Heredity (2005) 95, 96–104

www.nature.com/hdy

A penalized maximum likelihood method for estimating epistatic effects of QTL

Y-M Zhang and S Xu
Department of Botany and Plant Sciences, University of California, Riverside, CA, USA

Although epistasis is an important phenomenon in the genetics and evolution of complex traits, epistatic effects are hard to estimate. The main problem is due to the over-parameterized epistatic genetic models. An epistatic genetic model should include potential pair-wise interaction effects of all loci. However, the model is saturated quickly as the number of loci increases. Therefore, a variable selection technique is usually considered to exclude those interactions with negligible effects. With such techniques, we may run a high risk of missing some important interaction effects by not fully exploring the extremely large parameter space of models. We develop a penalized maximum likelihood method. The method developed here adopts a penalty that depends on the values of the parameters. The penalized likelihood method allows spurious QTL effects to be shrunk towards zero, while QTL with large effects are estimated with virtually no shrinkage. A simulation study shows that the new method can handle a model with a number of effects 15 times larger than the sample size. Simulation studies also show that results of the penalized likelihood method are comparable to the Bayesian shrinkage analysis, but the computational speed of the penalized method is orders of magnitude faster.
Heredity (2005) **95,** 96–104. doi:10.1038/sj.hdy.6800702
Published online 25 May 2005

Keywords: epistatic effect; marker analysis; penalized maximum likelihood; quantitative trait loci

Introduction

Epistasis plays a fundamental role in the genetic control and evolution of complex traits. However, epistatic effects are hard to detect (Cheverud and Routman, 1995) because an epistatic genetic model potentially contains a large number of model effects. One choice is to use a model selection technique to eliminate the spurious effects so that the number of effects is reduced to a manageable level. Several model selection methods have been developed recently. A maximum likelihood (ML) based stepwise regression method has been developed by Kao *et al* (1999) for model selection. A one-dimensional search method, also based on ML, was developed by Jannink and Jansen (2001) and Boer *et al* (2002). More recently, a Bayesian method based on the stochastic search variable selection (SSVS) has been applied to QTL mapping (Oh *et al*, 2003; Yi *et al*, 2003). These various selection methods are still open to discussion because the criteria of variable inclusion and exclusion are somewhat subjective (Balding *et al*, 2002; Broman and Speed, 2002; Sillanpaa and Corander, 2002; Kadane and Lazar, 2004).

Ridge regression represents another class of methods for handling oversaturated models. Whittaker *et al* (2000) adopted the original ridge regression idea of Hoerl and Kennard (1970) to shrink marker effects proportionally in the context of marker-assisted selection. Gianola *et al* (2003) claimed that the mixed model analysis of genetic effects is the same as the ridge regression analysis, except that the ridge factors vary across model effects and can be estimated from the data. Xu (2003) found that ridge regression works only if the number of model effects is in the same order as the number of observations. Xu (2003) modified the ridge regression by allowing the ridge factor to vary across different model effects. The difference between Xu (2003) and Gianola *et al* (2003) is that Xu's (2003) method can estimate the QTL variance using only a single regression coefficient whereas the method of Gianola *et al* (2003) estimates the QTL variance using a batch of regression coefficients. The modified ridge regression methods turn out to be equivalent to the Bayesian analysis with different model effects taking different prior distributions. The model-selection-free method of Xu (2003) has successfully detected multiple QTL with main effects. Extension of the method to an epistatic effects model has not been explored, although it is straightforward. One concern about the extension is the intensive computing time because the Markov Chain Monte Carlo (MCMC) algorithm requires repeatedly sampling a huge number of model effects. If we incorporate the idea of estimating the parameters of the prior distribution from the data, a kind of empirical Bayesian analysis, the method becomes a penalized ML method (Boer *et al*, 2002).

In this study, we develop such a penalized likelihood method, with the penalty being a function of the parameters. The method can handle an oversaturated model, with the number of model effects many times larger than the number of observations. The method allows spurious effects to be shrunk towards zero, while QTL with large effects is subject to virtually no shrinkage. Therefore, model selection is no longer

Correspondence: S Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA.
E-mail: xu@genetics.ucr.edu
Received 24 November 2004; accepted 14 April 2005; published online 25 May 2005

needed, and all problems associated with model selection, for example, lack of full exploration of the parameter space of models and intensive computing time, are not of concern.

We use the backcross (BC) design as an example to demonstrate the new method. We also fix the QTL positions at markers so that estimation of QTL positions is irrelevant and we can concentrate on evaluating the performance of the method based on the estimated effects. The method is essentially the multiple marker analysis of Xu (2003) incorporating epistatic effects from the ML perspective.

Theory and methods

Epistatic effect model

Let y_i $(i=1,\ldots,n)$ be the phenotypic value of the ith individual in a BC mapping population of size n. The epistatic effect model is

$$y_i = a_0 + \sum_{j=1}^{p} z_{ij}a_j + \sum_{r<s}^{p} z_{ir}z_{is}a_{rs} + \varepsilon_i \quad (1)$$

where a_0 is the population mean, Z_{ij} is a dummy variable indicating the genotype of the jth marker for individual i, a_j is the effect of marker j $(j=1, \ldots, p)$, p is the total number of markers on the entire genome, a_{rs} is the epistatic effect between markers r and s $(r=1, \ldots, p-1; s=r+1, \ldots, p)$, and ε_j is the residual error with a $N(0, \sigma^2)$ distribution. In a BC population, an individual can take one of two genotypes, heterozygote and homozygote. The dummy variable is defined as $Z_{ij}=1$ for heterozygote and $Z_{ij}=-1$ for homozygote.

Methods of estimating the main effects and interaction effects are the same. For the sake of clarity of notation, we redefine the design matrix and the regression coefficients as follows. Let $b_0=a_0$, $b_j=a_j$ $(j=1, \ldots, p)$, and

$$b_{j+p} = a_{rs} \ (r = 1, \ldots, p-1; s = r+1, \ldots, p; \ j = 1, \ldots, q-p),$$

where $q=p(p+1)/2$. Similarly, we define $x_{ij}=z_{ij}$ $(j=1, \ldots, p)$ and

$$x_{i(j+p)} = z_{ir}z_{is} \ (r = 1, \ldots, p-1; s = r+1, \ldots, p; \ j = 1, \ldots, q-p)$$

Model (1) is now rewritten as

$$y_i = b_0 + \sum_{j=1}^{q} x_{ij}b_j + \varepsilon_i \quad (2)$$

We now have a simple model that includes both the main and the interaction effects.

Penalized likelihood function

The penalized likelihood is similar to the posterior distribution of the parameters, with the prior distribution of the parameters serving as the penalty. The difference between the penalized likelihood method and the Bayesian method is that the parameters in the prior distributions are estimated simultaneously along with the parameters of interest. Let $\theta=\{b_0, b_1, \ldots, b_q, \sigma^2\}$ be the vector of parameters of interest. The log likelihood function is

$$L(\theta) = \sum_{i=1}^{n} \log\phi(y_i; \beta_i, \sigma^2) \quad (3)$$

where $\beta_i = b_0 + \sum_{j=1}^{q} x_{ij}b_j$ and $\phi(y_i; \beta_i, \sigma^2)$ is the normal density with mean β_i and variance σ^2.

We now introduce a factor to penalize the large number of model effects. This penalty should be a function of the parameters. The prior density of the parameters in the Bayesian framework is an ideal choice for the penalty factor. Let us introduce the following prior density for each of the parameters. Parameters b_0 and σ^2 are always included in the model and thus their inclusion should not be penalized. We introduce a normal prior for each of the regression coefficients,

$$p(b_j) = \phi(b_j; \mu_j, \sigma_j^2) \quad \text{for } j = 1, \ldots, q \quad (4)$$

In classical Bayesian regression analysis, μ_j and σ_j^2 are hyperparameters. In the oversaturated model, however, the choice of the parameters in the prior distribution is very important. Therefore, we will estimate these hyperparameters from the data. Our experience shows that a prior distribution should also be assigned to μ_j and the normal prior given below is necessary

$$p(\mu_j) = \phi(\mu_j; 0, \sigma_j^2/\eta) \quad \text{for } j = 1, ., q \quad (5)$$

where $\eta>0$ serves as a prior sample size for accessing μ_j. Let $\xi = \{\mu_1, \ldots, \mu_q, \sigma_1^2, \ldots, \sigma_q^2\}$ be the hyperparameters that are subject to estimation. The logarithm of the prior density is used as the penalty and it has the following form;

$$P(\theta, \xi) = \log p(\theta, \xi) = \sum_{j=1}^{q} \left[\log \phi(b_j; \mu_j, \sigma_j^2) + \log \phi(\mu_j; 0, \sigma_j^2/\eta)\right] \quad (6)$$

The penalized log likelihood is defined as

$$\psi(\theta, \xi) = L(\theta) + P(\theta, \xi) \quad (7)$$

Parameter estimation

The parameters are estimated by maximizing $\psi(\theta, \xi)$ with respect to θ and ξ simultaneously. The solutions are called the penalized maximum likelihood estimates (PMLE) of the parameters. The PMLE of ξ are not of direct interest, but provided estimates of nuisance parameters. We now describe an iterative algorithm to solve the PMLE of the parameters.

The PMLE of the intercept is found by setting

$$\frac{\partial}{\partial b_0}\psi(\theta, \xi) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{j=1}^{q} x_{ij}b_j\right)(-2) = 0 \quad (8)$$

and solving for b_0, which is

$$b_0 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{q} x_{ij}b_j\right) \quad (9)$$

Setting

$$\frac{\partial}{\partial b_j}\psi(\theta,\xi) = -\frac{\sum_{i=1}^{n}(-2x_{ij})(y_i - b_0 - \sum_{k=1}^{q} x_{ik}b_k)}{2\sigma^2} - \frac{2(b_j-\mu_j)}{2\sigma_j^2} = 0 \quad (10)$$

and solving for b_j, we get

$$b_j = \left[\sum_{i=1}^{n} x_{ij}^2 + \sigma^2/\sigma_j^2\right]^{-1} \times \left[\sum_{i=1}^{n} x_{ij}(y_i - b_0 - \sum_{k\neq j}^{q} x_{ik}b_k) + \mu_j\sigma^2/\sigma_j^2\right] \quad (j=1,\ldots,q) \quad (11)$$

The residual variance is estimated by setting

$$\frac{\partial}{\partial \sigma^2}\psi(\theta,\xi) = -\frac{n}{2\sigma^2} - \frac{\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{j=1}^{q} x_{ij}b_j\right)^2(-1)}{2(\sigma^2)^2} = 0 \quad (12)$$

and solving for σ^2, which is

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{j=1}^{q} x_{ij}b_j\right)^2 \quad (13)$$

Although the nuisance parameters ξ are not of direct interest, they must be estimated from the data. The PMLE of ξ are all obtained by setting $(\partial/\partial\xi_j)\psi(\theta,\xi)=0$, where ξ_j is the jth component of the nuisance parameters. The PMLE of the nuisance parameters are

$$\mu_j = b_j/(\eta+1) \quad \text{for } j=1,\ldots,q \quad (14)$$

and

$$\sigma_j^2 = \frac{1}{2}\left[(b_j-\mu_j)^2 + \eta\mu_j^2\right] \quad \text{for } j=1,\ldots,q \quad (15)$$

Summary of iterations

1. Set $\eta>0$ and provide initial values for θ and ξ;
2. Update b_0 using Eq. (9);
3. Update b_j ($j=1,\ldots,q$) using Eq. (11);
4. Update σ^2 using Eq. (13);
5. Update μ_j ($j=1,\ldots,q$) using Eq. (14);
6. Update σ_j^2 ($j=1,\ldots,q$) using Eq. (15);
7. Repeat step 2 to step 6 until a certain criterion of convergence is satisfied.

The initial values $\theta^{(0)} = \{b_0^{(0)}, b_1^{(0)}, \ldots, b_q^{(0)}, \sigma^{2(0)}\} = \{\bar{y}, 0, \ldots, 0, s_y^2\}$ are suggested for θ, where $\bar{y}$ and s_y^2 are sample mean and variance of the phenotypic values. The initial values $\xi^{(0)} = \{\mu_1^{(0)}, \ldots, \mu_q^{(0)}, \sigma_1^{2(0)}, \ldots, \sigma_q^{2(0)}\} = \{0,\ldots,0,0.5,\ldots,0.5\}$ are suggested for ξ.

Statistical test

The proposed penalized likelihood method is intended to include all markers in a single model. Theoretically, no statistical tests are necessary because all marker effects are estimated and none of them are missing. However, investigators may only want to report markers with relatively large effects. Fortunately, the penalized likelihood method often provides extremely small estimates for markers not closely linked to QTL. The signals (estimated effects) of these null markers are almost negligible compared to those of the markers that are linked to QTL (Xu, 2003). It is not hard to pick up 'significant' markers just by visually scanning on the genome. The majority of the markers should have no effects and their estimated effects should be close to zero. When we plot the estimated marker effects against the genome location, the effects of the null markers should serve as background. Our simulation experiments show that the background noise is indeed very close to zero, making the signals of markers linked to QTL very clear. How large an estimated marker effect is large enough to warrant a spot in the final list of markers associated with the phenotype? An objective statistical test may be helpful. Unfortunately, a usual likelihood ratio test cannot be performed with the penalized likelihood method because of the overparameterization. Therefore, we propose the following two-stage selection process to screen the markers. All markers with $|\hat{b}_j|/\hat{\sigma} > 10^{-6}$ are deemed to have passed the first round of selection. If $|\hat{b}_j|/\hat{\sigma} \leq 10^{-6}$, even if it was significant statistically, it would not be interesting biologically. In a BC population, a QTL of this size would explain less than $10^{-10}\%$ of the phenotypic variance. This selection criterion is already quite stringent because very few spurious markers will survive this selection owing to the enforced stringent penalty (shrinkage). In the second stage of the selection, we are more careful on choosing the criterion of selection. We now modify our epistatic model so that only effects that have passed the first round of selection are included in the model because the dimensionality of such a model is quite small compared to the original oversaturated model. Owing to the dimension of the modified model being small, we can use a regular (unpenalized) ML method to reanalyze the data and perform a likelihood ratio test for each QTL. The estimated QTL effects from the penalized likelihood using the oversaturated model are almost identical to the effects estimated from the likelihood analysis using the modified model that includes only the QTL surviving the first round of selection (see results of simulation).

Let s be the total number of QTL effects that have passed the first round of selection and $\theta = \{b_0, b_1, \ldots, b_s, \sigma^2\}$ be the parameters that are subject to the ML analysis for significance test. To test the null hypothesis that $H_0: b_j = 0$, that is, the jth surviving QTL (passed the first round of selection) is not true, we use the following likelihood ratio test statistic,

$$LR_j = -2[L(\theta_{-j}) - L(\theta)] \quad (16)$$

(Lander and Botstein, 1989), where $\theta_{-j} = \{b_0, b_1, \ldots b_{j-1}, b_{j+1}, \ldots, b_s, \sigma^2\}$ is the vector of parameters that excludes b_j. As pointed out by Kao *et al* (1999), the choice of critical value for claiming a significant QTL becomes complicated for multiple QTL tests. For simplicity, we use the usual $LOD_j \geq 3$ as the criterion, where $LOD_j = LR_j/$

$(2\ln 10) \approx LR_j/4.61$. Application of the permutation test (Churchill and Doerge, 1994) is discussed later.

Simulation studies

We conducted three simulation experiments to evaluate the performance of the method. In the first experiment, we simulated a single genome of 200 cM long with 21 evenly spaced markers, with equal marker distance of 10 cM. We put four main QTL effects and four pairwise interaction effects, all of which overlap with markers. The positions and effects of the simulated QTL are given in Table 1 along with the simulated residual variance. We simulated a BC population with sample size of $n = 200$ for one case and $n = 500$ for the other case. Each case was replicated 100 times to evaluate the accuracy, the precision, and the statistical power for each estimated QTL effect. The total number of QTL effects included in the model is $21(21+1)/2 = 231$. We used the two-stage screening process to select markers and further tested the selected marker using the likelihood ratio tests. For each simulated QTL, we counted the samples in which the LOD statistic had passed 3. The ratio of the number of such samples to the total number of replicates (100 in this case) represented the empirical power for this QTL. When the sample size was small, we noticed that a marker with a simulated QTL effect was not always significant, but a significant LOD occurred in a nearby marker. This reflected the uncertainty of the estimated QTL position. In this case, the simulated QTL was also counted as significant (detected). This is why there is an average estimate of QTL position shown in Table 1. The table shows that the larger sample size does have a higher power than the smaller sample size. QTL with small effects tend to be associated with lower powers. The method can detect the smallest QTL (explaining 2.5% of the phenotypic variance) with 63% power even when $n = 200$.

The prior value $\eta = 5$ was used in the simulation experiment. We also tried $\eta = 10$ and 30, which had virtually no effect on the result. The convergence criterion was chosen as $||\theta^{(t)} - \theta^{(t-1)}|| \leq 10^{-4}$, where $\theta^{(t)}$ is the vector of parameter values at the iteration. The convergence was usually very fast, taking only 40–70 iterations to the convergence criterion.

In the second simulation experiment, we doubled the chromosome size (400 cM long) but simulated the same number of QTL with the same positions and effects as those given in the first experiment. The total number of markers was 41, with the total number of marker effects (including all pairwise interactions) being $41(41+1) = 861$. The sample size was now 300 in the second simulation experiment. We also simulated the chromosome length 200 cM (the same as that in the first experiment) with $n = 300$ for comparison. Again, all QTL resided at marker positions. The objective of this experiment was to evaluate the performance of the new method on a more saturated model. Our prediction was that the method would still have a satisfactory performance, even though the number of model effects was almost three times as large as the sample size. The results are given in Table 2 and consistent with the above prediction.

Finally, we simulated a single large chromosome 1800 cM long, covered by 121 evenly spaced markers with a 15 cM per marker interval. The total number of QTL effects included in the model was $121(121+1)/2 = 7381$. We increased the sample size to 600. The number of model effects was about 12 times as large as the number of observations. The simulated parameters (positions and effects of QTL) are given in Table 3 for the main effects and Table 4 for the epistatic effects. We simulated nine main-effect QTL and 13 interacting QTL effects. The sizes of QTL (measured by the proportions of phenotypic variance explained by QTL) varied from 0.5 to 20%. Residual variance was set at 10. The data were analyzed with two methods: a Bayesian method and the penalized likelihood method. The Bayesian method was implemented via the MCMC algorithm and it is a simple extension of Xu (2003) by incorporating epistatic effects into the oversaturated model. The initial values and prior distribution of the parameters for the Bayesian analysis were the same as those given by Xu (2003). The length of the Markov chain was of 20 000 iterations, excluding 4000

Table 1 Effect of sample sizes on the results of epistatic QTL analysis (100 replicates)

	QTL	*Main effect*			*Interaction*				σ^2
		Power	*Position*	*Effect*	*Power*	*Position 1*	*Position 2*	*Effect*	
1	True value	—	0.00	1.0000	—	0.00	80.00	1.0000	10.0000
	$n = 200$	63	3.02(5.86)	1.0507(0.2852)	51	4.86(9.61)	82.38(25.81)	0.9975(0.2506)	10.7410(1.4764)
	$n = 500$	92	1.22(4.41)	0.9104(0.1859)	93	0.72(3.05)	78.44(8.77)	0.9427(0.1879)	9.9434(0.7291)
2	True value	—	60.00	1.4142	—	40.00	140.00	1.4142	
	$n = 200$	85	58.71(5.30)	1.4993(0.2908)	87	42.17(13.50)	137.74(14.33)	1.4178(0.3755)	
	$n = 500$	100	60.23(1.79)	1.4247(0.1804)	100	39.24(4.78)	139.50(5.14)	1.3787(0.1856)	
3	True value	—	140.00	2.0000	—	80.00	200.00	2.0000	
	$n = 200$	100	139.75(2.65)	1.9811(0.2906)	98	79.53(3.65)	199.37(2.71)	1.9659(0.3050)	
	$n = 500$	100	140.01(0.09)	1.9865(0.1634)	100	79.98(0.27)	200.00(0.00)	2.0079(0.1684)	
4	True value	—	180.00	2.8284	—	100.00	120.00	2.8284	
	$n = 200$	100	179.89(1.86)	2.8074(0.3012)	98	99.30(3.61)	120.10(1.67)	2.6130(0.4750)	
	$n = 500$	100	179.99(0.08)	2.8205(0.1856)	100	99.98(0.16)	120.00(0.00)	2.8166(0.2048)	

$n = 200$ and $n = 500$ represent estimates from sample sizes 200 and 500, respectively.
The standard deviations of estimates are calculated from only the significant samples and are given in parentheses.

Table 2 Effect of the number of variables on the results of epistatic QTL analysis (100 replicates)

	QTL	*Main effect*			*Interaction*				σ^2
		Power	*Position*	*Effect*	*Power*	*Position 1*	*Position 2*	*Effect*	
1	True value	—	0.00	1.0000	—	0.00	80.00	1.0000	10.0000
	Estimate 1	87	4.48(10.76)	0.9628(0.2269)	80	2.38(7.33)	80.20(20.87)	0.9488(0.2148)	10.4211(1.2610)
	Estimate 2	90	2.88(8.89)	0.9789(0.2149)	88	2.80(9.68)	76.80(14.12)	0.9727(0.2456)	10.1275(1.0965)
2	True value	—	60.00	1.4142	—	40.00	140.00	1.4142	
	Estimate 1	97	58.94(3.95)	1.4021(0.2295)	97	40.30(6.63)	136.28(9.18)	1.3206(0.2705)	
	Estimate 2	98	59.06(4.12)	1.3890(0.2458)	99	39.53(3.50)	138.30(6.85)	1.3617(0.2751)	
3	True value	—	140.00	2.0000	—	80.00	200.00	2.0000	
	Estimate 1	100	140.07(1.20)	2.0031(0.2516)	99	79.96(1.34)	199.67(1.53)	1.9716(0.2229)	
	Estimate 2	100	140.03(0.26)	1.9890(0.2377)	100	79.50(2.63)	199.60(1.91)	1.9556(0.2276)	
4	True value	—	180.00	2.8284	—	100.00	120.00	2.8284	
	Estimate 1	100	180.00(0.23)	2.7873(0.2390)	100	99.98(0.17)	119.99(0.12)	2.7541(0.2713)	
	Estimate 2	100	180.08(0.76)	2.8132(0.1702)	100	99.69(2.23)	119.98(0.18)	2.7695(0.3069)	

Estimates 1 and 2 represent estimates from the models of 231 and 861 QTL effects, respectively. Position 1 and 2: positions of QTL_1 and QTL_2, respectively.
The standard deviations of estimates are calculated from only the significant samples and are given in parentheses.

Table 3 Simulated and estimated QTL positions and effects from a single dataset of a large genome

Marker	*True parameters*			*Bayesian analysis*		*Penalized likelihood*	
	Position	*Effect*	*Proportion*	*Position*	*Effect*	*Position*	*Effect*
1	0	4.47	0.200	0	4.4593 (0.1507)	0	4.5760 (0.1477)
21	300	3.16	0.100	300	3.1493 (0.1462)	300	3.2344 (0.1475)
31	450	2.24	0.050	450	2.2770 (0.1510)	450	2.3337 (0.1471)
51	750	1.58	0.025	750	1.3133 (0.1644)	750	1.4163 (0.1459)
71	1050	1.58	0.025	1050	1.5325 (0.1470)	1050	1.5996 (0.1463)
91	1350	1.10	0.012	1350	0.9083 (0.1556)	1350	0.9644 (0.1438)
101	1500	1.10	0.012	1500	1.2145 (0.1557)	1500	1.2391 (0.1454)
111	1650	0.77	0.006	1650	0.5948 (0.2990)	1635	0.5456 (0.1363)
121	1800	0.77	0.006	1800	0.4200 (0.3479)	1800	0.5801 (0.1375)

Table 4 Simulated and estimated positions and effects of interacting QTL from a single data set of a large genome

Marker pair	*True parameter*			*Bayesian analysis*		*Penalized likelihood*	
	Positions 1 & 2	*Effect*	*Proportion*	*Positions 1 & 2*	*Effect*	*Positions 1 & 2*	*Effect*
1–11	0 & 150	1.00	0.010	0 & 150	0.7374 (0.1618)	0 & 150	0.8894 (0.1432)
2–119	15 & 1770	3.87	0.150	15 & 1770	3.8497 (0.1648)	15 & 1770	3.7274 (0.1476)
10–91	135 & 1350	1.30	0.017	135 & 1350	1.2942 (0.1759)	135 & 1350	1.3931 (0.1459)
15–75	210 & 1110	1.73	0.030	210 & 1110	1.5068 (0.2593)	210 & 1110	1.6738 (0.1465)
20–46	285 & 675	1.00	0.010	285 & 675	0.9463 (0.1739)	285 & 660	0.7630 (0.1416)
21–22	300 & 315	1.00	0.010	300 & 315	0.9371 (0.2765)	Missing	
26–91	375 & 1350	1.00	0.010	375 & 1350	1.2616 (0.1712)	360 & 1350	0.8232 (0.1424)
41–61	600 & 900	0.71	0.005	600 & 915	0.2545 (0.3102)	Missing	
56–91	825 & 1350	3.16	0.100	825 & 1350	3.1118 (0.1603)	825 & 1350	3.0166 (0.1474)
65–85	960 & 1260	2.24	0.050	960 & 1260	2.4575 (0.1543)	900 & 1275	0.7848 (0.1492)
						960 & 1245	1.8898 (0.1468)
86–96	1275 & 1425	0.89	0.008	1275 & 1425	0.9811 (0.2198)	1275 & 1425	1.0248 (0.1443)
101–105	1500 & 1560	1.00	0.010	1500 & 1560	0.9895 (0.1645)	Missing	
111–121	1650 & 1800	2.24	0.050	1650 & 1800	1.8096 (0.9778)	1650 & 1800	2.3618 (0.1472)

iterations for the burn-in period. The chain was trimmed by keeping one observation in every 20 iterations so that the posterior sample size for the post-MCMC analysis was 1000. The penalized likelihood method took about 58 iterations to converge and the total computing time with our (SAS, 1999) program run at PC Dell Optiplex GX 400 was about 14 h. The Bayesian analysis, however, took about 3 weeks in the same computer. The estimated main QTL effects are plotted in Figure 1 and the interaction effects are plotted (3D plot) in Figure 2. For the penalized likelihood method, the number of estimated effects (including both the main and interaction

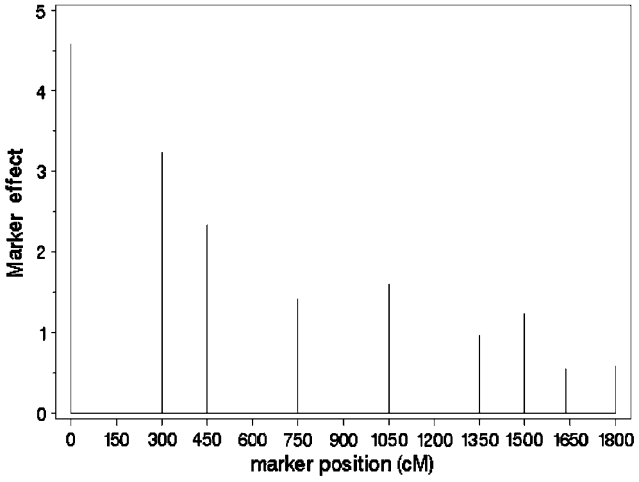

Figure 1 Plot of the estimated main QTL effects against the genome location using the penalized likelihood method.

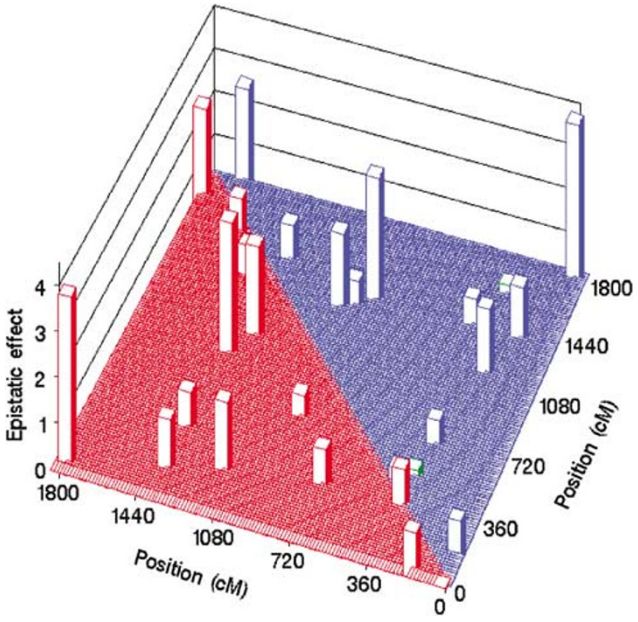

Figure 2 Plot (3-D) of the interaction effects against the genome location. The left-hand side of the figure shows the true effects (red) and the right-hand side of the figure shows the estimated effects (blue or green). The blue and green colored prisms represent positive and negative estimates of the interaction effects, respectively.

effects) that had passed the first round of selection was 23 (nine main effects and 14 interaction effects). In the second round of selection, all the nine main effects and 11 of the interaction effects surpassed the critical value of the test statistic (Table 5). One interaction effect was partitioned into two significant effects (Table 5) and three interaction effects were not detected (Table 4). The penalized likelihood analysis generated results that are quite consistent with the Bayesian analysis, except that the former missed a few QTL effects. This observation was expected because we gain the fast speed with the cost of failing to detect a few very small QTL effects (all explaining $<1\%$ of the phenotypic variance). Among the three undetected epistatic QTL effects, two are the interactions between two closely linked markers, that is, marker pairs 20–21 and 101–105, and one is an effect explaining only 0.5% of the phenotypic variance.

If we compare results of Tables 3 and 4 (penalized likelihood estimates, last columns) with the results of Table 5 (the reduced likelihood estimates, columns 2 and 5) methods), we can see that the penalized estimates and the reduced likelihood estimates are almost identical (differ only by $<10^{-6}$). Therefore, the two-stage analysis did not change the original estimates of the genetic effects other than facilitate a way to test the significances of the estimated genetic effects.

Discussion

We used a BC design as an example to demonstrate the penalized likelihood analysis. The method can be directly applied to double haploids (DH) and recombinant inbred lines (RIL). Extension to more complicated designs is possible. For example, to analyze data for an F_2 design, we need to partition the genetic effect of a single locus into an additive effect (a) and a dominance effect (d). Correspondingly, the epistatic effect can be partitioned into additive-by-additive (aa), additive-by-dominance (ad), dominance-by-additive (da), and dominance-by-dominance (dd). The dimensionality of the model will be doubled for the main effect QTL (a and d) and quadrupled for the epistatic effect QTL (aa, ad, da, and dd), but the method remains the same. The model appears to be

$$y_i = a_0 + \sum_{j=1}^{p}(z_{ij}a_j + w_{ij}d_j) + \sum_{r<s}^{p}[z_{ir}z_{is}(aa)_{rs} + z_{ir}w_{is}(ad)_{rs} + w_{ir}z_{is}(da)_{rs} + w_{ir}w_{is}(dd)_{rs}] + \varepsilon_i \quad (17)$$

where $Z_{ij}=\{1,\ 0,\ -1\}$ and $W_{ij}=\{0,\ 1,\ 0\}$ for the three genotypes {QQ, Qq, qq}. One can adopt the $b_0+\sum x_{ij}b_j$ notation, as given in equation (2), for the above model and thereafter use the same method to estimate all the model effects. For a four-way cross design, the genetic effect of a single locus can be partitioned into three terms, allelic effect from the male parent (a^m), allelic effect from the female parent (a^f), and the dominance effect (δ). Correspondingly, the epistatic effect can be partitioned into $3\times3=9$ terms. The incidence variables (coefficients of the main QTL effects) given by Xu *et al* (2003) for a four-way cross design may be adopted here. The incidence variables for the epistatis effects simply take the products of the corresponding incidence variables of the main effects. No additional theory and methods are involved other than that the dimension of the model should be expanded. Extension of the method to pedigree data is not obvious and deserves further investigation.

Kao and Zeng (2002) proposed the use of Cockerham's (1954) model to define the epistatic and other model effects. Cockerham's model in its original form only applies to two loci. For interactions involving multiple loci, substantial additional work may be required to construct a Cockerham's model. The key of the Cockerham's model is the orthogonality of the model effects. Cockerham (1954) defined model effects by orthogonal linear contrasts of the original genotypic effects. The

Table 5 Likelihood ratio test for the significance of the estimated QTL effects

QTL effect	*Estimate*	*LOD*	*Parameter*	*Estimate*	*LOD*
b_1	4.5760	121.98	$b_{15\times75}$	1.6738	24.85
b_{21}	3.2344	73.36	$b_{19\times106}$	−0.4299	2.88^{NS}
b_{31}	2.3337	43.71	$b_{2\times45}$	0.7630	6.53
b_{51}	1.4163	18.46	$b_{22\times32}$	−0.3499	2.33^{NS}
b_{71}	1.5996	23.25	$b_{25\times91}$	0.8232	7.26
b_{91}	0.9644	9.48	$b_{26\times29}$	0.4798	2.77^{NS}
B_{11}	1.2391	13.92	$b_{56\times91}$	3.0166	66.43
b_{111}	0.5456	3.82	$b_{61\times86}$	0.7848	6.53
B_{11}	0.5801	4.01	$b_{65\times84}$	1.8898	29.91
$b_{1\times11}$	0.8894	8.50	$b_{86\times96}$	1.0248	10.92
$b_{2\times119}$	3.7274	91.19	$b_{111\times121}$	2.3618	45.24
$b_{10\times91}$	1.3931	17.76			

^{NS}These tests are not significant based on the LOD 3 criterion.

linear contrasts may be called the statistical parameters, while the genotypic effects may be called the genetic parameters. The transformation from genetic parameters into statistical parameters has several desirable properties, including additivity of the variance components and ease of parameter estimation. However, the additive effects (defined as the linear contrasts) also contain dominance and epistatic effects defined in the original genotypic scale if the population is in linkage disequilibrium. This, unfortunately, has caused some problem in interpreting the biological meanings of the statistical parameters. Nonetheless, we might use the term Cockerham's model if orthogonal linear contrasts are estimated as the model effects. In that case, Cockerham's model in the context of multiple QTL should be described using the following orthogonal transformations. Let us rewrite model (2) as

$$y_i = \beta_0 + \sum_{j=1}^{q} \psi_{ij}\beta_j + \varepsilon_i$$

which can be further expressed in matrix notation by

$$y_i = \beta_0 + \psi_i\beta + \varepsilon_i \tag{18}$$

where $\psi_i = [\psi_{i1}, \ldots, \psi_{iq}]$ and $\beta = [\beta_1, \ldots, \beta_q]^T$. Model (18) may be interpreted as the Cockerham's model if the following constraints are enforced,

$$\sum_{i=1}^{n} \psi_i = 0_q \text{ and } \sum_{i=1}^{n} \psi_i^T\psi_i = nI_{q\times q}$$

where 0_q is a $1\times q$ vector of zeros and $I_{q\times q}$ is an identity matrix with dimension q. Such a ψ_i can be found using

$$\psi_i = (x_i - \bar{x})L,$$

where L is a generalized inverse of the Choleskey decomposition (upper triangular matrix) of

$$\sum = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x}),$$

ie, $L = \sum^{-1/2}$ and $L^TL = \sum^{-1}$. By taking such an orthogonal transformation, we can see that $\beta_0 = b_0 + \bar{x}b_0$ and $\beta = L^{-1}b$. Substituting ψ_i, β_0 and β into equation (18), we get

$$\begin{aligned} y_i &= \beta_0 + \psi_i\beta + \varepsilon_i \\ &= b_0 + \bar{x}b + (x_i - \bar{x})LL^{-1}b + \varepsilon_i \\ &= b_0 + x_ib + \varepsilon_i \end{aligned} \tag{19}$$

and thus we have recovered model (2). One can directly deal with the Cockerham's model using our penalized likelihood method for estimating β, or estimate the original genetic parameters $\hat{b}$ with our method and then convert them into Cockerham's statistical parameters using $\hat{\beta} = L^{-1}\hat{b}$. With the Cockerham's orthogonal transformation, we can see that the total phenotypic variance can be partitioned into independent variance components, as shown below,

$$\begin{aligned} Var(y_i) &= \sum_{j=1}^{q} Var(\psi_{ij}\beta_j) + Var(\varepsilon_i) \\ &= \sum_{j=1}^{q} Var(\psi_{ij})\beta_j^2 + \sigma^2 \\ &= \sum_{j=1}^{q} \beta_j^2 + \sigma^2 \end{aligned} \tag{20}$$

The Bayesian method developed by Xu (2003) is a model-selection-free method for multiple QTL mapping. The method is very simple so that it can be easily extended to mapping epistatic effects with very little additional effort. However, because the method is implemented via the MCMC, computing time then becomes a major concern for that extension. Within each iteration, a huge number of parameters need to be updated (sampled) and the posterior sample size should be of the order of tens of thousands. Although the computing times spent on updating parameters in each iterations are almost the same for the proposed penalized method and the Bayesian method, the former only takes a few (<70 usually) iterations to converge, while the latter takes several orders of magnitude of iterations to converge to a stationary distribution. The time-saving factor was the major motivation for developing the current method. In addition, the method facilitates an approximate hypothesis test for each putative QTL effect, while the Bayesian method of Xu (2003) does not.

The penalized likelihood method bears all the shrinkage property of the method of Xu (2003). This explains why both methods can handle an extremely oversaturated linear model. However, the shrinkage factor is even more selective in the penalized method. This is reflected by the additional prior distribution assigned to μ_j, which is the prior mean for b_j and is equal to zero in the Bayesian method. Here, in the penalized analysis, the prior for μ_j is $p(\mu_j) = N(0, \sigma_j^2/\eta)$ with $\eta > 0$, which facilitates a mechanism to estimate μ_j from the data. It is different from Gianola *et al* (2003). With an estimated μ_j away from zero, the shrinkage for large QTL effects is not as stringent as when $\mu_j = 0$. However, when the QTL effects are indeed very small, the estimated μ_j is closer to zero than b_j is, and thus the shrinkage becomes more stringent for smaller b_j than for larger b_j. This more selectively different shrinkage is one of the major differences of the penalized likelihood method from the Bayesian method of Xu (2003). The subjective parameter $\eta > 0$ does not have a major influence on the result as long as $\eta < \infty$. We tested a wide range of values for $\eta > 0$, for example, 10, 30, 50, and so on; the results were all comparable to that when $\eta = 5$. We understand that if $\eta \to \infty$, this is of no difference from setting a prior $p(b_j) = N(0, \sigma_j^2)$ for the regression coefficient, and the

result is not as good as that when $\eta < \infty$. Our experience shows that $1 \leq \eta \leq 30$ works well for all the simulated data we have examined.

The proposed penalized likelihood method also differs from that of Boer *et al* (2002), who defined the epistatic effect of each marker as a variance component of the current marker with all other markers (background). Our method actually pinpoints the occurrence of the interaction between specific markers. Is it possible to include higher-order interactions in the epistatic model? Theoretically, it is possible because our model does not have restriction on the number of model effects. Practically, however, there is a concern with the lengthy computing time. Based on the assumption of higher-order interactions being less important than lower-order interactions (Falconer, 1989), the pairwise epistatic genetic model should suffice. The proposed penalized likelihood method also differs from that of Kao and Zeng (2002). Although the latter makes use of the orthogonal property of Cockerham's model, variable selection is still needed.

The penalized likelihood method is similar to the BIC or other information-criterion-based methods for parameter estimation (Akaike, 1973; Schwarz, 1978; Jansen, 1994; Broman and Speed, 2002; Sillanpaa and Corander, 2002). However, these methods were designed mainly for model selection. We adopted a similar idea, but allowed the penalty to be a function of the parameters. This modification appears to be trivial, but it has played an important role in improving the performance of the method. All variables are included in the model and no variable selection is performed, and thus we take no risk of missing any important QTL effects.

We used LOD ≥ 3 as the criterion to declare statistical significance. We found no false positive QTL. LOD $= 3$ is somewhat arbitrary. We then took a permutation approach (Churchill and Doerge, 1994) to find the empirical critical values. These empirical 95% critical values tend to be small (all less than 1.0, data not shown). Based on the empirical critical value, we did find a few false-positive QTL. For the large-genome simulation experiment (the third simulation experiment), three false-positive QTL were found. Considering a total number of 7381 effects included in the model, we think that the false positive rate is extremely low. In the first and second simulation experiments (small genome simulations), we only found eight false-positive QTL. Therefore, the probability of false positive is very low. Our experience indicates that the penalized likelihood analysis usually generates such clear signals of QTL effects that a statistical test may not even be required.

The method was validated using simulated data. When applied to real data for estimating epistatic effects, most of the favorable properties will remain. However, some minor modifications are required before the method is applied to real data. (1) In reality, markers may not be evenly distributed along the genome. Although our method does not depend on the uniformity of marker distribution, very tightly linked markers may cause poor estimates of the marker effects due to high degree of multicollinearity. Therefore, it is recommended to use only one marker from a cluster of markers. (2) Missing marker genotypes may occur in real data analysis. Multiple imputations for the missing marker genotypes (Sen and Churchill, 2001) may be adopted here to simulate the missing genotypes. This requires multiple analyses of the data, each for one imputed data set. In all, 10–20 imputed data sets may suffice (Sen and Churchill, 2001). Alternatively, we may replace the indicator variables for the missing marker genotypes by their conditional expectations calculated with the multipoint method (Rao and Xu, 1998). (3) When two markers are far away, it is possible to insert a virtual marker in the middle of the interval bracketed by the markers. The genotypes of the virtual markers are missing across all individuals. Imputations of the virtual marker genotypes are required to complete the data analysis. (4) In real-data analysis, the expected outcome will be slightly different from what was observed in the simulation study in that the background noise of the plots (see Figure 2) will be larger in the real data analysis than in the simulation study. This is not a deficit of the method; rather, it is due to the polygenic nature of quantitative traits. The high 'noise' may not be the true noise but caused by the polygenic effects. Excluding these background polygenic effects from the model, as done in any model selection approach, may be detrimental because the polygenic effects, collectively, may significantly contribute to the residual variance.

Finally, we have paid all our attention to developing the penalized likelihood method for handling oversaturated model. We only evaluated marker effects so that we could focus on the performance of the penalized likelihood method on an oversaturated model rather than digressing to estimating QTL positions. Broman and Speed (2002) took the same approach when they investigated various model selection algorithms on QTL mapping. They fixed QTL at marker positions so that they could concentrate on the main issue of model selection rather than addressing estimation of QTL positions. The natural next step would be to develop a true QTL-mapping method incorporating the penalized likelihood method, in which we would allow QTL to move away from markers positions. This has been carried out in the Bayesian shrinkage analysis of Wang *et al* (2005) for main effect QTL. Extension to QTL with epistatic effects, making use of this penalized likelihood framework, is underway and will be reported in a subsequent paper.

Acknowledgements

We thank two anonymous reviewers and the associate editor for their constructive comments on an earlier version of the manuscript. The research was supported by the National Institute of Health Grant R01-GM55321 SX.

References

Akaike H (1973). Second international symposium on information theory. In: Petrox BN, Caski F (eds) *Information Theory and an Extension of the Maximum Likelihood Principle*. Akademiai Kiado: Budapest. p 267.

Balding DJ (2002). Discussion on the meeting on 'statistical modeling and analysis of genetic data'. *J R Statist Soc: Ser B* **64**: 737–775.

Boer MP, Braak CJF, Jansen RC (2002). A penalized likelihood method for mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* **162**: 951–960.

Broman KW, Speed TP (2002). A model selection approach for the identification of quantitative trait loci in experiment crosses. *J R Statist Soc B* **64**: 641–656.
Cheverud JM, Routman EJ (1995). Epistasis and its contribution to genetic variance components. *Genetics* **139**: 1455–1461.
Churchill GA, Doerge RW (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.
Cockerham CC (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* **39**: 859–882.
Falconer DS (1989). *Introduction to Quantitative Genetics*, 3rd edn. John Wiley and Sons: New York.
Gianola D, Perez-Enciso M, Toro MA (2003). On marker-assisted prediction of genetic value: beyond the ridge. *Genetics* **163**: 347–365.
Hoerl AE, Kennard RW (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**: 55–67.
Jannink JL, Jansen RC (2001). Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* **157**: 445–454.
Jansen RC (1994). Controlling the type I and II errors in mapping quantitative trait loci. *Genetics* **138**: 871–881.
Kadane JB, Lazar NA (2004). Methods and criteria for model selection. *J Am Statist Assoc* **99**: 279–290.
Kao CH, Zeng Z-B (2002). Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics* **160**: 1243–1261.
Kao CH, Zeng Z-B, Teasdale RD (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**: 1203–1216.
Lander ES, Botstein D (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.
Oh C, Ye KQ, He QM, Mendell NR (2003). Locating disease genes using Bayesian variable selection with the Haseman–Elston method. *BMC Genet* **4**(Suppl. 1): S69.
Rao SQ, Xu S (1998). Mapping quantitative trait loci for categorical traits in four-way crosses. *Heredity* **81**: 214–224.
SAS institute (1999). *SAS/IML User's Guide*, Version 8. SAS Institute Inc.: Cary.
Schwarz G (1978). Estimating the dimension of a model. *Ann Statist* **6**: 461–464.
Sen S, Churchill GA (2001). A statistical framework for quantitative trait mapping. *Genetics* **159**: 371–387.
Sillanpaa MJ, Corander J (2002). Model choice in gene mapping: what and why. *Trends Genet* **18**: 301–307.
Wang H, Zhang Y–M, Li X, Masinde GL, Mohan S, Baylink DJ, Xu S (2005). Bayesian shrinkage estimation of QTL parameters. *Genetics* (in press), doi:10.1534/genetics.104.039362.
Whittaker JC, Thompson R, Denham MG (2000). Marker-assisted selection using ridge regression. *Genet Res* **75**: 249–252.
Xu S (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**: 789–801.
Xu S, Yi N, Burke D, Galecki A, Miller RA (2003). An EM algorithm for mapping binary disease loci: application to fibrosarcoma in a four-way cross mouse family. *Genet Res* **82**: 127–138.
Yi NJ, George V, Allison DB (2003). Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* **164**: 1129–1138.