Systems Mapping of Complex Traits Using Kinetic Models

Jiguo Cao[1,2*], Zhongwen Huang[3], Junyi Gai[4], C. Eduardo Vallejos[5] and Rongling Wu[2,6*]

[1] *Department of Statistics & Actuarial Science, Simon Fraser University, Burnaby, B.C. Canada V5A 1S6*

[2] *Center for Computational Biology, Beijing Forestry University, Beijing 100083, China*

[3] *Department of Agronomy, Henan Institute of Science and Technology, Xinxiang 453003, China*

[4] *National Center for Soybean Improvement, National Key Laboratory for Crop Genetics and Germplasm Enhancement, Soybean Research Institute, Nanjing Agricultural University, Nanjing 210095, China*

[5] *Department of Horticultural Sciences, University of Florida, Gainesville, FA 32611, USA*

[6] *Center for Statistical Genetics, Pennsylvania State University, Hershey, PA 17033, USA*

*Corresponding authors: jiguo_cao@sfu.ca, rwu@hes.hmc.psu.edu

SUMMARY: Every phenotypic trait can be viewed as a "system" in which a group of independent but interconnected components function together to comprise a unified whole. Once a system's components and interactions have been delineated according to biological principles, therefore, we can change the phenotype of the system by altering those components and interactions that are functionally relevant to the system. Here, by incorporating a system of differential equations that quantifies how alterations of different components lead to the global change of trait development, we devise a computational model for genetic mapping of complex traits, equipped with a capacity to characterize genes and genetic interactions involved in the regulation of biological circuits. The model provides a quantitative and testable platform for assessing the in-

terplay between gene action and development. Using Monte Carlo simulation, we tested the statistical behavior and power of the new model. The usefulness and utilization of the new model is validated through analyzing a real data set for genetic dissection of biomass growth in soybeans. The new model should enable geneticists to shed light on the genetic control mechanisms of a biological system and predict its physiological and pathological states.

Keywords: Differential equations, dynamic system, quantitative trait loci, soybean, systems mapping

1 Introduction

How to predict the phenotype from the genotype of complex organisms is one of the most important and challenging questions we face in modern biology and biomedicine (Mackay et al. 2009). Genetic mapping, aimed to dissect a phenotypic trait into its underlying genes, known as quantitative trait loci (QTLs), using molecular markers, has proven powerful for constructing the genotype-phenotype relationship and predicting phenotypic changes of individual organisms based on their QTL genotypes responsible for the trait (Lynch and Walsh 1998). The success of the prediction of phenotype from genotype depends on how we can precisely map the underlying QTLs and characterize complex interactions of these QTLs with each other and environmental factors. In the past two decades, tremendous efforts have been made worldwide to detect QTLs and study their biological function for various phenotypic traits by developing powerful statistical models (Lander and Botstein 1989; Zeng 1994; Xu and Atchley 1995; Wu et al. 2002; Yi and Xu 2008; Zou et al. 2009) and performing molecular experiments from which a large amount of data have been collected (Steinmetz et al. 2002; Mackay et al. 2009; Balasubramanian et al. 2009; Ehrenreich et al. 2010). Unfortunately, despite hundreds of thousands of QTLs detected in a diversity of organisms, the actual biological function of these QTLs remains elusive. Because of paucity of

this knowledge, very few QTLs can have been cloned thus far in practice (see Frary et al. 2000; Li et al. 2006).

Perhaps one of the reasons for this gap is that most of the QTLs detected, although significant by a statistical test, are not biologically relevant, since existing strategies for QTL mapping were built on a direct association between genotype and final phenotype. Although such strategies are simple and have been widely accepted, they neglect the biology inside the "black box" that lies in the process of trait formation (Mackay et al. 2009). To attempt to fill this gap, a statistical model, called functional mapping (Ma et al. 2002; Wu and Lin 2006; Li and Wu 2010), has been developed to study the interplay between genetics and the developmental process of a phenotypic trait by integrating mathematical models and computational algorithms. If a trait is treated as a "system" that is composed of many underlying biological components, we should be in a better position to comprehend the process and behavior of trait formation based on interactive relationships among different components. Through mapping and using those QTLs that govern design principles of a biological system, a new trait that is able to maximize resource-use efficiency can be generated.

In agriculture, whole-plant biomass, as one of the most important target traits for plant breeding, can be partitioned into leaves, stem, and roots. From a mechanistic perspective, plant biomass growth is not simply the addition of these individual parts, and more importantly, entails the coordination of these parts through natural laws. Chen and Reynolds (1997) used coordination theory to model the dynamic allocation of carbon to different organs by a group of differential equations. A series of allometric studies by West et al. (1997, 1999, 2001) explains a power relationship existing between parts and the whole from fundamental biophysical and biochemical principles; i.e., plants tend to maximize leaf surface area for photosynthesis and minimize the transport distance for water, nutrients, and carbon. By integrating Chen and Reynolds's coordination theory and West et al.'s allometry, we construct a tripled group of ordinary differential equations (ODEs) to specify the coordination of leaf,

stem, and root biomass for a plant:

$$\begin{aligned} \frac{dM_L}{dt} &= \alpha_L W^{\beta_L} - \gamma_L M_L \\ \frac{dM_S}{dt} &= \alpha_S W^{\beta_S} \\ \frac{dM_R}{dt} &= \alpha_R W^{\beta_R} - \gamma_R M_R \end{aligned} \tag{1}$$

where M_L, M_S, and M_R are the biomasses of the leaves (L), the stem (S), and the roots (R), respectively, with whole-plant biomass $W = M_L + M_S + M_R$; α and β are the constant and exponent power of an organ biomass scaling as whole-plant biomass (West et al. 1997, 1999); and γ is the rate of eliminating ageing leaves and roots. The interactions between different parts of a plant can be modeled and studied by estimating and testing the ODE parameters $(\alpha_L, \beta_L, \lambda_L, \alpha_S, \beta_S, \alpha_R, \beta_R, \lambda_R)$.

The motivation of this study is to integrate a group of differential equations like (1) into a statistical framework for QTL mapping. Much work in solving ODEs has focused on the simulation and analysis of the behavior of state variables for a dynamic system, but the estimation of ODE parameters that define the system based on the measurement of state variables at multiple time points is relatively a new area. Yet, in the recent years, many statisticians have made great attempts to develop statistical approaches for estimating ODE parameters by modeling the structure of measurement errors. In summary, these statistical approaches include: (a) hierarchical Bayesian models for estimating dynamic parameters (Putter et al. 2002; Huang and Wu 2006; Huang et al. 200); (b) spline-based approaches for estimating time-varying parameters (Li et al. 2002); (c) principal differential analysis (Ramsay 1996; Ramsay and Silverman 2005); (d) a penalized spline method for estimating constant dynamic parameters (Ramsay et al. 2007); and (e) a local smoothing approach based on a pseudo-least square principle (Liang and Wu 2008).

Our problem for systems mapping with ODE models is different from those considered in current literature. First, our model is constructed within a mixture-based framework because QTL genotypes that define the DE mod-

els are missing. Second, our model incorporates genotypic data which are categorical or binary. These two characteristics determine the high complexity of our statistical model and computational algorithm. The rest of the paper is organized as follows. Section 2 formulates the principle and method of systems mapping, computational algorithms and hypothesis tests for QTL effects. Section 3 demonstrates an application of the model by analyzing a real data set for soybean mapping as well as results from computer simulations illustrating the statistical properties of the model. The conclusions and discussion are given in Section 4.

2 Model

2.1 Genetic Design

Consider a mapping plant population of n recombinant inbred lines (RIL) or doubled haploids (DH), initiated with two contrasting homozygous inbred parents. For an RIL or DH population, two genotypes at each locus are homozygous for the alternative alleles each derived from one original parent. For this reason, the progeny of each RIL or DH has an identical genotype which allows the same genotype to be repeated. This population is genotyped for a panel of molecular markers from which a linkage map covering the entire genome is constructed through a pair-wise linkage analysis. The recombination fraction between two markers is converted to the genetic distance in centiMorgan (cM) through a map function, such as the Haldane or Kosambi map function. The map constructed is used to locate QTLs that control a quantitative trait of interest. Unlike a traditional mapping project, we will partition the trait into its different components based on biological knowledge. Consider whole-plant biomass as an example. We partition it into leaf, stem and root biomasses for each plant. By repeating the progeny of the same genotype with multiple replicates, we use a destructive sampling approach to measure biomasses from the three organs, respectively, at multiple time points in a given period of plant growth. Our goal is to map QTLs that control whole-plant biomass

through the dynamic coordination of three components.

2.2 Likelihood

Let $\mathbf{y}_{ki} = (y_{ki}(t_{i1}), \cdots, y_{ki}(t_{im_i}))^T$ denote the vector of phenotypic values for trait k ($k = 1$ for leaf biomass (L), 2 for stem biomass (S), and 3 for root biomass (R)) measured on progeny i at time points $(t_{i1}, \cdots, t_{im_i})^T$. Note that the number of time points measured may be progeny-specific, expressed as m_i for progeny i. Assuming that multiple QTLs (segregating with J genotypes), each bracketed by two flanking markers $\mathbf{M}$, affects these three traits, we construct a mixture model-based likelihood as

$$L(\mathbf{z}, \mathbf{M}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \left[\omega_{j|i} f_j(\mathbf{z}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi})\right] \tag{2}$$

where $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3)$ is a joint vector of phenotypic values for the three traits, with $\mathbf{z}_i = (\mathbf{y}_{1i}, \mathbf{y}_{2i}, \mathbf{y}_{3i})$ presenting the $\mathbf{z}$-vector for progeny i; $\omega_{j|i}$ is the conditional probability of QTL genotype j ($j = 1, ..., J$) given the marker genotype of progeny i , which can be expressed as a function of the recombination fractions between the QTL and markers (Wu et al. 2007), and $f_j(\mathbf{z}_i; \boldsymbol{\Theta}_j, \boldsymbol{\Psi})$ is an MVN of leaf, stem and root biomasses for progeny i which carries QTL genotype j, with mean vectors

$$\boldsymbol{\mu}_j = (\boldsymbol{\mu}_{1j}, \boldsymbol{\mu}_{2j}, \boldsymbol{\mu}_{3j}), \quad j = 1, ..., J$$

specified by $\boldsymbol{\Theta}_j$, and covariance matrix specified by $\boldsymbol{\Psi}$. If a system of differential equations (1) is used to jointly model QTL genotype-specific means vectors for the three traits, then we have $\boldsymbol{\Theta}_j = (\alpha_{Lj}, \beta_{Lj}, \lambda_{jL}, \alpha_{Sj}, \beta_{Sj}, \alpha_{Rj}, \beta_{Rj}, \lambda_{Rj})$ for genotype j.

2.3 Estimation

Unlike a traditional mixture model for QTL mapping, we will model the genotypic values of each QTL genotype in likelihood (2) characterized by a group of nonlinear ODEs. While an analytical solution is not available, we will implement numerical approaches to solve these ODEs. Let $\mu_{kj}(t)$ denote the

genotypic value of the kth trait at time t for a QTL genotype j. Thus, the dynamic system of the traits and their interactions, regulated by QTL genotypes, can be modeled by a system of ODE (1),

$$\frac{d\mu_{kj}(t)}{dt} = g_k(\boldsymbol{\mu}_j(t), \boldsymbol{\Theta}_j),\ k = 1, \cdots, 3, \tag{3}$$

where $\boldsymbol{\mu}_j(t) = (\mu_{1j}(t), \cdots, \mu_{3j}(t))^T$, and $\boldsymbol{\Theta}_j$ is a vector of ODE parameters associated with QTL genotype j. For J possible genotypes in the mapping population, we have $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1^T, ..., \boldsymbol{\Theta}_J^T)^T$. The question now is how to estimate $\boldsymbol{\Theta}$ from noisy measurements. The functional mean $\mu_{kj}(t)$ may be represented as a linear combination of basis functions:

$$\mu_{kj}(t) = \sum_{r=1}^{R} c_{kjr}\phi_{kjr}(t) = \mathbf{c}_{kj}^T \boldsymbol{\phi}_{kj}(t)$$

where $\boldsymbol{\phi}_{kj}(t) = (\phi_{kj1}(t), \cdots, \phi_{kjR}(t))^T$ is a vector of basis functions with R orders and $\mathbf{c}_{kj} = (c_{kj1}, \cdots, c_{kjR})^T$ is a vector of basis coefficients. Define $\mathbf{c} = (\{\mathbf{c}_{kj}\}_{k=1;j=1}^{3;J})$ as a length $(R \times 3 \times J)$ vector of basis coefficients. The cubic B-splines are often chosen as basis functions, since any B-spline basis function is only positive over a short subinterval and zero elsewhere. This is called the *compact support* property, and is essential for efficient computation. The flexibility of the B-spline basis functions depend on the number and location of knots we choose. It is an infinite-dimension optimization problem to choose the optimal number of knots and their locations. A popular approach to avoid this dilemma is choosing a saturated number of knots and using a roughness penalty to control the smoothness of the fitted curve and avoid over-fitting (Ramsay and Silverman 2005).

We estimate the basis coefficient $\mathbf{c}$ and ODE parameter $\boldsymbol{\Theta}$ based on a two-nested level of optimization. In the inner level of optimization, $\mathbf{c}$ is estimated by optimizing a criterion $U(\mathbf{c}|\boldsymbol{\Theta})$, given any value of $\boldsymbol{\Theta}$. Therefore, the estimate $\widehat{\mathbf{c}}$ may be viewed as a function of $\boldsymbol{\Theta}$, which is denoted as $\widehat{\mathbf{c}}(\boldsymbol{\Theta})$. Since no analytic formula for $\widehat{\mathbf{c}}$ is available, $\widehat{\mathbf{c}}(\boldsymbol{\Theta})$ is an implicit function. In the outer level of optimization, $\boldsymbol{\Theta}$ is estimated by optimizing a criterion $H(\boldsymbol{\Theta}, \widehat{\mathbf{c}}(\boldsymbol{\Theta}))$.

The parameter $\widehat{\mathbf{c}}$ is removed in the parameter space in the outer level by treating it as an implicit function of $\boldsymbol{\Theta}$. Although $\widehat{\mathbf{c}}(\boldsymbol{\Theta})$ does not have an analytic formula, the outer level of optimization only requires to calculate the derivative $d\widehat{\mathbf{c}}/d\boldsymbol{\Theta}$, which can be obtained by using the implicit function theorem. The above optimization procedure is called the parameter cascading method. Note that when the two criteria $U(\mathbf{c}|\boldsymbol{\Theta})$ and $H(\boldsymbol{\Theta}, \widehat{\mathbf{c}}(\boldsymbol{\Theta}))$ are the same, the parameter cascading method is equivalent to the profiling method. The next two subsections will introduce the two optimization criteria and implementation details.

2.4 Inner Level of Optimization

The basis coefficient $\mathbf{c}$ may be estimated by maximizing the log likelihood function of the data. However, some important information may then be missed. Although the ODE parameter $\boldsymbol{\Theta}$ are often unknown, the structure form of the ODEs (3) is often given directly by the investigators based on their knowledge of the dynamic system. Therefore, the following optimization criterion is proposed to estimate the basis coefficients, given the value of ODE parameters:

$$\begin{aligned} U(\mathbf{c}|\boldsymbol{\Theta}) = & - \sum_{k=1}^{3}\sum_{i=1}^{n} \log\left[\sum_{j=1}^{J} \omega_{j|i} f_j(\mathbf{y}_{ki}|\mathbf{c}_{kj})\right] \\ & + \sum_{k=1}^{3}\sum_{j=1}^{J} \lambda_{kj} \int \left[\frac{d\mu_{kj}(t)}{dt} - g_k\Big(\boldsymbol{\mu}_j(t), \boldsymbol{\Theta}_j\Big)\right]^2 dt \end{aligned} \tag{4}$$

where the first term is the minus log-likelihood function having the mean $E(\mathbf{y}_{ki}) = \boldsymbol{\Phi}_{kj}\mathbf{c}_{kj}$ with $\boldsymbol{\Phi}_{kj}$ being an $(m_i \times R)$ matrix whose τr-th element is $\phi_{kjr}(t_{i\tau})$; and the second term in (4) measures the infidelity of the functional mean $\mu_{kj}(t)$ to the ODE model (4). The smoothing parameter λ_{kj} controls the trade-off between the fit to the data and fidelity to the ODE model. Because the second term has the first derivative of $\mu_{kj}(t)$ with respect to t, it also penalizes the roughness of the functional mean.

Since the first term in (4) has the structure of the mixture pdf, it is often

impossible to obtain the analytic formula for the estimate $\hat{\mathbf{c}}$. The basis coefficients $\mathbf{c}$ are estimated by the Newton-Raphson iteration method for which the Hessian matrix is given in the Appendix. Let $\mathbf{c}^0$ be the starting value, the iteration becomes

$$\mathbf{c}^{v+1} = \mathbf{c}^v - \left(\frac{\partial^2 U}{\partial \mathbf{c}^2}\bigg|_{\mathbf{c}^v} \right)^{-1} \left(\frac{\partial U}{\partial \mathbf{c}}\bigg|_{\mathbf{c}^v} \right)$$

until convergence. The above two derivatives are can be obtained analytically, which facilitates the stability and computing speed of the iteration. Once the analytic first and second derivative of $U(\mathbf{c}|\mathbf{\Theta})$ with respect to $\mathbf{c}$ are derived (with a detail given in the Appendix), an outer level of optimization can be used to estimate $\mathbf{\Theta}$.

In practice, the integration term in (4) is evaluated numerically. We use the composite Simpson's rule, which provides an adequate approximation to the exact integral (Burden and Faires 2000). For an arbitrary function $\rho(t)$, the composite Simpson's rule is given by

$$\int_{t_1}^{t_n} \rho(t)dt \approx \frac{h}{3}[\rho(s_0) + 2\sum_{q=1}^{Q/2-1} \rho(s_{2q} + 4\sum_{q=1}^{Q/2} \rho(s_{2q-1}) + \rho(s_Q)]$$

where the quadrature points $s_q = t_1 + qh$, $q = 0, \cdots, Q$, and $h = (t_n - t_1)/Q$.

2.5 Outer Level of Optimization

The ODE parameter $\mathbf{\Theta}$ is then estimated by maximizing the log likelihood function

$$\ell(\mathbf{\Theta}) = \sum_{k=1}^{3}\sum_{i=1}^{n} \log \left[\sum_{j=1}^{J} \omega_{j|i} f_j\Big(\mathbf{y}_{ki}|\widehat{\mathbf{c}}(\mathbf{\Theta})\Big) \right] \tag{5}$$

The optimization is implemented using the Newton-Raphson iteration method, which can run faster and more stable if the gradient is given analytically. The analytic gradient is derived with the chain rule to accommodate $\widehat{\mathbf{c}}$ which is a function of $\mathbf{\Theta}$:

$$\frac{d\ell}{d\mathbf{\Theta}} = \left(\frac{d\widehat{\mathbf{c}}}{d\mathbf{\Theta}} \right)^T \frac{d\ell}{d\widehat{\mathbf{c}}}.$$

Although $\widehat{\mathbf{c}}(\mathbf{\Theta})$ is an implicit function, the analytic formula of the derivative $d\widehat{\mathbf{c}}/d\mathbf{\Theta}$ can be obtained by using the implicit function theorem as follows. Taking the $\mathbf{\Theta}$-derivative on both sides of the identity $\frac{\partial U}{\partial \mathbf{c}}|_{\widehat{\mathbf{c}}} = 0$:

$$\frac{d}{d\mathbf{\Theta}}\left(\left.\frac{\partial U}{\partial \mathbf{c}}\right|_{\widehat{\mathbf{c}}}\right) = \left.\frac{\partial^2 U}{\partial \mathbf{c}\partial\mathbf{\Theta}}\right|_{\widehat{\mathbf{c}}} + \left.\frac{\partial^2 U}{\partial \mathbf{c}^2}\right|_{\widehat{\mathbf{c}}}\frac{d\widehat{\mathbf{c}}}{d\mathbf{\Theta}} = 0.$$

Assuming that $\frac{\partial^2 U}{\partial \mathbf{c}^2}|_{\widehat{\mathbf{c}}}$ is not singular, we get

$$\frac{d\widehat{\mathbf{c}}}{d\mathbf{\Theta}} = -\left[\left.\frac{\partial^2 U}{\partial \mathbf{c}^2}\right|_{\widehat{\mathbf{c}}}\right]^{-1}\left[\left.\frac{\partial^2 U}{\partial \mathbf{c}\partial\mathbf{\Theta}}\right|_{\widehat{\mathbf{c}}}\right].$$

2.6 Likelihood Ratio Test

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

$$H_0 : \mathbf{\Theta}_j \equiv \mathbf{\Theta} \; vs. \; H_1 : \mathbf{\Theta}_j \neq \mathbf{\Theta}, \text{ for } j = 1, ..., J \tag{6}$$

The likelihoods under the null and alternative hypotheses are calculated, from which the log-likelihood ratio (LR) is computed. Suppose the ODE parameters are estimated as $\widetilde{\mathbf{\Theta}}$ and $\widehat{\mathbf{\Theta}}$ under the H_0 and H_1, respectively, the LR test statistic is calculated as

$$\text{LR} = -2[\ell(\widetilde{\mathbf{\Theta}}) - \ell(\widehat{\mathbf{\Theta}})],$$

where the log-likelihood function $\ell(\cdot)$ is defined in (5). There is significant evidence for QTLs located at different positions if the calculated LR $> T$, where T is a significance threshold. Since the distribution of the LR values under the null hypothesis is unknown, empirical permutation tests are usually used to determine the threshold (Churchill and Doerge 1994).

3 Results

3.1 Worked Example

The mapping population includes $n = 184$ recombinant inbred lines (RIL) derived from two cultivars, Kefeng No. 1 and Nannong 1138-2, and was geno-

typed for 950 molecular markers locating in 25 linkage groups (Zhang et al. 2004; Li et al. 2007). The plants were grown in a simple lattice design with two replicates in a plot at Jiangpu Soybean Experiment Station, Nanjing Agricultural University, China. Ten serial plants in the second row of a plot were randomly selected for measuring leaf, stem and root biomass at each time in the whole growing season. After 20 days of seedling emergence, dry weights separately for the leaves, stem and roots were measured once every 5 to 10 days until most plants stopped growth. A total of 6 to 8 measurements were taken for the RILs studied. Great efforts were made to control measurement errors for such a large-scale field trial, with phenotyping precision beyond 95%.

Figure 1 illustrates the growth trajectories of leaf, stem and root biomass for individual RILs. By using a system of ODEs (1) to fit growth trajectories of leaf, stem and root biomass, we obtained a mean curve for all the 184 RILs. It can be seen that growth trajectories can be well modeled by three interconnecting ODEs (1) that were derived from coordination theory (Chen and Reynolds 1997) and allometric scaling (West et al. 1997, 1999, 2001). The models delineate the decay of leaf and root biomass because of ageing, which has been well reflected in fitted curves of leaf (Fig. 1A) and root biomass (Fig. 1C). As expected, stem biomass growth does not experience a decay (Fig. 1B) although it tends to be stationary in a late stage of development.

By assuming a possible QTL locating at every 2 cM throughout the 25 linkage groups (LG), we calculated the LR values at each position using systems mapping and then plotted them against the length of genome to get a smooth QTL support LR curve (Fig. 2). Two significant QTLs were identified, one residing between two markers GMKF167a and GMKF167b on LG 12 and the other residing between two markers sat_274 and BE801128 on LG 24, with the corresponding LR values beyond the significance threshold 1073.2 obtained from 100 permutation tests. Table 1 shows the maximum likelihood estimates of the parameters in ODE (1) for the detected QTLs. The standard errors for the parameter estimates were obtained by the parameter bootstrap (Efron and Tibshirani 1993). The estimated ODE parameters $\mathbf{\Theta}_j$ for two different

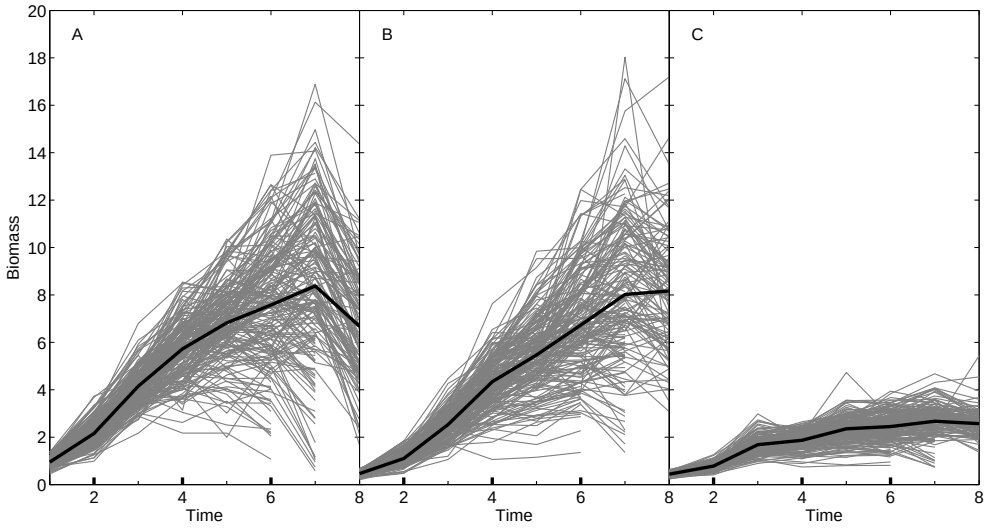

Figure 1: Growth trajectories of leaf (A), stem (B) and root biomass (C) measured at multiple time points in a growing season of soybeans. Each grey line presents the growth trajectory of one of 184 RILs, whereas black lines are the mean growth trajectories of all RILs fitted using a system of ODEs (1).

genotypes at each QTL were used to dictate the biomass growth trajectories of soybean leaves, stem, and roots in the growing season (Fig. 3).

The genetic effects of the QTLs displayed different temporal patterns for three organs. There is a steeper rate of increase of allelic expression at the QTLs with time for the stem than for the leaves and roots. At the LG 12 QTL, parent Nannong 1138-2 contributes favorable alleles to leaf and stem growth (Fig. 3A and 3B), but the favorable alleles are derived from parent Kefeng No. 1 for root growth (Fig. 1C). At the LG 24 QTL, the alleles inherited from parent Nannong 1138-2 are favorable for better leaf growth (Fig. 3D), whereas the allele from parent Kefeng No. 1 provides favorable alleles for stem and root growth (Fig. 3E and 3F). It can be seen that the QTL on LG 12 makes leaf and stem growth favorably correlated, whereas the QTL on LG 24 makes stem and root growth favorably correlated.

The functional relationships among leaf, stem and root biomass were deter-

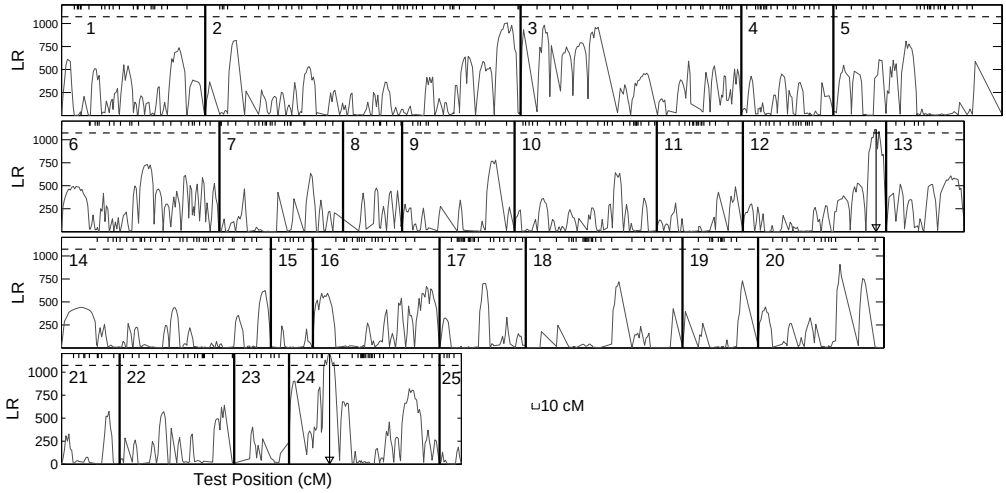

Figure 2: The profiles of log-likelihood ratios (LR) between the full model (there is a QTL) and reduced model (there is no QTL) for growth trajectories of three related traits, leaf biomass, stem biomass and root biomass, through a system of ODEs throughout the soybean genome composed of 25 linkage groups. Genomic positions at the peaks of the profile beyond the threshold are the maximum-likelihood estimates of the QTL locations. Tick marks on the ceilings of each panel represent the positions of molecular markers on each linkage groups (bar, 10 cM). The critical threshold for acclaiming the genome-wide existence of a QTL was obtained from permutation tests. The 95th percentile (indicated by broken horizontal lines) of the distribution of the maximum LR values obtained from 100 permutation tests is used as an empirical critical value to declare genome-wide existence of a QTL at the 5% significance level.

Table 1: The maximum likelihood point estimates (PEs) of ODE parameters and standard errors (SEs) of the estimates for the QTLs located on LG 12 and 24. Model M1 assumes two different genotypes at a QTL (under the H_1), whereas Model M0 assumes a single genotype (under the H_0).

Model	Genotype	Estimate	α_L	β_L	γ_L	α_S	β_S	α_R	β_R	γ_R
QTL between markers GMKF167a and GMKF167b on LG 12										
M1	*QQ*	PE	2.09	0.16	0.43	0.93	0.07	1.52	0.66	2.91
		SE	0.02	1e-3	3e-3	0.01	2e-4	0.01	6e-3	0.02
	qq	PE	2.53	0.11	0.36	0.92	0.04	1.57	0.54	3.90
		SE	0.01	4e-4	2e-3	0.01	1e-4	0.01	5e-3	0.01
M0		PE	2.30	0.13	0.39	0.92	0.05	1.56	0.60	3.37
		SE	0.07	2e-3	0.01	0.02	4e-4	0.05	0.02	0.08
QTL between markers sat_274 and BE801128 on LG 24										
M1	*QQ*	PE	1.89	0.14	0.44	1.04	0.07	1.11	0.56	1.85
		SE	0.04	1e-3	0.01	0.01	1e-4	0.01	5e-3	0.01
	qq	PE	2.55	0.10	0.31	0.98	0.04	1.11	0.51	2.18
		SE	0.04	1e-3	0.01	0.01	4e-4	0.01	5e-3	0.02
M0		PE	2.25	0.12	0.37	1.03	0.05	1.12	0.55	2.06
		SE	0.08	3e-3	0.02	0.02	6e-4	0.02	0.01	0.05

Note: The alleles of genotype *QQ* are derived from Kefeng No. 1 and those of genotype *qq* derived from Nannong 1138-2.

mined by the QTLs detected (Fig. 4). For the QTL on LG 12, two genotypes are not only different in biomass growth trajectories of individual organs, but also display pronounced discrepancies in whole-plant biomass trajectory (Fig. 4A and 4B). This means that this QTL affects the dynamics of both plant size and biomass partitioning. For the QTL on LG 24, two genotypes are similar in total plant size during the entire growing period, but they have a striking distinction in biomass partitioning (Fig. 4C and 4D). It appears that this QTL affects structural trajectories of plants through altering biomass partitioning rather than the total amount of biomass.

The alleles at the LG-12 QTL derived from parent Kefeng No. 1 tend to allocate more biomass to leaves than the stem in lifetime, although this difference decreases with developmental time (Fig. 5A). A similar pattern of biomass partitioning is observed for the alleles at the LG-24 QTL from parent Nannong 1138-2 (Fig. 5D). The alleles at the LG-12 QTL from Nannong 1138-2 and those at the LG-24 QTL from parent Kefeng No. 1 allocates more biomass into leaves than the stem at the early part of the growing season but alter this pattern of biomass allocation at the late part (Fig. 5B and 5C). In general, the stem receives a proportionally larger amount of biomass with development, accompanying a proportional decrease of biomass to the leaves and roots. It can be seen that this allometric change with development is under genetic control. The alleles from Kefeng No. 1 tend to amplify this contrast in development-dependent biomass partitioning, as compared to those from Nannong 1138-2. Results about the genetic control of biomass partitioning help to gain more insights into the mechanistic response of plant size and shape to developmental and environmental signals and, also, provide guidance to select an ideotype of crop cultivars with optimal shape and structure suited to a particular environment.

3.2 Simulation

The performance of the newly derived systems mapping model is examined through a simulation study that mimics the example used in the preceding

Table 2: Summary of the estimates for the ODE parameters from 100 simulation replicates.

Genotype	Estimate	α_L	β_L	γ_L	α_S	β_S	α_R	β_R	γ_R
QQ	TRUE	2.55	0.10	0.31	0.98	0.04	1.11	0.51	2.18
	MEAN	2.55	0.10	0.31	0.98	0.04	1.11	0.51	2.18
	BIAS$*10^3$	2.10	-0.07	0.72	0.27	0.01	3.49	-0.88	-2.89
	STD$*10^2$	3.75	0.08	0.60	0.94	0.01	1.29	0.49	1.11
	RMSE$*10^2$	3.75	0.08	0.61	0.94	0.01	1.34	0.50	1.15
qq	TRUE	1.89	0.14	0.44	1.04	0.07	1.11	0.56	1.85
	MEAN	1.89	0.14	0.44	1.04	0.07	1.11	0.56	1.85
	BIAS$*10^3$	-4.23	0.11	0.55	-0.61	-0.05	1.62	0.92	-2.02
	STD$*10^2$	3.60	0.14	0.92	0.99	0.04	1.38	0.45	1.66
	RMSE$*10^2$	3.62	0.14	0.92	0.99	0.04	1.39	0.46	1.67

section. The simulated mapping population contains the same genotype data for the 184 RILs. Phenotype data were simulated as the summation of genotypic data and residual errors. In order to mimic the real data as much as possible, genotypic values at different time points are described by the estimates of ODE parameters estimated for two different genotypes of a significant QTL (say one on LG 24). The phenotypic values for the k-th trait were simulated by adding white noise with variances σ_k^2 to the ODE curve for the j-th QTL genotype with the probability $w_{j|i}$, i.e., the conditional probability of the i-th RIL that carries the $j$the QTL genotype, given the two markers genotypes of this RIL. The values of the noise variance σ_k^2 were set as the estimates from the real data, which are $\sigma_1^2 = 2.42$, $\sigma_2^2 = 1.72$ and $\sigma_3^2 = 0.14$ for leaf, stem, and root biomass, respectively. The parameters in the ODE (1) are then estimated from the simulated data with the parameter cascading method. The simulation was implemented with 100 replicates from which the means, biases, standard deviations, and root mean square errors were calculated (Table 2).

It can be seen from Table 1 that the model can provide reasonably accurate

and precise estimates of QTL genotype-specific ODE parameters from the real data set provided with a modest sample size ($n = 184$). The biases of the estimates are negligible, compared with the scale of the standard deviations. Such good estimates may be due to low measurement errors during trait phenotyping. We performed an additional simulation to investigate the power and false positive rate of the model by changing levels of noises. In general, the power of the model is high, reaching 0.80 even when the heritability of growth curves is low (0.05). Basically, a QTL can be fully detected when the heritability is 0.10 or larger. In any case, the false positive rate is not beyond 0.10, mostly being less than 0.05.

4 Discussion

With the completion of the Human Genome Project, followed by other species, it has been possible to comprehend the genetic architecture of complex traits and, ultimately, integrate genetic information into genetic improvement programs or clinical therapies for disease treatment and prevention. To achieve this goal, we need to detect genes and genetic interactions that determine the complex coordination of different parts of a phenotype. Previous work by implementing the mathematical aspects of trait formation leads to a dynamic model called functional mapping (Ma et al. 2002; Wu and Lin 2006; Li and Wu 2009). The central innovation of functional mapping lies in the study of the temporal pattern of genetic variation for a complex trait in a time course and the identification of specific QTLs that determine such a time-dependent change of the trait. With increasing recognition of the importance of trait partitioning in systems and synthetic biology, however, it is now a time to push functional mapping toward the level of systems mapping at which the genetic control of part-part interactions as well as the pathways to the final phenotype can be studied.

In this article, we developed a statistical model, called systems mapping, aimed to map genes or quantitative trait loci (QTLs) that control a dynamic

system through regulating a web of interactions among different components of the system. By integrating the mathematical aspects of the biological mechanisms and processes of the system into a mixture model framework, systems mapping can address many questions, such as the patterns of genetic control over development, the duration of QTL effects, as well as what causes developmental trajectories to change or stop changing. From a biological viewpoint, systems mapping can test the genetic control of several mechanistically meaningful relationships. Specifically, they are

(1) Size-shape relationship: Parameters $(\alpha_L, \beta_L, \lambda_L)$, (α_S, β_S) and $(\alpha_R, \beta_R, \lambda_R)$ describe the allometric relationships between size and shape for a plant. For example, a big plant may be due to many leaves, a large stem, and/or a large root system. We can test if such a relationship is controlled by the QTL. The null hypothesis is formulated as $H_0 : (\alpha_L, \beta_L, \lambda_L) \equiv (\alpha, \beta, \lambda)$ for the leaf trait, $H_0 : (\alpha_S, \beta_S) \equiv (\alpha_S, \beta_S)$ for the stem trait, and $H_0 : (\alpha_R, \beta_R, \lambda_R) \equiv (\alpha_R, \beta_R, \lambda_R)$ for the root trait.

(2) Structural-functional relationship: Similarly, we can test whether the QTL affects the structural-functional relationship of leaf and stem traits with whole-plant growth.

(3) Cause-effect relationship: How leaf traits (causes) determine other traits (effects), such as whole-plant biomass and seed production (if any), can be readily tested from the likelihood model (2).

(4) Pleiotropic relationship: We are modeling a dynamic system in which different parts or components function jointly to determine final outcome in time course. Thus, it is of paramount importance to test whether the QTL pleiotropically determines different aspects of the system. We can formulate different tests based on the tripled ODEs (1).

Systems mapping provides a quantitative and testable framework for assessing the interplay between gene action and development. The genetic dissection of a biological process through systems mapping will not only identify genes that regulate the final form of the trait, but also characterize the genes responsible for the dynamic pattern of part-part interactions that cause the final

form. Furthermore, systems mapping capitalizes on the parsimonious modeling of longitudinal mean-covariance structures and, therefore, substantially increases the statistical power and robustness of genetic mapping.

Acknowledgements

This work is supported by DMS/NIGMS-0540745, NSF/IOS-0923975, Changjiang Scholars Award, and "One-thousand Person" Award, and a discovery grant of the Natural Sciences and Engineering Research Council of Canada (NSERC) to J. Cao.

Appendix

If $y_{ki}(t)$, given the QTL genotype j, is assumed to be normally distribution with the mean $\mu_{kj}(t)$ and variance σ_k^2. Denote $\mathbf{y}_{ki} = (y_{ki}(t_{i1}), \cdots, y_{ki}(t_{im_{ki}}))^T$, $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1^T, \cdots, \boldsymbol{\Theta}_J^T)^T$, and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \cdots, \sigma_H^2)^T$. The optimization criterion (4) may be expressed as

$$\begin{aligned} U(\mathbf{c}_{kj}|\boldsymbol{\Theta}, \boldsymbol{\sigma}^2) = & - \sum_{k=1}^{3}\sum_{i=1}^{n} \log\left[\sum_{j=1}^{J} \omega_{j|i} f_j\left(\mathbf{y}_{ki}|\mathbf{c}_{kj}, \sigma_k^2\right)\right] \\ & + \sum_{j=1}^{J}\sum_{k=1}^{3} \lambda_k \int \left[\frac{d\mu_{kj}(t)}{dt} - g_k\left(\boldsymbol{\mu}_j(t), \boldsymbol{\Theta}_j\right)\right]^2 dt. \end{aligned}$$

where $f_j(\mathbf{y}_{ki}|\mathbf{c}_{kj}, \sigma_k^2) = (2\pi\sigma_k^2)^{-m_i/2} \exp(-(\mathbf{y}_{ki} - \boldsymbol{\Phi}_{kj}\mathbf{c}_{kj})^T(\mathbf{y}_{ki} - \boldsymbol{\Phi}_{kj}\mathbf{c}_{kj})/(2\sigma_k^2))$ $(j = 1, \cdots, J)$ indicates the normal distribution of genotype j and $\boldsymbol{\Phi}_{kj}$ is a $m_i \times R$ matrix with the τr-th element $\phi_{kjr}(t_{i\tau})$.

The gradient is

$$\begin{aligned} \frac{dU}{d\mathbf{c}_{kj}} = & -\sum_{k=1}^{3}\sum_{i=1}^{n}\sum_{j=1}^{J} \frac{P_{ijk}}{\sigma_k^2} \boldsymbol{\Phi}_{kj}^T(\mathbf{y}_i - \boldsymbol{\Phi}_{kj}\mathbf{c}_{kj}) \\ & +2\lambda_k \int \left[\frac{d\mu_{kj}(t)}{dt} - g_k(\boldsymbol{\mu}_j(t), \boldsymbol{\Theta}_j)\right]\left[\frac{d\boldsymbol{\phi}_{kj}(t)}{dt} - \frac{dg_h(\boldsymbol{\mu}_j(t), \boldsymbol{\Theta}_j)}{d\mu_k}\boldsymbol{\phi}_{kj}(t)\right] dt \\ & +2\sum_{s=1, s\neq k}^{K} \lambda_s \int \left[\frac{d\mu_{sj}(t)}{dt} - g_s(\boldsymbol{\mu}_j(t), \boldsymbol{\Theta}_j)\right]\left[-\frac{dg_s(\boldsymbol{\mu}_j(t), \boldsymbol{\Theta}_j)}{d\mu_k}\boldsymbol{\phi}_{kj}(t)\right] dt \end{aligned}$$

where

$$P_{ijk} = \frac{\omega_{j|i} f_j(\mathbf{y}_{ki}|\mathbf{c}_{kj}, \sigma_k^2)}{\sum_{j=1}^{J} \omega_{j|i} f_j(\mathbf{y}_{ki}|\mathbf{c}_{kj}, \sigma_k^2)}.$$

The Hessian matrix is

$$\begin{aligned}
&\frac{d^2U}{d\mathbf{c}_{kj}^T d\mathbf{c}_{kj}} \\
= &\sum_{k=1}^{R}\sum_{i=1}^{n}\sum_{j=1}^{J} \frac{P_{ijk}}{\sigma_k^2}\mathbf{\Phi}_{kj}^T\mathbf{\Phi}_{kj} \\
&+2\lambda_k \int \left[\frac{d\boldsymbol{\phi}_{kj}(t)}{dt} - \frac{dg_h(\boldsymbol{\mu}_j(t), \mathbf{\Theta}_j)}{d\mu_k}\boldsymbol{\phi}_{kj}(t)\right]\left[\frac{d\boldsymbol{\phi}_{kj}(t)}{dt} - \frac{dg_h(\boldsymbol{\mu}_j(t), \mathbf{\Theta}_j)}{d\mu_k}\boldsymbol{\phi}_{kj}(t)\right]^T dt \\
&+2\lambda_k \int \left[\frac{d\mu_{kj}(t)}{dt} - g_k(\boldsymbol{\mu}_j(t), \mathbf{\Theta}_j)\right]\left[-\frac{dg_h(\boldsymbol{\mu}_j(t), \mathbf{\Theta}_j)}{d{\mu_k}^2}\boldsymbol{\phi}_{kj}(t)\boldsymbol{\phi}_{kj}^T(t)\right] dt \\
&+2\sum_{s=1,s\neq k}^{K}\lambda_s \int \left[-\frac{dg_s(\boldsymbol{\mu}_j(t), \mathbf{\Theta}_j)}{d\mu_k}\boldsymbol{\phi}_{kj}(t)\right]\left[-\frac{dg_s(\boldsymbol{\mu}_j(t), \mathbf{\Theta}_j)}{d\mu_k}\boldsymbol{\phi}_{kj}(t)\right]^T dt \\
&+2\sum_{s=1,s\neq k}^{K}\lambda_s \int \left[\frac{d\mu_{sj}(t)}{dt} - g_s(\boldsymbol{\mu}_j(t), \mathbf{\Theta}_j)\right]\left[-\frac{dg_s(\boldsymbol{\mu}_j(t), \mathbf{\Theta}_j)}{d{\mu_k}^2}\boldsymbol{\phi}_{kj}(t)\boldsymbol{\phi}_{kj}^T(t)\right] dt
\end{aligned}$$

The variance σ_k^2 is estimated by maximizing the log likelihood function, and the estimate is

$$\hat{\sigma}_k^2(\mathbf{\Theta}) = \sum_{i=1}^{n}\sum_{j=1}^{J}\left[\frac{P_{ijk}}{m_{ki}}(\mathbf{y}_{ki} - \mathbf{\Phi}_{kj}\widehat{\mathbf{c}}_{kj}(\mathbf{\Theta}))^T(\mathbf{y}_{ki} - \mathbf{\Phi}_{kj}\widehat{\mathbf{c}}_{kj}(\mathbf{\Theta}))\right].$$

References

Balasubramanian S, Schwartz C, Singh A, Warthmann N, Kim MC, et al. (2009) QTL Mapping in New *Arabidopsis thaliana* Advanced Intercross-Recombinant Inbred Lines. PLoS ONE 4(2): e4318.

Burden RL, Faires JD (2000) Numerical Analysis (7th ed.). Brooks/Cole. Pacific Grove, CA.

Chen J, Reynolds J (1997) A coordination model of carbon allocation in relation to water supply. Ann Bot 80: 45–55

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138: 963–971

Efron B, Tibshirani RJ (1993) An Introduction to the Bootstrap, Chapman & Hall.

Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, et al. (2010) Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature 464: 1039–1042

Frary A, Nesbitt TC, Grandillo S, Knaap E, Cong B, et al. (2000) fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289: 85–88

Huang Y, Wu H (2006) A Bayesian approach for estimating antiviral efficacy in HIV dynamic models. Journal of Applied Statistics 33: 155–174

Huang JZ, Liu N, Pourahmadi M, Liu L (2006) Covariance selection and estimation via penalized normal likelihood. Biometrika 93: 85–98

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185–199

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

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

Li H, Huang Z, Gai J, Wu S, Zeng Y, et al. (2007) A Conceptual Framework for Mapping Quantitative Trait Loci Regulating Ontogenetic Allometry. PLoS ONE 2(11): e1245.

Li L, Brown MB, Lee KH, Gupta S (2002) Estimation and inference for a spline-enhanced population pharmacokinetic model. Biometrics 58: 601–611

Liang H, Wu HL (2008) Parameter estimation for differential equation models using a framework of measurement error in regression model. Journal of American Statistical Association 103: 1570–1583

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

Ma CX, Casella G, Wu RL (2002) Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics 161: 1751–1762

Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics 10: 565–577

Putter H, Heisterkamp SH, Lange JMA, De Wolf F (2002) A Bayesian approach to parameter estimation in HIV dynamical models. Statistics in Medicine 21: 2199–2214

Ramsay JO (1996) Principal differential analysis: Data reduction by differential operators. Journal of Royal Statistical Society Series B 58: 495–508

Ramsay JO, Silverman BW (2005) Functional Data Analysis (2nd ed.). Springer, New York

Silverman Ramsay JO, Hooker G, Campbell D, Cao JG (2007) Parameter estimation for differential equations: a generalized smoothing approach (with discussion). Journal of Royal Statistical Society Series B 69: 741–796

Steinmetz LM, Sinha H, Richards DR, Spiegelman JI, Oefner PJ, et al. (2002) Dissecting the complex architecture of a quantitative trait locus in yeast. Nature 416: 326–330

West GB, Brown JH, Enquist BJ (1997) A general model for the origin of allometric scaling laws in biology. Science 276: 122–126

West GB, Brown JH, Enquist BJ (1999) The fourth dimension of life: Fractal geometry and allometric scaling of organisms. Science 284: 1677–1679

West GB, Brown JH, Enquist BJ (2001) A general model for ontogenetic growth. Nature 413: 628–631

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

Wu RL, Ma CX, Casella G (2002) Joint linkage and linkage disequilibrium mapping of quantitative trait loci in natural populations. Genetics 160: 779–792

Xu S, Atchley WR (1995) A random model approach to interval mapping of

quantitative genes. Genetics 141: 1189–1197

Yi N, Xu S (2008) Bayesian lasso for quantitative trait loci mapping. Genetics 179: 1045–1055

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136: 1457–1468

Zhang W-K, Wang Y-J, Luo G-Z, Zhang J-S, He C-Y, et al. (2004) QTL mapping of ten agronomic traits on the soybean (*Glycine* max L. Merr.) genetic map and their association with EST markers. Theoretical and Applied Genetics 108: 1131–1139

Zou F, Nie L, Wright FA, Sen PK (2009) A robust QTL mapping procedure. Journal of Statistical Planning and Inference 139: 978–989

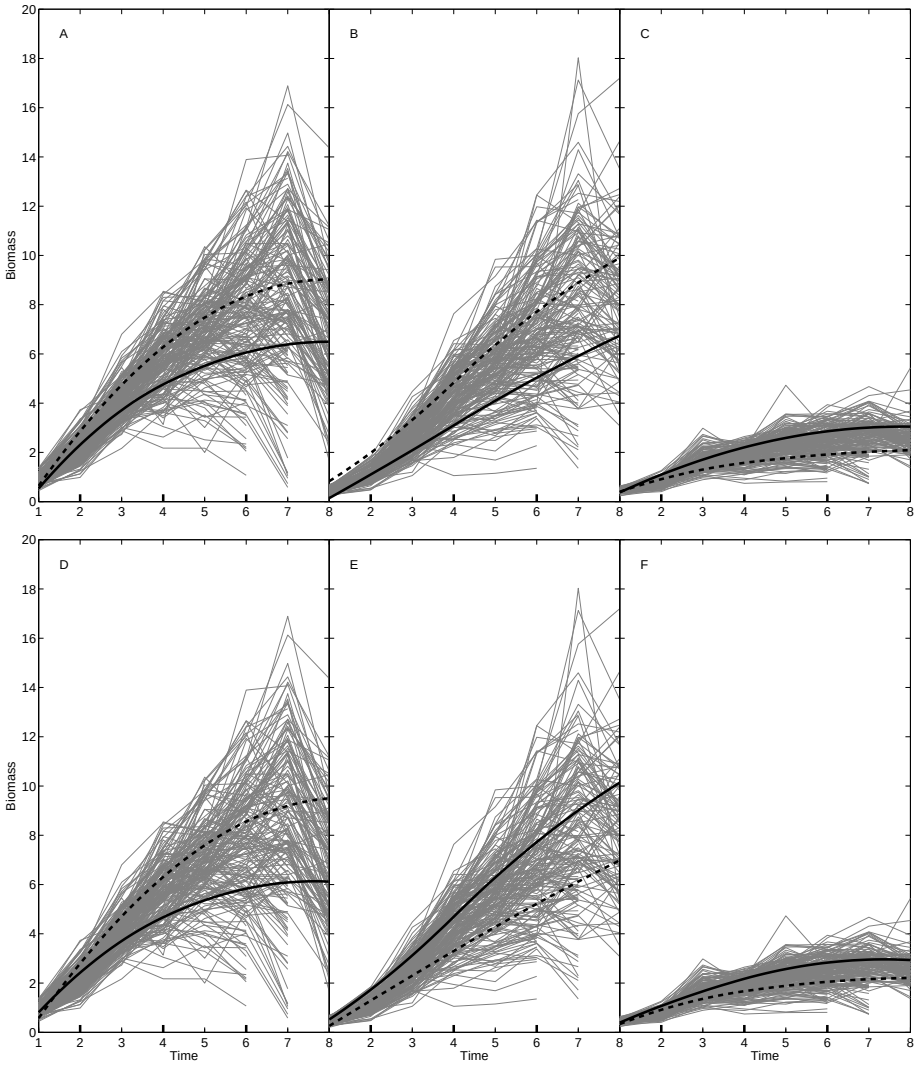

Figure 3: Growth trajectories of leaf (A, D), stem (B, E) and root biomass (C, F) for two different genotypes (presented by solid and broken black curves) at a QTL detected on linkage group 12 (upper panel) and 24 (lower panel), respectively. Two genotypes at a QTL are the homozygote for the alleles inherited from Kefeng No.1 (broken) and the homozygote for the allele from Nannong 1138-2 (solid). Curves in grey are growth trajectories of 184 RILs.

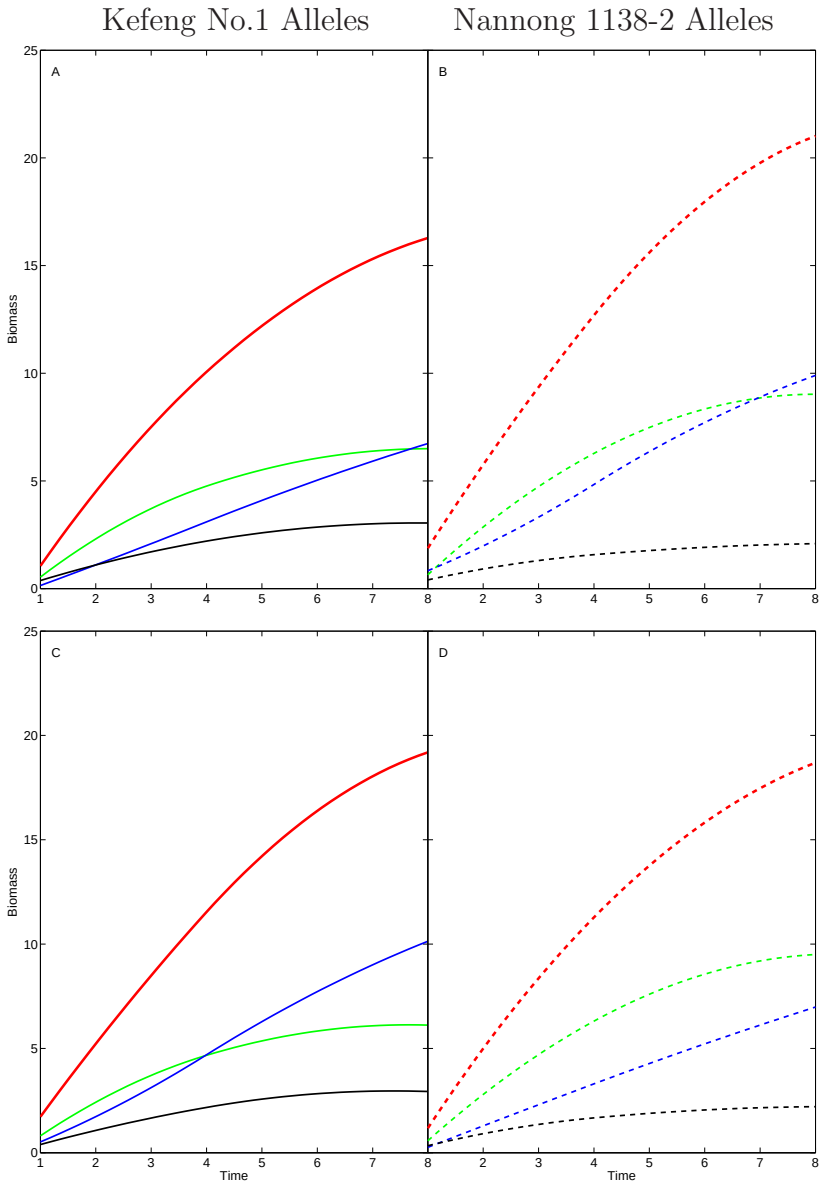

Figure 4: Growth trajectories of whole-plant (red), leaf (green), stem (blue) and root biomass (black) for two different genotypes at a QTL detected on linkage group 12 (upper panel) and 24 (lower panel). Two genotypes at a QTL are the homozygote for the alleles inherited from Kefeng No.1 (A, C) and the homozygote for alleles inherited from Nannong 1138-2 (B, D).

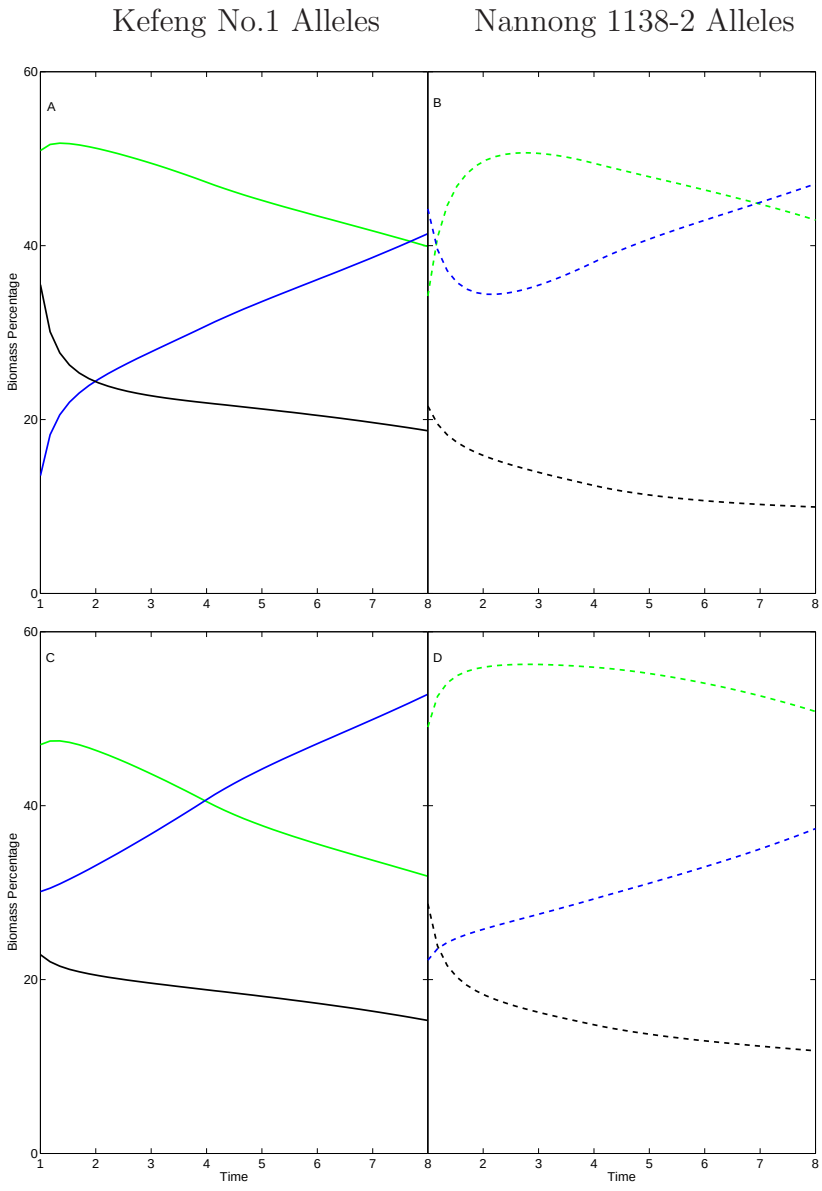

Figure 5: Time-dependent percentages of leaf (green), stem (blue) and root biomass (thin black) for two different genotypes at a QTL detected on linkage group 12 (upper panel) and 24 (lower panel). Two genotypes at a QTL are the homozygote for the alleles inherited from Kefeng No.1 (A, D) and the homozygote for alleles from Nannong 1138-2 (B, C).