Genet. Res., Camb. (2002), **79**, *pp.* 235–245. *With* 4 *figures.*
DOI: 10.1017/S0016672302005633 Printed in the United Kingdom

A logistic mixture model for characterizing genetic determinants causing differentiation in growth trajectories

RONGLINGWU[1]*†, CHANG-XINGMA[1, 2]†, MYRONCHANG[1], RAMON C.LITTELL[1], SAMUEL S.WU[1], TONGMINGYIN[3], MINRENHUANG[3], MINGXIUWANG[3] AND GEORGECASELLA[1]

[1] *Department of Statistics, University of Florida, Gainesville, FL 32611, USA*
[2] *Department of Statistics, Nankai University, Tianjian 300071, China*
[3] *The Key National Laboratory of Forest Genetics and Gene Engineering, Nanjing Forestry University, Nanjing, Jinagsu 210037, China*

(*Received 21 November 2001 and in revised form 5 February 2002*)

Summary

The logistic or S-shaped curve of growth is one of the few universal laws in biology. It is certain that there exist specific genes affecting growth curves, but, due to a lack of statistical models, it is unclear how these genes cause phenotypic differentiation in growth and developmental trajectories. In this paper we present a statistical model for detecting major genes responsible for growth trajectories. This model is incorporated with pervasive logistic growth curves under the maximum likelihood framework and, thus, is expected to improve over previous models in both parameter estimation and inference. The power of this model is demonstrated by an example using forest tree data, in which evidence of major genes affecting stem growth processes is successfully detected. The implications for this model and its extensions are discussed.

1. Introduction

The phenotype of an individual, including its size, shape, anatomy and metabolic rate, changes with age. The genetic analysis of these age-dependent phenotypes, termed growth trajectories, has long been of interest to students in different disciplines of biology and genetics. Plant and animal breeders, faced with dramatic growth and morphological changes of domestic species over times, are interested in the genetic mechanism underlying these changes. With this information, the yields of food or fibre can be improved by altering growth patterns through artificial selection (e.g. Wu *et al*., 1992). Human geneticists are concerned with age-specific gene expression because of the potential to design specific drugs that can remove human diseases at younger ages (Roses, 2000). The genetic basis of age-dependent fitness components is essential for evolutionary biologists to estimate the origin of trait evolution and predict the evolutionary and developmental change of phenotypes within particular environmental contexts (Atchley, 1984; Rice, 1997).

Classic quantitative genetics based on a univariate analysis provides a simple method for comparing the genetic control of growth at different ages in a variety of organisms (Cheverud *et al*., 1983; Kremer, 1992; Wu *et al*., 1992; Dieters *et al*., 1995; Lynch & Walsh, 1998). By treating growth for each age as a different trait, a multivariate analysis approach is developed to capture the covariance information of growth among different ages (Hughes & Charlesworth, 1994; Pletcher *et al*., 1998). However, this approach is problematic when the number of ages is large and when measurements are taken at irregular intervals. Recognizing the limit of the classical approach, Kirkpatrick and colleagues developed an infinite-dimensional model for estimating genetic parameters of growth by treating growth trajectories as an infinite number of measurements (Kirkpatrick & Hackman, 1989; Kirkpatrick *et al*., 1990, 1994), rather than a finite number of measurements, as can be manipulated by a classic genetic model. Similar theoretical models for developmental genetic studies were also put forth using random regression theory (Meyer, 1998) and stochastic process theory (Pletcher & Geyer, 1999). Based

* Corresponding author. Tel: +1 (352) 392 3806. Fax: +1 (352) 392 8555. e-mail: rwu@stat.ufl.edu
† These two authors contributed equally to this work.

on these studies, a widely accepted view of the genetic basis of growth proposes that a given set of genes affecting growth is progressively modified; after each season a portion of the set is replaced and, several years later, the original set has been totally modified (Kremer, 1992; Meyer & Hill, 1997; Kirkpatrick, 1997; Pletcher & Geyer, 1999). This view is generally confirmed by recent quantitative trait locus (QTL) analyses from genetic maps, which further show that growth variations may result from the activation and repression of genes responsible for changes in growth (Nuzhdin *et al.*, 1997; Vaughn *et al.*, 1999; Wu *et al.*, 2002).

Although current quantitative genetic and molecular mapping approaches are useful in detecting development-dependent genetic components, they have not been incorporated with universal growth-curve laws within a statistical framework. Every cell, organ, tissue, organism or population, and every species, in a range in body size from microbes (10^{-13} g) to blue whales (10^{8} g), follows as exponential growth law or curve (von Bertalanffy, 1957). The objective of this study is to embed one of the most commonly used growth curves – logistic or sigmoid (Niklas, 1994) – in the quantitative genetic analysis of growth trajectories. A maximum-likelihood-based method implemented with the EM algorithm is used to detect major genes that are responsible for growth differentiation. In an example using forest tree data, the logistic-based genetic model put forth in this paper successfully provides evidence of the existence of major genes affecting stem height and diameter growth. The differential patterns of the expression of these major genes are consistent with the findings from earlier quantitative and ecological genetic analysis of the same material (Wu & Stettler, 1996).

2. Growth laws

A growth law can be visualized as the 'force field' propelling a point through a phenotype space, tracing out the ontogenetic path. If the size of an organism is denoted by y, its *ontogenetic trajectory* $y(t)$ can be generated through the differential dy/dt, which models the growth rate. Many differential functions have been established to describe growth trajectory. Basically, they are sorted into three categories: (1) exponential, (2) saturating and (3) sigmoidal (von Bertalanffy, 1957; Niklas, 1994). Each of these growth models has a common feature that the development of ontogenetic trajectory is regulated by a set of 'control parameters' such as onset age of growth, offset signal for growth, growth rate during the period of growth and initial size at the commencement of the growth period. Also, each of these growth models exhibits an initial phase of exponential growth simply due to the geometrically multiplying population of newly differentiated cells. This initial growth phase has the property that small perturbations in growth rate or onset age are amplified enormously during ontogeny. Thus, it is easy to find examples of how a small 'mutation' in a growth parameter causes a series of developmental alterations that produce a phenotype qualitatively different from the normal one.

The sigmoidal (or logistic) growth function, detected for the first time by Pearl (1925), 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

$$g(t) = \frac{a}{1 + b\mathrm{e}^{-rt}}, \tag{1}$$

where a is the asymptotic or limit value of g when $t \to \infty$, $a/(1+b)$ is the initial value of g when $t = 0$ and r is the relative rate of growth (von Bertalanffy, 1957). The logistic growth curve consists of two phases: an exponential phase and an asymptotic phase. The overall form of the curve is determined by different combinations of parameters a, b and r. 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.

The logistic growth curve described in (1) can be used to determine the coordinates of a biologically important point in the entire growth trajectory – the *inflection point* – where the exponential phase ends and the asymptotic phase begins (Niklas, 1994). The time at the inflection point corresponds to the time point at which a maximum growth rate occurs. The time (t_I) and growth [$g(t_I)$] at the inflection point for a QTL genotype can be derived as

$$\begin{cases} t_I = \dfrac{\log b}{r} \\ g(t_I) = \dfrac{a}{2} \end{cases}. \tag{2}$$

The difference in the coordinates between different genotypes provides important information about the genetics and evolution of growth trajectories (Niklas, 1994). Moreover, the time at the inflection point, together with the initial growth and asymptotic growth, determine exclusively the difference of two growth curves. Any two curves will not be distinguishable if they have the same values for these three variables.

Many of the established growth models can only describe the growth that has occurred, but cannot be used to predict growth *per se*. From a mechanistic perspective, the scaling of growth can be described in

https://doi.org/10.1017/S0016672302005633 Published online by Cambridge University Press

terms of the balance between the rate at which metabolities are synthesized (anabolism) and the rate at which metabolites are consumed (catabolism) (West *et al*., 2001). Using this principle, Beverton & Holt (1957) formulated the mathematical equations for predicting the length $L(t)$ and mass $M(t)$ of an organism at any time t:

$$L(t) = L_\infty - (L_\infty - L_0)\mathrm{e}^{-rt},$$
$$M(t) = [M_\infty^{1/3} - (M_\infty^{1/3} - M_0^{1/3})\mathrm{e}^{-rt}]^{1/3},$$

where L_∞ and M_∞ are the length and mass at time ∞, L_0 and M_0 are the length and mass at time 0, and k is the specific growth rate. These equations successfully describe the growth curves of a variety of marine fish (Beverton & Holt, 1957) and some plants under natural conditions (Blackman, 1961). Other growth functions (Richards, 1959), like the Gompertz equation

$$G(t) = a\mathrm{e}^{-b\exp(-rt)},$$

can describe growth curves for particular kinds of organisms or under particular environments.

3. Genetic design

To make our logistic-based genetic model more applicable, we suppose a general full-sib family derived from two arbitrarily heterozygous parent. With two such assumed parents, there may be a varying number of genotype classes in the progeny for a major gene affecting growth trajectories. This major gene may form a total of four different genotypes segregating 1:1:1:1 in the progeny if these two parents carry different alleles, e.g. $Q_1Q_2 \times Q_3Q_4$ (*outcrossing design*). But three genotypes segregating 1:2:1 will result if the two parents carry the same allelle system and both are heterozygous, i.e. $Q_1Q_2 \times Q_1Q_2$ (F_2 *design*). Alternatively, there are only two genotypes segregating 1:1 if one parent is heterozygous but the other is homozygous, like $Q_1Q_2 \times Q_2Q_2$ (*backcross design*). We use k to denote the number of genotypes in the progeny, $k = 2$ for the backcross design, $k = 3$ for the F_2 design and $k = 4$ for the outcrossing design. For a particular genotype j, the parameters describing its logistic curve are denoted by a_j, b_j and r_j, $j = 1,\ldots, k$ for any possible design. The comparisons of these parameters among different genotypes can determine whether and how this putative gene affects growth trajectories.

4. Statistical method

Assume that all N progeny in the pedigree are measured for a quantitative trait at each of m times. The trait phenotypes of progeny i measured at time t can be expressed by a linear statistical model (Kirkpatrick & Heckman, 1989; Pletcher & Geyer, 1999):

$$y_i(t) = \sum_{j=1}^{k} x_{ij} g_j(t) + e_i(t), \quad k = 2 \text{ or } 3,$$

where x_{ij} is an indicator variable describing a possible genotype j of the major gene for progeny i and defined as 1 if a particular genotype is observed and 0 otherwise, $g_j(t)$ is the genotypic value of the trait 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 phenotypes of the trait at all time points 1, 2,…, m for each genotype group follow a multivariate normal density,

$$f_j(\mathbf{y}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{-(\mathbf{y}-\mathbf{g}_j)^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{g}_j)/2\},$$

where $\mathbf{g}_j$ is the vector of the expected genotypic values of the trait for QTL genotype j measured for t times and $\boldsymbol{\Sigma}$ is the residual variance–covariance matrix of $\mathbf{y}$. Unlike a usual single-trait analysis, the expected genotypic values of an age-specific trait and its residual variance–covariances among different time points are fitted by mathematical or statistical models built upon biological backgrounds.

Indeed, $\mathbf{g}_j$ can be modelled by one growth equation as described in Section 2. The choice of an appropriate growth equation depends on its goodness-of-fit to observational data. Although exponential-based growth models have been found to fit a broad variety of species (Niklas, 1994), the Beverton–Holt equation (1957) and the Gompertz equation (Richards, 1959) may be more suitable for the prediction of the growth that has not occurred. To simplify our formulations, we use a logistic curve (Eq. (1); West *et al*., 2001) as an example to incorporate growth models into the detection of a major gene affecting growth trajectories. Thus, we model $\mathbf{g}_j$ by

$$\mathbf{g}_j = (g_j(t))_{1\times m} = \left(\frac{a_j}{1+b_j\mathrm{e}^{-r_j t}}\right)_{1\times m}.$$

In statistics, many approaches have been proposed to model the structure of $\boldsymbol{\Sigma}$. For simplicity, $\boldsymbol{\Sigma}$ can be assumed identical among different genotypes and modelled using AR(1) repeated measurement errors (Davidian & Giltinan, 1995; Verbeke & Molenberghs, 2000):

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho^{m-1} \\ \rho & 1 & \cdots & \rho^{m-2} \\ \cdots & \cdots & \cdots & \cdots \\ \rho^{m-1} & \rho^{m-2} & \cdots & 1 \end{bmatrix}. \tag{3}$$

The matrix $\boldsymbol{\Sigma}$ of (3) assumes variance stationarity, i.e. there is the same residual variance (σ^2) for the trait at

https://doi.org/10.1017/S0016672302005633 Published online by Cambridge University Press

each time, and correlation stationarity, i.e. the covariance between different measurements decreases proportionally (in ρ) with increased time interval (see also Pletcher & Geyer, 1999). The inverse matrix of $\mathbf{\Sigma}$ and its determinant $|\mathbf{\Sigma}|$ are derived in Appendix A. It should be pointed out that a more accurate fit of $\mathbf{\Sigma}$ than (3) may exist. In practice, selecting an appropriate fit of $\mathbf{\Sigma}$ should follow two rules: (1) an optimal approximation to biological reality, and (2) ease of mathematical manipulation (e.g. inversion or factorization).

Let vector $\mathbf{\Omega} = (a_j, b_j, c_j, \rho, \sigma^2)^T (j = 1,\ldots, k)$ denote unknown parameters to be estimated. The likelihood of a sample of N progeny, whose growth trajectories are measured for k times and controlled by a pleiotropic major gene, can be represented by a mixture model

$$L(\mathbf{\Omega}) = \prod_{i=1}^{N} \left[\sum_{j=1}^{k} \gamma_{ij} f_j(\mathbf{y}_i) \right], \tag{4}$$

where γ_{ij} is the proportion of each mixture normal (i.e. genotypes) for a progeny i. This is $(\frac{1}{2}\ \frac{1}{2})$ for the backcross, $(\frac{1}{4}\ \frac{1}{2}\ \frac{1}{4})$ for the F_2 design, or $(\frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4})$ for the outcrossing design.

The maximum likelihood estimates (MLEs) of the unknown parameters under the pleiotropic model can be computed by implementing an EM algorithm (Dempster *et al.*, 1977; Meng & Rubin, 1993). The log-likelihood is given by

$$\log L(\mathbf{\Omega}) = \sum_{i=1}^{N} \log \left[\sum_{j=1}^{k} \gamma_{ij} f_j(\mathbf{y}_i) \right], \tag{5}$$

with derivatives

$$\begin{aligned} \frac{\partial}{\partial \mathbf{\Omega}_m} \log L(\mathbf{\Omega}) &= \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{\gamma_{ij} \frac{\partial}{\partial \mathbf{\Omega}_m} f_j(\mathbf{y}_i)}{\sum_{j=1}^{k} \gamma_{ij} f_j(\mathbf{y}_i)} \\ &= \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{\gamma_{ij} f_j(\mathbf{y}_i)}{\sum_{j=1}^{k} \gamma_{ij} f_j(\mathbf{y}_i)} \frac{\partial}{\partial \mathbf{\Omega}_m} \log f_j(\mathbf{y}_i) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{k} \Gamma_{ij} \frac{\partial}{\partial \mathbf{\Omega}_m} \log f_j(\mathbf{y}_i), \end{aligned}$$

where we define

$$\Gamma_{ij} = \frac{\gamma_{ij} f_j(\mathbf{y}_i)}{\sum_{j=1}^{k} \gamma_{ij} f_j(\mathbf{y}_i)}, \tag{6}$$

which could be thought of as a posterior probability that progeny i have QTL genotype j. We then implement the EM algorithm with the expanded parameter set $\{\mathbf{\Omega}, \mathbf{\Gamma}\}$, where $\mathbf{\Gamma} = \{\mathbf{\Gamma}_{ij}, j = 1,\ldots, k; i = 1,\ldots, N\}$. Conditional on $\mathbf{\Gamma}$, we solve for the zeros of $\frac{\partial}{\partial \mathbf{\Omega}_m} \log L(\mathbf{\Omega})$ to get out estimates of $\mathbf{\Omega}$ (the M step). The estimates are then used to update $\mathbf{\Gamma}$ (the E step), and the process is repeated until convergence. The values at convergence are the MLEs (see Appendix B for another variant on this model). The standard errors of the MLEs are estimated using the inverse of the Fisher information matrix. The derivation of this matrix needs the second derivatives of the log-likelihood in (5) with respect to the unknown parameters.

5. Hypothesis tests

A number of biologically meaningful hypotheses can be tested based on the logistic-based genetic model. The hypothesis about the existence of a major gene affecting an overall growth curve can be formulated as

$$\begin{cases} H_0: & a_1 = \cdots = a_k, b_1 = \cdots = b_k, r_1 = \cdots = r_k \\ H_1: & \text{at least one of the equalities above does not hold.} \end{cases} \tag{7}$$

The test statistic for testing the above hypotheses is calculated as the log-likelihood ratio of the full model (H_1) over the reduced model (H_0):

$$LR = -2 \log \left[\frac{L(\tilde{\mathbf{\Omega}})}{L(\hat{\mathbf{\Omega}})} \right],$$

where $\tilde{\mathbf{\Omega}}$ and $\hat{\mathbf{\Omega}}$ denote the ML estimates of the unknown parameters under H_0 and H_1, respectively. Unlike a usual situation, in which the LR is approximately χ^2-distributed with 3 degrees of freedom for the backcross or 6 for the F_2 design, the distribution of the likelihood ratio test for the detection of a segregating major gene is a mixture of χ^2 and Dirac distributions (Loisel *et al.*, 1994). Here we use a simulation study to calculate the threshold for acclaiming the existence of a gene affecting the overall growth curves (Lynch & Walsh, 1998).

For a particular full-sib family, one should test which design – backcross, F_2 or outcrossing – is the best fit of the data. Because the three possible designs are not nested, AIC based on Akaike's (1974) information criterion should be used:

$$\begin{aligned} \text{AIC} = &-2 \ln(\text{maximum likelihood}) \\ &+2(\text{number of fitted parameters}). \end{aligned}$$

The design with the smallest AIC is chosen as the most parsimonious.

A hypothesis test can also be performed on the time that the detected major gene turns on or off to affect growth trajectories, by comparing the difference of the expected means between different genotypes at various time points. At a given time t^*, the hypothesis for a general design is

$$\begin{cases} H_0: & g_1(t^*) = \cdots = g_k(t^*) \\ H_1: & \text{at least one of the equalities above does not hold.} \end{cases} \tag{8}$$

https://doi.org/10.1017/S0016672302005633 Published online by Cambridge University Press

If H_1 is accepted, this means that the major gene has a significant effect on variation in growth at time t^*. Testing the hypotheses (8) is equivalent to testing the difference between the model with no restriction and the model with the restriction

$$\frac{a_1}{1+b_1 e^{-r_1 t^*}} = \cdots = \frac{a_k}{1+b_k e^{-r_k t^*}}.$$

Because t^* is given, one of the six logistic parameters can be expressed as a function of the other five and, thus, there is $(k-1)$ fewer parameters to be estimated for the model with the above restriction (the reduced model) than the model with no restriction (the full model). The test statistic for (8) is approximately χ^2-distributed with k-1 degrees of freedom. By scanning time points from 1 to m, one can find the time points at which the major gene starts or ceases to exert an effect on growth.

For the F_2 design, one can typically test whether the additive or dominant effects are significant on growth trajectories. The test is formulated as:

$$\begin{cases} H_0: & g_1(t^*) = g_3(t^*) \\ H_1: & g_1(t^*) \neq g_3(t^*) \end{cases}, \tag{9}$$

for the additive effect, and

$$\begin{cases} H_0: & [g_1(t^*)+g_3(t^*)]/2 = g_2(t^*) \\ H_1: & [g_1(t^*)+g_3(t^*)]/2 \neq g_2(t^*) \end{cases}, \tag{10}$$

for the dominant effect of the major gene, where genotypes 1 and 3 are the homozygotes and genotype 2 is the heterozygote.

In addition, the genotypic differences in time (t_I) and growth ($G(t_I)$) at the inflection point can be tested. The test for the genotypic difference is based on the restriction

$$\frac{\log b_1}{r_1} = \cdots = \frac{\log b_k}{r_k}, \tag{11}$$

for t_I at maximum growth rate, and

$$\frac{a_1}{2} = \cdots = \frac{a_k}{2}, \tag{12}$$

for $(G(t_I))$ at maximum growth rate.

6. Example

(i) *Plant material*

We use an example from an outcrossing forest tree to demonstrate the power of our statistical model. As one of our continuing genome projects, this example was derived from the triple hybridization of *Populus* (poplar). A *P. deltoides* clone designated I-69 was used as a female parent to mate with a *P. deltoides* × *P. nigra* interspecific clone designated I-45 as a male parent (Wu *et al.*, 1992). The hybrids between *P. deltoides* and *P. nigra* are called Euramerica poplar (*P. euramericana*). Both I-69 and I-45 were selected at the Research Institute for Poplars in Italy in 1950s and were introduced to China in 1972. In spring 1988, a total of 450 one-year-old rooted three-way hybrid seedlings were planted at a spacing of 4 × 5 m at a forest farm near Xuchou City, Jiangsu Province, China. The total stem height and diameter growth were measured at the end of each of 14 growing seasons. In this study, we use all available 14 year measurements for a subset of 125 trees randomly selected from this hybrid population.

By plotting annual measurements against year, we observe marked evidence that each of the genotypes followed the S-shaped growth curve (Fig. 1; $P < 0{\cdot}001$, results not shown). The statistical model built upon this universal growth law (West *et al.*, 2001) is used to detect major genes responsible for these growth trajectories. Since our hybrid material is highly heterozygous, the segregation pattern of genes may not be fixed. Hence, we will test all three designs – the backcross, F_2 and outcrossing – from which a most likely one is chosen using AIC.

(ii) *Threshold value*

The empirical estimate of the critical value for testing the existence of a major gene is obtained from the distribution of the *LR* values calculated from the simulated phenotypic data assuming no QTL. The experimental design used in this simulation mimics the example, assuming 125 progeny and 14 measurement points. A set of phenotypic values for these progeny are simulated to follow a single logistic curve under the assumption of no QTL involved in growth trajectories. The three parameters describing the assumed uniform logistic curve are given as $a = 19{\cdot}23$, $b = 5{\cdot}89$ and $r = 0{\cdot}38$ for height and $a = 28{\cdot}74$, $b = 13{\cdot}96$ and $r = 0{\cdot}55$ for diameter, which are the corresponding MLEs of the logistic parameters under the null hypothesis of no QTL for these two traits, respectively. The phenotypic values simulated at a total of 14 time points are constrained to be correlated with the residual covariance matrix,

$$\mathbf{\Sigma} = \begin{bmatrix} 2{\cdot}04 & 1{\cdot}84 & \cdots & 0{\cdot}52 \\ 1{\cdot}84 & 2{\cdot}04 & \cdots & 0{\cdot}58 \\ \cdots & \cdots & \cdots & \cdots \\ 0{\cdot}52 & 0{\cdot}58 & \cdots & 2{\cdot}04 \end{bmatrix},$$

for height, and

$$\mathbf{\Sigma} = \begin{bmatrix} 5{\cdot}15 & 4{\cdot}94 & \cdots & 3{\cdot}02 \\ 4{\cdot}94 & 5{\cdot}15 & \cdots & 3{\cdot}15 \\ \cdots & \cdots & \cdots & \cdots \\ 3{\cdot}02 & 3{\cdot}15 & \cdots & 5{\cdot}15 \end{bmatrix},$$

https://doi.org/10.1017/S0016672302005633 Published online by Cambridge University Press

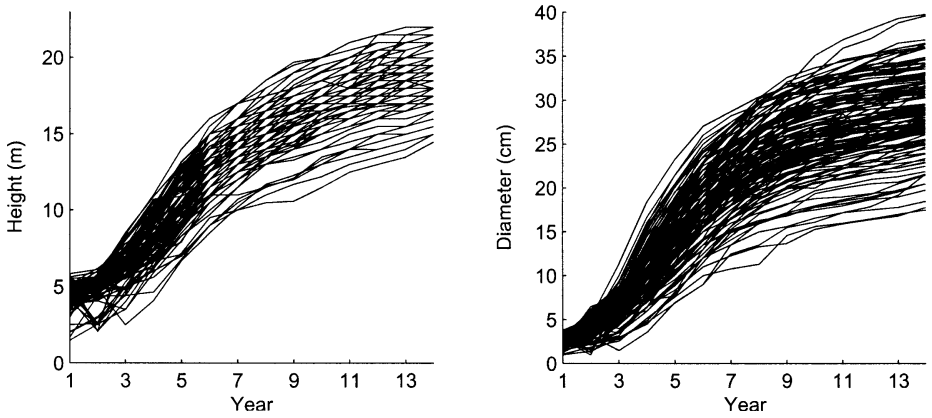

Fig. 1. Plots of height and diameter growth versus time for poplar hybrids.

Table 1. *The 99th and 99·9th percentiles of the distribution of the LR values under three different designs used as empirical critical values to declare the existence of a QTL for growth trajectories in height and diameter*

Significant level (α)	Height			Diameter		
	Backcross	F_2	Outcrossing	Backcross	F_2	Outcrossing
0·01	12·96	17·15	20·41	11·39	14·51	16·79
0·001	18·90	24·48	28·17	16·86	20·67	21·27

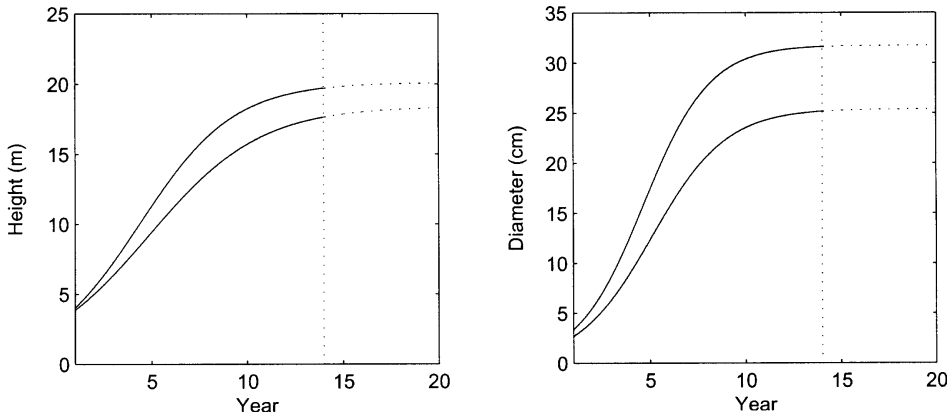

Fig. 2. Two growth curves each presenting a group of genotypes for both height and diameter when the backcross design is assumed.

for diameter. Similarly, these are the MLEs of the between-year residual covariance matrices for the two traits, respectively, under the null hypothesis of no QTL.

The simulated data are analysed by three different genetic models under the backcross, F_2 and outcrossing designs. For each design, the distribution of the *LR* values over 1000 simulation replicates can be approximated by a χ^2 distribution. The 99th and 99·9th percentiles of the distribution of the maximum are used as empirical critical values to declare the existence of a QTL for growth trajectories at the significance levels $\alpha = 0{\cdot}01$ and 0·001. Table 1 lists these percentiles simulated for height and diameter, respectively, under the three different genetic designs.

(iii) *Backcross design*

Assuming that one parent used for the three-way hybridization is heterozygous, whereas the other parent is homozygous, the progeny has two different genotypes segregating in a 1:1 ratio. Under this backcross model, a major gene with strong effect on growth trajectories was detected (Fig. 2). The test statistics for testing the existence of a major gene are 71·8 for heights and 281·0 for diameters, both

https://doi.org/10.1017/S0016672302005633 Published online by Cambridge University Press

Table 2. *Parameter estimates* (*with standard errors in parentheses*) *of a major gene affecting logistic curves in a poplar hybrid progeny under the backcross design*

	Height		Diameter	
Parameters	Genotype 1	Genotype 2	Genotype 1	Genotype 2
a	20·1207 (0·1666)	18·4219 (0·2124)	31·7783 (0·3103)	25·4435 (0·3206)
b	5·9866 (0·2566)	5·3136 (0·2720)	14·5797 (0·7205)	14·0162 (0·8983)
r	0·4066 (0·0094)	0·3431 (0·0100)	0·5797 (0·0097)	0·5168 (0·0115)
σ^2	1·2058 (0·0957)		5·2336 (0·4568)	
ρ	0·8390 (0·0136)		0·4568 (0·0061)	
LR	71·74***		281·07***	

***$P < 0.001$.

Table 3. *Parameter estimates* (*with standard errors in parentheses*) *of a major gene affecting logistic curves in a poplar hybrid progeny under the F_2 design*

	Height			Diameter		
Parameters	Genotype 1	Genotype 2	Genotype 3	Genotype 1	Genotype 2	Genotype 3
a	20·6283 (0·2469)	19·2419 (0·1900)	17·6023 (0·3706)	32·7026 (0·3272)	27·8490 (0·3188)	22·0987 (0·5485)
b	6·2018 (0·3197)	5·4560 (0·2256)	4·9612 (0·4341)	15·1740 (0·7866)	14·3835 (0·7736)	15·8718 (2·0699)
r	0·4180 (0·0126)	0·3684 (0·0100)	0·3189 (0·0180)	0·5898 (0·0116)	0·5370 (0·0105)	0·5130 (0·0228)
σ^2		0·9513 (0·0746)			3·4207 (0·3028)	
ρ		0·7976 (0·0161)			0·9000 (0·0092)	
LR		95·83***			363·25***	

***$P < 0.001$.

significantly greater than empirical critical thresholds at $\alpha = 0{\cdot}001$, 18·90 and 16·86, respectively. Table 2 gives the ML estimates and their standard errors for the logistic parameters for each genotype group at the major gene. Small standard errors imply that our estimates have high precision.

(iv) *F_2 design*

Using the F_2 design, we detect a significant major gene heterozygous in both parents. The segregation of three genotypes in the progeny under the F_2 design leads to three distinct growth trajectories, as seen from the test statistics (Table 3). It appears that the F_2 design has a strikingly increased fit to diameter growth ($LR = 363{\cdot}2$) compared with the backcross design ($LR = 281{\cdot}0$).

The major gene detected from the F_2 design seems to be additive, because the heterozygote (genotype 2) is intermediate between the two homozygotes (genotypes 1 and 3 (Fig. 3). This major gene displays different effects on stem height and diameter growth trajectories.

(v) *Outcrossing design*

Statistically, it is also significant that growth trajectories are controlled by a quadri-allelic major gene

https://doi.org/10.1017/S0016672302005633 Published online by Cambridge University Press

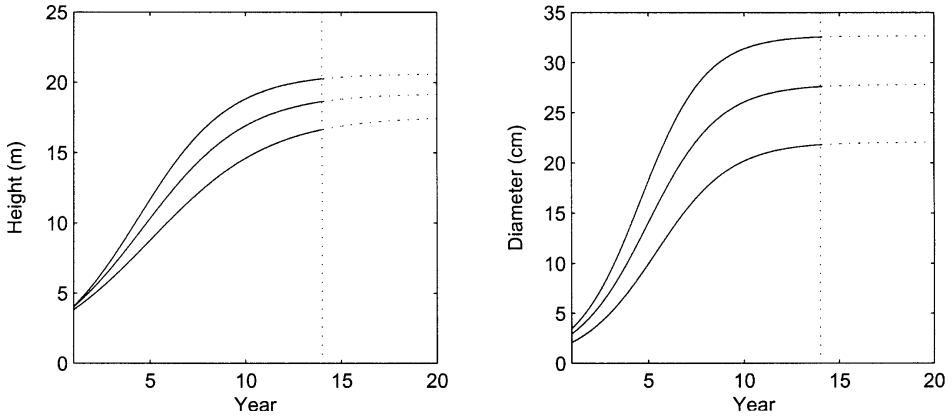

Fig. 3. Three growth curves each presenting a group of genotypes for both height and diameter when the F_2 design is assumed.

Table 4. *Parameter estimates of a major gene affecting logistic curves in a poplar hybrid progeny under the outcrossing design*

	Height				Diameter			
Parameters	Genotype 1	Genotype 2	Genotype 3	Genotype 4	Genotype 1	Genotype 2	Genotype 3	Genotype 4
a	18·7483	20·8837	19·3391	16·8294	31·2871	33·1284	27·2570	22·2130
b	5·4871	6·1716	5·3658	4·4931	16·3684	16·6571	13·8161	16·1576
r	0·3490	0·4091	0·3879	0·2911	0·6572	0·5424	0·5330	0·5148
σ^2	0·7610				3·1284			
ρ	0·7488				0·8990			
LR	104·02***				445·03***			

***$P < 0.001$.

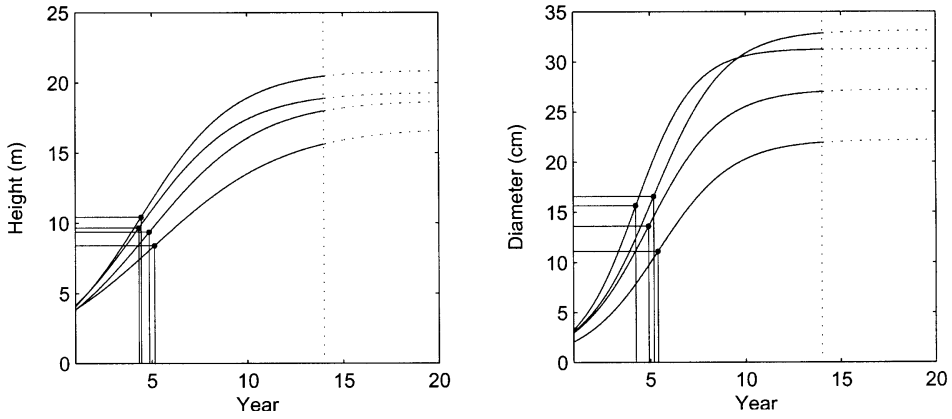

Fig. 4. Four growth curves each presenting a group of genotypes for both height and diameter when the outcrossing design is assumed. The times (t_I) and growth ($g(t_I)$) at the inflection point are indicated by four coordinates for genotypes 1–4 at the major gene identified.

segregating in the poplar hybrid family studied (Table 4; Fig. 4). Similar to the backcross and F_2 design, the outcrossing design displays a better fit to diameter than height growth.

According to Akaike's (1974) information criterion (AIC), a most likely design used to fit both height and diameter growth is the outcrossing design, whose AIC value is 1728·0 for height, smaller than 1738·1 in the backcross and 1729·1 in the F_2, and is 1628·6 for diameter, smaller than 2704·5 in the backcross and 2666·5 in the F_2. In this study, we do not know whether this is the same major gene that affects height and diameter growth in the poplar hybrids. We will use the outcrossing design to analyse age-specific differentiation of the effect of the major gene on height and diameter growth.

https://doi.org/10.1017/S0016672302005633 Published online by Cambridge University Press

The gene detected for the overall curve of height growth from the outcrossing design significantly affects the time at which trees grow most rapidly in height (the inflection point), as indicated by the test statistics of $LR = 21{\cdot}65$ which is calculated as the log-likelihood ratio of the full model and the reduced model with a restriction of Eq. 11 ($\chi^2_{0{\cdot}001(3)} = 16{\cdot}27$). The gene detected for diameter growth from the outcrossing design also affects the inflection point of this trait, given $LR = 30{\cdot}32$ larger than $\chi^2_{0{\cdot}001(3)} = 16{\cdot}27$. The genetic control of the inflection point suggests that growth trajectory can be genetically modified to increase a tree's capacity to effectively acquire spatial resources.

Although at a similar time point (both around age 2 years) the major gene detected exerts an effect on height and diameter according to results from test (8) (data not shown), it generates different shapes of growth curves for these two growth traits. It appears that height tends to have reduced differentiation after year 14, whereas diameter maintains a high degree of differentiation after this time point. This finding is consistent with ecological viewpoints of allometric scaling (Wu & Stettler, 1996). From year 14, strong competitive interactions occur among poplar trees due to the closure of canopies. Thus, to maintain vigorous growth, larger individuals must allocate more stem biomass to radial growth than to height growth, whereas smaller individuals tend to emphasize height growth at the cost of radial growth to gain access to light. Such changes in biomass allocation during canopy closure result in reduced differences in height but increased differences in diameter. It is not surprising that genes are responsible for this transition of growth phases.

7. Discussion

We have incorporated ubiquitous growth curve laws into a statistical framework for detecting a major gene governing growth trajectories. While many current statistical models are criticized for their loose ties with biological realities, our model represents an attempt to explore biological questions in a way that integrates biology and statistics. The statistical and biological power of this new model was well demonstrated by the successful detection of major genes affecting growth trajectories in an example from a forest tree. Compared with earlier attempts to adapt quantitative genetics to growth trajectories, our model offers significant advantages and has potential to improve current quantitative developmental genetic studies.

First, this model views growth traits as a full trajectory rather than a set of landmark ages. Classic quantitative genetic methods, as widely used in the literature (Kremer, 1992; Wu *et al*., 1992; Hughes & Charlesworth, 1994), have no power to analyse and model the data of growth as a continuous function of time and measured at uneven time intervals. Currently, three alternative methods have been suggested for genetic analyses of growth trajectories: random regression (RR; Meyer, 1998), orthogonal polynomials (OR; Kirkpatrick & Heckman, 1989) and character process (CP; Pletcher & Geyer, 1999; Jaffrezic & Pletcher, 2000). The RR method employs a particular function to model the age-dependent deviation from the population mean due to an individual's genotype, whereas the other two attempt to model the structure of covariances among different ages using Legendre polynomials or stochastic processes. The OP and CP models may be limited in practical data analysis because the pattern with which covariances change with age is not observable. Our model is similar in spirit to, but an improvement on, the RR model built upon genotypic deviations. Our model implements universal growth curve laws and allows for the detection of individual genes (rather than overall gene effects) responsible for growth differentiation.

Second, our model provides a method for analysing patterns of genetic variation that reveal evolutionary changes in growth trajectory. Currently, the study of difference in ontogenetic trajectories between a descendant and its ancestor, namely heterochrony, is an active area for integrating development and evolution to shed light on fundamental biological questions (Rice, 1997). The characterization of genetic factors underlying ontogenetic trajectories from our model helps to unravel the origin of morphological novelties. Third, the model can increase the precision of parameter estimates by reducing the number of unknowns. For example, using our logistic model there are only 8 unknown parameters for a 10-year measurement data set under the backcross design, whereas based on classic models one would estimate 10 overall effects, 10 additive effects, and 55 residual variances and covariances for the same question. In this study, it is assumed that residual variances and covariances among different ages are stationary. This assumption simplifies the mathematical manipulation of the residual variance–covariance matrix (inversion, factorization, etc.), but may deviate from reality. The extension of our analysis to non-stationary variance–covariance structures is possible, as proposed by Nunez-Anton (1997) and Nunez-Anton & Zimmerman (2000) in their structured antedependent models. Other more complicated methods for modelling age-specific covariances (e.g. Davidian & Giltinan, 1995) also deserve exploration.

Our model can be extended to more general situations. In the current model, we assume that growth at time t is decomposed into the effect due to a major gene and residual effect confounded by polygenes and random errors. It is essential to split the

residual effect into the polygene and error components and examine the relative role of a major gene and polygenes in determining growth trajectories. Given the complexity of growth trajectories, it is worthwhile modelling two or more major genes, their epistatic interactions and their dependence upon environments in which organisms are grown. The incorporation of molecular markers into our growth analysis can provide additional insights into the genetic mechanisms underlying growth trajectories.

Finally, our model can be improved from knowledge about growth equations. We have based our analysis on a commonly used growth curve – logistic – in this study. The choice of an appropriate growth equation can be based on the goodness of fit to observed data. An increasing interest now is to derive a growth model from a mechanistic perspective of biological processes (West *et al.*, 2001). Despite possible technical complexities, it is worthwhile integrating our idea for detecting biologically meaningful major genes with the derivations of the mechanistic models specifying developmental and physiological processes. With such integration, we are truly in an interplay between genetics, development, physiology and statistics to push the frontiers of biological research forward.

We are grateful to two anonymous referees for their constructive comments on this manuscript. This work is partially supported by grants from the National Science Foundation to G.C. (DMS9971586) and from the Outstanding Young Investigator Award of the National Natural Science Foundation of China to R.W. (30128017). The publication of this manuscript is approved as journal series R-08642 by the Florida Agricultural Experiment Station.

Appendix A

Three results about the matrix $\mathbf{\Sigma}$ are:

(1) $\mathbf{\Sigma}^{-1}$ is a tridiagonal symmetric matrix. Its diagonal elements are $(1, 1+\rho^2, 1+\rho^2,\ldots, 1+\rho^2, 1+\rho^2, 1)'/[\sigma^2(1-\rho^2)]$ and its second diagonal elements are all $-\rho/[\sigma^2(1-\rho^2)]$.

(2) $|\mathbf{\Sigma}| = [\sigma^2(1-\rho^2)]^{m-1}\sigma^2$.

(3) Let $\boldsymbol{z} = (z_1,\ldots, z_m)' = \mathbf{y}-\mathbf{g}^{(i)}$, $i = 1, 2$ for the backcross or 1, 2, 3 for the F_2, then

$$\boldsymbol{z}'\mathbf{\Sigma}^{-1}\boldsymbol{z} = \frac{\sum_{i=1}^{m-1}(z_i-\rho z_{i+1})^2+(1-\rho^2)z_m^2}{\sigma^2(1-\rho^2)}.$$

Appendix B

The EM algorithm of Section 4 can be thought of as an average of a completed-data EM algorithm as follows: For $i = 1,\ldots, n$, define the augmented variable $\mathbf{Z}_i$ to satisfy

$$\mathbf{Y}_i \,|\, \mathbf{Z}_i = j' \sim f_{j'}(\mathbf{y}),$$
$$P(\mathbf{Z}_i = j') = \gamma_{ij'}.$$

Then (4) is the observed data likelihood, and the complete-data likelihood is given by

$$L(\mathbf{\Omega}\,|\,\mathbf{Y}, \mathbf{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{k} f\mathbf{z}_i(\mathbf{y}_i). \tag{B1}$$

We could now estimate $\mathbf{\Omega}$ using a Gibbs sampler that generates $\mathbf{\Omega}\,|\,\mathbf{Z}$ from (B1), and $\mathbf{Z}\,|\,\mathbf{\Omega}$ from

$$P(\mathbf{Z}_i = j'\,|\,\mathbf{\Omega}) = \frac{\gamma_{ij'}f_{j'}(\mathbf{y}_i)}{\sum_{j'=1}^{k}\gamma_{ij'}f_{j'}(\mathbf{y}_i)}.$$

The posterior models obtained from this Gibbs sampler are the MLEs of the EM algorithm of Section 4 as $\Gamma_{ij} = P(\mathbf{Z}_i = j\,|\,\mathbf{\Omega})$.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Atchley, W. R. (1984). Ontogeny, timing of development, and genetic variance–covariance structure. *American Naturalist* **123**, 519–540.

Beverton, R. J. H. & Holt, S. J. (1957). *On the Dynamics of Exploited Fish Population*. London: Her Majesty's Stationary Office.

Blackman, G. E. (1961). Responses to environmental factors by plants in the vegetative phase. In *Growth in Living Systems* (ed. M. X. Zarrow), pp. 525–556. New York: Basic Books.

Cheverud, J. M., Rutledge, J. J. & Atchley, W. R. (1983). Quantitative genetics of development: genetic correlations among age-specific trait values and the evolution of ontogeny. *Evolution* **37**, 895–905.

Davidian, M. & Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. London: Chapman and Hall.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society Series B* **39**, 1–38.

Dieters, M. J., White, T. L. & Hodge, G. R. (1995). Genetic parameter estimates for volume from full-sib tests of slash pine (*Pinus elliottii*). *Canadian Journal of Forest Research* **25**, 1397–1408.

Hughes, K. A. & Charlesworth, B. (1994). A genetic analysis of senescence in *Drosophila*. *Nature* **367**, 64–66.

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.

Kirkpatrick, M. (1997). Genetic improvement of livestock growth using infinite-dimensional analysis. *Animal Biotechnology* **8**, 55–61.

Kirkpatrick, M. & Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal of Mathematical Biology* **27**, 429–450.

Kirkpatrick, M., Lofsvold, D. & Bulmen, M. (1990). Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics* **124**, 979–993.

Kirkpatrick, M., Hill, W. G. & Thompson, R. (1994). Estimating the covariance structure of traits during growth and aging, illustrated with lactation in dairy cattle. *Genetical Research* **64**, 57–69.

Kremer, A. (1992). Prediction of age–age correlations of total height based on serial correlations between height increments in maritime pine (*Pinus pinaster* Ait.). *Theoretical and Applied Genetics* **85**, 152–158.

Loisel, P., Goffinet, B., Monod, H. & Deoca, G. M. (1994). Detecting a major gene in an F2 population. *Biometrics* **50**, 512–516.

Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates.

Meng, X. L. & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267–278.

Meyer, K. (1998). Estimating covariance functions for longitudinal data using a random regression model. *Genetics*, *Selection*, *Evolution* **30**, 221–240.

Meyer, K. & Hill, W. G. (1997). Estimation of genetic and phenotypic covariance functions for longitudinal or 'repeated' records by restricted maximum likelihood. *Livestock Production Science* **47**, 185–200.

Niklas, K. L. (1994). *Plant Allometry: The Scaling of Form and Process*. Chicago, IL: University of Chicago.

Nunez-Anton, V. (1997). Longitudinal data analysis: nonstationary error structures and antedependent models. *Applied Stochastic Models and Data Analysis* **13**, 279–287.

Nunez-Anton, V. & Zimmerman, D. L. (2000). Modeling nonstationary longitudinal data. *Biometrics* **56**, 699–705.

Nuzhdin, S. V., Pasyukova, E. G., Dilda, C. L., Zeng, Z. B. & Mackay, T. F. C. (1997). Sex-specific quantitative trait loci affecting longevity in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the USA* **94**, 9734–9739.

Pearl, R. (1925). *The Biology of Population Growth*. New York: Knopf.

Pletcher, S. D. & Geyer, C. J. (1999). The genetic analysis of age-dependent traits: modeling the character process. *Genetics* **153**, 825–835.

Pletcher, S. D., Houle, D. & Curtsinger, J. W. (1998). Age-specific properties of spontaneous mutations affecting mortality in *Drosophila melanogaster*. *Genetics* **148**, 287–303.

Pletcher, S. D., Houle, D. & Curtsinger, J. W. (1999). The evolution of age-specific mortality rates in *Drosophila melanogaster*: genetic divergence among unselected lines. *Genetics* **153**, 813–823.

Rice, S. H. (1997). The analysis of ontogenetic trajectories: when a change in size or shape is not heterochrony. *Proceedings of the National Academy of Sciences of the USA* **94**, 907–912.

Richards, F. J. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany* **10**, 290–300.

Roderick, M. L. (2000). On the measurement of growth with applications to the modelling and analysis of plant growth. *Functional Ecology* **14**, 244–251.

Roses, A. D. (2000). Pharmacogenetics and the practice of medicine. *Nature* **405**, 857–865.

Vaughn, T. T., Pletscher, L. S., Peripato, A., King-Ellison, K., Adams, E., Erikson, C. & Cheverud, J. M. (1999). Mapping quantitative trait loci for murine growth: a closer look at genetic architecture. *Genetical Research* **74**, 313–322.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Berlin: Springer.

von Bertalanffy, L. (1957). Quantitative laws in metabolism and growth. *Quarterly Review of Biology* **32**, 217–231.

West, G. B., Brown, J. H. & Enquist, B. J. (2001). A general model for ontogenetic growth. *Nature* **413**, 628–631.

Wu, R. & Stettler, R. F. (1996). The genetic resolution of juvenile canopy structure and function in a three-generation pedigree of *Populus*. *Trees – Structure and Function* **11**, 99–108.

Wu, R. L., Wang, M. X. & Huang, M. R. (1992). Quantitative genetics of yield breeding for *Populus* short rotation culture. I. Dynamics of genetic control and selection models of yield traits. *Canadian Journal of Forest Research* **22**, 175–182.

Wu, R., Ma, C.-X., Zhu, J. & Casella, G. (2002). Mapping epigenetic QTLs affecting developmental trajectories. *Genome* **45**: 28–33.