A statistical procedure to map high-order epistasis for complex traits


Submitted: 22nd February 2012; Received (in revised form): 27th April 2012

Abstract

Genetic interactions or epistasis have been thought to play a pivotal role in shaping the formation, development and evolution of life. Previous work focused on lower-order interactions between a pair of genes, but it is obviously inadequate to explain a complex network of genetic interactions and pathways. We review and assess a statistical model for characterizing high-order epistasis among more than two genes or quantitative trait loci (QTLs) that control a complex trait. The model includes a series of start-of-the-art standard procedures for estimating and testing the nature and magnitude of QTL interactions. Results from simulation studies and real data analysis warrant the
statistical properties of the model and its usefulness in practice. High-order epistatic mapping will provide a routine procedure for charting a detailed picture of the genetic regulation mechanisms underlying the phenotypic variation of complex traits.

**Keywords:** Epistasis; high-order interactions; quantitative trait loci; EM algorithm

**INTRODUCTION**

The past decade has been a critical period in which some phenomena related to genetic architecture are rerecognized. For example, epistasis has been thought to be an important force for evolution and speciation [1, 2], but recent genetic studies from vast quantities of molecular data have increasingly indicated that epistasis critically affects the pathogenesis of most inherited human diseases, such as cancer or cardiovascular disease [3–5], the developmental process and pattern of traits [6–8], susceptibility to HIV virus [9, 10] and virus drug resistance [11]. The expression of an interconnected network of genes is contingent upon environmental conditions, often with the elements and connections of the network displaying non-linear relationships with environmental factors [12]. Not only do these elements interact with each other in a pair-wise manner, they also form a complicated web of high-order interactions [13]. Because the embryonic expression pattern of a complex trait undergoes a sequence of metabolic pathways, such an interaction web should involve multiple interacting gene products and regulatory loci [8, 14–19].

Methodologically, it is a challenge to test and quantify genetic interactions among multiple genes. Genetic mapping with molecular linkage maps has proven to be powerful for the genome-wide detection of specific genes or quantitative trait loci (QTLs) for complex traits [20–23]. This approach has now been extended to search for epistatic interactions between different genes in controlled crosses [24], nuclear families [25], natural populations [26] and case-control designs [19]. Wu et al. [7] incorporated an epistatic model to study the genetic control of developmental trajectories for a complex trait. Several Bayesian approaches that allow an efficient search for pair-wise epistasis throughout the genome have been developed [27]. However, these strategies for genetic mapping can be equipped with genetic interactions among more than two QTLs, making it possible to elucidate a detailed picture of the genetic architecture of complex traits.

In a theoretical exploration by computer simulation, Stich et al. [16] found that genetic mapping has adequate power for the detection of three-way interactions while with a low false positive rate. Several authors showed the mathematical description of high-order epistasis from regulatory networks [28–30]. These advances in mathematical and statistical modeling of high-order epistasis provide an incentive to study this complex genetic phenomenon. The purpose of this article is to describe and assess a general procedure for a genome-wide search for high-order epistasis involving more than two QTLs using a genetic mapping strategy. This procedure integrates traditional quantitative genetic theory into genetic mapping, allowing the discernment of epistasis at different orders. The procedure is reviewed and tested in a genetic mapping study of rice with a doubled haploid population [31], in which significant three-way additive × additive × additive epistasis was identified. Computer simulation was used to investigate the statistical behavior of the model and algorithm for three-way epistatic mapping.

**HIGH-DIMENSIONAL GENETIC MODELING**

**Why high-order epistasis?**

Epistasis is the masking of the phenotype of one allele by the phenotype of an allele in another locus [32, 33]. Since a phenotypic trait involves an intricate network of biochemical reactions affected by multiple interacting gene products and regulatory loci, it is likely that genes generate higher-order epistatic interactions [6, 8, 15, 16, 30]. For example, maize (*Zea mays* L.) resists the corn earworm *Helicoverpa zea* (Boddie), a major insect pest of crops in the United States and elsewhere in the Western Hemisphere, because of the C-glycosyl flavones maysin, apimaysin and methoxymaysin synthesized in silks [34]. As a resistance phenotype, the biosynthesis of maysin, apimaysin and methoxymaysin undergoes a complex network of metabolic pathways. Figure 1 illustrates a branch of the well
characterized flavonoid pathway in which each step and reaction are regulated by genes [15]. In order for maysin to be synthesized, alleles at the following genes should coordinate appropriately, \( p_1 \), \( c_2 \) and/or \( whpl \) (encoding chalcone synthases [35]), \( chi1 \) (encoding chalcone isomerase [36]), \( pr1 \) (controlling the 3’-hydroxylation of the flavonoid B-ring to convert monohydroxy to dihydroxy compounds [37]) and unidentified additional loci encoding flavone synthase, C-glycosyl transferase, glucose oxidase, rhamnosyl transferase and an enzyme such as glutathione S-transferase for transport to the vacuole [38, 39] (Figure 1). McMullen et al. [15] argued that higher-order epistatic interactions among multiple genes at different levels of biochemical pathways are a determinant of final maysin synthesis. The occurrence of high-order epistasis entails the development of high-dimensional model for gene detection.

**Figure 1:** A branch of biochemical pathways for flavones synthesis in maize. CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone-3-hydroxylase; DFR, dihydroflavanone reductase; F3’H, flavanone-3’-hydroxylase; FNS, flavone synthase; RT, rhamnosyl transferase. Adapted from [15].

---

**Quantitative genetic model for high-order epistasis**

The formation of a final phenotype is the consequence of sequential genetic interactions involved in biochemical and metabolic networks. Quantitative genetic theory has well been established to describe pair-wise epistasis by partitioning it into different components [40, 41]. Here, we extend this theory to study high-order epistasis among three or more genes. Consider three QTLs \( Q_1, Q_2, Q_3 \), each with two alleles \( Q \) and \( q \), which control a complex trait. Let \( j_k \) denote one of three genotypes at QTL \( Q_k (k = 1, 2, 3) \). The genotypic value of a 3-QTL genotype, denoted as \( y_{j_1j_2j_3} \) (\( j_k = 0 \) for \( q_kq_k \), 1 for \( Q_kq_k \), 2 for \( Q_kQ_k \)), can be partitioned into different components including the main effects, two-way interaction effects and three-way interaction effects [19], i.e.
where $\mu$ is the overall mean; $a_1$, $a_2$ and $a_3$ are the main additive genetic effects, $d_1$, $d_2$ and $d_3$ are the main dominant effects at QTLs $Q_1$, $Q_2$, $Q_3$, respectively; $i_{a_1a_2}$, $i_{a_1d_2}$, $i_{a_2a_3}$, $i_{a_2d_3}$, $i_{a_3a_1}$, $i_{a_3d_1}$, $i_{d_1a_1}$, $i_{d_1d_2}$, $i_{d_2a_2}$, $i_{d_2d_3}$, $i_{d_3a_3}$, $i_{d_3d_1}$, $i_{d_1a_2}$, $i_{d_2a_3}$ are the two-way additive × additive, additive × dominant, dominant × additive and dominant × dominant epistasis between QTLs $Q_1$ and $Q_2$; between QTLs $Q_1$ and $Q_3$, and between QTLs $Q_2$ and $Q_3$, respectively; $i_{a_1a_2a_3}$, $i_{a_1d_2d_3}$, $i_{a_2a_3d_1}$, $i_{a_2d_3d_1}$, $i_{d_1a_1d_2}$, $i_{d_1d_2d_3}$, $i_{d_2a_2d_3}$, $i_{d_2d_3d_1}$, $i_{d_3a_3d_1}$, $i_{d_3d_1d_2}$ are the three-way additive × additive × additive, additive × additive × dominant, additive × dominant × additive, dominant × additive × additive, dominant × additive × dominant, dominant × additive × dominant, dominant × dominant × additive and dominant × dominant × dominant epistasis among QTLs $Q_1$, $Q_2$, $Q_3$, respectively.

The genetic effect parameters are then solved from the genotypic values:

$$
\mathbf{p}^T = \{p_{a_1a_2a_3}, p_{a_1d_2d_3}, p_{a_1d_1d_2}, p_{a_1d_1d_3}, p_{a_1d_2d_3}, p_{a_2a_3d_1}, p_{a_2d_3d_1}, p_{d_1a_1d_2}, p_{d_1d_2d_3}, p_{d_2a_2d_3}, p_{d_2d_3d_1}, p_{d_3a_3d_1}, p_{d_3d_1d_2}\}
$$
Using these expressions, we can test the significance of each genetic effect. The model can be extended to characterize high-order epistasis among any number of QTLs.

**Approaches for mapping high-order epistasis**

Genetic mapping founded on quantitative genetic theory has been used to study genetic interactions. Consider an F2 mapping population, derived from two inbred lines, in which all progeny is genotyped for a panel of molecular markers to construct a genetic linkage map and phenotyped for a complex trait [23]. The likelihood of trait values, determined by the model composed of 3^3 genotype components, i.e.

\[ L(y) = \prod_{i=1}^{n} \sum_{j_1=0}^{2} \sum_{j_2=0}^{2} \sum_{j_3=0}^{2} \left( \sum_{j_{ij_1j_2j_3}} \omega_{ij_1j_2j_3} \varphi_{ij_1j_2j_3}(y_i) \right) \]  

where \( y_i \) is the phenotypic value of the trait for progeny \( i \), \( \omega_{ij_1j_2j_3} \) is the probability at which an arbitrary progeny \( i \) is QTL genotype \( j_{ij_1j_2j_3} \) and \( \varphi_{ij_1j_2j_3}(y_i) \) is the normal density function of progeny \( i \) with genotypic mean \( \mu_{ij_1j_2j_3} \) and variance \( \sigma^2 \). Since the QTL genotype of a progeny is unknown but can be inferred from its marker genotype, \( \omega_{ij_1j_2j_3} \) is actually a conditional probability of QTL genotype \( j_{ij_1j_2j_3} \) given the marker genotype of progeny \( i \), which can be expressed in terms of the recombination fractions between the QTLs and markers [23]. In the likelihood (3), we have the unknown parameters, arrayed in \( \Theta \), including QTL positions, genotypic means and variance.

The parameters can be estimated by maximizing the likelihood (3). This can be done by differentiating the likelihood with respect of individual parameters \( \theta \) (\( \theta \in \Theta \)), i.e.

\[ \frac{\partial}{\partial \theta} \log L(y) = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log \sum_{j_{ij_1j_2j_3}} \omega_{ij_1j_2j_3} \varphi_{ij_1j_2j_3}(y_i) \]

\[ = \sum_{i=1}^{n} \sum_{j_{ij_1j_2j_3}} \sum_{j_{ij_1j_2j_3}} \sum_{j_{ij_1j_2j_3}} \omega_{ij_1j_2j_3} \varphi_{ij_1j_2j_3}(y_i) \frac{\partial}{\partial \theta} \varphi_{ij_1j_2j_3}(y_i) \]

\[ = \sum_{i=1}^{n} \sum_{j_{ij_1j_2j_3}} \sum_{j_{ij_1j_2j_3}} \sum_{j_{ij_1j_2j_3}} \Omega_{ij_1j_2j_3} \frac{\partial}{\partial \theta} \varphi_{ij_1j_2j_3}(y_i) \]

where

\[ \Omega_{ij_1j_2j_3} = \frac{\omega_{ij_1j_2j_3} \varphi_{ij_1j_2j_3}(y_i)}{\sum_{j_{ij_1j_2j_3}} \sum_{j_{ij_1j_2j_3}} \sum_{j_{ij_1j_2j_3}} \omega_{ij_1j_2j_3} \varphi_{ij_1j_2j_3}(y_i)} \]  

is interpreted as the posterior probability that progeny \( i \) has QTL genotype \( j_{ij_1j_2j_3} \). Substituting (5) into (4), we obtain the formulas to estimate genotypic means and variance expressed as

\[ \mu_{ij_1j_2j_3} = \frac{\sum_{i=1}^{n} \Omega_{ij_1j_2j_3} y_i}{\sum_{i=1}^{n} \Omega_{ij_1j_2j_3}} \]  

\[ \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j_{ij_1j_2j_3}} \sum_{j_{ij_1j_2j_3}} \Omega_{ij_1j_2j_3} (y_i - \mu_{ij_1j_2j_3})^2 \]  

The EM algorithm [20] is implemented to estimate \( \mu_{ij_1j_2j_3} \) and \( \sigma^2 \) by using an iterative procedure between the E step (5) and M steps (6) and (7). The values at convergence are the maximum-likelihood estimates (MLEs) of \( \mu_{ij_1j_2j_3} \) and \( \sigma^2 \). After obtaining the estimates of genotypic means, we can solve for the estimates of the genetic effect parameters using Equation (2). In practice, the QTL positions are estimated by treating \( \Omega_{ij_1j_2j_3} \) as a fixed parameter, scanning the entire genome and detecting the largest likelihood that corresponds to the best estimation of QTL positions.

**Hypothesis tests**

When no QTL is segregating, only one normal density can describe the population in which case no EM algorithm is needed for parameter estimation. The existence of QTLs can be tested by calculating and comparing the likelihoods under the null hypothesis \( H_0 \): there is no QTL, \( L(\tilde{\Theta}) \) and the alternative hypothesis \( H_1 \): there is at least one QTL, \( L(\hat{\Theta}) \). The resulting log-likelihood ratio (LR) test statistics is

\[ LR = 2[\log L(\hat{\Theta}) - \log L(\tilde{\Theta})] \]

where \( \tilde{\Theta} \) and \( \hat{\Theta} \) are the MLEs of unknown parameters under the \( H_0 \) and \( H_1 \), respectively. The significance of the result can be tested by using permutation tests [42]. By reshuffling the phenotypic data and calculating the LR genome-wide for each permutation, a critical threshold is obtained at a particular significance level.
A procedure can also be given to test different components of genotypic values including the additive \((a_1, a_2, a_3)\) and dominant main genetic effects \((d_1, d_2, d_3)\) at individual QTLs, two-way epistatic interactions of 12 different types \((i_{a_1a_2}, i_{a_1d_2}, i_{a_2d_1}, i_{d_1d_2}, i_{a_3a_2}, i_{d_1d_3}, i_{a_3d_1}, i_{a_2d_1}, i_{a_3d_2}, i_{d_2d_3}, i_{d_3d_1}, i_{d_3d_2})\), and three-way epistatic interactions of eight different types \((i_{a_1a_2a_3}, i_{a_1a_2d_2}, i_{a_1a_2d_3}, i_{a_1a_3d_2}, i_{a_1d_2d_3}, i_{a_2a_3d_2}, i_{a_2a_3d_3}, i_{a_2d_2d_3}, i_{d_1d_2d_3})\). All these components are calculated from genotypic means using a group of equations (1). When we want to test whether one or more of the 26 effects equals zero, we will only need to estimate the remaining effects. In such a reduced model, we will use the same EM algorithm for parameter estimation described for the full model (5)–(7), but with the constraint(s) that poses on the \(t\)imate the remaining effects. In such a reduced model, represented by \(\mathbf{y}_A = \mathbf{X}_A \mathbf{\beta}\), i.e.

\[
y_{ij\bar{b}j|i} = \Omega_{ij\bar{b}j|i} y_i,
\]

where \(\Omega_{ij\bar{b}j|i}\) is the posterior probability obtained from the E-step (5);

(2) Define a vector of dummy variable \(X_{ij\bar{b}j|i}\) that meets

\[
E(y_{ij\bar{b}j|i}) = X_{ij\bar{b}j|i} \mathbf{\beta},
\]

where \(\mathbf{\beta}\) is the vector of genetic effect parameters (2);

(3) By arranging the augmented data in a linear model framework, we have

\[
y_A = X_A \mathbf{\beta},
\]

where \(y_A = \{y_{ij\bar{b}j|i}\}\) and \(X_A = \{X_{ij\bar{b}j|i}\}\). For a given reduced model, represented by \(K^T \mathbf{\beta} = 0\), where \(K\) is a vector that constrains a certain effects to be equal to zero, we have

\[
\hat{\mathbf{\beta}}_K = \hat{\mathbf{\beta}} - (X_A^T X_A)^{-1}K^T(X_A^T \mathbf{\beta} - X_A^T \mathbf{\beta}),
\]

where \(\mathbf{\beta} = (X_A^T X_A)^{-1}(X_A^T y_A);\) (4) The variance in the reduced model is estimated by

\[
\hat{\sigma}_K^2 = \frac{1}{n} \sum_{i=1}^n \Omega_{ij\bar{b}j|i} (y_i - X_{ij\bar{b}j|i} \hat{\mathbf{\beta}}_K)^2,
\]

where \(n\) is the total number of progeny in the mapping population.

The iteration is made between the E step (5) and M step (8) and (9) until the stable values are obtained. These stable values are the MLEs of the parameters under the reduced model. In each case of testing the significance of effect parameters, we calculate the likelihoods under the null and alternative hypotheses and therefore the LRs. The critical thresholds for testing each effect can be obtained from simulation approaches [23].

To reduce the computing burden for threshold determination, several formulae have been derived for computing approximate critical thresholds to control the type I error rate at a chromosome- or genome-wide level [44, 45]. Chang et al. [46] proposed a score test statistic for QTL mapping. The score test is computationally simpler than the LR test, since it only uses the MLEs of parameters under the null hypothesis. More importantly, the maximum of the square of score statistics are asymptotically equivalent to the maximum of the LR test statistics under the null hypothesis, thus the critical threshold for the score test can also be used for the LR test, which can improve the computing efficiency of threshold determination.

**MODEL VALIDATION**

**Worked example**

Two rice cultivars, semi-dwarf IR64 and tall Azucena, was crossed to generate a doubled-haploid (DH) population. Using 135 DH lines from this population, a genetic linkage map was constructed, covering 12 chromosomes with 175 molecular markers [31]. The DH population was grown in a randomized complete design with two replicates at a spacing of 15 × 20 cm in a field near Hangzhou, China. Final plant heights were measured for each plant. To reduce random errors in height phenotypes, we took the mean of two replicates for each DH line, used for QTL mapping. Indeed, two replicates can be incorporated into the model in a way, as shown by Wu et al. [47], which takes into account
the spatial correlation of phenotypic values due to microenvironmental factors.

Marker and QTL co-segregation in a DH population follows a backcross pattern. In a DH population, each locus has two homozygotes (denoted as 1 and 0, respectively) each inheriting two alleles from a different parent. Thus, three QTLs form eight different homozygotes, denoted as $j_1j_2j_3$ ($j_1$, $j_2$, $j_3 = 1$, 0), with genotypic value $\mu (j_1j_2j_3)$ which can be partitioned into different components as follows:

\[
\begin{align*}
\mu_{111} &= \mu + a_1 + a_2 + a_3 + i_{a_1a_2} + i_{a_1a_3} + i_{a_2a_3} \\
\mu_{110} &= \mu + a_1 + a_2 - a_3 + i_{a_1a_2} - i_{a_1a_3} - i_{a_2a_3} \\
\mu_{101} &= \mu + a_1 - a_2 + a_3 - i_{a_1a_2} + i_{a_1a_3} - i_{a_2a_3} \\
\mu_{100} &= \mu + a_1 - a_2 - a_3 - i_{a_1a_2} - i_{a_1a_3} - i_{a_2a_3} \\
\mu_{011} &= \mu - a_1 + a_2 + a_3 - i_{a_1a_2} - i_{a_1a_3} + i_{a_2a_3} \\
\mu_{010} &= \mu - a_1 + a_2 - a_3 - i_{a_1a_2} + i_{a_1a_3} - i_{a_2a_3} \\
\mu_{001} &= \mu - a_1 - a_2 + a_3 - i_{a_1a_2} + i_{a_1a_3} + i_{a_2a_3} \\
\mu_{000} &= \mu - a_1 - a_2 - a_3 + i_{a_1a_2} + i_{a_1a_3} + i_{a_2a_3} - i_{a_1a_2a_3}
\end{align*}
\]

where $\mu$ is the overall mean; $a$‘s are the additive effects at different QTLs, and $i$‘s are two- or three-way epistatic interactions between the QTLs. The procedure for mapping high-order epistasis was used to estimate the additive, additive $\times$ additive, additive $\times$ additive $\times$ additive genetic effects on plant height in this DH population.

By simultaneously searching for three QTLs at every 4 cm throughout the entire genome, we detected three locations, markers RG403–RG229 on chromosome 5, markers RZ37B–CDO497 on chromosome 7, and markers RG667–RG451 on chromosome 9, which jointly affect plant heights. Figure 2 shows a portion of the overall LR plot against the searched positions S, with the LR peak (64, 288, 440) that corresponds to the locations of three QTLs on Chromosomes 5, 7 and 9. Three-dimensional plots are displayed for cycles 1–4 in Figure 3, where the $x$, $y$-coordinates are the $b$ and $c$.
search positions, respectively, and the z-coordinate is the LR value.

Table 1 gives the MLEs of the locations and genetic effects of the three QTLs detected. The QTL on chromosome 9, linked to marker RG667, triggers a highly significant additive genetic effect \( a_3 \) on plant height, but the additive effects of the two QTLs on chromosomes 5 and 7 are not significant. The genomic region of marker RG667 was shown to harbor a QTL that affects a plant height-corrected trait, number of productive tillers in rice [48]. The alleles derived from the tall parent Azucena contribute favorably to height growth. There is a significant two-way interaction epistasis \( i_{a_1a_2} \) between the additive effects at the significant QTL on chromosome 5 and the non-significant QTL on chromosome 9, but the ‘tall’ alleles (derived from parent Azucena) at each of these two QTLs interact to inhibit the plant height growth of rice. It is interesting to find that a highly significant three-way interaction epistasis \( i_{a_1a_2a_3} \) occurs among the additive effects at the three QTLs by favorably increasing rice height growth.

### Computer simulation

We simulated a backcross design with two genotypes 1 and 0 at each locus and randomly generated nine markers equally spaced in a linkage map of 225 cM. Let three putative QTLs be located at 10 cM each from the third, sixth and eighth markers, respectively. Phenotypic values were then simulated by summing the genotypic value of a specific three-QTL genotype and residual errors that follow a normal distribution with mean zero and variance \( \sigma^2 \). The simulation studies were designed for different sample sizes \( n = 100 \) and 400) and different heritabilities \( H^2 = 0.1 \) and 0.4). The values of the residual variance were determined, depending on the level of heritability.

Tables 2 and 3 tabulate the estimates of QTL locations and effect parameters from the simulated data based on 100 simulation replicates. In general, the locations and genotypic values of the QTLs and residual variance can well be estimated even when there is a modest sample size (100) and heritability (0.1) (Table 2). Increasing sample sizes and heritabilities can remarkably improve the estimation precision of these parameters. The estimation precision of the genetic effect parameters depends on the type of the parameters (Table 3). The additive effect can be estimated most precisely, followed by two-way interaction effects and three-way interaction effects. Also, the estimates of the genetic effect parameters were found to be sensitive to sample size and/or heritability. Yet, the increase of heritability from 0.1 to 0.4 produces a much better efficiency in improving estimation precision than that of sample size from 100 to 400. This suggests that a better management of plants, aimed to minimize experimental errors, will contribute more substantially to mapping precision than a simple increase of sample size.

In general, if a trait has a high heritability, a sample size of 100 is adequately enough for the reasonable estimation of the additive and epistatic effects. For a modest heritability (say 0.1), 400 samples are needed. By increasing sample size to 1000, it was found that all estimates can be improved even when the trait has a low heritability \( \leq 0.05 \). It is always important to investigate the power of the model to identify significant genetic effects given a particular sample size. We calculated the power using computer simulation. With the value of a particular genetic effect, which is used to determine the magnitude of residual variance for a given heritability, phenotypic and marker are simulated. The proportion of the number of significant simulation replicates (i.e. those in which the effect is found to be significant) over the total number of simulation replicates is empirically regarded as statistical power for identifying this genetic

### Table 1: The MLEs of the QTL locations and effects for plant height growth in a DH mapping population of rice

<table>
<thead>
<tr>
<th>Chr</th>
<th>Marker interval</th>
<th>Map distance</th>
<th>Main effect</th>
<th>Two-way epistasis</th>
<th>Three-way epistasis</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>RG403-RG229</td>
<td>2.1 cM</td>
<td>( a_1 = 1.05 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>RZ337B-CDO497</td>
<td>10.7 cM</td>
<td>( a_2 = 1.05 )</td>
<td>( i_{a_1a_2} = -4.76 )</td>
<td>( i_{a_1a_2} = -8.77 )</td>
</tr>
<tr>
<td>9</td>
<td>RG667-RG451</td>
<td>15.6 cM</td>
<td>( a_3 = 11.52 )</td>
<td>( i_{a_1a_2} = -4.76 )</td>
<td>( i_{a_1a_2} = 15.06 )</td>
</tr>
</tbody>
</table>

Map distance means the distance of the QTL from the left marker for an interval.
We ran an additional simulation to examine the power of the model for detecting three-way interactions. For a quantitative trait with a modest heritability (say 0.1), we detected power of ~0.70 for detecting three-way epistasis if a sample size of 400 was used. Under the same heritability, the power increases to >0.9 if the sample size increases to 800.

**DISCUSSION**

Our understanding of how the information contained in genotypes is translated into complex phenotypic traits represents a major challenge in biological research. Although its precise description has not been clear yet, existing evidence shows that this process undergoes a multilayered hierarchy of regulatory networks in which genes and products from different levels or stages interact and coordinate to form a final phenotype. It is highly likely that interactions involving more than genes, i.e. so-called high-order interactions should play a central role in coordinating the networks [29, 30]. Although the contribution of epistatic interactions to quantitative genetic variation has been increasingly recognized by population and evolutionary biologists [1, 2, 4] and medical geneticists [5, 33], the impact of high-order epistasis on phenotypic diversity has not been carefully explored. Results from a limited number of quantitative genetic studies show that high-order epistasis could be correlated with some certain cytological phenomena [49] and growth traits [50].

In this article, we review a mapping model for characterizing genetic interactions of multiple orders that are responsible for complex traits. The model was founded on the general framework of genetic mapping with molecular maps, allowing the genome-wide search for multilocus interactions throughout the genome. The model shows several advantages. First, it can test the relative importance of different types of genetic effects including the main additive, low-order epistasis and high-order epistasis and, thus, provides an unprecedented opportunity to study the detailed atlas of genetic control mechanisms for complex phenotypes. From a biological perspective, it is possible that a single gene does not trigger a significant effect on a phenotype, but exerts a remarkable impact on the phenotype through epistasis with other genes involved in key pathways that form the final phenotype [29, 30, 33]. Second, we derived a closed form for estimating the genetic effects of different types,
facilitating the computational efficiency and implementation into a package of computer software. More interesting, the closed form for parameter estimation exist for reduced models in which our test is to focus on a particular subset of parameters. The model was validated by reanalyzing a real data set for genetic mapping of plant heights in rice [31]. On one hand, this reanalysis has well warranted the usefulness and utilization of the model in practice.

On the other hand, new discoveries for the genetic control of plant height growth in rice have been made by using the new model. In previous studies, several significant QTLs have been detected for height growth in this mapping population [51, 52]. For example, Zhao et al. [52] identified these QTLs on chromosomes 1, 3, 7, 9 and 11. In addition to the same QTLs identified on chromosomes 7 and 9, the new model has also detected a QTL on chromosome

### Table 3: The averaged MLEs of the QTL positions and genetic effects and their standard errors (given in parentheses) under different sample sizes ($n$) and heritabilities ($H^2$) based on 100 simulation replicates.

<table>
<thead>
<tr>
<th>$H^2$</th>
<th>$n$</th>
<th>$\mu$</th>
<th>Additive effect</th>
<th>Interaction effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$a_1$</td>
<td>$a_2$</td>
</tr>
<tr>
<td>0.1</td>
<td>100</td>
<td>150.20(0.15)</td>
<td>-0.17(0.14)</td>
<td>-1.08(0.16)</td>
</tr>
<tr>
<td>0.1</td>
<td>400</td>
<td>150.15(0.11)</td>
<td>-0.51(0.13)</td>
<td>-1.01(0.12)</td>
</tr>
<tr>
<td>0.4</td>
<td>100</td>
<td>150.13(0.06)</td>
<td>-0.78(0.06)</td>
<td>-0.95(0.07)</td>
</tr>
<tr>
<td>0.4</td>
<td>400</td>
<td>150.03(0.02)</td>
<td>-0.97(0.02)</td>
<td>-0.97(0.02)</td>
</tr>
<tr>
<td>True Value</td>
<td></td>
<td>150</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

**Figure 3:** The LR value in three dimensions. The four plots correspond to the four cycles in Figure 2. The rectangle is the 200 permutation cutoff.
5. Although this new QTL has no significant main additive effect, it functions epistatically with one on chromosome 9 to determine the final plant height of rice. It is very interesting to find that these two QTLs, along with one on chromosome 7, display a highly significant three-way additive × additive × additive epistatic effect through increasing or decreasing ~15 cm in plant height. It should be noted that the detection of more QTLs by Zhao et al. [52] than the new model may be due to the fact that the former makes use of height data at multiple time points. To test the statistical behavior of the new model, simulation studies were performed, suggesting that three-way epistasis can well be estimated when an adequately large sample size is used.

With the model described in this article, the investigation of whether three-way epistasis is a widespread phenomenon in plant height can now be made possible by reanalyzing published mapping data or designing new mapping experiments. In practice, it is impossible to precisely estimate a genetic effect from a single study. The uncertainty of chromosomal locations and genetic effects for QTLs can be overcome through replicating the same experiments. The estimates of these QTL parameters from multiple replicates are closer to the reality of the parameters.

As genome-wide association studies (GWAS) have emerged as a useful tool for plant, animal and human genetics [53–57], it is crucial to incorporate the multilocus epistasis detection model to illustrate a network of genetic interactions throughout the genome. In the current model specification, we do not consider environmental factors. Given its importance, genotype × environment interaction should be integrated into the high-order epistasis model [12]. Also, it is worthwhile to model the pleiotropic effects of high-order epistasis on different aspects of phenotypic traits [58, 59]. The major factor of limiting these extensions is the combinatorial search of too many interactions on a much smaller number of samples. However, the recent availability of feature selection methods [60], equipped with efficient computing algorithms, such as genetic programming [13], provides an unprecedented opportunity to produce a useful statistical toolbox for dissecting complex phenotypes into their genetic components at different levels. The computer code used to detect and test high-order epistasis is available at http://statgen.psu.edu.

**Key Points**

- Despite considerable efforts to dissect the genetic architecture of complex traits, much still remains unclear including the distribution, mechanisms and importance of genetic interactions.
- High-order epistasis due to multilevel interactions of genes is thought to be the hidden genetic variation that has not been utilized in agriculture and biomedicine.
- Genetic mapping, now used as a routine approach for studying the genetic regulation of quantitative traits, has a unique power to characterize pair-wise epistatic interactions. However, there is still no in-depth exploration to estimate and test the genetic effects due to interactions among three or more genes.
- We formulate and assess a state-of-the-art statistical procedure of implementing genetic mapping to detect high-order epistasis, filling a gap that occurs in quantitative genetics, evolutionary genetics and medical genetics. We argue that high-order epistatic mapping can serve as a routine tool to comprehend the genetic architecture of complex traits.

**ACKNOWLEDGEMENTS**

The authors thank Prof. Jun Zhu at Zhejiang University for providing the rice data to validate the high-order epistatic mapping model.

**FUNDING**

The Special Fund for Forestry Scientific Research in the Public Interest (No. 201004017), NSF/IOS-0923975; the Changjiang Scholarship Award; and ‘Thousand-person Plan’ Award.

**References**


50. Wu RL. Detecting epistatic genetic variance with a clonally replicated design: models for low-


