Original Paper

Human Hered 2011;71:161–170
DOI: 10.1159/000327299

Received: October 27, 2010
Accepted after revision: March 10, 2011
Published online: July 20, 2011

A Non-Parametric Method for Building Predictive Genetic Tests on High-Dimensional Data

Chengyin Ye a, b Yuehua Cui c Changshuai Wei b Robert C. Elston d Jun Zhu e Qing Lu b

a College of Life Sciences, Zhejiang University, Hangzhou, Zhejiang, PR China; Departments of 
b Epidemiology and c Statistics and Probability, Michigan State University, East Lansing, Mich., and 
d Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA; 
e Institute of Bioinformatics, Zhejiang University, Hangzhou, Zhejiang, PR China

Key Words
High-dimensional data · Disease prediction · Optimal ROC curve

Abstract
Objective: Predictive tests that capitalize on emerging genetic findings hold great promise for enhanced personalized healthcare. With the emergence of a large amount of data from genome-wide association studies (GWAS), interest has shifted towards high-dimensional risk prediction. Methods: To form predictive genetic tests on high-dimensional data, we propose a non-parametric method, called the ‘forward ROC method’. The method adopts a computationally efficient algorithm to search for environment risk factors, genetic predictors on the entire genome, and their possible interactions for an optimal risk prediction model, without relying on prior knowledge of known risk factors. An efficient yet powerful procedure is also incorporated into the method to handle missing data. Results: Through simulations and real data applications, we found our proposed method outperformed the existing approaches. We applied the new method to the Wellcome Trust rheumatoid arthritis GWAS dataset with a total of 460,547 markers. The results from the risk prediction analysis suggested important roles of HLA-DRB1 and PTPN22 in predicting rheumatoid arthritis. Conclusion: We proposed a powerful and robust approach for high-dimensional risk prediction. The new method will facilitate future risk prediction that considers a large number of predictors and their interaction for improved performance.

Introduction

Genetic variants have been recognized as having a significant impact on human diseases. In the field of medical care and public health, predictive genetic tests formed with multiple genetic variants are attracting more interest. These tests can be conducted to assess an individual’s future risk of disease, so that early preventive intervention can be adopted to reduce morbidity and mortality [1]. Predictive genetic tests have been recognized as the cornerstone of much of genomic medicine, and are expected to play a key role in transforming the emphasis in the future of healthcare from disease treatment to disease prevention [2]. In the last decade, predictive genetic tests have been successfully built for several human diseases.
with significant hereditary effects, including, for example, multiple endocrine neoplasia type 2 and hereditary breast cancer [3, 4].

Common complex diseases are likely caused by the interplay of a large number of genetic and environmental risk factors, each with a medium- or small-effect size. It has been shown elsewhere that, for common complex diseases, combining multiple genetic variants could substantially increase a test’s predictive value [5]. With the rapid advances in high-throughput technologies, genome-wide association studies (GWAS) have been widely undertaken to study the genetic etiology of common complex diseases. The novel loci discovered through these studies (e.g., more than 30 loci have been identified as predisposing to Crohn’s disease [6]) can be used in forming predictive genetic tests. Alternatively, considering a much wider range of potential risk loci (e.g., those with medium- to small-effect sizes), or even the entire genome, could lead to the formation of more effective tests. Combining such a large number of predictors brings both new opportunities for better disease prediction and statistical challenges.

In previous studies, several existing statistical methods have been adopted to form predictive genetic tests based on multiple genetic variants. Among these, the allele-counting method has been widely used, owing to its easy implementation and ability to handle a large number of loci [7, 8]. This method summarizes the number of risk alleles at each locus into a one-dimensional risk score, and uses this risk score for disease prediction. However, it assumes an equal effect size for all of the risk alleles within or across loci. Such an assumption will not hold when disease-associated loci follow dominant/recessive modes of inheritance, are in linkage disequilibrium, or interact with each other. Furthermore, the allele-counting method cannot be applied directly to whole genome data, since the majority of genetic loci are not risk loci and should be excluded. The classification and regression tree (CART) method is another non-parametric method commonly used for prediction and classification [9–11]. The CART method has the advantage of identifying important interactions and providing easily interpretable results, but has been found to have a low prediction value [12]. Furthermore, the CART method currently lacks the software for dealing with a whole genome dataset.

Recently, methods using optimal receiver operating characteristic (ROC) curves have been proposed [13, 14] as a way to combine multiple risk factors for disease prediction. These methods suggest that tests building on likelihood ratios (LRs) reach the highest performance [15–17]. However, the methods were designed to deal with only a relatively small number of loci and are not capable of handling high-dimensional data (e.g., whole genome-wide data). In addition to this limitation, the methods also lack ways of dealing with missing genotypes. Although this is less of an issue for a small number of loci, it becomes critical for high-dimensional data, which are likely to have a significant proportion of the samples missing at least one genotype. Simply discarding those individuals could dramatically reduce the study sample size, while applying commonly used statistical tools to a large number of loci, such as multiple imputation [18], would be time consuming and computationally intensive. To target these issues, we propose in this work a new non-parametric method we call the ‘forward ROC method’. This method is based on the concept of optimal ROC curves, and thus theoretically forms a test with the highest accuracy. In contrast to previous methods, it adopts a forward-selection algorithm and efficiently searches for the best prediction model on the basis of a large number of genetic and environmental risk factors, or even all the available loci across the entire genome. Moreover, we have introduced a computationally efficient, yet powerful, approach to handle missing values into the method, thus making it feasible to handle missing values in high-dimensional data.

We evaluated our proposed forward ROC method through simulation studies and found that the new method performed better than the commonly used CART and allele-counting methods in a variety of scenarios. Finally, we applied our proposed method to a real data application, the Wellcome Trust Case Control Consortium (WTCCC) rheumatoid arthritis (RA) GWAS dataset, and built a whole genome-wide predictive genetic test for RA.

Methods

Optimal ROC Curve

The ROC curve has been widely used in the medical field to measure the accuracy of a test. The ROC curve plots the sensitivity of a test against 1 – specificity, by continually varying threshold values over the entire range of possible test results. By considering all possible pairs of sensitivity and specificity, the ROC curve gives a global description of a test’s accuracy. For convenience, the ROC curve is often summarized by a one-dimensional summary accuracy index – the area under the ROC curve (AUC) – that ranges from 0.5 to 1. A test with an AUC value of 0.5 has no classification accuracy, while one with an AUC value of 1 has the highest possible classification accuracy [19, 20].
The optimality of a ROC curve can be attained by plotting the LRs of test results from their largest to smallest value [15, 17]. The LR of a particular test result is defined as

\[
LR(x) = P(x \mid S = 1)/P(x \mid S = 0),
\]

where \(x\) indicates the test result and \(S\) represents the disease status: \(S = 1\) (\(S = 0\)) if an individual develops (does not develop) the disease. The test formed by the optimal ROC curve has the highest accuracy in terms of the AUC, owing to the ideal properties of each point on the optimal ROC curve, i.e., it has the highest sensitivity (specificity) for a fixed value of specificity (sensitivity).

The Forward ROC Method

Based on the concept of the optimal ROC curve, we now introduce the forward ROC method for building predictive genetic tests. The method adopts a computationally efficient algorithm – the forward selection algorithm – to perform the model selection. Moreover, it incorporates a computationally efficient, yet powerful, way of handling missing values, and ultimately uses the maximum amount of data information available.

Forward Selection Algorithm

The forward selection algorithm is an iterative algorithm that gradually searches and adds impactful predictors into the prediction model to classify individuals into different risk groups. To illustrate the method, assume that we have \(n\) individuals and \(p\) disease-related loci. For a particular locus \(j\), we assume there are a total of \(K\) genotypes, denoted by \(g^j_1, g^j_2, \ldots, g^j_K\). Note that, although the illustration below is specified in a genetic setting, the same procedure is also valid for environmental risk factors, where the \(K\) different risk levels of the \(j\)-th particular environmental factor can be denoted by \(e^j_1, e^j_2, \ldots, e^j_K\). We assume risk factors can be categorized. Typically, little information is lost if continuous data are divided into a small number of categories, such as quantiles. In fact, this categorization may be desirable if there are outliers, or if the measurements are imprecise [21]. We initiate the process by treating all individuals as belonging to one risk group.

In step 1, we consider all of the possible ways of classifying the individuals into two risk groups, by evaluating the effect of their carrying or not carrying a particular genotype at just one locus. For instance, assuming locus \(j\) has three possible genotypes – \(g^j_1, g^j_2, \text{and} g^j_3\) – there are three possible non-degenerate ways to classify the individuals into two risk groups: (1) \((g^j_1, g^j_3), (g^j_2)\); (2) \((g^j_1, g^j_2), (g^j_3)\), and (3) \((g^j_1), (g^j_2, g^j_3)\). For each one of these three classifications, we calculate the LR for the two risk groups as

\[
LR(G) = P(G_i \mid S = 1)/P(G_i \mid S = 0), \quad i = 1, 2,
\]

where \(G_i\) represents a risk group (e.g., \((g^j_1, g^j_3), (g^j_2)\)) and \(i\) represents the risk group index. By ordering the LRs of the two risk groups, we can calculate the AUC value of this particular classification by

\[
AUC = \frac{1}{2} \sum_{\xi = 1}^{L} \left[ \sum_{i = 1}^{L} P(G_i \mid S = 1) + \sum_{i = 1}^{L} P(G_i \mid S = 1) \right]
\]

\[
\times \left[ \sum_{i = 1}^{L} P(G_i \mid S = 0) - \sum_{i = 1}^{L} P(G_i \mid S = 0) \right]
\]

\(\xi = 1, \ldots, L, \quad l = 1, \ldots, L,\)

where \(G_{(i)}\) is the multi-locus genotype of the \(\xi\)-th risk group after ordering. \(L\) represents the total number of risk groups (e.g., 2 at step 1) and \(P(G_{(0)} \mid S = 1) = P(G_{(0)} \mid S = 0) = 0\). If there are \(p\) loci, this leads to \(3^p\) non-degenerate possible classifications for generating two risk groups at this stage of step 1. We choose the classification with the highest AUC value (i.e., the one giving the most accurate discrimination) as our candidate prediction model at step 1, denoted by \(M_1\). Assuming the model reaches the highest AUC at locus \(j\), it simply classifies individuals into two risk groups, \(G^j_1, G^j_2\), according to their genotypes at that locus \(j\).

In step 2, in order to construct a more complex prediction model with higher classification accuracy, we again scan over all of the \(p\) loci, and search for a second locus to further split the existing two risk groups \(G^j_1, G^j_2\) in a binary fashion. Assuming, among all loci, locus \(j'\) further splits the individuals into the four risk groups \(G^{j'j}_1, G^{j'j}_2, G^{j'j}_3, G^{j'j}_4\) and attains the highest AUC, we choose this partition as the candidate prediction model at step 2, denoted by \(M_2\).

A special case occurs when we use the same locus \(j\) to further split the risk groups \(G^j_1, G^j_2\) (i.e., \(j' = j\)) and there are only three genotypes at locus \(j\). In this particular case, the prediction model contains only three risk groups, \(G^j_1, G^j_2, G^j_3\), and no further split can be processed at locus \(j\). As a result, the algorithm will stop selecting among genotypes at locus \(j\) in the remaining steps. It is easy to show that \(M_2\) has higher classification accuracy than \(M_1\) in terms of the AUC value; but it is more complex than \(M_1\) in terms of the number of risk groups.

The whole splitting process could be continued until the AUC value reaches 1 and a series of prediction models, \(M_1, M_2, \ldots, M_N\), with different levels of model complexity are obtained. Among these models, \(M_1\) would be the simplest model with the lowest AUC value, and \(M_N\) would represent the full model with the largest number of risk groups and the highest possible AUC value of 1. Although the more complex models will appear to have higher classification accuracy, with more risk groups they are more likely to overfit the data. To determine the appropriate model complexity while at the same time increasing computational efficiency, we randomly divide the original dataset into \(K\) subsets, and conduct \(K\)-fold cross-validation (details are described in online supplementary methods, for all online supplementary material, see www.karger.com/doi/10.1159/000327299). The iterative algorithm stops as soon as there is no increase in the average AUC value calculated from the validation subsets, with the result that the final prediction model built has the appropriate model complexity.

For the forward selection algorithm, incorporating the \(K\)-fold cross-validation process is an intuitive way to address the issue of overfitting. However, when a large number of predictors are considered, the \(K\)-fold cross-validation procedure may tend to be unstable. Therefore, as an alternative to the \(K\)-fold cross-validation procedure, we also developed a cross-validation procedure with bootstrap aggregating (bagging) to further improve the method’s performance (for details, see online suppl. methods).

Non-binary classifications can sometimes outperform binary classifications [22]. For such circumstances, the proposed forward ROC method can easily be extended to use non-binary classifications, i.e., at each step of the forward selection process, for a certain locus \(j\), we could consider the possibility of classifying the individuals into three risk groups by splitting at each genotype of locus \(j\). For instance, in step 1, we could consider classifying the
individuals into three risk groups: \( \{g_1^j\}, \{g_2^j\}, \{g_3^j\}\). However, whereas this kind of non-binary split may be reasonable for environmental categories, it can lead to the formation of very unlikely complex genotype groups.

Procedure for Handling Missing Data

Since a large number of predictors are included in the analysis, the issue of missing values needs to be addressed. We incorporate a procedure into our method to allow for the presence of missing values. The principle of the procedure is that, at each step, we assign individuals with missing genotypes the LR values obtained at the previous step (i.e., the closest LR values we could obtain for the individuals), so that these individuals could be used in the analysis. Thus suppose, after step 1 \( G_t \) \((t = 1, 2, \ldots)\), we classify individuals into \( K \) groups, \( \{G_1, G_2, \ldots, G_K\} \) with their corresponding LRs, \( \{LR(G_1), LR(G_2), \ldots, LR(G_K)\} \). In step \( t + 1 \), assuming some individuals belonging to \( G_t \) have missing genotypes at locus \( j^t \), instead of excluding these individuals from the analysis, we treat them as a separate risk group \( G_t^{j_t} \) and assign the value of LR(\( G_t \)) obtained from step \( t \) to this group. Based on the LR values, we rank this group, \( G_t^{j_t} \), as well as other risk groups, and estimate the AUC value for step \( t + 1 \). In the remaining steps, we keep this risk group, \( G_t^{j_t} \), and its LR value for the AUC calculations. Compared to existing missing data procedures, the proposed procedure is computationally efficient and naturally fits the forward selection algorithm. It infers the risk of an individual with missing data by using all the available genotypes, and thus maximally uses the data information. Should there be any individuals with missing genotypes at the locus used to obtain the best classification at step 1, there are no LR values from the previous step that can be assigned to them. We therefore exclude these individuals from the analysis (note that a locus at which there are missing values has, on average, a smaller chance of resulting in the classification with the highest AUC at this first step, so it is unlikely that this will lead to many, if any, individuals being excluded).

### Results

#### Simulation Studies

We compared the forward ROC method with two other methods – the CART [9] and the allele-counting method [7]. For our comparison, the CART method was implemented using the `tree` package in R, while the allele-counting method was performed using R code that we have written. For each simulation setting, 1,000 replicates were simulated, each consisting of 500 cases and 500 controls. In each replicate, we split the whole sample into a learning set and a validation set, with a ratio of 2:1. We applied each method to the learning set to build the prediction models, and then applied the prediction model to the validation set to calculate the predicted AUC. For the forward ROC method, we evaluated its performance using 10-fold cross-validation.

#### Scenario I

In this simple simulation, we considered three disease-related loci (denoted by \( a, b \) and \( c \)), one environmental risk factor (denoted by \( e \)) and several non-causal loci. The simulation parameters were all picked to mimic common complex disease scenarios. The three disease loci – \( a, b \) and \( c \) – were simulated under additive, recessive and dominant modes of inheritance, respectively. We also introduced a two-way gene-environment interaction, \( a \times e \), and a two-way gene-gene interaction, \( b \times c \), into the simulation. The details of the parameter settings are listed in online supplementary table 1. We set the disease prevalence to be 0.035 and calculated the true AUC value, which was 0.651. To assess the impact of ‘noise’ loci on the performance of the three methods, we simulated 10, 15 and 20 ‘noise’ loci when evaluating these methods under each simulation setting. We assumed no missing data in this first simulation.

The results from simulating scenario I are summarized in table 1. In all three simulation settings, the proposed forward ROC method attained higher AUC means and smaller AUC mean square errors than CART and the allele-counting method. For example, in the 10-noise-locus setting, the test built by the proposed forward ROC method had an average classification accuracy of 0.6198, a 5.88% accuracy increase over CART and a 9.47% increase over the allele-counting method. As expected,

Table 1. Summary of the results from simulation scenario I

<table>
<thead>
<tr>
<th></th>
<th>Forward ROC mean²</th>
<th>CART mean</th>
<th>Allele counting mean²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bias SD</td>
<td>bias SD</td>
<td>bias SD</td>
</tr>
<tr>
<td>10 noise loci</td>
<td>0.6198 –0.0312 0.0305 0.0019</td>
<td>0.5854 –0.0656 0.0365 0.0056</td>
<td>0.5662 –0.0848 0.0310 0.0082</td>
</tr>
<tr>
<td>15 noise loci</td>
<td>0.6018 –0.0492 0.0343 0.0036</td>
<td>0.5770 –0.0740 0.0419 0.0072</td>
<td>0.5560 –0.1001 0.0274 0.0108</td>
</tr>
<tr>
<td>20 noise loci</td>
<td>0.5985 –0.0525 0.0398 0.0043</td>
<td>0.5647 –0.0863 0.0423 0.0092</td>
<td>0.5585 –0.0925 0.0324 0.0096</td>
</tr>
</tbody>
</table>

² SD = Standard deviation; MSE = mean square error; CART = classification and regression tree. a AUC estimator.
with the increasing number of noise loci, the tests formed by all these methods tended to decrease in accuracy. Nevertheless, the forward ROC method always maintained a similar advantage over the CART and allele-counting method (table 1).

Scenario II
Simulations were also conducted to evaluate the performance of the three methods when there are missing genotypes. We used the multiple imputation method with \( m = 50 \) [18, 23] to handle the missing data for the allele-counting method, while for CART we used its built-in missing-data procedure. We adopted the same simulation settings as in scenario I, but limited the number of noise loci to 10. Three situations were simulated, in which the proportions of missing values were 0, 10 and 15%, respectively.

The results from simulating scenario II are summarized in table 2. Consistent with the first simulation, the forward ROC method performed better than the CART and the allele-counting method in all three situations. For example, in the case where 10% of the data was missing, the AUC estimated using the forward ROC method had a mean squared error of 0.0029, while those from CART and the allele-counting methods were 0.0075 and 0.0085, respectively. This simulation also showed that, with an increasing amount of missing data, tests built from all three methods tend to have lower accuracy, owing to the loss of data information (table 2).

Cross-Validation Procedure with Bootstrap Aggregating
Through simulation scenarios I and II, we also evaluated the forward ROC method using bagging cross-validation with 50 bootstrap samples, and compared its classification accuracy with the one we had conducted using 10-fold cross-validation (for details, see online suppl. simulation 1). Overall, these two methods had a similar performance. The bagging cross-validation tended to perform better than the 10-fold cross-validation when the number of loci was large or the data contained more missing genotypes. We would expect that the performance of the bagging cross-validation could be further improved by increasing the number of bootstrap samples (i.e., >50). However, bagging cross-validation is time consuming. For instance, in simulation scenario I with 10 noise loci, it took 50 min to finish each replicate, while the cross-validation required only 1.5 min. Thus, when compared with 10-fold cross-validation, bagging cross-validation is not as feasible for handling high-dimensional data (e.g., a genome-wide dataset).

Predictive Genetic Tests for RA
RA is a chronic inflammatory joint disease caused by the interplay of genetic and environmental factors. The association of candidate genes – such as HLA-DRB1 and PTPN22 – with RA has been well replicated in previous genetic association studies [24, 25]. To evaluate the accuracy of the prediction models, we split the entire dataset into a learning set and a validation set, with a 2:1 ratio.

Predictive Genetic Tests Based on Currently Known Risk Factors
In this study, we first assessed the combined effect of gender [25] and 35 RA-associated genetic variants (see online suppl. table 5) in RA prediction. Among them, 18 SNPs can be found directly in the WTCCC dataset, and the other 17 SNPs were imputed using the model-based imputation software IMPUTE [26]. The original WTCCC GWAS treated the genotype as missing when the most

<table>
<thead>
<tr>
<th>Missing values, %</th>
<th>Forward ROC</th>
<th>CART</th>
<th>Allele counting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>meana</td>
<td>bias</td>
<td>SD</td>
</tr>
<tr>
<td>0</td>
<td>0.6198</td>
<td>-0.0312</td>
<td>0.0305</td>
</tr>
<tr>
<td>10</td>
<td>0.6078</td>
<td>-0.0432</td>
<td>0.0317</td>
</tr>
<tr>
<td>15</td>
<td>0.6007</td>
<td>-0.0504</td>
<td>0.0327</td>
</tr>
</tbody>
</table>

In this simulation, with an increasing amount of missing data, tests built from all of these methods tended to have lower accuracy. For example, when the missing rate increased from 10% to 15%, the AUC values decreased from 0.6078 to 0.6007 for the forward ROC method (a 1.17% decrease), while the AUC values decreased from 0.5740 to 0.5611 (a 2.25% decrease) for CART and from 0.5627 to 0.5612 (a 0.27% decrease) for the allele-counting method.

SD = Standard deviation; MSE = mean squared error; CART = classification and regression tree. a AUC estimator.
probable call fell below 0.9 [25]. To be consistent with the original study, we here used the same criterion to call the 18 existing SNPs and the 17 imputed SNPs. Based on this criterion, we obtained a dataset with a missing rate of 4.67%, where missing values were present for both the 18 existing SNPs and the 17 imputed SNPs. For forward ROC and the allele-counting method, we adopted the same missing-data procedures as described in the simulations. For the CART method, we used both the built-in missing-data procedure in CART and the multiple imputation method.

The classification accuracy of these methods is presented in figure 1, which is consistent with the findings from the simulation studies. The forward ROC method (AUC = 0.7220) attained a higher accuracy than the other two methods, CART (AUC = 0.7012) and the allele-counting method (AUC = 0.6827). We also note that, in our case, the CART method had a much lower predicted AUC value (0.6552) when using its own missing value method. The results also showed that the forward ROC method using cross-validation with bagging formed a test with the same accuracy as the one using 10-fold cross-validation.

The forward ROC method selected four predictors – gender, rs660895, rs6679677 and rs6457617 – into the final model, which classified individuals into 24 distinct risk groups. rs6457617 is located in the region of the major histocompatibility complex (MHC) and rs6679677 is perfectly correlated (r² = 1 from HapMap CEU data) with a functional SNP of PTPN22 (rs2476601) [25], both of which showed the strongest associations with RA in the WTCCC GWAS study [25]. Another SNP, rs660895, which is a tag SNP for the HLA-DRB1*0401 allele, has also been reported as being highly associated with RA [24]. All three SNPs are involved in the pathway of anti-CCP-positive RA [27–29]. Note that the CART method selected gender and two

Fig. 1. The predictive genetic test formed based on currently known RA risk factors. a The ROC curves, from top to bottom, correspond to the fitted ROC curves built by the forward ROC method with 10-fold cross-validation, the forward ROC method with bagging cross-validation, CART and the allele-counting method. b The ROC curves were obtained by applying the built prediction models to the validation set.
A Forward ROC Method for High-Dimensional Risk Prediction

SNPs – rs660895 and rs6457617 – for the final prediction model, whereas an additional SNP was selected by the forward ROC method.

A Predictive Genetic Test Formed Based on Whole Genome-Wide Data

Taking only the currently known risk factors into consideration may limit our ability to form a powerful test. To consider more unknown, but potentially important, risk factors, we also sought a whole genome-wide predictive genetic test using all the 460,547 SNPs available from the Wellcome Trust RA GWAS. Owing to the limitations of CART and the allele-counting method on high-dimensional data, and the intensive computational requirements of cross-validation with bagging, we only investigated the forward ROC method with 10-fold cross-validation. The whole genome-wide risk prediction analysis took 43 h on a 4G-memory computer. By applying a C++ program of the forward ROC method on the Wellcome Trust RA GWAS dataset, we formed a genome-wide predictive test with a predicted AUC of 0.7101 (fig. 2).

Three predictors – rs6457620, gender and rs6679677 – were selected in the final model. The SNP rs6679677 was selected in the above analysis of currently known risk factors. The other SNP, rs6457620, is a SNP within the HLA-DRB1 gene and has been previously found to be associated with RA [30]; it is in perfect linkage disequilibrium with the SNP rs6457617 ($r^2 = 1$ in the HapMap CEU Founders dataset), which was also selected in the above analysis. This shows that the findings from our genome-wide prediction modeling were consistent overall with the prediction modeling based on the known risk factors, suggesting the important role of HLA-DRB1 and PTPN22 in predicting RA. Both loci are involved in the pathway of anti-CCP-positive RA [27–29]. For anti-CCP-positive RA before the symptom of synovitis, where biomarkers such as anti-CCP antibodies and the rheumatoid factor may have less value for prediction [27, 31], genetic markers, such as the HLA-DRBI- and PTPN22-related loci, could be included for earlier RA prediction [32].

Discussion

We have introduced a non-parametric method, the forward ROC method, for constructing predictive genetic tests on high-dimensional data. The development of the method was motivated by the increasing findings from the ongoing WTCCC GWAS and the emerging need for translating these discoveries into clinical use. As stated above, the forward ROC method is specifically designed to construct predictive genetic tests on high-dimensional data. It models the complex relationship between phenotype and genotype by considering gene/gene/gene-environment interactions, and is capable of exploring a large number of potential predictors for improved performance. Moreover, it uses a computationally efficient procedure to deal with the missing values existing in high-dimensional data. As illustrated by our simulations and a real-data application, the missing-data procedure in the forward ROC method attains a more robust performance than other existing methods. The implementation of the forward ROC method is written in both R and C++, and these programs are available upon request.

We conducted simulation studies and a real data application to evaluate the forward ROC method and then compared it with two other commonly used methods, the allele-counting method and CART. The proposed forward ROC method is non-parametric. It makes no assumptions about the mode of inheritance at a locus, and thus attains a robust performance under common complex disease scenarios when the underlying mode of inheritance is unknown. The method clusters individuals into different multi-locus genotype groups. By evaluating the risk differences among these groups, it takes interaction into account and thus attains a better performance when interactions are present. The allele-counting meth-

Fig. 2. The predicted ROC curve built based on whole genome-wide RA data using the forward ROC method.
od, on the other hand, assumes that the effect sizes of risk alleles are additive and the same within and across loci, and could be subject to low performance if this assumption fails. As illustrated by the simulations, the forward ROC method performed better than the allele-counting method when interactions are present and the disease-related loci follow dominant/recessive modes of inheritance. Furthermore, unlike the allele-counting method, the new method adopts a variable selection procedure to choose only a few important predictors for risk prediction. This makes the results more interpretable and easy to investigate in a follow-up study.

An improved allele-counting method, the weighted allele-counting method, has also been introduced to address the limitations of the allele-counting method (i.e., assuming equal effect size within and across loci) [33, 34]. Using the data on 35 RA risk predictors, we compared the weighted allele-counting method with the forward ROC method. We found that the weighted allele-counting method (AUC = 0.6966) outperformed the original allele-counting method (AUC = 0.6827). Nevertheless, it had a lower accuracy when compared to the forward ROC method (AUC = 0.7220). Although the weighted allele-counting method relaxes the assumption of equal effect sizes, it still assumes that all the predictors are disease-related and are independent of each other. It is, therefore, subject to low performance when non-causal predictors or interactions are present. Moreover, the empirical weighting process can potentially introduce bias if there are variations across studies or insufficient information in previous studies for reliable parameter estimation [34].

We also evaluated a prediction model for the RA data by using logistic regression on the known risk factors. In the logistic regression model, we considered all main effects and two-way interaction effects, and used forward selection to choose the most parsimonious model. The result showed that the prediction model formed by logistic regression reached a lower AUC value (0.6666) than our method (0.7220). This may indicate that, even with a moderate number of risk factors, logistic regression could still be subject to low performance, especially when the mode of inheritances and interactions are unknown and a large number of parameters are involved. However, unlike parametric methods such as logistic regression, non-parametric methods – as the forward ROC method and CART – do not suffer from issues relating to an increasing number of parameters. This makes them more suitable for high-dimensional data and the consideration of high-order interactions. We found that, in the real data application, both the forward ROC and CART methods could detect important interactions. However, the forward ROC method detected more risk factors, and thus formed a predictive genetic test with a higher validated accuracy than CART.

Most common complex diseases, such as RA, are usually caused by the interplay of a large number of genetic and environmental risk factors, each with a medium- or small-effect size. However, some diseases, such as age-related macular degeneration (AMD) [35], are caused by a small set of variants, each with a large effect size. We also conducted a simulation to evaluate the new method in such a scenario. Through the simulation, we found that tests built by all three methods achieved medium-to-high accuracy, but the proposed forward ROC method performed much better than either CART or the allele-counting method (for details, see online suppl. simulation 3).

Another commonly used non-parametric method, the random forest (RF) method, uses an ensemble of individual classification or regression trees for disease prediction. The RF method can be used for risk prediction on a relatively large number of genetic/environmental risk predictors, each with a small effect [36, 37]. We simulated the data under different effect sizes and causal/noise locus ratios, and compared the forward ROC method with the RF method (for details, see online suppl. simulation 2). The results suggest that the RF method attains higher classification accuracy than the forward ROC method when the data contain a large proportion of disease susceptibility loci (e.g., a risk prediction study based on known risk predictors) each with a relatively small effect size; while the forward ROC method performs better when the data comprise a small proportion of disease susceptibility loci (e.g., a genome-wide risk prediction study) each with a relatively large effect size (for details of the simulation settings, see online suppl. table 1). The different performance of the two methods may be explained by their own unique features. The RFs comprised a large number of fully grown trees, each formed by a set of candidate loci. Thus, compared to the forward ROC method that only selects a single set of disease susceptibility loci, the RF method is capable of considering multiple sets of a large number of loci for disease prediction, potentially involving different pathways. However, unlike the forward ROC method, the RF method grows a full size tree, and could be subject to low performance (i.e., decreased accuracy and increased variation) when a large proportion of the loci are noise loci. Therefore, it is less suitable for genome-wide risk prediction, where we expect a large number of loci to be noise. Unlike the RF method, the forward ROC method is computationally efficient and leads to an easy interpretation.
of the prediction model, which makes it more suitable for whole genome-wide risk prediction.

Using the proposed method, we investigated a predictive genetic test based on known RA risk factors (AUC = 0.7220), as well as a whole genome-wide predictive genetic test (AUC = 0.7101) based on approximately 500,000 SNPs. The small AUC difference between the two predictive genetic tests is mainly due to rs660895, an imputed marker, which was not included in the genome-wide risk prediction analysis. We also ran additional risk prediction analyses to evaluate the role of currently known risk predictors and other potential risk predictors on RA risk prediction. Following a strategy used by Wei et al. [19], we evaluated different sets of potential risk predictors by varying the p value cutoff threshold (e.g., p value <0.0001). For each threshold, we selected a set of potential genetic predictors, and combined them with 36 known risk predictors to form a risk prediction model. The results from these additional analyses were consistent with the findings from the genome-wide analysis and known risk factors analysis. For instance, a risk prediction model built on 36 known risk predictors and 48 genetic predictors of marginal significance (<0.0001) attained the same AUC value of 0.7220. This suggests there is no strong predictor that can be identified in the given data, which contains no environmental variables. Nevertheless, additional environmental conditions might be measured in other studies, and could be combined with the genetic data to further improve the RA predictive genetic test. The accuracy of a RA predictive genetic test may also be improved by considering disease heterogeneity. In the presence of disease heterogeneity, it would be more ideal to predict sub-phenotypes (e.g., anti-CCP-negative RA [38]). Genome-wide risk prediction analysis does not depend on previous association findings – which makes it appealing for the analysis of less studied sub-phenotypes.

Aside from the currently known environmental risk factors and common genetic variants, rare genetic variants provide another potential resource for early disease prediction [39]. When dealing with rare variants, however, issues of sparseness and missing values could arise. Thus, new methods may need to be developed in the future that incorporate into the forward ROC method, the collapsing or the aggregating methods for rare variants. By considering together common variants, rare variants, and environmental risk factors, we expect that predictive genetic tests for common complex diseases will be further improved, and will eventually be used for clinical purposes.

Acknowledgements

We would like to thank the anonymous reviewers for their suggestions that greatly improved our paper. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available at www.wtccc.org.uk.

References
