Bayesian Inference of Epistatic Interactions in Case-Control Studies

Yu Zhang and Jun S. Liu*

* corresponding author

Supplementary Methods

Simulate Disease Associated Markers

We use the following procedure to simulate genotypes at k disease associated markers:

1. Specify the MAF of each disease locus in controls;
2. Under the independence and HWE assumption, calculate the genotype vector frequencies Θ_u of disease loci in controls, where the frequency for each of 3^k possible genotype vectors is the product of the corresponding marginal genotype frequency of each disease locus;
3. Calculate the genotype vector frequencies Θ_c of disease loci in cases, by $\Theta_u \cdot \mathrm{B}$, where B is a 3^k dimensional vector;
4. Determine the conditional probability $\{Q_{ij}\}$ of the allele (j=0,1) of the associated marker given the allele (i=0,1) of each disease locus, such that a specified LD is obtained (either measured by r^2 or by D');
5. Simulate genotypes in cases and controls according to Θ_c, Θ_u, $\{Q_{ij}\}$, for each disease associated marker respectively.

In step (3), B is proportional to the risks of genotype vectors, and the risks are represented as functions of θ in Table 1. We numerically solve θ, and thus B, so as to

obtain a desired marginal effect size λ per disease locus. B is normalized such that $|\mathrm{B}| = 1$. Let A/a denote the major/minor allele of a SNP, and let $D/\overline{D}$ denote the affected/unaffected status. The effect size of a disease locus is defined as

$$\lambda = \frac{P(D \mid Aa)/P(D \mid AA)}{P(\overline{D} \mid Aa)/P(\overline{D} \mid AA)} - 1 = \frac{P(Aa \mid D)/P(AA \mid D)}{P(Aa \mid \overline{D})/P(AA \mid \overline{D})} - 1$$

which can be derived from Θ_u and Θ_c. For example, if we use Model 2 with the MAFs of the two disease loci fixed at 0.1, we obtain the values of θ as 0.95, 1.40, 2.25, and 3.04 for λ = 0.2, 0.3, 0.5, and 0.7, respectively. In comparison, if we fix the MAFs at 0.5, the corresponding values of θ become 0.19, 0.28, 0.45, 0.61, respectively.

In step (4), we numerically determine $\{Q_{ij}\}$ so that a desired D' (or r^2) is obtained. For dataset in Figure 2, we fixed the MAFs at the disease loci and the associated markers. For other simulation datasets, we only fixed the MAFs at disease loci, but restricted $\{Q_{ij}\}$ to be symmetric. In either case, we only need to find Q_{00} such that $\{Q_{ij}\}$ can be determined for a specific D' (or r^2) value for each disease locus. We start randomly and iteratively modify Q_{00} by a small amount towards the desired D' (or r^2) value. The step size was reduced to half whenever two adjacent updates yielded one larger LD and one smaller LD with respect to the desired LD. The iterative update of Q_{00} is terminated when a certain precision (0.001) is achieved.

Model Linkage Disequilibrium

A first-order Markov model is used to account for LD between adjacent markers. With transition probabilities $\{p_{ij}\}_{i,j=1,2,3}$ between two adjacent markers, the likelihood of genotypes at the second marker conditional on the first marker is $\prod_{i=1}^{3}\prod_{j=1}^{3} p_{ij}^{n_{ij}}$, where n_{ij} is the number of transitions from genotype *i* of the first marker to genotype *j* of the second marker. We assign a Dirichlet prior to $\{p_{ij}\}$ with parameter $\{\gamma_{ij}\}_{i,j=1,2,3}$=0.5, and integrate out $\{p_{ij}\}$ to obtain the likelihood function $\prod_{i=1}^{3}\left[\left(\prod_{j=1}^{3}\frac{\Gamma(n_{ij}+\gamma_{ij})}{\Gamma(\gamma_{ij})}\right)\frac{\Gamma(|\gamma|)}{\Gamma(N+|\gamma|)}\right]$.

We further introduce a binary label $\varphi = 0,1$ indicating whether or not markers in group 2 are truly associated with the disease. We iteratively update the value of φ and record markers in group 2 only when $\varphi = 1$. Let D_2 and U_2 denote the genotypes of markers in group2, in cases and controls, respectively. Using the Bayes rule, we have

$$P(\varphi \mid D_2, U_2) = \frac{P(D_2, U_2 \mid \varphi)P(\varphi)}{P(D_2, U_2 \mid \varphi = 0)P(\varphi = 0) + P(D_2, U_2 \mid \varphi = 1)P(\varphi = 1)}$$

Since $\varphi = 0$ indicates that the dependence between markers in group 2 is non-specific to the disease population, we have

$$P(D_2, U_2 \mid \varphi = 0) = \left(\prod_{l=1}^{3^{l_2}}\frac{\Gamma(n_l + m_l + \beta_l)}{\Gamma(\beta_l)}\right)\frac{\Gamma(|\beta|)}{\Gamma\left(N_d + N_u + |\beta|\right)}$$

Here, l_2 denotes the number of markers in group2, n_l, m_l denotes the number of genotype vector *l* in cases and controls, respectively, and $\{\beta_l\}$ are Dirichlet parameters with $\beta_l = 0.5, \forall l = 1,...,3^{l_2}$. If $\varphi = 1$, the dependence between markers in group 2 is associated with the disease, and we define $P(D_2, U_2 \mid \varphi = 1) = P(D_2)P(U_2)$, where

$P(D_2)$, $P(U_2)$ are given in formula (2) and formula (3) in Methods, respectively. Finally, we set the prior for $\varphi = 0,1$ by 0.5 each.

The Null Distribution of the B-statistic and the Conditional B-statistic

Under the null hypothesis that a set *M* of *k* markers are not associated with the disease, we prove that the asymptotic distribution of $2B_M$ is a shifted chi-square with $(3^k - 1)$ degrees of freedom. In general, under the null hypothesis that markers in $M \backslash T$ are not associated with the disease, we can show that $2B_{M|T}$ follows asymptotically a shifted chi-square distribution with $(3^k - 3^t)$ degrees of freedom.

We first consider B_M. For conciseness, we only show here the case where markers are unlinked in controls. The linked case can be proved similarly and more easily. When markers in controls are unlinked, we have

$$B_M = \ln \frac{P_{join}(D_M)P_{ind}(U_M)}{P_{ind}(D_M, U_M)} + \ln \frac{1 + P_{join}(U_M)/P_{ind}(U_M)}{1 + P_{join}(D_M, U_M)/P_{ind}(D_M, U_M)} \quad \text{(S1)}$$

As the sample size increased, it is standard to show that $\frac{P_{join}(D_M, U_M)}{P_{ind}(D_M, U_M)} \to 0$ if the markers are truly unlinked. Thus, the second term in (S1) converges to zero, whereas the first term in (S1) can be further written as

$$\ln \frac{P_{join}(D_M)P_{ind}(U_M)}{P_{ind}(D_M, U_M)} = \ln \frac{\prod_{l=1}^{d} \Gamma(x_l + \alpha_l) \prod_{i=1}^{k} \prod_{j=1}^{3} \Gamma(m_{ij} + \beta_j)}{\prod_{i=1}^{k} \prod_{j=1}^{3} \Gamma(n_{ij} + m_{ij} + \gamma_j)} + c, \quad \text{(S2)}$$

where x_l denotes the number of diplotypes l observed at the k markers in cases, n_{ij} and m_{ij} denote the number of allele j observed at marker i from cases and controls, respectively, α, β, γ are the Dirichlet parameters, and c is a generic constant, which may take different values in different formulas. Let N_d and N_u denote the numbers of cases and controls, respectively. Let $\omega = \frac{N_d}{N_d + N_u}$, $\hat{\eta}_l = \frac{x_l}{N_d}$, $\hat{p}_{ij} = \frac{n_{ij}}{N_d}$, $\hat{q}_{ij} = \frac{m_{ij}}{N_u}$, and $\tilde{p}_{ij} = \omega \hat{p}_{ij} + (1-\omega)\hat{q}_{ij}$. Using Stirling's formula, we can approximate (S2) by

$$B_M \approx N_d \sum_{l=1}^{d} \hat{\eta}_l \ln \hat{\eta}_l + N_u \sum_{i=1}^{k} \sum_{j=1}^{3} \hat{q}_{ij} \ln \hat{q}_{ij} - (N_d + N_u) \sum_{i=1}^{k} \sum_{j=1}^{3} \tilde{p}_{ij} \ln \tilde{p}_{ij} + c \qquad \text{(S3)}$$

Under the null hypothesis, the allele frequency is $p_{ij} = q_{ij}$ for cases and controls. Since markers are unlinked, the expectation of $\hat{\eta}$ is simply $\eta_l = \prod_{i=1}^{k} q_{ij}$, where j is the allele of marker i, and j at each disease marker is uniquely determined by the diplotype l.

By further Taylor-expanding (S2) with respect to $\hat{\eta}_l, \tilde{q}_{ij}, \tilde{p}_{ij}$ around their means, i.e., using the formula $x \ln x \approx \mu \ln \mu + (\ln \mu + 1)(x - \mu) + \frac{(x-\mu)^2}{2\mu}$, we have

$$2B_M \approx N_d \sum_{l=1}^{d} \frac{(\hat{\eta}_l - \eta_l)^2}{\eta_l} + N_u \sum_{i=1}^{k} \sum_{j=1}^{3} \frac{(\hat{q}_{ij} - q_{ij})^2}{q_{ij}} - (N_d + N_u) \sum_{i=1}^{k} \sum_{j=1}^{3} \frac{(\tilde{p}_{ij} - p_{ij})^2}{p_{ij}} + c \qquad \text{(S4)}$$

Note that since the markers are unlinked under the null, the estimated diplotype frequency $\hat{\eta}_l$ is very close to $\tilde{\eta}_l = \prod_{i=1}^{k} \hat{p}_{ij} = \prod_{i=1}^{k} p_{ij} \left(1 + \frac{\hat{p}_{ij} - p_{ij}}{p_{ij}}\right) \approx \eta_l \left(1 + \sum_{i=1}^{k} \frac{\hat{p}_{ij} - p_{ij}}{p_{ij}}\right)$, which can be written as $\tilde{\eta}_l \approx \eta_l + \delta$. Thus, since $\sum_{l=1}^{d} \frac{(\hat{\eta}_l - \tilde{\eta}_l)\delta_l}{\eta_l} = 0$, we have

$$\sum_{l=1}^{d}\frac{(\hat{\eta}_l-\eta_l)^2}{\eta_l}\approx\sum_{l=1}^{d}\frac{(\hat{\eta}_l-\tilde{\eta}_l+\delta_l)^2}{\eta_l}=\sum_{l=1}^{d}\frac{(\hat{\eta}_l-\tilde{\eta}_l)^2}{\eta_l}+\sum_{l=1}^{d}\frac{\delta_l^{\,2}}{\eta_l},$$

It is easy to see that $\sum_{l=1}^{d}\frac{\delta_l^{\,2}}{\eta_l}=\sum_{i=1}^{k}\sum_{j=1}^{3}\frac{(\hat{p}_{ij}-p_{ij})^2}{p_{ij}}$. Consequently, if we define

$$X=N_d\sum_{l=1}^{d}\frac{(\hat{\eta}_l-\tilde{\eta}_l)^2}{\eta_l}=N_d\sum_{l=1}^{d}\frac{(\hat{\eta}_l-\prod_{i=1}^{k}\hat{p}_{ij})^2}{\eta_l}, \text{ and}$$

$$Y=N_d\sum_{i=1}^{k}\sum_{j=1}^{3}\frac{(\hat{p}_{ij}-p_{ij})^2}{p_{ij}}+N_u\sum_{i=1}^{k}\sum_{j=1}^{3}\frac{(\hat{q}_{ij}-q_{ij})^2}{q_{ij}}-(N_d+N_u)\sum_{i=1}^{k}\sum_{j=1}^{3}\frac{(\tilde{p}_{ij}-p_{ij})^2}{p_{ij}},$$

we have $2B_M\approx X+Y+c$.

Under the null, *X*'s asymptotic distribution is clearly a chi-square with 3^k-2k-1 df. It is also easy to follow a standard proof used for contingency tables to show that *Y* has the asymptotic chi-square distribution with *2k* df. Furthermore, since the asymptotic distribution of *X* conditional on the marginal frequencies $\hat{p}_{ij}$ and $\hat{q}_{ij}$ is always the same chi-square distribution, *X* must be independent of $\hat{p}_{ij}$ and $\hat{q}_{ij}$. On the other hand, *Y* is a function of $\hat{p}_{ij}$ and $\hat{q}_{ij}$. Thus, X must be independent of *Y*. This shows that the asymptotic distribution of *X*+*Y* is chi-square with 3^k-1 df.

Following similar arguments, we can show that the conditional B-statistic $B_{M|T}$ also has the stated asymptotic distribution. The shift parameter can be calculated as $c=2c_1$, where

$$c_1 = -\frac{d-1}{2}\ln\left[\frac{\omega(1-\omega)}{2\pi}(N_d + N_u)\right] + \ln\frac{\Gamma(|\alpha|)\Gamma(|\beta|)\prod_{i=1}^{d}\Gamma(\gamma_i)}{\Gamma(|\gamma|)\prod_{i=1}^{d}\Gamma(\alpha_i)\prod_{i=1}^{d}\Gamma(\beta_i)} \quad \text{(S5)}$$

for linked markers, and $c = 2c_1 + (3^k - 2k - 1)\ln(1-\omega)$ for unlinked markers. If we set $\alpha_i = \beta_j = \gamma_k = 0.5, \forall i, j, k$, the shift parameter is unrelated with allele frequencies. Whether or not the markers in M are linked can be determined by the posterior probability $P(linkage \mid U_M) = \frac{P_{join}(U_M)}{P_{ind}(U_M) + P_{join}(U_M)}$.

An alternative and asymptotically equivalent approach to the above B-statistic is to first determine whether or not markers in controls are linked according to either a log-likelihood ratio test or some Bayesian test, and then model the controls accordingly and use another log-likelihood ratio test to further test for associations. A comparison of the two methods is under investigation.

A Hierarchical Procedure to Declare Significance

We developed a hierarchical approach to evaluate the statistical significance for interactions of various sizes. Given a case-control dataset, significant markers were detected as follows. (1) We report all markers with significant marginal associations after a Bonferroni correction for the total number of markers, L. (2) We report all novel 2-way interactions (i.e., neither markers has been reported earlier) that are significant after the Bonferroni correction for $L(L-1)/2$ tests. For those significant 2-way interactions of which one marker has been reported earlier, we compute its conditional B-statistic (or the

conditional log likelihood ratio (LLR) for logistic regression), estimate the p-value, and report the interaction if significant after a Bonferroni correction. We ignore the case when both markers have been reported significant because this will not affect the detection power. (3) We report all novel 3-way interactions that are significant after a Bonferroni correction for $L(L-1)(L-2)/6$ tests, If t=1 or 2 markers in a significant 3-way interaction were already found significant, we calculate the conditional B-statistic (or the conditional LLR), estimate the p-value, and report the interaction if it is still significant after a Bonferroni correction. All p-values were estimated by a chi-square distribution with $d=3^k-3^t$ degrees of freedom, for k=1,2,3, t=0,1,2, $t<k$, and adjusted by Bonferroni corrections. The asymptotic null distribution of the conditional B-statistic is indeed the stated chi-square. Thus, the hierarchical procedure controlled the type I errors under the expected level.

A Hidden Markov Process to Simulate Descendants of the 146 individuals in the AMD Dataset

To simulate genotypes of the 100K SNPs for one diseased descendent (patient) according to model 2, for example, we 1) randomly select two SNPs in the AMD dataset as disease SNPs (dSNPs); 2) compute the joint genotype frequency vector for the dSNPs according to the model; 3) sample a genotype configuration for the dSNPs according to the calculated frequencies and assign it to the patient; 4) generate genotypes of the remaining SNPs of the patient according to a hidden Markov process. As shown in Supplementary Figure 4, each SNP locus has a hidden state that follows a Markov chain (the bottom chain). The state of chain indicates one of the 146 individuals in the original AMD

dataset, and the transition probability to another individual chosen at random is 0.3 per Mb. The two red circles represent the two disease loci, with its genotype data inherited from individual #3. This individual is randomly chosen among those in the AMD dataset who carry the selected genotype configuration at the disease loci. If no individuals carry the selected genotype configuration at the disease loci, we return to step (1) and choose another set of disease loci

Since the disease loci (with their states pre-selected) partition the genome into intervals, within each interval, either both ends are disease loci or one end is a disease locus. We first perform forward summation to calculate the probability of states at each marker from left to right. By default, the initial distribution of forward summation is uniform. However, if the left-most marker is a disease marker, the initial probability is 1 for the disease state, and 0 for other states. We then perform backward sampling to sample the states of each marker from right to left, and if the right-most marker is a disease marker, its state is the pre-determined value. Once the hidden chain is generated, we just copy down the genotypes from the individuals indicated by the chain to the patient. The same procedure, but with different joint genotype frequencies in step 2, is used to simulate individuals in controls. Finally, we remove the disease SNPs from the simulated dataset as if they are unobserved.

Supplementary Notes

Simulation Results for the Null Distribution of B-statistics

We conducted simulation studies to verify our chi-square approximation. We randomly selected 3000 sets of *k*-markers, for *k*=1, 2, 3, from a null dataset (without disease markers) and compared the empirical distribution of $2B_M - c$ to a $(3^k - 1)$ -df chi-square distribution in a quantile-quantile plot (QQ-plot). We simulated datasets with different numbers of cases and controls: 500 cases against 1000 controls, 1000 cases against 1000 controls, and 1000 cases against 500 controls. As shown in Figure A1, the empirical distributions of B-statistics agreed well with the chi-square distributions. Alternatively, we can use a histogram to visualize the estimated p-values of the B-statistic based on its asymptotic chi-square distribution, which should be uniform between [0,1] if the asymptotics kicks in early enough. Figure A2(a) shows that our approximated p-values for 3-way interactions was conservative when sample sizes are 1000; whereas these p-values became very well behaved when the sample size was increased to 10000 for both cases and controls (Fig A2(b)).

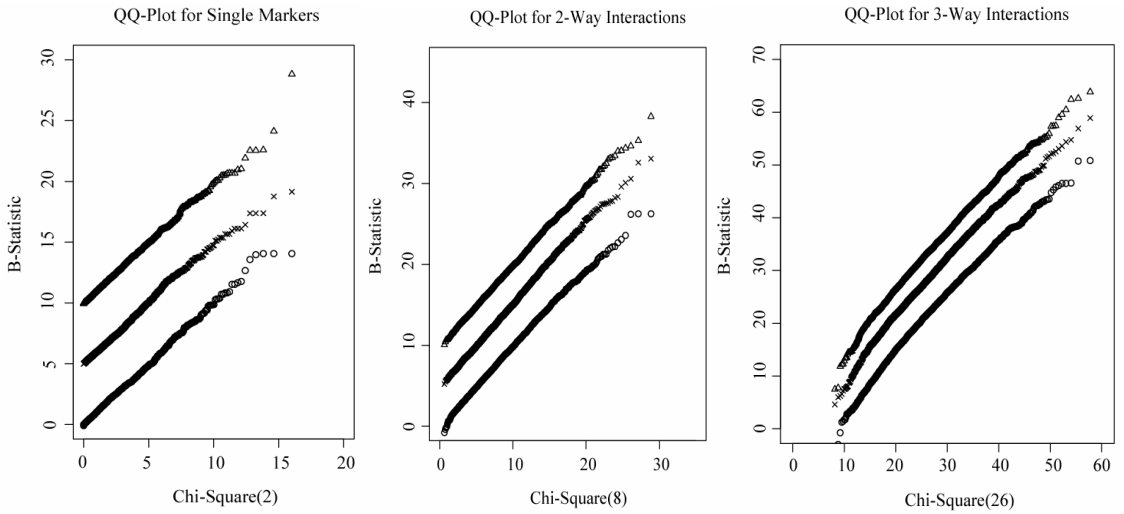
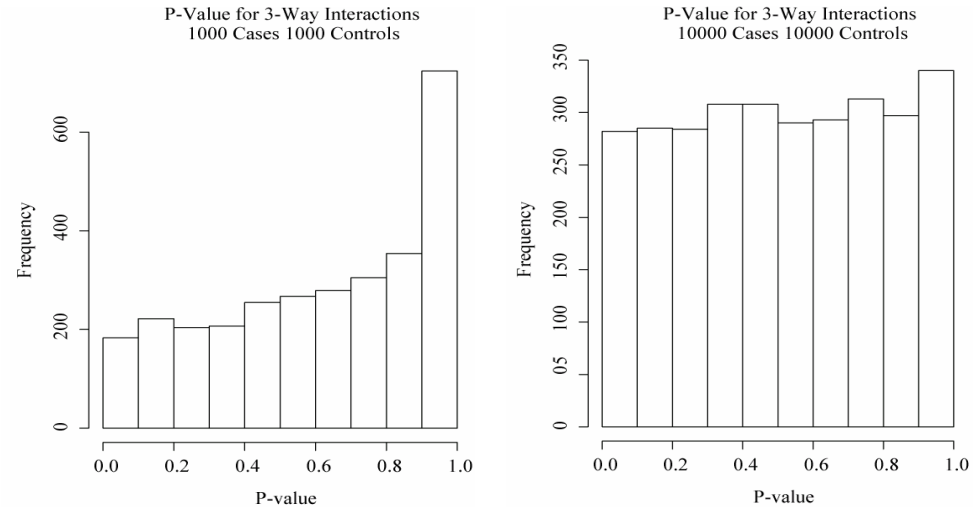


Figure A1. QQ-plots for the simulated B-statistics against chi-square distributions. In each plot, results for three different sample sizes are shown. 'o': (500 cases, 1000 controls); '$\times$': (1000 cases, 1000 controls); and 'Δ': (1000 cases, 500 controls). For clarity, we arbitrarily shifted data of each sample size by a step size of 5 along the y-axis.



(a) (b)

Figure A2. Histogram of p-values estimated for 3-way B-statistics. (a) Results from 1000 cases and 1000 controls; (b) results from 10000 cases and 10000 controls.

Next, we check the empirical distribution of the conditional B-statistic, which has been used in a hierarchical approach to identity interactions. In particular, when a subset T of $t(<k)$ markers in a set M of k markers are significantly associated with the disease, we demonstrate that the null distribution of $2B_{M|T} - c$ can be approximated by a $(3^k - 3^t)$-df chi-square distribution. We first simulated a dataset containing a 2-way interaction according to disease Model 2, with marginal effect size $\lambda = 5$ and MAF=0.4 per marker. We then randomly selected 3000 sets of k markers. In each set, t markers were selected from the disease markers. We chose a large sample size of 10000 cases and 10000 controls. For each set, we calculated $B_{M|T}$, and then used a QQ-plot to compare

$2B_{M|T} - c$ with a $(3^k - 3^t)$-df chi-square distribution. The constant c is computed from using (S5) as the difference between constants for M and T. We did this for (k,t)=(2,1),(3,1),(3,2), respectively, and the results are shown in Figure A3. The empirical distributions agreed very well with chi-square distributions.

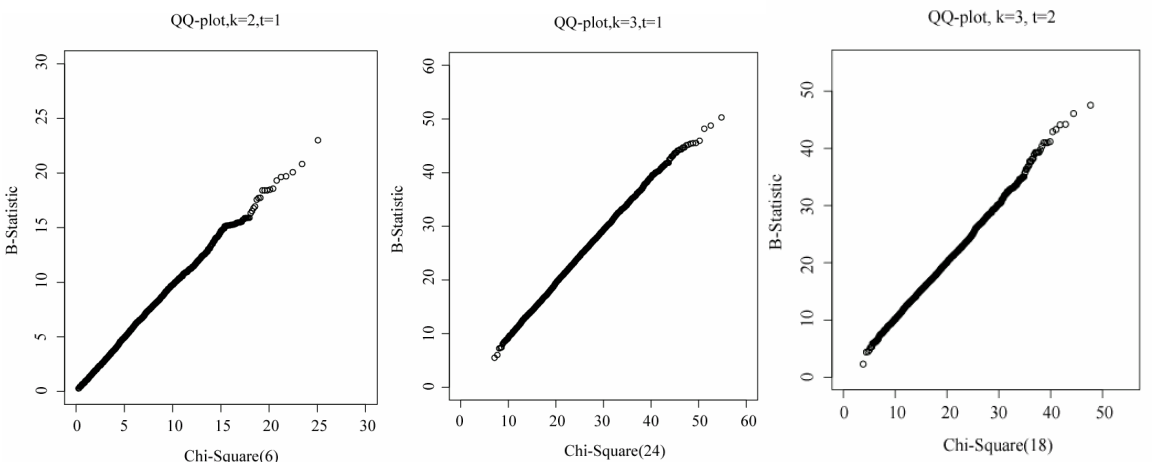


Figure A3. QQ-plots of empirical distributions of $B_{M|T}$ versus their approximating chi-square distributions, 10000 cases and 10000 controls. Left: k=2, t=1; Middle: k=3, t=1; Right: k=3, t=2.

Type I Error Rates

Here we examined the type I error rates of BEAM, stepwise B-stat, and stepwise Logistic Regression based on datasets shown in Figure 1, at the significance level 0.1 after Bonferroni corrections. The type I error rates for each method and each interaction size k can be calculated as the number of significant k-way interactions that contain at least one unassociated marker, divided by the number of datasets tested. The results calculated in the unit of interaction were summarized in Table A1. An alternative way to calculate type I error rates is to count the number of unassociated markers being significant either

marginally or jointly with others, divided by the number of datasets tested. The results calculated in the unit of marker were summarized in Table A2. Since there are only 50 datasets for each disease model under each setting, the type I error rates were averaged over various settings for each disease model.

Table A1. Type I Error Rates of BEAM (B), stepwise B-stat (S), and stepwise Logistic Regression (L), calculated in the unit of interaction.

	$k=1$			$k=2$			$k=3$		
	B	S	L	B	S	L	B	S	L
Model 1	.100	.100	.104	.018	.018	.025	.000	.000	.000
Model 2	.114	.114	.115	.013	.008	.026	.000	.000	.000
Model 3	.103	.103	.103	.010	.004	.010	.000	.000	.000
Model 4	.094	.094	.095	.019	.008	.029	.003	.003	.006
Model 5	.113	.113	.114	.020	.023	.041	.016	.045	.020
Model 6	.093	.093	.098	.015	.021	.019	.044	.023	.005
Average	.103	.103	.105	.016	.014	.025	.010	.014	.005

Table A2. Type I Error Rates of BEAM (B), stepwise B-stat (S), and stepwise Logistic Regression (L), calculated in the unit of marker.

	$k=1$			$k=2$			$k=3$		
	B	S	L	B	S	L	B	S	L
Model 1	.100	.100	.104	.030	.031	.046	.000	.000	.000
Model 2	.114	.114	.115	.024	.014	.045	.000	.000	.000
Model 3	.103	.103	.103	.018	.005	.020	.000	.000	.000
Model 4	.094	.094	.095	.034	.013	.054	.008	.008	.016
Model 5	.113	.113	.114	.039	.041	.075	.019	.053	.049
Model 6	.093	.093	.098	.029	.040	.033	.044	.029	.013
Average	.103	.103	.105	.029	.024	.045	.012	.015	.013

We observe that BEAM, stepwise B-stat, and stepwise Logistic Regression all produced similar amount of type I errors for datasets in Figure 1. In particular, the type I error rates for single markers were about 10%, as expected at the 0.1 significance level. For interactions, however, all methods produced fewer type I errors, which may be due to the fact that the Bonferroni correction is too conservative especially for the highly correlated *k*-way interaction statistics. Interestingly, under Model 5 and Model 6, the type I errors of

BEAM and stepwise B-stat for 3-way interactions were relatively large. We manually checked those false positives. Most of the false positive 3-way interactions contained exactly two true disease markers! This is perhaps a desirable feature since it can help get back some weakly associated markers even if they do not really interact. Similar increase of type I errors were also found for stepwise Logistic Regression, but more so from 2-way interactions. The overall type I error rate measured in the unit of marker, summed over all interaction sizes and averaged over all datasets, was 0.143 for BEAM, 0.141 for stepwise B-stat, and 0.163 for stepwise Logistic Regression.

Since it is a possibility that our under-commitment of type-I errors is caused by the methods' not being able to exhaustively search all possible interactions, we performed additional simulation studies in which we exhaustively examined all interaction combinations. We simulated 500 datasets without associations. Each dataset contains 100 markers in 1000 cases and 1000 controls. We exhaustively tested all 100 one-way, 4950 two-way, 161700 three-way interactions in each dataset at the significance level of 0.1 after Bonferroni corrections using both the B-statistic and the log-likelihood ratio statistic in Logistic Regression. The type I error rates for each interaction size (k) were calculated in the unit of interaction and summarized in Table A3. For marginal associations, both B-statistics and Logistic Regression have type-I error rates slightly lower than 10%. For 2-way and 3-way interactions, both methods appeared to be very conservative, further confirming our intuition that the Bonferroni correction is too conservative.

Table A3. Type I error rates for B-statistics and Logistic Regression, calculated in the unit of interaction.

Type I Error Rate	Single Marker	2-Way Interaction	3-Way Interaction
B-Statistic	0.084	0.039	0.023
Logistic Regression	0.088	0.047	0.043

MCMC Convergence

A simple diagnostic of the convergence of BEAM is to run multiple MCMC chains independently, and monitor the ratio between the within-chain variance and the between-chain variance of the log-posterior probability[1]. In our experience, for datasets with 1000 markers, the MCMC chains mix well after the first few thousands of iterations. In general, the number of burn-in iterations for each chain in BEAM should be a few times the total number of markers (L) considered. After burn-in, running BEAM for cL^2 iterations, with c between 1-5, is often sufficient to obtain accurate estimates of posterior distributions. We also observe that the auto-correlation time[2] of the log-posterior probability is in the order of L, which intuitively means that one obtains the equivalence of one independent posterior sample for every cL iterations. Supplementary Figure 2 displays the trace and autocorrelation plots of the log-joint posterior probability for a simulated data and the AMD data.

Posterior Distributions

BEAM outputs posterior probabilities for each marker and/or set of markers to be associated with the disease. To illustrate, we used the same dataset as shown in Supplementary Figure 2, which contained 1000 markers in 1000 cases and 1000 controls. Three markers were selected to be interacting among the cases (Model 4), each with

disease MAF=0.1 and marginal effect size $\lambda = 0.4$. Based on 3 independent chains with 35,000 iterations each, we show in Supplementary Figure 3(a) the posterior probability for each marker to be associated with the disease, either independently or jointly with some other markers. Supplementary Figure 3 also shows the posterior distributions for the number of marginal associations (b) as well as the interaction size (c). Furthermore, when we down-tune the prior probability for association from 1/3 to 1/300, we observe in Supplementary Figure 3(d,e,f) that those sporadic "wrong" markers with high posterior probabilities disappeared. Although the prior effect will diminish as sample sizes increases, in practice it is often insightful to examine the effects of priors and other critical model assumptions.

Reference

1. Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7:457-72
2. Liu, J.S. *Monte Carlo Strategies in Scientific Computing*. Springer, New York (2001).

Supplementary Table 1: Simulated disease models

Model 1

Risk	*A/A*	*A/a*	*a/a*
B/B	*1*	$1+\theta$	$(1+\theta)^2$
B/b	$1+\theta$	$(1+\theta)^2$	$(1+\theta)^3$
b/b	$(1+\theta)^2$	$(1+\theta)^3$	$(1+\theta)^4$

Model 2

Risk	*A/A*	*A/a*	*a/a*
B/B	*1*	*1*	*1*
B/b	*1*	$(1+\theta)^2$	$(1+\theta)^3$
b/b	*1*	$(1+\theta)^3$	$(1+\theta)^4$

Model 3

Risk	*A/A*	*A/a*	*a/a*
B/B	*1*	*1*	*1*
B/b	*1*	$1+\theta$	$1+\theta$
b/b	*1*	$1+\theta$	$1+\theta$

Model 4

	A/A			*A/a*			*a/a*		
Risk	*C/C*	*C/c*	*c/c*	*C/C*	*C/c*	*c/c*	*C/C*	*C/c*	*c/c*
B/B	*1*	*1*	*1*	*1*	*1*	$1+\theta$	*1*	$1+\theta$	*1*
B/b	*1*	*1*	$1+\theta$	*1*	$1+\alpha\theta$	*1*	$1+\theta$	*1*	*1*
b/b	*1*	$1+\theta$	*1*	$1+\theta$	*1*	*1*	*1*	*1*	*1*

Model 5

Risk	*A/A*	*A/a*	*a/a*	Risk	*C/C*	*C/c*	*c/c*
B/B	*1*	*1*	*1*	*D/D*	*1*	*1*	*1*
B/b	*1*	*1*	$1+\theta$	*D/d*	*1*	*1*	$1+\theta$
b/b	*1*	$1+\theta$	*1*	*d/d*	*1*	$1+\theta$	*1*

Table legend: Each table cell lists the relative risk of the corresponding genotype combination. Genotypes with risks equal to 1 have no effects to the disease (phenocopies). The parameter θ is computed conditional on specified marginal effects and disease MAFs. For Model 4, we choose α=10, 4, 1.5, 0.5 when MAF=0.05, 0.1, 0.2, 0.5, respectively. Since there were no marginal effects in Model 4 when MAF=0.5, we chose θ=7 arbitrarily. Model 5 is a 4-locus model consisting of two 2-way interactions. Each of the 2-way interaction contributes to the disease risk but the effects of two interactions do not add. We further constructed a 6-way interaction model (Model 6) as follows: denote the genotypes of each SNP by 0, 1, and 2 and code each genotype combination over 6 disease loci by integers between 0~728; assign disease effect θ= 50 to genotype combinations 4, 5, 7, 111, 114, 253, 254, 360, 387, 603, and 630. We chose θ= 50 so that these genotype combinations can explain a non-trivial portion (>10%) of cases.

Supplementary Figure 1: Prior calibration

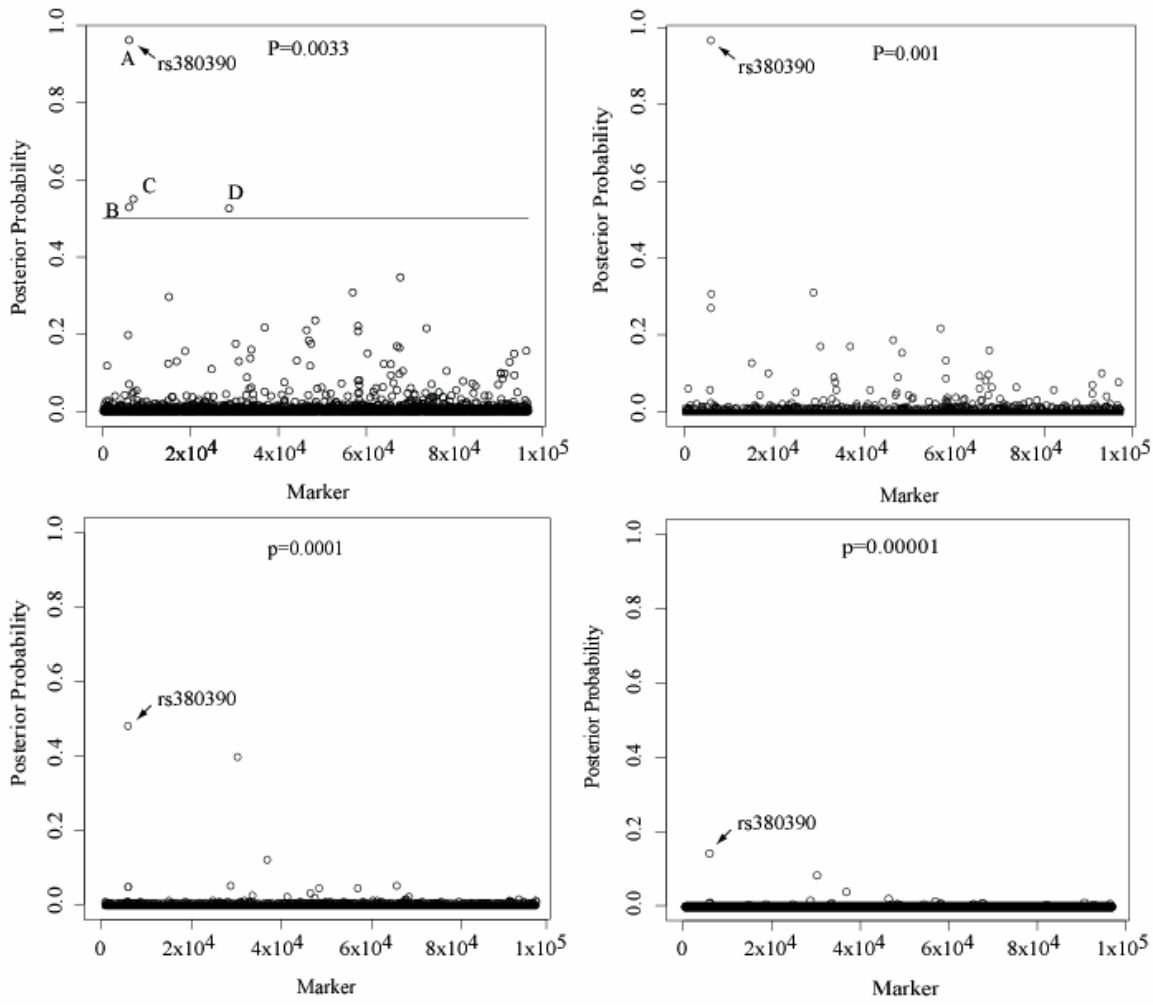


Figure legend: Posterior probability of associations for each marker in the AMD dataset. From upper left to bottom right clockwise: p=0.0033, p=0.001, p=0.0001, and p=0.00001, where p=p1=p2 are the priors for each marker to belong to group 1 and group 2, respectively. With decreasing priors, the posterior distribution shrinks to zero, but the relative order of markers ranked by posterior probabilities remains almost the same. In the upper-left panel, markers A and B corresponds to rs380390 and rs1329428, the two markers previously reported by Klein et al. Marker C (rs1296210) is located 2Mb downstream of A and B, while marker D (rs3775640) is on chromosome 4. Marker C's

marginal significance is partly due to a joint effect it has with its preceding marker. Although the prior effect will diminish as sample sizes increases, in practice it is often insightful to examine the effects of priors and other critical model assumptions.

Supplementary Figure 2: Trace and autocorrelation plots

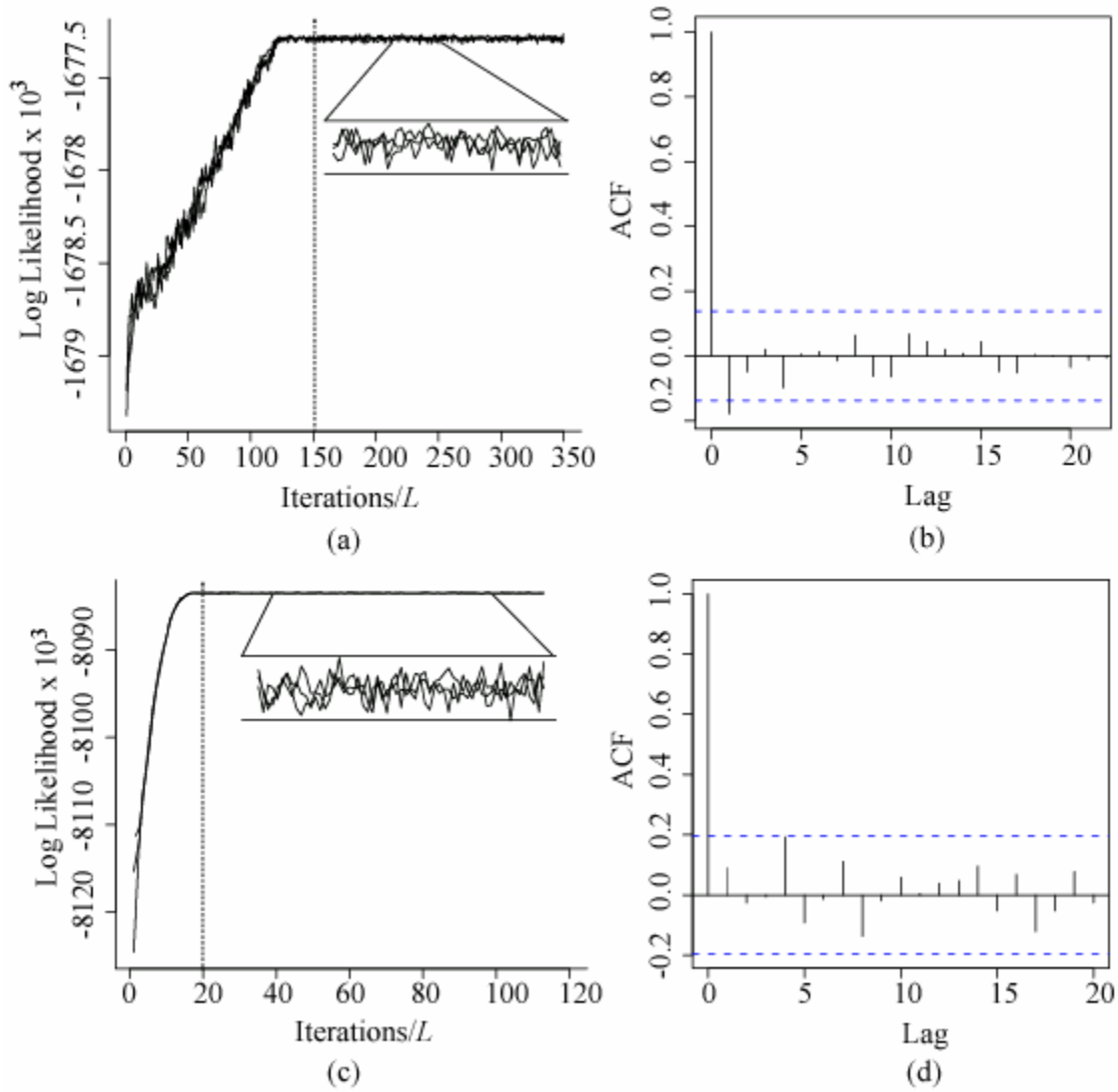


Figure legend: Trace plots and autocorrelation plots for the log-joint posterior probability for a simulated dataset and the AMD dataset. The simulated dataset contains 1000 markers from 1000 cases and 1000 controls. The disease model is Model 4, with MAF 0.1 and marginal effect size 0.4 per locus. Trace plots are obtained from three independent chains, and the autocorrelation plots are obtained from the first chain. We ran BEAM for 150,000 burn-ins plus 200,000 samplings in three chains for the simulated dataset, with prior $p_1 = p_2 = 1/3$. We ran BEAM for 2,000,000 burn-ins plus

10,000,000 samplings in three chains for the AMD dataset, with prior $p_1 = p_2 = 0.001$. (a) & (b): trace and autocorrelation plots for the simulated data, respectively; (c) & (d): for the AMD data. Autocorrelation plots are generated using samples from every L iterations, where L=1000 for the simulated dataset, and L=96,932 for the AMD dataset. Part of trace plots are amplified in order to show three chains.

Supplementary Figure 3: Posterior distribution of associations

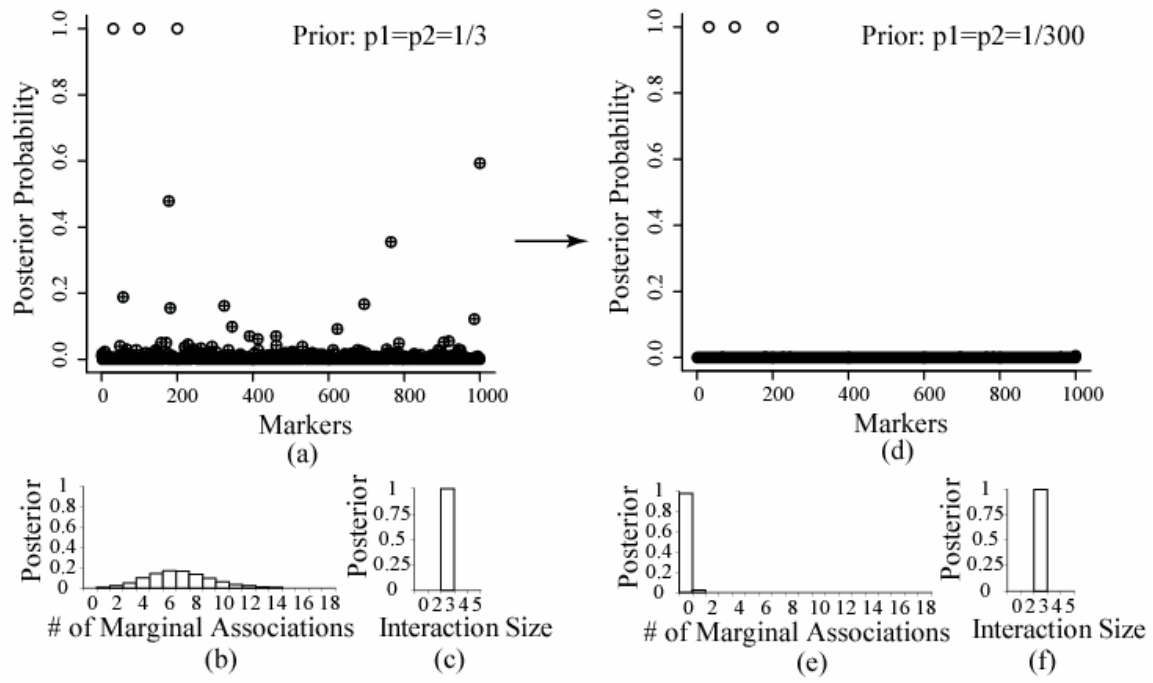


Figure legend: Posterior probability of associations for each marker estimated from the same dataset used in Supplementary Figure 2(a). Circles denote the overall posterior probabilities of associations, with marginal and joint associations combined. Plus signs denote posterior probabilities of marginal associations. Three circles on the top of (a) and (d) correspond to the three simulated disease markers having interaction effects. (a) posterior probability of associations for each marker; (b) posterior distribution for the number of marginally associated markers in the dataset, where 0 is the true value; (c) posterior distribution for the interaction size, where 3 is the true value. Plots (a,b,c) are obtained using $(p_0, p_1, p_2) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, where p_i is the prior probability for each marker to belong to group i. Plots (d,e,f) are the corresponding results using $(p_0, p_1, p_2) = (\frac{298}{300}, \frac{1}{300}, \frac{1}{300})$.

Supplementary Figure 4: HMM process for simulating descendants from the AMD dataset

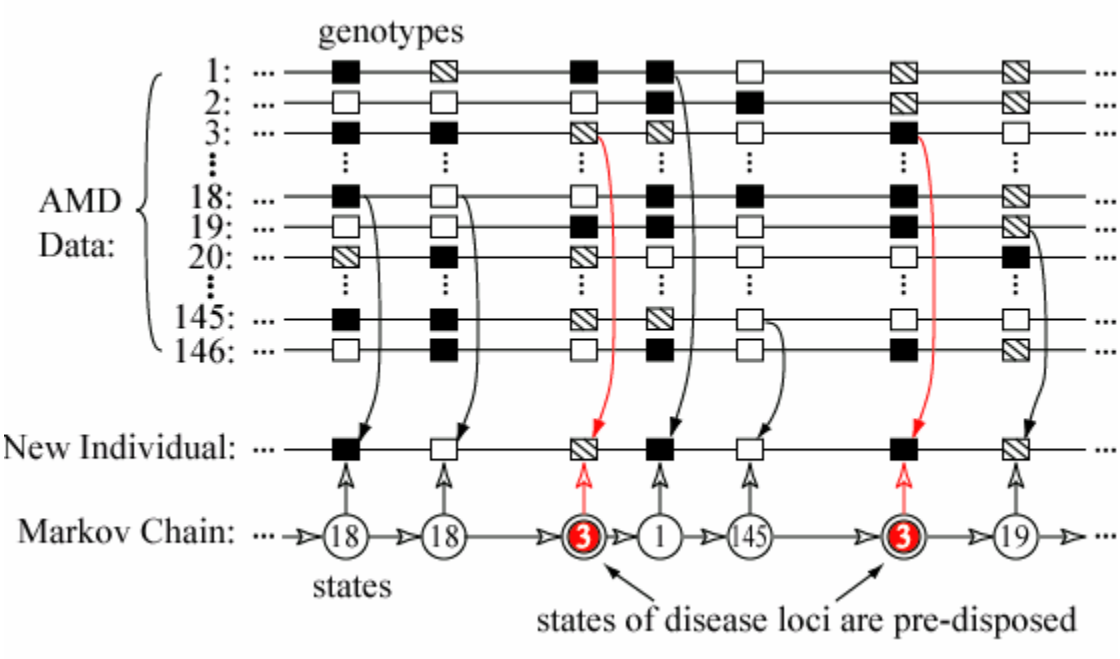


Figure legend: The hidden Markov chain used to simulate descendants of the 146 individuals in the AMD dataset. Each SNP locus has a hidden state that follows a Markov chain (bottom). The state of chain indicates one of the 146 AMD individuals (top). The two red circles represent the two disease loci, with its genotype data inherited from individual #3. The states of the two disease loci are selected before determining the states of other loci.