EpiSmokEr: A robust classifier to determine smoking status from DNA methylation data

Sailalitha Bollepalli[1,2], Tellervo Korhonen[1,3], Jaakko Kaprio[1,2], Miina Ollikainen[1,2*] and Simon Anders[1,4*]

Author Affiliations:
[1]Institute for Molecular Medicine (FIMM), University of Helsinki, Finland;
[2]Department of Public Health, University of Helsinki, Finland;
[3]National Institute for Health and Welfare (THL), Finland;
[4]Center for Molecular Biology (ZMBH), University of Heidelberg, Germany.

Corresponding Author: miina.ollikainen@helsinki.fi

* Equal contribution

Keywords: DNA methylation, epigenetic smoking status, smoking status classifier, multinomial LASSO

Abstract

Self-reported smoking status is prone to misclassification due to under-reporting, while biomarkers like cotinine can only measure recent exposure. Smoking strongly influences DNA methylation, with current, former and never smokers exhibiting different methylation profiles. Recently, two approaches were proposed to calculate scores based on smoking-responsive DNA methylation loci, to serve as reliable indicators of long-term exposure and potential biomarkers to estimate smoking behavior. However, these two methodologies need significant improvements to make them globally applicable to all populations and to achieve an optimal classification of individuals with unknown smoking habits. To advance the practical applicability of the smoking-associated methylation signals, we used machine learning methodology to train a classifier for smoking status prediction. We show the prediction performance of our classifier on three independent whole-blood test datasets demonstrating its robustness and global applicability. Furthermore, we show the cross-tissue functionality of our classifier in tissues other than blood. Additionally, we provide the community with an R package, *EpiSmokEr*, facilitating implementation of our classifier to predict smoking status in future studies.

Introduction

Cigarette smoking has widespread and deleterious effects on the human body as a whole and within cells, with a significant impact at the genetic and epigenetic level. Various studies [1–9] have assessed the effect of smoking on DNA methylation. The largest epigenome-wide association study (EWAS) on smoking to date [8] has identified 18,760 smoking associated CpG sites with statistical significance on a single-locus level, but the overall differences between current and never smokers are arguably even more extensive. The well-established impact of smoking on epigenetic and expression patterns necessitates the inclusion of smoking status as a confounder in DNA methylation or gene expression studies. However, information on the study subjects' smoking habits is not always available, and where it is, it is not necessarily reliable as subjects often under-report their smoking [10]. Additionally, current smokers also self-report as non-smokers, which is particularly evident when comparing self-reported smoking with measurements of cotinine, a reliable short-term metabolic marker of nicotine exposure [10,11]. Hence, it has been proposed that determining a subjects' smoking habits from DNA methylation data itself and subsequently using this information as a covariate in the association study of interest to reduce the potential confounding due to smoking [12].

To this end, Elliot *et al* [2] have suggested calculating a weighted score referred to as "smoking score" (SSc) by adding up the methylation levels of 187 CpG sites found to be significantly associated with smoking by Zellinger *et al* [1], after first multiplying each methylation value by its effect size in Zellinger et al.'s EWAS study. This ad hoc procedure yields a score that is associated with smoking behavior, yet there is no clear way to find a suitable threshold score that best separates smokers from non-smokers. Indeed, Elliot *et al*. were required to use different threshold values for two different cohorts on which they tested the score. Additionally, there is no evidence that the effect sizes from marginal (i.e., single-locus) association tests are appropriate for a predictive score. On the contrary, regression theory assumes signals from the loci to be independent, which certainly does not hold true in the case of DNA methylation.

The subsequent work by Zhang *et al* [4] used a more standard regression approach. They first determined significantly associated CpG sites in their cohort, and then employed stepwise logistic regression with forward selection. They found that after including the 4 most significant loci, the addition

of the fifth locus no longer resulted in a significant reduction of the fit deviance. Therefore, their proposed “methylation score” (MS) only requires four CpG sites.

The question of how to derive a predictor of class membership (here: smoker or non-smoker) from high-dimensional observational data (here: the methylation values for hundreds of thousands of CpG sites) is a task thoroughly studied in statistics research and may be considered the prototypical task of the field of machine learning (see Friedmann *et al* [13] for a standard textbook). We, therefore, investigated whether we could gain further improvements by employing some established techniques from this discipline, especially as these have been used with great success in other areas of genomics, most importantly in deriving "signatures" from gene expression data to sub-stratify diseases or provide a prognosis of disease progression (e.g., [14]).

The aforementioned previous works focused only on CpG sites that have been shown to be significantly associated with smoking in a prior GWAS-style analysis. There is, however, no need for this in a machine learning setting; in fact, one typically finds that features that, taken on their own, are not statistically significant can still contribute to improving the classifier. Instead of assuring statistical significance by marginal (i.e., feature by feature) testing, one uses cross-validation to avoid overfitting.

Perhaps the most commonly used approach is L1-penalized regression, commonly called LASSO (least absolute shrinkage and selection operator) [15]. Here we use LASSO in the classification setting (also known as multinomial or logistic regression): Each subject has one of three class labels -- never smoker, former smoker, or current smoker --, and we seek to find a linear combination of methylation values of selected probes that give (via a logistic link function) a probability value for each of the labels.

We note already here three differences in our method compared to the two previous approaches: First, we aim at a three-level classification, while the previous study [4] discriminated only current from never smokers and former from never smokers. Second, the "smoking score" [2] needs to be compared to a specific threshold value to call a person a smoker or a non-smoker. An appealing feature of logistic regression is that it yields probabilities for the class labels, and the class with the greatest probability will be the classifier's output. Alternatively, the probabilities can be transformed to logarithmic odds, with a log odd above zero indicating a probability above 0.5. Hence, multinomial regression implicitly sets the threshold to zero. Third, our approach uses internal cross-validation (CV) to decide how many features to include. This means that the penalty parameter, which controls the size of the selected feature set, is determined by testing a range of candidate values by repeatedly training models on a subset of the data and testing them on a held-out set (see Methods for details). Instead, the previous work [4] has used forward stepwise regression with an arbitrary significance threshold as a stopping criterion. As CV directly estimates the classifiers predictive performance, it is preferable over a significance-based criterion for deciding on the size of the feature set. This also circumvents the often discussed (see, e.g., [16]) issues of lack of type-I error control inherent to stepwise regression.

Results

Training data

Our training dataset (DILGOM) comprised 474 individuals representative of the Finnish adult population, with a broad age distribution and extensive smoking behavior information (Table 1 and Supplementary Methods). We used peripheral blood leukocyte DNA methylation data, assessed by Infinium HumanMethylation450 BeadChips, and self-reported smoking status information.

Table 1: Characteristics of the training dataset

Training Dataset	**Self-reported Smoking Status**		
DILGOM	**Current Smokers**	**Former Smokers***	**Never Smokers**
N	113	118	243
Age (mean±SD)	48.8±12.7	56.3±11.6	51.95±14.4
Age range (years)	25-72	25-74	25-74
Sex (M/F)	57/56	61/57	97/146
Cotinine information (%)	76	26	1
Nationality	Finnish		
Source	Whole blood		
Platform	Infinium HumanMethylation450 BeadChip		

*Only former smokers who quit smoking at least a year ago were considered.

An optimal penalization parameter was determined through internal cross-validation. This resulted in a classifier using 121 CpG sites corresponding to 92 genes (Table 2). Figure 1 illustrates the workflow of the classifier with respect to the training and test datasets.

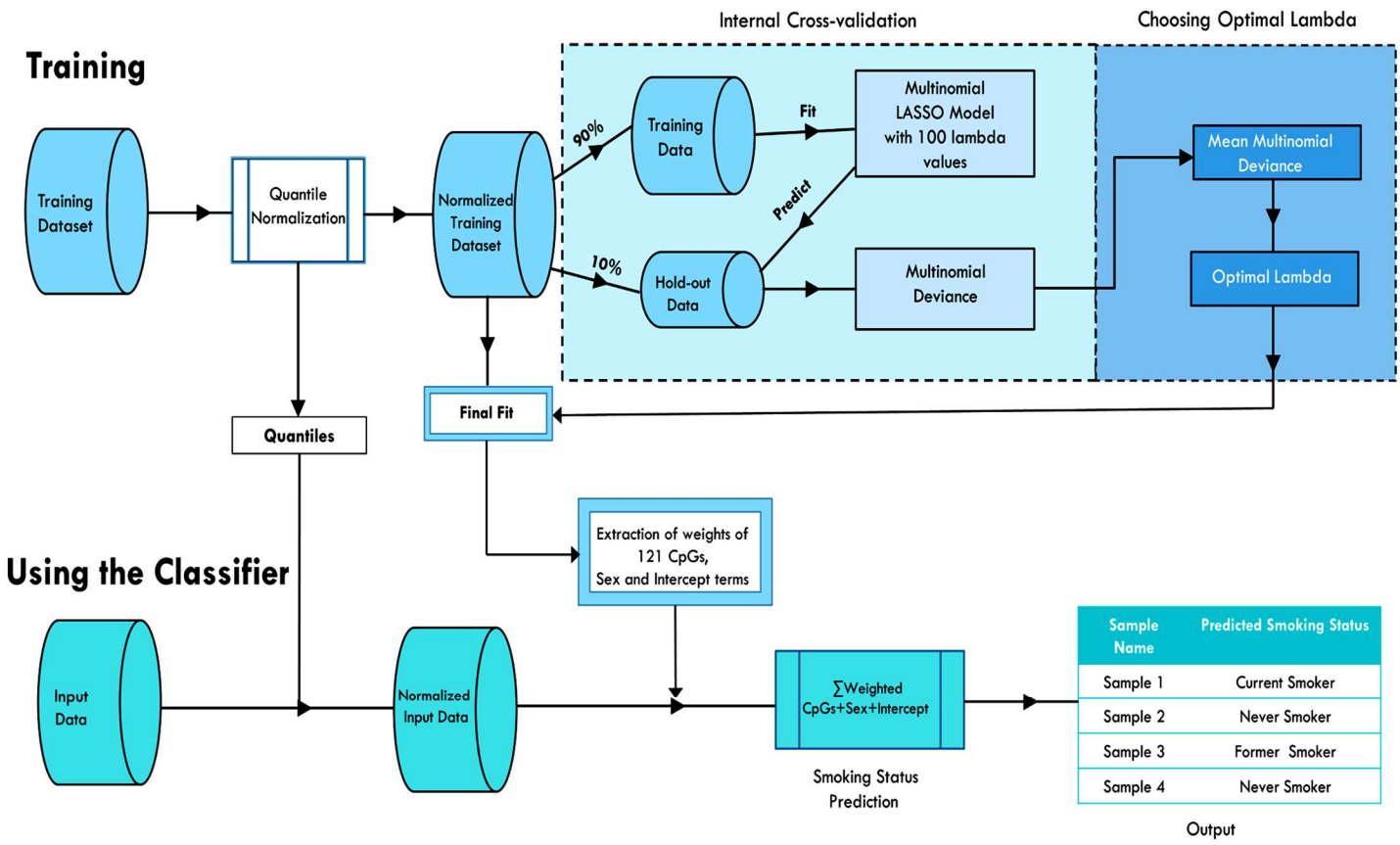

Figure 1: Schematic representation of the workflow of the classifier

Training: *Quantile normalized training data was used to perform 100 iterations of internal cross-validation to choose the optimal lambda. Using that lambda, a final fit was performed on the full training data, yielding 123 non-zero coefficients.*
Using the Classifier: *The input methylation data is first normalized and then probabilities for each smoking status are calculated. The smoking status with the highest probability is taken as the call.*

We used sex as an additional covariate owing to differences in smoking prevalence among men and women. We did not adjust for the blood cell type composition as SVD analysis revealed no strong association between top principal components and blood cell subtypes (Supplementary Methods). Moreover, several studies have reported that adjusting for blood cell type composition does not alter the smoking-associated methylation signals [2,3,7,8,12].

Validation and performance comparison

For validation, we have chosen three external, independent test datasets: FTC, EIRA and CARDIOGENICS, originating from different populations to demonstrate transferability and global applicability of our classifier (Table 3). To visualize the performance of our classifier we generated confusion matrices by comparing the self-reported with predicted smoking behavior (Figure 2 A-C). In order to quantify the performance of the classifier, we determined sensitivity and specificity (Table 4), as well as area under the receiver-operating curve (AUROC) (Figures 3 A-C). Owing to the multinomial

model with three smoking status categories, these values were calculated for each category, comparing that one versus the other two categories (See Methods).

Table 3: Characteristics of three whole blood test datasets

Description of test datasets used in this study to assess the performance of smoking status classifier.

Test Datasets	Dataset	N	Sex (M/F)	Self-reported Smoking Status			Nationality	Age (mean±SD)	Age range (years)
				Current	Former	Never			
FTC (Finnish Twin Cohort)	Available upon a reasonable request	408	166/242	67	141	200	Finnish	62.2±4.3	32.3-69.7
EIRA (Epidemiological Investigation of Rheumatoid Arthritis)	GSE42861	687	196/491	266	228	193	Swedish	51.9± 11.8	18-70
CARDIOGENICS	GSE50660	464	327/137	22	263	179	French and British	55.4±6.7	38-67

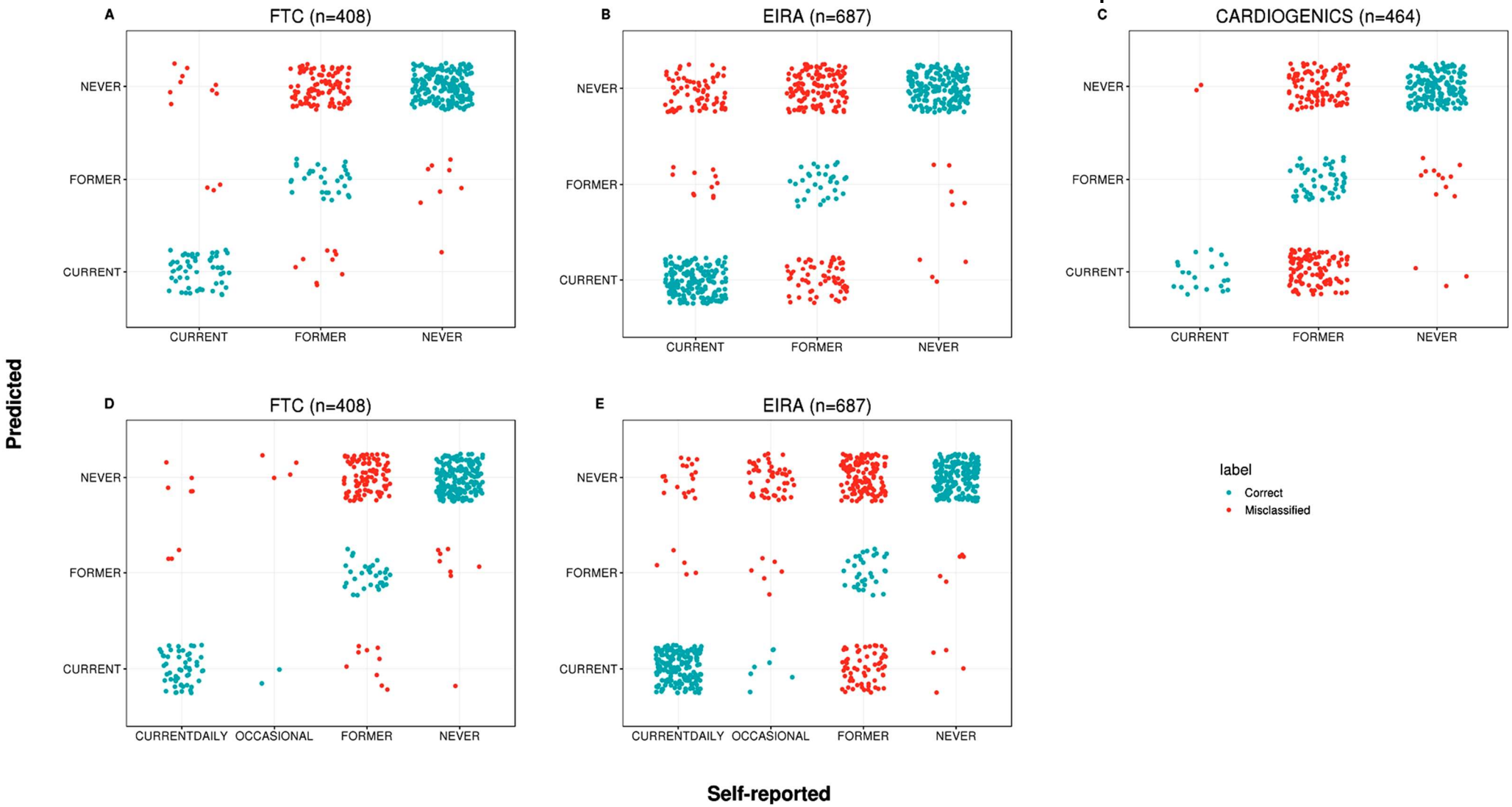

Figure 2: Confusion matrices comparing self-reported versus predicted smoking behavior

Confusion matrices for FTC (A), EIRA (B) and CARDIOGENICS (C) datasets. Self-reported smoking behavior is compared with the predicted smoking status determined by the classifier. Points in cyan represent agreement between self-reported and predicted smoking statuses. Points in salmon represent misclassification. Based on the questionnaire data, the class of current smokers from the FTC and EIRA datasets could be further subdivided into current daily and occasional smokers. The confusion matrices D and E from FTC and EIRA show that the majority of the misclassification arises from the occasional smokers, who were often predicted as former or never smokers by the classifier.

Table 4: Prediction estimates from Smoking status (SSt) approach, across the training and test datasets.

Smoking status categories were predicted using the classifier, based on the 121 CpGs chosen by multinomial LASSO. Predicted smoking status was compared with self-reported smoking status. For each smoking status category, the relative sensitivity, specificity and AUROC values were calculated by comparing that one with the other two categories (one vs all approach).

Datasets	Sensitivity (%)	Specificity (%)	AUROC
Training Dataset (DILGOM) (N=474)			
Current vs others	75	98	0.95
Former vs others	60	99	0.97
Never vs others	99	72	0.97
Test Datasets			
Finnish Twin Cohort (FTC) (N=408)			
Current vs others	82	97	0.96
Former vs others	22	96	0.58
Never vs others	96	47	0.78
EIRA (N=687)			
Current vs others	69	84	0.82
Former vs others	14	97	0.61
Never vs others	95	58	0.88
CARDIOGENICS (N=464)			
Current vs others	91	73	0.86
Former vs others	19	95	0.49
Never vs others	92	65	0.87

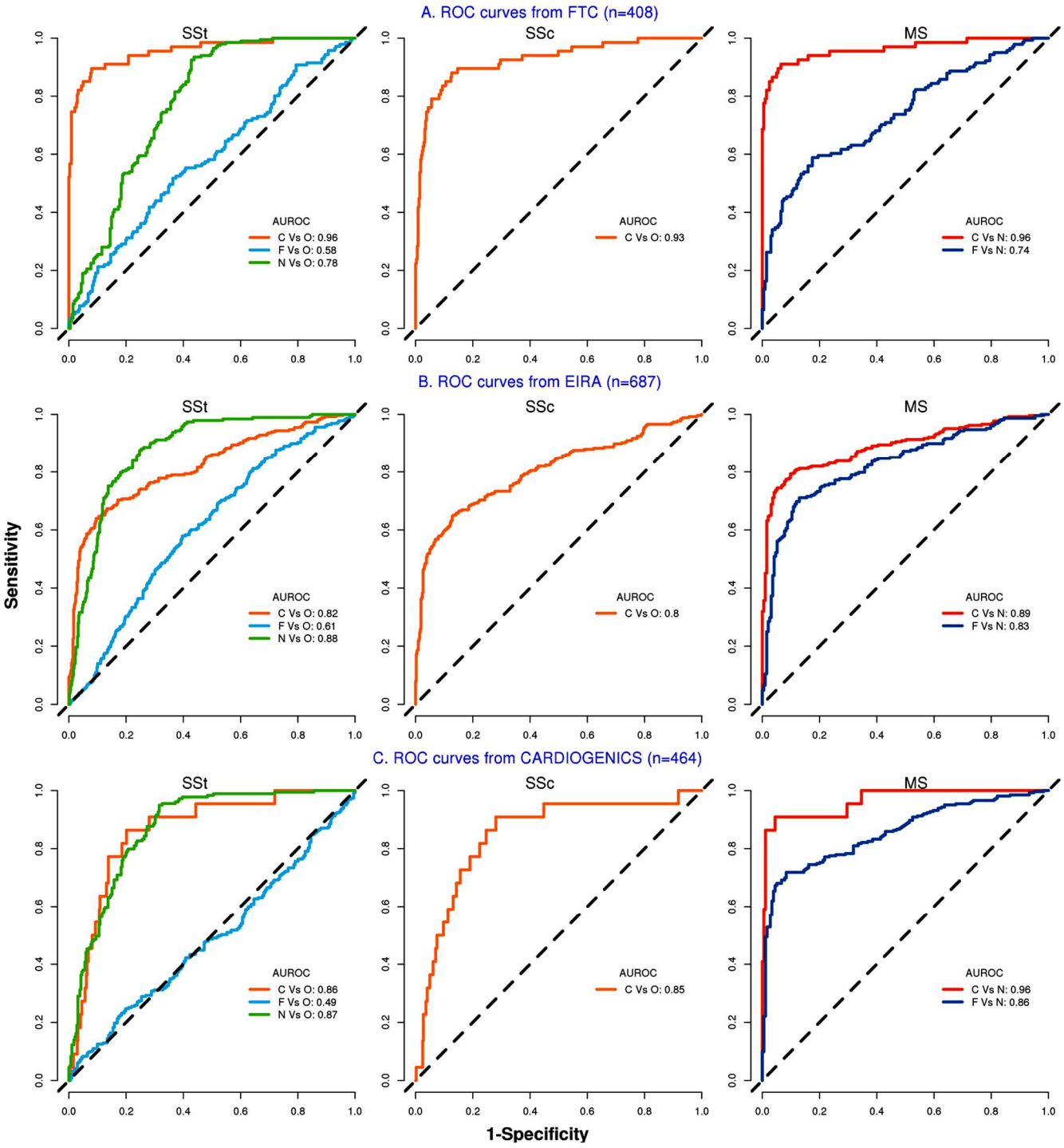

Figure 3: ROC curves from the test datasets

Receiver operating characteristic (ROC) curves were plotted to illustrate the discriminative ability of Smoking Status (SSt, the method presented here), Smoking Score (SSc) and Methylation Score (MS)) when varying the threshold values. For SSt ROC curves were obtained for each smoking category by comparing with the other two categories (one vs all approach). SSc can discriminate only current smokers from other two categories. MS performs only binary comparisons (current versus never and former versus never). ***A*** *ROC curves for the FTC dataset (n=408),* ***B*** *the EIRA dataset (n=687), and* ***C*** *the CARDIOGENICS dataset (n=464). Here C: Current, F: Former, N: Never and O: other two categories.*

At the actual threshold of zero, on average in the three test datasets, self-reported current smokers were identified with 81% sensitivity and 85% specificity, while never smokers were identified with 94% sensitivity and 57% specificity (Table 4). Sensitivity values were highest in the "never vs other" comparison across all three test datasets. For former smokers, the classifier showed low sensitivity across the test datasets, with an average sensitivity of only 18%. However, the classifier was able to identify individuals who did not belong to the former smoker category with a high average specificity of 96%.

We show the performance of our classifier along with the other two methods (Table 5) using ROC curves (Figures 3 A-C). However, the underlying differences in the approaches from the three methods limited our ability to comprehensively compare their performances across the test datasets. Our classifier's multinomial model allowed us to discriminate each smoking status category from the other two (current versus others, former versus others and never versus others). The Smoking Score (SSc) method [2] only discriminates current smokers from the other two, while the Methylation Score (MS) method [4] allows only binary comparisons (current versus never and former versus never). Our classifier had the highest AUROC values in discriminating current smokers from the other two categories across all datasets in comparison with SSc. In addition, we observed strong correlation between our method's log odds and the SSc values, for the current smoker category across the training and test datasets (Supplementary Figure 1). Slightly lower AUROC values were observed for never smokers in the FTC dataset than in the other test datasets. We note that our classifier showed moderate to marginal ability in discriminating former smokers from others across the test datasets. MS showed good performance across all the test datasets, if used only for binary classification comparisons, i.e. discriminating current from never smokers and former from never smokers.

Table 5: Methylation-based smoking status estimation approaches compared in this study

We list the three smoking status estimation approaches compared in this study, specific normalization method employed in that approach, and the reference.

Outcome	Approach	Normalisation	Reference
Smoking Status (SSt)	Weighted sum of 121 CpGs, Sex and intercept terms. These CpGs were identified by Multinomial LASSO (ML).	Quantile Normalization (QN)	Our classifier
Smoking Score (SSc)	Weighted sum of 187 CpGs. These CpGs were identified in an EWAS [1].	Subset Quantile Normalization (SQN)	Elliott *et al* [2]
Methylation Score (MS)	Weighted sum of 4 CpGs. These CpGs were selected via stepwise regression from results of an EWAS [4].	Illumina Normalization (ILN)	Zhang *et al* [4]

To visualize the results from all the three methods we generated bee swarm plots of SSc and MS across all the test datasets (Figures 4A and 4B). We illustrated the results from our classifier as ternary plots based on the estimated smoking probabilities (Figures 4C-4E). Each corner of a ternary plot corresponds to a smoking status category. Each point in the ternary plot represents an individual with a corresponding triplet of estimated smoking probabilities that add up to 100%. So the higher the probability for a specific smoking status category the closer is the point to the corresponding corner. Individuals with mixed profiles showing similar composition to more than one category are on one of the sidelines or in the

central part of the ternary plot. Hence, we observe most of the former smokers in the middle of the ternary plot as they exhibit mixed profiles of current and never smokers. We also show that the threshold of 17.55 proposed by SSc [2] for European ethnicity failed to distinguish the majority of the current smokers from other smoking status categories (Figure 4 A). We note that our classifier uses an implicit threshold of zero which curtails the need for a population-specific threshold. For instance, the training dataset used for building the classifier was from the Finnish population and yet the classifier performed well in the test datasets originating from different populations.

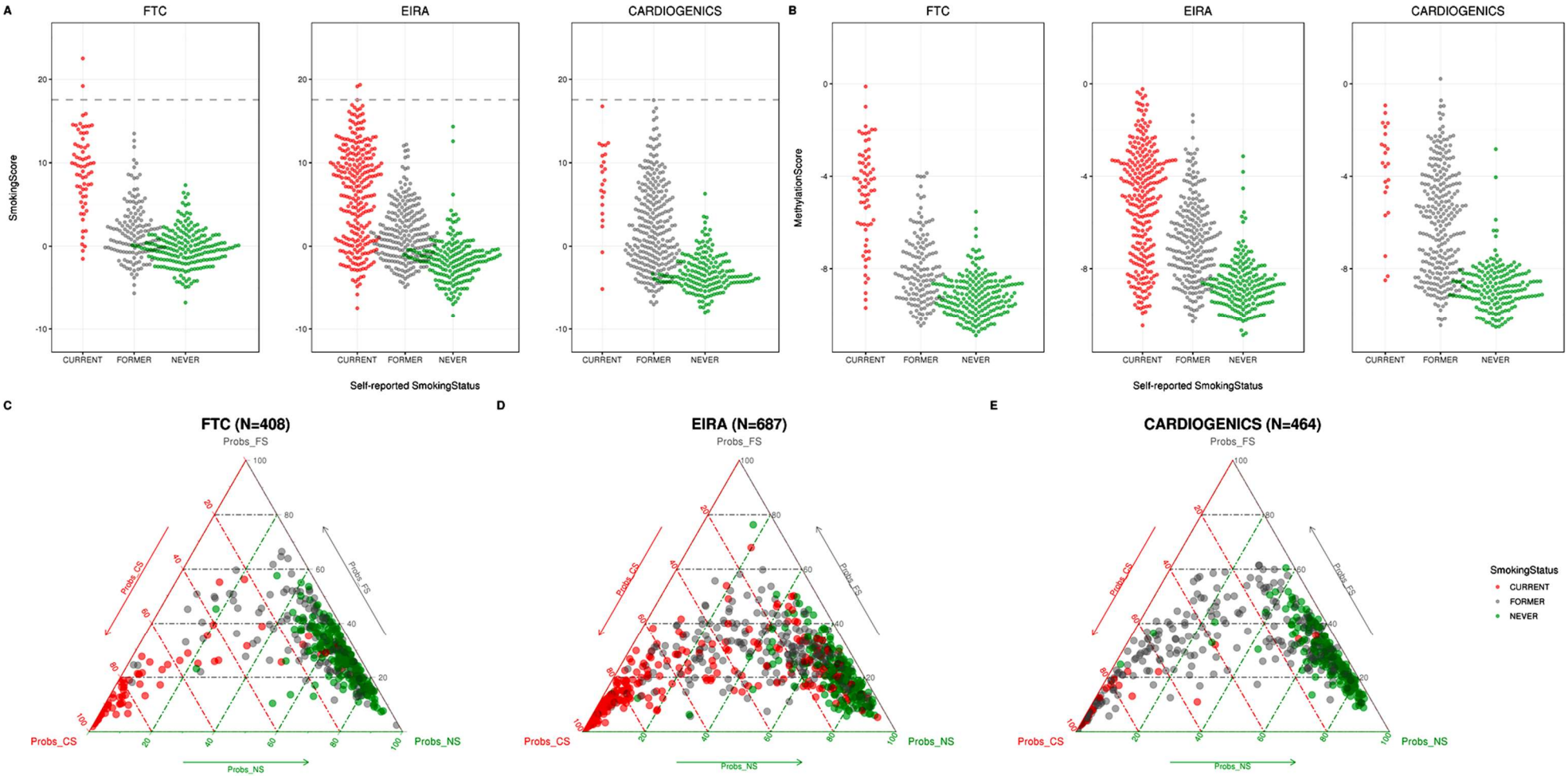
A
FTC
EIRA
CARDIOGENICS
SmokingScore
CURRENT
FORMER
NEVER
Self-reported SmokingStatus
B
MethylationScore
C
FTC (N=408)
D
EIRA (N=687)
E
CARDIOGENICS (N=464)
Probs_FS
Probs_CS
Probs_NS
SmokingStatus
CURRENT
FORMER
NEVER

Figure 4: Comparison of smoking score, methylation score and classifier results across three datasets

Bee swarm plots and ternary plots comparing smoking score, methylation score and classifier results from three datasets: FTC, EIRA and CARDIOGENICS. ***A*** *Smoking score computed for current, former and never classes of the three test datasets. The grey dashed line represents the European threshold of 17.55 proposed by Elliott et al. Only two individuals from FTC and EIRA datasets have a smoking score above 17.55.* ***B*** *Methylation score computed for current, former and never classes of three datasets. Each point in the bee swarm plots corresponds to an individual, with the x-axes representing the self-reported smoking behavior category and the y-axes corresponding to the smoking or methylation score. The color of the points corresponds to self-reported smoking behavior.*

Figures C-E represent ternary plots based on the predicted probabilities given by our classifier in the test datasets. The ternary plot is an equilateral triangle with each corner corresponding to a smoking status category. Each point in the ternary plot represents an individual with the corresponding triplet of probabilities that adds up to 100%. So the higher the probability for a category the closer is the point to the corresponding corner. The color of the points corresponds to self-reported smoking behavior.

The effect of normalization method on smoking behavior prediction

Each of the three approaches under consideration (Table 5) used a different method to normalize DNA methylation data. To assess the effect of normalization on smoking behavior prediction, we normalized the test datasets using three different normalization methods (quantile normalization (QN), subset quantile normalization (SQN) and Illumina normalization (ILN); described in Methods). We then estimated smoking scores, methylation scores and smoking status using two test datasets after each of the normalization methods. In the third test dataset (CARDIOGENICS) due to unavailability of raw data, only QN was performed. When comparing the performance of smoking status (SSt) approach for different normalization methods, QN had better average sensitivity and specificity values (Supplementary Table 2).

Reasons for misclassification

To discern the possible reasons for misclassification we performed a comprehensive phenotypic assessment on the two test datasets with additional phenotypic information. In FTC and EIRA, the "current smoker" category was further subdivided into current daily smokers and occasional smokers. To highlight the impact of occasional smokers on the performance of the classifier, we have also shown the confusion matrices with the occasional smokers as a separate category in the FTC and EIRA datasets (Figure 2 D-E). For instance, in the EIRA dataset, of the 66 occasional smokers, 53 were predicted as never smokers and 6 as former smokers (Figure 2 E). This misclassification can be attributed to the similar methylation profiles of occasional smokers to that of never and former smokers. Therefore, we expect a decrease in the classification accuracy when current daily and occasional smokers are combined into current smoker category.

We also observed an increase in the sensitivity values when occasional smokers were excluded from the current smokers. For example, the exclusion of 66 occasional smokers in the EIRA dataset improved the sensitivity from 69% to 88% (Supplementary Table 2). Similarly, in the FTC dataset exclusion of 6 occasional smokers increased the sensitivity from 82% to 87%. As expected, the exclusion of occasional smokers further improved discrimination of AUROC values of current smokers across all the methods (Supplementary Figure 2 A-B).

To also understand the misclassification of former smokers, we used the extensive smoking behavior information available from the FTC dataset. In FTC, the former smokers class had the highest misclassification rate, with 72% (n=101) of former smokers identified as never smokers (Figure 2D and

Supplementary Figure 3 A-C). Phenotypic verification revealed that 85 of 101 individuals had quit smoking for more than 10 years prior to blood sampling for DNA methylation measurement (Supplementary Figure 3 A-B). The 6% (n=9) of former smokers identified as current smokers had quit smoking recently and had higher mean pack-years compared to other former smokers (Supplementary Figure 3 B-C). We also observed lower mean and range for pack-years among current smokers identified as never smokers (Supplementary Figure 3B).

In addition to active smoking, every class of smokers were exposed to passive (second-hand) smoking to a certain extent (Figure 4D). We also noticed that some of the individuals were exposed to both active and passive smoking for prolonged periods (Figure 4D). Our classifier results indicated that methylation levels reflect the cumulative smoking exposure. However, from our analyses, it is hard to determine the extent of passive smoking which leads to a change in the methylation levels.

Cross-tissue performance

We have used methylation data from buccal tissue [17], peripheral blood mononuclear cells (PBMCs) [18] and lung cancer [19] to assess the cross-tissue performance of the classifier. Current smokers were identified with highest sensitivity (95%) and specificity (97%) in the buccal tissue dataset (Table 6). Based on the self-reported smoking status, this dataset was labelled as “cigarette smoker” and “non-tobacco smoker”. Specificity drops by 37% when considering samples that were identified as former smokers by our classifier. This suggests that around 38% (n=15) of the non-tobacco smokers are former smokers based on their methylation profiles. This is in line with the specific cohort definition of non-tobacco smokers [17] being individuals abstinent from tobacco or nicotine-containing products for at least 5 years.

Similarly, in the PBMC dataset with 50 smokers and 61 non-smokers, sensitivity will be reduced to 76% if we were to consider the inclusion of samples that were classified as former smokers by our classifier. Notably, both the buccal tissue and PBMC datasets comprised individuals of African-American ethnicity, also demonstrating the global applicability of our classifier across different ethnicities.

Table 6: Characteristics and performance measures from tissues other than blood
Three datasets from tissues other than blood were used to demonstrate the cross-tissue applicability of the classifier.

Source Tissue	Dataset	N	Self-reported Smoking Status		Sex (M/F)	Age (mean)	Age range (years)	Current Vs Never	
			Current	Never				Sensitivity (%)	Specificity (%)
Buccal	GSE94876	80[a]	40[b]	40[b]	80/0	47.2±7.9	35-60	95	97
PBMCs	GSE53045	111[c]	50	61	0/111	48.4 ± 10	NA	89	96
Lung adenocarcinoma	GSE3927[d]	297	254[d]	43	152/145	65± 10.4	40-90	96	10
Lung squamous cell carcinoma	GSE39279[d]	84	80[d]	4	61/23	65.5±10.5	35-83	90	25

[a] Of these 80 individuals, 58 were Caucasian, 21 were African-American and 1 unknown ethnicity. [b] In buccal tissue data smoking behavior was defined as cigarette smoker and non-tobacco smoker. [c] African-American ancestry. [d] In lung tumor data smoking behavior was defined as smoker and non-smoker, where smoker includes both current and former. NA: Not available.

The R package EpiSmokEr

We implemented the smoking status classifier as an R package. To estimate smoking behavior, *EpiSmokEr* (*Epi*genetic *Smok*ing status *E*stimator) expects methylation data from the Illumina HumanMethylation450 (or compatible) arrays as an input. Input data can be either raw methylation data in the form of intensity data (IDAT) files or a normalized methylation matrix in the beta scale ranging from 0 to 1. A sample sheet with sex is required to complement the methylation data. The *normaliseData* function has a suite of customized internal functions to normalize and calculate beta values from the IDAT files. To perform SQN and ILN normalization we use functionality from the minfi package [20], and for the QN we use a custom function. Quantiles from our training dataset are used to adjust the distribution of the input data.

Furthermore, the R package also provides functions to calculate the SSc by Elliott *et al* [2] and the MS by Zhang *et al* [4], allowing users to select the most appropriate method and output for their purposes. Extensive documentation with examples is provided in the package vignette. A typical calculation starting with the IDAT files to the estimation of smoking status takes only a few minutes.

EpiSmokEr is available from GitHub: https://github.com/sailalithabollepalli/EpiSmokEr

Discussion

The Smoking Score proposed by Elliott *et al* [2] focuses on distinguishing self-reported current smokers from other classes, while the Methylation Score proposed by Zhang *et al* [4] can only distinguish current from never smokers and former from never smokers. To improve the practical applicability of smoking-associated methylation signals, we applied a machine learning based approach to build a classifier that predicts smoking status. Our classifier estimates the probabilities of an individual being a current, former or never smoker. The smoking status category with the highest probability is determined as the predicted smoking status. A major advantage of our approach is its global applicability without need for a population-specific threshold to predict the smoking status. Also, the straightforward output from the classifier i.e. a label of certain smoking status category, is easy to interpret.

Performance evaluation

We calculated sensitivity and specificity of our classifier across test datasets using zero as a threshold. As the calculation of sensitivity and specificity requires a threshold applicable to all populations and smoking status categories we cannot compute these values for the other two methods. Instead, we used ROC curves to illustrate the discriminative ability of our classifier in comparison with other methods. Plotting the sensitivity versus 1-specificity allows us to see the performance of each method at every threshold. Our classifier showed good discriminative ability for identifying current smokers and never smokers across all the test datasets. MS [4] performed well across the test datasets, however, this approach can only distinguish between two smoking status categories. Also, a good correlation between our classifier's log odds and smoking score values indicates that the methods tend to agree. Notably, the European ethnic smoking score threshold of 17.55 proposed earlier [2] was not applicable in any of the datasets used here. This suggests that the threshold is dataset-specific and may be partly governed by technical or genetic differences which cannot be discerned. It is also possible that smoking score can only be used for distinguishing heavy current smokers from the other classes. Also, in the case of MS, no threshold has been given, which makes it difficult to interpret the score when calculated for individual samples as there is no reference value to compare with. In addition, the performance of smoking status estimation approaches could be marginally dependent and influenced by the choice of normalization method. For instance, SQN normalized data showed lower sensitivity for current and never smoking status categories, compared to the other two normalization methods (Supplementary Table 2). This difference might be a

result of SQN normalization which adjusts the signal intensities of Infinium II probes with respect to the Infinium I probes. Taken together, using our classifier provides a two-fold advantage as it does not require a threshold to identify smoking status [2] and can handle three smoking status categories [4].

Examining the reasons for performance differences

In addition to the differences in the training schemes and normalization methods used in the three smoking behavior prediction methods we have compared in this paper (smoking status (SSt), methylation score (MS) and smoking score (SSc)), it is essential to consider other factors that might have affected their performance. Sensitivity and specificity measures from our classifier are lower than those reported for the other two methods considered here [2,4] and a comparable method for alcohol status [21]. However, these methylation biomarkers identified for smoking and alcohol were mainly focused on classifying heavy smoking and drinking, respectively [2,21]. To make a fair comparison we must consider factors such as cohort composition, gender, age, and the smoking-status definitions considered by the other classifiers. For deriving the European ethnic smoking score threshold of 17.55 to discriminate current smokers from others [2], 16 heavy smokers from a dataset of 95 men were considered. To develop the methylation score [4] older adults aged 50–75 years were used. On the other hand, we have trained our classifier on a dataset with a broad age spectrum, including both men and women. Zhang *et al [4]* considered only binary comparisons, while we have focused on identifying the smoking behavior of an individual giving equal priority to all the three smoking classes.

Primarily, we calculate the prediction accuracy estimates by comparing the predicted status to the self-reported smoking behavior which is prone to misreporting and poor recall of long-term smoking history [10]. Furthermore, approaches are based on the unrealistic independence (mutually exclusive) assumption that no two smoking status categories will have any overlapping methylation profiles. However, this assumption may not be valid for smoking-associated methylation profiles and also for smoking behavior in general. To elaborate, DNA methylation profiles of former smokers might resemble either current or never smokers, depending on how long they have been abstinent (total time of abstinence between cessation and sample taken) and how much and for how long they have smoked before quitting (pack-year history). Also, current smokers can greatly vary based on their current (cigarettes per day) and cumulative past (pack-year history) smoking behavior, as well as use of tobacco products other than cigarettes. Apart from misreporting and passive smoking, the never smoker category is expected to be more homogenous compared to the other two categories. The heterogeneity within each category leads to mixed DNA methylation profiles, which is biologically and perhaps clinically relevant, and as expected. However, the same affects the performance of the classifier, which is seen as misclassification. This in turn could be falsely interpreted as poor performance of the classifier.

Results from the secondary analyses in the FTC dataset using total time of abstinence and cumulative pack-years showed that misclassifications by the classifier are often biologically meaningful. For instance, misclassification of former smokers with higher cessation time as never smokers was reported by several earlier studies [1–3,5,7,8]. This specific misclassification can be attributed to the reversal of methylation levels in former smokers to the levels similar to never smokers with cessation time.

Passive smoking was prevalent across all the smoking status categories in the FTC dataset. However, based on the results from our classifier, we cannot quantify the extent of the impact of passive smoking on the methylation profiles and prediction of smoking status.

Also, verifying the classifier results regarding smoking-status definitions aided in understanding misclassifications. We observed that occasional smokers are often misclassified as never smokers especially if they have smoked rarely. This was clear from the FTC and EIRA datasets, where the

classifier performance was affected by including occasional smokers. To sum up, it is important to consider the biological relevance and the aforementioned factors while interpreting the results and performance metrics from a classifier. Further phenotypic information on occasional smokers show clearly that classification to a smoking category depends on when they have last smoked, how much they typically smoke on days that they smoke and the frequency of smoking.

Prediction of former smokers

Former smokers are increasing worldwide as a result of successful anti-smoking campaigns. There is an increasing necessity for biomarkers to identify former smokers accurately for assessing the extent of irreversible damage caused by smoking and to estimate disease risk (e.g.: lung cancer or COPD) after smoking cessation. Yet, distinguishing former from never and current smokers by DNA methylation has proven challenging for several reasons. The plastic nature of DNA methylation which makes it suitable as a dynamic biomarker to capture the recent and long-term exposure poses an inherent limitation in classifying former smokers. Though reversible, DNA methylation levels of former smokers might never reach the levels of never smokers, even after several decades of cessation [1,3,5]. Hence, based on the duration of abstinence, former smokers may have varying methylation profiles, overlapping with either current or never smokers [1,3–8,22,23]. The extent of overlap also depends on the smoking intensity of the former smokers and the magnitude of change in the methylation levels [5]. Additionally, site-specific reversal of DNA methylation, ranging from sites with very slow or no reversibility to CpGs with quick reversibility [1,3,5] further complicates the classification problem. Furthermore, the presence of genetic variants in the vicinity of the smoking-associated CpG sites might also impact the classification [24]. All these aforementioned factors greatly limit the ability to accurately distinguish former smokers from never and current smokers, thereby reducing the prediction accuracy.

By far the best approach to distinguish former smokers is to consider only a binary classification, by comparing former with never smokers [4,23]. Further, stratifying the former smokers based on the cessation time lead to the identification of persistent smoking-associated CpGs [5,7]. Although our classifier considers three smoking status categories, the intrinsic limitations associated with former smokers and DNA methylation was also reflected in our results. Varying definitions of former smokers across different datasets might also impact the sensitivity of the classifier. Additionally, former smokers exposed to passive smoking may have methylation profiles more similar to current smokers, leading to misclassification. Given the complexity of former smokers both biologically and behaviorally, we caution that former smokers may remain as an ultimate challenge for any DNA methylation-based prediction algorithm. Nevertheless, from a clinical/biological perspective, despite the misclassification, results from the classifier are still significant as they reflect the impact of smoking on DNA methylation and potentially gene function of the individual.

Cross-tissue applicability

We have considered three tissues: buccal, PBMCs and lung cancers to assess the cross-tissue performance of our classifier. Only current and never smoking were self-reported in these datasets. Therefore, sensitivity and specificity measures were calculated only from the samples which were classified as current or never smokers by our classifier.

Given the tissue-specific nature of DNA methylation, the good performance of our blood data trained classifier on the buccal tissue data was perhaps surprising. However, a recent study [25] implemented the smoking index (SI) based on 66 blood-derived smoking-related CpGs [26] on the same buccal dataset and showed successful discrimination of smokers from non-smokers. SI uses smoking-related CpGs associated with DNA methylation age (DNAm-age) [27]. However, SI is study-sample specific as it uses

the non-smokers from the study sample as a reference and cannot distinguish former smokers. Our results are also in line with another study that showed a substantial overlap between smoking-associated CpGs identified in blood and buccal tissues obtained from matched samples [28].

Similarly, results from the PBMCs are also reassuring, as these cells are extracted from whole blood. An EWAS conducted on this dataset [18] identified highest methylation difference between smokers and non-smokers for two loci in the AHRR gene (cg05575921 and cg23576855), which were also a part of 121 CpGs of our classifier. In lung carcinoma samples low specificity rates can be partially attributed to changes in methylation levels of CpGs that are associated with both lung cancer and smoking.

Considerable performance of our classifier in tissues other than blood suggests a broader and systematic impact of smoking on the methylome, across various tissues. Interestingly, the buccal tissue and PBMC datasets used here comprised samples of African-American origins. Hence, these two datasets served as case studies to demonstrate the cross-tissue applicability and global applicability of our classifier.

Practical implications

Identification of smoking status based on methylation is more robust than other traditional biomarkers with short half-lives in body fluids. Our classifier provides an objective measure of smoking status and is globally applicable across all populations curtailing the need for computing population/ethnic-specific thresholds.

Predicted smoking status from our classifier can be useful when self-reported data is missing or not available or is presumed to be highly inaccurate. This is advantageous for larger consortia studies where a smoking phenotype was not collected at the time of sample collection. Also, predicted smoking status can validate and provide precision to the self-reported smoking status and can reduce misreporting bias. Methylation-derived smoking status can also serve as a tool in forensic settings and can also be used as a covariate to adjust for smoking-associated confounding in association analyses like EWAS and GWAS. Finally, the reasonable cross-tissue performance of our classifier shows that it can be applied to tissues other than blood to assess smoking status.

Conclusions

We have developed a robust DNA methylation-based predictor providing an objective measure of smoking status. Our novel approach considers three smoking status categories and is applicable to datasets from different populations and tissues. We also examined reasons for biologically meaningful misclassifications through comprehensive phenotypic evaluation. We provide the community with an R package, *EpiSmokEr*, where raw or normalized DNA methylation data can be used to predict smoking status. *EpiSmokEr* includes functionality to also calculate smoking score and methylation score allowing users to select the most appropriate method and output for their purposes.

Methods

Pre-processing and Quality control

All statistical analyses were performed in R. To ensure inclusion of high-quality samples and probes, stringent quality control measures were implemented. On the training dataset, we performed quality control following the pipeline by Lehne *et al* [29]. Probes with SNPs, or detection P value above 10^{-16},

non-CpG probes and probes at sex chromosomes were set to missing. Samples with missing data above 5% and probes with missing data above 2% were excluded. After the QC, four samples were excluded owing to low-quality.

Quantile normalization

To normalize the training dataset, we separated the probe intensity values into six categories based on the color channel, probe-type and subtype. Within each probe category, quantile normalization was performed by first sorting the actual intensity values within each sample. After sorting average intensity values were calculated across all the samples. Then the actual intensity values were replaced with the corresponding averaged intensity values (quantiles) thereby forcing all the samples to have the same distribution. The quantiles obtained from the six probe categories were saved. Beta values were calculated for each CpG probe using normalized intensity values (Methylated/(Methylated+Unmethylated+100)). We note that we do not rescale the intensities of Infinium II probes based on Infinium I probes, as it is nonessential for our classification purposes.

Following normalization, cross-reactive probes from Chen *et al* [30] and low varying probes with variance < 0.002 across all the individuals were removed. The final training dataset consisted of 52421 probes from 474 samples.

Test datasets were normalized using the same six sets of quantiles (corresponding to probe categories) obtained from the training dataset using a custom function. This approach ensures cross-study performance by fitting the distribution of the test dataset to the training dataset and can be called a "frozen" quantile normalization, following McCall et al [31].

Multinomial LASSO regression

For training the predictor we used the set of 52421 probes available from the DILGOM data after quality control and filtering, to perform LASSO regression using the R package glmnet [32].

We fitted a LASSO-penalized generalized linear model of the multinomial family, for which we briefly review the mathematics here. We write i for the index of the CpG probe on the array and j for the subject (sample). Our input data are the quantile normalized methylation values x_{ij}, and the subjects' smoking status categories (class labels) $k_j \in K$ = {never-smoker, former-smoker, current-smoker}. In multinomial regression, the model fits a linear predictor η_{jk} which is understood as a multinomial equivalent to the log odds of logistic regression, i.e., the predictor corresponds to the probability p_{jk} which the classifier assigns to the call that subject j has smoking status k via the logistic transformation,

$$p_{jk} = \frac{e^{\eta_{jk}}}{\sum_{k' \in K} e^{\eta_{jk'}}} \tag{1}$$

The denominator ensures that the probabilities for the three smoking statuses add up to 1.

The linear predictor η_{jk} is given as a linear combination of the fitted model coefficients β_{ik},

$$\eta_{jk} = \beta_{0k} + \beta_{\mathrm{S}k}\mathrm{X}_{i\mathrm{S}} + \sum_i \beta_{ik}\mathrm{X}_{ij},$$

where each probe *i* has three coefficients β_{ik}. The three intercepts β_{0k} effectively take up the role of thresholds. As we have included sex as an additional covariate, a sex coefficient β_{Sk} appears in the model which is added to the intercept if subject *i* is male ($x_{iS} = 1$) and omitted if subject *i* is female ($x_{iS} = 0$).
The coefficients are found by the model fitting procedure, which maximizes the penalized log likelihood,

$$l(\beta) = \sum_j \log p_{jk_j} + \lambda \, \|\beta\|_G = \frac{1}{N} \sum_{j=1}^{N} (\eta_{jk} - \log \sum_{k' \in K} e^{\eta_{jk'}}) + \lambda \sum_j (\sum_k \beta_{jk}^2)^{1/2}.$$

The penalty term, $\lambda \, \|\beta\|_G$, serves to prevent overfitting. It causes the fit to select only those features (probes) *i* that contribute most to the prediction and shrinks the coefficients β_{ik} for all other probes to zero. We are using here the "grouped" shrinkage option, which the *glmnet* function offers for multinomial regression. It links the three coefficients for a given probe *i* such that they are selected or not selected together (which is reflected in the penalty term by the penalty acting on $(\Sigma_k \beta_{jk}^2)^{1/2}$).

The optimal value for the penalty parameter λ, which regulates the strength of this selection force, is chosen by cross-validation, which we describe next.

Cross-validation

An optimal lambda of 0.055 was chosen from a sequence of 100 candidate values by performing 100 iterations of internal cross-validation. In each iteration, the original training dataset was randomly subdivided into 90% training data and 10% hold-out data. A multinomial LASSO regression model was fit on the training dataset, using the candidate values for λ. For each lambda value, the fit of the prediction is tested on the validation dataset, to obtain estimated probabilities per each class of smoking status.

In each iteration, for every lambda value the multinomial deviance

$$D(\lambda) = -2 \sum_j \log p_{jk_j}$$

was computed, where p_{jk_j} is the probability that the classifier has assigned (using Equation (1)) to held-out subject *j* for that subject's true smoking status k_j when using the coefficients obtained from training with the given penalty parameter λ. We when averaged the deviances for each λ value over the 100 iterations and chose the λ value with lowest average deviation as optimal λ.

We then performed the final fit on the full training dataset and obtained non-zero coefficients for 121 probes. These coefficients (for each of the three statuses: 121 CpG coefficients, one sex and one intercept coefficients) were used in the calculation of smoking statuses for the test data sets (Supplementary Table 1).

Smoking status prediction

When we use the classifier on a given test data set, that data set first had to be quantile normalized. As stated above, this is done reusing the quantiles obtained in the training set. Then, for each subject in the test data set, a probability is calculated for each of the three smoking statuses using Equation (1). The smoking status category with highest probability is reported as the classifier's call.

Data preparation

To compare the performance of our classifier with other methods we have normalized the test datasets accordingly.

For calculating the smoking score (SSc) proposed by Elliott *et al* [2] we have used the subset quantile normalization (SQN) [33]. In SQN, first the reference quantiles are computed for each probe category of Infinium I signals based on CpG annotations S shore, S shelf, N shore, N shelf and distant), to which the type II probes are normalized.

The datasets were normalized using Illumina normalization method (ILN) to calculate methylation score proposed by Zhang *et al* [4]. In ILN, internal controls are used as a reference to which the data were normalized. No additional background correction was performed.

Smoking scores (SSc) and methylation scores (MS)

Smoking scores are calculated as described by the Elliot *et al.* [2], which we briefly recapitulate here: for a given subject *j* in a test data set, the smoking score is calculated as

$$\mathrm{SSc}_j = \sum_{i=1}^{187} w_i (x_{ij} - x_i^{\mathrm{NS}})$$

where the index *i* runs over the 187 CpGs listed in Zellinger et al.'s Table S2 [1]. As before, we denote with x_{ij} the normalized methylation value of this subject for probe *j*. The reference methylation value, x_i^{NS}, for probe *i*, is Zellinger et al.'s median methylation values of never smokers, averaged over discovery and replication cohorts (columns P and S of the Table S2). The probe weight w_i is given by

$$w_i = \frac{x_i^{\mathrm{CS}} - x_i^{\mathrm{NS}}}{\sum_{i'=1}^{187} (x_{i'}^{\mathrm{CS}} - x_{i'}^{\mathrm{NS}})},$$

where x_i^{CS} is the median methylation values of current smokers, also averaged over discovery and replication cohorts (columns Q and T of the Table S2).

For methylation score calculation, the methylation values of the 4 CpG probes *cg05575921, cg05951221, cg02451831* and *cg06126421* are multiplied with their corresponding weights given in the Figure 4 of Zhang *et al [4]* and then summed up.

R packages used

The *preprocessIllumina* and *preprocessQuantile* functions from the minfi package [20] were used to perform ILN and SQN normalization methods, respectively. The Biobase package [34] was used to save the methylation and phenotype data as an *eSet* object. The glmnet package [32] was used to perform multinomial LASSO. The following R packages were used to generate figures: ggplot2 [35], ROCR [36], ggtern [37], ggpubr [38], ggbeeswarm [39] and cowplot [40].

Datasets

Training dataset

We used DILGOM (Dietary, Lifestyle and Genetic determinants of Obesity and Metabolic syndrome) data originated from the Finnish population-based FINRISK 2007 study as the training dataset. A detailed description of the DILGOM dataset is provided in greater detail elsewhere [41,42]. Genome-wide DNA methylation data and self-reported smoking status were available from 514 individuals. Based on the self-reported smoking behavior, three major classes of smoking statuses were identified: current, former and never smokers. In addition, plasma cotinine measures determined by gas chromatography-mass spectrometry [43] were available from 86 current, 31 former and 3 never smokers. For the individuals with cotinine measures available, self-reported smoking status was validated using cotinine levels. Self-ascribed current daily smokers and cotinine verified current smokers (with cotinine levels above 10 ng/ml) were considered as current smokers. Current occasional smokers with either high cotinine values or who smoked their last cigarette less than a half a year ago were also considered as current smokers. Former smokers were defined as individuals who quit smoking for at least a year ago. Following these criteria, 28 former and 2 never smokers with high cotinine values were excluded from the analyses.

Test datasets

We have considered three whole-blood datasets quantified on Infinium HumanMethylation450 BeadChip for validation of our classifier. The first test dataset consisted of 405 twins from the Finnish Twin Cohort (FTC) [44]. The second dataset from the EIRA study comprised 687 participants with 354 rheumatoid arthritis cases [45]. Raw intensity files (IDAT) were available for these two datasets. The third dataset consisted of 464 samples from the CARDIOGENICS consortium, comprising samples from healthy subjects and from subjects with coronary artery disease [3]. This dataset was provided in the methylated and unmethylated signal intensity format, suitable only for performing quantile normalization (color channel, probe type and subtype). Self-reported smoking status was available from all these test datasets. In the FTC and EIRA datasets, the current smoking category was further divided into current daily and occasional smokers. These sub-categories were used to assess the performance of our classifier.

To investigate and assess the cross-tissue performance of our classifier, we used three publicly available normalized datasets from buccal tissue [17], PBMCs [18] and lung tissue [19] (Table 6). No normalization step was performed on these datasets owing to the different tissue types. The buccal tissue dataset comprised samples from Caucasian and African-American ethnicities. The PBMC dataset consisted of adult African-American females.

For the detailed smoking-status definitions from all the datasets please refer to the Supplementary Methods.

DNA methylation analysis

For all whole-blood datasets, [3,41,42,44,45] bisulfite conversion had been performed using the EZ-96 DNA Methylation-Gold Kit (Zymo Research, Irvine, CA, USA) and genome-wide DNA methylation was quantified on the Infinium HumanMethylation450 BeadChip (Illumina, San Diego, CA, USA).

Secondary analyses

To comprehensively scrutinize the misclassifications of our classifier we used the well-annotated FTC dataset. We examined the duration of smoking abstinence (years since quitting) and cumulative smoking

exposure (pack-years), the two most informative smoking behavior variables. Also, we examined the effects of passive smoking on the misclassification.

Abbreviations

AUROC: Area Under the Receiver Operating Characteristic curve
COPD: Chronic Obstructive Pulmonary Disease
DILGOM: Dietary, Lifestyle and Genetic determinants of Obesity and Metabolic syndrome
EIRA: Epidemiological Investigation of Rheumatoid Arthritis
FTC: Finnish Twin Cohort
IDAT: Intensity Data
ILN: Illumina normalization
LASSO: least absolute shrinkage and selection operator
QN: Quantile normalization
ROC: Receiver Operating Characteristic curve
SQN: Subset quantile normalization
SSt: Smoking status, result from our classifier.
SS: Smoking Score, as defined by Elliot et al. [2]
MS: Methylation Score, as defined by Zhang et al. [4]

Declarations

Ethical approval and consent to participate

The DILGOM/FINRISK participants have provided written informed consent. The FINRISK study was designed and conducted according to the principles of the Declaration of Helsinki. Approval for the FINRISK study protocol was obtained from ethics committees of the hospital districts of Helsinki and Uusimaa, Finland.

Ethical permissions for the Finnish Twin Cohort study were obtained from the Ethics committee of the Department of Public Health, University of Helsinki and Helsinki University Central Hospital. Informed consent was obtained from all the participants.

All the other datasets used in the study have obtained permission from their respective ethical boards.

Consent for publication

Not applicable

Availability of data and material

Samples and data from participants of the DILGOM 2007survey are available through The National Institute for Health and Welfare's (THL) Biobank. For further information please contact the THL Biobank (finriski@thl.fi).

The FTC data is not publicly available due to the restrictions of informed consent. However, the FTC data is available through the Institute for Molecular Medicine Finland (FIMM) Data Access Committee (DAC) (fimm-dac@helsinki.fi) for authorized researchers who have IRB/ethics approval and an institutionally approved study plan. Note that anonymized individual-level data can only be released after study has been approved by the Research Ethics Committee of the University of Helsinki and must be carried out in collaboration with FTC investigators. To ensure the protection of privacy and compliance with national data protection legislation, a data use/transfer agreement is needed, the content and specific clauses of

which will depend on the nature of the requested data. For further information please contact Jaakko Kaprio (jaakko.kaprio@helsinki.fi).

All the other test datasets used in this study are publicly available from the Gene Expression Omnibus (GEO).

Competing interests

JK and TK have provided consultation to Pfizer on nicotine dependence and its treatment during 2011–2014 and 2011–2015, respectively. The remaining authors declare no conflict of interest.

Funding

This work was supported by the Academy of Finland [297908 to MO], EC MSC ITN Project EPITRAIN (MO), and the University of Helsinki Research Funds to MO, and Sigrid Juselius Foundation to MO. SA is funded by the Deutsche Forschungsgemeinschaft (DFG) via SFB 1036.

The DILGOM study has been supported by the Academy of Finland Research (Grants 255935 to Prof Markus Perola) for DNA methylation data. DNA methylation data for the DILGOM sample was generated at the Estonian Genome Center, University of Tartu, under the supervision of Prof Andres Metspalu. Data collection of the Finnish twin cohort samples has been supported by the Academy of Finland Center of Excellence in Complex Disease Genetics (Grants 213506 and 129680), the Academy of Finland (Grants 265240 and 263278 to JK), NIH Grant DA12854 (to PAFM), Sigrid Juselius Foundation (to JK), Global Research Award for Nicotine Dependence, Pfizer (to JK), the Wellcome Trust Sanger Institute, UK and the Broad Institute of MIT and Harvard, USA.

Author's contributions

MO, SA and SB conceived the project. SA supervised the development of the classifier. SB implemented the analyses and designed the R package. SB and SA drafted the manuscript. MO provided critical feedback and helped shape the analysis and manuscript. SA and MO jointly supervised this work. TK contributed to the FTC and DILGOM smoking phenotypes. JK is the PI of the FTC and co-designed phenotype and DNA collection for the FTC. All authors discussed the results and commented on and approved the final manuscript.

Acknowledgements

We thank all the researchers for making their datasets publicly available. We are grateful to the DILGOM and the FTC for providing the data. We thank all the participants who contributed to different cohorts. We warmly acknowledge Dr Milla Kibble and Dr Emma Cazaly for providing constructive feedback on the manuscript.

References

1. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS One. 2013;8:e63812.

2. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clin Epigenetics. 2014;6:4.

3. Tsaprouni LG, Yang T-P, Bell J, Dick KJ, Kanoni S, Nisbet J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9:1382–96.

4. Zhang Y, Florath I, Saum K-U, Brenner H. Self-reported smoking, serum cotinine, and blood DNA methylation. Environ Res. 2016;146:395–403.

5. Guida F, Sandanger TM, Castagné R, Campanella G, Polidoro S, Palli D, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. Hum Mol Genet. 2015;24:2349–59.

6. Zhang Y, Yang R, Burwinkel B, Breitling LP, Brenner H. F2RL3 methylation as a biomarker of current and lifetime smoking exposures. Environ Health Perspect. 2014;122:131–7.

7. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics. 2016;8:599–618.

8. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet. 2016;9:436–47.

9. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. Clin Epigenetics. 2015;7:113.

10. Benowitz NL, Schultz KE, Haller CA, Wu AHB, Dains KM, Jacob P. Prevalence of Smoking Assessed Biochemically in an Urban Public Hospital: A Rationale for Routine Cotinine Screening. Am J Epidemiol. 2009;170:885–91.

11. Hsieh SJ, Ware LB, Eisner MD, Yu L, Jacob P 3rd, Havel C, et al. Biomarkers increase detection of active smoking and secondhand smoke exposure in critically ill patients. Crit Care Med. 2011;39:40–5.

12. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet. 2013;22:843–51.

13. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Science & Business Media; 2009.

14. van ’t Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6.

15. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Series B Stat Methodol. [Royal Statistical Society, Wiley]; 1996;58:267–88.

16. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. Br J Math Stat Psychol. 1992;45:265–82.

17. Prasad GL, Jones BA, Chen P, Gregg EO. A cross-sectional study of biomarkers of exposure and effect in smokers and moist snuff consumers. Clin Chem Lab Med. 2016;54:633–42.

18. Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, Simons R, et al. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics. 2014;15:151.

19. Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, et al. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. J Clin Oncol. 2013;31:4140–7.

20. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9.

21. Liu C, Marioni RE, Hedman ÅK, Pfeiffer L, Tsai P-C, Reynolds LM, et al. A DNA methylation biomarker of alcohol consumption. Mol Psychiatry. 2018;23:422–33.

22. Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, et al. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Hum Mol Genet. 2012;21:3073–82.

23. Shenker NS, Ueland PM, Polidoro S, van Veldhoven K, Ricceri F, Brown R, et al. DNA methylation as a long-term biomarker of exposure to tobacco smoke. Epidemiology. 2013;24:712–6.

24. Gao X, Thomsen H, Zhang Y, Breitling LP, Brenner H. The impact of methylation quantitative trait loci (mQTLs) on active smoking-related DNA methylation changes. Clin Epigenetics. 2017;9:87.

25. Nwanaji-Enwerem JC, Cardenas A, Chai PR, Weisskopf MG, Baccarelli AA, Boyer EW. Relationships of Long-term Smoking and Moist Snuff Consumption with a DNA Methylation Age Relevant Smoking Index: An Analysis in Buccal Cells. Nicotine Tob Res [Internet]. 2018; Available from: http://dx.doi.org/10.1093/ntr/nty156

26. Gao X, Zhang Y, Breitling LP, Brenner H. Relationship of tobacco smoking and smoking-related DNA methylation with epigenetic age acceleration. Oncotarget. 2016;7:46878–89.

27. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115.

28. Teschendorff AE, Yang Z, Wong A, Pipinikas CP, Jiao Y, Jones A, et al. Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer. JAMA Oncol. 2015;1:476–85.

29. Lehne B, Drong AW, Loh M, Zhang W, Scott WR, Tan S-T, et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biol. 2015;16:37.

30. Chen Y-A, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8:203–9.

31. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics. 2010;11:242–53.

32. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33:1–22.

33. Touleimat N, Tost J. Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics. 2012;4:325–41.

34. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12:115–21.

35. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer; 2016.

36. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21:3940–1.

37. Hamilton N. ggtern: An Extension to “ggplot2”, for the Creation of Ternary Diagrams [Internet]. 2018. Available from: https://CRAN.R-project.org/package=ggtern

38. Kassambara A. ggpubr: “ggplot2” Based Publication Ready Plots [Internet]. 2018. Available from: http://www.sthda.com/english/rpkgs/ggpubr

39. Clarke E, Sherrill-Mix S. ggbeeswarm: Categorical Scatter (Violin Point) Plots [Internet]. 2017. Available from: https://CRAN.R-project.org/package=ggbeeswarm

40. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for “ggplot2” [Internet]. 2018. Available from: https://CRAN.R-project.org/package=cowplot

41. Inouye M, Silander K, Hamalainen E, Salomaa V, Harald K, Jousilahti P, et al. An immune response network associated with blood lipid levels. PLoS Genet. 2010;6:e1001113.

42. Inouye M, Kettunen J, Soininen P, Silander K, Ripatti S, Kumpula LS, et al. Metabonomic, transcriptomic, and genomic variation of a population cohort. Mol Syst Biol. 2010;6:441.

43. Broms U, Pennanen M, Patja K, Ollila H, Korhonen T, Kankaanpää A, et al. Diurnal Evening Type is Associated with Current Smoking, Nicotine Dependence and Nicotine Intake in the Population Based National FINRISK 2007 Study. J Addict Res Ther [Internet]. 2012;S2. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22905332

44. Kaprio J. The Finnish Twin Cohort Study: an update. Twin Res Hum Genet. 2013;16:157–62.

45. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31:142–7.