

## Genetics and population analysis

## A non-stationary model for functional mapping of complex traits

Wei Zhao<sup>1</sup>, Ying Q. Chen<sup>1</sup>, George Casella<sup>1</sup>, James M. Cheverud<sup>2</sup> and Rongling Wu<sup>1,\*</sup><sup>1</sup>Department of Statistics, University of Florida, Gainesville, FL 32611, USA and <sup>2</sup>Department of Anatomy and Neurobiology, Washington University Medical School, St. Louis, MO 63110, USA

Received on October 6, 2004; revised on February 23, 2005; accepted on March 7, 2005

Advance Access publication March 15, 2005

## ABSTRACT

**Summary:** Understanding the genetic control of growth is fundamental to agricultural, evolutionary and biomedical genetic research. In this article, we present a statistical model for mapping quantitative trait loci (QTL) that are responsible for genetic differences in growth trajectories during ontogenetic development. This model is derived within the maximum likelihood context, implemented with the expectation–maximization algorithm. We incorporate mathematical aspects of growth processes to model the mean vector and structured antedependence models to approximate time-dependent covariance matrices for longitudinal traits. Our model has been employed to map QTL that affect body mass growth trajectories in both male and female mice of an F<sub>2</sub> population derived from the Large and Small mouse strains. The results from this model are compared with those from the autoregressive-based functional mapping approach. Based on results from computer simulation studies, we suggest that these two models are alternative to one another and should be used simultaneously for the same dataset.

**Contact:** rwu@mail.ifas.ufl.edu

## 1 INTRODUCTION

The traits whose phenotypes change with time or any other independent variable are important in agriculture, biological and medical research. For this reason, the genetic analysis of these so-called longitudinal traits has been a focus of a number of statistical and genetic studies aimed at predicting the dynamic change of genetic control at the genotype (Kirkpatrick and Heckman, 1989; Pletcher and Jaffrézic, 2002; Jaffrézic *et al.*, 2003) or individual quantitative trait locus (QTL) levels (Wu *et al.*, 2002).

More recently, a collection of statistical methods implemented with growth model theories have been proposed to map QTL that govern growth trajectories using molecular linkage maps (Wu *et al.*, 2002, 2004a,b; Ma *et al.*, 2002). The basic principle of this method, called functional mapping, is to express the genotypic means of a QTL at different time points in terms of a continuous growth function with respect to time  $t$ . Under this principle, the parameters describing the shape of growth curves, rather than the genotypic means as expected in traditional mapping strategies, are estimated within a maximum likelihood framework. Also unlike traditional mapping strategies, functional mapping estimates the parameters that model the structure of the (co)variance matrix among multiple different time points and, therefore, largely reduces the number of parameters

being estimated for variances and covariances, especially when the number of time points is large.

Functional mapping is, in spirit, a statistical problem of jointly modelling mean–covariance structures in longitudinal studies, an area that has recently received considerable interest in the statistical literature (Pourahmadi, 1999, 2000; Pan and Mackenzie, 2003; Daniels and Pourahmadi, 2002; Wu and Pourahmadi, 2003). However, in contrast to general longitudinal modelling, functional mapping integrates the estimation and test process of its underlying parameters within a mixture-based likelihood framework. Each mixture component in the likelihood model is given a particular biological rationale. For a finite mixture model, each observation is assumed to have arisen from one of a known or unknown number of components, each component being modelled by a density from the parametric family. Assuming that there are  $J$  QTL genotypes contributing to a longitudinal trait measured at  $\tau$  time points (denoted by  $\mathbf{y}$ ), this mixture model is expressed as

$$\mathbf{y} \sim p(\mathbf{y}|\varpi, \mathbf{m}_j, \Sigma) = \varpi_1 f_1(\mathbf{y}; \mathbf{m}_1, \Sigma) + \dots + \varpi_J f_J(\mathbf{y}; \mathbf{m}_J, \Sigma), \quad (1)$$

where  $\varpi = (\varpi_1, \dots, \varpi_J)$  are the mixture proportions (i.e. QTL genotype frequencies) which are constrained to be non-negative and sum to unity,  $\mathbf{m}_j$  is a vector that contains the parameters specific to component (or QTL genotype)  $j$ , and  $\Sigma$  includes the parameters common to all components (residual variances and covariances). We use the multivariate normal distribution to model each density, and for individual  $i$  it is expressed as

$$f(\mathbf{y}_i; \Omega_j) = \frac{1}{(2\pi)^{\tau/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{y}_i - \mathbf{m}_j) \Sigma^{-1} (\mathbf{y}_i - \mathbf{m}_j)' \right], \quad (2)$$

where  $\mathbf{y}_i = \{y_i(t)\}_{t=1}^{\tau} = [y_i(1), \dots, y_i(\tau)]$  is a vector of observation measured at  $\tau$  time points and  $\mathbf{m}_j = \{\mu_j(t)\}_{t=1}^{\tau} = [\mu_j(1), \dots, \mu_j(\tau)]$  is a vector of expected values for QTL genotype  $j$  at different points. At a particular time  $t$ , the relationship between the observation and expected mean can be described by a regression model,

$$y_i(t) = \sum_{j=1}^J \xi_{ij} \mu_j(t) + e_i(t), \quad (3)$$

where  $\xi_{ij}$  is the indicator variable denoted as 1 if a QTL genotype  $j$  is considered for individual  $i$  and 0 otherwise;  $e_i(t)$  is the residual error (i.e., the accumulative effect of polygenes and errors) that is iid

\*To whom correspondence should be addressed.

(independently and identically distributed) normal with mean zero and variance  $\sigma^2(t)$ . The errors at two different time points,  $t_1$  and  $t_2$ , are correlated with covariance  $\sigma(t_1, t_2)$ .

The mixture proportions,  $\varpi_1, \dots, \varpi_J$ , in Equation (1) can be viewed as the prior probabilities of QTL genotypes in the mapping population. When known markers that are co-segregating with the putative QTL are incorporated into the mixture model, these mixture proportions are substituted by the conditional probabilities of QTL genotypes given the marker genotypes. Because a given individual  $i$  has known marker genotype, these conditional probabilities can be simply denoted as  $\varpi_{1|i}, \dots, \varpi_{J|i}$ . These conditional probabilities can be derived, with expression depending on the type of the population. For a structured pedigree, they are expressed in terms of the recombination fractions, whereas for a natural population, they can be expressed in terms of linkage disequilibria. The derivations of the conditional probabilities of QTL genotypes are given in general QTL mapping literature (Wu and Casella, 2005).

To estimate the parameters of the likelihood function implemented with growth data measured at multiple time points, one can extend the traditional interval mapping approach to accommodate the multivariate nature of time-dependent traits. However, this extension is limited in three aspects: (1) Individual expected means of different QTL genotypes at all points and all elements in the matrix  $\Sigma$  need to be estimated, resulting in substantial computational difficulties when the vector and matrix dimensions are large. (2) The result from this approach may not be biologically meaningful because the underlying biological principle for growth is not incorporated. (3) This approach cannot be well deployed in a practical scheme because of (2). Thus, some biologically interesting questions cannot be asked and answered.

A new statistical framework has been developed to detect QTL affecting growth trajectories (Wu et al., 2002, 2004a,b; Ma et al., 2002). This framework included two tasks:

(1) Model the time-dependent expected means of QTL genotype  $j$  using a growth equation,

$$\mu_j(t) = g(t; \Omega_{m_j}), \quad (4)$$

which is specified by a set of curve parameters arrayed in  $\Omega_{m_j}$ . All living entities are characterized by growth defined as the irreversible increase of size with time. A number of mathematical models, such as logistic or S-shaped curves, have been proposed to describe growth trajectories. The fundamental biological aspects of mathematical modelling of growth curves have been founded by earlier mathematical biologists, e.g. von Bertalanffy et al. (1957), and recently revisited by West et al. (2001).

The overall form of the growth curve of QTL genotype  $j$  is determined by the set of curve parameters contained in  $\Omega_{m_j}$ . If different genotypes at a putative QTL have different combinations of these parameters, this implies that this QTL plays a role in governing the differentiation of growth trajectories. Thus, by testing for the difference of  $\Omega_{m_j}$  among different genotypes, we can determine whether there exists a specific QTL that confers an effect on growth curves.

The time-dependent growth curves for different genotypes,  $\mu_j(t)$ , can be used to estimate the dynamic changes of various genetic effects, including the additive, dominant and epistatic effects as well as the interaction between these effects and environmental factors. For example, an  $F_2$  population containing three genotypes at a QTL, coded as 2 for  $QQ$ , 1 for  $Qq$  and 0 for  $qq$ , allows for the estimates

of the time-dependent additive effect,  $a(t) = \frac{1}{2}[\mu_2(t) - \mu_0(t)]$ , and dominant effects,  $d(t) = \mu_1(t) - \frac{1}{2}[\mu_2(t) + \mu_0(t)]$ , during growth trajectories. The time-dependent epistatic and genotype  $\times$  environment effects can be characterized in a similar way if a multi-QTL model is assumed or if an experimental design with multiple environments is used.

(2) Model the structure of the within-subject (co)variance matrix using the first-order autoregressive [AR(1)] model (Diggle et al., 2002), expressed as

$$\sigma^2(1) = \dots = \sigma^2(\tau) = \sigma^2$$

for the variance, and

$$\sigma(t_1, t_2) = \sigma^2 \rho^{|t_2 - t_1|}$$

for the covariance between any two time intervals  $t_1$  and  $t_2$ , where  $0 < \rho < 1$  is the proportion parameter with which the correlation decays with time lag. The parameters that model the structure of the (co)variance matrix are arrayed in  $\Omega_v$ .

We implemented the expectation-maximization (EM) algorithm, originally proposed by Dempster et al. (1977), to obtain the maximum likelihood estimates (MLEs) of three groups of unknown parameters in a QTL mapping model, that is, the QTL-segregating parameters ( $\Omega_\ell$ ) that specify the co-segregation patterns of QTL and markers in the population, the curve parameters ( $\Omega_{m_j}$ ) that model the mean vector and the parameters ( $\Omega_v$ ) that model the structure of the (co)variance matrix (Ma et al., 2002; Wu et al., 2004a,b). These unknowns, denoted by  $\Omega = (\Omega_\ell, \Omega_{m_j}, \Omega_v)$ , are contained within the mixture model described by Equation (1). A detailed description of the EM algorithm was given in Wu et al. (2002) and Ma et al. (2002).

After the MLEs of the parameters are obtained, the existence of a QTL affecting an overall growth curve should be tested by formulating two alternative hypotheses,

$$\begin{cases} H_0: \Omega_{m_j} \equiv \Omega_m \\ H_1: \text{at least one of the equalities above does not hold,} \end{cases} \quad (5)$$

where  $H_0$  corresponds to the reduced model, in which the data can be fit by a single growth curve, and  $H_1$  corresponds to the full model, in which there exist different growth curves to fit the data. The test statistic for testing the hypotheses in Equation (5) is calculated as the log-likelihood (LR) ratio of the reduced to the full model:

$$\text{LR} = -2 [\log L(\tilde{\Omega}|\mathbf{y}) - \log L(\hat{\Omega}|\mathbf{y})], \quad (6)$$

where  $\tilde{\Omega}$  and  $\hat{\Omega}$  denote the MLEs of the unknown parameters under  $H_0$  and  $H_1$ , respectively. After the existence of QTL is tested, a number of biologically meaningful hypotheses regarding the interplay between gene action and development can be formulated (Wu et al., 2004a).

To remove the heteroscedastic problem of the residual variance, which violates a basic assumption of the simple AR(1) model, two approaches can be used. The first approach is to model the residual variance by a parametric function of time, as originally proposed by Pletcher and Geyer (1999). But this approach needs to implement additional parameters for characterizing the age-dependent change of the variance. The second approach is to embed Carroll Rupert's (1984) transform-both-sides (TBS) model into the growth-incorporated finite mixture model (Wu et al., 2004b), which does

not need any more parameters. Both empirical analyses with real examples and computer simulations suggest that the TBS-based model can increase the precision of parameter estimation and computational efficiency. Furthermore, the TBS model preserves original biological means of the curve parameters although statistical analyses are based on transformed data.

The TBS-based model displays the potential to relax the assumption of variance stationarity, but the covariance stationarity issue remains unsolved. Zimmerman and Núñez-Antón (1997) proposed a so-called structured antedependence (SAD) model to model the age-specific change of correlation in the analysis of longitudinal traits. The SAD model has been employed in several studies and displays many favorable properties (Zimmerman and Núñez-Antón, 2001).

In this article, we will incorporate the SAD model within the mixture model for functional mapping to take advantages of it in adequately modeling the (co)variance structure. In Section 2, we describe the statistical model for deriving the SAD model. Section 3 provides a discussion of model selection for different orders. The method is illustrated in Section 4 using growth data of body mass measured at 10 different time points in an F<sub>2</sub> progeny derived from the Large and Small mouse strains. The implications and extensions of the model are discussed in Section 5.

## 2 THE STRUCTURED ANTEDEPENDENCE MODEL

The antedependence model was originally proposed by Gabriel (1962). It states that an observation at a particular time  $t$  depends on the previous ones, with the degree of dependence decaying with time lag. If an observation at time  $t$  is independent of all observations before  $t - r$ , this antedependence model is thought to be of order  $r$ . The antedependence model is extended to fit the structure of time-dependent variance and correlation, leading to the SAD model (Núñez-Antón and Zimmerman, 2000). The SAD model can be incorporated to the QTL mapping of longitudinal growth traits.

Let us consider an F<sub>2</sub> design of  $n$  progeny derived from two contrasting homozygous inbred lines, in which there are three QTL genotypes ( $J = 3$ ). For this F<sub>2</sub>, a genetic linkage map is constructed with molecular markers and a growth trait ( $y$ ) is measured for a finite set of times,  $1, \dots, \tau$ .  $y(1), \dots, y(\tau)$  are  $r$ th-order antedependent if the conditional distribution of  $y(t)$ , given  $y(t-1), \dots, y(1)$ , depends on  $y(t-1), \dots, y(t-r)$ , for all  $t \geq r$  (Gabriel, 1962). This concept is equivalent to  $y(1), \dots, y(\tau)$  having a Markovian dependence of order  $r$ . The order  $r$  serves as a memory gauge, where  $r = 0$  corresponds to independence and  $r = \tau - 1$  to arbitrary multivariate dependence. A parametrically specified definition for individual  $i$  is given on the basis of Equation (3), which can be expressed as

$$y_i(t) = \sum_{j=1}^3 \xi_{ij} \mu_j(t) + \sum_{t'=1}^{r^*} \sum_{j=1}^3 \phi_{i,t-t'} [y_i(t-t') - \xi_{ij} \mu_j(t-t')] + e_i(t), \quad t = 1, \dots, \tau, \quad (7)$$

where  $r^* = \min(r, t - 1)$ ,  $\phi_{i,t-t'}$ s are unrestricted antedependence parameters, and independent normal random variable  $e_i(t)$  may have time-dependent variances,  $v^2(t)$ , termed innovation variances.

Like AR models, this model allows for serial correlation within subjects, but unlike AR models, it does not assume that the variances are constant nor that correlations between measurements equidistant in time are equal. The antedependence model (7) is called the unstructured antedependence model of order  $r$  (UAD( $r$ )) because  $(r + 1)(2\tau - r)/2$  parameters, including the variances  $\sigma^2(t)$  and covariances  $\sigma(t, t - t')$ , among measurements at  $\tau$  different time points are not expressed as a function of a smaller set of parameters (Núñez-Antón and Zimmerman, 2000).

To make the UAD( $r$ ) model more parsimonious, Núñez-Antón (1997) and Núñez-Antón and Zimmerman (2000) proposed the so-called SAD models. One useful class models the autoregressive coefficients with the Box-Cox power law and models the innovation variances with a parametric function, i.e.

$$\begin{aligned} \phi_{i,t-t'} &= \phi_{t'}^{w(T_i;\lambda_{t'}) - w(T_{t-t'};\lambda_{t'})}, \quad t = r + 1, \dots, \tau; \\ t' &= 1, \dots, r, \quad \phi_{t'} > 0, \\ v^2(t) &= v^2 v(T_i; \psi), \quad \sigma^2 > 0, \quad \{\psi: v(T_i; \psi) > 0\}, \end{aligned} \quad (8)$$

where  $T_i$  and  $T_{t-t'}$  are measurement times,  $w(T; \lambda)$  equals  $(T^\lambda - 1)/\lambda$  if  $\lambda \neq 0$  and equals  $\log T$  if  $\lambda = 0$ , and  $v(\cdot)$  is a function of relatively few parameters (e.g. a low-order polynomial). Thus, different from Gabriel's (1962) original treatment, we only need to estimate three parameters to model the innovation variances if a quadratic polynomial is used, regardless of the number of time points.

As a simplified example with the SAD(1) model in which innovation variances are constant over time points, Jaffrézic *et al.* (2003) derived the analytical forms for variance and covariance functions among time-dependent measurements, expressed, respectively, as

$$\sigma^2(t) = \frac{1 - \phi^{2t}}{1 - \phi^2} v^2, \quad (9)$$

$$\sigma(t_1, t_2) = \phi^{t_2 - t_1} \frac{1 - \phi^{2t_1}}{1 - \phi^2} v^2, \quad t_2 \geq t_1, \quad (10)$$

for equally spaced repeated measurements. It can be seen that although constant innovation variances are assumed, the residual variance can change with time (Jaffrézic *et al.*, 2003). Also, for the simplest SAD model, the correlation function is non-stationary because the correlation does not depend only on the time interval  $t_2 - t_1$  but also depends on the start and end points of the interval  $t_1$  and  $t_2$ .

## 3 MODEL SELECTION

Jaffrézic *et al.* (2003) proposed an *ad hoc* approach for model selection. Their strategy is to increase the antedependence order until the additional antedependence coefficient is close to zero.

Núñez-Antón and Zimmerman (2000) proposed using the AIC information criterion to select the best model. Hurvich and Tsai (1989) showed that AIC can drastically underestimate the expected Kullback-Leibler information when only few repeated measurements are available. Instead, they derived a corrected AIC, expressed as

$$\text{AIC}_C = \tau \log \hat{v}^2 + \tau \frac{1 + r/\tau}{1 - (r + 2)/\tau} \quad (11)$$

where  $\hat{v}^2$  is the white noise variance,  $\tau$  is the number of repeated measurements and  $r$  is the order of the model. The number of

parameters is heavily penalized so that models selected by  $AIC_C$  are typically much more parsimonious than those selected by AIC (Hurvich and Tsai, 1989). An alternative criterion, BIC, that selects a model with a maximum posterior probability (Schwarz, 1978) can also be used to determine the best antedependence order.

## 4 APPLICATION

### 4.1 Mouse data

We used the joint statistical model to map sex-specific QTL that affect growth trajectories in an animal model system—mouse. Vaughn *et al.* (1999) constructed a linkage map with 96 microsatellite markers for 502  $F_2$  mice (259 males and 243 females) derived from two strains, the Large (LG/J) and Small (SM/J). This map has a total map distance of  $\sim 1780$  cM (in Haldane's units) and an average interval length of  $\sim 23$  cM. The  $F_2$  progeny were measured for their body mass at 10 weekly intervals starting at age 7 days. The raw weights were corrected for the effects of each covariate due to dam, litter size at birth and parity, but not for the effect due to sex (Vaughn *et al.*, 1999).

### 4.2 The partitioning of sex-specific genotypic values

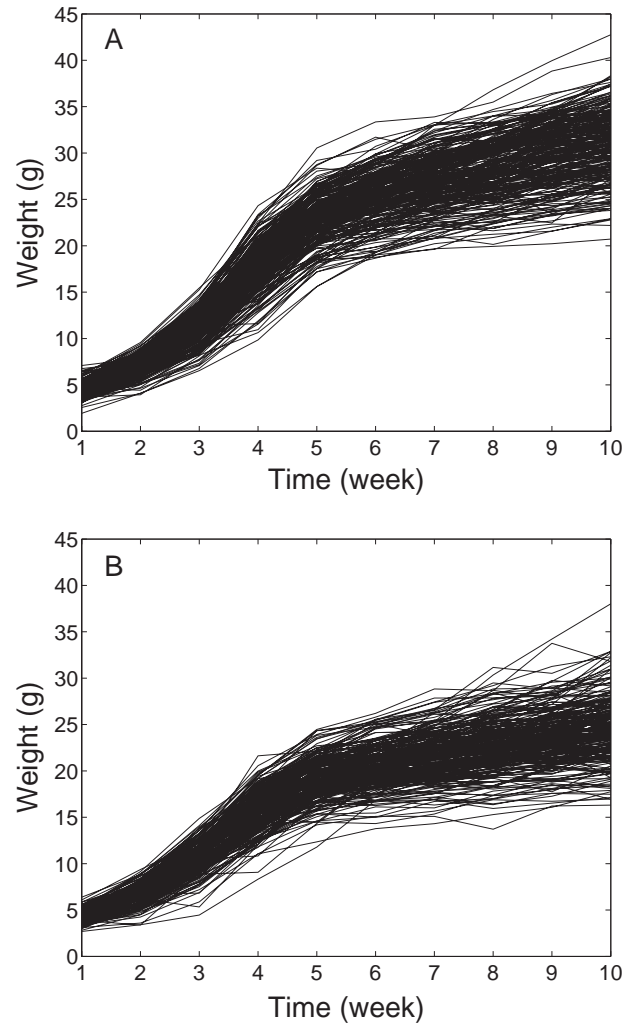
The  $F_2$  population used here displayed a marked difference between the two sexes, with the male mice growing faster than female mice (Fig. 1). Our model allows for the tests of the genetic, sex and their interaction effects on growth trajectories. The statistical model for specifying the growth phenotype for individual  $i$  that contains the sex effect is expressed as

$$y_{ik}(t) = \sum_{j=1}^3 \xi_{ijk} \mu_{jk}(t) + \sum_{t'=1}^{t-1} \phi_{i,t-t'} [y_{ik}(t-t') - x_i \mu_{jk}(t-t')] + e_{ik}(t),$$

where  $y_{ik}(t)$  is the time-dependent growth for individual  $i$  that has sex  $k$  ( $k = 1$  for the male and 0 for the female),  $\xi_{ijk}$  is the indicator variable representing QTL genotype  $j$  for individual  $i$  with sex  $k$ ,  $\mu_{jk}(t)$  is the time-dependent genotypic value for QTL genotype  $j$  having sex  $k$ , and  $e_{ik}(t)$  is the time-dependent sex-specific residual error.

Based on quantitative genetic theory, we partition time-dependent sex-specific genotypic value,  $\mu_{jk}(t)$ , into the additive and dominant effects due to the QTL and the interaction effects between these two genetic effects and sex. Let  $a(t)$  and  $d(t)$  be the time-dependent additive and dominant effects of the QTL,  $s(t)$  be the time-dependent sex effect and  $I_{as}(t)$  and  $I_{ds}(t)$  be the time-dependent additive  $\times$  sex and dominant  $\times$  sex interaction effects, respectively. We tabulate time-dependent genotypic value  $\mu_{jk}(t)$  for each QTL genotype  $j$  ( $j = 2, 1, 0$ ) with sex  $k$  ( $k = 1, 2$ ) in terms of its genetic compositions as follows:

QTL genotype	Sex Male	Female
$QQ$	$\mu_{21}(t) = \mu(t) + a(t) + \frac{1}{2}s(t) + \frac{1}{2}I_{as}(t)$	$\mu_{22}(t) = \mu(t) + a(t) - \frac{1}{2}s(t) - \frac{1}{2}I_{as}(t)$
$Qq$	$\mu_{11}(t) = \mu(t) + d(t) + \frac{1}{2}s(t) + \frac{1}{2}I_{ds}(t)$	$\mu_{12}(t) = \mu(t) + d(t) - \frac{1}{2}s(t) - \frac{1}{2}I_{ds}(t)$
$qq$	$\mu_{01}(t) = \mu(t) - a(t) + \frac{1}{2}s(t) - \frac{1}{2}I_{as}(t)$	$\mu_{02}(t) = \mu(t) - a(t) - \frac{1}{2}s(t) + \frac{1}{2}I_{as}(t)$



**Fig. 1.** Plots of body mass versus ages for 259 male (A) and 243 female (B) mice in an  $F_2$  progeny derived from LG/J and SM/J strains (Vaughn *et al.*, 1999). To display sex-specific differences, body mass was not corrected for the sex effect.

Here  $\mu(t)$  is the time-dependent overall mean. The dynamic changes of these different effects can be derived, which are expressed as

$$s(t) = \frac{1}{2}[\mu_{21}(t) + \mu_{01}(t) - \mu_{22}(t) - \mu_{02}(t)], \quad (12)$$

for the sex effect;

$$a(t) = \frac{1}{4}[\mu_{21}(t) + \mu_{22}(t) - \mu_{01}(t) - \mu_{02}(t)], \quad (13)$$

for the additive genetic effect of the QTL;

$$d(t) = \frac{1}{4}[2\mu_{11}(t) + 2\mu_{12}(t) - \mu_{21}(t) - \mu_{01}(t) - \mu_{22}(t) - \mu_{02}(t)], \quad (14)$$

for the dominant genetic effect of the QTL;

$$I_{as}(t) = \frac{1}{2}[\mu_{21}(t) + \mu_{02}(t) - \mu_{01}(t) - \mu_{22}(t)], \quad (15)$$

for the additive  $\times$  sex interaction effect; and

$$I_{ds}(t) = \frac{1}{2}[2\mu_{11}(t) - 2\mu_{12}(t) - \mu_{21}(t) - \mu_{01}(t) + \mu_{22}(t) + \mu_{02}(t)], \quad (16)$$

for the dominant  $\times$  sex interaction effect. All these effects can be estimated and tested.

### 4.3 The growth law

The sigmoidal (or logistic) growth function is regarded as being nearly universal in living systems to capture age-specific change in growth (West *et al.*, 2001). The logistic growth curve as a biological law can be mathematically described by

$$g(t) = \frac{\alpha}{1 + \beta e^{-\gamma t}} \quad (17)$$

where  $\alpha$  is the asymptotic or limiting value of  $g$  when  $t \rightarrow \infty$ ,  $\alpha/(1 + \beta)$  is the initial value of  $g$  when  $t = 0$  and  $\gamma$  is the relative rate of growth (von Bertalanffy *et al.*, 1957). If different genotypes at a putative QTL have different combinations of these parameters, this implies that this QTL plays a role in governing the difference of growth trajectories.

By plotting body mass growth against time (Fig. 1), it is observed that each of the mapped  $F_2$  mice follows the S-shaped (logistic) growth curve. A non-linear least squares approach was used to fit the growth trajectory of body mass with the logistic curve of Equation (17) for each mouse. Based on statistical tests, all the mice can be well fit by a logistic curve ( $r^2 > 0.95$ ). Therefore, we use Equation (17) to fit the mean vector for QTL genotype  $j$  with sex  $k$ , with  $\Omega_{m_{jk}} = (\alpha_{jk}, \beta_{jk}, \gamma_{jk})$ , in our mapping model.

### 4.4 Computational algorithm

We adopt our previous algorithm developed on the basis of Equation (1) (Ma *et al.*, 2002; Wu *et al.*, 2004a,b) to estimate the QTL position ( $\Omega_\ell$ ), the curve parameters modelling the mean vector ( $\Omega_{m_{jk}}$ ) and the parameters modeling the (co)variance matrix ( $\Omega_v$ ). In practical computations, the QTL position parameter can be viewed as a fixed parameter because a putative QTL can be searched at every 1 or 2 cM on a map interval bracketed by two markers throughout the entire linkage map. The amount of support for a QTL at a particular map position is often displayed graphically through the use of likelihood maps or profiles, which plot the likelihood ratio test statistic as a function of the map position of the putative QTL.

The Nelder–Mead simplex algorithm, originally proposed by Nelder and Mead (1965), can be used to estimate  $\Omega_{m_{jk}}$  and  $\Omega_v$  contained in Equations (1) and (2) (Zhao *et al.*, 2004b). It is a direct search method for non-linear unconstrained optimization. It attempts to minimize a scalar-valued non-linear function using only function values, without any derivative information (explicit or implicit). The algorithm uses linear adjustment of the parameters until some convergence criterion is met. The term ‘simplex’ arises because the feasible solutions for the parameters may be represented by a polytope figure called a ‘simplex’. The simplex is a line in one dimension, triangle in two dimensions and tetrahedron in three dimensions, respectively. To increase the computation efficiency, we derived the closed forms of the determinant and inverse of the residual variance matrix [in Equation (2)] fitted by the SAD(1) model and incorporated these forms to estimate the mixture model (1).

After the point estimates of parameters are obtained by the EM algorithm, we derive the approximate variance–covariance matrix and evaluate the sampling errors of the estimates ( $\widehat{\Omega}_\ell, \widehat{\Omega}_{m_{jk}}, \widehat{\Omega}_v$ ). The techniques for so doing involve the calculation of the incomplete-data information matrix which is approximated by the negative second-order derivative of the incomplete-data log-likelihood. The incomplete-data information can be calculated by extracting the information for the missing data from the information for the complete data (Louis, 1982). A different so-called supplemented EM algorithm or SEM algorithm was proposed by Meng and Rubin (1991) to estimate the asymptotic variance–covariance matrices, which can also be used to calculate approximate sampling errors for the MLEs of the parameters ( $\Omega_\ell, \Omega_{m_{jk}}, \Omega_v$ ) in our mixture model setting.

### 4.5 Results

To determine the best order of the SAD models for the detection of QTL hidden in this mouse dataset, we should estimate the residual variances under the full model of hypothesis (5) and further calculate the AIC<sub>C</sub> values using Equation (11) for different orders. However, this would be computationally expensive because at each antedependence order, we need to estimate numerous parameters. Here, we instead based our order determination on the reduced model of hypothesis (5) in which there is only one mean curve that explains the  $F_2$  growth data. In fact, from a small simulation study, the result about order determination was found to be consistent based on the full and reduced models (data not shown).

To simplify the statistical analysis, we assume that antedependence coefficients between measurements equidistant in time are equal. Table 1 lists the AIC<sub>C</sub> values from models SAD(1) to SAD(6), which are found to increase with order. Also, for the SAD model of higher orders, the first antedependence coefficient is markedly higher than the rest coefficients. Thus, we think that the SAD(1) model should be reasonable to fit the  $F_2$  mouse data in this example. Along with the constant innovation variance, this model was incorporated to approximate the structure of the (co)variance matrix for growth trajectories in the  $F_2$  mouse progeny.

The profile of the LR of the full versus reduced model across the entire genome estimated from the SAD-based model has three clear peaks on chromosomes 6, 7 and 10 (Fig. 2). These peaks correspond to the locations of the detected QTL, with the LR values 64, 60 and 40, respectively, well beyond the genomewide critical threshold, 32, at the significance level  $P = 0.05$ . The critical value for claiming the existence of QTL can be determined on the basis of the Bonferroni argument for the sparse-map case (Lander and Botstein, 1989) or by permutation tests proposed by Churchill and Doerge (1994). In this example, the empirical estimate of the critical value is obtained from 100 permutation tests.

This SAD-based model provided estimates of the curve parameters as well as the innovation variance and antedependent coefficient that model the structure of the (co)variance matrix for each sex of the  $F_2$  progeny (Table 2). Estimated small sampling errors suggest that our model provides precise estimates of all the model parameters. Figure 3 illustrates the growth curves of three genotypes at each of the detected QTL separately for two different sexes drawn from the estimates of curve parameters in Table 2. Our model can test the additive and dominant effect of the QTL as well as their interaction effects with sex. All these age-dependent

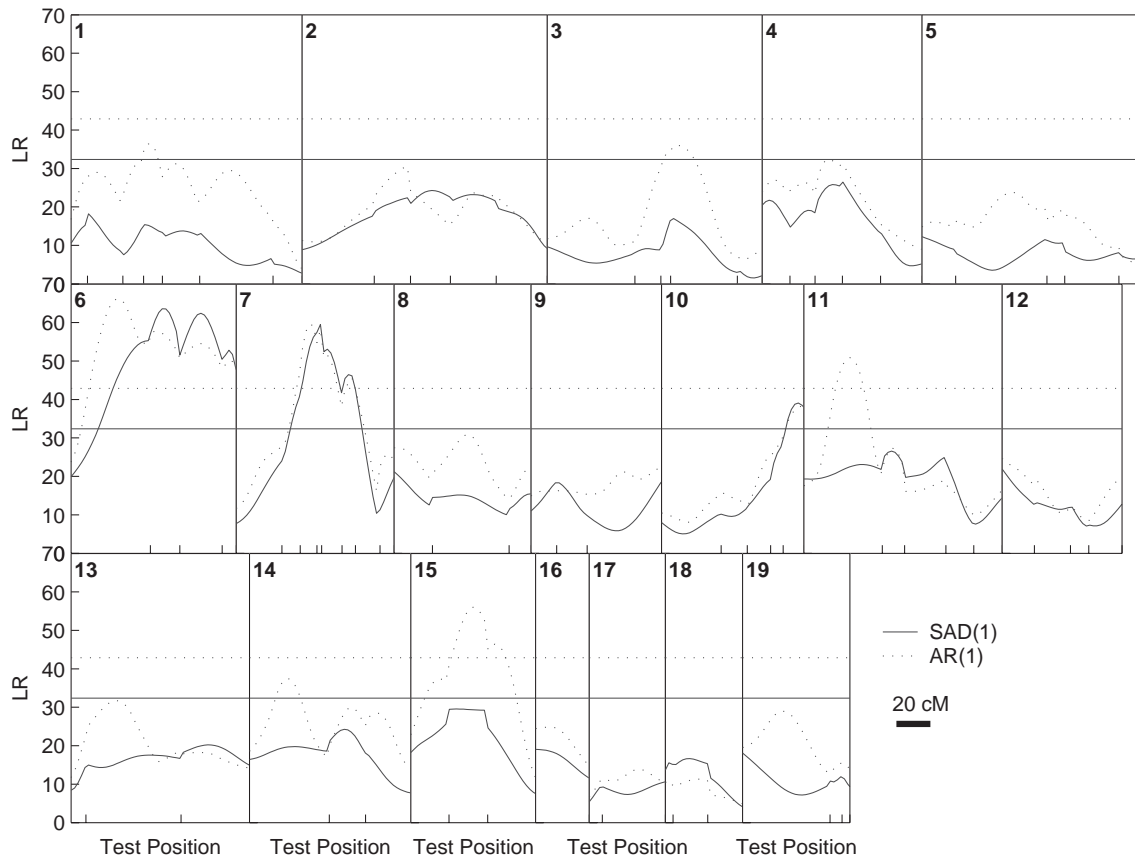
**Table 1.**  $AIC_C$  information criteria for female mice data<sup>a</sup>

Structure	$AIC_C$	$\hat{\nu}^{2b}$	$\hat{\phi}_1^c$	$\hat{\phi}_2^c$	$\hat{\phi}_3^c$	$\hat{\phi}_4^c$	$\hat{\phi}_5^c$	$\hat{\phi}_6^c$
SAD(1)	18.98	1.386	0.987					
SAD(2)	23.26	1.386	1.003	-0.021				
SAD(3)	29.15	1.370	1.007	-0.126	0.149			
SAD(4)	38.10	1.363	0.997	-0.118	0.065	0.123		
SAD(5)	53.10	1.363	0.996	-0.120	0.067	0.116	0.012	
SAD(6)	83.06	1.359	0.996	-0.115	0.070	0.101	0.085	-0.140

<sup>a</sup>The best model is the one with the minimum corrected AIC value.

<sup>b</sup> $\hat{\nu}^2$  is the estimated white noise variance.

<sup>c</sup> $\hat{\phi}_s$  are the estimated coefficients of the SAD models.



**Fig. 2.** The profiles of the LR between the full and reduced (no QTL) model estimated from the SAD(1) (solid) and AR(1) models (dot) for body mass growth trajectories across the entire genome using the linkage map constructed from microsatellite markers (Vaughn *et al.*, 1999). The genomic positions corresponding to the peaks of the curves are the MLEs of the QTL positions. The genome-wide threshold values for claiming the existence of QTL are given as the horizontal lines. Tick marks on the x-axis represent the positions of markers on the linkage group, whose names are given in Vaughn *et al.* (1999).

changes of genetic effects estimated by Equations (12)–(16) are illustrated in Figure 4. The sex effect,  $s(t)$ , increases rapidly with age, but the other effects appear to change with age at a much lesser extent. As expected, the LG/J strain contributes the body-increasing allele to the  $F_2$  progeny at all the three detected QTL. The additive effects,  $a(t)$ , at these QTL increase with age (Fig. 4). The dominant effects of the QTL,  $d(t)$  and its interaction effects with the sex,  $I_{as}(t)$  and  $I_{ds}(t)$ , appear to change with age at lesser extents.

Also, they display different dynamic patterns, depending on the QTL detected.

## 5 METHODOLOGICAL COMPARISONS: A SIMULATION

The statistical behavior of our new model described in this article can be well investigated by comparing it with the AR(1)-based functional

**Table 2.** The MLEs of the model parameters and their asymptotic standard errors in the parentheses under the SAD(1) model for different sexes of an F<sub>2</sub> mouse progeny (composed of 259 males and 243 females)<sup>a</sup>

Chr. no.	Location (cM)	Sex	QQ			Qq			qq			ν <sup>2</sup>	ϕ
			α <sub>2</sub>	β <sub>2</sub>	γ <sub>2</sub>	α <sub>1</sub>	β <sub>1</sub>	γ <sub>1</sub>	α <sub>0</sub>	β <sub>0</sub>	γ <sub>0</sub>		
6	52	Female	25.33 (0.46)	5.03 (0.20)	0.73 (0.020)	23.78 (0.33)	4.62 (0.14)	0.72 (0.013)	22.55 (0.41)	4.48 (0.18)	0.74 (0.016)	1.36 (0.040)	0.97 (0.013)
		Male	31.53 (0.48)	6.27 (0.26)	0.74 (0.017)	31.25 (0.33)	6.11 (0.16)	0.71 (0.010)	28.36 (0.45)	5.67 (0.22)	0.70 (0.015)	1.63 (0.046)	0.96 (0.012)
7	48	Female	22.21 (0.42)	4.48 (0.19)	0.76 (0.020)	23.82 (0.31)	4.62 (0.13)	0.73 (0.013)	25.30 (0.40)	5.03 (0.18)	0.71 (0.015)	1.33 (0.040)	0.98 (0.013)
		Male	31.92 (0.50)	6.33 (0.25)	0.71 (0.015)	30.72 (0.34)	6.01 (0.17)	0.72 (0.011)	28.75 (0.45)	6.04 (0.24)	0.73 (0.015)	1.65 (0.049)	0.97 (0.012)
10	78	Female	24.78 (0.42)	5.02 (0.20)	0.74 (0.018)	24.44 (0.30)	4.74 (0.13)	0.72 (0.012)	22.39 (0.39)	4.36 (0.17)	0.73 (0.018)	1.37 (0.040)	0.96 (0.012)
		Male	31.47 (0.43)	6.28 (0.22)	0.72 (0.014)	30.11 (0.36)	5.93 (0.17)	0.71 (0.011)	30.00 (0.46)	6.09 (0.23)	0.72 (0.015)	1.67 (0.047)	0.97 (0.011)

<sup>a</sup>The location of QTL is described in the distance (cM) from the first marker on a chromosome.

**Table 3.** Three different patterns of the covariance matrix used to simulate the multivariate phenotypic data

Pattern	Matrix structure		Given value	
	Variance σ <sup>2</sup> (t)	Covariance σ(t <sub>1</sub> , t <sub>2</sub> )	ν <sup>2</sup>	ϕ or ρ
A	ν <sup>2</sup>	ρ <sup> t<sub>1</sub>-t<sub>2</sub>  ν<sup>2</sup></sup>	0.015	0.8
B	$\frac{1-\phi^{2t}}{1-\phi^2}\nu^2$	$\phi^{t_2-t_1}\frac{1-\phi^{2t_1}}{1-\phi^2}\nu^2, t_2 > t_1$	0.5	1.12
C	t <sup>(3/2)</sup> ν <sup>2</sup>	ρ <sup>(1/2) t<sub>1</sub>-t<sub>2</sub>  <math>\sqrt{t_1^{(3/2)}t_2^{(3/2)}}\nu^2</math></sup>	0.5	0.8

mapping model. The statistical properties of the AR(1)-based model have been studied through simulation and empirically (Ma *et al.*, 2002; Wu *et al.*, 2004a,b). Here, we perform a series of simulation studies to compare these two models.

Consider an F<sub>2</sub> population with which a 40-cM long linkage group composed of three equidistant markers is constructed. A QTL that affects growth curves is located at 10 cM from the first marker on the linkage group. Assume that there are 500 F<sub>2</sub> progeny, each measured at 10 equally spaced time points. Our simulation was based on four different patterns of covariance structures (Table 3). Patterns A and B conform to the AR(1) and SAD(1) model, respectively, whereas Pattern C does not follow either of these two models. The phenotypic and marker data simulated under each of these patterns were analyzed by both the AR(1)- and SAD(1)-based models, with the MLEs of unknown parameters (Ω<sub>t</sub>, Ω<sub>m<sub>jk</sub></sub>, Ω<sub>v</sub>) and their sampling errors based on 100 simulation replicates given in Table 4.

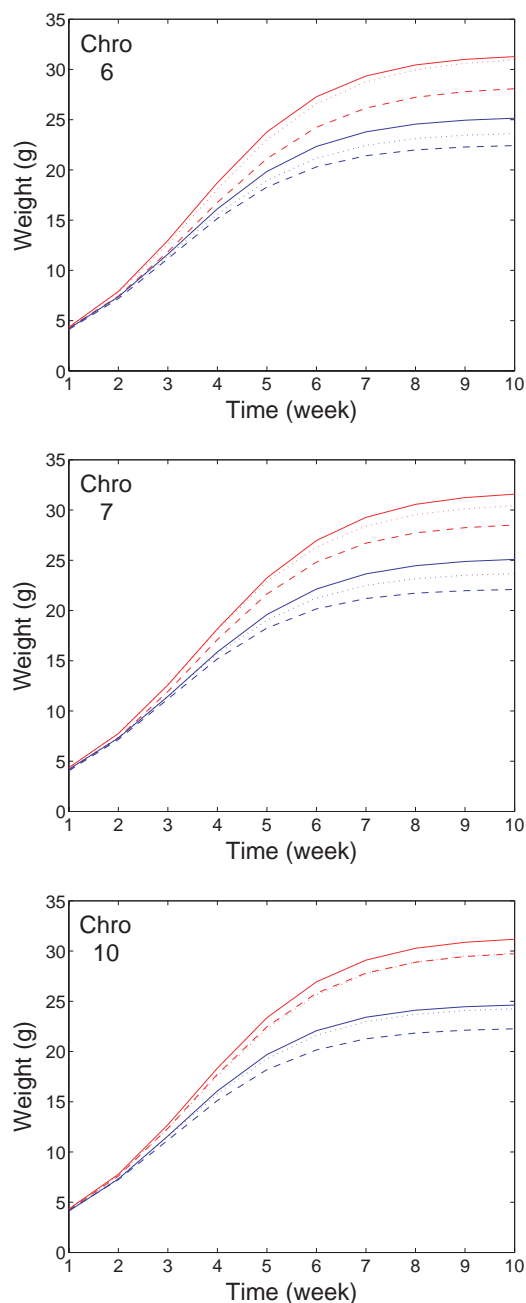
The QTL location, growth curve parameters and covariance-structuring parameters can be equally precisely estimated for the data simulated under the AR(1) model by the SAD(1) and AR(1) models (Table 4). However, if the data structure follows the SAD(1) model, the estimation precision of the parameters is much better from the SAD(1) than from the AR(1) model. These comparisons suggest that the SAD(1)-based functional mapping model is statistically more robust than the AR(1) model; i.e. the results from the SAD(1) model are less data-dependent than the AR(1) model. It is interesting to note that both the models display similar estimation

precisions for the data that follow either the SAD(1) model or the AR(1) model. This implies that for a practical dataset, whose covariance structure is unknown, both the models should be tested and that the growth QTL detected from the two models, if they are different, should be considered to be reasonable.

## 6 DISCUSSION

We have developed a new statistical model for functional mapping of QTL that affect growth curves. Functional mapping based on a longitudinal model has proven to be more powerful for QTL detection compared to a simple univariate analysis (Ma *et al.*, 2002). In this article, functional mapping is incorporated by the SAD model (Núñez-Antón and Zimmerman, 2000; Pourahmadi, 1999) that specifies the non-stationarity of the (co)variances for longitudinal traits. It is intriguing to compare this SAD-based model with our previous AR(1)-based model under the assumptions of stationary variance and covariance (Wu *et al.*, 2002; Ma *et al.*, 2002).

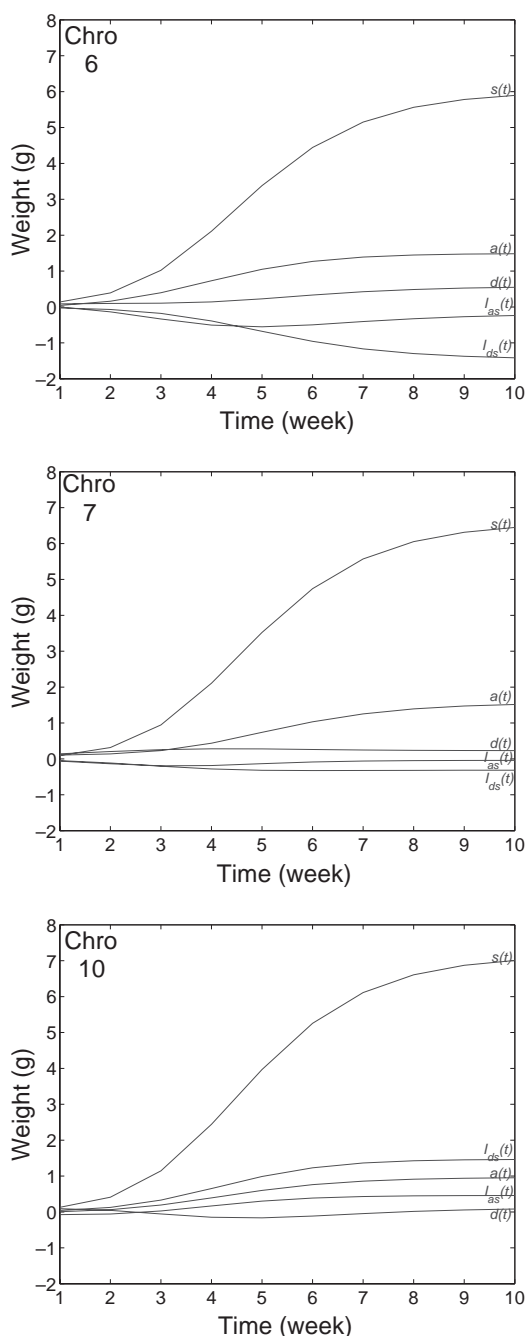
Zhao *et al.* (2004a) reported the results of QTL mapping for growth trajectories in the same F<sub>2</sub> mouse progeny as used in this study based on the functional mapping model integrated by the AR(1) structure. To compare the results from the two models, we use dot curves to draw the LR profile for Zhao *et al.*'s AR(1) model (Fig. 2). The AR(1) model discovered four QTL, as opposed to three detected by the SAD(1) model. Both the models identified two QTL on chromosomes 6 and 7, although the estimated locations of the QTL on chromosome 6 are different in the two models. The third QTL detected by the SAD(1) model is located on chromosome 10, whereas the other two QTL detected by the AR(1) model are located on chromosomes 11 and 15. Such pronounced discrepancies may be due to the difference in the nature of QTL control. For instance, if a time-increasing total variance is explained by a time-increasing differentiation triggered by a QTL, with the residual variance constant across time, then the AR model would work better than the SAD model. On the other hand, if the effects of QTL considered are not the sole source contributing to increased overall variances with time, the SAD model has more power to detect such QTL. We have used simulation studies to explain their differences. Although the SAD model is more robust than the AR model, they should can be



**Fig. 3.** Three growth curves each presenting a group of genotype,  $QQ$  (solid curves),  $Qq$  (dot curves) and  $qq$  (broken curves), in the male (red) and female (blue) mice at the QTL, detected by our SAD-based model, on chromosomes 6, 7 and 10.

viewed as alternative to one another. In practice, these two models should be used simultaneously to monitor the existence of growth QTL for the same dataset.

One of the major advantages of our model is that it can discern the changes in genetic actions and interactions of QTL for complex traits during ontogeny and has the potential to integrate developmental biology into quantitative genetics theory and methodology. Used in an  $F_2$  mouse population, our model has discovered the age-related



**Fig. 4.** Dynamic changes of the sex effect,  $s(t)$ , the additive,  $a(t)$ , and dominant effect,  $d(t)$ , of the QTL and the interaction effects,  $I_{as}$  and  $I_{ds}$ , between the QTL and sex, detected by our SAD-based model, on chromosomes 6, 7 and 10.

change of genetic influence on body size during development. Similar findings in mice have been obtained by Atchley and Zhu (1997). Our model can be generalized to a dynamic mixture model to handle longitudinal data measured at uneven time intervals and with different measurement patterns among different individuals. It can be anticipated that the model proposed in this article and its extensions will handle various complexities of longitudinal data. It will have great

**Table 4.** Reciprocal comparisons between the MLEs and sampling errors (in parentheses) of the parameters ( $\Omega_\ell$ ,  $\Omega_{m_{jk}}$ ,  $\Omega_v$ ) from the SAD(1)- and AR(1)-based model

Analytical model	Location	$QQ$			$Qq$			$qq$			Residuals	
		$\alpha_2 = 26$	$\beta_2 = 6.04$	$\gamma_2 = 0.5$	$\alpha_1 = 28$	$\beta_1 = 6.04$	$\gamma_1 = 0.5$	$\alpha_0 = 30$	$\beta_0 = 6.04$	$\gamma_0 = 0.5$	$v^2$	$\phi$ or $\rho$
(A) Data simulated by the AR(1) model												
SAD(1)	10.22 (3.2)	26.21 (0.47)	5.98 (0.13)	0.50 (0.009)	28.11 (0.37)	6.02 (0.09)	0.50 (0.007)	30.47 (0.58)	6.14 (0.14)	0.50 (0.01)	1.82 (0.053)	0.89 (0.012)
AR(1)	10.04 (2.2)	26.09 (0.41)	6.06 (0.13)	0.50 (0.007)	27.97 (0.29)	6.03 (0.08)	0.50 (0.004)	29.98 (0.43)	6.03 (0.12)	0.50 (0.006)	0.015 (0.0006)	0.8 (0.009)
(B) Data simulated by the SAD(1) model												
SAD(1)	10.42 (2.90)	25.98 (0.46)	6.03 (0.10)	0.50 (0.005)	27.98 (0.29)	6.05 (0.07)	0.50 (0.004)	29.90 (0.39)	6.04 (0.087)	0.50 (0.004)	0.50 (0.011)	1.12 (0.004)
AR(1)	12.78 (7.20)	25.58 (0.55)	6.10 (0.18)	0.51 (0.018)	27.67 (0.40)	6.08 (0.13)	0.50 (0.013)	29.52 (0.46)	6.05 (0.11)	0.50 (0.010)	0.027 (0.002)	0.92 (0.005)
(C) Data simulated by neither the AR(1) model nor the SAD(1) model												
SAD(1)	10.44 (5.41)	26.15 (0.55)	6.02 (0.16)	0.50 (0.01)	27.97 (0.37)	6.03 (0.10)	0.50 (0.006)	29.97 (0.51)	6.07 (0.13)	0.50 (0.008)	1.49 (0.040)	1.01 (0.008)
AR(1)	10.86 (4.93)	25.89 (0.51)	6.19 (0.20)	0.50 (0.008)	27.75 (0.33)	6.08 (0.11)	0.50 (0.005)	29.61 (0.43)	6.05 (0.13)	0.50 (0.006)	0.029 (0.002)	0.89 (0.006)

implications for the design of an efficient early selection program in plant and animal breeding and for asking and addressing biological questions at the interface of genetics, development and evolution.

### ACKNOWLEDGEMENTS

We thank three anonymous referees for their constructive comments on this manuscript. This work was partially supported by a NIH grant DK52514 to J.M.C., and by the National Science Foundation of China to R.W. (09 95671). The publication of this manuscript is approved as journal series No.-10587 by the Florida Agricultural Experimental Station.

### REFERENCES

Atchley,W.R. and Zhu,J. (1997) Developmental quantitative genetics, conditional epigenetic variability and growth in mice. *Genetics*, **147**, 765–776.  
 Carroll,R.J. and Ruppert,D. (1984) Power-transformations when fitting theoretical models to data. *J. Am. Statist. Assoc.*, **79**, 321–328.  
 Churchill,G.A. and Doerge,R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.  
 Daniels,M.J. and Pourahmadi,M. (2002) Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, **89**, 553–566.  
 Dempster,A.P. et al. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, **39**, 1–38.  
 Diggle,P.J., Heagerty,P., Liang,K.Y. and Zeger,S.L. (2002) *Analysis of Longitudinal Data*. Oxford University Press, Oxford, UK.  
 Gabriel,K.R. (1962) Ante-dependence analysis of an ordered set of variables. *Ann. Math. Statist.*, **33**, 201–212.  
 Hurvich,C.M. and Tsai,C.-L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.  
 Jaffrézic,F. et al. (2003) Structured antedependence models for genetic analysis of repeated measures on multiple quantitative traits. *Genet. Res.*, **82**, 55–65.  
 Kirkpatrick,M. and Heckman,N. (1989) A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *J. Math. Biol.*, **27**, 429–450.  
 Lander,E.S. and Botstein,D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.  
 Louis,T.A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. B*, **44**, 226–233.  
 Ma,C.-X. et al. (2002) Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics*, **161**, 1751–1762.

Meng,X.-L. and Rubin,D.B. (1991) Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Am. Statist. Assoc.*, **86**, 899–909.  
 Nelder,J.A. and Mead, R. (1965). A simplex method for function minimization. *Comput. J.*, **7**, 308–313.  
 Núñez-Antón,V. (1997) Longitudinal data analysis: non-stationary error structures and antedependence models. *Appl. Stochast. Models Data Anal.*, **13**, 279–287.  
 Núñez-Antón,V. and Zimmerman,D.L. (2000) Modeling nonstationary longitudinal data. *Biometrics*, **56**, 699–705.  
 Pan,J.X. and Mackenzie,G. (2003) On modelling mean-covariance structures in longitudinal studies. *Biometrika*, **90**, 239–244.  
 Pletcher,S.D. and Geyer,C.J. (1999) The genetic analysis of age-dependent traits: modeling the character process. *Genetics*, **153**, 825–835.  
 Pletcher,S.D. and Jaffrézic,F. (2002) Generalized character process models: estimating the genetic basis of traits that cannot be observed and that change with age or environmental conditions. *Biometrics*, **58**, 157–162.  
 Pourahmadi,M. (1999) Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, **86**, 677–690.  
 Pourahmadi,M. (2000) Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, **87**, 425–435.  
 Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.  
 Vaughn,T.T. et al. (1999) Mapping quantitative trait loci for murine growth: a closer look at genetic architecture. *Genet. Res.*, **74**, 313–322.  
 von Bertalanffy,L. (1957) Quantitative laws in metabolism and growth. *Q. Rev. Biol.*, **32**, 217–231.  
 West,G.B. et al. (2001) A general model for ontogenetic growth. *Nature*, **413**, 628–631.  
 Wu,R.L. et al. (2002) A logistic mixture model for characterizing genetic determinants causing differentiation in growth trajectories. *Genet. Res.*, **19**, 235–245.  
 Wu,R.L. et al. (2004a) A general framework for analyzing the genetic architecture of developmental characteristics. *Genetics*, **166**, 1541–1551.  
 Wu,R.L. et al. (2004b) Functional mapping of quantitative trait loci underlying growth trajectories using a transform-both-sides logistic model. *Biometrics*, **60**, 729–738.  
 Wu,W.B. and Pourahmadi,M. (2003) Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, **90**, 831–844.  
 Zhao,W. et al. (2004a) A unifying statistical model for QTL mapping of genotype-sex interaction for developmental trajectories. *Physiol. Genom.*, **19**, 218–227.  
 Zhao,W. et al. (2004b) A fast algorithm for functional mapping of complex traits. *Genetics*, **167**, 2133–2137.  
 Zimmerman,D.L. and Núñez-Antón,V. (1997) Structured antedependence models for longitudinal data. In Gregoire,T.G., Brillinger,D.R., Diggle,P.J., Russek-Cohen,E., Warren,W.G. and Wolfinger,R. (eds), *Modelling Longitudinal and Spatially Correlated Data. Methods, Applications, and Future Directions*. Springer-Verlag, New York, pp. 63–76.  
 Zimmerman,D.L. and Núñez-Antón,V. (2001) Parametric modeling of growth curve data: an overview (with discussion). *Test*, **10**, 1–73.