Epigenetic prediction of smoking status using machine-learning methods

Tianxiao Liu[1], Yunfeng Huang[1], Qin Hui[1] and Yan V Sun[1]

[1] Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA 30322, USA

E-mail: tianxiao.liu@emory.edu

Abstract

Background: Tobacco smoking has been recognized as a major risk factor for many adverse health outcomes. Although many DNA methylation sites have been reported to be associated with tobacco smoking, few studies have focused on establishing prediction models of smoking status from DNA methylation data. This study aims at smoking status prediction using machine learning algorithms with precision, generalizability and a small number of predictors. **Methods:** An epigenetic prediction analysis of smoking status was performed on 218 male Caucasian twins, using DNA methylation data and two machine learning methods, random forests and elastic net. Training and testing of the prediction models were performed in two non-overlapping subsets. **Results:** Accuracy of the prediction model is higher in differentiating current and non-current smokers, than that in differentiating past and never smokers. In predicting past and never smokers, elastic net has a higher accuracy for smaller predictor sets compared with random forests. After variable tuning and predictor selection, the performance of random forests in predicting past and never smokers increases for all predictor sets. **Conclusion:** This study suggested that machine learning approaches could be utilized in understanding smoking risks using DNA methylation data with a relatively small set of DNA methylation data.

Keywords: DNA methylation, machine learning, random forests, elastic net, smoking status, predictor selection

1. Background

Tobacco smoking is one of the major preventable causes of premature death. It is associated with fatal diseases such as coronary heart diseases, stroke, lung cancer, chronic obstructive pulmonary disease, miscarriage and underdevelopment of foetus

(1). Despite the decreasing prevalence in recent years with the internationalization of tobacco control, it remains the top risk factor for many disease outcomes (2).

DNA methylation (DNAm), a major form of epigenetic modifications, is strongly associated with tobacco smoking and smoking-related diseases. Since the first epigenome-wide association study of cigarette smoking (3), thousands of smoking-related DNAm sites have been discovered using earlier Infinium Human Methylation Beadchip (27 and 450 K) assays, such as in Zeilinger et al (4) and recent Methylation EPIC BeadChip (850K), such as in Barcelona et al (5). Although the strong association of tobacco smoking and DNAm has been well documented, few studies have quantified these observations to computational models that could predict smoking status. A recent study by Bollepalli et al. used lasso regression to build smoking status prediction models using DNAm profiles and applied the results to three independent whole-blood samples to demonstrate its robustness and global applicability (6). The authors developed an R package EpiSmokEr (6), and identified 121 DNAm sites that contribute to the prediction model of smoking status.

In this study, we aimed to improve the precision and generalizability of predicting smoking status and to reduce the number of DNAm sites in the prediction model using two established machine learning methods – random forests and elastic net regularization. Random forests method takes root from decision trees with robustness in dealing with the problem of input variable noise, overfitting and a vast improvement in prediction accuracy (7). Elastic net regularization combines the strengths of lasso regression and ridge regression to produce a prediction model with small sample size and a large number of predictors. Elastic net can deal with correlated predictors such as corrected DNAm sites in epigenetic prediction.

2. Methods

2.1 Study population

The DNA methylomic data and smoking status were obtained from samples of the Emory Twin Study (ETS). The ETS consists of 307 middle-aged male Caucasian monozygotic and dizygotic twin pairs from the Vietnam Era Twin Registry (8) born between 1946 and 1956 (9, 10). All participants were examined at the Emory University General Clinical Research Center between 2002 and 2010. A subset of 218 ETS participants was epityped and used in subsequent prediction analysis. The analysis dataset consists of 108 twin pairs and 2 singletons, which were randomly split into two sets with each set containing only one individual from a twin pair. The two non-overlapping subsets were used as the training set and the testing set. Demographic information of the study population was summarized in Table 1.

2.2 DNA methylation data

Genomic DNA was extracted from peripheral blood leukocytes (PBL) samples, epityped using the Illumina HumanMethylation450 BeadChip (450 K). Genomic DNA was bisulfite converted, then whole-genome amplified, enzymatically fragmented and purified. DNA samples were randomly hybridized to the array, which were then fluorescently stained, scanned, and assessed for fluorescence intensities (11). DNAm sites were quantified with β-values, which represents the proportion of methylation level (12, 13). After excluding DNAm sites that overlap with SNP or are not uniquely mapped to the reference genome, 473,864 DNAm sites were available for analysis (11).

2.3 DNAm sites selection

From previous publications, three sets of DNAm sites were selected as candidate smoking-related DNAm sites to build the prediction models. Those studies were selected as they covered different ethnicities, sexes, source tissue types of extracted DNA, smoking status classification, and the number of DNAm sites reported.

Barcelona et al reported 26 DNAm sites that are significantly associated with current smoking status among African American women using saliva samples and replicated in African American men and women using blood cells (4). Among those DNAm sites, 18 were available for model fitting in our 450K data. Gao et al provided a systematic review of DNAm studies in blood DNA from 17 earlier studies with a total of 1,460 smoking-associated DNAm sites, among which 61 were reported more than 3 times (14). These DNAm sites could potentially help quantify a more precise long-term smoking exposure assessment (14). Bollepalli et al built a smoking status prediction model using lasso regression and included 121 DNAm sites that were considered significantly associated with smoking status including current, past and never smokers (6).

This study constructed the prediction models based on the above three sets of DNAm sites that will be further referred to by the number of DNAm sites obtained from each of them (18, 61, and 121).

2.4 Machine learning methods and data analysis

Random forests and elastic net regularization are machine learning algorithms used in this study. In this study, we focused on two classifications: 1. current smoker vs. non-current smoker; 2. past smoker vs. never smoker, and evaluated the prediction accuracy and other performance measures of the prediction models using three subsets of DNAm sites. Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) curve was used to assess the predictive ability of each machine learning model. A prediction model was built based on DNAm and smoking status data of one subset of the study participants (training set). The model's prediction performance was then assessed using the second subset (testing set). All statistical analyses were done in R Version 3.5.3 (https://www.r-project.org/).

2.4.1 Random forests

Random forests model consists of a large number of decision trees that are uncorrelated, where each decision tree is built by a random subset of predictors (i.e. DNAm data in this study) (15). Each decision tree makes its own class prediction and the class with the most votes becomes the prediction of the overall model. In R, a package 'randomForest' that provides functions that could make random forest predictions was used in this analysis (16).

The predictive ability of predictors in random forests is measured by "importance". In this analysis, we used the mean decrease in accuracy (MDA), referring to the mean accuracy decrease when the predictors are excluded across all trees in one random forest run, to assess whether a predictor should be included in the model. Negative values in MDA means that the performance of the model increases after excluding a certain preditor, indicating that this predictor does no better than a random guess in the classification of smoking status, thus it is recommended to be eliminated from the set of preditors for the model (17).

Random forest method also implements out-of-bag (OOB) error rate that measures the prediction error. It is the average error for a training sample that does not contain the OOB sample in the bootstrap sample during model building (18). This measure helps determine the optimal number of trees to grow in a random forest run, where the error rate converges and stabilizes.

We used the averaged MDA and error rate of 50 random forest runs to determine which predictors are of significance in each model and the optimal number of trees to grow for each model to achieve prediction accuracy and modeling optimization, with respect to maximum efficiency on time and space complexity.

2.4.2 Elastic net regularization

Elastic net regularization combines both ridge penalty and lasso penalty in a single model of regularized regression (19). In R, package 'glmnet' provides a function that could perform elastic net regularization (20). It's formulated as:

$$(\sum e^2) + \lambda\ [\alpha(|V_1|+|V_2|+\ldots+|V_n|) + (1-\alpha)\ (V_1^2+V_2^2+\ldots+V_n^2)]$$

Where e is the residual, λ is the shrinkage parameter for ridge and lasso regression penalty, and α could be customized as any value from 0 to 1 when calling the regression model to adjust the weight of ridge and lasso regression penalty in the modeling process. V is the coefficient for each predictor (i.e. DNAm sites), and n equals the number of predictors used in the model. Optimal λ is determined using 10-fold cross-validation, and optimal α will be determined using the one that produces the largest AUC.

3. Results

3.1 Sample characteristics

As summarized in Table 1, demographic information of each set for training and testing is similar, if not identical, since study participants in each set came from one individual of a twin pair except for the two singletons. Smoking behavior for the two sets is also comparable.

3.2 Classification of current and non-current smoking

In the classification of current and non-current smokers, random forests perform similarly as elastic net regularization (Figure 1). For random forests, AUC is 0.922 for the 18 predictor set, 0.904 for the 61 predictor set, and 0.892 for the 121 predictor set. For elastic net, AUC is 0.929 for the 18 predictor set, 0.938 for the 61 predictor set, and 0.922 for the 121 predictor set. Reversing training and testing sets results in prediction with AUC ± <0.05. Specifically, for random forests, AUC is 0.912 for the 18 predictor set, 0.917 for the 61 predictor set, and 0.939 for the 121 predictor set. For elastic net, AUC is 0.936 for the 18 predictor set, 0.933 for the 61 predictor set, and 0.941 for the 121 predictor set.

3.3 Classification of past and never smoking

In the classification of past and never smokers, the performance of both models built from random forests and elastic net regularization decreased. For random forests, the model using 18 predictors has an AUC of 0.660, the model using 61 predictors has an AUC of 0.742, and the model using 121 predictors has an AUC of 0.774 (Figure 2A). For elastic net regularization, the performance is better compared with the corresponding random forest model (Figure 2B). And when the model was built using the 121 predictor set, the performance is further improved, but still lower than that in the prediction of current and non-current smoking. Specifically, the AUCs are 0.757, 0.808 and 0.887 for the 18, 61 and 121 predictor sets, respectively. Reversing training and testing sets results in prediction with AUC ± <0.05. For random forests, AUCs are 0.695, 0.788 and 0.813 for the 18, 61 and 121 predictor sets, respectively. For elastic net, AUCs are 0.766, 0.811 and 0.903 for the 18, 61 and 121 predictor sets, respectively.

3.4 Optimization of tuning parameters

The OOB error rate of different models using the random forest approach was also calculated and plotted, which could help determine the optimal number of trees to grow and whether the model's performance is stabilized after the number of trees grown reached some point (Figure 3). It appears that using 18 predictors, performance of the models both in predicting current vs. non-current smokers and past vs. never smokers get slightly worse when number of trees keeps growing, from 0.15 to 0.16 for current vs. non-current smokers and from 0.38 to 0.40 for past vs. never smokers, indicating by more sampling we are adding more noise to the model. When using 61 predictors, there is also a slight increase in error rate when modeling past vs. never smokers form 0.33 to 0.34, but the performance stabilizes when modeling current vs. non-current smokers. When using 121 predictors, the error rate keeps decreasing as the number of trees grown increases up to 3000 in models that predict current vs. non-current smokers; it stabilizes in models that predict past vs. never smokers when the number of trees grown exceeds 500 approximately.

For elastic net, models reached maximum performance (defined as reaching maximum AUC in this analysis) when tuning parameter α is around 0.02 when differentiating past and never smokers, but optimal α varies for classification of current and non-current smokers. λ that generates the minimum mean cross-validated error differs across different runs, predictor sets, and classification types (current vs. non-current or past vs. never), with a range in [0.03, 0.40]. For current and non-current smoker classification, number of coefficients (V) in the optimal model is around 7, 14 and 96 for the originally 18, 61 and 121 predictor set. For past and never smoker classification, number of coefficients in the optimal model is around 16, 55 and 97 for the originally 18, 61 and 121 predictor set.

3.5 Predictor selection

MDA for each predictor in the random forest approach was examined using 50 runs of random forest prediction. Predictors with MDA < 0 were excluded. In models predicting current vs non-current smokers, 14 out 18, 35 out of 61, and 42 out of 121 predictors remained in the prediction model with positive MDA; in models predicting past vs. never smokers, 10 out of 18, 32 out of 61, 57 out of 121 predictors remained in the prediction model with positive MDA. DNAm sites that were calculated as significant were summarized in Table 2 and Table 3 with their values of mean importance and MDA in the respective models. After such exclusion, precision improvement for each predictor set was summarized in Table 4. AUC slightly increases to 0.6925, 0.7539 and 0.8289 for the originally 18, 61 and 121 predictors respectively.

3.6 Summary

In summary, sample characteristics are comparable for training and testing sets. Classification of current and non-current smokers achieves more accuracy than the classification of past and never smokers for both methods, but elastic net performs better in differentiating past and never smokers than random forests. Generally, OOB error rates in random forests fluctuate for smaller predictor sets and stabilize for larger predictor sets as the number of trees grows. After exclusion of insignificant predictors, random forests' performance in classification of past and never smokers increases, but only to a small extent.

4. Discussion

In this study, quantified prediction of tobacco smoking using DNAm data was conducted. Random forests and elastic net regularization models were built using selected DNAm sites from three previous studies of smoking-related DNAm using different tissue types, ethnicities of participants and number of predictors. Using non-overlapping training and testing datasets from study samples, we were able to evaluate the predictive ability, predictor selection, and potential overfitting.

Evaluation criteria for a "good" AUC could depend on the situation, but generally, an AUC of 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered good, and more than 0.9 is outstanding (21). Overall, models built from all predictor sets could predict current and non-current smokers accurately with AUCs around 0.9. The prediction models were less accurate in differentiating past and never smokers, compared with differentiating current and non-current smokers, especially when using the smallest set of predictors that were identified in an African American female cohort with saliva samples that focused on current smoking status. As a comparison, the DNAm data used in the present study was obtained from blood samples of a group of Caucasian males. The performance of both random forests and elastic net regularization models had consistent prediction accuracy in predicting current vs. non-current smokers. In predicting past and never smokers, random forest models were less accurate than elastic net models measured by AUC of ROC curves. The prediction accuracy for both algorithms improves as the number of predictors increases in predicting past vs. never smokers, and the improvement is more substantial using the elastic net method. Performance of the elastic net model that predicts past vs. never smokers approached the performance of its current vs. non-current counterpart when more predictors were included (121 predictors).

The less accurate prediction in differentiating past and never smokers could be possibly attributed to 1) markers used in the analysis are less predictive in past and never smokers compared with current and non-current smokers, especially for the 18 predictor set where the original DNAm sites were identified based on current smoking status; 2) accurate prediction of past and never smokers may require more DNAm markers due to weaker predictive ability of individual markers.

It is notable that when elastic net reaches its best performance for classification of past and never smokers, α value is close to 0, meaning that ridge regression weights more in producing an optimal model for smoking status than lasso regression in the elastic net formula. Ridge regression has the advantage of dealing with correlated variables in modeling. This could indicate a possible explanation to as why elastic net performs better when predicting past and never smokers when some correlated DNAm markers contribute to the prediction. Models predicting current and non-current smokers could be generalized across different types of samples, gender and races; as for models predicting past and never smokers, different machine learning approaches and predictor sets generate different performances.

In the variable selection process of past and never smoker differentiation, we noticed two major characteristics of the predictor sets. With larger predictor sets, the performance improves. On the other hand, the exclusion of insignificant variables improved the prediction accuracy. This might be attributed to different variables' different contribution and significance in constructing the models. Meanwhile, OOB error rates for smaller predictor sets fluctuated and stabilized for larger predictor sets. Thus, variable selection should be considered and evaluated to achieve optimal prediction accuracy for these machine learning methods.

This study demonstrated the potential application of machine learning prediction models of smoking status using DNAm data that could be generalized to different populations, types of samples, and smoking status categorization, with acceptable precision. Instead of examining hundreds of thousands of DNAm sites, only less than 100 preditors are needed in building the model with relatively high accuracy. Random forests and elastic net are both capable of dealing with a large number of predictors, but time and space efficiency are always important considerations in evaluating how good a model is. By comparing the performance using different machine learning algorithms, the generalizability of the DNAm sites selected by the models could be examined.

This study has several notable limitations. First, the dataset is restricted in Caucasian males only, whereas the results could possibly be different in other demographic populations. Secondly, the DNAm was measured in blood samples - while this is the most commonly used sample in DNAm retrieving, performance could be different for samples from other tissue types. Also, smoking status measured in the study population is based on self-reported status which is prone to reporting errors. Thus, the prediction accuracy of the smoking status could be underestimated in the present study. Using a gold standard of smoking status (e.g., blood-based cotinine levels), as well as examining other variables of smoking behaviors such as time from smoking cessation or smoking dosage (such as measured in pack-time) would potentially benefit the accuracy and evaluation of the results.

Application of the study results could lead to a better prediction of smoking status with a relatively small number of predictors needed. The DNAm-predicted smoking status can minimize the missing data of smoking status, and be a more accurate measurement of smoking behavior than self-reported data in risk prediction of disease outcomes. Studies have shown that the accuracy of self-reported smoking status could be less than 80%, and missing data in such statistics is partially due to unwillingness to report and partially due to passive smoking that was not considered in questionnaires (22, 23). Compared with the performance of models built in this study, the predicted smoking status with more completeness and objectiveness may replace self-reported smoking status in risk prediction of smoking-related diseases such as coronary artery diseases, coronary heart diseases, hypertension.

References

1. West R. Tobacco smoking: Health impact, prevalence, correlates and interventions. Psychol Health. 32(8): 1018–1036 (2017).
2. Reubi D, Berridge V. The Internationalisation of Tobacco Control, 1950–2010. Med Hist. 60(4): 453–472 (2016).
3. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27 K discovery and replication. Am J Hum Genet. 2011;88(4):450–7. doi:10.1016/j.ajhg.2011.03.003
4. Zeilinger S, Kuhnel B, Klopp N et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS One. 8(5):e63812 (2013).
5. Barcelona V, Huang Y, Brown K et al. Novel DNA methylation sites associated with cigarette smoking among African Americans. Epigenetics 14, 4 (2019).
6. Bollepalli S, Korhonen T, Kaprio J. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. Epigenomics11(13):1469-1486 (2019).
7. Sun YV. Multigenic modeling of complex disease by random forests. Adv Genet. 72:73-99 (2010).
8. Goldberg J, Curran B, Vitek ME, Henderson WG, Boyko EJ. The Vietnam Era Twin Registry. Twin Res. 5(5), 476–481 (2002)
9. Vaccarino V, Brennan M-L, Miller AH et al. Association of major depressive disorder with serum myeloperoxidase and other markers of inflammation: a twin study. Biol. Psychiatry 64(6), 476–483 (2008).
10. Vaccarino V, Lampert R, Bremner JD et al. Depressive symptoms and heart rate variability: evidence for a shared genetic substrate in a study of twins. Psychosom. Med. 70(6), 628–636 (2008).

11. Huang Y, Hui Q, Walker D et al. Untargeted metabolomics reveals multiple metabolites influencing smoking-related DNA methylation. Epigenomics. 10(4):379-393 (2018).

12. Klebaner D, Huang Y, Hui Q et al. X chromosome-wide analysis identifies DNA methylation sites influenced by cigarette smoking. Clin. Epigenetics 8, 20 (2016).

13. Chen Y-A, Lemire M, Choufani S et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics 8(2), 203–209 (2013).

14. Gao X, Jia M, Zhang Y et al. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. Clinical Epigenetics (2015) 7:113.

15. Breiman L. Random forests. Machine learning. 45:5 (2001).

16. Liaw A, Wiener M. Breiman and Cutler's Random Forests for Classification and Regression. CRAN (2018).

17. Robin Genuer, Jean-Michel Poggi, Christine Tuleau-Malot. Variable selection using Random Forests. Pattern Recognition Letters, Elsevier, 2010, 31 (14), pp.2225-2236. ffhal-00755489f.

18. Hastie T, Tibshirani R, Friedman J. Elements of Statistical Learning Ed. 2. Springer. p592-593 (2009).

19. Zou H, Hastie T. Regularization and variable selection via the elastic net.J. R. Statist. Soc. 67, Part 2, pp. 301–320 (2005)

20. Friedman J, Hastie T, Tibshirani R et al. Lasso and Elastic-Net Regularized Generalized Linear Models. CRAN (2019)

21. Tharwat A. Classification assessment methods. Applied Computing and Informatics (2018).

22. Spencer K, Cowans NJ. Accuracy of self-reported smoking status in first trimester aneuploidy screening. Prenat Diagn. 33(3):245-50 (2013).

23. Hwang JH, Kim JY, Lee DH et al. Underestimation of Self-Reported Smoking Prevalence in Korean Adolescents: Evidence from Gold Standard by Combined Method. Int J Environ Res Public Health. 15(4). pii: E689 (2018).

Table 1. Demographic information for the study sample of 218 Caucasian male twins

	Set 1 (n=109)	Set 2 (n=109)	Total (n=218)
Age			
Mean (SD)	55.6 (3.30)	55.6 (3.29)	55.6 (3.29)
Median [Min, Max]	56.0 [48.0, 63.0]	56.0 [48.0, 63.0]	56.0 [48.0, 63.0]
Obesity			
No	63 (57.8%)	64 (58.7%)	127 (58.3%)
Yes	46 (42.2%)	45 (41.3%)	91 (41.7%)
Education Level			
Less than High School	6 (5.5%)	3 (2.8%)	9 (4.1%)
High School	42 (38.5%)	34 (31.2%)	76 (34.9%)
Some College	33 (30.3%)	39 (35.8%)	72 (33.0%)
College Degree or above	28 (25.7%)	33 (30.3%)	61 (28.0%)
Smoking Status			
Never	25 (22.9%)	27 (24.8%)	52 (23.9%)
Past	52 (47.7%)	50 (45.9%)	102 (46.8%)
Current	32 (29.4%)	32 (29.4%)	64 (29.4%)
Twin Type			
Monozygotic	82 (75.2%)	82 (75.2%)	164 (75.2%)
Dizygotic	27 (24.8%)	27 (24.8%)	54 (24.8%)

Figure 1. ROC in predicting current and non-current smokers using model built from 50 random forest runs averaged (A) and elastic net regularization (B)

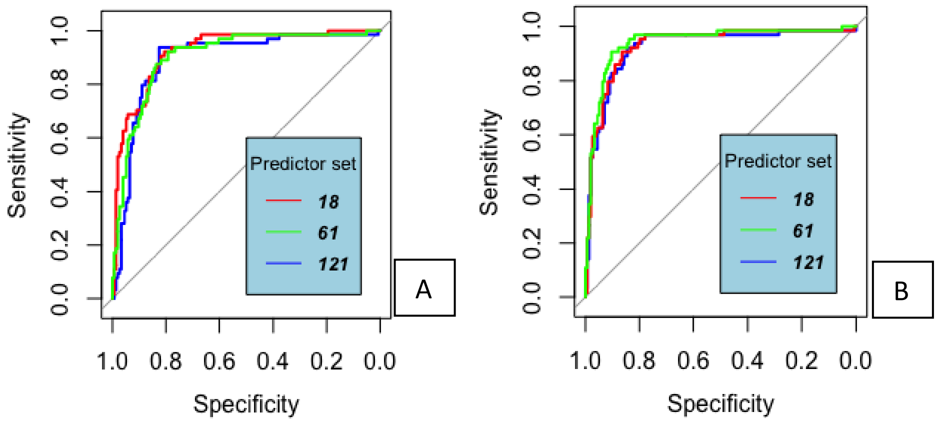

Figure 2. ROC in predicting past and never smokers using model built from 50 random forest runs averaged (A) and elastic net regularization (B)

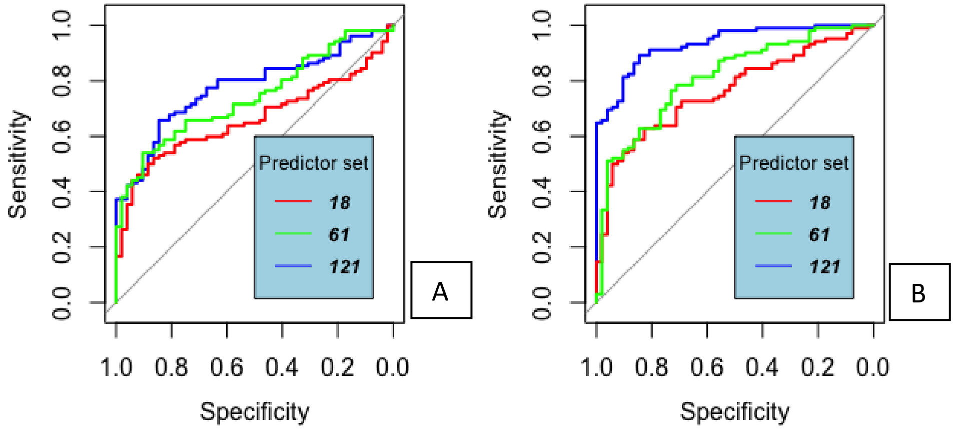

Figure 3. Mean and 95% confidence interval of out-of-bag (OBB) error rate in predicting current and non-current smokers using 50 runs of random forests

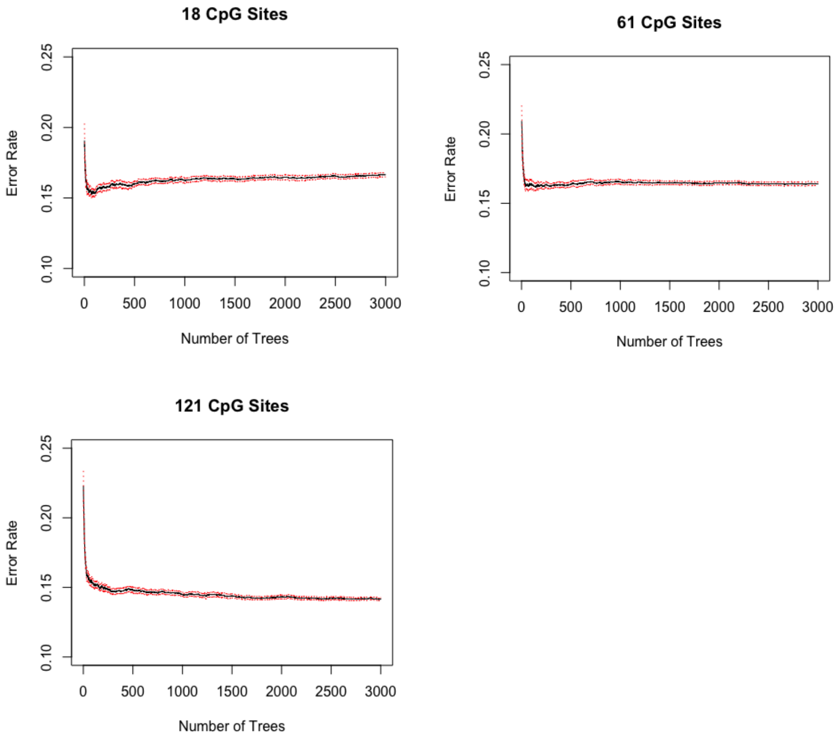

Figure 4. Mean and 95% confidence interval of out-of-bag (OBB) error rate in predicting past and never smokers using 50 runs of random forests

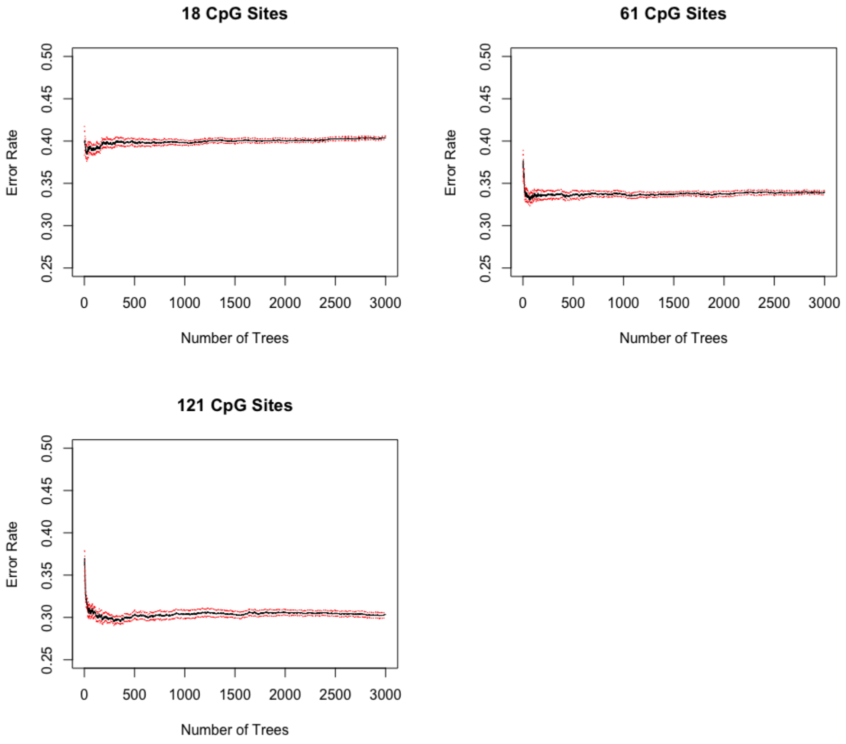

Table 2. Importance and mean decrease in accuracy (MDA) of DNAm sites that are significant in models predicting current and non-current smokers

DNAm Sites	121 Importance	121 MDA	18 Importance	18 MDA	61 Importance	61 MDA
cg05575921	3.056E-01	1.136E-01	2.561E-01	1.104E-01	2.005E-01	6.800E-02
cg16117605	1.500E-03	3.700E-03				
cg21566642	2.482E-03	2.236E-03	1.271E-02	1.190E-03	1.452E-03	1.422E-03
cg07721625	3.843E-03	2.077E-03				
cg21733098	1.417E-05	1.467E-03				
cg10957001	-1.665E-03	1.419E-03				
cg13944838	-5.102E-04	9.512E-04				
cg26048448	-3.348E-04	9.210E-04				
cg18877361	-6.325E-04	6.198E-04				
cg26029902	9.065E-04	5.505E-04				
cg09173768	-3.127E-04	5.288E-04				
cg26103168	-6.308E-04	4.869E-04				
cg13771313	-2.191E-04	3.937E-04				
cg06715410	1.320E-03	3.077E-04				
cg23942311	-9.741E-04	2.942E-04				
cg21594961	-9.568E-05	2.340E-04				
cg26086649	1.725E-03	1.644E-04				
cg16113156	-6.891E-04	1.371E-04				

cg20839206	3.206E-05	1.188E-04				
cg02243946	1.065E-03	9.463E-05				
cg00593900	-6.483E-04	7.616E-05				
cg06677021	5.977E-04	6.929E-05				
cg02725398	-5.815E-04	6.491E-05				
cg25242471	-8.366E-04	5.787E-05				
cg16775095	-3.259E-04	5.138E-05				
cg18268547	-2.762E-04	5.097E-05				
cg06644428	-8.914E-04	4.959E-05			-1.308E-03	3.226E-04
cg01080924	5.023E-04	4.645E-05				
cg18369516	-8.162E-05	3.817E-05				
cg06120313	3.208E-04	2.997E-05				
cg06442199	-5.751E-04	2.777E-05				
cg17619755	-5.900E-04	2.454E-05				
cg22947000	-6.270E-04	2.152E-05				
cg17453416	-1.357E-03	1.885E-05				
cg20738735	-1.383E-04	1.622E-05				
cg18106898	1.222E-04	1.593E-05				
cg26169299	1.304E-05	1.502E-05				
cg03245590	-8.464E-04	1.367E-05				
cg17535283	-1.227E-04	4.586E-06				
cg18161956	2.282E-04	2.847E-06				
cg15064086	-1.602E-04	9.189E-07				
cg00075467	3.401E-04	2.973E-08				
cg00073090			1.327E-02	3.738E-04		
cg00748718			1.627E-03	1.005E-03		
cg01731783					-1.762E-04	3.522E-05
cg01899089					9.755E-04	6.343E-04
cg01901332					-3.649E-04	2.679E-04
cg01940273			5.017E-02	4.321E-02	2.378E-02	3.029E-02
cg02451831					1.115E-02	4.204E-03
cg03547355					2.629E-05	1.311E-05
cg03991871					-8.242E-04	5.205E-07
cg04885881					-1.201E-03	1.490E-04
cg05284742					2.429E-04	8.698E-04
cg05644151			-8.517E-04	5.639E-04		
cg05951221					1.428E-02	3.210E-03
cg06126421					-7.980E-04	1.693E-05
cg07824483			-5.247E-05	1.205E-04		
cg11207515					5.122E-03	6.090E-04
cg11314684					-1.760E-03	1.139E-04
cg11660018					-7.706E-04	6.628E-04
cg13976502					-1.232E-03	9.588E-05
cg14389122			1.999E-02	6.909E-03		
cg14580211					8.291E-04	6.353E-05
cg15342087					9.803E-04	1.079E-03
cg16937168			1.444E-03	5.743E-05		
cg19859270					2.232E-03	1.562E-04

cg20295214					5.277E-03	3.372E-03
cg21121843					1.016E-02	2.357E-03
cg21161138					3.878E-03	3.306E-03
cg21611682					2.205E-04	7.375E-05
cg22132788					3.026E-03	7.335E-04
cg23079012			-1.002E-02	6.299E-03		
cg23161492					9.557E-03	8.847E-04
cg23771366					-3.467E-03	1.836E-03
cg23916896					-1.075E-03	1.269E-04
cg24859433					1.504E-03	1.111E-03
cg24996979					7.377E-03	3.524E-03
cg25189904			-6.536E-03	1.238E-02	-1.087E-03	2.408E-03
cg26271591					1.702E-04	1.469E-04
cg26703534			5.502E-02	7.604E-03	1.993E-02	5.267E-03
cg27174698			-1.374E-03	3.581E-05		
cg27241845			4.839E-03	1.057E-02	1.397E-03	1.031E-02

Table 3. Importance and mean decrease in accuracy (MDA) of DNAm sites that are significant in models predicting past and never smokers

DNAm Sites	**121 Importance**	**121 MDA**	**18 Importance**	**18 MDA**	**61 Importance**	**61 MDA**
cg05951221	4.368E-02	6.838E-02			3.011E-02	3.328E-02
cg06644428	9.963E-03	1.346E-02			1.321E-02	6.330E-03
cg06126421	7.262E-03	1.258E-02			4.758E-03	5.324E-03
cg27650870	7.566E-03	1.076E-02				
cg09173768	2.922E-03	3.605E-03				
cg22947000	4.406E-06	2.773E-03				
cg09068031	2.519E-03	2.415E-03				
cg24629356	7.260E-04	2.042E-03				
cg00066239	-7.259E-05	1.973E-03				
cg13626582	1.278E-03	1.939E-03				
cg13619177	4.674E-04	1.862E-03				
cg01080924	2.255E-03	1.515E-03				
cg16113156	-5.776E-05	1.404E-03				
cg13910813	4.707E-04	1.298E-03				
cg07499182	1.763E-03	1.015E-03				
cg18315060	-3.539E-04	9.415E-04				
cg10525394	2.587E-04	8.736E-04				
cg19572487	-3.103E-04	8.329E-04			-6.486E-04	1.930E-03
cg15064086	-1.917E-04	7.737E-04				
cg05323345	-3.635E-04	7.722E-04				
cg19091257	4.831E-04	7.633E-04				
cg16702083	8.945E-06	7.093E-04				
cg23576855	1.827E-03	6.918E-04			3.372E-03	6.019E-04
cg20618441	-7.835E-05	6.813E-04				
cg06597652	-1.060E-04	6.722E-04				

cg13791092	1.665E-04	6.570E-04				
cg19643109	2.951E-04	6.442E-04				
cg23942311	-8.424E-05	5.893E-04				
cg00593900	1.746E-04	4.995E-04				
cg17619755	7.310E-04	4.859E-04				
cg26048448	-1.295E-04	4.266E-04				
cg21733098	-1.696E-04	2.706E-04				
cg26558023	-2.297E-04	2.674E-04				
cg23126342	1.380E-05	2.671E-04				
cg13771313	6.946E-05	2.655E-04				
cg18877361	5.090E-05	2.492E-04				
cg12438330	1.735E-04	2.481E-04				
cg03133799	-2.947E-04	2.217E-04				
cg25221984	-3.829E-04	1.946E-04				
cg12589188	6.139E-05	1.932E-04				
cg10957001	-2.056E-04	1.881E-04				
cg13944838	3.447E-04	1.569E-04				
cg21450627	1.552E-04	1.465E-04				
cg09298273	-3.657E-04	1.350E-04				
cg01273991	-1.971E-05	1.192E-04				
cg22587600	-1.305E-04	1.145E-04				
cg02431260	-2.644E-04	9.458E-05				
cg03847932	1.140E-04	8.222E-05				
cg06442199	2.359E-04	7.843E-05				
cg18161956	-1.686E-04	7.141E-05				
cg26169299	-6.639E-05	7.050E-05				
cg03936870	-1.318E-04	6.436E-05				
cg20738735	3.608E-04	5.878E-05				
cg05293490	-2.413E-04	4.246E-05				
cg10006428	-3.079E-05	2.027E-05				
cg22331349	4.890E-04	5.558E-06				
cg00846554	-9.710E-05	4.132E-06				
cg01899089					3.851E-03	1.803E-03
cg01940273			2.027E-02	8.836E-03		
cg02451831					-4.765E-04	6.080E-03
cg02657160					-7.016E-04	2.688E-04
cg03991871					-6.995E-04	6.216E-04
cg05575921			6.184E-03	1.021E-03		
cg07123182					-7.516E-04	1.209E-03
cg11231349					4.572E-04	3.917E-04
cg11660018					6.808E-03	1.141E-02
cg12803068					2.368E-03	4.673E-03
cg12806681					2.346E-04	7.760E-04
cg13976502					-7.903E-04	9.301E-04
cg14389122			-5.744E-03	2.004E-02		
cg14753356			-6.341E-04	8.107E-03	-4.460E-04	1.609E-03
cg14817490					-3.297E-03	9.875E-03
cg15342087					4.143E-03	7.132E-03

cg19695041	-2.971E-03	4.148E-03		
cg19859270			2.322E-03	6.751E-03
cg20295214			-4.291E-05	6.507E-04
cg21161138			-4.456E-03	1.666E-02
cg21566642	3.070E-02	7.478E-03		
cg21913886			-7.760E-04	2.373E-03
cg22132788			1.889E-03	2.988E-03
cg23771366			5.994E-03	4.473E-03
cg23916896			2.329E-03	3.378E-03
cg24090911			-2.239E-03	4.531E-04
cg24859433			1.552E-04	4.616E-03
cg24996979			3.302E-04	7.791E-04
cg25189904	1.246E-02	1.981E-02	1.330E-03	1.252E-03
cg25648203	-6.856E-03	1.151E-02	-1.309E-03	1.851E-03
cg26703534	-1.281E-04	2.911E-03	2.588E-04	1.041E-03
cg26963277			-3.596E-04	6.627E-04
cg27174698	1.368E-03	6.773E-03		

Table 4. Mean performance improvement with standard deviation for models predicting past vs. never smokers after exclusion of insignificant predictors from the three predictor sets using 50 runs of random forests

Number of predictors: Full (Reduced)	**18 (10)**	**61 (32)**	**121 (57)**
Full set AUC (SD)	0.6604 (0.0029)	0.7423 (0.0034)	0.7744 (0.0040)
Reduced set AUC (SD)	0.6925 (0.0029)	0.7539 (0.0032)	0.8289 (0.0032)