BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Variable selection for large *p* small *n* regression models with incomplete data: Mapping QTL with epistases

Min Zhang[1], Dabao Zhang*[1] and Martin T Wells[2]

Address: [1]Department of Statistics, Purdue University, West Lafayette, IN 47907 USA and [2]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853 USA

Email: Min Zhang - minzhang@stat.purdue.edu; Dabao Zhang* - zhangdb@stat.purdue.edu; Martin T Wells - mtw1@cornell.edu

* Corresponding author

Published: 29 May 2008

BMC Bioinformatics 2008, **9**:251 doi:10.1186/1471-2105-9-251

Received: 26 November 2007
Accepted: 29 May 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/251

© 2008 Zhang et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Identifying quantitative trait loci (QTL) for both additive and epistatic effects raises the statistical issue of selecting variables from a large number of candidates using a small number of observations. Missing trait and/or marker values prevent one from directly applying the classical model selection criteria such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC).

Results: We propose a two-step Bayesian variable selection method which deals with the sparse parameter space and the small sample size issues. The regression coefficient priors are flexible enough to incorporate the characteristic of "large *p* small *n*" data. Specifically, sparseness and possible asymmetry of the significant coefficients are dealt with by developing a Gibbs sampling algorithm to stochastically search through low-dimensional subspaces for significant variables. The superior performance of the approach is demonstrated via simulation study. We also applied it to real QTL mapping datasets.

Conclusion: The two-step procedure coupled with Bayesian classification offers flexibility in modeling "large p small n" data, especially for the sparse and asymmetric parameter space. This approach can be extended to other settings characterized by high dimension and low sample size.

Background

With the advent of high-throughput biotechnologies to genotype dense molecular markers throughout the genome, statistical methodologies are crucial in understanding the genetic architecture of complex traits, and in locating genes underlying important traits. Since the pioneering statistical work by Lander and Botstein [1], much effort has been devoted to improving the efficiency and accuracy of QTL mapping. Traditional approaches to QTL mapping test each of dense grid loci on chromosomes via the likelihood ratios of linear regression models (see the reviews by Doerge et al. [2] and Broman and Speed [3]), and Wang et al. [4] also proposed a Bayesian shrinkage estimation of QTL parameters allowing varying shrinkage factors across different effects.

Epistases (that is, interactions between genes) are ubiquitous in biological systems [5] and may even play a more important role than additive effects, as have been shown in human population [6,7] and other organisms [8-12]. However, even a moderate number of markers implies a large number of pairwise combinations, thus creating statistical issues in QTL mapping. Due to the small sample sizes and the lack of efficient statistical tools, the number of identified genes is limited although the existence of epistasis has been recognized for nearly a hundred years [13]. To detect epistatic effects, Kao and Zeng [14] proposed modeling epistasis via orthogonal contrast scales using Cockerham's model; Yi and Xu [15] developed a Bayesian method to detect epistasis using reversible jump Markov chain Monte Carlo (MCMC) algorithm; Yi et al. [16-18] then proposed a Bayesian model selection approach to detect genome-wide epistasis (with the software described in [19]); Bogdan et al. [20] modified Bayesian information criterion (mBIC) to permit the identification of additive effects as well as pairwise interactions; and Cui and Wu [21] also proposed a statistical framework to detect genetic interactions derived from different genomes in self-pollinated plants. Recently, *Żak* et al. [22] developed a rank-based model selection and Shi et al. [23] developed a LASSO-type penalized likelihood method to locate interacting QTL while Bogdan et. al [24] extended mBIC for strongly correlated markers and multiple interval mapping.

Consider Y_i as the trait value of strain $i = 1, \ldots, n$, and let X_{ij} be the genotypic value of marker $j = 1, \ldots, p_\beta$ within the i-th strain. Here we focus on the populations with binary markers X_{ij} (coded as -0.5 and 0.5), such as doubled-haploid, backcross or recombinant inbred lines. With available markers (either observed or imputed) densely located on chromosomes, we assume the putative QTL co-transmit with some of the markers. Let $\mathcal{I}\{X\}$ denote the set including all pairwise epistases of interest, and define $Z_{ij} = X_{ik} X_{il}$ for the j-th candidate epistasis $(k, l) \in \mathcal{I}\{X\}$, $j = 1, \ldots, p_\gamma$. We investigate the additive effects of putative QTL and the epistatic interactions between them through the following multiple regression model,

$$Y_i = \mu + \sum_{j=1}^{p_\beta} \beta_j X_{ij} + \sum_{j=1}^{p_\gamma} \gamma_j Z_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2), \tag{1}$$

where μ is the overall mean, β_j is the additive effect of marker j, γ_j represents the j-th epistatic effect, and ε_i is the random error.

QTL mapping with this multiple regression model can be viewed as a model selection procedure [3,25-27]. However, several characteristics of the data complicate the application of classical statistical methodologies. First, a large amount of missing molecular markers, due to failure in genotyping or selective genotyping, is common in practice. When markers are sparse, the missing genotype information between markers must be inferred. Second, the molecular markers in the same linkage group may be highly correlated. Third, the total number of molecular markers and putative epistases, i.e., $p = p_\beta + p_\gamma$, is usually much larger than the sample size n. Because of these issues, the efficiency and accuracy are usually compromised for easy development of statistical approaches. Characteristics of the "large p small n" data with missing values require further attention via extensions of traditional model selection approaches. We extend the Bayesian classification approach in Zhang et al. [28] to map QTL with epistases. Spike and slab priors have been used by, for example, Mitchell and Beauchamp [29], George and McCulloch [30], and Ishwaran and Rao [31] to develop Bayesian variable selection approaches. The spike and slab priors consist of two components, with one modeling zero coefficients and the other modeling nonzero ones.

Furthermore, the mixing weight plays a crucial role in condensing the searchable parameter space and enforcing a stochastic search within low-dimensional spaces. When only a limited number of covariates are being investigated, a uniform distribution on [0, 1] or even a fixed value (e.g., 0.5) is usually chosen for the mixing weight. However, when $n << p$, it is unrealistic to expect half of the variables to be selected because the final model may still be unidentifiable. Instead, we expect that, for a successful variable selection, the prior distributions of the mixing weights depend on both n and p.

We investigate the predictability of a model developed for a dataset of sample size n, and tackle the aforementioned issues. We then construct a two-step Bayesian variable selection approach for model (1) in the case that $n << (p_\beta + p_\gamma)$. In the first step, we employ a restrictive prior for each of the coefficients in model (1) in order to enforce stochastic filtering of the large number of candidate variables. This prior also allows flexibility for the possible different numbers and/or scales of positive and negative coefficients (see [32] for more details on its advantage over symmetric priors). A Gibbs sampling algorithm is developed to compute the posterior distributions and to implement the stochastic search. Only a limited number of variables are filtered to go through the second step, which repeats the first step but with much fewer candidate variables. The second step is necessary to model (1) when $n << (p_\beta + p_\gamma)$, as the priors in the first step could potentially be too restrictive. The performance of our approach is evaluated via a simulation study and application to real datasets.

Results and Discussion

Simulation

Simulation studies were performed to evaluate the performance of our method in the case of $p \gg n$. We simu-

Table 1: Design of the simulation studies.

Type of effect	Marker(s)	Effect size	Heritability
Additive effect	2–47.8	0.8	0.0989
	3–141.5	1.2	0.2225
Interaction effect	(1–24.7, 2–47.8)	1.7321	0.1159
	(2–47.8, 3–141.5)	1.7321	0.1159
	(2–133.8, 3–56.7)	-1.7321	0.1159

Each marker is referred by its chromosome index and its location on the chromosome.

lated 56 markers across 3 chromosomes, with each having 10, 20, and 26 markers, and being 56.7 cM, 133.5 cM and 171.6 cM long respectively. We specify σ_{ε}^{2} = 0.5415, and the locations of 28 markers are chosen based on the Drosophila data [28], which include 221 inbred introgression lines between two closely related species. The other 28 markers are chosen such that the neighboring markers are at least 5 cM away. Table 1 shows the detailed information of the non-zero effects specified in the simulation, including two additive effects and three epistatic effects. To assess whether our method is able to identify different types of epistatic effects, we include all three possible interactions in the simulation: (1) neither of the two markers has additive effects (that is, 2–133.8 and 3–56.7); (2) one of them has additive effects (that is, 1–24.7 and 2–47.8); (3) both have additive effects (that is, 2–47.8 and 3–141.5). All epistatic effects were set at the same size to avoid its effects on detectability. Due to the intensive computation involved in Gibbs sampling, a total of 100 complete data sets were simulated. Each of the 100 data sets was analyzed using two models, one model with both additive and epistatic effects while the other with additive effects only. When mapping QTL with epistases, we have a total number of 1596 variables (56 additive-effect loci and 1540 epistases) versus 221 observations in the model.

For the model without epistases, both markers can be detected in most of the 100 simulated datasets even when the false discovery rate (FDR) is controlled as low as 0 (via setting the Bayes factor higher than 3.2), see Table 2. When modeling the epistases, all (additive and interaction) effects are still detected in more than 90% of the data sets for all levels of Bayes factor (BF) though the FDRs are higher. For those data sets with any effect not identified, the immediate neighbors of the corresponding marker locus are mostly detected instead. As expected, it is more difficult to detect epistases than to detect additive effects. The epistasis of markers both having additive effects is the easiest to be detected among all epistases. The true parameter values are included in their 95% credible intervals with the associated posterior probabilities being very close to one (results not shown).

Application

We apply the developed method to the *simulans* backcross II (BS2) data and the *mauritiana* backcross II (BM2) data [33,34]. An F$_1$ population was first produced by females from an inbred line of *D. simulans* and males from an inbred line of *D. mauritiana*. Then the F$_1$ females were backcrossed to the parental line of *D. simulans*, which was fixed for different alleles at 45 marker loci, to produce a *simulans* backcross (BS) population. A *mauritiana* backcross (BM) population was also produced by backcrossing the F$_1$ females to the other parental line. Based on the two different times of crossing, a total of four data sets were obtained, namely, BS1 (*n* = 186), BS2 (*n* = 288), BM1 (*n* = 192), and BM2 (*n* = 299). The phenotypic value of an individual is a morphometric descriptor of the posterior lobe, obtained by averaging both sides of the first principal component (PC1) of the Fourier coefficients of the

Table 2: Simulation results on the basis of model (1).

Model	Marker(s)	Mean	SE	BF ≥ 1	BF ≥ 3.2	BF ≥ 10	BF ≥ 100
Without Epistases	2–47.8	0.7453	0.1340	94 (5)	93 (5)	92 (5)	90 (3)
	3–141.5	1.1222	0.231	100 (0)	100 (0)	100 (0)	100 (0)
	FDR (additive)	--	--	0.0067	0	0	0
With Epistases	2–47.8	0.7610	0.1439	94 (6)	93 (6)	93 (6)	93 (6)
	3–141.5	1.1316	0.1402	100 (0)	100 (0)	100 (0)	100 (0)
	(1–24.7, 2–47.8)	1.5607	0.3921	92 (6)	91 (7)	91 (7)	90 (7)
	(2–47.8, 3–141.5)	1.5558	0.3054	97 (3)	96 (4)	96 (4)	96 (4)
	(2–133.8, 3–56.7)	-1.6204	0.3875	92 (8)	92 (8)	92 (8)	90 (10)
	FDR (additive)	--	--	0.0408	0.0333	0.0133	0.0067
	FDR (epistatic)	--	--	0.4872	0.3251	0.2283	0.1122

Out of 100 simulated data sets, the total numbers of data sets that correctly identify the true additive and interaction effects (in the brackets, their neighboring ones when the true ones are missed) are counted respectively when thresholding the Bayes factor (BF) at different levels. Also listed are the mean and standard error (SE) of the estimated effect sizes.

posterior lobe. The genotypes of males were determined at each marker locus, and genetic map positions were estimated from gametes produced by the F_1 females in this study. Further information about the data is referred to Liu et al. [33] and Zeng et al. [34].

Employing multiple interval mapping (MIM) [25,35] to the BS2 data, Zeng et al. [34] detected a total of 16 additive effects and no epistatic effect. Pooling all four data sets, Zeng et al. [34] detected three extra additive effects and six epistatic effects. These epistatic effects appeared to be relatively unimportant for PC1 in the interspecific backcross populations, which carried an observation difficult to interpret biologically. Of the 19 additive effects, 18 additive effect estimates have the same sign [34]. Zeng et al. [34] explained this interesting phenomena as an unusually strong directional selection, although Tanksley [36] suggested that transgressive segregation usually followed a mixture of plus and minus alleles in each species as demonstrated by most previous analyses of quantitative traits.

We focused our analysis on the BS2 and BM2 data with the standardized phenotypic values. Of the 19 putative QTL reported by Zeng et al. [34], only nine are at least 1 cM away from the 45 marker loci. Therefore, we analyzed both datasets with these 54 additive effects (nine putative QTL and 45 markers) and all possible pairwise interactions (that is, 1431 putative epistases). When controlling BF ≥ 1, the analysis of the BS2 data reported a total of 25 additive effects (see Table 3), including all nine putative QTL, but no epistatic effect. The analysis of the BM2 data instead reported a total of 20 additive effects (see Table 4), including three of the nine putative QTL, and 18 epistatic effects (see Table 5). On the basis of the simulation study, we may expect less than 0.67% FDR for those 17 and 16 additive effects reported with BF ≥ 100 in analyzing the BS2 and BM2 data respectively. Similarly, three epistatic effects reported in analyzing the BM2 data have BF ≥ 100, less than 12% of which may be false discoveries.

Table 3: Additive effects with BF ≥ 1 in analyzing the BS2 data.

Marker	Coefficient	S.D.	BF
1–3.6	-0.3797	0.0707	> 1000
1–23.4	-0.3462	0.0426	> 1000
2-0	-0.2284	0.0493	> 1000
2–17.08	-0.1906	0.1055	> 1000
2–27	-0.1262	0.1491	> 1000
2–28.53	-0.1618	0.1387	> 1000
2–69	-0.2969	0.1382	> 1000
2–113.92	-0.0682	0.0487	4.38
2–143	-0.0454	0.0592	1.72
2–145.85	-0.0322	0.0648	1.29
3-0	-0.1726	0.0880	> 1000
3–21.3	-0.3100	0.0569	> 1000
3–43.2	-0.1482	0.1052	> 1000
3–47	-0.1261	0.1571	> 1000
3–49.99	-0.2164	0.0992	> 1000
3–75	-0.4018	0.1072	> 1000
3–94	-0.2147	0.1360	> 1000
3–101.29	-0.0520	0.0904	2.03
3–117	-0.0941	0.0960	29.55
3–126.62	-0.0378	0.0780	1.20
3–134.6	-0.0724	0.1255	4.21
3–139	-0.2624	0.1604	> 1000
3–147.69	-0.0420	0.0833	1.29
3–160	-0.1847	0.1154	> 1000
3–171.22	-0.3295	0.0567	> 1000

The position of each significant additive effect is specified by an index of the corresponding chromosome and its location on this chromosome (cM). The estimated sizes of additive effects and the standard deviations of the Markov chains are also shown in the columns of coefficient and S.D., respectively.

Interestingly, the 25 additive effects detected from the BS2 data include all those detected by Zeng et al. [34] except the 2–135, 3–5 and 3–83 (we consider the markers within 1 cM to be same), but the 20 additive effects detected from the BM2 data only include nine of those detected by Zeng et al. [34]. On the other hand, nine additive effects (i.e., 2–28.53, 2–145.85, 3-0, 3–43.2, 3–49.99, 3–101.29, 3–126.62, 3–134.6, 3–147.69) from the BS2 data are not reported by Zeng et al. [34], and eleven additive effects from the BM2 data (i.e., 1-0, 2–6.98, 2–67.96, 2–145.85, 3–14.33, 3–28.74, 3–43.2, 3–49.99, 3–126.62, 3–147.69, 3–161.43) are not reported by Zeng et al. [34]. Note that

Table 4: Additive effects with BF ≥ 1 in analyzing the BM2 data.

Marker	Coefficient	S.D.	BF
1-0	-0.2181	0.1426	> 1000
1–3.6	-0.1438	0.1506	920.66
1–23.4	-0.1909	0.0654	> 1000
2–6.98	-0.2393	0.0809	> 1000
2–27	-0.3361	0.0855	> 1000
2–67.96	-0.0561	0.1093	1.29
2–69	-0.1473	0.1146	> 1000
2–113.92	-0.2496	0.0509	> 1000
2–145.85	-0.1145	0.0856	79.06
3–4.99	-0.1973	0.0954	> 1000
3–14.33	-0.2855	0.0928	> 1000
3–28.74	-0.1754	0.0934	> 1000
3–43.2	-0.0586	0.1077	1.80
3–47	-0.2213	0.1648	> 1000
3–49.99	-0.1749	0.1355	> 1000
3–83.15	-0.5978	0.0781	> 1000
3–126.62	-0.1970	0.1066	> 1000
3–147.69	-0.0698	0.0826	3.05
3–161.43	-0.1982	0.0950	> 1000
3–171.22	-0.2385	0.1028	> 1000

The position of each significant additive effect is specified by an index of the corresponding chromosome and its location on this chromosome (cM). The estimated sizes of additive effects and the standard deviations of the Markov chains are also shown in the columns of coefficient and S.D., respectively.

Table 5: Epistatic effects with BF ≥ 1 in analyzing the BM2 data.

Markers	Coefficient	S.D.	BF
(1–3.6, 3–14.34)	0.0601	0.1077	1.36
(1–3.6, 3–101.29)	0.0383	0.0994	1.05
(1–14.2, 2–28.53)	0.0156	0.0792	1.07
(1–14.2, 3–134.6)	0.2231	0.1552	116.30
(1–14.2, 3–139)	0.1689	0.1453	13.40
(2–17.08, 3–157.73)	0.3304	0.0806	> 1000
(2–28.53, 3–101.29)	0.2688	0.1063	> 1000
(2–34.72, 3–76.3)	0.1307	0.0960	5.34
(2–113.92, 3–83.15)	0.0779	0.0911	1.89
(2–138.82, 3–147.69)	0.1678	0.0943	12.23
(2–143, 3–101.29)	0.0463	0.0972	1.25
(2–145.85, 3–28.74)	0.0896	0.0909	2.61
(2–145.85, 3–43.2)	0.0330	0.0980	1.07
(2–145.85, 3–101.29)	0.0419	0.0921	1.29
(3–21.3, 3–76.3)	0.0797	0.0856	1.97
(3–28.74, 3–53.54)	0.0487	0.0999	1.19
(3–43.2, 3–123.32)	0.0400	0.1014	1.04
(3–53.54, 3–123.33)	0.1925	0.1226	43.36

The QTL positions of each significant epistatic effect are specified by the indices of the corresponding chromosomes and the locations on the chromosomes (cM). The estimated sizes of the epistatic effects and the standard deviations of the Markov chains are also shown in the columns of coefficient and S.D., respectively.

almost each additive effect uniquely detected by Zeng et al. [34] has a neighboring one (within 10 cM) in our lists except 2–135 and 3–94 for the BM2 dataset, and almost each additive effect unique in our lists has a neighboring one (within 10 cM) detected by Zeng et al. [34]. Per the discussion on the precision of QTL location by Bogdan and Doerge [37] and Bogdan et al. [24], these effects of close neighbors may be due to identical QTL. Our analysis reported R2 = 0.934 and R2 = 0.902 for the BS2 and BM2 data respectively.

Conclusion

This article extends the Bayesian framework in Zhang et al. [28] to identify both additive and epistatic effects of QTL based on model (1). The advantage of this approach mainly lies in the flexible priors for the regression coefficients by accounting for some characteristics of "large p small n" data, the predictability of a model constructed with size n data, and the two step strategy for dimension reduction. A Gibbs sampler is developed to draw Markov chain samples from the posterior distributions, which can be considered as a stochastic search for an optimal model. Unlike information criteria based model selections which require calculation of the effective sample size for incomplete data, missing values can be naturally imputed within the Gibbs sampling scheme. The corresponding algorithm has been implemented in Matlab and is available as QTLBayes http://www.stat.purdue.edu/~zhangdb/QTLBayes/.

Bayesian variable selections can be viewed as penalized likelihood approaches, which have been studied recently [38,39]. With "large p small n" data, it is not clear how to set up the penalty properly such that it will neither overpenalize nor underpenalize the likelihood. An overpenalized likelihood will lose some significant variables of particular interest, while an underpenalized likelihood will introduce false positives. The predictability of size n data sheds light on the choice of this penalty. Since a size n data set will allow us to understand the variation of the trait explained by only $p_n = O(\sqrt{n})$ QTL with accuracy $O(n^{-1/2})$, selecting too many variables into the model will ruin this practice of QTL mapping. As shown by Bogdan and Doerge [37], severely biased estimates can be resulted from large genome and/or marker number in QTL mapping. We propose a Bayesian framework to resolve the bias problem. We have illustrated our approach by application to the BS2 and BM2 data [33,34], both of which have 45 markers observed across three chromosomes. The disadvantage of this approach is the heavy computation involved as the computation-intensive Markov chain Monte Carlo algorithm is utilized. For example, the analysis of a dataset with more than 200 markers from 1,000 subjects take almost 24 hours using one Intel® Xeon™ CPU at 2.80 GHz.

Coding binary markers with -0.5 and 0.5 has been commonly utilized in QTL mapping as it does not introduce correlation between additive effects and interactive effects, and such uncorrelation benefits the identification of additive effects. On the other hand, coding binary markers with 0 and 1 introduces correlation and thus is not preferred for QTL mapping with epistases [40,41]. Although developed for QTL mapping, this approach is completely general and can be applied to other settings with "large p small n" data, such as associating genomic features to clinical outcomes or phenotypes of biological interest. Unlike QTL mapping data with known missing structure from the linkage information, genomic data with imaging and microarray may require more information to impute missing values because of the unknown missing mechanism. Even though the missing values are usually imputed with a nearest-neighbor approach [42], Gibbs samplers allow natural multiple imputation under the assumption of missing at random (MAR, see Little and Rubin, [43]).

Methods

Predictability and Sample Size

Suppose, for a sample of size n, we select up to p_n (assuming $p_n < n$) significant variables into the following regression model,

$$\mathbf{Y}_n = \mathbf{X}_n \quad + \quad _n, \qquad _n \sim N(0, \quad ^2 I_n),$$

where $\boldsymbol{Y}_n$ is an n-dimensional column vector; $\boldsymbol{X}_n$ is an $n \times p_n$ design matrix such that $\boldsymbol{X}_n^T \boldsymbol{X}_n = n \times I_{p_n}$. The best linear unbiased estimator (BLUE) of β is

$$\hat{\ }_n = \quad + \frac{1}{n} \boldsymbol{X}_n^T \quad _n.$$

Let $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_{p_n})$ include p_n predictors for $\tilde{y}$ such that $\max_{1 \le j \le p_n} | \tilde{x}_j | = O(1)$. Since $\text{trace}\{Var(\quad _n)\} = \frac{p_n}{n} \quad ^2, \tilde{x}$ can be consistently estimated by $\tilde{\boldsymbol{x}} \hat{\ }_n$. When using $\tilde{\boldsymbol{x}} \hat{\ }_n$ to predict $\tilde{y}$, the mean squared prediction error is

$$E[(\tilde{y} - \tilde{\boldsymbol{x}} \quad _n)^2] = \quad ^2 + \frac{p_n}{n} \frac{\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T}{p_n} \quad ^2.$$

If $p_n = o(n)$, the mean squared prediction error asymptotically achieves the minimum variance, and thus the prediction is asymptotically efficient.

This illustration implies that, with a sample of size n and $p_n = O(\sqrt{n})$ predictors, the mean squared prediction error can reach the minimum prediction error at rate $O(n^{-1/2})$. Suppose that all p_n significant variables could be perfectly selected out of p candidates, we still need $p_n = o(n)$ in order to have a chance to correctly understand the variation of the dependent variable explained by the selected predictors. Therefore, we always assume that there are at most $p_n = O(\sqrt{n})$ significant variables among a total of p candidates in the case of $p \gg n$. Indeed, the study of consistency in a triangular array setting for regression problems was conducted by Huber [44-46]. In examining the underlying theory of 'model-selection' and 'variable-selection' procedures that choose p_n explanatory variables from an initial set of variables, Greenshtein and Ritov [46] proved that one may expect consistency for the choice of p_n with an order between $o(\sqrt{n/\log(n)})$ and $o(n/\log(n))$. Our choice of $p_n = O(\sqrt{n})$ satisfies the Greenshtein and Ritov [46] conditions for consistency.

Bayesian Variable Selection

Here we propose a two-step Bayesian variable selection approach to map QTL with epistases through model (1). With the following Bayesian framework, we first select $c\sqrt{n}$ out of p_β additive effects and $c\sqrt{n}$ out of p_γ epistatic effects (e.g., we use $c = 2$), respectively, using a restrictive prior for each coefficient. We then apply the same Bayesian framework to stochastically select the filtered variables, using a non-restrictive prior for each coefficient. Gibbs sampling algorithms are developed to stochastically search low-dimensional subspaces, as implied by the predictability of a size n data set.

Prior Specification

For a two-state marker system, both additive effects β_j, $j = 1, \quad , p_\beta$, and epistatic effects γ_j, $j = 1, \quad , p_\gamma$, are the primary focus of QTL mapping. As is often the case $p = (p_\beta + p_\gamma) \gg n$, many of these coefficients are zero, either because the variation of the trait can be explained by only a few QTL or because the limited sample size precludes selecting too many variables (otherwise the constructed model is not reliable as shown in the previous section). It is also possible that the number and/or scale of the positive coefficients may be different from those of the negative ones. To account for these properties, a three-component mixture prior is specified for each coefficient β_j or γ_j. More specifically,

$$\begin{cases} \beta_j \overset{iid}{\sim} (1 - w_{\beta+} - w_{\beta-})\delta(\cdot) + w_{\beta+} N_+(\mu_{\beta+}, \sigma^2_{\beta+}) + w_{\beta-} N_-(\mu_{\beta-}, \sigma^2_{\beta-}), \\ \gamma_j \overset{iid}{\sim} (1 - w_{\gamma+} - w_{\gamma-})\delta(\cdot) + w_{\gamma+} N_+(\mu_{\gamma+}, \sigma^2_{\gamma+}) + w_{\gamma-} N_-(\mu_{\gamma-}, \sigma^2_{\gamma-}), \end{cases} \tag{2}$$

where $\delta(\cdot)$ is a Dirac function with mass one at zero; $N_+(\mu, \sigma^2)$ and $N_-(\mu, \sigma^2)$ positively and negatively truncate the normal distribution, i.e., $N(\mu, \sigma^2)$, respectively. Therefore, $w_{\beta+}$ (or $w_{\beta-}$) is the probability for any single marker, and $w_{\gamma+}$ (or $w_{\gamma-}$) is the probability for any pair of markers in $\mathcal{I}\{X\}$, to have positive (or negative) interactive effect on the trait.

The hyperparameters, $\sigma^2_{\beta+}, \sigma^2_{\beta-}, \sigma^2_{\gamma+}$ and $\sigma^2_{\gamma-}$, are assumed to have priors as inverse gamma distributions, that is, $IG(\theta_{\beta+}, \phi_{\beta+})$, $IG(\theta_{\beta-}, \phi_{\beta-})$, $IG(\theta_{\gamma+}, \phi_{\gamma+})$, and $IG(\theta_{\gamma-}, \phi_{\gamma-}$), respectively (e.g., setting $\theta_{\beta+} = \theta_{\beta-} = \theta_{\gamma+} = \theta_{\gamma-} = 0.1$ and $\phi_{\beta+} = \phi_{\beta-} = \phi_{\gamma+} = \phi_{\gamma-} = 10$). As a result, the prior on β (and γ) is essentially a mixture of a point mass at zero and some truncated t-distributions, which shrinks the smaller effects towards zero and allows sufficient flexibility for non-zero effects. Furthermore, t-type prior distributions yield Bayes rules with desirable decision-theoretic frequentist properties [47]. The hyperparameters, $\mu_{\beta+}$, $\mu_{\gamma+}$, $\mu_{\beta-}$ and $\mu_{\gamma-}$, are assumed to have diffuse priors, and the prior distribution for σ_{ε}^{2} is proportional to $1/\sigma_{\varepsilon}^{2}$.

As suggested by the predictability of a size n data set, we expect to select at most $p_n = O(\sqrt{n}\,)$ out of the p variables for the final model. Therefore, we specify the priors for $(w_{\beta+}, w_{\beta-})$ and $(w_{\gamma+}, w_{\gamma-})$ as

$$w_{\beta+} + w_{\beta-} \sim U(0, c\sqrt{n}\,/\,p_{\beta}), \quad w_{\gamma+} + w_{\gamma-} \sim U(0, c\sqrt{n}\,/\,p_{\gamma}), \tag{3}$$

that is, expecting at most $c\sqrt{n}$ significant additive effects and epistatic effects, respectively. Gaffney [48] and Yi et al. [17], among others, employed similar ideas to rescale the priors based on the number of possible effects. Apparently, when $n << (p_{\beta} + p_{\gamma})$, either $c\sqrt{n}\,/p_{\beta}$ or $c\sqrt{n}\,/p_{\gamma}$ is very small, which implies a restrictive prior on each corresponding coefficient. Therefore, we usually select $c\sqrt{n}$ additive effects and $c\sqrt{n}$ epistatic effects during the first run of Bayesian analysis. We then apply the same Bayesian analysis to these pre-selected variables. The second run of Bayesian analysis has both $w_{\beta+} + w_{\beta-}$ and $w_{\gamma+} + w_{\gamma-}$, *a priori*, uniformly distributed on [0, 1].

Likelihood

Let $\boldsymbol{Y}_n$ be the column vector including the trait values of all strains under investigation, let $\boldsymbol{X}_i$ be the vector of all marker values of the i-th strain and $X_n = (X_1^T, \cdots, X_n^T)^T$, and let $\boldsymbol{Z}_i$ be the vector of all epistatic candidate values of the i-th strain. Denote the marginal distribution of A as $[A]$, and the conditional distribution of A given B as $[A|B]$. With data $(\mathbf{Y}_n, \mathbf{X}_n)$ and the prior specification in Section 3.1, we have the likelihood function, that is, the joint distribution function of the data $(\boldsymbol{Y}_n, \boldsymbol{X}_n)$, the parameters (μ, β, γ), σ_{ε}^{2}, and all hyperparameters

$$(w_{\ +}, w_{\ -}, w_{\ +}, w_{\ -}, \ {}_{+}, \ {}_{+}, \ {}^{2}{}_{+}, \ {}^{2}{}_{+}, \ {}_{+}, \ {}_{-}, \ {}^{2}{}_{-}, \ {}^{2}{}_{-}),$$

$$\begin{aligned}
L &\propto [\mathbf{Y}_n \mid \mathbf{X}_n, \ , \ , \ , \ {}^{2}] \times [\ \ ] \times [\ {}_{+}] \times [\ {}^{2}{}_{+}] \times [\ {}_{-}] \times [\ {}^{2}{}_{-}] \times [w_{\ +}, w_{\ -}] \\
&\quad \times [\ \mid w_{\ +}, w_{\ -}, \ {}_{+}, \ {}^{2}{}_{+}, \ {}_{-}, \ {}^{2}{}_{-}] \times [\ {}_{+}] \times [\ {}^{2}{}_{+}] \times [\ {}_{-}] \times [\ {}^{2}{}_{-}] \\
&\quad \times [w_{\ +}, w_{\ -}] \times [\ \mid w_{\ +}, w_{\ -}, \ {}_{+}, \ {}^{2}{}_{+}, \ {}_{-}, \ {}^{2}{}_{-}] \times [\ {}^{2}] \times [\mathbf{X}_n] \\
&\propto \ {}^{-n-2} \exp\left\{ -\sum_{i=1}^{n} \frac{(Y_i - \ - X_i \ - Z_i \ )^2}{2 \ {}^{2}} \right\} \times \exp\left(-\frac{\ {}^{2}{}_{+}}{\ {}_{+}} - \frac{\ {}^{2}{}_{-}}{\ {}_{-}} \right) \\
&\quad \times (\ {}^{2}{}_{+})^{- \ {}_{+}-1} \times (\ {}^{2}{}_{-})^{- \ {}_{-}-1} \times \exp\left(-\frac{\ {}^{2}{}_{+}}{\ {}_{+}} - \frac{\ {}^{2}{}_{-}}{\ {}_{-}} \right) \times (\ {}^{2}{}_{+})^{- \ {}_{+}-1} \\
&\quad \times (\ {}^{2}{}_{-})^{- \ {}_{-}-1} \times [\ \mid w_{\ +}, w_{\ -}, \ {}_{+}, \ {}^{2}{}_{+}, \ {}_{-}, \ {}^{2}{}_{-}] \times I\left[w_{\ +} + w_{\ -} \le c\frac{\sqrt{n}}{p} \right] \\
&\quad \times [\ \mid w_{\ +}, w_{\ -}, \ {}_{+}, \ {}^{2}{}_{+}, \ {}_{-}, \ {}^{2}{}_{-}] \times I\left[w_{\ +} + w_{\ -} \le c\frac{\sqrt{n}}{p} \right] \times [\mathbf{X}_n].
\end{aligned} \tag{4}$$

The distribution of $\boldsymbol{X}_n$ can be specified based on the available linkage map information [2]. The conditional distribution of $[\boldsymbol{\beta} \mid w_{\beta+}, w_{\beta-}, \mu_{\beta+}, \sigma_{\beta+}^{2}, \mu_{\beta-}, \sigma_{\beta-}^{2}]$ is a product of the prior distribution for each β_j. Similarly, the conditional distribution of $[\boldsymbol{\gamma} \mid w_{\gamma+}, w_{\gamma-}, \mu_{\gamma+}, \sigma_{\gamma+}^{2}, \mu_{\gamma-}, \sigma_{\gamma-}^{2}]$ is a product of the prior distribution for each γ_j. The priors of the hyperparameters, $\theta_{\beta+}$, $\theta_{\gamma+}$, $\phi_{\beta+}$, $\phi_{\gamma+}$, $\theta_{\beta-}$, $\theta_{\gamma-}$, $\phi_{\beta-}$ and $\phi_{\gamma-}$, are specified to be as noninformative as possible.

Gibbs Sampling

Since the specified priors are conditionally conjugate, Bayesian variable selection can be implemented with a Gibbs sampling algorithm. We initialize the algorithm by imputing missing genotypic values based on the observed genotypes and linkage information. The initial value of μ is set as the mean of the observed trait values. Then, with individuals having fully observed trait values, each component of β and γ is initially estimated using recursive univariate regression. Other parameters, $w_{\beta+}, w_{\beta-}, \mu_{\beta+}, \sigma_{\beta+}^{2}, \mu_{\beta-}$ and $\sigma_{\beta-}^{2}$, are simply initialized based on the initial value of β, and similarly, the initial values for $w_{\gamma+}, w_{\gamma-}, \mu_{\gamma+}, \sigma_{\gamma+}^{2}, \mu_{\gamma-}$, and $\sigma_{\gamma-}^{2}$ can be specified using the information from γ. For example, we can initialize $\sigma_{\beta+}^{2} = \sigma_{\beta-}^{2}$ with an estimate from the initial value β, and then use $\max\{\#\{j : \beta_j > 2\sigma_{\beta+}\}, 1\}/p_{\beta}$ to initialize $w_{\beta+}$.

Let $\boldsymbol{X}_{i,-j}$ be $\boldsymbol{X}_i$ excluding the j-th component, and define β_{-j} and γ_{-j} similarly. Based on the likelihood function in (4), the Gibbs sampler can be developed by recursively drawing the missing genotypic values, the missing trait values, and the model parameters from their full conditional posterior distributions as follows.

Sample missing values: Sample each missing genotypic value X_{ij} from its full conditional posterior distribution,

$$[X_{ij} \mid Y_i, \boldsymbol{X}_{i,-j}, \;\;,\;\;,\;\;,\;\; {}^2] \propto [Y_i \mid \boldsymbol{X}_{i,-j}, X_{ij}, \;\;,\;\;,\;\;,\;\; {}^2] \times [X_{ij} \mid X_{i,j-1}, X_{i,j+1}],$$

and then sample each missing trait value Y_i from its full conditional posterior distribution $[Y_i|X_i, \mu, \beta, \gamma, \sigma_\varepsilon^2]$.

Sample μ: Sample μ from its full conditional posterior distribution,

$$\mid \boldsymbol{Y}_n, \boldsymbol{X}_n, \;\;,\;\;,\;\; {}^2 \sim N\left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i \;\; - Z_i \;\;), \frac{2}{n} \right).$$

Sample β **and** γ: Sample each β_j and γ_j from their full conditional posterior distributions,

$$\begin{aligned}
&[\;\; {}_j \mid \mathbf{Y}_n, \mathbf{X}_n, \;\;,\;\; {}_{-j}, \;\;, w \;\; {}_+, w \;\; {}_-, \;\; {}^2, \;\; {}^2_+, \;\; {}^2_-] \\
&\quad \sim (1 - \tilde{w} \;\; {}_{j+} - \tilde{w} \;\; {}_{j-}) \;\; (\;\; {}_j) + \tilde{w} \;\; {}_{j+} N_+(\tilde{} \;\; {}_{j+}, \tilde{} \, {}^2 {}_{j+}) + \tilde{w} \;\; {}_{j-} N_-(\tilde{} \;\; {}_{j-}, \tilde{} \, {}^2 {}_{j-}), \\
&[\;\; {}_j \mid \mathbf{Y}_n, \mathbf{X}_n, \;\;,\;\;,\;\; {}_{-j}, w \;\; {}_+, w \;\; {}_-, \;\; {}^2, \;\; {}^2_+, \;\; {}^2_-] \\
&\quad \sim (1 - \tilde{w} \;\; {}_{j+} - \tilde{w} \;\; {}_{j-}) \;\; (\;\; {}_j) + \tilde{w} \;\; {}_{j+} N_+(\tilde{} \;\; {}_{j+}, \tilde{} \, {}^2 {}_{j+}) + \tilde{w} \;\; {}_{j-} N_-(\tilde{} \;\; {}_{j-}, \tilde{} \, {}^2 {}_{j-}),
\end{aligned}$$

where $\tilde{w}_{\beta j+}, \tilde{w}_{\beta j-}, \tilde{\mu}_{\beta j+}, \tilde{\sigma}^2_{\beta j+}, \tilde{\mu}_{\beta j-}$, and $\tilde{\sigma}^2_{\beta j-}$ are specified in the APPENDIX. In addition, $\tilde{w}_{\gamma j+}, \tilde{w}_{\gamma j-}, \tilde{\mu}_{\gamma j+}, \tilde{\sigma}^2_{\gamma j+}, \tilde{\mu}_{\gamma j-}$, and $\tilde{\sigma}^2_{\gamma j-}$ can be obtained similarly.

Sample $w_{\beta+}$, $w_{\gamma+}$, $w_{\beta-}$, **and** $w_{\gamma-}$: These parameters can be sampled from the conditional posterior distributions,

$$\begin{aligned}
&(w_{\beta+}, w_{\beta-}, 1 - w_{\beta+} - w_{\beta-}) \mid \beta \\
&\quad \sim Dirichlet(\tilde{p}_{\beta+} + 1, \tilde{p}_{\beta-} + 1, p_\beta - \tilde{p}_{\beta+} - \tilde{p}_{\beta-} + 1), w_{\beta+} + w_{\beta-} \le \frac{c\sqrt{n}}{p_\beta}, \\
&(w_{\gamma+}, w_{\gamma-}, 1 - w_{\gamma+} - w_{\gamma-}) \mid \gamma \\
&\quad \sim Dirichlet(\tilde{p}_{\gamma+} + 1, \tilde{p}_{\gamma-} + 1, p_\gamma - \tilde{p}_{\gamma+} - \tilde{p}_{\gamma-} + 1), w_{\gamma+} + w_{\gamma-} \le \frac{c\sqrt{n}}{p_\gamma},
\end{aligned}$$

where $\tilde{p}_{\beta+} = \#\{\beta_j > 0 : 1 \le j \le p_\beta\}$ and $\tilde{p}_{\beta-} = \#\{\beta_j < 0 : 1 \le j \le p_\beta\}$; $\tilde{p}_{\gamma+} = \#\{\gamma_j > 0 : 1 \le j \le p_\gamma\}$ and $\tilde{p}_{\gamma-} = \#\{\gamma_j < 0 : 1 \le j \le p_\gamma\}$.

Sample $\sigma^2_\varepsilon, \sigma^2_{\beta+}, \sigma^2_{\gamma+}, \sigma^2_{\beta-}$, **and** $\sigma^2_{\gamma-}$: With conditionally conjugate priors, the posterior for all variance parameters are still inverse gamma distributions. Specifically,

$$\begin{aligned}
&{}^2 \mid \boldsymbol{Y}_n, \boldsymbol{X}_n, \;\;,\;\;,\;\; \sim IG\left(\frac{n}{2}, \frac{2}{\sum_{i=1}^n (Y_i - \;\; - X_i \;\; - Z_i \;\;)^2} \right), \\
&{}^2_+ \mid \;\; \sim IG\left(\;\; {}_+ + \frac{\tilde{p} \;\; +}{2}, \frac{2}{\frac{2}{+} + \sum_{j=1}^p {}^2_j I[\;\; j>0]} \right), \\
&{}^2_+ \mid \;\; \sim IG\left(\;\; {}_+ + \frac{\tilde{p} \;\; +}{2}, \frac{2}{\frac{2}{+} + \sum_{j=1}^p {}^2_j I[\;\; j>0]} \right), \\
&{}^2_- \mid \;\; \sim IG\left(\;\; {}_- + \frac{\tilde{p} \;\; -}{2}, \frac{2}{\frac{2}{-} + \sum_{j=1}^p {}^2_j I[\;\; j<0]} \right), \\
&{}^2_- \mid \;\; \sim IG\left(\;\; {}_- + \frac{\tilde{p} \;\; -}{2}, \frac{2}{\frac{2}{-} + \sum_{j=1}^p {}^2_j I[\;\; j<0]} \right).
\end{aligned}$$

Bayesian Inference

For each variable in model (1), one pair of parameters is used to select the corresponding variable. They are, for the j-th additive effect, the posterior probabilities $w_{\beta j+} = P(\beta_j > 0|\mathbf{Y}_n, \mathbf{X}_n)$ and $w_{\beta j-} = P(\beta_j < 0|\mathbf{Y}_n, \mathbf{X}_n)$. With the full conditional posterior distribution of β_j and all the notations in the APPENDIX, we have

$$w \;\; {}_{j+} = E[\tilde{w} \;\; {}_{j+} \mid \mathbf{Y}_n, \mathbf{X}_n], \quad w \;\; {}_{j-} = E[\tilde{w} \;\; {}_{j-} \mid \mathbf{Y}_n, \mathbf{X}_n].$$

Therefore, the two parameters $w_{\beta j+}$ and $w_{\beta j-}$ can be estimated with the Markov chains of $\tilde{w}_{\beta j+}$ and $\tilde{w}_{\beta j-}$ drawn from the above Gibbs sampler. If and only if both $w_{\beta j+}$ and $w_{\beta j-}$ are less than 0.5, the median of the posterior distribution of β_j is zero. Similarly, the posterior probabilities $w_{\gamma j+} = P(\gamma_j > 0|\boldsymbol{Y}_n, \boldsymbol{X}_n)$ and $w_{\gamma j-} = P(\gamma_j < 0|\boldsymbol{Y}_n, \boldsymbol{X}_n)$ can be estimated with the Markov chains of $\tilde{w}_{\gamma j+}$ and $\tilde{w}_{\gamma j-}$ drawn from the above Gibbs sampler.

We propose to select variables twice under the above Bayesian framework for model (1). At the first step, we use a restrictive prior for each coefficient to ensure an identifiable Bayesian model and enforce to stochastically search for an optimal low-dimensional parameter subspace. We then rank the j-th additive effect based on $\max\{w_{\beta j+}, w_{\beta j-}\}$, and rank the j-th epistatic effect based on $\max\{w_{\gamma j+}, w_{\gamma j-}\}$. The top $c\sqrt{n}$ out of p_β additive effects, and the top $c\sqrt{n}$ out of p_γ epistatic effects are selected, respectively. At the second step, we select variables out of those selected $c\sqrt{n}$ additive effects and $c\sqrt{n}$ epistatic effects, under the above Bayesian framework for model (1). Obviously, we have a non-restrictive prior for each coefficient at the second step, and therefore avoid possible over-penalization due to restrictive priors.

Following Jeffreys [49,50], we test the hypothesis $H_0 : \beta_j = 0$ vs. $H_1: \beta_j \neq 0$ on the basis of the Bayes factor, which was defined as

$$B_{10}(\beta_j) = \frac{P(Data|\beta j{\neq}0)}{P(Data|\beta j{=}0)} = \frac{P(\beta j{\neq}0|Data)}{P(\beta j{=}0|Data)} \times \frac{\pi(\beta j{=}0)}{\pi(\beta j{\neq}0)} = \frac{w\beta j{+}{+}w\beta j{-}}{1{-}w\beta j{+}{-}w\beta j{-}},$$

where $\pi(\beta_j = 0)$ and $\pi(\beta_j \neq 0)$ are the *a priori* probabilities, and the last equality follows the fact that $\pi(\beta_j = 0) = \pi(\beta_j \neq 0)$ at the second step of our Bayesian Classification. As suggested by Jeffreys [50], a $B_{10}(\beta_j)$ with value between 1 and $\sqrt{10} \approx 3.2$ provides "not worth more than a bare mention" evidence against H_0; a $B_{10}(\beta_j)$ with value from $\sqrt{10}$ to 10 provides "substantial" evidence against H_0; a $B_{10}(\beta_j)$ with value from 10 to 100 provides "strong" evidence against H_0; and a $B_{10}(\beta_j)$ with value larger than 100 provides "decisive" evidence against H_0. Similarly, we can test the hypothesis H_0: $\gamma_j = 0$ vs. H_1: $j \neq 0$ using the following Bayes factor

$$B_{10}(\gamma_j) = \frac{w\gamma j{+}{+}w\gamma j{-}}{1{-}w\gamma j{+}{-}w\gamma j{-}}. \qquad (5)$$

Authors' contributions

MZ and DZ both conceived the study and developed the method. DZ wrote the MATLAB® code and did the simulation study. MZ analyzed the real data. MTW participated in conceptual development, writing, reviewing and editing the manuscript. All authors read and approved the final manuscript.

Appendix

Fully Conditional Posterior Distribution of β_j

For each $j = 1, \quad , p_\beta$, the fully conditional posterior distribution of β_j is

$$\begin{aligned} &_j \mid \mathbf{Y}_n, \mathbf{X}_n, \ \ , \ \ _{-j}, \ \ , w\ _+, w\ _-, \ \ ^2, \ \ ^2_+, \ \ ^2_- \\ &\sim (1-\tilde{w}\ \ _{j+}-\tilde{w}\ \ _{j-})\ \ _{(0)}+\tilde{w}\ \ _{j+}N_+(\tilde{}\ \ _{j+},\tilde{}\ \ ^2_{j+})+\tilde{w}\ \ _{j-}N_-(\tilde{}\ \ _{j-},\tilde{}\ \ ^2_{j-}), \end{aligned}$$

where the updated parameter values are

$$\begin{aligned}
\tilde{\mu}_{\beta j+} &= \sigma^2_{\beta+}\sum_{i=1}^{n} X_{ij}(Y_i-\mu-X_{i,-j}\beta_{-j}-Z_i\gamma)\Bigg/\left(\sigma^2_\varepsilon+\sigma^2_{\beta+}\sum_{i=1}^{n}X^2_{ij}\right),\\
\tilde{\sigma}^2_{\beta j+} &= \sigma^2_{\beta+}\sigma^2_\varepsilon\Bigg/\left(\sigma^2_\varepsilon+\sigma^2_{\beta+}\sum_{i=1}^{n}X^2_{ij}\right),\\
\tilde{\mu}_{\beta j-} &= \sigma^2_{\beta-}\sum_{i=1}^{n} X_{ij}(Y_i-\mu-X_{i,-j}\beta_{-j}-Z_i\gamma)\Bigg/\left(\sigma^2_\varepsilon+\sigma^2_{\beta-}\sum_{i=1}^{n}X^2_{ij}\right),\\
\tilde{\sigma}^2_{\beta j-} &= \sigma^2_{\beta-}\sigma^2_\varepsilon\Bigg/\left(\sigma^2_\varepsilon+\sigma^2_{\beta-}\sum_{i=1}^{n}X^2_{ij}\right),\\
\tilde{w}_{\beta j+} &= \frac{\tilde{\sigma}\beta j+}{\sigma\beta+}\Phi\left(\frac{\tilde{\mu}\beta j+}{\tilde{\sigma}\beta j+}\right)\Bigg/\Bigg[\frac{1-w_{\beta+}-w_{\beta-}}{2w_{\beta+}}\exp\left(-\frac{\tilde{\mu}^2_{\beta j+}}{2\tilde{\sigma}^2_{\beta j+}}\right)+\frac{\tilde{\sigma}_{\beta j+}}{\sigma_{\beta+}}\\
&\quad\times\Phi\left(\frac{\tilde{\mu}_{\beta j+}}{\tilde{\sigma}_{\beta j+}}\right)+\frac{w\beta_-\tilde{\sigma}\beta j-}{w\beta_+\sigma\beta-}\Phi\left(-\frac{\tilde{\mu}\beta j-}{\tilde{\sigma}\beta j-}\right)\exp\left(\frac{\tilde{\mu}^2_{\beta j-}}{2\tilde{\sigma}^2_{\beta j-}}-\frac{\tilde{\mu}^2_{\beta j+}}{2\tilde{\sigma}^2_{\beta j+}}\right)\Bigg],\\
\tilde{w}_{\beta j-} &= \frac{\tilde{\sigma}\beta j-}{\sigma\beta-}\Phi\left(-\frac{\tilde{\mu}\beta j-}{\tilde{\sigma}\beta j-}\right)\Bigg/\Bigg[\frac{1-w_{\beta+}-w_{\beta-}}{2w_{\beta-}}\exp\left(-\frac{\tilde{\mu}^2_{\beta j-}}{2\tilde{\sigma}^2_{\beta j-}}\right)+\frac{\tilde{\sigma}_{\beta j-}}{\sigma_{\beta j-}}\\
&\quad\times\Phi\left(-\frac{\tilde{\mu}_{\beta j-}}{\tilde{\sigma}_{\beta j-}}\right)+\frac{w\beta_+\tilde{\sigma}\beta j+}{w\beta_-\sigma\beta+}\Phi\left(\frac{\tilde{\mu}\beta j+}{\tilde{\sigma}\beta j+}\right)\exp\left(\frac{\tilde{\mu}^2_{j+}}{2\tilde{\sigma}^2_{\beta j+}}-\frac{\tilde{\mu}^2_{\beta j-}}{2\tilde{\sigma}^2_{\beta j-}}\right)\Bigg].
\end{aligned}$$

Acknowledgements

We thank Rebecca Doerge and three anonymous referees for suggestions and comments that significantly improved this article. This research was supported by NSF grant 0612031 and NIH-NIGMS 1R01GM083606-01 to MTW, Purdue Research Foundation grants to MZ and DZ, and Purdue Alumni Association Faculty Incentive Grant to MZ.

References

1. Lander ES, Botstein D: **Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps.** *Genetics* 1989, **121:**185-199.
2. Doerge RW, Zeng ZB, Weir BS: **Statistical issues in the search for genes affecting quantitative traits in experimental populations.** *Statistical Science* 1997, **12:**195-219.
3. Broman KW, Speed TP: **A model selection approach for the identification of quantitative trait loci in experimental crosses.** *Journal of the Royal Statistical Society Series B* 2002, **64:**641-656.
4. Wang H, Zhang YM, Li X, Masinde GL, Mohan S, Baylink DJ, Xu S: **Bayesian shrinkage estimation of quantitative trait loci parameters.** *Genetics* 2005, **170:**465-480.
5. Carlborg d, Haley CS: **Epistasis: too often neglected in complex trait studies?** *Natuer Review Genetics* 2004, **5:**618-625.
6. Moore JH: **The ubiquitous nature of epistasis in determining susceptibility to common human disease.** *Human Heredity* 2003, **56:**73-82.
7. Williams SM, Addy JH, Phillips JAI, Dai M, Kpodonu J, Afful J, Jackson H, Joseph K, Eason F, Murray MM, Epperson P, Aduonum A, Wong LJ, Jose PA, Felder RA: **Combinations of variation in multiple genes are associated with hypertension.** *Hypertension* 2000, **36:**2-6.

8. Leamy LJ, Routman EJ, Cheverud JM: **An epistatic genetic basis for fluctuating asymmetry of mandible size in mice.** *Evolution* 2002, **56:**642-653.
9. Wagner A: **Robustness against mutations in genetic networks of yeast.** *Nature Genetics* 2000, **24:**355-361.
10. Sanjuán R, Cuevas JM, Moya A, Elena SF: **Epistasis and the adaptability of an RNA virus.** *Genetics* 2005, **170:**1001-1008.
11. Eshed Y, Zamir D: **Less-than-additive epistatic interactions of quantitative trait loci in tomato.** *Genetics* 1996, **143:**1807-1817.
12. Xu S, Jia Z: **Genomewide analysis of epistatic effects for quantitative traits in Barley.** *Genetics* 2007, **175:**1955-1963.
13. Bateson W: *Mendel's Principles of Heredity* Cambridge: Cambridge University Press; 1909.
14. Kao CH, Zeng ZB: **Modeling epistasis of quantitative trait loci using Cockerham's model.** *Genetics* 2002, **160:**1243-1261.
15. Yi N, Xu S: **Mapping quantitative trait loci with epistatic effects.** *Genetical Research* 2002, **79:**185-198.
16. Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D: **Bayesian model selection for genome-wide epistatic quantitative trait loci analysis.** *Genetics* 2005, **170:**s1333-1344.
17. Yi N, Banerjee S, Pomp D, Yandell BS: **Bayesian mapping of genomewide interacting quantitative trait loci for ordinal traits.** *Genetics* 2007, **176:**1855-1864.
18. Yi N, Shriner D, Banerjee S, Mehta T, Pomp D, Yandell BS: **An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects.** *Genetics* 2007, **176:**1865-1877.
19. Yandell BS, Mehta T, Banerjee S, Shriner D, Venkataraman R, Moon JY, Neely WW, Wu H, von Smith R, Yi N: **R/qtlbim: QTL with Bayesian interval mapping in experimental crosses.** *Bioinformatics* 2007, **23:**641-643.
20. Bogdan M, Ghosh JK, Doerge RW: **Modifying the Schwartz Bayesian information criterion to locate multiple interacting quantitative trait loci.** *Genetics* 2004, **167:**989-999.
21. Cui YH, Wu R: **Mapping genome-genome epistasis: a high-dimensional model.** *Bioinformatics* 2005, **21:**2447-2455.
22. Żak M, Baierl A, Bogdan M, Futschik A: **Locating multiple interacting quantitative trait loci using rank-based model selection.** *Genetics* 2007, **176:**1845-1854.
23. Shi W, Lee KE, Wahba G: **Detecing disease-causing genes by LASSO-Patternsearch algorithm.** *BMC Proceedings* 2007, **1(Suppl 1):**S60.
24. Bogdan M, Frommlet F, Biecek P, Cheng R, Ghosh JK, Doerge RW: **Extending the modified Bayesian information criterion (mBIC) to dense markers and multiple interval mapping.** *Biometrics* in press.
25. Kao CH, Zeng ZB, Teasdale RD: **Multiple interval mapping for quantitative trait loci.** *Genetics* 1999, **152:**1203-1216.
26. Zeng ZB, Kao CH, Basten CJ: **Estimating the genetic architecture of quantitative traits.** *Genetical Research* 1999, **74:**279-289.
27. Ball RD: **Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion.** *Genetics* 2001, **159:**1351-1364.
28. Zhang M, Montooth KL, Wells MT, Clark AG, Zhang D: **Mapping multiple quantitative trait loci by Bayesian classification.** *Genetics* 2005, **169:**2305-2318.
29. Mitchell TJ, Beauchamp JJ: **Bayesian variable selection in linear regression (with discussion).** *Journal of the American Statistical Association* 1988, **83:**1023-1036.
30. George EI, McCulloch RE: **Variable selection via Gibbs sampling.** *Journal of the American Statistical Association* 1993, **88:**881-889.
31. Ishwaran H, Rao JS: **Spike and slab variable selection: frequentist and Bayesian strategies.** *The Annals of Statistics* 2005, **33:**730-773.
32. Zhang M, Zhang D, Wells MT: **Generalized Shrinkage Estimators Adpative to Sparsity and Asymmetry of High Dimensional Parameter Spaces.** *Technical Reports, Department of Statistics, Purdue University* 2008:08-01.
33. Liu J, Mercer JM, Stam LF, Gibson G, Zeng ZB, Laurie CC: **Genetic analysis of a morphological shape difference in the male genitalia of *Drosophila simulans* and *D. mauritiana*.** *Genetics* 1996, **142:**1129-1145.
34. Zeng ZB, Liu J, Stam LF, Kao CH, Mercer JM, Laurie CC: **Genetic architecture of a morphological shape difference between two drosophila species.** *Genetics* 2000, **154:**299-310.
35. Kao CH, Zeng ZB: **General formula for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm.** *Biometrics* 1997, **53:**653-665.
36. Tanksley SD: **Mapping polygenes.** *Annual Review Genetics* 1993, **27:**205-233.
37. Bogdan M, Doerge RW: **Biased estimators of quantitative trait locus heritability and location in interval mapping.** *Heredity* 2005, **95:**476-484.
38. Tibshirani RJ: **Regression shrinkage and selection via the lasso.** *Journal of the Royal Statistical Society Series B* 1996, **58:**267-288.
39. Fan J, Peng H: **Nonconcave penalized likelihood with a diverging number of parameters.** *The Annals of Statistics* 2004, **32:**928-961.
40. Álvarez-Castro JM, Carlborg O: **A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis.** *Genetics* 2007, **176:**1151-1167.
41. Zeng ZB, Wang T, Zou W: **Modeling quantitative trait loci and interpretation of models.** *Genetics* 2005, **169:**1711-1725.
42. Hastie H, Tibshirani R, Sherlock G, Eisen M, Brown P, Botstein D: **Imputing missing data for gene expression arrays.** In *PhD thesis* Stanford University, Statistics Department; 1999.
43. Little RJA, Rubin DB: *Statistical Analysis with Missing Data* New York: John Wiley; 2002.
44. Huber P: **Robust regression: asymptotics, conjectures, and Monte Carlo.** *The Annals of Statistics* 1973, **1:**799-821.
45. Portnoy S: **Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large, I. Consistency.** *Annals of Statistics* 1984, **12:**1298-1309.
46. Greenshtein E, Ritov Y: **Persistence in high-dimensional linear predictor selection and the virtue of overparametrization.** *Bernoulli* 2004, **10:**971-988.
47. Fourdrinier D, Strawderman WE, Wells MT: **On the construction of Bayes minimax estimators.** *The Annals of Statistics* 1998, **26:**660-671.
48. Gaffney PJ: **An efficient reversible jump Markov chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses.** In *PhD thesis* Department of Statistics, University of Wisconsin, Madison, WI; 2001.
49. Jeffreys H: **Some tests of significance, treated by the theory of probability.** *Proceedings of the Cambridge Philosophy Society* 1935, **31:**201-222.
50. Jeffreys H: *Theory of Probability* Oxford: Clarendon Press; 1961.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral