European Journal of Epidemiology (2019) 34:1055–1074
https://doi.org/10.1007/s10654-019-00555-w

METHODS

Validated inference of smoking habits from blood with a finite DNA methylation marker set

Silvana C. E. Maas[1,2] · Athina Vidaki[2] · Rory Wilson[3,4] · Alexander Teumer[5,6] · Fan Liu[2,7,8] · Joyce B. J. van Meurs[1,9] · André G. Uitterlinden[1,9] · Dorret I. Boomsma[10] · Eco J. C. de Geus[10] · Gonneke Willemsen[10] · Jenny van Dongen[10] · Carla J. H. van der Kallen[11,12] · P. Eline Slagboom[13] · Marian Beekman[13] · Diana van Heemst[14] · Leonard H. van den Berg[15] · BIOS Consortium · Liesbeth Duijts[16] · Vincent W. V. Jaddoe[1,17,18] · Karl-Heinz Ladwig[4] · Sonja Kunze[3,4] · Annette Peters[3,4,19,20] · M. Arfan Ikram[1] · Hans J. Grabe[21] · Janine F. Felix[1,17,18] · Melanie Waldenberger[3,4,19] · Oscar H. Franco[1] · Mohsen Ghanbari[1,22] · Manfred Kayser[2]

Received: 4 May 2019 / Accepted: 22 August 2019 / Published online: 7 September 2019

Abstract

Inferring a person's smoking habit and history from blood is relevant for complementing or replacing self-reports in epidemiological and public health research, and for forensic applications. However, a finite DNA methylation marker set and a validated statistical model based on a large dataset are not yet available. Employing 14 epigenome-wide association studies for marker discovery, and using data from six population-based cohorts (N = 3764) for model building, we identified 13 CpGs most suitable for inferring smoking versus non-smoking status from blood with a cumulative Area Under the Curve (AUC) of 0.901. Internal fivefold cross-validation yielded an average AUC of 0.897 ± 0.137, while external model validation in an independent population-based cohort (N = 1608) achieved an AUC of 0.911. These 13 CpGs also provided accurate inference of current (average $AUC_{crossvalidation}$ 0.925 ± 0.021, $AUC_{externalvalidation}$ 0.914), former (0.766 ± 0.023, 0.699) and never smoking (0.830 ± 0.019, 0.781) status, allowed inferring pack-years in current smokers (10 pack-years 0.800 ± 0.068, 0.796; 15 pack-years 0.767 ± 0.102, 0.752) and inferring smoking cessation time in former smokers (5 years 0.774 ± 0.024, 0.760; 10 years 0.766 ± 0.033, 0.764; 15 years 0.767 ± 0.020, 0.754). Model application to children revealed highly accurate inference of the true non-smoking status (6 years of age: accuracy 0.994, N = 355; 10 years: 0.994, N = 309), suggesting prenatal and passive smoking exposure having no impact on model applications in adults. The finite set of DNA methylation markers allow accurate inference of smoking habit, with comparable accuracy as plasma cotinine use, and smoking history from blood, which we envision becoming useful in epidemiology and public health research, and in medical and forensic applications.

Keywords Epigenetics · DNA methylation · Smoking inference · Epidemiology · Forensics

Introduction

Several studies suggest that tobacco smoking impacts the human epigenome, particularly by changing DNA methylation patterns [1, 2]. DNA methylation is catalyzed by DNA methyltransferases (DNMT's); the carcinogens in cigarette smoke cause double-strand DNA breaks and the DNA repair sites recruit DNMT1 [3], which methylates cytosines in CpGs adjacent to the repaired nucleotides [4]. Nicotine was shown to down-regulate DNMT1, and mRNA and protein expression [5]. Furthermore, cigarette smoke condensate increases expression of Sp1, a transcription factor that binds to GC-rich motifs in gene promoters, preventing de novo methylation [6–9]. In recent years, various epigenome-wide

Mohsen Ghanbari and Manfred Kayser have contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10654-019-00555-w) contains supplementary material, which is available to authorized users.

✉ Mohsen Ghanbari
m.ghanbari@erasmusm.nl

✉ Manfred Kayser
m.kayser@erasmusmc.nl

Extended author information available on the last page of the article

association studies (EWASs) have provided a long list of CpGs significantly associated with tobacco smoking habits in blood [10]. Although there are strong smoking associations across the epigenome, some studies suggest that after smoking cessation, DNA methylation patterns can return back to those found in never smokers [11, 12].

Smoking is a well-known risk factor for the development of several diseases [13, 14]. Therefore, studies that investigate smoking and its effect on mortality and morbidity rely on accurate assessments of smoking exposure. These studies use mainly self-reported smoking questionnaires to collect this information, which could result in underestimation and misrepresent the degree of the true smoking exposure [15]. In particular, it is possible that specific groups of participants, for instance pregnant women, are more reluctant to confide that they smoke [16]. Hence, the ability to reliably and accurately infer a person's smoking habit from blood is relevant in epidemiology and public health research as well as in medical practice, because such an approach could complement, or even replace, self-reported smoking questionnaires.

Moreover, inference of a person's smoking habit from blood traces found at crime scenes would allow the broadening of DNA investigative intelligence beyond the currently considered parameters of appearance, bio-geographic ancestry and age, thus helping to better find unknown perpetrators of crime who are not identifiable via standard forensic DNA profiling [17]. Blood-based toxicological tests for measurement of tobacco exposure exist; however, they assess current and acute, rather than habitual, smoking [18]. In addition, biomarkers used include nicotine itself or its metabolite cotinine, and their accurate detection of current smokers is affected by their short half-lives (2–3 h vs. 15–19 h for nicotine and cotinine, respectively) and individual variation in metabolic rates [19]. Therefore, when using the cotinine-based approach false-negatives can be easily obtained, and also false-positives may occur in former smokers that use nicotine replacement therapy [20]. Given these constrains of current toxicology blood measures, and considering the recent progress in understanding the impact of smoking on epigenetic variation, we envision DNA methylation from blood as a promising approach for long-term habitual smoking behaviour.

Although progress has been made in understanding the epigenetic impact of smoking [1], only a limited number of studies have explored the inference of smoking habits from blood with DNA methylation markers, albeit with various limitations such as small sample size, limited validation, restricting to smokers and non-smokers and not considering former smokers in the model building, and/or utilizing large numbers of CpGs [21–27]. Reliable studies on the validated inference of a person's smoking habits and history from blood with a finite set of DNA methylation markers and based on statistical models with large underlying data are not available as of yet. A finite number of DNA methylation markers achieving maximal prediction accuracy would be especially beneficial for those practical applications where—due to limited DNA quality and quantity, a common problem in forensics—it is impossible to apply standard DNA methylation microarray technology [17].

With this study, we aimed to identify a robust, finite set of DNA methylation markers in blood and, based on this finite biomarker set, develop accurate, reliable and validated statistical models for inferring a person's tobacco smoking habits and history from blood, which we envision becoming useful in future epidemiology and public health research as well as medical and forensic applications.

Materials and methods

Study population

This study was embedded within the Biobank-based Integrative Omics Study (BIOS) Consortium [28], which consists of six Dutch cohorts (N = 3118), including the Rotterdam Study (RS) (N = 584) [29], Cohort on Diabetes and Atherosclerosis Maastricht (CODAM) (N = 156) [30], The Netherlands Twin Register (NTR) (N = 894) [31], Leiden Longevity Study (LLS) (N = 625) [32], Prospective ALS Study Netherlands (PAN) (N = 167) [33] and LifeLines (LL) (N = 692) [34]. Additionally, we included another 646 unrelated participants from the Rotterdam Study (RS-III-1) not included in BIOS. We externally validated our model in the Kooperative Gesundheitsforschung in der Region Augsburg (KORA) study (F4, N = 1608) [35], as well as in the Study of Health in Pomerania (SHIP)-Trend (N = 244) [36] cohort. Characteristics of all cohorts used can be found in Online Resource 1: Table S1. We additionally tested our model in samples from children included in the Generation R Study [37], in particular, we used data from children participating at birth (N = 1111), at the age of 6 years (N = 355), and at the age of 10 years (N = 309), of which 197 overlapped between all three time points, providing longitudinal data (Online Resource 1: Table S2). The smoking status information was obtained using questionnaires. The study characteristics are described in detail in Online Resource 2: Supplemental methods.

DNA methylation quantification

DNA was extracted from whole peripheral blood in all studies using standard procedures. All studies used the Illumina Infinium Human Methylation 450 K BeadChip (Illumina Inc, San Diego, CA, USA) for epigenome-wide DNA methylation measurements, except the SHIP-Trend study, which

used the more recent Infinium MethylationEPIC BeadChip (Illumina Inc, San Diego, CA, USA). DNA methylation data pre-processing for cohorts included in the BIOS consortium were conducted together via the pipeline created by Tobi et al. [38, 39]. The DNA methylation data pre-processing in the external validation cohorts and the Generation R Study were done independently. The methylation proportion of a CpG site was reported as a methylation β-value in the range of 0 (representing completely non-methylated sites) to 1 (representing completely methylated sites). Further study-specific methods can be found in Online Resource 2: Supplemental methods.

Ascertainment of smoking-associated CpGs

EWASs using the Illumina Infinium Human Methylation 27 K or 450 K BeadChip investigating smoking-induced changes in DNA methylation patterns were reviewed [2, 21, 40–50]. We excluded studies [11] that used cohorts included in our model-building dataset, to avoid over-estimation of our model. Envisioning future laboratory tool development, we only selected robust CpGs that were (1) highlighted in two or more studies, (2) with at least 10% difference in mean or median (depending on availability per EWAS) β-values between current smokers and never-smokers (or non-smokers when non-smoking data was available) in at least one of the studies, and (3) with the same direction in β-value difference between current smokers and never/non-smokers in all studies investigated.

Statistical modeling for current smoking habits

Of the total participants considered for model building ($N_{total} = 5178$), we excluded those with (1) missing data for smoking habits (1206 participants), (2) missing β-values for the predictive CpGs (82 participants), or (3) extreme outliers for one or more CpGs (mean ± 4 SD) (126 participants). In the end, we included 3764 participants in the final model building set, who were then categorized based on their smoking habits as (1) current smokers or (2) former and never smokers combined. The association between the candidate CpGs and smoking habits (smokers vs. non-smokers) was replicated in our model building dataset using binomial regression analysis adjusted for age and sex using the "glm" function with "binomial" as family and "logit" as link. To identify the most informative set of DNA methylation predictors from the candidate CpGs, the association between the complete set of predictive CpGs and smoking habits was assessed in a binary logistic regression analysis, using the "glm" function with "binomial" as family and "logit" as link. Backward elimination procedures were used for the marker selection process. We excluded the CpGs one by one based on their absolute z-statistic per regression (calculated by dividing the regression coefficient by its standard error) assessed using the "VarImp" function (r-package "caret"). The predictive CpG with the lowest absolute z-statistic in the regression was removed. The model was applied to the dataset with the "predict" function (type = "response") and the confusion matrix (r-package "caret") was conducted using a probability threshold of 0.5. The prediction performance of the model was additionally assessed using "prediction" and "performance" (r-package "ROCR"), the Area Under the Curve (AUC) per model was calculated (r-package "ROCR") and a cumulative AUC profile was conducted for each model to obtain a cumulative AUC profile. We selected the best-fit prediction model using a combination of the backward elimination approach and the Chi squared test. In particular, we compared the model including all CpGs ($model_{FULL}$) with the model excluding one CpGs, ($model_{FULL\text{-}1CpG}$), this model $_{FULL\text{-}1CpG}$ was then compared with the model excluding another CpG ($model_{FULL\text{-}2CpGs}$), following the same order as conducted via the backward approach, and so on until we noticed a statistically significant difference between two models in the backward approach. Subsequently, we tested the inclusion of age, sex and cell counts to the final model.

Former smokers as additional category

Participants included in the model building dataset (N = 3764) without additional smoking data, including the age someone stopped smoking (former smokers) or the age someone started smoking or the number of cigarettes someone smokes per day (current smokers), were excluded, resulting in a dataset including 2939 participants. The association between the previously selected predictive CpGs and the three smoking categories was assessed in a multinomial regression analysis, using the "multinom" function (r-package "nnet"). We predicted the smoking categories using the "predict" function (type = "class") and the confusion matrix (r-package "caret") was conducted. The AUC per category was conducted using the "predict" function (type = "probs") and "roc" function (r-package "pROC").

Smoking cessation time inference in former smokers

In the former smokers (N = 1332), smoking cessation time was calculated as one's age minus the age one stopped smoking. The participants were split into two categories for three models. For model 1, ≥ 5 years cessation time were coded as "1" and < 5 years smoking cessation were coded as "0", for model 2, ≥ 10 years cessation time were coded as "1" and < 10 years smoking cessation were coded as "0", and for model 3, ≥ 15 years cessation time were coded as "1" and < 15 years smoking cessation were coded as "0". The predictions were conducted using the same method as described for the current versus non-smokers model.

Probability thresholds were set to 0.8733, 0.7650 and 0.6397 respectively.

Pack-year inference in current smokers

For the current smokers (N = 364) the pack-years were calculated as the number of cigarettes smoked per day divided by 20, multiplied by the total years of smoking. The participants were categorized into two categories for two models. For model 1, ≥ 15 pack-years were coded as "1" and < 15 pack-years coded as "0", for model 2, ≥ 10 pack-years were coded as "1" and < 10 pack-years coded as "0". The predictions were conducted using the same method as described for the current *vs* non-smokers model.

Pack-years (current-), smoking cessation time (former-) and never smokers

We combined the pack-year inference in current smokers with the cessation time in former and never smokers, resulting into five categories in two models (N = 2939) for inferring life-time smoking information. For model 1, the current smokers ≥ 15 pack-years were coded as "5", with < 15 pack-years were coded as "4", the former smokers ≤ 10 years smoking cessation were coded as "3", with > 10 years smoking cessation were coded as "2" and never smokers were coded as "1". In the second model the same categories were used except for the pack-years which were now divided in ≥ 10 pack-years (coded as "5") and < 10 pack-years (coded as "4"). The predictions were conducted using the same method as described for the current *vs* former *vs* never smokers model.

Internal validation of the developed prediction models

For internal validation of the developed predictive models, we adopted a fivefold cross-validation scheme [51], in which the whole dataset is first randomly distributed into five equal and non-overlapping subsets. Four of the subsets (80% of the data) are combined to form a dataset used to train the logistic regression model which is then tested by inferring the smoking habits in the remaining dataset (20% of the data). This resulted in five different training (80%) and testing (20%) sets. The model was trained in the five training sets and applied to corresponding testing sets, resulting in five logistic regression models. Subsequently, we used the bootstrap method (r-packages "boot" and "parallel") as additional internal validation to correct for potential overestimation of the prediction, since we use the same data for model building and predictions. We generated 1000 bootstrap samples, with replacement from the dataset for which we estimated the model and applied each fitted model to the original sample, resulting in 1000 AUC estimates. Thereafter, we recalculated the prediction accuracy by applying the fitted model to the bootstrap sample itself. The performance in the bootstrap sample represents an estimation of the apparent performance, and the performance in the original sample represents test performance. The difference between the average of the two conducted AUCs is a stable estimate of the optimism. We corrected for prediction overestimation by subtracting the optimism from the apparent AUC, to obtain an improved estimate of the prediction AUC [52, 53].

External validation of the developed prediction models

We externally validated our prediction models in two independent cohorts from German-European origin. The full models were validated in the KORA F4 study (N = 1608). Additionally, we externally validated our models in the SHIP-Trend study (N = 244). In this cohort, the EPIC methylation array was used which does not include all CpGs of the 450 K array. We therefore first generated the prediction models based on the overlapping CpGs in the model building dataset and subsequently externally validated them in the SHIP-Trend dataset.

Comparing performance of CpG-based model with cotinine level cut-off

We compared the outcomes of the CpG model to infer current *vs* non-smokers with the outcomes using a cotinine level cut-off of 50 ng/mL [54, 55] and applied smoking information from self-reports as reference. We employed a subset of our model building dataset (N = 488 participants included in NTR [56]) in which both DNA methylation levels and cotinine levels were available. First, participants were categorized as smokers when their plasma cotinine levels were > 50 ng/mL, or as non-smokers with cotinine levels ≤ 50 ng/mL, threshold according to previous studies including the used cotinine data [54, 55]. Second, the current versus non-smokers CpG model was applied to this subset, obtaining the inferred smoking status for the participants. Third, we compared the obtained smoking status for both models with the information obtained from the self-reported questionnaires and computed the sensitivity and specificity per model.

Application of the developed prediction model in newborns and young children

Studies have shown the impact of prenatal smoking exposure on the DNA methylation pattern of the offspring [57] and the ability of predicting maternal smoking status using these patterns [58]. In this context, we wanted to test the effect of prenatal exposure on model application in adults.

Hence, when an adult does not smoke, but was exposed to prenatal smoking, do we predict this person indeed as a true non-smoker? To test for this putative impact of exposure to prenatal smoking on epigenetic inference of smoking habits using our model, we tested our model in umbilical cord blood of newborns (N = 1111), and in whole blood of children at the ages of six (N = 355) and 10 years (N = 309). We used five different analyses to evaluate the effects of active smoking of the mothers and passive smoking of the mothers (i.e. smoking of others in the mother's home and work environment) during pregnancy on smoking habit inference using our model. In our first analysis, we did not take the smoking habits of the pregnant mothers or others in the pregnant mother's home and work environment into account and all children were coded as non-smokers. The proportion of accurately predicted cases was calculated using a probability threshold of 0.5. In each of the following analyses, we coded the children "1" if their parents met the smoking habit criteria, otherwise they were coded as "0". So, in the second analysis, only sustained maternal smoking throughout pregnancy was considered. Therefore, the children of mothers that smoked during the whole pregnancy were coded as "1". In the third analysis, we additionally included the children of mothers who stopped smoking when they realized that they were pregnant by coding these children as "1". In the fourth analysis, we additionally included smoking of the father and/or others in the mother's household/at work (> 1 h per day) during pregnancy (i.e. passive smoking). In the fifth analysis, we assessed the sole effect of passive smoking i.e., where the mother did not smoke but the father or someone else in the house or at work (> 1 h per day) smoked during the pregnancy of the mother. For 197 children, DNA methylation levels were measured at all three time points, i.e. birth, 6 years of age and 10 years of age; hence, we repeated the previous models again in these children to allow a direct comparison of the findings at these three time points in the same individuals.

Results

Ascertaining candidate DNA methylation markers for inferring smoking habits from blood

We inspected 14 published EWASs on tobacco smoking habits (N_{total} = 7015) [2, 21, 40–50] to identify smoking-associated CpGs as candidate DNA methylation markers for prediction modeling of smoking habits. CpGs were selected as candidate prediction markers if they met three criteria as mentioned in the method section. This procedure highlighted 20 top smoking-associated CpGs as candidate markers used for further analyses (Table 1). The differences in β-values between smokers and never-/non-smokers reported previously for these 20 top smoking-associated CpGs are illustrated in Fig. 1.

Building CpG-based models for inferring smoking habit and history from blood

Following the replication of the association between the CpGs and smoking habits (smokers vs. non-smokers) after adjusting for age and sex (Online source Table 3), we assessed the predictive effect of the selected 20 candidate markers in the model building dataset (N = 3764). Starting with a model including all 20 CpGs, the CpG with the lowest z-value per model was sequentially removed, and the AUC was calculated for each model to obtain a cumulative AUC profile (Table 1; Fig. 2).

To identify the minimal number of CpGs required to achieve maximum prediction accuracy, we additionally used Chi squared tests. Applying this backward approach, the first significant difference between two models was noted when we compared the model with and without cg09935388 (Table 1; Fig. 2). The combined marker elimination approach resulted in a finite set of DNA methylation markers comprising 13 CpGs (Table 1; Fig. 2). The AUC for the identified 13-CpG model was 0.901 for distinguishing between smokers versus non-smokers (for other prediction accuracy measures, see Table 2). The remaining 7 CpGs raised the cumulative AUC only on the 4th decimal i.e. from 0.9010 to 0.9016 (Table 1; Fig. 2). Hence, this finite set of 13 CpGs was used for subsequent prediction modeling. Using the 13-CpG model, we inferred the smoking status of the participants included in our model building dataset; the inferred probabilities are presented in a histogram in Fig. 3, where each probability bin is overlaid with the percentage of accurately inferred smoking habits in that probability range.

Adjusting the prediction model for age resulted in a minor AUC increase from 0.901 to 0.907, adjusting for sex from 0.901 to 0.903 and including both age and sex in the model increased the AUC slightly from 0.901 to 0.911 (Online Resource 1: Table S4). Additionally, we tested the influence of cell counts on the model accuracy. In the subset of participants for which cell count measures were available (N = 3402), our 13-CpG model without cell counts achieved an AUC of 0.906. Including the cell count measurements for monocytes, granulocytes and lymphocytes in our 13-CpG model, the AUC was almost identical at 0.907 (Online Resource 1: Table S5). Since age, sex and cell counts only had a minor impact on the prediction accuracy, these three non-epigenetic factors were not considered in the final model used in the subsequent analyses.

Next, we considered former smokers as an additional, separate category in the prediction model building based on the finite set of 13 CpGs, resulting in a three-category prediction model. To this end, we considered a subset of

Table 1 Top 20 smoking-associated CpGs from 14 previous EWASs considered here for marker sub-selection and their contribution to smoking inference from blood

CpG ID	Chr:position[b]	Gene ID[c]	Location of CpG	Cumulative AUC
cg05575921[a]	5:373,378	*AHRR*	Gene body	0.8801
cg13039251[a]	5:32,018,601	*PDZD2*	Gene body	0.8888
cg03636183[a]	19:17,000,585	*F2RL3*	Gene body	0.8883
cg12803068[a]	7:45,002,919	*MYO1G*	Gene body	0.8889
cg22132788[a]	7:45,002,486	*MYO1G*	Gene body	0.8934
cg06126421[a]	6:30,720,080	NA	–	0.8929
cg21566642[a]	2:233,284,661	NA	–	0.8957
cg23576855[a]	5:373,299	*AHRR*	Gene body	0.8967
cg15693572[a]	3:22,412,385	NA	–	0.8982
cg05951221[a]	2:233,284,402	NA	–	0.8989
cg01940273[a]	2:233,284,934	NA	–	0.8998
cg12876356[a]	1:92,946,825	*GFI1*	Gene body	0.9005
cg09935388[a]	1:92,947,588	*GFI1*	Gene body	0.9010
cg19572487	17:38,476,024	*RARA*	5′UTR	0.9012
cg19859270	3:98,251,294	*GPR15*	Gene body (1st Exon)	0.9015
cg18146737	1:92,946,700	*GFI1*	Gene body	0.9015
cg21161138	5:399,360	*AHRR*	Gene body	0.9015
cg23480021	3:22,412,746	NA	–	0.9016
cg21188533	3:53,700,263	*CACNA1D*	Gene body	0.9015
cg03274391	3:22,413,232	NA	–	0.9015

NA not annotated to any gene according to the Illumina Infinium Human Methylation 450 K annotation file

AUC Area under the curve

[a]CpGs included in our final 13 CpG-model

[b]Genome coordinates provided by Illumina (GRCh37/hg19)

[c]According to the Illumina Infinium Human Methylation 450 K annotation file

2939 participants for which the relevant smoking habit information was available. We obtained for the current smokers (N = 364) an AUC of 0.928, for the former smokers (N = 1332) 0.772, and for the never smokers (N = 1243) 0.835 (for other accuracy measures, see Table 3). Additionally, we calculated smoking cessation time for the former smokers (N = 1332), and used the 13-CpGs to infer smoking cessation for ≥ 5 years (N = 1160) versus < 5 years (N = 172), which resulted in an AUC of 0.793, for ≥ 10 versus < 10 years smoking cessation time (N = 1028 and N = 304, respectively) an AUC of 0.778 was obtained and for ≥ 15 versus < 15 years smoking cessation time (N = 887 and N = 445, respectively) an AUC of 0.779 was obtained (Table 4).

Furthermore, for the current smokers (N = 364) we calculated the pack-years (see methods) and used the 13 CpG markers to infer pack-years for ≥ 15 pack-years (N = 210) versus < 15 pack-years (N = 154), which resulted in an AUC of 0.815. For ≥ 10 versus < 10 pack-years (N = 246 and N = 118, respectively) an AUC of 0.846 was obtained (Table 5).

Finally, we combined the pack-years in current smokers, smoking cessation in former smokers with the never smokers (N = 2939) into one model for life-time smoking information inferring. We obtained for the current smokers with ≥ 15 pack-years (N = 210) an AUC of 0.949, < 15 pack-years (N = 154) an AUC of 0.869, in former smokers with ≤ 10 years smoking cessation (N = 311) an AUC of 0.793, with > 10 years smoking cessation (N = 1021) an AUC of 0.739 and the never smokers (N = 1243) an AUC of 0.835 (Table 6). We obtained for the current smokers with ≥ 10 pack-years (N = 246) an AUC of 0.948, < 10 pack-years (N = 118) an AUC of 0.863, former smokers with ≤ 10 years smoking cessation (N = 311) an AUC of 0.794, with > 10 years smoking cessation (N = 1021) an AUC of 0.739, and the never smokers (N = 1243) an AUC of 0.835 (Table 6).

Validating CpG-based models for inferring smoking habit and history from blood

We validated the newly developed prediction models based on the 13 selected CpGs via both internal and external validation procedures. Internal validation was carried out in the model building set using fivefold cross-validation and bootstrapping. For the two-category model (smokers vs. non-smokers), the optimism from bootstrap internal validation was 0.0032, resulting in a bootstrap-adjusted AUC

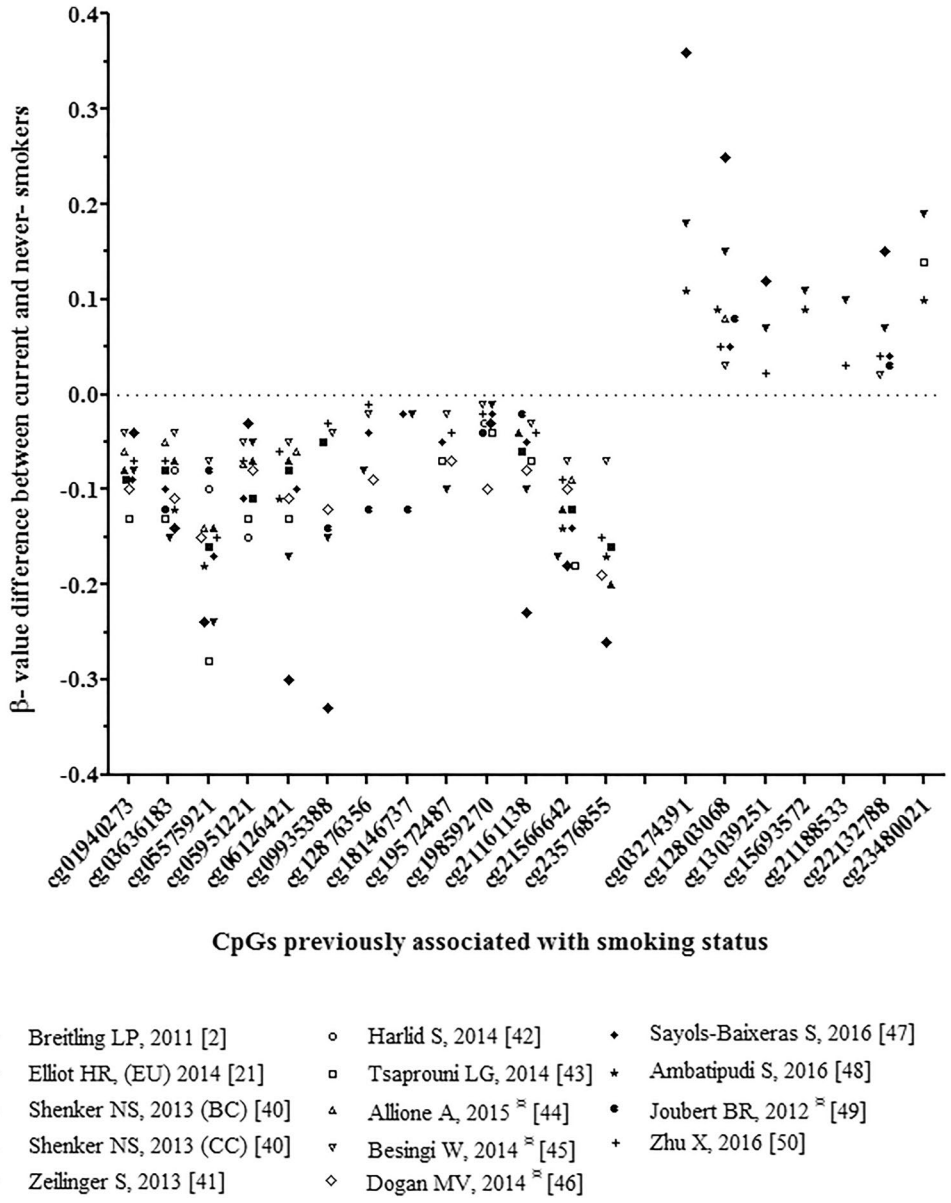

Fig. 1 DNA methylation β-value differences between smokers and never-smokers for the top 20 smoking-associated CpGs. Previously reported differences in β-values in mean or median (depending on availability per EWAS) between smokers and never-smokers (¤ or non-smokers, when non-smoking data was available) for the selected 20 top-associated CpGs obtained from the 14 reviewed EWASs on smoking habits that did not include samples used here for model building

of 0.898 (0.901–0.0032), see Table 2 for other accuracy measures and cross-validation results. For the three-category model (smokers vs. former smokers vs. never smokers) the bootstrap conducted optimisms are 0.0032 for current smokers, 0.0063 for former smokers and 0.0036 for never smokers resulting in bootstrap adjusted AUCs of 0.925 (0.928–0.0032) for current smokers, 0.766 (0.772–0.0063) for former smokers and 0.831 (0.835–0.0036) for never smokers (Table 3). For the smoking cessation time inference in former smoker, (1) for ≥ 5 versus < 5 years smoking cessation the bootstrap optimism was 0.0170 resulting in a bootstrap-adjusted AUC of 0.776 (0.793–0.0170); (2) for ≥ 10 versus < 10 years smoking cessation the bootstrap resulted in an optimism of 0.0112, giving a bootstrap-adjusted AUC of 0.767 (0.778–0.0112); (3) ≥ 15 versus < 15 years smoking cessation the bootstrap resulted in an optimism of 0.0096, giving a bootstrap-adjusted AUC of 0.769 (0.779–0.0096) (Table 4). For the two pack-year models, (1) the bootstrap optimism for ≥ 15 versus < 15 pack—was 0.029 resulting in a bootstrap-adjusted AUC of 0.786 (0.815–0.029); and (2) for ≥ 10 versus < 10 pack-years the bootstrap resulted in an optimism of 0.026, giving a bootstrap-adjusted AUC of 0.820 (0.846–0.026) (Table 5). Finally, for the life-time smoking information inferring, we obtained for ≥ 15 pack-years a bootstrap optimism of 0.0034 resulting in a bootstrap-adjusted AUC of 0.946 (0.949–0.0034), for < 15 pack-years a bootstrap-adjusted AUC of 0.860 (0.869–0.0091), for ≤ 10 smoking cessation a bootstrap-adjusted AUC of 0.782 (0.793–0.0106), > 10 years smoking cessation a bootstrap optimism of 0.0075 resulting in a bootstrap-adjusted AUC

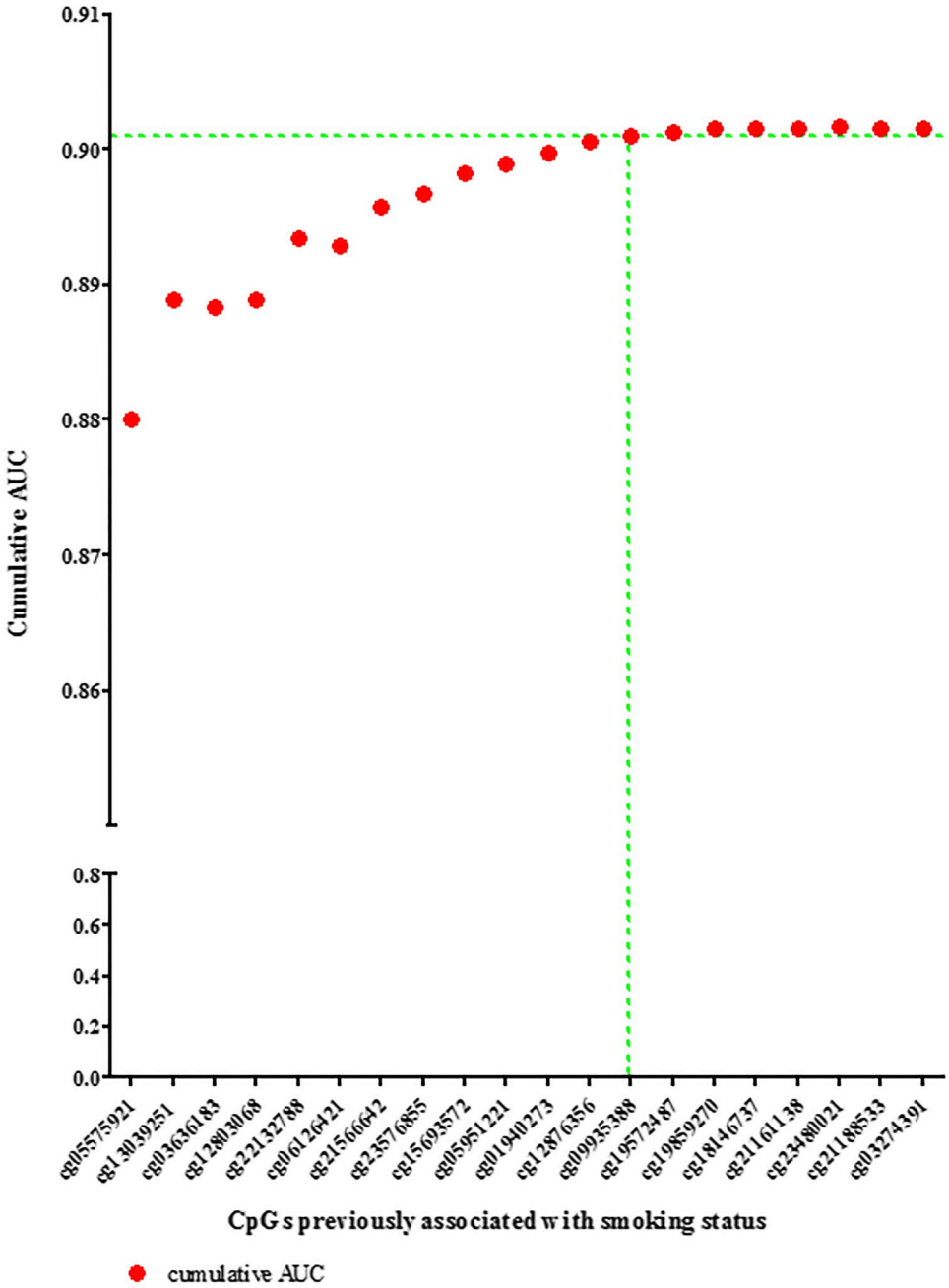

Fig. 2 Cumulative AUC profile for smoking habit inference from blood based on the top 20 CpGs. The 20 CpGs were selected from previous EWASs on smoking habits (see Fig. 1) and were tested in the model-building set (N = 3764). Presented is the cumulative contribution of each of the selected 20 CpGs to the model-based smoking habit inference, shown as the AUC plotted against the number of CpGs included in the binary logistic regression model. In the model selection process, first all CpGs were included, and using backward elimination procedures, those with the lowest z-statistic per model were removed one by one. After 13 CpGs, the AUC plateaus; therefore, and by considering the results from Chi squared testing, these 13 CpGs were used for further analyses

of 0.732 (0.739–0.0075) and for never smokers a bootstrap-adjusted AUC of 0.831 (0.835–0.0037) (Table 6). For the second five-category model, very similar results were obtained (Table 6).

External validation was performed in independent samples of two population-based studies, KORA and SHIP-Trend. In KORA (F4, N = 1608), an AUC of 0.911 was achieved for the full 13-CpG two-category model (Table 2). In SHIP-Trend (N = 244), an AUC of 0.888 was obtained for the two-category model based on a subset of ten CpGs, since the EPIC-array applied for SHIP-Trend is missing three of the 13 CpGs (cg06126421, cg22132788 and cg05951221). This 10-CpG model in the model building set gave a cross-validated average AUC of 0.893 ± 0.012 (Table 2). External validation of the three-category model in the KORA study (F4, N = 1608) achieved an AUC of 0.914 for the current smokers (N = 226), 0.699 for the former smokers (N = 707), and 0.781 for the never smokers (N = 675) (Table 3). The three-category model validation in SHIP-Trend for the 10-CpG model resulted in an AUC of 0.882 for current smokers (N = 51), 0.654 for former smokers (N = 92), and 0.778 for never smokers (N = 101) (Table 3). For comparison, in the model building set, this three category 10-CpG model

Table 2 Outcomes of the two-category-model (smokers vs. non-smokers) for inferring smoking habits from blood based on CpGs

	13-CpG model			10-CpG model[a]		
	Model building data set (N = 3764)		External validation	Model building data set (N = 3764)		External validation
	Model building	Fivefold cross-validation	KORA (N = 1608)	Model building	Fivefold cross-validation	SHIP-Trend (N = 244)
Accuracy[b] (95% CI) ± SD	0.923 (0.914, 0.931)	0.921 ± 0.008	0.926 (0.912, 0.938)	0.917 (0.908, 0.926)	0.917 ± 0.011	0.873 (0.825, 0.912)
Specificity	0.976	0.976 ± 0.005	0.983	0.975	0.975 ± 0.006	0.995
Sensitivity	0.585	0.577 ± 0.044	0.580	0.548	0.551 ± 0.042	0.412
AUC	0.901	0.897 ± 0.137	0.911	0.896	0.893 ± 0.012	0.888

Cross-validation analysis results are presented as mean ± standard deviation

AUC Area under the curve

[a]Three CpGs (cg06126421, cg22132788 and cg05951221) are not included in the EPIC methylation microarray dataset from SHIP-Trend, this model is included here to demonstrate a second external validation in SHIP next to KORA with the full 13-CpG model

[b]Proportion accurately inferred smoking habits, 95% confidence interval (CI)

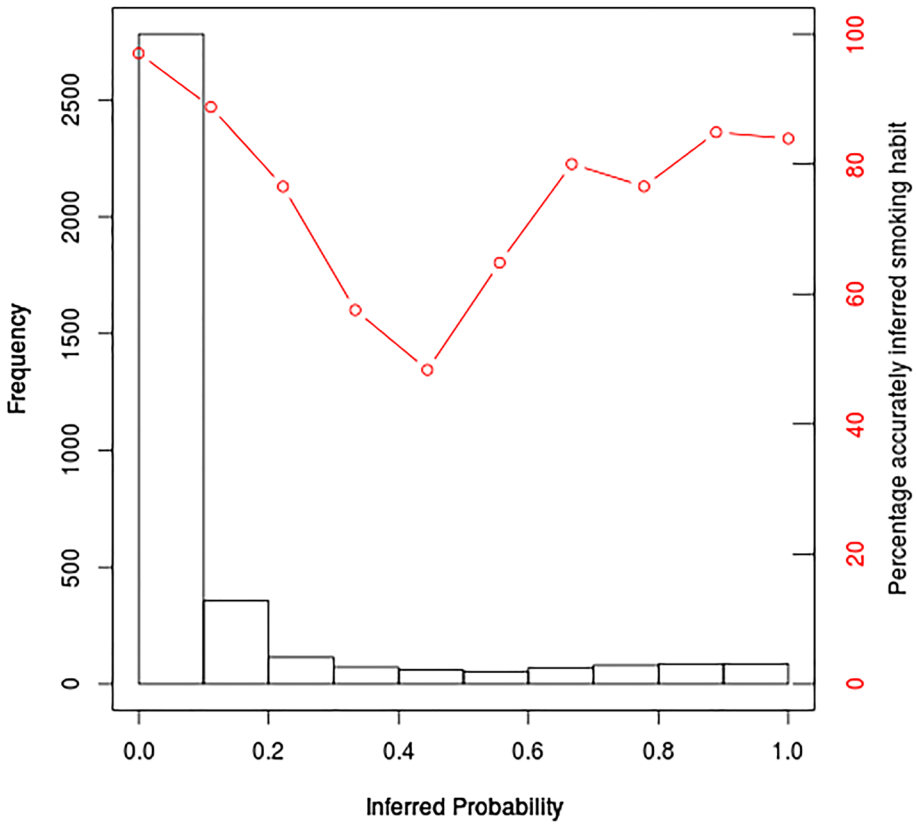

Fig. 3 Inferred probability of being a smoker versus the percentage of correctly inferred smoking habits. Histogram of predicted probabilities in our model building dataset (N = 3764), probabilities determined using the 13 CpGs included in the final prediction model. The y-axis presents the number of individuals for whom the predicted probability of being a smoker was within the given probability range (x-axis). The red dots present the percentage of individuals in each probability bin that were accurately inferred using a > 0.5 probability threshold for being a smoker

gave a cross-validated average AUC of 0.919 ± 0.019 for current smokers, 0.748 ± 0.023 for former smokers, and 0.823 ± 0.018 for never smokers (Table 3). External validation of smoking cessation time inference in former smokers in the KORA study (N = 652) resulted in an AUC of 0.760 for ≥ 5 versus < 5 years of smoking cessation time, an AUC of 0.764 for ≥ 10 versus < 10 years of smoking cessation time, and of 0.754 for ≥ 15 versus < 15 years of smoking cessation time (Table 4). Furthermore, we externally validated the prediction of pack-years in the current smokers of the KORA study (F4, N = 224) and obtained an AUC of 0.752 for inferring ≥ 15 versus < 15 pack-years and an AUC of 0.796 for ≥ 10 versus < 10 pack-years (Table 5). The pack-year validation in the

Table 3 Outcomes of the three-category-model (current smokers vs. former smokers vs. never smokers) for inferring smoking habits from blood based on CpGs

Model building data set (N = 2939): model building 13-CpG model	Never (N = 1243)	Former (N = 1332)	Current (N = 364)
Specificity	0.746	0.770	0.997
Sensitivity	0.780	0.652	0.668
AUC	0.835	0.772	0.928
Fivefold cross-validation			
Specificity	0.739 ± 0.017	0.766 ± 0.053	0.975 ± 0.008
Sensitivity	0.769 ± 0.060	0.643 ± 0.039	0.669 ± 0.056
AUC	0.830 ± 0.019	0.766 ± 0.023	0.925 ± 0.021
External replication in KORA (N = 1608): 13-CpG model	Never (N = 675)	Former (N = 707)	Current (N = 226)
Specificity	0.539	0.870	0.980
Sensitivity	0.916	0.392	0.615
AUC	0.781	0.699	0.914
Model building data set (N = 2939): model building 10-CpG model[a]	Never (N = 1243)	Former (N = 1332)	Current (N = 364)
Specificity	0.749	0.737	0.974
Sensitivity	0.751	0.648	0.626
AUC	0.825	0.753	0.922
Fivefold cross-validation			
Specificity	0.745 ± 0.013	0.735 ± 0.042	0.975 ± 0.010
Sensitivity	0.747 ± 0.050	0.645 ± 0.026	0.627 ± 0.025
AUC	0.823 ± 0.018	0.748 ± 0.023	0.919 ± 0.019
External replication in SHIP-Trend (N = 244): 10-CpG model[a]	Never (N = 101)	Former (N = 92)	Current (N = 51)
Specificity	0.490	0.822	0.990
Sensitivity	0.891	0.315	0.451
AUC	0.778	0.654	0.882

Cross-validation analysis results are presented as mean ± standard deviation

AUC Area under the Curve

[a]Three CpGs (cg06126421, cg22132788 and cg05951221) are not included in the EPIC methylation microarray dataset from SHIP-Trend

current smokers of SHIP-Trend (N = 41) for the 10-CpG model resulted in an AUC of 0.779 for ≥ 15 versus < 15 pack-years (AUC of 0.757 ± 0.077 in the model building set) and an AUC of 0.837 for ≥ 10 versus < 10 pack-years (AUC of 0.809 ± 0.039 in the model building) (Table 5). The external validation of the five-category models in the KORA study resulted for the current smokers with ≥ 15 pack-years in an AUC of 0.955, for < 15 pack-years an AUC of 0.710, for ≤ 10 years smoking cessation an AUC of 0.791, > 10 years smoking cessation an AUC of 0.650 and for never smokers an AUC of 0.788. For the second five-category model, we obtained in the KORA study an AUC of 0.943 for ≥ 10 pack-years, of 0.694 for < 15 pack-years, an AUC of 0.791 for ≤ 10 years smoking cessation, of 0.651 ≥ 10 years smoking cessation and an AUC of 0.788 for never smokers (Table 6).

Comparing CpG-based with cotinine-based inference of smoking habit

In a subset of 488 participants for which we had CpG, cotinine and smoking information available, we compared our validated CpG-based prediction model for current versus non-smokers with the use of a cotinine cut-off to determine current smoking, using smoking information from self-reported questionnaires as reference. Using our CpG-model, we accurately inferred 87 of the 140 smokers and 344 of the 348 non-smokers (sensitivity of 0.621 and specificity of 0.989) compared to 105 of the 140 smokers

Table 4 Outcomes of the two-category models for inferring smoking history (years of cessation time) in former smokers from blood based on 13 CpGs

	Former < 5 year versus Former ≥ 5 year cessation time			Former < 10 year versus Former ≥ 10 year cessation time			Former < 15 year versus Former ≥ 15 year cessation time		
	Model building data set (N = 1332)		External validation	Model building data set (N = 1332)		External validation	Model building data set (N = 1332)		External validation
	Model building	Fivefold cross-validation	KORA (N = 652)	Model building	Fivefold cross-validation	KORA (N = 652)	Model building	Fivefold cross-validation	KORA (N = 652)
Accuracy[a] (95% CI) ± SD	0.725 (0.700, 0.749)	0.715 ± 0.020	0.830 (0.799, 0.858)	0.730 (0.705, 0.753)	0.721 ± 0.029	0.799 (0.766, 0.829)	0.732 (0.707, 0.756)	0.718 ± 0.016	0.759
Specificity	0.715	0.691 ± 0.090	0.494	0.694	0.682 ± 0.063	0.471	0.663	0.644 ± 0.033	0.449
Sensitivity	0.727	0.718 ± 0.026	0.879	0.740	0.733 ± 0.026	0.900	0.767	0.756 ± 0.015	0.902
AUC	0.793	0.774 ± 0.024	0.760	0.778	0.766 ± 0.033	0.764	0.779	0.767 ± 0.020	0.754

Cross-validation analysis results are presented as mean ± standard deviation

AUC Area under the curve

[a]Proportion accurately inferred smoking habits, 95% confidence interval (CI)

Table 5 Outcomes of model applications to infer smoking history (pack-years) in current smokers (N = 364) from blood based on CpGs

	13-CpG model			10-CpG model[a]		
	Model Building (N = 364)	Fivefold Cross-validation	KORA F4 (N = 224)	Model Building (N = 364)	Fivefold Cross-validation	SHIP-Trend (N = 41)
More or less than 10 pack-years						
Accuracy (95% CI)[b]	0.824 (0.781, 0.862)	0.783 ± 0.05	0.813 (0.755, 0.861)	0.808 (0.76, 0.847)	0.770 ± 0.035	0.805 (0.651, 0.912)
Specificity	0.644	0.577 ± 0.131	0.343	0.602	0.548 ± 0.14	0.778
Sensitivity	0.911	0.882 ± 0.045	0.899	0.907	0.879 ± 0.046	0.813
AUC	0.846	0.800 ± 0.068	0.796	0.834	0.809 ± 0.039	0.837
More or less than 15 pack-years						
Accuracy (95% CI)[b]	0.733 (0.685, 0.778)	0.719 ± 0.093	0.786 (0.726, 0.838)	0.728 (0.679, 0.773)	0.709 ± 0.059	0.659 (0.494, 0.799)
Specificity	0.617	0.600 ± 0.204	0.455	0.597	0.575 ± 0.143	0.533
Sensitivity	0.819	0.805 ± 0.042	0.894	0.824	0.808 ± 0.035	0.731
AUC	0.815	0.767 ± 0.102	0.752	0.786	0.757 ± 0.077	0.779

Cross-validation analysis results are presented as mean ± standard deviation

Pack-years were calculated as the number of cigarettes smoked per day divided by 20, multiplied by the total years of smoking

[a]Three CpGs (cg06126421, cg22132788 and cg05951221) are not included in the EPIC methylation microarray dataset from SHIP-Trend

[b]Proportion accurately inferred smoking habits; 95% CI, confidence interval; AUC, Area under the Curve

and 342 of the 348 non-smokers using the cotinine level cut-off of 50 ng/mL (sensitivity of 0.750 and specificity of 0.983). Out of the 87 accurately inferred smokers with our CpG model, 75 (86%) were also accurately selected as smokers based on cotinine, and out of the 105 participants correctly selected with cotinine as smokers, 75 (71%) were accurately inferred as smokers with our CpG model. For the non-smokers, out of the 344 accurately inferred with our CpG model, 340 (99%) were also selected with cotinine as non-smokers, and 340 (99%) out of the 342 accurately selected non-smokers with cotinine, were accurately inferred as non-smokers with our CpG model. Finally, when comparing all three methods(questionnaires/cotinine levels/DNA methylation prediction), 340 participants were highlighted as non-smokers and 75 as smokers with all three methods, 12 were selected as smokers based on questionnaires and DNA

Table 6 Outcomes of the five-category-model for inferring smoking habits and smoking history from blood based on 13 CpGs

Never versus former > 10 years cessation time versus former ≤ 10 years cessation time versus < 15 pack-years versus ≥ 15 pack-years					
Model building data set (N = 2939)	Never (N = 1243)	F > 10 year (N = 1021)	F ≤ 10 year (N = 311)	< 15PY (N = 154)	≥ 15PY (N = 210)
Specificity	0.712	0.777	0.979	0.987	0.967
Sensitivity	0.817	0.554	0.206	0.299	0.724
AUC	0.835	0.739	0.793	0.869	0.949
Fivefold cross-validation					
Specificity	0.711 ± 0.022	0.775 ± 0.036	0.977 ± 0.009	0.984 ± 0.009	0.963 ± 0.014
Sensitivity	0.809 ± 0.047	0.545 ± 0.040	0.199 ± 0.042	0.274 ± 0.128	0.695 ± 0.064
AUC	0.832 ± 0.014	0.731 ± 0.026	0.779 ± 0.018	0.855 ± 0.046	0.947 ± 0.016
External replication in KORA (N = 1551)	Never (N = 675)	F > 10 year (N = 488)	F ≤ 10 year (N = 164)	< 15 PY (N = 55)	≥ 15PY (N = 169)
Specificity	0.534	0.830	0.994	0.994	0.979
Sensitivity	0.927	0.299	0.122	0.018	0.728
AUC	0.788	0.650	0.791	0.710	0.955
Never versus former > 10 years cessation versus former ≤ 10 years cessation versus < 10 pack-years versus ≥ 10 pack-years					
Model building data set (N = 2939)	Never (N = 1243)	F > 10 year (N = 1021)	F ≤ 10 year (N = 311)	< 10 PY (N = 118)	≥ 10PY (N = 246)
Specificity	0.714	0.776	0.981	0.994	0.963
Sensitivity	0.817	0.554	0.193	0.220	0.772
AUC	0.835	0.739	0.794	0.863	0.948
Fivefold cross-validation					
Specificity	0.709 ± 0.023	0.774 ± 0.034	0.980 ± 0.006	0.992 ± 0.003	0.960 ± 0.008
Sensitivity	0.808 ± 0.045	0.542 ± 0.042	0.194 ± 0.043	0.206 ± 0.066	0.758 ± 0.067
AUC	0.831 ± 0.014	0.730 ± 0.027	0.780 ± 0.018	0.847 ± 0.047	0.946 ± 0.023
External replication in KORA (N = 1551)	Never (N = 675)	F > 10 year (N = 488)	F ≤ 10 year (N = 164)	< 10 PY (N = 35)	≥ 10PY (N = 189)
Specificity	0.535	0.827	0.994	0.998	0.977
Sensitivity	0.926	0.299	0.110	0.000	0.683
AUC	0.788	0.651	0.791	0.694	0.943

Cross-validation analysis results are presented as mean ± standard deviation

AUC area under the curve, *F* former smokers in years cessation time, *PY* pack-years

methylation inference, 30 as smokers with both questionnaires and cotinine, 2 were determined as smokers with both cotinine and DNA methylation inference, whereas 23 were determined as smokers with questionnaires only, 2 as smokers with DNA methylation inference only, and 4 as smokers with cotinine only.

Investigating prenatal smoking exposure effects on CpG-based inference of smoking habit

Next, we investigated the putative effect of prenatal smoking exposure and passive smoking on the epigenetic inference of smoking habits achievable with our validated model. When applying our model to the DNA methylation data at time of birth collected from cord blood, the proportion of children accurately inferred as non-smokers was surprisingly low at 0.114 (N = 1111) (Online Resource 1: Table S6). We then classified children whose mothers smoked throughout pregnancy as "smokers", and obtained an AUC of 0.773, with a high sensitivity of 0.981 and a low specificity of 0.131. The AUC decreased to 0.664 when additionally considering mothers who stopped smoking when they became aware of their pregnancy (generally in the first trimester), and decreased further to 0.591 when additionally considering

passive smoking of the mother during pregnancy; assessing the latter solely, an AUC of 0.460 was obtained, reflecting random prediction.

Additionally, we applied our model to data of children from the Generation R Study obtained from blood collected at the ages of six (N = 355) and ten (N = 309) years. In contrast to the results for newborns obtained from cord blood, we found that the proportion of 6- and 10-year-old children accurately inferred as non-smokers with our model was very high at 0.994 for both age groups (Table 7). This suggests no impact of prenatal smoking exposure nor passive smoking exposure during early childhood on the model performance. Subsequently, we applied our model to those 197 children for which epigenetic data were available from serial samples collected at birth, 6, and 10 years of age. The proportion of children that with our model accurately inferred as non-smokers at birth was 0.112, whereas it was 0.994 at six and 0.995 at 10 years of age, which was highly similar to the results obtained from the total datasets available for these three time points. The β-values per CpG for the model building set and the three time points in Generation R are shown in Online Resource 3: Figures S1–15.

Discussion

In this study, we introduce a robust, finite set of DNA methylation markers and carefully validated statistical models based on reasonably large population-based data, which together allow accurate and reliable inference of a person's tobacco smoking habit and history from blood DNA.

Previous studies have identified numerous CpGs associated with tobacco smoking in blood, and showed that DNA methylation patterns of specific genes are modified by smoking habits [2, 21, 40–50]; here we took advantage of these EWASs as a marker discovery resource. From the 20 top smoking-associated CpGs consistently highlighted in previous EWASs and by using new population-based cohort data not overlapping with these previous EWASs, we identified a robust, finite set of 13 CpG markers as being most suitable for inferring a person's smoking habit from blood DNA. Eight of these 13 CpGs are annotated to five known genes i.e., *AHRR* (2 CpGs), *GFI1* (2), *MYO1G* (2), *F2RL3* (1) and *PDZD2* (1), while the remaining 5 CpGs are not annotated to any coding regions. The highest AUC (0.880) for a given CpG among the 13 biomarkers in the model was achieved for cg05575921, which, together with one other CpG in the model (cg23576855), is located in the *AHRR* gene. The *AHRR* gene was shown to interact with the aryl hydrocarbon receptor (AHR), the induction point for the xenobiotic pathway, which includes several P450 enzymes, and is responsible for the degradation of environmental toxins [59–61]. Notably, *AHRR* provides the strongest epigenetic response to tobacco smoking known today [59, 62].

While a few previous studies have investigated DNA methylation markers for inferring smoking habits from blood, they all suffered from one or more limitations, including small sample size, limited model validation, exclusion of the former smoker category from the prediction model building, using a large number of CpGs and others [21–26]. For example, Philibert et al. [23] reported on the performance of five CpGs yielding AUCs 0.86–0.99 but only using 61 subjects. Notably, all five CpGs were among the 20 markers investigated in our study and are also included in our final 13-CpG model. For cg05575921, Philibert et al. estimated an AUC of 0.99 [23]; when testing this DNA methylation marker in our model building set of 3764 samples, a considerably lower AUC of 0.8801 was achieved. In another study, Elliot et al. [21] reported a methylation score based on 183 CpGs to distinguish between current, former and never smokers, with a sensitivity of 100% and a specificity of 97% using 96 subjects only. When generating the

Table 7 Model application to children from the Generation R study at 6 and 10 years of age

	Six years old Whole dataset (N = 355)	Six years old Serial samples (N = 197)	Ten years old Whole dataset (N = 309)	Ten years old Serial samples (N = 197)
Child non-smoking (all "0")				
Accuracy[a]	0.994	0.994	0.994	0.995
Sustained prenatal smoking of mother throughout pregnancy				
N	0:309	0:173	0:274	0:173
	1:46	1:24	1:35	1:24
Specificity	0.997	0.994	0.993	0.994
Sensitivity	0.022	0.0	0.0	0.0
AUC	0.649	0.650	0.606	0.592

AUC area under the curve

[a]Proportion of children correctly predicted as non-smokers

methylation score using the methods described by Elliot et al., and applying it to our model building set (N = 3764), we obtained a specificity of 0.864 and sensitivity of 0.747 with an AUC of 0.806, considerably lower than reported by Elliot et al. These two examples illustrate that previously reported prediction accuracies obtained from studies using small sample size likely reflect overestimation caused by small sample size. Given the relatively larger sample size for model building and internal validation, and for external validation with independent samples as utilized here, our results demonstrate that the new 13-CpG model introduced here provides more robust and reliable accuracy outcomes than previously reported models.

Previous studies have shown that DNA methylation patterns can be altered by age, sex and various lifestyle factors other than tobacco smoking [63, 64]. Additionally, recent papers suggest that the change in DNA methylation measurements due to smoking are mainly caused by the smoking induced changes in cell types [65–68]. We therefore tested the impact of age, sex and cell counts on the model performance and found that these covariates only provide a slight increase in the prediction accuracy our model provides. Notably, a model that does not consider sex, age and cell counts is beneficial for those applications where (some of) this information is not easily available, such as in forensics.

A recent study reported that the DNA methylation of most CpGs returns to never smoker levels within 5 years of smoking cessation, while some do not go back completely [11]. Also, previous work demonstrated that there is an association between smoking cessation time and smoking pack-years with DNA methylation scores [65, 69]. We therefore tested to what degree the 13 selected CpGs can distinguish former smokers from current smokers and never-smokers, and how well they allow inferring smoking history such as smoking cessation time and pack-years. Our results demonstrate that our 3-category model allows as first the inference of the former smoking category (smoking cessation between 0.1 and 58.86 years) together with current smokers and never smokers and also a more in depth inference possibility for cessation time categories as of more versus less than 5, 10 and 15 years of smoking cessation, although not as accurately as current and never smokers, as may be expected. The 13 CpGs also allowed accurate prediction of the pack-years in current smokers with a high AUCs for distinguishing between more or less than 10 pack-years, and for distinguishing between more or less than 15 pack-years. Finally, we show, to the best of our knowledge, for the first time an inference model able of inferring life-time smoking information in one model including the never smokers, cessation time in former smokers and pack-years in current smokers. Thus, the finite set of 13 DNA methylation markers and models we introduce here not only allow inferring information on current smoking or non-smoking status, but additionally provide information on former smoking and cessation time, smoking intensity in current smokers, and can additionally, as the first model to date, also provide complete life-time smoking information as of five different smoking categories.

Cotinine is the primary metabolite of nicotine and is therefore used as a reliable measurement for current smoking [19]. However, due to the short half-live of cotinine (between 15 and 19 h), a false-negative prediction of current smoking can be easily obtained when there is a long time between the last cigarette and blood drawn [19]. In addition, former smokers that use nicotine replacement therapy to reduce the motivation to smoke and for nicotine withdrawal symptoms, might result in false-positive predictions since cotinine, nicotine's metabolite, will still be traceable [20, 70]. Finally, due to protein instability over time, cotinine levels would only be accurately measurable in fresh blood samples, which are not always available such as in forensic investigations. Zhang et al. [24] showed that both DNA methylation and cotinine can accurately distinguish current from never smokers, but also emphasized that only DNA methylation is able to provide more in depth life-time smoking information. In line with this, we show in the current study that using both cotinine (sensitivity 0.750, specificity 0.983) and DNA methylation (sensitivity 0.621, specificity 0.989) we can infer current smokers with high accuracy. However, the sensitivity of our CpG model is slightly lower than the use of the cotinine cut-off in this subset. Nonetheless, with the upcoming availability of DNA methylation data in large cohort studies, the availability of a reliable smoking inference model, giving extending life-time smoking information inference, would be more widely accessible than information on cotinine levels.

Maternal smoking during pregnancy has been shown to influence fetal DNA methylation patterns [57, 71], which in principle could affect epigenetic inference of smoking habits in adults. Additionally, it is shown that maternal smoking status can be predicted from DNA methylation retrieved from newborns [72, 73]. Therefore, we employed data from the Generation R study to test the influence of prenatal smoking exposure on the inference of smoking status in adolescence. Hence, we tested our prediction model using epigenetic data from cord blood collected at time of birth, and peripheral blood collected at 6 and 10 years of age [37]. Our results showed that at the age of 6 years, 353 of the 355 children were correctly inferred as non-smokers (accuracy of 0.994), and at the age of 10 years 307 of the 309 children (accuracy of 0.994) were correctly inferred as non-smokers. This might indicate that prenatal smoking exposure and passive smoking exposure does not affect DNA methylation levels to such an extent that they are detected with our inference model. At time of birth, our model incorrectly inferred 984 (88.57%) of the 1111 children as smokers (accuracy of

0.114). To test whether the newborns were inferred wrongly as smokers due to prenatal smoking exposure, we further classified the newborns as smokers when their mothers smoked throughout pregnancy (N = 161). This resulted in a high AUC (0.773), with high sensitivity (0.981) but low specificity (0.131). Retrieving this low specificity while correcting for prenatal smoking exposure may indicate that the incorrect smoking inference of newborns achieved with our model can only in part be explained by smoking exposure during pregnancy. Other explanations may be developmental effects, and perhaps the tissue difference between whole blood and cord blood and therefore the difference in cell composition, given that the applied model was developed using whole blood [74]. Previous studies have shown specific changes in DNA methylation during early childhood that were explained by developmental effects [71, 75]. In any case, given that envisioned applications of epigenetic inference of smoking habit in medical and forensic practice, as well as in most epidemiological and public health research, are typically performed in adults, our findings in children of advanced age imply that our model will indeed deliver smoking habit information of the adult individual tested, independent of prenatal smoking exposure or other effects.

The main strengths of our study are (1) the use of robust DNA methylation markers highlighted in multiple epigenome-wide association studies, (2) the use of independent population-based studies for marker discovery, model building and external model validation, and (3) the employment of thousands of samples for model building and validation. We therefore expect that the high prediction accuracy (AUC of 0.911) obtained from the full 13-CpG model in the KORA samples used for external validation reflects a realistic characterization of the performance of our model. This is also supported in part by the SHIP-Trend outcomes (AUC of 0.888) of the partial 10-CpG model. As the Illumina 450 K array on which our marker selection was initially based is no longer available, the SHIP-Trend results using 10-CpG subset from the current Infinium MethylationEPIC BeadChip indicate that this sub-model would be applicable to new studies moving forward.

This study, however, does not come without limitations. Our model is based on smoking habit data retrieved from self-reported questionnaires, which are generally considered unreliable in terms of underestimating actual smoking levels [15]. Regarding the putative inaccuracy of self-reported smoking habits used here as phenotypes, we cannot know how error-prone these reports are. In particular, it is possible that specific groups of volunteers, for instance pregnant women such as those involved in the Generation R Study, are more reluctant to confide that they smoke [16]. However, we did not use the Generation R Study data for model building or validation purposes. Moreover, we included cotinine data to confirm the self-reported smoking habits for subset of participants (N = 488). Overall, we expect that smoking phenotype inaccuracy did not strongly impact the performance outcomes of our models. Lastly, all but one of the studies included in the model building and model validation are population-based studies, which therefore can include participants with various diseases. Though, due to the large sample sizes used for model building and validation, we expect that disease status does not strongly impact our model performance. Another limitation for the pack-year model is the formula used to calculate the pack-years. For this estimation, the number of cigarettes the participant currently smokes is used, which might have changed over the life span, and if so, this phenotypic variation is not considered. Additionally, the start-age is used to calculate the number of years someone smoked or has been smoking, which might be prone to recall bias especially for elderly people.

We envision that future works may provide targeted laboratory tools for analysing the 13 CpGs included in our final model in different types of blood samples and possible translation to different tissues, as is recently already shown to be promising for our top hit CpG (cg0557592) in saliva [76]. This would enhance the spectrum of practical applications of epigenetic smoking habit inference. Given the finite set of DNA methylation markers introduced here, it is impractical to apply genome-wide DNA methylation microarrays just for the purpose of analyzing 13 CpGs. Moreover, there can be blood samples where microarrays do not produce reliable DNA methylation data, such as when the amount of DNA is low and/or the DNA is degraded such as DNA obtained from crime scene traces [17]. Hence, the future development of a fast and cheap laboratory tool that allows the reliable targeted analysis of the 13 CpGs highlighted here by employing a technology that can handle low quality and/or quantity DNA would be valuable. Foreseeing the future development of such a lab tool, we only included CpGs with at least a β-value difference ≥ 10% in mean or median (depending on availability per EWAS) in at least one published EWAS, to ensure detectability of the DNA methylation differences with targeted analysis technologies currently available [77, 78]. We view the positive results on epigenetic inference of smoking habits from blood presented here as a promising starting point for inferring more lifestyle factors using DNA methylation markers within the concept of epigenetic fingerprinting [17]. This requires continuous progress in identifying candidate DNA methylation predictors of lifestyle factors via dedicated EWASs, the subsequent use of these biomarkers in prediction modeling and validation studies to generate reliable and accurate models such as that reported here for tobacco smoking, and the development of robust and sensitive lab tools that allow the successful analysis of the DNA samples of interest, including those of limited quality and quantity.

Acknowledgements The authors are grateful to the participants of the cohorts used: LifeLines (http://lifelines.nl/lifelines-research/general), the Leiden Longevity Study (http://www.leidenlangleven.nl), the Netherlands Twin Registry (http://www.tweelingenregister.org), the Rotterdam studies (http://www.erasmus epidemiology.nl/research/ergo.htm), the CODAM study (http://www.carimmaastricht.nl/), and the PAN study (http://www.alsonderzoek.nl/), the KORA study (https://www.helmholtz muenchen.de/en/kora/index.html), SHIP-Trend (http://www.medizin.uni greifswald.de/cm/fv/ship.html), Generation R (https://www.generationr.nl/). We also thank Dr. Hannah R Elliott for kindly sharing the R script, and Michael Verbiest, Mila Jhamai, Sarah Higgins, Marijn Verkerk and Dr. Lisette Stolk for their help in creating the EWAS database for RS and Generation R Study.

Funding This work was performed within the framework of the Biobank-Based Integrative Omics Studies (BIOS) Consortium funded by BBMRI-NL, a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO 184.021.007). This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant agreements No. 633595 (DynaHEALTH) and 733206 (LIFECYCLE). SCEM was supported by Netherlands Institute for Health Sciences scholarship. AV and MK were supported by the Erasmus MC University Medical Center Rotterdam. AV was additionally supported with an EUR fellowship by Erasmus University Rotterdam. LD received funding from the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 696295; 2017) co-funded by ERA-Net on Biomarkers for Nutrition and Health (ERA HDHL) and ZonMW The Netherlands (No. 529051014; 2017) (ALPHABET project). VWVJ received funding from the Netherlands Organization for Health Research and Development (VIDI 016.136.361) and a Consolidator Grant from the European Research Council (ERC-2014-CoG-648916). MW has received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under Grant agreements n°603288 (SysVasc) and n°602736 (PAIN-OMICS). The establishment of the RS EWAS data was funded by the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, and by the Netherlands Organization for Scientific Research (NWO; Project Number 184021007). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The general design of the Generation R Study is made possible by financial support from the Erasmus MC, the Erasmus University Rotterdam, the Netherlands Organization for Health Research and Development, and the Ministry of Health, Welfare and Sport. The generation and management of the Illumina 450 K methylation array data was funded by a grant to VWJ from the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA; Project No. 050-060-810), by funds from the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, and by a grant from the National Institute of Child and Human Development (R01HD068437). CODAM was supported by Grants of the Netherlands Organization for Scientific Research (940–35–034) and the Dutch Diabetes Research Foundation (98.901). Funding for the NTR was obtained from the Netherlands Organization for Scientific Research (NWO) and The Netherlands Organisation for Health Research and Development (ZonMW) Grants 904-61-090, 985-10-002, 912-10-020, 904-61-193,480-04-004, 463-06-001, 451-04-034, 400-05-717, Addiction-31160008, 016-115-035, 481-08-011, 056-32-010, Middelgroot-911-09-032, and NWO-Groot 480-15-001/674. The KORA study was initiated and financed by the Helmholtz Zentrum München –German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (Grants No. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine (GANI_MED)' funded by the Federal Ministry of Education and Research (Grant 03IS2061A). DNA methylation data have been supported by the DZHK (Grant 81X3400104). The University of Greifswald is a member of the Caché Campus program of the InterSystems GmbH. The researchers are independent from the funders. The study sponsors had no role in the study design, data collection, data analysis, interpretation of data, and preparation, review or approval of the manuscript.

Compliance with Ethical Standards

Conflict of interest H.J. Grabe has received funding from Fresenius Medical Care and speaker's honoraria as well as travel funds from Fresenius Medical Care, Neuraxpharm and Janssen-Cilag. Other than that, the authors declared no conflict of interest.

Ethics approval The study was approved by the institutional review boards of the participating medical centers: CODAM, Medical Ethical Committee of the Maastricht University; LL, Ethics committee of The University Medical Centre Groningen; LLS, Ethical committee of the Leiden University Medical Center; PAN, Institutional review board of the University Medical Centre Utrecht; NTR, Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre; RS, Dutch Ministry of Health; KORA, Institutional review board Ethics Committee of the Bavarian Medical Association (Bayrische Landesärztekammer); SHIP-Trend, Institutional review board Ethics committee of the University of Greifswald. The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC (Registration Number MEC 02.1015) and by the Dutch Ministry of Health, Welfare and Sport (Population Screening Act WBO, License Number 1071272-159521-PG). The Rotterdam Study has been entered into the Netherlands National Trial Register (NTR; www.trialregister.nl) and into the WHO International Clinical Trials Registry Platform (ICTRP; www.who.int/ictrp/network/primary/en/) under shared catalogue Number NTR6831. The experimental methods comply with the Helsinki Declaration.

Informed consent All participants, in case of children their parents, provided written informed consent to participate in the study and to have their information obtained from treating physicians.

References

1. Lee KW, Pausova Z. Cigarette smoking and DNA methylation. Front Genet. 2013;4:132.
2. 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.

3. Mortusewicz O, Schermelleh L, Walter J, Cardoso MC, Leonhardt H. Recruitment of DNA methyltransferase I to DNA repair sites. Proc Natl Acad Sci USA. 2005;102(25):8905–9.
4. Cuozzo C, Porcellini A, Angrisano T, Morano A, Lee B, Di Pardo A, et al. DNA damage, homology-directed repair, and DNA methylation. PLoS Genet. 2007;3(7):e110.
5. Satta R, Maloku E, Zhubi A, Pibiri F, Hajos M, Costa E, et al. Nicotine decreases DNA methyltransferase 1 expression and glutamic acid decarboxylase 67 promoter methylation in GABAergic interneurons. Proc Natl Acad Sci USA. 2008;105(42):16356–61.
6. Mercer BA, Wallace AM, Brinckerhoff CE, D'Armiento JM. Identification of a cigarette smoke-responsive region in the distal MMP-1 promoter. Am J Respir Cell Mol Biol. 2009;40(1):4–12.
7. Kadonaga JT, Carner KR, Masiarz FR, Tjian R. Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. Cell. 1987;51(6):1079–90.
8. Di YP, Zhao J, Harper R. Cigarette smoke induces MUC5AC protein expression through the activation of Sp1. J Biol Chem. 2012;287(33):27948–58.
9. Han L, Lin IG, Hsieh CL. Protein binding protects sites on stable episomes and in the chromosome from de novo methylation. Mol Cell Biol. 2001;21(10):3416–24.
10. 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.
11. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic signatures of cigarette smoking. Circ Cardiovasc Genet. 2016;9(5):436–47.
12. 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(13):3073–82.
13. Ligthart S, Steenaard RV, Peters MJ, van Meurs JB, Sijbrands EJ, Uitterlinden AG, et al. Tobacco smoking is associated with DNA methylation of diabetes susceptibility genes. Diabetologia. 2016;59:998–1006.
14. Steenaard RV, Ligthart S, Stolk L, Peters MJ, van Meurs JB, Uitterlinden AG, et al. Tobacco smoking is associated with methylation of genes related to coronary artery disease. Clin Epigenetics. 2015;7(1):54.
15. Connor Gorber S, Schofield-Hurwitz S, Hardt J, Levasseur G, Tremblay M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. Nicotine Tob Res. 2009;11(1):12–24.
16. Shipton D, Tappin DM, Vadiveloo T, Crossley JA, Aitken DA, Chalmers J. Reliability of self reported smoking status by pregnant women for estimating smoking prevalence: a retrospective, cross sectional study. BMJ. 2009;339:b4347.
17. Vidaki A, Kayser M. From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. Genome Biol. 2017;18(1):238.
18. Florescu A, Ferrence R, Einarson T, Selby P, Soldin O, Koren G. Methods for quantification of exposure to cigarette smoking and environmental tobacco smoke: focus on developmental toxicology. Ther Drug Monit. 2009;31(1):14–30.
19. Benowitz NL. Cotinine as a biomarker of environmental tobacco smoke exposure. Epidemiol Rev. 1996;18(2):188–204.
20. Benowitz NL, Hukkanen J, Jacob P 3rd. Nicotine chemistry, metabolism, kinetics and biomarkers. Handb Exp Pharmacol. 2009;192:29–60.
21. 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(1):4.
22. 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(5):712–6.
23. Philibert R, Hollenbeck N, Andersen E, Osborn T, Gerrard M, Gibbons FX, et al. A quantitative epigenetic approach for the assessment of cigarette consumption. Front Psychol. 2015;6:656.
24. Zhang Y, Florath I, Saum KU, Brenner H. Self-reported smoking, serum cotinine, and blood DNA methylation. Environ Res. 2016;146:395–403.
25. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. Genome Biol. 2018;19(1):136.
26. Kondratyev N, Golov A, Alfimova M, Lezheiko T, Golimbet V. Prediction of smoking by multiplex bisulfite PCR with long amplicons considering allele-specific effects on DNA methylation. Clin Epigenetics. 2018;10(1):130.
27. Sugden K, Hannon EJ, Arseneault L, Belsky DW, Broadbent JM, Corcoran DL, et al. Establishing a generalized polyepigenetic biomarker for tobacco smoking. Transl Psychiatry. 2019;9(1):92.
28. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49(1):131–8.
29. Ikram MA, Brusselle GGO, Murad SD, van Duijn CM, Franco OH, Goedegebure A, et al. The Rotterdam study: 2018 update on objectives, design and main results. Eur J Epidemiol. 2017;32(9):807–50.
30. van Greevenbroek MM, Jacobs M, van der Kallen CJ, Vermeulen VM, Jansen EH, Schalkwijk CG, et al. The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). Eur J Clin Invest. 2011;41(4):372–9.
31. Willemsen G, Vink JM, Abdellaoui A, den Braber A, van Beek JH, Draisma HH, et al. The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. Twin Res Hum Genet. 2013;16(1):271–81.
32. Schoenmaker M, de Craen AJ, de Meijer PH, Beekman M, Blauw GJ, Slagboom PE, et al. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. Eur J Hum Genet. 2006;14(1):79–84.
33. Huisman MH, de Jong SW, van Doormaal PT, Weinreich SS, Schelhaas HJ, van der Kooi AJ, et al. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. J Neurol Neurosurg Psychiatry. 2011;82(10):1165–70.
34. Tigchelaar EF, Zhernakova A, Dekens JA, Hermes G, Baranska A, Mujagic Z, et al. Cohort profile: lifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open. 2015;5(8):e006772.
35. Holle R, Happich M, Lowel H, Wichmann HE, Group MKS. KORA—a research platform for population based health research. Gesundheitswesen. 2005;67(Suppl 1):S19–25.
36. Jurgens C, Volzke H, Tost F. [Study of health in Pomerania (SHIP-Trend): important aspects for healthcare research in ophthalmology] Study of Health in Pomerania (SHIP-Trend): Wichtige Aspekte fur die ophthalmologische Versorgungsforschung. Ophthalmologe. 2014;111(5):443–7.
37. Kooijman MN, Kruithof CJ, van Duijn CM, Duijts L, Franco OH, van IJzendoorn MH, et al. The Generation R Study: design and cohort update 2017. Eur J Epidemiol. 2016;31(12):1243–64.
38. Tobi EW, Slieker RC, Stein AD, Suchiman HE, Slagboom PE, van Zwet EW, et al. Early gestation as the critical time-window for changes in the prenatal environment to affect the adult human blood methylome. Int J Epidemiol. 2015;44(4):1211–23.

39. van Iterson M, Tobi EW, Slieker RC, den Hollander W, Luijk R, Slagboom PE, et al. MethylAid: visual and interactive quality control of large Illumina 450 k datasets. Bioinformatics. 2014;30(23):3435–7.
40. 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(5):843–51.
41. Zeilinger S, Kuhnel 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(5):e63812.
42. Harlid S, Xu Z, Panduri V, Sandler DP, Taylor JA. CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the Sister Study. Environ Health Perspect. 2014;122(7):673–8.
43. Tsaprouni LG, Yang TP, 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(10):1382–96.
44. Allione A, Marcon F, Fiorito G, Guarrera S, Siniscalchi E, Zijno A, et al. Novel epigenetic changes unveiled by monozygotic twins discordant for smoking habits. PLoS One. 2015;10(6):e0128265.
45. Besingi W, Johansson A. Smoke-related DNA methylation changes in the etiology of human disease. Hum Mol Genet. 2014;23(9):2290–7.
46. 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 Genom. 2014;15:151.
47. Sayols-Baixeras S, Lluis-Ganella C, Subirana I, Salas LA, Vilahur N, Corella D, et al. Corrigendum. Identification of a new locus and validation of previously reported loci showing differential methylation associated with smoking. The REGICOR study. Epigenetics. 2016;11(2):174.
48. 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(5):599–618.
49. Joubert BR, Haberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 450 K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect. 2012;120(10):1425–31.
50. Zhu X, Li J, Deng S, Yu K, Liu X, Deng Q, et al. Genome-wide analysis of DNA methylation and cigarette smoking in a Chinese population. Environ Health Perspect. 2016;124(7):966–73.
51. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning; data mining, inference, and prediction. 2nd ed. Berlin: Springer; 2009.
52. Bradley E, Tibshirani RJ. An introduction to the bootstrap. Boca Raton: Chapman and Hall/CRC; 1994.
53. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54(8):774–81.
54. Ware JJ, Chen X, Vink J, Loukola A, Minica C, Pool R, et al. Genome-wide meta-analysis of cotinine levels in cigarette smokers identifies locus at 4q13.2. Sci Rep. 2016;6:20092.
55. Gupta R, van Dongen J, Fu Y, Abdellaoui A, Tyndale RF, Velagapudi V, et al. Epigenome-wide association study of serum cotinine in current smokers reveals novel genetically driven loci. Clin Epigenetics. 2019;11(1):1.
56. Bot M, Vink JM, Willemsen G, Smit JH, Neuteboom J, Kluft C, et al. Exposure to secondhand smoke and depression and anxiety: a report from two studies in the Netherlands. J Psychosom Res. 2013;75(5):431–6.
57. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. DNA methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. Am J Hum Genet. 2016;98(4):680–96.
58. Richmond RC, Suderman M, Langdon R, Relton CL, Davey Smith G. DNA methylation as a marker for prenatal smoke exposure in adults. Int J Epidemiol. 2018;47(4):1120–30.
59. Philibert RA, Beach SR, Lei MK, Brody GH. Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. Clin Epigenetics. 2013;5(1):19.
60. Esser C. Biology and function of the aryl hydrocarbon receptor: report of an international and interdisciplinary conference. Arch Toxicol. 2012;86(8):1323–9.
61. Nguyen TA, Hoivik D, Lee JE, Safe S. Interactions of nuclear receptor coactivator/corepressor proteins with the aryl hydrocarbon receptor complex. Arch Biochem Biophys. 1999;367(2):250–7.
62. Bojesen SE, Timpson N, Relton C, Davey Smith G, Nordestgaard BG. AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. Thorax. 2017;72(7):646–53.
63. Zaghlool SB, Al-Shafai M, Al Muftah WA, Kumar P, Falchi M, Suhre K. Association of DNA methylation with age, gender, and smoking in an Arab population. Clin Epigenetics. 2015;7:6.
64. Lim U, Song MA. Dietary and lifestyle factors of DNA methylation. Methods Mol Biol. 2012;863:359–76.
65. Su D, Wang X, Campbell MR, Porter DK, Pittman GS, Bennett BD, et al. Distinct epigenetic effects of tobacco smoking in whole blood and among leukocyte subtypes. PLoS ONE. 2016;11(12):e0166486.
66. Bauer M, Fink B, Thurmann L, Eszlinger M, Herberth G, Lehmann I. Tobacco smoking differently influences cell types of the innate and adaptive immune system-indications from CpG site methylation. Clin Epigenetics. 2015;7:83.
67. Bauer M, Linsel G, Fink B, Offenberg K, Hahn AM, Sack U, et al. A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. Clin Epigenetics. 2015;7:81.
68. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. Epigenomics. 2017;9(5):757–68.
69. Zhang Y, Schottker B, Florath I, Stock C, Butterbach K, Holleczek B, et al. Smoking-associated DNA methylation biomarkers and their predictive value for all-cause and cardiovascular mortality. Environ Health Perspect. 2016;124(1):67–74.
70. Stead LF, Perera R, Bullen C, Mant D, Hartmann-Boyce J, Cahill K, et al. Nicotine replacement therapy for smoking cessation. Cochrane Database Syst Rev. 2012;11:CD000146.
71. Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL, et al. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). Hum Mol Genet. 2015;24(8):2201–17.
72. Ladd-Acosta C, Shu C, Lee BK, Gidaya N, Singer A, Schieve LA, et al. Presence of an epigenetic signature of prenatal cigarette smoke exposure in childhood. Environ Res. 2016;144(Pt A):139–48.
73. Reese SE, Zhao S, Wu MC, Joubert BR, Parr CL, Haberg SE, et al. DNA methylation score as a biomarker in newborns for sustained maternal smoking during pregnancy. Environ Health Perspect. 2017;125(4):760–6.
74. Bergens MA, Pittman GS, Thompson IJB, Campbell MR, Wang X, Hoyo C, et al. Smoking-associated AHRR demethylation in cord blood DNA: impact of CD235a + nucleated red blood cells. Clin Epigenetics. 2019;11(1):87.
75. Xu CJ, Bonder MJ, Soderhall C, Bustamante M, Baiz N, Gehring U, et al. The emerging landscape of dynamic DNA methylation in early childhood. BMC Genom. 2017;18(1):25.

76. Dawes K, Andersen A, Vercande K, Papworth E, Philibert W, Beach SRH, et al. Saliva DNA methylation detects nascent smoking in adolescents. J Child Adolesc Psychopharmacol. 2019. https://doi.org/10.1089/cap.2018.0176.
77. Vidaki A, Johansson C, Giangasparo F, Denise Syndercombe C. Differentially methylated embryonal Fyn-associated substrate (EFS) gene as a blood-specific epigenetic marker and its potential application in forensic casework. Forensic Sci Int Genet. 2017;29:165–73.
78. Vidaki A, Ballard D, Aliferi A, Miller TH, Barron LP, Syndercombe Court D. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. Forensic Sci Int Genet. 2017;28:225–36.

Affiliations

Silvana C. E. Maas[1,2] · Athina Vidaki[2] · Rory Wilson[3,4] · Alexander Teumer[5,6] · Fan Liu[2,7,8] · Joyce B. J. van Meurs[1,9] · André G. Uitterlinden[1,9] · Dorret I. Boomsma[10] · Eco J. C. de Geus[10] · Gonneke Willemsen[10] · Jenny van Dongen[10] · Carla J. H. van der Kallen[11,12] · P. Eline Slagboom[13] · Marian Beekman[13] · Diana van Heemst[14] · Leonard H. van den Berg[15] · BIOS Consortium · Liesbeth Duijts[16] · Vincent W. V. Jaddoe[1,17,18] · Karl-Heinz Ladwig[4] · Sonja Kunze[3,4] · Annette Peters[3,4,19,20] · M. Arfan Ikram[1] · Hans J. Grabe[21] · Janine F. Felix[1,17,18] · Melanie Waldenberger[3,4,19] · Oscar H. Franco[1] · Mohsen Ghanbari[1,22] · Manfred Kayser[2]

Silvana C. E. Maas
s.maas@erasmusmc.nl

Athina Vidaki
a.vidaki@erasmusmc.nl

Rory Wilson
rory.wilson@helmholtz-muenchen.de

Alexander Teumer
ateumer@uni-greifswald.de

Fan Liu
liufan@big.ac.cn

Joyce B. J. van Meurs
j.vanmeurs@erasmusmc.nl

André G. Uitterlinden
a.g.uitterlinden@erasmusmc.nl

Dorret I. Boomsma
di.boomsma@vu.nl

Eco J. C. de Geus
eco.de.geus@vu.nl

Gonneke Willemsen
a.h.m.willemsen@vu.nl

Jenny van Dongen
j.van.dongen@vu.nl

Carla J. H. van der Kallen
c.vanderkallen@maastrichtuniversity.nl

P. Eline Slagboom
P.Slagboom@lumc.nl

Marian Beekman
M.Beekman@lumc.nl

Diana van Heemst
d.van_heemst@lumc.nl

Leonard H. van den Berg
L.H.vandenBerg@umcutrecht.nl

BIOS Consortium
info@bbmri.nl

Liesbeth Duijts
l.duijts@erasmusmc.nl

Vincent W. V. Jaddoe
v.jaddoe@erasmusmc.nl

Karl-Heinz Ladwig
ladwig@helmholtz-muenchen.de

Sonja Kunze
sonja.kunze@helmholtz-muenchen.de

Annette Peters
peters@helmholtz-muenchen.de

M. Arfan Ikram
m.a.ikram@erasmusmc.nl

Hans J. Grabe
grabeh@uni-greifswald.de

Janine F. Felix
j.felix@erasmusmc.nl

Melanie Waldenberger
waldenberger@helmholtz-muenchen.de

Oscar H. Franco
o.franco@erasmusmc.nl

[1] Department of Epidemiology, Erasmus MC University Medical Center Rotterdam, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands

[2] Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands

[3] Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Ingolstädter Landstr. 1, 85764 Neuherberg, Germany

[4] Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Ingolstädter Landstr. 1, 85764 Neuherberg, Germany

[5] Institute for Community Medicine, University Medicine Greifswald, Walther-Rathenau-Str. 48, 17475 Greifswald, Germany

[6] German Center for Cardiovascular Research (DZHK), Partner Site Greifswald, 17475 Greifswald, Germany

[7] Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, NO.1 Beichen West Road, Chaoyang District, 100101 Beijing, People's Republic of China

[8] University of Chinese Academy of Sciences, No.19A Yuquan Road, Shijingshan District, 100049 Beijing, People's Republic of China

[9] Department of Internal Medicine, Erasmus MC University Medical Center Rotterdam, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands

[10] Department of Biological Psychology, Vrije Universiteit, Van der Boechorststraat 7-9, 1081 BT Amsterdam, The Netherlands

[11] Department of Internal Medicine, Maastricht University Medical Center, Randwycksingel 35, 6229 EG Maastricht, The Netherlands

[12] Cardiovascular Research Institute Maastricht (CARIM), Maastricht University, Universiteitssingel 50, 6229 ER Maastricht, The Netherlands

[13] Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, P.O. box 9600, 2300 RC Leiden, The Netherlands

[14] Gerontology and Geriatrics, Department of Internal Medicine, Leiden University Medical Center, P.O. box 9600, 2300 RC Leiden, The Netherlands

[15] Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Postbus 85500, 3508 GA Utrecht, The Netherlands

[16] Division of Respiratory Medicine and Allergology and Division of Neonatology, Department of Pediatrics, Erasmus MC University Medical Center Rotterdam, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands

[17] The Generation R Study Group, Erasmus MC University Medical Center Rotterdam, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands

[18] Department of Pediatrics, Erasmus MC University Medical Center Rotterdam, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands

[19] German Center for Cardiovascular Research (DZHK), Partner Site Munich Heart Alliance, 80802 Munich, Germany

[20] Institute for Medical Informatics, Biometrics and Epidemiology, Ludwig-Maximilians-Universität (LMU) Munich, Marchioninistr. 15, 81377 Munich, Germany

[21] Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Ellernholzstraße 1-2, 17475 Greifswald, Germany

[22] Department of Genetics, School of Medicine, Mashhad University of Medical Science, PO Box 91735-951, 9133913716 Mashhad, Iran