nature genetics

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

Bayesian inference of epistatic interactions in case-control studies

Yu Zhang[1] & Jun S Liu[2]

Epistatic interactions among multiple genetic variants in the human genome may be important in determining individual susceptibility to common diseases. Although some existing computational methods for identifying genetic interactions have been effective for small-scale studies, we here propose a method, denoted 'bayesian epistasis association mapping' (BEAM), for genome-wide case-control studies. BEAM treats the disease-associated markers and their interactions via a bayesian partitioning model and computes, via Markov chain Monte Carlo, the posterior probability that each marker set is associated with the disease. Testing this on an age-related macular degeneration genome-wide association data set, we demonstrate that the method is significantly more powerful than existing approaches and that genome-wide case-control epistasis mapping with many thousands of markers is both computationally and statistically feasible.

In the past century, scientists have made great progresses in mapping genes responsible for mendelian diseases. However, genetic variants underlying most common (or 'complex') diseases are non-mendelian. These variants are typically not rare in the population ($>2\%$). They show very little effect independently with low penetrance, but they may interact with each other in complex ways. The joint behavior of genetic variants is often referred to as epistasis or multilocus interaction. It has been speculated that epistasis ubiquitously contributes to complex traits partly because of the sophisticated regulatory mechanisms encoded in the human genome[1]. An increasing number of reports have indicated the presence of multilocus interactions in many human complex traits, such as breast cancer[2], post-PTCA stenosis[3], essential hypertension[4], atrial fibrillation[5] and type 2 diabetes[6].

As the number of possible interaction combinations among the genotyped markers is astronomical for a large-scale case-control genetic association study, it is a daunting task to 'catch' one or a very few disease-related interactions among all these combinations. Several approaches have been developed to detect epistasis, including the combinatorial partitioning method (CPM)[7], the restricted partitioning method (RPM)[8], multifactor-dimensionality reduction (MDR)[2], multivariate adaptive regression spline (MARS)[9], the logic regression method[10] and backward genotype-trait association (BGTA)[11]. Although these methods all showed promise, they have been tested only on small data sets. For example, logic regression and BGTA have been tested on data sets with 89 and 80 biallelic markers, respectively; and methods based on brute-force searches such as CPM and MDR are impractical for large data sets. Recently, a simulation study[12] explored the use of a stepwise logistic regression approach to identify two-way and three-way interactions. The authors demonstrated that searching for interactions in genome-wide association mapping can be more fruitful than traditional approaches that exclusively focus on marginal effects.

Here we introduce the bayesian epistasis association mapping (BEAM) algorithm for identifying both single-marker and epistasis associations in population-based case-control studies. Our method uses Markov chain Monte Carlo (MCMC) to 'interrogate' each marker conditional on the current status of other markers iteratively and outputs the posterior probability that each marker and/or epistasis is associated with the disease. Using extensive simulations, we demonstrate that BEAM is considerably more powerful than existing methods for epistasis mapping. We also applied BEAM to an association study of age-related macular degeneration (AMD)[13], which included ~100,000 SNP markers. Although BEAM did not find significant interactions in the AMD data set, it was able to discover two-way or three-way interactions among the ~100,000 SNPs simulated based on the AMD data. Our study indicates that a genome-scale epistasis mapping is both feasible and desirable: it does not lose much power when epistasis is not present and can often be more powerful than the single-marker approach.

RESULTS

The BEAM algorithm

The BEAM algorithm takes case-control genotype marker data as input and produces, via MCMC simulations, posterior probabilities that each marker is associated with the disease and involved with other markers in epistasis. The input genotyped markers should be in their natural genomic order when there is linkage disequilibrium (LD) among some of them. The method can be used either in a 'pure' bayesian sense or just as a tool to discover potential 'hits'. For the former, one relies on the reported posterior probabilities to make

[1]Department of Statistics, the Pennsylvania State University, Thomas Building 422A, University Park, Pennsylvania 16802, USA. [2]Department of Statistics, Harvard University, Science Center 715, 1 Oxford Street, Cambridge, Massachusetts 02138, USA. Correspondence should be addressed to J.S.L. (jliu@stat.harvard.edu).

Received 21 February; accepted 2 July; published online 26 August 2007; doi:10.1038/ng2110

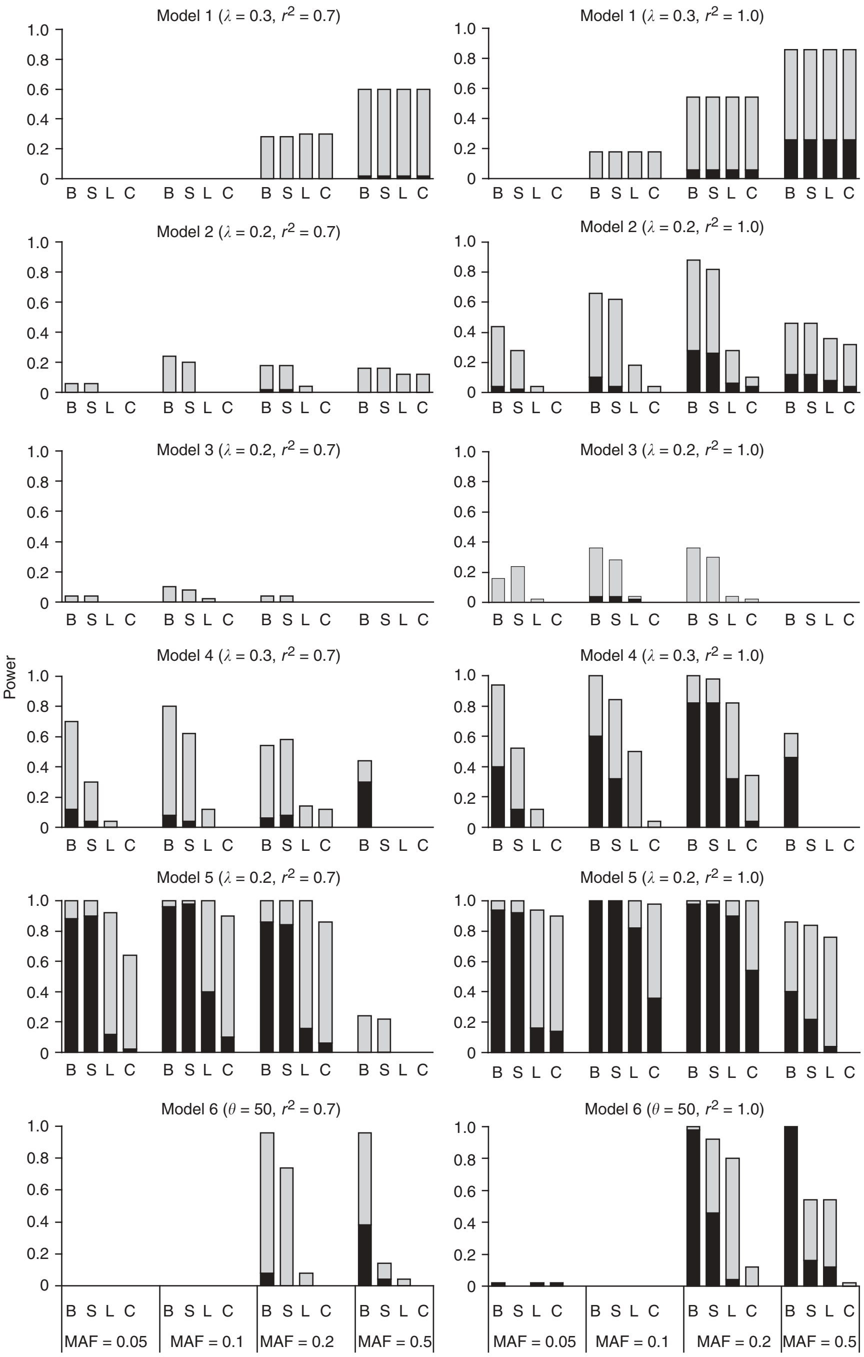

Figure 1 Comparison between BEAM (B), the stepwise B-stat (S), the stepwise logistic regression (L) and the 2-d.f. χ^2 test (C) on six disease models. Under each setting, the power is calculated as the proportion of 50 data sets in which all associated markers are identified at a 重要性阈值 significance threshold of 0.1 after Bonferroni correction. Each data set contains 1,000 markers. Black bars represent the power for 1,000 cases and 1,000 controls, and gray bars represent the power for 2,000 cases and 2,000 controls. The absence of bars indicates zero power. 连锁不平衡 LD between each unobserved disease locus and the associated genotyped marker is r^2:测量未发现的疾病位点和关联性的基因位点 measured by r^2. λ:每一个疾病位点的边界效应的影响度 The marginal effect per disease locus is measured in effect size λ. For model 6, θ:相互影响程度 the interaction effect size $\theta = 50$.

inferential statements; as for the latter, one can take the reported hits and use another procedure to test whether these hits are 统计显著 statistically significant. The latter approach is more robust to 选型 model selection and 是先假设 prior assumptions (such as 狄利克雷先验 Dirichlet priors with arbitrary parameters) and is less 倾向 prone to the slow mixing problem in the MCMC computational 计算过程 procedure. We also propose the *B* statistic to facilitate the latter approach and show that it is more powerful than the standard χ^2 statistic for epistasis detections.

Epistasis models and simulations

There is a wide 宽频谱 spectrum of interaction models in which the disease risks at single markers are small[14]. Here we consider six models with different characteristics (see **Supplementary Table 1** online for details). Model 1 contains two 疾病位点 disease loci, each of which 导致疾病的风险是相互独立的 contributes to the disease risk independently (that 即 is, their effects 影响叠加 are additive). Model 2 is similar to model 1, but the disease risk is present only when both loci have at least one disease allele. Model 3 is a threshold model in which additional disease alleles at each locus do not further increase the disease risk. Model 4 contains three disease loci. Increased disease risk is assigned to certain genotype combinations, and marginal effect of each disease locus ranges from very small to zero. The interaction effects in these models are determined such that the marginal effect, measured by the effect size (defined as the odds ratio minus 1) of each disease locus, equals a specified value. Model 5 is constructed to mimic multiple causal epistasis by a mixture of two two-way interactions. Each two-way interaction can increase the disease risk, but the disease risk is not further increased when both two-way interactions are present. Model 6 is designed to have a six-way interaction. Unless specified otherwise, 50 data sets for each disease model were simulated under each setting, with marker minor allele frequencies (MAF) chosen uniformly in [0.05, 0.5]. Each untyped disease locus is linked to one genotyped marker, and the remaining markers are unlinked. More simulation details can be found in **Supplementary Methods** online.

Comparison with stepwise logistic regression

The stepwise logistic regression approach of ref. 12 works as follows: (i) all markers are individually tested and ranked for marginal associations with the disease; (ii) the top 10% of markers are selected, among which all *k*-way ($k = 2$ or 3) interactions are tested and ranked for associations. The authors of ref. 12 also proposed an exhaustive logistic regression testing approach, which we choose not to consider in this study because of its prohibitive computational cost. Note that even their stepwise approach can become computationally intractable for high-order interactions. As a benchmark, we also implemented a

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

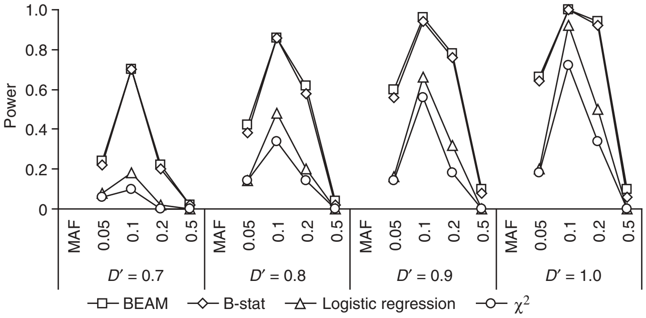

Figure 2 Impact of MAF discrepancy and LD on the powers of BEAM (B), the stepwise B-stat (S), the stepwise logistic regression (L) and the 2-d.f. χ^2 test (C). The comparison is based on model 2, where the allele frequencies of the second disease locus are unmatched by that of the associated marker. The marginal effect size per disease locus is 0.5. Under each setting, the power is calculated from 50 data sets containing 1,000 markers genotyped from 1,000 cases and 1,000 controls. The power is the proportion of 50 data sets in which all associated markers are identified at a significance threshold of 0.1 after Bonferroni correction.

χ^2 test with two degrees of freedom (2 d.f.; for three possible genotypes of a biallelic marker) to test for single-marker associations.

We used both BEAM and stepwise logistic regression to search for significant associations of up to three-way interactions and used the 2-d.f. χ^2 test to search for marginal associations among 1,000 markers. To better compare the MCMC and the stepwise searching strategies, we further implemented a stepwise method called 'stepwise B-stat', which uses the same search strategy as stepwise logistic regression but uses the B statistic proposed in this article for testing significance. We used Bonferroni correction to account for multiple comparisons. We define the power of each method as the proportion of 50 data sets in which all truly associated markers are identified and show statistically significant associations (adjusted P values below 0.1) with the disease.

For the non-epistasis model (model 1), all three epistasis mapping methods performed similarly to the single-marker χ^2 test (**Fig. 1**), indicating that the power for detecting marginal associations was not compromised by using the more complex models. For epistasis models (models 2–6), BEAM (and often the stepwise B-stat) significantly outperformed the stepwise logistic regression, which in turn outperformed the single-marker χ^2. The difference was most notable when either disease allele frequencies or marginal effects were small, consistent with the observation that some single-marker association mapping studies were not reproducible. Notably, results for model 4 suggest that stepwise methods can miss markers with small or no marginal effects, whereas BEAM can get these markers back through iterations. Although the power of all methods decreases with the decay of the LD (measured in r^2) between disease loci and associated markers, doubling the sample size can significantly increase the power.

The extra power gained by BEAM relative to that of the stepwise logistic regression is attributable to two factors incorporated in BEAM: the MCMC sampling recipe and the B statistic. The stepwise B-stat achieved a better power than the stepwise logistic regression (**Fig. 1**). We also observed that the stepwise B-stat was less powerful than BEAM, particularly for detecting high-order interactions, indicating the benefit of using an MCMC scheme to search for interactions.

All three epistasis mapping methods made similar amounts of type I errors. At the 0.1 significance level, they all made ~10% type I errors (after Bonferroni correction) when searching only for marginally significant markers. All methods made much fewer than 10% type I errors when searching for interactions. Type I error results and detailed analyses are presented in the **Supplementary Note** online.

Impact of mismatch in allele frequencies and LD

The power of association mapping can be greatly hampered by the discrepancy of allele frequencies between unobserved disease loci and associated genotyped markers[15]. We investigated the impact of such a discrepancy on epistasis mapping by simulating data sets based on model 2, where MAFs at two interacting disease loci were both 0.1, and the marginal effect size per disease locus was 0.5. Two genotyped markers were linked with the two disease loci. One linked marker had the matched MAF, whereas the other had an MAF ranging from 0.05 to 0.5. The LD between disease loci and associated markers was controlled to range from $D' = 0.7$ to $D' = 1$.

BEAM and the stepwise B-stat achieved the highest power for data sets with small frequency mismatch (MAF = 0.05, 0.1 and 0.2) and high LD between disease loci and associated markers ($D' \geq 0.8$) (**Fig. 2**). For data sets with large MAF discrepancies and moderate LD, the power of all methods suffered. BEAM and the stepwise B-stat were clearly more robust to MAF discrepancy and LD decay compared with the other two methods. At the extreme case when the MAF discrepancy was maximized (that is, MAF = 0.5), all methods had little power in detecting interaction associations. The impact of LD on power seemed to be less profound than the effect of MAF discrepancy. Given a small MAF discrepancy and moderate to high LD, epistasis can generally be identified, and the power can be further increased using larger, but feasible, sample sizes (for example, ~1,000 cases and controls).

Genome-wide association study of AMD

Studies have shown that a real genome-wide case-control association study may require genotyping of 30,000–500,000 common SNPs[16,17]. To our knowledge, epistasis mapping at such a scale has yet to become practical, owing to computational and statistical issues. We demonstrate the potential of BEAM in genome-wide association studies by analyzing an AMD data set[13]. The data set contains 116,204 SNPs genotyped for 96 affected individuals and 50 controls. We removed nonpolymorphic SNPs and those that significantly deviated from Hardy-Weinberg Equilibrium (HWE), as suggested in ref. 13. We removed additional SNPs containing more than five missing genotypes. After the filtration, 96,932 SNPs remained.

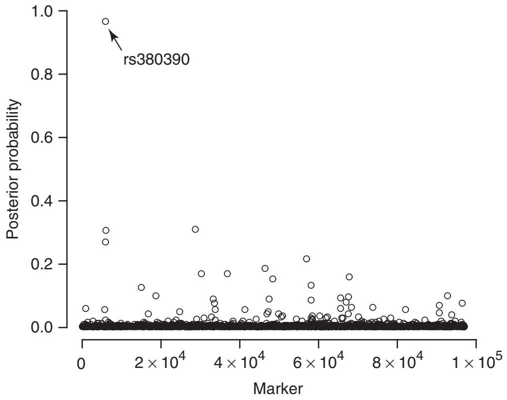

Figure 3 Posterior probabilities of association for each marker in the AMD data set, obtained by running BEAM for 10^8 iterations and taking samples at every 10^5 iterations. Priors for each marker to belong to group 1 (markers contributing independently to the disease risk) and group 2 (markers that jointly influence disease risk) were 0.001 each. Only one marker, rs380390 (reported in ref. 13), has a posterior probability above 0.5.

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

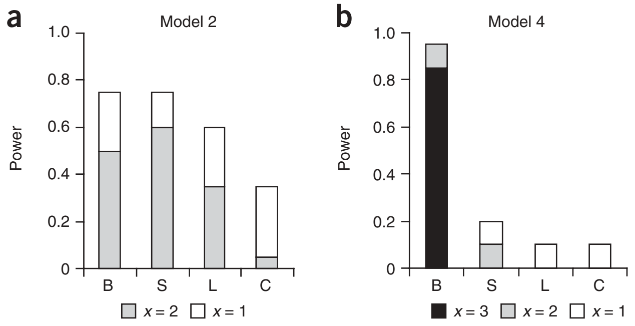

Figure 4 Comparison of BEAM (B), the stepwise B-stat (S), the stepwise logistic regression (L) and the 2-d.f. χ^2 test (C) on the ~100,000-SNP data sets. (**a**) Model 2, with MAF = 0.1 and $\lambda = 0.7$. (**b**) Model 4, with MAF = {0.5,0.5,0.4} at three disease loci and $\theta = 7$. For each disease model, the power is calculated as the proportion of 20 ~100,000-SNP data sets in which $x = 1$, 2 or 3 associated markers are identified at a statistical significance threshold of 0.1 after Bonferroni correction. Each data set contains 500 cases and 500 controls simulated from the original AMD data set.

The authors of ref. 13 found a significant association between AMD and SNP rs380390, with a *P* value of 0.004 based on a Bonferroni-corrected 1-d.f. χ^2 test under the assumption of HWE. BEAM identified rs380390 (**Fig. 3**) with posterior probability above 0.5 (the prior probability was 1/1,000). The other reported marker (rs1329428) did not reach our posterior probability cut-off. As a calibration, we also computed various Bonferroni-corrected *P* values for marker rs380390. The *P* value based on our *B* statistic and its asymptotic χ^2 distribution is 0.06. Using permutations instead of the asymptotic distribution, we obtained a *P* value of 0.047. The *P* value based on the logistic regression (the log-likelihood-ratio test) is 0.059, and that based on the standard χ^2 is 0.156. These *P* values differ significantly from those of ref. 13 because the methods we used here do not assume HWE. In addition, we observed that with only 146 individuals and ~100,000 SNPs, the posterior probability of associations for each marker is strongly influenced by the choice of priors, although the order of these probabilities is nearly invariant (**Supplementary Fig. 1** online).

BEAM found no significant interactions associated with AMD from this data set. It is possible that the small sample size of 146 individuals is insufficient for detecting subtle epistasis interactions. To demonstrate the feasibility of BEAM in genome-wide epistasis mapping, and to provide an example on how to trade off sample sizes and genetic effect sizes, we performed a simulation study based on the AMD data set. We simulated 20 data sets, each containing ~100,000 SNPs genotyped from 500 cases and 500 controls, under model 2 and under model 4. SNPs simulated in each data set have similar genotype distributions and LD structures as in the original AMD data set. For model 2, the MAFs at the two interacting disease loci were both 0.1, and the marginal effect size per disease locus was 0.7. For model 4, the MAFs at the three interacting disease loci were 0.5, 0.5 and 0.4, respectively, and the effect size θ of interactions was set as 7 so that the third locus had a marginal effect of 0.67, and the other two loci had no marginal effects. We ran BEAM for 5 × 10^7 iterations for each data set. The iteration number is the same as the number of all possible two-way interactions among 10% of ~100,000 SNPs. Therefore, BEAM and the stepwise logistic regression took approximately the same computational time to detect two-way interactions, but BEAM took much less computational time to detect three-way interactions.

BEAM and the stepwise B-stat achieved higher power than the stepwise logistic regression and the single-marker χ^2 test for model 2 (**Fig. 4**), and BEAM achieved significantly higher power than all other methods for model 4 in identifying three-way interaction among the ~100,000 SNPs. In both cases, the single-marker χ^2 test performed the worst. The result for model 4 demonstrates that BEAM can make use of weak marginal effects or low-order interactions to gradually work its way toward the correct solution. Intuitively, BEAM starts by biasing in favor of SNPs that show weak marginal or low-order interaction effects. Once it has obtained more disease-associated SNPs by chance, BEAM immediately 'crystallizes' on the true interaction set.

Comparison with other epistasis mapping approaches

We compared the performance of BEAM with those of three other recently developed epistasis mapping algorithms: MDR[2], logic regression[10] and BGTA[11]. MDR identifies *k*-way interactions through an exhaustive search and evaluates the association between each interaction and the disease by cross-validations. Logic regression infers a tree-based relationship between the disease status and a set of markers. It evaluates the detected associations by permutation tests. BGTA uses a bootstrap-type resampling screening procedure to select markers, and those markers with return frequencies greater than the third quartile plus 1.8 times the interquartile range are deemed disease-associated markers.

The comparison was based on simulated data sets under model 4, in which each of the three disease loci is perfectly linked to one genotyped marker, and the marginal effect size per disease locus is 0.5 (except for MAF = 0.5, in which case the marginal effect size is 0). As MDR is computationally expensive, we simulated genotypes at only 40 markers for 400 cases and 400 controls. We also ran the 2-d.f. χ^2 test for each marker as a benchmark of single-marker methods.

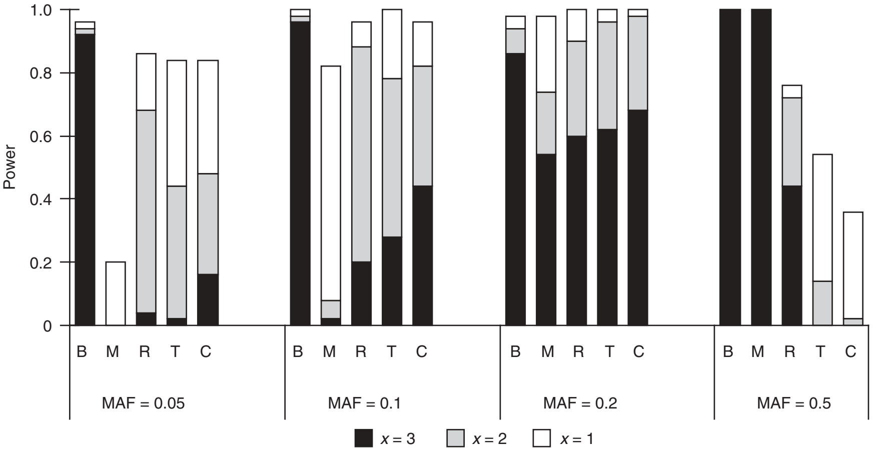

Figure 5 Comparison of BEAM (B), MDR (M), logic regression (R), BGTA (T) and the 2-d.f. χ^2 test (C) on model 4. The marginal effect size per disease locus is 0.5, and LD is 1.0. Under each setting, the power is calculated as the proportion of 50 data sets in which $x = 1$, 2 and 3 associated markers is identified among the top five candidate markers reported by each method. Each data set contains 40 markers genotyped from 400 cases and 400 controls.

Because these methods output differently and have different ways of assessing significance, we calculated the proportion of 50 data sets in which $x = 1,2,3$ associated markers were among the best models output by MDR and logic regression and among the top five ranked markers by BGTA, the single-marker χ^2 test and BEAM, respectively. Based on this criterion, each method will yield the same number of falsely associated markers in worst-case scenarios.

BEAM outperformed all other methods significantly (**Fig. 5**), especially for data sets with small disease allele frequencies (for example, MAF = 0.05 and 0.1). MDR performed better than logic regression for common disease alleles but had little power when disease allele frequencies were small. The power loss for MDR may be due to the large proportion of phenocopies (up to 95%) when disease alleles are rare. Logic regression had a much reduced power when several genotype combinations contributed equally to the disease risk with a small effect. BGTA performed similarly to logic regression, except that it was less powerful when the model was purely epistatic. We were surprised that the χ^2 test performed similarly to all methods but BEAM.

DISCUSSION

The BEAM algorithm has two essential components: a bayesian epistasis inference tool implemented via MCMC and a novel test statistic for evaluating statistical significance. Although these two parts come from opposing schools of statistics, they can provide complementary statistical insights to the scientist and help reconfirm each other. A natural advantage of the bayesian approach is its ability to incorporate prior knowledge about each marker (for example, whether it is in a coding or regulatory region) and to quantify all information and uncertainties in the form of posterior distributions. However, evaluating the statistical significance of a candidate finding via P values is more robust to model choice and prior assumptions and can give the scientist peace of mind. It is worth mentioning that the new test statistic we developed is particularly suitable for detecting epistasis associations, as it is adaptive to the correlation structure of the candidate markers.

We have shown not only that BEAM performs uniformly better than all the existing epistasis mapping methods tested but also that genome-scale epistasis mapping is feasible with BEAM. Consistent with previous reports[15,18], our study demonstrates that, given limited resources and knowledge about the disease of interest, statistical approaches that account for epistasis can greatly increase the chance to identify significant associations. However, the power of such mappings depends critically on sample size, effects of disease mutations and any discrepancy in allele frequencies between disease loci and associated markers.

There are several issues that we have not addressed that may affect the accuracy of our method, such as population substructures, genotyping errors and disease heterogeneities. In principle, the population substructure may be either accommodated directly in our bayesian model or corrected *a priori*, but disease heterogeneities can severely affect any population-based genetic study. In addition, further generalizations and improvements of the bayesian model as well as the MCMC algorithm used here are needed to effectively handle the ~500,000-SNP data sets commonly found in recent genome-wide association studies.

METHODS

Notations. Suppose N_d cases and N_u controls were genotyped at L SNP markers. Let case genotypes be $D = (d_1, \ldots, d_{N_d})$ with $d_i = (d_{i1}, \ldots, d_{iL})$ representing genotypes of patient I at L markers, and let control genotypes be $U = (u_1, \ldots, u_{N_u})$ with $u_i = (u_{i1}, \ldots, u_{iL})$. The L markers are partitioned into three groups: group 0 contains markers unlinked to the disease, group 1 contains markers contributing independently to the disease risk and group 2 contains markers that jointly influence the disease risk (interactions). Let $I = (I_1, \ldots, I_L)$ indicate the membership of the markers with $I_j = 0$, 1 and 2, respectively. Our goal is to infer the set of markers that are associated with the disease (that is, the set $\{j : I_j > 0\}$). Let l_0, l_1, l_2 denote the number of markers in each group ($l_0 + l_1 + l_2 = L$), and let D_0, D_1 and D_2 denote case genotypes of markers in group 0, 1 and 2, respectively.

The bayesian marker partition model. Case genotypes at associated markers should show different distributions when compared with control genotypes. For simplicity, we describe the likelihood model assuming independence between markers in the control population (see **Supplementary Methods** for a generalized model to account for LD). Let $\Theta_1 = \{(\theta_{j1}, \theta_{j2}, \theta_{j3}) : I_j = 1\}$ be the genotype frequencies of each biallelic marker in group 1 in the disease population; we write the likelihood of D_1 as

$$P(D_1|\Theta_1) = \prod_{j:I=1} \prod_{k=1}^{3} \theta_{jk}^{n_{jk}},$$

where $\{n_{j1}, n_{j2}, n_{j3}\}$ are genotype counts of each marker j in group 1. Assuming a Dirichlet(α) prior for $\{\theta_{j1}, \theta_{j2}, \theta_{j3}\}$, where $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, we integrate out Θ_1 and obtain the marginal probability:

$$P(D_1|I) = \prod_{j:I=1} \left(\left(\prod_{k=1}^{3} \frac{\Gamma(n_{jk}+\alpha_k)}{\Gamma(\alpha_k)} \right) \frac{\Gamma(|\alpha|)}{\Gamma(N_d+|\alpha|)} \right) \qquad (1)$$

Here the notation $|\alpha|$ represents the sum of all elements in α.

Markers in group 2 influence the disease risk through interactions. Thus, each genotype combination over the l_2 markers in this group represents a potential interaction. There are 3^{l_2} possible genotype combinations with frequency $\Theta_2 = (\rho_1, \ldots, \rho_{3^{l_2}})$ in the disease population. Let n_k be the number of genotype combination k in D_2. Again, with a Dirichlet(β) prior distribution of Θ_2, $\beta = (\beta_1, \ldots, \beta_{3^l})$, we integrate out Θ_2 so that

$$P(D_2|I) = \left(\prod_{k=1}^{3^{l_2}} \frac{\Gamma(n_k+\beta_k)}{\Gamma(\beta_k)} \right) \frac{\Gamma(|\beta|)}{\Gamma(N_d+|\beta|)} \qquad (2)$$

The remaining data D_0 consist of markers that follow the same distributions as in the control population. Let $\Theta = (\theta_1, \ldots, \theta_L)$ denote the genotype frequencies of the L markers in the control population, and let n_{jk} and m_{jk} be the number of individuals with genotype k at marker j in D and U, respectively. Assuming Dirichlet priors with parameters $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ for θ_j, $j = 1, \ldots, L$, we integrate out Θ and obtain

$$P(D_0, U|I) = \prod_{j=1}^{L} \left(\left(\prod_{k=1}^{3} \frac{\Gamma(n_{jk}+m_{jk}+\gamma_k)}{\Gamma(\gamma_k)} \right) \frac{\Gamma(|\gamma|)}{\Gamma\left(\sum_{k=1}^{3}(n_{jk}+m_{jk})+|\gamma|\right)} \right) \qquad (3)$$

Combining formulas (1), (2) and (3), we obtain the posterior distribution of I as

$$P(I|D,U) \propto P(D_1|I)P(D_2|I)P(D_0,U|I)P(I) \qquad (4)$$

Note that I determines the configuration of D_i. We let $P(I) \propto p_1^{l_1} p_2^{l_1} (1 - p_1 - p_2)^{L-l_1-l_2}$ which may be modified to reflect our prior knowledge of each marker being associated with the disease. As sample sizes dictate our capability in identifying high-order interactions, we restrict that $l_2 \leq \log_3(N_d) - 1$. By default (in the available software), we set $p_1 = p_2 = 0.01$. When BEAM is used as a search tool, these priors can be set quite liberally without affecting the results. For example, **Figures 1**, **2**, **4** and **5** resulted from using $p_1 = p_2 = \frac{1}{3}$ However, if we need to use the posterior probabilities for decision making, the priors need to be calibrated with our prior knowledge. We further set the parameters for the Dirichlet priors as $\alpha_i = \beta_j = \gamma_k = 0.5, \forall i, j, k$.

MCMC sampling. Our goal is to draw the indicator I from distribution (4). We initialize I according to the prior $P(I)$ and use the Metropolis-Hastings (MH)

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

algorithm[19] to update I. Two types of proposals are used: (i) 随机改变一个位点群成员 randomly change a marker's group membership, or (ii) randomly 在三个群体之间随机交换两个位点 exchange two markers between groups 0, 1 and 2. The proposed move is accepted according to the MH ratio, which is just a 比率 ratio of Gamma functions. The output is the posterior distribution of 位点和在与疾病有关联的互作为条件的后验概率 makers and interactions associated with the disease. To improve the sampling efficiency, we first set a 低边界 lower bound on the 位点的数量 number of markers in group 2 and 逐步减少 gradually reduce this bound to 0 during burn-in. This forces the algorithm to explore the space of 高阶 high-order interactions. We also used an 锻炼 annealing strategy in burn-in iterations with a temperature set high initially and gradually reduced to 1. The trace and 自相关 autocorrelation plots shown in **Supplementary Figure 2** online for our analyses of a simulated data set and the AMD data set demonstrate that the MCMC chains 取得 attained their respective 固定的 stationary distributions after first few thousand iterations. An example of posterior estimation is shown in **Supplementary Figure 3** online. More details about MCMC convergence and posterior analysis can be found in the **Supplementary Note**.

***B* statistic and conditional *B* statistic.** Although we can make statistical 统计推断 inferences directly from the posterior probabilities of associations output by BEAM, we can also analyze the results in a 频率论的方法 frequentist way. We developed a 假设检验流程 hypothesis-testing procedure to check each marker or set of markers for 重要的关联 significant associations, where the marker set is selected based on 'hits' output by BEAM. This validation procedure yields results that are more robust to model selection and prior misspecifications and avoids the slow mixing problem often encountered in MCMC.

For each set M of k markers to be tested, the null hypothesis is that markers in M are not associated with the disease. Here, $k = 1,2,3,\ldots$ represents single-marker, two-way and three-way interactions, etc. We define the B statistic for the marker set M as:

$$B_M = \ln \frac{P_A(D_M, U_M)}{P_0(D_M, U_M)} = \ln \frac{P_{\text{join}}(D_M)[P_{\text{ind}}(U_M)+P_{\text{join}}(U_M)]}{P_{\text{ind}}(D_M, U_M)+P_{\text{join}}(D_M, U_M)}$$

Here, D_M and U_M denote the genotype data for M in cases and controls, and P_0 (D_M, U_M) and P_A (D_M, U_M) are really the Bayes factors (that is, the marginal probabilities of the data with parameters integrated out from our bayesian model, under the null and the alternative models, respectively). Under the null model, genotypes in both cases and controls follow a common distribution, whereas under the alternative model they follow different distributions. We choose both P_0 (D_M, U_M) and P_A (U_M) as an equal mixture of two distributions: one that assumes independence among markers in M, P_{ind} (X), of which the form is given in equation (1), and the other a saturated joint distribution of genotype combinations across all markers in M, P_{join} (X), as in equation (2). Under the null hypothesis that M is not associated with the disease, the B statistic is asymptotically distributed as a shifted χ^2 with $3^k - 1$ degrees of freedom (**Supplementary Methods**). The shifting parameter of the distribution can be computed explicitly. Simulations confirm that this asymptotic approximation is quite accurate for reasonably sized data sets (**Supplementary Note**).

A notable feature of the B statistic is its use of a mixture distribution to accommodate the possibilities that the markers in the controls may or may not be in linkage equilibrium. The use of the Bayes factors instead of the typical maximum likelihood function is that the former can automatically penalize the larger model when the smaller model is true. An alternative, and asymptotically equivalent, approach is to first determine whether markers in controls are linked, and then use a corresponding log likelihood ratio test to test for associations.

When testing for interaction associations, a set of k $(= 2,3,\ldots)$ markers may include $t(<k)$ markers that are significant through either marginal or partial interaction associations. In this case, we want to test for the additional association effects conditional on the t associated markers. Let T denote the t associated markers in a set M of k markers; then, the conditional B statistic for the marker set M is defined as

$$B_{M|T} = \ln \frac{P_{\text{join}}(D_M|D_T)[P_{\text{ind}}(U_{M\setminus T})+P_{\text{join}}(U_M|U_T)]}{P_{\text{ind}}(D_{M\setminus T}, U_{M\setminus T})+P_{\text{join}}(D_M, U_M|D_T, U_T)}$$

Here, D_X and U_X denote the genotype data for the marker set X in cases and controls, respectively. Note that the nonconditional B statistic B_M corresponds to the conditional B statistic $B_{M\,|\,T}$ when T is an empty set. We can also show that the asymptotic null distribution of $B_{M\,|\,T}$ is a shifted χ^2, with $3^k - 3^t$ degrees of freedom.

Power calculation. To evaluate the statistical significance of interactions, we used the B statistic for BEAM and the stepwise B-stat, and the log-likelihood ratio for the stepwise logistic regression. Under the null hypothesis, both statistics have the same asymptotic χ^2 distribution. Based on this, we developed a 分层的 hierarchical approach to evaluate the statistical significance for interactions of various sizes. Details of the hierarchical significance declaration procedure can be found in **Supplementary Methods**. Power results shown in **Figures 1, 2** and **4** represent the percentage of data sets in which all disease markers were identified at the significance level 0.1 after Bonferroni corrections.

For results shown in **Figure 5**, we calculated the number of truly disease-associated markers in the top five candidates selected by each method. For BEAM, BGTA and the 2-d.f. χ^2 test, markers were ranked by their posterior probabilities, backward selection frequencies and χ^2 statistics, respectively, for each program. For MDR, the best models of one-, two- and three-way interactions were used and overall included up to six different markers. For logic regression, we specified a search for at most three trees consisting of five leaves. We then took the five leaves in the best logic regression model as the top five candidate markers. Because every marker in logic regression was represented by two variables, it is possible that we underestimated the power of logic regression (but the amount of underestimation should be very small).

Simulation of ~100,000 SNPs from the AMD data set. Intuitively, our procedure simulates fictitious sets of descendents of the 146 individuals in the original AMD data set. To simulate genotypes for one diseased descendent (patient) according to model 2, for example, we (i) randomly selected two SNPs in the AMD data set as disease SNPs (dSNPs), (ii) computed the joint genotype frequency vector for the dSNPs according to the model, (iii) sampled a genotype configuration for the dSNPs according to the calculated frequencies and assign it to the patient and (iv) generated genotypes of the remaining SNPs of the patient according to a hidden Markov process. Details of the hidden Markov process are presented in **Supplementary Methods**, and a depiction of the simulation process is shown in **Supplementary Figure 4**. We used the same procedure, but with different joint genotype frequencies in step (ii), to simulate individuals in controls. Finally, we removed the disease SNPs from the simulated data set as if they were unobserved.

Computation time. The computation time of BEAM depends on the number of MCMC iterations and the number of individuals genotyped. For a data set of 1,000 markers in 1,000 cases and 1,000 controls, BEAM took 4 min to run 500,000 MCMC iterations on a Pentium M 1.6GHz laptop with 512 Mb memory. When the sample size was halved or doubled, the computation times were 2 min and 8 min, respectively. For the AMD data set containing 96,932 markers genotyped from 146 individuals, BEAM ran for about 5 h, for a total of 10^8 iterations.

URLs. BEAM: http://www.fas.harvard.edu/~junliu/BEAM/. MDR (version 1.00rc1): http://www.epistasis.org/mdr.html. Logic regression (version 1.41 for R): http://bear.fhcrc.org/~ingor/logic.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This work was supported in part by US National Institutes of Health grant R01HG002518-06, US National Science Foundation grant DMS-0204674 and a grant from the National Science Foundation of China (10228102). We thank J. Hoh for providing us the AMD data set and T. Niu for discussions.

AUTHOR CONTRIBUTIONS

Y.Z. and J.S.L. designed the statistical models and simulation studies together. Y.Z. implemented the method and wrote the software. Both authors contributed to the writing of the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Moore, J.H. & Williams, S.M. New strategies for identifying gene-gene interactions in hypertension. *Ann. Med.* **34**, 88–95 (2002).
2. Ritchie, M.D. *et al.* Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **69**, 138–147 (2001).
3. Zee, R.Y. *et al.* Multi-locus interactions predict risk for post-PTCA restenosis: an approach to the genetic analysis of common complex disease. *Pharmacogenomics J.* **2**, 197–201 (2002).
4. Williams, S.M. *et al.* Multilocus analysis of hypertension: a hierarchical approach. *Hum. Hered.* **57**, 28–38 (2004).
5. Tsai, C.T. *et al.* Renin-angiotensin system gene polymorphisms and atrial fibrillation. *Circulation* **109**, 1640–1646 (2004).
6. Cho, Y.M. *et al.* Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia* **47**, 549–554 (2004).
7. Nelson, M.R., Kardia, S.L., Ferrell, R.E. & Sing, C.F. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* **11**, 458–470 (2001).
8. Culverhouse, R., Klein, T. & Shannon, W. Detecting epistatic interactions contributing to quantitative traits. *Genet. Epidemiol.* **27**, 141–152 (2004).
9. Cook, N.R., Zee, R.Y. & Ridker, P.M. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat. Med.* **23**, 1439–1453 (2004).
10. Kooperberg, C. & Ruczinski, I. Identifying interaction SNPs using Monte Carlo logic regression. *Genet. Epidemiol.* **28**, 157–170 (2005).
11. Zheng, T., Wang, H. & Lo, S.H. Backward genotype-trait association (BGTA) - based dissection of complex traits in case-control design. *Hum. Hered.* **62**, 196–212 (2006).
12. Marchini, J., Donnelly, P. & Cardon, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37**, 413–417 (2005).
13. Klein, R.J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
14. Culverhouse, R., Suarez, B.K., Lin, J. & Reich, T. A perspective on epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet.* **70**, 461–471 (2002).
15. Zondervan, K.T. & Cardon, L.R. The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5**, 89–100 (2004).
16. Collins, A., Lonjou, C. & Morton, N.E. Genetic epidemiology of single-nucleotide polymorphism. *Proc. Natl. Acad. Sci. USA* **96**, 15173–15177 (1999).
17. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22**, 139–144 (1999).
18. Wang, W.Y.S., Barratt, B.J., Clayton, D.G. & Todd, J.A. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* **6**, 109–118 (2005).
19. Liu, J.S. *Monte Carlo Strategies in Scientific Computing* (Springer, New York, 2001).