denoted by \( H_2: \alpha = 0 \). The test statistic is

\[
\lambda_2 = -2 \left[ \ln L(\hat{\theta}_2) - \ln L(\hat{\theta}) \right],
\]

where \( \hat{\theta}_2 = \{\hat{\mu}, 0, \hat{\delta}, \hat{\Sigma}, \hat{\sigma}^2\} \). Each one of \( \lambda_1 \) and \( \lambda_2 \) will approximately follow a chi-square distribution with \( r + 1 \) degrees of freedom.

It is also possible to test other hypotheses, such as whether a polynomial of order \( r + 1 \) is sufficiently better than that of order \( r \) and whether the structured \( V \) matrix sufficiently explains the covariances of the experimental errors compared to an unstructured \( V \) matrix. We defer these hypothesis tests to the sections on data analysis and discussion.

So far we have only discussed hypothesis tests at one location of the genome. To scan the entire genome for QTL, we test every putative position of the genome with a one or two cM increment and plot the test statistic against the genome location to form a test statistic profile. Positions that correspond to significant peaks of the profile provide estimates of the QTL locations. The critical value of the test statistic used to declare significance is obtained via the permutation test (Churchill and Doerge 1994).
DATA ANALYSIS

Simulation studies

We simulated an F\textsubscript{2} population under several different sample sizes. A single chromosome of 200 cM was simulated with 21 evenly spaced markers (10 cM per interval). The position of the first marker was designated as position zero and positions of all other markers were measured as the distances from the first marker. We put three QTL at positions 23 cM, 94 cM and 187 cM, respectively. The range of the original time points \((t)\) was [1,150]. The cumulative heritability of the three QTL were 0.24, 0.24 and 0.15, respectively, each of which was defined as

\[ h^2 = \int_{0}^{\tau} h^2(\tau) d\tau, \quad (40) \]

where \( h^2(\tau) \) is the proportion of the phenotypic variance explained by the QTL at time point \( \tau \) [see equation (4) for the definition of \( h^2(\tau) \)]. The QTL effects that generate such values of \( h^2 \) were found by trial and error and are given in Table 1 with Legendre polynomial of order three \((r = 3)\). The simulated population mean was

\[ \mu = [45 \text{ } 44 \text{ } -1 \text{ } -7]^T. \]

The functional curve for the mean was \( \mu(\tau) = \psi(\tau)\mu \). The functions of population mean, additive effect, dominance effect and heritability of the three QTL plotted against time are given in Figure 1(a), (b), (c) and (d), respectively. The individual specific and time-dependent random effect was simulated using \( \xi_i(\tau) = \psi(\tau)\xi_i \), where \( \xi_i \) was generated from a \( N(0, \Sigma) \) distribution with
\[ \Sigma = \begin{bmatrix}
1.042 & 0.171 & -0.315 & 0.100 \\
0.171 & 1.586 & -0.035 & 0.041 \\
-0.315 & -0.035 & 1.736 & 0.287 \\
0.100 & 0.041 & 0.287 & 0.772
\end{bmatrix}, \]

which is a \((3+1) \times (3+1)\) positive definite covariance matrix. The residual variance was \(\sigma^2 = 2.0\). The model to generate the phenotypic values was

\[ y_i(\tau) = \psi(\tau)\mu + \sum^3_{i=1} [x_i\psi(\tau)\alpha + z_i\psi(\tau)\delta] + \psi(\tau)\xi_i + \varepsilon_i, \quad (41) \]

where the summation indicated the sum of three QTL.

A F_2 population was simulated with three different sample sizes, \(n = 50, 80, 100\). The frequency of measurement of the trait was simulated at two levels, \(m = 5, 10\). When the trait was measured 5 times, the interval between two consecutive measurements was 30 days, whereas \(m = 10\) corresponded to an interval of measurement of 15 days. All six combinations of \(n\) and \(m\) were simulated for comparison. Each of the six combinations (scenarios) was replicated 100 times.

The critical values used to declare statistical significance for QTL in the simulation experiments were obtained empirically by simulating 500 additional replicates under the null model where all QTL effects were set to zero. The empirical critical values under the six scenarios are given in Table 2.

The average estimated QTL positions, the likelihood ratio test statistics and the empirical powers of the three QTL are given in Tables 3. The estimated QTL effects (including additive and dominance effects) are given in Table 4. The purpose of the simulations was neither to select the optimal design of the experiment nor to choose the
optimal model of QTL mapping; rather, it was to demonstrate the general behavior of the mapping procedure and validate the computer program. Therefore, the simulations were not exhaustive. The results of the simulations did show the expected behaviors of a mapping procedure: (1) the power increases as the sample size increases, (2) the standard error of each estimated parameter decreases as the sample size increases, (3) the increased power and decreased error are also associated with the increased frequency of the trait measurements. Therefore, the algorithm converges as expected and the computer program is logically valid. The average likelihood ratio profiles of the 100 replicates under $n = 80$ and $m = 10$ are demonstrated in Figure 2. The three peaks in the likelihood ratio profile clearly demonstrated the efficiency of the method for scanning multiple QTL. Again, we are not looking at model selection here; rather, we are trying to verify the model fit (WILTSHIRE et al. 2005). We developed a single QTL model but implemented the method with data containing multiple QTL. This is a typical approach for genome scanning with interval mapping (LANDER and BOTSTEIN 1989). The estimated positions and effects of the three simulated QTL were very close to the true values (see Table 4). This demonstrated the robustness of the single QTL model for mapping multiple additive QTL. The robustness may be credited to the covariance matrix $\Sigma$ included in the model. When QTL1 was evaluated, QTL2 and QTL3 were not taken into account by the model. These two neglected QTL would ordinarily cause extra correlation among the repeated measurements. The extra correlation, however, were absorbed by matrix $\Sigma$. Therefore, the estimated $\mu$, $\Sigma$ and $\sigma^2$ from the single QTL model were not comparable to the true values of $\mu$, $\Sigma$ and $\sigma^2$ were used to simulate the three-QTL-controlled longitudinal trait.
Normally, one would expect to see some bias in the estimated effects associated with the Legendre polynomials due to the Beavis’ effect (Beavis 1994; Xu 2003). The lack of such biases in the simulation experiment is due to the fact that we reported all effects, regardless their significance of the test. The Beavis’ effect would be expected to happen if the mean effects were calculated based on a censored sample (containing only the replicates in which significant QTL were detected).

To verify the efficiency of the method for estimating $\mu$, $\Sigma$ and $\sigma^2$, we simulated an additional dataset with exactly the same setup as the original simulation experiment except that we now only kept the first QTL (see Table 1 for the parameters of QTL1) in the simulation under $n = 80$ and $m = 10$. The experiment was replicated 100 times. The average estimated $\sigma^2$ was $2.081 \pm 0.092$, close to the true value of 2.0. The average estimated $\mu$ and $\Sigma$ were

$$\hat{\mu} = \begin{bmatrix} 44.95 \pm 0.18 \\ 44.05 \pm 0.12 \\ -0.997 \pm 0.031 \\ -7.06 \pm 0.086 \end{bmatrix}$$

and

$$\hat{\Sigma} = \begin{bmatrix} 1.062 \pm 0.024 & 0.198 \pm 0.020 & -0.364 \pm 0.025 & 0.125 \pm 0.020 \\ 0.198 \pm 0.020 & 1.533 \pm 0.034 & -0.043 \pm 0.020 & 0.042 \pm 0.024 \\ -0.364 \pm 0.025 & -0.043 \pm 0.020 & 1.718 \pm 0.058 & 0.243 \pm 0.031 \\ 0.125 \pm 0.020 & 0.042 \pm 0.024 & 0.243 \pm 0.031 & 0.795 \pm 0.035 \end{bmatrix},$$

respectively, where the numbers following $\pm$ are the standard deviations obtained from the 100 replicates. The estimated $\mu$ and $\Sigma$ were also close to the true values of $\mu$ and $\Sigma$. Overall, the simulation experiment demonstrated that we have developed a working procedure of QTL mapping for longitudinal traits.
For data generated by the single QTL model, we also performed the logistic regression analysis using FunMap (Ma et al. 2002; Ma et al. 2004). Unfortunately, no QTL was detected in any of the replicates (see Figure 3 for the comparison of the two methods). This comparison is not quite fair because the method of Ma et al. (2002) was not designed to handle curves of arbitrary shape with a general covariance structure. This only demonstrates that extreme departure from the assumption of Ma et al. (2002) will cause the method they developed to fail, and so experimentalists should take care in its application. A fair comparison is deferred to the real data analysis.

We now demonstrate that, with a little modification, the method can handle models with epistatic (interactive) effects of QTL. Wu et al. (2004) recently extended their logistic model to map epistatic effects and published the epistatic model alone in a separate study. For simplicity of presentation, we simulated a BC population of 100 individuals with test frequencies of 5 or 10 for the period of 150 days. A single chromosome of 100 cM was simulated and 11 markers were placed evenly on the chromosome (10 cM per interval). Two QTL were simulated at positions 23 cM and 75 cM, respectively. The effects of these two QTL were defined under the following two models: (1) Additive plus epistatic effect model (A+E model), in which both QTL have their own additive effects and they also act interactively to determine the phenotype of the trait, and (2) Epistatic effect only model (E model), in which the two QTL act interactively to determine the phenotype and neither QTL has its own additive effect. These two models were used to generate the data, but the analytical model always assumed the presence of both the additive and epistatic effects and thus always presented estimated values for both types of effects. The epistatic
The estimated positions of the two QTL were obtained in a two-dimension grid search for the maximum value of the likelihood function. Once the positions were found, the estimated QTL effects corresponding to the optimal positions were reported as the MLE of the QTL effects. The critical value used for declaration of statistical significance was obtained in the same way as those in the previous simulation experiment.

Table 5 shows the estimated positions of the two QTL along with the empirical statistical powers. The A + E model reached 100% power in both test frequencies examined.
The E model, however, has less power. The estimated errors of parameters of the two models under different test frequencies also behave as expected. Table 6 gives the average estimated QTL effects, which were all close to the true values used to simulate the data. Overall, the simulation result has validated the method and the program.

**Example of real data analysis**

We used the data published by MA et al. (2002) as an example to demonstrate the application of the new method to real data from a longitudinal trait. The trait was the growth of the stem diameter of poplar trees measured annually for 11 years ($m = 11$). The mapping population consisted of 78 progeny of a pseudo backcross family ($n = 78$ progeny of a hybrid backcrossed to a parent of the hybrid). The method developed for F2 design was revised by including two genotypes only per locus. The conditional probability of the QTL genotype given marker information was calculated using the multipoint method of JIANG and ZENG (1997). The data contained markers of chromosome 10 only, and thus only one chromosome was scanned.

Before conducting QTL mapping, we took the least-squares method and fit the phenotypic values of diameter growth to the logistic curve and Legendre polynomials of orders 2, 3 and 4, respectively, for each individual. The averages goodness of fit ($R^2$) for the growth trajectories of 78 individuals showed that Legendre polynomial with $r \geq 3$ fit the data equally well as, or better than, the logistic curve. This justifies the orders chosen for the polynomial analysis in this study and the logistic analysis in MA et al. (2002). Theoretically, the polynomial approach and the logistic analysis can only be compared if the shape of the growth curve is sigmoid. When the order of the polynomial is $r \leq 2$, the
curve is linear or quadratic, which fits poorly to the tree diameter growth curve. The tree diameter growth curve appeared to be s-shaped and thus a valid comparison of the polynomial and the logistic analyses can be made.

There are a maximum of 10 different orders of the polynomial to choose for the time points. We fit each of the orders and chose the one with the minimum BIC (Bayesian Information Criterion, SCHWARZ 1978). The BIC for order $r$ was calculated using

$$BIC(r) = -2 \ln L(\hat{\theta}|r) + \text{dim}(\hat{\theta}|r) \ln[n(m+1)]$$

where $\hat{\theta}|r$ is the MLE of parameters for order $r$ and $\text{dim}(\hat{\theta}|r)$ is the dimension of $\theta$ (the number of independent parameters in the model). The BIC scores for the Legendre polynomial of orders 2, 3, 4 and 5 are 1332.6, 920.3, 768.0 and 809.6, respectively. Clearly, $r = 4$ is the best order. Therefore, all subsequent reports are based on the result of $r = 4$ Legendre polynomial. Results of fitting other orders of the polynomial were not reported due to space limitations. However, the results of $r = 3$ and 5 were very close to that of $r = 4$ (data not shown).

Figure 4 gives the likelihood ratio test statistic profile for chromosome 10. Two peaks are evident in the profile. We conducted a permutation test (with 500 randomly reshuffled samples) to determine the empirical critical values for significance declaration. Both peaks passed the critical values, and thus both QTL were significant. The positions of the two QTL were estimated at 32 cM and 54 cM, respectively. The two QTL are designated as QTL32 and QTL54, respectively. The estimated QTL effects were

$$\hat{\alpha} = [-0.269, -0.419, -0.438, -0.056, 0.193]^T$$
for QTL32 and
\[ \hat{\alpha} = [-0.346 \quad -0.530 \quad -0.436 \quad -0.054 \quad 0.198]^T \]
for QTL54. The QTL effect functions are expressed as
\[ \hat{\alpha}(\tau) = \psi(\tau)\hat{\alpha} = -0.269 - 0.419\psi_1(\tau) - 0.438\psi_2(\tau) - 0.056\psi_3(\tau) + 0.193\psi_4(\tau) \]
for QTL32 and
\[ \hat{\alpha}(\tau) = \psi(\tau)\hat{\alpha} = -0.346 - 0.530\psi_1(\tau) - 0.436\psi_2(\tau) - 0.054\psi_3(\tau) + 0.198\psi_4(\tau) \]
for QTL54. Note that dominance effects cannot be separated from the additive effects in a backcross design. Therefore, the additive effects estimated are actually confounded with the dominance effects. The functions of the two QTL effects are depicted in Figure 5, which show similar dynamic patterns of change. Both QTL appeared to be inactive until year seven and then the effects increased until year 10 when a plateau is reached.

For comparison, we analyzed the same data with the method of MA et al. (2002) using FunMap (MA et al. 2004), a public program for mapping longitudinal traits with the logistical regression method. The covariance structure specified in the program was AR(1), first order autoregressive model. The likelihood ratio profile is given in Figure 6 (the light curve). Only one peak passed the critical value and the location was at 14 cM. We calculated the BIC value for this analysis and found that the BIC was 915.8, higher than 768.0, which is the BIC of the Legendre polynomial of order 4. Therefore, the two different methods seem to generate somewhat different results.

Because the difference between the two methods may be due to the different covariance structures specified, we then modified our method by using the AR(1) covariance structure but still fit the QTL effects to the Legendre polynomials. We found that the minimum BIC value still occurred when \( r = 4 \) (the BIC score was 906.0). This value is larger than 768.0,
which is the BIC of the Legendre polynomial of order 4 with the general covariance structure
\( (\psi^T \Sigma \psi + \mathbf{I} \sigma^2) \). The likelihood ratio test statistic profile is depicted in Figure 6 (the dark curve). The profile has exactly the same shape as that of Ma et al. (2002). This showed that the difference in results from the two methods was partially due to the difference in the covariance structure. Judging by the BIC values, we may conclude that this hybrid method of Legendre polynomial with AR(1) covariance structure is slightly better than the method of Ma et al. (2002).

DISCUSSION

We presented a polynomial approach to QTL mapping for longitudinal traits as opposed to the logistic regression approach (Ma et al. 2002). Both methods may be considered as dimension reduction techniques for multivariate analysis. With the polynomial approach, the dimension of the model can vary according to the model goodness of fit to the actual data. The BIC value was used to determine the optimal dimension. As a result, the polynomial model is more general and flexible than the logistic regression approach, which has a fixed dimension of three. More importantly, the polynomial model is linear in parameters, which allows well developed linear model methodology to be fully utilized in the longitudinal trait study. Both methods benefit more when the trait is measured more frequently. There is another method for functional data analysis that is quite similar to the polynomial approach, namely, the method of B-splines (De Boor 2001). The B-splines are more commonly used in non-parametric data analysis to infer the empirical distribution of a variable, and is rarely applied to quantitative genetics. We think that the polynomial approach is more intuitive and easy to understand. Once the
polynomial approach is accepted by geneticists, we may further explore the B-splines approach and compare the two competing but similar methods.

For simplicity, we fit all model effects to the same basis (order) of the polynomials. In fact, different types of model effects may have different functional relationships with time. For example, the population mean may be linear in time, whereas the additive effect may be quadratic or cubic in time. This can be taken into account by describing each type of model effect by a polynomial with its own basis. The model with this flexibility may look like

$$y_i(\tau) = \psi(\tau | r_\mu) \mu + x_i \psi(\tau | r_\alpha) \alpha + z_i \psi(\tau | r_\beta) \beta + \psi(\tau | r_\xi) \xi_i + e_i$$

(44)

where $\psi(\tau | r_\mu) = [\psi_0(\tau) \ldots \psi_{r_\mu}(\tau)]$ is the basis with order $r_\mu$, which may differ from $r_\alpha$ and the orders of other effects. Allowing the random effects to have a different basis from other effects ($r_\xi \neq r_\mu = r_\alpha = r_\beta$) may be very desirable because this will introduce a different structured residual covariance matrix. Further discussion on the structured covariance matrix is given later.

If a major gene model sufficiently describes the data, the residual covariance matrix should have a simple form, e.g., $V = \text{diagonal}$ or simply $V = I \sigma^2$. Polygenic effects and other factors not included in the model may cause a more dense structure of the covariance matrix. Wu et al. (2002b, 2003, 2004) used AR(1), the first order autoregressive covariance matrix, because AR(1) is one of the simplest structures that allows the authors to develop a nice EM algorithm. Extension to more complicated structures is either impossible or would take substantial effort. With the polynomial analysis, however, we are able to use
\( V = \psi^T \Sigma \psi + I \sigma^2 \), a general structured covariance matrix. We can choose the dimension of \( \xi_i \) as low as zero so that \( V = I \sigma^2 \) and as high as \( m + 1 \) so that \( V = \Sigma_{(m+1)(m+1)} \) (a fully unstructured covariance matrix). This flexibility cannot be enjoyed by the logistic growth curve fitting approach. The optimal dimension of \( \xi_i \) may be found by evaluating the BIC estimates of different dimensions. Again, there is no theoretical difficulty with the polynomial analysis.

Many aspects of agricultural experiments are subject to high levels of uncertainty. Not all plants may have full measurements at all the calibrated time points. Occasionally, a plant may fail to be measured for the phenotype at a certain time point due to oversight of the researchers. This is a typical missing value problem. Some plants may die before they are fully measured for the entire term of the experiment. This is a typical problem of longitudinal study (DIGGLE et al. 1994). The EM algorithm was particularly designed to handle problems like this. So, there is no additional technique required to handle these missing value problems. One should still include the missing phenotype (as a symbolic quantity) in the model but simply replace all terms involving the missing phenotype by their conditional expectations and proceed with the usual EM iterations. The conditional expectation for the term involving a missing phenotype, say value \( y_j \), should be calculated based on the parameters at the current iteration and a restriction \( y_{t(j-1)} \leq y_j \leq y_{t(j+1)} \), where \( y_{t(j-1)} \) and \( y_{t(j+1)} \) are observed values before and after time point \( j \). If the plant dies at time point \( j \) (longitudinal data analysis), the restriction should be \( y_{t(j-1)} \leq y_j \) for all \( j-1 < j \leq m+1 \). These restrictions only apply to growth traits. For other types of traits, say daily gain, such restrictions are not necessary.
An alternative approach for dealing with missing values may be more powerful than
the aforementioned one. Instead of using the conditional expectations of the terms related
to the missing phenotypes, we may only model the observed phenotypes and completely
ignore the missing phenotypes. The complication encountered in this alternative approach
is that the vector of phenotypic values for an individual with incomplete measurement has a
dimension less than \( m + 1 \). The general situation is that the dimension of vector \( y_i \) varies
across individuals. This will cause the dimension of matrix \( V \) to vary, which will be
reflected by using \( V_i \) in place of \( V \). There is no theoretical difficulty in incorporating the
property of variable dimension into the likelihood analysis. The model bearing this
flexibility may be

\[
y_i(\tau) = \varphi(\tau \mid i)\mu + x_i\varphi(\tau \mid i)\alpha + z_i\varphi(\tau \mid i)\delta + \phi(\tau \mid i)\xi_i + \epsilon_i
\]

(45)

where \( \varphi(\tau \mid i) \) now represents the basis of polynomial with order \( r \) but constructed using
only the time points with actual measurements for individual \( i \). With the polynomial basis
varying across individuals, we can fully take into account the missing value problem. This
approach is sufficiently general to handle even more complicated experiments, such as
different individuals are measured at different time points during the entire term of the
experiment. One should be cautious about standardizing the time point from the original
scale to the [-1, +1] scale (see APPENDIX B for the definition of standardized scale). The
initial and ending time points used should be constant across individuals. In other words,
one should use \( t_0 \) as the time that the measurement starts and \( t_m \) the time that the
measurement ends during the entire experiment for all individual plants. Other technical
problems may occur, which deserve further investigation. The primary objective of this
study was to lay the foundation for the polynomial approach to longitudinal trait study. Technical improvements, which will digress from the main focus, are deferred to subsequent studies.

We have demonstrated that if the growth curve is indeed sigmoid, the proposed polynomial analysis can still be used due to the great generality of the method. The flexibility of the polynomial analysis in terms of incorporating structured residual covariance matrices makes the method more preferable to the logistic approach, which only fits an RA(1) covariance structure. One criticism on the polynomial approach when applied to the sigmoid curve is the interpretability of the polynomial regression coefficients. In logistic models, each parameter has a biological meaning, whereas that is not the case when using a polynomial. However, a logistic curve can be described by a polynomial with order three \((r = 3)\). If a growth curve is indeed fitted with such a polynomial, we can find the functional relationships of the parameters in the logistic curve with respect to the parameters in the polynomial. By doing so, we can interpret the biological meanings of the polynomial parameters. We take the following approaches to infer the functional relationships between the logistic parameters and the polynomial parameters. First, we understand that the actual shape of a logistic curve is also determined by the initial value of the curve at time zero and the value of the curve at the inflection point. This implies that we can express the logistic parameters \((a, b, r)\) as functions of the two coordinates (the initial and the inflection points) because both determine the shape of the curve. We then find the functional relationships of the values of the two coordinates with respect to the polynomial parameters \((\beta_0, \beta_1, \beta_2, \beta_3)\). Combining the two steps, we can eventually find the
relationships between parameters of the two types of curves, i.e., \( a = f_a(\beta_0, \beta_1, \beta_2, \beta_3), \)
\( b = f_b(\beta_0, \beta_1, \beta_2, \beta_3) \) and \( r = f_r(\beta_0, \beta_1, \beta_2, \beta_3). \) Derivations of the three functions are provided in APPENDIX C. We now have expressed the logistic parameters as functions of the polynomial parameters. Although the biological meaning of each polynomial regression coefficient alone is not obvious, different combinations (functions) of the polynomial parameters determine the parameters of the logistic curve, which are well interpretable biologically.

ACKNOWLEDGMENTS

We are grateful to two anonymous reviewers for their comments and suggestions, which greatly assisted the revision of the manuscript. The research was supported by the National Institute of Health Grant R01-GM55321, the National Science Foundation Grant DBI-0345205 to SX and the Chinese National Natural Science Foundation Grant 30471236 to RY.
APPENDIX A
DERIVATIONS OF QTL VARIANCES IN A F2 POPULATION

The linear model for the phenotypic value at time \( \tau \) is

\[
y_i(\tau) = \mu(\tau) + x_i\alpha(\tau) + z_i\delta(\tau) + \xi_i(\tau) + \epsilon_i.
\]  
(A1)

Definitions of the symbols are given in the main text, see equation (1). When \( \alpha(\tau) \) and \( \delta(\tau) \) are treated as fixed effects, the genetic effect,

\[
g_i(\tau) = x_i\alpha(\tau) + z_i\delta(\tau),
\]  
(A2)

has a variance of

\[
\text{var}[g_i(\tau)] = \text{var}(x_i)\alpha^2(\tau) + \text{var}(z_i)\delta^2(\tau) + 2\text{cov}(x_i, x_i)\alpha(\tau)\delta(\tau). 
\]  
(A3)

Define \( \sigma^2_g(\tau) = \text{var}[g_i(\tau)] \), \( \sigma^2_x = \text{var}(x_i) \) and \( \sigma^2_z = \text{var}(z_i) \) as the theoretical variances of the effects in question across individuals within the F2 population. Under the assumption of no segregation distortion, the ratio of the three genotypes (\( AA \), \( Aa \) and \( aa \)) is expected to be \( 1:2:1 \). Based on the definitions of variables \( x \) and \( z \) presented in the text, we can show that

\[
\sigma^2_x = E(x^2) - E^2(x) = \left[\frac{1}{4}(1)^2 + \frac{1}{2}(0)^2 + \frac{1}{4}(-1)^2\right] - \left[\frac{1}{4}(1) + \frac{1}{2}(0) + \frac{1}{4}(-1)\right]^2 = \frac{1}{2}, \]  
(A4)

\[
\sigma^2_z = E(z^2) - E^2(z) = \left[\frac{1}{4}(-1)^2 + \frac{1}{2}(1)^2 + \frac{1}{4}(-1)^2\right] - \left[\frac{1}{4}(-1) + \frac{1}{2}(1) + \frac{1}{4}(-1)\right]^2 = 1 \]  
(A5)

and

\[
\text{cov}(x, z) = E(xz) - E(x)E(z) = \left[\frac{1}{4}(1)(-1) + \frac{1}{2}(0)(1) + \frac{1}{4}(-1)(-1)\right] - \left[\frac{1}{4}(1) + \frac{1}{2}(0) + \frac{1}{4}(-1)\right] \left[\frac{1}{4}(-1) + \frac{1}{2}(1) + \frac{1}{4}(-1)\right]. \]  
(A6)

Therefore, the genetic variance contributed by the major gene at time \( \tau \) is

\[
\sigma^2_g(\tau) = \sigma^2_x\alpha^2(\tau) + \sigma^2_z\delta^2(\tau) = \frac{1}{2}\alpha^2(\tau) + \delta^2(\tau). \]  
(A7)
The particular values of \( \sigma_x^2 = \frac{1}{2} \) and \( \sigma_z^2 = 1 \) defined for a F2 population are due to the special way we defined the values of variables \( x \) and \( z \) for the three genotypes. One can define the values of \( x \) and \( z \) in any arbitrary ways without affecting the results of the analysis. The way we chose for the definitions of \( x \) and \( z \) is adopted from Xu (1998). In fact, a mathematically more attractive way to define \( x \) and \( z \) is \( x = \{ \sqrt{2}, 0, -\sqrt{2} \} \) and \( z = \{-1,1,-1\} \) for genotypes \( \{ AA, Aa, aa \} \). The \( x \) defined this way appears to be odd, which explains why this system has not been adopted in the literature. However, one can prove that \( \sigma_x^2 = \sigma_z^2 = 1 \) and \( \text{cov}(x,z) = 0 \), leading to

\[
\sigma_x^2(\tau) = \alpha^2(\tau) + \delta^2(\tau), \tag{A8}
\]

which is the mathematical attractiveness of the new system. One should be cautious in adopting a new system for the definitions of \( x \) and \( z \). If \( x \) and \( z \) are not orthogonal, then \( \text{cov}(x,z) \neq 0 \), must be incorporated into the formula of genetic variance. Furthermore, if segregation distortion has occurred, the theoretical variances and covariance of variables \( x \) and \( z \) are not known. One may estimate the variances and covariance from the observed data and use the estimated values instead.
APPENDIX B
BASIS OF THE ORTHOGONAL POLYNOMIAL

Orthogonal polynomial analysis is a general way to describe the functional relationship of one variable relative to another variable. The dependent variable is usually stochastic and follows some kind of distribution. The independent variable is usually calibrated by some fixed time points, which are controlled by the investigator. One can choose the order of the polynomial as high as the total number of fixed time points. The model can fully describe the data with no error, but such a model has no general use. In contrast, one can choose the order of the polynomial as low as 0, leading to a constant, again a trivial result. If one chooses the order of 1, the functional relationship becomes linear in time. The order of 2 describes a quadratic function and so on. The rule of thumb is to choose the order as low as possible but the model with this order must be able to describe the data sufficiently. In biological science, most curves may be described by the order of less than five. A curve with an order greater than five is difficult to interpret.

The first step in the orthogonal polynomial analysis is to convert the original time point \( t \) into a standardized time point \( \tau \) using the following expression,

\[
\tau = 2 \times \frac{t - t_0}{t_m - t_0} - 1, \quad \tau \in [-1, +1],
\]

(B1)

where, \( t_0 \) is the initial time point and \( t_m \) is the ending time point. Therefore, time point \( \tau \) used in the entire text is the standardized time point. Recall that if we choose a basis with order \( r \), then

\[
\psi(\tau) = [\psi_0(\tau) \quad \psi_1(\tau) \quad \cdots \quad \psi_r(\tau)].
\]

(B2)
The $k$th ($0 \leq k \leq r$) component of $\psi(\tau)$ is defined as

$$
\psi_k(\tau) = \frac{1}{2^l} \sum_{l=0}^{\text{int}(k/2)} (-1)^l (2k - 2l)! \frac{(k-2l)!}{l!(k-l)!(k-2l)!} \tau^{k-2l},
$$

(B3)

where $\text{int}(k/2)$ is an integer function. The first five terms of the orthogonal polynomial are $\psi_0(\tau) = 1$, $\psi_1(\tau) = \tau$, $\psi_2(\tau) = \frac{1}{2} (3\tau^2 - 1)$, $\psi_3(\tau) = \frac{1}{2} (5\tau^3 - 3\tau)$ and $\psi_4(\tau) = \frac{1}{3} (34\tau^4 - 10\tau^2 + 1)$, respectively (SCHAEFFER 2004).
APPENDIX C

RELATIONSHIP BETWEEN LOGISTIC PARAMETERS AND POLYNOMIAL PARAMETERS

The logistic curve is

\[ y(t) = \frac{a}{1 + be^{-rt}} \quad \text{for} \quad t \in [t_0, t_m], \quad (C1) \]

where \( t_0 = 0 \) is chosen for convenience. The initial coordinate is \((t_0, y_0) = (0, \frac{a}{1+b})\) and the coordinate at the inflection point is \((t_\zeta, y_\zeta) = (\frac{\ln b}{r}, \frac{a}{2})\), which is found by setting \(\frac{d^2 y(t)}{dt^2} = 0\) and solving for \(t\). Let \(t_\zeta\) be the solution for equation \(\frac{d^2 y(t)}{dt^2} = 0\), we obtain \(y_\zeta = y(t_\zeta)\) using equation (C1). Given \((y_0, t_\zeta, y_\zeta)\), we now have three equations with three unknowns (logistic parameters),

\[
\begin{cases}
\frac{1}{1+b} = y_0 \\
\ln b = t_\zeta \\
\frac{a}{2} = y_\zeta 
\end{cases}, \quad (C2)
\]

The solutions of equations (C2) are

\[
\begin{cases}
a = 2y_\zeta \\
b = \frac{2y_\zeta}{y_0} - 1 \\
r = \frac{1}{t_\zeta} \ln\left(\frac{2y_\zeta}{y_0} - 1\right)
\end{cases}, \quad (C3)
\]

We now evaluate the polynomial curve of order three,

\[ y(\tau) = \beta_0 + \beta_1 \tau + \frac{1}{2} \beta_2 (3\tau^2 - 1) + \frac{1}{6} \beta_3 (5\tau^3 - 3\tau) \quad \text{for} \quad \tau \in [-1, +1]. \quad (C4) \]

The coordinate of the initial point is \((\tau_0, y_0) = (-1, \beta_0 - \beta_1 + \beta_2 - \beta_3)\), i.e.,

\[ y_0 = \beta_0 - \beta_1 + \beta_2 - \beta_3. \quad (C5) \]

The second derivative of \(y(\tau)\) with respect to \(\tau\) is
\[
\frac{\partial^2}{\partial \tau^2} y(\tau) = 3\beta_2 + 15\beta_3 \tau.
\]  
(C6)

Setting \( \frac{\partial^2 y(\tau)}{\partial \tau^2} = 0 \) and solving for \( \tau \), we get \( \tau_c = -\frac{\beta_2}{5\beta_3} \). Substituting \( \tau_c \) into equation (C4), we have

\[
y_c = \beta_0 + \beta_1 \tau_c + \frac{1}{2} \beta_2 (3\tau_c^2 - 1) + \frac{1}{2} \beta_3 (5\tau_c^3 - 3\tau_c)
\]

\[
= \beta_0 - \beta_1 \left( \frac{\beta_2}{5\beta_3} \right) + \frac{1}{2} \beta_2 \left[ 3 \left( \frac{\beta_2}{5\beta_3} \right)^2 - 1 \right] + \frac{1}{2} \beta_3 \left[ 3 \left( \frac{\beta_2}{5\beta_3} \right) - 5 \left( \frac{\beta_2}{5\beta_3} \right)^3 \right].
\]  
(C7)

We now convert \( \tau_c \) (time point in the standardized scale) back into \( t_c \) (time point in the original scale) by

\[
t_c = \frac{1}{2} (t_m - t_0)(\tau_c - 1) + t_0 = \frac{1}{2} t_m (\tau_c - 1) = -\frac{1}{2} t_m \left( \frac{\beta_2}{5\beta_3} + 1 \right).
\]  
(C8)

The simplification is due to \( t_0 = 0 \). Substituting \( y_0 \) (equation C5), \( y_c \) (equation C7) and \( t_c \) (equation C8) into equations (C3), we obtain

\[
\begin{align*}
a &= 2\beta_0 - 2\beta_1 \left( \frac{\beta_2}{5\beta_3} \right) + \beta_2 \left[ 3 \left( \frac{\beta_2}{5\beta_3} \right)^2 - 1 \right] + \beta_3 \left[ 3 \left( \frac{\beta_2}{5\beta_3} \right) - 5 \left( \frac{\beta_2}{5\beta_3} \right)^3 \right] \\
b &= \frac{2\beta_0 - 2\beta_1 \left( \frac{\beta_2}{5\beta_3} \right) + \beta_2 \left[ 3 \left( \frac{\beta_2}{5\beta_3} \right)^2 - 1 \right] + \beta_3 \left[ 3 \left( \frac{\beta_2}{5\beta_3} \right) - 5 \left( \frac{\beta_2}{5\beta_3} \right)^3 \right]}{\beta_0 - \beta_1 + \beta_2 - \beta_3} - 1 \\
r &= -\frac{2}{t_m \left( \frac{\beta_2}{5\beta_3} + 1 \right)} \log_e \left\{ \frac{2\beta_0 - 2\beta_1 \left( \frac{\beta_2}{5\beta_3} \right) + \beta_2 \left[ 3 \left( \frac{\beta_2}{5\beta_3} \right)^2 - 1 \right] + \beta_3 \left[ 3 \left( \frac{\beta_2}{5\beta_3} \right) - 5 \left( \frac{\beta_2}{5\beta_3} \right)^3 \right]}{\beta_0 - \beta_1 + \beta_2 - \beta_3} - 1 \right\}
\end{align*}
\]  
(C9)
LITERATURE CITED


GUO, Z., and L. R. SCHAEFFER, 2002 Random Regression Submodel Comparison. 7th WCGALP, France.


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


MISZTAL, I., T. STRABEL, J. JAMROZIK, J. E. A. MANTYSAARI, and T. H. MEUWISSEN, 2000 Strategies for estimating the parameters needed for different test day models. J.


SCHAEFFER, L. R., 2001 Animal models. Courses at CGIL, University of Guelph.

SCHAEFFER, L. R., and J. C. M. DEKKERS, 1994 Random regressions in animal models for test-day production in dairy cattle. Proc. 5th WCGALP, Canada.


Wiltshire, S., A. P. Morris, M. I. McCarthy and L. R. Cardon, 2005 How useful is the fine-scale mapping of complex trait linkage peaks? Evaluating the impact of additional microsatellite genotyping on the posterior probability of linkage. Genet. Epidemiol. 28:1-10


FIGURE LEGENDS

**Figure 1.** The simulated parameters of QTL1, QTL2 and QTL3 expressed as functions of time: (a) Population means $\mu(\tau)$, (b) Additive effects $\alpha(\tau)$, (c) Dominance effects $\delta(\tau)$ and (d) Heritability (proportion of phenotypic variance explained by a locus $h^2(\tau)$).

**Figure 2.** The average likelihood ratio test statistic profile obtained from 100 replicated simulations under scenario 5 ($n = 80, m = 10$) with three QTL. The horizontal line indicates the empirical critical value when the Type I error rate was 1%.

**Figure 3.** The average likelihood ratio test statistic profile obtained from 100 replicated simulations under scenario 5 ($n = 80, m = 10$) with 1 QTL using the Legendre polynomial of order 4 proposed in this study (bold lines) as well as the logistic regression method (thin lines). The two horizontal line indicates the empirical critical values of the two methods when the Type I error rate was set at 1%.

**Figure 4.** The likelihood ratio test statistic profile of QTL detection for the diameter growth trajectory of poplar trees using Legendre polynomial of order 4 proposed in this study. The two horizontal lines represent the empirical critical values at Type I error rates of 0.01 (top) and 0.05 (bottom), respectively.

**Figure 5.** Additive effects of the two detected QTL expressed as functions of time.

**Figure 6.** The likelihood ratio test statistic profiles of QTL detection for the diameter growth trajectory of poplar trees using the Legendre polynomial of order 4 for QTL effects but fitted with AR(1) covariance structure for the residual errors (bold lines) and the logistic regression method (thin lines). The two horizontal lines represent
the empirical critical values at Type I error rates of 0.01, for the two models, respectively.
Figure 1
Figure 2
Figure 3
Figure 4
Figure 5
Figure 6
<table>
<thead>
<tr>
<th>QTL</th>
<th>Additive Effect</th>
<th>Dominance Effect</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha_0$</td>
<td>$\alpha_1$</td>
<td>$\alpha_2$</td>
</tr>
<tr>
<td>1</td>
<td>1.28</td>
<td>0.65</td>
<td>1.52</td>
</tr>
<tr>
<td>2</td>
<td>1.30</td>
<td>-0.39</td>
<td>1.31</td>
</tr>
<tr>
<td>3</td>
<td>0.94</td>
<td>-0.64</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 1. QTL parameters used in the simulation experiment.
Table 2. Six different experimental designs (scenarios) of experiment (combinations of sample size and frequency of trait measurement) and the empirical critical values of the test statistics when the Type I error is 0.05 or 0.01.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Sample size (n)</th>
<th>Frequency of test (m)</th>
<th>Critical value (0.05)</th>
<th>Critical value (0.01)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>5</td>
<td>93.5</td>
<td>104.1</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>5</td>
<td>104.0</td>
<td>112.0</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>5</td>
<td>111.0</td>
<td>122.8</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>10</td>
<td>96.8</td>
<td>105.9</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>10</td>
<td>110.1</td>
<td>120.31</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>10</td>
<td>114.3</td>
<td>126.3</td>
</tr>
</tbody>
</table>
Table 3. Average estimates of QTL positions (cM), the likelihood ratio (LR) test statistics and the empirical statistical powers of three QTL calculated from 100 replicated simulations. The standard deviations obtained from the 100 replicates are given in parentheses.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>QTL1 cM</th>
<th>LR</th>
<th>Power (5%)</th>
<th>Power (1%)</th>
<th>cM</th>
<th>LR</th>
<th>Power (5%)</th>
<th>Power (1%)</th>
<th>cM</th>
<th>LR</th>
<th>Power (5%)</th>
<th>Power (1%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24.06</td>
<td>211.13</td>
<td>99</td>
<td>95</td>
<td>95.00</td>
<td>169.97</td>
<td>92</td>
<td>82</td>
<td>179.79</td>
<td>121.65 (3.22)</td>
<td>80</td>
<td>72</td>
</tr>
<tr>
<td>2</td>
<td>22.61</td>
<td>318.09</td>
<td>100</td>
<td>100</td>
<td>93.86</td>
<td>242.57</td>
<td>99</td>
<td>92</td>
<td>183.41</td>
<td>176.19</td>
<td>98</td>
<td>93</td>
</tr>
<tr>
<td>3</td>
<td>22.68</td>
<td>413.37</td>
<td>100</td>
<td>100</td>
<td>94.61</td>
<td>309.74</td>
<td>100</td>
<td>100</td>
<td>184.17</td>
<td>212.84</td>
<td>99</td>
<td>94</td>
</tr>
<tr>
<td>4</td>
<td>24.04</td>
<td>251.24</td>
<td>100</td>
<td>100</td>
<td>93.20</td>
<td>205.46</td>
<td>99</td>
<td>99</td>
<td>182.60</td>
<td>140.28</td>
<td>87</td>
<td>79</td>
</tr>
<tr>
<td>5</td>
<td>23.40</td>
<td>397.70</td>
<td>100</td>
<td>100</td>
<td>94.25</td>
<td>299.65</td>
<td>99</td>
<td>99</td>
<td>184.62</td>
<td>208.00</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>6</td>
<td>22.82</td>
<td>512.58</td>
<td>100</td>
<td>100</td>
<td>94.51</td>
<td>353.63</td>
<td>100</td>
<td>100</td>
<td>186.80</td>
<td>250.41</td>
<td>100</td>
<td>99</td>
</tr>
</tbody>
</table>
Table 4. Average estimates of the additive ($\alpha$) and dominance ($\delta$) effects of three QTL calculated from 100 replicated simulations. The standard deviations obtained from the 100 replicates are given in parentheses. The three rows in each scenario represent the three simulated QTL.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$\alpha_0$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>$\delta_0$</th>
<th>$\delta_1$</th>
<th>$\delta_2$</th>
<th>$\delta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.57(0.06)</td>
<td>0.59(0.05)</td>
<td>1.40(0.07)</td>
<td>1.18(0.07)</td>
<td>1.05(0.04)</td>
<td>0.27(0.03)</td>
<td>-0.99(0.06)</td>
<td>0.84(0.06)</td>
</tr>
<tr>
<td></td>
<td>1.62(0.07)</td>
<td>-0.31(0.05)</td>
<td>1.16(0.08)</td>
<td>0.99(0.08)</td>
<td>1.05(0.04)</td>
<td>-0.19(0.04)</td>
<td>1.05(0.05)</td>
<td>0.71(0.05)</td>
</tr>
<tr>
<td></td>
<td>1.27(0.06)</td>
<td>-0.61(0.04)</td>
<td>0.97(0.08)</td>
<td>0.91(0.07)</td>
<td>0.72(0.05)</td>
<td>-0.44(0.03)</td>
<td>0.56(0.06)</td>
<td>0.95(0.06)</td>
</tr>
<tr>
<td>2</td>
<td>1.62(0.05)</td>
<td>0.55(0.04)</td>
<td>1.49(0.06)</td>
<td>1.10(0.05)</td>
<td>0.98(0.03)</td>
<td>0.35(0.03)</td>
<td>-1.05(0.03)</td>
<td>0.81(0.04)</td>
</tr>
<tr>
<td></td>
<td>1.74(0.05)</td>
<td>-0.28(0.03)</td>
<td>1.36(0.06)</td>
<td>0.93(0.07)</td>
<td>1.03(0.04)</td>
<td>-0.22(0.02)</td>
<td>1.06(0.04)</td>
<td>0.70(0.05)</td>
</tr>
<tr>
<td></td>
<td>1.23(0.04)</td>
<td>-0.66(0.03)</td>
<td>0.96(0.05)</td>
<td>0.91(0.05)</td>
<td>0.89(0.03)</td>
<td>-0.46(0.02)</td>
<td>0.73(0.04)</td>
<td>1.18(0.03)</td>
</tr>
<tr>
<td>3</td>
<td>1.58(0.04)</td>
<td>0.55(0.03)</td>
<td>1.43(0.05)</td>
<td>1.20(0.05)</td>
<td>1.07(0.03)</td>
<td>0.34(0.02)</td>
<td>-1.08(0.03)</td>
<td>0.87(0.03)</td>
</tr>
<tr>
<td></td>
<td>1.76(0.04)</td>
<td>-0.30(0.03)</td>
<td>1.20(0.05)</td>
<td>0.93(0.05)</td>
<td>1.10(0.03)</td>
<td>-0.25(0.02)</td>
<td>1.11(0.04)</td>
<td>0.70(0.03)</td>
</tr>
<tr>
<td></td>
<td>1.22(0.04)</td>
<td>-0.67(0.03)</td>
<td>0.95(0.05)</td>
<td>0.82(0.05)</td>
<td>0.77(0.03)</td>
<td>-0.45(0.02)</td>
<td>0.64(0.04)</td>
<td>1.07(0.03)</td>
</tr>
<tr>
<td></td>
<td>1.64(0.07)</td>
<td>0.50(0.04)</td>
<td>1.46(0.06)</td>
<td>1.03(0.07)</td>
<td>0.93(0.05)</td>
<td>0.32(0.03)</td>
<td>-1.11(0.05)</td>
<td>0.71(0.05)</td>
</tr>
<tr>
<td>---</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td>1.70(0.05)</td>
<td>-0.30(0.04)</td>
<td>1.28(0.06)</td>
<td>0.96(0.07)</td>
<td>1.03(0.04)</td>
<td>-0.25(0.03)</td>
<td>1.11(0.05)</td>
<td>0.74(0.04)</td>
</tr>
<tr>
<td></td>
<td>1.21(0.06)</td>
<td>-0.64(0.04)</td>
<td>0.89(0.07)</td>
<td>0.74(0.06)</td>
<td>0.84(0.05)</td>
<td>-0.45(0.03)</td>
<td>0.66(0.05)</td>
<td>1.07(0.05)</td>
</tr>
<tr>
<td>4</td>
<td>1.66(0.05)</td>
<td>0.50(0.03)</td>
<td>1.45(0.04)</td>
<td>1.15(0.05)</td>
<td>1.00(0.04)</td>
<td>0.36(0.02)</td>
<td>-1.17(0.04)</td>
<td>0.76(0.04)</td>
</tr>
<tr>
<td></td>
<td>1.75(0.05)</td>
<td>-0.32(0.03)</td>
<td>1.23(0.04)</td>
<td>0.99(0.06)</td>
<td>1.09(0.04)</td>
<td>-0.22(0.02)</td>
<td>1.15(0.04)</td>
<td>0.72(0.03)</td>
</tr>
<tr>
<td></td>
<td>1.21(0.05)</td>
<td>-0.71(0.03)</td>
<td>0.93(0.05)</td>
<td>0.94(0.05)</td>
<td>0.84(0.04)</td>
<td>-0.46(0.02)</td>
<td>0.71(0.04)</td>
<td>1.18(0.03)</td>
</tr>
<tr>
<td>5</td>
<td>1.60(0.04)</td>
<td>0.53(0.03)</td>
<td>1.40(0.04)</td>
<td>1.19(0.05)</td>
<td>1.06(0.03)</td>
<td>0.35(0.02)</td>
<td>-1.11(0.03)</td>
<td>0.80(0.03)</td>
</tr>
<tr>
<td></td>
<td>1.74(0.04)</td>
<td>-0.33(0.03)</td>
<td>1.22(0.04)</td>
<td>0.97(0.05)</td>
<td>1.07(0.03)</td>
<td>-0.25(0.02)</td>
<td>1.15(0.03)</td>
<td>0.73(0.03)</td>
</tr>
<tr>
<td></td>
<td>1.19(0.04)</td>
<td>-0.71(0.03)</td>
<td>0.88(0.04)</td>
<td>0.90(0.05)</td>
<td>0.77(0.03)</td>
<td>-0.48(0.02)</td>
<td>0.66(0.03)</td>
<td>1.09(0.03)</td>
</tr>
</tbody>
</table>
Table 5. Means and standard deviations (in parentheses) of the estimated QTL positions under the epistatic genetic effect models obtained from 100 replicated simulations. The empirical statistical powers at Type I error of 0.05 are given in the last column of the table. The true positions of QTL(1) and QTL(2) are 23cM and 75cM, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>Test frequency (m)</th>
<th>QTL(1)</th>
<th>QTL(2)</th>
<th>Power (0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive + Epistatic</td>
<td>5</td>
<td>23.23(0.15)</td>
<td>75.26(0.14)</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>23.26(0.13)</td>
<td>75.32(0.12)</td>
<td>100%</td>
</tr>
<tr>
<td>Epistatic effect only</td>
<td>5</td>
<td>23.2(0.26)</td>
<td>74.69(0.27)</td>
<td>60%</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>23.43(0.23)</td>
<td>75.01(0.27)</td>
<td>94%</td>
</tr>
</tbody>
</table>
Table 6. Means and standard deviations (in parentheses) of the estimated QTL effects obtained from 100 replicated simulations under the A+E model and the E model. QTL(1) and QTL(2) represent the additive effects of the two QTL and QTL(1,2) represents the epistatic effects between the two QTL. Estimate \((m = 5)\) and Estimate \((m=10)\) represent the estimated effects from data of the two different test frequencies.

<table>
<thead>
<tr>
<th>Model</th>
<th>QTL</th>
<th>True value</th>
<th>Estimate ((m=5))</th>
<th>Estimate ((m=10))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add. and Epi. (A+E)</td>
<td>QTL(1)</td>
<td>1.28 0.65 1.52 1.20</td>
<td>1.28 0.65 1.52 1.20</td>
<td>1.28 0.65 1.52 1.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.01) (0.01) (0.02) (0.02)</td>
<td>(0.01) (0.01) (0.02) (0.02)</td>
</tr>
<tr>
<td></td>
<td>QTL(2)</td>
<td>0.94 -0.64 0.73 0.72</td>
<td>0.95 -0.59 0.72 0.71</td>
<td>0.95 -0.65 0.72 0.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.01) (0.01) (0.02) (0.02)</td>
<td>(0.01) (0.01) (0.02) (0.02)</td>
</tr>
<tr>
<td></td>
<td>QTL(1,2)</td>
<td>0.65 -0.20 0.66 0.43</td>
<td>0.64 -0.18 0.6 0.37</td>
<td>0.66 -0.17 0.67 0.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.01) (0.01) (0.02) (0.02)</td>
<td>(0.01) (0.01) (0.02) (0.02)</td>
</tr>
<tr>
<td></td>
<td>QTL(1)</td>
<td>0.00 0.00 0.00 0.00</td>
<td>0.01 0.00 0.01 0.01</td>
<td>0.01 -0.01 0.02 0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.01) (0.01) (0.02) (0.02)</td>
<td>(0.01) (0.01) (0.02) (0.02)</td>
</tr>
<tr>
<td>Epistatic (E)</td>
<td>QTL(2)</td>
<td>QTL(1,2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>--------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>-0.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.01</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>-0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.03</td>
<td>0.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>-0.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.01</td>
<td>0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.42</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
</tbody>
</table>