BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 15 2014, pages 2171–2178
doi:10.1093/bioinformatics/btu186

Genetics and population analysis

Advance Access publication April 9, 2014

A fast and powerful tree-based association test for detecting complex joint effects in case–control studies

Han Zhang[1], William Wheeler[2], Zhaoming Wang[1,3], Philip R. Taylor[1] and Kai Yu[1,*]

[1]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20850, USA, [2]Information Management Services, Inc., Silver Spring, Maryland 20904, USA, and [3]Cancer Genomics Research Laboratory, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Gaithersburg, Maryland 20877, USA

Associate Editor: Jeffrey Barrrett

ABSTRACT

Motivation: Multivariate tests derived from the logistic regression model are widely used to assess the joint effect of multiple predictors on a disease outcome in case–control studies. These tests become less optimal if the joint effect cannot be approximated adequately by the additive model. The tree-structure model is an attractive alternative, as it is more apt to capture non-additive effects. However, the tree model is used most commonly for prediction and seldom for hypothesis testing, mainly because of the computational burden associated with the resampling-based procedure required for estimating the significance level.

Results: We designed a fast algorithm for building the tree-structure model and proposed a robust TREe-based Association Test (TREAT) that incorporates an adaptive model selection procedure to identify the optimal tree model representing the joint effect. We applied TREAT as a multilocus association test on >20 000 genes/regions in a study of esophageal squamous cell carcinoma (ESCC) and detected a highly significant novel association between the gene *CDKN2B* and ESCC ($P = 6.0 \times 10^{-8}$). We also demonstrated, through simulation studies, the power advantage of TREAT over other commonly used tests.

Availability and implementation: The package TREAT is freely available for download at http://www.hanzhang.name/softwares/treat, implemented in C++ and R and supported on 64-bit Linux and 64-bit MS Windows.

Contact: yuka@mail.nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 24, 2013; revised on March 26, 2014; accepted on April 2, 2014

1 INTRODUCTION

The logistic regression model is the most widely used approach for studying the relationship between a binary outcome and a set of explanatory variables. Under the logistic regression model, the cumulative effect of all the considered explanatory variables is assumed to be additive on the logit scale. In situations where the summation of the main effects from all considered variables is not adequate, the addition of interaction terms, each of which is a product of two chosen main effects, can be considered. The logistic regression model is a convenient tool for studying the joint effect of multiple risk factors if their effects are nearly additive, but it can be less efficient if the joint effect is more complicated and cannot be approximated adequately by the additive model, even with the inclusion of interaction terms.

The tree-structure model (Breiman *et al.*, 1984; Zhang and Singer, 1999), which hierarchically partitions the data into multiple exclusive subsets and models each subset individually, can be an attractive alternative to the linear model approach, especially in the situation where the joint effect of multiple risk factors (predictors) is non-additive. Tree-structure models are used most often for outcome prediction but seldom for hypothesis testing. When building a tree model, a predictor is chosen to split each subset (a node of a tree) according to certain criteria. Also, the tree size is decided through a model selection procedure, such as cross-validation. Thus, there are extensive model and variable selections in the tree-model-building process. As a result, it is hard to obtain the asymptotic distribution for the test statistic derived from the final tree model. Instead, we have to resort to a computationally intensive resampling-based procedure to evaluate the statistical significance (Chen *et al.*, 2007; Yu *et al.*, 2007). This greatly limits the application of the tree-structure model in the setting of hypothesis testing.

Our goal is to develop a robust hypothesis-testing procedure based on the tree-structure model that is computationally efficient, even with the resampling-based procedure. The procedure is for testing the null hypothesis that none of the considered risk factors is associated with the outcome. One motivated application is to conduct a gene-level association test on a genome-wide association (GWA) study. GWA studies typically measure genotypes on 200 000–2 million single-nucleotide polymorphisms (SNPs) on a group of cases and controls. One way to identify loci that are associated with the disease condition is to perform an association test on each SNP separately. Another strategy is to conduct a gene-level analysis by treating each chromosomal region as a testing unit and testing whether all SNPs in the considered region are associated with the disease condition. A considered region can contain a few to a few hundred SNPs measured by a commercial genotype array. When conducting an agnostic screen over the genome, we usually need to test >20 000 genes or annotated regions. When the Bonferroni correction for multiple comparisons is used, the *P*-value threshold for declaring the global (genome-wide) significance at the family-wise type I error rate of 0.05 is about 2.5×10^{-6} ($= 0.05/20\ 000$). This creates a huge computational burden on existing gene-based

*To whom correspondence should be addressed.

tests that rely on a resampling-based procedure to estimate their significance levels. For example, it requires about 10^8 iterations for a resampling-based procedure to achieve a reliable estimate for a *P*-value at the level of 1.0×10^{-6}, which would normally occur when >20 000 tests are performed.

Our proposed method is a general testing procedure that can be used as a multivariate test for the association between a set of predictors and a binary outcome. The test statistic is derived from a tree-structure model that approximates the joint effect of a set of predictors on the disease risk and relies on a permutation procedure to evaluate its significance level. We reduce the computational burden considerably by developing memory- and CPU-efficient algorithms for evaluating splitting rules in the tree-building process. We further speed up the testing procedure by deriving closed-form score test statistics for assessing the joint effect summarized by the tree-structure model. We use this new procedure to perform a genome-wide gene-based analysis on >20 000 genes/regions in a GWA study of esophageal squamous cell carcinoma (ESCC) based on about 2000 cases and 2000 controls. We also demonstrate the advantage of the proposed procedure over existing ones through extensive simulation studies.

2 METHODS

2.1 Notations

Here, we describe the proposed testing procedure as a multilocus test for the association between a disease condition and a set of genetic markers within a chromosomal region or gene (henceforth we use 'gene' for brevity, but everything applies equally to annotated regions). But the procedure is a general joint testing procedure and is applicable in areas beyond genetics. We consider a case–control study with n_1 cases and n_0 controls. The total sample size is $n = n_0 + n_1$.The observation for a subject can be represented as $(y,\ \mathbf{g},\ \mathbf{x})$, where the outcome $y = 1$ for an affected subject and $y = 0$ for a subject without the disease condition, $\mathbf{g} = (g_1, \ldots, g_p)$ is a vector of measured genotypes (encoded as the counts of the minor allele) at p SNPs within the considered gene, and $\mathbf{x}$ is the vector of all adjusted covariates. We consider a binary tree-structure model, in which a node is always split into two off-spring nodes by a binary splitting rule. We consider splitting rules defined by dichotomized genotypes. For a given SNP, we can create three binary variables, with each of them being the indicator for one type of the genotype g (0, 1 or 2) at a given SNP. We can also define other types of binary variables as candidate splitting rules. For example, to be more interpretable, we can restrict the splitting rules to two types of indicator variables, $g^{(1)}$ and $g^{(2)}$, where $g^{(1)} = \mathrm{I}(g = 1 \text{ or } 2)$ and $g^{(2)} = \mathrm{I}(g = 2)$, $\mathrm{I}(\cdot)$ is the indicator function. Regardless of how the set of binary variables is defined, we can represent all possible binary variables derived from genotypes of a set of SNPs as $\mathbf{f} = (f_1, \ldots, f_J)$. For example, there are $2p$ binary variables in $\mathbf{f}$, if only $g^{(1)}$ and $g^{(2)}$ are considered for each SNP. In the following real data application and in numerical experiments, we use splitting rules defined by $g^{(1)}$ and $g^{(2)}$.

2.2 The statistical test based on the tree-structure model

We assume that the subjects can be assigned into C unknown genetic risk groups according to the genetic profile characterized by the joint genotypes within a considered gene. The latent assignment can be denoted by a set of binary variables $\{z_k, k = 1, \ldots, C\}$, with $z_k = 1$ if the subject's genotype profile belongs to risk group k and $z_k = 0$ otherwise. We further assume that the underlying risk model has the form:

$$\log\left(\frac{\Pr(y = 1|\mathbf{x}, \mathbf{g})}{\Pr(y = 0|\mathbf{x}, \mathbf{g})}\right) = \alpha + \mathbf{x}'\gamma + \sum_{k=1}^{C-1} z_k \beta_k \tag{1}$$

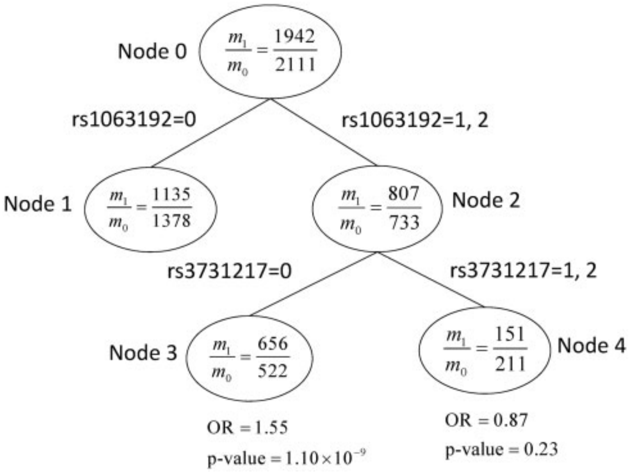
Node 0
$\frac{m_1}{m_0} = \frac{1942}{2111}$
rs1063192=0
rs1063192=1, 2
Node 1
$\frac{m_1}{m_0} = \frac{1135}{1378}$
$\frac{m_1}{m_0} = \frac{807}{733}$
Node 2
rs3731217=0
rs3731217=1, 2
Node 3
$\frac{m_1}{m_0} = \frac{656}{522}$
$\frac{m_1}{m_0} = \frac{151}{211}$
Node 4
OR = 1.55
p-value = 1.10×10^{-9}
OR = 0.87
p-value = 0.23

Fig. 1. The optimal tree model for the gene *CDKN2B* identified by TREAT. m_1/m_0 are the numbers of cases and controls in the node. The odds ratios (ORs) and *P*-values are post hoc estimates based on the standard logistic regression model, with node 1 being the reference group

where γ and β_k, $1 \le k \le C - 1$, are regression coefficients. We need only $C - 1$ dummy variables in (1) to represent the genetic risk factor with C categories. In practice, because the exact number of risk groups and their definition are unknown, the standard $C - 1$ degree-of-freedom chi-squared test for testing the global null hypothesis $H_0 : \beta_k = 0$, $1 \le k \le C - 1$, is not applicable. Instead, we propose to use a tree-structure model to approximate the latent risk group assignment. For example, in the tree-structure model shown in Figure 1, subjects are assigned into three risk groups according to their genotypes.

We begin by describing the testing procedure in the absence of covariates and then extend it to account for covariates. We build the tree model by a sequential binary partitioning of the samples. Figure 1 illustrates a tree model built on the gene *CDKN2B* detected by our test. All subjects in the root node (node 0) are first divided into two offspring nodes (node 1 and 2) according to the splitting rule defined by a binary variable chosen from $\mathbf{f}$. Next, according to the criteria described later, we expand the tree by splitting one of the two offspring nodes (node 1 or 2). In this example, node 2 is chosen. This procedure is continued until certain pre-specified stopping rules are met (described later).

We want to choose a binary variable from $\mathbf{f}$ to split a node in such a way that the difference in the disease risk between the two resultant offspring nodes is maximized. Suppose that we want to expand a tree T_K with K leaf nodes (terminal nodes) $A_k, k = 1, \ldots, K$. For any given pair (A_k, f_j), the samples in A_k can be summarized as the 2×2 contingency table shown in Table 1. We use the following 1-df chi-squared test statistic to measure the 'goodness-of-split' when the node A_k is split by f_j (Agresti, 2007):

$$t^{(kj)} = \frac{n^{(k)}(n_{00}^{(kj)} n_{11}^{(kj)} - n_{01}^{(kj)} n_{10}^{(kj)})^2}{n_0^{(k)} n_1^{(k)} m_0^{(kj)} m_1^{(kj)}} \tag{2}$$

The best pair (the node and the corresponding splitting rule) is therefore defined as:

$$(k_{\mathrm{opt}}, j_{\mathrm{opt}}) = \underset{(k,j)}{\operatorname{argmax}}\, t^{(kj)} \tag{3}$$

The tree T_K is then grown by splitting the leaf node $A_{k_{\mathrm{opt}}}$ according to the rule defined by $f_{j_{\mathrm{opt}}}$, which results to a new tree T_{K+1} with $K + 1$ leaf nodes. This binary partitioning procedure is continued until the number

Table 1. The 2 × 2 contingency table for the samples in the leaf node A_k, cross-tabulated by the outcome y and the binary variable f_j

	$f_j = 0$	$f_j = 1$	Total
$y = 0$	$n_{00}^{(kj)}$	$n_{01}^{(kj)}$	$n_0^{(k)}$
$y = 1$	$n_{10}^{(kj)}$	$n_{11}^{(kj)}$	$n_1^{(k)}$
Total	$m_0^{(kj)}$	$m_1^{(kj)}$	$n^{(k)}$

Table 2. The 2 × k contingency table for all subjects, cross-tabulated by the outcome y and the set of binary variables $F_1, \ldots, F_k$ defined by the leaf nodes of the tree T_k

	$F_1 = 1$	$F_2 = 1$	...	$F_k = 1$	Total
$y = 0$	$n_{01}^{(k)}$	$n_{02}^{(k)}$	...	$n_{0k}^{(k)}$	n_0
$y = 1$	$n_{11}^{(k)}$	$n_{12}^{(k)}$	...	$n_{1k}^{(k)}$	n_1
Total	$m_1^{(k)}$	$m_2^{(k)}$	...	$m_k^{(k)}$	n

of leaf nodes reaches a pre-specified value K_0. To make the procedure more robust, a leaf node that is eligible for splitting has to contain at least N_0 subjects. We set $K_0 = 5$ and $N_0 = 50$ in the simulation studies and real data application. Issues involved in choosing values for K_0 and N_0 will be examined in Section 4.

By growing the tree with the partitioning algorithm described earlier, we can obtain a sequence of $K_0 - 1$ nested tree models $T_2, \ldots, T_{K_0}$, with T_k having k leaf nodes. Each tree model in the sequence provides an estimate for the latent genetic risk groups. For a given tree model T_k, we define a set of binary variables $F_1, ..., F_k$ indicating a subject's assignment to one of the k leaf nodes of T_k, with $F_j = 1, j = 1, \ldots, k$, if the subject falls in the j-th leaf, and $F_j = 0$ otherwise. We model the association between gene and disease as:

$$\text{logit P}(y = 1 | F_1, \ldots, F_{k-1}) = \alpha + \sum_{j=1}^{k-1} \beta_j F_j \tag{4}$$

The departure from the null hypothesis $H_0 : \beta_j = 0, j = 1, \ldots, k - 1$, can be measured by the score test statistic. Table 2 provides notation for the frequency distribution for samples in each leaf node in T_k. We show in Supplementary Materials that the score test statistic based on (4) can be represented as:

$$s_0^{(k)} = \frac{n^2}{n_0 n_1} \sum_{j=1}^{k} \frac{1}{m_j^{(k)}} \left(n_{1j}^{(k)} - m_j^{(k)} \frac{n_1}{n} \right)^2, \quad k = 2, \ldots, K_0 \tag{5}$$

Because the building of the tree model is 'supervised' by the outcome, the definition of $F_1, \ldots, F_k$ is partially driven by the outcome. As a result, $s_0^{(k)}$ no longer has an asymptotic chi-squared distribution. Instead, the empirical significance level of $s_0^{(k)}$, $2 \le k \le K_0$, denoted by $p_0^{(k)}$, can be evaluated through a permutation procedure that generates null datasets by randomly shuffling the outcome among subjects, while keeping their genotypes unchanged. Let the number of permutation steps be M. For the m-th permuted dataset, the same tree growing algorithm is applied, and the score test statistics $(s_m^{(2)}, \ldots, s_m^{(K_0)})$ defined in (5) are obtained for the sequence of the $K_0 - 1$ nested tree models. We can estimate the empirical significance level of $s_0^{(k)}$ by:

$$p_0^{(k)} = \frac{\sum_{m=0}^{M} \text{I}(s_m^{(k)} \ge s_0^{(k)})}{M + 1}, \quad k = 2, \ldots, K_0 \tag{6}$$

Because we have $K_0 - 1$ candidate risk models for assessing the disease–gene association, each having empirical P-value $p_0^{(k)}, k = 2, \ldots, K_0$, a natural choice for the final test statistic is $T_0 = \min\{p_0^{(k)}; k = 2, \ldots, K_0\}$. We identify the tree model with the smallest empirical P-value $p_0^{(k)}$ as the optimal tree model representing the gene's effect. In principle, the P-value of T_0 could be estimated by a computationally intensive two-layer permutation procedure. The inner layer is required for getting the empirical P-value [i.e. $p_0^{(k)}$ in (6)] for the tree models fitted from the observed and permuted datasets, and the outer layer is required for getting the empirical null distribution of T_0. Instead, we adopt the minP algorithm proposed by Ge *et al.* (2003), where only a single-layer permutation procedure is needed. We describe Ge's algorithm in more detail in Supplementary Materials.

2.3 Accelerating the growth of the tree by Boolean operations

As mentioned in the Introduction, testing procedures based on the tree model can be computationally intensive, as the evaluation of the P-value requires a permutation procedure. Even though the computational burden is reduced considerably by the use of closed-form formula for the evaluation of splitting rules (2) and the score test statistics (5), as well as Ge's algorithm, it is still not practical to use the proposed procedure as a gene-based test to screen >20 000 genes, some of which have P-values $< 1.0 \times 10^{-6}$. The main bottleneck is to build the contingency table shown in Table 1, which is required for the evaluation of every candidate-splitting rule in the tree-model-building process on observed and permuted datasets. For any given pair (A_k, f_j), $O(nJ)$ if-else operations on y and $\mathbf{f}$ are required to obtain Table 1. Thus, about $O(nJK_0)$ if-else operations are performed during the growth of a tree with K_0 leaf nodes. The computational overhead for calculating the score statistics in (5) is negligible, as Table 2 can be obtained from Table 1 directly. When the permutation procedure is used with M iterations, the total number of the if-else operations scales up to the order of $O(nJK_0M)$, whereas M can be as large as 10^8 to estimate the P-value for genome-wide significance reliably.

To make our proposed test computationally feasible for large-scale genome-wide gene-based analysis, we adopt the BOolean Operation (BOO) method, which was originally proposed by Wan *et al.* (2010) for detecting SNP–SNP interaction at the genome-wide scale, to accelerate the calculation of contingency table, e.g. Table 1. The BOO method is a memory- and CPU-efficient strategy for constructing tens of millions of contingency tables swiftly. To be more specific, we illustrate the use of the BOO algorithm in our setting. For the binary outcome y, instead of storing it as a vector with n standard integers (a standard integer would be represented by 64 binary digits in a 64-bit computing system), we encode 64 binary outcomes (say $y_1, \ldots, y_{64}$) into one 64-bit integer w according to the following rule:

$$w = \sum_{i=1}^{64} 2^{64-i} y_i$$

Thus, the binary outcome y can be stored as a length-$\lceil n/64 \rceil$ vector y^{b} consisting of 64-bit integers. We also need to maintain a binary vector θ_k, with its i-th entry indicating whether the i-th subject falls in the k-th leaf node. Similar representation can be applied to θ_k and each of the dichotomized genotypes f_j, $1 \le j \le J$. Hence, θ_k and f_j are converted to length-$\lceil n/64 \rceil$ vectors θ_k^{b} and f_j^b as well. We can obtain a length-$\lceil n/64 \rceil$ vector $v_{kj}^{\text{b}} = y^{\text{b}} \& \theta_k^{\text{b}} \& f_j^{\text{b}}$, where & represents the CPU-efficient logical

Table 3. The 2×2 contingency table for the subjects in the leaf node A_k with $x_l = 1$, cross-tabulated by the outcome y and the binary variable f_j

	$f_j = 0$	$f_j = 1$	Total
$y = 0$	$n_{00}^{(kjl)}$	$n_{01}^{(kjl)}$	$n_0^{(kl)}$
$y = 1$	$n_{10}^{(kjl)}$	$n_{11}^{(kjl)}$	$n_1^{(kl)}$
Total	$m_0^{(kjl)}$	$m_1^{(kjl)}$	$n^{(kl)}$

AND operation in an element-wise manner. The numbers in Table 1 can be obtained by counting and summing over the '1' bits in each of the entries of v_{kj}^{b}. Supplementary Figure S1 illustrates an example showing how the BOO method works when the total sample size is 64. The tree-based testing procedure incorporating the BOO method is roughly 60-fold faster than the one using the standard if-else operations. More discussion of the computing time will be given in Supplementary Materials.

2.4 Adjusting covariates

In this section, we extend the proposed testing procedure by allowing the adjustment of covariates. To make the BOO method applicable in this situation, we need to extend the statistics in (2) and (5) accordingly, so that they are closed-form functions of subject frequencies under various restrictions. We assume that the adjusted covariates can be discretized and represented by a multilevel categorical variable. For example, if we wanted to adjust two covariates with two and three levels, respectively, we would adjust these two covariates jointly as a categorical variable with six levels. Equivalently, both main and interactive effects from the two adjusted covariates are adjusted. Under such a strategy, we can represent all adjusted covariates as a factor encoded by a set of dummy variables $\mathbf{x} = (x_1, \ldots, x_L)$.

For any candidate leaf node A_k to be split, the samples in A_k can be summarized as L 2×2 contingency tables (Table 3 shows the l-th table, for those samples with $x_l = 1$). We extend (2) to allowing the adjustment of $\mathbf{x}$ by use of the Cochran–Mantel–Haenszel test statistic (Mantel, 1963; Mantel and Haenszel, 1959). Define the test statistic:

$$t^{(kj)} = \frac{\left(\sum_{l=1}^{L}\left(n_{11}^{(kjl)} - m_1^{(kjl)}\frac{n_1^{(kl)}}{n^{(kl)}}\right)\right)^2}{\sum_{l=1}^{L} m_1^{(kjl)} m_0^{(kjl)} \frac{n_0^{(kl)} n_1^{(kl)}}{(n^{(kl)})^3}} \tag{7}$$

The best split is again chosen as $(k_{\mathrm{opt}}, j_{\mathrm{opt}}) = \mathrm{argmax}_{(k,j)}\, t^{(kj)}$.

We also modify the score statistics $s_0^{(k)}$ in (5) for testing the association between the gene and the disease condition by assuming the following risk model:

$$\mathrm{logit}\ \mathrm{P}(y = 1 | x_1, \ldots, x_L, F_1, \ldots, F_k) = \alpha + \sum_{l=1}^{L-1} \gamma_l x_l + \sum_{j=1}^{k-1} \beta_j F_j \tag{8}$$

For any subtree T_k with k leaves, $k = 2, \ldots, K_0$, the samples can be summarized as L $2 \times k$ contingency tables (Table 4 shows the l-th table for those samples with $x_l = 1$). To test the null hypothesis $H_0 : \beta_j = 0$, $j = 1, \ldots, k-1$, we can derive the score test statistic under the risk model (8) as:

$$s_0^{(k)} = U_k' V_k^{-1} U_k \tag{9}$$

where

$$U_k = (u_{k1}, \ldots, u_{k,k-1})', \quad V_k = \left(v_{j_1 j_2}\right)_{(k-1)\times(k-1)}$$

Table 4. The $2 \times k$ contingency table for subjects with $x_l = 1$, cross-tabulated by the outcome y and the binary variables $F_1, \ldots, F_k$ defined by the leaf nodes of the tree T_k

	$F_1 = 1$	$F_2 = 1$	...	$F_k = 1$	Total
$y = 0$	$n_{01}^{(kl)}$	$n_{02}^{(kl)}$	...	$n_{0k}^{(kl)}$	$n_0^{(l)}$
$y = 1$	$n_{11}^{(kl)}$	$n_{12}^{(kl)}$	...	$n_{1k}^{(kl)}$	$n_1^{(l)}$
Total	$m_1^{(kl)}$	$m_2^{(kl)}$	...	$m_k^{(kl)}$	$n^{(l)}$

$$u_{kj} = \sum_{l=1}^{L}\left(n_{1j}^{(kl)} - m_j^{(kl)}\frac{n_1^{(l)}}{n^{(l)}}\right), \quad j = 1, \ldots, k-1$$

$$v_{j_1 j_2} = -\sum_{l=1}^{L} m_{j_1}^{(kl)} m_{j_2}^{(kl)} \frac{n_0^{(l)} n_1^{(l)}}{(n^{(l)})^3}, \quad 1 \le j_1 \ne j_2 \le k-1$$

$$v_{jj} = \sum_{l=1}^{L} m_j^{(kl)}\left(n^{(l)} - m_j^{(kl)}\right)\frac{n_0^{(l)} n_1^{(l)}}{(n^{(l)})^3}, \quad j = 1, \ldots, k-1$$

The significance of this extended tree-based multilocus test can be evaluated by a permutation procedure in which we randomly shuffle the outcome y within each stratum defined by the dummy covariate $\mathbf{x}$, while keeping the genotypes and $\mathbf{x}$ the same within each subject. The steps for estimating the P-value remain the same. With minor modification, the BOO method described earlier can be applied. For example, instead of applying the CPU-efficient logical AND operations on y^{b}, θ_k^{b} and f_j^{b}, we apply them on y^{b}, θ_k^{b}, f_j^{b} and x_l^{b}, with x_l^{b} being the vector of 64-bit integers converted from the vector of x_l.

3 RESULTS

3.1 Application to a GWA study of ESCC

We demonstrated the application of our TREe-based Association Test (TREAT) as a gene-based test by applying it on a GWA study of ESCC. We focused on 1942 cases and 2111 controls taken from the Shanxi Upper Gastrointestinal Cancer Genetics Project and the Linxian Nutrition Intervention Trials (Abnet *et al.*, 2010). Both studies were conducted in the Taihang mountain area in China. The analysis was conducted on 22 193 genes or annotated regions extracted by the software GLU (http://code.google.com/p/glu-genetics/). We set the threshold for genome-wide significance at 2.3×10^{-6} ($\approx 0.05/22\,193$) according to the Bonferroni correction adjusting for all 22 193 tests. With the 1000 Genomes data (version 3) as the reference (1000 Genomes Project Consortium *et al.*, 2012), we used the software IMPUTE2 (Howie *et al.*, 2009) to impute the missing genotypes. Besides TREAT, we also applied several other multilocus tests, such as the Min-P test (Seaman and Muller-Myhsok, 2005), the Adaptive Joint test (AdaJoint) (Zhang *et al.*, 2013), the Adaptive Rank Truncated Product test (ARTP) (Yu *et al.*, 2009), the SNP set association test (SKAT) (Wu *et al.*, 2010) and its variation (SKAT-O) (Lee *et al.*, 2012).

All tests considered here were adjusted for study, sex and an indicator for whether a subject was younger than 40 years. Those covariates were chosen because of their significant marginal

Table 5. Testing results on the GWA studies of ESCC

Gene	Location	Size	Min *P*-value	Strongest SNP	TREAT	Min-*P*	ARTP	AdaJoint	SKAT	SKAT-O-5	SKAT-O-10
CDKN2B	9p21.3	6	3.8e-6	rs1063192	**6.0e-8**	4.1e-5	3.0e-5	2.2e-5	2.4e-4	2.8e-4	2.8e-4
CDKN2A	9p21.3	10	3.8e-6	rs1063192	**1.9e-7**	6.8e-5	2.2e-5	5.8e-5	5.1e-5	4.4e-5	4.4e-5
MTAP	9p21.3	30	3.8e-6	rs1063192	**1.3e-6**	1.9e-4	9.3e-5	4.6e-5	1.2e-3	2.4e-3	2.4e-3
PLCE1	10q23.33	34	5.5e-8	rs3781264	6.4e-5	3.4e-6	**2.0e-6**	8.5e-6	9.7e-4	1.7e-3	1.7e-3
CR599144	10q23.33	3	1.4e-7	rs12263737	1.3e-5	**6.9e-7**	**1.2e-7**	**1.5e-6**	**7.1e-8**	**9.1e-8**	**3.2e-8**
KIAA1516	10q23.33	5	5.5e-8	rs3781264	8.9e-6	**5.7e-7**	**1.0e-7**	**1.2e-6**	**1.1e-7**	**4.1e-8**	**5.1e-8**
HSCB	22q12.1	6	1.3e-7	rs738722	**2.1e-6**	**1.7e-6**	**1.8e-6**	**1.3e-7**	3.7e-5	3.4e-5	3.4e-5
CHEK2/Chk2	22q12.1	8	1.3e-7	rs738722	4.5e-5	2.4e-6	**1.9e-6**	**1.2e-6**	1.2e-4	2.2e-4	2.3e-4
CASP8	2q33.1	8	3.3e-6	rs13016963	4.2e-5	2.7e-5	5.8e-5	**2.4e-7**	1.0e-3	1.7e-3	1.7e-3

Note: These are genes on which at least one of the considered tests produces a $P < 2.3 \times 10^{-6}$ (bold values). The *P*-values of TREAT, Min-*P*, ARTP and AdaJoint tests are estimated with 10^9 replicates of permutation.

effects ($P < 0.05$) on the outcome. After imputation, SNPs with missing rate >5% or minor allele frequencies (MAFs) <5% were excluded from the analysis. For two SNPs with pairwise linkage disequilibrium (LD) coefficient >0.95, the one with a smaller MAF was abandoned. In the end, we had 159 046 unique SNPs in 22 193 genes and/or annotated regions. The inflation factor of the single-marker test on those SNPs is 1.037, which suggests that there is no major effect due to population stratification after adjusting for the three covariates. Adjusting for the top 10 eigenvectors yields a similar inflation factor of 1.038. Therefore, we did not adjust eigenvectors in subsequent analysis. For TREAT, we considered a tree with up to five leaf nodes. A node was considered for splitting if it contained at least 50 samples. For AdaJoint and ARTP, the optimal models were built by checking the combinations of the top 1, 2,..., 5 SNPs. For SKAT-O, the *P*-values were calculated using 5 or 10 points of equal-sized grids searching for the weight parameter from 0 to 1 (denoted as SKAT-O-5 and SKAT-O-10, respectively). The *P*-values of the Min-*P*, AdaJoint, ARTP and TREAT were calculated by 10^5 replicates of permutation. For genes with initially estimated *P*-values $< 1.0 \times 10^{-4}$, we further refined their *P*-values by 10^9 replicates of permutation.

Table 5 lists the genes with at least one *P*-value produced by the considered tests less than the genome-wide threshold 2.3×10^{-6}. TREAT identified a novel gene *CDKN2B* in the region 9p21.3 with *P*-value 6.0×10^{-8}, much lower than the threshold for the family-wise significance level. For this gene, the Min-*P*, ARTP, AdaJoint, SKAT, SKAT-O-5 and SKAT-O-10 tests reported far less impressive *P*-values, which were 4.1×10^{-5}, 3.0×10^{-5}, 2.2×10^{-5}, 2.4×10^{-4}, 2.8×10^{-4} and 2.8×10^{-4}, respectively. Figure 1 illustrates the optimal tree model for the gene *CDKN2B*. The optimal tree model defines a risk factor with three levels, corresponding to leaf nodes 1, 3 and 4. We fitted a post hoc logistic regression model on this derived risk factor. With leaf node 1 as the reference group, the ORs for leaf nodes 3 and 4 were 1.55 ($P = 1.10 \times 10^{-9}$) and 0.87 ($P = 0.23$), respectively. Although these estimates are biased, as the risk factor is defined and analyzed within the same dataset, it nonetheless suggests the genotype group defined by the leaf node 3 is the high-risk group. In the same region, TREAT also identified another two genes *CDKN2A* ($P = 1.9 \times 10^{-7}$) and *MTAP* ($P = 1.3 \times 10^{-6}$), one having four overlapping SNPs with *CDKN2B*, the other containing *CDKN2B*. The other six tests failed to detect them. Among the three known ESCC-associated regions at 10q23.33, 22q12.1 (Abnet *et al.*, 2010) and 2q33.1 (Abnet *et al.*, 2012), TREAT identified gene *HSCB* ($P = 2.1 \times 10^{-6}$) at 22q12.1. In general, the ARTP and AdaJoint tests, each of which found five associated genes that exceeded the genome-wide threshold, appear to have the most success in these three established regions.

3.2 Simulation Studies

We conducted extensive simulation studies to compare the performance between the proposed TREAT and several commonly used approaches, including Min-*P*, ARTP, AdaJoint and SKAT. Results of SKAT-O-5 and SKAT-O-10 are similar to that of SKAT, and thus are not shown here.

First, we used observed genotypes at the six SNPs within the gene *CDKN2B* in the GWA study of ESCC described earlier as a template to simulate data. Each simulated dataset consisted of 3000 cases and 3000 controls. Genotypes of controls were directly sampled from the joint genotype distribution observed in the ESCC study. We generated the genotypes of cases by sampling from the ESCC data with the weight for each subject specified by the assumed risk model [see Yu *et al.* (2009) for more details on how the genotypes were assigned]. We considered the risk model used in Figure 1, which can be represented as:

$$\text{logit}\Pr(y = 1 \mid I_3, I_4) = \alpha + \beta_3 I_3 + \beta_4 I_4 \qquad (10)$$

where I_3 (I_4) is the indicator variable on whether a sample falls in the third (fourth) leaf node in the tree shown in Figure 1. We set $\beta_3 = \beta_4 = 0$ in (10) under the null hypothesis, and generated 10 000 datasets for assessing type I errors of all considered tests under the significance level at 0.05 and 0.001. Table 6 shows that all considered tests can properly maintain their empirical type I errors. More results about the empirical type I error of TREAT applied to genes with various sizes are given in the Supplementary Table S1.

We set $\beta_3 = 0.5$ and $\beta_4 = -0.2$ in (10) and generated 1000 datasets for the power comparison. Power for various tests at

Table 6. The empirical type I errors at the levels of 0.05 and 0.001

Level	TREAT	Min-p	ARTP	AdaJoint	SKAT
.05	.047	.046	.045	.052	.047
.001	.0007	.0009	.0007	.0012	.0011

Table 7. Power comparisons under the risk model (10)

Level	TREAT	Min-p	ARTP	AdaJoint	SKAT
.05	.855	.756	.749	.728	.690
.001	.483	.237	.229	.233	.116

the levels of 0.05 and 0.001 is summarized in Table 7. It is obvious that TREAT has a clear power advantage over other tests under this scenario.

Next, we conducted simulations by generating genotypes under various LD structures. Instead of relying on observed joint genotype distribution in an existing study, we simulated genotypes at multiple SNPs in a study population according to the algorithm in the study by Wang and Elston (2007). We considered a gene with 20 SNPs, and a study with 2000 cases and 2000 controls. We generated vectors of continuous variables $U = (U_1, \ldots, U_{20})'$ under a multivariate normal distribution with mean zero and the covariance matrix $\Sigma = (\sigma_{ij})_{20\times 20}$, where $\sigma_{ij} = \rho^{|i-j|}$. We then generated genotypes at the i-th SNP by discretizing U_i into three levels corresponding to genotypes 0, 1 and 2, so that the resultant SNP had a minor allele frequency of 0.4. The correlation among genotypes of the 20 SNPs was controlled by ρ. In the simulation, we considered $\rho = 0$, 0.5 and 0.9 to represent a low, median and high LD relationship among SNPs, respectively. We used this procedure to generate genotypes for controls, and we used the weighted sampling procedure to generate genotypes for cases according to the assumed risk model.

We designated the 10th and 11th SNPs as the risk SNPs, and we defined the two following indicator variables to represent the way that the two risk SNPs affect the outcome:

$$I_1 = \begin{cases} 1, & \text{if } G_{10} = 1 \text{ or } 2 \text{ and } G_{11} = 0 \text{ or } 1 \\ 0, & \text{otherwise,} \end{cases}$$

and

$$I_2 = \begin{cases} 1, & \text{if } G_{10} = 1 \text{ or } 2 \text{ and } G_{11} = 2 \\ 0, & \text{otherwise.} \end{cases}$$

The risk model was defined as

$$\text{logit } \Pr(y = 1 \mid G_{10}, G_{11}) = \alpha + \beta_1 I_1 + \beta_2 I_2 \tag{11}$$

This model can be represented as a tree model with three leaf nodes. Table 8 compares the power of all considered tests under various scenarios, which were specified by ρ, β_1 and β_2. The power was evaluated at the level 0.05 based on 1000 replicated datasets simulated under each scenario. It is evident from Table 8 that TREAT is the most powerful under model (11). The Min-P, ARTP and AdaJoint tests performed similarly to each other and showed some advantage over SKAT.

Table 8. Power comparisons under the risk model (11)

Setting	TREAT	Min-p	ARTP	AdaJoint	SKAT
$\beta_1 = 0,\ \beta_2 = \log(1.5)$					
$\rho = .0$	.644	.481	.450	.456	.359
$\rho = .5$	.936	.783	.790	.736	.674
$\rho = .9$	.991	.927	.946	.882	.872
$\beta_1 = -\log(1.2),\ \beta_2 = \log(1.2)$					
$\rho = .0$	.568	.314	.320	.340	.283
$\rho = .5$	.691	.233	.257	.354	.137
$\rho = .9$	.812	.195	.166	.212	.122

Table 9. Power comparisons under the risk model (12)

Setting	TREAT	Min-p	ARTP	AdaJoint	SKAT
$\rho = .0$	.338	.201	.243	.251	.202
$\rho = .5$	.732	.633	.650	.544	.541
$\rho = .9$	.960	.887	.900	.866	.886

The risk models considered in Table 8 are essentially tree models. We then conducted simulation under the following two-locus epistasis risk model:

$$\begin{aligned} \text{logit } \Pr(y = 1 \mid G_{10}, G_{11}) = \alpha + \beta_1 I(G_{10} = 2) + \beta_2 I(G_{11} = 2) \\ + \gamma I(G_{10} = G_{11} = 2) \end{aligned} \tag{12}$$

with $\beta_1 = \beta_2 = \gamma = 0.15$. Power results based on 1000 replicates are summarized in Table 9. Again, TREAT outperformed all other tests under the risk model (12).

Finally, we compared the performance of all considered tests under traditional linear additive models (on the logit scale) with one, two or three disease risk SNPs. The results are summarized in Supplementary Table S2. As TREAT is not designed to detect additive effects, it is not a surprise to see that TREAT has a less favorable performance compared with other tests, such as ARTP, AdaJoint and SKAT, which aim at the linear additive risk model. In practice, we recommend the use of TREAT in conjunction with one of the tests that target detection of additive joint effects.

3.3 Computing time

Here, we provide a summary of computing time when applying TREAT on a gene with 5, 10, 20, 50 and 100 SNPs in a study consisting of 2000 cases and 2000 controls. We assume that the adjusted covariates are treated as a 5-level categorical variable.

We set $K_0 = 5$ and $N_0 = 50$ and used 10^5 replicates of permutation when applying TREAT. In Supplementary Table S3, we report the averaged running time over 100 experiments for each scenario on a Linux machine with a 2.8-GHz Xeon CPU. It appears that it is practical to run TREAT as a routine test, as the computing time is reasonable.

4 DISCUSSION

The linear regression model, such as the logistic regression model, has been the dominant tool for assessing the joint effect of multiple predictors. The following three factors contribute to its popularity. First, the linear regression is convenient to use, as the asymptotic distributions have been established for derived test statistics. Second, it is generally accepted that the linear model is usually adequate for depicting the joint effect. Third, there are not many alternative options available. There is clearly a demand for more robust testing procedures, especially in situations when the cumulative effect from multiple predictors is non-additive. Here, we have proposed a general testing procedure based on the tree-structure model. Leveraging a computationally efficient tree-building algorithm, the procedure has limited requirements for CPU time and memory. For a gene with 20 SNPs in a study of 2000 cases and 2000 controls, it takes 45 s to run the procedure with 10^5 permutation steps on a Linux machine with a 2.8-GHz Xeon CPU. We have demonstrated the power advantage for the proposed procedure in the setting of gene-based association analysis through real and simulated datasets.

The proposed testing procedure is general and can be used for testing the association between a set of risk factors and a binary outcome. It can adjust the covariates' effects by treating them as a categorical variable with multiple levels. It has two major tuning parameters that have to be set by the user: the maximum tree sizes K_0 (the number of leaf nodes) and the minimum sample size N_0 required for each leaf node to be split. In practice, the maximum tree size depends on the sample size of the study, the number of risk factors and the prior knowledge on the complexity of the underlying risk model. For the purpose of hypothesis testing, we favor a relatively small tree size, as we want to limit the number of considered candidate risk models, which are subtrees of the full tree with the maximum number of leaf nodes, to control multiple-comparison effects. On the other hand, the maximum tree size has to be large enough to allow the searching space to be sufficiently broad to cover a model that adequately approximates the true risk model. In the setting of gene-based analysis, we recommend choosing K_0 in the range of 5 to 10, under the assumption that the number of true disease-associated loci within a gene tends to be relatively small. The choice of the minimum sample size N_0 required in each leaf node is driven mainly by the total sample size. Because each leaf node represents a candidate risk group, we suggest $N_0 \geq 50$ for robust testing results.

The computational burden associated with the tree-structure model is one major barrier to its use for hypothesis testing. We overcame this obstacle by combining several strategies to speed up the proposed procedure, including the derivation of closed-form formulas for evaluating splitting rules and score test statistics for assessing the joint effect summarized by the tree-structure model, the use of Ge's one-layer permutation algorithm (Ge *et al.*, 2003) for the evaluation of the *P*-value and the adaptation of the BOO method (Wan *et al.*, 2010) for swift scanning over candidate splitting rules. We demonstrated that it is feasible to apply the proposed procedure as a multilocus association test in large-scale gene-based association studies with >20 000 candidate genes or annotated regions. The tree-building algorithm can be accelerated further by noticing that the evaluation of splitting rules at each given tree node is parallelizable. We are currently working on a GPU (graphic processing unit) implementation to take advantage of the parallelizability (Greene *et al.*, 2010; Yung *et al.*, 2011).

We developed a general testing procedure based on the tree-structure model and presented a fast algorithm for building the tree-structure model by adopting the computationally efficient Boolean operator for swift evaluation of the splitting rule. This procedure has broad applications, and it is a valuable tool that can be run routinely in conjunction with other tests that target an additive joint effect. We have created an R package implementing the procedure (http://www.hanzhang.name/softwares/treat).

ACKNOWLEDGEMENT

We thank the anonymous referees for their helpful comments. This study used the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD. (http://biowulf.nih.gov).

Funding: The work of H.Z., P.R.T. and K.Y. was supported by the Intramural Program of the National Institutes of Health, the National Cancer Institute and the Division of Cancer Epidemiology and Genetics.

Conflict of Interest: none declared.

REFERENCES

Abnet,C.C. *et al.* (2010) A shared susceptibility locus in *PLCE1* at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nat. Genet.*, **42**, 764–767.

Abnet,C.C. *et al.* (2012) Genotypic variants at 2q33 and risk of esophageal squamous cell carcinoma in china: a meta-analysis of genome-wide association studies. *Hum. Mol. Genet.*, **21**, 2132–2141.

Agresti,A. (2007) An introduction to categorical data analysis. In: *Wiley Series in Probability and Mathematical Statistics*. Vol. 423, 2 edn. John Wiley & Sons, Hoboken, NJ.

Breiman,L. *et al.* (1984) *Classification and Regression Trees.* Wadsworth & Brooks, Monterey, CA.

Chen,J. *et al.* (2007) A partially linear tree-based regression model for assessing complex joint gene-gene and gene-environment effects. *Genet. Epidemiol.*, **31**, 238–251.

Ge,Y. *et al.* (2003) Resampling-based multiple testing for microarray data analysis. *Test*, **12**, 1–77.

Greene,C.S. *et al.* (2010) Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic als. *Bioinformatics*, **26**, 694–695.

Howie,B.N. *et al.* (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.

Lee,S. *et al.* (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, **13**, 762–775.

Mantel,N. (1963) Chi-square tests with one degree of freedom; extensions of mantel-haenszel procedure. *J. Am. Stat. Assoc.*, **58**, 690–700.

Mantel,N. and Haenszel,W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl Cancer Inst.*, **22**, 719–748.

Seaman,S.R. and Muller-Myhsok,B. (2005) Rapid simulation of *P* values for product methods and multiple-testing adjustment in association studies. *Am. J. Hum. Genet.*, **76**, 399–408.

Wan,X. *et al.* (2010) BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.*, **87**, 325–340.

Wang,T. and Elston,R.C. (2007) Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am. J. Hum. Genet.*, **80**, 353–360.

Wu,M.C. *et al.* (2010) Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 929–942.

Yu,K. *et al.* (2009) Pathway analysis by adaptive combination of *P*-values. *Genet. Epidemiol.*, **33**, 700–709.

Yu,K. *et al.* (2007) Two-sample comparison based on prediction error, with applications to candidate gene association studies. *Ann. Hum. Genet.*, **71**, 107–118.

Yung,L.S. *et al.* (2011) GBOOST: a GPU-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics*, **27**, 1309–1310.

Zhang,H. *et al.* (2014) A fast multilocus test with adaptive SNP selection for large-scale genetic association studies. *Eur. J. Hum. Genet.*, **22**, 696–702.

Zhang,H. and Singer,B. (1999) *Recursive Partitioning in the Health Sciences*. Springer Verlag, New York.

1000 Genomes Project Consortium *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.