

Ensemble learning for detecting gene-gene interactions in colorectal cancer

Faramarz Dorani^{1,*}, Ting Hu^{1,*}, Michael O. Woods²
and Guangju Zhai²

¹ Department of Computer Science, Memorial University, St. John's, Newfoundland and Labrador, Canada

² Faculty of Medicine, Memorial University, St. John's, Newfoundland and Labrador, Canada

* These authors contributed equally to this work.

ABSTRACT

Colorectal cancer (CRC) has a high incident rate in both men and women and is affecting millions of people every year. Genome-wide association studies (GWAS) on CRC have successfully revealed common single-nucleotide polymorphisms (SNPs) associated with CRC risk. However, they can only explain a very limited fraction of the disease heritability. One reason may be the common uni-variable analyses in GWAS where genetic variants are examined one at a time. Given the complexity of cancers, the non-additive interaction effects among multiple genetic variants have a potential of explaining the missing heritability. In this study, we employed two powerful ensemble learning algorithms, random forests and gradient boosting machine (GBM), to search for SNPs that contribute to the disease risk through non-additive gene-gene interactions. We were able to find 44 possible susceptibility SNPs that were ranked most significant by both algorithms. Out of those 44 SNPs, 29 are in coding regions. The 29 genes include *ARRDC5*, *DCC*, *ALK*, and *ITGA1*, which have been found previously associated with CRC, and *E2F3* and *NID2*, which are potentially related to CRC since they have known associations with other types of cancer. We performed pairwise and three-way interaction analysis on the 44 SNPs using information theoretical techniques and found 17 pairwise ($p < 0.02$) and 16 three-way ($p \leq 0.001$) interactions among them. Moreover, functional enrichment analysis suggested 16 functional terms or biological pathways that may help us better understand the etiology of the disease.

Submitted 5 June 2018

Accepted 28 September 2018

Published 29 October 2018

Corresponding author

Ting Hu, ting.hu@mun.ca

Academic editor

Jin-Kao Hao

Additional Information and
Declarations can be found on
page 18

DOI 10.7717/peerj.5854

© Copyright

2018 Dorani et al.

Distributed under

Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Bioinformatics, Computational Biology, Genetics, Epidemiology, Data Mining and Machine Learning

Keywords Ensemble learning, Gene-gene interaction, Colorectal cancer, Random forests, Gradient boosting machine, Genetic marker discovery, Epistasis, Complex diseases

INTRODUCTION

The goal of most genome-wide association studies (GWAS) is to examine common genetic variations across the entire human genome in order to identify single-nucleotide polymorphisms (SNPs) that are associated with diseases or phenotypic traits (*McCarthy et al., 2008; Hindorff et al., 2009; Foulkes, 2009; Lek et al., 2016*). GWAS survey the human genome for causal factors (*Hirschhorn & Daly, 2005*), and have successfully identified genetic variants that influence the risks of many complex diseases, including

cardiovascular diseases (Mohlke, Boehnke & Abecasis, 2008; Nikpay et al., 2015), autoimmune diseases (Samani et al., 2007), and cancers (Easton & Eeles, 2008; Lettre & Rioux, 2008; Michailidou et al., 2013).

Various statistical and computational approaches can be used to identify genetic variants associated with human diseases. Those approaches can be divided into two categories: uni-variable and multi-variable analyses. In the uni-variable analysis, single-locus tests examine each SNP independently for its association with the phenotype. The effect size (or penetrance) of a variant is calculated and scored based on its significance of association with the disease phenotype. Many single-SNP-based methods were used for initial GWAS analyses, but had limited success in detecting genetic risk factors and explaining disease heritability (Moore & Williams, 2002; Balding, 2006; Szymczak et al., 2009; Manolio et al., 2009).

Recent research has shifted toward using or developing multi-variable approaches that examine interactions among genetic variants (Friedman, Hastie & Tibshirani, 2001; Balding, 2006; Zhang & Liu, 2007; Frazer et al., 2009; Han & Chen, 2011; Bush & Moore, 2012; Beam, Motsinger-Reif & Doyle, 2014; Jing & Shen, 2015; Sun et al., 2017). It is more plausible that interactions of multiple factors rather than individual genetic variants, can better explain susceptibility of complex diseases (Moore & Williams, 2009; Mackay & Moore, 2014). Multi-locus methods analyze combinations of SNPs and are able to capture interactions among multiple SNPs in GWAS data. However, they present significant statistical and computational challenges in terms of developing powerful methods to model non-linear, non-additive SNP interactions, selecting most relevant genetic variables, and interpreting discovered gene-gene interaction models (Moore & Ritchie, 2004).

Machine learning methods that are able to derive predictive models through training from historical data have seen increasing applications in finding genetic markers for GWAS (Szymczak et al., 2009; Okser et al., 2014; Libbrecht & Noble, 2015; Niel et al., 2015). For instance, the Ridge regression method was employed to identify a risk susceptibility SNP in rheumatoid arthritis near the *HLA-B* gene (Sun et al., 2009). D'Angelo, Rao & Gu (2009) combined the least absolute shrinkage and selection operator with the principal component analysis and detected two significant gene-gene interactions in rheumatoid arthritis. Although successful in characterizing gene-gene interactions, the common regression methods are not able to search for combinations of SNPs in high-dimensional and large-scale GWAS data (Szymczak et al., 2009).

Meanwhile, ensemble learning methods, especially the random forests algorithm, have been explored to search for and characterize combinatorial and non-linear interactions in microarray gene expression data (Huynh-Thu et al., 2010) and in GWAS data (Yoshida & Koike, 2011; Wright, Ziegler & König, 2016; Niel et al., 2018; Niel & Sinoquet, 2018). For instance, random forests were shown to be able to rank both main-effect and interacting SNPs using its Gini Index variable importance measure (Kim et al., 2009; Chen & Ishwaran, 2012), and to identify many known and several new susceptibility SNPs associated with human diseases (Tang et al., 2009; Wang et al., 2009). In a comprehensive comparative study (Olson et al., 2017), it was suggested that when

applied to classification problems in bioinformatics data, tree-based ensemble learning methods outperform others, including support vector machine and naive Bayes methods. A total of 13 machine learning methods were investigated and applied to 165 benchmarking bioinformatics datasets. The comparison results showed that gradient tree boosting and subsequently random forests performed the best in terms of achieving the best cross-validation (CV) classification accuracy.

In this article, we investigated two ensemble learning algorithms, random forests and gradient boosting machine (GBM), to search for risk susceptibility SNPs associated with colorectal cancer (CRC). We used a CRC GWAS dataset collected from the Canadian province of Newfoundland and Labrador. Both algorithms were able to capture the non-linear gene-gene interactions associated with the disease status, and to rank SNPs based on their disease associations through either main effects or interaction effects. We identified 44 SNPs ranked as the most significant by both algorithms, including both known CRC related genes such as *DCC* and *ALK*, and SNPs that have not been found previously associated with CRC. We further performed a gene-gene interaction analysis on the 44 SNPs using an information gain method, and were able to validate significant pairwise and three-way synergistic interactions among them. The functional enrichment analysis on the 29 genes mapped from the 44 SNPs also suggested several significant biological pathways that may help explain the risk of CRC.

METHODOLOGY

CRC GWAS data and preprocessing

The colorectal (CRC) GWAS case-control data were collected from CRC patients and healthy individuals with matching age, gender, and geographical distributions within the province of Newfoundland and Labrador, Canada. The CRC Transdisciplinary (CORECT) consortium coordinated the genotyping of data. Genotyping was conducted using a custom Affymetrix genome-wide platform (the Axiom CORECT Set) on two physical genotyping chips (pegs) for two datasets with around 1.2 and 1.1 million SNPs (*Schumacher et al., 2015*). The first dataset has 1,236,084 SNPs and 696 samples with 200 cases and 496 controls and the genotyping rate was 0.997. In GWAS, genotyping rate is computed as the percentage of samples (including both cases and controls) that are successfully genotyped. The second dataset has 1,134,514 SNPs and 656 cases with a genotyping rate of 0.888. Using PLINK (*Purcell et al., 2007*), a tool for analyzing genetic data, we merged these two datasets based on their common SNPs and obtained a dataset of 265,195 SNPs and 1,152 unique samples. Among the samples, 656 were cases and 496 were controls.

Next, we performed per-sample and per-marker quality control steps and linkage disequilibrium (LD) pruning on the CRC data using PLINK. In the per-sample quality control, sex check was performed, sex chromosomes were excluded, and samples with more than 1% missing genotypes and outlier heterozygosity rate were removed. In the per-marker quality control, SNPs with missing call rates higher than 5%, with minor allele frequencies (MAFs) less than 5%, or with Hardy-Weinberg equilibrium values greater than 0.0001 were removed. Then, we pruned SNPs that were dependent on

each other (correlation coefficient $r^2 > 0.6$, LD window size 2,000). Missing genotypes were imputed using the most frequent alleles in the population. Last, the genotypes were re-coded using numerical values {0, 1, 2} with 0 standing for homozygous reference, 1 for heterozygous variant, and 2 for homozygous variant. After the steps of quality control and imputation, the final preprocessed and balanced dataset had 186,251 SNPs and 944 samples (472 being cases and 472 being controls).

The original data collection was approved by Memorial University Health Research Ethics Authority (HREA) with the approval number HIC 01.70. Our study was recognized by HREA as the use of secondary data which have already been collected and de-identified, and did not require a clearance.

Feature selection

Most computational methods in informatics find it prohibiting to analyze high dimensional GWAS data (*Moore, Asselbergs & Williams, 2010*) due to the massive number of attributes, that is, hundreds of thousands of SNPs in the data. It is impossible for machine learning methods to detect interactions by enumerating all combinations of SNPs in a GWAS dataset. Furthermore, the existence of redundant and irrelevant attributes hinders machine learning methods to reveal actual gene-gene interactions in the data. These together set the stage for the necessity of dimensionality reduction or feature (attribute) selection for analyzing GWAS data (*Moore, Asselbergs & Williams, 2010*).

Feature selection is frequently used as a data filtering step in machine learning when the original data contain noisy or irrelevant features, or attributes, that could compromise the prediction power of learning algorithms (*Yu & Liu, 2003*). Feature selection methods choose only a subset of the most important features, and thus reduce the dimensionality of the data, speed up the learning process, simplify the learned model, and improve the prediction performance (*Dash & Liu, 1997; Guyon & Elisseeff, 2003*).

In our previous study (*Dorani & Hu, 2018*), six feature selection algorithms, including chi-square, logistic regression, odds ratio, and three Relief-based algorithms (Relieff, Tuned Relieff (TuRF), and Spatial Uniform Relieff) were compared based on how they rank the most important SNP attributes that contribute to the phenotypic outcome through gene-gene interactions. We applied the feature selection methods to both simulated and real GWAS datasets and showed that Relief-based methods, specifically TuRF, performed the best in filtering SNPs associated with a disease through interaction effects. This observation was also supported by other independent studies (*Beretta & Santaniello, 2011; Urbanowicz et al., 2018*). Therefore, in this study, we used the TuRF feature selection method to reduce the CRC GWAS dataset to be of a manageable size such that it can be analyzed using the downstream classification algorithms in a reasonable time frame.

Ensemble learning

Many preliminary GWAS employed uni-variable approaches where interactions between multiple variables could be overlooked (*Moore & Ritchie, 2004*). Moreover, the parametric linear statistical models look at multiple variables but have limitations for detecting

non-linear interactions (Moore, Asselbergs & Williams, 2010). Thanks to the intrinsic multi-variable and non-linear properties, tree-based ensemble learning methods have been proved to be a powerful analysis tool for detecting interacting genes in GWAS (Moore, Asselbergs & Williams, 2010; Upstill-Goddard et al., 2012). They can be used to train highly accurate classifiers, as well as to discover new genetic markers by ranking genetic variables based on their importance in classification.

Ensemble learning methods use an aggregation of predictors known as base learners. To produce a final prediction, the predictions of the base learners are weighted and the overall predictions are decided through majority voting for classifications and averaging for regressions. It has been shown that ensemble learning methods are powerful at reducing variance and overfitting by utilizing a collection of diverse base learners such as classification and regression trees (CART) (Breiman, 1996; Dietterich, 2000). In this article, we adopted two tree-based ensemble learning algorithms, random forests and GBM, to explore their power in detecting non-linear gene-gene interactions for GWAS.

Random forests

One popular ensemble learning method is *bagging* (short for bootstrap aggregating) that uses bootstrapped samples of the training data to train classification or regression models separately (Breiman, 1996). Bagging reduces the variance of an estimated prediction function, and works especially well for high-variance, low-bias procedures, such as trees (Friedman, Hastie & Tibshirani, 2001). The random forests algorithm is a special case of bagging where the variables are randomly selected to determine the optimal split at each node of the tree (Breiman, 2001). Random forests are shown to be a very powerful regression and classification method which utilizes a large collection of possibly uncorrelated decision trees (Breiman, 2001; Szymczak et al., 2009; Ziegler, DeStefano & König, 2007; Schwarz, Inke & Ziegler, 2010). Each tree is grown using the CART methodology (Breiman, 1996).

In random forests, n trees are grown independently using bootstrapped training samples. While constructing an individual tree, for each node, m predictor variables are randomly picked from the total M variables in the original data. Then the best variable/partition-point among the m variables is picked and used to split the node into two daughter nodes. Each tree grows to its maximum size, that is, when there is only one sample in the leaves. To make a classification for a new testing sample, each tree casts a vote for the predicted class and the majority vote will be the final prediction result. Intuitively, reducing m will reduce the correlation between any pair of trees in the ensemble. For classification, the default value of m is $\lfloor \sqrt{M} \rfloor$; however, the best value for this parameter will depend on the application problem, and it should be treated as a tuning parameter (Breiman, 2001; Friedman, Hastie & Tibshirani, 2001).

The random forests algorithm is effective in uncovering interactions among genes that do not exhibit strong main effects (Moore, Asselbergs & Williams, 2010). The algorithm has been employed in various studies to predict rheumatoid arthritis risks (Sun et al., 2007), to rank SNP predictors (Schwarz et al., 2007; Sun et al., 2008), and to

detect the epistatic effects associated with human diseases ([Garca-Magariños et al., 2009](#); [Pan et al., 2014](#)).

[Goldstein et al. \(2010\)](#) indicated that using the random forests algorithm with default settings of hyper-parameters would not yield optimal results for large GWAS datasets. In contrast, tuning the hyper-parameters, including *mtry* (number of random variables to make best split at each tree node) and *ntree* (number of trees), and using greater values, work well generally for large GWAS datasets.

In this study, we used a very fast implementation of random forests provided in an R package called “ranger” ([Wright & Ziegler, 2017](#)). The “ranger” package provides all functionalities similar to the RandomForest package in R with much greater speed. Therefore, we can use it for the GWAS datasets with a large number of SNPs. We performed parameter tuning on *mtry* and *ntree* in our implementation. For *mtry* we selected the values of {100, 200, 300, 500, 1,000}, and the selected values for *ntree* were {500, 1,000, 2,000}. Other parameters used the package default settings. Therefore, we had 15 different combinations of *mtry* and *ntree*.

In addition, we performed a 10-fold CV. The original dataset was partitioned into 10 equal subsets, and for each of the 10 iterations, one subset was picked as the testing data and the rest nine subsets served as the training data. Thus, each sample was used in the testing data exactly once, and the CV accuracy was computed as the percentage of correctly classifying all the testing samples. For each parameter combination and each fold of the CV (150 iterations in total), we repeated the random forests algorithm 10 times given the stochastic nature of bootstrapping and selecting variables to grow the trees. Since each execution of the random forests will yield a unique predictive model, we computed the accuracy of classifying each testing sample as the fraction of it being correctly predicted by the 10 models resulted by the 10 executions of the algorithm. Then the *average CV accuracy* using a specific combination of parameters was computed as the averaged accuracy across all testing samples. The parameter combination with the best average CV accuracy will be picked.

When the best combination of parameters was picked, the predictive models learned using it were then used to rank features according to the importance of the features in classification. For random forests, we used Gini Index as the quantification of feature importance ([Breiman, 2017](#)). Gini Index measures the inequality of dividing samples of two classes using a feature at a node of the trees. A higher Gini Index of a feature infers a better ability of using this feature to differentiate the two classes of the samples. Since we had multiple learned predictive models as a result of CV and repetitive runs, the final importance score of a feature was computed as the average Gini Index across all models.

Gradient boosting machine

Another method to generate an ensemble is *boosting* where the base learners evolve over time and make weighted votes ([Friedman, Hastie & Tibshirani, 2001](#)). In a boosting algorithm, many base learners are built and the new learners improve on the previous ones. The learners are trained sequentially, which result in building a

“committee” of complex predictors ([Friedman, Hastie & Tibshirani, 2001](#); [Friedman, 2001](#)).

Gradient boosting machine is a boosting machine learning algorithm in which a weighted combination of predictors are used to make the final prediction ([Friedman, 2001](#)). A set of $n.trees$ base learner trees are created in an iterative fashion, where a new one learns from the previous one. Let $F_m(\mathbf{x})$, where $m = 1, 2, \dots, n.trees$, denote the approximation that maps the input variables \mathbf{x} to the desired output y . At iteration m , a new approximation $F_m(\mathbf{x})$ is constructed through improving on the previous one $F_{m-1}(\mathbf{x})$ by adjusting it using the gradient of the loss function $\nabla L(y, F_{m-1}(\mathbf{x}))$.

For implementation, we used an R package called “gbm” ([Ridgeway, 2007](#)). Similar to random forests, we chose a range of values for each hyper-parameter in order to search for the best combination. GBM has three main parameters, $n.trees$ (the number of trees), $interaction.depth$ (the complexity of interactions between nodes, i.e., features), and $shrinkage$ (the learning rate or step-size reduction). The testing values of these parameters were as follows: $n.trees = \{100, 500, 1,000, 2,000\}$, $interaction.depth = \{1, 2, 10\}$, and $shrinkage = \{0.001, 0.01, 0.1\}$. Therefore, we had 36 different combinations of the three parameters. Other parameters of GBM, such as $n.minobsinnode$ (minimum number of samples in the tree terminal nodes), $bag.fraction$ (the fraction of the training set samples randomly selected to construct a new tree), and $train.fraction$ (the fraction of samples used to fit a GBM model), were set to default values as 10, 0.5, and 0.5, respectively.

Similar to random forests implementation, we used a 10-fold CV and repeated the algorithm 10 times for each combination of parameters and each fold of the CV. The best combination of parameters found was then used to rank the features. Feature Importance was estimated based on the number of times a feature is selected for splitting and the improvement to the model as a result of each split, weighted by the number of observations each split operates, and averaged over all trees ([Friedman, Hastie & Tibshirani, 2001](#)). Again, since we repeated GBM algorithm 10 times for each iteration of the 10-fold CV, the final importance score of a feature was averaged over 10 runs and 10-folds.

Statistical interaction analysis

Both the random forests and GBM algorithms provide importance estimations on each SNP based on their contribution to the accurate classification. Such importance scores by the ensemble algorithms help to identify the top-ranked SNPs that are most possibly associated with the disease of CRC. It has been suggested that tree-based ensemble learning algorithms are able to detect interaction effects among multiple SNP features. Therefore, we employ an information gain method ([Hu et al., 2011](#); [Hu et al., 2013a, 2013b](#)) to quantitatively investigate the statistical evidence of pairwise and three-way interactions among the top identified SNPs.

The information gain method considers the genotype of each SNP and the disease outcome as random variables. It first computes the mutual information $I(A; D)$ between SNP A and the disease outcome D as the main effect of the individual SNP A , since it captures how much information of the disease status the genotype of SNP A provides. Then the information gain $IG(A; B; D)$ of combining two SNPs A and B on explaining the

disease status D is computed by subtracting the main effects of A and B , that is, $I(A; D)$ and $I(B; D)$ from the total mutual information $I(A; B; D)$ by joining A and B to explain D . Essentially, the information gain $IG(A; B; D)$ captures the *gained* information of combining A and B on explaining D , that is, the synergistic and non-additive interaction effect between SNPs A and B .

Similarly, when three SNPs, A , B , and C , are considered, the information gain $IG(A; B; C; D)$ is computed by subtracting all the main effects and pairwise interaction effects from the total mutual information $I(A; B; C; D)$ between joining A , B , and C together and the disease status D . Such an information gain metric measures the pure three-way synergistic interaction among three SNPs on explaining the disease outcome.

Permutation testing can be used to assess the significance level of a computed information gain value. In each permutation, the disease status labels of the data are randomly shuffled, and the information gain values are computed again for the permuted data. In this study, we performed a 1,000-fold permutation testing, that is, 1,000 permutations were collected to provide a null distribution of the assumption that there was no association between the genotypes and the disease outcome. Then the p -value was assessed for each computed information gain by counting the portion of permuted datasets that had a greater value of the information gain than the real data.

Functional enrichment analysis

We also annotated the identified top SNPs with functional information using online resources including the ENSEMBL (<http://www.ensembl.org>) and the National Center for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov/>) databases. The databases provide biological information on the allele, chromosome, and gene information for each SNP.

Then we used the Database for Annotation, Visualization, and Integrated Discovery (DAVID) bioinformatics tool (*Dennis et al., 2003; Huang, Sherman & Lempicki, 2008*) to perform a functional enrichment analysis using the genes from the identified top SNPs. Functional enrichment analysis is a method to identify gene functional categories or associations with diseases that are over-represented in a set of genes. We compared the set of identified genes to the background of the entire human genome, in order to test which functional categories were significantly over-represented. The enriched functional categories may help better understand the disease of CRC. DAVID is widely used for enrichment analysis, specifically for exploring the functions of genes. It is a web-based tool which takes a list of genes as an input and produces annotation tables and charts to show diseases and relevant gene ontology (GO) terms enriched by a given list of genes.

RESULTS

In this section, we first show the result of feature selection and then the application results of the two ensemble learning methods on the CRC GWAS dataset. We performed the parameter tuning and used the parameter combinations of the random forest and the GBM algorithms that yielded the best classification accuracies. Meanwhile, both algorithms can rank the SNP variables based on their contributions to the classification.

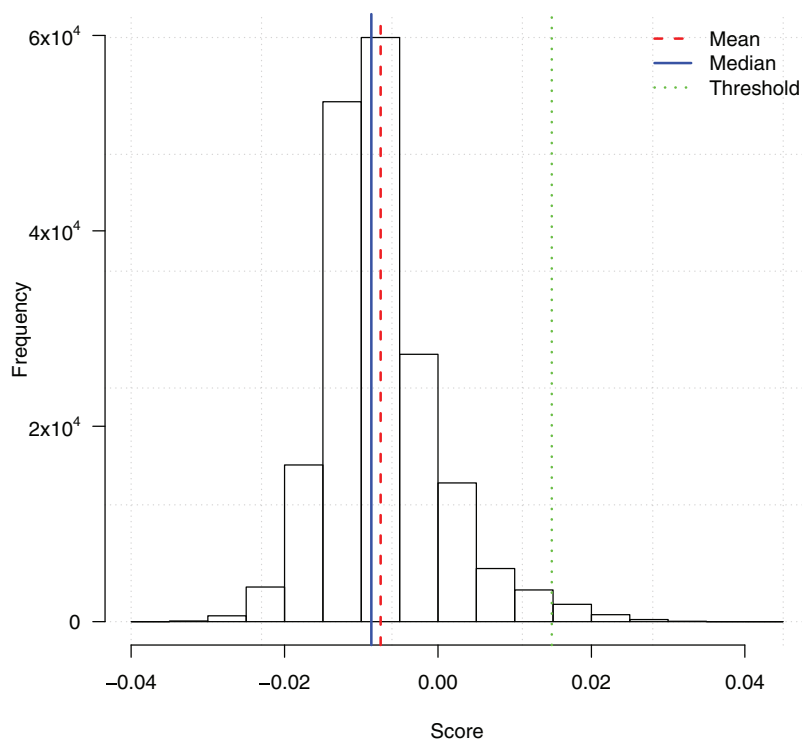


Figure 1 Histogram of SNP importance scores computed by the TuRF feature selection method. A cutoff threshold $mean + 3SD$ is used to select the subset of the top 2,798 SNPs for the downstream ensemble learning analysis. [Full-size](#) [DOI: 10.7717/peerj.5854/fig-1](https://doi.org/10.7717/peerj.5854/fig-1)

We identified the set of the best-ranked SNPs by both algorithms as our key possible susceptibility genetic markers, which were further investigated and validated through gene-gene interaction and pathway enrichment analyses.

Filtered SNPs using TuRF

We ran the TuRF feature selection method on all 186,251 SNPs. Each SNP was assigned a score based on its contribution to the disease status. Since TuRF is a multi-variable feature selection algorithm, such a contribution may include both the individual main effect or the effect of interacting with other SNPs. [Fig. 1](#) shows the distribution of the TuRF scores for all the SNPs.

To reduce the computational overhead of the downstream ensemble learning analysis, we then picked a score threshold and only included the potential SNPs with TuRF importance scores higher than the threshold. The threshold $mean + 3SD$ was chosen which filtered the top 2,798 SNPs. The feature subset was an appropriate size such that the training of the ensemble learning algorithms could be finished in a reasonable time frame, meanwhile, enough features were included in the subsequent analysis to identify possible gene-gene interactions.

Optimized parameters for the ensemble learning algorithms

The parameter tuning for the random forests algorithm included adjusting the values of two parameters, $mtry = \{100, 200, 300, 500, 1,000\}$ and $ntree = \{500, 1,000, 2,000\}$.

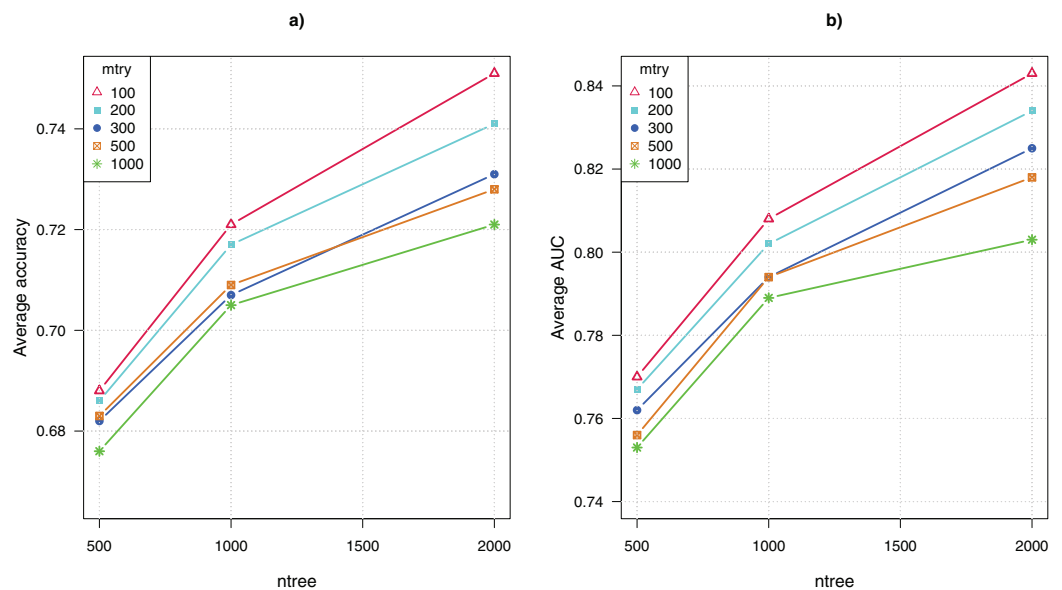


Figure 2 Parameter tuning for random forests. (A) Average CV accuracy using different parameter configurations. (B) Average CV area under the curve (AUC). [Full-size !\[\]\(1663bb69f307a960345edb0e712f8c02_img.jpg\) DOI: 10.7717/peerj.5854/fig-2](https://doi.org/10.7717/peerj.5854/fig-2)

Thus, there were 15 combinations of parameter values. Recall that we ran the algorithm 100 times for each combination, the average CV accuracy and area under the curve (AUC) were computed over 100 runs.

Figure 2 shows the comparative results for the 15 configurations of the random forests algorithm. The average CV accuracy and AUC are shown as a function of the parameter *ntree*, and different curves represent using different values of the other parameter *mtry*. The highest CV accuracy of 75% and AUC of 0.84 were achieved when *mtry* = 100 and *ntree* = 2,000. That is, the algorithm performed the best when *ntree* was set to the maximum value and *mtry* was set to the minimum value. This suggests that greater values of *ntree* and lower values of *mtry* are preferable for analyzing GWAS data. Both increasing the number of trees and reducing the number of randomly picked variables to construct each node of a tree enhance the diversity of trees in the ensemble. The result suggests that the genetic explanation of complex diseases may be highly heterogeneous, that is, a large number of diverse classification models that use different combinations of SNP attributes are needed to approximate the mapping from the genotypes to the disease outcome.

The GBM algorithm had three parameters, that is, $n.trees = \{100, 500, 1,000, 2,000\}$, $interaction.depth = \{1, 2, 10\}$, and $shrinkage = \{0.001, 0.01, 0.1\}$. Again, for each of the 36 combinations of parameter combinations, 10 runs were collected for each of the 10-fold CV, and the average CV accuracy and AUC were computed over 100 runs.

Figure 3 shows the comparative results for the parameter tuning. Each sub-figure uses a *shrinkage* value, and shows the average CV accuracy (left) and AUC (right) as a function of the parameter *n.trees*. The highest CV accuracy was 74% and the CV AUC was 0.82, achieved when *n.trees* was set to 2,000, *interaction.depth* was 10, and *shrinkage*

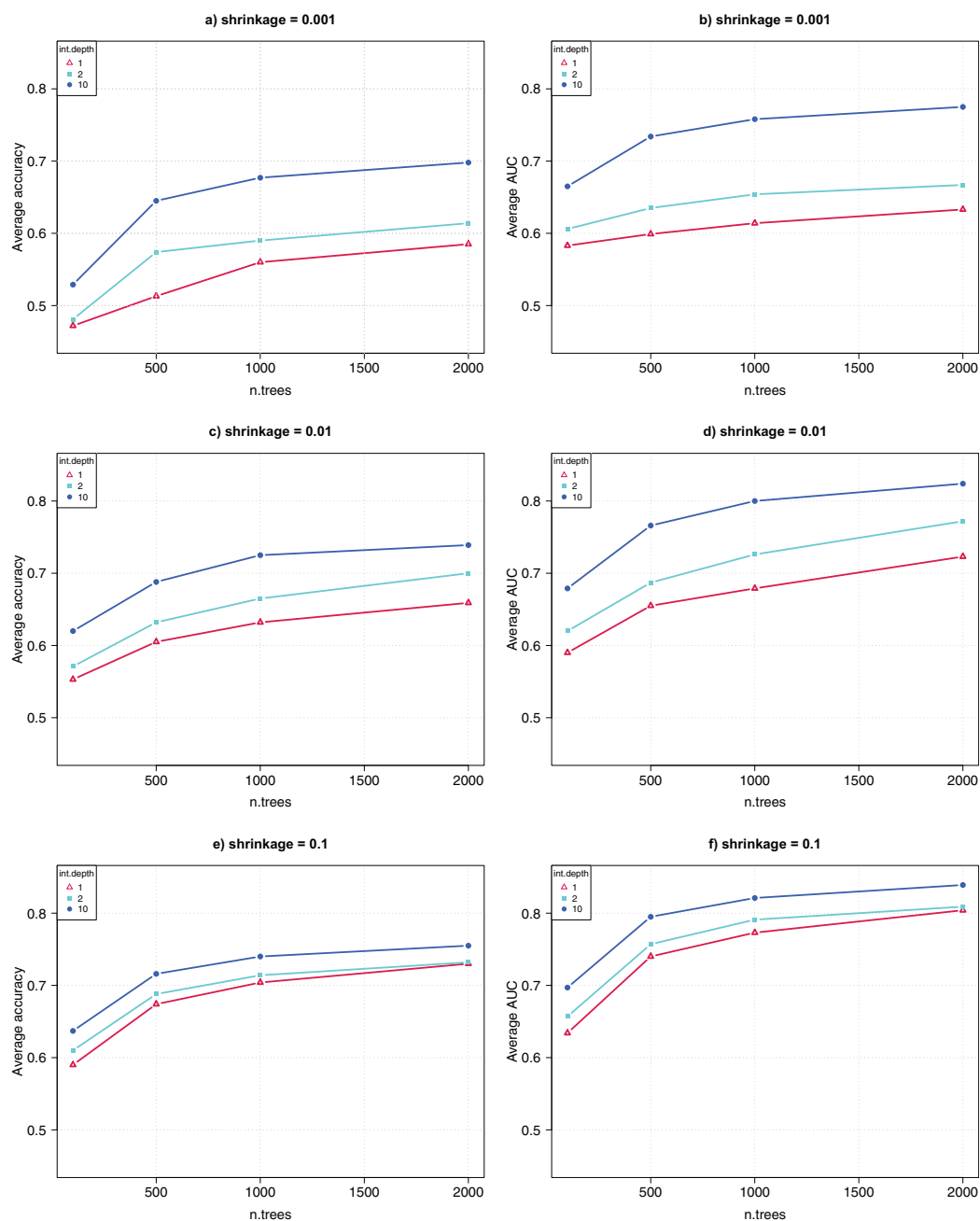


Figure 3 Parameter tuning for GBM. (A, C, E) show the average CV accuracy for *shrinkage* of 0.001, 0.01, and 0.1. (B, D, F) show the average CV AUC for *shrinkage* of 0.001, 0.01, and 0.1, respectively.

Full-size DOI: [10.7717/peerj.5854/fig-3](https://doi.org/10.7717/peerj.5854/fig-3)

was 0.1. This suggests that, for GWAS data analysis, GBM performs better when greater values of *interaction.depth*, *n.trees*, and *shrinkage* are used.

Top-ranked SNPs

After parameter tuning, we chose the optimal parameter combinations of the random forests algorithm, that is, *mtry* = 100 and *n.tree* = 2,000, and of the GBM algorithm, that is,

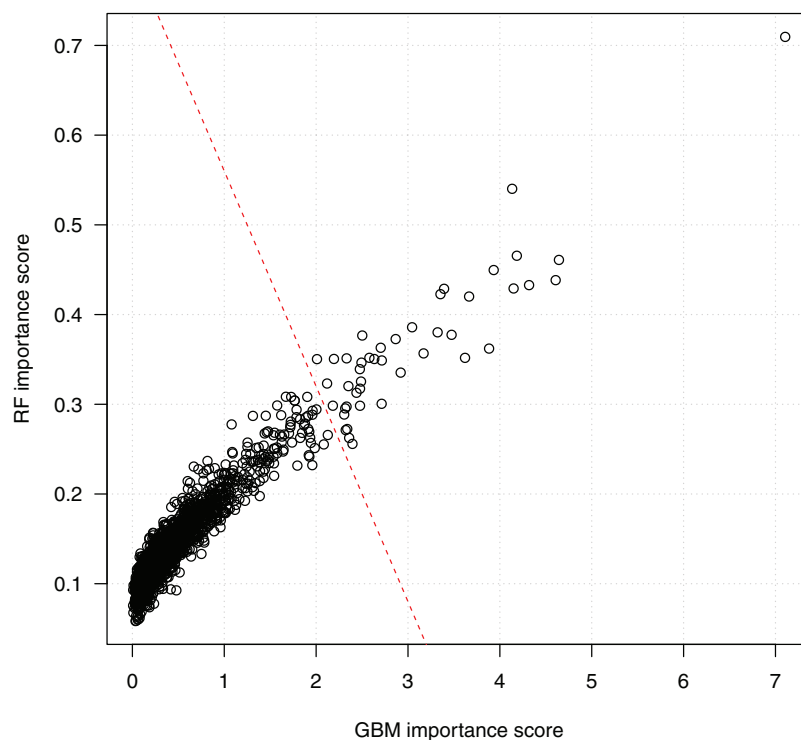


Figure 4 Scatter plot of SNP importance scores by the two ensemble learning algorithms. Each data point represents a SNP, and the x-axis shows its GBM importance score and the y-axis shows its random forests (RF) importance score. A separating line (red dashed) is used to show the most important SNPs identified by both algorithm. [Full-size !\[\]\(5f471a71b78d7676bc356df190b88ab4_img.jpg\) DOI: 10.7717/peerj.5854/fig-4](https://doi.org/10.7717/peerj.5854/fig-4)

$n.trees = 2,000$, $interaction.depth = 10$, and $shrinkage = 0.1$, to assess the importance of the SNP features. In random forests, the Gini Index importance of a feature was estimated based on the impurity of splitting testing samples on the tree node using the feature. Since we repeated the algorithm 100 times (10 times for each fold of the 10-fold CV), the final importance score (between 0 and 1) of a feature was computed as the averaged impurity score over 100 runs. The averaged importance score of each feature using GBM was computed similarly.

Figure 4 shows the scores of the 2,798 SNPs assessed by both algorithms. We see that the scorings by random forests and GBM are highly correlated, which indicates the consistency of the two algorithms. Since a higher score indicates a higher importance of a SNP in classification, the top-right corner of the figure presents the SNPs ranked high by both the random forests and the GBM algorithms. The bottom-left corner of the figure includes a dense cluster of SNPs with low importance scores by both algorithms. We draw a line to separate SNPs into two clusters. The 44 SNPs on the right side of the separating line are the most important SNPs identified by both algorithms, and will be investigated subsequently on their biological functions and statistical interactions. Among them, SNP rs3760948 from gene *ARRDC5* is ranked as the most important by both algorithms.

These 44 SNPs and their chromosome and gene information are listed in Table 1. Gene information of 29 SNPs was found using the ENSEMBL and the NCBI databases.

Table 1 List of the 44 identified top ranking SNPs.

Chromosome	SNP	A1	MAF	p-value	p-value rank	Gene
1	rs12407198	G	0.338	5.675×10^{-3}	1,128	<i>C1orf101</i>
1	rs647831	G	0.345	1.041×10^{-2}	2,018	-
2	rs1367128	G	0.192	3.527×10^{-3}	724	<i>THSD7B</i>
3	rs1505229	T	0.391	9.559×10^{-4}	208	<i>LRRTM4</i>
2	rs1816647	T	0.391	3.521×10^{-2}	6,803	-
2	rs7594717	G	0.335	4.632×10^{-5}	10	<i>ALK</i>
2	rs9288684	T	0.078	1.041×10^{-4}	25	<i>INPP5D</i>
3	rs11185516	A	0.427	6.691×10^{-1}	125,071	<i>ZDHHC19</i>
3	rs12695485	T	0.111	2.966×10^{-2}	5,777	<i>LOC107986044</i>
3	rs6782709	G	0.345	1.885×10^{-2}	3,663	<i>LOC105374217</i>
4	rs10016091	G	0.464	2.475×10^{-3}	514	<i>SCFD2</i>
4	rs1991915	T	0.366	6.061×10^{-3}	1,213	<i>OTOP1</i>
4	rs2010907	G	0.276	8.263×10^{-4}	167	-
4	rs2736486	C	0.326	1.471×10^{-2}	2,888	-
5	rs2406370	G	0.436	1.630×10^{-1}	31,039	<i>ITGA1</i>
5	rs9688110	A	0.356	9.197×10^{-4}	198	<i>FAT2</i>
6	rs7747931	A	0.433	4.637×10^{-2}	8,963	<i>E2F3</i>
6	rs952880	C	0.485	2.318×10^{-1}	43,974	<i>KCNQ5</i>
7	rs17162736	A	0.140	1.466×10^{-2}	2,880	<i>STEAP2-AS1</i>
7	rs17379465	A	0.316	1.253×10^{-1}	23,999	-
8	rs11783793	T	0.417	6.035×10^{-4}	118	-
8	rs11985944	T	0.269	1.981×10^{-3}	432	-
8	rs13263313	T	0.347	8.509×10^{-3}	1,670	<i>JPH1</i>
8	rs1495008	C	0.170	3.740×10^{-3}	765	<i>LOC101929628</i>
8	rs17831158	A	0.390	1.942×10^{-3}	417	<i>LINC00968</i>
8	rs721619	G	0.335	1.755×10^{-1}	33,327	<i>EPHX2</i>
9	rs10814848	G	0.507	1.000×10^{-3}	223	<i>GLIS3</i>
9	rs3912454	C	0.471	3.876×10^{-2}	7,479	-
9	rs4625115	T	0.402	6.728×10^{-4}	138	-
9	rs4961513	A	0.291	3.435×10^{-4}	62	-
11	rs6578849	G	0.379	5.830×10^{-4}	111	<i>SYT9</i>
12	rs11610311	C	0.265	2.581×10^{-3}	528	-
14	rs1212694	A	0.234	1.030×10^{-3}	225	<i>ACTR10</i>
14	rs2645737	C	0.450	1.135×10^{-2}	2,223	<i>NID2</i>
14	rs8022574	A	0.406	2.786×10^{-3}	585	-
18	rs2571219	G	0.387	1.339×10^{-3}	288	<i>ATP8B1</i>
18	rs3844138	A	0.250	1.120×10^{-2}	2,193	-
18	rs658836	C	0.211	4.416×10^{-4}	82	-
18	rs898438	G	0.366	9.927×10^{-4}	218	<i>DCC</i>
19	rs344570	T	0.089	2.718×10^{-4}	53	<i>TNFSF14</i>
19	rs3760948	T	0.371	2.007×10^{-4}	46	<i>ARRDC5</i>
20	rs2179321	T	0.512	4.872×10^{-2}	9,380	<i>PLCB4</i>

(Continued)

Table 1 (continued).

Chromosome	SNP	A1	MAF	<i>p</i> -value	<i>p</i> -value rank	Gene
20	rs2386946	A	0.210	5.192×10^{-3}	1,035	<i>CDH4</i>
21	rs3842986	T	0.224	4.063×10^{-2}	7,875	–

Note:

A1 is the minor allele of a SNP, and MAF stands for the minor allele frequency. The *p*-value shows the confidence of a SNP's association with the disease, computed using PLINK, as well as how it ranks among all 186,251 SNPs in the dataset (from the most to the least significant).

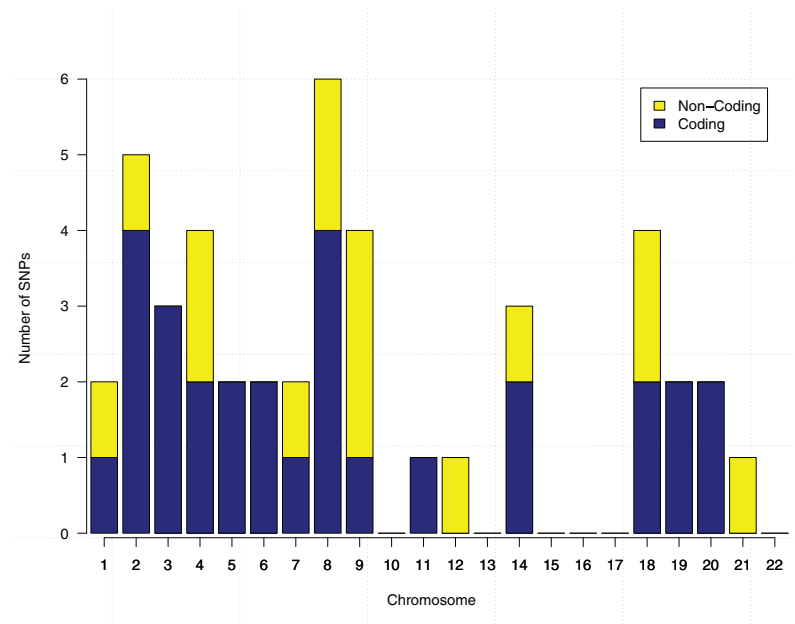


Figure 5 Chromosome distribution and the coding vs. non-coding regions of the identified 44 SNPs.

Full-size DOI: 10.7717/peerj.5854/fig-5

In the table, 15 SNPs without any gene information are from non-coding DNA regions. The *p*-values of SNPs and their ranking among all 186,251 SNPs are also shown in the table. Figure 5 further shows the distribution of the 44 identified SNPs in each chromosome. The comparison of coding vs. non-coding regions is also shown. There were no identified SNPs in chromosomes {10, 13, 15, 16, 17, 22}. Chromosome eight contains six identified SNPs in which four of them belong to coding regions. Chromosome two has four out of five identified SNPs from the coding regions of DNA.

Statistical interactions among the identified SNPs

We calculated the information gain between all $\binom{44}{2}$ pairs of the 44 identified SNPs. A 1,000-fold permutation test was performed to assess the significance level of each pair. Table 2 lists the 17 most significant SNP pairs ($p < 0.02$) with their information gain measures and the *p*-values. The maximum information gain was 1.3% for the interaction between SNPs rs2010907 and rs3760948 ($p = 0.002$), meaning that the synergistic interaction effect between the two SNPs explained 1.3% of the disease outcome. Six pairs of coding SNPs, that is, SNPs from genes, had significant interaction effects.

Table 2 Most significant pairwise interactions of the 44 identified SNPs.

SNP ₁ (gene ₁)	SNP ₂ (gene ₂)	Information gain (%)	p-value
rs2010907	rs3760948 (<i>ARRDC5</i>)	1.30	0.002
rs11185516 (<i>ZDHHC19</i>)	rs344570 (<i>TNFSF14</i>)	1.15	0.008
rs9688110 (<i>FAT2</i>)	rs658836	1.12	0.005
rs4625115	rs344570 (<i>TNFSF14</i>)	1.07	0.004
rs9288684 (<i>INPP5D</i>)	rs2179321 (<i>PLCB4</i>)	1.04	0.015
rs1367128 (<i>THSD7B</i>)	rs8022574	1.02	0.016
rs11185516 (<i>ZDHHC19</i>)	rs3842986	1.01	0.010
rs10814848 (<i>GLIS3</i>)	rs6578849 (<i>SYT9</i>)	0.98	0.008
rs1505229 (<i>LRRTM4</i>)	rs952880 (<i>KCNQ5</i>)	0.98	0.011
rs1505229 (<i>LRRTM4</i>)	rs11610311	0.98	0.013
rs9288684 (<i>INPP5D</i>)	rs2386946 (<i>CDH4</i>)	0.96	0.018
rs1991915 (<i>OTOP1</i>)	rs3842986	0.95	0.019
rs11783793	rs11610311	0.93	0.012
rs721619 (<i>EPHX2</i>)	rs4961513	0.92	0.017
rs4625115	rs2571219 (<i>ATP8B1</i>)	0.91	0.017
rs12695485 (<i>LOC107986044</i>)	rs898438 (<i>DCC</i>)	0.79	0.016
rs9288684 (<i>INPP5D</i>)	rs11985944	0.70	0.015

Table 3 Most significant three-way interactions of the 44 identified SNPs.

SNP ₁ (gene ₁)	SNP ₂ (gene ₂)	SNP ₃ (gene ₃)	Information gain (%)	p-value
rs1991915 (<i>OTOP1</i>)	rs721619 (<i>EPHX2</i>)	rs17831158 (<i>LINC</i>)	2.55	<0.001
rs1816647	rs10814848 (<i>GLIS3</i>)	rs3760948 (<i>ARRDC5</i>)	2.43	<0.001
rs11185516 (<i>ZDHHC</i>)	rs11985944	rs3844138	2.41	<0.001
rs2010907	rs9688110 (<i>FAT2</i>)	rs4625115	2.27	<0.001
rs7594717 (<i>ALK</i>)	rs721619 (<i>EPHX2</i>)	rs3760948 (<i>ARRDC5</i>)	2.25	<0.001
rs12695485 (<i>LOC107</i>)	rs17831158 (<i>LINC</i>)	rs13263313 (<i>JPH1</i>)	2.25	<0.001
rs647831	rs2736486	rs952880 (<i>KCNQ5</i>)	2.23	<0.001
rs1991915 (<i>OTOP1</i>)	rs8022574	rs2571219 (<i>ATP8B1</i>)	2.22	<0.001
rs11185516 (<i>ZDHHC</i>)	rs2736486	rs10814848 (<i>GLIS3</i>)	2.11	<0.001
rs7594717 (<i>ALK</i>)	rs13263313 (<i>JPH1</i>)	rs898438 (<i>DCC</i>)	1.93	0.001
rs17379465	rs2645737 (<i>NID2</i>)	rs658836	1.90	0.001
rs1367128 (<i>THSD7B</i>)	rs17831158 (<i>LINC</i>)	rs2386946 (<i>CDH4</i>)	1.89	0.001
rs12407198 (<i>C1orf101</i>)	rs10016091 (<i>SCFD2</i>)	rs2010907	1.86	0.001
rs1816647	rs6782709 (<i>LOC105</i>)	rs4961513	1.82	0.001
rs4961513	rs11610311	rs3842986	1.82	0.001
rs1367128 (<i>THSD7B</i>)	rs6578849 (<i>SYT9</i>)	rs344570 (<i>TNFSF14</i>)	1.40	0.001

We then calculated the three-way information gain among all $\binom{44}{3}$ possible trios of the 44 identified SNPs. Table 3 lists the 16 most significant ($p \leq 0.001$) SNP trios with their information gain measures and p-values. The strongest three-interaction was

Table 4 Enriched gene ontology (GO) terms on the 29 identified genes.

Category	Term	Gene count	p-value
GAD_DISEASE	Tobacco use disorder	14	3.9×10^{-5}
GAD_DISEASE_CLASS	Chemdependency	14	3.0×10^{-4}
GAD_DISEASE_CLASS	Metabolic	15	3.6×10^{-3}
GOTERM_MF_DIRECT	Calcium ion binding	5	6.1×10^{-3}
GAD_DISEASE_CLASS	Cardiovascular	13	6.5×10^{-3}
GOTERM_MF_FAT	Calcium ion binding	5	6.7×10^{-3}
GOTERM_BP_FAT	Movement of cell	7	1.2×10^{-2}
GOTERM_CC_DIRECT	Integral component of membrane	12	1.7×10^{-2}
GOTERM_MF_FAT	Metal ion binding	10	2.0×10^{-2}
GOTERM_BP_FAT	Neuron development	5	2.2×10^{-2}
GOTERM_MF_FAT	Cation binding	10	2.2×10^{-2}
GOTERM_BP_FAT	Locomotion	6	2.5×10^{-2}
GOTERM_MF_FAT	Ion binding	10	2.8×10^{-2}
GOTERM_CC_DIRECT	Plasma membrane	10	3.1×10^{-2}
GOTERM_BP_FAT	Cell migration	5	4.2×10^{-2}
GOTERM_BP_FAT	Neuron differentiation	5	4.7×10^{-2}

found among SNPs from genes *OTOPI*, *EPHX2*, and *LINC* with a strength of 2.55% and a significance level of $p < 0.001$, suggesting that the pure three-way synergistic effect among the three SNPs can explain 2.55% of the disease outcome. There were six trios of significant interacting SNPs all from coding regions of DNA.

Enriched gene functional terms

We submitted the identified 29 genes to the DAVID software, and chose categories of Disease, GO, and Pathway for a functional enrichment analysis. We set the gene count threshold as five, that is, only functional categories including more than 4 of the 29 identified genes were considered, and the significance cutoff as 0.05, that is, Fisher's exact test $p < 0.05$. Table 4 lists the most significantly enriched functional categories. There were 16 enriched terms, four of which were diseases, and the rest were GO terms. The most significantly enriched term was the disease *tobacco use disorder* with 14 out of 29 genes in the category and a significance level of 3.9×10^{-5} .

DISCUSSION

Identifying genetic markers associated with complex human diseases helps us better understand the disease etiology in order to better diagnose, treat, and even prevent diseases. Given the complexity of human diseases, especially cancers, the causing factors are more plausibly interactions among multiple genetic attributes instead of individual contributions. However, searching for combinations of attributes imposes a significant challenge for bioinformatics and genome-wide association research, since thousands to a million possible genetic attributes can be included for investigation.

Powerful machine learning algorithms have been used for mining high-volume data in fields such as engineering, finance, and social sciences, and have started to see applications in analyzing high-dimensional biomedical data as well.

In this article, we explored the application of two ensemble learning algorithms, random forests, and GBM, in identifying interacting genetic attributes associated with CRC. We studied a GWAS CRC dataset collected from the Canadian province of Newfoundland. We performed data preprocessing and filtering using the TuRF feature selection algorithm. By parameter tuning, we optimized the parameters for both random forests and GBM for GWAS data analyses. Both ensemble learning algorithms produced rankings on the importance of SNP contribution to the disease classification. By comparing the rankings provided by both algorithms, we identified a set of 44 top ranked SNPs from both coding and non-coding regions of DNA (Table 1). The coding SNPs mapped to 29 genes, which included both known CRC association genes: *DCC*, *ALK*, *ITGA1*, *E2F3*, and *NID2*, and unknown but potential CRC association genes.

We performed the statistical interaction analysis of the 44 identified SNPs and were able to validate strong and significant pairwise and three-way gene-gene interactions (Tables 2 and 3). These included a three-way interaction among *ALK*, *JPH1*, and *DCC*. In addition, functional enrichment analysis on the set of the 29 identified genes suggested 16 significantly enriched functional terms including four diseases: *tobacco use disorder*, *chemical dependency*, *metabolic diseases*, and *cardiovascular diseases*, and 12 GO terms (Table 4). We highlight some important findings in the following paragraphs.

SNP rs3760948 from gene arrestin domain-containing 5 (*ARRDC5*) was ranked the highest by both random forests and GBM (Fig. 4). It was detected with the strongest pairwise interaction with a non-coding SNP rs2010907 (Table 2) and the second strongest three-way interaction with SNP rs10814848 from gene GLIS family zinc finger 3 (*GLIS3*) and non-coding SNP rs1816647 (Table 3). This is a supportive evidence that both ensemble algorithms can detect interacting genetic attributes.

The gene “deleted in CRC” (*DCC*) is a tumor suppressor in CRC (Castets et al., 2012) and is well known to be associated with CRC (ENSEMBL database: http://www.ensembl.org/Homos_sapiens/Gene/Phenotype?db=core;g=ENSG00000187323). Anaplastic lymphoma kinase (*ALK*) gene is directly related to colorectal adenocarcinoma, and may affect treatments for advanced CRC (Lipson et al., 2012; Aisner et al., 2014; Pietrantonio et al., 2014). Genes *DCC* and *ALK*, along with junctophilin-1 *JPH1*, were found having a strong (1.93%) and significant ($p = 0.001$) three-way interaction. Gene $\alpha 1$ -integrin (*ITGA1*) and transcription factor *E2F3* have been found related to the prevention of tumor progression in colon cancer patients (Van Slambrouck et al., 2007; Akao et al., 2011; Boudjadi et al., 2016). Gene nidogen-2 (*NID2*) has been found to be associated with lung cancer (Zhang et al., 2014).

Our results also suggested novel genes associated with the disease of CRC. For instance, enzyme gene phosphatidylinositol-3,4,5-trisphosphate 5-phosphatase 1 (*INPP5D*) had strong and significant pairwise interactions with another enzyme

gene 1-Phosphatidylinositol-4,5-bisphosphate phosphodiesterase beta-4 (*PLCB4*) and gene cadherin-4 (*CDH4*), as well as a non-coding SNP rs11985944. We have not seen any published literature associating *INPP5D* with CRC and related diseases, and it can be a possible risk association gene to investigate in biology.

We found 14 genes associated with the functional term *tobacco use disorder* ($p = 3.9 \times 10^{-5}$) (Table 4). In a study on tobacco use and the risk of CRC using a retrospective cohort study of germ-line mutants, tobacco use was found to significantly increase the risk of CRC (Watson *et al.*, 2004). Another significantly enriched functional term was *cell migration*, which has been reported to be related to cancers (Friedl & Wolf, 2003; Paul, Mistriotis & Konstantopoulos, 2017). The tumor cells migrate and enter the blood and will go to other tissues and spread cancer. When CRC cells spread, they most often spread to the liver. Our results suggest the possible importance of confining cell mobility in CRC treatment and prevention.

Our future studies include: (1) utilizing more computational power in order to include more SNP attributes in the classifier training and feature importance analysis and (2) exploring or combining other machine learning algorithms, including neural networks and evolutionary algorithms, for the search of gene-gene interactions.

CONCLUSIONS

Machine learning algorithms have seen increasing applications in bioinformatics and computational biology thanks to their powerful abilities of automatic learning and modeling complex relationships among a large number of features. Applications of machine learning techniques require domain-specific tailoring and careful design since an application area has unique problem definitions and challenges. In the context of GWAS, we need algorithms that are able to detect the non-linear, non-additive interactions among multiple genetic factors that contribute to the disease outcome. This article demonstrates a novel design of an informatics framework of using two ensemble learning algorithms to search for interacting genetic factors associated with cancer and validating the results through statistical interaction analysis and biological functional enrichment analysis. With further biological experiments, our bioinformatics findings may help us better understand the disease etiology of CRC. We also hope our study can inspire more bioinformatics tool developments for human disease association studies.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Ignite Grant from the Research and Development Corporation of Newfoundland and Labrador and the Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

Ignite Grant from the Research and Development Corporation of Newfoundland and Labrador.

Natural Sciences and Engineering Research Council of Canada (NSERC).

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Faramarz Dorani conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Ting Hu conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Michael O. Woods contributed reagents/materials/analysis tools, approved the final draft, data collection and preliminary processing.
- Guangju Zhai contributed reagents/materials/analysis tools, approved the final draft, data collection and preliminary processing.

Human Ethics

The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):

Memorial University Granted Ethical approval to carry out the study within its facilities (approval number HIC 01.70).

Data Availability

The following information was supplied regarding data availability:

Colorectal Transdisciplinary (CORECT) Study: <https://research.fhcrc.org/peters/en/corect-study.html>

GitHub: <https://github.com/MIBlab-MUN/ensemble-learning-GWAS.git>.

REFERENCES

- Aisner DL, Nguyen TT, Paskulin DD, Le AT, Haney J, Schulte N, Chionh F, Hardingham J, Mariadason J, Tebbutt N, Doebele RC, Weickhardt AJ, Varella-Garcia M. 2014. ROS1 and ALK fusions in colorectal cancer, with evidence of intratumoral heterogeneity for molecular drivers. *Molecular Cancer Research* **12**(1):111–118 DOI [10.1158/1541-7786.mcr-13-0479-t](https://doi.org/10.1158/1541-7786.mcr-13-0479-t).
- Akao Y, Noguchi S, Iio A, Kojima K, Takagi T, Naoe T. 2011. Dysregulation of microRNA-34a expression causes drug-resistance to 5-fu in human colon cancer dld-1 cells. *Cancer Letters* **300**(2):197–204 DOI [10.1016/j.canlet.2010.10.006](https://doi.org/10.1016/j.canlet.2010.10.006).
- Balding DJ. 2006. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7**(10):781–791 DOI [10.1038/nrg1916](https://doi.org/10.1038/nrg1916).
- Beam AL, Motsinger-Reif A, Doyle J. 2014. Bayesian neural networks for detecting epistasis in genetic association studies. *BMC Bioinformatics* **15**(1):368 DOI [10.1186/s12859-014-0368-0](https://doi.org/10.1186/s12859-014-0368-0).

- Beretta L, Santaniello A. 2011.** Implementing relief filters to extract meaningful features from genetic lifetime datasets. *Journal of Biomedical Informatics* **44**(2):361–369 DOI [10.1016/j.jbi.2010.12.003](https://doi.org/10.1016/j.jbi.2010.12.003).
- Boudjadi S, Carrier J, Groulx JF, Beaulieu JF. 2016.** Integrin $\alpha 1\beta 1$ expression is controlled by c-myc in colorectal cancer cells. *Oncogene* **35**(13):1671–1678 DOI [10.1038/onc.2015.231](https://doi.org/10.1038/onc.2015.231).
- Breiman L. 1996.** Bagging predictors. *Machine Learning* **24**(2):123–140 DOI [10.1007/bf00058655](https://doi.org/10.1007/bf00058655).
- Breiman L. 2001.** Random forests. *Machine Learning* **45**(1):5–32.
- Breiman L, Friedman JH, Olshen RA, Stone CJ. 2017.** *Classification and regression trees*. Boca Raton, FL, USA: Chapman & Hall/CRC.
- Bush WS, Moore JH. 2012.** Genome-wide association studies. *PLOS Computational Biology* **8**(12):e1002822.
- Castets M, Broutier L, Molin Y, Brevet M, Chazot G, Gadot N, Paquet A, Mazelin L, Jarrosson-Wuilleme L, Scoazec J-Y, Bernet A, Mehlen P. 2012.** DCC constrains tumour progression via its dependence receptor activity. *Nature* **482**(7386):534–537 DOI [10.1038/nature10708](https://doi.org/10.1038/nature10708).
- Chen X, Ishwaran H. 2012.** Random forests for genomic data analysis. *Genomics* **99**(6):323–329 DOI [10.1016/j.ygeno.2012.04.003](https://doi.org/10.1016/j.ygeno.2012.04.003).
- D'Angelo GM, Rao DC, Gu CC. 2009.** Combining least absolute shrinkage and selection operator (lasso) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. *BMC Proceedings* **3**(Suppl 7):S62 DOI [10.1186/1753-6561-3-s7-s62](https://doi.org/10.1186/1753-6561-3-s7-s62).
- Dash M, Liu H. 1997.** Feature selection for classification. *Intelligent Data Analysis* **1**(1–4):131–156.
- Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003.** DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology* **4**(9):R60. Available at <https://genomebiology.biomedcentral.com/track/pdf/10.1186/gb-2003-4-9-r60>.
- Dietterich TG. 2000.** Ensemble methods in machine learning. *Multiple Classifier Systems* **1857**:1–15.
- Dorani F, Hu T. 2018.** Feature selection for detecting gene-gene interactions in genome-wide association studies. In: *Proceedings of 21st European Conference on the Applications of Evolutionary Computation (EvoApplications), Vol. 10784 of Lecture Notes in Computer Science*. Heidelberg, Germany: Springer, 33–46.
- Easton DF, Eeles RA. 2008.** Genome-wide association studies in cancer. *Human Molecular Genetics* **17**(R2):R109–R115 DOI [10.1093/hmg/ddn287](https://doi.org/10.1093/hmg/ddn287).
- Foulkes AS. 2009.** *Applied Statistical Genetics With R: For Population-Based Association Studies*. New York, NY, USA: Springer Science & Business Media.
- Frazer KA, Murray SS, Schork NJ, Topol EJ. 2009.** Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* **10**(4):241–251 DOI [10.1038/nrg2554](https://doi.org/10.1038/nrg2554).
- Friedl P, Wolf K. 2003.** Tumour-cell invasion and migration: diversity and escape mechanisms. *Nature Reviews Cancer* **3**(5):362–374 DOI [10.1038/nrc1075](https://doi.org/10.1038/nrc1075).
- Friedman JH. 2001.** Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**(5):1189–1232.
- Friedman J, Hastie T, Tibshirani R. 2001.** *The Elements of Statistical Learning, Vol. 1 of Springer Series in Statistics*. Berlin: Springer.
- Garca-Magariños M, López-De Ullibarri I, Cao R, Salas A. 2009.** Evaluating the ability of tree-based methods and logistic regression for the detection of snp-snp interaction. *Annals of Human Genetics* **73**(3):360–369 DOI [10.1111/j.1469-1809.2009.00511.x](https://doi.org/10.1111/j.1469-1809.2009.00511.x).

- Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. 2010.** An application of random forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genetics* **11(1)**:49 DOI [10.1186/1471-2156-11-49](https://doi.org/10.1186/1471-2156-11-49).
- Guyon I, Elisseeff A. 2003.** An introduction to variable and feature selection. *Journal of Machine Learning Research* **3(Mar)**:1157–1182.
- Han B, Chen XW. 2011.** bNEAT: a Bayesian network method for detecting epistatic interactions in genome-wide association studies. *BMC Genomics* **12(suppl 2)**:S9 DOI [10.1186/1471-2164-12-s2-s9](https://doi.org/10.1186/1471-2164-12-s2-s9).
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009.** Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United State of America* **106(23)**:9362–9367 DOI [10.1073/pnas.0903103106](https://doi.org/10.1073/pnas.0903103106).
- Hirschhorn JN, Daly MJ. 2005.** Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6(2)**:95–108 DOI [10.1038/nrg1521](https://doi.org/10.1038/nrg1521).
- Hu T, Chen Y, Kiralis JW, Collins RL, Wejse C, Sirugo G, Williams SM, Moore JH. 2013a.** An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *Journal of the American Medical Informatics Association* **20(4)**:630–636 DOI [10.1136/amiajnl-2012-001525](https://doi.org/10.1136/amiajnl-2012-001525).
- Hu T, Chen Y, Kiralis JW, Moore JH. 2013b.** ViSEN: methodology and software for visualization of statistical epistasis networks. *Genetic Epidemiology* **37(3)**:283–285 DOI [10.1002/gepi.21718](https://doi.org/10.1002/gepi.21718).
- Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, Moore JH. 2011.** Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics* **12(1)**:364 DOI [10.1186/1471-2105-12-364](https://doi.org/10.1186/1471-2105-12-364).
- Huang DW, Sherman BT, Lempicki RA. 2008.** Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37(1)**:1–13 DOI [10.1093/nar/gkn923](https://doi.org/10.1093/nar/gkn923).
- Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. 2010.** Inferring regulatory networks from expression data using tree-based methods. *PLOS ONE* **5(9)**:e12776 DOI [10.1371/journal.pone.0012776](https://doi.org/10.1371/journal.pone.0012776).
- Jing P-J, Shen H-B. 2015.** MACOED: a multi-objective ant colony optimization algorithm for snp epistasis detection in genome-wide association studies. *Bioinformatics* **31(5)**:634–641 DOI [10.1093/bioinformatics/btu702](https://doi.org/10.1093/bioinformatics/btu702).
- Kim Y, Wojciechowski R, Sung H, Mathias RA, Wang L, Klein AP, Lenroot RK, Malley J, Bailey-Wilson JE. 2009.** Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. *BMC Proceedings* **3(Suppl 7)**:S64 DOI [10.1186/1753-6561-3-s7-s64](https://doi.org/10.1186/1753-6561-3-s7-s64).
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarrroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P,**

- Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–291.
- Lettre G, Rioux JD. 2008. Autoimmune diseases: insights from genome-wide association studies. *Human Molecular Genetics* 17(R2):R116–R121 DOI 10.1093/hmg/ddn246.
- Libbrecht MW, Noble WS. 2015. Machine learning applications in genetics and genomics. *Nature Reviews Genetics* 16(6):321–332 DOI 10.1038/nrg3920.
- Lipson D, Capelletti M, Yelensky R, Otto G, Parker A, Jarosz M, Curran JA, Balasubramanian S, Bloom T, Brennan KW, Donahue A, Downing SR, Frampton GM, Garcia L, Juhn F, Mitchell KC, White E, White J, Zwirko Z, Peretz T, Nechushtan H, Soussan-Gutman L, Kim J, Sasaki H, Kim HR, il Park S, Ercan D, Sheehan CE, Ross JS, Cronin MT, Jänne PA, Stephens PJ. 2012. Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. *Nature Medicine* 18(3):382–384 DOI 10.1038/nm.2673.
- Mackay TFC, Moore JH. 2014. Why epistasis is important for tackling complex human disease genetics. *Genome Medicine* 6(6):125 DOI 10.1186/gm561.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. *Nature* 461(7265):747–753 DOI 10.1038/nature08494.
- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 9(5):356–369 DOI 10.1038/nrg2344.
- Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, Schmidt MK, Chang-Claude J, Bojesen SE, Bolla MK, Wang Q, Dicks E, Lee A, Turnbull C, Rahman N, Breast T, Collaboration OCS, Fletcher O, Peto J, Gibson L, Dos Santos Silva I, Nevanlinna H, Muranen TA, Aittomäki K, Blomqvist C, Czene K, Irwanto A, Liu J, Waisfisz Q, Meijers-Heijboer H, Adank M, Breast H, (HEBON) OCRGN, Van Der Luijt RB, Hein R, Dahmen N, Beckman L, Meindl A, Schmutzler RK, Müller-Myhsok B, Lichtner P, Hopper JL, Southey MC, Makalic E, Schmidt DF, Uitterlinden AG, Hofman A, Hunter DJ, Chanock SJ, Vincent D, Bacot F, Tessier DC, Canisius S, Wessels LFA, Haiman CA, Shah M, Luben R, Brown J, Luccarini C, Schoof N, Humphreys K, Li J, Nordestgaard BG, Nielsen SF, Flyger H, Couch FJ, Wang X, Vachon C, Stevens KN, Lambrechts D, Moisse M, Paridaens R, Christiaens M-R, Rudolph A, Nickels S, Flesch-Janys D, Johnson N, Aitken Z, Aaltonen K, Heikkinen T, Broeks A, Veer LJV, Van Der Schoot CE, Guénel P, Truong T, Laurent-Puig P, Menegaux F, Marme F, Schneeweiss A, Sohn C, Burwinkel B, Zamora MP, Perez JIA, Pita G, Alonso MR, Cox A, Brock IW, Cross SS, Reed MWR, Sawyer EJ, Tomlinson I, Kerin MJ, Miller N, Henderson BE, Schumacher F, Marchand LL, Andrulis IL, Knight JA, Glendon G, Mulligan AM, Lindblom A, Margolin S, Hooning MJ, Hollestelle A, Van Den Ouweland AMW, Jager A, Bui QM, Stone J, Dite GS, Apicella C, Tsimiklis H, Giles GG, Severi G, Baglietto L, Fasching PA, Haerle L, Ekici AB, Beckmann MW, Brenner H, Müller H, Arndt V, Stegmaier C, Swerdlow A, Ashworth A, Orr N, Jones M, Figueroa J, Lissowska J, Brinton L, Goldberg MS, Labrèche F, Dumont M, Winqvist R, Pylkäs K, Jukkola-Vuorinen A, Grip M, Brauch H, Hamann U, Brüning T, Radice P, Peterlongo P, Manoukian S, Bonanni B, Devilee P, Tollenaar RAEM, Seynaeve C, Van Asperen CJ, Jakubowska A, Lubinski J, Jaworska K, Durda K, Mannermaa A, Kataja V, Kosma V-M, Hartikainen JM, Bogdanova NV, Antonenkova NN, Dörk T, Kristensen VN, Anton-Culver H, Slager S, Toland AE, Edge S,

- Fostira F, Kang D, Yoo K-Y, Noh D-Y, Matsuo K, Ito H, Iwata H, Sueta A, Wu AH, Tseng C-C, Berg DVD, Stram DO, Shu X-O, Lu W, Gao Y-T, Cai H, Teo SH, Yip CH, Phuah SY, Cornes BK, Hartman M, Miao H, Lim WY, Sng J-H, Muir K, Lophatananon A, Stewart-Brown S, Siriwanarangsana P, Shen C-Y, Hsiung C-N, Wu P-E, Ding S-L, Sangrajrang S, Gaborieau V, Brennan P, McKay J, Blot WJ, Signorello LB, Cai Q, Zheng W, Deming-Halverson S, Shrubsole M, Long J, Simard J, Garcia-Closas M, Pharoah PDP, Chenevix-Trench G, Dunning AM, Benitez J, Easton DF, Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON), kConFab Investigators, Australian Ovarian Cancer Study Group, The GENICA (Gene Environment Interaction and Breast Cancer in Germany) Network. 2013. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genetics* 45(4):353–361 DOI 10.1038/ng.2563.
- Mohlke KL, Boehnke M, Abecasis GR. 2008. Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. *Human Molecular Genetics* 17(R2):R102–R108 DOI 10.1093/hmg/ddn275.
- Moore JH, Asselbergs FW, Williams SM. 2010. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26(4):445–455 DOI 10.1093/bioinformatics/btp713.
- Moore JH, Ritchie MD. 2004. The challenges of whole-genome approaches to common diseases. *JAMA: Journal of the American Medical Association* 291(13):1642–1643 DOI 10.1001/jama.291.13.1642.
- Moore JH, Williams SM. 2002. New strategies for identifying gene-gene interactions in hypertension. *Annals of Medicine* 34(2):88–95 DOI 10.1080/07853890252953473.
- Moore JH, Williams SM. 2009. Epistasis and its implications for personal genetics. *American Journal of Human Genetics* 85(3):309–320 DOI 10.1016/j.ajhg.2009.08.006.
- Niel C, Sinoquet C. 2018. Enhancement of a stochastic Markov blanket framework with ant colony optimization, to uncover epistasis in genetic association studies. *Proceedings of the 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Belgium: Ottignies-Louvain-la-Neuve, 673–678.
- Niel C, Sinoquet C, Dina C, Rocheleau G. 2015. A survey about methods dedicated to epistasis detection. *Frontiers in Genetics* 6:285 DOI 10.3389/fgene.2015.00285.
- Niel C, Sinoquet C, Dina C, Rocheleau G. 2018. SMMB: a stochastic Markov blanket framework strategy for epistasis detection in GWAS. *Bioinformatics* 34(16):2773–2780 DOI 10.1093/bioinformatics/bty154.
- Nikpay M, Goel A, Won H-H, Hall LM, Willenborg C, Kanoni S, Saleheen D, Kyriakou T, Nelson CP, Hopewell JC, Webb TR, Zeng L, Dehghan A, Alver M, Armasu SM, Auro K, Bjornnes A, Chasman DI, Chen S, Ford I, Franceschini N, Gieger C, Grace C, Gustafsson S, Huang J, Hwang S-J, Kim YK, Kleber ME, Lau KW, Lu X, Lu Y, Lyytikäinen L-P, Mihailov E, Morrison AC, Pervjakova N, Qu L, Rose LM, Salfati E, Saxena R, Scholz M, Smith AV, Tikkanen E, Uitterlinden A, Yang X, Zhang W, Zhao W, De Andrade M, De Vries PS, Van Zuydam NR, Anand SS, Bertram L, Beutner F, Dedoussis G, Frossard P, Gauguier D, Goodall AH, Gottesman O, Haber M, Han B-G, Huang J, Jalilzadeh S, Kessler T, König IR, Lannfelt L, Lieb W, Lind L, Lindgren CM, Lokki M-L, Magnusson PK, Mallick NH, Mehra N, Meitinger T, Memon F-ur-R, Morris AP, Nieminen MS, Pedersen NL, Peters A, Rallidis LS, Rasheed A, Samuel M, Shah SH, Sinisalo J, Stirrups KE, Trompet S, Wang L, Zaman KS, Ardissino D, Boerwinkle E, Borecki IB, Bottinger EP, Buring JE, Chambers JC, Collins R, Cupples LA, Danesh J, Demuth I, Elosua R, Epstein SE, Esko T, Feitosa MF, Franco OH, Franzosi MG, Granger CB, Gu D, Gudnason V, Hall AS, Hamsten A, Harris TB, Hazen SL, Hengstenberg C, Hofman A, Ingelsson E, Iribarren C, Jukema JW, Karhunen PJ, Kim B-J, Kooner JS, Kullo IJ, Lehtimäki T, Loos RJF, Melander O,

- Metspalu A, März W, Palmer CN, Perola M, Quertermous T, Rader DJ, Ridker PM, Ripatti S, Roberts R, Salomaa V, Sanghera DK, Schwartz SM, Seedorf U, Stewart AF, Stott DJ, Thiery J, Zalloua PA, O'Donnell CJ, Reilly MP, Assimes TL, Thompson JR, Erdmann J, Clarke R, Watkins H, Kathiresan S, McPherson R, Deloukas P, Schunkert H, Samani NJ, Farrall M. 2015. A comprehensive 1,000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* 47(10):1121–1130 DOI 10.1038/ng.3396.
- Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. 2014. Regularized machine learning in the genetic prediction of complex traits. *PLOS Genetics* 10(11):e1004754 DOI 10.1371/journal.pgen.1004754.
- Olson RS, La Cava W, Mustahsan Z, Varik A, Moore JH. 2017. Data-driven advice for applying machine learning to bioinformatics problems. arXiv:1708.05070 preprint DOI 10.1142/9789813235533_0018.
- Pan Q, Hu T, Malley JD, Andrew AS, Karagas MR, Moore JH. 2014. A system-level pathway-phenotype association analysis using synthetic feature random forest. *Genetic Epidemiology* 38(3):209–219 DOI 10.1002/gepi.21794.
- Paul CD, Mistriotis P, Konstantopoulos K. 2017. Cancer cell motility: lessons from migration in confined spaces. *Nature Reviews Cancer* 17(2):131–140 DOI 10.1038/nrc.2016.123.
- Pietrantonio F, Maggi C, Di Bartolomeo M, Facciorusso M, Perrone F, Testi A, Lacovelli R, Miceli R, Bossi I, Leone G, Milione M, Pelosi G, De Braud F. 2014. Gain of ALK gene copy number may predict lack of benefit from anti-EGFR treatment in patients with advanced colorectal cancer and RAS-RAF-PI3KCA wild-type status. *PLOS ONE* 9(4):e92147 DOI 10.1371/journal.pone.0092147.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81(3):559–575 DOI 10.1086/519795.
- Ridgeway G. 2007. *Generalized boosted regression models: a guide to the GBM package*. Available at <https://cran.r-project.org/web/packages/gbm/>.
- Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann H-E, Barrett JH, König IR, Stevens SE, Szymczak S, Tregouet D-A, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Brønne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H. 2007. Genomewide association analysis of coronary artery disease. *New England Journal of Medicine* 357(5):443–453.
- Schumacher FR, Schmit SL, Jiao S, Edlund CK, Wang H, Zhang B, Hsu L, Huang S-C, Fischer CP, Harju JF, Idos GE, Lejbkovicz F, Manion FJ, McDonnell K, McNeil CE, Melas M, Rennert HS, Shi W, Thomas DC, Berg DJVD, Hutter CM, Aragaki AK, Butterbach K, Caan BJ, Carlson CS, Chanock SJ, Curtis KR, Fuchs CS, Gala M, Giovannucci EL, Gogarten SM, Hayes RB, Henderson B, Hunter DJ, Jackson RD, Kolonel LN, Kooperberg C, Küry S, LaCroix A, Laurie CC, Laurie CA, Lemire M, Levine D, Ma J, Makar KW, Qu C, Taverna D, Ulrich CM, Wu K, Kono S, West DW, Berndt SI, Bezieau S, Brenner H, Campbell PT, Chan AT, Chang-Claude J, Coetzee GA, Conti DV, Duggan D, Figueiredo JC, Fortini BK, Gallinger SJ, Gauderman WJ, Giles G, Green R, Haile R, Harrison TA, Hoffmeister M, Hopper JL, Hudson TJ, Jacobs E, Iwasaki M, Jee SH, Jenkins M, Jia W-H, Joshi A, Li L, Lindor NM, Matsuo K, Moreno V, Mukherjee B, Newcomb PA, Potter JD, Raskin L, Rennert G, Rosse S, Severi G, Schoen RE, Seminara D, Shu X-O, Slattery ML, Tsugane S, White E, Xiang Y-B, Zanke BW, Zheng W, Marchand LL,

- Casey G, Gruber SB, Peters U. 2015. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nature Communications* 6(1):7138 DOI 10.1038/ncomms8138.
- Schwarz DF, Inke KR, Ziegler A. 2010. On safari to random jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics* 26(14):1752–1758 DOI 10.1093/bioinformatics/btq257.
- Schwarz DF, Szymczak S, Ziegler A, König IR. 2007. Picking single-nucleotide polymorphisms in forests. *BMC Proceedings* 1(Suppl 1):S59 DOI 10.1186/1753-6561-1-s1-s59.
- Sun YV, Bielak LF, Peyser PA, Turner ST, Sheedy PF, Boerwinkle E, Kardia SL. 2008. Application of machine learning algorithms to predict coronary artery calcification with a sibship-based design. *Genetic Epidemiology* 32(4):350–360 DOI 10.1002/gepi.20309.
- Sun YV, Cai Z, Desai K, Lawrance R, Leff R, Jawaid A, Kardia SL, Yang H. 2007. Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests. *BMC Proceedings* 1(Suppl 1):S62 DOI 10.1186/1753-6561-1-s1-s62.
- Sun Y, Shang J, Xing Liu J, Li S, Zheng C-H. 2017. epiACO—a method for identifying epistasis based on ant colony optimization algorithm. *BioData Mining* 10(1):23 DOI 10.1186/s13040-017-0143-7.
- Sun YV, Shedden KA, Zhu J, Choi N-H, Kardia SL. 2009. Identification of correlated genetic variants jointly associated with rheumatoid arthritis using ridge regression. *BMC Proceedings* 3(Suppl 7):S67 DOI 10.1186/1753-6561-3-s7-s67.
- Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, Zhang H, Sun YV. 2009. Machine learning in genome-wide association studies. *Genetic Epidemiology* 33(S1):S51–S57 DOI 10.1002/gepi.20473.
- Tang R, Sinnwell JP, Li J, Rider DN, De Andrade M, Biernacka JM. 2009. Identification of genes and haplotypes that predict rheumatoid arthritis using random forests. *BMC Proceedings* 3(Suppl 7):S68 DOI 10.1186/1753-6561-3-s7-s68.
- Upstill-Goddard R, Eccles D, Fliege J, Collins A. 2012. Machine learning approaches for the discovery of gene-gene interactions in disease data. *Briefings in Bioinformatics* 14(2):251–260 DOI 10.1093/bib/bbs024.
- Urbanowicz RJ, Olson RS, Schmitt P, Meeker M, Moore JH. 2018. Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of Biomedical Informatics* 85:168–188 DOI 10.1016/j.jbi.2018.07.015.
- Van Slambrouck S, Grijelmo C, De Wever O, Bruyneel E, Emami S, Gespach C, Steelant WF. 2007. Activation of the fak-src molecular scaffolds and p130cas-jnk signaling cascades by α 1-integrins during colon cancer cell invasion. *International Journal of Oncology* 31(6):1501–1508 DOI 10.3892/ijo.31.6.1501.
- Wang M, Chen X, Zhang M, Zhu W, Cho K, Zhang H. 2009. Detecting significant single-nucleotide polymorphisms in a rheumatoid arthritis study using random forests. *BMC Proceedings* 3(Suppl 7):S69 DOI 10.1186/1753-6561-3-s7-s69.
- Watson P, Ashwathnarayan R, Lynch HT, Roy HK. 2004. Tobacco use and increased colorectal cancer risk in patients with hereditary nonpolyposis colorectal cancer (lynch syndrome). *Archives of Internal Medicine* 164(22):2429–2431 DOI 10.1001/archinte.164.22.2429.
- Wright MN, Ziegler A. 2017. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77(1):1–17 DOI 10.18637/jss.v077.i01.

- Wright MN, Ziegler A, König IR. 2016.** Do little interactions get lost in dark random forests? *BMC Bioinformatics* **17**(1):145 DOI [10.1186/s12859-016-0995-8](https://doi.org/10.1186/s12859-016-0995-8).
- Yoshida M, Koike A. 2011.** Snpinterforest: a new method for detecting epistatic interactions. *BMC Bioinformatics* **12**(1):469 DOI [10.1186/1471-2105-12-469](https://doi.org/10.1186/1471-2105-12-469).
- Yu L, Liu H. 2003.** Feature selection for high-dimensional data: a fast correlation-based filter solution. *ICML 3*. Palo Alto, CA, USA: AAAI Press, 856–863.
- Zhang R, Chu M, Zhao Y, Wu C, Guo H, Shi Y, Dai J, Wei Y, Jin G, Ma H, Dong J, Yi H, Bai J, Gong J, Sun C, Zhu M, Wu T, Hu Z, Lin D, Shen H, Chen F. 2014.** A genome-wide gene-environment interaction analysis for tobacco smoke and lung cancer susceptibility. *Carcinogenesis* **35**(7):1528–1535 DOI [10.1093/carcin/bgu076](https://doi.org/10.1093/carcin/bgu076).
- Zhang Y, Liu JS. 2007.** Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics* **39**(9):1167–1173 DOI [10.1038/ng2110](https://doi.org/10.1038/ng2110).
- Ziegler A, DeStefano AL, König IR. 2007.** Data mining, neural nets, trees—problems 2 and 3 of genetic analysis workshop 15. *Genetic Epidemiology* **31**(S1):S51–S60 DOI [10.1002/gepi.20280](https://doi.org/10.1002/gepi.20280).