RESEARCH ARTICLE

An Ultrahigh-Dimensional Mapping Model of High-order Epistatic Networks for Complex Traits

Kirk Gosik, Lidan Sun, Vernon M. Chinchilli and Rongling Wu*

Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033, USA

Abstract: ***Background*:** Genetic interactions involving more than two loci have been thought to affect quantitatively inherited traits and diseases more pervasively than previously appreciated. However, the detection of such high-order interactions to chart a complete portrait of genetic architecture has not been well explored.

***Methods*:** We present an ultrahigh-dimensional model to systematically characterize genetic main effects and interaction effects of various orders among all possible markers in a genetic mapping or association study. The model was built on the extension of a variable selection procedure, called iFORM, derived from forward selection. The model shows its unique power to estimate the magnitudes and signs of high-order epistatic effects, in addition to those of main effects and pairwise epistatic effects.

***Results*:** The statistical properties of the model were tested and validated through simulation studies. By analyzing a real data for shoot growth in a mapping population of woody plant, mei (*Prunus mume*), we demonstrated the usefulness and utility of the model in practical genetic studies. The model has identified important high-order interactions that contribute to shoot growth for mei.

***Conclusion*:** The model provides a tool to precisely construct genotype-phenotype maps for quantitative traits by identifying any possible high-order epistasis which is often ignored in the current genetic literature.

ARTICLE HISTORY

Received: January 02, 2017
Revised: March 28, 2017
Accepted: May 04, 2017

DOI:
10.2174/1389202919666171218162210

Keywords: Variable selection, iFORM, Epistasis, High-order interactions, Quantitative trait, Woody plant, *Prunus mume.*

1. INTRODUCTION

Quantitative traits are very difficult to study because these traits are controlled by many genes that interact in a complicated way [1, 2]. Genome-wide mapping and association studies increasingly available due to next-generation high-throughput genotyping techniques have proven to be useful for characterizing gene-gene interactions, coined epistasis, that contribute to phenotypic variation [3-5]. Powerful statistical methods have been developed to analyze all possible markers simultaneously, from which to search for a complete set of epistasis for quantitative traits [6, 7]. The joint analysis of all markers is particularly needed to chart an overall picture of genetic interactions, in comparison with computationally less expensive marginal analysis.

Epistasis reported in the current literature is mostly due to interactions between two genes. However, a growing body of evidence shows that genetic interactions involving more than two loci play a pivotal role in regulating the genetic variation of traits [8-11]. For example, in a mapping population deriving from crossing two chicken lines, three-locus interactions were detected to determine body weight [12]. A mapping study established by two yeast strains identified genetic interactions involving five or more loci for colony morphology [13]. Other studies have demonstrated that high-order epistasis is of critical importance in regulating metabolic networks in yeast [14] and *Escherichia coli* and *Saccharomyces cerevisiae* [15, 16], whereas lower-order (pairwise) epistasis may be insufficient to explain metabolic variation for these organisms.

The theoretical models of high-order epistasis have well been established by mathematical biologists [17, 18]. These models provided a foundation to interpret high-order epistasis from a biological standpoint. A few statistical models have been derived to estimate and test high-order epistasis in case-control designs [6, 19] and population-based mapping settings [10], which are suitable for mapping disease traits and quantitative traits, respectively [20]. Wang *et al.* [21] developed a Bayesian version of detecting high-order interactions for both continuous and discrete phenotypes. However, these models were based on a marginal analysis, thus less powerful to illustrate a global view of genetic control mechanisms due to high-order epistasis.

In this article, we deploy a variable selection procedure within a genetic mapping or association setting to character-

*Address correspondence to this author at the Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033, USA;
Tel: 717-531-2037; Fax: 717-531-0484; E-mail: rwu@phs.psu.edu

1389-2029/18 $58.00+.00

ize the genetic architecture of complex traits composed of main effects of individual genes, pairwise epistasis between two genes, and three-way epistasis among three genes. The model was built on Hao and Zhang's [22] iFORM, greedy interaction screening forward selection developed under the marginality principle. This approach pursues forward selection on the main effects and incorporates interactions into the model once the main effects of the corresponding covariates are selected. iFORM has theoretically proved to possess sure screening property for ultrahigh-dimensional modeling and impressive computational efficiency. Haris *et al.* [23] formulated a general framework for fitting a regression model through convex modeling of interactions with strong heredity. iFORM has been implemented to model the genetic architecture of main effects and pairwise epistasis due to eQTLs for gene transcripts, showing convincing utility to quantitative genetic studies [7]. Here, we extend the implementation of iFORM to systematically capture three-way interactions that are expressed among all possible markers studied. To show the statistical power of the extended model, we performed computer simulation studies. The model was further validated through analyzing a real data of genetic mapping for shoot growth in a woody plant, mei (*Prunus mume*). The model should be used in any other mapping or association studies of quantitative traits.

2. MODEL

2.1. Mapping and Association Studies

Genetic mapping and association studies are two types of designs used to dissect quantitative traits. The former is based on a controlled cross derived from distinct parents, whereas the latter samples different genotypes from a pool of accessions or a natural population. In both types of design, a set of individuals are sampled to be phenotyped for quantitative traits of interest and genotyped by molecular markers distributed throughout the entire genome. For a particular genetic experiment, the number of markers is much larger than that of samples, thus, it is impossible to estimate the genetic effects of all markers simultaneously using traditional regression models. This issue becomes more intractable when we aim to estimate genetic interactions of different orders. To tackle the issue of the number of predictors >> the number of samples, several variable selection approaches have been implemented in association studies. One approach is forward selection which was shown to be robust for estimating pairwise interactions of predictors. With sure screening properties and controlling for false positives, this approach, named iFORM , performs very well in capturing important information in explaining the response variable [22, 23]. On top of these nice theoretical properties it is computationally efficient by using ordinary least squares calculations and only requiring a predetermined set up steps. Here, we extended the iFORM procedure to include high-order genetic interactions to capture more relevant information. In the following sections, the notation and model set-up will be introduced, followed by the investigation of theoretical properties of the model.

2.2. Epistatic Model

Consider a linear model that underlies the true genotype-phenotype relationship. Assume that the phenotype, as the response of the model, is controlled by a set of p SNPs that act singly and/or interact with each other. These main and interaction effects of markers, *i.e.*, the predictors of the model, need to be estimated. Let $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ denote the phenotypic values of n samples from a mapping or association population. If pairwise and three-way interactions are considered, the linear model of predicting the phenotypic values is expressed as

$$\boldsymbol{Y} = \boldsymbol{\alpha} + \boldsymbol{X^T\beta} + \boldsymbol{Z^T\gamma} + \boldsymbol{W^T\eta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{X} = \left(X_1, \ldots, X_p\right)^T$ is the design matrix that specifies the genetic effects of each marker $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$; $\boldsymbol{Z} = \left(X_jX_k\right)^T$ $(1 \le j \le k \le p)$ is the design matrix that specifies the epistatic effects between two markers, expressed in $\boldsymbol{\gamma}$; $\boldsymbol{W} = \left(X_jX_kX_l\right)^T$ $(1 \le j \le k \le l \le p)$ is the design matrix that specifies the epistatic effects among three markers, expressed in $\boldsymbol{\eta}$; and $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$ is the residual error normally distributed with mean zero and variance σ^2.

We denote the index sets for the linear, order-2 and order-3 effects in equation (**1**), respectively, as

$$\mathcal{P}_1 = \{1,2, \ldots, p\},$$

$$\mathcal{P}_2 = \{(j,k): 1 \le j \le k \le p\},$$

$$\mathcal{P}_3 = \{(j,k,l): 1 \le j \le k \le l \le p\},$$

with the significant main, order-2 interaction and order-3 interaction effect sets being,

$$\mathcal{T}_1 = \left\{j: \boldsymbol{\beta_j} \ne 0, j \in \mathcal{P}_1\right\},$$

$$\mathcal{T}_2 = \left\{(j,k): \boldsymbol{\gamma_{jk}} \ne 0, (j,k) \in \mathcal{P}_2\right\},$$

$$\mathcal{T}_3 = \left\{(j,k,l): \boldsymbol{\eta_{jkl}} \ne 0, (j,k,l) \in \mathcal{P}_3\right\}.$$

The true sizes of $\mathcal{T}_1$, $\mathcal{T}_2$ and $\mathcal{T}_3$ are p_1, p_2 and p_3, respectively. There will be a total of 3 sets referred to throughout the procedure, the candidate set $\mathcal{C}$, the selection set $\mathcal{S}$ and the model set, $\mathcal{M}$. The candidate set is the set of all possible predictors at a given step in the selection process. The selection set contains the predictors that have previously been selected from the candidate set from each iteration of the procedure. Finally, the model set is the final model that is fit from the selection set at the end of the procedure. The BIC is used to determine the optimal cutoff for the final model size.

2.3. iFORM with High-order Epistasis

The iFORM procedure is a forward selecting procedure. In traditional forward selection the procedure starts with the empty set and then iterates through the entire set of possible predictors in $\mathcal{C}$ and selects the best predictor and includes it in $\mathcal{S}$ at the end of each step. The best predictor can be determined in many ways but usually is defined by the predictor that results in the least amount of error. For our purposes we use the residual sum of squares. This continues with selecting the best predictor from $\mathcal{C}$ at each step until a designated stopping criterion is met or until some information criterion is met. Common information criteria used for selecting predictors to be in $\mathcal{M}$ are AIC, BIC, R^2 and Mallow's C_p statistic.

The iFORM procedure for high-order epistatic detection parallels the forward selection procedure, but $\mathcal{C}$ will grow dynamically with the creation of order-2 and order-3 interaction effects between main effects that were included from previous iterations of the procedure. There are three steps to the model selection. The first step is to initialize the 3 sets mentioned above. The sets, $\mathcal{S}\ and\ \mathcal{M}$ are set to the empty set while the candidate set, $\mathcal{C}$, is first set to $\mathcal{P}_1$, all the main effects. The next step starts the forward selection procedure selecting predictors from $\mathcal{C}$. The selected predictor will be a main effect at the first step. At subsequent steps, after interaction effects are included, selected predictors could be either be a main genetic effect, pairwise or three-way genetic interaction effect. The final step involves repeating the second step until a designated stopping criterion is met. This can be a certain amount of predictors to be considered in the final model, or it can be based off of other factors such as the sample size. The designated stopping criterion will be denoted as d. For our purposes we use d as a function of the sample size, $d = n/\,log_2(n)$. The procedure will run up until d iterations, and the optimal model will then be constructed from the selection set. This is done by an information criterion. Here we used the Bayesian Information Criterion proposed by Chen and Chen (2008) denoted as the BIC_2. This was derived by them to control the false discovery rate in high dimensional model selections.

$$BIC_2(\widehat{\mathcal{M}}) = \log(\hat{\sigma}^2_{\widehat{\mathcal{M}}}) + n^{-1}|\widehat{\mathcal{M}}| * (\log(n) + 2 * \log(d^*)) \quad (2)$$

Once the selection procedure is done and there are d predictors in the selection set the BIC is used to determine the cutoff value for the optimum number of predictors in the model set. Then linear regression is performed on the model set.

Two guiding principles are used to help dynamically select the main effects and epistasis effects throughout the procedure. The first is the marginality principle, which states that an effect will not be removed from the model once it has been selected. A previous selected effect may become marginal by the inclusion of subsequent effects. This especially can be the case when an interaction effect is included. One of the parent effects may become less significant or even not significant at all by considering both in the model. The next principle we state as the heredity principle but has also been referred to in other work as the hierarchy principle [24, 25]. There are two cases of the heredity principle considered. The strong case would not allow for an order-2 epistasis effect to be included into the candidate set without both the parent main effects that make up the interaction are first included in the model. More formally this can be written as, $\gamma_{jk} \neq 0 \text{ only if } \beta_j, \beta_k \neq 0\ \forall\ 1 \leq j, k \leq p$. Similarly with order-3 epistasis, you would need to have all order-2 epistatic parent effects included in the model before including as a candidate predictor. This would translate to, $\eta_{jkl} \neq 0 \text{ only if } \gamma_{jk}, \gamma_{jl}, \gamma_{kl} \neq 0\ \forall\ 1 \leq j, k, l \leq p$. The weak case relaxes the need for all parent effects to be included in the model before considering the epistatic effects as candidates. Only one parent effect would be required to be in the model for candidates to be included. In the scenario with order-2 epistatic effects we would need, $\gamma_{jk} \neq 0 \text{ only if } \beta_j^2 +, \beta_k^2 \neq 0\ \forall\ 1 \leq j, k \leq p$ and with order-3 epistatic effects to be considered as a candidate we would need, $\eta_{jkl} \neq 0 \text{ only if } \gamma_{jk}^2 + \beta_l^2 \neq 0\ \forall\ 1 \leq j, k, l \leq p$.

The heredity (hierarchy) principle help reduce the search space by making the assumption that previously selected main effects would be involved in the interaction effects. By considering this principle it substantially reduces the search space making this feasible for ultra-high dimensional situations. The weak version of the heredity principle for three-way interactions states that at least one of the main effects needs to be selected into the model to consider an interaction effect that contains that predictor. Considering a moderately high set of predictors say p = 5000, if trying to include all pairwise interactions upfront, will make the candidate set be as high as 12,498,000. This alone could exceed most ram requirements of standard computers. This is before even stepping up to three-way interactions. The weak heredity principle would decrease the candidate set substantially. Assuming a sample size of n = 200, would give a cut off of $n/\,log_2(n) = 200/\,log_2(200) = 26$ steps in the procedure. The 5000 original predictors plus up to 5000 epistatic predictors included in the candidate set at each step in the procedure would give a maximum of approximately 135,000 candidate predictors. This would give a maximum of approximately 135,000 candidate predictors. This gives a 100 fold decrease in the candidate set. This could substantially make ultra-high dimensional analysis more feasible and also speed it up in the process. This is the weak case. If considering the strong case the decrease in candidate space is even more apparent. Aside from the efficiency by lowering the search space of the candidate set, the heredity principle is usually taken into account by researchers when selecting models involving the consideration for interaction effects.

3. THEORETICAL PROPERTIES

The theoretical properties of the iFORM procedure with high-order epistasis follow closely with the forward selection procedure. Hao and Zhang [22] summarize forward selection nicely as follows. At each step, the response is regressed on the most correlated covariate, and the residual is calculated and used as the new response in next step. After the most correlated covariate (say, X_1) is selected, all other covariates are regressed on X_1, and then the covariates are substituted by the corresponding normalized residuals, which are used as the new covariates in next step. By viewing forward selection in this sense the computational complexity of the procedure depends upon the size of the candidate set. The candidate set in the iFORM's case does grow dynamically at each step, by at most the number of predictors currently selected in $\mathcal{C}$ for each step. If we denote the current size of the candidate set as m then each iteration of the procedure grows with complexity of O(nm), where n is the sample size. Leaving the selection unrestricted we would not be able to fit more than n predictors for a linear model and therefore n would be the most main effects that would be able to be selected. Considering the weakest form of the heredity principle at the current iteration there would be at most $p + \frac{n(n-1)(n-2)}{6}$ predictors in the candidate set. This would make the total complexity of the selection procedure to be $nO\left(n\left(p + n(n-1)(n-2)\right)\right) = O(n^3p + n^5)$. This makes the total complexity grow linearly as p grows.

The theoretical properties of the iFORM procedure show sure screening properties [26]. By this we mean that all the important predictors, whether that is a main effect or epistatic effect will be selected with probability tending to 1. This is important to capture as much of the signal as possible through all the noise that comes with $p >> n$ or ultra-high dimensional situations. It is also important not to 'over-fit' the model with unnecessary predictors that actually explain more noise in the data that the model is being fitted on than the actual signal you would like to pick up on.

To show the property from above the following conditions would need to be met. Hao and Zhang [22] showed how under these conditions sure screening properties for interaction models like FS2 and iFORM are satisfied. This also applies to three-way interaction models like FS3 and iFORM with higher order epistasis, like we do with the high-order epistasis model. The following assumptions need to be met for these conditions. The first is that the $\boldsymbol{X} = \left(X_1, \dots, X_p\right)^T$ are jointly and marginally normal with independent normally distributed error. Next we would need the eigenvalues of the covariance matrix to be positive and bounded by two constants $0 < \tau_{min} < 1 < \tau_{max} < \infty$, such that $\sqrt{\tau}_{min} < \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) < \sqrt{\tau}_{max}/4$. Also, the genetic effects, β needs a certain level of signal strength. This we would assume to be $\left||\beta|\right| \leq C_\beta$ for some positive constant$C_\beta \; and \; \beta_{min} \geq \nu\beta\eta^{-\xi_{min}}, with \; \beta_{min} = \min{(\beta)}$. Lastly, there needs to be a certain level of sparsity to the number of important effects. Denoting the total number of important effects as d_0, and positive constants $\xi, \xi_0 \; and \; \nu$ we would need $\log(p) \leq \nu n^{\xi}, d_0 \leq \nu n^{\xi_0} \; and \; \xi + 6\xi_0 + 12\xi_{min} < \frac{1}{2}$. The conditions stated are accepted standards in the literature when studying ultra-high dimensional situations [22, 26, 27].

4. SIMULATION STUDIES

To study the numeric properties of the selection procedure, simulation studies were conducted. Data was generated using R 3.1. The $X_i's$ were all independently and identically distributed realizations generated from $Binomial(0.5)$ and the true effects for both the main and epistatic effects were included following different heredity scenarios. The phenotype was generated from the linear model setup described previously. To capture relevant data structures, there were several different scenarios considered. For each scenario 50 predictors were generated with a sample size of 300 observations. The data was split into training and a testing set to study both the fitted properties of the model as well as the generalizability of the model. There were a variety of metrics obtained to assess the suitability of each model utilized in the simulations. The first metrics that were taken into account were the rates for the true positives, false positives, true negatives and false negatives. Since we have a variety of levels to each of the models each of the rates were evaluated for the different hierarchical levels. Some of the models only have main effects and/or two-way interactions, therefore the rates were only given for the area applicable to model and the rest were reported as NA. The generalizability of the models was also assessed by withholding 100 random observations as a test set. All the data was generated from the same scenario and then 100 of the observations were randomly selected and stored for out of sample measures. The data was generated from the given scenario and randomly split before assessing the models. The exact same training and testing sets were used to fit and assess each of the models in order to make as fair of a comparison as possible. Each scenario was replicated 100 times and measures were averaged over all replicates. The two measures assessed were mean square error and the coefficient of determination. The analogous in-sample measures were also calculated for comparison. The models being compared in the simulation studies are Forward Selection, Forward Selection with all pairwise interactions (FS2), Forward Selection with all three-way interactions (FS3), iFORM strong heredity two-way, iFORM weak heredity two-way, iFORM strong heredity three-way, iFORM weak heredity three-way, Glinternet [25], and finally hierNet [24].

Covering a variety of settings the following scenarios were evaluated and compared.

Scenario 1:

$$Y = \beta_1 x_1 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7 + \gamma_{1,4} x_1 x_4 + \gamma_{1,6} x_1 x_6 + \gamma_{17} x_1 x_7 \\ + \gamma_{6,7} x_6 x_7 + \eta_{1,6,7} x_1 x_6 x_7$$

The first is where the data were generated from the interactions of the model follow a strong heredity (hierarchy) with sigma = 1. Notice we have all parent effects of the order-2 epistatic effects and also all parent effects of the order-3 epistatic effect are also in the model.

Scenario 2:

$$Y = \beta_1 x_1 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7 + \gamma_{1,4} x_1 x_4 + \gamma_{1,6} x_1 x_6 + \gamma_{1,9} x_1 x_9 \\ + \gamma_{6,7} x_6 x_7 + \eta_{1,6,9} x_1 x_6 x_9$$

The second, the data is generated to have the interactions in follow a weak heredity (hierarchy) with sigma = 1. In this scenario the main effect of x_9 is not included in the model but you can see it is part of both an order-2 and the order-3 effect.

Scenario 3:

$$Y = \beta_1 x_1 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7 + \gamma_{2,3} x_2 x_3 + \gamma_{3,8} x_3 x_8 \\ + \gamma_{5,8} x_5 x_8 + \gamma_{5,9} x_5 x_9 + \eta_{3,9,11} x_3 x_9 x_{11}$$

The third scenario is anti-heredity (hierarchical) where the interaction effects are only among predictors not present as main effects in the model. We still have main effects and epistatic effects in the model. However, the parent effects of the interactions are not the main effects included in the model.

Scenario 4:

$$Y = \gamma_{1,4} x_1 x_4 + \gamma_{1,6} x_1 x_6 + \gamma_{1,9} x_1 x_9 + \gamma_{6,7} x_6 x_7 + \eta_{1,6,9} x_1 x_6 x_9$$

Finally the last scenario only generates data that come from pure interactions between predictors with no main effects present in the model used to generate the data.

For the first scenarios where the truth obeys strong heredity where all of the parent main effects need to be selected before interactions are selected. The models that appeared to do the best in this simulation were forward selection on all three-way interactions included from the beginning (FS3), iFORM three-way weak heredity and iFORM three-way strong heredity (Table **1**). The FS3 took over a 40 fold increase in time to run. The other comparison models, glinternet and hierNet seemed to perform well on the training set but not as well on the testing set. This would indicate that some overfitting was occurring with those types of regularization models. The next scenario was when the truth obeys weak heredity. With the underlying model obeying the weak heredity, the iFORM tree-way strong heredity version dropped off in performance slightly. However, the FS3 and iFORM three-way remained as top performers (Table **2**). The third scenario assessed was from an underlying model with an anti-heredity structure. Both main effects and interaction effects were used in the model to generate the data. However, the interactions included in the model were of combinations of main effects in the candidate set, which were not in the model. The iFORM seems to drop in performance with this scenario (Table **3**). This is to be expected because it is in direct violation of the underlying assumptions of the model hierarchy. Even with these violations of the heredity it still performed reasonably well. Lastly, making the scenario a little more extreme, the underlying model generating the data was only of interactions. There were no main effects included in the model. The results of this scenario are shown in Table **4**. Performance appeared to drop off for all models explored in the simulation.

In the scenarios where the data was assumed to follow some form of a hierarchical structure for the epistasis effects the iFORM procedure for higher-order epistasis effects appeared to perform the best. Not only did it result in selecting the correct model, the false positive rate was also among the lowest. The out of sample error was also among the lowest between each of the models compared. With the procedure using OLS calculations, it also performed the fastest out of the models including epistasis effects. All of the combined show the promise of the iFORM procedure for GWAS type studies. With the other scenarios, the underlying structure of the data does not follow a typical intuition about the structure of data in biology.

5. WORKED EXAMPLE

We validated the biological usefulness of the model by analyzing mapping data for a woody plant, mei (*Prunus mume*). Originated in China, mei has been cultivated for its ornamental flowers for thousands of years [28, 29]. Its many desirable properties, such as cold-hardiness, colors and flavors, are appraised as a symbol of persistence and beauty in Chinese culture. Recent sequencing of its genome has made it an ideal model system to study the genetics and evolution of woody plants [30]. To improve the growth rigor and form of mei important to its ornamental value, a cross was made between two distinct cultivars, Fenban (female parent) and Kouzi Yudie (male parent), aimed to select superior genotypes from hybrids. To the end, an F_1 mapping population of 190 hybrids was established and further genotyped for 4,934 SNP markers over eight linkage groups which correspond to 8 chromosomes across the entire genome.

To test genotypic differences in growth performance, each of these hybrids was grafted on an established root stock using multiple budding scions. Next spring, buds on the scions sprouted into shoots. The lengths and diameters of 10 randomly selected shoots were measured once every two weeks during an entire growth season from March to October. It was found that both shoot length and diameter growth was well fitted to the three-parameter growth equation expressed as

$$g(t) = a/[1 + b\exp(-rt)], \tag{3}$$

where $g(t)$ is the amount of shoot growth at time t, a is the asymptotic value of growth when time tends to be infinite, b is a parameter that reflects the amount of growth at time 0, and r is the relative growth rate. These three parameters determine the overall form of growth curve jointly, although they function differently. Thus, by estimating these parameters for individual hybrids using a nonlinear least squares approach, we can draw the growth curve of each hybrid. Differences in growth curve among hybrids may be controlled by specific genes or Quantitative Trait Loci (QTLs). Although tremendous efforts have been made to map growth QTLs and their epistasis [31-33], none has characterized the contribution of high-order epistasis although it has been thought to regulate growth processes.

By treating the estimates of growth parameters for individual hybrids as “phenotypic traits,” we used iFORM to map growth QTLs and QTL-QTL interactions. Of 4,934 markers, 2,100 are the testcross markers at which markers are segregating due to only one heterozygous parent and 2,834 are the intercross markers whose segregation results from the heterozygosity of both parents. For a testcross marker, there is only one main genetic effect, whereas an intercross marker contains additive and dominant main effects. Thus, a pair of testcross markers produces only type of epistasis, but a pair of intercross markers forms four types of epistasis, additive × additive, additive × dominant, dominant × additive and dominant × dominant. For two markers with one from the testcross and the other from the intercross, there are two types of epistasis, *i.e.*, additive × additive and additive × dominant [34]. The number and type of epistasis can be characterized for any three markers accordingly. Here, the iFORM was implemented in a way that allows both marker markers to be modeled and analyzed simultaneously.

To demonstrate the possible importance of high-order epistasis, we analyze the data by assuming that growth parameters are controlled by low-order epistasis only and by both low- and high-order epistasis, respectively. The weak heredity (hierarchical) was used to screen every SNP and possible interaction of the main effects selected and the rest of the SNPs left in the candidate set. It was not restricted to the strong case where both main effects had to be in the model for the interaction to be considered. For the pairwise epistatic model, this grew the candidate set to almost 20,000 predictors to choose from. It turned out that 5 predictors were chosen, *i.e.*, four main additive effects of markers, AATTC_nn_np_2517, AATTC_nn_np_2815, CATG_nn_np_3479 and CATG_nn_np_1284 and one epistatic effect due to markers AATTC_nn_np_2815 and AATTC_lm_ll_3034, for growth parameter r of shoot length (Table **5**). The main effect of marker AATTC_lm_ll_3034 was detected to be insignificant. These main and epistatic effects together explained 32.41% of the total variance of parameter r.

Table 1. Simulation results when the truth obeys strong heredity.

$$Y = \beta_1 x_1 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7 + \gamma_{1,4} x_1 x_4 + \gamma_{1,6} x_1 x_6 + \gamma_{17} x_1 x_7 + \gamma_{6,7} x_6 x_7 + \eta_{1,6,7} x_1 x_6 x_7$$

Model	T1 tpr	T1 fpr	T2 tpr	T2 fpr	T3 tpr	T3 fpr	Train MSE	Train Rsq	Test MSE	Test Rsq	Model Size	Run Time
Forward select	1.00	0.001	0.000	0.000	0.000	0.000	3.330	0.727	3.490	0.711	4.04	0.757
Iform weak(2)	1.00	0.000	0.000	0.000	0.000	0.000	1.128	0.907	1.252	0.895	8.08	5.896
Iform strong(2)	1.00	0.000	0.000	0.000	0.000	0.000	1.102	0.909	1.198	0.900	8	1.557
Forward select(2)	1.00	0.000	0.000	0.000	0.000	0.000	1.086	0.910	1.198	0.900	8.02	25.481
Forward select(3)	1.00	0.000	0.000	0.000	0.000	0.000	0.992	0.918	1.121	0.906	8.56	471.88
Iform weak(3)	1.00	0.000	0.000	0.000	0.000	0.000	1.020	0.916	1.135	0.905	9.13	11.346
Iform strong(3)	1.00	0.000	0.000	0.000	0.000	0.000	0.968	0.920	1.060	0.911	8.95	1.872
glinternet	1.00	0.441	1.000	0.018	0.000	0.000	1.246	0.898	1.446	0.880	29.9	208.17
hierNet	1.00	0.303	1.000	0.024	0.000	0.000	0.906	0.925	1.421	0.882	40.99	27.521
Oracle	NA	NA	NA	NA	NA	NA	0.953	0.921	1.050	0.912	9	NA

Table 1: shows simulation results under the first simulation scenario described. Results for the true positive rate(tpr) and ralse positive rate(fpr) are given for each level of hierarchy in the effects (T1 - main effects, T2 - order2 and T3 - order3). The Mean Square Error (MSE) is given for both the training and testing set generated. The coefficient of determination (Rsq) is also give for both training and testing set for comparison across models. The average final model size and the average run time in seconds of each model are presented as well.

Table 2. Simulation results when the truth obeys weak heredity.

$$Y = \beta_1 x_1 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7 + \gamma_{1,4} x_1 x_4 + \gamma_{1,6} x_1 x_6 + \gamma_{1,9} x_1 x_9 + \gamma_{6,7} x_6 x_7 + \eta_{1,6,9} x_1 x_6 x_9$$

Model	T1 tpr	T1 fpr	T2 tpr	T2 fpr	T3 tpr	T3 fpr	Train MSE	Train Rsq	Test MSE	Test Rsq	Model Size	Run Time
Forward select	1.000	0.001	0.000	0.000	0.000	0.000	3.326	0.731	3.480	0.716	4.03	4.355
Iform weak(2)	1.000	0.000	0.000	0.000	0.000	0.000	1.119	0.910	1.200	0.901	8.07	8.342
Iform strong(2)	1.00	0.008	0.000	0.000	0.000	0.000	1.580	0.872	1.707	0.859	7.54	2.952
Forward select(2)	1.00	0.000	0.000	0.000	0.000	0.000	1.083	0.912	1.167	0.904	8	38.872
Forward select(3)	1.00	0.000	0.000	0.000	0.000	0.000	0.979	0.921	1.089	0.910	8.58	569.98
Iform weak(3)	1.00	0.000	0.000	0.000	0.000	0.000	1.003	0.919	1.079	0.911	9.03	13.054
Iform strong(3)	1.00	0.008	0.000	0.000	0.000	0.000	1.578	0.872	1.705	0.859	7.58	2.787
Glinternet	1.00	0.531	1.000	0.020	0.000	0.000	0.906	0.927	1.425	0.883	33.18	29.975
hierNet	1.00	0.343	1.000	0.027	0.000	0.000	0.856	0.931	1.412	0.884	43.43	33.302
Oracle	NA	NA	NA	NA	NA	NA	0.940	0.924	1.034	0.915	9	NA

Table 2. shows simulation results under the second simulation scenario described. Results for the true positive rate(tpr) and ralse positive rate(fpr) are given for each level of hierarchy in the effects (T1 - main effects, T2 - order2 and T3 - order3). The Mean Square Error (MSE) is given for both the training and testing set generated. The coefficient of determination (Rsq) is also give for both training and testing set for comparison across models. The average final model size and the average run time in seconds of each model are presented as well.

Table 3. Simulation results when the truth is anti-heredity.

$$Y = \beta_1 x_1 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7 + \gamma_{2,3} x_2 x_3 + \gamma_{3,8} x_3 x_8 + \gamma_{5,8} x_5 x_8 + \gamma_{5,9} x_5 x_9 + \eta_{3,9,11} x_3 x_9 x_{11}$$

Model	T1 tpr	T1 fpr	T2 tpr	T2 fpr	T3 tpr	T3 fpr	Train MSE	Train Rsq	Test MSE	Test Rsq	Model Size	Run Time
Forward select	1.00	0.000	0.000	0.000	0.000	0.000	3.284	0.729	3.510	0.714	4.02	1.005
Iform weak(2)	1.00	0.004	0.000	0.000	0.000	0.000	3.140	0.741	3.435	0.719	4.77	7.866
Iform strong(2)	1.00	0.000	0.000	0.000	0.000	0.000	3.284	0.729	3.510	0.714	4.02	2.386
Forward select(2)	1.00	0.000	0.000	0.000	0.000	0.000	1.081	0.911	1.171	0.904	8.04	29.095
Forward select(3)	1.00	0.000	0.000	0.000	0.000	0.000	0.989	0.918	1.095	0.910	8.59	548.62
Iform weak(3)	1.00	0.003	0.000	0.000	0.000	0.000	3.155	0.739	3.448	0.719	4.57	13.216
Iform strong(3)	1.00	0.000	0.000	0.000	0.000	0.000	3.284	0.729	3.510	0.714	4.02	2.703
glinternet	1.00	0.710	1.000	0.029	0.000	0.000	0.844	0.931	1.578	0.871	44.59	26.564
hierNet	1.00	0.858	1.000	0.085	0.000	0.000	0.307	0.975	2.216	0.819	119.73	3.417
Oracle	NA	NA	NA	NA	NA	NA	0.952	0.921	1.031	0.915	9	NA

Table 3. shows simulation results under the third simulation scenario described. Results for the true positive rate(tpr) and ralse positive rate(fpr) are given for each level of hierarchy in the effects (T1 - main effects, T2 - order2 and T3 - order3). The Mean Square Error (MSE) is given for both the training and testing set generated. The coefficient of determination (Rsq) is also give for both training and testing set for comparison across models. The average final model size and the average run time in seconds of each model are presented as well.

Table 4. Simulation results when the truth is constructed of pure interactions.

$$Y = \gamma_{1,4} x_1 x_4 + \gamma_{1,6} x_1 x_6 + \gamma_{1,9} x_1 x_9 + \gamma_{6,7} x_6 x_7 + \eta_{1,6,9} x_1 x_6 x_9$$

Model	T1 tpr	T1 fpr	T2 tpr	T2 fpr	T3 tpr	T3 fpr	Train MSE	Train Rsq	Test MSE	Test Rsq	Model Size	Run Time
Forward select	NaN	0.020	0.000	0.000	0.000	0.000	3.316	0.025	3.445	-0.039	1	1.177
Iform weak(2)	NaN	0.028	0.000	0.000	0.000	0.000	3.007	0.115	3.181	0.040	2.27	5.840
Iform strong(2)	NaN	0.021	0.000	0.000	0.000	0.000	3.294	0.031	3.429	-0.034	1.08	2.081
Forward select(2)	NaN	0.000	0.000	0.000	0.000	0.000	1.117	0.669	1.170	0.644	4.01	26.396
Forward Select(3)	NaN	0.000	0.000	0.000	0.000	0.000	1.005	0.703	1.081	0.671	4.62	530.36
Iform weak(3)	NaN	0.025	0.000	0.000	0.000	0.000	3.043	0.106	3.209	0.032	1.86	9.461
Iform strong(3)	NaN	0.021	0.000	0.000	0.000	0.000	3.294	0.031	3.429	-0.034	1.08	2.265
glinternet	NaN	0.571	1.000	0.017	0.000	0.000	1.002	0.699	1.445	0.561	27.53	145.08
hierNet	NaN	0.853	1.000	0.045	0.000	0.000	0.672	0.802	1.758	0.467	92.52	4.491
Oracle	NA	NA	NA	NA	NA	NA	0.968	0.713	1.022	0.689	5	NA

Table 4. shows simulation results under the fourth simulation scenario described. Results for the true positive rate(tpr) and ralse positive rate(fpr) are given for each level of hierarchy in the effects (T1 - main effects, T2 - order2 and T3 - order3). The Mean Square Error (MSE) is given for both the training and testing set generated. The coefficient of determination (Rsq) is also give for both training and testing set for comparison across models. The average final model size and the average run time in seconds of each model are presented as well.

Table 5. The detection of epistasis for the relative growth rate (*r*) of shoot length in the full-sib family of mei tree by a low-order epistatic model.

Coefficient	Estimate	SE	T-value	P-value
(Intercept)	0.18285	0.07613	2.402	0.0174 *
AATTC_nn_np_2517_a	0.40013	0.06509	6.147	5.13e-09 ***
AATTC_nn_np_2815_a	0.15792	0.06837	2.310	0.0221 *
CATG_nn_np_3479_a	0.23433	0.05285	4.434	1.63e-05 ***
CATG_nn_np_1284_a	0.22200	0.05313	4.179	4.61e-05 ***
AATTC_nn_np_2815_a×AATTC_lm_ll_3034_a	0.45783	0.09244	4.953	1.71e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3504 on 176 degrees of freedom

Multiple R-squared: 0.3428, Adjusted R-squared: 0.3241

F-statistic: 18.36 on 5 and 176 DF, p-value: 1.189e-14

When opening up the iFORM procedure to the possibility to creating higher order interactions to be placed into the candidate set, a more complete picture of the phenotypical variation was revealed. The amount of predictors included in the final model grew to 12, with one of them being three-way interactions among markers AATTC_nn_np_2815, AATTC_lm_ll_3034 and AATTC_nn_np_1615. The adjusted R^2 jumped up to over 70% (Table **6**). This astonishing jump in predictive power is an exemplar case as to the importance of higher-order interactions in genetic models. Not only did higher-order interactions become one of the most significant predictors in the model selected, it also allowed for other order-two interactions and main effects to be kept in the model that were previously left out. At the next step of the iteration the new candidate effect was conditioned on everything previously selected. With the conditional effect of the higher-order interaction it enabled for other lost effects to be modeled as well.

The purpose of the mei genetic project is to study the genetic control of shoot growth form. Here, we further analyze how three-way interactions detected by our model affect growth form. Assume that there are three testcross markers, **A** (with two alleles *A*, *a*), **B** (with two alleles *B*, *b*), and **C** (with two alleles *C*, *c*), which interact jointly to affect shoot growth. The three markers form eight genotypes *AABBCC*, *AABBCc*, *AABbCC*, *AABbCc*, *AaBBCC*, *AaBBCc*, *AaBbCC* and *AaBbCc* whose genotypic means at time *t* are partitioned into different components, respectively, expressed as

$$\mu_{111}(t) = \mu(t) + \alpha_1(t) + \alpha_2(t) + \alpha_3(t) + i_{12}(t) + i_{13}(t) + i_{23}(t) + i_{123}(t)$$

$$\mu_{112}(t) = \mu(t) + \alpha_1(t) + \alpha_2(t) - \alpha_3(t) + i_{12}(t) - i_{13}(t) - i_{23}(t) - i_{123}(t)$$

$$\mu_{121}(t) = \mu(t) + \alpha_1(t) - \alpha_2(t) + \alpha_3(t) - i_{12}(t) + i_{13}(t) - i_{23}(t) - i_{123}(t)$$

$$\mu_{122}(t) = \mu(t) + \alpha_1(t) - \alpha_2(t) - \alpha_3(t) - i_{12}(t) - i_{13}(t) + i_{23}(t) + i_{123}(t)$$

$$\mu_{211}(t) = \mu(t) - \alpha_1(t) + \alpha_2(t) + \alpha_3(t) - i_{12}(t) - i_{13}(t) + i_{23}(t) - i_{123}(t)$$

$$\mu_{212}(t) = \mu(t) - \alpha_1(t) + \alpha_2(t) - \alpha_3(t) - i_{12}(t) + i_{13}(t) - i_{23}(t) + i_{123}(t)$$

$$\mu_{221}(t) = \mu(t) - \alpha_1(t) - \alpha_2(t) + \alpha_3(t) + i_{12}(t) - i_{13}(t) - i_{23}(t) + i_{123}(t)$$

$$\mu_{222}(t) = \mu(t) - \alpha_1(t) - \alpha_2(t) - \alpha_3(t) + i_{12}(t) + i_{13}(t) + i_{23}(t) - i_{123}(t) \tag{4}$$

where $\mu(t)$ is the population mean at time t; $\alpha_1(t)$, $\alpha_2(t)$ and $\alpha_3(t)$ are the genetic effects of markers **A**, **B** and **C** at time t, respectively; $i_{12}(t)$, $i_{13}(t)$ and $i_{23}(t)$ are the pairwise epistatic effects between markers **A** and **B**, **A** and **C** and **B** and **C** at time t, respectively; and $i_{123}(t)$ is the three-way epistatic effect among three the markers at time t. From the above equations, we solve the pairwise and three-way epistatic effects as

$$i_{12}(\mathrm{t}) = \frac{1}{8}\,[(\mu_{111}(t) + \mu_{112}(t) + \mu_{221}(t) + \mu_{222}(t)) - (\mu_{121}(t) + \mu_{122}(t) + \mu_{211}(t) + \mu_{212}(t))]$$

$$i_{13}(\mathrm{t}) = \frac{1}{8}\,[(\mu_{111}(t) + \mu_{121}(t) + \mu_{212}(t) + \mu_{222}(t)) - (\mu_{112}(t) + \mu_{122}(t) + \mu_{211}(t) + \mu_{221}(t))]$$

$$i_{23}(\mathrm{t}) = \frac{1}{8}\,[(\mu_{111}(t) + \mu_{122}(t) + \mu_{211}(t) + \mu_{222}(t)) - (\mu_{112}(t) + \mu_{121}(t) + \mu_{212}(t) + \mu_{221}(t))]$$

$$i_{123}(t) = \frac{1}{8}\,[(\mu_{111}(t) + \mu_{122}(t) + \mu_{212}(t) + \mu_{122}(t)) - (\mu_{112}(t) + \mu_{121}(t) + \mu_{211}(t) + \mu_{222}(t))] \tag{5}$$

Each genotype can draw a growth curve using its growth parameters (*a*, *b*, *r*) estimated from raw data, from which we can chart the curves of pairwise and three-way epistatic effects using equation (**4**). Three markers AATTC_nn_np_2815 (AA/Aa), AATTC_lm_ll_3034 (BB/Bb) and AATTC_nn_np_1615 (CC/Cc) that produce a significant three-way interaction for parameter x of shoot length display pronounced differences in growth curve (Fig. **1**). The epistasis of low- and high-order performs differently to affect growth form, with three-way interactions playing a more remarkable role than pairwise epistasis (Fig. **2**).

The figures display the variation between each of the growth curves for the eight combinations of the three marker genotypes focused on Fig. (**1**). Differences of each of the growth parameters can be observed when studying the figures. There is clear separation in the shoot length that is observed at the end of the 16 weeks. This difference can be visually grouped into four clusters that show the effect a genotype combination can have on the asymptotic growth parameter, a. Another noticeable different between the curves displayed is the rate at which the growth developed. At the earlier weeks of development you can see some of the genotype combinations grew faster, manifesting in a steeper

Table 6. The detection of epistasis for the relative growth rate (*r*) of shoot length in the full-sib family of mei tree by a high-order epistatic model.

Coefficient	Estimate	SE	T-value	P-value
(Intercept)	0.16859	0.05801	2.906	0.00415 **
AATTC_nn_np_2517_a	0.27773	0.04396	6.318	2.27e-09 ***
AATTC_nn_np_2815_a	0.26382	0.05295	4.983	1.54e-06 ***
CATG_nn_np_3479_a	0.20767	0.03467	5.990	1.23e-08 ***
CATG_nn_np_1284_a	0.04522	0.04265	1.060	0.29055
AATTC_nn_np_2815_a×AATTC_lm_ll_3034_a	1.82572	0.17925	10.185	< 2e-16 ***
AATTC_nn_np_2815_a×AATTC_hk_hk_278_a	0.25935	0.03888	6.671	3.48e-10 ***
CATG_lm_ll_3153_a	0.14877	0.03491	4.262	3.36e-05 ***
CATG_nn_np_1284_a×AATTC_nn_np_554_a	0.22994	0.05104	4.505	1.23e-05 ***
AATTC_nn_np_2815_a.AATTC_lm_ll_3034_a×AATTC_nn_np_1615_a	-1.51714	0.19060	-7.960	2.39e-13 ***
AATTC_nn_np_2815_a×AATTC_nn_np_929_a	-0.30805	0.05477	-5.624	7.57e-08 ***
AATTC_hk_hk_479_d	0.16044	0.03443	4.660	6.37e-06 ***
AATTC_nn_np_2517_a×CATG_hk_hk_648_a	0.14537	0.02840	5.118	8.33e-07 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2268 on 169 degrees of freedom

Multiple R-squared: 0.7356, Adjusted R-squared: 0.7168

F-statistic: 39.19 on 12 and 169 DF, p-value: < 2.2e-16

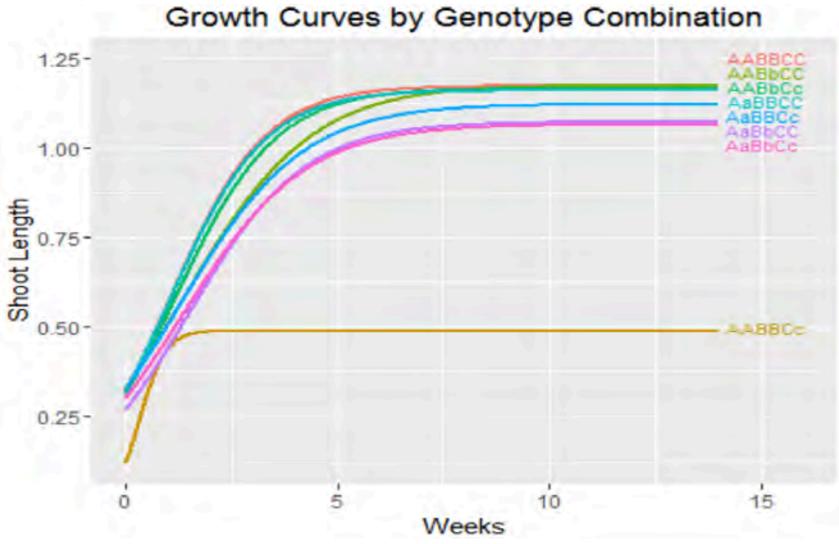

Fig. (1). Growth curves of shoot length in mei drawn from estimated growth parameters at three loci of significant high-order epistasis.

slope and other genotypes had shallower slopes. All of these visually show what was picked up on when modeling the shoot length growth and the impact of the higher-order interactions between the genotypes have on such growth. By solving the system of linear equations in (**5**) we can dissect the epistatic effects of the genotype combinations. The effects over time are displayed (Fig. **2**) and in this you can see the non-linear influence of the interactions between the markers included.

DISCUSSION AND CONCLUSION

Genetic interactions have been thought to contribute to a significant portion of genetic variance for quantitative traits of critical importance to evolutionary biology, agriculture and medicine [1, 2]. While pairwise interactions have been a major focus of quantitative genetic studies, there has been growing evidence that genetic interactions involving three or more loci play an important role in affecting the phenotypic differentiation of traits [9-14]. Because of its complexity due

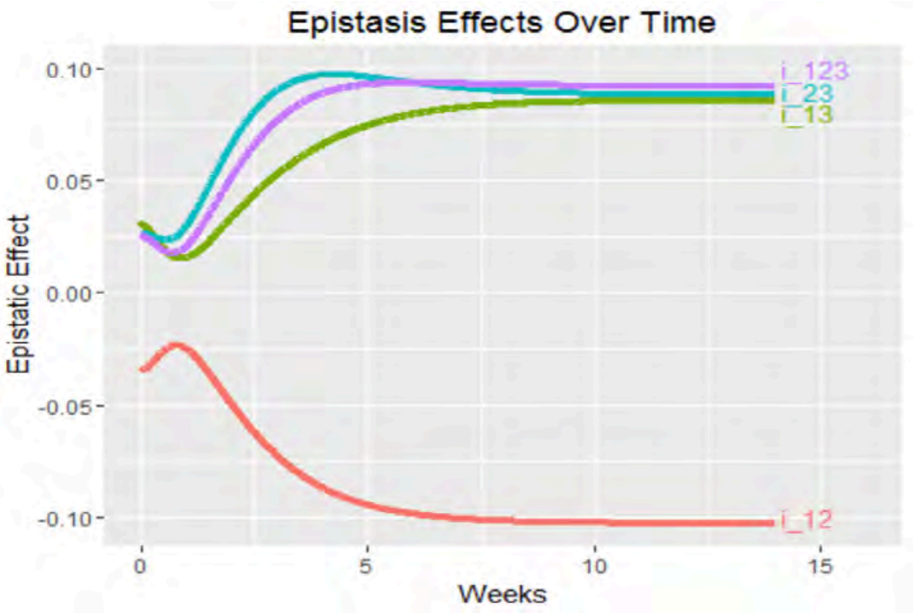

Fig. (2). Curves of epistatic effects on shoot length growth in mei at three significant loci.

to a network of interactions, the detection of high-order epistasis is extremely difficult [2]. More importantly, interpretation of high-order epistasis and its contribution to overall genetic architecture can be better made by jointly analyzing all possible low- and high-order interactions among genes. This has added an extra challenge to statistical modeling and detection of this important phenomenon. Thanks to the recent development of statistical models for high-dimensional variable selection, we have reformed a statistical modeling framework for detecting high-order epistasis by focusing on three-way interactions.

Our model extends Hao and Zhang's [22] forward selection-based algorithm iFORM that has proven to be robust and efficient for computing and detecting two-way interactions between predictors (including continuous predictors). A favorable property of iFORM is its capacity to detect interactions even if the dimension of predictors is extremely high relative to a sample size used. The fundamental assumption used by iFORM is the heredity principle, *i.e.*, the existence of interactions between a pair of variables that each has at least weak main effects. After extending it to characterize three-way interactions, this assumption can be relaxed for the third variable; *i.e.*, even if there is no detectable main effect for the third marker, then extended iFORM can still detect the three-way interaction. This property may explain the reason why high-order epistatic model outperforms low-order epistatic model, as demonstrated from the detection of significant genetic interactions in a real data of a woody plant, mei (*Prunus mume*). It was found from a recent study that loci participating in high-order genetic interactions may not individually have measurable effects [35]. As a result, our model can be used as a general tool to detect genetic interactions of various orders and, therefore, elucidate the overall picture of genetic architecture by capturing the so-called missing heritability.

The model was investigated by simulation studies whose result help users to determine an optimal design of mapping or association studies in terms of sample size, phenotyping precision and the number of markers. Its application to *P. mume* genetic mapping leads to the detection of key loci and their interactions expressed at the low- and high-order levels for the growth form of shoots. The curve of three-way epistasis on mei shoot length growth was observed to increase exponentially during the first five weeks of shoot sprouting and become stable after five weeks. Such integration of the model into growth equation shed light on the developmental mechanisms of growth processes through epistasis, a question that has evoked a tremendous interest of researchers globally in the area of evolutionary developmental biology [36-38]. We have created an R package that has implemented the model which adds a function to allow epistasis of any orders to be searched. The package can be uploaded at http://statgen.psu.edu/software/ and will be made available through CRAN (Comprehensive R Archive Network).

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are base of this research.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Nelson, R.M.; Pettersson, M.E.; Carlborg, Ö. A century after Fisher: Time for a new paradigm in quantitative genetics. *Trends Genet.,* **2013**, *29*(12), 669-676.

[2] Mackay, T.F. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.*, **2014**, *15*(1), 22-33.

[3] Cordell, H.J. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.,* **2009**, *10*(6), 392-404.

[4] Van Steen, K. Travelling the world of gene-gene interactions. *Brief Bioinform.,* **2012**, *13*(1), 1-19.

[5] Wei, W.H.; Hemani, G.; Haley, C.S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.*, **2014**, *15*(11), 722-733.

[6] Li, J.H.; Zhong, W.; Li, R.; Wu, R.L. A fast algorithm for detecting gene-gene interactions in genome-wide association studies. *Ann. Appl. Stat.*, **2014**, *8*(4)*,* 2292-2318.

[7] Gosik, K.; Kong, L.; Chinchilli, V.M.; Wu, R.L. iFORM /eQTL: An ultrahigh-dimensional platform for inferring the global genetic architecture of gene transcripts. *Brief Bioinform.,* **2017**, *18*(2), 250-259.

[8] Wang, Z.; Liu, T.; Lin, Z.W.; Hegarty, J.; Koltun, W.A.; Wu, R.L. A general model for multilocus epistatic interactions in case-control studies. *PLoS One*, **2010**, *5*(8), e11384. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0011384

[9] Dowell, R.D.; Ryan, O.; Jansen, A.; Cheung, D.; Agarwala, S.; Danford, T.; Berstein, D.A.; Rolfe, P.A.; Heisler, L.E.; Chin, B.; Nislow, C.; Giaever, G.; Phillips, P.; Fink, G.; Gifford, D.; Boone, C. Genotype to phenotype: A complex problem. *Science*, **2010**, *328*(5977), 469. Available from: http://science.sciencemag.org/content/328/5977/469

[10] Pang, X.M.; Wang, Z.; Yap, J.S.; Bo, W.H.; Lv, Y.F.; Xu, F.; Zhou, T.; Peng, S.; Shen, D.; Wu, R. A statistical procedure to map high-order epistasis for complex traits. *Brief Bioinform.,* **2013**, *14*(3), 302-314.

[11] Taylor, M.B.; Ehrenreich, I.M. Higher-order genetic interactions and their contribution to complex traits. *Trends Genet.*, **2015**, *31*(1), 34-40.

[12] Pettersson, M.; Besnier, F.; Siegel, P.B.; Carlborg, Ö. Replication and explorations of high-order epistasis using a large advanced intercross line pedigree. *PLoS Genet.*, **2011**, *7*(7), e1002180. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002180

[13] Taylor, M.B.; Ehrenreich, I.M. Genetic interactions involving five or more genes contribute to a complex trait in yeast. *PLoS Genet.,* **2014**, *10*(5), e1004324. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004324

[14] Weinreich, D.M.; Lan, Y.H.; Wylie, C.S.; Heckendorn, R.B. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.,* **2013**, *23*(6), 700-707.

[15] Imielinski, M.; Belta, C. Exploiting the pathway structure of metabolism to reveal high-order epistasis. *BMC Syst. Biol.*, **2008**, *2*, 40. Available from: https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-2-40

[16] He, X.L.; Qian, W.F.; Wang, Z.; Li, Y.; Zhang, J.Z. Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks. *Nat. Genet.,* **2010**, *42*(3), 272-276. Available from: https://www.nature.com/articles/ng.524

[17] Hansen, T.; Wagner, G. Epistasis and the mutation load: A measurement-theoretical approach. *Genetics*, **2001**, *158*(1), 477-485.

[18] Beerenwinkel, N.; Pachter, L.; Sturmfels, B.; Elena, S.; Lenski, R. Analysis of epistatic interactions and fitness landscapes using a new geometric approach. *BMC Evol. Biol.,* **2007**, *7*, 60. Available from: https://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-7-60

[19] Liu, T.; Thalamuthu, B.; Liu, C.; Chen, J.; Wu, R.L. Asymptotic distribution for epistatic tests in case-control studies. *Genomics*, **2011**, *98*(2), 145-151.

[20] Upton, A.; Trelles, O.; Cornejo-García, J.A.; Perkins, J.R. High-performance computing to detect epistasis in genome scale data sets. *Brief Bioinform.,* **2016**, *17*(3)*,* 368-379.

[21] Wang, J.; Joshi, T.; Valliyodan, B.; Shi, H.; Liang, Y.; Nguyen, H.T.; Zhang, J.; Xu, D. A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies. *BMC Genom*., **2015**, *16*, 1011. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-2217-6

[22] Hao, N.; Zhang, H.H. Interaction screening for ultrahigh-dimensional data. *J. Am. Stat. Assoc.,* **2014**, *109*(507), 1285-1301.

[23] Haris, A.; Witten, D.; Simon, N. Convex modeling of interactions with strong heredity. *J. Comput. Graph. Stat.*, **2016**, *25*(4), 981-1004.

[24] Bien, J.; Taylor, J.; Tibshirani, R. A lasso for hierarchical interactions. *Ann. Stat.,* **2013**, *41*(3), 1111-1114.

[25] Lim, M.; Hastie, T. Learning interactions *via* hierarchical group-lasso regularization. *J. Comput. Graph. Stat.,* **2015**, *24*(3)*,* 626-654.

[26] Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Stat. Soc. Ser. B,* **2008**, *70*(5), 849-911.

[27] Zhang, C.H.; Huang, J. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Stat.*, **2008**, *36*(4)*,* 1567-1594.

[28] Sun, L.; Wang, Y.; Yan, X.; Cheng, T.R.; Ma, K.F.; Yang, W.R.; Pan, H.; Zheng, C.; Zhu, X.; Wang, J.; Wu, R.L.; Zhang, Q. Genetic control of juvenile growth and botanical architecture in an ornamental woody plant, *Prunus mume*Sieb. et Zucc. as revealed by a high-density linkage map. *BMC Genet*., **2014**, *15*, S1. Available from: https://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-15-S1-S1

[29] Sun, L.D.; Yang, W.; Zhang, Q.; Cheng, T.; Pan, H.; Xu, Z.; Zhang, J.; Chen, C. Genome-wide characterization and linkage mapping of simple sequence repeats in mei (*Prunus mume* Sieb. et Zucc.). *PLoS One*, **2013**, *8*(3)*,* e59562. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0059562

[30] Zhang, Q.; Chen, W.; Sun, L.D.; Zhao, F.; Huang, B. The genome of *Prunus mume*. *Nat. Commun*., **2012**, *3*, 1318. Available from: https://www.nature.com/articles/ncomms2290

[31] Ma, C.X.; Casella, G.; Wu, R.L. Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. *Genetics*, **2002**, *161*(4), 1751-1762.

[32] Wu, R.L.; Lin, M. Functional mapping - how to map and study the genetic architecture of dynamic complex traits. *Nat. Rev. Genet*., **2006**, *7*(3), 229-237.

[33] Li, Z.; Sillanpaa, M.J. Dynamic quantitative trait locus analysis of plant phenomic data. *Trends Plant Sci*., **2015**, *20*(12), 822-833.

[34] Tong, C.F.; Wang, Z.; Zhang, B.; Shi, J.S.; Wu, R.L. 3FunMap: Full-sib family functional mapping of dynamic traits. *Bioinformatics*, **2011**, *27*(14), 2006-2008.

[35] Bloom, J.S.; Ehrenreich, I.M.; Loo, W.T.; Lite, T.L.; Kruglyak, L. Finding the sources of missing heritability in a yeast cross. *Nature*, **2013**, *494*(7436), 234-237. Available from: https://www.nature.com/articles/nature11867

[36] Franks, S.J.; Sim, S.; Weis, A.E. Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. *Proc. Natl. Acad. Sci. U.S.A*., **2007**, *104*(4), 1278-1282.

[37] Cartolano, M.; Pieper, B.; Lempe, J.; Tattersall, A.; Huijser, P.; Tresch, A.; Darrah, P.; Hay, A.; Tsiantis, M. Heterochrony underpins natural variation in *Cardamine hirsuta* leaf form. *Proc. Natl. Acad. Sci. U.S.A*., **2015**, *112*(33), 10539-10544.

[38] Nishino, J.; Kim, S.; Zhu, Y.; Zhu, H.; Morrison, S.J. A network of heterochronic genes including *Imp1* regulates temporal changes in stem cell properties. *eLife*, 2013, *2*, e00924. Available from: https://elifesciences.org/articles/00924