Article
Chinese Science Bulletin
SPECIAL TOPIC
Quantitative Genetics
July 2012 Vol.57 No.21: 2666–2674
doi: 10.1007/s11434-012-5195-y

Bin-based model construction and analytical strategies for dissecting complex traits with chromosome segment substitution lines

TANG ZaiXiang[1,2], XIAO Jing[3], HU WenMing[2], YU Bo[1] & XU ChenWu[2*]

[1] *School of Public Health, Medical College of Soochow University, Suzhou 215123, China;*
[2] *Jiangsu Provincial Key Laboratory of Crop Genetics and Physiology, Key Laboratory of Plant Functional Genomics of Ministry of Education, Yangzhou University, Yangzhou 225009, China;*
[3] *Department of Epidemiology and Biostatistics, School of Public Health, Nantong University, Nantong 226019, China*

Received January 14, 2012; accepted April 1, 2012; published online May 1, 2012

Chromosome segment substitution lines have been created in several experimental models, including many plant and animal species, and are useful tools for the genetic analysis and mapping of complex traits. The traditional *t*-test is usually applied to identify a quantitative trait locus (QTL) that is contained within a chromosome segment to estimate the QTL's effect. However, current methods cannot uncover the entire genetic structure of complex traits. For example, current methods cannot distinguish between main effects and epistatic effects. In this paper, a linear epistatic model was constructed to dissect complex traits. First, all the long substituted segments were divided into overlapping small bins, and each small bin was considered a unique independent variable. The genetic model for complex traits was then constructed. When considering all the possible main effects and epistatic effects, the dimensions of the linear model can become extremely high. Therefore, variable selection via stepwise regression (Bin-REG) was proposed for the epistatic QTL analysis in the present study. Furthermore, we tested the feasibility of using the LASSO (least absolute shrinkage and selection operator) algorithm to estimate epistatic effects, examined the fully Bayesian SSVS (stochastic search variable selection) approach, tested the empirical Bayes (E-BAYES) method, and evaluated the penalized likelihood (PENAL) method for mapping epistatic QTLs. Simulation studies suggested that all of the above methods, excluding the LASSO and PENAL approaches, performed satisfactorily. The Bin-REG method appears to outperform all other methods in terms of estimating positions and effects.

complex trait, chromosome segment substitution line (CSSL), epistasis, stepwise regression, Bayesian statistics

Citation: Tang Z X, Xiao J, Hu W M, et al. Bin-based model construction and analytical strategies for dissecting complex traits with chromosome segment substitution lines. Chin Sci Bull, 2012, 57: 2666–2674, doi: 10.1007/s11434-012-5195-y

Since the landmark study by Lander and Botstein [1] in the field of quantitative genetics, interest in the genetic analysis of complex traits has increased enormously. During the past century, numerous investigators have inferred the action of a polygene to be the underlying cause of a complex phenotype with continuous variation. The quantitative trait loci (QTLs) responsible for phenotypic variation can be mapped within a chromosome interval using conventional segregation populations, such as the F_2 and backcross mapping populations. Some QTLs have been cloned in rice, mouse, and other model organisms [2–4]. These achievements have enhanced our understanding of complex traits and enabled us to use marker-assisted selection (MAS) and genetic engineering to introduce valuable alleles into crops and improve crop breeding more effectively.

Genetic studies can provide important insights into the detailed molecular mechanisms that underlie the variation of complex traits. However, conventional populations have several limitations for accurate identification and fine mapping of QTLs [5,6]. One of the shortcomings of these populations is that a major QTL can overshadow a small-effect QTL by increasing the total phenotypic variation; thus, the

*Corresponding author (email: qtls@yzu.edu.cn)

small-effect QTL cannot reach the threshold of detection. An additional restriction of conventional populations is the background noise that results from the complex epistatic interactions of different loci. A number of studies have shown that these interactions substantially contribute to the genetic control and evolution of complex traits [7,8]. However, many other studies that attempted to explore the genetic basis of complex traits ignored the possibility that loci interact [9]. Furthermore, in existing mapping populations, the wide variation in plant growth rate and morphology strongly influence the effects of QTLs. It is also difficult to perform repeated tests, because each individual has a unique genotype. These problems pose major challenges for detecting and mapping QTLs in detail. Therefore, the development of new resources is necessary to facilitate fine mapping and cloning of QTLs.

To address these problems, Eshed and Zamir [5] pioneered research on the fine mapping of QTLs in tomato using introgression lines (ILs). Later, these novel mapping populations were further developed through successive introgression backcrosses and marker-assisted selection to produce chromosome substitution lines in *Arabidopsis* [10], chromosome segment substitution lines (CSSLs) and ILs in rice [11–14], recombinant chromosome substitution lines in barley [15], and backcross inbreed lines in lettuce [16] and tomato [17]. In animal and human genetic studies, Matin et al. [18] were the first to use a chromosome-substitution strain for QTL mapping. Singer et al. [19] described the construction of the first complete set of chromosome-substitution strains and their application in genome-wide QTL mapping. The advantages of these libraries are clear. As homozygous immortal lines, CSSLs can be phenotyped repeatedly and used for the simultaneous mapping of many traits. Additionally, each line contains a single chromosome segment that originates from the donor parent in an otherwise uniform genetic background [20]. As a result, the background genetic noise is reduced, so that each QTL can explain a greater proportion of the total phenotypic variation, which allows for a more detailed and reliable QTL identification [5,21] as well as permitting fine mapping [22] and cloning of the QTL [23]. In addition, CSSLs containing the QTL of interest can be backcrossed to various lines to investigate interactive effects and gene network effects [24]. Moreover, these libraries will improve our understanding of genetic traits that have biological and economic importance in plants and animals [25].

Although several CSSL libraries have recently been created in a number of plant and animal experimental models, an appropriate method based on these populations has not been developed for dissecting complex traits. Currently, most researchers use the standard *t*-test to identify QTLs [12,26]. In an ideal case, each substitution line carries one donor segment; thus, the genetic difference between the substitution line and the recurrent line can only be caused by the donor segment. However, false positives are often produced when the multiple *t*-tests are performed because of imprecise error estimates. In addition, the *t*-test is not suitable for CSSLs that each carry several donor segments. An alternative method to the standard *t*-test is Dunnett's test [17,27,28] and the *RSTEP-LRT* mapping method recently proposed by Wang et al. [29]. However, these approaches can only detect the major effect of a QTL on the basis of the novel population. A method that can distinguish between main and epistatic QTL effects has not yet been developed. In this article, we proposed a bin-based epistatic model and evaluated several methods in variable selection. Additionally, we performed a series of simulation experiments to test these methods. The goal of this study is to provide a suitable method for exploring the genetic basis of complex traits and, in doing so, to improve crop breeding.

1 Model construction

Figure 1 depicts the hypothetical CSSLs that were used in the present study. We present two types of CSSLs. Each CSSL contains one or several small, homozygous donor segments, shown as red bars in Figure 1. There was some overlap of donor segments between neighboring lines, such as between CSSL1 and CSSL2. High resolution QTL mapping will benefit from these overlaps. To localize a QTL to a smaller interval within a donor segment, the donor segment was divided into smaller segments according to the overlapping of different segments in each line (indicated by the vertical dashed lines). These smaller segments are referred to as bins [4]. A commonly used mapping scheme involves phenotyping all the lines in randomized replicate trials and presenting the results as a difference from the recurrent parent. The lines that contain a significant contribution can then be identified [28]. In our study, each bin is considered an independent variable. For example, bins from A to P in Figure 1(a) or bins from A to I in Figure 1(b) can be defined as x_1 to x_{16} and x_1 to x_9. As such, the main effect model can be described by the following multiple linear model:

$$y_i = b_0 + \sum_{k=1}^{m} b_k x_{ik} + e_i \,, \tag{1}$$

where y_i is the mean value of the ith line of a CSSLs library comprising l lines; b_0 is the overall mean of the population; m is the total number of bins in the entire genome; b_k is the main effect associated with bin k; x_{ik} is an indicator variable, denoting $x_{ik}=1$ for the donor parent bin and $x_{ik}=-1$ for the recurrent parent bin; and e_j denotes the residual error following a normal distribution.

The epistatic model can be easily derived from the above model and can be written as

$$y_i = b_0 + \sum_{k=1}^{m} b_k x_{ik} + \sum_{p,q\ p\neq q}^{m} b_{p\bullet q} x_{ip} x_{iq} + e_i, \tag{2}$$

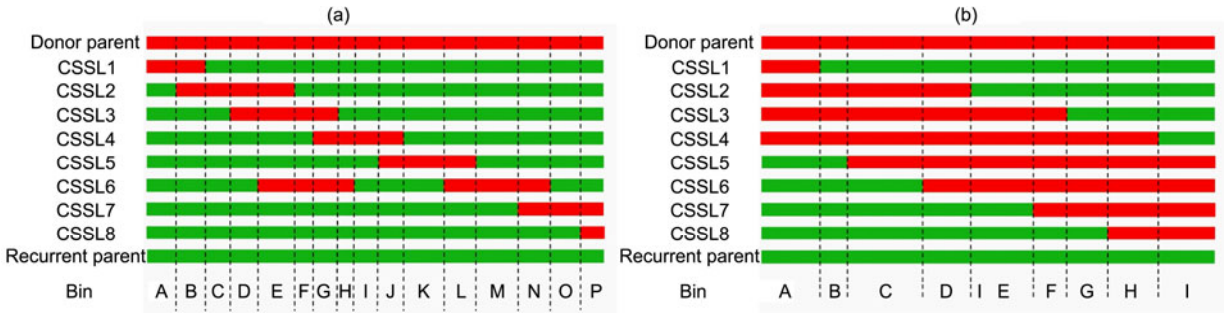

Figure 1 Model of the CSSL library. (a) The entire mapping population, consisting of individual lines each containing a single or a few homozygous donor segments in a uniform genetic background. (b) The introgression line (IL) library, consisting of a series of lines harboring a single, long, homozygous donor segment introgressed into a homogenous genetic background.

where $b_{p\cdot q}$ is the epistatic effect between the pth and qth bin. x_{ip} and x_{iq} have the same definition as x_{ik}. For the sake of clarity and notation, we redefined the design matrix and the partial regression coefficients as follows: let $b_j=b_k$ and $x_{ij}=x_{ik}$ when $j=k=1,\ldots, m$; let $b_j=b_{p\cdot q}$ and $x_{ij}=x_{ip}x_{iq}$ when $j=m+1,\ldots, m(m+1)/2$, $p=1,\ldots, m-1$ and $q=p+1,\ldots, m$. Then, model (2) can be rewritten as

$$y_i = b_0 + \sum_{j=1}^{m(m+1)/2} b_j x_{ij} + e_i \,. \tag{3}$$

When model (2) and model (3) are compared, it is apparent that $b_j=b_k$ if the jth effect is the main effect, and $b_j=b_{p\cdot q}$ if the jth effect is the epistatic effect; therefore, we used a general linear model (3) to describe both the main effect and the epistatic effect. In terms of the method of estimation, a distinction between a main effect and an epistatic effect is unnecessary.

However, not all of the bin pair interactions can be estimated with this model. Statistically, an interaction can be defined as the variation among the differences between the means for different levels of one factor over different levels of the other factor. Therefore, to consider the interaction between two bins, the mean value $y_{(1,1)}$ of substitution lines in bin pair ($x_{ip}=1$, $x_{iq}=1$), $y_{(1,-1)}$ in ($x_{ip}=1$, $x_{iq}=-1$), $y_{(-1,1)}$ in ($x_{ip}=-1$, $x_{iq}=1$), and $y_{(-1,-1)}$ in ($x_{ip}=-1$, $x_{iq}=-1$) should be evaluated, and the interaction effect estimated as $(y_{(1,1)}+y_{(-1,-1)})-(y_{(1,-1)}+y_{(-1,1)})$. If any of the means cannot be calculated, the interaction between the corresponding bin pairs cannot be estimated and must be removed from the model when analyzing the data set. For example, the interaction between bins A and B in Figure 1(a) could not be calculated because of the missing mean value of substitution lines in bin pair ($x_{iA}=1$, $x_{iB}=-1$).

The genotype indicator variable for a QTL is not observable. However, in our model, we assume that each QTL was placed in a bin; thus, the effect of the individual QTL is replaced by that of the corresponding bin. Therefore, the proposed method is essentially a multiple bin analysis. For this reason, we will use the terms bin and QTL interchangeably.

2 Variable selection strategy

This study considers the estimation of the coefficients of a linear regression model with a dependent variable y and a large number regressor x. Usually, a model is said to be "large" if $m(m+1)/2>l$. Here, model (3) saturates quickly as the number of bins increases. Thus, the ordinary least-square approach does not have a unique solution. We assume that the number of variables that are known to affect phenotypic value y_i is less than the number of substitution lines. Although the number of parameters m or $m(m+1)/2$ can be very large, most of them will be zero. Therefore, research has focused on selecting the variables that significantly affect y_i. In this section, several methods capable of dealing with the large regression model are presented and evaluated in the following simulation study.

2.1 Bin-based stepwise regression method

For simplicity, we call this method Bin-REG here. In statistics, stepwise regression is the most intuitive method for choosing predictive variables and is carried out by an automatic procedure. The main approach includes the following steps: (i) forward selection, which involves starting out with zero variables in the model and testing the variables one by one. Variables that are statistically significant are then included in the analysis. (ii) Backward selection involves starting out with all of the candidate variables, evaluating them one by one for statistical significance and eliminating any that meet the criterion for removal. (iii) Stepwise selection, a method that is a combination of (i) and (ii), involves testing at each stage for variables that will be included or excluded. In a stepwise regression analysis, the selection criterion is one of the key issues in variable selection. Usually, this takes the form of a sequence of F-tests, but other techniques are also possible.

2.2 Spike and slab variable selection

Stochastic search variable selection (SSVS) is a fully Bayesian

variable selection method implemented via Markov chain Monte Carlo (MCMC) and was originally proposed by George and McCulloch [30]. Yi et al. [31] applied this method to multiple QTL analyses. In SSVS, the dimensionality of the model is not changed by limiting the posterior distribution of nonsignificant variables to a small value near zero, instead of removing them from the model. Therefore, SSVS can be easily implemented via the Gibbs sampler and can provide the posterior probability of each variable that is included in the model. The effect of each variable can then be evaluated.

2.3 LASSO algorithm

Least absolute shrinkage and selection operator, or LASSO, is a shrinkage and selection method for linear regression [32]. As a method of model selection designed to be used when the number of variables is larger than the number of observables, LASSO minimizes the residual sum of squared errors, with a limit on the sum of the absolute values of the coefficients. The nature of this constraint allows LASSO to produce some coefficients that are exactly zero and, as a result, provides an interpretable model.

2.4 Penalized maximum likelihood method

The penalized maximum likelihood (PENAL) method was originally developed by Zhang and Xu [33] and does not remove all the nonsignificant variables from the model; thus, PENAL can also handle supersaturated models. The proposed method adopts a penalty that depends on the values of the parameters and allows spurious QTL effects to be minimized toward zero, while QTLs with large effects are estimated with virtually no shrinkage. Under the shrinkage estimation framework for a supersaturated model, PENAL can produce similar results to the fully Bayesian shrinkage approach and is quickly computable.

2.5 Empirical Bayes method

Empirical Bayes (E-BAYES) methods use empirical data to evaluate the conditional probability distributions and combine Bayesian and frequentist approaches in the estimation. These methods have been introduced into statistical genomics by Beasley et al. [34] and Zhang et al. [35]. Recently, Xu [36] proposed an empirical Bayes method that can simultaneously estimate the main effects of all individual markers and the epistatic effects of all pairs of markers. This method does not require MCMC samplings, but can still estimate the variance parameters for prior regression coefficients. The method is intended for estimating epistatic effects in situations where many of them are actually zero. More recently, Xu and Jia [37] applied the empirical Bayes method to simultaneously estimate the main effects for all markers and interaction effects for all marker pairs in a single model.

3 Simulation study

To illustrate the application of the above methods, we simulated two data sets according to the CSSLs model depicted in Figure 1, which denote two types of substitution lines. For simplicity, we call the two novel populations librarys A and B, respectively. Library A, which consists of 62 lines including two parent lines, was simulated for a genome of 1970 cm with 12 chromosomes. Similarly, Library B, which consists of 135 lines, including two parent lines, was simulated for a single giant chromosome of 1660 cm. The length of the substituted component from the donor parent in each line was generated randomly. According to the overlap between each donor segment, we created 112 and 107 mapping bins for Library A and Library B, respectively. Four of the bins overlapped with the main effect QTL and three out of all possible bin pairs had interaction effects. The genetic variance σ_g^2 was approximately 3.83 and 14.5 for library A and library B, respectively, which was calculated by $\sigma_g^2 = \sum_{q=1}^{4} \sigma_q^2 + \sum_{i=1}^{3} \sigma_i^2$, where $\sigma_q^2 = a^2 - (l-2f)^2 a^2/l^2$ and $\sigma_i^2 = a_i^2 - (l-2f)^2 a_i^2/l^2$. a and a_i represent the main effect and the epistatic effect, respectively, and f equals the number of x_{ij}=1 in the column associated with a QTL and QTL interaction in the design matrix. The residual variance σ_e^2 was defined as $\sigma_e^2 = (1-H^2)\sigma_p^2$, where H^2 was the total hereditary capacity. H^2=0.8 was chosen for our simulation study, and σ_e^2 for the two populations was 0.96 and 3.6 for library A and library B, respectively. The phenotypic variance σ_p^2 was about 4.79 and 18.1 for library A and library B, respectively. The theoretical proportion of the phenotypic variance contributed by each individual QTL and the interaction were simply defined by $h^2 = \sigma_q^2/\sigma_p^2$ and h^2= σ_i^2/σ_p^2, respectively. In our simulation experiment, the h^2 of an individual QTL and a pair of QTLs ranged from 2.6% to 20.2%. Some of the QTLs had main effects only, while others had both main and epistatic effects. Additionally, some QTLs with epistatic effects had no main effects. In total, the models contained 6383 and 5779 effects for library A and library B, respectively, which is approximately 102 and 42 times as large as the sample size for library A and library B, respectively.

To evaluate the different methods, we analyzed simulated dataset based on libraries A and B. The Bin-REG method was adopted to analyze the simulated data through the SAS/IML program. The LASSO method was implemented using the 'lasso' option of GLMSELECT in SAS/STAT. The SAS/IML procedures implementing SSVS, PENAL, and E-BAYES methods are available under the Paper In-

formation link at the *Biometris* website (http://www.statgen.ucr.edu/) and were written by Xu [36].

In the Bin-REG method, the significance levels required for a variable to enter and stay in the regression are specified by the 'sle' and 'sls' options, respectively. The default for both parameters is 0.15, and 0.01 was chosen as the significance level. In the GLMSELECT procedure of the SAS/STAT model, 'lasso' was chosen as the model selection method to implement the LASSO algorithm. A modification of SSVS was chosen to determine a prior probability of δ_j=1, where δ_j is an indicator variable used to include or exclude the *j*th bin effect. Typically, the original SSVS method $p(\delta_j)=\rho^{\delta_j}(1-\rho)^{1-\delta_j}$, where 0<$\rho$<1 is a constant and $\rho=0.5$, is used [30]. However, Xu [36] suggested that $\rho=0.5$ was only suitable when the number of predictors was relatively small, such as in the main effect QTL model. For the epistatic effect model, ρ should be very small. For this reason, $\rho=0.1$ was chosen for our experiment. The hyper parameters of E-BAYES were (τ, ω)=(−1, 0.003), in accordance with Xu [36]. We also tested other hyper parameters and found that other values did not shrink the parameters properly.

The results for Library B are presented in Table 1 and plotted in Figure 2. Simulation results for library A are not presented because the results were uninterruptible. One possible reason for this result may be that the design matrix for library A was not suitable for dissecting statistical epistatic effects. In comparing the same data sets from Library B (Table 1 and Figure 2), we found that the Bin-REG method generated better results than LASSO, SSVS, E-BAYES, and PENAL on the estimates of QTL position. SSVS and E-BAYES produced similar results, with the exception that the SSVS method missed one QTL with a large epistatic effect. The E-BAYES method, using the hyper parameters setting of (τ, ω)=(−1, 0.003) reported a confounded result. Bin pairs (3, 100) and (27, 50) interfered with each other and generated spurious interactions as bin pairs (3, 50) and (27, 100). In addition, exact effect estimation could also be observed. Overall, Bin-REG, SSVS, and E-BAYES produced satisfactory results. However, LASSO and PENAL differed from the other methods in that: (i) most of the large effect was overshrunk; (ii) the estimated QTLs with main effects only were biased in position estimates, while the other methods detected the position of these QTLs exactly, and (iii) the epistatic QTLs were not detected.

4 Discussion

Novel CSSLs have been developed in several plant and animal species for fine mapping, cloning, and functional research on QTLs [11,19,24]. Ideal CSSLs carrying one donor segment can be analyzed efficiently using the *t*-test method, which is commonly used by researchers. However, the disadvantage of the *t*-test is apparent. When multiple *t*-tests are performed, researchers run the risk of greatly inflating the family wise error rate (FWER) or the false positive rate (FPR), defined as the probability of making one or more Type I errors [38]. Thus, some donor segments containing no QTLs would show a significant difference from the background parent. Non-ideal CSSLs may contain two or more donor segments per substitution line. Under this situation, the *t*-test method cannot distinguish which segment contains the QTL of interest. In contrast, the target QTL can be localized within a smaller segment by a bin mapping method [4,28]. However, the problem of high FPR still exists for bin mapping methods. To control the high FPR in multiple *t*-tests, Dunnett's test has been suggested [17,27,28]. In reality, the critical value of Dunnett's test is always higher than that of *t*-tests, which means that stricter criteria are used in Dunnett's test, and as a result, some QTLs with small effects may not be detected.

To address this problem, Tang and Xu [39] proposed an improved *t*-test method. First, the variations within each CSSL

Table 1 Simulated QTL positions and effects, and estimates using various methods on the design matrix of Library B

Bins (i, j)	True value	h^2 (%)	Bin-REG		LASSO[a)]		SSVS		E-BAYES		PENAL[a)]	
			Position	Estimated effect	Position	Estimated effect	Position	Estimated effect	Position	Estimated effect	Position	Estimated effect
(3, 3)	1.8	17.9	(3, 3)	1.68±0.08	(3, 3)	1.70	(3, 3)	1.66±0.17	(3, 3)	1.53±0.09	(1,1)	2.40±0.24
(27, 27)	1.3	9.3	(27, 27)	1.47±0.08	(27, 27)	1.16	(27, 27)	1.23±0.70	(27, 27)	2.00±0.08	(25, 25)	2.55±0.24
(39, 39)	1.5	12.4	(39, 39)	1.33±0.07	(42, 42)	0.19	(39, 39)	1.27±0.57	(41,41)[b)]	0.98±0.15	–	–
(100, 100)	−1.3	9.3	(100, 100)	−1.31±0.07	(102, 102)	−0.06	(100, 100)	−1.20±0.54[c)]	(100, 100)	−1.61±0.08	–	–
(3, 100)	2.3	10.1	(3, 100)	2.35±0.07	–[d)]	–	(3, 100)	2.57±0.33	(3, 50)[e)]	2.26±0.39	–	–
(27, 50)	2	15.3	(27, 50)	2.01±0.05	(50, 50)	0.60	(27, 49)	1.40±0.47	(27, 100)[e)]	1.89±0.38	–	–
(66, 80)	1.6	5.6	(66, 80)	1.56±0.07	–	–	–	–	(66, 80)	1.78±0.38	–	–

a) Some of the estimated bins are confused, and some of them are obtained from the neighboring bins using the LASSO and PENAL algorithms for the bins pair (i, j)). b) The estimated bin is shifted to the neighboring position. c) The estimated effect was obtained from both the assumed bin and a neighboring bin. d) The short horizontal lines denote that the assumed bins or bin pairs have not been detected. e) The two bin pairs interfere with each other.

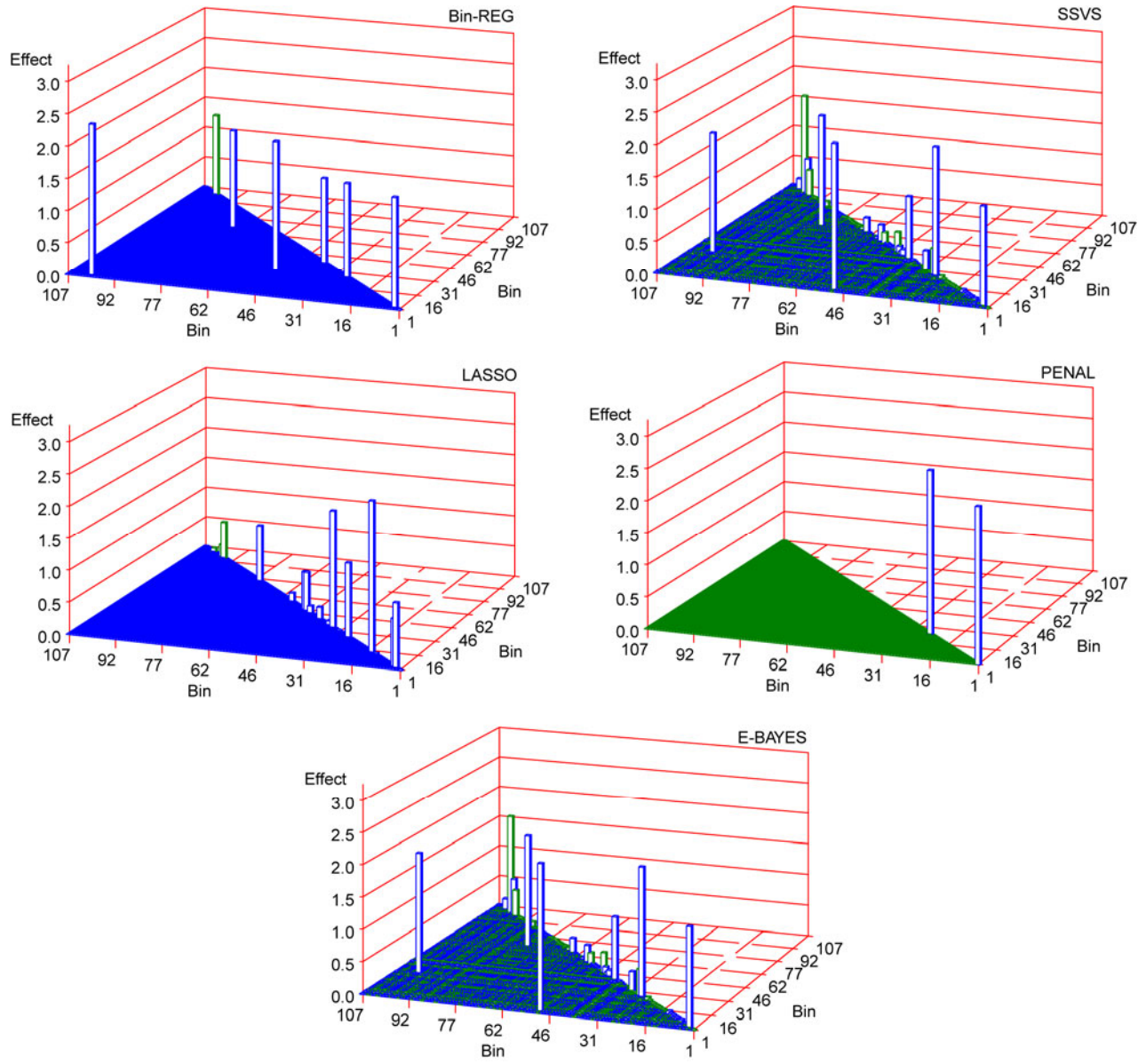

Figure 2 Estimated main and epistatic effects of bin pairs from library B using the following four methods and the same simulated data set: the Bin-REG method proposed in this study, SSVS, LASSO, PENAL, and E-BAYES. The green- and blue-colored prisms represent estimated positive and negative effects, respectively.

are combined to estimate environmental variance *Mse*; the improved protected least significant difference (IPLSD) can then be used as a new criterion to examine the difference between each CSSL and the background parent. The IPLSD is calculated using the formula $\mathrm{IPLSD}_{\alpha}=t_{\alpha,df}\times\sqrt{\frac{2MSe}{n}}$, where n is the number of individuals of each line. For non-ideal CSSLs classified as Library A in our paper, we also proposed a bin-based regression method to detect a small chromosome region of interest using the main effect model (Eq. (1)) [39]. We assumed that several bins contained only main effects in libraries A and B and analyzed the simulated data using the main effect model; the results show that bins of interest can be detected with high statistical power. Further analyses revealed that increasing the number of individuals within each line can improve both the statistical power of QTL detection and QTL effects estimation, especially for a QTL with heritability lower than 5%, which is consistent with our general expectations.

We used the proposed main effect model to analyze the rice dataset [40], with the aim of identifying loci that influence heading data, which is an important agronomic characteristic in any rice breeding program. A collection of SSSLs consisting of 52 lines was created from six donors with an elite cultivar 'Huajingxian 74' genetic background; 20 of these individuals were included in each line. On the

basis of the linkage map information, we constructed the design matrix and chromosome bin map. A total of 162 markers covering the genome were used in the analysis, and 56 bins were generated. The proposed approach (Eq. (1)) was used for data analysis using the regression options 'sls=0.1' and 'sle=0.1'. As we expected, our result was consistent with the *t*-test result obtained by He et al. [40], with the exception that three of the 30 donor segments could not be detected. Detailed examination showed that these three segments had small effects and relatively smaller contributions. However, our simulation studies suggest that under options 'sls=0.1' and 'sle=0.1', spurious effects will be generated because of the low significance level.

In investigating the genetic basis of complex traits, the extent to which epistasis controls variation in complex traits can never be explored using the main effect model. Hence, simultaneous mapping of QTLs using an epistatic model is needed, as it can detect the loci that mainly affect the quantitative trait through epistatic interactions with another locus. In our study, donor segments were divided into small bins according to the overlap of segments in the CSSLs, and each bin was considered a different indicator variable describing the different parental origin. Based on this design matrix, an epistatic QTL mapping model (Eqs. (2) and (3)) was constructed that is more flexible than that (Eq. (1)) for individual QTLs.

The epistatic model is essentially an oversaturated linear model. Our primary interest concerned the regression coefficients (additive and epistatic effects, both denoted by b_j). In this study, we evaluated several approaches capable of dealing with the oversaturated linear model. The classic regression method was sufficiently robust to analyze the dataset. We chose stepwise selection with regression options 'sls=0.01' and 'sle=0.01'. We also tested a significance level ranging from 0.01 to 0.15, and the result produced the target QTL in addition to many spurious QTLs that were also generated. Furthermore, we performed a simulation study using different sample sizes and different levels of heritability, and the results were consistent with our general expectations. Large sample sizes and high heritability produced accurate estimates with small estimation errors. The simulated results suggested that a guarantee of high heritability is important in detecting the target QTL and QTL interactions correctly. Our simulation also showed that the forward selection strategy produces a similar result; the backward selection strategy does not, because of the different strategy used in variable selection.

In an oversaturated model, a heuristic search is possible, but may not ensure the generation of an optimal model within a reasonable time frame, even when using a super computer [37]. Bayesian model selection, by taking advantage of the MCMC method, is a more efficient algorithm than both exhaustive and heuristic searches [30]. We employed SSVS, which is a Bayesian model selection algorithm, and E-BAYES to analyze the simulated dataset. Parameter $\rho = 0.1$ [36] was used in the SSVS algorithm, and the hyper parameters (τ, ω)=(−1, 0.003) were used for E-BAYES. According to our simulation studies, other parameter settings cannot guarantee satisfactory results based on the design matrix of library B. However, running SSVS takes a considerable amount of time on any computer. We also tried E-BAYES using the hyper parameters setting (τ, ω)=(−2, 0), which means a uniform prior was used for each variance component, as suggested by Xu [36]. The results from this analysis showed that numerous spurious effects were generated; however, bins of interest were still detected, though the analysis misidentified bin pairs (3, 100) and (27, 50) as (3, 50) and (27, 100), respectively.

The comparison of different algorithms applied to the data from library B revealed that the PENAL and LASSO algorithms failed to estimate the genetic parameters. The precise reason for this failure is unclear, but multicollinearity of the oversaturated independent variables has contributed to this result. Overall, the Bin-REG method proposed in this study provides a direct approach to locate QTLs with main or epistatic effects in small chromosome bins. Our simulation studies suggest that this approach outperforms LASSO, PENAL, SSVS, and E-BAYES in terms of estimating position and effects. Additional simulation studies also suggest that the statistical power of Bin-REG is very high for both main effects and epistatic effects (Table 2). Surprisingly, the statistical power is close to 100% for the

Table 2 Simulated QTL positions and effects, and the estimated values from Library B

Bins (i, j)[a)]	True value	h^2 (%)	Estimated bins	Power[b)]	Estimated value[c)]	Standard error
(3, 3)	1.8	17.9	(3, 3)	100	1.90	0.51
(27, 27)	1.3	9.3	(27, 27)	100	1.30	0.13
(39, 39)	1.5	12.4	(39, 39)	100	1.51	0.14
(100, 100)	−1.3	9.3	(100, 100)	100	−1.29	0.15
(3, 100)	2.3	10.1	(3, 100)	100	2.48	0.49
(27, 50)	2.0	15.3	(27, 50)	99	2.00	0.21
(66, 80)	1.6	5.6	(66, 80)	99	1.62	0.17

a) When $i=j$, the QTL is a main effect; otherwise, it is an epistatic effect. b) The power was obtained from both the assumed bin pairs and a neighboring bin pair, rather than only from simulated bin pairs. c) The estimated effect is expressed as the weighted mean multiplied by the power and mean value of significant QTLs or QTL interaction.

bin pair (66, 80), with a heritability of about 5%. In conclusion, accurate and precise estimates of QTL main effects and epistatic effects can be produced using the Bin-REG method with Library B (Table 2).

The epistatic model could not be used to analyze the data from library A through the epistatic regression method. Some possible reasons for this failure include: (1) strong collinearity or multicollinearity of the independent variables; and (2) the small variation of each independent variable. Figure 1(a) illustrates these two possibilities clearly. The failure of these algorithms was reflected by a confused result. However, that does not mean this population is not suitable for mapping genes. Each line of library A consists of a series of lines harboring a single homozygous donor segment introgressed into a uniform genetic background. The genetic noise from the genomic background is well controlled. The benefit is that even a QTL with a small effect can be identified significantly. Furthermore, crosses between individual introgression lines, each bearing one of the interacting alleles, can be set up to investigate the extent of the interaction [24,41,42].

This work was supported by the National Basic Research Program of China (2011CB100106), the National Natural Science Foundation of China (30971846 and 31171187), and the Vital Project of Natural Science of Universities in Jiangsu Province (09KJA210002) to C. Xu, the National Natural Science Foundation of China (31100882) to Z. Tang, and National Natural Science Foundation of China (31000539) to J. Xiao.

1 Lander E S, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 1989, 121: 185–199
2 Salvi S, Tuberosa R. To clone or not to clone plant QTLs: Present and future challenges. Trends Plant Sci, 2005, 10: 297–304
3 Flint J, Valdar W, Shifman S, et al. Strategies for mapping and cloning quantitative trait genes in rodents. Nat Rev Genet, 2005, 6: 271–286
4 Paran I, Zamir D. Quantitative traits in plants: Beyond the QTL. Trends Genet, 2003, 19: 303–316
5 Eshed Y, Zamir D. An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics, 1995, 141: 1147–1162
6 Doerge R W. Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet, 2002, 3: 43–52
7 Cheverud J M. Detecting epistasis among quantitative trait loci. In: Wade M, Brodie B, Wolf J, eds. Epistasis and the Evolutionary Process. New York: Oxford University Press, 2000. 58–81
8 Malmberg R L, Held S, Waits A, et al. Epistasis for fitness-related quantitative traits in *Arabidopsis thaliana* grown in the field and in the greenhouse. Genetics, 2005, 171: 2013–2027
9 Carlborg O, Haley C S. Epistasis: Too often neglected in complex trait studies? Nat Rev Genet, 2004, 5: 618–625
10 Koumproglou R, Wilkes T M, Townson P, et al. STAIRS: A new genetic resource for functional genomic studies of *Arabidopsis*. Plant J, 2002, 31: 355–364
11 Xi Z Y, He F H, Zeng R Z, et al. Development of a wide population of chromosome single-segment substitution lines in the genetic background of an elite cultivar of rice (*Oryza sativa* L.). Genome, 2006, 49: 476–484
12 Ebitani T, Takeuchi Y, Nonoue Y, et al. Construction and evaluation of chromosome segment substitution lines carrying overlapping chromosome segments of indica rice cultivar kasalath in a genetic background of *japonica* elite cultivar koshihikari. Breed Sci, 2005, 55: 65–73
13 Kubo T, Aida Y, Nakamura K, et al. Reciprocal chromosome segment substitution series derived from *japonica* and *indica* cross of rice (*Oryza sativa* L.). Breed Sci, 2002, 52: 319–325
14 Doi K, Iwata N, Yoshimura A. The construction of chromosome substitution lines of African rice (*Oryza glaberrima* Steud.) in the background of Japonica rice (*O. sativa* L.). Rice Genet Newslett, 1997, 14: 39–41
15 Matus I, Corey A, Filichkin T, et al. Development and characterization of recombinant chromosome substitution lines (RCSLs) using *Hordeum vulgare* subsp. *spontaneum* as a source of donor alleles in a *Hordeum vulgare* subsp. *vulgare* background. Genome, 2003, 46: 1010–1023
16 Jeuken M J, Lindhout P. The development of lettuce backcross inbred lines (BILs) for exploitation of the *Lactuca saligna* (wild lettuce) germplasm. Theor Appl Genet, 2004, 109: 394–401
17 Finkers R, van Heusden A W, Meijer-Dekens F, et al. The construction of a *Solanum habrochaites* LYC4 introgression line population and the identification of QTLs for resistance to *Botrytis cinerea*. Theor Appl Genet, 2007, 114: 1071–1080
18 Matin A, Collin G B, Asada Y, et al. Susceptibility to testicular germ-cell tumours in a 129.MOLF-Chr 19 chromosome substitution strain. Nat Genet, 1999, 23: 237–240
19 Singer J B, Hill A E, Burrage L C, et al. Genetic dissection of complex traits with chromosome substitution strains of mice. Science, 2004, 304: 445–448
20 Zamir D. Improving plant breeding with exotic genetic libraries. Nat Rev Genet, 2001, 2: 983–989
21 Rousseaux M C, Jones C M, Adams D, et al. QTL analysis of fruit antioxidants in tomato using *Lycopersicon pennellii* introgression lines. Theor Appl Genet, 2005, 111: 1396–1408
22 Monforte A J, Tanksley S D. Fine mapping of a quantitative trait locus (QTL) from *Lycopersicon hirsutum* chromosome 1 affecting fruit characteristics and agronomic traits: Breaking linkage among QTLs affecting different traits and dissection of heterosis for yield. Theor Appl Genet, 2000, 100: 471–479
23 Frary A, Nesbitt T C, Grandillo S, et al. *fw2.2*: A quantitative trait locus key to the evolution of tomato fruit size. Science, 2000, 289: 85–88
24 Lin H X, Yamamoto T, Sasaki T, et al. Characterization and detection of epistatic interactions of 3 QTLs, Hd1, Hd2, and Hd3, controlling heading date in rice using nearly isogenic lines. Theor Appl Genet, 2000, 101: 1021–1028
25 Peleman J D, van der Voort J R. Breeding by design. Trends Plant Sci, 2003, 8: 330–834
26 Wan X Y, Wan J M, Su C C, et al. QTL detection for eating quality of cooked rice in a population of chromosome segment substitution lines. Theor Appl Genet, 2004, 110: 71–79
27 Dunnett C W. A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc, 1955, 50: 1096–1121
28 Fridman E, Liu Y S, Carmel-Goren L, et al. Two tightly linked QTLs modify tomato sugar content via different physiological pathways. Mol Genet Genomics, 2002, 266: 821–826
29 Wang J, Wan X, Crossa J, et al. QTL mapping of grain length in rice (*Oryza sativa* L.) using chromosome segment substitution lines. Genet Res, 2006, 88: 93–104
30 George E I, McCulloch R E. Variable selection via Gibbs sampling. J Am Stat Assoc, 1993, 88: 881–889
31 Yi N, George V, Allison D B. Stochastic search variable selection for identifying multiple quantitative trait loci. Genetics, 2003, 164: 1129–1138
32 Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B, 1996, 58: 267–288
33 Zhang Y M, Xu S. A penalized maximum likelihood method for estimating epistatic effects of QTL. Heredity, 2005, 95: 96–104
34 Beasley T M, Wiener H, Zhang K, et al. Empirical bayes method for incorporating data from multiple genome scans. Hum Hered, 2005, 60: 36–42

35 Zhang K, Wiener H, Beasley M, et al. An empirical Bayes method for updating inferences in analysis of quantitative trait loci using information from related genome scans. Genetics, 2006, 173: 2283–2296
36 Xu S. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics, 2007, 63: 513–521
37 Xu S, Jia Z. Genomewide analysis of epistatic effects for quantitative traits in barley. Genetics, 2007, 175: 1955–1963
38 Bekiroğlu N. Multiple *t*-tests or ANOVA (analysis of variance)? Turk Respir J, 2001, 2: 21–22
39 Tang Z X, Xu C. A preliminary study of mapping genes underlying complex traits based on chromosome segment substitution lines. Mol Plant Breed, 2007, 5: 242–244
40 He F, Xi Z, Zeng R, et al. Mapping of heading date *qtls* in rice (*Oryza sativa* L.) using single segment substitution lines. Sci Agricul Sin, 2005, 38: 1505–1513
41 Peleman J D, van der Voort J R. Breeding by design. Trends Plant Sci, 2003, 8: 330–334
42 Eshed Y, Zamir D. Less than additive epistatic interactions of QTL in tomato. Genetics, 1996, 143: 1807–1817