Published in final edited form as:

Genet Epidemiol. 2008 May ; 32(4): 285–300. doi:10.1002/gepi.20304.

An Ensemble Learning Approach Jointly Modeling Main and Interaction Effects in Genetic Association Studies

Zhaogong Zhang^{1,2}, Shuanglin Zhang^{1,2}, Man-Yu Wong³, Nicholas J. Wareham⁴, and Qiuying Sha^{1,*}

¹Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan

²Heilongjiang University, Harbin, China

³Department of Mathematics, Hong Kong University of Sciences and Technology, Hong Kong, China

⁴Department of Public Health and Primary Care, University of Cambridge Institute of Public Health, Cambridge, United Kingdom

Abstract

Complex diseases are presumed to be the results of interactions of several genes and environmental factors, with each gene only having a small effect on the disease. Thus, the methods that can account for gene-gene interactions to search for a set of marker loci in different genes or across genome and to analyze these loci jointly are critical. In this article, we propose an ensemble learning approach (ELA) to detect a set of loci whose main and interaction effects jointly have a significant association with the trait. In the ELA, we first search for “base learners” and then combine the effects of the base learners by a linear model. Each base learner represents a main effect or an interaction effect. The result of the ELA is easy to interpret. When the ELA is applied to analyze a data set, we can get a final model, an overall P -value of the association test between the set of loci involved in the final model and the trait, and an importance measure for each base learner and each marker involved in the final model. The final model is a linear combination of some base learners. We know which base learner represents a main effect and which one represents an interaction effect. The importance measure of each base learner or marker can tell us the relative importance of the base learner or marker in the final model. We used intensive simulation studies as well as a real data set to evaluate the performance of the ELA. Our simulation studies demonstrated that the ELA is more powerful than the single-marker test in all the simulation scenarios. The ELA also outperformed the other three existing multi-locus methods in almost all cases. In an application to a large-scale case-control study for Type 2 diabetes, the ELA identified 11 single nucleotide polymorphisms that have a significant multi-locus effect (P -value = 0.01), while none of the single nucleotide polymorphisms showed significant marginal effects and none of the two-locus combinations showed significant two-locus interaction effects.

Keywords

epistasis; association study; complex disease; Type 2 diabetes

INTRODUCTION

There is increasing evidence suggesting that complex diseases are the results of interactions of many genes and environmental factors, with each gene having a small effect [Risch, 2000; Risch et al., 1999; Nicolae and Cox, 2002; Carrasquillo et al., 2002; Olson et al., 2002; Hoh and Ott, 2003]. Furthermore, recent human and animal studies of complex diseases have identified susceptibility genes that marginally contribute to a common trait, to a minor extent only or not at all, but that interact significantly in combined analyses [Barlassina et al., 2002; De Miglio et al., 2004; Yanchina et al., 2004; Yang et al., 2004; Aston et al., 2005; Dong et al., 2005; Roldan et al., 2005; Millstein et al., 2006]. Thus, methods that can account for gene-gene interactions in searching for a set of marker loci in different genes and can analyze these loci jointly are critical.

Most association studies in practice essentially evaluate one locus at a time. These methods make the implicit assumption that susceptibility loci can be identified through their independent, marginal contributions to the trait variability. This simplified approach ignores the possibility that effects of multi-locus functional genetic units play a larger role than the single-locus effect in determining the trait variability [Templeton, 2000; Nelson et al., 2001; Hoh et al., 2001; Sha et al., 2006]. Forming haplotypes over multiple neighboring loci in one gene can increase the power of gene mapping studies [Zhao et al., 2000; Fallin et al., 2001; Schaid et al., 2002; Zhang et al., 2003], but these methods only work locally in a given genomic region. Although various authors have postulated the need for methods that investigate multiple interacting genes jointly [Tiwari and Elston, 1998; Cox et al., 1999; Templeton, 2000; Wilson, 2001; Cordell et al., 2001; Cordell, 2002; Culverhouse et al., 2002; Moore and Williams, 2002; Moore, 2003], only a few viable approaches in this direction exist [Hoh et al., 2001; Xiong et al., 2002; Potter, 2006; Dudbridge et al., 2006; Ritchie et al., 2001, 2003; Moore, 2004; Nelson et al., 2001; Culverhouse et al., 2004; Sha et al., 2006; Millstein et al., 2006].

The existing methods of searching for a set of disease-susceptibility genes and analyzing them jointly in association studies can be roughly divided into two groups. The first group is called conditional approaches. In conditional approaches, a new locus is searched for, given good evidence for an existing locus or a set of loci. The conditional approaches include: truncated product of P -values [Potter, 2006; Dudbridge et al., 2006], step-wise regression [Hoh and Ott, 2003], random forest [Bureau et al., 2005; Lunetta et al., 2004], multivariate adaptive regression splines [Cook et al., 2004], and Hotelling's T^2 test among others. While these methods can be powerful for finding many markers with small effects that combine to have an important effect on the phenotype, they do not explicitly account for possible epistatic interactions.

The second group is called exhaustive searching approaches in which each of the 1-locus, 2-locus, ..., and L -locus combinations is considered, and different models for each locus combination are also considered. This group includes the multifactor dimensionality reduction (MDR) method proposed by Ritchie et al. [2001, 2003] and reviewed recently by Moore [2004], the combinatorial partitioning method proposed by Nelson et al. [2001], the restrict partitioning method proposed by Culverhouse et al. [2004], the Combinatorial searching method (CSM) proposed by Sha et al. [2006], and the focused interaction testing framework (FITF) proposed by Millstein et al. [2006] among others. The methods in this group can search for a set of interacting loci that jointly have significant effects on the disease. However, the exhaustive searching approach is designed for a small number of markers in candidate gene studies. When the total number of markers is large, more than 10,000, for example, we should probably only consider two-locus combinations. Considering the locus combinations of more than two markers will be computationally

infeasible. In other words, when the number of markers is large, the existing methods that use exhaustive searching approaches cannot jointly analyze more than two markers.

In this paper, we present an alternative method, an ensemble learning approach (ELA) to analyze multi-marker jointly while accounting gene-gene interactions. The ELA is based on the ensemble learning framework in which many “base-learners” (with weak effects) are combined by a linear model. Each base learner represents a main effect or a low-order interaction effect. There are two steps in deriving base learners. In the first step, we search for candidate locus combinations with modest effects on the disease. Then, we find the “best” model for each candidate locus combination and derive base learners from the “best” models. Compared with the existing exhaustive searching approaches in searching up to two-locus combinations, the ELA first finds candidate single-marker effects and two-locus interaction effects and then evaluates statistical significance by jointly considering all the candidate effects. The FITF, on the other hand, evaluates the statistical significance of each single-marker and each two-locus combination separately, and the MDR, combinatorial partitioning method, restrict partitioning method, and CSM first find a “best” marker or a “best” two-locus combination and then evaluate the statistical significance of the “best” single-marker or the “best” two-locus combination. Our simulation studies demonstrated that the ELA is more powerful than the single-marker test in all the simulation scenarios. The ELA also outperformed the other methods using exhaustive searching approaches in almost all cases. In an application to a large-scale case-control study for Type 2 diabetes, the ELA identified 11 SNPs that have a significant multi-locus effect (P -value = 0.01), while none of the SNPs showed significant marginal effects and none of two-locus combinations showed significant two-locus interaction effects.

METHODS

The basic idea of the ELA is jointly modeling the main effects and interaction effects of many markers through the ensemble learning framework. Ensemble learning methods have emerged as being among the most powerful learning approaches [Breiman, 1996, 2001; Freund and Schapire, 1996; Friedman, 2001]. The structural model of an ensemble learning method takes the form

$$y_i = \alpha_0 + \sum_{j=1}^J \alpha_j f_j(X_i) + \varepsilon_i, i=1, \dots, n, \quad (1)$$

where each ensemble member or “base learner”, $f_j(X)$ is a function of the input variable X ; in the situation of genetic association studies, X_i is the numerical code of a multi-marker genotype and y_i is the trait value (for a qualitative trait, denote affected as 1 and unaffected as 0) of the i th individual; ε_i is a random error. Based on the model given in equation (1), as outlined in Figure 1, the ELA includes the following four parts: (1) deriving base learners, where each base learner may represent a main effect or an interaction effect; (2) using model selection and parameter estimation methods to get a final model, that combines many main effects and interaction effects or the effects of many base learners; (3) using a permutation test to evaluate the significance of association between the set of markers (or base learners) involved in the final model and the disease; and (4) calculating the importance measures of the base learners and markers to help to interpret the results if the permutation test shows significant results. Consider a sample of n individuals, and suppose that M single nucleotide polymorphisms (SNPs) are genotyped for each of the sampled individuals. Let y_i denote the trait value and X_i denote the numerical code of the multi-marker genotype of the i th individual. We give the details in the following sections.

DERIVING BASE LEARNERS

The real art of the ensemble learning approach used here is in the derivation of the base learners. To derive base learners, as shown in Figure 1, we use three steps: searching for candidate locus combinations, finding the “best model” for each candidate locus-combination, and transferring each “best model” to one or more base learners.

Step 1. Search for candidate locus combinations with modest effects—In this step, we consider each single marker and each two-locus combination, and retain those with modest effects. The reasons we consider only up to two-locus combinations are that, for a large M , the computation is infeasible and the genotype data will be very sparse for higher-order locus combinations. To measure the effect of a locus combination, we use a generalized Hotelling’s T^2 test that was recently proposed for the analysis of quantitative and qualitative traits with the use of multiple markers [Xiong et al., 2002; Chapman et al., 2003; Wallace et al., 2006]. Let individual i have trait value y_i and genotype scores X_i , respectively. Let $\mathcal{U} = \frac{1}{n} \sum_i (y_i - \bar{y})(X_i - \bar{X})$ and $V = \hat{\sigma}_y^2 \sum_i (X_i - \bar{X})(X_i - \bar{X})^T$, where $\bar{y} = \frac{1}{n} \sum_i y_i$, $\bar{X} = \frac{1}{n} \sum_i X_i$, $\hat{\sigma}_y^2 = (1/(n-1)) \sum_i (y_i - \bar{y})^2$, and n is the sample size. The score test statistic (the generalized Hotelling’s T^2 test statistic) is given by

$$T^2 = \mathcal{U}^T V^\oplus \mathcal{U}, \quad (2)$$

where $V^\oplus$ denotes the generalized inverse of V . The score test statistic T^2 asymptotically has a χ^2 distribution with degrees of freedom equal to the rank of V [Rao, 1962].

This test depends on the genotype coding scheme. Denote the three genotypes aa, aA and AA in each single marker by 0, 1, and 2. For a l -locus combination, Xiong et al. [2002] coded the genotype of individual i by $X_i = (x_{i1}, \dots, x_{il})$, where $x_{ij} = 0, 1$, or 2 according to the genotype at the j th marker. This kind of coding can only model additive effects of the markers, and the corresponding test will lose power on non-additive interactions and will have no power at all on pure interactions (no marginal effects). In this step, we use the following coding scheme. For a l -locus combination, let $g_1, \dots, g_{m+1}$ denote all the distinct l -locus genotypes observed in the sample. Define the numerical code for the l -locus genotype of individual i by $X_i = (x_{i1}, \dots, x_{im})$, where

$$x_{ij} = \begin{cases} 1 & \text{if the genotype of individual } i \text{ is } g_j, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

This coding scheme is equivalent to the mature model in Chapman et al. [2003], which includes all orders of interactions. Under this coding scheme and for a qualitative trait, the test statistic T^2 is the same as the goodness-of-fit test statistic given by

$$G_o = \sum_{j=1}^{m+1} \frac{(\hat{p}_j - \hat{q}_j)^2}{\frac{\hat{p}_j}{n_c} + \frac{\hat{q}_j}{n_a}}, \quad (4)$$

where n_a and n_c are the number of cases and controls, and $\hat{p}_j$ and $\hat{q}_j$ are the frequencies of genotype g_j in cases and controls, respectively (see Appendix for details).

With a total number of markers M , there are $M_1 = M$ one-locus tests, $M_2 = M(M-1)/2$ two-locus tests, and so on. A l -locus combination will be a candidate locus combination if the P -value of the l -locus test is less than a threshold δ_l . The threshold δ_l is determined by controlling the false discovery rates [FDRs; Benjamini and Hochberg, 1995], the ratio of the

number of falsely rejected null hypotheses to the total number of rejected null hypotheses. Take $l = 2$ as an example. To control the FDR α , we can choose the cutoff δ_2 as follows. Let $p_{(1)}, \dots, p_{(M_2)}$ be the ordered P -values (unadjusted) of the M_2 two-locus tests. Then,

$$\delta_2 = \max \left\{ p_{(i)} : p_{(i)} \leq \frac{i\alpha}{M_2} \right\}.$$

In the simulation studies and real data analysis, we use $\alpha = 0.75$ to determine the cutoff δ_l to retain the locus combinations with modest effects.

In this step, we aim to search for the candidate locus combinations with the modest effects. Thus, we use a large FDR ($\alpha = 0.75$). If a small FDR such as $\alpha = 0.05$ is used, we may miss all the locus combinations with the modest effects. Instead of FDR, we can also use Bonferroni correction to determine the cutoff δ_l . However, Bonferroni correction is very conservative in this case because of the strong correlation among locus combinations. As an example, we consider $M = 1,000$ markers, and thus $M_2 \approx 500,000$. By Bonferroni correction, $\delta_2 = \alpha/M_2 \approx 2\alpha \times 10^{-6}$. In this case, our simulation studies show that we will miss almost all of the two-locus combinations with the modest effects, even if $\alpha = 1$. To determine the cutoff δ_l , we believe that FDR is more appropriate than Bonferroni correction.

Step 2. Find the “best model” for each candidate locus combination—In this step, we use Nelson et al.’s [2001] idea to partition or group multi-locus genotypes and find the “best partition” as the “best model.” Nelson et al. [2001] proposed to evaluate every possible partition of the genotypes. This procedure makes the method computationally intensive even for a two-locus combination. As noted by Culverhouse et al. [2004], a large part of the partitions is unnecessary to be evaluated. In fact, a good partition should have the property that genotypes with similar effects on the disease will be in the same group. Following Sha et al.’s [2006] method of finding the approximate “best model” by using a clustering method, we first define a quantity for each genotype as its effect on the trait and then cluster the genotypes according to the similarities of their effects on the trait.

For a l -locus combination, let $g_1, \dots, g_{m+1}$ denote all the distinct l -locus genotypes observed in the sample. Define

$$Y_s = \frac{\bar{y}_s - \bar{y}}{\hat{\sigma}_{(\bar{y}_s - \bar{y})}}$$

as the effect of genotype g_s ($s = 1, \dots, m+1$) on the trait, where $\bar{y}_s$ is the average trait value of individuals with genotype g_s , $\bar{y}$ is the overall average of the trait values,

$\hat{\sigma}_{(\bar{y}_s - \bar{y})}^2 = ((n - n_s)/n_s n^2) \sum_{i=1}^n (y_i - \bar{y})^2$, and n_s is the number of individuals with genotype g_s . It is easy to see that a large positive Y_s means that genotype g_s has a large positive effect on the trait (g_s causes a high trait value), and a large negative Y_s means that genotype g_s has a large negative effect on the trait (g_s causes a low trait value). For a case-control study, $Y_s = (\hat{p}_s - \hat{q}_s)/\hat{\sigma}_{(\hat{p}_s - \hat{q}_s)}$, where $\hat{p}_s$ and $\hat{q}_s$ are the sample frequencies of genotype g_s in cases and in controls, respectively, and $\hat{\sigma}_{(\hat{p}_s - \hat{q}_s)}^2$ is the variance estimate of $\hat{p}_s - \hat{q}_s$.

Next, by using the K -mean clustering method [Johnson and Wichern, 1998; Sha et al., 2006], we cluster the $m+1$ genotypes, $g_1, \dots, g_{m+1}$, according to their effects on the trait; that is, $Y_1, \dots, Y_{m+1}$. Genotypes with similar effects on the trait will be clustered into the same group. We cluster the $m+1$ genotypes into k groups for $k = 2, 3, \dots, m+1$. For a given

number of k , we denote $G_1, \dots, G_k$ to be the k genotype groups found by the K -mean clustering method. We take $G_1, \dots, G_k$ as if they were k different genotypes and define a numerical code for the multi-locus genotype of the i th individual as a vector $X_i = (x_{i,1}, \dots, x_{i,k-1})$, where

$$x_{i,j} = \begin{cases} 1 & \text{if the genotype of individual } i \text{ belongs to } G_j, \\ 0 & \text{otherwise,} \end{cases}$$

and the value of the T^2 test statistic denoted by T_k^2 is given by equation (2). Let P_k denote the P -value of the test associated with test statistic T_k^2 ($k=2, \dots, m+1$). We propose to choose k with the smallest P -value as the “best partition”; that is, we choose k such that $P_k = \min\{P_2, \dots, P_{m+1}\}$. Because T_k^2 does not follow a χ^2 distribution, we use a permutation test to evaluate the P -values. For each permutation, we randomly permute the trait values. Based on the permuted data, we re-calculate Y_s for each genotype, re-do the clustering, and re-calculate the values of the T^2 test statistic for $k=2, \dots, m+1$. For the l th permutation, let $T_{2l}^2, \dots, T_{m+1,l}^2$ denote the values of the T^2 test statistic for $k=2, \dots, m+1$, respectively. The empirical P -value of the test associated with the test statistic T_k^2 by B permutations ($B=1,000$ in our simulations) is given by

$$\hat{P}_k = \frac{\sum_{l=1}^B I_{\{T_{kl}^2 > T_k^2\}}}{B}.$$

In this step, we use the K -mean clustering method to group genotypes and we are aware that there are many other clustering methods available to do this job. In fact, we have compared the performance of the K -mean with that of the other clustering methods which include hierarchical clustering method and mixture model method in our previous paper [Sha et al., 2006] and this paper. Our comparisons (results not shown) indicate that the K -mean clustering method is one of the best clustering methods in this situation.

Step 3. From the “best model” to base learners—Each “best model” in step 2 corresponds to a partition of the genotypes of a locus combination. If there are k groups in the partition of the “best model”, we use $k-1$ dummy variables, $x_1, \dots, x_{k-1}$, to denote the k groups, where

$$x_i = \begin{cases} 1 & \text{if the genotype is in the } i\text{th group,} \\ 0 & \text{otherwise.} \end{cases}$$

Then, the $k-1$ dummy variables will be the $k-1$ base learners corresponding to the “best model”. In this way, each “best model” of a locus combination corresponds to one or more base learners.

To illustrate the procedure of deriving base learners, let us consider an example of 10 markers. In step 1, we first test each single marker. Suppose that among the 10 markers, we retain marker 1 and marker 2 according to the criterion. Then, we test each of the two-locus combinations. Suppose that among the 45 two-locus combinations, we retain one of them, say $\{3, 4\}$ according to the criterion. Using the method described in step 2, we get one “best partition” for each retained single marker or locus combination. Suppose the “best partition” of the three genotypes of marker 1 contains two groups, $G_{A1} = \{AA, Aa\}$ and $G_{A2} = \{aa\}$,

the “best partition” of the three genotypes of marker 2 contains two groups, $G_{B1} = \{BB\}$ and $G_{B2} = \{Bb, bb\}$, and the “best partition” of the nine genotypes of two-locus combination $\{3, 4\}$ contains three groups, $G_{C1} = \{CCDD, CcDD, ccDD\}$, $G_{C2} = \{CCDd, CCdd\}$, and $G_{C3} = \{CcDd, Ccdd, ccDd, ccdd\}$. Then, we get four base learners given by the four indicator functions, $f_1(x) = I_{\{g_1 \in G_{A1}\}}$, $f_2(x) = I_{\{g_2 \in G_{B1}\}}$, $f_3(x) = I_{\{g_{\{3,4\}} \in G_{C1}\}}$, and $f_4(x) = I_{\{g_{\{3,4\}} \in G_{C2}\}}$, where g_1 , g_2 and $g_{\{3,4\}}$ denote the genotypes at marker 1, marker 2, and two-locus genotype at marker 3 and 4, respectively.

PARAMETER ESTIMATION IN THE STRUCTURAL MODEL

Under model (1) and given the base learners $f_1(x), \dots, f_J(x)$ we use the Lasso method [Tibshirani, 1996] to estimate the parameters $\{\alpha_j\}_0^J$. The estimators $\{\hat{\alpha}_j\}_0^J$ are given by

$$\{\hat{\alpha}_j\}_0^J = \arg \min_{\{\alpha_j\}_0^J} \left[\sum_{i=1}^n \left(y_i - \alpha_0 - \sum_{j=1}^J \alpha_j f_j(X_i) \right)^2 + \lambda \sum_{j=1}^J |\alpha_j| \right] \quad (5)$$

based on the Lasso method. The first term in (5) measures the prediction error on the training sample, and the second term penalizes large values of the coefficients of the base learners. The Lasso penalty (the second term in (5)) produces shrinkage among the estimators $\{\hat{\alpha}_j\}_0^J$ and forces many estimators to be zero. Thus, the Lasso estimation procedure can perform parameter estimation and model selection at the same time.

We use the Least angle regression (LAR) to implement the Lasso estimation. The LAR, recently proposed by Efron et al. [2004], is a useful and less greedy version of the traditional forward selection algorithm. A simple modification of the LAR algorithm implements the Lasso, and the computational time of the LAR is in the same order as the calculation of an ordinary least square estimator. The LAR procedure works roughly as follows. As the classic forward selection, the LAR selects one predictor at a time for the “most correlated” set. Suppose that predictors $f_{j_1}(x), \dots, f_{j_k}(x)$ are already in the “most correlated” set, the LAR proceeds in a direction equiangular between the k predictors until a $k+1$ predictor $f_{j_{k+1}}(x)$ earns its way into the “most correlated” set. After all predictors have been added to the “most correlated” set, we choose the model with the smallest $C_p(k)$ ($k = 1, \dots, J$) as the best model, where

$$C_p(k) = \frac{SSE_k}{\hat{\sigma}^2} - n - 2k, \quad (6)$$

SSE_k is the sum of squared errors of the LAR prediction at the k th step, and $\hat{\sigma}^2$ is the ordinary least square estimator of the variance of the response under the full model. Suppose that the “best model” contains J_0 predictors. The prediction equation from the “best model”

is given by $\hat{y}_i = \hat{\alpha}_0 + \sum_{l=1}^{J_0} \hat{\alpha}_l f_l(X_i)$, which we simply write as

$$\hat{y}_i = \hat{\alpha}_0 + \sum_{l=1}^{J_0} \hat{\alpha}_l f_l(X_i). \quad (7)$$

We use the LAR to implement the Lasso in estimating parameters and doing model selection based on the following considerations. First, the LAR is the fastest method in doing model selection. Second, the Lasso is one of the best model selection methods [Tibshirani, 1996; Efron et al., 2004].

PERMUTATION TEST

In the previous step, we ended with a final model. However, the statistical significance of the association between the markers involved in the model and the trait is still not clear. As pointed out by Devlin et al. [2003], under the null hypothesis of no markers associated with the trait, the model selection procedure may still select several predictors to the model. Thus, $J_0 = 0$ in final model (7) does not necessarily mean that the markers involved in the model jointly have a significant effect on the trait. To test the significance of the association between the set of markers involved in the model and the trait, we propose to use a permutation test based on the test statistic

$$T = \frac{C_p(0) - C_p(J_0)}{C_p(0)}, \quad (8)$$

where $C_p(J_0)$ and $C_p(0)$ are the values of $C_p(k)$ in equation (6) for the “best model” and the null model ($k = 0$ or no predictors in the model), respectively. Denote the value of the statistic given by equation (8) from the original sample as T_0 . In each permutation, we randomly permute trait values of the individuals and repeat all steps of the ELA to calculate the statistic T based on the permuted data. We repeated this permutation procedure B times ($B = 1,000$ in our simulation) and denote values of the statistic for the B permuted samples as $T_1, \dots, T_B$. The estimated P -value of the test for the association between the set of markers involved in the final model and the trait is

$$P\text{-value} = \frac{\#\{i : T_i > T_0\}}{B}. \quad (9)$$

This P -value is an overall P -value.

IMPORTANCE MEASURES

To help interpret the results, we give an importance measure for each of the base learners and each of the markers involved in the final model. Similar to the idea discussed by Hastie et al. [2001] and Friedman and Popescu [2005], using the notation in equation (7) we define the importance measure of the base learner $f_l(x)$ as

$$\hat{\beta}_l = \frac{|\hat{\alpha}_l|}{\hat{\sigma}(f_l)},$$

where $\hat{\sigma}^2(f_l) = 1/n \sum_{i=1}^n (f_l(X_i) - \bar{f}_l)^2$ and $\bar{f}_l = 1/n \sum_{i=1}^n f_l(X_i)$. Based on the importance measures of the base learners, we define the importance measure of marker l by

$$\text{Imp}_l = \sum_{j=1}^{J_0} \frac{\hat{\beta}_j}{m_j} I_{\{\text{marker } l \in f_j\}},$$

where $I_{\{\cdot\}}$ is an indicator function and m_j is the number of markers involved in base learner f_j . Thus, the markers involved in the final model have positive and all other markers have zero importance measures. If a marker has a large importance measure, this marker makes a large contribution to the final model or to the association between the set of markers involved in the final model and the trait. In the following discussion, all importance measures are re-scaled as the percentage of the sum of all importance measures. For example, Imp_l will be rescaled as $\text{Imp}_l / \sum_k \text{Imp}_k \times 100\%$.

MISSING DATA

In the above discussion, we assume that no genotypes are missing. However, missing values are not unusual in practice. One basic method to deal with missing genotypes is to delete the individuals with missing genotypes. If the probability of an individual having missing genotypes at one marker is 1%, then the probabilities of an individual with missing genotypes at one or more markers are 1, 2, 63, and ~99% for 1, 2, 100, and 500 markers, respectively. Deleting individuals with missing genotypes is a viable method when we consider only a few markers. However, this method will lose power if we consider many markers simultaneously, because a large portion of the individuals will be deleted in this case. Based on this consideration we propose the following approach to deal with missing genotypes. In the step when base learners are derived, we use the method to delete the individuals with missing genotypes because, in this step, we consider only one or two markers at a time. However, the method of deleting individuals with missing genotypes is not appropriate in the parameter estimation step because there may be many markers involved in the model given by equation (1). In this step, we use multiple imputations to deal with missing data [Souverein et al., 2006]. Multiple imputations allow us to fill each missing genotype randomly with one of the three genotypes according to the genotype frequencies in the sample. After we fill in all the missing genotypes, we apply the LAR to the filled data set to determine the statistic. By repeating the filling procedure D times ($D = 10$ in our data analysis), we get an average value of the statistic. In the permutation test, we also repeat the filling procedure D times to obtain the average value of the test statistic for each permutation. By replacing the value of the test statistic T by the average value, the estimated P -value is also calculated by equation (9). The importance measure of each base learner or each marker is the average of those in the filling procedures.

OTHER METHODS COMPARED

We use simulation studies as well as application of a real data set to evaluate the performance of our method and compare it with four other methods. The four methods include the FITF proposed by Millstein et al. [2006], CSM proposed by Sha et al. [2006], Hotelling's T^2 with the sequence-forward-selection (HT-SFS) proposed by Xiong et al. [2002], and single-marker test (SMT). Originally, we planned to compare our method with MDR [Ritchie et al., 2001]. The user-friendly interface of the currently available version of the MDR makes it very easy to analyze one data set, but it is hard to analyze many data sets automatically. In our simulation, we need to analyze more than 10,000 data sets. It is not practical to use the MDR to analyze one data set at a time manually. Thus, we used the CSM instead of MDR. The results of Sha et al. [2006] showed that the CSM and MDR have similar power. The Hotelling's T^2 statistic proposed by Xiong et al. [2002] is equivalent to the T^2 given by equation (2) by coding the m -marker genotype as $X = (x_1, \dots, x_m)$, where $x_k = 0, 1$, or 2 corresponding to genotype aa, Aa, or AA at the k th marker. The HT-SFS works as follows. Using the T^2 test, we test each of the markers and choose the marker with the smallest P -value, say marker 1. Then, we test all possible two-locus combinations that contain marker 1 by using the T^2 test and choose the two-locus combination with the smallest P -value, say markers 1 and 2. We continue to add one marker at a time to the winner set until a pre-specified number of loci is reached, or the P -value begins to increase. In this way, we get a final marker-set and an associated P -value.

However, this associated P -value is not the overall P -value. We use a permutation test to evaluate the overall P -value of the final marker-set. For the SMT, we use the Hotelling's T^2 test to test each of the M markers and denote the ordered P -values as $p_{(1)}, \dots, p_{(M)}$. To control for FDR α we choose the cutoff δ as

$$\delta = \max \left\{ p_{(i)} : p_{(i)} \leq \frac{i\alpha}{M} \right\}.$$

All the markers with associated P -values less than δ are considered to be associated with the trait. The set that consists of all the markers associated with the trait is called the final marker-set.

SIMULATION STUDIES

Although the ELA proposed in this article is applicable to both qualitative traits and quantitative traits, we considered only qualitative traits in our simulations because the FITF is only applicable to qualitative traits.

ASSESSING THE TYPE I ERROR

To assess the type I error rate of the ELA for testing associations between the set of markers involved in the final model and the trait, we generate alleles for each marker independently according to their allele frequencies, and randomly assign one individual as a case or a control independent of the genotypes. The frequency of the minor allele at each marker is drawn from a uniform distribution on the interval (a, b) . We consider two distributions of minor allele frequencies: $(a, b) = (0.1, 0.25)$ and $(a, b) = (0.25, 0.4)$ three sample sizes: 200, 400, and 800 (half cases and half controls), and three different numbers of markers: 10, 20, and 100. There are 18 different scenarios in total. For each scenario, we generate 1,000 replicated samples to evaluate the type I error rate. The estimated type I errors for the 18 scenarios are summarized in Table I. With 1,000 replicated samples, the standard deviations for the type I error rate are $\sqrt{0.05 \times 0.95/1000} \approx 0.007$ and $\sqrt{0.01 \times 0.99/1000} \approx 0.0031$ for nominal levels of 0.05 and 0.01 respectively. The 95% confidence intervals are (0.036, 0.064) and (0.004, 0.016) for type I error rates 0.05 and 0.01, respectively. It is easy to see from Table I that the estimated type I errors of the test are not significantly different from the nominal levels.

SIMULATION STUDIES FOR EVALUATING POWER

We use simulation studies to compare the power of the ELA with four existing methods: the FITF, CSM, HT-SFS, and SMT. First, we clarify the meaning of power for different methods.

Power and power calculation—To estimate the power of the five methods, we use 100 replicated samples in each simulated scenario. Suppose that there are M biallelic markers and among the M markers there are m disease loci. For the ELA, CSM, and HT-SFS, there is a final marker-set and an overall P -value of the test for testing the association between the final marker-set and the trait. Let s_j and p_j denote the number of disease loci contained in the final marker-set and the overall P -value of the test to test the association between the final marker-set and the trait, respectively, for the j th replicated sample. Then, the estimated

power of either the ELA, CSM, or HT-SFS is given by $power = (1/100m) \sum_{i=1}^{100} s_i I_{\{p_i \leq \alpha\}}$, where $I_{\{\cdot\}}$ is an indicator function and α is a significance level. The FITF or SMT also gives a final marker-set that contains all the markers significantly associated with the trait. Let s_j denote the number of disease loci contained in the final marker-set of either the FITF or

SMT. Then, the estimated power is given by $power = (1/100m) \sum_{i=1}^{100} s_i$. In other words, the power of either the ELA, CSM, or HT-SFS is the percentage of disease loci contained in the

final marker-sets that have significant association with the trait, and the power of either the FITF or SMT is the percentage of disease loci contained in the final marker-sets.

To reduce the computational burden, instead of evaluating the value of p_i by a permutation procedure for each replicated sample for the ELA, CSM, and HT-SFS, we used a significance threshold T_α for all 100 replicated samples in each simulation scenario. Take the ELA as an example. For the i th replicated sample, let T_i denote the value of the statistic T (given by equation (8)). Then $p_i \leq \alpha$ is equivalent to $T_i \leq T_\alpha$. To calculate the significance threshold T_α we apply the ELA to 1,000 replicated samples generated under the null hypothesis in each simulation scenario, which yields 1,000 values of the statistic T given by equation (8). The significance threshold T_α is the α percentile of the 1,000 values of T . We have compared the performance of using permutation to evaluate p_i with that of using a significance threshold T_α in each simulation scenario. Our comparisons for a small number of markers (10 and 20) show that these two methods have almost identical results.

In the power evaluations of the ELA and FITF, we search up to two-locus combinations in all simulation scenarios. Although we search only up to two-locus combinations, the final marker-set of either the ELA or FITF can consist of many number of markers. However, the final marker-set of the CSM can contain at most two markers if we search up to two-locus combinations. If there are more than two disease loci (there were 10 disease loci in one group of our simulations), it will not be fair to the CSM in the power comparison. In order to be fair to the CSM, we search up to m -locus combinations (m is the number of disease loci) when using the CSM.

Power comparisons—To compare power of the five methods, we generate M independent biallelic markers ($M = 20, 100$ or $1,000$) with minor allele frequencies randomly sampled between 0.1 and 0.33. To generate the genotypes at disease loci, we consider three sets of disease models. In the first set, we consider eight three-locus disease models, denoted as models $L1$ – $L8$, which are similar to those used by Millstein et al. [2006] in their simulation studies. Let A_i denote the minor allele at the i th disease locus (alternative allele is a_i) and the population frequency of A_i is a random number between 0.1 and 0.33. A logistic model is used to relate genotypes at disease loci to the trait. Let $p = P(\text{affected} | \text{genotype})$ and x_1, x_2 , and x_3 be the numerical codes of the genotypes at the three disease loci. The relationship between p and x_1, x_2, x_3 is given by the logistic model

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{123} x_1 x_2 x_3.$$

Assume that the overall population prevalence is 10%. Then the value of β_0 can be determined by the value of the other parameters. The eight different models are determined by the different values of the parameters and different coding schemes of genotypes. The values of the parameters are given in Table II. In models $L1$ – $L4$, $x_k = 0, 1$, or 2 corresponding to genotypes $a_k a_k$, $A_k a_k$, or $A_k A_k$ at the k th disease loci ($k = 1, 2, 3$), an additive coding of the genotypes. The coding schemes in models $L5$ – $L8$ are also given in Table II. The values of x_k are defined by

$$x_k = \begin{cases} 1, & A_k a_k \text{ or } A_k A_k \\ 0, & a_k a_k \end{cases} \quad \text{for dominant and}$$

$$x_k = \begin{cases} 0, & A_k a_k \text{ or } A_k A_k \\ 1, & a_k a_k \end{cases} \quad \text{for recessive,}$$

where we use the major allele a_k as the high-risk allele for the recessive coding.

The power comparisons under the eight logistic models are summarized in Table III. It can be seen clearly that, the ELA is the most powerful method among the five except for model $L4$. Under model $L4$ the HT-SFS is the most powerful method as expected, as the assumption of model $L4$ is exactly the same as the assumption of the HT-SFS; that is, additive effects of the two alleles at each locus and additive effects of different markers under a logistic model. Under model $L4$, the ELA is the second most powerful method among the five. Generally speaking, among the five methods, the ELA is the most powerful one, the FITF and HT-SFS have very similar power and are in the second most powerful group, and the CSM and SMT are in the least powerful group. Between the CSM and SMT, the SMT is more powerful than the CSM in models $L1$ – $L4$, which assume additive effects of the two alleles at each locus, and the CSM is more powerful than the SMT in models $L5$ – $L8$ which assume dominant or recessive effects of the two alleles at each locus.

In the second set of simulations, we consider eight two-locus epistatic models as described in Table IV. The first four models, Ep1–Ep4, modified from the models used by Becker et al. [2005], are two-locus epistatic models that exhibit interaction effects as well as main effects (marginal effects). The last four models, P1–P4 as described by Ritchie et al. [2003] and Sha et al. [2006], are four pure interaction models that exhibit interaction effects in the absence of any main effects when the Hardy-Weinberg equilibrium is assumed.

The power comparisons of the eight epistatic models are summarized in Table V. Among the first four models, Ep1–Ep4, with interaction effects as well as main effects, the ELA is the most powerful in models Ep1–Ep3, and the HT-SFS is the most powerful in model Ep4 in which two disease loci have an additive effect. Even in model Ep4, an additive model and favorable to the HT-SFS and FITF, the ELA is still the second most powerful method. Of the second four models, P1–P4, with interaction effects in the absence of any main effects, the SMT and HT-SFS have no power at all as expected, since these two methods only model main effects. Although these four models are among the most favorable models for the CSM, the ELA is still the most powerful method under models P2, P3, and P4. The CSM is the most powerful and ELA is the second powerful method in model P1.

In the third set of simulations, we consider eight 10-locus disease models. The 10 disease loci were divided into four groups, $S_1 = \{1,2,3\}$, $S_2 = \{4,5,6\}$, $S_3 = \{7,8\}$ and $S_4 = \{9,10\}$. We model the disease loci in S_1 by either model $L1$ or $L4$, S_2 by either model $L6$ or $L8$, S_3 by either model $Ep1$ or $Ep3$, S_4 by either model $P2$ or $P3$. There are 16 model combinations. We choose eight model combinations as listed in Table VI. For example, $L1+L6+Ep1+P2$ means that $L1, L6, Ep1$ and $P2$ are used to model the relationship between the trait and the genotypes in S_1, S_2, S_3 , and S_4 , respectively. We assume that the genotypes in the four groups of disease loci are independent in both cases and controls, that is,

$$\Pr(G_{s_1} G_{s_2} G_{s_3} G_{s_4} | \text{Affected}) = \prod_{i=1}^4 \Pr(G_{s_i} | \text{Affected})$$

and

$$\Pr(G_{S_1} G_{S_2} G_{S_3} G_{S_3} | \text{Normal}) = \prod_{i=1}^4 \Pr(G_{S_i} | \text{Normal}),$$

where G_{S_i} denotes the multi-marker genotype of the markers in S_i . The independence is favorable to the FITF and SMT, because it allows these two methods to consider the statistical significance separately for each single marker or each two-locus combination without losing much of power.

The results of the power comparisons of the eight 10-locus disease models are summarized in Table VI. The results show that the ELA is clearly the most powerful method. On average, the ELA is about 30% more powerful than the FITF, about 50% more powerful than the HT-SFS, and about 200% more powerful than the CSM and the SMT. The increase in power of the ELA over the other four methods indicates that considering the statistical significance of the main effects and the interaction effects in several marker-combinations jointly can be more powerful than considering the statistical significance of each of the single-marker or two-marker combinations separately.

IMPORTANCE MEASURE

We also evaluated the performance of the importance measure proposed in this paper. For each simulation scenario, an average importance measure is calculated over 100 replicated samples for each marker. The average importance measures of the 20 SNPs under models $L1$ – $L4$ are given in Figure 2. Under these four models, the first three SNPs are the disease loci. Figure 2 shows clearly that the importance measure of each disease locus is larger than that of the non-disease loci. In model $L1$ and model $L4$, the three disease loci have the same effect on the disease, and we can see from Figure 2 that the importance measures of the three loci are also similar. In model $L2$, the total effect of the first disease locus on the disease (both the main effect and the effect of the first disease locus interacted with other loci) is larger than that of the second and third disease loci. Figure 2 shows that the importance measure of the first disease locus is also larger than that of the second and third disease loci. In model $L3$, the first disease locus has the largest total effect on the disease and the second disease locus has the smallest total effect on the disease. We can see from Figure 2 that the importance measures of the three disease loci also have the same order; that is, the first disease locus has the largest importance measure and the second disease locus has the smallest importance measure. In summary, the disease locus with a large effect on the disease will have a large importance measure. The importance measures under other simulation scenarios showed similar patterns (results are not shown). These results show that the importance measure of a locus proposed in this paper can reflect the effect of the locus on the disease.

ANALYSIS OF TYPE 2 DIABETES DATA

The data set came from a case-control study of Type 2 diabetes. In this study, 152 SNPs in 71 candidate genes were genotyped in a population-based cohort of 517 unrelated Caucasians in the United Kingdom with Type 2 diabetes and an equal number of controls with normal glycated hemoglobin (HbA1c) levels, individually matched to cases by age, sex, and geographical location. The 71 genes were subdivided into three broad groups: (1) 15 genes primarily involved in pancreatic b-cell function; (2) 35 genes primarily influencing insulin action and glucose metabolism in the main target tissues, muscle, liver, and fat; and (3) 21 other genes. The third group includes genes that influence processes potentially relevant to diabetes, such as energy intake, energy expenditure, and lipid metabolism. The great majority of the 152 SNPs in the 71 genes had a minor allele frequency of greater than

5%. This data set has been described in detail and analyzed by Barroso et al. [2003]. Barroso et al. [2003] did single-marker association tests with each marker tested for dominant, additive, and recessive effects. They found 20 SNPs whose marker-specific P -values (unadjusted) were less than 5% in at least one model. The smallest unadjusted marker-specific P -value was 0.002. In $152 \times 3 = 456$ tests, the P -value of 0.002 is not necessarily less than 0.05 after adjusted for multiple tests.

Besides the SMT, we have used four multi-locus methods to analyze this data set. All of the four multi-locus methods, the HT-SFS, CSM, FITF, and ELA, can jointly analyze markers in different genes. The results are summarized in Table VII. Four SNPs with the smallest unadjusted P -values and adjusted P -values by the SMT are listed in Table VII. rs5210 had the smallest P -value (unadjusted P -value = 0.0014). However, after adjusting for multiple tests by the permutation approach, the adjusted P -value of rs5210 was 0.14. This means that the SMT did not find any significant association. The HT-SFS found a marker-set with 10 SNPs to be the “best” marker-set. However, the overall P -value between the “best” marker-set and Type 2 diabetes was 0.08. The “best” marker-set found by using the CSM (we searched up to three-locus combinations) contained two SNPs, rs5210 and rs2276931. However, the association between the two-locus set and Type 2 diabetes is not statistically significant at a 5% level (overall or adjusted P -value = 0.145). We used the FITF to test main effects and two-locus interactions (stage 2 in the FITF). The two SNPs with the strongest main effects and the two SNPs with the strongest two-locus effect with their unadjusted and adjusted P -values are listed in Table VII. It can be seen that the FITF did not find any statistically significant association at a 5% level.

When using the ELA on the data set, we found a final model with 11 SNPs. The permutation test in the ELA showed that the 11 SNPs have a significant joint effect on Type 2 diabetes (the overall or adjusted P -value is 0.01). The final model is given by

$$\hat{y} = \alpha_0 + \sum_{k=1}^{10} \alpha_k f_k(x),$$

where the 10 base learners, $f_1, \dots, f_{10}$ and the coefficients, $\alpha_1, \dots, \alpha_{10}$ are given in Table SI. Among the 10 base learners, each of $f_7 - f_{10}$ involved one SNP, and each of $f_1 - f_6$ involved two SNPs. Among the 11 SNPs, four SNPs—rs7577088, rs5400, rs8192680, and rs7252268 contributed to Type 2 diabetes additively. The other seven SNPs contributed to Type 2 diabetes through the effects of two-locus combinations (or two-locus interactions): rs1801282 $\times$ rs5218, rs283696 $\times$ rs736824, rs5210 $\times$ rs2304592, and rs5210 $\times$ rs2276931. The relationship of the 11 SNPs and Type 2 diabetes is depicted in Figure 3. The 11 SNPs belong to nine genes. Rs5210 and rs5218, both in the CPE gene and 0.82 kb apart, had a strong linkage disequilibrium ($r^2 = 0.98$). Rs2304592 and rs2276931, both in gene KCNJ11 and 1.7 kb apart, also had a strong linkage disequilibrium ($r^2 = 0.95$). The detailed information of the 11 SNPs involved in the final model is given in Table SII. Among the 11 SNPs, five SNPs, rs283696, rs1801282, rs2304592, rs2276931, and rs736824, are not in Barroso et al.'s [2003] Table II, which contains 20 SNPs with unadjusted P -values less than 5% under at least one of the three models. These five markers are involved in the final model of the ELA because of the effects of their interactions with each other or with the other SNPs.

In summary, the results of the data analysis showed that either the single SNPs (the results of the SMT) or two-locus combinations (the results of the CSM and FITF) or linear combinations of many single-marker effects (the results of HT-SFS) does not have a significant effect on Type 2 diabetes. The results also showed that the 11 SNPs found by the

ELA have a significant joint effect on Type 2 diabetes through the combination of 10 base learners, which represented four single-marker effects and six two-locus interaction effects. The relative importance of the 11 SNPs and the relative importance of the 10 base learners are shown in Figure 4. If we consider a base learner involved with one SNP as a single-marker effect and a base learner involved with two SNPs as a two-locus interaction effect, we can see from Figure 4 that there are no major single-marker effects and no major two-locus interaction effects. These findings are consistent with the complex nature of the genetic architecture of common diseases.

The ELA method has been implemented in the ELA program written in C programming language. The program is available at Shuanglin Zhang's homepage <http://www.math.mtu.edu/~shuzhang/software.html>.

DISCUSSION

The ELA is proposed to find SNPs with small main effects and low-order interaction effects that have an important joint effect on the trait. It is also proposed for analyzing data sets that may contain a large number of markers, especially data sets from genome-wide association studies. The development of the method was motivated by the consideration that the genetic factors of complex diseases may include many genes and their interactions, and it would be desirable to combine many small main effects as well as interaction effects in a test for marker/disease association. The existing methods of searching for a set of disease-susceptibility genes and analyzing them jointly are either jointly consider main effect only (conditional approaches) or mainly consider interaction effects (exhaustive searching approaches). Both the approaches may be not optimal due to the nature of complex diseases. The causes of complex diseases probably are the combinations of many main effects and interaction effects. As an example, consider the genetic factors we found for Type 2 diabetes as depicted in Figure 3. In this example, both the conditional approach and the exhaustive searching approach did not find significant results, which means that there are no significant main effects, no significant two-locus interaction effects, and no significant main effect combinations. The ELA, by jointly considering the main effects and two-locus interaction effects, found that the combination of four main effects and six two-locus interaction effects makes an important contribution to Type 2 diabetes. This example and our simulation studies show that the ELA can be more powerful than the existing methods in detecting disease-susceptibility genes of complex diseases.

Interpretation of multi-locus models with interactions is always a difficult task due to the complex nature of gene-gene interactions plus the different meanings of interaction in statistics and genetics literatures [Cordell et al., 2001; Moore and Williams, 2005]. The ELA uses base learners, the importance measure of the base learners, and the importance measure of markers to help interpret the multi-marker model, that is, the final model of the ELA. If a base learner involves one marker, we say that this base learner represents a main effect. If a base learner involves two markers, we say that the base learner represents a two-locus interaction effect. The importance measure of a base learner or a marker represents the relative importance of the base learner or the marker.

One drawback of the proposed ELA is the intensity of the computations due to the permutation test used to evaluate the overall P -value of the association test between the set of loci involved in the final model and the trait. Analyzing a data set with 1,000 markers and 1,000 individuals, it will take about 5 hr for 1,000 permutations on a PC computer with 3 hr GHz CPU. To analyze a large data set in the current version of the program, we first filter markers using a single-marker test; that is, we do a single-marker test for each individual marker and retain the markers with P -values less than a certain value, 1% or 5% for

example. We have used this method to analyze a genome-wide association data set of sporadic amyotrophic lateral sclerosis (ALS). This data set was made available to the public by Schymick et al. [2007] through the SNP database at the NINDS Human Genetics Resource center DNA and Cell Line Repository (<http://ccr.coriell.org/ninds>). In total 555,352 unique SNPs were genotyped across the genome in 276 patients with sporadic ALS and 271 neurologically normal controls in this data set. Schymick et al.'s [2007] analysis showed that 34 SNPs have P -values less than 0.0001; the smallest was found to be 6.8×10^{-7} . None of these SNPs reached a significance level of 0.05 after Bonferroni correction. We use our ELA program to analyze this data set (perform a single-marker test first to screen SNPs and retain 5,000 SNPs with the smallest P -values on the original data and each permuted data). We have found that 20 SNPs jointly have significant association with sporadic ALS. The details of this analysis will be discussed elsewhere. Another way to reduce the computational burden is the sample splitting technique. We divide the sample into two or more groups, and use one group to build the model and use the sample in the other groups to test the significance. The feasibility and the power of this technique need further investigations.

Acknowledgments

Contract grant sponsor: NIH; Contract grant numbers: R01 GM069940; R03 HG 003613; R01 HG003054; R03 AG024491; Contract grant sponsor: Hong Kong Earmarked Research Grant; Contract grant numbers: 603506; Contract grant sponsor: Overseas-Returned Scholars Foundation of Department of Education of Heilongjiang Province; Contract grant numbers: 1152HZ01.

REFERENCES

- Aston CE, Ralph DA, Lalo DP, Manjeshwar S, Gramling BA, Defreese DC, West AD, Branam DE, Thompson LF, Craft MA, Mitchell DS, Shimasaki CD, Mulvihill JJ, Jupe ER. Oligogenic combinations associated with breast cancer risk in women under 53 years of age. *Hum Genet.* 2005; 116:208–221. [PubMed: 15611867]
- Barlassina C, Lanzani C, Manunta P, Bianchi G. Genetics of essential hypertension: from families to genes. *J Am Soc Nephrol.* 2002; 13(Suppl 3):S155–S164. [PubMed: 12466306]
- Barroso I, Luan J, Middelberg RP, Harding AH, Franks PW, Jakes RW, Clayton D, Schafer AJ, O'Rahilly S, Wareham NJ. Candidate gene association study in Type 2 Diabetes indicates a role for genes involved in beta-cell function as well as insulin Action. *PLoS Biol.* 2003; 1:E20. [PubMed: 14551916]
- Becker T, Schumacher J, Cichon S, Baur MP, Knapp M. Haplotype interaction analysis of unlinked regions. *Genet Epidemiol.* 2005; 29:313–322. [PubMed: 16240441]
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B.* 1995; 57:289–300.
- Breiman L. Bagging predictor. *Mach Learn.* 1996; 26:123–140.
- Breiman L. Random Forests. *Mach Learn.* 2001; 45:5–32.
- Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Eerdewegh PV. Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol.* 2005; 28:171–182. [PubMed: 15593090]
- Carrasquillo MM, McCallion AS, Puffenberger EG, Kashuk CS, Nouri N, Chakravarti A. Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nat Genet.* 2002; 32:237–244. [PubMed: 12355085]
- Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered.* 2003; 56:18–31. [PubMed: 14614235]
- Cook NR, Zee RY, 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. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet.* 2002; 11:2463–2468. [PubMed: 12351582]
- Cordell HJ, Todd JA, Hill NJ, Lord CJ, Lyons PA, Peterson LB, Wicker LS, Clayton DG. Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics.* 2001; 158:357–367. [PubMed: 11333244]
- Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A. Loci on chromosome 2 (NIDDM1) and 15 interact to disease susceptibility to diabetes in Mexican American. *Nat Genet.* 1999; 21:213–215. [PubMed: 9988276]
- 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]
- Culverhouse R, Klein T, Shannon W. Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol.* 2004; 27:141–152. [PubMed: 15305330]
- De Miglio MR, Pascale RM, Simile MM, Muroi MR, Virdis P, Kwong KM, Wong LK, Bosinco GM, Pulina FR, Calvisi DF, Frau M, Wood GA, Archer MC, Feo F. Polygenic control of hepatocarcinogenesis in Copenhagen #F344 rats. *Int J Cancer.* 2004; 111:9–16. [PubMed: 15185337]
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999; 55:997–1004. [PubMed: 11315092]
- Devlin B, Roeder K, Wasserman L. Analysis of multilocus models of association. *Genet Epidemiol.* 2003; 25:36–47. [PubMed: 12813725]
- Dong C, Li WD, Li D, Price RA. Interaction between obesity-susceptibility loci in chromosome regions 2p25-p24 and 13q13-q21. *Eur J Hum Genet.* 2005; 13:102–108. [PubMed: 15470360]
- Dudbridge F, Gusnanto A, Koeleman B. Detecting multiple associations in genome-wide studies. *Hum Genomics.* 2006; 2:310–317. [PubMed: 16595075]
- Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat.* 2004; 32:407–499.
- Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res.* 2001; 11:143–151. [PubMed: 11156623]
- Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat.* 2001; 29:1189–1232.
- Friedman, JH.; Popescu, BE. Technical Report. Stanford University, Department of Statistics; 2005. Predictive learning via rule ensembles.
- Freund, Y.; Schapire, RE. Experiments with a new boosting algorithm. *Machine Learning: Proceeding of the Thirteenth International Conference*; Morgan Kaufman; San Francisco. 1996. p. 148-156.
- Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer Verlag; 2001.
- Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet.* 2003; 4:701–709. [PubMed: 12951571]
- Hoh J, Wille A, Ott J. Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res.* 2001; 11:2115–2119. [PubMed: 11731502]
- Johnson, RA.; Wichern, DW. Applied multivariate statistical analysis. New Jersey: Prentice-Hall; 1998.
- Lunetta KL, Hayward LB, Segal J, Eerdewegh PV. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.* 2004; 5:32. [PubMed: 15588316]
- Millstein J, Conti DV, Gilliland FD, W. James Gauderman WJ. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet.* 2006; 78:15–27. [PubMed: 16385446]
- Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered.* 2003; 56:73–82. [PubMed: 14614241]
- Moore JH. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert Rev Mol Diagn.* 2004; 4:795–803. [PubMed: 15525222]
- Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension. *Annu Med.* 2002; 34:88–95.

- Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays*. 2005; 27:637–646. [PubMed: 15892116]
- Nelson MR, Kardia SL, 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]
- Nicolae DL, Cox NJ. MERLIN....and the geneticist's stone? *Nat Genet*. 2002; 30:3–4. [PubMed: 11753377]
- Olson JM, Goddard KA, Dudek DM. A second locus for very-late-onset Alzheimer disease: a genome scan reveals linkage to 20p and epistasis between 20p and the amyloid precursor protein region. *Am J Hum Genet*. 2002; 71:154–161. [PubMed: 12016588]
- Potter DM. Omnibus permutation tests of association of an ensemble of genetic markers with disease in case-control studies. *Genet Epidemiol*. 2006; 30:438–446. [PubMed: 16671109]
- Rao CR. A note on a generalized inverse of a matrix with applications to problems in mathematical statistics. *J R Stat Soc Ser B*. 1962; 24:152–158.
- Risch NJ. Search for genetic determinations in the new millenium. *Nature*. 2000; 405:847–856. [PubMed: 10866211]
- Risch N, Spiker D, Lotspeich L, Nouri N, Hinds D, Hallmayer J, Kalaydjieva L, McCague P, Dimiceli S, Pitts T, Nguyen L, Yang J, Harper C, Thorpe D, Vermeer S, Young H, Hebert J, Lin A, Ferguson J, Chiotti C, Wiese-Slater S, Rogers T, Salmon B, Nicholas P, Myers RM. A genomic screen of autism: evidence for a multilocus etiology. *Am J Hum Genet*. 1999; 65:493–507. [PubMed: 10417292]
- 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]
- Ritchie MD, Hahn LW, Moore JH. Power of multifactor-dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol*. 2003; 24:150–157. [PubMed: 12548676]
- Roldan V, Gonzalez-Conejero R, Marin F, Pineda J, Vicente V, Corral J. Five prothrombotic polymorphisms and the prevalence of premature myocardial infarction. *Haematologica*. 2005; 90:421–423. [PubMed: 15749685]
- Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score test for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet*. 2002; 70:425–434. [PubMed: 11791212]
- Schymick JC, Scholz SW, Fung H-C, Britto A, Arepall S, Gibbs JR, Lombardo F, Matarin M, Kasperaviciute D, Hernandez DG, Crews C, Bruijn L, Rothstein J, Mora G, Restagno G, Chiò A, Singleton A, Hardy J, Traynor BJ. Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol*. 2007; 6:322–328. [PubMed: 17362836]
- Sha Q, Zhu X, Zuo Y, Cooper R, Zhang S. A combinatorial searching method for detecting a set of interacting loci associated with complex traits. *Ann Hum Genet*. 2006; 70:677–692. [PubMed: 16907712]
- Souverein OW, Zwinderman AH, Tanck MWT. Multiple imputation of missing genotype data for unrelated individuals. *Ann Hum Genet*. 2006; 70:372–381. [PubMed: 16674559]
- Templeton AR. Epistasis and complex trait. In: Wade, M.; Brodie, B., III; Wolf, J., editors. *Epistasis and the evolutionary process*. Oxford: Oxford University Press; 2000.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B*. 1996; 58:267–288.
- Tiwari H, Elston RC. Restrictions on components of variance for epistatic models. *Theor Popul Biol*. 1998; 54:161–174. [PubMed: 9733657]
- Wilson SR. Epistasis and its possible effects on transmission disequilibrium tests. *Ann Hum Genet*. 2001; 62:565–575. [PubMed: 11851986]
- Wallace C, Chapman JM, Clayton DG. Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *Am J Hum Genet*. 2006; 78:498–504. [PubMed: 16465623]

- Xiong M, Zhao J, Boerwinkle E. Generalized T^2 test for genome association studies. *Am J Hum Genet.* 2002; 70:1257–1268. [PubMed: 11923914]
- Yanchina ED, Ivchik TV, Shvarts EI, Kokosov AN, Khodzhayantz NE. Gene-gene interactions between glutathione-s transferaseM1 and matrix metalloproteinase 9 in the formation of hereditary predisposition to chronic obstructive pulmonary disease. *Bull Exp Biol Med.* 2004; 137:64–66. [PubMed: 15085249]
- Yang P, Bamlet WR, Ebbert JO, Taylor WR, de Andrade M. Glutathione pathway genes and lung cancer risk in young and old populations. *Carcinogenesis.* 2004; 25:1935–1944. [PubMed: 15192016]
- Zhang S, Sha Q, Chen HS, Dong J, Jiang R. Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. *Am J Hum Genet.* 2003; 73:566–579. [PubMed: 12929082]
- Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK. Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet.* 2000; 67:936–946. [PubMed: 10968775]

APPENDIX

For a l -locus combination, let $g_1, \dots, g_{m+1}$ denote all the distinct l -locus genotypes observed in the sample. If we code the l -locus genotype of individual i by $X_i = (x_{i1}, \dots, x_{im})$, where

$$x_{ij} = \begin{cases} 1 & \text{if the genotype of individual } i \text{ is } g_j, \\ 0 & \text{otherwise,} \end{cases}$$

then, for sample size n , the score test statistic $T^2 = \mathcal{U}^T V^{\oplus} \mathcal{U}$ given by equation (2) can be written as

$$T^2 = \frac{1}{\hat{\sigma}^2} \sum_{l=1}^{m+1} n_l (\bar{y}_l - \bar{y})^2, \quad (\text{A1})$$

where $\hat{\sigma}^2 = (1/n) \sum_{i=1}^n (y_i - \bar{y})^2$, n_l and $\bar{y}_l$ are the number and the average trait values of individuals who have genotype g_l . For a qualitative trait, suppose that among the n individuals there are N cases and M controls, and N_l cases and M_l controls with genotype g_l . Then

$$T^2 = \sum_{l=1}^{m+1} \frac{(\hat{p}_l - \hat{q}_l)^2}{\hat{q}_l/N + \hat{p}_l/M}, \quad (\text{A2})$$

where $\hat{p}_l = N_l/N$ and $\hat{q}_l = M_l/M$.

To prove the above statement, we re-index the n trait values, $y_1, \dots, y_n$, and the n genotype scores, $X_1, \dots, X_n$, as $\{y_{lj}, X_{lj}; l = 1, \dots, m+1; j = 1, \dots, n_l\}$, where y_{lj} and X_{lj} ($j = 1, \dots, n_l$) are the traits and genotype scores of all the individuals who have genotype g_l . It can be seen that X_{lj} is a m -dimensional vector with j th element being 1 and all others being zero. It follows that

$$\mathcal{U} = \sum_{i=1}^n (y_i - \bar{y}) X_i = \sum_{l=1}^{m+1} \sum_{j=1}^{n_l} (y_{lj} - \bar{y}) X_{lj}$$

$$=(n_1(\bar{y}_1 - \bar{y}), \dots, n_m(\bar{y}_m - \bar{y}))^T,$$

where $\bar{y}_l = (1/n_l) \sum_{j=1}^{n_l} y_{lj}$, and

$$\frac{V}{\hat{\sigma}^2} = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T = \sum_{i=1}^n X_i X_i^T - n \bar{X} \bar{X}^T.$$

Let $A = \sum_{i=1}^n X_i X_i^T$. It is easy to verify that

$$A = \text{diag}(n_1, \dots, n_m), \bar{X} = \frac{1}{n}(n_1, \dots, n_m)^T, \quad \text{and } (A - n \bar{X} \bar{X}^T)^{-1} = A^{-1} + \frac{1}{n_{m+1}} \mathbf{J} \mathbf{J}^T,$$

where $\mathbf{J}$ is a m -dimensional vector with all elements being 1. Note that $n = n_1 + \dots + n_{m+1}$. Then,

$$\begin{aligned} \mathcal{U}^T V^{-1} \mathcal{U} &= \frac{1}{\hat{\sigma}^2} \mathcal{U}^T \left(A^{-1} + \frac{1}{n_{m+1}} \mathbf{J} \mathbf{J}^T \right) \mathcal{U} \\ &= \frac{1}{\hat{\sigma}^2} \sum_{l=1}^{m+1} n_l (\bar{y}_l - \bar{y})^2 \end{aligned}$$

For a qualitative trait, $\bar{y} = N/(N+M)$ and $\hat{\sigma}^2 = NM/((N+M)^2)$. From equation (A.1), we have

$$\begin{aligned} \mathcal{U}^T V^{-1} \mathcal{U} &= \frac{1}{\hat{\sigma}^2} \sum_{l=1}^{m+1} n_l (\bar{y}_l - \bar{y})^2 \\ &= \frac{1}{\hat{\sigma}^2} \sum_{l=1}^{m+1} (N_l + M_l) \left(\frac{N_l}{M_l + N_l} - \frac{N}{M + N} \right)^2 \\ &= \frac{1}{\hat{\sigma}^2} \sum_{l=1}^{m+1} (N_l + M_l) \left(\frac{N_l(M+N) - N(M_l + N_l)}{(M_l + N_l)(M+N)} \right)^2 \\ &= \sum_{l=1}^{m+1} \frac{(N_l M - N M_l)^2}{(M_l + N_l) M N} \end{aligned}$$

$$= \sum_{l=1}^{m+1} \frac{(\hat{p}_l - \hat{q}_l)^2}{\hat{q}_l/N + \hat{p}_l/M}.$$

The last term is the statistic of the goodness-of-fit test.

Fig. 1.
Outline of the ensemble learning approach procedure.

Fig. 2. The average importance measures over 100 replicated samples of the 20 single nucleotide polymorphisms for models L1–L4. Each sample consisted of 200 cases and 200 controls.

Fig. 3.
The relationship between the 11 single nucleotide polymorphisms in the final marker-set of the ensemble learning approach and Type 2 diabetes.

Fig. 4.

Importance measures of the 11 single nucleotide polymorphisms and the 10 base learners involved in the final model of the ensemble learning approach in analyzing Type 2 diabetes data set.

TABLE I

Type I error rates of the ELA

Minor allele frequency	Sample size	Significance level 5%			Significance level 1%		
		Number of markers			Number of markers		
		20	50	100	20	50	100
0.1–0.25	200	0.052	0.05	0.045	0.007	0.011	0.003
	400	0.041	0.057	0.064	0.007	0.012	0.009
	600	0.037	0.045	0.043	0.008	0.016	0.006
0.25–0.5	200	0.04	0.049	0.063	0.008	0.012	0.015
	400	0.035	0.053	0.050	0.006	0.016	0.01
	600	0.056	0.037	0.053	0.013	0.013	0.012

TABLE II

Eight logistic disease models

Models	Parameters
<i>L1</i>	$\beta_{123} = \log(3)$
<i>L2</i>	$\beta_1 = \log(1.5); \beta_{123} = \log(3)$
<i>L3</i>	$\beta_1 = \log(1.5); \beta_2 = \log(0.65); \beta_{123} = \log(3)$
<i>L4</i>	$\beta_1 = \beta_2 = \beta_3 = \log(1.5)$
	Parameters Coding schemes
<i>L5</i>	$\beta_{123} = \log(6)$ dominant, dominant, dominant
<i>L6</i>	$\beta_{123} = \log(6)$ dominant, dominant, recessive
<i>L7</i>	$\beta_{123} = \log(6)$ dominant, recessive, recessive
<i>L8</i>	$\beta_{123} = \log(6)$ recessive, recessive, recessive

Penetrance $p = \text{Pr}(\text{affected}|\text{genotype})$ is modeled by a logistic model $\log \frac{p}{(1-p)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{123} x_1 x_2 x_3$, where x_1 , x_2 , and x_3 are the numerical codes of genotypes at three functional SNPs as described in the text. In models *L1–L4*, x_j ($j = 1, 2, 3$) is an additive code of genotype at the i th disease locus, while in models *L5–L8*, the coding scheme at each disease locus is either dominant or recessive as given in the table. Minor allele frequencies were set between 0.1 and 0.33, and disease prevalence was set to be 0.1. Except for β_0 , all other parameters not in the table are zero, and the value of β_0 can be determined by the values of the other parameters.

TABLE III

Power (in percentage) comparisons of the eight logistic models

Model	No. of markers	ELA	FITF	CSM	SMT	HT-SFS
L1	20	38.3	26.6	19	17.6	32
	100	22	19	7.3	11.3	13
	1000	31			17.3	24
L2	20	60.6	47	34	42.3	55.3
	100	45.3	37.3	25.3	33.3	42
	1000	57.6			41.6	48.6
L3	20	58	41.3	33.6	30.3	40.6
	100	38.3	27	18.3	22	25.6
	1000	46.3			33.6	35.3
L4	20	51.6	27	10	29	56
	100	20	13.3	5.3	15.6	33
	1000	34.6			26.6	48
L5	20	45	25.3	21	12	30.6
	100	22.6	17	13	7.6	12.3
	1000	35			17.6	24.3
L6	20	68.3	51	42	32.3	51
	100	45.6	38.3	25	23.3	29.3
	1000	62.6			39.3	52
L7	20	91.3	72.3	60	54.3	70.3
	100	76.6	56	45.6	43.6	50.6
	1000	94			64.3	63
L8	20	98.3	87.6	73	75.3	90
	100	95	75.3	63.6	67.6	76
	1000	98.3			87	78.6

Each data set consisted of 200 cases and 200 controls (400 cases and 400 controls for the cases of 1,000 markers). Power corresponding to blank cells was not calculated because of being too computationally intensive. The combinatorial searching method searched up to three-locus combinations. For each scenario, the highest power is in bold.

TABLE IV

Eight two-locus epistatic models

Model	f_{22}	f_{21}	f_{20}	f_{12}	f_{11}	f_{10}	f_{02}	f_{01}	f_{00}	p_1	p_2
Ep1	0.222	0.222	0.08	0.222	0.222	0.08	0.08	0.08	0.08	0.21	0.21
Ep2	0.26	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.577	0.577
Ep3	0.24	0.24	0.08	0.24	0.08	0.08	0.08	0.08	0.08	0.349	0.349
Ep4	0.267	0.267	0.18	0.267	0.267	0.18	0.18	0.18	0.08	0.053	0.053
P1	0.08	0.07	0.05	0.1	0	0.1	0.08	0.1	0.04	0.25	0.25
P2	0	0	0.1	0	0.05	0	0.1	0	0	0.5	0.5
P3	0.07	0.05	0.02	0.05	0.09	0.01	0.02	0.01	0.08	0.1	0.1
P4	0.09	0.001	0.02	0.08	0.07	0.005	0.003	0.007	0.02	0.1	0.1

f_{ij} is the penetrance of a genotype carrying i high-risk alleles at locus 1 and j high-risk alleles at locus 2. p_1 and p_2 are the allele frequencies of high-risk alleles at locus 1 and locus 2, respectively.

TABLE V

Power (in percentage) comparisons of the eight two-locus epistatic models

Model	No. of markers	ELA	FITF	CSM	SMT	HT-SFS
Ep1	20	53.5	54.5	48	36	51.5
	100	39	21	21	15	23.5
	1,000	66		48	34.5	43.5
Ep2	20	77.5	65	49.5	40	53.5
	100	64.5	23	26	21	22.5
	1,000	84		68	46	37
Ep3	20	58	49	57.5	43	57.5
	100	49.5	15.3	25	22.5	27
	1,000	70.5		64.5	43.5	50.5
Ep4	20	51	44.5	49	43.5	69
	100	32	17	24.5	23.5	32.5
	1,000	55.5		37	41	62
P1	20	100	29.2	100	0.5	0
	100	87	5	97	0	0
	1,000	96		100	0	0
P2	20	100	100	100	0	1
	100	100	100	100	0	0.5
	1,000	100		100	0	0
P3	20	90.5	89.5	82.5	0.5	0.5
	100	58	41	48	0	0.5
	1,000	96		77	0	0
P4	20	95	93.5	89	0.5	1.5
	100	84	75	78	0	1
	1,000	100		99	0	0.5

Each data set consisted of 200 cases and 200 controls (400 cases and 400 controls for the cases of 1,000 markers). Power corresponding to blank cells was not calculated because of being too computationally intensive. The CSM searched up to two-locus combinations. For each scenario, the highest power is in bold. ELA, ensemble learning approach; FITF, focused interaction testing framework; CSM, combinatorial searching method; SMT, single-marker test; HT-SFS, Hotelling's T with the sequence-forward-selection.

TABLE VI

Power (in percentage) comparisons of the eight 10-locus diseased models

Model	No. of markers	ELA	FTTF	CSM	SMT	HT-SFS
L1+L6+EP1+P2	20	76.3	61.8	28	23.5	51.6
	100	61.9	44.6		15.5	36.5
	1,000	66.5			24.3	35
L1+L6+EP3+P2	20	78.2	62.7	40	25.4	51.1
	100	66	44.6		13.5	38.3
	1,000	69.2			23.6	36.4
L1+L8+EP1+P3	20	76.8	68.5	2	36.7	54.6
	100	64.5	43.8		28.9	46.4
	1,000	61.6			38	38.1
L1+L8+EP3+P3	20	73	70.8	9.6	37.3	54.4
	100	61.7	43.8		30.2	47
	1,000	63.3			41.6	38.3
L4+L6+EP1+P3	20	79.9	57	2.9	23.4	55.3
	100	60.1	28.2		14.6	45.2
	1,000	76.1			24.4	40.1
L4+L6+EP3+P3	20	81.1	57.7	15.6	27.8	54.2
	100	60.6	28.2		14.8	46.8
	1,000	74.2			24.8	39
L4+L8+EP1+P2	20	82.2	71.1	24	37	62.1
	100	72.7	58.4		29	53.8
	1,000	78.3			39.3	38.2
L4+L8+EP3+P2	20	81.7	73.1	34.7	41.3	61.1
	100	73.7	59.5		28.8	54.9
	1,000	78.8			45.1	38.3

Each data set consisted of 200 cases and 200 controls (400 cases and 400 controls for the cases of 1,000 markers). Power corresponding to blank cells was not calculated because of being too computationally intensive. The CSM searched up to 10-locus combinations. For each scenario, the highest power is in bold. ELA, ensemble learning approach; FTTF, focused interaction testing framework; CSM, combinatorial searching method; SMT, single-marker test; HT-SFS, Hotelling's T with the sequence-forward-selection.

TABLE VII

The results of analyzing Type 2 diabetes data set by using the five methods

Methods	Results			
SMT	The four most significant SNPs			
	rs5210	rs7577088	rs7252268	rs8192691
	Unadjusted <i>P</i> -value	0.0014	0.0023	0.01
	Adjusted <i>P</i> -value	0.14	0.25	0.75
HT-SFS	Final marker-set contains 10 SNPs with IDs rs7577088, rs5400, rs3755863, rs8192680, rs1042044, rs5210, rs3136540, rs1051690, rs7252268, rs2071197; <i>P</i> -value = 0.08 for testing the association between the 10 SNPs and the trait			
CSM	The best marker-set contains two SNPs, rs5210 and rs2276931; <i>P</i> -value = 0.145			
FITF	Two most significant SNPs in marginal tests and tests in stage 2			
	Marginal tests		Tests in stage 2	
	rs5210	rs7577088	rs3842752 and rs7577088	
	Unadjusted <i>P</i> -value	9×10^{-4}	1.7×10^{-3}	
	5% significance cutoff	1.6×10^{-4}	3.36×10^{-4}	
	Adjusted <i>P</i> -value	>0.05	1.1×10^{-5} >0.05	
ELA	11 SNPs (see Fig. 3, Tables SI and SII for details) in the final marker-set; <i>P</i> -value = 0.01 for testing the association between the 11 SNPs and the trait			

SNPs, single nucleotide polymorphisms; SMT, single-marker test; HT-SFS, Hotelling's T with the sequence-forward-selection; CSM, combinatorial searching method; FITF, focused interaction testing framework; ELA, ensemble learning approach.