PLANT METHODS

METHODOLOGY   **Open Access**

Functional mapping of genotype-environment interactions for soybean growth by a semiparametric approach

Qin Li[1†], Zhongwen Huang[2,3†], Meng Xu[4], Chenguang Wang[1], Junyi Gai[3*], Youjun Huang[5], Xiaoming Pang[6,7], Rongling Wu[6,7*]

Abstract

Background: Functional mapping is a powerful approach for mapping quantitative trait loci (QTLs) that control biological processes. Functional mapping incorporates mathematical aspects of growth and development into a general QTL mapping framework and has been recently integrated with composite interval mapping to build up a so-called composite functional mapping model, aimed to separate multiple linked QTLs on the same chromosomal region.

Results: This article reports the principle of using composite functional mapping to estimate the effects of QTL-environment interactions on growth trajectories by parametrically modeling the tested QTL in a marker interval and nonparametrically modeling the markers outside the interval as co-factors. With this new model, we can characterize the dynamic patterns of the genetic effects of QTLs governing growth trajectories, estimate the global effects of the underlying QTLs during the course of growth and development, and test the differentiation in the shapes of QTL genotype-specific growth curves between different environments. By analyzing a real example from a soybean genome project, our model detects several QTLs that cause significant genotype-environment interactions for plant height growth processes.

Conclusions: The model provides a basis for deciphering the genetic architecture of trait expression adjusted to different biotic and abiotic environments for any organism.

Background

In nature, any biological trait of an organism is never isolated from other traits or variables, but rather all of them are integrated under the premise that natural selection has tended to optimize energy absorption and transport within fractal-like distribution networks [1,2]. Such a tendency has led to several universal biological laws; for example, the change of metabolic rate or surface area scales as the 3/4-power of body mass [3,4], organismic growth as a function of age follows a sigmoidal shape [5,6] and metabolic rate increases exponentially with temperature [7]. A number of explicit mathematical equations have been established to describe these biological laws that hold true for all manners of life forms under natural selection.

These mathematical functions have now played an important role in incorporating biological laws into a mapping framework to detect specific quantitative trait loci (QTLs) for growth trajectories and developmental events. R. Wu and group pioneered a series of statistical models, called functional mapping, to map growth and development QTLs through estimating genotype-specific mathematical parameters that define a biological process (reviewed in [8]). In statistics, functional mapping has proven powerful and stable by modeling the patterns of trait development and autocorrelations among different time points measured [9-12]. For some dynamic traits whose expression cannot be mathematically described, nonparametric modeling based on the Legendre polynomial orthogonal has been proposed, thus enhancing the

* Correspondence: sri@njau.edu.cn; rwu@hes.hmc.psu.edu
† Contributed equally
[3]National Center for Soybean Improvement, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China
[6]Center for Computational Biology, Beijing Forestry University, Beijing 100083, China

flexibility of functional mapping [13,14]. Functional mapping is genetically relevant, allowing the formulation and test of numerous biologically meaningful hypotheses about the genetic control of growth [15,16]. Some of the most important hypotheses include the timing of a QTL to switch on or off, the duration of its genetic effect, and the pleiotropic effect of this QTL on different aspects of development.

Functional mapping has been further extended to model the genetic control of more complicated biological problems related to QTL-QTL interactions [15,17] and QTL-environment interactions on growth trajectories [10,11]. [15] provided a general procedure within the functional mapping framework for testing the effects of individual genetic components, including the additive, dominance, additive-additive, additive-dominance, dominance-additive and dominance-dominance, on different stages of development. [10,11] incorporated environmental regimes, such as climates or sexes, into functional mapping to explore the consequence of interactions between the QTL and environments in shaping developmental trajectories. Original functional mapping and its extensions were founded on simple interval mapping, whose utilization may be limited in a situation where there are more than one QTL located on a similar region of a chromosome. When such multiple QTL occur, interval mapping-based approaches may find some "ghost" QTLs which provide significant signals for their existence although they do not exist in reality. [18,19] developed a so-called composite interval mapping to separate multiple linked QTLs by integrating the principle of interval mapping for the testing markers and partial regression analysis of all possible other markers as co-factors. Composite interval mapping has been instrumental for the identification of QTLs that are responsible for different traits [20].

The combination of functional mapping and composite interval mapping can be expected to improve the estimation of multiple QTLs for growth trajectories, but this computationally presents a high challenge in terms of the complexity of estimating growth parameters associated with genotypes at each marker as a co-factor. Fortunately, some of these issues have been solved by R. Wu and group [14]. These authors integrate parametrical interval mapping (aimed to test a hypothesized QTL in a test interval) and Legendre polynomial-based nonparametric longitudinal regression analysis (aimed to control the genome background by choosing a proper set of markers as co-factors), facilitating computation, estimation and tests of functional mapping for multiple QTLs. In this article, we extend [14]'s semiparametric idea to illustrate an analytical strategy for mapping QTLs that affect growth trajectories through their main effects or QTL-environment interaction effects. An instructive procedure is described to test the impacts of each of these components on the timing of development and stages of growth in a time course. We report on the detection of QTLs that affect plant height growth trajectories in soybeans by using the approach developed in this article.

Model
Experimental Design and Regression Model
Consider a recombinant inbred line (RIL) population in which there are two homozygous genotypes for alternative alleles at each locus. Molecular markers are genotyped for each RIL progeny and analyzed to construct a genetic linkage map used to identify quantitative trait loci (QTLs) that affect growth trajectories. The RIL design allows the same genotype to be replicated in time and space. Assume that there are n RILs in a QTL mapping study which are grown in a randomized complete block design in L different environments with multiple replicates per environment. Each of the individuals studied is measured for a growth trait at a series of time points, say T. We will take means at each time point over replicates in each environment to describe growth trajectories for an RIL. Alternatively, we can use individual plants per RIL to conduct the functional mapping of growth trajectories by modeling the spatial structure of a covariance [21].

Suppose there is a quantitative trait locus (QTL) with genotypes QQ (symbolized as 1) and qq (symbolized as 2) that controls the dynamic expression of the trait measured. This QTL is located on somewhere on the genetic linkage map constructed. While a pair of markers is used to map a hypothesized QTL on this test interval, m markers within a given window length (in cM) are chosen as co-factors to associate with growth trajectories through partial regression analysis. Thus, according to the principle of composite interval mapping [18,19], the phenotypic value of the growth trait, $y_{il}(t)$, for individual i measured at time t in environment l, affected by the putative QTL, is expressed as

$$y_{il}(t) = \mu_l(t) + x_i^* \alpha_l(t) + \sum_{k=1}^{m} x_{ik} a_{kl}(t) + e_{il}(t) \quad (1)$$

where and $\mu_l(t)$ and $\alpha_l(t)$ are the environment-specific population mean and additive genetic effect for the QTL at time t, respectively; x_i^* is the indicator variables for individual i that specify the QTL genotypes, which is defined as 1 for genotype QQ and 0 for genotype qq; $\alpha_{kl}(t)$ is the environment-dependent additive effect at time t associated with marker k (except for the interval constructed by the two markers); x_{ik} is the indicator variables that specify the additive effect of marker k for individual i, respectively, which is defined similarly as

x_i^*; and $e_{il}(t)$ is the environment-dependent residual error at time t, normally distributed as $N(0, \sigma_l^2(t))$. The covariance between the residual errors at different time points t_1 and t_2 within environment l is denoted as $\sigma_l(t_1, t_2)$. All the variances and covariances in environment l form a $(T \times T)$ covariance matrix Σ_l.

Likelihood

Let $\mathbf{y}_{il} = (y_{il}(1), ..., y_{il}(T))$ be the phenotypic vector at different time points for individual RIL i at environment l. A mixture model-based likelihood function of the across-replicate henotypic mean vector for the growth trait ($\mathbf{y}$) and marker data ($\mathbf{M}$) can be written, by assuming that different environments are independent, as

$$L(\Omega \mid \mathbf{y}, \mathbf{M}) = \sum_{l=1}^{L} \left[\prod_{i=1}^{n} \sum_{j=1}^{2} w_{j|i} f_{jl}(\mathbf{y}_{il}; \Theta_{jl}, \Psi_l) \right], \quad (2)$$

where $\Omega = (\omega_{j|i}, \Theta_{jl}, \Psi_l; j = 1, 2, l = 1, ..., L)$ is a vector of known parameters, that is, the genomic location of the QTL, QTL genotype-specific parameters, and parameters common to all genotypes. In statistics, the location parameter $\omega_{j|i}$ is the proportion of different mixture normals in equation (2), reflecting the segregation of the QTL in the population, which can be inferred from known marker genotypes at the test interval. For a RIL mapping population, n progeny can be classified into four different groups of two-marker genotypes. In each group, the mixture proportion or frequency of a QTL genotype is RIL-specific and can be expressed as the conditional probability of QTL genotype j for RIL i given its marker genotype [21]. QTL genotype- and environment-specific distribution density, $f_{jl}(\mathbf{y}_{il}; \Theta_{jl}, \Psi_l)$, is assumed to be multivariate normal with expected mean vector $\mathbf{u}_{jl} = (u_{jl}(1), ..., u_{jl}(T))$ for QTL genotype j and covariance matrix Σ_l.

Modeling Time-Dependent Genetic Effects

To solve the likelihood function (2), functional mapping models the genotypic means for an assumed QTL at different time points by a biologically meaningful mathematical function. Since a growth trait is considered, the mathematical function used can be a logistic curve, expressed as

$$g(t) = \frac{a}{1 + be^{-rt}}, \quad (3)$$

where the set of parameters (a, b, r) defines the shape of the genotypic curve. Thus, by estimating the three curve parameters, functional mapping can test the differences in geno-typic means over times. Depending on the nature of experimental materials, many other growth equations can also be used [4]. By estimating QTL

genotype-specific growth curve parameters for different environments (a_{jl}, b_{jl}, r_{jl}) (j = 1, 2; l = 1, ..., L), the time-dependent additive effect of the QTL in environment l is the difference in growth curve between the two genotypes, i.e.

$$\alpha_1(t) = g_{1l}(t) - g_{2l}(t). \quad (4)$$

We will use [14]'s nonparametric approach to model time-dependent effects of the markers outside the test interval used as co-factors for partial regression analysis. This approach based on a Legendre polynomial is flexible for curve fitting [22] and has a closed form solution for the marker effects. The Legendre polynomial of order $\mathcal{S}$ implemented to model the genetic effects of individual markers is generally expressed as $\mathbf{L}_S(t) = (L_0(\tau), L_1(\tau), \cdots, L_S(\tau))$ where

$$L_s(\tau) = \frac{1}{2^s} \sum_{r=0}^{s/2} \frac{(-1)^r (2s - 2r)!}{r!(s-r)!(s-2r)!} \tau^{s-2r}$$

where s = 0, 1, ..., $\mathcal{S}$ and $\tau = -1 + \frac{2(t - \min(t))}{\max(t) - \min(t)}$ with $\min(t)$ and $\max(t)$ being the first and last time point, respectively.

Combining the parametric model for the QTL effect and the Legendre-based nonparametric model for the marker effects together, we rewrite the time-dependent expected means of different QTL genotypes for RIL i in environment l as

$$u_{1l}(t) = \mathbf{L}_S(t)\boldsymbol{\mu}_l' + \alpha_l(t) + \sum_{k=1}^{m} x_{ik} \mathbf{L}_S(t)\mathbf{a}_{kl}'$$

$$\quad (5)$$

$$u_{2l}(t) = \mathbf{L}_S(t)\boldsymbol{\mu}_l' + \sum_{k=1}^{m} x_{ik} \mathbf{L}_S(t)\mathbf{a}_{kl}'$$

where $\boldsymbol{\mu}_l = (\mu_{0l}, \mu_{1l}, \cdots, \mu_{Sl})$ and $\mathbf{a}_{kl} = (a_{k0l}, a_{k1l}, \cdots, a_{kSl})$ are the base population mean vector and the base additive effect vector for marker k as a co-factor, respectively.

Combining the QTL and marker effects in composite functional mapping, time-dependent expected means for different QTL genotypes in environment l are modelled by $\Theta_{jl} = (a_{jl}, b_{jl}, r_{jl}, \boldsymbol{\mu}_l, \{\mathbf{a}_{kl}\}_{k=1}^m)$ if a logistic curve (3) is considered. The model allows the fitting of any other growth curves.

Modeling the Covariance Matrix

The covariance structure of serial measurements can be modeled by a number of approaches. One of the commonly used approaches for structuring the covariance is the first order au-toregressive (AR(1)) model [23]. Its advantage lies in the existence of a general expression for calculating the determinant and inverse of the matrix for any number of time points measured.

In practice, the assumptions of variance stationarity and correlation stationarity, i.e., the residual variance at different time points is the same, expressed as σ^2, and the correlation between two different time points t_1 and t_2 decreases exponentially in ρ with time lag, expressed as $(t_1, t_2) = \rho^{|t_1 - t_2|}$, may not hold. [24] used [25] transform-both-sides (TBS) model to meet the first assumption because heteroscedastic variances can be stable after the data are log-transformed. For the AR(1) model, we only need to use $\Psi_l = (\rho_l, \sigma_1^2)$ to model the structure of environment-specific covariance matrix Σ_l. Some other approaches for modeling the covariance structure include the structured antedepedence model [26,27].

Computing Algorithm

By fixing the position of a hypothesized QTL (in terms of the proportion of recombinant homozygotes) within a test interval, we obtain the maximum likelihood estimates (MLEs) of mean-modeling parameters (Θ_{jl}) and covariance-structuring parameters (Ψ_l) with the EM algorithm. In the E step, the posterior probabilities of each QTL genotype for RIL i in environment l are calculated by

$$\Gamma_{jl|i} = \frac{w_{jl|i} f_j(\mathbf{y}_{il}; \mathbf{\Theta}_{jl}, \mathbf{\Psi}_l)}{\sum_{j'=1}^2 w_{j'|i} f_{j'}(\mathbf{y}_{il}; \mathbf{\Theta}_{j'l}, \mathbf{\Psi}_l)}$$

In the M step, the log-likelihood equations are derived in terms of $\Omega_{jl|i}$ to estimate the growth parameters associated with the QTL effect, basis function parameters associated with the population means and marker effects, and parameters modeling the covariance structure. It is possible to derive the closed forms for the base population means and base additive effects for m markers as cofactors, which are expressed as

$$\begin{bmatrix} \mu_1 \\ a_{1l} \\ \vdots \\ a_{ml} \end{bmatrix} = \begin{bmatrix} A & \cdots & B \\ \vdots & \ddots & \vdots \\ C & \cdots & D \end{bmatrix} \times E$$

where

$$A = \begin{bmatrix} \sum_{i=1}^n \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & \sum_{i=1}^n x_{i1} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & \sum_{i=1}^n z_{i1} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s \\ \sum_{i=1}^n x_{i1} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & \sum_{i=1}^n x_{i1}^2 \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & 0 \\ \sum_{i=1}^n z_{i1} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & 0 & \sum_{i=1}^n z_{i1}^2 \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s \end{bmatrix}$$

$$B = \begin{bmatrix} \sum_{i=1}^n x_{im} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & \sum_{i=1}^n z_{im} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s \\ \sum_{i=1}^n x_{i1} x_{im} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & \sum_{i=1}^n x_{i1} z_{im} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s \\ \sum_{i=1}^n z_{i1} x_{im} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & \sum_{i=1}^n z_{i1} z_{im} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s \end{bmatrix}$$

$$C = \begin{bmatrix} \sum_{i=1}^n x_{im} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & \sum_{i=1}^n x_{i1} x_{im} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & \sum_{i=1}^n z_{i1} x_{im} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s \\ \sum_{i=1}^n z_{im} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & \sum_{i=1}^n x_{i1} z_{im} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & \sum_{i=1}^n z_{i1} z_{im} \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s \end{bmatrix}$$

$$D = \begin{bmatrix} \sum_{i=1}^n x_{im}^2 \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s & 0 \\ 0 & \sum_{i=1}^n z_{im}^2 \mathbf{L}_s' \Sigma_l^{-1} \mathbf{L}_s \end{bmatrix}$$

$$E = \begin{bmatrix} \sum_{i=1}^n x_{i1} \mathbf{L}_s' \Sigma_i^{-1} \mathbf{L}_s \\ \sum_{i=1}^n x_{i1} \mathbf{L}_s' \Sigma_i^{-1} [\mathbf{y}_i - (\Gamma_{1l|i} - \Gamma_{2l|i}) \mathbf{L}_s \alpha] \\ \sum_{i=1}^n z_{i1} \mathbf{L}_s' \Sigma_i^{-1} [\mathbf{y}_i - (\Gamma_{1l|i} - \Gamma_{2l|i}) \mathbf{L}_s \alpha] \\ \vdots \\ \sum_{i=1}^n x_{im} \mathbf{L}_s' \Sigma_i^{-1} [\mathbf{y}_i - (\Gamma_{1l|i} - \Gamma_{2l|i}) \mathbf{L}_s \alpha] \\ \sum_{i=1}^n z_{im} \mathbf{L}_s' \Sigma_i^{-1} [\mathbf{y}_i - (\Gamma_{1l|i} - \Gamma_{2l|i}) \mathbf{L}_s \alpha] \end{bmatrix}$$

with $\mathbf{L}_S = (L_S'(1), \cdots, L_S'(T))$ and $\alpha_l = (\alpha_l(1), ..., \alpha_l(T))$ both for environment l.

To estimate the MLEs of the parameters that model the time-dependent QTL effects and covariance matrix,

we implement the simplex or Newton-Raphson algorithm in the estimation process with the EM algorithm [28,29]. As shown in [9], the closed forms for estimating the determinant and inverse of the AR(1) covariance matrix can be derived, whose implementation will increase the computational efficiency of functional mapping.

A numerical derivative approach was derived to calculate the sampling errors for the MLEs of the parameters contained within the mixture model (2). The partial derivative of a given function with respect to some parameters x and z is given by

$$\frac{\partial^2 \ell}{\partial x \partial z} = \{[\ell(x+\delta_1, z+\delta_2) - f(x+\delta_1, z-\delta_2)] - [f(x-\delta_1, z+\delta_2) - f(x-\delta_1, z-\delta_2)]\} / (\delta_1, \delta_2)$$

If x and z are the same parameters, the above formula simplifies to

$$\frac{\partial^2 \ell}{\partial x^2} = \frac{\ell(x+\delta) - 2f(x-\delta) + f(x)}{\delta^2}$$

where ℓ is the log-likelihood function and x and z are any parameters of interest in the current model. Taking the inverse of the negative values of the partial derivative matrix, the estimate of the asymptotic sample covariance matrix can be obtained.

Model Selection

To obtain the best fit of the data, the optimal number of markers involved in the partial regression analysis of composite functional mapping and optimal order of the Legendre polynomial to model the marker effects should be determined. We used the Bayesian information criterion (BIC) [30] as the model selection criterion of the optimal marker number and polynomial order. The BIC is defined as

$$\mathrm{BIC} = -2\ln L(\{\hat{\mathbf{\Theta}}_{jl}, \hat{\mathbf{\Psi}}_l\}_{j=1;l=1}^{2;L} \mid \mathcal{S})$$
$$+ \dimension(\{\mathbf{\Theta}_{jl}, \mathbf{\Psi}_l\}_{j=1;l=1}^{2;L} \mid \mathcal{S})\ln(n).$$

where $\{\hat{\mathbf{\Theta}}_{jl}, \hat{\mathbf{\Psi}}_l\}_{j=1;l=1}^{2;L}$ are the MLEs of parameters under the Legendre polynomial of order $\mathcal{S}$ dimension, $\{\mathbf{\Theta}_{jl}, \mathbf{\Psi}_l\}_{j=1;l=1}^{2;L}$ represents the number of independent parameters under order $\mathcal{S}$, and n is the total number of observation at a particular time point. The optimal model is one that displays the minimum BIC value. Note that the model selection procedure also incorporates the choice of the best number of markers (flanking a given test interval) as co-factors in composite functional mapping.

Hypothesis Tests
Existence of a QTL

Within the framework of functional mapping, a number of biologically meaningful hypotheses can be tested [15]. To understand the genetic architecture of a growth trait, we need to first test the existence of a QTL that affects the dynamic process and shape of the trait. The following hypotheses are formulated to test the genetic control over the entire dynamic process of a trait:

$$\begin{aligned}
H_0 &: (a_{11}, b_{11}, r_{11}) = (a_{21}, b_{21}, r_{21}) \\
&\equiv (a_1, b_1, r_1), \dots, \\
&(a_{1L}, b_{1L}, r_{1L}) = (a_{2L}, b_{2L}, r_{2L}) \\
&\equiv (a_L, b_L, r_L) \\
H_1 &: \text{at least one of the above} \\
&\quad \text{equalities does not hold.}
\end{aligned} \tag{6}$$

The H_0 states that there are no QTL affecting growth curves (the reduced model), whereas the H_1 proposes that such QTL do exist (the full model). The test statistic for the hypotheses (6) is calculated as the log-likelihood ratio of the reduced to the full model:

$$\begin{aligned}
\mathrm{LR} = -2[&\ln L(\{\bar{\mathbf{\Theta}}_{jl}, \bar{\mathbf{\Psi}}_l\}_{j=1;l=1}^{2;L} \mid \mathbf{y}, \mathbf{M}_{-2}) \\
&- \ln L(\{\hat{\omega}_{j|i}, \hat{\mathbf{\Theta}}_{jl}, \hat{\mathbf{\Psi}}_l\}_{j=1;l=1}^{2;L} \mid \mathbf{y}, \mathbf{M})],
\end{aligned} \tag{7}$$

where the tildes and hats denote the MLEs of the unknown parameters under the H_0 and H_1, respectively, and $\mathbf{M}_{-2}$ is the marker information excluding the two tested markers. The critical value of the LR test statistic can be determined by estimating its behavior under the null hypothesis for a whole genome. An empirical approach based on permutation tests by destroying the relationships between the phenotypic values and tested marker interval genotypes [31-33] is usually used to determine the critical threshold of the LR for interval mapping. But this approach cannot be directly used for composite interval mapping in which additional markers (excluding the two tested markers) serve as co-factors to be associated with the phenotypic values. [19] proposed a simulation approach to examine the distribution of the LR values under the null hypothesis. The phenotypic values simulated under the null hypothesis should reflect the effects of the markers as co-factors. This can be done by assuming that the time-dependent phenotypic values in environment ℓ follow a multivariate normal distribution with mean vector

$$\mathbf{L}_s(t)\hat{\boldsymbol{\mu}}_l' + \sum_{k=1}^{m} x_{ik}\mathbf{L}_s(t)\hat{\mathbf{a}}_{kl}'$$

and the covariance matrix with the AR(1) structure. The threshold for the tested interval is estimated as the 5% percentile of the LR values from 1000 simulation replicates. A genome-wide critical threshold is determined by scanning through the entire linkage map, although this process is computationally extensive.

Pleiotropic Effect of a QTL

If a significant QTL is found, then we can test whether this QTL has a pleiotropic effect on growth trajectories in any two different environments l_1 and l_2. This test is formulated as

$$
\begin{aligned}
H_0 &: (a_{1l_1}, b_{1l_1}, r_{1l_1}) = (a_{2l_1}, b_{2l_1}, r_{2l_1}) \\
&\equiv (a_{l_1}, b_{l_1}, r_{l_1}) \\
H_1 &: \text{at least one of the above} \\
&\quad \text{equalities does not hold}
\end{aligned} \tag{8}
$$

for environment l_1, and

$$
\begin{aligned}
H_0 &: (a_{1l_1}, b_{1l_1}, r_{1l_1}) = (a_{2l_2}, b_{2l_2}, r_{2l_2}) \\
&\equiv (a_{l_2}, b_{l_2}, r_{l_2}) \\
H_1 &: \text{at least one of the above} \\
&\quad \text{equalities does not hold}
\end{aligned} \tag{9}
$$

for environment l_2. If the null hypotheses above (8 and 9) are both rejected, this implies that the significant QTL detected affects pleiotropically growth trajectories in the two environments. Otherwise, this QTL is operational only in one environment, which causes genotype by environment interactions for growth trajectories. The thresholds for each of the hypotheses (8 and 9) corresponding to two different environments can be determined using the simulation approach as described above.

QTL by Environment Interaction

In practice, although the QTL is pleiotropic, its effect on growth trajectories may depend on the environment due to QTL by environment interactions. This can be tested for any two environments l_1 and l_2 using

$$
\begin{aligned}
H_0 &: \frac{a_{1l_1}}{1+b_{1l_1}e^{r_{1l_1}t}} - \frac{a_{2l_1}}{1+b_{2l_1}e^{r_{2l_1}t}} \\
&\equiv \frac{a_{1l_2}}{1+b_{1l_2}e^{r_{1l_2}t}} - \frac{a_{2l_2}}{1+b_{2l_2}e^{r_{2l_2}t}} \\
H_1 &: \frac{a_{1l_1}}{1+b_{1l_1}e^{r_{1l_1}t}} - \frac{a_{2l_1}}{1+b_{2l_1}e^{r_{2l_1}t}} \\
&\neq \frac{a_{1l_2}}{1+b_{1l_2}e^{r_{1l_2}t}} - \frac{a_{2l_2}}{1+b_{2l_2}e^{r_{2l_2}t}}
\end{aligned} \tag{10}
$$

The rejection of the above null hypothesis means that a significant genotype by environment interaction exists due to allelic sensitivity to a varying environment. The log-likelihood ratio test statistics for hypothesis (10) can be determined from simulation studies.

Results

The model described above was used to analyze a real example from a soybean genome project at Nanjing Agricultural University, China. Two original inbred lines of soybean, Kefeng No. 1 and Nannong 1138-2, as parents were crossed to generate an F_1 population which was selfed for 7 generations to produce an RIL population composed of two groups of homozygous genotypes each containing two identical alleles from a different parental line. Let 1 and 2 denote the homozygotes derived from the Kefeng No. 1 alleles and Nannong 1138-2 alleles, respectively. A total of 184 RILs were genotyped for 488 molecular markers (restricted fragment length polymorphisms, simple sequence repeats and amplified fragment length polymorphsim) that construct a linkage map with 25 linkage groups covering 4,151.2 cM of the soybean genome [34].

The RILs were planted in a simple lattice design with multiple replicates randomly grown in a plot at the Jiangpu Agricultural Experiment Station of Nanjing Agricultural University, China. The plants were measured for their plant height growth for six to eight times with the first time at the 28th day after emergence and successive seven times every 10 days thereafter. The same study was repeated for year 2005 and 2006. Thus, different years are viewed as two different environments.

By taking the means for the same RIL over replicates, plant heights are plotted against time in two different years (Fig. 1). The variation among RILs tends to increase with time for raw data (Fig. 1, upper panel), but seems to be constant over time when the data are log-transformed (Fig. 1, lower panel). Thus, we can use AR(1)-based functional mapping to analyze the transformed growth data. Through incorporating the TBS model [24], biological meanings of the growth parameters can be preserved. From Fig. 1, it is seen that plant height growth in each year can well be fit by a logistic curve (3) containing parameters (a, b, r).

Composite functional mapping will map QTL × year interactions for plant height growth trajectories of soybeans planted in two different years. However, because this approach is computationally very expensive, we will first run traditional functional mapping by [9] and [10,11], aimed to scan growth QTLs throughout the genome. If multiple peaks on the same chromosome are detected from functional mapping, composite functional mapping will be performed to resolve the existence of
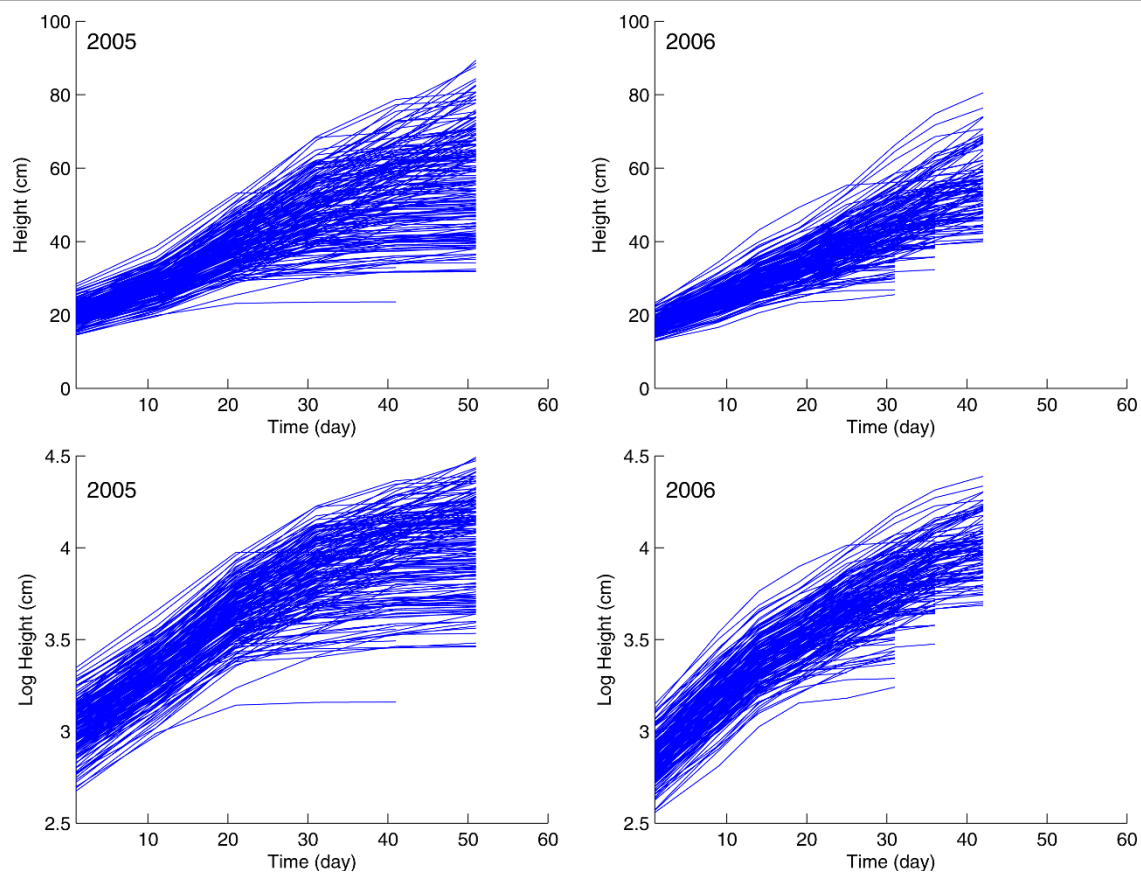
Figure 1 Growth curves for plant heights in a recombinant inbred lines of soybeans planted in years 2005 and 2006.

multiple possible linked QTLs. Log-likelihood ratios from functional mapping were plotted against 25 linkage groups (Fig. 2), from which several peaks were detected on chromosomes 3, 6, and 24. As compared with the critical threshold determined from 100 permutation tests, these peaks were thought to harbor significant QTLs. Yet, these linkage groups have multiple peaks, which is thus subjected to a reanalysis by composite functional mapping.

In composite functional mapping, the growth equation was incorporated to specify the QTL effect, whereas the Legendre polynomial was used to model the marker effects by choosing different numbers of markers as cofactors within a changing window length (in cM) of the test interval of the QTL. The best number of co-factors is determined under different orders of the Legendre polynomial using the BIC criterion (Table 1). The optimal order is found to be 5 for the soybean data. Under these optimal circumstances, only one QTL on each linkage group (3, 6, and 24) was detected by composite functional mapping. The estimates of the locations of these QTLs are given in Table 2, along with the

estimates of growth curve parameters for different genotypes at each of the QTL detected (Table 2). It seems that these estimates are reasonably precise because their standard errors estimates are small.

Figure 3 illustrates the growth curves of QTL genotypes in years 2005 and 2006 drawn with the estimates of curve parameters in Table 2. All the QTLs detected display increasing additive effects with time. At the QTLs on linkage groups 3 and 6, parent Kefeng No. 1 contributes favorable alleles to increasing plant growth, whereas such favorable alleles are derived from parent Nannong 1138-2 at the QTL on linkage group 24. After significant QTLs were detected, hypothesis tests (8 and 9) were used to investigate whether these QTLs trigger pleiotropic effects on growth trajectories for soybeans in different years. At each of the detected QTLs, the null hypotheses of both (8 and 9) are rejected, which suggests that they are all pleiotropic QTLs affecting growth trajectories in both years. A further test with hypothesis (10) suggests that, although two of these QTLs (on chromosomes 3 and 6) are pleiotropic, the magnitude and temporal pattern of their additive genetic effects on
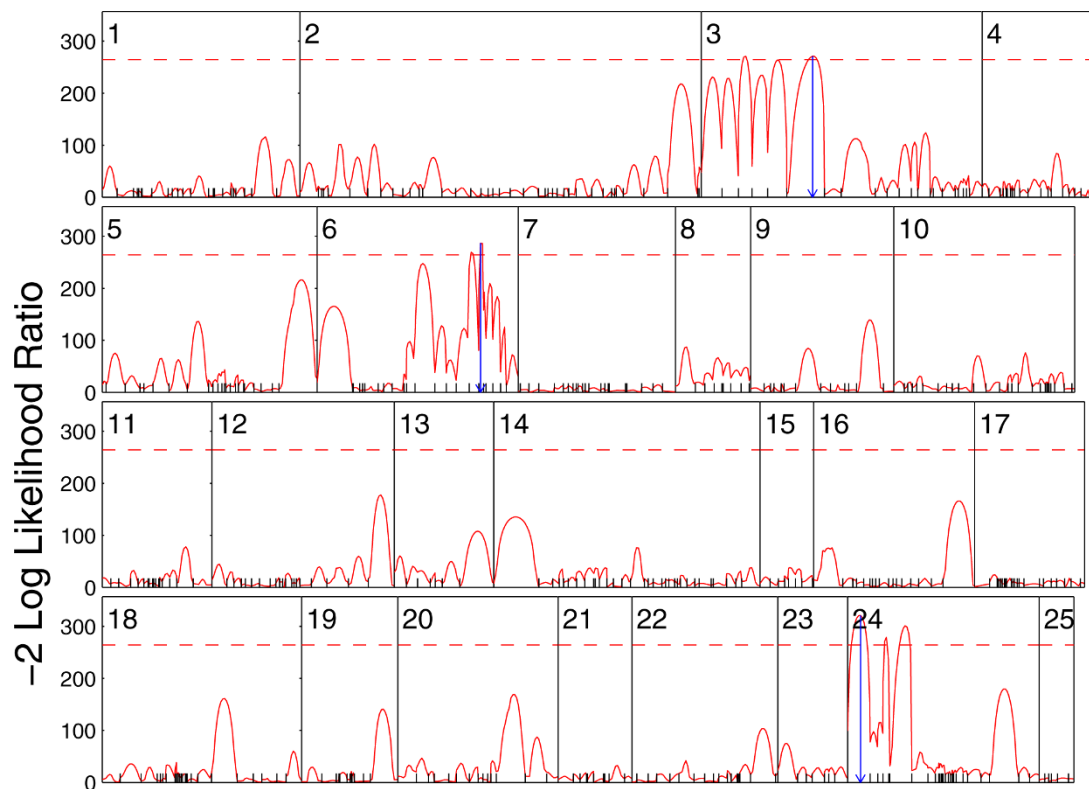
Figure 2 The profile of the log-likelihood ratios (*LR*) between the full model (there is a QTL) and reduced (there is no QTL) model for plant height growth trajectories across the genome in soybeans planted in years 2005 and 2006. The genomic position corresponding to the peak of the curve is the maximum likelihood estimate of the QTL localization. The vertical broken lines indicate the positions of markers on this chromosome shown beneath. The map distances (in centiMorgan) between two markers are calculated using the Haldane mapping function. The thresholds for acclaiming the genome-wide existence of a QTL are obtained from 100 permutation tests.

growth traits vary from year to year (see Fig. 3). Such interactions between the QTLs and years determine year-specific differences in growth trajectories of plant heights in soybeans. The QTL on chromosome 24 affect plant growth trajectories in a consistent pattern over years, displaying no genotype × year interactions.

By estimating and testing the inflection point of a logistic curve using ln b/r, we further investigate how the detected QTLs exert significant impacts on the timing of maximum growth rate (Fig. 3). The inflection point

of growth curves in 2005 displays a difference of three days between the two genotypes (32 vs. 29) at the QTL on chromosome 3, whereas this difference in 2006 is as large as 21 days (45 vs. 24). For the QTL on chromosome 6, we detect a similar pattern for the difference in the inflection point, i.e., 4 days (31 vs. 27) in 2005 and 25 days (41 vs. 16) in 2006. Yet, unlike these two QTLs, there is no year-specific difference in the inflection point for the QTL on chromosome 24, which does not exhibit a significant effect on this developmental characteristic in both years. Although favorable alleles that increase plant growth are contributed by parent Kefeng No. 1 for QTLs detected on chromosomes 3 and 6, parent Nannong 1138-2 contributes such a favorable allele for the QTL on chromosome 24 (Fig. 3).

Table 1 BIC values under different models of composite functional mapping by choosing markers within different window lengths (in cM) as co-factors. Different orders of the Legendre polynomial are considered under each model.

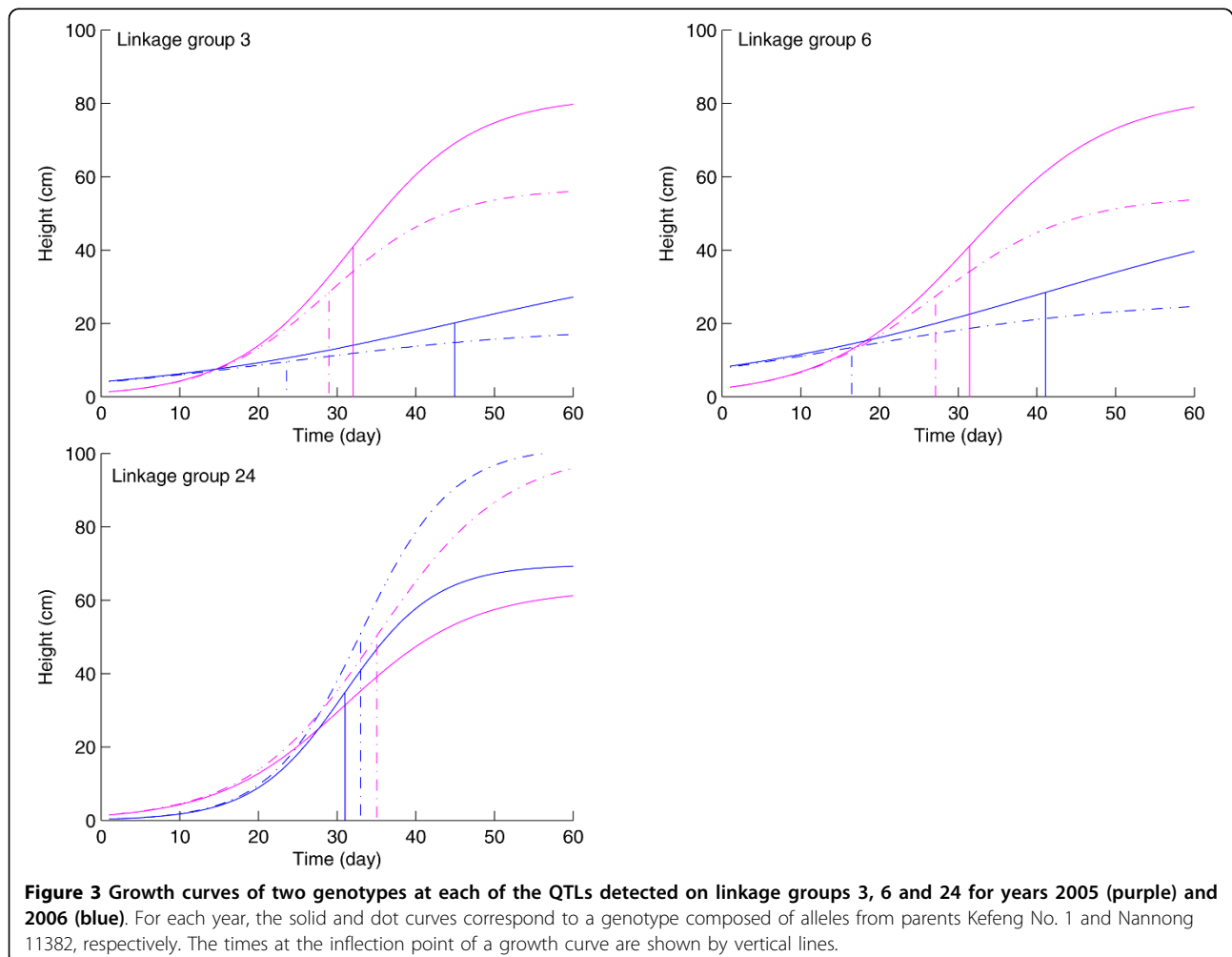| Chromosome	Window Width (cM)	Order of Legendre Polynomials			
		2	3	4	5
3	25	-5833.70	-5815.68	-5925.53	-5941.03
6	10	-7264.87	-7286.02	-7434.39	-7441.61
24	5	-7056.95	-7062.54	-7211.4	-7292.07

Discussion

Genetic mapping has proven to be a powerful approach for map individual genes or quantitative trait loci (QTLs) that control a quantitatively inherited trait [35]. However, this approach would not gain too much insight into the genetic control mechanisms for

Table 2 The MLEs of genotypic curve parameters at the two QTLs detected on chromosomes 3, 6, and 24 and sampling errors (in parentheses) of the estimates by composite functional mapping under the optimal order of the Legendre polynomial and the optimal number of co-factors (see Table 1) for plant height growth trajectories of soybeans in two different years (2005 and 2006).

			QTL Genotype					
			QQ			qq		
Chromosome	Marker Interval	Year	*a*	*b*	*r*	*a*	*b*	*r*
3	GMKF104b-GMKF177	2005	81.790	66.583	0.131	56.889	47.148	0.133
			(4.013)	(5.148)	(0.006)	(3.283)	(2.520)	(0.007)
		2006	40.314	8.799	0.048	19.148	3.865	0.057
			(3.142)	(1.538)	(0.006)	(3.398)	(1.081)	(0.004)
6	A748V-A397I	2005	82.223	33.853	0.112	54.908	22.763	0.115
			(10.290)	(1.364)	(0.001)	(4.534)	(1.422)	(0.002)
		2006	56.924	6.072	0.044	26.879	2.460	0.055
			(3.711)	(0.417)	(0.001)	(2.540)	(0.193)	(0.003)
24	sat_231- LE23T	2005	62.894	46.764	0.124	100.571	72.397	0.122
			(7.006)	(9.020)	(0.006)	(10.245)	(13.981)	(0.006)
		2006	69.695	223.809	0.175	101.781	308.94	0.174
			(9.148)	(52.231)	(0.010)	(12.711)	(69.850)	(0.010)

Figure 3 Growth curves of two genotypes at each of the QTLs detected on linkage groups 3, 6 and 24 for years 2005 (purple) and 2006 (blue). For each year, the solid and dot curves correspond to a genotype composed of alleles from parents Kefeng No. 1 and Nannong 11382, respectively. The times at the inflection point of a growth curve are shown by vertical lines.

phenotypic variation if some statistical and biological issues related to the approach are not resolved. Zeng and others are among the first who systematically investigate the effects of multiple linked QTLs on the power of genetic mapping, and further proposed so-called composite interval mapping to separate linked QTLs through using the markers outside the test interval to make the background control of the genome [18,19,36]. Many other approaches that take into account a specific statistical, computational or genetic issue of genetic mapping have been developed [37-43]. All these approaches have been instrumental for the characterization of QTLs that control quantitative traits of interest to agriculture, biology and health sciences [44,20,45,46].

More recently, genetic mapping has been integrated with some fundamental biological principles, aimed to generate biologically more meaningful discoveries related to trait formation and development. One of the most important products for this integration is the formulation of a series of statistical models, called functional mapping [9,15,16,24,8]. Functional mapping capitalizes on the mathematical aspects of biological processes to model the temporal pattern of genetic effects exerted by a QTL in time course. It offers tremendous advantages in the generation of testable biological hypotheses and synthesis of different disciplines for a more comprehensive understanding of biology. Functional mapping have now been extended to explore the roles gene-gene, gene-environment, gene-sex interactions play in directing growth trajectories of a complex trait [15,10,11]. Combined with composite interval mapping, functional mapping has been shown to have more power to separate linked QTLs on the same chromosome [14]. This composite functional mapping uses a parametric approach to model the temporal effect of the QTL effect and a nonparametric approach based on the Lengendre polynomial [22,47,13] to model the temporal effects of different markers as co-factors. Parametric modeling preserves the biological relevance of the original functional mapping, whereas nonparametric modeling increases the flexibility of functional mapping and its computational efficiency.

In this report, we incorporate composite functional mapping to map QTLs that interact with environments to regulate the process of trait development. While traditional functional mapping detected significant signals for the existence of QTLs for plant height growth trajectories in a recombinant inbred line (RIL) population of soybeans, the new composite functional mapping shows great power to separate multiple linked QTLs on the same chromosome. In total, we detected three significant QTLs on chromosomes 3, 6 and 24 that affect height growth in different years. These QTLs were also found by an allometric model [29]. By integrating environmental factors, the new model detected these QTLs to be operational in both years. However, the magnitudes of their effects vary between different years, showing significant QTL by year interactions. It seems that plant height growth is under strong genetic control. In the functional mapping of rice, [27] found several significant QTLs for plant heights in two contrasting climates although the expression of these QTLs is not environment-dependent.

Currently, there are three different hypotheses proposed to explain the mechanisms of genotype × environment interactions-heterozygosity, allelic sensitivity, and gene regulation [48,49]. The new model provides a genome-wide search for QTLs that act in terms of the second hypothesis, i.e., the environment-dependent change of a trait is caused by differentiate expression of a QTL in different environments. However, it is possible to modify the model to test the other hypotheses. For example, the gene regulation hypothesis can be tested by assuming different but epistatically interacting QTLs for mean growth curves and individual growth curves across different environments [50,51]. Our approach combines powerful statistics and molecular genetics with developmental and ecological mechanisms underlying biological features, relationships and processes to shed light on the genetic basis of complex traits. Such a *mechanistic* strategy will be powerful to address fundamental questions about plant development and plastic response to changing environments.

Author details

[1]Department of Statistics, University of Florida, Gainesville, FL 32611 USA. [2]Department of Agronomy, Henan Institute of Science and Technology, Xinxiang, Henan 453003, China. [3]National Center for Soybean Improvement, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China. [4]Key Laboratory of Forest Genetics and Tree Engineering, Nanjing Forestry University, Nanjing, Jiangsu 210037, China. [5]School of Forestry and Biotechnology, Zhejiang Forestry University, Lin'an, Zhejiang 311300, China. [6]Center for Computational Biology, Beijing Forestry University, Beijing 100083, China. [7]National Engineering Laboratory for Tree Breeding, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Beijing Forestry University, Beijing 100083, China.

Authors' contributions

QL, CW, and YH formulated the model and performed simulation studies and data analysis, ZH and JG designed and performed the experimental work. MX and XP provided biological inputs. JG and RW coordinated the work. RW conceived the model of functional mapping and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. West GB, Brown JH, Enquist BJ: **A general model for the origin of allometric scaling laws in biology.** *Science* 1997, **276**:122-126.
2. West GB, Brown JH, Enquist BJ: **The fourth dimension of life: Fractal geometry and allometric scaling of organisms.** *Science* 1999, **284**:1677-1679.
3. Kleiber M: **Body size and metabolism.** *Hilgardia* 1932, **6**:315-353.
4. Niklas KJ: **Plant Allometry: The Scaling of Form and Process.** University of Chicago Press, Chicago, IL 1994.
5. Bertalanffy von L: **Quantitative laws for metabolism and growth.** *Quart Rev Biol* 1957, **32**:217-231.
6. West GB, Brown JH, Enquist BJ: **A general model for ontogenetic growth.** *Nature* 2001, **413**:628-631.
7. Gillooly JF, Brown JH, West GB, Savage VM, Charnov EL: **Effects of Size and Temperature on Metabolic Rate.** *Science* 2001, **293**:2248-2251.
8. Wu RL, Lin M: **Functional mapping –; how to map and study the genetic architecture of dynamic complex traits.** *Nat Rev Genet* 2006, **7**:229-237.
9. Ma CX, Casella G, Wu RL: **Functional mapping of quantitative trait loci underlying the character process: A theoretical framework.** *Genetics* 2002, **161**:1751-62.
10. Zhao W, Ma CX, Cheverud JM, Wu RL: **A unifying statistical model for QTL mapping of genotype-sex interaction for developmental trajectories.** *Physiol Genom* 2004, **19**:218-227.
11. Zhao W, Zhu J, Gallo-Meagher M, Wu RL: **A unified statistical model for functional mapping of genotype × environment interactions for ontogenetic development.** *Genetics* 2004, **168**:1751-1762.
12. Yang RQ, Tian Q, Xu SZ: **Mapping quantitative trait loci for longitudinal traits in line crosses.** *Genetics* 2006, **173**:2339-2356.
13. Lin M, Wu RL: **A joint model for nonparametric functional mapping of longitudinal trajectories and time-to-events.** *BMC Bioinformatics* 2006, **7**:138.
14. Yang RQ, Gao J, Wang X, Zhang J, Zeng ZB, Wu RL: **A semiparametric model for composite functional mapping of dynamic quantitative traits.** *Genetics* 2007, **177**:1859-1870.
15. Wu RL, Ma CX, Lin M, Casella G: **A general framework for analyzing the genetic architecture of developmental characteristics.** *Genetics* 2004, **166**:1541-1551.
16. Wu RL, Ma CX, Lin M, Wang ZH, Casella G: **Functional mapping of growth QTL using a transform-both-sides logistic model.** *Biometrics* 2004, **60**:729-738.
17. Lin M, Hou W, Li HY, Johnson JA, Wu RL: **Modeling interactive quantitative trait nucleotides for drug response.** *Bioinformatics* 2007, **23**:1251-1257.
18. Zeng ZB: **Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci.** *Proc Natl Acad Sci USA* 1993, **90**:10972-10976.
19. Zeng ZB: **Precision mapping of quantitative trait loci.** *Genetics* 1994, **136**:1457-1468.
20. Mackay TFC: **Quantitative trait loci in Drosophila.** *Nat Rev Genet* 2001, **2**:11-20.
21. Wu RL, Ma CX, Casella G: **Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL.** Springer-Verlag, New York 2007.
22. Kirkpatrick M, Heckman N: **A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters.** *J Math Biol* 1989, **27**:429-450.
23. Verbeke G, Molenberghs G: **Linear Mixed Models for Longitudinal Data.** New York: Springer-Verlag 2000.
24. Wu RL, Wang ZH, Zhao W, Cheverud JM: **A mechanistic model for genetic machinery of ontogenetic growth.** *Genetics* 2004, **168**:2383-2394.
25. Carroll RJ, Ruppert D: **Power-transformations when fitting theoretical models to data.** *J Am Stat Assoc* 1984, **79**:321-328.
26. Zimmerman DL, Núñez-Antón V: **Parametric modeling of growth curve data: An overview (with discussion).** *Test* 2001, **10**:1-73.
27. Zhao W, Chen YQ, Casella G, Cheverud JM, Wu RL: **A non-stationary model for functional mapping of longitudinal quantitative traits.** *Bioinformatics* 2005, **21**:2469-2477.
28. Zhao W, Wu RL, Ma CX, Casella G: **A fast algorithm for functional mapping of complex traits.** *Genetics* 2004, **167**:2133-2137.
29. Li HY, Huang ZW, Wu S, Gai JY, Zeng YR, Wu RL: **A conceptual framework to mapping quantitative trait loci governing ontogenetic allometry.** *PLoS ONE* 2007, **2(8)**:e1245, doi:10.1371/journal.pone.0001245.
30. Schwarz G: **Estimating the dimension of a model.** *Ann Stat* 1978, **6**:461-464.
31. Doerge RW, Churchill GA: **Permutation tests for multiple loci affecting a quantitative character.** *Genetics* 1996, **142**:285-294.
32. Zou F, Fine JP, Hu J, Lin DY: **An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci.** *Genetics* 2004, **168**:2307-2316.
33. Jin C, Fine JP, Yandell BS: **A unified semiparametric framework for QTL analyses, with application to spike phenotypes.** *J Am Stat Assoc* 2007, **102**:56-67.
34. Zhang WK, Wang YJ, Luo GZ, Zhang JS, He CY, Wu XL, Gai JY, Chen SY: **QTL mapping of ten agronomic traits on the soybean (Glycine max L. Merr.) genetic map and their association with EST markers.** *Theor Appl Genet* 2004, **108**:1131-1139.
35. Lander ES, Botstein D: **Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps.** *Genetics* 1989, **121**:185-199.
36. Jansen RC, Stam P: **High resolution mapping of quantitative traits into multiple loci via interval mapping.** *Genetics* 1994, **136**:1447-1455.
37. Kao CH, Zeng ZB: **General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm.** *Biometrics* 1997, **53**:653-665.
38. Kao CH, Zeng ZB: **Modeling epistasis of quantitative trait loci using Cocker-ham's model.** *Genetics* 2002, **160**:1243-1261.
39. Kao CH, Zeng ZB, Teasdale RD: **Multiple interval mapping for quantitative trait loci.** *Genetics* 1999, **152**:1203-1216.
40. Sen S, Churchill GA: **A statistical framework for quantitative trait mapping.** *Genetics* 2001, **159**:371-387.
41. Broman KW, Speed TP: **A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion).** *J Roy Stat Soc B* 2002, **64**:641-656.
42. Liu T, Todhunter RJ, Wu S, Hou W, Mateescu R, Zhang ZW, Burton-Wurster NI, Acland GM, Lust G, Wu RL: **A random model for mapping imprinted quantitative trait loci in a structured pedigree: An implication for mapping canine hip dysplasia.** *Genomics* 2007, **90**:276-284.
43. Wu J, Zhang B, Cui Y, Zhao W, Xu L, Huang M, Zeng Y, Zhu J, Wu RL: **Genetic mapping of developmental instability: Design, model and algorithm.** *Genetics* 2007, **176**:1187-1196.
44. Zeng ZB, Liu J, Stam LF, Kao CH, Mercer JM, Laurie CC: **Genetic architecture of a morphological shape difference between two Drosophila species.** *Genetics* 2000, **154**:299-310.
45. Frary A, Nesbitt TC, Frary A, Grandillo S, van der Knaap E, Cong B, Liu JP, Meller J, Elber R, Alpert KB, Tanksley SD: **fw2.2: A quantitative trait locus key to the evolution of tomato fruit size.** *Science* 2000, **289**:85-88.
46. Li CB, Zhou AL, Sang T: **Rice domestication by reducing shattering.** *Science* 2006, **311**:1936-1939.
47. Cui YH, Zhu J, Wu RL: **Functional mapping for genetic control of programmed cell death.** *Physiol Genom* 2006, **25**:458-469.
48. Via S, Gomulkievicz R, de Jong G, Scheiner SM, *et al*: **Adaptive phenotypic plasticity: Consensus and controversy.** *Trend Ecol Evol* 1995, **10**:212-217.
49. Via S, Lande R: **Genotype-environment interactions and the evolution of phe-notypic plasticity.** *Evolution* 1985, **39**:505-522.
50. Scheiner SM, Lyman RF: **The Genetics of Phenotypic Plasticity. 1. Heritability.** *J Evol Biol* 1989, **2**:95-107.
51. Weber SL, Scheiner SM: **The genetics of phenotypic plasticity. 4. Chromosomal localization.** *J Evol Biol* 1992, **5**:109-120.