Functional Mapping of Multiple Dynamic Traits

Jiguo Cao[1], Liangliang Wang[1], Zhongwen Huang[2,3],

Junyi Gai[3], Rongling Wu[4]

[1]Department of Statistics and Actuarial Science, Simon Fraser University,
Burnaby, BC, V6T1N8, Canada. Email: jca76@sfu.ca

[2]Department of Agronomy, Henan Institute of Science and Technology,
Xinxiang 453003, China.

[3]Soybean Research Institute of Nanjing Agricultural University,
National Key Laboratory for Crop Genetics and Germplasm Enhancement,
Nanjing 210095, China

[4]Center for Statistical Genetics, Pennsylvania State University
Hershey, PA 17033, USA. Email: rwu@hes.hmc.psu.edu

Abstract Many biological phenomena undergo developmental changes in time and space. Functional mapping, which is aimed at mapping genes that affect developmental patterns, is instrumental for studying the genetic architecture of biological changes. Often biological processes are mediated by a network of developmental and physiological components and, therefore, are better described by multiple phenotypes. In this article, we develop a multivariate model for functional mapping that can detect and characterize quantitative trait loci (QTLs) that simultaneously control multiple dynamic traits. Because the true genotypes of QTLs are unknown, the measurements for the multiple dynamic traits are modeled using a mixture distribution. The functional means of the multiple dynamic traits are estimated using the nonparametric regression method, which avoids any parametric assumption on the functional means. We propose the profile likelihood method to estimate the mixture model. A likelihood ratio test is exploited to test for the existence of pleiotropic effects on distinct but developmentally correlated traits. A simulation study is implemented to illustrate the finite sample performance of our proposed method. We also demonstrate our method by identifying QTLs that simultaneously control three dynamic traits of soybeans. The three dynamic traits are the time-course biomass of the leaf, the stem, and the root of the whole soybean. The genetic linkage map is constructed with 950 microsatellite markers. The new model can aid in our comprehension of the genetic control mechanisms of complex dynamic traits over time.

Keywords B-spline, functional mapping, multivariate model, quantitative trait loci, soybean

1 Introduction

The past two decades have witnessed the rapid development of genomic technologies that allow for molecular characterization of polymorphic markers throughout the entire genome. Nowadays molecular markers are readily available for a diversity of species. These markers are used to identify and localize quantitative trait loci (QTLs) that control phenotypic variation in a complex trait of interest. For a variety of traits hundreds of thousands of QTLs have been discovered, which play an important role in explaining the genetic control of biological characteristics. Among many successful examples of genetic mapping, it has been observed that QTLs are responsible for: branching, florescence, and grain architecture in maize (Doebley et al. 1997; Gallavotti et al. 2004; Wang et al. 2005); fruit size and shape in tomatoes (Paterson et al. 1988; Frary et al. 2000); the reduction of grain shattering (Li et al. 2006); complex behaviors in Drosophila (Anholt and Mackay 2004); whole blood serotonin level in humans (Weiss et al. 2005); as well as height growth and 16 other quantitative traits in the Hutterites, a founder human population (Weiss et al. 2006).

Genetic mapping of QTLs using molecular markers is founded on a statistical model pioneered by Weller (1986) and Lander and Botstein (1989). This model implements the EM algorithm to estimate the chromosomal positions and genetic effects of individual QTLs on a phenotypic trait. The publication of the first mapping model has led to an explosion of new statistical methods, which can precisely and accurately map QTLs under a variety of circumstances (Knapp 1991; Haley and Knott 1992; Jansen and Stam 1994; Zeng 1994; Sen and Churchill 2001; Kao and Zeng 2002). By considering the developmental complexity of several complex traits, such as growth, cell cycles and drug response, Ma et al. (2002) developed a new statistical method called functional mapping. This method maps that QTLs that influence the dynamic behavior of phenotypic values in time and space. Functional mapping capitalizes on mathematical aspects of biological and biochemical principles to model the temporal-spatial pattern of genetic effects triggered by QTLs. It has been proven to be a powerful method for studying and mapping the genetic architecture of dynamic trajectories across time and space, and for testing the genetic mechanisms underlying developmental alterations (Wu and Lin 2006; Li and Wu 2010; He et al. 2010). Statistically, functional mapping displays an increased power to detect QTLs, because fewer parameters are used to describe dynamic traits.

To broaden the range of applications for functional mapping, in which no parametric forms are available to specify the dynamic behavior of a trait, several nonparametric versions of functional mapping have been proposed, such as Legendre orthogonal polynomials (Lin and Carroll 2006; Yang and Xu 2007; Das et al. 2011) and B-splines (Yang et al. 2009). These nonparametric functions are flexible enough to represent dynamic or longitudinal traits in various shapes. The nonparametric functions are estimated directly from repeated measurements of dynamic traits, thus avoiding biases arising from inaccurate parametric assumptions.

With an increasing interest in systems mapping, which aims to elucidate a comprehensive picture of trait development, some studies have started to map phenotypic changes of multiple traits over time and space (Zhao et al. 2005; Li et al. 2006; Wu et al. 2011). Zhao et al. (2005) developed a growth equation approach for mapping two correlated growth traits. In a recent study, Wu et al. (2011) implemented a system of differential equations to model the temporal change of QTL effects on multiple traits that constitute a dynamic system. However, since these approaches require explicit mathematical equations to specify the dynamic traits, they are not as well suited to applications where no explicit equations exist.

The purpose of this article is to develop a flexible functional mapping method that can detect QTLs responsible for multiple dynamic traits. Each dynamic trait is represented by a nonparametric function, which is expressed as a linear combination of basis functions. We propose the profile likelihood method to estimate the nonparametric functions, as well as the correlations among multiple dynamic traits. A likelihood ratio test is implemented to identify QTLs at a grid of possible QTL locations. The significance threshold of the likelihood ratio test is derived using the permutation test. Since the exact genotype of a potential QTL at a grid point is unknown, dynamic traits are assumed to be in a mixture normal distribution. The Cholesky decomposition (Trefethen and Bau 1997) is used to parameterize the variance-covariance matrix of the multiple dynamic traits to ensure that the estimated variance-covariance matrix is both symmetric and positive definite. Our method can also accommodate irregular sparse data such as data collected at different time for different subjects or missing data at any time points.

The remainder of this article is organized as follows. Section 2 introduces our statistical model for functional mapping. The parameter estimation method for our statistical model is introduced in Section 3. In Section 4, our functional mapping method is applied to detect QTLs that control multiple dynamic traits of soybeans. Section 5 presents simulation studies implemented to evaluate the finite sample performance of our functional mapping method. The discussion of this model is given in Section 6.

2 A Mixture Model

Suppose multiple dynamic traits are measured at a series of time points. Let $Y_{hi}(t_{ir})$ be the measured h-th dynamic trait at the r-th time point t_{ir} for the i-th subject, $h = 1, \cdots, H$, $i = 1, \cdots, n_h$, $r = 1, \cdots, m_i$. Let $\mathbf{Y}_i(t) = (Y_{1i}(t), \cdots, Y_{Hi}(t))^T$ denote the vector of measurements for multiple dynamic traits at the time point t. Given that the i-th subject has QTL genotype j, $j = 1, \cdots, J$, $\mathbf{Y}_i(t)$ is assumed to have a multivariate normal distribution with mean $\boldsymbol{\mu}_j(t) = (\mu_{1j}(t), \cdots, \mu_{Hj}(t))^T$ and a variance-covariance matrix $\boldsymbol{\Sigma}$.

In practice, the true genotypes of QTLs are unknown. But we can calculate the conditional probabilities of QTL genotypes given marker genotypes as a function of the recombination fractions between the QTL and markers (Wu et al. 2007). Let ω_{ij} be the line origin probability of the i-th subject having the QTL genotype j, $j = 1, \cdots, J$. The line origin probability ω_{ij} can be calculated in advance based on experimental population designs such as inbreed, outbreed and backcross. The vector of measurements for multiple dynamic traits, $\mathbf{Y}_i(t)$, is modeled using a mixture distribution:

$$\mathbf{Y}_i(t) \sim \sum_{j=1}^{J} \omega_{ij} f(\mathbf{Y}_i(t)|\boldsymbol{\mu}_j(t), \boldsymbol{\Sigma}) , \tag{1}$$

where $f(\mathbf{Y}_i(t)|\boldsymbol{\mu}_j(t), \boldsymbol{\Sigma})$ is the probability density function (pdf) of the multivariate normal distribution with mean $\boldsymbol{\mu}_j(t)$ and a variance-covariance matrix $\boldsymbol{\Sigma}$, which is expressed as follows:

$$f(\mathbf{Y}_i(t)|\boldsymbol{\mu}_j(t), \boldsymbol{\Sigma}) \;=\; (2\pi)^{-H/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\{-(\mathbf{Y}_i(t) - \boldsymbol{\mu}_j(t))^T \boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i(t) - \boldsymbol{\mu}_j(t))/2\}. \tag{2}$$

In order to avoid any parametric constraints on the functional mean of the h-th dynamic trait $\mu_{hj}(t)$, given the QTL genotype j, $\mu_{hj}(t)$ is estimated using the nonparametric smoothing method (Ramsay and Silverman 2005). In our article, the functional mean $\mu_{hj}(t)$ is represented as a linear combination of basis functions,

$$\mu_{hj}(t) = \sum_{k=1}^{K} c_{hjk}\phi_{hjk}(t) = \mathbf{c}_{hj}^T \boldsymbol{\phi}_{hj}(t) ,$$

where $\boldsymbol{\phi}_{hj}(t) = (\phi_{hj1}(t), \cdots, \phi_{hjK}(t))^T$ is a vector of basis functions, and $\mathbf{c}_{hj} = (c_{hj1}, \cdots, c_{hjK})^T$ is a vector of basis coefficients. Cubic B-splines are often chosen as basis functions since any B-spline basis function is only positive over a short interval and zero elsewhere. This is called the *compact support* property, and is essential for efficient computation (de Boor 2001).

The variance-covariance matrix, $\boldsymbol{\Sigma}$, must be symmetric and positive-definite. Therefore it may be estimated using a constrained optimization method. Alternatively, $\boldsymbol{\Sigma}$ can be decomposed as

$$\boldsymbol{\Sigma}^{-1} = \mathbf{L}\mathbf{L}^T ,$$

where $\mathbf{L}$ is a lower triangular matrix with strictly positive diagonal entries . This is called the Cholesky decomposition (Trefethen and Bau 1997). By employing the Cholesky decomposition, we can estimate the lower triangular matrix $\mathbf{L}$ directly, without considering the usual constraints on $\mathbf{\Sigma}$, since $\widehat{\mathbf{\Sigma}} = (\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^T)^{-1}$ will automatically be symmetric and positive-definite (Cao and Ramsay 2012). Essentially the Cholesky decomposition allows us to convert a constrained optimization problem to an unconstrained optimization. The Cholesky decomposition is not uniquely defined for a given positive-definite matrix, but it can be made unique by requiring the diagonal elements in $\mathbf{L}$ to be all positive. Consequently the diagonal elements in $\mathbf{L}$ are parameterized in terms of their logarithms in our article.

3 QTL Mapping Method for Multiple Dynamic Traits

The mixture model (1) has two types of parameters to estimate, the basis coefficient, $\mathbf{c}_{hj}$, and the lower triangular matrix, $\mathbf{L}$. Define the long vector of basis coefficients $\mathbf{c}_j = (\mathbf{c}_{1j}^T, \cdots, \mathbf{c}_{Hj}^T)^T$, and $\mathbf{c} = (\mathbf{c}_1^T, \ldots, \mathbf{c}_J^T)^T$. We propose to estimate the basis coefficient, $\mathbf{c}$, and the lower triangular matrix, $\mathbf{L}$, using the profile likelihood method. The method estimates the two parameters in two nested levels of optimization. In the inner level of optimization, the basis coefficient, $\mathbf{c}$, is estimated by maximizing the log likelihood function for a given lower triangular matrix, $\mathbf{L}$. There is no analytic expression for the estimated basis coefficient, $\widehat{\mathbf{c}}$, but it can be viewed as an implicit function of $\mathbf{L}$. In the outer level of optimization, the lower triangular matrix, $\mathbf{L}$, is estimated by maximizing the profile likelihood function, in which the basis coefficient is removed from the parameter space by treating it as a function of $\mathbf{L}$. Although the estimated basis coefficient, $\widehat{\mathbf{c}}$, has no analytic formula, we use the implicit function theorem to obtain analytic gradients for the optimization iteration process, which makes computation faster and more stable. We outline the details of the profile likelihood method below.

3.1 Profile Likelihood Method for Estimating the Mixture Model

To simply notation, we first define some vectors and matrices to be used in the likelihood function. These matrix representations also help to significantly increase computational efficiency in MATLAB (MATLAB 2015). Define the long vector of data $\mathbf{Y}_i = (\mathbf{Y}_i(t_{i1})^T, \cdots, \mathbf{Y}_i(t_{im_i})^T)^T$, and the long vector of basis coefficients $\mathbf{c}_j = (\mathbf{c}_{1j}^T, \cdots, \mathbf{c}_{Hj}^T)^T$. Then the distribution of $\mathbf{Y}_i$, given the QTL genotype j, is the multivariate normal distribution with mean $\mathbf{\Psi}_{ij}\mathbf{c}_j$ and a variance-covariance matrix $\mathbf{\Gamma}$, where $\mathbf{\Gamma} = \mathbf{I}_{m_i} \otimes \mathbf{\Sigma}$, $\mathbf{\Psi}_{ij}$ is a $Hm_i \times HK$ matrix, and $\mathbf{A}_{ijr}$, $r = 1, \dots, m_i$, is a block diagonal matrix with the h-th diagonal block as $\boldsymbol{\phi}_{hj}^T(t_{ir})$ defined below:

$$\mathbf{\Psi}_{ij} = \begin{pmatrix} \mathbf{A}_{ij1} \\ \mathbf{A}_{ij2} \\ \vdots \\ \mathbf{A}_{ijm_i} \end{pmatrix}, \mathbf{A}_{ijr} = \begin{pmatrix} \boldsymbol{\phi}_{1j}^T(t_{ir}) & \mathbf{0}_{1\times K} & \dots & \mathbf{0}_{1\times K} \\ \mathbf{0}_{1\times K} & \boldsymbol{\phi}_{2j}^T(t_{ir}) & \dots & \mathbf{0}_{1\times K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{1\times K} & \mathbf{0}_{1\times K} & \dots & \boldsymbol{\phi}_{Hj}^T(t_{ir}) \end{pmatrix}$$

The vector of basis coefficients $\mathbf{c}$ is estimated by maximizing the log likelihood function for a given lower triangular matrix, $\mathbf{L}$:

$$J(\mathbf{c}|\mathbf{L}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{J} \omega_{ij} f(\mathbf{Y}_i|\mathbf{c}_j, \mathbf{L}) \right], \tag{3}$$

where

$$\begin{aligned} f(\mathbf{Y}_i|\mathbf{c}_j, \mathbf{L}) &= (2\pi)^{-H*m_i/2}|\mathbf{\Gamma}|^{-1/2} \exp\{-(\mathbf{Y}_i - \mathbf{\Psi}_i\mathbf{c}_j)^T\mathbf{\Gamma}^{-1}(\mathbf{Y}_i - \mathbf{\Psi}_i\mathbf{c}_j)/2\}, \qquad (4) \\ \mathbf{\Gamma}^{-1} &= \mathbf{I}_{m_i} \otimes \mathbf{L}\mathbf{L}^T, \\ |\mathbf{\Gamma}| &= |\mathbf{\Sigma}|^{m_i} = |(\mathbf{L}\mathbf{L})^{-1}|^{m_i} = |(\mathbf{L}\mathbf{L})|^{-m_i} = |\mathbf{L}|^{-2m_i}. \end{aligned}$$

Since the log likelihood function (3) is structured as a mixture distribution, it is impossible to obtain an analytic formula for the estimate $\widehat{\mathbf{c}}$. We use the Newton-Raphson method to maximize $J(\mathbf{c}|\mathbf{L})$ as follows. Let $\mathbf{c}^{(0)}$ be the starting value for $\mathbf{c}$, the v-th iteration step updates $\mathbf{c}$ by

$$\mathbf{c}^{(v)} = \mathbf{c}^{(v-1)} - \left(\frac{d^2 J}{d\mathbf{c}^2}\bigg|_{\mathbf{c}^{(v-1)}} \right)^{-1} \left(\frac{dJ}{d\mathbf{c}}\bigg|_{\mathbf{c}^{(v-1)}} \right),$$

$v = 1, 2, \ldots,$ until convergence occurs. To ensure the Newton-Raphson method is both stable and fast, the first derivative of $J(\mathbf{c}|\mathbf{L})$ with respect to $\mathbf{c}$ is derived analytically as follows.

$$\frac{dJ}{d\mathbf{c}_j} = \sum_{i=1}^{n} P_{ij} \boldsymbol{\Psi}_i^T \boldsymbol{\Gamma}^{-1} (\mathbf{Y}_i - \boldsymbol{\Psi}_i \mathbf{c}_j), \tag{5}$$

where

$$P_{ij} = \frac{\omega_{ij} f(\mathbf{Y}_i|\mathbf{c}_j, \mathbf{L})}{\sum_{j=1}^{J} \omega_{ij} f(\mathbf{Y}_i|\mathbf{c}_j, \mathbf{L})},$$

with $\sum_{j=1}^{J} P_{ij} = 1$. The second derivative of $J(\mathbf{c}|\mathbf{L})$ with respect to $\mathbf{c}$ is hard to obtain analytically, hence we apply the finite-difference method to approximate the second derivative by using the analytic first derivative given in (5).

The estimate for the basis coefficient, $\widehat{\mathbf{c}}$, is obtained for any given value of the lower triangular matrix, $\mathbf{L}$, so $\widehat{\mathbf{c}}$ may be viewed as an implicit function of $\mathbf{L}$, which is denoted as $\widehat{\mathbf{c}}(\mathbf{L})$. The lower triangular matrix $\mathbf{L}$ is then estimated by maximizing the log profile likelihood function

$$F(\mathbf{L}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{J} \omega_{ij} f(\mathbf{Y}_i|\widehat{\mathbf{c}}_j(\mathbf{L}), \mathbf{L}) \right]. \tag{6}$$

by using the Newton-Raphson method. In the optimization process, the entries below or in the main diagonal of the $H \times H$ lower triangular matrix, $\mathbf{L}$, are combined in a vector, ℓ, with length $H(H+1)/2$. To ensure the Newton-Raphson method is both stable and fast, the first derivative of $F(\mathbf{L})$ is derived analytically using the chain rule after considering $\widehat{\mathbf{c}}$ as a function of ℓ:

$$\frac{dF(\mathbf{L})}{d\ell} = \frac{\partial F(\mathbf{L})}{\partial \ell} + \frac{\partial F(\mathbf{L})}{\partial \widehat{\mathbf{c}}} \frac{d\widehat{\mathbf{c}}}{d\ell}\,.$$

Since $\widehat{\mathbf{c}}$ is an implicit function of ℓ, the derivative $d\widehat{\mathbf{c}}/d\ell$ can be derived analytically by applying the implicit function theorem as follows. We take advantage of the fact that the estimate $\widehat{\mathbf{c}}$ satisfies

$$\left.\frac{\partial J}{\partial \mathbf{c}^T}\right|_{\widehat{\mathbf{c}}} \equiv 0$$

Taking the ℓ-derivative on both sides of the above identity, we obtain

$$\frac{d}{d\ell}\left.\frac{\partial J}{\partial \mathbf{c}^T}\right|_{\widehat{\mathbf{c}}} = \left\{ \left.\frac{\partial^2 J}{\partial \mathbf{c}^T \partial \ell}\right|_{\widehat{\mathbf{c}}} \right\} + \left\{ \left.\frac{\partial^2 J}{\partial \mathbf{c}^T \partial \mathbf{c}}\right|_{\widehat{\mathbf{c}}} \right\} \left\{ \frac{d\widehat{\mathbf{c}}^T}{d\ell} \right\} \equiv 0,$$

which yields

$$\frac{d\widehat{\mathbf{c}}^T}{d\ell} = -\left\{ \left.\frac{\partial^2 J}{\partial \mathbf{c}^T \partial \mathbf{c}}\right|_{\widehat{\mathbf{c}}} \right\}^{-1} \left\{ \left.\frac{\partial^2 J}{\partial \mathbf{c}^T \partial \ell}\right|_{\widehat{\mathbf{c}}} \right\},$$

provided that $\partial^2 J/\partial \mathbf{c}^T \partial \mathbf{c}$ is non-singular at $\mathbf{c} = \widehat{\mathbf{c}}$.

The algorithm of our proposed profile likelihood method can be summarized as follows:

The algorithm of the profile likelihood method

1. Choose an initial value, $\mathbf{L}^{(0)}$, for the lower triangular matrix $\mathbf{L}$.

2. For a given $\mathbf{L}^{(\tau)}$,

2.1 Estimate $\mathbf{c}$ by maximizing

$$J(\mathbf{c}|\mathbf{L}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{J} \omega_{ij} f(\mathbf{Y}_i | \mathbf{c}_j, \mathbf{L}) \right]$$

using the Newton-Raphson method.

2.2 After obtaining the estimate $\widehat{\mathbf{c}}$, calculate $dF(\mathbf{L})/d\ell$ and $d^2F(\mathbf{L})/d\ell^2$.

2.3 Update ℓ by the Newton-Raphson method:

$$\ell^{(\tau)} = \ell^{(\tau-1)} - \left(\left. \frac{d^2 F}{d\ell^2} \right|_{\ell^{(\tau-1)}} \right)^{-1} \left(\left. \frac{dF}{d\ell} \right|_{\ell^{(\tau-1)}} \right),$$

3. $\tau = \tau + 1$. Go to Step 2 until the Newton-Raphson iteration procedure for maximizing $F(\mathbf{L})$ converges.

3.2 Likelihood Ratio Test

For a given linkage map, we will search at any possible position in the genome for QTLs that simultaneously control multiple dynamic traits. The significance test for the existence of a QTL can be performed by formulating hypotheses as follows:

$$\begin{aligned} H_0 &: \mathbf{c}_1 = \mathbf{c}_2 = \cdots = \mathbf{c}_J, \\ H_1 &: \text{at least two of } \mathbf{c}_j, j = 1, \cdots, J, \text{ are not equal to each other.} \end{aligned}$$

Under the alternative hypothesis, H_1, means of dynamic traits are different for at least two of J QTL genotypes, i.e., at least two of $\boldsymbol{\mu}_j$, $j = 1, \ldots, J$, are not equivalent. At any possible position of a QTL, we calculate the conditional probability, ω_{ij}, of QTL genotypes given marker genotypes as a function of the recombination fractions between the QTL and markers (Wu et al. 2007). The mixture model (1) is then estimated with our proposed profile likelihood method.

Under the null hypothesis, H_0, means of dynamic traits are the same for different QTL genotypes, i.e. $\boldsymbol{\mu}_1(t) = \boldsymbol{\mu}_2(t) = \cdots = \boldsymbol{\mu}_J(t)$. Therefore, the vector of measurements for multiple dynamic traits, $\mathbf{Y}_i(t)$, is modelled by a multivariate normal distribution with mean, $\boldsymbol{\mu}^{H_0}(t)$, and a variance-covariance matrix, $\boldsymbol{\Sigma}^{H_0}$:

$$\mathbf{Y}_i(t) \sim f(\mathbf{Y}_i(t) | \boldsymbol{\mu}^{H_0}(t), \boldsymbol{\Sigma}^{H_0}), \tag{7}$$

where $f(\cdot)$ is the probability density function (pdf) of multivariate normal distribution as expressed in (2). The mean, $\boldsymbol{\mu}^{H_0}(t)$, can be represented as a linear combination of basis functions, and the variance-covariance matrix, $\boldsymbol{\Sigma}^{H_0}$, can be reparameterized using the Cholesky decomposition, as described in Section 2. Then $\boldsymbol{\mu}^{H_0}(t)$ and $\boldsymbol{\Sigma}^{H_0}$ are estimated using the profile likelihood method as introduced in Section 3.

The likelihoods under the null and alternative hypotheses are calculated, from which the log-likelihood ratio (LR) is computed. Let $\widehat{\mathbf{c}}^{H0}$, $\widehat{\mathbf{L}}^{H0}$, and $\widehat{\mathbf{c}}_j^{H1}$, $\widehat{\mathbf{L}}^{H1}$ be the parameter estimates obtained under H_0 and H_1, respectively. The LR test statistic is given by

$$\mathrm{LR} = -2\left[\sum_{i=1}^{n} \log f(\mathbf{Y}_i|\widehat{\mathbf{c}}^{H0}, \widehat{\mathbf{L}}^{H0}) - \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{J} \omega_{ij} f(\mathbf{Y}_i|\widehat{\mathbf{c}}_j^{H1}, \widehat{\mathbf{L}}^{H1})\right\}\right], \tag{8}$$

where $f(\cdot)$ is the probability density function (pdf) of $H \times m_i$-dimensional multivariate normal distribution as expressed in (4). Given a significance threshold T, there is significant evidence that a QTL exists at a certain position if LR $> T$. Since the distribution of the LR values under the null hypothesis is unknown, empirical permutation tests are usually used to determine the threshold (Churchill and Doerge 1994). In our article, we keep all multiple dynamic trait data for each individual in its entirety, and permute the phenotypic data among all individuals. Note that we do not permute the phenotypic data measured at different time points for the same individual.

4 Application

We use the proposed functional mapping method to identify QTLs that simultaneously control three dynamic traits of soybeans. These dynamic traits are the time-course biomass of the whole-plant leaf, the whole-plant stem, and the whole-plant root of the soybean. A mapping population composed of 184 recombinant inbred lines (RILs) is derived from the cross of two cultivars, Kefeng No. 1 and Nannong 1138-2. In this RIL population, there are two homozygous genotypes, one containing two Kefeng No. 1 alleles and the other containing

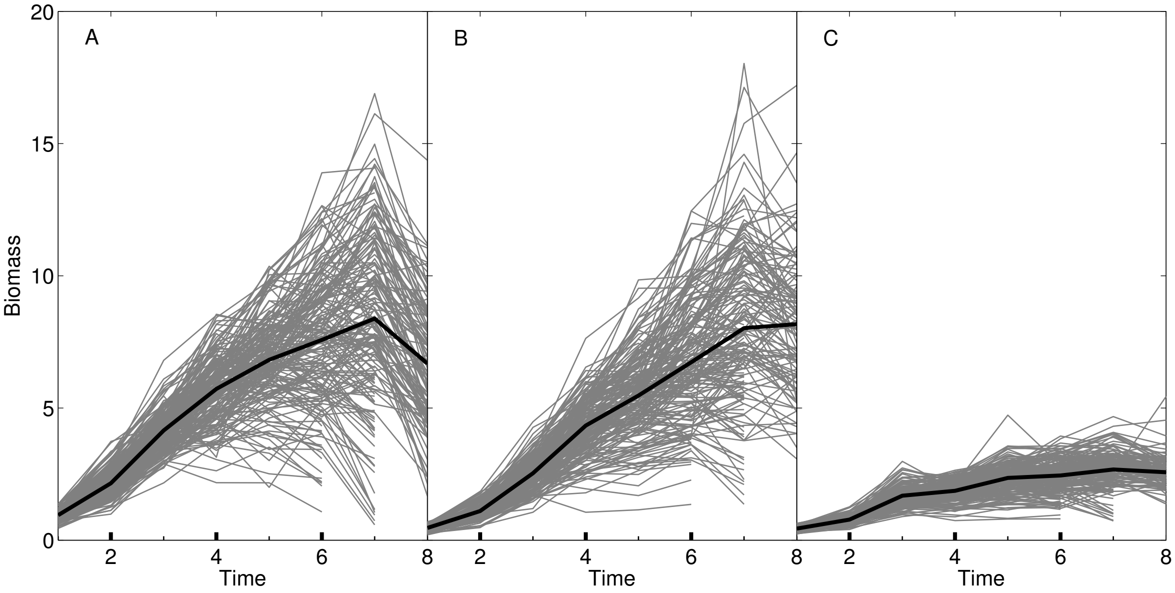

Figure 1: Dynamic behavior in the whole-plant leaf biomass (A), the whole-plant stem biomass (B), and whole-plant root biomass (C) in a growing season. Each grey line represents the trajectory curve of one of 184 recombinant inbred lines (RILs) and black lines are mean trajectory curves.

two Nannong 1138-2 alleles. A genetic linkage map is constructed with 950 microsatellite markers. The whole-plant leaf biomass, the whole-plant stem biomass, and the whole-plant root biomass were measured for each RIL of the mapping population weekly for eight weeks in a growing season of soybeans.

Figure 1 illustrates the age-dependent trajectories of the three biomass traits; each of which displays considerable variation. As seen in mature organs with strong cell turnover rates, leaf and root biomass may also undergo reduction as a plant ages. This phenomenon is evident in the age-dependent trajectories of the leaf and root biomasses (see Fig. 1). Nonparametric regression modeling is employed to fit the dynamic biomass traits in this soybean example, as explained in Section 2.

We scan for possible QTLs by assuming their positions at every 2 cM within a given marker interval in each of 25 linkage groups. For simplicity, we assume a single QTL at a time. For any given position of the QTL, the conditional probability, ω_{ij}, of QTL genotypes

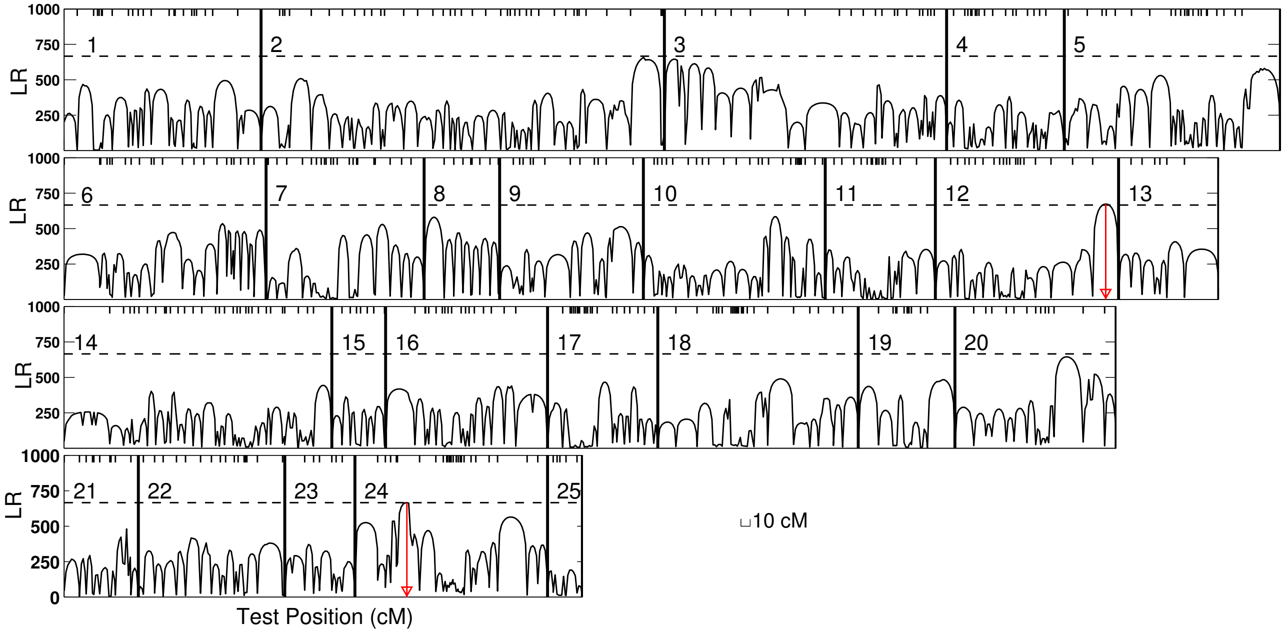

Figure 2: The log-likelihood ratios (LR) used to test for the existence of QTLs at every 2 cM along the genetic linkage map composed of 950 molecular markers. The number in each panel indicates the 25 linkage groups. The horizontal dashed line indicates the significant threshold for confirming the genome-wide existence of a QTL. Tick marks on the ceilings of each panel represent the positions of molecular markers in each linkage group. The arrowed red lines indicate the locations of the QTLs detected by our method.

is calculated based on the marker genotypes as a function of the recombination fractions between the QTL and markers (Wu et al. 2007). We then use our proposed profile likelihood method to estimate the mixture model (1). Finally, the LR test statistic (8) is calculated at a grid of possible QTL positions. The total computation time for scanning the entire genetic linkage map is around 31 hours by using a MacBook Pro laptop with a 2.5 GHz Intel Core i7 processor and a 16 GB 1600 MHz DDR3 Memory.

Figure 2 displays the log-likelihood ratio (LR) profile at every 2 cM along the whole genetic linkage map. To determine the significance threshold for confirming the genome-wide existence of a QTL, a permutation test with 100 permutation replicates was conducted. The 95th percentile of the distribution of the maximum LR values obtained from the permutation test is 665.61, which is used as the empirical critical value to declare genome-wide existence of a QTL at the 5% significance level. The QTL location is estimated by the genomic positions of the peaks of the LR profile that extends beyond the threshold.

Two significant QTLs are found: one (called `QTL1`) located between markers GMKF167a and GMKF167b on linkage group 12 and the other (called `QTL2`) located between markers sat_274 and BE801128 on linkage group 24. The estimated curve parameters for each QTL genotype ($\mathbf{c}_{hj}$) allow us to draw mean biomass trajectories for leaves, the stem and roots, which are displayed in Figure 3. At `QTL1`, the biomasses of the two homozygous genotypes increase in a similar way, but then diverge after the third time of measurement. This suggests that the QTL remains inactive until a particular time point in the growing season. The genotype composed of the parent Nannong 1138-2 alleles displays a much faster rate of increase, especially for leaf and stem biomass, than that of the parent Kefeng No. 1 alleles. The leaf and root biomass of both genotypes decay at a later stage in life. The leaf biomass starts to decay at 6.1 and 6.4 weeks for the genotypes composed of the parent Kefeng No. 1 alleles and Nannong 1138-2 alleles, respectively. A similar pattern is present in the root biomass, which starts to decay at 7.3 and 6.8 weeks, respectively, for the genotypes composed of the parent Kefeng No. 1 alleles and Nannong 1138-2 alleles. Since the leaf and root decay times are slightly different for the two QTL genotypes, this suggests that `QTL1` may impact the starting time of both leaf biomass and root biomass decay. A similar result is observed for `QTL2`, except here we see a reversal in the direction of the genetic effects for the two original parents (Fig. 3D,E,F). It is interesting to see that the two QTLs detected by our new model have also been observed by our previous model, which integrates allometric scaling through a system of differential equations (Wu et al. 2011). The distinction between the two models is that our proposed functional mapping method considers the correlation among multiple dynamic traits.

Table 1 displays the standard deviation and correlation coefficient estimates for the whole-plant leaf biomass, whole-plant stem biomass, and the whole-plant root biomass at `QTL1` and

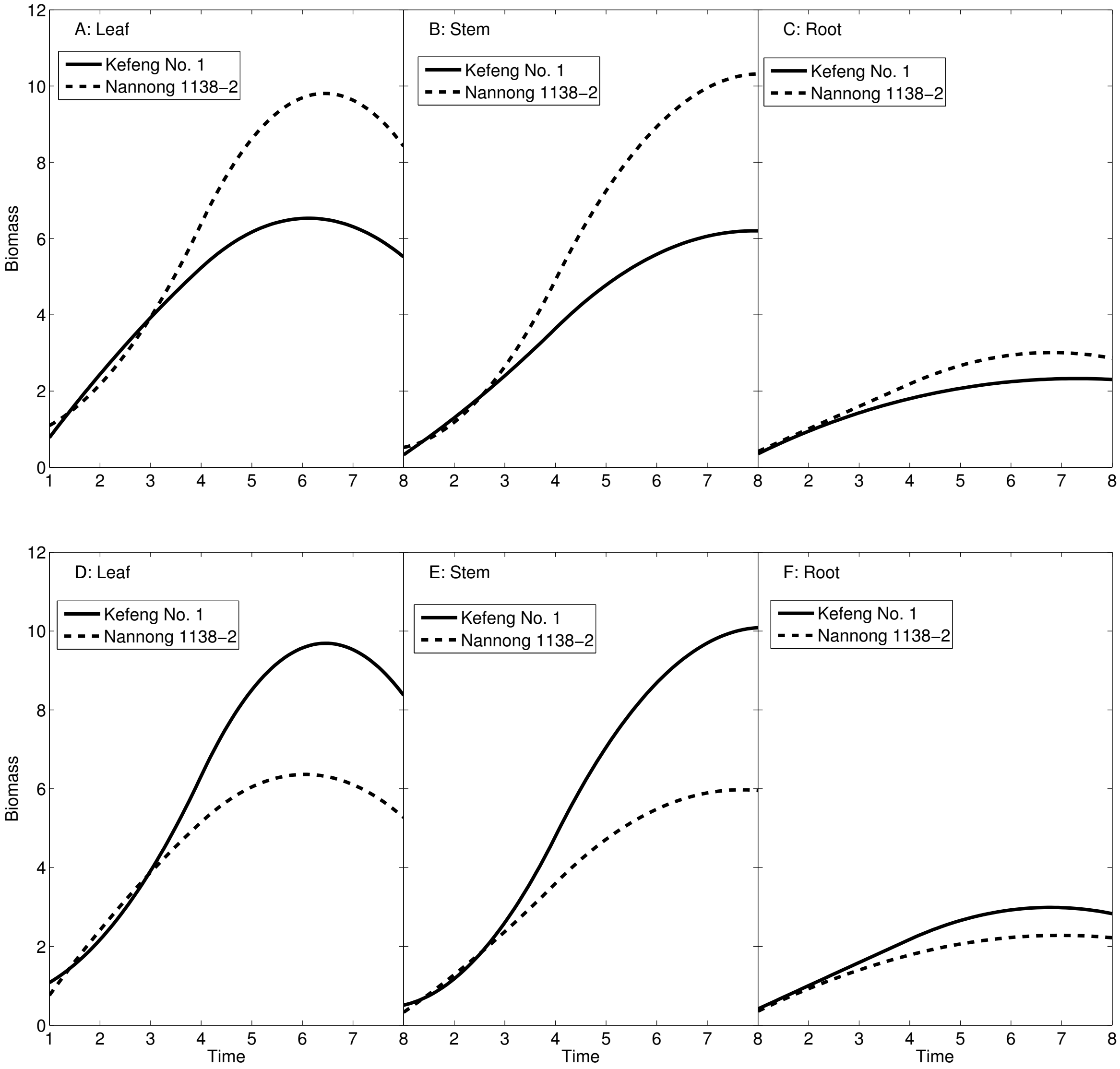

Figure 3: Estimated mean trajectory curves of whole-plant leaf biomass, stem biomass and whole-plant root biomass for two different genotypes of `QTL1` detected in linkage group 12 (A, B, C, respectively) and `QTL2` detected in linkage group 24 (D, E, F, respectively).

Table 1: The estimates and standard errors (SEs) for the standard deviations and correlation coefficients of the biomass of leaves (L), stems (S), and roots (R) for QTL1 and QTL2, which are detected on linkage group 12 and 24, respectively. Here, σ denotes the standard deviations, and ρ denotes the correlation coefficients. H1 is the alternative hypothesis, which assumes two different mean growth curves for two different genotypes of QTL. H0 is the null hypothesis, which assumes the same mean growth curves for two different genotypes of QTL.

			σ_{LL}	σ_{SS}	σ_{RR}	ρ_{LS}	ρ_{LR}	ρ_{SR}
QTL1	H1	Estimate	1.49	1.19	0.41	0.79	0.66	0.65
		SE	0.03	0.02	0.01	0.01	0.01	0.02
	H0	Estimate	1.80	1.67	0.47	0.85	0.74	0.73
		SE	0.03	0.03	0.01	0.01	0.01	0.01
QTL2	H1	Estimate	1.45	1.20	0.40	0.77	0.67	0.66
		SE	0.03	0.02	0.01	0.01	0.01	0.01
	H0	Estimate	1.79	1.66	0.47	0.86	0.76	0.75
		SE	0.03	0.03	0.01	0.01	0.01	0.01

QTL2. As expected these three traits are positively correlated due to allometric scaling. Strong trait-trait correlations imply the necessity of jointly modeling the three traits. The standard deviations of each trait under the full model (H1, there is a QTL) are 17.2%, 28.7%, and 12.8% smaller than under the reduced model (H0, there is no QTL). Also, comparing the reduced model to the full model, trait-trait correlations tend to decrease under the assumptions of the reduced model. The standard errors of the estimates are obtained by the parametric bootstrap method as follows. For the parameter estimates under H1 and H0, the phenotypic data are generated using the mixture distribution (1) and the multivariate normal distribution, respectively, where the means and the variance-covariance matrices are set to be the same as the estimates from the real data. The standard deviations, $\sigma_{LL}, \sigma_{SS}, \sigma_{RR}$, and correlation coefficients, ρ_{LS}, ρ_{LR}, ρ_{SR}, are then estimated with the profile likelihood method. The above process is replicated 100 times. The sample standard deviation of the 100 replicative estimates for the six parameters are used as the standard errors of the

parameter estimates. It shows the standard errors of the estimates are very small, which may indicate that the correlation coefficient estimates are statistically different from zero.

5 Simulations

We implement a simulation study to investigate the statistical properties of our functional mapping method. The data are simulated using the same marker information as the twelfth linkage group of the soybean, which is analyzed as a real application in Section 4. This linkage group has 21 markers in total with a length of 196 cM. A QTL, named `QTL1`, is located at 182.6 cM from the first marker in this group. We assume that three dynamic traits are measured at eight equally-spaced time points, which is also consistent with the real data. The phenotypic data are generated for 184 RILs based on the mixture distribution (1), where means, $\boldsymbol{\mu}_1(t)$ and $\boldsymbol{\mu}_2(t)$, and a variance-covariance matrix, $\boldsymbol{\Sigma}$, are set to be the same as the estimates from the real data.

For each simulation data set, we scan for possible QTLs by assuming their positions at every 2 cM within a given marker interval in the twelfth linkage group. At each possible location, the conditional probability, ω_{ij}, of QTL genotypes is calculated based on the marker genotypes as a function of the recombination fractions between the QTL and markers (Wu et al. 2007). Next, we use our proposed profile likelihood method to estimate the mixture model (1), and finally the LR test statistic (8) is calculated at any given position of a QTL. We use the permutation test with 100 permutation replicates to obtain the significance threshold for confirming the existence of a QTL. The 95th percentile of the distribution of the maximum LR values is obtained from the permutation test, which is then set to be the empirical critical value for declaring the existence of a QTL at the 5% significance level. The above simulation procedure is repeated 100 times.

We compare our functional mapping method, which accounts for correlations, with the traditional method, which does not consider the correlations between multiple dynamic

Table 2: Means, biases, standard deviations (STDs), and root mean squared errors (RMSEs) of the estimated location of a QTL in 100 simulation replicates using the functional mapping method with or without considering the correlation of multiple dynamic traits.

	True	Mean	Bias	STD	RMSE	Confidence Interval
Correlated Model	182.6	182.4	0.2	1.9	1.9	[178.7,186.2]
Uncorrelated Model	182.6	180.6	2.0	18.3	18.4	[144.7, 216.6]

traits. Table 2 summarizes the estimated QTL locations in 100 simulation replicates for each of these two methods. When considering the correlation between multiple dynamic traits, we see that the biases, standard deviations (STDs) and root mean squared errors (RMSEs) of the estimated QTL locations are reasonably small, which indicates that our functional mapping method provides an accurate estimate of the QTL location for this sample size. Alternatively, if we do not consider the correlation between multiple dynamic traits, the estimate of the QTL position is very biased. The standard deviation and root mean squared error of the estimated QTL locations also substantially increase by not considering the correlation between multiple dynamic traits. Notably, the RMSE of the estimated QTL locations is decreased by 89.7% by considering the correlation between multiple dynamic traits.

Figure 4 displays the point-wise biases and standard deviations (STDs) of the estimated mean trajectory curves of the whole-plant leaf biomass, the whole-plant stem biomass, and the whole-plant root biomass for the two different genotypes of the QTL located at 182 cM from the first marker in the twelfth linkage group. It can be seen that the biases of the estimated mean trajectory curves are negligible for the whole-plant leaf, stem, and root.

A power study is implemented to evaluate the power of the proposed likelihood ratio test to determine the existence of a QTL. Assuming a QTL is located at 182.6 cM from the first marker in the twelfth linkage group of the soybean, the phenotypic data are generated for 184 RILs from the mixture distribution with the true parameter values same to the estimates from the real data. When a QTL exists, i.e., the alternative hypothesis is true, we detect a

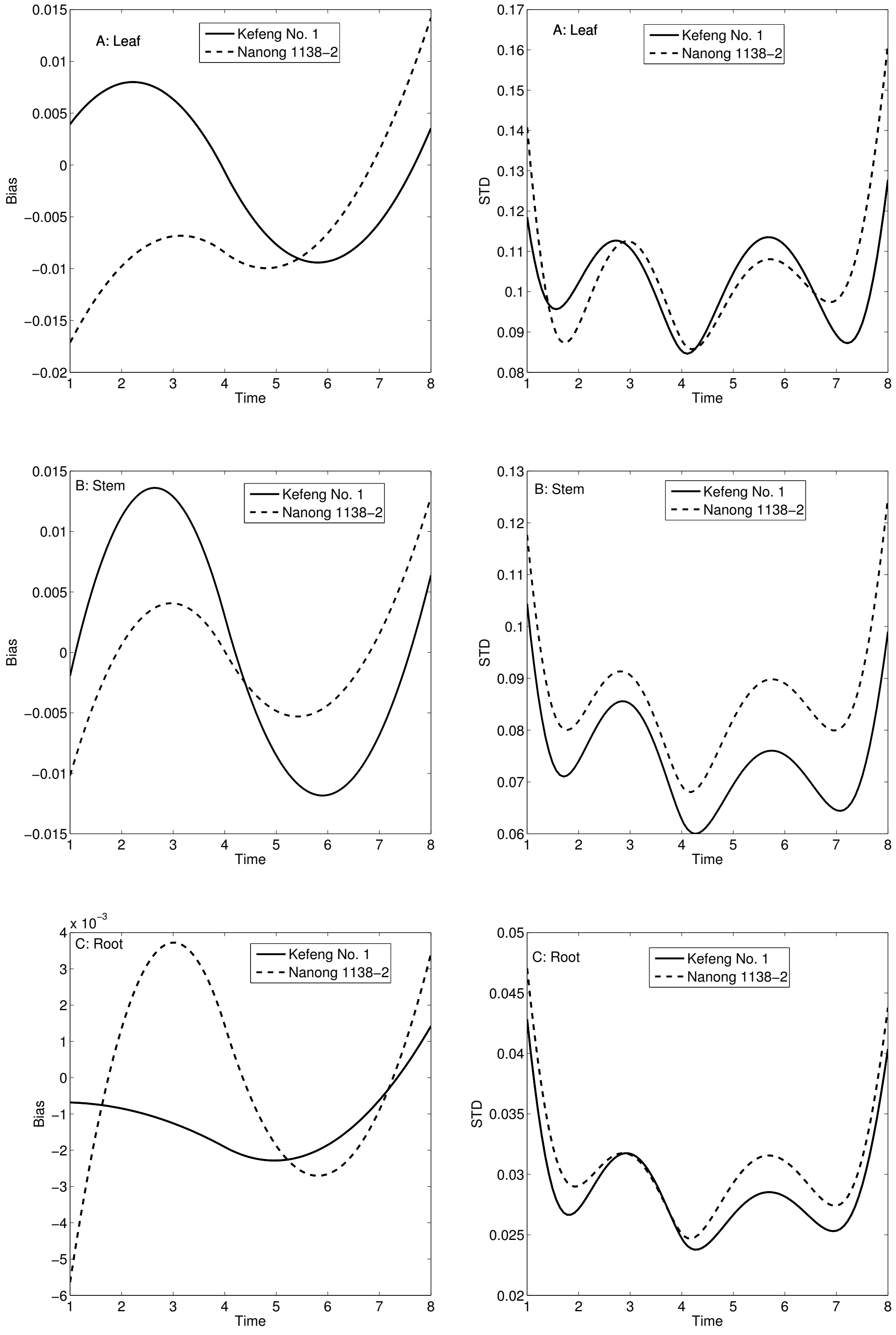

Figure 4: The point-wise biases and standard deviations (STDs) of the estimated mean trajectory curves of whole-plant leaf biomass, stem biomass and whole-plant root biomass (marked by A, B, C, respectively) for two different genotypes (plotted in solid and dashed lines, respectively,) of `QTL` located at 182 cM from the first marker in linkage group 12 in our simulation study.

QTL in all simulation replicates. So the power of the proposed test is 100%.

6 Discussion

In this article, we develop an innovative version of functional mapping by implementing a multivariate mixture model. The biological merit of this innovation is two-fold: (1) it is highly flexible to fit any form of trajectory curves and (2) it allows multiple dynamic traits to be analyzed simultaneously, providing a general way to test for pleiotropic control of QTLs. We have for the first time both derived a statistical method for estimating the multivariate mixture model and studied its statistical properties.

Perhaps, the most significant part of this study lies in its scientific validation and application to a real data set for QTL mapping in soybeans. The two significant QTLs detected by our new model have very intuitive interpretations that agree with developmental principles of trait formation and progression. It is impossible to obtain such an in-depth understanding of trait control by traditional QTL mapping approaches based on static traits. The new model allows numerous versatile tests for when and how a QTL exerts its effect on trait development. If a trait, such as leaf biomass or root biomass, experiences growth and senescence stages, the new model is able to test whether the detected QTLs determine the timing of a trait's developmental transitions.

A linear or semiparametric mixed model is a possible alternative to test for the effects of genetic markers on the multivariate traits (Thiébaut et al. 2002; Sithole and Jones 2007; Ghosh and Hanson 2010; Das and Daniels 2014). However, this framework is based on the assumption that QTLs controlling multivariate traits are observed directly, and are included in observed genetic markers. This assumption is beyond the scope of this manuscript.

References

Anholt, R. R. and T. F. C. Mackay (2004). Quantitative genetic analysis of complex behaviors in drosophila. *Nature Review: Genetics 5*, 838–849.

Cao, J. and J. Ramsay (2012). Linear mixed-effects modeling by parameter cascading. *Journal of the American Statistical Association 105*, 365–374.

Das, K. and M. J. Daniels (2014). A semiparametric approach to simultaneous covariance estimation for bivariate sparse longitudinal data. *Biometrics 70*(1), 33–43.

Das, K., J. Li, Z. Wang, C. Tong, G. Fu, Y. Li, M. Xu, K. Ahn, D. Mauger, R. Li, and R. Wu (2011). A dynamic model for genome-wide association studies. *Human genetics 129*(6), 629–639.

de Boor, C. (2001). *A Practical Guide to Splines*. New York: Springer.

Doebley, J., A. Stec, and L. Hubbard (1997). The evolution of apical dominance in maize. *Nature 386*, 485–488.

Frary, A., T. C. Nesbitt, A. Frary, S. Grandillo, E. van der Knaap, B. Cong, J. P. Liu, J. Meller, R. Elber, K. B. Alpert, and S. D. Tanksley (2000). fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. *Science 289*, 85–88.

Gallavotti, A., Q. Zhao, J. Kyozuka, R. B. Meeley, M. K. Ritter, J. F. Doebley, M. E. Pe, and R. J. Schmidt (2004). The role of barren stalk1 in the architecture of maize. *Nature 432*, 630–635.

Ghosh, P. and T. Hanson (2010). A semiparametric bayesian approach to multivariate longitudinal data. *Australian & New Zealand journal of statistics 52*(3), 275–288.

Haley, C. S. and S. A. Knott (1992). A simple method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity 69*, 315–324.

He, Q. L., A. Berg, Y. Li, C. E. Vallejos, and R. L. Wu (2010). Modeling genes for plant structure, development and evolution: Functional mapping meets plant ontology.

Trends in Genetics 26, 39–46.

Jansen, R. C. and P. Stam (1994). High resolution mapping of quantitative traits into multiple loci via interval mapping. *Genetics 136*, 1447–1455.

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

Knapp, S. J. (1991). Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred, and doubled haploid progeny. *Theor. Appl. Genet. 81*, 333–338.

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

Li, C. B., A. L. Zhou, and T. Sang (2006). Rice domestication by reducing shattering. *Science 311*, 1936–1939.

Li, R., S. W. Tsaih, K. Shockley, I. M. Stylianou, J. Wergedahl, B. Paigen, and G. A. Churchill (2006). Structural model analysis of multiple quantitative traits. *PLoS Genetics 2*(7), e114.

Li, Y. and R. L. Wu (2010). Functional mapping of growth and development. *Biological Reviews 85*, 207–216.

Lin, X. and R. Carroll (2006). Semiparamtric estimation in general repeated measures problems. *Journal of the Royal Statistical Society, Series B 68*, 69–88.

Ma, C. X., G. Casella, and R. L. Wu (2002). Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics 161*, 1751–1762.

MATLAB (2015). *version 8.6 (R2015b).* Natick, Massachusetts: The MathWorks Inc.

Paterson, A. H., E. S. LANDER, J. D. Hewitt, S. PETERSON, S. E. LINCOLN, and S. D. Tanksley (1988). Resolution of quantitative traits into mendelian factors by using a complete linkage map of restriction fragment polymorphisms. *Nature 335*, 721–726.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (Second ed.). New York: Springer.

Sen, S. and G. A. Churchill (2001). A statistical framework for quantitative trait mapping. *Genetics 159*, 371–387.

Sithole, J. S. and P. W. Jones (2007). Bivariate longitudinal model for detecting prescribing change in two drugs simultaneously with correlated errors. *Journal of Applied Statistics 34*(3), 339–352.

Thiébaut, R., H. Jacqmin-Gadda, G. Chêne, C. Leport, and D. Commenges (2002). Bivariate linear mixed models using sas proc mixed. *Computer methods and programs in biomedicine 69*(3), 249–256.

Trefethen, L. N. and D. Bau (1997). *Numerical linear angebra.* Philadelphia: Society for Industrial and Applied Mathematics (SIAM).

Wang, H., T. Nussbaum-Wagler, B. L. Li, Q. Zhao, Y. Vigouroux, L. L. M. Faller, K. Bomblies, and J. F. Doebley (2005). The origin of the naked grains of maize. *Nature 436*, 714–719.

Weiss, L. A., M. Abney, E. H. Cook, and C. Ober (2005). Sex-specific genetic architecture of whole blood serotonin levels. *American Journal of Human Genetics 76*, 33–41.

Weiss, L. A., L. Pan, M. Abney, and C. Ober (2006). The sex-specific genetic architecture of quantitative traits in humans. *Nature Genetics 38*, 218–222.

Weller, J. I. (1986). Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics 42*, 627–640.

Wu, R. L., J. Cao, Z. W. Huang, Z. Wang, J. Y. Gai, and C. E. Vallejos (2011). Systems mapping: How to improve the genetic mapping of complex traits through design principles of biological systems. *BMC Systems Biology 5*(84), 1–11.

Wu, R. L. and M. Lin (2006). Functional mapping - how to map and study the genetic architecture of dynamic complex traits. *Nature Reviews Genetics 7*, 229–237.

Wu, R. L., C.-X. Ma, and G. Casella (2007). *Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL*. New York: Springer-Verlag.

Yang, J., R. Wu, and G. Casella (2009). Nonparametric functional mapping of quantitative trait loci. *Biometrics 65*, 30–39.

Yang, R. Q. and S. Z. Xu (2007). Bayesian shrinkage analysis of quantitative trait loci for dynamic traits. *Genetics 176*, 1169–1185.

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

Zhao, W., W. Hou, R. C. Littell, and R. L. Wu (2005). Structured antedependence models for functional mapping of multivariate longitudinal quantitative traits. *Statistical Methods in Molecular Genetics and Biology 4*(1).