

An optimal strategy for functional mapping of dynamic trait loci

TIANBO JIN¹, JIAHAN LI², YING GUO³, XIAOJING ZHOU³, RUNQING YANG^{3,4*}
AND RONGLING WU^{2*}

¹ Department of Biology, Northwest University, National Engineering Research Center for Miniaturized Detection System, Xi'an 710069, China

² Department of Statistics, University of Florida, Gainesville, FL 32611 USA

³ Department of Mathematics, Heilongjiang Bayi Agricultural University, Daqing, 163319, People's Republic of China

⁴ School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai 201101, People's Republic of China

(Received 10 June 2009 and in revised form 25 November 2009; first published online 3 March 2010)

Summary

As an emerging powerful approach for mapping quantitative trait loci (QTLs) responsible for dynamic traits, functional mapping models the time-dependent mean vector with biologically meaningful equations and are likely to generate biologically relevant and interpretable results. Given the autocorrelation nature of a dynamic trait, functional mapping needs the implementation of the models for the structure of the covariance matrix. In this article, we have provided a comprehensive set of approaches for modelling the covariance structure and incorporated each of these approaches into the framework of functional mapping. The Bayesian information criterion (BIC) values are used as a model selection criterion to choose the optimal combination of the submodels for the mean vector and covariance structure. In an example for leaf age growth from a rice molecular genetic project, the best submodel combination was found between the Gaussian model for the correlation structure, power equation of order 1 for the variance and the power curve for the mean vector. Under this combination, several significant QTLs for leaf age growth trajectories were detected on different chromosomes. Our model can be well used to study the genetic architecture of dynamic traits of agricultural values.

1. Introduction

Functional mapping has proven to be a powerful approach for mapping quantitative trait loci (QTLs) responsible for dynamic traits whose phenotypic values change with time or other independent variables (Ma *et al.*, 2002; Wu *et al.*, 2004*a, b, c*; reviewed in Wu & Lin, 2006). A general strategy for QTL mapping is to construct a finite mixture model (Lander & Botstein, 1989; Jansen & Stam, 1994; Zeng, 1994) in which any individual from a mapping population is assumed to arise from one of the multiple possible QTL genotypes. For a quantitatively inherited trait, each group of QTL genotypes is

thought to follow a normal distribution with a mean expressed as the genotypic value and a variance common to each group. The central idea of functional mapping is that the mean vectors of QTL genotypes in a time limit are modelled by a biologically meaningful mathematical equation, whereas the covariance matrix is modelled in terms of its time-series autocorrelation structure (Ma *et al.*, 2002).

From a biological viewpoint, functional mapping is advantageous because it incorporates fundamental principles behind biological processes or networks into a QTL mapping framework. Much attempt has been made to embed various mathematical and differential equations in the mapping framework to reveal dynamic changes of genetic control over a biological or medical trait given its developmental nature. For example, the Richards growth equation has been used to model growth trajectories for various biological entities, bi-exponential curves to approximate HIV-1 dynamics (Perelson *et al.*, 1996), Emax

* Corresponding authors: School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai 200240, People's Republic of China. Tel: (8621) 34206146. Fax: (8621) 34206146. e-mail: runqingyang@sjtu.edu.cn
Department of Statistics, University of Florida, Gainesville, FL 32611. Tel: (352)392 3806. Fax: (352)392 8555. e-mail: rwu@stat.ufl.edu

models to fit drug response curve over a range of concentrations (Giraldo, 2003) and power curve to quantify the power required for a bird to fly at a particular speed (Tobalske *et al.*, 2003). The parameters in these equations have particular biological or clinical means. Thus, the results from functional mapping are biologically more interpretable and relevant than those from general approaches without biological incorporation. Also, through testing, individually or in combination, for the curves parameters in functional mapping, many biologically interesting questions can be asked and addressed at the interplay between genetic actions and developmental patterns (Wu *et al.*, 2004a).

Functional mapping attempts to estimate mathematical parameters that model the time-dependent vector of QTL effects and residual covariance matrix rather than estimate all elements in the vector and matrix and, therefore, displays strong statistical robustness due to the estimation of fewer parameters (Ma *et al.*, 2002). In previous functional mapping, approaches for modelling the structure of the covariance matrix include the autoregressive (AR) (Diggle *et al.*, 2002) and structured antedependence (SAD) model (Núñez-Antón & Zimmerman, 2000). Whereas the AR model assumes variance stationarity and correlation stationarity, these two assumptions are relaxed for the SAD model that approximates the variance and correlation as a function of time. Zhao *et al.* (2005) found that the AR and SAD models cannot be replaced by each other, although the SAD model seems to be more general in terms of its non-stationary nature. In statistical modelling of longitudinal variables, many other models have been available to structure the covariance matrix, all of which can be potentially used for functional mapping. Thus far, a comprehensive treatment of modelling the covariance matrix has not been fully explored. Meyer (2001) systematically reviewed various approaches for estimating covariance functions related to cattle genetic and breeding studies. The purpose of this study is to incorporate several most commonly used covariance-structuring approaches into the framework of functional mapping, aimed at providing a general strategy for the choice of an optimal approach for structuring the covariance matrix for practical data sets.

2. Model and method

(i) Genetic model

We start with a simple F_2 population from two homozygous lines. Consider one segregating QTL, with three genotypes, QQ (2), Qq (1) and qq (0), responsible for dynamic changes of a trait. All n progeny in the pedigree are measured for the dynamic trait at each of T time points. The trait

phenotype (y_i) of progeny i measured at time t can be expressed as

$$y_i(t) = \sum_{j=0}^2 \xi_{ij} \mu_j(t) + e_i(t), \quad (1)$$

where ξ_{ij} is an indicator variable describing a possible genotype j ($j=2,1,0$) of the QTL for progeny i and defined as 1 if a particular genotype is observed and 0 otherwise, $\mu_j(t)$ is the genotypic value of the trait for QTL genotype j at time t , and $e_i(t)$ is the residual including the aggregate effect of polygenes and error effect and distributed as $N(0, \sigma^2(t))$. The covariance between residuals at two different time points, t_1 and t_2 , is expressed as $\sigma(t_1, t_2)$. The variances and covariances form a residual covariance matrix Σ . The probability of ξ_{ij} to take 1 or 0 can be inferred from the two-locus genotypes of the flanking markers that bracket the QTL, expressed as the conditional probability of the QTL genotype given marker genotypes described in terms of the recombination fractions between the QTL and the two markers.

(ii) Likelihood function

The likelihood of the dynamic trait with T -dimensional measurements $y_i = (y_i(1), \dots, y_i(T))$, and marker data, M , can be represented by a multivariate mixture model

$$L(\theta, u_j, \Sigma | y, M) = \prod_{i=1}^n \left[\sum_{j=0}^2 \varpi_{ji} f_j(y_i) \right] \quad (2)$$

where the mixture proportion, ϖ_{ji} , is the conditional probability of QTL genotype given a marker genotype for progeny i , which describes the position of the QTL (θ) within the marker interval, and $f_j(y_i)$ is the multivariate normal density with mean vector $u_j = (\mu_j(1), \dots, \mu_j(T))$ and variance matrix Σ .

Several numerical methods can be used to solve the maximum likelihood estimates (MLEs) of the parameters (θ, u_j, Σ) that define the likelihood function (2). Like a general treatment, functional mapping estimates θ with a grid approach, i.e., the likelihood values are plotted against the length of a linkage group and the peak of the likelihood profile is then regarded as the location of QTL. However, functional mapping attempts to estimate and test the parameters that model the structures of (u_j, Σ), instead of estimating all the elements contained within (u_j, Σ). The parameter vectors for modelling the mean vector and covariance matrix are denoted as Ω_{u_j} for QTL genotype j and Ω_{Σ} , respectively.

(iii) Submodel for structuring the mean vector

If the traits studied follow a certain pattern of development, mathematical equations can be used to

model the mean vector. For example, growth trajectories for biological entities at different organizational levels can be approximated by the Richards growth equation. The power required for a bird to fly at a particular speed can be identified by incorporating power curves. Bi-exponential curves and Emax models have been utilized to approximate HIV-1 dynamics after drug treatment and drug response at different concentrations, respectively. In general, the time-dependent mean vector of a dynamic trait is described by a mathematical function

$$u_i(t) = g(t; \Omega_{u_i}), \tag{3}$$

where Ω_{u_i} is a set of curve parameters that define the dynamic mechanism of a particular trait.

In practice, many dynamic traits may not be described by a mechanistic model, relying on statistical fitting. Functional mapping for these kinds of traits can be based on non-parametric modelling. Statistically, many non-parametric approaches have been available; for example, B-spline, P-spline or orthogonal polynomials.

(iv) *Submodel for structuring the covariance matrix*

An unstructured matrix would be best to fit the residual covariance since it can approximate the true structure, but with the measured time points increased, the unstructured matrix is too large to be solved. The structured models with more parameters are more appropriate than fewer ones, but the number of parameters in structured models is more, and the structured models is more closed to unstructured ones. So, the covariance matrix Σ should be structured to consider the inherent pattern of time-dependent variances and covariances, reduce the number of unknown parameters being estimated and decrease the effect of noises on the precision of QTL mapping. The simplest approach for structuring the matrix is based on the AR model, in which the variance and covariance assumes the stationarity, i.e. there is the same residual variance for the trait at each time point and the covariance between different measurements decays purely with time interval. The variance stationarity assumption has been relaxed by implementing the transform-both-sides (TBS) model (Carroll & Puppert, 1984) within the context of functional mapping (Wu *et al.*, 2004*b*). The SAD model was used to characterize the non-stationary variance and correlation (Núñez-Antón & Zimmerman, 2000).

However, the AR and SAD models can be only used in their respective situations. The residual variances and covariances characterized by both the models include contributions by other QTLs and errors. Because the residual may also change with time interval in a particular manner, the AR and SAD

models will display different powers to detect QTL. Here, we propose a general model for structuring the residual covariance matrix, which covers the features of the AR and SAD models.

According to the theory of matrix algebra, we can decompose Σ into the product of standard deviations and correlations expressed as

$$\Sigma = V^{1/2} R V^{1/2}, \tag{4}$$

where $V^{1/2}$ is the diagonal matrix that is composed of the residual standard deviation $\sigma(t)$ for $t = 1, 2, \dots, T$, and R is the correlation matrix. Approaches have been available to model the structures of V and R .

(a) *Standard deviation function*

A number of variance functions have been proposed to model heterogeneous variances across different time points (Foulley *et al.*, 1998). These functions include a step function or a polynomial function as follows:

$$\sigma^2(t) = \begin{cases} \sigma^2(1 + \sum_{r=1}^v b_r t^r), \\ \sigma^2 e^{(1 + \sum_{r=1}^v b_r t^r)}, \\ e^{(\sigma^2 + \sum_{r=1}^v b_r t^r)}, \text{ or } \ln \sigma^2(t) = \sigma^2 + \sum_{r=1}^v b_r t^r, \end{cases} \tag{5}$$

where σ^2 is the variance at the intercept, b_r is the coefficient of the variance function and v is the order of polynomial fit. The standard deviation function is the square root of $\sigma^2(t)$.

(b) *Correlation function*

Correlations in R between different time points can be modelled as a function of time interval. Correlation functions with one or more parameters can be either stationary or non-stationary model without variance. The simplest correlation function is those specified by a single parameter. Table 1 lists several stationary models for correlation functions with a single parameter. Equation (4) can be reduced to the first-order AR (AR(1)) covariance structure when $\sigma^2(t)$ is equal to σ^2 and when the correlation function is specified by the AR(1) model. All the parameters that describe the standard deviation and correlation functions are arrayed in Ω_v .

More complicated models, such as autoregressive moving average (ARMA), can be considered to model correlation functions. The ARMA model typically contains two parts, an AR and a moving average (MA). The model is usually referred to as an ARMA (p, q) model, where p is the order of the AR part and q is the order of the MA part. For some data, non-stationary models perform better than stationary

Table 1. Common stationary models for correlation functions with a single parameter

Model	Correlation function	Range of parameter
Compound symmetry (CS)	ρ	$-1 < \rho < 1$
First-order AR	ρ^l	$-1 < \rho < 1$
Exponential (EXP)	$e^{-\rho l}$	$\rho > 0$
Gaussian (GAU)	$e^{-\rho^2 l}$	$\rho > 0$
Cauchy (CAU)	$(1 + \rho^2 l^2)^{-1}$	$\rho > 0$
Hyperbolic cosine (HYP)	$(\cosh(\pi \rho l / 2))^{-1}$	$\rho > 0$
Uniform distribution (UNI)	$\sin(\rho l) / (\rho l)$	$\rho > 0$
Triangular distribution (TRI)	$(1 - \cos(\rho l)) / \rho l$	$\rho > 0$
'Damped' exponential (DEX)	$e^{-\rho l^\phi}$	$\rho > 0, 0 < \phi < 2$

Note: l denotes the lag in time for a pair of time points.

models in the functional mapping (Zhao *et al.*, 2005; Lin & Wu, 2006; Cui *et al.*, 2008). The SAD model is one class of non-stationary models (Zimmerman & Núñez-Antón, 2001). The form of the SAD model was described in Zhao *et al.* (2005) who provided a closed form of the SAD-structured covariance matrix.

(v) Hypothesis tests

Whether there is a significant QTL that controls the developmental process of a dynamic trait should be tested by calculating the likelihood ratio statistic, expressed as

$$LR = 2[\ln L(\tilde{\Omega}_u, \tilde{\Omega}_v | y) - \ln L(\hat{\theta}, \hat{\Omega}_{u_j}, \hat{\Omega}_v | y, M)], \quad (6)$$

where $\tilde{\Omega}_u$ and $\tilde{\Omega}_v$ are the MLEs of the curve parameters and the parameters that structure the residual variance matrix under the null hypothesis (there is no QTL) and $\hat{\theta}$, $\hat{\Omega}_{u_j}$ and $\hat{\Omega}_v$ are the MLEs of the QTL position, genotype-specific curve parameters and matrix-structuring parameters under the alternative hypothesis (there is a QTL), respectively.

The plug-in LR value calculated by eqn (6) follows asymptotically a chi-square distribution with degrees of freedoms equal to the difference in the numbers of parameters between the models with the two different hypotheses. The critical threshold for claiming the existence of a QTL can be empirically determined on the basis of permutation tests.

Functional mapping allows for the test of various biologically relevant hypotheses regarding the genetic control of developmental timing. Wu *et al.* (2004a) formulated a number of such hypotheses and applied them to a practical example in poplar trees.

(vi) Model selection

For a practical data set, we should know which model is optimal to explain the structure of the covariance matrix. The selection of an optimal

covariance-structuring model is crucial for increasing the power of functional mapping and its estimation precision. In statistics, there are several criteria for model selection although none of them perform the best under all circumstances. Here, we used the commonly used Bayesian information criterion (BIC) (Schwarz, 1978) as the model selection criterion of the optimal combination of two submodels with parsimonious parameters. The BIC is defined as

$$BIC = -2 \ln L(\hat{\Omega}_{u_j}, \hat{\Omega}_v | M) + \text{dimension}(\Omega_{u_j}, \Omega_v | M) \ln(n), \text{ for } j = 2, 1, 0.$$

where $\hat{\Omega}_{u_j}$ and $\hat{\Omega}_v$ are the MLEs of parameters under a particular model (M), $\text{dimension}(\Omega_{u_j}, \Omega_v | r)$ represents the number of independent parameters under this model and n is the total number of observations on all individuals. The optimal model is the one that displays the minimum BIC value.

3. Results

A doubled-haploid (DH) population with 111 lines was generated by crossing an indica rice variety Gui-630 and a japonica rice variety Taiwanjing (Weng *et al.*, 2000). A linkage map composed of 175 RFLP markers was constructed for the DH population, covering a total length of 1225 cM. This DH population was grown with replicates in a field trial (Zhou *et al.*, 2001). For each plant, the number of developed leaves on the main stem was counted, and the length of the developing leaf was measured every 5–7 days from day 30 after sowing until the full development of the leaf. These measured data were used to estimate the leaf age of a plant (y) using

$$y = \text{Number of developed leaves} + \frac{\text{Length of the developing leaf}}{\text{Final length of the developing leaf}}$$

Table 2. BIC information criteria for leaf age data from different combinations of sub-models

Variance function ^a	Mean vector	Correlation structure										
		CS	AR	EXP	GAU	CAU	HYP	UNI	TRI	DEX	SAD ^b	ARMA ^c
0	RIC	1537.7	2072.0	2390.4	420.3	1888.9	1790.8	452.2	946.9	2534.3	1848.1	2401.2
	POW	1582.5	2260.5	2507.1	407.9	1920.6	1789.7	434.7	987.4	2653.2	1729.3	2565.6
	PAR	1599.8	2319.8	2566.2	423.5	1883.3	1806.3	452.4	995.5	2704.2	1773.9	2617.4
	LEG	1535.1	2069.3	2387.7	400.4	1886.7	1788.1	449.3	963.6	2531.6	1844.3	2398.3
1	RIC	2378.7	2565.7	3206.5	313.1	2416.1	1880.5	635.3	963.5	2466.6	1894.4	3065.5
	POW	2411.5	2815.9	3033.7	308.9	1642.2	1847.7	653.1	1138.7	3221.3	1776.8	2545.9
	PAR	2436.7	2766.1	3012.9	379.6	2420.3	2812.4	770.3	1096.2	2792.3	1803.9	2604.7
	LEG	2357.3	2881.3	3167.5	375.1	2280.9	2186.9	807.4	884.6	2850.9	1891.2	2351.9
2	RIC	2037.4	2323.4	2665.8	323.2	1851.5	1735.6	464.7	1117.5	2692.2	1779.8	2532.2
	POW	2095.6	2385.8	2712.8	425.1	1815.4	1818.5	485.6	1103.1	2734.8	1806.9	2747.8
	PAR	2003.3	2159.0	2516.2	394.3	1915.4	1807.0	485.2	1108.5	2613.6	1894.1	2532.2
	LEG	2003.3	2159.0	2597.2	397.2	1909.5	1807.0	485.2	1115.6	2613.6	1897.4	2175.2

^a The variance structure is modelled by a power equation of different orders 0–2.

^b SAD is specified as the first-order SAD model.

^c ARMA is specified as the autoregressive moving average model with the first-order AR part and the first-order MA part.

The growth of leaf age of a plant can be fit by a mathematical equation. We will use three equations to model leaf age growth which include:

(1) Richards’ (RIC) model expressed as

$$g(t) = a(1 + be^{-ct})^{1/(1-d)},$$

(2) A power curve (POW) expressed as

$$g(t) = c + at^b,$$

(3) A parabolic curve (PAR) expressed as

$$g(t) = \frac{a + bt}{1 + ct + dt^2},$$

where the parameters a , b , c and/or d in the three above equations can be explained in a biological sense. By estimating these parameters and testing their differences among different QTL genotypes, we can determine whether and how a QTL triggers an effect on leaf age growth trajectories.

Considering the data of leaf age growth may not follow any curve, we use a non-parametric approach based on Legendre (LEG) polynomials (Meyer, 2001; Yang *et al.*, 2006) to approximate the dynamic change of leaf age in development. In this study, the Legendre polynomial of order 2 is considered and expressed as

$$g(t) = b_0 + b_1\tau + b_2(1.5\tau^2 - 0.5),$$

where $\tau = -1 + 2 \times \frac{t-5}{34}$.

These parametric and non-parametric approaches are incorporated into the framework of functional mapping, along with modelling the structure of the covariance matrix by the standard deviation function, $\sigma^2(t) = \sigma^2(1 + \sum_{r=1}^v b_r t^r)$ with different orders (see eqn (5)), and the correlation function (Table 1). All

possible combinations between standard deviation functions and correlation functions are each incorporated into functional mapping. By calculating the BIC values, an optimal combination can be determined. Table 2 tabulates the BIC values for each combination of standard deviation and correlation functions under different submodels of the mean vector. It can be seen that different combinations give markedly different BIC values. In general, the Gaussian model of the correlation function performs best among all the models, followed by the uniform distribution model. Combining the Gaussian model and the standard deviation function of order 1 with the power curve for the mean vector provides an optimal fit of the data (Table 2).

Figure 1 gives the profile of LR test statistics across all the 12 chromosomes under the optimal combination. Permutation tests based on 1000 replicates have been used to determine the genome-wide empirical critical threshold for significance declaration. There are a total of six peaks for the LR profile that surpass the threshold, suggesting the detection of QTL for leaf age growth. These QTLs are located on different chromosomes 1, 2, 6, 7, 10 and 11, respectively (Fig. 1). Table 3 tabulates the estimates of the curve parameters for each QTL genotype and the parameters that model the structure of the covariance matrix. As shown by the estimates of the standard errors obtained with the bootstrapping approach, the parameter estimation has reasonably high precision. Using the estimated curve parameters, two curves for leaf age growth are drawn for each QTL genotype (Fig. 2). Based on the differences in curve shape between the two QTL genotypes, all the six QTLs detected display a substantial effect on leaf age growth during the entire development of plants.

Table 3. The MLEs of the model parameters and their asymptotic standard errors in the parentheses under the combination of POW and GAU distribution models for leaf age growth trajectories

QTL	Chr	Position Marker interval	QQ			qq			Covariance matrix		
			a_2	b_2	c_2	a_0	b_0	c_0	σ^2	ρ	$b_1(\times 10^{-3})$
1	1	C178-RZ776	0.967 (0.134)	0.592 (0.026)	6.269 (0.179)	0.692 (0.135)	0.642 (0.027)	5.864 (0.150)	0.224 (0.006)	0.628 (0.001)	5.121 (0.311)
2	3	C1032-RZ142	0.688 (0.149)	0.643 (0.029)	5.867 (0.170)	0.942 (0.143)	0.598 (0.030)	6.234 (0.216)	0.224 (0.007)	0.629 (0.002)	0.013 (0.003)
3	6	RG4241-C358	0.948 (0.102)	0.595 (0.020)	6.298 (0.143)	0.667 (0.085)	0.650 (0.017)	5.897 (0.104)	0.219 (0.007)	0.627 (0.002)	1.101 (0.211)
4	7	RG128-R1394	0.649 (0.150)	0.655 (0.031)	5.926 (0.228)	0.998 (0.153)	0.584 (0.032)	6.135 (0.198)	0.225 (0.006)	0.628 (0.001)	0.0112 (0.001)
5	10	RZ892-RZ561	0.918 (0.162)	0.602 (0.033)	6.301 (0.212)	0.701 (0.139)	0.639 (0.033)	5.807 (0.228)	0.2187 (0.014)	0.628 (0.002)	43.222 (7.489)
6	11	C189-C6A	0.666 (0.151)	0.650 (0.031)	5.895 (0.158)	0.991 (0.688)	0.587 (0.061)	6.190 (0.866)	0.217 (0.012)	0.627 (0.002)	0.086 (0.001)

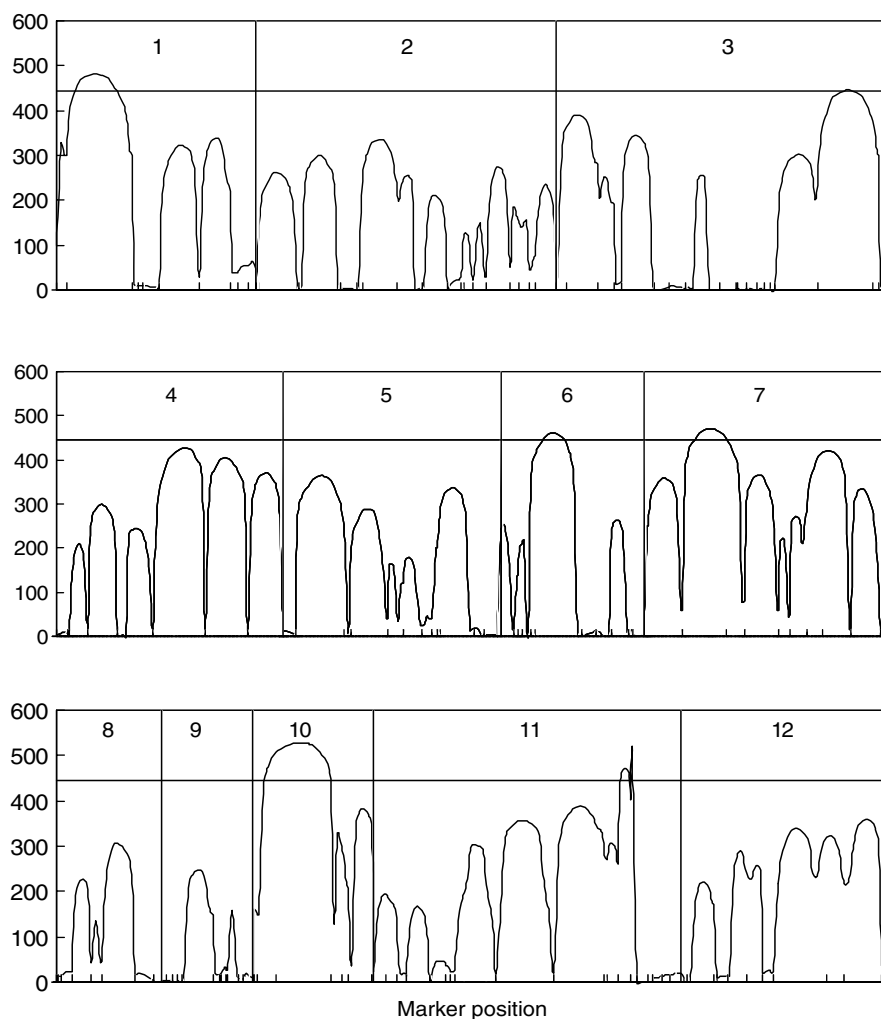


Fig. 1. The profile of the log-likelihood ratios between the full and reduced (no QTL) model for leaf age growth trajectories across the genome. The genome-wide threshold value is given as a horizontal line. Ticks indicate the positions of markers on linkage groups.

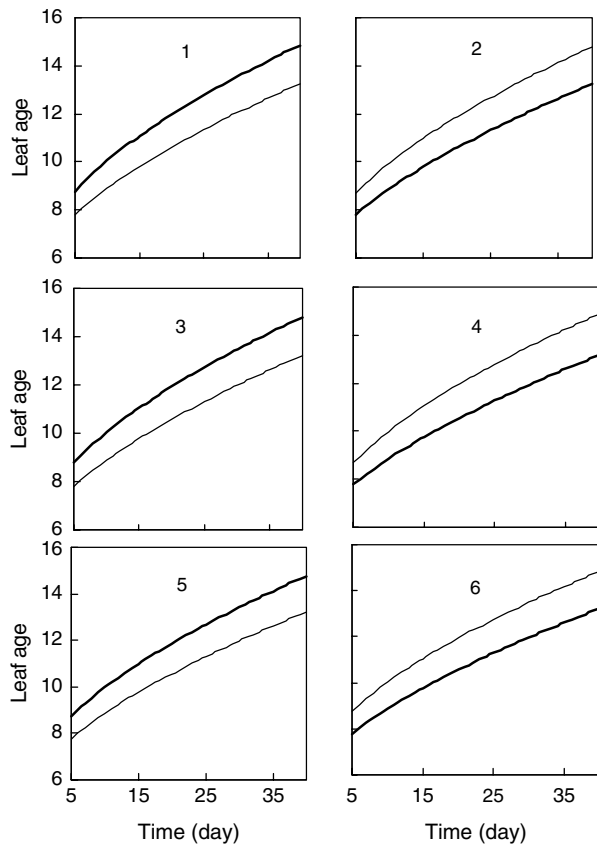


Fig. 2. Two growth curves each presenting two groups of genotypes (bold lines are for QQ) at the six QTLs, symbolized by 1, 2, ..., 6, detected on chromosome 1, 3, 6, 7, 10 and 11, respectively.

4. Discussion

Functional mapping incorporates fundamental principles behind biological processes or networks that are bridged with mathematical functions into a QTL mapping framework (Ma *et al.*, 2002; Wu & Lin, 2006). Functional mapping involves not only modelling time-dependent expected mean vectors of QTL genotypes but also the structure of the within-subject residual covariance matrix. The mean vectors of QTL genotypes were usually modelled by biological meaningful functions including a few key parameters that define the shape and function of a particular biological network (see also Wu *et al.*, 2002). Because of this implementation with biological principles, functional mapping provides more sensible and interpretable results than previous QTL mapping approaches (Wu & Lin, 2006).

Functional mapping is also advantageous in statistics because the structure, rather than all the elements, of the mean vector and the covariance matrix is estimated, which can largely reduce the number of parameters to be estimated. Original functional mapping capitalizes on a simple statistical process, such as the AR process, to model the

covariance structure (Ma *et al.*, 2002). In subsequent modelling, several more complicated processes, such as SAD, have been implemented, leading to significant results for QTL detection (Zhao *et al.*, 2005). However, there are a number of other statistical processes (see Henderson, 1982; Jaffrezic & Pletcher, 2000; Meyer, 2001) that may better suit functional mapping than the AR or SAD models given a particular data set. In this article, we implement a comprehensive set of statistical models for the structure of the covariance matrix, provide a systematic procedure for selecting an optimal approach for the mean-covariance structure, and then exemplify the utility of our model selection idea using a real example from the rice genetic mapping project (Weng *et al.*, 2000; Zhou *et al.*, 2001).

Our implementation includes the modelling of the covariance matrix with parametric or parametric variance and correlation functions. Depending on the complexity of a problem, parametric functions may be specified by one single or more parameters, or can be stationary and non-stationary. In some situations, no parametric functions are available to capture the pattern of the covariance matrix, in which we use non-parametric modelling based on the Legendre polynomial to fit changes of the variances and covariances (Gao & Yang, 2006; Yang *et al.*, 2006). Incorporated into a random regression model (Kirkpatrick *et al.*, 1994; Meyer, 2001), the Legendre polynomial displays the changes of the covariance matrix in a linear way, thus largely facilitating the result interpretation and computation of functional mapping. In a data analysis using a published example for leaf age growth in rice, all possible combinations of the submodels for the mean vector and covariance matrix have been considered. An optimal combination based on the model selection criterion, BIC, was determined, which includes the Gaussian model for the correlation structure, the power equation of order 1 for the variance structure and power curve for the mean vector (Table 2). Although the result was not shown in Table 2, such a combination is in fact better than the fitting of the covariance matrix with the Legendre polynomial for this particular example. For example, the Legendre polynomial of the best order 2 gives the BIC values of 2227 to 2491.

The model framework proposed in this article will provide a powerful tool for functional mapping of QTL that determine the developmental pattern of dynamic traits. The framework allows for the choice of the best fitness of the mean-covariance structure for a particular data. With this framework, we are in an excellent position to study and map the detailed genetic architecture of complex traits that undergo developmental changes in an organism's lifetime.

The preparation of this manuscript was partially supported by NSF grant (No. 0540745) and the National Natural Science Foundation of China (No. 30972077).

References

- Carroll, R. J. & Puppert, D. (1984). Power-transformations when fitting theoretical models to data. *Journal of the American Statistical Association* **79**, 321–328.
- Cui, Y. H., Wu, R. L., Casella, G. & Zhu, J. (2008). Nonparametric functional mapping of quantitative trait loci underlying programmed cell death. Statistical applications in genetics and molecular. *Biology* **1**, 1–30.
- Diggle, P. J., Heagerty, P., Liang, K. Y. & Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford, UK: Oxford University Press.
- Foulley, J. L., Quaas, R. L. & D'Arnoldi, T. C. (1998). A link function approach to heterogeneous variance components. *Genetics, Selection, Evolution* **30**, 27–44.
- Gao, H. J. & Yang, R. Q. (2006). Composite interval mapping of QTL for dynamic traits. *Chinese Science Bulletin* **51**, 1857–1862.
- Giraldo, J. (2003). Empirical models and Hill coefficients. *Trends Pharmacological Sciences* **24**, 63–65.
- Henderson, C. R. Jr. (1982). Analysis of covariance in the mixed model: higher level, no homogenous, and random regressions. *Biometrics* **38**, 623–640.
- Jaffrezic, F. & Pletcher, S. D. (2000). Statistical models for estimating the genetic basis of repeated measures and other function-valued traits. *Genetics* **156**, 913–922.
- Jansen, R. C. & Stam, P. (1994). High resolution mapping of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447–1455.
- Kirkpatrick, M., Hill, W. G. & Thompson, R. (1994). Estimating the covariance structure of traits during growth and ageing, illustrated with lactating dairy cattle. *Genetical Research* **64**, 57–69.
- Lander, E. S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.
- Lin, M. & Wu, R. (2006). A joint model for nonparametric functional mapping of longitudinal trajectory and time-to-event. *BMC Bioinformatics* **7**, 138.
- Ma, C. X., Casella, G. & Wu, R. L. (2002). Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics* **61**, 1751–1762.
- Meyer, K. (2001). Estimating genetic covariance functions assuming a parametric correlation structure for environmental effects. *Genetics, Selection, Evolution* **33**, 557–585.
- Núñez-Antón, V. & Zimmerman, D. L. (2000). Modeling nonstationary longitudinal data. *Biometrics* **56**, 699–705.
- Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. & Ho, D. D. (1996). HIV-1 dynamics *in vivo*: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**, 1582–1586.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.
- Tobalske, B. W., Hedrick, T. L., Dial, K. P. & Biewener, A. A. (2003). Comparative power curves in bird flight. *Nature* **421**, 363–366.
- Weng, Q., Wu, W., Li, W., Liu, H., Tang, D., Zhou, Y. & Zhang, Q. (2000). Construction of an RFLP linkage map of rice using DNA probes from two different sources. *Journal of the Fujian Agricultural University* **29**, 129–133.
- Wu, R. L. & Lin, M. (2006). Functional mapping -How to study the genetic architecture of dynamic complex traits. *Nature Reviews Genetics* **7**, 229–237.
- Wu, R. L., Ma, C. X., Lin, M. & Casella, G. (2004a). A general framework for analyzing the genetic architecture of developmental characteristics. *Genetics* **166**, 1541–1551.
- Wu, R. L., Ma, C. X., Lin, M., Wang, Z. H. & Casella, G. (2004b). Functional mapping of growth QTL using a transform-both-sides logistic model. *Biometrics* **60**, 729–738.
- Wu, R. L., Wang, Z. H., Zhao, W. & Cheverud, J. M. (2004c). A mechanistic model for genetic machinery of ontogenetic growth. *Genetics* **168**, 2383–2394.
- Wu, W. R., Zhou, Y., Li, W., Mao, D. & Chen, Q. (2002). Mapping of quantitative trait loci based on growth models. *Theoretical and Applied Genetics* **105**, 1043–1049.
- Yang, R. Q., Tian, Q. & Xu, S. Z. (2006). Mapping quantitative trait loci for longitudinal traits in line crosses. *Genetics* **173**, 2339–2356.
- Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.
- Zhao, W., Chen, Y. Q., Casella, G., Cheverud, J. M. & Wu, R. L. (2005). A non-stationary model for functional mapping of longitudinal quantitative traits. *Bioinformatics* **21**, 2469–2477.
- Zhou, Y., Li, W., Wu, W., Chen, Q., Mao, D. & Worland, A. J. (2001). Genetic dissection of heading time and its components in rice. *Theoretical Applied Genetics* **102**, 1236–1242.
- Zimmerman, D. L. & Núñez-Antón, V. (2001). Parametric modeling of growth curve data: an overview (with discussion). *Test* **10**, 1–73.