Published in final edited form as:
Ann Hum Genet. 2011 January ; 75(1): 112–121. doi:10.1111/j.1469-1809.2010.00627.x.

Detecting Epistatic SNPs Associated with Complex Diseases via a Bayesian Classification Tree Search Method

Min Chen[1], **Judy Cho**[2], and **Hongyu Zhao**[3,*]

[1]Division of Biostatistics, Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX

[2]Internal Medicine, Yale University, New Haven, CT

[3]Center for Statistical Genomics and Proteomics, Department of Epidemiology and Public Health, Yale University, New Haven, CT

Summary

Complex phenotypes are known to be associated with interactions among genetic factors. A growing body of evidence suggests that gene–gene interactions contribute to many common human diseases. Identifying potential interactions of multiple polymorphisms thus may be important to understand the biology and biochemical processes of the disease etiology. However, despite the great success of genome-wide association studies that mostly focus on single locus analysis, it is challenging to detect these interactions, especially when the marginal effects of the susceptible loci are weak and/or they involve several genetic factors. Here we describe a Bayesian classification tree model to detect such interactions in case-control association studies. We show that this method has the potential to uncover interactions involving polymorphisms showing weak to moderate marginal effects as well as multi-factorial interactions involving more than two loci.

Keywords

Epistasis; GWAS; Bayesian CART; MCMC; logistic regression; Crohn's disease

Introduction

To identify genetic variants that are responsible for complex human diseases, genome wide association studies (GWAS) often scan single nucleotide polymorphisms (SNPs) across the genome of individuals in case and control groups. Most GWAS focus on single locus analysis by testing the main effect of one SNP at a time. However, single-locus analysis may fail to detect SNPs that do not have significant marginal effects individually but have strong effects collectively (Culverhouse et al. 2002). It is known that genes interact with one another in biological processes like gene regulations, metabolisms, signal transduction, and various development related pathways. So genetic variants in multiple genomic loci may jointly contribute to complex phenotypes (Moore 2003). In the literature it has been reported that many complex diseases, such as sporadic Alzheimer's disease (Combarros et al. 2009), type 2 diabetes (Wiltshire et al. 2006), breast cancer (Ritchie et al. 2001), among others, are associated with interactions of multiple polymorphisms. The phenomenon that the effect of one variant in one gene may depend on those in other genomic loci is known as epistasis.

*Corresponding author: Hongyu Zhao, Department of Epidemiology and Public Health, Yale University, 60 College Street, New Haven, Connecticut 06520-8034. Tel: (203) 785-6271; Fax: (203) 785-6912; hongyu.zhao@yale.edu.

Despite the potential importance of their roles in uncovering the disease etiology, it is difficult to identify epistatic effects in genome-wide settings. To address this challenge, many statistical methods have been proposed and a recent review paper provides a good survey of methods and related software packages for detecting epistasis (Cordell 2009). According to the author, these methods include exhaustive search algorithms, data-mining and machine learning related approaches, and Bayesian model selection methods.

Among the class of machine learning methods is recursive partitioning that produces tree-structure models (Breiman et al. 1984; Zhang & Bonney 2000; Nelson et al. 2001; Cook et al. 2004). Figure 1 depicts an example of a tree model. In tree models, each nonterminal node defines a splitting rule based on a predictor variable. A path from the top node to each terminal node corresponds to a unique mapping from the predictor space to a specific outcome, depending on the values of all predictor variables along that path. Therefore, each terminal node represents a particular combination of values for all variables on the path, and thus naturally allows epistatic effects of those variables in the model. In addition, due to the fact that there can be multiple levels of nodes involving two or more variables, tree based models also allow detection of multi-way interactions. Since the partition of the predictor space is constructed in a recursive manner, the splitting of a variable is conditional on the values of other variables in its ancestral nodes in the tree.

A common practice of searching the tree space is through a greedy algorithm where at each node the splitting variable and its corresponding partition rule is determined by choosing the one, from the pool of all available variables and splitting values, that maximizes the separation of the resulting partition. Thus, this type of algorithms have the limitation that they may fail to identify those interactions that do not display substantial marginal effects (Cordell 2009). To alleviate this problem, algorithms based on Bayesian modeling were proposed to stochastically search promising classification trees through Markov chain Monte Carlo (MCMC) modelling (Chipman et al. 1998; Denison et al. 1998). Moreover, methods based on Bayesian analysis to detect epistasis association have been proposed by several authors (Lunn et al. 2006; Zhang & Liu 2007). The idea of Bayesian classification trees is closely related to Bayesian model selection in which a prior is assigned to all tree models and it serves the purpose of controlling the sizes of trees. One advantage of such prior specification is that it ensures splitting of a variable with a certain probability even though it does not exhibit a strong marginal effect. As a result, this method may enhance the probability of finding epistatic effects whose marginal effects are weak. Besides, the MCMC algorithm also has the adaptive property, where it tends to search more thoroughly in the vicinity of trees containing the interacting variables already found in previous iterations. Thus it allows the detection of multiway interactions. This desirable feature is distinct compared to other methods based on ensemble trees, like the ones using random forests (Breiman 2001; Lunetta et al. 2004; Bureau et al. 2005), in which trees are constructed independently and so are ‘memoryless’ of promising trees visited previously. As a result, potentially important interactions may be diluted in the ensemble consisting of a large number of trees, making it difficult to uncover possible multi-way interactions.

In the next section, we will provide detailed description of binary classification trees and the Bayesian treatment of model search, followed by illustrations of the approach through simulation studies and a real data example.

Materials and Methods

Binary Classification Trees

There are two types of nodes in a binary classification tree– internal nodes represented by ovals and terminal nodes represented by rectangles as shown in Figure 1. Each internal node

has an associated splitting rule and exactly two offspring called child nodes. The splitting rule uses a feature or variable, like the genotype of a SNP or the age of an individual, to assign an observation to either the left or right child nodes. The classification process starts from the top node that is called the root node. At each internal node, an observation is classified to one of the two child nodes according to its feature value and the splitting rule. After moving down along the branches of the tree, the observation finally reaches one of the terminal nodes. Therefore, all terminal nodes represent a partition of the feature space. A general principle of the partitioning process is to make individuals in a terminal node as homogeneous as possible in terms of the outcome, while different terminal nodes are heterogeneous. For instance, in the example shown in Figure 1, there are 500 cases and 500 controls and two SNPs X_1 and X_2, each taking one of the values 0, 1 and 2 corresponding to the three genotypes. The rule in the root node classifies individuals with $X_2 \in \{1, 2\}$ to the left child and all others to the right child, which is a terminal node. After this step, terminal node E contains 280 individuals who are all controls. Thus the misclassification rate at this node is 0/280. For those in the left child node, 500 are cases while the remaining 220 are controls, which are further split based on their X_1 values. The partitions take place iteratively and each individual eventually reaches one of the five terminal nodes.

Note that a classification tree can naturally represent the epistatic interaction among features. For example, terminal nodes C and D represent the interactions of $X_1 \in \{0, 1\} \cap X_2 \in \{2\}$ and $X_1 \in \{0, 1\} \cap X_2 \in \{1\}$, respectively. It shows that the effect of X_1 depends on the genotype of X_2 – individuals with $X_1 \in \{0, 1\}$ have low risk when $X_2 = 1$ but their risk is very high when $X_2 = 2$. Similarly individuals with $X_1 = 2$ belong to the high risk group only if they also carry genotype 1 or 2 in X_2; otherwise their risk is low.

Bayesian Classification Tree Search Method

Consider a case-control sample of n subjects. For individual i, y_i is a binary response taking values 0 (control) and 1 (case); and $x_i = (x_{i1}, \cdots, x_{ik})$ are genotypes of k SNPs. Let T denote a binary tree like the one shown in Figure 1. Note that T is the parameter of interest on which we will assign a prior distribution. Recall that paths from the root node down in tree T naturally represent interaction relationships among features so that inferences of epistatic interactions can be made from the posterior distribution of T. Define function $m(T)$ to be the number of terminal nodes of T, and for notation simplicity we write m in places of $m(T)$. For a given tree T, in terminal j, let t_j be the set of all individuals in j and p_j be the mean response of y. Here the p_j are model parameters, upon which we will put prior distributions and which will be integrated out, as will be shown next. Noting that the distribution of y_i for all individuals in j is i.i.d. Bernoulli, the likelihood function can be written as

$$L(y; p_1, \cdots, p_m, T) = \prod_{j=1}^{m} L_j(y; p_j, T),$$

where $L_j(y; p_j, T) = \prod_{i \in t_j} p_j^{y_i} (1 - p_j)^{1 - y_i}$.

Now we proceed to address the problem of choosing priors for parameters $(p_1, \cdots, p_m, T)$, which can be expressed in a conditional form $\pi(p_1, \cdots, p_m, T) = \pi(p_1, \cdots, p_m | T)\pi(T)$. Note that we are interested in the posterior of T and want to integrate out all p_j's. A natural choice for the conditional prior distribution of $\pi(p_1, \cdots, p_m | T)$ is, assuming conditional independence, a Beta(γ_1, γ_2) conjugate prior for p_j given T. Under this prior, the marginal likelihood is:

$$p(y|x,T)=\int_0^1\cdots\int_0^1\prod_{j=1}^{m}L_j(y;p_j,T)\mathrm{Beta}(\gamma_1,\gamma_2)dp_1\cdots dp_k$$
$$=\left(\frac{\Gamma(\gamma_1+\gamma_2)}{\Gamma(\gamma_1)\Gamma(\gamma_2)}\right)^m\prod_{j=1}^{m}\frac{\Gamma(n_{j0}+\gamma_1)\Gamma(n_{j1}+\gamma_2)}{\Gamma(n_{j0}+n_{j1}+\gamma_1+\gamma_2)}, \quad (1)$$

where n_{j0} and n_{j1} are the number of controls and cases in node j, respectively. Note that setting γ_1 and γ_2 to 1 yields the uniform prior on p_j.

Next we consider specifying prior distributions for the tree model T. In the literature several priors have been proposed including the prior derived from a stochastic tree-generating process (Chipman et al. 1998), a truncated Poisson prior on the number of terminal nodes (Denison et al. 1998), and the pinball prior (Wu et al. 2007). Here we follow Chipman et al. (1998) because it is intuitive to understand and has the advantage of easy implementation. This prior distribution, denoted by $\pi(T)$, does not have a closed-form expression; rather, drawing from $\pi(T)$ is through a stochastic tree generating process as follows. Starting from the trivial tree that has only one singleton node, a terminal node η will split with probability

$$\alpha(1+d_\eta)^{-\beta}, \quad (2)$$

where d_η is the depth of η, and α and β are hyperparameters. If it splits, to form a splitting rule, a SNP is randomly drawn from the pool of all available ones, followed by a random selection of available genotypes of that SNP to determine its left and right child nodes. Finally, we set η to the newly created left and right child nodes, and repeat the procedure recursively. Note that the left and right nodes are constructed independently. The splitting probability is α for the root node and it decreases at a rate $(1 + d_\eta)^{-\beta}$ as the tree becomes large. As a consequence, this prior penalizes unbalanced or large trees, and tree sizes are controlled by hyperparameters α and β. Also, the splitting rule ensures equal probabilities for all SNPs under consideration and thus is non-informative.

The posterior of T is $p(T|x, y) \propto p(y|x, T)\pi(T)$. Although the space of T is finite, it is infeasible to exhaustively evaluate all possible trees in the genome-wide setting. Here the Metropolis–Hastings algorithm described below can be applied to draw from the posterior distribution.

1. Set an initial tree T^0. It can be any tree and the simplest choice is the trivial singleton tree;
2. At step i, propose a candidate tree T^* from a transition function $q(T^i, T^*)$, and set $T^{i+1} = T^*$ with probability

$$\min\left(\frac{p(y|x,T^*)\pi(T^*)q(T^*,T^i)}{p(y|x,T^i)\pi(T^i)q(T^i,T^*)},1\right). \quad (3)$$

Otherwise keep the current tree, i.e., set $T^{i+1} = T^i$.

In step 2, it is important to specify the transition function q. Here we follow Chipman et al. (1998) and consider the transition $q(T^i, T^*)$ that randomly chooses one from four operations: Grow, Prune, Change and Swap. Details can be found in Chipman et al. (1998), and here we provide a brief description. In the Grow step a randomly chosen terminal node η is split into two child nodes according to the similar procedure as in the prior drawing step. The Prune

step is exactly the reverse operation of Grow in which two randomly selected sibling nodes are pruned. In the Change step one randomly picks an internal node and changes its splitting rule at random. The Swap step randomly selects a parent-child pair that are both internal nodes and swaps their splitting rules. Note that the Grow and Prune steps are counterparts of each other, as mentioned above, and the Change and Swap operations are counterparts of themselves. This feature is appealing because it results in a reversible Markov chain that will ensure the convergence to the posterior distribution. Moreover, it also can greatly simplify the evaluation of (3) (Chipman et al. 1998).

Results

Simulated Data

To evaluate the performance of the Bayesian tree model, first we conduct simulation studies. We use 12 two-locus interaction models considered by several authors (e.g. Neuman & Rice 1992; Schork et al. 1993; Knapp et al. 1994; Becker et al. 2005) plus 2 additive models used by Chen et al. (2007) in their publications for epistasis detection. We simulate 500 cases and 500 controls according to the diseases models listed in Tables 1 and 2. We assume there are two disease loci that are in linkage equilibrium, and 98 non-disease loci. Each locus is diallelic, in Hardy-Weinberg equilibrium (HWE), and unlinked to any other. The two disease loci are denoted by SNP1 with alleles A and a, and SNP2 with alleles B and b, respectively. The two-locus penetrance and relative risks (RR) are shown in Table 1. The minor allele frequencies (MAF) and marginal relative risks of the two disease loci are listed in Table 2. Note that the disease prevalence and the percentage of phenotypic variance explained by the two disease loci, shown in Table 2, are fully determined by the penetrance and MAF parameters specified in the two tables, given the model assumptions. MAFs of the 98 non-disease loci are simulated at random from a uniform distribution *Uni f* [0.05, 0.50]. Detailed information can be found in Knapp et al. (1994) and Chen et al. (2007).

For the Bayesian classification model, we test four different sets of values for the hyperparameters (α, β), namely (0.8,1), (0.8,1.8), (0.95,1) and (0.95,1.8). We draw trees from their prior distributions and plot the prior distribution of the number of terminal nodes in Figure 2. To understand the effect of hyperparameters, we use $\alpha = 0.95$ and $\beta = 1.8$ as an example. This prior specifies that any SNP can be split at the root node (the first level node) with a prior probability 0.95; but this probability decreases rapidly to 0.27 and 0.13 at the second and third level, respectively. In general, large values of α will reduce the probability of getting a singleton tree (i.e., trees with only one node), whereas large values of β will prevent a tree from growing, reducing the probability on large trees. Indeed this is what we observe from Figure 2. It is clear that the priors with $\alpha = 0.95$ generate fewer singleton trees than $\alpha = 0.8$. The priors with $\beta = 1$ have more prior weights on larger trees than $\beta = 1.8$. Note that the posterior probability of each split depends on both the prior probability and conditional effect. Thus, at the root node a SNP would have a certain posterior probability for splitting, even if its marginal effect is weak. However, at the second level, a SNP would be split only if it has a reasonably large conditional effect given the first SNP. In other words the epistatic effect of these two SNPs must be large.

For comparison, we consider an exhaustive search of all two-way interactions using PLINK (Purcell et al. 2007) with the fast epistasis option ‘fast-epistasis,’ which is known to yield very similar results to logistic regression with all pairs of SNPs, but is more computationally efficient. For each disease model we simulate 50 case-control data sets. Table 3 shows the comparison in terms of power and false positive rate (FPR) based on these 50 simulation runs. For the PLINK method, the power is defined as the proportion of the interaction between SNP1 and SNP2 being significant at 0.10 level after the Bonferroni correction for 4,950 comparisons, i.e., p value is less than 0.10/4950. The FPR is the proportion of

detecting false two-way interactions. For the Bayesian classification tree, we run MCMC with three random restarts and each has 4000 iterations. The tree with the largest posterior probability is reported. The power of the interaction is defined as the proportion of having at least one terminal node that involves spliting on both SNP1 and SNP2. Finally the FPR is defined as the proportion of having at least one terminal node that involves splitting on two or more SNPs other than SNP1 and SNP2. From the table we can see that the Bayesian tree is powerful in detecting the epistasis. The performance of different hyperparameters is quite similar, suggesting that it is not sensitive to the choice of hyperparameters in this case. On the other hand, the PLINK fast epistasis search fails to identify the epistasis in half of the 14 models, and has lower power than the Bayesian classification tree in the other half of the models. We also note that we conducted a two-stage search algorithm using logistic regression (results not reported here), in which single-locus analysis was done in the first stage to identify the top 10 most significant SNPs, and then in stage 2 we performed exhaustive search of all possible two-locus models (with two-way interactions included) involving these 10 SNPs. The power of detecting epistasis of this two-stage search approach was low. The reason is that in many cases the marginal effects are elusive and are missed in the first stage, which leads to a poor level of power in identifying the epistatic effect.

Crohn's Disease Data

Next we use a case-control data set of Crohn's disease (Duerr et al. 2006) to demonstrate the use of the Bayesian classification model. Crohn's disease is an ongoing autoimmune disorder that causes discontinuous and transmural inflammation in the digestive tract. It most commonly affects the lower part of the small intestine called the ileum. Crohn's disease has been found to have a strong genetic component (Peeters et al. 1996). For example, relatives have a 20–30 fold increased risk compared to non-relatives, and monozygotic twins have a 10–50 fold increased risk compared to dizygotic twins. The disease is believed to involve the interaction of several factors such as genetic susceptibility, the intestinal microbial flora inside the patient, the immune response to these microbiota, and triggers involving environmental factors (Sartor 2006).

Here we apply the Bayesian tree to the cohort containing 401 cases and 433 controls. For quality control, we exclude SNPs with a call rate lower than 0.99, minor allele frequency lower than 0.05, or HWE p value lower than 0.001. In addition, all subjects with a call rate less than 0.95 are removed from the analysis. Finally a total of 397 cases and 431 controls pass the quality threshold and are kept in the analysis. We first conduct single SNP association tests and select the top 5000 ones based on the p values, and apply the Bayesian classification tree to those 5000 SNPs. The main reason for choosing the top 5000 SNPs is a balance between statistical power and computational efficiency. A premise of this selection is that most interactions would involve genes with weak to moderate marginal effects, so focusing on the top ones would likely capture most interactions unless the interaction patterns are such that there is no main effect at all. In the simple case of two-way interactions, our previous analytical work (Wu & Zhao 2009) and a follow-up study (Wu & Zhao 2010, unpublished data) suggest that the two-stage analysis is among the most efficient approaches, at least in the models considered. The top 5000 SNPs contain many SNPs with weak marginal effects. Actually the p values of these 5000 SNPs are roughly uniformly distributed from 0 to 0.02. Table 4 lists the top 20 SNPs, among others, from these 5000 ones that have the smallest p values of association tests. We run the MCMC with five restarts, each of which has 50,000 iterations. The hyperparameters are set to $\alpha = 0.95$ and $\beta = 0.5$. The best singleton tree picks rs1343151 on chromosome 1 that belongs to IL23R. As can be seen from Table 4, IL23R is among the top genes in the list and has been previously confirmed to be associated with Crohn's disease (Barrett et al. 2008). The best tree having three terminal nodes involves rs2463031 on chromosome 19, ranked number 8 in Table 4,

and rs3213255 on chromosome 19. SNP rs2463031 is in the intergenic region between LOC345571 and EFNA5 while rs3213255 is in the intron region of XRCC1. An interesting case is the best tree with 4 terminal nodes, which is plotted in Figure 3. A contingency table of the first two SNPs are shown in Table 5. To examine the performance of classification error of this tree, we test epistasis by an exhaustive search of all 2-locus models, and keep the ones whose *p* values are below 0.001. Then for these kept models we calculate the misclassification rates and the histogram is shown in Figure 4. In addition, we also put the misclassification rate of the best Bayesian tree with 4 terminal nodes on the same plot. It is clear that the Bayesian method gives an error rate close to the lower bound of all models by exhaustive search. The tree shown in Figure 3 identifies a possible epistasis between rs13611 $\in \{1, 2\}$ and rs178900 $\in \{1\}$, where individuals carrying the combination of these genotypes have significantly lower risk (0.27) than others. We notice that this interaction is missed by the exhaustive logistic regression search because the *p* value of the two-way interaction in the logistic regression model is 0.15. Now we look at functional annotations of these SNPs. In the classification tree, the SNP at the root node is rs136211 that is located on chromosome 22 in the gene region of MYH9, which encodes a non-muscle myosin IIA (NM IIA) heavy chain. Recently NM IIA was found to regulate intestinal epithelial cell restitution and matrix invasion (Babbin et al. 2009). Intestinal epithelial restitution is the closure of mucosal wounds that is heavily influenced by epithelial migration. Epithelial cell migration is known to have a significant contribution to the pathophysiology of intestinal disorders like inflammatory bowel disease. The findings by Babbin et al. (2009) suggest that NM IIA promotes 2-D epithelial cell migration but antagonizes 3-D invasion. The second SNP found in the tree is rs178900 that is located in the intron region of RAB11FIP4 on chromosome 17. RAB11FIP4 is RAB11 family interacting protein 4 that plays regulatory roles in the formation, targeting, and fusion of intracellular transport vesicles (Entrez Gene). The last SNP is rs8055192 that is located on chromosome 16 but it is not in a gene region. Its functional annotation remains unclear to us at this time. The tree model suggests that there may be epistasis among these three loci.

Discussion

In this paper we have described a Bayesian classification tree model to identify the epistatic SNPs in GWAS. In Bayesian treatment for the classification tree model, there are two key components that determine the posterior model search, that is, the prior specification of all trees and the transition kernel in MCMC. With the prior in (2) derived from a tree generating process, the model allows a SNP to split with a certain posterior probability, even when the marginal effect is not significant. This feature can enhance the power of identifying interactions among genes, as demonstrated in the simulation studies. In addition, due to the adaptive property of the MCMC algorithm, the Bayesian model search also can detect higher-order interactions.

In the real data example, we find that the MCMC algorithm moves rapidly from its initial state toward regions with high posterior probabilities, and tends to make local moves thereafter. This is not very surprising because the transition function proposes trees in the local regions and can hardly move to another mode in the tree space. This finding is consistent with the original authors (Chipman et al. 1998), who proposed to run the MCMC with repeated random starts. Based on our experience this does help to find models that fit the data better. The slow convergence in MCMC may be problematic in some cases, for instance, if one wants to do model averaging or use the posterior distribution to assess the importance of all epistatic interactions. However, it is not of a major concern if our purpose is to find some potentially important interactions instead of the most important one.

We also tested genome-wide search using approximately 260,400 SNPs that pass certain quality control thresholds. The program is written in C++ and runs very fast. It took about 70 minutes to run 1,000,000 MCMC iterations on a PC with 2.5 GHz Intel Core 2 Duo CPU and 4G memory. However, due to the huge number of trees in the search space, the mixing of the Markov chain is slow. Nonetheless, it is still feasible to apply our method to GWAS data on a cluster of servers and allow it to run for a large number of iterations to be more inclusive. However, this may not be the statistically most efficient approach under most interaction models due to the substantially increased model space.

Acknowledgments

This work was supported in part by NIH grants GM 59507, U01 DK062422, 1R01DK072373, and UL1 RR024139, and NSF grant DMS-0714817.

References

Babbin BA, Koch S, Bachar M, Conti MA, Parkos CA, Adelstein RS, Nusrat A, Ivanov AI. Non-muscle myosin IIA differentially regulates intestinal epithelial cell restitution and matrix invasion. Am J Pathol. 2009; 174:436–448. [PubMed: 19147824]

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Gossum AV, Zelenika D, Franchimont D, Hugot J-P, de Vos M, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. the NIDDK IBD Genetics Consortium; the Belgian-French IBD Consortium, the Wellcome Trust Case Control Consortium. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet. 2008; 40:955–962. [PubMed: 18587394]

Becker T, Schumacher J, Cichon S, Baur MP, Knapp M. Haplotype interaction analysis of unlinked regions. Genet Epidemiol. 2005; 29:313–322. [PubMed: 16240441]

Breiman L. Random forests. Mach Learn. 2001; 45:5–32.

Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and Regression Trees. New York: Chapman and Hall; 1984.

Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P. Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol. 2005; 28:171–182. [PubMed: 15593090]

Chen X, Liu CT, Zhang M, Zhang H. A forest-based approach to identifying gene and gene–gene interactions. Proc Natl Acad Sci USA. 2007; 104:19199–19203. [PubMed: 18048322]

Chipman H, George E, McCulloch R. Bayesian CART model search. J Am Stat Assoc. 1998; 93:935–948.

Combarros O, Cortina-Borja M, Smith AD, Lehmann DJ. Epistasis in sporadic Alzheimer's disease. Neurobiol Aging. 2009; 30:1333–1349. [PubMed: 18206267]

Cook NR, Zee RYL, Ridker PM. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. Stat Med. 2004; 23:1439–1453. [PubMed: 15116352]

Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 2009; 10:392–404. [PubMed: 19434077]

Culverhouse R, Suarez BK, Lin J, Reich T. A perspective on epistasis: Limits of models displaying no main effect. Am J Hum Genet. 2002; 70:461–471. [PubMed: 11791213]

Denison DGT, Mallick BK, Smith AFM. A Bayesian CART algorithm. Biometrika. 1998; 85:363–377.

Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, Dassopoulos T, Bitton A, Yang H, Targan S, Datta LW, Kistner EO,

Schumm LP, Lee AT, Gregersen PK, Barmada MM, Rotter JI, Nicolae DL, Cho JH. A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. Science. 2006; 314:1461–1463. [PubMed: 17068223]

Knapp M, Seuchter SA, Baur MP. Two-locus disease models with two marker loci: The power of affected-sib-pair tests. Am J Hum Genet. 1994; 55:1030–1041. [PubMed: 7977340]

Lunetta KL, Hayward LB, Segal J, van Eerdewegh P. Screening large-scale association study data: Exploiting interactions using random forests. BMC Genet. 2004; 5:32.10.1186/1471-2156-5-32 [PubMed: 15588316]

Lunn DJ, Whittaker JC, Best N. A Bayesian toolkit for genetic association studies. Genet Epidemiol. 2006; 30:231–247. [PubMed: 16544290]

Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered. 2003; 56:73–82. [PubMed: 14614241]

Nelson MR, Kardia SLR, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res. 2001; 11:458–470. [PubMed: 11230170]

Neuman RJ, Rice JP. Two-locus models of disease. Genet Epidemiol. 1992; 9:347–365. [PubMed: 1427023]

Peeters M, Nevens H, Baert F, Hiele M, de Meyer A, Vlietinck R, Rutgeerts P. Familial aggregation in crohn's disease: Increased age-adjusted risk and concordance in clinical characteristics. Gastroenterology. 1996; 111:597–603. [PubMed: 8780562]

Purcell, S. Plink (v1.07). 2009. http://pngu.mgh.harvard.edu/purcell/plink/

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet. 2001; 69:138–147. [PubMed: 11404819]

Sartor RB. Mechanisms of disease: pathogenesis of crohn's disease and ulcerative colitis. Nat Clin Pract Gastroenterol Hepatol. 2006; 3:390–407. [PubMed: 16819502]

Schork NJ, Boehnke M, Terwilliger JD, Ott J. Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am J Hum Genet. 1993; 53:1127–1136. [PubMed: 8213836]

Wiltshire S, Bell JT, Groves CJ, Dina C, Hattersley AT, Frayling TM, Walker M, Hitman GA, Vaxillaire M, Farrall M, Froguel P, Mccarthy MI. Epistasis between type 2 diabetes susceptibility loci on chromosomes 1q21-25 and 10q23-26 in Northern Europeans. Ann Hum Genet. 2006; 70:726–737. [PubMed: 17044847]

Wu Y, Tjelmeland H, West M. Bayesian CART. J Comput Graph Stat. 2007; 16:44–66.

Wu Z, Zhao H. Statistical power of model selection strategies for genome-wide association studies. PLoS Genet. 2009; 5:e1000582. [PubMed: 19649321]

Zhang H, Bonney G. Use of classification trees for association studies. Genet Epidemiol. 2000; 19:323–332. [PubMed: 11108642]

Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. Nat Genet. 2007; 39:1167–1173. [PubMed: 17721534]

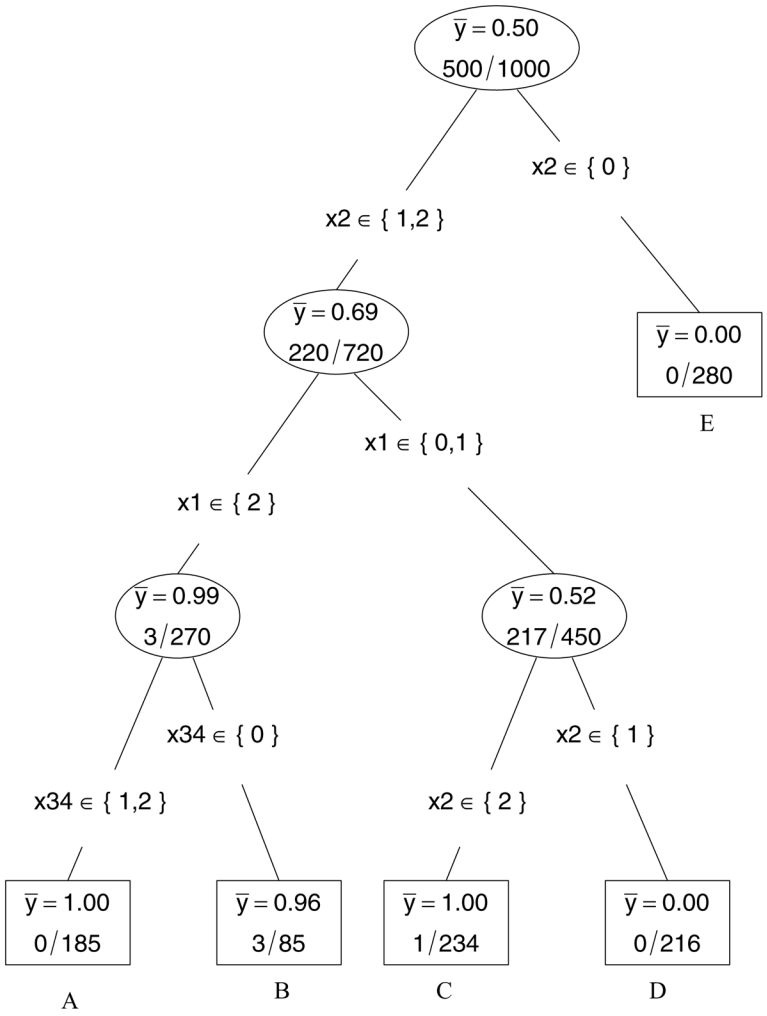

Figure 1.

An example of a classification tree: $\bar{y}$ is the proportion of cases; the fraction n_1/n is the misclassification rate where n is the total number of individuals in this node and n_1 is the number of mis-classified ones.

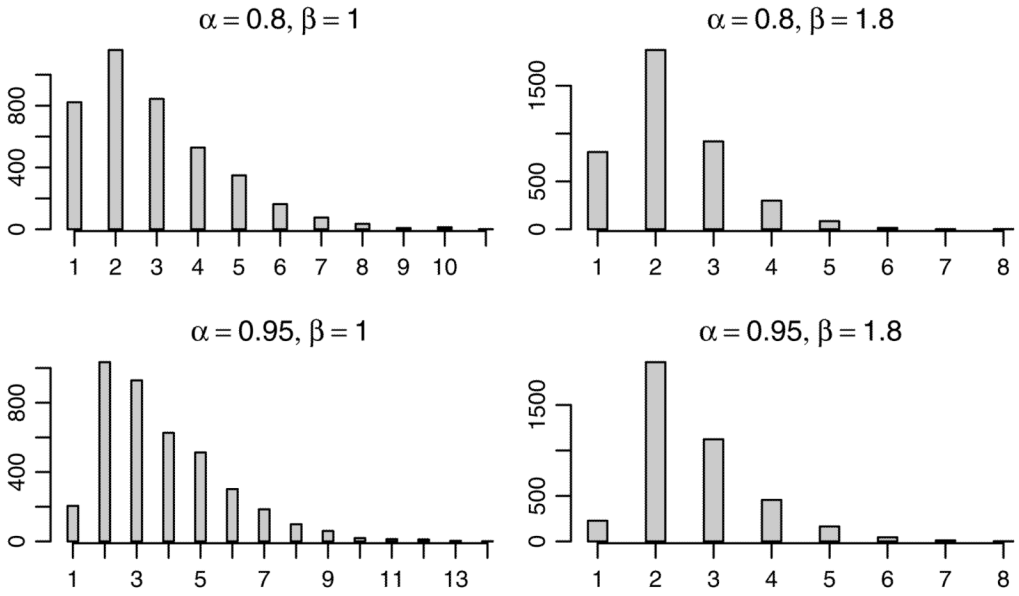

Figure 2.
Prior distribution of number of terminal nodes with various hyperparameters.

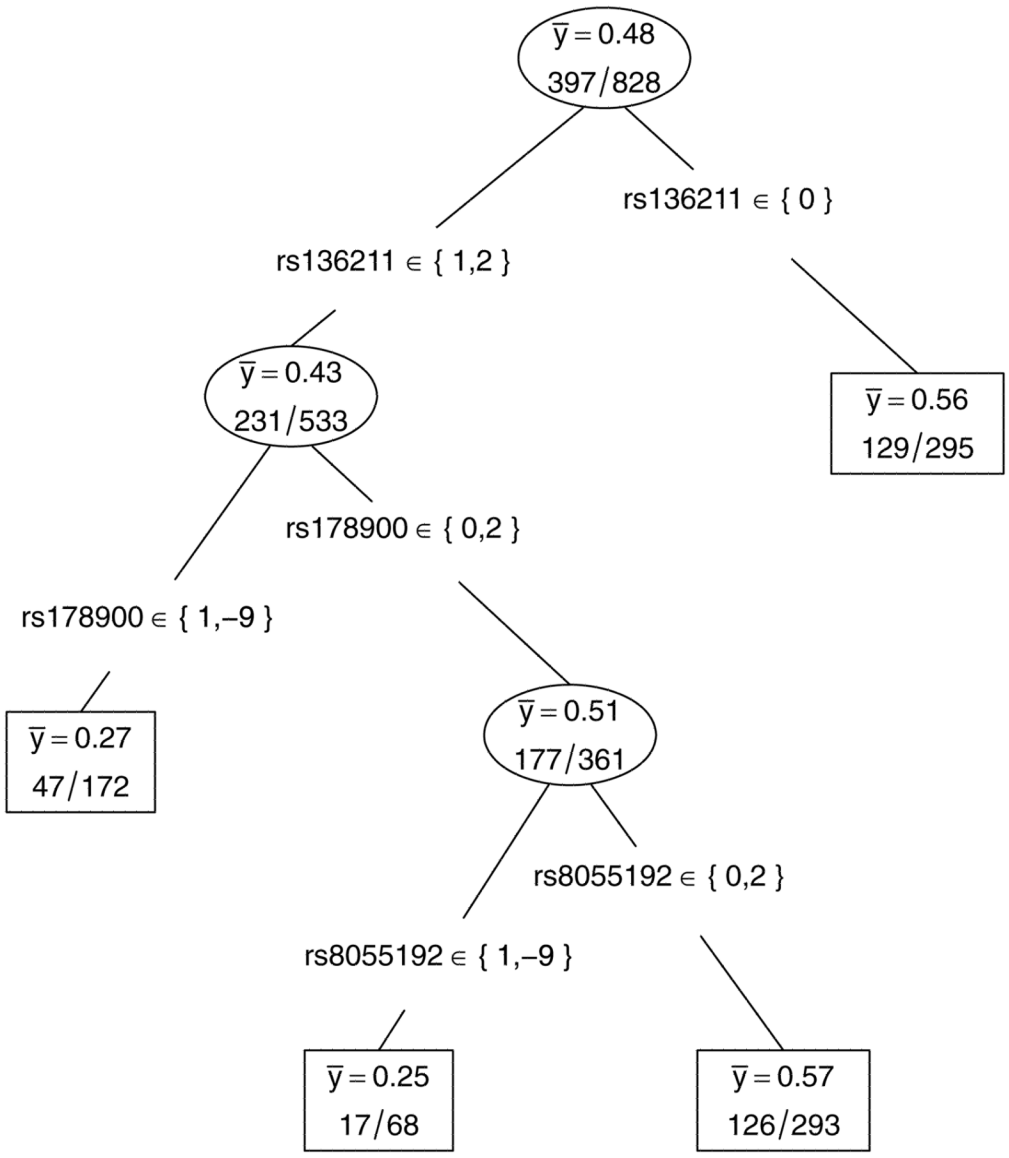

Figure 3.
The best 4-terminal-node tree for Crohn's disease data. Note: −9 denotes the missing genotype.

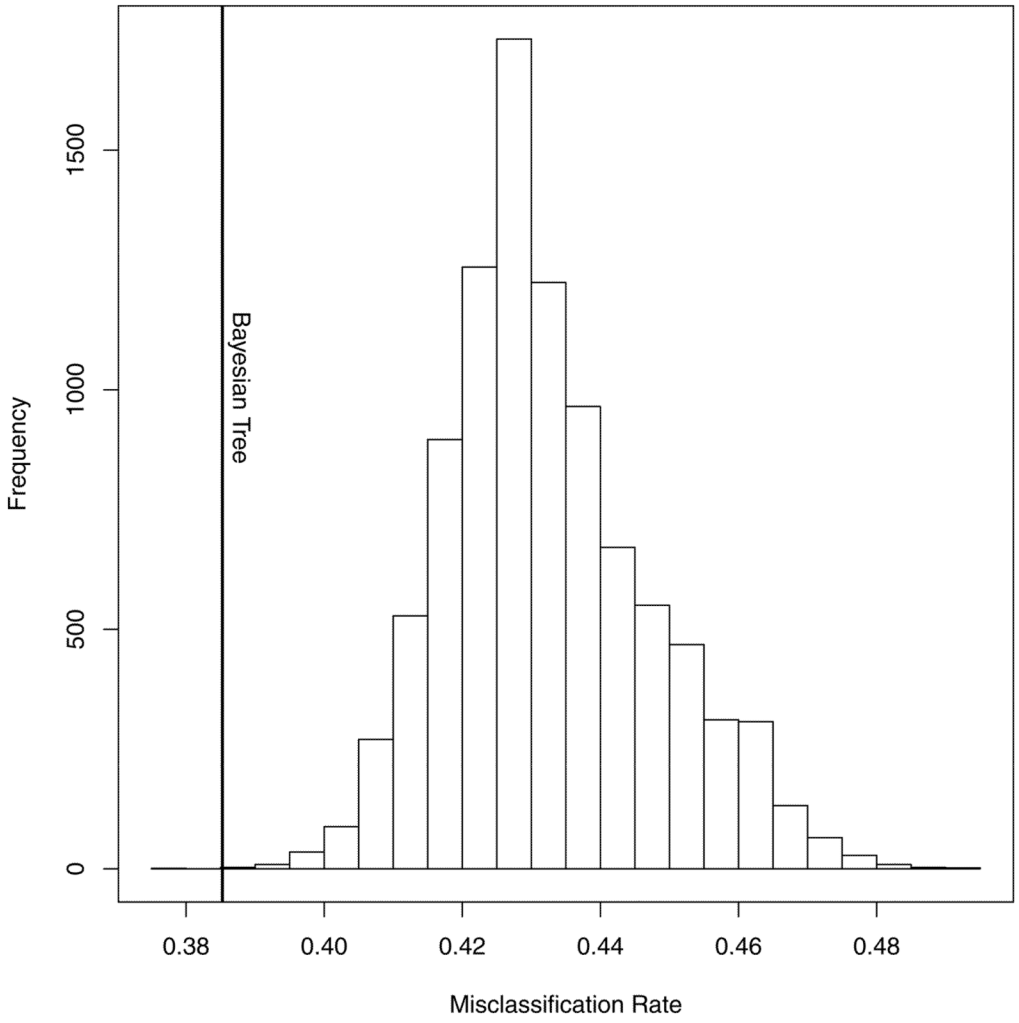

Figure 4.
Misclassification rate.

Table 1

Prevalence and odds ratio of two-locus epistatic models.

		Penetrance			RR†		
Model*		**BB**	**Bb**	**bb**	**BB**	**Bb**	**bb**
Ep-1[1,s]	AA	0	0	0	0.0	0.0	0.0
	Aa	0	0.71	0.71	0.0	7.1	7.1
	aa	0	0.71	0.71	0.0	7.1	7.1
Ep-2[1,u]	AA	0	0	0	0.0	0.0	0.0
	Aa	0	0	0	0.0	0.0	0.0
	aa	0	0.78	0.78	0.0	7.8	7.8
Ep-3[1,s]	AA	0	0	0	0.0	0.0	0.0
	Aa	0	0	0	0.0	0.0	0.0
	aa	0	0	0.9	0.0	0.0	9.0
Ep-4[1,u]	AA	0	0	0.91	0.0	0.0	9.1
	Aa	0	0	0.91	0.0	0.0	9.1
	aa	0	0.91	0.91	0.0	9.1	9.1
Ep-5[1,s]	AA	0	0	0	0.0	0.0	0.0
	Aa	0	0	0.80	0.0	0.0	8.0
	aa	0	0.80	0.80	0.0	8.0	8.0
Ep-6[1,s]	AA	0	0	1	0.0	0.0	14.3
	Aa	0	0	1	0.0	0.0	14.3
	aa	1	1	0	14.3	14.3	0.0
Het-1[2,s]	AA	0	0.50	0.50	0.0	5.0	5.0
	Aa	0.5	0.75	0.75	5.0	7.5	7.5
	aa	0.5	0.75	0.75	5.0	7.5	7.5
Het-2[2,u]	AA	0	0.66	0.66	0.0	6.6	6.6
	Aa	0	0.66	0.66	0.0	6.6	6.6
	aa	0.66	0.88	0.88	6.6	8.8	8.8
Het-3[2,s]	AA	0	0	1	0.0	0.0	13.5
	Aa	0	0	1	0.0	0.0	13.5
	aa	1	1	1	13.5	13.5	13.5

Model*		Penetrance			RR†		
		BB	Bb	bb	BB	Bb	bb
S-1^{2,s}	AA	0	0.52	0.52	0.0	5.2	5.2
	Aa	0.52	0.52	0.52	5.2	5.2	5.2
	aa	0.52	0.52	0.52	5.2	5.2	5.2
S-2^{2,u}	AA	0	0.57	0.57	0.0	5.7	5.7
	Aa	0	0.57	0.57	0.0	5.7	5.7
	aa	1	1	1	10.0	10.0	10.0
S-3^{1,s}	AA	0	0	0.51	0.0	0.0	5.1
	Aa	0	0.51	1	0.0	5.1	10.0
	aa	0.51	1	1	5.1	10.0	10.0
Ad-1^{3,u}	AA	0.01	0.01	0.01	0.1	0.1	0.1
	Aa	0.02	0.30	0.80	0.1	1.7	4.6
	aa	0.04	0.80	0.80	0.2	4.6	4.6
Ad-2^{3,u}	AA	0.05	0.05	0.05	0.2	0.2	0.2
	Aa	0.10	0.32	0.80	0.5	1.5	3.7
	aa	0.15	0.80	0.80	0.7	3.7	3.7

*1–Epistasis model; 2–Heterogeneity model; 3–Additive model; s–symmetrical; u–unsymmetrical.

†The baseline risk is the population disease prevalence listed in Table 2.

Table 2

MAF, prevalence, and odds ratio of two-locus epistatic models.

	MAF			RR of SNP1†			RR of SNP2†			
Model*	**SNP1**	**SNP2**	**Disease Prevalence**	**AA**	**Aa**	**aa**	**BB**	**Bb**	**bb**	**Phenotypic Variance Explained by 2 SNPs**
Ep-1[1,s]	0.210	0.210	0.100	0.0	2.7	2.7	0.0	2.7	2.7	67.4%
Ep-2[1,u]	0.600	0.199	0.100	0.0	0.0	2.8	0.0	2.8	2.8	75.3%
Ep-3[1,s]	0.577	0.577	0.100	0.0	0.0	3.0	0.0	0.0	3.0	88.9%
Ep-4[1,u]	0.372	0.243	0.100	0.5	0.5	3.9	0.0	1.3	9.1	90.1%
Ep-5[1,s]	0.349	0.349	0.100	0.0	1.0	4.6	0.0	1.0	4.6	77.7%
Ep-6[1,s]	0.190	0.190	0.070	0.5	0.5	13.9	0.5	0.5	13.9	100%
Het-1[2,s]	0.053	0.053	0.100	0.5	5.2	5.2	0.5	5.2	5.2	46.1%
Het-2[2,u]	0.279	0.040	0.100	0.5	0.5	6.7	0.5	6.7	6.7	63.5%
Het-3[2,s]	0.194	0.194	0.074	0.5	0.5	13.5	0.5	0.5	13.5	100%
S-1[2,s]	0.052	0.052	0.100	0.5	5.2	5.2	0.5	5.2	5.2	46.9%
S-2[2,u]	0.228	0.045	0.100	0.5	0.5	10.0	0.5	6.0	6.0	77.3%
S-3[1,s]	0.194	0.194	0.100	0.2	2.0	6.8	0.2	2.0	6.8	59.3%
Ad-1[3,u]	0.349	0.349	0.173	0.1	1.4	2.8	0.1	1.4	2.7	48.5%
Ad-2[3,u]	0.349	0.349	0.215	0.2	1.3	2.4	0.4	1.2	2.2	35.2%

*1–Epistasis model; 2–Heterogeneity model; 3–Additive model; s–symmetrical; u–unsymmetrical.

†The baseline risk is the population disease prevalence listed in column 4.

Table 3

Power and FPR comparison of detecting epistasis.

	Bayesian classification tree with various hyperparameters								Plink	
	α = 0.80, β= 1.0		α = 0.80, β= 1.8		α = 0.95, β= 1.0		α = 0.95, β= 1.8		Fast Epistasis	
Model	**Power**	**FPR**	**Power**	**FPR**	**Power**	**FPR**	**Power**	**FPR**	**Power**	**FPR**
Ep-1	0.96	0.02	0.92	0.08	0.92	0.08	0.98	0.14	0.32	0.06
Ep-2	0.98	0.10	0.94	0.12	0.98	0.12	0.94	0.06	0	0.12
Ep-3	1	0.10	0.98	0.14	0.98	0.14	0.98	0.14	0	0.06
Ep-4	0.98	0.04	1	0.04	1	0.06	1	0.08	0	0.14
Ep-5	1	0.04	1	0.04	1	0.06	1	0.04	0	0.04
Ep-6	0.98	0.04	1	0.06	0.98	0.02	1	0	0	0.04
Het-1	0.9	0.06	0.94	0.10	0.94	0.10	0.98	0.08	0.12	0.22
Het-2	1	0.12	1	0.08	0.98	0.20	1	0.06	0.36	0.12
Het-3	0.98	0.02	1	0.04	0.96	0.08	0.98	0.06	0	0
S-1	0.98	0.08	0.96	0.10	0.92	0.12	0.86	0.06	0.58	0.06
S-2	1	0.02	0.98	0.10	0.96	0.06	0.96	0.10	0.62	0.06
S-3	1	0.10	1	0.16	1	0.14	1	0.16	0	0.08
Ad-1	0.98	0.26	0.98	0.10	1	0.22	0.98	0.08	0.10	0.12
Ad-2	0.98	0.14	0.96	0.12	0.98	0.08	0.98	0.16	0.40	0.10

Table 4

Top SNPs by single-locus analysis.

	SNP*	CHR	Unadj. P	Gene	Description
1	rs7517847	1	7.74E-07	IL23R	interleukin 23 receptor
2	rs1343151[1]	1	1.31E-06	IL23R	interleukin 23 receptor
3	rs7302601	12	1.73E-06		
4	rs2076756	16	3.06E-06	NOD2	nucleotide-binding oligomerization domain containing 2
5	rs10489629	1	3.32E-06	IL23R	interleukin 23 receptor
6	rs9315762	13	5.89E-06		
7	rs933534	17	9.88E-06	MSI2	musashi homolog 2 (Drosophila)
8	rs2463031[2]	5	1.49E-05		
9	rs6538370	12	1.53E-05		
10	rs925530	12	1.61E-05	LOC144404	hypothetical LOC144404
11	rs7135617	12	1.73E-05	TMEM142A	transmembrane protein 142A
12	rs10889677	1	1.85E-05	IL23R	interleukin 23 receptor
13	rs12320939	12	2.21E-05		
14	rs3934658	12	2.69E-05		
15	rs4820972	22	2.71E-05	EIF4ENIF1	eukaryotic translation initiation factor 4E nuclear import factor 1
16	rs2201841	1	2.94E-05	IL23R	interleukin 23 receptor
17	rs1028863	9	3.02E-05		
18	rs7398558	12	3.13E-05		
19	rs4760516	12	3.14E-05		
20	rs2066843	16	3.49E-05	NOD2	nucleotide-binding oligomerization domain containing 2
⋮	⋮	⋮	⋮	⋮	⋮
90	rs178900[3]	17	0.000320	RAB11FIP4	RAB11 family interacting protein 4 (class II)
153	rs3213255[2]	19	0.000537	XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1
1593	rs136211[3]	22	0.005566	MYH9	myosin, heavy chain 9, non-muscle
2370	rs8055192[3]	16	0.008312		

*1: Best 2-terminal-node tree; 2: Best 3-terminal-node tree; 3: Best 4-terminal-node tree.

Table 5

Contingency table of rs136211 and 178900.

rs136211	rs178900				Total
	0	1	2	−9 (Missing)	
0	*0.58*	*0.54*	*0.54*	*1.00*	*0.56*
	71/98(169)	52/60(112)	6/7(13)	0/1(1)	129/166(295)
1	*0.51*	*0.27*	*0.38*	*0.50*	*0.42*
	121/124(245)	89/33(122)	15/9(24)	1/1(2)	226/167(393)
2	*0.55*	*0.27*	*0.75*		*0.46*
	40/48(88)	35/13(48)	1/3(4)	0/0(0)	76/64(140)
Total	*0.54*	*0.38*	*0.46*	*0.67*	*0.48*
	232/270(502)	176/106(282)	22/19(41)	1/2(3)	431/397(828)

In each cell the top row is the proportion of cases; the numbers on the bottom row represent the numbers of controls, cases, and total observations, respectively.