Genetic Epidemiology 28: 171–182 (2005)

Identifying SNPs Predictive of Phenotype Using Random Forests

Alexandre Bureau,[1,2*] Josée Dupuis,[3] Kathleen Falls,[1,4] Kathryn L. Lunetta,[1,3] Brooke Hayward,[1] Tim P. Keith,[1,5] and Paul Van Eerdewegh[1,5,6]

[1]*Department of Human Genetics, Oscient Pharmaceuticals, Waltham, Massachusetts*
[2]*School of Health Sciences, University of Lethbridge, Lethbridge, Alberta, Canada*
[3]*Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts*
[4]*Department of Neurology, Boston University School of Medicine, Boston, Massachusetts*
[5]*Galileo Genomics Inc., Saint-Laurent, Québec, Canada*
[6]*Department of Psychiatry, Harvard Medical School, Boston, Massachusetts*

There has been a great interest and a few successes in the identification of complex disease susceptibility genes in recent years. Association studies, where a large number of single-nucleotide polymorphisms (SNPs) are typed in a sample of cases and controls to determine which genes are associated with a specific disease, provide a powerful approach for complex disease gene mapping. Genes of interest in those studies may contain large numbers of SNPs that classical statistical methods cannot handle simultaneously without requiring prohibitively large sample sizes. By contrast, high-dimensional nonparametric methods thrive on large numbers of predictors. This work explores the application of one such method, random forests, to the problem of identifying SNPs predictive of the phenotype in the case-control study design. A random forest is a collection of classification trees grown on bootstrap samples of observations, using a random subset of predictors to define the best split at each node. The observations left out of the bootstrap samples are used to estimate prediction error. The importance of a predictor is quantified by the increase in misclassification occurring when the values of the predictor are randomly permuted. We extend the concept of importance to pairs of predictors, to capture joint effects, and we explore the behavior of importance measures over a range of two-locus disease models in the presence of a varying number of SNPs unassociated with the phenotype. We illustrate the application of random forests with a data set of asthma cases and unaffected controls genotyped at 42 SNPs in ADAM33, a previously identified asthma susceptibility gene. SNPs and SNP pairs highly associated with asthma tend to have the highest importance index value, but predictive importance and association do not always coincide. *Genet. Epidemiol.* 28:171–182, 2005. © 2004 Wiley-Liss, Inc.

Key words: genotype-phenotype association; predictive importance; classification trees; case-control study

*Correspondence to: Alexandre Bureau, Ph.D., Département de Médecine Sociale et Préventive, Pavillon de l'Est, Université Laval, 2180 Chemin Sainte-Foy, Quebec City, Quebec G1K 7P4, Canada. E-mail: alexandre.bureau@msp.ulaval.ca
Received 3 March 2004; Accepted 15 July 2004
Published online 8 December 2004 in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20041

INTRODUCTION

Until recently, the identification of disease genes had been restricted to Mendelian traits. However, new technologies and more comprehensive genome resources are making the identification of complex disease genes possible [Glazier et al., 2002]. Genetic association studies are a powerful and widely used approach for complex disease gene mapping. The case-control design, where a large number of cases and ethnicity-matched controls are collected, is widely used due to its efficiency [Morton and Collins, 1998]. Nevertheless, the identification of polymorphisms associated with a disease among a large number of SNPs can be daunting. When applying classical statistical methods, such as logistic regression, one faces the "curse of dimensionality:" the model becomes unstable as more SNP main effects and interaction terms are added, in the sense that the variance of the parameter estimates becomes excessively large. By contrast, high-dimensional nonparametric predictive models thrive on large numbers of predictors. Comparisons of such models are based on prediction accuracy instead of model fit.

There is growing interest in applying predictive models to the analysis of genetic association studies. The explicit representation of interaction effects in classification and regression trees (CART) [Breiman, 1984] has motivated their use in recent years to model interaction between selected genes [Cook et al., 2004; Pociot et al.,

2004] and also between genes and environmental factors [Horng et al., 2004; Kim et al., 2004]. These authors compared CART to other predictive models on the basis of prediction error. Trees were also applied to identify, among large numbers of genetic polymorphisms, those associated with a phenotype [Pociot et al., 2004; Zhang and Bonney, 2000]. The interest in predictive models has led to the development of methods designed for genetic data, such as multifactor-dimensionality reduction [Ritchie et al., 2001] and extensions of classification and regression trees to take advantage of the structure of genetic data, such as using inferred haplotypes as predictors in the program Helix Tree [Golden Helix, 2002].

Random forests [Breiman, 2001] are a type of high-dimensional nonparametric predictive model consisting of a collection of classification or regression trees. The main attractive features of random forests are their top performance for prediction, built-in estimates of prediction error, and measures of predictive importance of variables to discover which ones are most predictive of the response. In the context of linkage mapping of quantitative traits on data simulated for Genetic Analysis Workshop 13 [Bureau et al., 2003], we built on the regression method of Haseman and Elston [1972] for mapping quantitative trait loci and used the estimated identity-by-descent at 400 markers along the genome to predict the absolute sibling pair trait difference using random forests regression. The measure of predictive importance successfully identified some of the quantitative trait loci [Bureau et al., 2003]. In the present work, we investigate the application of random forests of classification trees in the context of a case-control association study between genotype and phenotype. Random forests were recently applied in this context for the purpose of classification of individuals [Schwender et al., 2004]. In contrast, our primary motivation for applying random forests is discovering which polymorphisms or markers are predictive of a phenotype, and hence likely to affect disease susceptibility, and we focus on measures of predictive importance.

Below, we review the random forests method, focusing on measures of the predictive importance of variables. Then we extend a predictive importance index to pairs of variables and explore its behavior over a range of two-locus models. We then illustrate the application of random forests to case-control genetic association studies on a data set of SNP genotypes from a study of the association between asthma and the ADAM33 gene.

RANDOM FORESTS

A random forest is a collection of trees with variations in structure generated using two modifications to the deterministic tree-growing algorithm [Breiman, 2001]. First, the best split at each node is selected from among a random subset of the predictor variables. Second, the training set used to grow each tree is a bootstrap sample of the observations, i.e., a sample of size N drawn with replacement from the original sample of N observations. Some observations are represented multiple times, while others are left out. The left-out observations are called "out-of-bag" and are used to estimate prediction error. The trees in a random forest are grown to their full extent, i.e., they are not pruned. Breiman [2001] showed that the prediction error converges to a limiting value as the number of trees tends to infinity. Different variables are used at each split in different trees. To predict the class of an observation, the observation is assigned to a terminal node, or leaf, based on its predictor values. The class of the majority of training set observations in the leaf is selected as the class prediction for the observation. In the current implementation [Breiman and Cutler, 2003], when two classes are tied for most frequent in a leaf, the class with the lowest label is selected. With a forest of classification trees, each tree gets one vote for each out-of-bag observation, and for a given observation the class receiving the most votes is the forest prediction. Again, ties are resolved by selecting the class with the lowest label. The probability of a tie is very small with large numbers of trees. The random forests prediction for an observation is computed by averaging the tree predictions over trees for which the given observation was out-of-bag.

The primary goal of a random forest analysis in the context of genetic association studies is to identify SNPs that may increase or decrease susceptibility to a disease. This can be achieved by quantifying how much each SNP contributes to the predictive accuracy of a random forest by measuring its predictive importance. Finding that an SNP helps differentiate between cases and controls is an indication that the SNP either contributes to the phenotype or is in linkage disequilibrium with SNPs contributing to the phenotype. We describe measures of predictive importance in the context of a categorical response, such as the case or control status of individuals in a genetic study. For individual i, let

$\mathbf{X}_i$ represent the vector of predictor variable values, y_i represent its true class, $V_j(\mathbf{X}_i)$ represent the vote of tree j, and t_{ij} represent an indicator taking value 1 when individual i is out-of-bag for tree j and 0 otherwise. Let

$$T_i = \sum_{j=1}^{T} t_{ij}$$

be the number of trees for which individual i is out-of-bag. The margin of votes $mg(\mathbf{X}_i, y_i)$ is the difference between the proportion of votes for the true class and the largest proportion of votes among the other classes for a given individual. With only two classes, such as cases and controls, the margin becomes the difference between the proportion of votes for the true class and the proportion of votes for the wrong class. Letting $1(V_j(\mathbf{X}_i)=y_i)$ denote the indicator function taking value 1 when $V_j(\mathbf{X}_i)=y_i$ and 0 otherwise, the margin can be written:

$$mg(\mathbf{X}_i, y_i) = \frac{1}{T_i}\sum_{j=1}^{T} 1(V_j(\mathbf{X}_i) = y_i)t_{ij} - \max_{k\neq y_i}\left\{\frac{1}{T_i}\sum_{j=1}^{T} 1(V_j(\mathbf{X}_i) = k)t_{ij}\right\}.$$

With only two classes, 0 and 1, the margin simplifies to:

$$\begin{aligned} mg(\mathbf{X}_i, y_i) &= \frac{1}{T_i}\sum_{j=1}^{T} 1(V_j(\mathbf{X}_i) = y_i)t_{ij} - \frac{1}{T_i}\sum_{j=1}^{T} 1(V_j(\mathbf{X}_i) = 1 - y_i)t_{ij} \\ &= \frac{2}{T_i}\sum_{j=1}^{T} 1(V_j(\mathbf{X}_i) = y_i)t_{ij} - 1. \end{aligned}$$

The margin represents the level of ''confidence'' of the forest prediction. When most trees vote for the true class of an individual and the margin is close to 1, the pattern of predictor values for that individual unambiguously matches that of other individuals in the true class. When a large proportion of trees votes for another class and the margin is just above 0 or even negative, the pattern of predictor values has only weak similarity with other individuals in the same class and may point to another class. Now, consider randomly permuting the values of a predictor variable such as a SNP genotype among the individuals excluded from the bootstrap sample, such that the variable becomes independent of the response. If the variable is predictive of the response, it will be present in a large proportion of trees and near the root of those trees. A large proportion of out-of-bag individuals whose genotype has changed will be directed to the wrong side of the tree. The margin is then expected to decrease compared to what it was with the original variable values. Conversely, if the variable is not related to the response, it will be present in few trees and, when present, will be near the leaves. Few, if any, individuals whose genotype has changed will be affected. The margin is then expected to change little. The decrease in margin is therefore a measure of the predictive importance of the variable. Let $\mathbf{X}^{(\mathbf{A},j)} = (\mathbf{X}_1^{(\mathbf{A},j)}, \ldots, \mathbf{X}_N^{(\mathbf{A},j)})$ represent the vector of predictor variables with the value of variable **A** randomly permuted among the out-of-bag individuals for tree j, and $\mathbf{X}^{(\mathbf{A})}$ the collection of $\mathbf{X}^{(\mathbf{A},j)}$ for all trees, where N is the total number of individuals in the sample. The importance index I_M for variable **A** is defined as the average difference in margin between the original predictor vector and the predictor vector with random values of variable **A** [Breiman and Cutler, 2003]:

$$I_M(\mathbf{A}) = \frac{1}{N}\sum_{i=1}^{N}\left[mg(\mathbf{X}_i, y_i) - mg(\mathbf{X}_i^{(\mathbf{A})}, y_i)\right].$$

In the special case of two classes, I_M simplifies to:

$$I_M(\mathbf{A}) = \frac{1}{N}\sum_{i=1}^{N}\frac{2}{T_i}\sum_{j=1}^{T}\left[1(V_j(\mathbf{X}_i) = y_i) - 1(V_j(\mathbf{X}_i^{(\mathbf{A},j)}) = y_i)\right]t_{ij}.$$

The importance index I_M defined above is implemented in version 4 of the Random Forests software [Breiman and Cutler, 2003].

JOINT IMPORTANCE OF PAIRS OF VARIABLES

Importance indices capture the contribution of a variable to the prediction of the response in the presence of all other variables in the model. However, they are computed for individual variables and do not reveal the interactions between variables. We extend the notion of variable importance to multiple variables to capture their joint effects. We propose to measure the joint importance of multiple variables by

permuting the values of the variables jointly among the out-of-bag individuals. The difference in margin I_M for a pair of variables **A** and **B** is:

$$I_M(\mathbf{A},\mathbf{B}) = \frac{1}{N}\sum_{i=1}^{N}\left[mg(\mathbf{X}_i, y_i) - mg(\mathbf{X}_i^{(\mathbf{A},\mathbf{B})}, y_i)\right].$$

In the special case of two classes, we again have the simplification:

$$I_M(\mathbf{A},\mathbf{B}) = \frac{1}{N}\sum_{i=1}^{N}\frac{2}{T_i}\sum_{j=1}^{T}\Big[1(V_j(\mathbf{X}_i) = y_i) - 1(V_j(\mathbf{X}_i^{(\mathbf{A},\mathbf{B},j)}) = y_i)\Big]t_{ij}.$$

We have implemented the computation of I_M for pairs of predictors in the Random Forests software. Next we explore the properties of the joint importance of pairs of predictor variables through a simulation study and an application to an asthma case-control data set.

SIMULATION STUDY

We investigate the behavior of the importance index I_M for individual SNPs and SNP pairs in case-control studies empirically. We focus on genetic models where two SNPs are involved in the etiology of the disease, in order to understand the relationship between I_M for individual SNPs and SNP pairs in the simplest multilocus context. We simulate two risk SNPs **A** and **B** with dominant risk alleles *A* and *B*. Each SNP has a high- and low-risk genotype subset denoted, for SNP **A**, H_A={*AA*, *Aa*}, and L_A={*aa*}, respectively. We consider models with the two SNPs contributing an equal risk difference. Thus, the penetrances are a function of the number of high-risk genotypes, and we denote these penetrances as f_{HH}, f_{HL}, and f_{LL}, without referring to the specific loci. We explore additive and epistatic two-locus models. Let *d* be the risk difference $f_{HH}-f_{LL}$. In the additive model, the risk difference associated with a risk genotype is $d/2$, i.e., $f_{HL}-f_{LL}=f_{HH}-f_{HL}=d/2$. In the epistatic models, the risk is elevated by the value *d* only in the H_AH_B group (that is $f_{HL}=f_{LL}$). We consider values of *d* ranging from 0.1–0.3. The phenocopy rate f_{LL} is selected to achieve a specified disease prevalence K=0.10. For the model with d=0.2, a prevalence K=0.05 is also considered.

We simulate additional SNPs in linkage equilibrium with SNPs **A** and **B**, and unassociated with disease (''noise'' SNPs), denoted collectively by U. Noise SNPs are generated in sets of 8 with the frequency of one of the homozygous genotypes ranging from 0.05–0.45 by increments of 0.05, and added to data sets comprising only SNPs **A** and **B**. Genotypes at SNPs **A** and **B** are sampled from the population genotype frequencies under the assumed model. We simulate 400 samples of 100 cases and 100 controls for every setting considered. Random forests of 5,000 trees are generated for each data set, a number we found to be sufficient for convergence of importance indices. We increase with the number of noise SNPs the size of the subset of predictors from which the best split at each node is selected, as suggested by Breiman and Cutler [2003]. In addition, the association between genotypes at each SNP and case-control status was tested using Fisher's exact test.

RESULTS

We examine how the importance indices I_M(**A**) and I_M(**A**,**B**) at the risk SNPs behave as the number of noise SNPs included in the analysis increases, for varying risk genotype frequency. The ratio of I_M(**A**) to the maximum I_M for the noise SNPs ($\max_U I_M(U)$) gives an indication of the ability to select the true risk SNPs from among the noise SNPs. Figures 1 and 2 present results for additive and epistatic models where d=0.2 and K=0.10. We plot the mean values of I_M(**A**) and I_M(**A**,**B**), and the median ratio of I_M(**A**) to $\max_U I_M(U)$ and I_M(**A**,**B**) to $\max_{U1,U2} I_M(U_1,U_2)$, for individual noise SNPs and noise SNP pairs, respectively, and the proportion of replicates where I_M(**A**) and I_M(**A**,**B**) exceed the maximum of the noise. The plots extend over the range of frequencies of H_A that are consistent with the additive model. As expected, I_M(**B**)=I_M(**A**) for these simulations (not shown). Figures 1a,b and 2a,b show an attenuation of the signal as the number of noise SNPs increases. However, the ratios of I_M(**A**) to $\max_U I_M(U)$ and I_M(**A**,**B**) to $\max_{U1,U2} I_M(U_1,U_2)$ decrease much more slowly, despite the fact that the maximum of the noise would be expected to increase with increasing number of noise SNPs (Figs. 1c,d, 2c,d). This implies that I_M decreases with the number of SNPs in the analysis both for risk SNPs and noise SNPs. The proportion of replicates where I_M at the risk SNPs exceeds the maximum of the noise also decreases slowly with the number of noise SNPs (Figs. 1e,f, 2e,f).

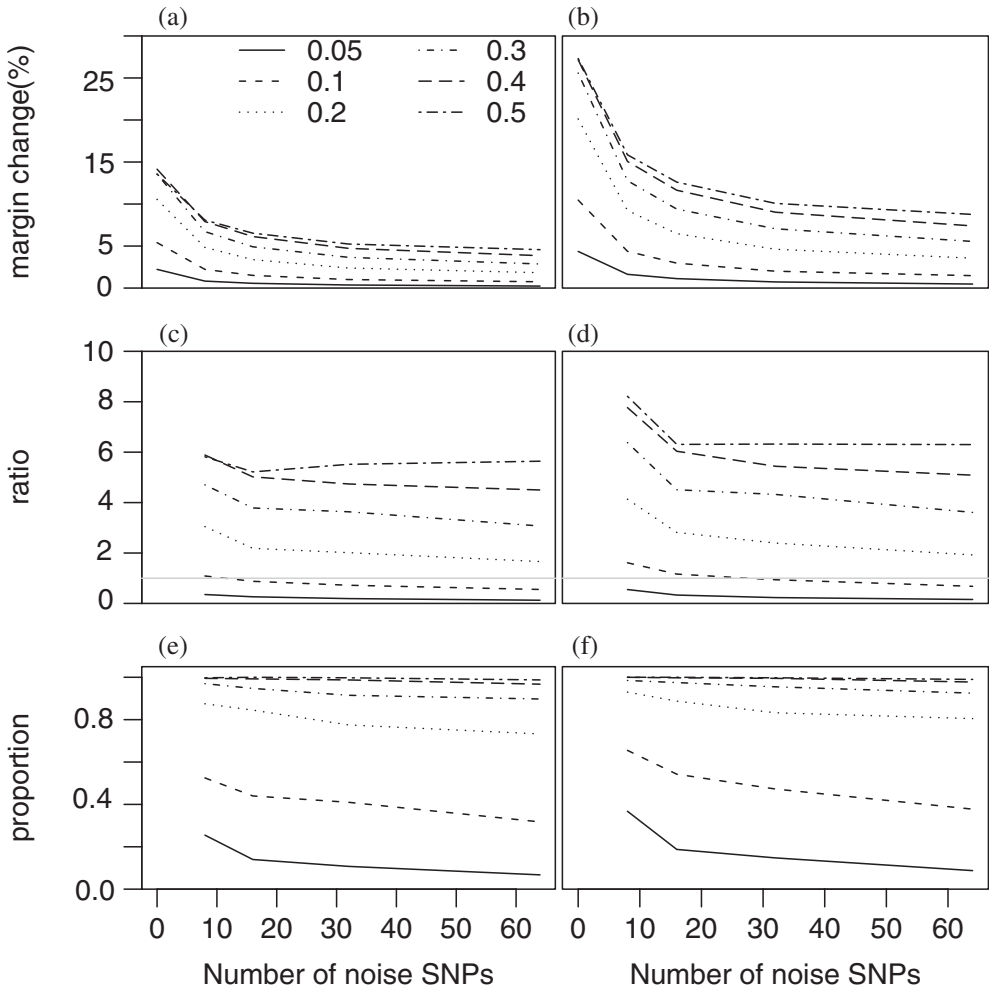

Fig. 1. Mean SNP importance under additive models. Top row: Mean values of (a) I_M(A) and (b) I_M(A,B). Middle row: Median ratio (c) of I_M(A) to maximum of I_M for individual noise SNPs $\max_U I_M(U)$ and median ratio (d) of I_M(A,B) to maximum of I_M for noise SNP pairs $\max_{U1,U2} I_M(U_1,U_2)$. Bottom row: Proportion of replicates where (e) I_M(A) is greater than $\max_U I_M(U)$ and (f) I_M(A,B) is greater than $\max_{U1,U2} I_M(U_1,U_2)$. Each line represents a different frequency of risk genotype, P[H_A].

The indices $I_M(\mathbf{A})$ and $I_M(\mathbf{A},\mathbf{B})$ increase with the risk genotype frequency from 0–0.5. Both $I_M(\mathbf{A})$ and $I_M(\mathbf{A},\mathbf{B})$ are frequently inferior to the maximum of the noise when the risk genotype frequencies are below 0.1 for the additive model, and 0.3 for the epistatic model.

The ratio of $I_M(\mathbf{A},\mathbf{B})$ to $\max_{U1,U2} I_M(U_1,U_2)$ for noise SNP pairs tends to be larger than the ratio of $I_M(\mathbf{A})$ to $\max_U I_M(U)$ for individual noise SNPs, suggesting a gain in signal-to-noise ratio by assessing the joint importance of the two-risk SNPs together compared to each one separately. The gain is particularly noticeable at high-risk genotype frequencies. The proportion of replicates where $I_M(\mathbf{A},\mathbf{B})$ is above the noise is also higher than the proportion of replicates where $I_M(\mathbf{A})$ is above the noise.

The proportion of replicates where SNP **A** has a higher importance than all noise SNPs is compared to the number of replicates where SNP **A** has the lowest *P*-value to assess the performance of the two methods (Fig. 3). While Fisher's exact test assigns the first rank to risk SNP **A** more often than the index I_M for low frequency of the high-risk genotype, the index I_M is superior to Fisher's exact test at high frequency of the high-risk genotype. The proportion of success of the index I_M decreases at a similar or slightly slower rate than that of Fisher's exact test as the number of noise SNPs increases. In particular, the proportion of success of the index I_M remains approximately stable with increasing noise when the risk genotype frequency is high, while it decreases for Fisher's exact test.

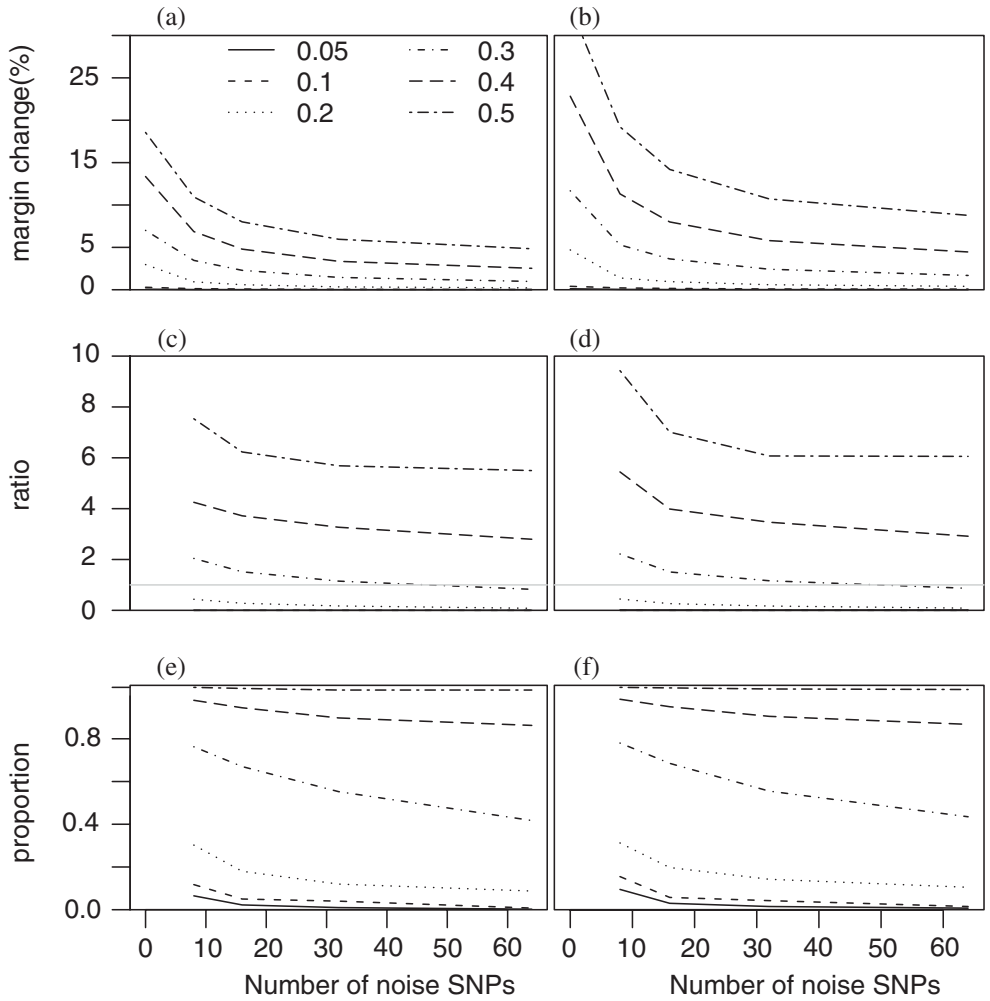

Fig. 2. Mean SNP importance under epistatic models. Plots as in Figure 1. Lines for P[H_A]=0.05 and 0.1 in c and d are too close to 0 to be visible.

The index I_M and Fisher's exact test did not show a strong tendency to assign the first rank to the risk SNP **A** for the same data sets. When we consider indicator variables equal to 1 if SNP **A** is ranked first and 0 otherwise, the coefficient of concordance kappa [Fleiss et al., 2003, Chapter 18] between those indicator variables for the index I_M and Fisher's exact test ranged from near 0 to 0.5, where 0 represents independence and 1 perfect concordance. The higher levels of agreement were observed with 64 noise SNPs.

We also investigated the effect of varying the size of the risk difference d. The values of I_M(**A**) and I_M(**A**,**B**) and the ratio of those indices to the maximum of the noise behave in a predictable fashion, becoming lower when d is set to 0.1 and higher when d is set to 0.3 (not shown). When we lower the prevalence of the disease to 0.05 while keeping the risk difference d at 0.2, the proportion of affected individuals with the low-risk genotype decreases considerably. This results in higher values of I_M(**A**), I_M(**A**,**B**), and signal-to-noise ratios. The proportions of replicates where I_M and Fisher's exact test P-value assign the first rank to risk SNP **A** increased or decreased in the same direction as I_M(**A**) in response to changes in the risk difference d and disease prevalence. In addition, the relative performances of I_M and Fisher's exact test remain similar, with I_M doing better at high frequency of the high-risk genotype and Fisher's exact test ahead at low frequency of the high-risk genotype.

We also consider the difference I_M(**A**,**B**) $-I_M$(**A**)$-I_M$(**B**). That difference is in general small compared to I_M(**A**) and I_M(**A**,**B**) (Fig. 4). For the models considered, I_M(**A**,**B**)$-I_M$(**A**)$-I_M$(**B**) tends

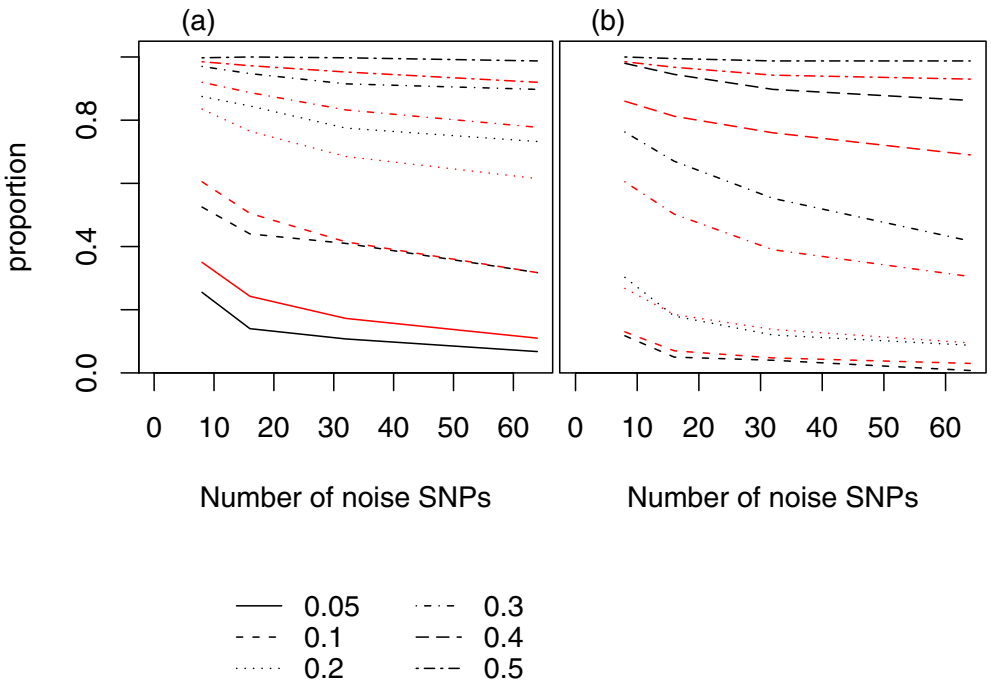

Fig. 3. Proportion of replicates where I_M(A) is greater than $\max_U I_M$(U) (lines without circles (black)) and where *P*-value of Fisher's exact test *P*(A) is lower than $\min_U$ *P*(U) (lines with circles (red)) for (a) additive models and (b) epistatic models. Each line represents different frequency of risk genotype, P[H_A]. Results for some risk genotype frequencies are not shown on one of two plots, to prevent overlapping lines. [Color figure can be viewed in the journal's online edition: www.interscience.wiley.com]

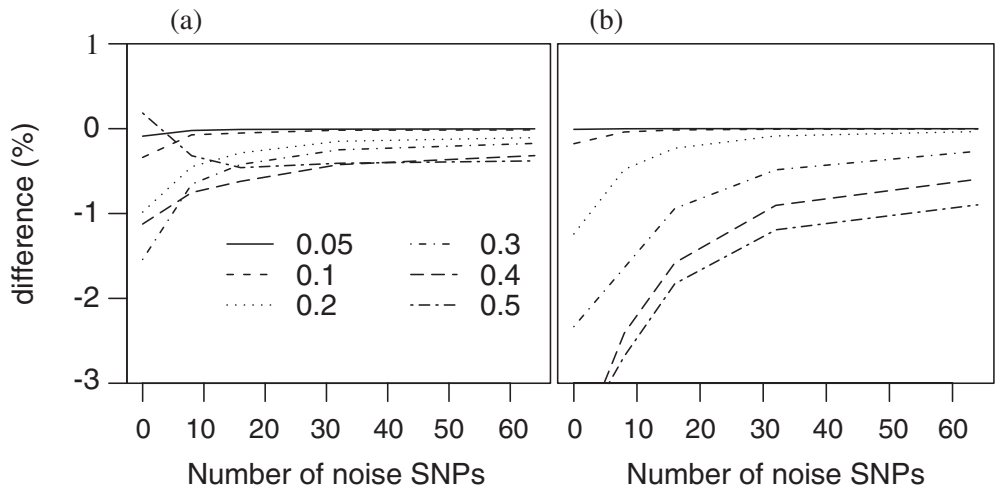

Fig. 4. Mean values of I_M(A,*B*)−I_M(A)−I_M(B) for (a) additive models and (b) epistatic models. Each line represents different frequency of risk genotype, P[H_A].

to be negative, and is lower under the epistatic model than under the additive model. The mean difference I_M(**A**,**B**)−I_M(**A**)−I_M(**B**) tends to become closer to 0 as the number of noise SNPs increases, except under the additive model when the genotype frequency is high, where it reaches positive values in the absence of noise SNPs before decreasing for 8 and 16 noise SNPs. In other simulations with only SNPs **A** and **B**, where the risk genotype frequencies at the two SNPs differed or where the risk alleles were positively associated in the population, we observed values of I_M(**A**,**B**)−I_M(**A**)−I_M(**B**) between −4% and 3% (not shown), indicating that I_M can be super-additive as well as subadditive. However, I_M(**A**,**B**) −I_M(**A**)−I_M(**B**) remained in most cases lower for the epistatic model than for the additive model.

APPLICATION TO THE ADAM33 ASTHMA DATA SET

We illustrate the application of the random forests importance index I_M for individual SNPs and SNP pairs on a data set of SNPs within the ADAM33 gene, typed on a sample of asthma cases and normal controls. ADAM33 is an asthma susceptibility gene identified using a positional cloning approach [Van Eerdewegh et al., 2002]. The linkage study was conducted on 460 Caucasian families with multiple asthmatic children from US and UK populations, and identified chromosome 20p13 as a region with significant linkage to asthma and bronchial hyperresponsiveness. A subset of 131 unrelated asthma cases from families showing evidence of linkage to 20p13 and 217 unrelated controls, of the same ethnicity and countries of origin as the cases, was subsequently selected for a case-control study on SNPs within the linked region. Multiple SNPs and SNP pairs among the 42 SNPs typed in ADAM33 were found to be associated with asthma [Van Eerdewegh et al., 2002].

We applied random forests to this data set to determine which of the 42 ADAM33 SNPs are predictive of asthma case-control status. Missing genotype data were imputed based on other typed SNPs for each individual, as described in the Appendix. Applying random forests to the 131 cases and 217 controls resulted in a disproportionately large number of individuals predicted to be controls. We explain this behavior from the observation that some cases and controls have exactly the same genotype at all SNPs used in a tree. Those cases and controls are present in the same leaves of the trees, where controls tend to be overrepresented due to their larger number in the data set overall, leading the trees to predict all carriers of the genotype to be controls. We opted to sample as many controls as we had cases in each of our populations (US and UK) and to build random forests on these balanced samples of 131 cases and 131 controls. The sampling of 131 controls out of the 217 available individuals was repeated multiple times, and the results were averaged over the replicate samples to reduce the variability due to the sampling of controls. We ran analyses with up to 100 replicates and found little difference with the results from 10 replicates. The results reported here used 10 replicate samples.

The size of the subsets of SNPs from which the best split at each node was selected was first set to 6, following the suggestion by Breiman and Cutler [2003] to use the square root of the number of variables, and then progressively increased up to 15. The prediction error was found to be insensitive to the value of that parameter. The number of trees needed for convergence of importance indices was higher with larger subsets of predictors. Results are reported here for a predictor subset of size 10. We monitored convergence of the index I_M by measuring the standard error of the importance index from 10 repetitions of the random forest analysis on the same balanced sample, progressively increasing the number of trees. With 20,000 trees and a subset size of 10, the maximum standard error of the index I_M for an individual SNP was 0.5% of the estimated value, and we considered that the index I_M had converged.

RESULTS

The average misclassification rate of random forests over the replicate samples was 44%. We applied the same analysis to 100 data sets created by randomly permuting the case and control status, and found that the lowest misclassification rate among those 100 replicates was 47.8%, indicating that the observed misclassification rate is below what can be expected by chance. Figure 5 shows the negative $\log_{10}$ P-value of Fisher's exact test of association of SNPs with asthma plotted against I_M. While there is a general concordance

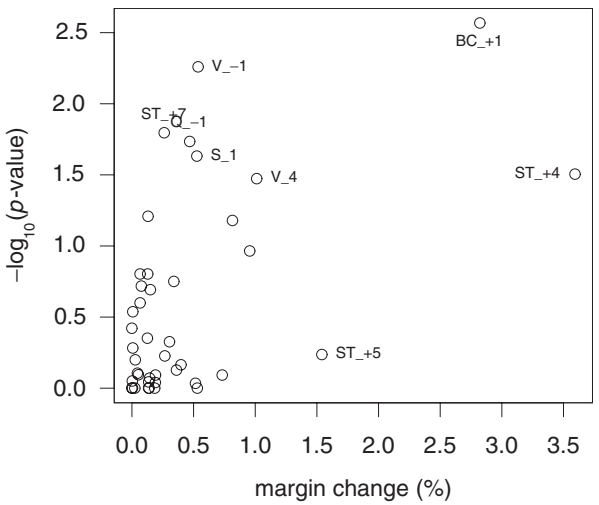

Fig. 5. Measures of association and predictive importance of individual SNPs in ADAM33. Predictive importance is measured with I_M. Association is measured by negative $\log_{10}$ P-value of Fisher's exact test on the 2×2 table of alleles vs. disease status.

between the importance index values and the strength of association, e.g., with SNPs BC+1 and ST+4 showing both strong association and high I_M values, some SNPs showing strong association to asthma had low importance index values, such as V−1, ST+7, and Q−1. In the control group, the frequency of the minor alleles of those three SNPs is lower than the frequency of the minor alleles of BC+1 and ST+4. Note that SNPs with high importance index values may also show little association with the phenotype, e.g., SNP ST+5.

Figure 6 compares the association of SNP pair haplotypes with asthma to the joint importance of SNP pairs as measured by I_M. The SNP pair with the highest I_M value is ST+4/BC+1, the two SNPs with the highest values of the individual variable I_M index. All the top SNP combinations include either of those two SNPs. Many of the SNP pairs with high I_M value are also among those with the most significant haplotype association to asthma. Other SNP pairs showing strong association to asthma do not have high I_M values, most notably those involving the SNP V−1.

A random forest built using only ST+4 and BC+1 as predictors gave a misclassification rate of 44%, equal to the rate for a random forest using all 42 SNPs. A random forest built using the 40 SNPs other than ST+4 and BC+1 had a misclassification rate of 45%, indicating that other SNPs in ADAM33 can also predict the phenotype almost as well as ST+4 and BC+1. However, when using as predictors only the three SNPs showing strong association but low importance (V−1, ST+7, and Q−1), the misclassification rate rose to 48%.

When we consider the difference between the joint I_M values and the sum of individual SNPs I_M values for all SNP pairs, a majority of joint I_M values are lower than the sum of the I_M values of the two SNPs forming the pair, including most pairs involving the SNPs ST+4 and BC+1 (not shown). The largest negative values are seen for combinations involving those two SNPs. Positive differences are all less than 0.05%, compared to negative differences of up to 0.51%.

DISCUSSION

Random forests are an alternative to classical statistical methods to identify SNPs predictive of a phenotype in a case-control study. We extended the index I_M based on the change in vote margin to jointly evaluate the predictive importance of pairs of SNPs, and studied the behavior of I_M for

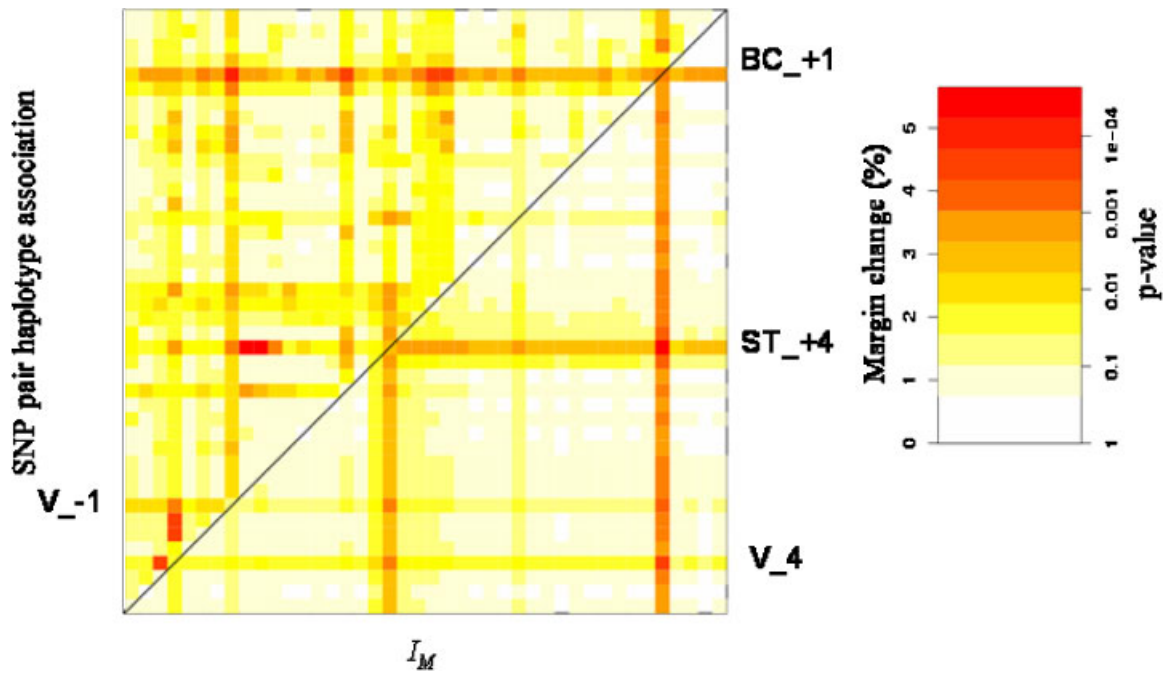

Fig. 6. Measures of association and predictive importance of SNP pairs in ADAM33. Predictive importance is measured with I_M (lower triangle). Association is measured by negative $\log_{10}$ *P*-value of likelihood ratio test of expected counts of haplotypes estimated separately in cases and controls over expected counts of haplotypes estimated by pooling cases and controls (upper triangle). [Color figure can be viewed in the journa's online edition: www.interscience.wiley.com]

individual SNPs and SNP pairs under simple two-locus disease models. We also presented the results of the analysis of a data set of SNPs within a candidate gene for asthma.

We first investigated the effect of noise variables on the random forests importance index values of true predictors of the response. Ideally, the presence of noise SNPs nonassociated with the disease would not affect the signal at disease-associated SNPs. In our study, we observed a decrease in both the signal at risk SNPs and the noise at SNPs unrelated to the disease with increasing numbers of noise SNPs. As long as the signal-to-noise ratio remains sufficiently high, the ability of the random forests importance index I_M retains its ability to detect risk SNPs in the presence of numerous noise SNPs. We indeed observed only a slow decrease of the ratio $I_M(\mathbf{A})$ to $\max_U I_M(U)$ with the number of noise SNPs in the models studied. The decrease in the proportion of replicates where the signal exceeded the noise was also slow, and was comparable to or slower than the decrease observed for Fisher's exact test applied to individual SNPs. Higher signal-to-noise ratios were observed for the joint importance of pairs of SNPs than for the importance of individual SNPs under our two-locus models. This suggests that, when multiple loci contribute to the risk of disease, measuring the importance of pairs of SNPs may be a more powerful approach than measuring the importance of each SNP individually.

It is straightforward to evaluate the joint importance of more than two variables by randomly permuting the values of any subset of variables jointly and capturing higher-order interactions. The rapid increase in the number of variable subsets with the size of the subset means that computational limitations will preclude evaluating every subset beyond a certain size. Strategies developed to search large discrete spaces could be applied to optimize the search for subsets with the highest importance index value.

The comparison in our simulation study of the values of I_M for SNP pairs to the sum of the values of I_M for the two SNPs individually revealed slight departures from additivity. The importance index I_M was subadditive in most cases. Interestingly, we observed that, when the value of I_M for SNP pairs differed from the sum of I_M for individual SNPs within ADAM33, it was usually less than the sum. Departures from additivity of I_M were present even when the risk differences were additive, but were of lesser magnitude than for epistatic models. Our results suggest that joint importance indices may provide information on interactions between SNPs, although more work is needed to establish whether the joint index I_M can be used to distinguish various forms of interactions. Our goal in developing the joint importance index was more to increase the power to detect disease-associated SNPs than to model particular pairwise interaction effects. Simpler models, including single trees, are likely to be more appropriate to confirm interaction effects, as attempted by Cook et al. [2004].

In the application to SNPs in the asthma susceptibility gene ADAM33, the misclassification rate of 44% represented a modest improvement over the rate of 50% expected for random predictions. This is not in contradiction with the observation of association between several of the SNPs in ADAM33 and asthma, since a significant association in a sample of the size of this study can be obtained with SNPs where a genotype more frequent in controls is also found in cases, and a genotype more frequent in cases is also found in controls, making the SNPs poor predictors of individual status. Limited predictive performance is expected for a complex trait like asthma where many genes other than ADAM33 as well as environmental factors are involved, and others obtained similar results in a similar context [Schwender et al., 2004]. The fact that the observed misclassification rate would be unlikely if the case and control status were unrelated to the SNP genotypes gives us some reassurance that the high changes in vote margin I_M for some of the SNPs and SNP pairs compared to the average level potentially represent reproducible signals. A random forest constructed using only the two SNPs with the highest importance index values achieved the same predictive accuracy as a forest constructed using all 42 SNPs, a performance that could not be matched using three SNPs also strongly associated with the asthma phenotype but with low I_M value. This confirms that the index I_M measures how much a SNP contributes to predicting the phenotype in independent (in this case, out-of-bag) individuals.

The random forest index I_M and Fisher's exact test of association did not agree perfectly in the SNPs they identified as related to asthma in ADAM33. The distinction between the two approaches is evidenced by their low-to-moderate concordance in identifying risk SNPs across simulated data replicates. The two methods can

therefore give different true- and false-positive results in the ADAM33 data set. The reasons for the differences in results are complex. One key factor that may explain the differences in results between the two methods is between-SNP interactions that influence disease risk. Fisher's exact test looks at each SNP individually, without taking into account any other genotypes. Thus, it measures the SNP's marginal effect in the data set. The random forest I_M for each SNP is a measure that takes into account the interactions that the SNP may have with other SNPs in the data set. Thus, when interactions exist, we would not expect the two measures to have high correlation. A second factor that we observed in our simulation study is a low risk allele frequency. Predictive importance depends on the proportion of individuals for which a polymorphism contributes to classification, and high-frequency polymorphisms can distinguish a larger proportion of individuals than low-frequency ones, even when the case sample is enriched for a rare risk allele. This suggests that the random forest approach may perform better at identifying frequent risk polymorphisms than rare ones. In the relatively simple scenarios we considered, the performance of a simple individual SNP association test was comparable to that of random forests. Future work will include the comparison of more complex scenarios.

The ultimate goal of a genetic study is to determine which genetic polymorphisms play a functional role in the etiology of a disease. Epidemiological data by themselves cannot answer that question, and different statistical methods applied to those epidemiological data may identify different polymorphisms, depending on the properties of the methods. Random forests identify SNPs that are predictive of a phenotype. These are likely to include some but not all of the SNPs involved in the etiology of a disease. We recommend random forests as a tool to screen large numbers of polymorphisms in association studies of complex diseases which should be used in conjunction with other statistical method to prioritize polymorphisms to be targeted in functional studies. In the case of ADAM33, it is only when comprehensive functional studies elucidate which polymorphisms play a role in the onset of asthma that an assessment of the success of random forests compared to other methods will be possible.

While we presented here an application to a data set of SNPs typed in a single gene, we anticipate that random forests will prove useful when applied to SNPs in multiple genes. Studying multiple polymorphisms in multiple genes simultaneously requires a method capable of capturing their interactions. In that respect, random forests is a promising tool.

REFERENCES

Breiman L. 1984. Classification and regression trees. Belmont, CA: Wadsworth International Group.

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

Breiman L, Cutler A. 2003. Random forests. Version 4.0. http://www.stat.berkeley.edu/users/breiman/RandomForests/.

Bureau A, Dupuis J, Hayward B, Falls K, Van Eerdewegh P. 2003. Mapping complex traits using random forests. BMC Genet [Suppl] 4:64.

Chiano MN, Clayton DG. 1998. Fine genetic mapping using haplotype analysis and the missing data problem. Ann Hum Genet 62:55–60.

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

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927.

Fleiss JL, Levin BA, Paik MC. 2003. Statistical methods for rates and proportions. Hoboken, NJ: John Wiley.

Glazier AM, Nadeau JH, Aitman TJ. 2002. Finding genes that underlie complex traits. Science 298:2345–2349.

Golden Helix, Inc. 2002. Helix tree manual. Version 2.2.0. Bozeman, MT: Golden Helix, Inc.

Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19.

Horng JT, Hu KC, Wu LC, Huang HD, Lin FM, Huang SL, Lai HC, Chu TY. 2004. Identifying the combination of genetic factors that determine susceptibility to cervical cancer. IEEE Trans Inf Technol Biomed 8:59–66.

Kim H, Neubert JK, San Miguel A, Xu K, Krishnaraju RK, Iadarola MJ, Goldman D, Dionne RA. 2004. Genetic influence on variability in human acute experimental pain sensitivity associated with gender, ethnicity and psychological temperament. Pain 109:488–496.

Morton NE, Collins A. 1998. Tests and estimates of allelic association in complex inheritance. Proc Natl Acad Sci USA 95:11389–11393.

Pociot F, Karlsen AE, Pedersen CB, Aalund M, Nerup J. 2004. Novel analytical methods applied to type 1 diabetes genome-scan data. Am J Hum Genet 74:647–660.

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

Schwender H, Zucknick M, Ickstadt K, Bolt HM. 2004. A pilot study on the application of statistical classification procedures to molecular epidemiological data. Toxicol Lett 151:291–299.

Van Eerdewegh P, Little RD, Dupuis J, Del Mastro RG, Falls K, Simon J, Torrey D, Pandit S, McKenny J, Braunschweiger K, Walsh A, Liu Z, Hayward B, Folz C, Manning SP, Bawa A, Saracino L, Thackston M, Benchekroun Y, Capparell N, Wang M, Adair R, Feng Y, Dubois J, Fitzgerald MG, Huang H, Gibson

R, Allen KM, Pedan A, Danzig MR, Umland SP, Egan RW, Cuss FM, Rorke S, Clough JB, Holloway JW, Holgate ST, Keith TP. 2002. Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. Nature 418:426–430.

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

APPENDIX: MISSING VALUES

We exploited the linkage disequilibrium (LD) present in the sample to infer missing SNP genotypes based on likely haplotypes formed with observed alleles at nearby SNPs. Windows of 19 consecutive SNPs were created, and haplotype frequencies over each 19-SNP combination in the pooled case and controls groups were estimated using our own implementation of the expectation maximization (EM) algorithm [Chiano and Clayton, 1998; Excoffier and Slatkin, 1995], allowing for missing genotype data at a subset of the SNPs. The two most likely haplotypes for each individual were determined over each window. The alleles at the SNP in the middle (10th) position of each window were used to impute the genotype of individuals whose genotype was missing at that position. Missing genotypes at the first and last nine SNPs were imputed using the alleles from the first and last haplotype window, respectively. The strength of this approach is its use of LD information to impute missing SNP genotypes. Version 4 of the Random Forests software [Breiman and Cutler, 2003] introduced a data imputation method based on a measure of proximity between the individuals which is derived from the forest. That method does not take advantage of the haplotype structure of genetic polymorphisms. Applying it resulted in very similar misclassification rates and importance index estimates. The limitation of both methods is that, by filling in missing SNP genotypes with one genotype value, the uncertainty in that genotype is ignored in the subsequent random forest analysis.