An allometric model for mapping seed development in plants

Zhongwen Huang, Chunfa Tong, Wenhao Bo, Xiaoming Pang, Zhong Wang, Jichen Xu, Junyi Gai and Rongling Wu

Abstract

Despite a tremendous effort to map quantitative trait loci (QTLs) responsible for agriculturally and biologically important traits in plants, our understanding of how a QTL governs the developmental process of plant seeds remains elusive. In this article, we address this issue by describing a model for functional mapping of seed development through the incorporation of the relationship between vegetative and reproductive growth. The time difference of reproductive from vegetative growth is described by Reeve and Huxley's allometric equation. Thus, the implementation of this equation into the framework of functional mapping allows dynamic QTLs for seed development to be identified more precisely. By estimating and testing mathematical parameters that define Reeve and Huxley's allometric equations of seed growth, the dynamic pattern of the genetic effects of the QTLs identified can be analyzed. We used the model to analyze a soybean data, leading to the detection of QTLs that control the growth of seed dry weight. Three dynamic QTLs, located in two different linkage groups, were detected to affect growth curves of seed dry weight. The QTLs detected may be used to improve seed yield with marker-assisted selection by altering the pattern of seed development in a hope to achieve a maximum size of seeds at a harvest time.

Keywords: allometry; functional mapping; quantitative trait loci (QTL); developmental trait

INTRODUCTION

Although traditional breeding strategies based on phenotypic selection have substantially contributed to improvement in plant yield, quality and resistance, their further use has proved to be limited for selecting superior varieties [1–3]. This is mainly because many traits are complex, controlled by polygenes and their interactions with developmental signals. In the past 2 decades, tremendous developments in molecular marker technologies and statistical models have given rise to the revolution of genetic analysis approaches with which any phenotypic trait can be
dissected into its underlying genetic components, known as quantitative trait loci (QTLs), with DNA-based linkage maps. With those identified QTLs, the improvement of economically important traits in soybeans can be made more efficient and effective.

There has been a wealth of literature on the construction of genetic linkage maps and detection of QTLs for different traits in plants [4–6]. As one of the most important traits, genetic mapping of seed traits has received considerable attention. For example, 94 QTLs related to seed weight have been reported in soybeans [7–12], three of which have been confirmed [13, 14]. Despite these efforts, most studies ignore dynamic and developmental changes implicated in seed growth. More recently, Teng et al. [6] performed QTL mapping for seed weight by measuring its developmental behavior, but they used a traditional mapping approach based on individual time points. A novel statistical method for mapping dynamic traits, called functional mapping, has been developed in the literature [15–22]. Functional mapping implements mathematical aspects of biological principles to describe the changes of gene actions and interactions triggered by QTLs during trait development. Practical applications of functional mapping can be found for diameter and rooting ability in poplars [15, 23], programmed cell death in rice [24], plant height in soybeans [25] and body mass growth in mice [26]. The QTLs detected in these examples display different temporal patterns in governing the formation and expression of a trait during development.

The earlier work of functional mapping dealt with the dynamic growth of vegetative traits. Because reproductively growth may be initiated from different stages of vegetative growth, functional mapping for seed development needs the adjustment for time differences of reproductive from vegetative growth. Thus, to better describe the developmental trajectories of seed traits, we capitalized on Reeve and Huxley’s [27] allometric equation, which allows the initial value of a trait not to go through the origin. A mixture multivariate normal model was formulated with a correlation matrix structured by an autoregressive model of order one. Maximum likelihood estimates (MLEs) of unknown parameters in the model were obtained by implementing the EM algorithm and Nelder–Mead simplex method. Results about QTL detection from a biologically meaningful functional mapping model should have potential to improve plant seed yield with marker-assisted selection.

**STATISTICAL MODEL**

We modified the original statistical model for functional mapping to characterize QTLs for seed development by incorporating Reeve and Huxley’s [27] allometric equation. Evolved from the widely used allometry equation, \( y = ax^\beta \) [28, 29], by adding an additional parameter, \( \gamma \), Reeve and Huxley’s equation is written as

\[
y = \alpha(x - \gamma)^\beta
\]

(1)

where \( y \) is a biological dependent variable, \( x \) is the body mass, \( \alpha \) is a constant parameter and \( \beta \) is a power component. The additional parameter in Reeve and Huxley’s equation is considered as the point in ontogeny where the development of \( y \) begins relative to \( x \) [30].

Seeds develop from sexually mature plants. In practice, it is possible to investigate the time to initiate seeds. Because there is considerable variation in the timing of seed formation, we used Reeve and Huxley’s equation to capture this variation through parameter \( \gamma \). Genetic mapping relies on a segregating population, such as the F2, backcross, double haploids or recombinant inbred lines (RIL). We have constructed a high-density linkage map for a mapping population using molecular markers. An RIL population allows its individual progeny to be replicated genotypically. By planting multiple replicates for each RIL in a randomized block design, we can measure its whole-plant biomass and seed biomass at a series of time points by destructive sampling. For RIL \( i \), \( T_i \) time points, which can be either even-spaced or uneven-spaced, are measured. In an RIL population, there are two homozygous genotypes \( QQ \) (1) and \( qq \) (2) at a QTL, with allele \( Q \) derived from one parent and allele \( q \) derived from the second parent. For an RIL \( i \), the phenotypic value of its seed biomass, \( y_{ij} \) at time \( t_{ij} \) (\( j = 1, \ldots, T_i \)), affected by a QTL, can be expressed by a nonlinear regression model, expressed as

\[
y_{ij} = \sum_{k=1}^{2} z_{ik} \theta_k (x_{ij} - \gamma_k)^{\beta_k} + e_{ij}
\]

(2)

where \( x_{ij} \) is the whole-plant biomass of RIL \( i \) at time \( t_{ij} \); \( z_{ik} \) is an indicator variable with \( z_{ik} = 1 \) if this line has QTL genotype \( k \), otherwise \( z_{ik} = 0 \); \( \theta_k \); \( \beta_k \) and \( \gamma_k \) are unknown parameters that correspond to QTL genotype \( k \); and \( e_{ij} \) is a residual error assumed to be normally distributed with mean 0 and variance \( \sigma^2 \). The two QTL genotypes have different growth curves of seeds if the parameter set \((\theta_k, \beta_k, \gamma_k)\)
is genotype-dependent. Thus, by testing how these parameters differ between the two genotypes, we can identify the pattern of QTL effects on seed development.

If we have $n$ RILs, the likelihood of unknown parameters given phenotypic ($y$) and marker data ($M$) can be expressed, in terms of a mixture model, as

$$L(\Theta | y, M) = \prod_{i=1}^{n} \left[ \omega_{k1}f_1(y_i, \mu_{ik}) + \omega_{k2}f_2(y_i, \mu_{ik}) \right]$$  \hspace{1cm} (3)

where $\Theta$ contains unknown model parameters to be estimated; $\omega_{kl}$ is the conditional probability of QTL genotype $k$, conditional on the genotype of two flanking markers of RIL $i$, which are referred to as Wu et al. [31]; and $f_k(y_i, \mu_{ik})$ is a multivariate normal function of RIL $i$ that carries QTL genotype $k$, expressed as

$$f_k(y_i, \mu_{ik}) = \frac{1}{(2\pi)^{n/2} |\Sigma_k|^{1/2}}$$
$$\times \exp \left[ -\frac{1}{2} (y_i - \mu_{ik})^{\prime} \Sigma_k^{-1} (y_i - \mu_{ik}) \right]$$  \hspace{1cm} (4)

where $y_i = (y_{i1}, \ldots, y_{in})$ is the phenotypic vector of RIL $i$ measured at $T_i$ time points; $\mu_{ik} = (\mu_{ik1}, \ldots, \mu_{ikn})$ is the mean vector of QTL genotype $k$ for RIL $i$ measured at $T_i$ time points; and $\Sigma_k$ is the residual covariance matrix of RIL $i$ with $T_i$ repeated measurements.

For a longitudinal covariance matrix $\Sigma$, it is suggested that an appropriate statistical model be used to model its structure. A number of models have been available to model the covariance structure within the functional mapping framework [32, 33]. In a real example for QTL mapping using an RIL population of soybeans (Figure 1), it appears that time-varying variability in seed developmental trajectories can be modeled by a simple autoregressive model of order one [AR(1)] for covariance structure. As an example, here, we use the AR(1) by assuming that variance and covariance are stationary, expressed as

$$\Sigma = \begin{bmatrix}
1 & \rho^{\tau_{1}} & \ldots & \rho^{\tau_{n-1}} \\
\rho^{\tau_{1}} & 1 & \ldots & \rho^{\tau_{n-2}} \\
\vdots & \vdots & \ddots & \vdots \\
\rho^{\tau_{n-1}} & \rho^{\tau_{n-2}} & \ldots & 1
\end{bmatrix} \sigma^2$$  \hspace{1cm} (5)

where $\sigma^2$ is the variance and $\rho$ is the proportion parameter with which the correlation decays with time lag. For growth data like one in this example, other parametric approaches for covariance structure can be used, such as structured antedependence models that relax the assumptions of variance stationarity and covariance stationarity [34].

Likelihood Equation (3) contains three types of unknown parameters, i.e. QTL location described by $\omega_{kl}$, genotype-specific curve parameters ($\alpha_k$, $\beta_k$, $\gamma_k$) and covariance-structuring parameters ($\rho, \sigma^2$). To obtain the MLEs of these unknown parameters, we used a hybrid of the EM algorithm and Nelder–Mead simplex method. The algorithm is a direct-search optimization method for non-linear functions in low dimensions and it does not need any derivative information.

The hypothesis to test whether there exists a QTL affecting seed growth curves at a specific genomic position can be formulated as

$$H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2, \gamma_1 = \gamma_2$$
$$H_1 : \alpha_1 \neq \alpha_2, \beta_1 \neq \beta_2, \gamma_1 \neq \gamma_2$$  \hspace{1cm} (6)

where the $H_0$ corresponds to the reduced model and the $H_1$ corresponds to the full model. The log-likelihood ratio of the full model over the reduced model is applied to test the aforementioned hypotheses,

$$LR = -2 \log \left[ \frac{L_0(\hat{\Theta})}{L_1(\hat{\Theta})} \right]$$  \hspace{1cm} (7)

where $\hat{\Theta}$ and $\hat{\Theta}$ denote the MLEs of the unknown parameters under the $H_0$ and $H_1$, respectively. Permutation tests were performed to determine genome-wide critical threshold [35], by which a QTL is asserted to exist in a position of chromosome if a high peak of LR profiles exceeds the threshold.

Each plant experiences reproductive behavior only when it reaches a particular size through vegetative growth. Parameter $\gamma$ can describe the time delay of reproductive behavior relative to vegetative growth. Whether this time delay is controlled by the QTL can be tested by the hypotheses

$$H_0 : \gamma_1 = \gamma_2$$
$$H_1 : \gamma_1 \neq \gamma_2$$  \hspace{1cm} (8)

The rejection of the null hypothesis implies that the QTL triggers a significant effect on the amount of vegetative growth that is ready to initiate seeds.

**APPLICATION**

**Plant materials**

An RIL population of soybean derived from the cross between cultivars Kefeng No.1 and Nannong1138-2 was used to validate the model for seed-development
mapping. The population consists of 184 RILs whose first linkage map constructed from 452 markers was published by Zhang et al. [36]. This map was recently updated by adding some new SSR makers and dumping some unreliable markers. The new map contains 834 molecular makers covering a length of 2308 cM in 24 linkage groups, with an average genetic distance of 2.85 cM between adjacent markers.

In 2006, the RILs and their parents were planted in a 14 × 14 simple lattice design with two replications, in the National Center of Soybean Experiment, Jiangsu, Nanjing Agricultural University, China. Each RIL was planted in a 4 × 2.5 m² plot with five rows spaced 0.5 m apart. The lattice design used can reduce the field experiment error. Ten plants in the second row of a plot for each RIL were randomly selected for measuring seed dry weight at multiple times in the whole growing season. Starting 20 August 2006, 100 seeds were sampled from a single plant to measure their dry weights and then calculate the average dry weight per seed once every a week until seeds stop growing. Four to eight repeated measurements were taken for the RILs studied.

RESULTS
Figure 1 illustrates the plot of seed biomass over time for 184 RILs in soybeans. The plots of two parents Kefeng No.1 (P1) and Nannong 1138-2 (P2) are indicated by thick lines.

![Figure 1: Plots of seed biomass over time for 184 RILs in soybeans. The plots of two parents Kefeng No.1 (P1) and Nannong 1138-2 (P2) are indicated by thick lines.](image)

By analyzing this data set of seed growth, we detected three significant QTLs on linkage groups B1 and O, as evidenced by their LR values beyond the genome-wide threshold determined by permutation tests (Figure 2). On linkage group B1, two QTLs are claimed because their peaks are well separated (>30 cM) from each other. A summary of the estimates of QTL locations, genotype-specific curve parameters and covariance-structuring parameters is listed in Table 1, where the standard errors of each estimate obtained by re-sampling are also given. It
seems that all these parameters can be reasonably precisely estimated, although the time of reproductive delay is more difficult to estimate. The estimated curve parameters were used to illustrate the developmental trajectories of seed biomass for each genotype at each QTL detected. As shown in Figure 3, two genotypes perform differently in the timing of seed formation and growth rate. At the two QTLs detected on linkage group B1, the time to form seeds was much earlier for the genotypes with alleles inherited from parent Kefeng No. 1 than those with alleles from parent Nannong1138–2. This is consistent with the parental difference, as parent Kefeng No. 1 is an early genotype, whereas parent Nannong1138–2 is a late genotype. However, at these two QTLs, small parent Kefeng No. 1 contributes favorable alleles to increasing seed sizes, leading to large seeds for the progeny composed of the Kefeng No. 1 alleles than the Nannong 1138–2 alleles. It is interesting to note that the time of reproductive delay differs significantly ($P < 0.001$) between two genotypes at each of the three QTLs detected, as

**Table 1:** Maximum likelihood estimates of seed biomass trajectory parameters and their standard errors (in brackets) for different genotypes at each of the QTLs detected in an RIL population of soybeans

<table>
<thead>
<tr>
<th>Parameter estimate</th>
<th>QTL</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linkage group</td>
<td></td>
<td>BI</td>
<td>BI</td>
<td>O</td>
</tr>
<tr>
<td>Maker interval</td>
<td></td>
<td>Satt426—GMKF080</td>
<td>GMKF082c—GMKF168</td>
<td>GNE035—Sat 231</td>
</tr>
<tr>
<td>$\log(a_1)$</td>
<td></td>
<td>3.703 (1.161)</td>
<td>4.685 (1.265)</td>
<td>-15.445 (4.725)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td></td>
<td>1.790 (0.304)</td>
<td>2.055 (0.319)</td>
<td>4.621 (1.006)</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td></td>
<td>-0.211 (2.672)</td>
<td>-1.783 (2.879)</td>
<td>-14.193 (9.086)</td>
</tr>
<tr>
<td>$\log(a_2)$</td>
<td></td>
<td>-14.713 (5.496)</td>
<td>-12.222 (3.877)</td>
<td>-3.937 (1.58)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td></td>
<td>4.535 (1.189)</td>
<td>4.103 (0.888)</td>
<td>1.897 (0.303)</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td></td>
<td>-10.753 (9.894)</td>
<td>-3.680 (6.362)</td>
<td>-0.051 (2.543)</td>
</tr>
<tr>
<td>$P$</td>
<td></td>
<td>0.991 (0.001)</td>
<td>0.993 (0.001)</td>
<td>0.992 (0.001)</td>
</tr>
<tr>
<td>$\alpha$</td>
<td></td>
<td>0.773 (0.097)</td>
<td>0.931 (0.022)</td>
<td>0.831 (0.009)</td>
</tr>
<tr>
<td>LR</td>
<td></td>
<td>41.54</td>
<td>45.61</td>
<td>3999</td>
</tr>
</tbody>
</table>

**Figure 2:** The log-likelihood ratios (LR) of the full model (there is a QTL) over the reduced model (there is no QTL) at every 2 cm along the genetic linkage map composed of 950 molecular markers. Tick marks on the ceilings of each panel represent the positions of molecular markers on each linkage group. The dashed horizontal line indicates the genome-wide threshold value for asserting the existence of a QTL at the significant level 0.01.
tested with hypothesis test (8), suggesting that this trait is under genetic control.

The QTL on linkage group O follows a different inheritance mode from those detected on linkage B1. Large parent Nannong 1138-2 contributes favorable alleles to large seed sizes, and also it contributes alleles to early seed formation, although it is a late genotype. Different combinations of genotypes at different QTLs determine different growth patterns of seed biomass in the RIL population of soybeans. In general, a QTL shows an increasing genetic effect with time, reaches a peak of effects in a middle growing season and then decreases its effect at the late stage of growth. The effects of the two QTLs on linkage group B1 disappear when seeds start to stop growth (Figure 3).

DISCUSSION
Almost every biological trait, including seed size, experiences a developmental change in particular stages of an organism’s lifetime. Although previous models map QTLs for a phenotypic trait using its static measure, functional mapping can dissect and identify

Figure 3: Estimated growth trajectories of seed biomass for two genotypes (QQ inherited from the Kefeng No.1 parent and qq inherited from the Nannong 1138-2 parent) at each of the three QTLs detected on linkage groups B1 (A and B) and O (C). Growth trajectories for all RILs are indicated in gray.
QTLs that control the process of trait growth and development [15, 16, 20, 21], shifting QTL mapping from a static to dynamic context. To map seed development, however, functional mapping needs to accommodate the lag of seed formation behind vegetative growth. In this article, we integrated functional mapping and Reeve and Huxley’s [27] allometric equation. This equation, modified from a simple allometric power, was equipped with an additional parameter that defines the time lag of trait development. By applying the model to a mapping population for soybeans, we identified three genome-wide significant QTLs that control seed developmental trajectories.

In the past two decades, a large number of QTL mapping studies have been conducted for many important traits of plants by only considering one stage of development [6–12]. Because of unavailability of a dynamic mapping model, these studies were not able to model the temporal behavior of QTL genetic effects during a specific growth period. There have been a few publications that report the dynamic change of QTL effects [6, 37, 38] by capitalizing on traditional developmental quantitative genetic models pioneered by Atchley [39]. These studies offer an important view that, in different stages of growth, a set of genes is selectively expressed and the expression of one gene may be modified by interaction with other genes and environment.

Apart from the integration of developmental quantitative genetic models, function mapping implements biological principles of trait growth and development using mathematical equations, treating time-dependent longitude data as a whole. Thus, it is possible that functional mapping provides biologically more relevant results about the developmental pattern of genetic control for complex traits and predicts the dynamic behavior of QTL effects at any time points or stages in lifetime. In statistics, functional mapping provides enhanced power and precision for QTL detection and estimation through parsimonious modeling of mean-covariance structures.

The merit of functional mapping over traditional approaches can be seen from the replication of QTL detection. Of many QTLs identified for seed weight in soybeans, only a small portion is consistent among different experiments. By comparing our detection with previous ones, we found some consistency. For example, the QTLs we detected for seed biomass detected in the marker interval between GMKF104b and Satt509 on linkage group B1 are consistent with that detected by both Teng et al. [6] and Zhang et al. [36]. Another QTL detected in the interval between GNE035 and Sat 231 on linkage group O is in a similar genomic region of a QTL detected by Liu et al. [40]. It should be noted that two QTLs were suggested to exist in the same linkage group B1 because they are distant from each other. Zeng’s [41] composite interval mapping (CIM) was incorporated to judge whether these three QTLs are each a real one (results not shown). Surprisingly, CIM also identified multiple peaks, which suggests that there are rich QTLs harbored in this linkage group.

In this study, we identified three significant QTLs, which seem to be low given the complexity of seed development. We attribute this to two reasons. First, we must admit that our sample size (184) is relatively small, making those QTLs of small effects undetected. Based on our simulation studies, the detection of small-sized QTLs (with <0.05 heritability) requires at least 400 progeny to map dynamic traits. The simulation also shows that, once a QTL is detected using the current sample size, the false positive rate of QTL detection is low (<0.05–0.07). Second, previous work did report many QTLs, but they were mostly based on a certain LOD threshold, for example, LOD = 3.0. This criterion may be too low to minimize false-positive rates of QTL detection [42]. In contrast, we have used a stringent threshold obtained from permutation tests. In the future, by augmenting this study through increasing a sample size to 400, we expect that functional mapping will generate more exciting results about QTL detection for seed development. By then, we will be in an excellent position to chart a picture of the genetic and developmental architecture of this important trait.

**Key Points**

- Seeds are important carriers for plants to transmit their genes from generation to generation. Also, seeds play a central role in nourishing humans.
- As an important organ, the genetic control of seeds is still unclear; thus limiting the efficiency of practical breeding and selection for seed production and quality.
- Functional mapping has proven to be powerful for genetic mapping of dynamic traits, but it needs a special treatment to map seed development given the unique property of this process.
- We show that Reeve and Huxley’s allometric model can be integrated with functional mapping to study the interplay between genes and seed development.
FUNDING
Doctor Foundation of Henan Institute of Science and Technology (2008005); Education Department of Henan Province Natural Science Project (2011A210005); NSF/IOS-0923975; Changjiang Scholars Award; ‘Thousand-person Plan’ Award; National Key Basic Research Program (2009CB1184, 2010CB1259, 2011CB1093); Foundation for University Key Teacher by He’nan Educational Committee (2011GGJS-133).

References


