Enhancing Selection of Alcohol Consumption Associated Genes by Random Forest

Chenglin Lyu [1,2], Roby Joehanes [3], Tianxiao Huan [3], Daniel Levy [3], Yi Li [1], Mengyao Wang [1], Xue Liu [1], Chunyu Liu [1*], Jiantao Ma [4*]

[1]Department of Biostatistics, Boston University School of Public Health, Boston, MA;
[2]Department of Anatomy and Neurobiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA;
[3]Framingham Heart Study and Population Sciences Branch, NHLBI, Framingham, MA;
[4]Nutrition Epidemiology and Data Science, Friedman School of Nutrition Science and Policy, Tufts University, Boston, MA

*****These authors contributed equally

Corresponding authors: Jiantao Ma, PhD, Nutrition Epidemiology Data Science, Friedman School, Tufts University, 155 Harrison Street, Boston, MA 02111, Email: jiantao.ma@tufts.edu
Chunyu Liu, PhD, Department of Biostatistics, Boston University, 715 Albany Street, Boston, MA 02118, Email: liuc@bu.edu

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

Abstract

Machine learning methods have been used in identifying omics markers for a variety of phenotypes. We aimed to examine whether a supervised machine learning algorithm can improve identification of alcohol-associated transcriptomic markers. In this study, we analyzed array-based, whole-blood derived expression data for 17,873 gene transcripts in 5,508 Framingham Heart Study participants. By using the Boruta algorithm, a supervised Random Forest (RF)-based feature selection method, we selected 25 alcohol-associated transcripts. In a testing set (30% of entire study participants), AUCs (area under the receiver operating characteristics curve) of these 25 transcripts were 0.73, 0.69, and 0.66 for nondrinkers vs. moderate drinkers, nondrinkers vs. heavy drinkers, and moderate drinkers vs. heavy drinkers, respectively. The AUCs of the selected transcripts by the Boruta method were comparable to those identified using conventional linear regression models, e.g., AUCs of 1,985 transcripts identified by conventional linear regression models (false discovery rate < 0.05) were 0.72, 0.68, and 0.68, respectively. With Bonferroni correction for the 25 Boruta method selected transcripts and three CVD risk factors (i.e., at P < 6.7e-4), we observed 13 transcripts were associated with obesity, 3 transcripts with type 2 diabetes, and 1 transcript with hypertension. For example, we observed that alcohol consumption was inversely associated with the expression of *DOCK4*, *IL4R*, and *SORT1*, and *DOCK4* and *SORT1* were positively associated with obesity and *IL4R* was inversely associated with hypertension. In conclusion, using a supervised machine learning method, the RF-based Boruta algorithm, we identified novel alcohol-associated gene transcripts.

Keywords: Alcohol consumption, gene expression, Cardiovascular disease, Machine learning, RandomForest, Boruta

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

Introduction

Alcohol consumption is an important lifestyle factor that has been associated with cardiovascular health. Excessive alcohol consumption leads to hypertension, dyslipidemia, and type 2 diabetes.(1, 2) Whereas moderate alcohol consumption may improve cardiovascular health despite that several recent studies suggest no beneficial relationship with reduction of cardiovascular disease (CVD).(3-5) The use of high-throughput transcriptomic analysis has been playing a significant role in investigating the pathogenesis of CVD.(6-8) In our previous study, using conventional linear regression models, we examined associations between alcohol consumption and transcriptomic markers in the community-based Framingham Heart Study (FHS).(9)

"Big Data" applications such as machine learning approaches provide new tools to discover novel biomarkers for better understanding of molecular mechanisms underlying diseases and to increase accuracy of disease predictions.(10) Random Forest (RF) is a supervised machine learning method that scores the importance of the features in a dataset.(11, 12) RF is a promising approach in prediction and classification for bias reduction.(11, 12) RF has been successfully applied in analyzing different types of omics biomarkers.(13-15) Boruta is an extension method based on RF to evaluate the importance of original features by comparing them with their randomized copies.(16) In essence, the Boruta method is an automatic feature selection method. The Boruta method has been used in over 100 studies in selecting omics biomarkers related to diseases or traits.(13) A recent study showed that, using simulated and published datasets, the Boruta method was a stable RF-based feature selection approach.(17)

Analysis using conventional linear regression may experience issues with multiple testing and cannot effectively handle high-order interactions among tested biomarkers.(18) Compared to conventional linear regression, RF method offers alternative analytical models that may have several advantages such as model flexibility.(19) RF-based approaches may improve the handling of high dimensional data by decorrelating the classifiers and minimizing the influence of over-fitting.(20) However, it is unclear whether using RF with automatic feature selection algorithms such as the Boruta method can identify additional alcohol-associated transcriptomic markers. To address this research question, we aimed to use the RF with the Boruta method to improve the identification of alcohol-associated gene transcripts and examine the associations of these gene transcripts with CVD risk factors in the FHS.

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

Methods

Study Participants. The FHS participants included in the present study are those who attended the eighth examination (2005 to 2008) of the Offspring cohort or the second examination (2008 to 2011) of the Third Generation cohort.(21, 22) The study sample of the present study was the same as that was used in our previous alcohol-gene transcripts analysis using conventional linear regression.(9) Briefly, after excluding participants with missing data on alcohol consumption and gene expression, we included 5,508 participants, 2,381 from the Offspring cohorts and 3,127 from the Third Generation cohort. The FHS protocols and procedures were approved by the Institutional Review Board for Human Research at Boston University Medical Center and all participants provided written informed consent. This study was conducted according to the guidelines laid down in the Declaration of Helsinki and all procedures involving human subjects were approved by the Institutional Review Board for Human Research at Boston University Medical Center (IRB number: H-41461). Written informed consent was obtained from all participants.

Alcohol consumption. Participants' alcohol consumption was measured by a technician administered questionnaire during the physical examination in the FHS clinic. Frequency of standard servings of beer, wine, and spirit consumed in a typical week or month were documented. We calculated the grams (g) of ethanol consumed each day using the following conversion factors: one 12 oz. beer has 14 g ethanol, one 4-5 oz. wine has 14 g ethanol, and one 1.5 oz. of 80 proof liquor has 14 g ethanol.(23) Based on the estimated daily alcohol consumption, we categorized our study participants into three groups, nondrinkers (n=1,729), moderate drinkers (0.1 to 28 g/day in women and 0.1 to 42 g/day in men; n=3,427), and heavy drinkers (> 28 g/day in women and > 42 g/day in men; n=352). We also split the moderate drinkers to light drinkers (0.1 to 14 g/day in women and 0.1 to 28 g/day in men; n=2806 and at-risk drinkers (14.1 to 28 g/day in women and 28.1 to 42 g/day in men; n=621) and conducted sensitivity analyses separately for the two groups.

Gene expression profiling. We analyzed gene expression levels that were measured using the GeneChip Human Exon 1.0 ST Array as described previously.(24) Briefly, fasting peripheral whole blood samples, from the same examinations that alcohol consumption was assessed, were

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

collected in PAXgene^{TM} tubes. Standard operating procedures were followed to isolate RNA using a KingFisher$^{®}$ 96 robot, and 50 ng RNA was amplified to create the cDNA library. The Affymetrix 7G GCS3000 scanner was used to measure gene expression levels, and the Human Exon 1.0 ST Array probeset was used to annotate gene transcripts. The final gene expression profiles were residuals of 17,873 transcripts of autosomal genes generated using linear mixed models with adjustment for technical covariates and other factors as fixed effects as well as batch as a random effect.(24)

CVD risk factors. Obesity, hypertension, and type 2 diabetes status at the same time for alcohol consumption and gene expression measurements were analyzed in the present study.(25) Obesity was defined as body mass index (BMI) ≥ 30 kg/m^2. Hypertension was defined as systolic blood pressure (SBP) ≥ 140 mm Hg or diastolic blood pressure (DBP) ≥ 90 mm Hg or taking antihypertensive drugs for high blood pressure. We also defined hypertension as SBP > 130 mm Hg or DBP > 80mm Hg or taking antihypertension drugs.(26) Type 2 diabetes was defined as fasting blood glucose level ≥ 126 mg/dL or taking antidiabetic drugs.

Statistical Analysis. We performed three main statistical analyses (**Figure 1**), including 1) using the Boruta method to select alcohol-associated gene transcripts, 2) using RF to examine the prediction capability of Boruta-selected transcripts for alcohol consumption categories, and 3) examining the cross-sectional associations of Boruta-selected transcripts with three CVD risk factors (obesity, hypertension, and type 2 diabetes). These analyses were performed by R studio (version 4.1.2).

Use Boruta algorithm for gene selection. RF method evaluates the importance of variables in the models by mean accuracy and Gini index.(11) However, the regular RF method does not provide cut-off values for these parameters for the purpose of variable selection. The Boruta algorithm extends the regular RF method by reporting the level of the predictors as “Confirmed”, “Tentative” and “Rejected”.(16, 27) We therefore used the Boruta method, implemented with the R *Boruta* package,(27) to facilitate automatic selection of alcohol-associated gene transcripts. In this analysis, alcohol consumption (g/day) was treated as outcome variable and gene transcripts were the main predictors, with sex and age as covariates. We used parameter doTrace = 2 to

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

obtain "confirmed" attributes, i.e., alcohol-associated gene transcripts. To achieve biological and statistical relevance of the transcripts determined by the Boruta algorithm, we applied two filtering methods, data-driven and pathway-based approaches, to choose transcripts to be tested. The first two sets were selected using the data-driven approach. The first set included 15,146 gene transcripts with absolute pair-wise Pearson $r < 0.6$ and the second set included 1,958 gene transcripts with false discovery rate (FDR) < 0.2 in the meta-analysis from our previous alcohol-gene transcript association analysis using conventional linear regression models.(9) The third to the fifth sets of gene transcripts were determined based on well-established gene pathway databases, including Wikipathways (n=6,890), Molecular Signatures Database (MSigDB) hallmark gene sets (H; 4,003 genes), and MSigDB immunologic signature gene sets (C7; 14,580 genes).(28-30) One at a time, we run Boruta models for these five sets of transcripts.

Gene ontology (GO) analysis. A web-based GO analysis (http://geneontology.org/) was performed to evaluate the biological process relevant to the Boruta method selected transcripts.(31) Fisher's exact tests were conducted using the default reference gene list. Similarly, GO term with FDR < 0.05 was considered statistically significant.

Exam prediction capability of selected gene transcripts. We used the RF models to examine whether the Boruta method-selected gene transcripts can distinguish different levels of alcohol consumption. Three comparisons were performed, including nondrinkers vs. moderate drinkers, nondrinkers vs. heavy drinkers, and moderate drinkers vs. heavy drinkers. The R *randomForest* package was used to perform these comparisons.(32) We randomly divided our study participants into a training set, which included 70% of the entire participants, and a testing set, which included 30% of the entire participants. The training data was used to train the RF model by default parameters: ntree (number of trees to grow) = 500 and mtry (number of variables randomly sampled as candidates at each split) = square root of number of attributes tested. The out-of-bag (OOB) error rate in the training set was used to determine the performance of the RF model, and the area under the receiver operating characteristic (ROC) curve (AUC) derived from the testing set was used to evaluate the prediction capability of the selected predictors.

Four sets of predictors were analyzed, including 1,958 transcripts with FDR < 0.2 in meta-analysis (set 1) and 25 alcohol-associated genes with significant Bonferroni-corrected *P* values (set 2) in our previous alcohol-gene transcript association analysis,(9) Boruta method-selected gene transcripts (set 3), and 144 alcohol consumption-associated CpGs (DNA methylation sites) identified from a previous epigenome-wide association analyses and meta-analysis (set 4).(23) We examined these four sets of predictors one at a time. In addition to these omics predictors, sex and age were covariates in all models. To determine the optimal threshold value for AUC calculation and avoid over- or under-sampling misclassification, we iterated each model ten times. The first iteration used default values. In the second iteration, using the *coords* function in R *pROC* package,(33) we calculated the maximum value of the sum of specificity and sensitivity using the Youden method based on the initial AUC calculation. This maximum value was used to derive the threshold for AUC calculation in this iteration. This process was repeated in the rest of iterations. We reported the AUC corresponding to the lowest OOB error rate after the initial iteration. Also, we compared the AUC calculated for the four different sets of predictors using the DeLong algorithm, implemented using the R *pROC* package. Code for Boruta method and AUC calculation using RF are in Supplemental materials.

Association analysis between the expression level of selected genes with CVD risk factors. We performed cross-sectional analyses between the Boruta method selected transcripts and obesity, hypertension, and type 2 diabetes. Covariates included age, sex, current smoking status, cohort (Offspring or Third Generation cohort), estimated blood cell compositions,(24) and BMI (only in analyses for hypertension and type 2 diabetes). Generalized estimation equations (GEE) were used to account for familial relationships. Bonferroni correction (i.e., 0.05 divided by the number of transcripts selected times three CVD risk factors) was applied to determine statistical significance.

Interaction analyses and stratification analyses. We examined potential interaction between alcohol consumption and sex and age (in continuous scale) in relation to gene expression for transcripts identified by the Boruta method. Linear mixed regression was performed accounting for family structure in FHS. A product term of alcohol consumption and sex or alcohol consumption and age were added in models. Covariates included sex, age, current smoking status, the FHS cohort index (Offspring versus Third Generation) and blood cell counts (counts of white cell, red cell, and platelet and proportion of neutrophils, lymphocytes, monocytes,

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

basophils and eosinophils).(9) We also performed interaction analysis between transcripts selected by the Boruta method and sex and age in relation to the three CVD risk factors. In these analyses, we used the same GEE modelling described above in the main effect analysis to test the statistical significance of the product term of transcripts and sex or age. Further, we stratified our study participants by sex and age (below or above median age 55 years) and reran the association analysis between transcripts and CVD risk factors in each stratum.

Results

Study Participants. About 54.3% participants were women and the average age of the participants was 55.4 (**Table 1)**, We classified the participants into three categories based on alcohol consumption levels: nondrinkers, moderate drinkers, and heavy drinkers. Nondrinkers tended to be older in age, followed by heavy drinkers and moderate drinkers. Men tended to drink more alcohol compared to women. More heavy drinkers were current smokers (19%) compared to nondrinkers (9%) and moderate drinkers (7%). The proportion of participants with obesity and type 2 diabetes was higher in nondrinkers (38% and 16%, respectively), while the proportion of participants with hypertension was higher in heavy drinkers (54%).

Use Boruta algorithm for gene selection. The Boruta method selected 6 gene transcripts (*SORT1*, *ODC1*, *CTSG*, *IL4R*, *MPO*, and *CYTH1*) from the Wikipathways set, 10 transcripts (*IFI44L*, *P2RY14*, *PLAGL1*, *DOCK4*, *GAPVD1*, *IFITM1*, *UTP20*, *MPO*, *ATP5F1D*, and *RBM38*) from the MSigDB hallmark pathway set, and 11 transcripts (*FCGR1A*, *IFI6*, *ABCA13*, *DOCK4*, *LCN2*, *DDX58*, *OLFM4*, *CTSG*, *MPO*, *CEACAM8*, and *BPI*) from the MSigDB immunologic signature sets (**Table 2**). Among transcripts that were associated with alcohol consumption at FDR < 0.2 in our previous analysis using linear regression models,(23) the Boruta method selected 4 transcripts (*OLFM4*, *CTSG*, *MPO*, and *CEACAM8*). From those with absolute pairwise $r < 0.6$, the Boruta method selected 3 transcripts (*SORT1*, *DOCK4*, and *TNFSF13B*). After removing duplicated transcripts (**Table 2**), we found 25 alcohol-associated transcripts using the Boruta method. We compared the differences of gene expression levels in moderate and heavy drinkers relative to nondrinkers (**Supplemental Figure 1)**. We found no substantial evidence supporting nonlinear relationships between alcohol consumption and these 25

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

transcripts. Also, we found no significant statistical interaction between the 25 transcripts and sex and age at $P < 0.002$ (Bonferroni correction for 25 transcripts; **Supplemental Table 7**).

Among these 25 Boruta method selected transcripts, 12 transcripts, (*MEIS1*, *ODC1*, *ABCA13*, *OLFM4*, *CTSG*, *CEACAM8*, *LCN2*, *UTP20*, *DOCK4*, *IL4R*, *MPO*, and *BPI*) had $P <$ 2.9e-6 (Bonferroni correction for 17,176 genes) in our previous meta-analysis based on linear regression models.(9) In these 12 transcripts, six (*MEIS1*, *ODC1*, *ABCA13*, *OLFM4*, *CTSG*, and *CEACAM8*; **Table 2**) were also among those (n=25) significant using discovery and replication strategy ($P <$ 8e-4 in the discovery analysis and $P <$ 1.9e-4 in the replication analysis).(9) The correlation between the 13 unique transcripts identified by the Boruta method and those identified by the conventional linear models (either using discovery and replication or meta-analysis; n=101) was largely modest, 97% pairs with Pearson $|r| < 0.3$ (**Supplemental Figure 3**). The pairwise correlation of the 25 Boruta method selected transcripts ranged from 0 to 0.84 (Pearson $|r|$) (**Supplemental Figure 2**). There were 240 pairs of transcripts with $|r| < 0.3$, 38 pairs of with $|r|$ between 0.3 and 0.6, and 22 pairs with $|r| > 0.6$. In these 22 pairs with $|r| > 0.6$, there were three clusters of transcripts (**Supplemental Figure 2**), including (1) *IFI6*, *DDX58*, and *IFITM1*, (2) *MPO*, *CTSG*; *LCN2*, *BPI*, *CEACAM8*, *ABCA13*, and *OLFM4*, and (3) *ODC1* and *RBM38.*

Gene ontology (GO) analysis. We found that the 25 Boruta method selected transcripts were enriched in 10 GO biological processes (**Supplemental Table 1**). The ancestor charts of these significant GO terms were shown in **Supplemental Figure 4**. These significant GO terms are primarily for defense response to bacterium (GO:0042742; P = 2.9e-5; FDR = 0.04) and immune response (GO:0006955; P = 4.4e-6; FDR = 0.009). We observed that several transcripts with $|r|$ > 0.6 were among the enriched genes, e.g., *IFI6* and *DDX58* from the first cluster (**Supplemental Figure 2**).

Exam prediction capability of selected gene transcripts. In **Figure 2**, we showed the ROC curves for the four sets of predictors derived from the present analysis and our previous studies, including 1,985 transcripts with FDR < 0.2 based on conventional regression,(9) 25 transcripts using discovery and replication strategy based on conventional regression,(9) the 25 Boruta method selected transcripts, and 144 alcohol associated CpGs.(23) In addition, we integrated

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

predictors from the latter three sets to test whether additively combining transcripts and CpGs might improve prediction. We calculated the AUC based on the lowest OOB error rate and the largest AUC from the 10 iterations (**Supplemental Table 2**). For all predictors, the AUC based on the lowest OOB error rate was slightly better in the analyses for nondrinkers vs. heavy drinkers (0.72 to 0.77) compared to that for nondrinkers vs. moderate drinkers (0.66 to 0.70) and moderate drinkers vs. heavy drinkers (0.65 to 0.70). In analysis to compare nondrinkers and heavy drinkers, the AUC of the 25 Boruta method selected transcripts was comparable (0.73) to that based on the conventional linear regression (0.74 for the 1,985 transcripts and 0.72 for the 25 transcripts) and lower than that using the 144 CpGs (0.77). We found the combining-predictors approach had a slightly better AUC than transcripts-based approaches and similar as that for CpGs. However, no significant statistical difference was detected between the 25 Boruta method selected transcripts and other sets of predictors using Delong tests in the above comparisons (**Supplemental Table 3**). The AUC from analyses based on light and as-risk drinkers was not substantially different from that in the primary analyses combining light and at-risk drinkers (**Supplemental Table 4**).

Cross-sectional association with CVD risk factors. With Bonferroni correction for the 25 Boruta selected transcripts and three CVD risk factors (i.e., at P < 6.7e-4), we observed that 13 transcripts were associated with obesity, 1 transcript with hypertension, and 3 transcripts with type 2 diabetes (**Table 3**). In analysis for hypertension defined as SBP > 130 mm Hg or DBP > 80mm Hg, the association was largely consistent. Nonetheless, two transcripts, *RBM38* (*P=1.7-4*) and *DOCK4* (*P=1.7e-4*), remained significant at P < 6.7e-4. Thus, taken together, 19 transcript-CVD risk factor pairs were observed. Among these 19 pairs, 5 pairs have been reported in our previous study,(9) and the other 14 pairs were unique in the present study (**Table 3; Supplemental Table 5**). In the FHS, we have observed that alcohol consumption was inversely associated with the risk of obesity and type 2 diabetes and positively associated with the risk of hypertension.(25) Therefore, if a transcript is positively associated with alcohol consumption, we expect that this transcript is inversely associated with obesity and diabetes and positively associated with hypertension, or vice versa. For the 14 novel pairs, the direction of the associations for four transcript-obesity pairs and one transcript-hypertension pair were consistent with our hypothesis. The association between alcohol consumption and these five transcripts

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

were shown in **Supplemental Table 6**. For example, alcohol consumption was inversely associated with the expression of *DOCK4*, *IL4R*, and *SORT1*, regression coefficients were -0.017 (95% CI: -0.024, -0.011; *P* = 1.8e-7), -0.016 (95% CI: -0.021, -0.011; *P* = 1.3e-10) and -0.007 (95% CI: -0.011, -0.003; *P* =0.0003) per 10 g/day higher alcohol consumption, respectively. Consistently, *DOCK4* and *SORT1* were positively associated with obesity and *IL4R* was inversely associated with hypertension (**Table 3**).

We found no significant interaction between the 25 transcripts and age (**Supplemental Table 8)**. We observed significant interaction between sex and three transcripts, including *DOCK4* (*P*=5.5e-5), *RBM38* (*P*=2.9e-4) and *MPO* (*P*=2.9e-5), in relation to obesity. Stratified analyses by sex and age are presented in **Supplemental Table 9-12**. For all the three transcripts, their association with obesity was in the same direction in both sex; however, the association strength varied in male and female participants. In male participants, the odds ratio (OR) for obesity was 1.30 (95%CI=1.03, 1.64; *P*=0.03) for *DOCK4*, 1.66 (95%CI=1.38, 2.00; *P*=7.9e-8) for *RBM38*, and 1.46 (95%CI=1.09, 1.96; *P*=0.01) for *MPO*. Whereas, in female participants, the OR was 2.48 (95%CI=1.98, 3.11; *P*=2.0e-15) for *DOCK4*, 2.65 (95%CI=2.17, 3.23; *P*=7.9e-22) for *RBM38*, 0.65 (95%CI=0.42, 1.00; *P*=0.05) for *MPO*.

Discussion

In the present analysis, we used the Boruta method and demonstrated that 25 gene transcripts were associated with alcohol consumption in FHS participants. Compared to our previous study based on conventional linear regression analysis, the present study identified 13 additional alcohol-associated transcripts. Several of the 13 transcripts such as *FCGR1A* and *SORT1* were further linked to CVD risk factors. We also showed that the Boruta method selected transcripts have comparable prediction capabilities as the transcripts identified by conventional linear regression analysis in the testing set (30% of entire study participants). Taken together, the present analysis suggests that the Boruta method can contribute to a better understanding of alcohol-associated transcriptomic changes. Taken together, the present analysis expanded the candidate list of gene transcripts for future validation studies, suggesting that the Boruta method can contribute to a better understanding of alcohol-associated transcriptomic changes.

RF is a commonly performed supervised machine learning method for transcriptomic data.(34) The RF-based Boruta method has been used in studies analyzing both array- and RNA-

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

sequencing-based transcriptomic data.(34-36) We used the Boruta method because of its stable feature selection capability relative to other approaches, e.g., a study reported that the Boruta method could identify important genes and achieved the highest ratio of self-consistent selections.(17) However, a recent study compared three feature selection algorithms, Boruta, Vita, and AUC-RF, and showed that the three approaches had a comparable performance regarding identification of transcriptomic signatures predicting colorectal cancer.(37) A recent study also compared several machine learning methods and showed the LASSO method identified more transcripts predicting asthma than the Boruta method.(38) It is difficult to directly compare these studies because of different study designs, data distribution, and phenotypes. Future studies to compare multiple machine learning methods are needed to explore at what conditions a certain method can perform better.

Because of the high dimensionality of the transcriptomic data, we applied two filtering methods, data-driven and pathway-based approaches before running the Boruta algorithm. Overall, the pathway-based approach performed better than the data-driven approach because the former identified more transcripts. This suggests that embedding biological knowledge may lead to a better performance of the Boruta method. To the best of our knowledge, machine learning approaches (such as RF with Boruta method) have not been extensively examined to study alcohol consumption related transcriptomic changes. The present study contributes novel information to the current literature; however, future studies are needed to establish a critical process for using machine learning methods in this research area, such as performing data harmonization and transformation, selecting appropriate machine learning methods, and conducting external validation.

In our previous study using conventional linear regression models,(9) we reported significant associations between 22 alcohol-associated transcripts and three CVD risk factors. The present study also showed several additional transcript-CVD risk factor pairs, particularly five pairs (for five transcripts; **Supplemental Table 6**) were in line with our previous observations on alcohol consumption and CVD risk factors.(25) Three of the five transcripts (*FCGR1A*, *IFITM1*, and *SORT1*) are among the 13 unique transcripts identified by the Boruta method. The three transcripts had low to moderate correlation with those identified by our previous study using conventional regression models.(9) GO analysis showed that *FCGR1A* (Fc gamma receptor Ia) and *IFITM1* (interferon induced transmembrane protein 1) were enriched in

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

nine GO terms related to defense or immune response (**Supplemental Table 1**), suggesting that alcohol consumption may trigger chronic inflammation and then affect CVD risk. A genetic variant (rs4970843-C) at intron of *SORT1* (sortilin 1) was associated with height,(39) which is consistent with the present observation on the *SORT1* and obesity (i.e., increased BMI). However, a study in the Danish PRISME study showed that heavy alcohol drinking was associated with an increased sortilin, which is opposite to the present observation on a negative association of alcohol consumption with *SORT1* expression levels (**Supplemental Table 6**). This may be due to most of our study participants (93%) are nondrinkers and moderate drinkers. Nonetheless, because of the cross-sectional and observational nature of the present analysis, we cannot infer causality. Future studies with large sample size and in diverse populations are warranted to validate the present findings.

In approximately 30% of our study participants (i.e., the testing set), we tested the prediction capabilities of the 25 Boruta method selected transcripts. Compared to the transcripts identified by conventional regression models, the 25 Boruta method selected transcripts had a comparable prediction capability. Although no statistical significance was detected, the overall prediction capabilities of selected gene transcripts were relatively weaker than DNA methylation markers (AUC 0.72 vs. 0.77). These DNA methylation markers were selected based on a large meta-analysis in 13 population-based cohorts;(23) therefore, this set of DNA methylation markers may be less noisy than the gene transcripts. The analysis combining gene transcripts and DNA methylation markers did not substantially increase the AUC, which also suggests that DNA methylation markers may have better prediction capabilities. However, the additive approach that was used to combine selected gene transcripts and CpGs may be biased because the potential interaction between different types of omics markers is not considered.(40) Thus, novel analytical approaches to integrating multiple omics markers are needed to comprehensively identify alcohol-associated markers. In addition, compared to array-based transcriptomic data, RNA sequencing (RNA-seq) has a better resolution and enables the identification of non-coding RNAs. Future studies utilizing RNA-seq data are needed to examine the alcohol-associated transcriptomic changes.

The advantages of the present study include using a well-established machine learning method and comprehensive data (alcohol consumption, transcriptomics, and clinical risk factors) collected from the well-characterized community based FHS. However, in addition to several weaknesses described above, other limitations warrant discussion. First, all study participants were Europeans, and most study participants were nondrinkers or moderate drinkers. This limits the generalizability of the present study to other more diverse populations. Second, interpretation of the transcripts selected by machine learning approaches is challenging. We explored their cross-sectional association with CVD risk factors. However, transcriptomic profiles may change over time. Prospective association analyses are therefore needed to provide more robust data regarding the relationship between alcohol, gene expression, and CVD risk factors. Third, different types of alcoholic beverages may have different responses in gene expression levels. Future studies with larger sample size are needed to examine specific transcriptomic characteristics associated with consumption of each type of alcoholic beverage. Fourth, questionnaires were used to collect self-reported alcohol consumption. Measurement errors may exist and affect transcript selection and prediction accuracy. Nonetheless, this also highlights the needs for future studies to comprehensively investigate surrogate markers for alcohol consumption.

The association of alcohol consumption and cardiovascular health is complex, mainly due to the uncertainty related to the potential impact of moderate alcohol drinking on cardiovascular health.(3-5) Majority of study participants are nondrinkers or moderate drinkers. Our previous study using conventional regression models did not find a clear protective effect of alcohol consumption on CVD risk factors through transcriptomic biomarkers. In the present study, we used a different analytical approach, yet the findings echo those from our previous study.(9) It should be noted that the present analysis only examined one commonly used machine learning algorithm. Other machine learning and deep learning algorithms,(41) together with profound bioinformatic knowledge, may facilitate the identification of true causal transcriptomic markers and improve the discrimination capacities of alcohol-associated transcriptomic biomarkers.

In conclusion, we applied a supervised machine learning approach, the RF-based Boruta method, and identified additional alcohol-associated gene transcripts, compared to analysis using the conventional linear regression models. These additional transcripts expand the candidate list for future validation studies; thus, our findings support the notion that machine learning approaches can contribute useful information to unravel the complex relationship between alcohol consumption and CVD risk. Our findings support the notion that machine learning approaches can contribute useful information to unraveling the complex relationship between

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

alcohol consumption and CVD risk. The present study also highlights that future studies in large and diverse samples are needed to comprehensively investigate the impact of alcohol consumption on transcriptomic changes and subsequent disease burden.

Acknowledgements: The Framingham Heart Study was supported by NIH contracts N01-HC-25195, HHSN268201500001I, and 75N92019D00031. Funding for SABRe gene expression was provided by Division of Intramural Research, NHLBI, and Center for Population Studies, NHLBI.

Financial Support: JM and CLiu are supported by NIH grant R01AA028263.

Conflict of Interest: The authors declare no conflicts of interest.

Authorship: The authors' contributions were as follows— JM and CLiu designed research and had primary responsibility for final content; CLyu conducted the analyses; JM, CLyu and CLiu interpreted the result; RJ conducted quality control and residual calculation for gene expression data; CLyu and JM wrote the manuscript; RJ, TH, DL, and CLiu critically reviewed the manuscript; and all authors read and approved the final manuscript.

Disclaimer: The views and opinions expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute, the National Institutes of Health, or the U.S. Department of Health and Human Services.

Data availability: The datasets analyzed in the present study are available at the dbGAP repository phs000007.v32.p13 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v30.p11).

Abbreviations: ABCA13: ATP binding cassette subfamily A member 13; ATP5F1D: ATP synthase F1 subunit delta; AUC: area under the receiver operating characteristics curve; BMI: body mass index; BPI: bactericidal permeability increasing protein; CEACAM8: CEA cell adhesion molecule 8; CpG: DNA methylation sites; CTSG: cathepsin G; CVD: cardiovascular

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

disease; CYTH1: cytohesin 1; DBP: diastolic blood pressure; DOCK4: dedicator of cytokinesis 4; FCGR1A: Fc gamma receptor Ia; FDR: false discovery rate; FHS: Framingham Heart Study; GAPVD1: GTPase activating protein and VPS9 domains 1; GEE: Generalized estimation equations; GO: Gene ontology; IFI44L: interferon induced protein 44 like; IFI6: interferon alpha inducible protein 6; IFITM1: interferon induced transmembrane protein 1; IL4R: interleukin 4 receptor; LCN2: lipocalin 2; MEIS1: Meis homeobox 1; MPO: myeloperoxidase; MSigDB: Molecular Signatures Database; ODC1: ornithine decarboxylase 1; OLFM4: olfactomedin 4; OOB: out-of-bag; P2RY14: purinergic receptor P2Y14; PLAGL1: PLAG1 like zinc finger 1; RBM38: RNA binding motif protein 38; RF: Random Forest; RIGI (DDX58): RNA sensor RIG-I; ROC: receiver operating characteristic; SBP: systolic blood pressure; SORT1: sortilin 1; TNFSF13B: TNF superfamily member 13b; UTP20: UTP20 small subunit processome component.

Reference

1. Emanuele NV, Swade TF, Emanuele MA. Consequences of alcohol use in diabetics. Alcohol Health Res World. 1998;22(3):211-9.
2. Chait A, Mancini M, February AW, et al. Clinical and metabolic study of alcoholic hyperlipidaemia. Lancet. 1972;2(7767):62-4.
3. Collaborators GBDA. Alcohol use and burden for 195 countries and territories, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2018;392(10152):1015-35.
4. Chikritzhs TN, Naimi TS, Stockwell TR, et al.Mendelian randomisation meta-analysis sheds doubt on protective associations between 'moderate' alcohol consumption and coronary heart disease. Evid Based Med. 2015;20(1):38.
5. Stockwell T, Zhao J, Panwar S, et al. Do "Moderate" Drinkers Have Reduced Mortality Risk? A Systematic Review and Meta-Analysis of Alcohol Consumption and All-Cause Mortality. J Stud Alcohol Drugs. 2016;77(2):185-98.
6. Huan T, Esko T, Peters MJ, et al. A meta-analysis of gene expression signatures of blood pressure and hypertension. PLoS Genet. 2015;11(3):e1005035.

7. Yao C, Chen BH, Joehanes R, et al. Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. Circulation. 2015;131(6):536-49.

8. Benton MC, Lea RA, Macartney-Coxson D, et al. Mapping eQTLs in the Norfolk Island genetic isolate identifies candidate genes for CVD risk traits. Am J Hum Genet. 2013;93(6):1087-99.

9. Ma J, Huang A, Yan K, et al. Blood transcriptomic biomarkers of alcohol consumption and cardiovascular disease risk factors: the Framingham Heart Study. Hum Mol Genet. 2023;32(4):649-58.

10. Luo J, Wu M, Gopukumar D, et al. Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomed Inform Insights. 2016;8:1-10.

11. Breiman L. Random Forests. Machine Learning. 2001;45.

12. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Brief Bioinform. 2023;24(2).

13. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform. 2019;20(2):492-503.

14. Cammarota C, Pinto A. Variable selection and importance in presence of high collinearity: an application to the prediction of lean body mass from multi-frequency bioelectrical impedance. J Appl Stat. 2021;48(9):1644-58.

15. Swan AL, Mobasheri A, Allaway D, et al. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. OMICS. 2013;17(12):595-610.

16. Kursa M, Jankowski A, Rudnicki W. Boruta - A System for Feature Selection. Fundam Inform. 2010;101:271-85.

17. Acharjee A, Larkman J, Xu Y, et al. A random forest based biomarker discovery and power analysis framework for diagnostics research. BMC Med Genomics. 2020;13(1):178.

18. Liu C, Ackerman HH, Carulli JP. A genome-wide screen of gene-gene interactions for rheumatoid arthritis susceptibility. Hum Genet. 2011;129(5):473-85.

19. Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. Biom J. 2014;56(4):601-6.

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

20. Polewko-Klim A, Lesinski W, Golinska AK, et al. Sensitivity analysis based on the random forest machine learning algorithm identifies candidate genes for regulation of innate and adaptive immune response of chicken. Poult Sci. 2020;99(12):6341-54.

21. Feinleib M, Kannel WB, Garrison RJ, et al. The Framingham Offspring Study. Design and preliminary data. Prev Med. 1975;4(4):518-25.

22. Splansky GL, Corey D, Yang Q, et al. The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. Am J Epidemiol. 2007;165(11):1328-35.

23. Liu C, Marioni RE, Hedman AK, et al. A DNA methylation biomarker of alcohol consumption. Mol Psychiatry. 2018;23(2):422-33.

24. Joehanes R, Ying S, Huan T, et al. Gene expression signatures of coronary heart disease. Arterioscler Thromb Vasc Biol. 2013;33(6):1418-26.

25. Sun X, Ho JE, Gao H, et al. Associations of Alcohol Consumption with Cardiovascular Disease-Related Proteomic Biomarkers: The Framingham Heart Study. J Nutr. 2021;151(9):2574-82.

26. Czuriga-Kovacs KR, Czuriga D, Kardos L, et al. Reply to letter: Reversibility of hypertension-induced subclinical vascular changes: Do the new ACC/AHA 2017 blood pressure guidelines and heart rate changes make a difference? J Clin Hypertens (Greenwich). 2019;21(8):1243-4.

27. Kursa M, Rudnicki W. Feature Selection with the Boruta Package. Journal of Statistical Software. 2010;36(11):13.

28. Martens M, Ammar A, Riutta A, et al. WikiPathways: connecting communities. Nucleic Acids Res. 2021;49(D1):D613-D21.

29. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34(3):267-73.

30. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50.

31. Thomas PD, Ebert D, Muruganujan A, et al. PANTHER: Making genome-scale phylogenetics accessible to all. Protein Sci. 2022;31(1):8-22.

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

32. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3).

33. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.

34. Kursa MB. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics. 2014;15:8.

35. Shen J, Qi L, Zou Z, et al. Identification of a novel gene signature for the prediction of recurrence in HCC patients by machine learning of genome-wide databases. Sci Rep. 2020;10(1):4435.

36. Lin MS, Jo SY, Luebeck J, et al. Transcriptional immune suppression and upregulation of double stranded DNA damage and repair repertoires in ecDNA-containing tumors. bioRxiv. 2023.

37. Long NP, Park S, Anh NH, et al. High-Throughput Omics and Statistical Learning Integration for the Discovery and Validation of Novel Diagnostic Signatures in Colorectal Cancer. Int J Mol Sci. 2019;20(2).

38. Dessie EY, Gautam Y, Ding L, et al. Development and validation of asthma risk prediction models using co-expression gene modules and machine learning methods. Sci Rep. 2023;13(1):11279.

39. Yengo L, Vedantam S, Marouli E, et al. A saturated map of common genetic variants associated with human height. Nature. 2022;610(7933):704-12.

40. Singh A, Shannon CP, Gautier B, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019;35(17):3055-62.

41. Wekesa JS, Kimwele M. A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. Front Genet. 2023;14:1199087.

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

Table 1. Participant characteristics

	Total (n=5508)	Nondrinkers (n=1729)	Moderate drinkers (n=3427)	Heavy drinkers (n=352)
Age	55.4 ± 13.1	60.8 ± 12.6	52.5 ±12.6	56.6 ± 11.5
Men	2516 (45.7%)	676 (39.1%)	1653 (48.2%)	187 (53.1%)
Obesity	1707 (31.0%)	657 (37.9%)	956 (27.9%)	94 (26.7%)
Hypertension	2112 (38.3%)	842 (48.7%)	1080 (31.5%)	190 (53.0%)
Type 2 Diabetes	487 (8.8%)	272 (15.7%)	193 (5.6%)	22 (6.3%)
Alcohol consumption(g/d)	4.7 (15)	0 (0)	9.1 (12.2)	51.3 (27.6)

Values are mean ± SD or n (%); alcohol consumption is presented as median (IQR)

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

Table 2. Boruta algorithm selected genes

					Gene sets				
Gene	**Chr**	**Start**	**Stop**	***P***	**Wikipathways**	**MSigDB hallmark**	**MSigDB immunologic signature**	**Association FDR <0.2**	**Pairwise Pearson r <0.6**
IFI44L	1	79086136	79108668	9.6e-1		√			
FCGR1A	1	149718521	149765367	5.8e-2			√		
IFI6	1	27992587	28359029	5.3e-1			√		
SORT1	1	109850942	109940573	3.4e-4	√				
MEIS1	2	66653313	66800441	7.2e-13					×
ODC1	2	10568023	10688889	8.4e-11	×				
P2RY14	3	150929912	150996391	2.3e-4		√			
PLAGL1	6	144261449	144385677	3.9e-6		√			
ABCA13	7	48237836	48700550	5.7e-10			×		
DOCK4	7	111365666	111846508	1.8e-7		√	√		√
GAPVD1	9	128022911	128191972	7.2e-1		√			
LCN2	9	130893682	130915718	4.2e-9			√		
DDX58	9	32455306	32732887	4.7e-1			√		
IFITM1	11	310891	315260	2.4e-3		√			
UTP20	12	101640624	101780384	2.2e-7		√			
OLFM4	13	53584428	53708870	5.1e-13			×	×	
TNFSF13B	13	108897127	108960825	5.0e-6					√
CTSG	14	25042724	25045559	8.1e-16	×		×	×	
IL4R	16	27325194	27385797	1.3e-10	√				
MPO	17	56347222	56358430	6.1e-11	√	√	√	√	
CYTH1	17	76670136	76778378	4.2e-2	√				
ATP5F1D	19	1239851	1244813	1.3e-1		√			
CEACAM8	19	43084395	43224500	3.6e-9			×	×	
BPI	20	36932545	36965907	5.0e-10			√		
RBM38	20	55966449	55984369	6.4e-5		√			

×: transcripts have been identified using conventional linear regression models (reference 9)
P values are from meta-analysis in reference 9.
Transcription start and stop position are based on GRCh37.

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

Table 3. Cross-sectional analysis of Bruta method selected genes with CVD risk factors

	Obesity			Hypertension			Type 2 diabetes		
Gene	OR	95% CI	*P*	OR	95% CI	*P*	OR	95% CI	*P*
FCGR1A	1.54	1.38, 1.72	3.0e-14						
SORT1	2.65	2.04, 3.45	3.7e-13						
ODC1	2.04	1.72, 2.41	2.2e-16				1.80	1.32, 2.44	1.6e-4
ABCA13	2.29	1.73, 3.01	4.5e-9						
DOCK4	1.84	1.56, 2.16	2.0e-13						
GAPVD1	3.02	2.31, 3.93	3.3e-16						
LCN2	1.71	1.54, 1.90	6.7e-24				1.32	1.14, 1.54	3.5e-4
IFITM1	1.32	1.13, 1.54	5.3e-4						
UTP20	2.24	1.58, 3.18	5.9e-6						
OLFM4	1.51	1.33, 1.71	1.7e-10						
IL4R				0.49	0.38, 0.62	3.3e-9			
CEACAM8	1.58	1.42, 1.76	2.1e-17						
BPI	1.31	1.17, 1.47	4.4e-6						
RBM38	2.05	1.79, 2.35	6.7e-25				1.72	1.35, 2.21	1.5e-5

Generalized estimation equations with adjustment for age, sex, current smoking status, FHS cohorts (the Offspring or Third Generation cohort), estimated blood cell compositions, and BMI (only in analyses for hypertension and type 2 diabetes)

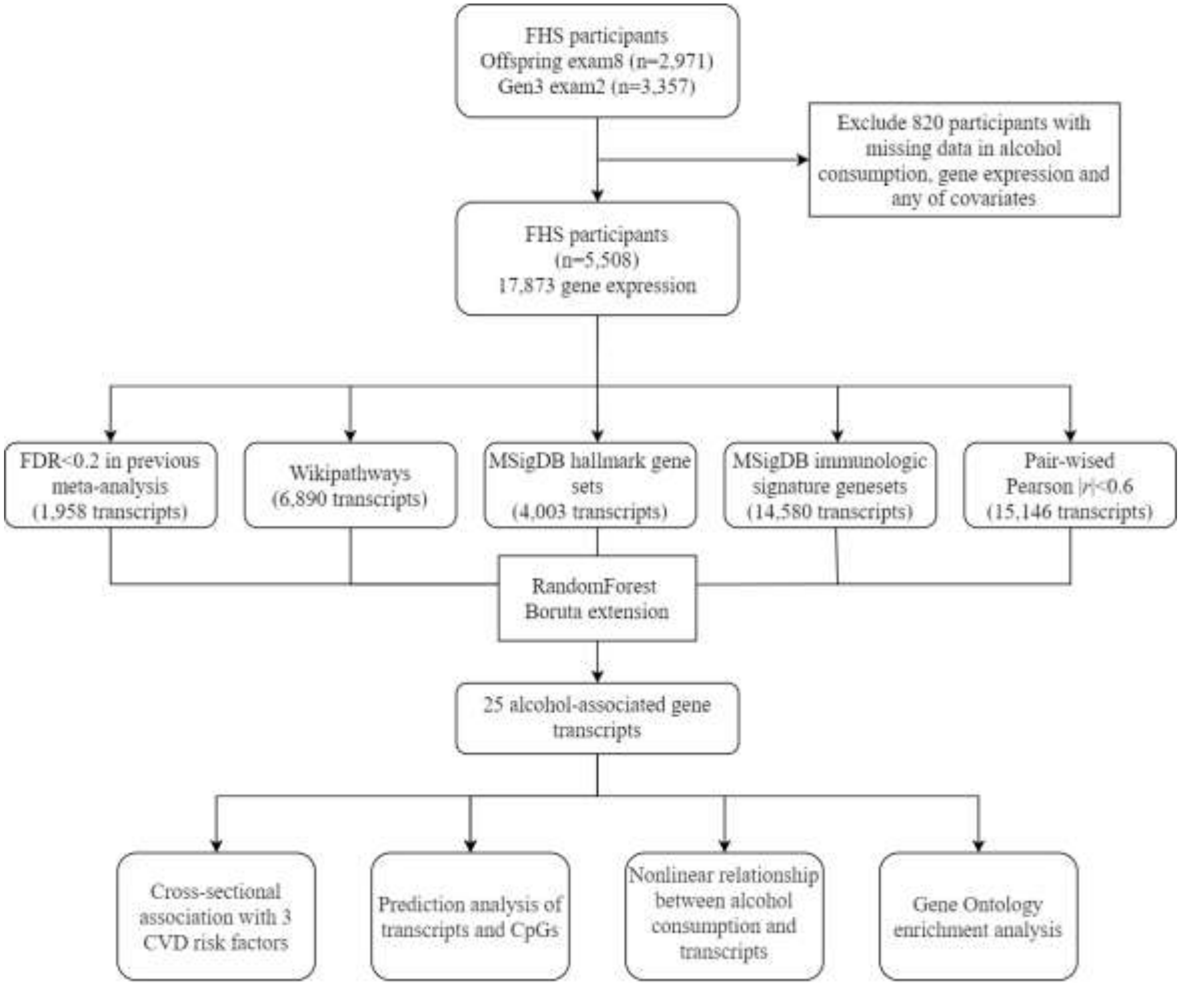
FHS participants
Offspring exam8 (n=2,971)
Gen3 exam2 (n=3,357)
Exclude 820 participants with missing data in alcohol consumption, gene expression and any of covariates
FHS participants
(n=5,508)
17,873 gene expression
FDR<0.2 in previous meta-analysis
(1,958 transcripts)
Wikipathways
(6,890 transcripts)
MSigDB hallmark gene sets
(4,003 transcripts)
MSigDB immunologic signature genesets
(14,580 transcripts)
Pair-wised Pearson |r|<0.6
(15,146 transcripts)
RandomForest
Boruta extension
25 alcohol-associated gene transcripts
Cross-sectional association with 3 CVD risk factors
Prediction analysis of transcripts and CpGs
Nonlinear relationship between alcohol consumption and transcripts
Gene Ontology enrichment analysis

Figure 1. Study Flow Chart

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press

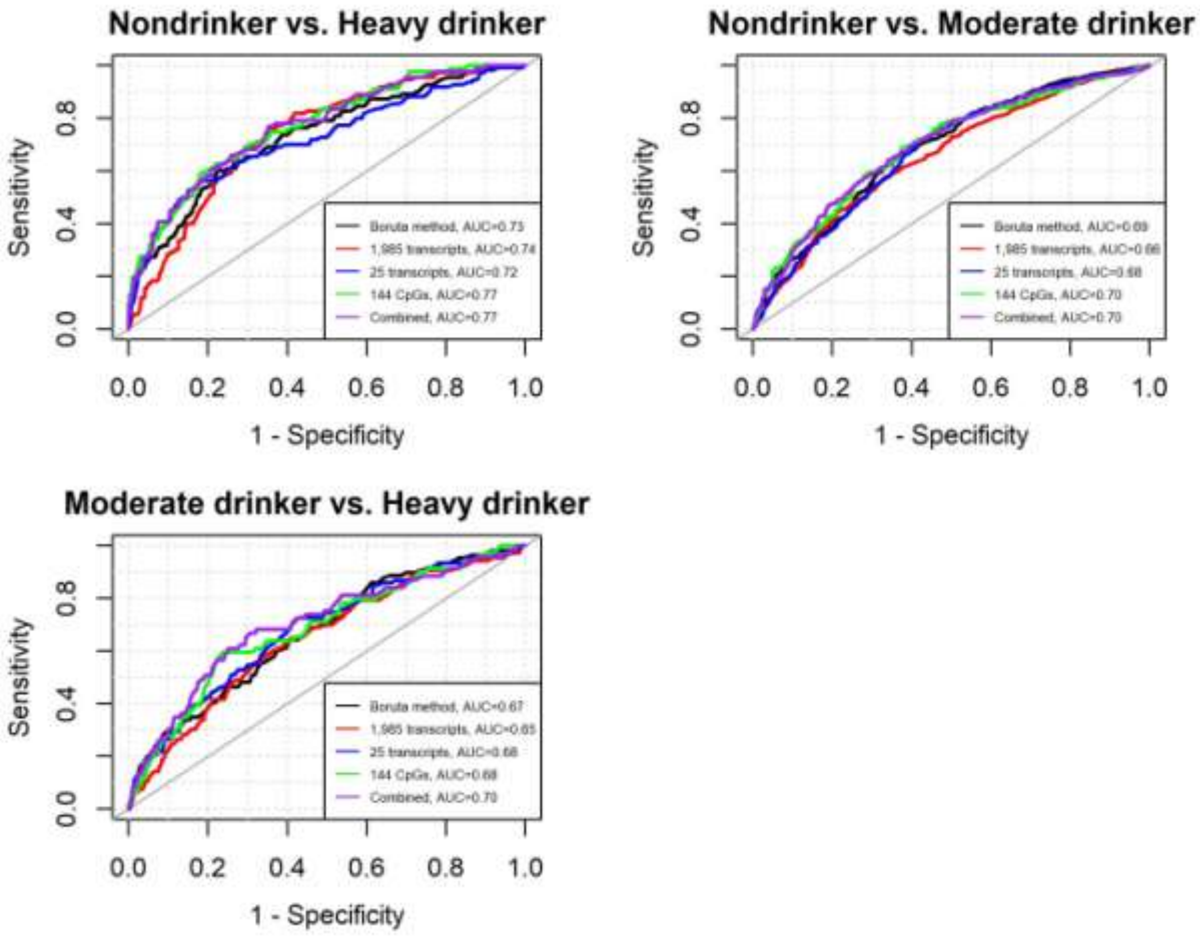

Figure 2. ROC of selected predictors. 1) Boruta method was based on the 25 Boruta method selected transcripts; 2) 1,985 transcripts and 3) 25 transcripts were from alcohol-gene expression analyses using conventional linear regression (reference 9); 4) 144 CpGs was from meta-analysis of alcohol associated DNA methylation markers (reference 21); 5) Combined predictors from sets 1, 3, and 4.

https://doi.org/10.1017/S0007114524000795 Published online by Cambridge University Press