Original Article

Epigenetic Signatures of Cigarette Smoking

Roby Joehanes, PhD*; Allan C. Just, PhD*; Riccardo E. Marioni, PhD*; Luke C. Pilling, PhD*; Lindsay M. Reynolds, PhD*; Pooja R. Mandaviya, MSc*; Weihua Guan, PhD*; Tao Xu, PhD*; Cathy E. Elks, PhD*; Stella Aslibekyan, PhD*; Hortensia Moreno-Macias, ScD*; Jennifer A. Smith, PhD, MPH*; Jennifer A. Brody, BA*; Radhika Dhingra, PhD*; Paul Yousefi, MPH; James S. Pankow, PhD; Sonja Kunze, PhD; Sonia H. Shah, PhD; Allan F. McRae, PhD; Kurt Lohman, MStat; Jin Sha, MS; Devin M. Absher, PhD; Luigi Ferrucci, MD, PhD; Wei Zhao, PhD; Ellen W. Demerath, PhD; Jan Bressler, PhD; Megan L. Grove, MS; Tianxiao Huan, PhD; Chunyu Liu, PhD; Michael M. Mendelson, MD; Chen Yao, PhD; Douglas P. Kiel, MD, MPH; Annette Peters, PhD; Rui Wang-Sattler, PhD; Peter M. Visscher, PhD; Naomi R. Wray, PhD; John M. Starr, PhD; Jingzhong Ding, PhD; Carlos J. Rodriguez, MD, MPH; Nicholas J. Wareham, PhD; Marguerite R. Irvin, PhD; Degui Zhi, PhD; Myrto Barrdahl, PhD; Paolo Vineis, MD; Srikant Ambatipudi, PhD; André G. Uitterlinden, PhD; Albert Hofman, MD, PhD; Joel Schwartz, PhD; Elena Colicino, PhD; Lifang Hou, MD, PhD; Pantel S. Vokonas, MD; Dena G. Hernandez, PhD; Andrew B. Singleton, PhD; Stefania Bandinelli, MD; Stephen T. Turner, MD; Erin B. Ware, PhD, MPH; Alicia K. Smith, PhD; Torsten Klengel, MD; Elisabeth B. Binder, MD, PhD; Bruce M. Psaty, MD, PhD; Kent D. Taylor, PhD; Sina A. Gharib, MD; Brenton R. Swenson, MPP; Liming Liang, PhD; Dawn L. DeMeo, MD, MPH; George T. O'Connor, MD, MS; Zdenko Herceg, DVM, MSc, PhD; Kerry J. Ressler, MD, PhD; Karen N. Conneely, PhD†; Nona Sotoodehnia, MD, MPH†; Sharon L. R. Kardia, PhD†; David Melzer, MBBCh, PhD†; Andrea A. Baccarelli, MD, PhD†; Joyce B. J. van Meurs, PhD†, Isabelle Romieu, MD, ScD†; Donna K. Arnett, PhD†; Ken K. Ong, MB BChir, PhD†; Yongmei Liu, MD, PhD†; Melanie Waldenberger, PhD†; Ian J. Deary, PhD†; Myriam Fornage, PhD†; Daniel Levy, MD†; Stephanie J. London, MD, DrPH†

Background—DNA methylation leaves a long-term signature of smoking exposure and is one potential mechanism by which tobacco exposure predisposes to adverse health outcomes, such as cancers, osteoporosis, lung, and cardiovascular disorders.

Methods and Results—To comprehensively determine the association between cigarette smoking and DNA methylation, we conducted a meta-analysis of genome-wide DNA methylation assessed using the Illumina BeadChip 450K array on 15 907 blood-derived DNA samples from participants in 16 cohorts (including 2433 current, 6518 former, and 6956 never smokers). Comparing current versus never smokers, 2623 cytosine–phosphate–guanine sites (CpGs), annotated to 1405 genes, were statistically significantly differentially methylated at Bonferroni threshold of $P<1\times10^{-7}$ (18 760 CpGs at false discovery rate <0.05). Genes annotated to these CpGs were enriched for associations with several smoking-related traits in genome-wide studies including pulmonary function, cancers, inflammatory diseases, and heart disease. Comparing former versus never smokers, 185 of the CpGs that differed between current and never smokers were significant $P<1\times10^{-7}$ (2623 CpGs at false discovery rate <0.05), indicating a pattern of persistent altered methylation, with attenuation, after smoking cessation. Transcriptomic integration identified effects on gene expression at many differentially methylated CpGs.

Received January 19, 2016; accepted August 16, 2016.

Guest Editor for this article was Christopher Semsarian, MBBS, PhD, MPH.

*Drs Joehanes, Just, Marioni, Pilling, Reynolds, Guan, Xu, Elks, Aslibekyan, Moreno-Macias, J.A. Smith, Brody, Dhingra, and P.R. Mandaviya contributed equally as first authors.

†Drs. Conneely, Sotoodehnia, Kardia, Melzer, Baccarelli, van Meurs, Romieu, Arnett, Ong, Y. Liu, Waldenberger, Deary, Fornage, Levy, and London contributed equally as senior authors.

The Data Supplement is available at http://circgenetics.ahajournals.org/lookup/suppl/doi:10.1161/CIRCGENETICS.116.001506/-/DC1.

Correspondence to Stephanie J. London, MD, DrPH, Epidemiology Branch, Department of Health and Human Services, National Institute of Environmental Health Sciences, National Institutes of Health, PO Box 12233, Room A306, Research Triangle Park, NC 27709. E-mail london2@niehs.nih.gov

***Circ Cardiovasc Genet* is available at http://circgenetics.ahajournals.org** **DOI: 10.1161/CIRCGENETICS.116.001506**

Conclusions—Cigarette smoking has a broad impact on genome-wide methylation that, at many loci, persists many years after smoking cessation. Many of the differentially methylated genes were novel genes with respect to biological effects of smoking and might represent therapeutic targets for prevention or treatment of tobacco-related diseases. Methylation at these sites could also serve as sensitive and stable biomarkers of lifetime exposure to tobacco smoke.

(*Circ Cardiovasc Genet*. 2016;9:436-447. DOI: 10.1161/CIRCGENETICS.116.001506.)

Key Words: biomarkers ■ genome-wide association study ■ meta-analysis ■ methylation ■ smoking

Cigarette smoking is a major causal risk factor for various diseases, including cancers, cardiovascular disease, chronic obstructive pulmonary disease,[1] and osteoporosis.[1] Worldwide cessation campaigns and legislative actions have been accompanied by a reduction in the number of cigarette smokers and corresponding increases in the number of former smokers. In the United States, there are more former smokers than current smokers.[1] Despite the decline in the prevalence of smoking in many countries, it remains the leading preventable cause of death in the world, accounting for ≈6 million deaths each year.[2]

Clinical Perspective on p 447

Even decades after cessation, cigarette smoking confers long-term risk of diseases including some cancers, chronic obstructive pulmonary disease, and stroke.[1] The mechanisms for these long-term effects are not well understood. DNA methylation changes have been proposed as one possible explanation.

DNA methylation seems to reflect exposure to a variety of lifestyle factors,[3] including cigarette smoking. Several studies have shown reproducible associations between tobacco smoking and altered DNA methylation at multiple cytosine–phosphate–guanine (CpG) sites (CpGs).[4–15] Some DNA methylation sites associated with tobacco smoking have also localized to genes related to coronary heart disease[5] and pulmonary disease.[16] Some studies have found differently associated CpGs in smokers versus nonsmokers.[8,11] Consortium-based meta-analyses have been extremely successful in identifying genetic variants associated with numerous phenotypes, but large-scale meta-analyses of genome-wide DNA methylation data have not yet been widely used. It is likely that additional novel loci differentially methylated in response to cigarette smoking remain to be discovered by meta-analyzing data across larger sample sizes comprising multiple cohorts. Differentially methylated loci with respect to smoking may serve as biomarkers of lifetime smoking exposure. They may also shed light on the molecular mechanisms by which tobacco exposure predisposes to multiple diseases.

A recent systematic review[13] analyzed published findings across 14 epigenome-wide association studies of smoking exposure across various DNA methylation platforms of varying degrees of coverage and varying phenotypic definitions. Among these were 12 studies (comprising 4750 subjects) that used the more comprehensive Illumina Human Methylation BeadChip 450K array (Illumina 450K), which includes and greatly expands on the coverage of the earlier 27K platform. The review compares only statistically significant published results and is not a meta-analysis that can identify signals that do not reach statistical significance in individual studies.[17]

In the current study, we meta-analyzed association results between DNA methylation and cigarette smoking in 15 907 individuals from 16 cohorts in the CHARGE consortium (*C*ohorts for *H*eart and *A*ging *R*esearch in *G*enomic *E*pidemiology) using a harmonized analysis. Methylation was measured on DNA extracted from blood samples using the Illumina Human Methylation BeadChip 450K array. In separate analyses, we compared current smokers and past smokers with nonsmokers and characterized the persistence of smoking-related CpG methylation associations with the duration of smoking cessation among former smokers. We integrated information from genome-wide association studies (GWAS) and gene expression data to gain insight into potential functional relevance of our findings for human diseases. Finally, we conducted analyses to identify pathways that may explain the molecular effects of cigarette exposure on tobacco-related diseases.

Materials and Methods

Study Participants

This study comprised a total of 15 907 participants from 16 cohorts of the Cohorts for Heart and Aging Research in Genetic Epidemiology Consortium (Table I in the Data Supplement). The 16 participating cohorts are ARIC, FHS Offspring, KORA F4, GOLDN, LBC 1921, LBC 1936, NAS, Rotterdam, Inchianti, GTP, CHS European Ancestry (EA), CHS African Ancestry (AA), GENOA, EPIC Norfolk, EPIC, and MESA (Multi-Ethnic Study of Atherosclerosis). Of these, 12 161 are of EA and 3746 are of AA. The study was approved by institutional review committees for each cohort, and all participants provided written informed consent for genetic research.

DNA Methylation Sample and Measurement

For most studies, methylation was measured on DNA extracted from whole blood, but some studies used CD4$^+$ T cells or monocytes (Table I in the Data Supplement). In all studies, DNA was bisulfite converted using the Zymo EZ DNA methylation kit and assayed for methylation using the Infinium HumanMethylation 450 BeadChip, which contains 485 512 CpG sites. Details of genomic DNA preparation, bisulfite conversion, and methylation assay for each cohort can be found in the Data Supplement.

Raw methylated and total probe intensities were extracted using the Illumina Genome Studio methylation module. Preprocessing of the methylated signal (M) and unmethylated signal (U) was conducted using various software tools, primarily DASEN of wateRmelon[18] and BMIQ,[19] both of which are R packages. The methylation beta (β) values were defined as $\beta=M/(M+U)$. Each cohort followed its own quality-control protocols, removing poor quality or outlier samples and excluding low-quality CpG sites (with detection P value >0.01). Each cohort evaluated batch effects and controlled for them in the analysis. Details of these processes can be found in the Data Supplement.

Smoking Phenotype Definition

Self-reported cigarette-smoking status was divided into 3 categories. Current smokers were defined as those who have smoked at least 1 cigarette a day within 12 months before the blood draw, former smokers were defined as those who had ever smoked at least 1 cigarette a day but had stopped at least 12 months before the blood draw, and never smokers reported never having smoked. Pack years was calculated based on self-report as the average number of cigarettes smoked per day divided by 20 multiplied by the number of years of smoking, with zero assigned to never smokers. A few cohorts recorded the number of years since each former smoker had stopped smoking.

Cohort-Specific Analyses and Meta-Analysis

Each cohort analyzed its data using at least 2 linear mixed-effect models. Each model was run separately for each CpG site. Model 1 is as follows:

$$\beta = \text{smoking phenotype} + \text{sex} + \text{age} + \text{blood count} + \text{technical covariates} \quad (1)$$

where blood count comprises the fractions of $CD4^+$ T cells, $CD8^+$ T cells, NK cells, monocyte, and eosinophils either measured or estimated using the Houseman et al method.[20] The blood count adjustment was performed only in cohorts with whole-blood and leukocyte samples. Familial relationship was also accounted for in the model when applicable (eg, for FHS, see Data Supplement for details). Acknowledging that each cohort may be influenced by a unique set of technical factors, we allow each cohort to choose its cohort-specific technical covariates. Model 2 added to model 1 body mass index because it is associated with methylation at some loci, making it a potential confounder.[21] Only 3 cohorts participated in model 2 analysis: FHS, KORA, and NAS. Model 3 substituted smoking phenotypes for pack years. Only 3 cohorts participated in model 3 analysis: FHS, Rotterdam, and Inchianti. The pack-year analysis was performed only on 2 subsets: current versus never smokers and former versus never smokers. Combining all 3 categories would require accurate records of time of quitting, which among the 3 cohorts was available for only FHS. To investigate cell type differences, we removed blood counts from model 1 and called it model 4. Only 3 cohorts participated in this analysis: FHS, KORA, and NAS. All models were run with the lme4 package[22] in R,[23] except for FHS (see Data Supplement for details).

Meta-analysis was performed to combine the results from all cohorts. Because of the variability of available CpG sites after quality-control steps, we excluded CpG sites that were available in <3 cohorts. The remaining 485 381 CpG sites were then meta-analyzed with a random-effects model using the following formula:

$$E_i = \mu + s_i + e_i \quad (2)$$

where E_i is the observed effect of study i, μ is the main smoking effect, s_i is the between-study error for study i, and e_i is the within-study error for study i, with both s_i and e_i are assumed to be normally distributed. The model is fitted using the restricted maximum likelihood criterion in R's metafor[24] package. Multiple-testing adjustment on the resulting P values was performed using the false discovery rate (FDR) method of Benjamini and Hochberg.[25] In addition, we also report results using the Bonferroni-corrected threshold of 1×10^{-7} (≈0.05/485 381).

The regression coefficient β (from meta-analysis) is interpretable as the difference in mean methylation between current and never smokers. We multiplied these by 100 to represent the percentage methylation difference where methylation ranges from 0% to 100%.

Literature Review to Identify Genes Previously Associated With Smoking and Methylation

We used the same literature search strategy published previously.[26] A broad query of NCBIs PubMed literature database using medical subject heading (MeSH) terms ("((((DNA Methylation[Mesh]) OR methylation)) AND ((Smoking[Mesh]) OR smoking))") yielded 775 results when initially performed on January 8, 2015, and 789 studies when repeated to update the results on March 1, 2015. Results were reviewed by abstract to determine whether studies met inclusion criteria: (1) performed in healthy human populations, (2) agnostically examined >1000 CpG sites at a time, (3) only cigarette exposure was considered, and (4) with public reporting of P values and gene annotations. A total of 25 publications met inclusion criteria, listed in the fourth supplementary table of Joubert et al.[26] CpG-level results (P values and gene annotations) for sites showing genome-wide statistically significant associations (FDR <0.05) were extracted and resulted in 1185 genes previously associated with adult or maternal smoking. All CpGs annotated to these 1185 genes were marked as previously found.

Gene-Set Enrichment Analysis

Gene-set enrichment analysis[27] was performed in the website (http://software.broadinstitute.org/gsea/msigdb/annotate.jsp) on significant findings to determine putative functions of the CpG sites. We selected gene ontology biological process (C5-BP) and collected all categories with FDR <0.05 (≤100 categories).

Enrichment Analysis for Localization to Different Genomic Features

Enrichment analysis on genomic features were performed using the annotation file supplied by Illumina (version 1.2; downloaded from manufacturer's website, http://support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit/downloads.html), which contains information of CpG location relative to gene (ie, body, first exon, 3′ UTR, 5′ UTR, within 200 base pairs of transcriptional start site [TSS200], and within 1500 base pairs of transcriptional start site [TSS1500], the relation of CpG site to a CpG island (*ie,* island, northern shelf, northern shore, southern shelf, and southern shore), whether the CpG site is known to be in differentially methylated regions, and whether the CpG site is known to be an enhancer or a DNAse I hypersensitive site. Enrichment analysis was performed using 1-sided Fisher exact set for each feature, using R's fisher.test.

GWAS Analysis

We intersected our results with single-nucleotide polymorphisms (SNPs) having GWAS P values $\leq5\times10^{-8}$ in the National Human Genome Research Institute GWAS catalog (accessed November 2, 2015).[28] The catalog contained 9777 SNPs annotated to 7075 genes associated with 865 phenotypes at $P\leq5\times10^{-8}$. To determine the genes, we looked up each significant CpG on the annotation file supplied by Illumina. Enrichment analysis was performed on a per-gene basis using 1-sided Fisher exact test.

For bone mineral phenotype enrichment, we included all SNPs containing terms bone mineral density or osteoporosis. For cardiovascular disease, we included all SNPs containing terms cardiovascular disease, stroke, coronary disease, cardiomyopathy, or myocardial infarction. For cardiovascular disease risk factors, we included all SNPs containing terms blood pressure, cholesterol, diabetes, obesity, or hypertension. For overall cancer enrichment, we included all SNPs containing terms cancer, carcinoma, or lymphoma, while removing those pertaining to cancer treatment effects. For overall pulmonary phenotype enrichment, we included all SNPs containing terms pulmonary disease, pulmonary function, emphysema, asthma, or airflow obstruction.

Analysis of Persistence of Methylation Signals With Time Since Quitting Smoking Among Former Smokers

We examined whether smoking methylation associations were attenuated over time in the FHS cohort, which had ascertained longitudinal smoking status of >35 years. The analysis was performed on 7 dichotomous variables, indicating cessation of smoking for 5, 10, 15, 20, 25, and 30 years versus never smokers. For example, for 5-year cessation variable, those who quit smoking before ≥5 years are marked as ones, whereas never smokers are marked as zeroes, and current smokers are excluded. For this analysis, we used the pedigreemm package[29] with the same set of covariates as in the primary analysis. Sites with $P<0.002$ across all 7 variables were deemed to be statistically significant compared with never-smoker levels.

Methylation by Expression Analysis

To determine transcriptomic association of each significant CpG site, we interrogated such CpG sites in the FHS gene-level methylation by expression database, at genome-wide FDR <0.05. The methylation by expression database was constructed from 2262 individuals from the FHS Offspring cohort attending examination cycle 8 (2005–2008) with both whole-blood DNA methylation and transcriptomic data based on the Affymetrix Human Exon Array ST 1.0. Enrichment analysis was performed using a 1-sided Fisher exact test. We defined that the methylation CpG site and the corresponding transcript are associated in *cis* if the location of the CpG site is within 500 kilobases of the transcript's start location.

Analysis of Ethnic Discrepancy Between AA and EA Cohorts

Meta-analysis of the current versus never smoker results of EA cohorts (FHS, KORA, GOLDN, LBC 1921, LBC 1936, NAS, Rotterdam, Inchianti, EPIC, EPIC Norfolk, MESA, and CHS-EA) was performed separately from those of AA cohorts (ARIC, GTP, GENOA, and CHS-AA).

Analysis of Sample Types for DNA Extraction

Meta-analysis was performed on the results from cohorts with whole blood/buffy coat samples (FHS, KORA, LBC 1921, LBC 1936, NAS, Rotterdam, Inchianti, GTP, CHS-EA, CHS-AA, ARIC, GENOA, EPIC, and EPIC Norfolk). $CD4^+$ samples in GOLDN and $CD14^+$ samples in MESA, because they comprise single cohorts, are not meta-analyzed. Correlations of results across different cell types were performed on CpG sites with FDR <0.05 in at least one cell type.

Results

Table 1 displays the characteristics of participants in the meta-analysis. The proportion of participants reporting current smoking ranged from 4% to 33% across the different study populations. The characteristics of the participants within each cohort are provided in Table I in the Data Supplement.

Current Versus Never Smokers

In the meta-analysis of current cigarette smokers (n=2433) versus never smokers (n=6956), 2623 CpGs annotated to 1405 genes met Bonferroni significance after correction for 485 381 tests ($P<1\times10^{-7}$). On the basis of genome-wide FDR< 0.05, 18 760 CpGs annotated to 7201 genes were differentially methylated. There was a moderate inflation factor[30] λ of 1.32 (Figure I in the Data Supplement), which is consistent with a large number of sites being impacted by smoking. Our results lend support to many previously reported loci,[7,8,11,13] including CpGs annotated to *AHRR*, *RARA*, *F2RL3*, and *LRRN3* (Table II in the Data Supplement). Not surprisingly, cg05575921 annotated to *AHRR*, the top CpG identified in most previous studies of smoking, was highly significant in our meta-analysis ($P=4.6\times10^{-26}$; ranked 36, Table II in the Data Supplement) and also had the largest effect size (−18% difference in methylation), which is comparable to effect sizes in previous studies.[18] Of the 18 760 significant CpGs at FDR <0.05, 16 673 (annotated to 6720 genes) have not been previously reported to be associated with cigarette smoking—these include 1500 of the 2623 CpGs that met Bonferroni significance. The 25 CpGs with lowest *P* values for both overall and novel findings are shown in Table 2. Table II in the Data Supplement provides the complete list of all CpGs that were significantly differentially methylated (FDR <0.05) in analysis of current versus never smokers. Adding body mass index into the model did not appreciably alter the results (Figure II in the Data Supplement).

Methylation can be either reduced or increased at CpG sites in response to smoking. For the 53.2% of FDR-significant CpGs with increased methylation in response to current smoking, the mean percentage difference in methylation between current and never smokers was 0.5% (SD=0.37%; range, 0.06–7.3%). For 46.8% of CpGs with decreased methylation in response to current smoking, the mean percentage difference was 0.65% (SD=0.56; range, 0.04–18%) The volcano plot can be found in Figure III in the Data Supplement.

We did not observe correlation between the number of significant CpGs and either the size of the gene or the number of exons or the coverage of the methylation platform. We performed a formal enrichment test for each of the 7201 genes in regard to the length of the gene or number of exons and found only 3 for which associations were observed (*AHRR*, *PRRT1*, and *TNF*). However, given the robust findings for a specific CpG in *AHRR* in multiple studies in the literature[4,7,9] and our own, and its key role in the AHR pathway, which is crucial in response to polyaromatic hydrocarbons, such as are produced by smoking,[31] it seems unlikely that the *AHRR* findings are false positives. Likewise, there is strong support in the literature for *PRRT1*[32] and *TNF*.[33] The enrichment results for methylation platform coverage also yielded the same 3 genes.

In a subset of 3 cohorts (1827 subjects), we investigated the association of the number of pack years smoked with the 18 760 CpGs that were differentially methylated (FDR <0.05) between current versus never smokers. Significant dose responses were observed for 11 267 CpGs (60.1%) at FDR <0.05 (Table III in the Data Supplement).

To investigate the pathways implicated by these genes, we performed a gene-set enrichment analysis[34] on the annotated genes. The results suggested that cigarette smoking is associated with potential changes in numerous vital molecular processes, such as signal transduction ($FDR=2.8\times10^{-79}$), protein metabolic processes ($FDR=1.2\times10^{-43}$), and transcription pathways ($FDR=8.4\times10^{-31}$). The complete list of 99 enriched molecular processes can be found in Table IV in the Data Supplement.

Former Versus Never Smokers

Meta-analysis of former (n=6518) versus never smokers (n=6956) restricted to the 18 760 CpG sites that were differentially methylated in current versus never smokers identified 2568 CpGs annotated to 1326 genes at FDR <0.05 (Table V in the Data Supplement). There were 185 CpGs (annotated to 149 genes) that also met Bonferroni correction ($P<0.05/18760\approx2.67\times10^{-6}$). There was no evidence of inflation[30] (λ=0.98) (Figure IV in the Data Supplement). We also confirmed previously reported findings for CpGs annotated to *AHRR*, *RARA*, and *LRRN3*.[7,8,11,13] Effect sizes of these CpGs were all weaker than that in the analysis of current versus never smokers (61.2%±15.3% weaker) for the 2568 CpGs

Table 1. Participant Characteristics

Characteristics	Current Smokers, n=2433	Former Smokers, n=6518	Never Smokers, n=6956
Sex (% men)	46.3	55.6	31.7
Age, y*	57.7±7.7	64.8±8.2	61.2±9.7
BMI, kg/m^2*	27.3±5.4	28.7±5.0	28.6±5.3

BMI indicates body mass index.
*Weighted mean±pooled SD across cohorts

that remained significantly differentially methylated in former versus never smokers compared with current versus never smokers. Results for the top 25 CpGs are displayed in Table 3. Adding body mass index to the model did not appreciably alter the results (Figure V in the Data Supplement). A volcano plot can be found in Figure VI in the Data Supplement. In a subset of 3 cohorts (3349 subjects), analyses using pack years confirmed a significant dose response for 1804 of the 2568 CpGs (70%) annotated to 942 genes at FDR <0.05 (Table VI in the Data Supplement).

The gene-set enrichment analysis[27] in the former versus never smoker analyses on all 1326 genes revealed enrichment for genes associated with protein metabolic processes (FDR=1.1×10^{-23}), RNA metabolic processes (FDR=1.4×10^{-17}), and transcription pathways (FDR=3.9×10^{-18}; Table VII in the Data Supplement). The gene-set enrichment analysis on the 942 genes for which the 1804 CpGs exhibited dose responses with pack years also revealed similar pathways to those summarized in Table VII in the Data Supplement, except with weaker enrichment FDR values.

In 2648 Framingham Heart Study participants with ≤30 years of prospectively collected smoking data, we examined the 2568 CpGs that were differentially methylated in meta-analysis of former versus never smokers and explored their associations with time since smoking cessation. Methylation levels of most CpGs returned toward that of never smokers within 5 years of smoking cessation. However, 36 CpGs annotated to 19 genes, including *TIAM2*, *PRRT1*, *AHRR*, *F2RL3*, *GNG12*, *LRRN3*, *APBA2*, *MACROD2*, and *PRSS23*, did not return to never-smoker levels even after 30 years of smoking cessation (Figure; Table 4).

The EPIC studies included cancer cases plus noncancer controls analyzed together, adjusting for cancer status. The other studies were population-based samples not selected for disease status. To evaluate residual confounding by cancer status after adjustment, we repeated the meta-analysis without the EPIC studies. The effect estimates were highly correlated: Pearson ρ=0.99 for current versus never smoking and 0.98 for former smoking versus never.

Enrichment Analysis for Genes Identified in GWAS of Smoking-Related Phenotypes

To identify potential relevance of the differentially methylated genes to smoking-related phenotypes, we determined whether these genes had been associated with smoking-related phenotypes in the National Human Genome Research Institute-EBI GWAS Catalog[28] (accessed November 2, 2015). The catalog contained 9777 SNPs annotated to 7075 genes associated with 865 phenotypes at $P\leq5\times10^{-8}$. Of the 7201 genes (mapped by 18 760 CpG sites) significantly differentially methylated in current versus never smokers, we found overlap with 1791 genes (4187 CpGs are mapped to these) associated in GWAS with 700 phenotypes (enrichment P=2.4×10^{-52}). We identified smoking-related traits using the 2014 US Surgeon General's report.[1] Enrichment results for a selection of smoking-related phenotypes, including coronary heart disease and its risk factors, various cancers, inflammatory diseases, osteoporosis, and pulmonary traits, are available in Table 5. We also performed the same enrichment analysis on the 2568 CpGs associated with former versus never-smoking status. We identified enrichment for coronary heart disease, pulmonary traits, and some cancers (Table 5). More detailed results are available in Tables VIII and IX in the Data Supplement. Differentially methylated genes in relation to smoking status that are associated in GWAS with coronary heart disease or coronary heart disease risk factors are available in Table X in the Data Supplement. We also performed enrichment analyses on phenotypes that have no clear relationships to smoking, such as male pattern baldness (P=0.0888), myopia (P=0.1070), thyroid cancer (P=0.2406), and testicular germ cell tumor (P=0.3602) and did not find significant enrichment.

Enrichment Analysis for Genomic Features

We examined the differentially methylated CpGs with respect to localization to different genomic regions including CpG islands, gene bodies, known differentially methylated regions, and sites identified as likely to be functionally important in the ENCODE project such as DNAse1 hypersensitivity sites and enhancers (refer to the Methods section for details). We performed this analysis separately for the CpGs related to current smoking and past smoking (Table XI in the Data Supplement). Trends were similar for the 2 sets of CpGs, although the power to identify enrichment was much greater for the larger set of 18 760 CpGs related to current smoking. There was no enrichment for CpG islands. In contrast, significant enrichment was observed for island shores, gene bodies, DNAse1 hypersensitivity sites, and enhancers.

Transcriptomic Integration

Of the 18 760 statistically significant CpG sites associated with current smoking in the meta-analysis, 1430 were significantly associated in *cis* with the expression of 924 genes at FDR <0.05 (enrichment P=3.6×10^{-215}; Table XII in the Data Supplement) using whole-blood samples from 2262 Framingham Heart Study participants. Of these, 424 CpGs associated with the expression of 285 genes were replicated at FDR <0.0001 in 1264 CD14$^+$ samples from the MESA.[35] These

Table 2. Most Statistically Significant CpG Sites That Were Associated With Current Vs Never-Smoker Status

Probe ID	Chromosome	Location	Gene Symbol*	Regression Coefficients	SE	P	FDR
25 most significant CpG sites							
cg16145216	1	42 385 662	*HIVEP3*	0.0298	0.0020	6.7×10^{-48}	3.3×10^{-42}
cg19406367	1	66 999 929	*SGIP1*	0.0175	0.0013	7×10^{-44}	1.7×10^{-38}
cg05603985	1	2 161 049	*SKI*	−0.0122	0.0009	1.8×10^{-43}	2.8×10^{-38}
cg14099685	11	47 546 068	*CUGBP1*	−0.0124	0.0009	1.5×10^{-42}	1.8×10^{-37}
cg12513616	5	177 370 977	—	−0.0262	0.0020	6.1×10^{-41}	5.9×10^{-36}
cg03792876†	16	73 243	—	−0.0182	0.0014	7.2×10^{-38}	5.9×10^{-33}
cg01097768	5	378 854	*AHRR*	−0.0166	0.0013	6.8×10^{-35}	4.7×10^{-30}
cg26856289	1	24 307 516	*SFRS13A*	−0.0163	0.0013	8.6×10^{-35}	5.2×10^{-30}
cg07954423	9	130 741 881	*FAM102A*	−0.0134	0.0011	1.2×10^{-34}	6.3×10^{-30}
cg01940273	2	233 284 934	—	−0.0815	0.0067	2×10^{-34}	9.8×10^{-30}
cg01083131	16	67 877 413	*THAP11;CENPT*	−0.0155	0.0013	3.7×10^{-34}	1.6×10^{-29}
cg01017464	18	47 018 095	*SNORD58A; SNORD58B; RPL17*	−0.0172	0.0014	1.9×10^{-33}	7.6×10^{-29}
cg06121808	2	113 404 678	*SLC20A1*	−0.0143	0.0012	2.1×10^{-32}	7.9×10^{-28}
cg10062919	17	38 503 802	*RARA*	−0.0128	0.0011	9.2×10^{-32}	3.2×10^{-27}
cg20066188	22	37 678 791	*CYTH4*	−0.0252	0.0022	1.6×10^{-31}	5.2×10^{-27}
cg04551776	5	393 366	*AHRR*	−0.0244	0.0021	5.8×10^{-31}	1.8×10^{-26}
cg11152412	15	74 927 688	*EDC3*	−0.0077	0.0007	1.8×10^{-30}	5×10^{-26}
cg00073090	19	1 265 879	—	−0.0196	0.0017	4.2×10^{-30}	1.1×10^{-25}
cg11902777	5	368 843	*AHRR*	−0.0201	0.0018	9.1×10^{-30}	2.3×10^{-25}
cg25212453	17	1 509 953	*SLC43A2*	−0.0101	0.0009	1.4×10^{-29}	3.5×10^{-25}
cg04956244	17	38 511 592	*RARA*	0.0122	0.0011	1.5×10^{-29}	3.5×10^{-25}
cg13951797	16	2 204 381	*TRAF7*	−0.0153	0.0014	1.6×10^{-29}	3.5×10^{-25}
cg11028075	10	97 200 911	*SORBS1*	0.0175	0.0016	1.7×10^{-29}	3.6×10^{-25}
cg11700584†	14	50 088 544	*RPL36AL;MGAT2*	−0.0151	0.0013	3.4×10^{-29}	6.8×10^{-25}
cg11263997	11	70 257 280	*CTTN*	0.0050	0.0005	4.3×10^{-29}	8.4×10^{-25}
25 most significant novel CpG sites							
cg11700584	14	50 088 544	*RPL36AL; MGAT2*	−0.0151	0.0013	3.4×10^{-29}	6.8×10^{-25}
cg22417733	6	153 303 409	*FBXO5*	−0.0171	0.0015	1.5×10^{-28}	2.7×10^{-24}
cg08118908	16	15 787 920	*NDE1*	0.0053	0.0005	5.4×10^{-26}	7.1×10^{-22}
cg14003265	9	139 796 499	*TRAF2*	−0.0106	0.0010	3.2×10^{-25}	3.7×10^{-21}
cg02556393	3	168 866 705	*MECOM*	−0.0162	0.0016	2.8×10^{-24}	2.6×10^{-20}
cg01218206	11	116 933 977	*SIK3*	−0.0150	0.0015	3.1×10^{-23}	2.5×10^{-19}
cg04987734	14	103 415 873	*CDC42BPB*	0.0149	0.0015	9.0×10^{-23}	6.8×10^{-19}
cg27118035	16	31 891 978	*ZNF267*	0.0136	0.0014	2.4×10^{-22}	1.7×10^{-18}
cg18450254	3	64 200 005	*PRICKLE2*	0.0120	0.0013	2.3×10^{-21}	1.3×10^{-17}
cg06753787	2	220 074 208	*ZFAND2B*	0.0063	0.0007	3.2×10^{-21}	1.8×10^{-17}
cg18158306	12	133 135 032	*FBRSL1*	0.0102	0.0011	6.2×10^{-21}	3.2×10^{-17}
cg19093370	17	17 110 180	*PLD6*	0.0198	0.0021	8.7×10^{-21}	4.4×10^{-17}
cg09182189	1	1 709 203	*NADK*	−0.0104	0.0011	2.0×10^{-20}	9.2×10^{-17}
cg18369990	2	112 941 244	*FBLN7*	0.0116	0.0013	2.3×10^{-20}	1.1×10^{-16}
cg24578857	17	17 110 207	*PLD6*	0.0200	0.0022	3.1×10^{-20}	1.4×10^{-16}

(*Continued*)

Table 2. Continued

Probe ID	Chromosome	Location	Gene Symbol*	Regression Coefficients	SE	P	FDR
cg20408402	10	72 362 452	*PRF1*	0.0085	0.0009	7.6×10^{-20}	3.1×10^{-16}
cg04673446	22	39 879 951	*MGAT3*	0.0060	0.0007	2.0×10^{-19}	8.0×10^{-16}
cg06803614	1	40 133 581	*NT5C1A*	−0.0088	0.0010	2.1×10^{-19}	8.3×10^{-16}
cg16274678	1	154 127 952	*TPM3; NUP210L*	−0.0152	0.0017	2.9×10^{-19}	1.1×10^{-15}
cg07286341	5	176 923 805	*PDLIM7*	−0.0077	0.0009	3.4×10^{-19}	1.3×10^{-15}
cg20674424	3	186 503 527	*MIR1248; EIF4A2; SNORA81*	−0.0091	0.0010	4.2×10^{-19}	1.5×10^{-15}
cg02279625	15	78 384 520	*SH2D7*	0.0105	0.0012	4.8×10^{-19}	1.7×10^{-15}
cg03485667	16	75 143 200	*ZNRF1*	−0.0168	0.0019	5.0×10^{-19}	1.8×10^{-15}
cg03531211	6	32 920 102	*HLA-DMA*	−0.0108	0.0012	7.5×10^{-19}	2.5×10^{-15}
cg09940677	14	103 415 458	*CDC42BPB*	0.0081	0.0009	1.0×10^{-18}	3.2×10^{-15}

CpG indicates cytosine–phosphate–guanine; and FDR, false discovery rate.
*CpG sites without gene names are intergenic. These are all included in all the analyses.
†Not previously discovered by other studies.

genes are associated with pathways similar to those described earlier (Table XIII in the Data Supplement).

Comparison Between AA and EA

Meta-analysis of the current versus never smokers in 11 cohorts with participants of EA (n=6750 subjects) yielded 10 977 CpGs annotated to 4940 genes at FDR <0.05. Meta-analysis of the results of the smaller data set of 4 cohorts with AA participants (n=2639) yielded 3945 CpGs annotated to 2088 genes at FDR <0.05. The effect estimates of the CpGs significant in at least one ancestry (12 927 CpGs) were highly correlated in the combined group of individuals of either ancestry (Spearman ρ=0.89). The results by ancestry are shown in Table XIV in the Data Supplement.

We performed the same ancestry-stratified analysis on former versus never smokers (Table XV in the Data Supplement). Meta-analysis of the results of EA participants yielded 2045 CpG sites annotated to 1081 genes at FDR <0.05. Meta-analysis of the results of AA participants yielded 329 CpG sites annotated to 178 genes at FDR <0.05. The effect estimates of the union of CpGs significant in at least one ancestry (2234 CpGs) were correlated in the combined group of individuals of either ancestry (Spearman ρ=0.75). Of note, one of CpG sites showing differential methylation in ancestry, cg00706683, mapped to gene *ECEL1P2*, did not return to never-smoker levels 30 years after smoking cessation (Table 4).

To more directly compare results by ethnicity, removing the effect of better statistical power in the larger EA sample size, we performed a meta-analysis on subset of EA cohorts: the Framingham Heart Study, Rotterdam Study, and KORA, such that the total number of smokers, the major determinant of power, would match that of AA cohorts. In this subset, similar correlations of the effect estimates were observed as in the complete analyses, suggesting that the differences in number of statistically significant CpGs are indeed because of better power in the EA cohorts (Spearman ρ=0.87 and 0.79 for current versus never smokers and former versus never smokers, respectively).

Cell Type Adjustment

We adjusted our main analyses for white blood cell fractions, in studies based on either whole blood or leukocytes from the buffy coat of whole blood, either measured or using a published method.[20] Reassuringly, results before and after cell type adjustment were highly comparable. The correlation of regression coefficients before and after adjustment is 0.85 for the current versus never-smoker analysis (Figure VII in the Data Supplement). Similarly for the analysis of former versus never smokers, the effect estimates were highly correlated before and after adjustment (ρ=0.93; Figure VIII in the Data Supplement). In addition, in 2 cohorts, we had results from specific cell fractions—CD4$^+$ cells in GOLDN and CD14$^+$ cells in MESA. The correlation of results between buffy coat and CD4$^+$ or CD14$^+$ for former versus never smokers are generally high (ρ>0.74; Table XVI in the Data Supplement).

Methylation Profile Across CpG Sites

We assessed methylation profile in FHS cohort as a representative cohort in the study. The profile of all 485 381 analyzed CpG sites can be found in Figure IX in the Data Supplement. The profile across 18 760 CpG sites significantly associated with current versus never smoking status can be found in Figure X in the Data Supplement. These plots indicate that most CpG sites with less dynamic range are largely not statistically significant in our results.

Discussion

We performed a genome-wide meta-analysis of blood-derived DNA methylation in 15 907 individuals across 16 cohorts and identified broad epigenome-wide impact of cigarette smoking, with 18 760 statistically significant CpGs (FDR <0.05) annotated to >7000 genes, or roughly one third of known human genes. These genes in turn affect multiple molecular mechanisms and are implicated in smoking-related phenotypes and diseases. In addition to confirming previous findings from smaller studies, we detected >16 000 novel differentially methylated CpGs in response to cigarette smoking. Many of

Table 3. Twenty-Five Most Statistically Significant CpG Sites That Were Associated With Former Versus Never Smoker Status

Probe ID	Chromosome	Location	Gene Symbol*	Regression Coefficients	SE	P	FDR
cg01940273	2	233 284 934		−0.0234	0.0013	9.6×10^{-73}	1.8×10^{-68}
cg25189904	1	68 299 493	*GNG12*	−0.0283	0.0021	3.5×10^{-40}	3.3×10^{-36}
cg12803068	7	45 002 919	*MYO1G*	0.0191	0.0017	9.3×10^{-31}	5.8×10^{-27}
cg19572487	17	38 476 024	*RARA*	−0.0159	0.0014	2.2×10^{-30}	1.0×10^{-26}
cg11554391	5	321 320	*AHRR*	−0.0091	0.0008	1.0×10^{-28}	3.9×10^{-25}
cg05951221	2	233 284 402	—	−0.0396	0.0036	1.1×10^{-27}	3.2×10^{-24}
cg23771366	11	86 510 998	*PRSS23*	−0.0167	0.0015	1.2×10^{-27}	3.2×10^{-24}
cg26764244	1	68 299 511	*GNG12*	−0.0119	0.0011	2.3×10^{-27}	5.4×10^{-24}
cg05575921	5	373 378	*AHRR*	−0.0406	0.0038	8.2×10^{-27}	1.7×10^{-23}
cg11660018	11	86 510 915	*PRSS23*	−0.0157	0.0015	4.3×10^{-26}	8.1×10^{-23}
cg21566642	2	233 284 661	—	−0.0434	0.0041	1.0×10^{-25}	1.7×10^{-22}
cg11902777	5	368 843	*AHRR*	−0.0063	0.0006	2.8×10^{-25}	4.3×10^{-22}
cg26850624	5	429 559	*AHRR*	0.0118	0.0011	3.1×10^{-25}	4.4×10^{-22}
cg03636183	19	17 000 585	*F2RL3*	−0.0267	0.0026	8.9×10^{-25}	1.2×10^{-21}
cg15693572	3	22 412 385	—	0.0190	0.0019	1.5×10^{-23}	1.9×10^{-20}
cg17924476	5	323 794	*AHRR*	0.0148	0.0016	4.0×10^{-20}	4.7×10^{-17}
cg12513616	5	177 370 977	—	−0.0072	0.0008	2.4×10^{-19}	2.7×10^{-16}
cg07339236	20	50 312 490	*ATP9A*	−0.0062	0.0007	1.4×10^{-18}	1.4×10^{-15}
cg06126421	6	30 720 080	—	−0.0365	0.0042	3.0×10^{-18}	3.0×10^{-15}
cg14624207	11	68 142 198	*LRP5*	−0.0070	0.0008	5.0×10^{-18}	4.7×10^{-15}
cg00706683	2	233 251 030	*ECEL1P2*	0.0101	0.0012	1.4×10^{-17}	1.2×10^{-14}
cg23351584	11	86 512 100	*PRSS23*	−0.0048	0.0006	7.0×10^{-17}	6.0×10^{-14}
cg02583484	12	54 677 008	*HNRNPA1*	−0.0062	0.0008	1.0×10^{-15}	8.5×10^{-13}
cg05302489	6	31 760 426	*VARS*	0.0079	0.0010	2.5×10^{-15}	2.0×10^{-12}
cg01442064	4	5 713 450	*EVC*	−0.0055	0.0007	3.3×10^{-15}	2.4×10^{-12}

CpG indicates cytosine–phosphate–guanine; and FDR, false discovery rate.

*CpG sites without gene names are intergenic. These are all included in all the analyses.

these genes have not been previously implicated in the biological effects of tobacco exposure. The large number of genes implicated in this well-powered meta-analysis might on first glance raise concerns about false positives. However, on further consideration, given the widespread impact of smoking on disease outcomes across many organ systems and across the life span,[1] the identification of a large number of genes at genome-wide significance is not surprising. In addition, our findings are robust and consistent across all 16 cohorts (Tables II and V in the Data Supplement) because we accounted for interstudy variability by using random-effect meta-analyses, which is conservative when heterogeneity is present.[36] The implicated genes are mainly involved in molecular machineries, such as transcription and translation. Furthermore, differential methylation of a subset of CpGs persisted, often for decades, after smoking cessation.

We found that genes differentially methylated in relation to smoking are enriched for variants associated in GWAS with smoking-related diseases,[1] including osteoporosis, colorectal cancers, chronic obstructive pulmonary disease, pulmonary function, cardiovascular disease, and rheumatoid arthritis. We find it noteworthy that there is enrichment of smoking-associated CpGs for genes associated with rheumatoid arthritis because DNA methylation is one of the proposed molecular mechanisms underlying this disease.[37] It is also interesting that the most significant association of smoking with methylation was for the gene *HIVEP3* (a.k.a. Schnurri3), the mammalian homolog of the Drosophila zinc finger adapter protein Shn.[38] This gene regulates bone formation, an important determinant to osteoporosis, which was one of the enriched GWAS phenotypes.

When we examined time since smoking cessation, we found that the majority of the differentially methylated CpG sites observed in analysis of current versus never smokers returned to the level of never smokers within 5 years of smoking cessation. This is consistent with the fact that risks of many smoking-related diseases revert to nonsmoking levels within this period of time. Our results also indicate that cigarette smoking induces long-lasting alterations in DNA methylation at some CpGs. Although speculative, it is possible that

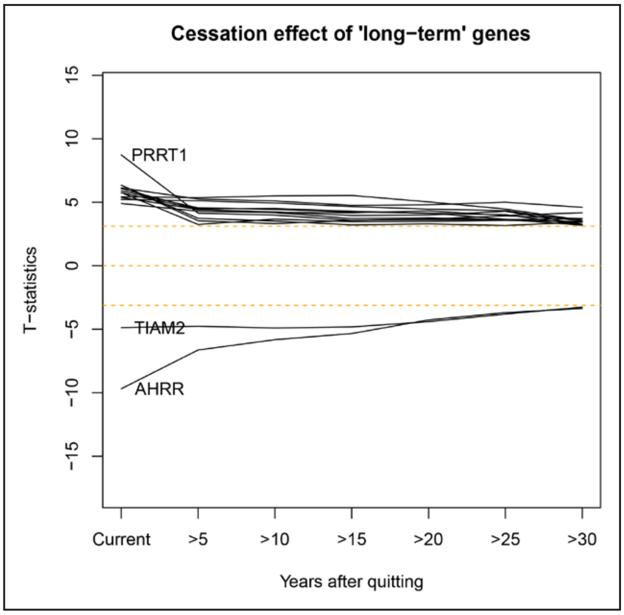


Figure. Trajectories of cytosine–phosphate–guanine (CpG) sites that did not return to never-smoker levels within 30 y after cessation.

persistent methylation changes at some loci might contribute to risks of some conditions that remain elevated after smoking cessation.

In all but 2 of our 14 cohorts, DNA was extracted from the entire circulating leukocyte population. Thus, there is the possibility of confounding by the effects of smoking on differential cell counts. We attempted to adjust for cell type and found that results were generally little changed by the adjustment.

Our significant results are highly enriched for CpG sites associated with the expression of nearby genes (ie, in *cis*) even though a single measurement of gene expression in blood is probably subject to considerably more within-subject variability than DNA methylation,[39] limiting our ability to find correlations. Differential DNA methylation at many of the CpGs we identified in relation to smoking status may have a functional impact on nearby gene expression. Our analysis of genomic regions further supports the potential functional impact of our findings on gene expression. We demonstrated enrichment for sites with greater functional impact, such as island shores, gene bodies, DNAse1 hypersensitivity sites, and enhancers, whereas we found no enrichment for CpG islands. These results reinforce previous findings showing that island shores, enhancers, and DNAse I hypersensitive sites are more dynamic (ie, susceptible to methylation changes) than CpG islands,[40] which may be more resistant to abrupt changes in DNA methylation in response to environmental exposures.[41] Thus, our results suggest that many of the smoking-associated CpG sites may have regulatory effects.

Although identification of changes in methylation patterns may suggest mechanisms by which exposure to tobacco smoke exerts its effects on several disease processes, DNA methylation profiles can also serve as biomarkers of exposure to tobacco smoke. Cotinine is a biomarker only of recent smoking; DNA methylation signals have the potential to serve as robust biomarkers of smoking history.[9,42] Indeed, several studies have identified several of such markers.[5,42,43] The large number of persistently modified CpGs may be

Table 4. The Top 36 Most Statistically Significant CpG Sites That Did Not Return to Never-Smoker Levels 30 Y After Smoking Cessation in the Framingham Heart Study (n=2648)

Probe ID	Chromosome	Location	Gene Symbol	P
cg05951221	2	233 284 402	—	3.2×10^{-15}
cg06644428	2	233 284 112	—	1.2×10^{-14}
cg05575921	5	373 378	*AHRR*	6.5×10^{-14}
cg21566642	2	233 284 661	—	8.6×10^{-10}
cg03636183	19	17 000 585	*F2RL3*	5.7×10^{-7}
cg06126421	6	30 720 080	—	1.3×10^{-6}
cg01940273	2	233 284 934	—	1.9×10^{-6}
cg23771366	11	86 510 998	*PRSS23*	3.1×10^{-6}
cg17272563	6	321 16 548	*PRRT1*	4.4×10^{-6}
cg23916896	5	368 804	*AHRR*	1.3×10^{-5}
cg11660018	11	86 510 915	*PRSS23*	1.3×10^{-5}
cg08118908	16	15 787 920	*NDE1*	3.0×10^{-5}
cg13937905	12	53 612 551	*RARG*	1.5×10^{-4}
cg24172324	2	232 258 363	—	1.7×10^{-4}
cg10780313	6	33 501 379	—	2.0×10^{-4}
cg14027333	6	32 116 317	*PRRT1*	2.1×10^{-4}
cg11245297	19	8 117 898	*CCL25*	2.1×10^{-4}
cg01692968	9	108 005 349	—	3.1×10^{-4}
cg00706683	2	233 251 030	*ECEL1P2*	3.4×10^{-4}
cg25317941	2	233 351 153	*ECEL1*	4.0×10^{-4}
cg25189904	1	68 299 493	*GNG12*	4.0×10^{-4}
cg14179389	1	92 947 961	*GFI1*	4.7×10^{-4}
cg13641317	3	127 255 552	—	4.9×10^{-4}
cg19847577	15	29 213 748	*APBA2*	5.1×10^{-4}
cg14239618	7	110 281 356	—	5.8×10^{-4}
cg25955180	6	32 116 538	*PRRT1*	6.3×10^{-4}
cg00774149	3	522 55 721	*TLR9*	6.4×10^{-4}
cg21351392	6	161 607 487	*AGPAT4*	7.1×10^{-4}
cg11902777	5	368 843	*AHRR*	7.6×10^{-4}
cg07251887	17	73 641 809	*LOC100130933;* *RECQL5*	7.7×10^{-4}
cg19382157	7	2 124 566	*MAD1L1*	8.9×10^{-4}
cg19925780	1	101 509 557	—	1.1×10^{-3}
cg03679544	6	155 537 972	*TIAM2*	1.1×10^{-3}
cg08559712	20	16 030 674	*MACROD2*	1.3×10^{-3}
cg09837977	7	110 731 201	*LRRN3; IMMP2L*	1.3×10^{-3}
cg00931843	6	155 442 993	*TIAM2*	1.4×10^{-3}

CpG indicates cytosine–phosphate–guanine.

*CpG sites without gene names are intergenic. These are all included in all the analyses.

Table 5. Enrichment of CpGs for Genome-Wide Association Study Phenotypes That Are Regarded as Causally Related to Cigarette Smoking[1]

GWAS Phenotype	Enrichment *P* Value
Current vs never smoking	
CHD and stroke	0.0028
Ischemic stroke	0.0095
CHD risk factors	1.2×10^{-12}
Blood pressure/hypertension	8.1×10^{-6}
Diastolic blood pressure	6.1×10^{-5}
Systolic blood pressure	0.0008
Hypertension	0.0150
Lipids	2.9×10^{-5}
High-density lipoprotein	0.0009
Type 2 diabetes mellitus	0.0106
Rheumatoid arthritis	2.9×10^{-5}
Bone mineral density and osteoporosis	0.0467
All pulmonary traits	2.8×10^{-6}
All COPD	0.0295
Moderate-to-severe COPD	0.0156
Pulmonary function	0.0044
Crohn disease	9.5×10^{-7}
Primary biliary cirrhosis	3.4×10^{-6}
Inflammation bowel disease	3.5×10^{-5}
Ulcerative colitis	9.8×10^{-5}
All cancer	8.0×10^{-15}
Lung adenocarcinoma	0.0015
Colorectal cancer	0.0014
Former vs never smoking	
CHD risk factors	7.6×10^{-5}
Blood pressure/hypertension	5.8×10^{-5}
Diastolic blood pressure	0.0021
Systolic blood pressure	0.0002
Hypertension	0.0023
Rheumatoid arthritis	6.3×10^{-5}
All pulmonary traits	0.0217
Inflammation bowel disease	5.2×10^{-6}
Crohn disease	0.0064
All cancer	7.8×10^{-6}

CHD indicates coronary heart disease; COPD, chronic obstructive pulmonary disease; and CpG, cytosine–phosphate–guanine.

useful to develop even more robust biomarkers to objectively quantify long-term cigarette-smoking exposure for prediction of risk for health outcomes in settings where smoking history is not available or is incomplete and to validate self-reported never-smoker status. Furthermore, our analyses of both former and current smokers show dose-dependent effects at many CpGs (Tables III and VII in the Data Supplement). Methylation-based biomarkers could be informative for investigating dose–response relationships with disease end points. This is useful because smokers often under-report the amount of smoking, both current and historical.

It is possible that smoking-related conditions or correlated exposures may contribute to some of the methylation signatures identified. However, our studies are nearly all population-based studies composed of predominantly healthy individuals, not selected for smoking-related disease. Given the number, strength, and robustness to replication of findings for smoking across the literature and among our diverse cohorts from various countries, the likelihood that these are confounded by other exposures or conditions related to smoking is greatly reduced.

There are several potential limitations to our study. First, the cross-sectional design limits our ability to study the time course of smoking effects. In addition, we analyzed methylation in DNA samples from blood, which is readily accessible. Although we demonstrated that blood-derived DNA reveals a strong and robust signature of cigarette-smoking exposure, studies in target tissues for smoking-related diseases (eg, heart and lung) would be of additional interest. In addition, our analyses could not distinguish direct effects of smoking from indirect effects of smoking because of smoking-induced changes in cell metabolism, organ function, inflammation, or injury that could in turn influence methylation. However, this is the largest examination to date of the effects of smoking on DNA methylation with 16 studies from different countries contributing.

In conclusion, we identify an order of magnitude more sites differentially methylated in relation to smoking across the genome than have been previously seen. Many of these signals persist long after smoking cessation, providing potential biomarkers of smoking history. These findings may provide new insights into molecular mechanisms underlying the protean effects of smoking on human health and disease.

Acknowledgments

We would like to thank Bonnie R. Joubert, PhD, and Frank Day, PhD, of the National Institute of Environmental Health Sciences (Research Triangle Park, NC) and Jianping Jin, PhD, of Westat (Durham, NC) for expert computational assistance. Additional Acknowledgments can be found in the Data Supplement.

Sources of Funding

Infrastructure for the CHARGE consortium is provided by the National Heart, Lung, and Blood Institute grant R01HL105756. This work was supported in part by the Intramural Research Program of the NIH; National Institute of Environmental Health Sciences; and the National Heart Lung and Blood Institute. Additional funding sources for each cohort can be found in the Data Supplement.

Disclosures

B.M. Psaty serves on Data Safety Monitoring Board (DSMB) of a clinical trial of a device funded by the manufacturer (Zoll LifeCor) and on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. C.E. Elks is currently employed by Astra Zeneca, although the work was completed before the employment. The other authors report no conflicts.

Appendix

From the Institute for Aging Research, Hebrew SeniorLife (R.J., D.P.K.), Department of Medicine, Beth Israel Deaconess Medical Center (R.J., D.P.K.), Channing Division of Network Medicine, Brigham and Women's Hospital (D.L.D.), and Department of Psychiatry (K.J.R.), Harvard Medical School, Boston, MA; Population Sciences Branch, National Heart, Lung, and Blood Institute (R.J., T.H., C.L., M.M.M., C.Y., D.L.) and Laboratory of Neurogenetics, National Institute on Aging (D.G.H., A.B.S.), National Institutes of Health, Bethesda, MD; Framingham Heart Study, MA (R.J., T.H., C.L., M.M.M., C.Y., D.L.); Department of Preventive Medicine, Icahn School of Medicine at Mount Sinai, New York, NY (A.C.J.); Centre for Cognitive Ageing and Cognitive Epidemiology (R.E.M., P.M.V., J.M.S., I.J.D.), Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine (R.E.M.), Alzheimer Scotland Dementia Research Centre (J.M.S.), and Department of Psychology, University of Edinburgh (I.J.D.), United Kingdom; Queensland Brain Institute (R.E.M., R.H.S., A.F.M., P.M.V., N.R.W.) and University of Queensland Diamantina Institute, Translational Research Institute (A.F.M., P.M.V.), University of Queensland, Brisbane, Australia; Epidemiology and Public Health Group, Institute of Biomedical and Clinical Science, University of Exeter Medical School, United Kingdom (L.C.P., D.M.); Department of Epidemiology and Prevention, Division of Public Health Sciences (L.M.R., C.J.R., Y.L.), Department of Biostatistical Sciences, Division of Public Health Sciences (K.L.), and Department of Internal Medicine (J.D.), Wake Forest School of Medicine, Winston-Salem, NC; Department of Internal Medicine (P.R.M., A.G.U., J.B.J.v.M.), Department of Clinical Chemistry (P.R.M.), and Department of Epidemiology (A.H.), Erasmus University Medical Center, Rotterdam, The Netherlands; Division of Biostatistics (W.G.) and Division of Epidemiology and Community Health (J.S.P.), School of Public Health, University of Minnesota, Minneapolis; Research Unit of Molecular Epidemiology, Institute of Epidemiology II, Helmhotz Zentrum Muenchen, Munich, Germany (T.X., S.K., A.P., R.W.-S., M.W.); MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, United Kingdom (C.E.E., N.J.W., K.K.O.); Department of Epidemiology, University of Alabama at Birmingham (S.A., J.S., M.R.I., D.K.A.); Autonomous Metropolitan University-Iztapalapa, Mexico City, Mexico (H.M.-M.); International Agency for Research on Cancer, Lyon, France (H.M.-M., S.A., Z.H., I.R.); Department of Epidemiology, School of Public Health (J.A.S., W.Z., E.B.W., S.L.R.K.) and Research Center for Group Dynamics, Institute for Social Research (E.B.W.), University of Michigan, Ann Arbor; Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services (J.A.B., B.M.P., B.R.S.), Center for Lung Biology, Division of Pulmonary and Critical Care Medicine, Department of Medicine (S.A.G.) and Cardiovascular Health Research Unit, Division of Cardiology, Department of Epidemiology (N.S.), University of Washington, Seattle; Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA (R.D.); School of Public Health, University of California, Berkeley (P.Y., E.W.D.); HudsonAlpha Institute for Biotechnology, Huntsville, AL (D.M.A.); Clinical Research Branch, National Institute on Aging, Baltimore, MD (L.F.); Human Genetics Center, School of Public Health (J.B., M.L.G., M.F.) and School of Biomedical Informatics (D.Z.), The University of Texas Health Science Center at Houston; Department of Cardiology, Boston Children's Hospital, Boston, MA (M.M.M.); Division of Cancer Epidemiology, German Cancer Research Center (DKFZ) Heidelberg (M.B.); MRC/PHE Centre for Environment and Health, School of Public Health, Imperial College London, United Kingdom (P.V.); HuGeF Foundation, Torino, Italy (P.V.); Department of Epidemiology (J.S., A.A.B.) and Department of Environmental Health (A.A.B.), Harvard T.H. Chan School of Public Health, Boston, MA; Department of Preventive Medicine and the Robert H. Lurie Comprehensive Cancer Center, Feinberg School of Medicine, Northwestern University, Chicago, IL (L.H.); VA Normative Aging Study, VA Boston Healthcare System & Department of Medicine, Boston University School of Medicine, Boston, MA (P.S.V.); Geriatric Unit, Azienda Sanitaria di Firenze, Florence, Italy (S.B.); Division of Nephrology & Hypertension, Mayo Clinic, Rochester, MN (S.T.T.); Department of Psychiatry and Behavioral Sciences (E.B.B., A.K.S., K.J.R.); Department of Human Genetics, Emory University School of Medicine, Atlanta, GA (K.N.C.); Department of Translational Research in Psychiatry, Max-Planck Institute of Psychiatry, Munich, Germany (T.K., E.B.B.); Division of Depression & Anxiety Disorders, McLean Hospital, Belmont, MA (T.K., K.J.R.); Group Health Research Institute, Group Health Cooperative, Seattle, WA (B.M.P.); Institute for Translational Genomics & Population Sciences, Los Angeles BioMedical Research Institute (K.D.T.), Division of Genomic Outcomes, Department of Pediatrics, Harbor-UCLA Medical Center, Torrance (K.D.T.); Departments of Pediatrics, Medicine, and Human Genetics, UCLA, Los Angeles, CA (K.D.T.); Harvard School of Public Health (L.L.); Boston University School of Medicine (G.T.O.); and Epidemiology Branch, Department of Health and Human Services, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC (S.J.L

References

1. National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Atlanta, GA: Centers for Disease Control and Prevention (US); 2014.
2. World Health Organization. WHO global report on trends in prevalence of tobacco smoking 2015. Available at http://apps.who.int/iris/bitstream/10665/156262/1/9789241564922_eng.pdf.
3. Szarc vel Szic K, Declerck K, Vidaković M, Vanden Berghe W. From inflammaging to healthy aging by dietary lifestyle choices: is epigenetics the key to personalized nutrition? *Clin Epigenetics*. 2015;7:33. doi: 10.1186/s13148-015-0068-2.
4. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet*. 2011;88:450–457. doi: 10.1016/j.ajhg.2011.03.003.
5. Breitling LP, Salzmann K, Rothenbacher D, Burwinkel B, Brenner H. Smoking, F2RL3 methylation, and prognosis in stable coronary heart disease. *Eur Heart J*. 2012;33:2841–2848. doi: 10.1093/eurheartj/ehs091.
6. 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;21:3073–3082. doi: 10.1093/hmg/dds135.
7. Wan ES, Qiu W, Carey VJ, Morrow J, Bacherman H, Foreman MG, et al. Smoking-associated site-specific differential methylation in buccal mucosa in the COPDGene study. *Am J Respir Cell Mol Biol*. 2015;53:246–254. doi: 10.1165/rcmb.2014-0103OC.
8. 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. doi: 10.1371/journal.pone.0063812.
9. 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;24:712–716. doi: 10.1097/EDE.0b013e31829d5cb3.
10. 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;22:843–851. doi: 10.1093/hmg/dds488.
11. 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;24:2349–2359. doi: 10.1093/hmg/ddu751.
12. Qiu W, Wan E, Morrow J, Cho MH, Crapo JD, Silverman EK, et al. The impact of genetic variation and cigarette smoke on DNA methylation in current and former smokers from the COPDGene study. *Epigenetics*. 2015;10:1064–1073. doi: 10.1080/15592294.2015.1106672.
13. 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. doi: 10.1186/s13148-015-0148-3.
14. Shah S, Bonder MJ, Marioni RE, Zhu Z, McRae AF, Zhernakova A, et al; BIOS Consortium. Improving phenotypic prediction by combining genetic and epigenetic associations. *Am J Hum Genet*. 2015;97:75–85. doi: 10.1016/j.ajhg.2015.05.014.

15. Beane J, Sebastiani P, Liu G, Brody JS, Lenburg ME, Spira A. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol*. 2007;8:R201. doi: 10.1186/gb-2007-8-9-r201.
16. Wauters E, Janssens W, Vansteenkiste J, Decaluwé H, Heulens N, Thienpont B, et al. DNA methylation profiling of non-small cell lung cancer reveals a COPD-driven immune-related signature. *Thorax*. 2015;70:1113–1122. doi: 10.1136/thoraxjnl-2015-207288.
17. Garg AX, Hackam D, Tonelli M. Systematic review and meta-analysis: when one study is just not enough. *Clin J Am Soc Nephrol*. 2008;3:253–260. doi: 10.2215/CJN.01430307.
18. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*. 2013;14:293. doi: 10.1186/1471-2164-14-293.
19. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*. 2013;29:189–196. doi: 10.1093/bioinformatics/bts680.
20. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13:86. doi: 10.1186/1471-2105-13-86.
21. Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aïssi D, Wahl S, et al. DNA methylation and body-mass index: a genome-wide analysis. *Lancet*. 2014;383:1990–1998. doi: 10.1016/S0140-6736(13)62674-4.
22. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67:1–48.
23. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Development Core Team; 2010.
24. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36:1–48.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JSSRB*. 1995;57:289–300. doi: 10.2307/2346101.
26. 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;98:680–696. doi: 10.1016/j.ajhg.2016.02.019.
27. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA*. 2004;101:6062–6067. doi: 10.1073/pnas.0400782101.
28. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*. 2009;106:9362–9367. doi: 10.1073/pnas.0903103106.
29. Vazquez AI, Bates DM, Rosa GJ, Gianola D, Weigel KA. Technical note: an R package for fitting generalized linear mixed models in animal breeding. *J Anim Sci*. 2010;88:497–504. doi: 10.2527/jas.2009-1952.
30. Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999;55:997–1004.
31. Martey CA, Baglole CJ, Gasiewicz TA, Sime PJ, Phipps RP. The aryl hydrocarbon receptor is a regulator of cigarette smoke induction of the cyclooxygenase and prostaglandin pathways in human lung fibroblasts. *Am J Physiol Lung Cell Mol Physiol*. 2005;289:L391–L399. doi: 10.1152/ajplung.00062.2005.
32. 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;1:476–485. doi: 10.1001/jamaoncol.2015.1053.
33. Campesi I, Carru C, Zinellu A, Occhioni S, Sanna M, Palermo M, et al. Regular cigarette smoking influences the transsulfuration pathway, endothelial function, and inflammation biomarkers in a sex-gender specific manner in healthy young humans. *Am J Transl Res*. 2013;5:497–509.
34. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005;102:15545–15550. doi: 10.1073/pnas.0506580102.
35. Liu Y, Ding J, Reynolds LM, Lohman K, Register TC, De La Fuente A, et al. Methylomics of gene expression in human monocytes. *Hum Mol Genet*. 2013;22:5065–5074. doi: 10.1093/hmg/ddt356.
36. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet*. 2011;88:586–598. doi: 10.1016/j.ajhg.2011.04.014.
37. 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;31:142–147. doi: 10.1038/nbt.2487.
38. Jones DC, Wein MN, Oukka M, Hofstaetter JG, Glimcher MJ, Glimcher LH. Regulation of adult bone mass by the zinc finger adapter protein Schnurri-3. *Science*. 2006;312:1223–1227. doi: 10.1126/science.1126313.
39. Suderman M, Pappas JJ, Borghol N, Buxton JL, McArdle WL, Ring SM, et al. Lymphoblastoid cell lines reveal associations of adult DNA methylation with childhood and current adversity that are distinct from whole blood associations. *Int J Epidemiol*. 2015;44:1331–1340. doi: 10.1093/ije/dyv168.
40. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature*. 2013;500:477–481. doi: 10.1038/nature12433.
41. Ivanova E, Chen JH, Segonds-Pichon A, Ozanne SE, Kelsey G. DNA methylation at differentially methylated regions of imprinted genes is resistant to developmental programming by maternal nutrition. *Epigenetics*. 2012;7:1200–1210. doi: 10.4161/epi.22141.
42. Zhang Y, Schöttker 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;124:67–74. doi: 10.1289/ehp.1409020.
43. 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;122:131–137. doi: 10.1289/ehp.1306937.

CLINICAL PERSPECTIVE

We combined data from 16 cohorts (15 907 individuals) examining genome-wide methylation, a type of epigenetic modification, in blood DNA, in relation to smoking status. In this large-scale meta-analysis, thousands of DNA methylation cytosine-P-guanine sites were associated with current versus never-smoking status. These methylation signals reside in genes that are associated with numerous diseases caused by cigarette smoking, such as cardiovascular diseases and certain cancers. Of the thousands of cytosine-P-guanine sites differentially methylated in current versus never smokers, >10% also were significantly associated with former versus never-smoking status. Although many of these former smoker methylation signals return to never-smoker levels with 5 years of quitting, a substantial proportion remain elevated even after 30 years of cessation. We also found widespread evidence that many differentially methylated sites also are related to gene expression, showing a functional impact on the genome. Furthermore, in our analyses, these cigarette-smoking DNA methylation signals affect genes important to fundamental molecular pathways, such as molecular signal transduction, protein metabolic processes, and transcription. In conclusion, cigarette smoking has a widespread and long-lasting impact on DNA methylation. DNA methylation is one potential mechanism by which tobacco exposure predisposes to numerous adverse health outcomes.

Circulation
Cardiovascular Genetics

American
Heart
Association®

Epigenetic Signatures of Cigarette Smoking

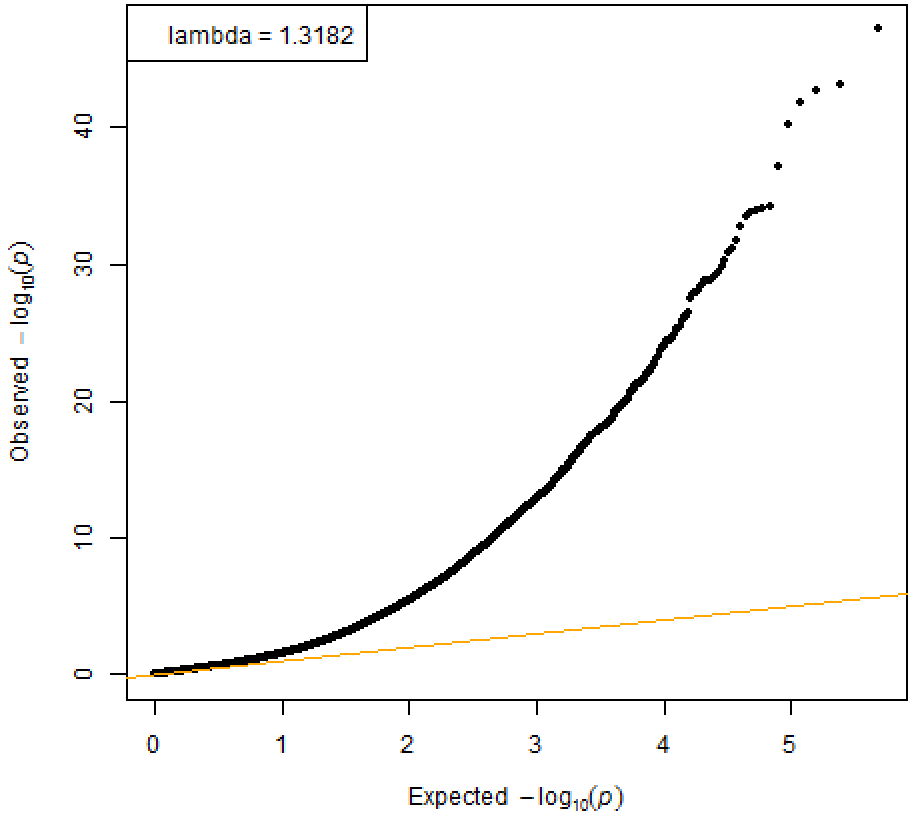
Roby Joehanes, Allan C. Just, Riccardo E. Marioni, Luke C. Pilling, Lindsay M. Reynolds, Pooja R. Mandaviya, Weihua Guan, Tao Xu, Cathy E. Elks, Stella Aslibekyan, Hortensia Moreno-Macias, Jennifer A. Smith, Jennifer A. Brody, Radhika Dhingra, Paul Yousefi, James S. Pankow, Sonja Kunze, Sonia H. Shah, Allan F. McRae, Kurt Lohman, Jin Sha, Devin M. Absher, Luigi Ferrucci, Wei Zhao, Ellen W. Demerath, Jan Bressler, Megan L. Grove, Tianxiao Huan, Chunyu Liu, Michael M. Mendelson, Chen Yao, Douglas P. Kiel, Annette Peters, Rui Wang-Sattler, Peter M. Visscher, Naomi R. Wray, John M. Starr, Jingzhong Ding, Carlos J. Rodriguez, Nicholas J. Wareham, Marguerite R. Irvin, Degui Zhi, Myrto Barrdahl, Paolo Vineis, Srikant Ambatipudi, André G. Uitterlinden, Albert Hofman, Joel Schwartz, Elena Colicino, Lifang Hou, Pantel S. Vokonas, Dena G. Hernandez, Andrew B. Singleton, Stefania Bandinelli, Stephen T. Turner, Erin B. Ware, Alicia K. Smith, Torsten Klengel, Elisabeth B. Binder, Bruce M. Psaty, Kent D. Taylor, Sina A. Gharib, Brenton R. Swenson, Liming Liang, Dawn L. DeMeo, George T. O'Connor, Zdenko Herceg, Kerry J. Ressler, Karen N. Conneely, Nona Sotoodehnia, Sharon L. R. Kardia, David Melzer, Andrea A. Baccarelli, Joyce B. J. van Meurs, Isabelle Romieu, Donna K. Arnett, Ken K. Ong, Yongmei Liu, Melanie Waldenberger, Ian J. Deary, Myriam Fornage, Daniel Levy and Stephanie J. London

Circ Cardiovasc Genet. 2016;9:436-447; originally published online September 20, 2016;
doi: 10.1161/CIRCGENETICS.116.001506
Circulation: Cardiovascular Genetics is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

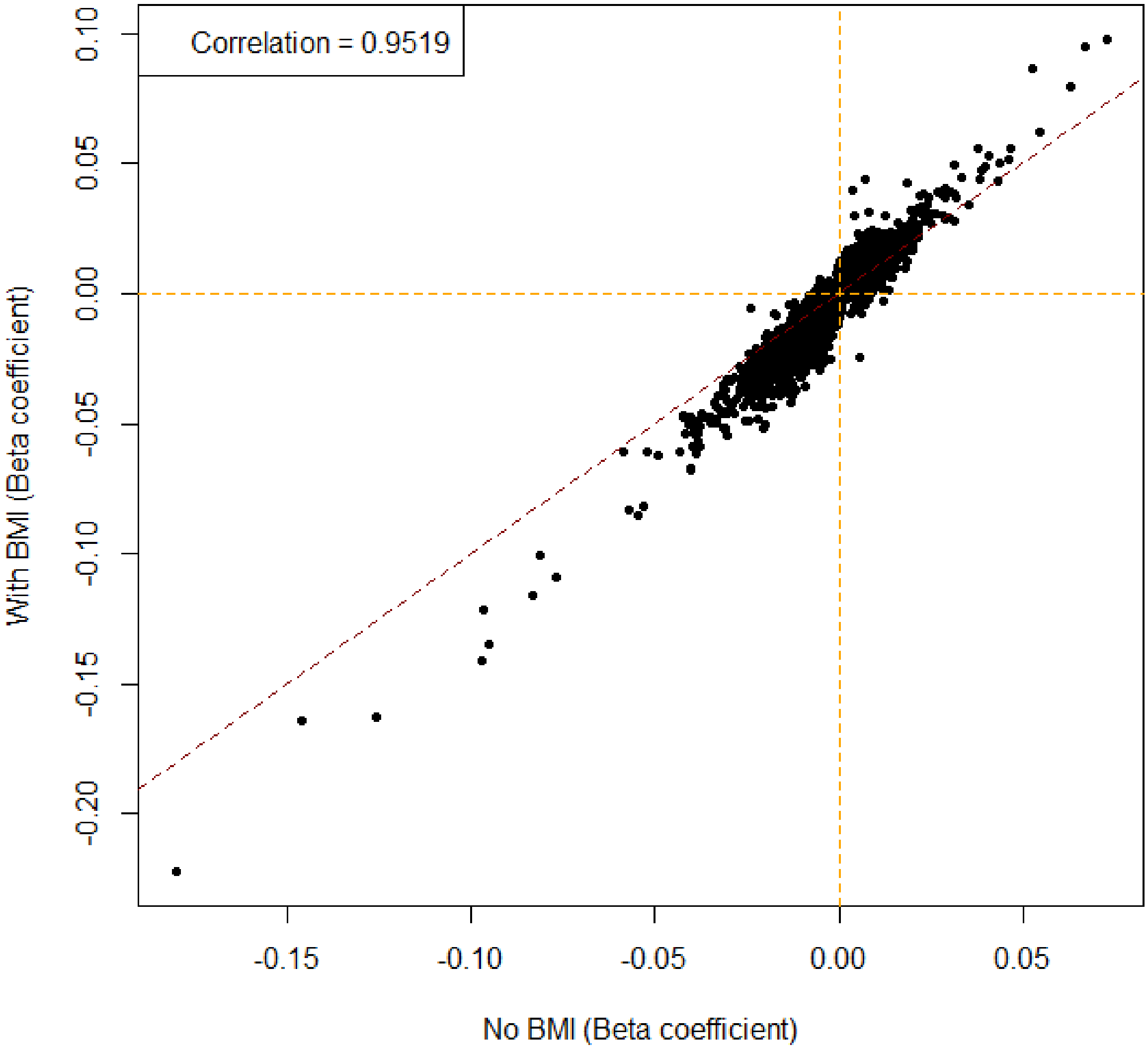
Print ISSN: 1942-325X. Online ISSN: 1942-3268

Supplemental Material

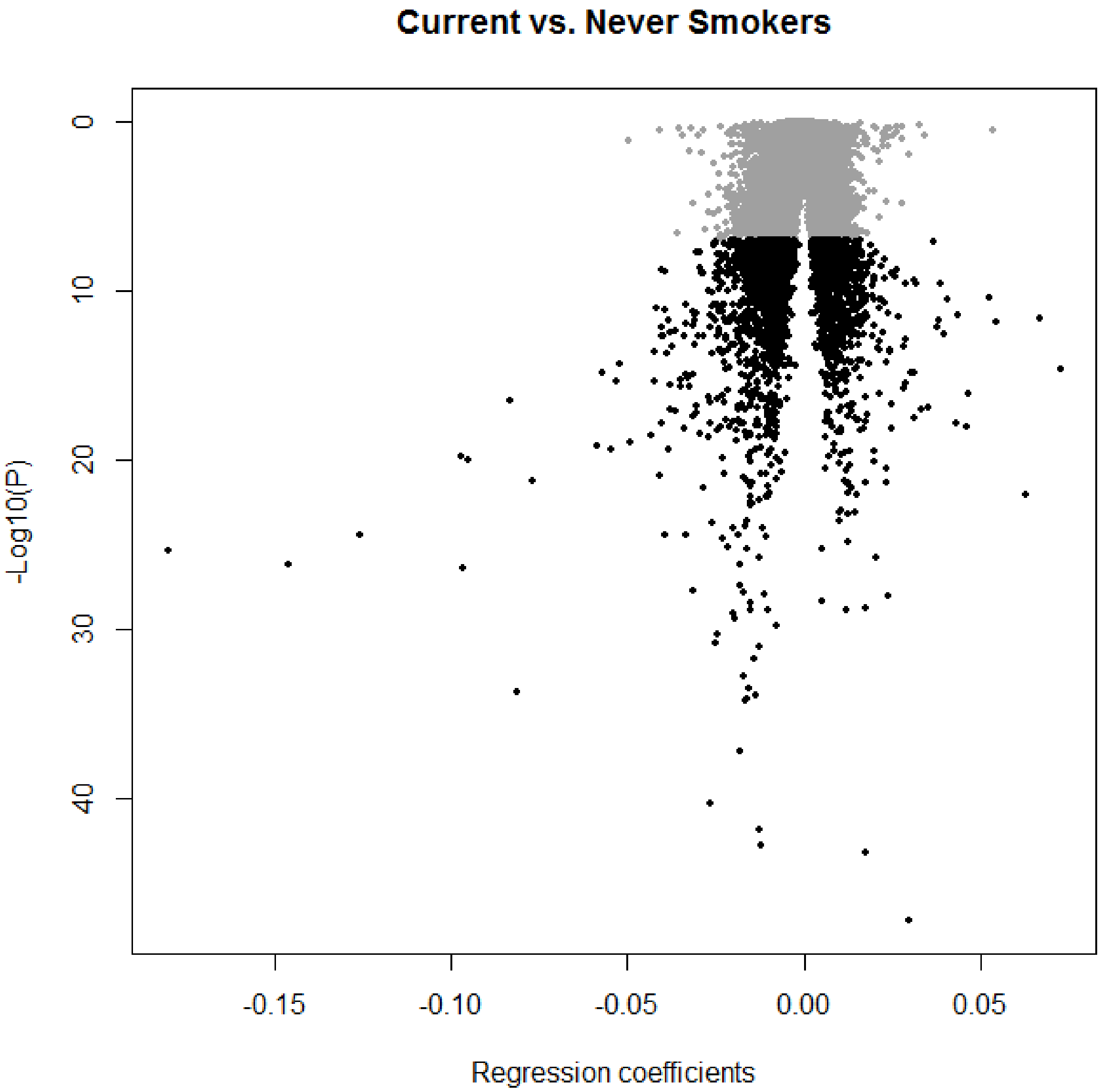
Supplemental Figure 1. Quantile-quantile (QQ) plot for CpG site association with respect to current versus never smoker



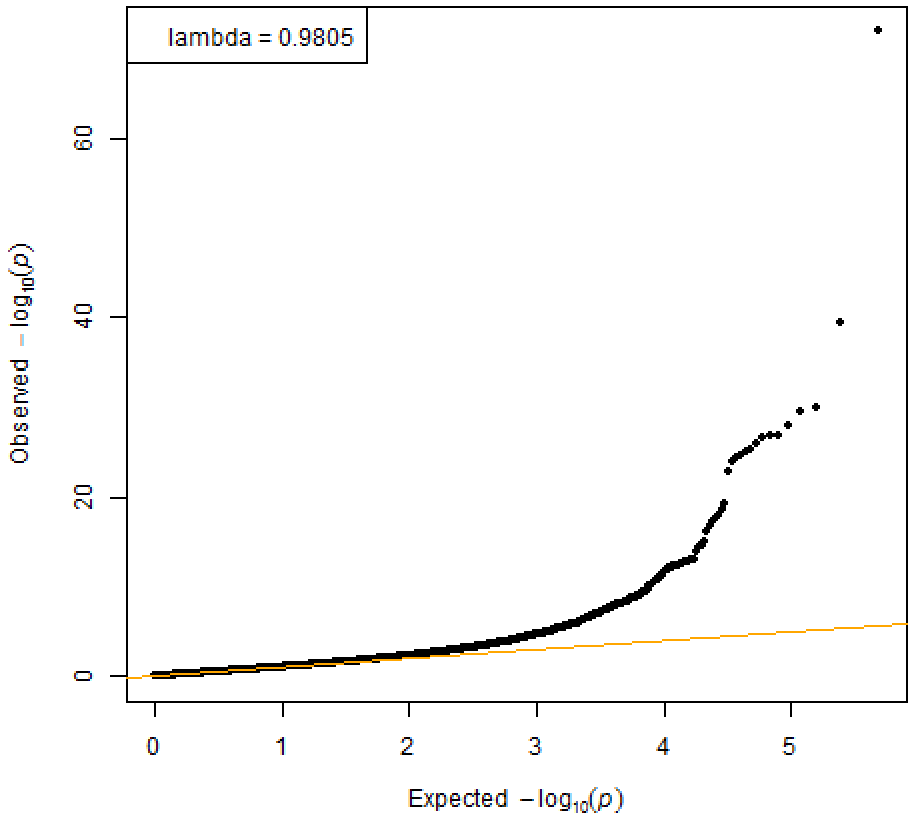
Supplemental Figure 2. Comparison of regression coefficients (beta) of significant 22,473 CpGs between two models (with and without BMI) in relation to current versus never smokers. The X axis indicates beta coefficients without body mass index (BMI). The y axis indicates beta coefficients with BMI added into the model. The CpGs are selected if they are significant (having false discovery rate < 0.05) in at least one of the models. The beta coefficients between the two models are correlated at 0.9519 level.



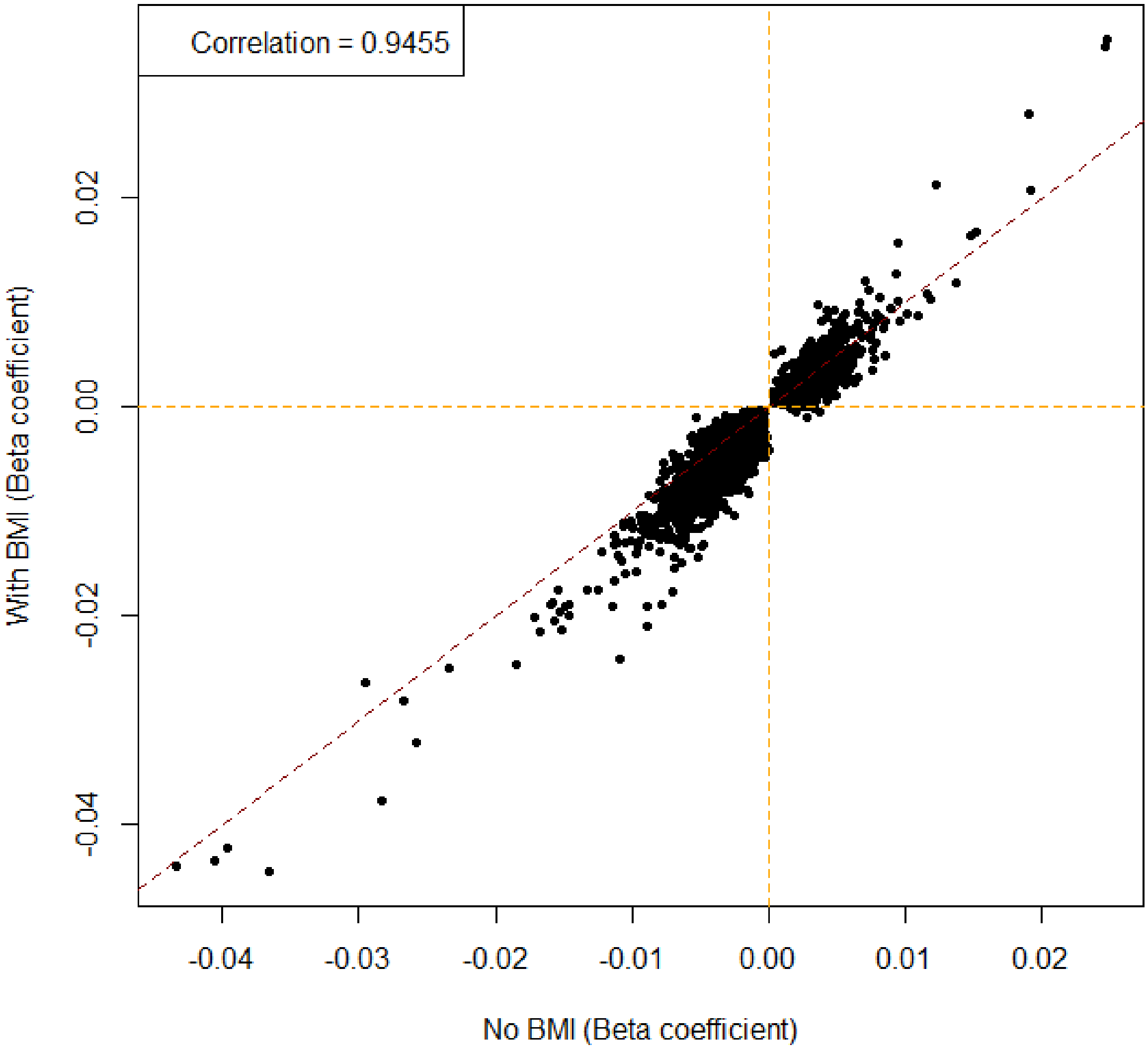
Supplemental Figure 3. Volcano plot for CpG site association with respect to current versus never smoker



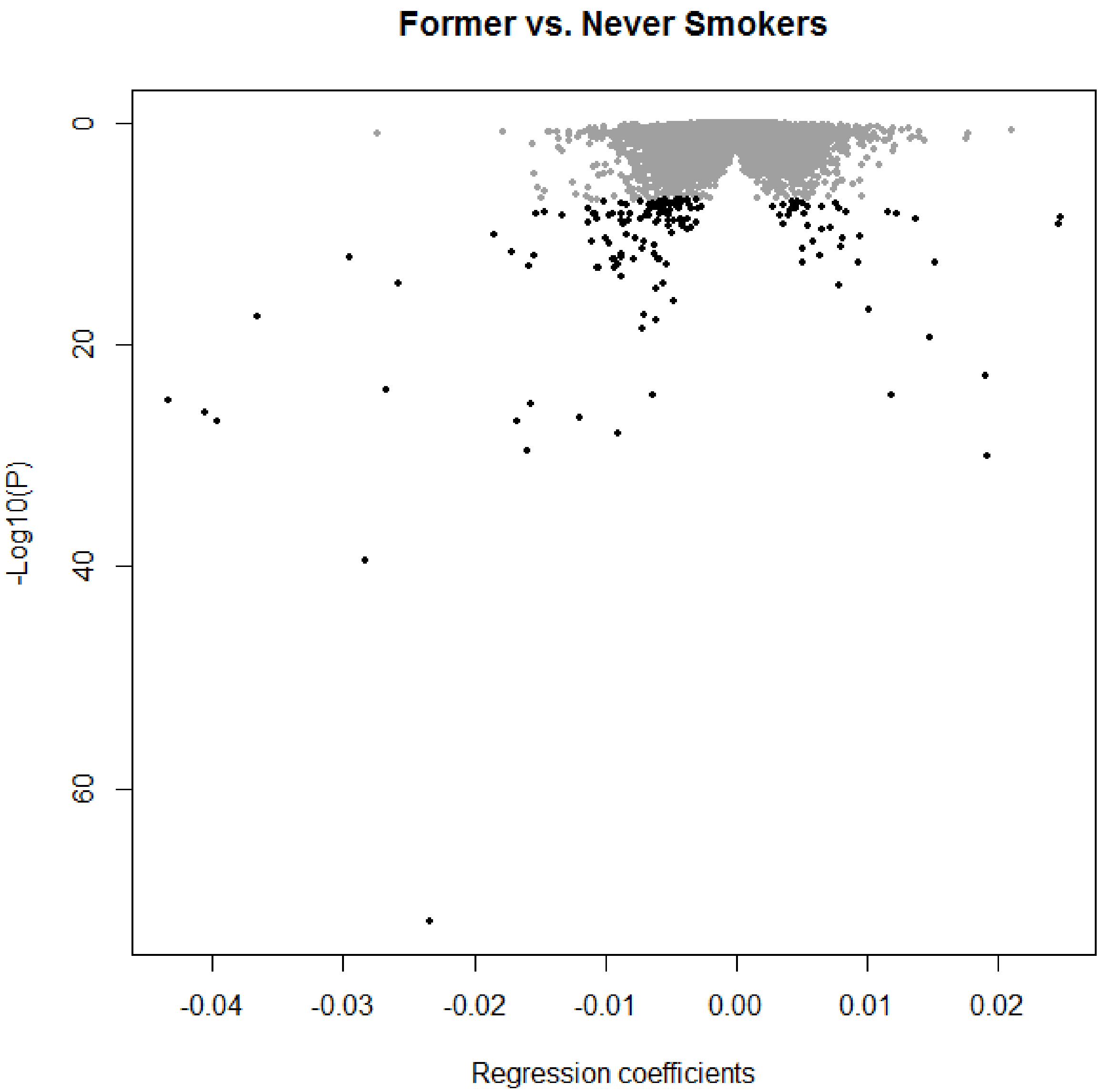
Supplemental Figure 4. Quantile-quantile (QQ) plot for CpG site association with respect to former versus never smoker



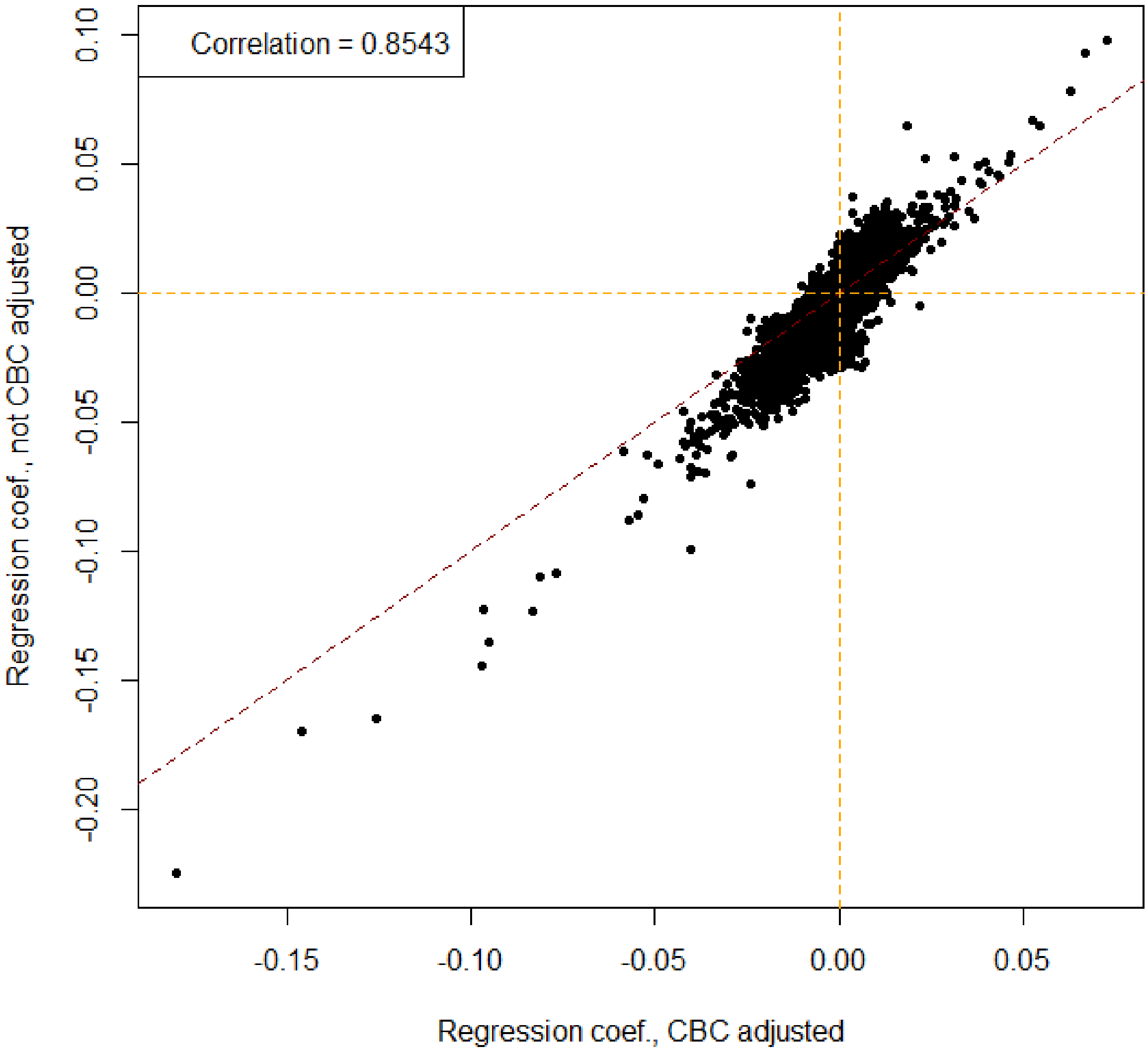
Supplemental Figure 5. Comparison of regression coefficients (beta) of significant 2,998 CpGs between two models (with and without BMI) in relation to former versus never smokers. The X axis indicates beta coefficients without body mass index (BMI). The y axis indicates beta coefficients with BMI added into the model. The CpGs are selected if they are significant (having false discovery rate < 0.05) in at least one of the models. The beta coefficients between the two models are correlated at 0.9455 level.



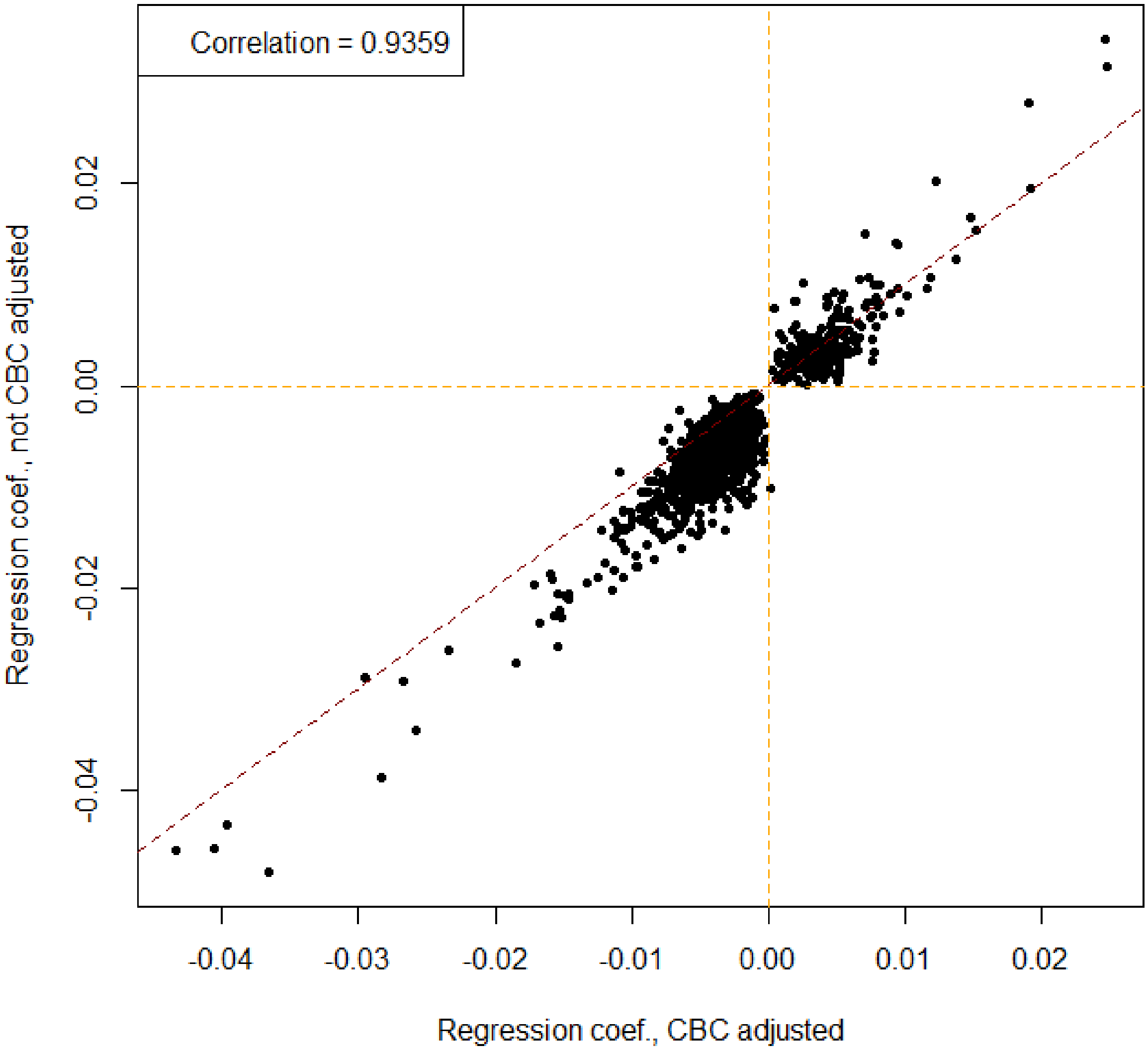
Supplemental Figure 6. Volcano plot for CpG site association with respect to former versus never smoker



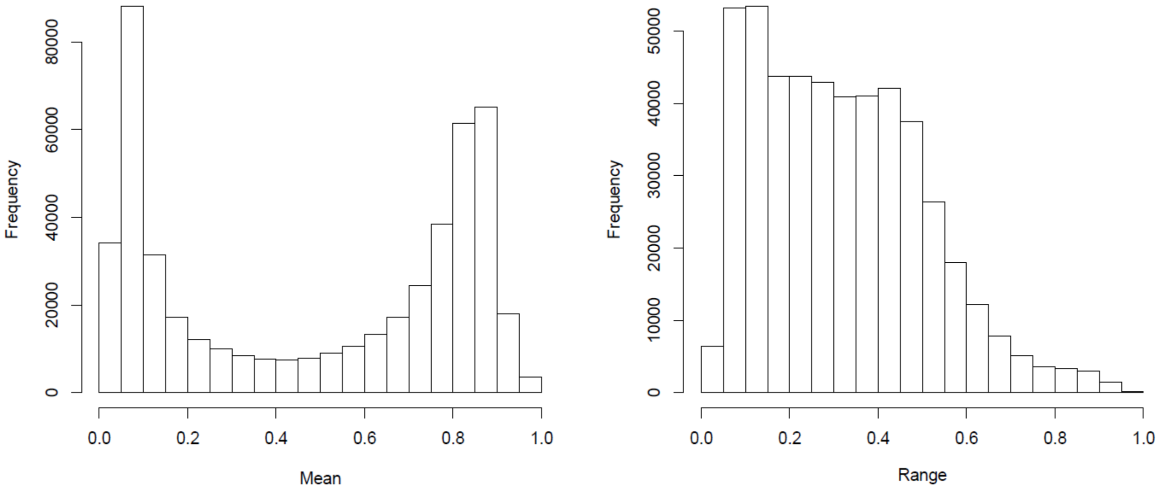
Supplemental Figure 7. Comparison of regression coefficients (beta) of significant 26,693 CpGs between two models (with and without blood cell type adjustment) in relation to current versus never smokers. The X axis indicates beta coefficients with complete blood count (CBC) adjustment. The y axis indicates beta coefficients with without CBC adjustment. The CpGs are selected if they are significant (having false discovery rate < 0.05) in at least one of the models. The beta coefficients between the two models are correlated at 0.8543 level.



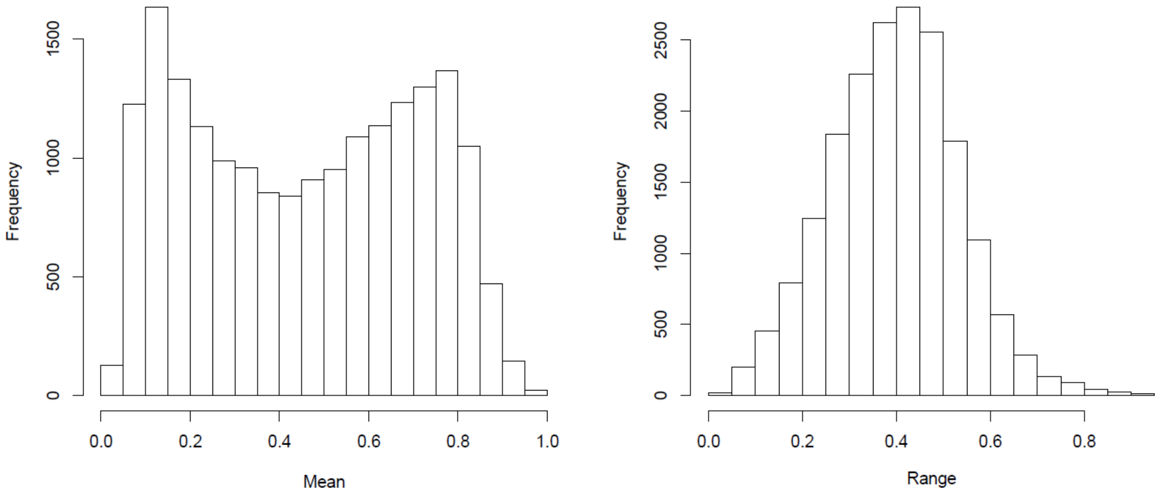
Supplemental Figure 8. Comparison of regression coefficients (beta) of significant 1,137 CpGs between two models (with and without blood cell type adjustment) in relation to former versus never smokers. The X axis indicates beta coefficients with complete blood count (CBC) adjustment. The y axis indicates beta coefficients with without CBC adjustment. The CpGs are selected if they are significant (having false discovery rate < 0.05) in at least one of the models. The beta coefficients between the two models are correlated at 0.9359 level.



Supplemental Figure 9. Histogram plot of mean and range of all 485,381 CpG sites in Framingham Heart Study (FHS) cohort, in methylation proportion (β) scale.



Supplemental Figure 10. Histogram plot of mean and range of 18,760 CpG sites significant in current *vs*. never smokers in Framingham Heart Study (FHS) cohort, in methylation proportion (β) scale.



Supplemental Tables

See separate Excel spreadsheet for all supplemental tables.

Supplemental Table 1. Detailed participant characteristics by cohort.

Supplemental Table 2. Statistically significant CpGs in relation to current *vs*. never smoking status at false discovery rate (FDR)<0.05.

Supplemental Table 3. Statistically significant CpGs in relation to current *vs*. never smoking status that exhibit dose-response relationship (via pack years) at FDR<0.05.

Supplemental Table 4. Gene Ontology pathways of genes whose CpGs are statistically significant in relation to current *vs*. never smoking status.

Supplemental Table 5. Statistically significant CpGs in relation to former *vs*. never smoking status at false discovery rate (FDR)<0.05.

Supplemental Table 6. Statistically significant CpGs in relation to former *vs*. never smoking status that exhibit dose-response relationship (via pack years) at FDR<0.05.

Supplemental Table 7. Gene Ontology pathways of genes whose CpGs are statistically significant in relation to former *vs*. never smoking status.

Supplemental Table 8. List of genome-wide association study (GWAS) phenotypes or diseases for which statistically significant CpGs in relation to current *vs*. never smoking status are enriched.

Supplemental Table 9. List of genome-wide association study (GWAS) phenotypes or diseases for which statistically significant CpGs in relation to former *vs*. never smoking status are enriched.

Supplemental Table 10. List of genes that are GWAS-associated with CVD-related diseases or risk factors that are differentially methylated in relation to current *vs*. never smoking status.

Supplemental Table 11. Enrichment results for genomic features for which differentially methylated CpGs in relation to smoking status are enriched.

Supplemental Table 12. Differentially methylated CpGs in relation to current *vs*. never smoking status that exhibit transcriptomic control in *cis*.

Supplemental Table 13. Gene Ontology pathways of genes whose transcripts are associated in *cis* with the differentially methylated CpGs in relation to current *vs*. never smoking status.

Supplemental Table 14. Comparison of differentially methylated CpGs in relation to current *vs*. never smoking status between cohorts of African Ancestry (AA) and European Ancestry (EA).

Supplemental Table 15. Comparison of differentially methylated CpGs in relation to former *vs*. never smoking status between cohorts of African Ancestry (AA) and European Ancestry (EA).

Supplemental Table 16. Comparison of differentially methylated CpGs in relation to current *vs*. never smoking status between cohorts of whole blood and leukocyte samples.

Supplemental Table 17. Comparison of differentially methylated CpGs in relation to former *vs*. never smoking status between cohorts of whole blood and leukocyte samples.

Supplemental Table 18. Correlation among regression coefficients of CpGs showing significant associations on smoking status across different cell types. Numbers above the diagonal line are for current *vs*. never smoker status, while those below are for former *vs*. never smoker status.

Supplemental Methods

Cohort overview

This study of Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) comprises a total of 15,907 participants from 16 cohorts ARIC, FHS Offspring, KORA F4, GOLDN, LBC 1921, LBC 1936, NAS, Rotterdam, Inchianti, GTP, CHS European Ancestry (EA), CHS African Ancestry (AA), GENOA, EPIC Norfolk, EPIC, and MESA. The study was approved by institutional review committees for each cohort and all participants provided written informed consent for genetic research.

Acknowledgements

S.J.L., M.F., and R.J. led the project in the CHARGE Epigenomic Working Group. M.F., S.J.L., R.J., J.S.P., E.W.D., J.B., J. Schwartz, P.S.V., and A.A.B. designed the study. D.L., I.J.D. M.W., Y.L., K.K.O., D.K.A., I.R., J.B.J.v.M., A.A.B., D.M., S.L.R.K., N.S., and K.N.C. initiated and coordinated the projects in their respective cohort. D.L., L.M.R., L.F., A.P., R.W-S., P.M.V., N.R.W., J.B., J.M.S., J.D., N.J.W., A.G.U., A.H., Z.H., and K.R. are cohort leaders and data coordinators. L.M.R., L.F., A.P., R.W-S., P.S.V., S.B., and E.B.B. provided additional coordination within cohort. R.J. and S.J.L. developed methods. R.J., A.C.J., R.E.M., L.C.P., L.M.R., P.R.M., W.G., T.X., C.E.E., S.A., H.M-M, J.A.S., J.A.B., R.D., S.K., S.S., A.F.M., K.L., J. Sha, M.R.I., E.C., E.B.W., and B.R.S. performed data analysis for each cohort. R.J. and A.C.J. performed meta analyses. R.J., J.S.P., D.M.A., E.W.D., J.B., and D.Z. developed analysis pipeline. R.J., T.X., S.K., D.M.A., M.L.G., A.P., P.M.V., N.R.W., J.M.S., J.D., M.B., P.V., S.A., A.H., E.C., L.H., D.G.H., A.B.S., S.T.T., T.K., E.B.B., B.M.P., K.D.T., L.L., Z.H., K.R., K.N.C., N.S., S.L.R.K., D.M., J.B.J.v.M., I.R., Y.L., M.W., I.J.D., collected the data. R.J., D.M.A., M.L.G., and L.L. performed quality control. P.Y. performed systematic review of the literature. T.H., C.L., M.M.M., and C.Y. provided valuable input to the analysis plan. R.J., S.J.L., and D.L. drafted the manuscript. D.P.K., C.J.R., J. Schwartz, B.M.P., R.E.M., S.A.G., P.M.V., D.L.D., G.T.O., A.A.B., and I.J.D. gave critical review of the manuscript. We would like to thank Bonnie R. Joubert, PhD of the National Institute of Environmental Health Sciences for providing additional literature search, George Chen of the Framingham Heart Study for additional proofreading, and Jianping Jin, PhD of Westat (Durham, NC) for expert computational assistance. Additional acknowledgement can be found in the supplementary materials. All authors provided input on drafts of the manuscript.

Framingham Heart Study (FHS)

Description

The Framingham Heart Study (FHS) is a population-based study that began in 1948. The offspring cohort, consisting of 5,124 participants of European ancestry, was recruited in 1971[1]. Excluding control samples, DNA methylation was measured on 2,792 offspring cohort participants who attended the eighth examination cycle (2005-2008). Of these, 2,648 had both measurements on methylation and smoking status (274 current, 1,538 former, and 836 never smokers). All participants provided written informed consent for genetic research.

DNA methylation sample, measurement, normalization, and quality control

Buffy coat fractions from peripheral whole blood samples were collected from 2,792 offspring cohort participants. Genomic DNA was extracted using the Puregene DNA extraction kit (Qiagen, Venlo, Netherlands) which subsequently bisulfite-converted using the EZ DNA Methylation kit (Zymo Research, Irvine, CA). The samples underwent whole genome amplification, fragmentation, array hybridization, single-base pair extension, and then assayed in two laboratories using the Infinium HumanMethylation 450 BeadChip, which contains 485,512 CpG sites in all. The first laboratory assayed 576 samples, while the second laboratory 2,270 samples.

Raw methylated and total probe intensities were extracted using the Illumina Genome Studio methylation module. Preprocessing of the methylated signal (M) and unmethylated signal (U) was conducted using DASEN of wateRmelon[2] version 3.0.2, an R package. The methylation beta (β) values were defined as $\beta = M/(M+U)$. We excluded low quality CpG sites (with detection p-value > 0.01). We excluded samples showing deviation from the first two principal components (PC1 and PC2), deviation from sex clusters (*i.e.*, male-labeled samples that cluster into female-sample cluster or vice versa), and deviation (>3*SD) from 5,997 SNPs showing the strongest cis methylation quantitative trait locus (mQTL).

Smoking phenotype

Smoking phenotype is as described in the main paper.

Analysis

In the first stage, we analyzed the data with two linear mixed effects models, as described in the main paper. For technical covariates, we included chip ID, row, and column effects as random effects, and PC1 and PC2 as fixed effects. The former factors are to account for technical artefact, the latter account for the inter-laboratory differences. As FHS is a cohort-based study, familial relationship was also included in the model. Thereby, we used pedigreemm package[3], instead of lme4. We also performed pack-year analysis, cessation analysis, and methylation by expression (MxE) analysis as described in the main paper.

Acknowledgements

The Framingham Heart Study is funded by National Institutes of Health contract N01-HC-25195. The laboratory work for this investigation was funded by the Division of

Intramural Research, National Heart, Lung, and Blood Institute, National Institutes of Health and an NIH Director's Challenge Award (D. Levy, Principal Investigator). Dr. Kiel's effort was supported by NIAMS R01 AR41398. Dr. Liang's effort was partially supported by P30 DK462000.

Genetics of Lipid Lowering Drugs and Diet Network (GOLDN)

Description

The GOLDN family study, recruited ~1300 Caucasian men and women with at least two siblings and three generational pedigrees from the participants of the National Heart, Lung, and Blood Institute Family Heart Study in two genetically homogenous centers in Minneapolis, MN and Salt Lake City, UT. The trial aimed to identify genetic factors that mediated response to lipid-raising (*i.e.*, postprandial lipemia challenge) or lipid-lowering (fenofibrate therapy) among metabolically healthy individuals. Participants were asked to discontinue the use of lipid-lowering agents for at least 4 weeks, to fast for at least 8 hours, and to abstain from alcohol and smoking for at least 24 hours prior to study visits. The study protocol was approved by the Institutional Review Boards at the University of Minnesota, University of Utah, Tufts University/New England Medical Center and the University of Alabama at Birmingham, and written informed consent was obtained from all participants[4].

DNA methylation sample, measurement, normalization, and quality control

Epigenetic Phenotyping

Details of the sample isolation are described in previous publications[5,6] and are as follows. CD4+ T-cells were isolated from frozen buffy coat samples using positive selection by antigen-specific magnetic beads (Invitrogen, Carlsbad, CA). DNA was isolated from the CD4+ T-cells using DNeasy kits (Qiagen, Venlo, Netherlands) (2). We used the Infinium Human Methylation 450 array (Illumina, San Diego, CA) to quantify genome-wide DNA methylation[5]. Prior to the standard manufacturer protocol steps of amplification, hybridization, and imaging steps, we treated 500ng of each DNA sample with sodium bisulfite (Zymo Research, Irvine, CA). We used IlluminaGenomeStudio software to estimate β scores, defined as the proportion of total signal from the methylation-specific probe or color channel, and detection p-values, defined as the probability that the total intensity for a given probe falls within the background signal intensity. β scores with an associated detection p-value greater than 0.01 were removed, as were samples with more than 1.5% missing data points across ~470,000 autosomal CpGs. Additionally, any CpG probes where more than 10% of samples failed to yield adequate intensity were removed[5]. Filtered β scores were normalized using the ComBat package for R software[7]. Normalization was performed on random subsets of 10,000 CpGs per run, where each array of 12 samples was used as a "batch." Separate normalization of probes from the Infinium I and II chemistries was performed and subsequently the β scores for Infinium II probes were adjusted using the equation derived from fitting a second order polynomial to the observed methylation values across all pairs of probes located <50bp apart (within-chemistry correlations > 0.99), where one probe was Infinium I and one was Infinium II. Finally, any CpGs where the probe sequence mapped either to a location that did not match the annotation file, or to more than one locus were eliminated. Such markers were identified by re-aligning all probes (with unconverted Cs) to the human reference genome (2). After quality control, we had data for 461,281 CpGs. Principal components based on the beta scores of all autosomal CpGs passing QC were generated using the *prcomp* function in R (V 2.12.1).

Genotyping

A hybrid data set of 2,543,887 single nucleotide polymorphisms (SNPs), of which 484,029 were typed using the Affymetrix Genome-Wide Human 6.0 Array (Affymetrix, Santa Clara, CA) and the rest were imputed using MACH software (Version 1.0.16, Ann Arbor, MI) with Human Genome Build 36 as a reference. Prior to imputation, SNPs were excluded if they were monomorphic, had a call rate of less than 96%, exhibited Mendelian errors, had a minor allele frequency of <1%, or failed the Hardy-Weinberg equilibrium (HWE) test at the P-value threshold of less than 10^{-6}.

Smoking phenotype

Data for smoking variables in this study were collected based on self-reported information. Smoking variables included current smoking status (smoke now, yes/no); number of pack years smoked (number of packs smoked per day x number of years smoked); ever smoked (current, past, or never).

Analysis

Linear mixed effect models were used for analyses:

β = Smoking phenotype + Sex + Age + center + PCs (to account for T-cell purity) + family (random effect).

Acknowledgements

This work has been funded by the NHLBI grant U01HL072524-04 and American Heart Association (AHA) Cardiovascular Genome-Phenome Study (CVGPS) grant 15GPSPG23760006.

Rotterdam Study (RS)

Description

The Rotterdam Study (RS) is a large prospective, population-based cohort study aimed at assessing the occurrence of and risk factors for chronic (cardiovascular, endocrine, hepatic, neurological, ophthalmic, psychiatric, dermatological, oncological, and respiratory) diseases in the elderly[8]. The study comprises 14,926 subjects in total, living in the well-defined Ommoord district in the city of Rotterdam in the Netherlands. In 1989, the first cohort, Rotterdam Study-I (RS-I) comprised of 7,983 subjects with age 55 years or above. In 2000, the second cohort, Rotterdam Study-II (RS-II) was included with 3,011 subjects who had reached an age of 45 years since 1989. In 2006, the third cohort, Rotterdam Study-III (RS-III) was further included with 3,932 subjects with age 45 years and above. Each participant gave an informed consent and the study was approved by the medical ethics committee of the Erasmus University Medical Center, Rotterdam, the Netherlands.

DNA methylation sample, measurement, normalization, and quality control

At the Genetic Laboratory (Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, the Netherlands), the DNA methylation dataset was generated for a subset of 747 individuals of RS-III at baseline. Genomic DNA was extracted from whole peripheral blood by standardized salting out methods. This was followed by a bisulfide conversion using the Zymo EZ-96 DNA-methylation kit (Zymo Research, Irvine, CA, USA). The genome for each sample was then amplified, fragmented and hybridized to the Infinium Illumina Human Methylation 450k arrays according to the manufacturer’s protocol.

The quality control for samples was performed using the Methylation Module of the GenomeStudio software (http://www.illumina.com/applications/microarrays/microarray-software/genomestudio.html). Data was extracted into beta values from raw IDAT files. We excluded samples based on the detection p-value criteria >99% (n=7), poor bisulfite conversion based on control dashboard check (n=5) and failed chromosome X & Y clustering (n=4).

The data preprocessing was additionally performed using an R programming pipeline which is based on the pipeline developed by Tost & Toulemat[9], which includes additional parameters and options to preprocess and normalize methylation data directly from idat files. The beta values were extracted using the R package methylumi. We excluded probes which had a detection p-value >0.01 in >95% of samples. 11648 probes at X and Y chromosomes were excluded to avoid gender bias. This filtering criteria left 731 samples and 463,456 probes. The raw beta values were then background corrected and normalized using the DASEN option of the WateRmelon R-package[2].

Smoking phenotype

Smoking phenotype is as described in the main paper.

Analysis

In the first stage, data were analyzed with linear mixed effects models, as described in the main paper. Pack-year analysis was also performed.

Acknowledgements

The generation and management of the Illumina 450K methylation array data (EWAS data) for the Rotterdam Study was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. The 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) and made available as a Rainbow Project (RP3; BIOS) of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL). We thank Mr. Michael Verbiest, Ms. Mila Jhamai, Ms. Sarah Higgins, Mr. Marijn Verkerk, and Lisette Stolk PhD for their help in creating the methylation database. 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 authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists.

InCHIANTI

Description

The InCHIANTI population is a large population-based study based in the Chianti region of Tuscany, Italy[10]. The participants are aged between 30-104 years and underwent thorough examination every three years from 1998-2000. Whole blood samples were collected using the PAXgene system in 2007[11]. Ethical approval was granted by the Instituto Nazionale Riposo e Cura Anziani institutional review board.

DNA methylation sample, measurement, normalization, and quality control

Genomic DNA was extracted from buffy coat samples using an AutoGen Flex and quantified on a Nanodrop1000 spectrophotometer prior to bisulfite conversion. Genomic DNA was bisulfite converted using Zymo EZ-96 DNA Methylation Kit per the manufacturer's protocol. CpG methylation status of 485,577 CpG sites was determined using Illumina Infinium HumanMethylation450 BeadChip per manufacturer's protocol and as previously described[12]. Initial data analysis was performed using GenomeStudio 2011.1 (Model M Version 1.9.0 Illumina, Inc. CA). Threshold call rate for inclusion of samples was 95%. Quality control of sample handling included comparison of clinically reported sex versus sex of the same samples determined by analysis of methylation levels of CpG sites on the X chromosome. Beta values were extracted for sites on the X chromosome. Subject mean methylation versus subject mean intensity levels were plotted in R V2.11.1. Based on methylation levels for chromosome X loci, these data split into two primary groups. Calls generated by this method were then compared with sample information reported by InChianti Study. Samples not matching between clinical reported sex and methylation data were excluded from analyses.

Quantile normalization of the methylation arrays was carried out using package “wateRmelon” for the R statistical computing language[2]. The DASEN method was applied, which performs the quantile normalization separately on M and U (methylated/un-methylated) values, and also separates the type 1 and type 2 Infinium probes. This minimises the technical variance between the arrays, whilst taking into consideration the different technologies present on the arrays. Methylation data was available for 506 InCHIANTI participants following quality control and data cleaning.

Analysis

Linear mixed effects models were applied to each 450k array probe in turn with the following cofactors included: fixed effects: age, sex, total white blood cell counts, lymphocyte, monocyte, eosinophil and basophil proportions, platelet counts: included as random effects: sentrix ID, sentrix position, and array batch. Analyses were performed on current vs. never, former vs. never, pack-years smoked, and years since quitting (cessation).

Acknowledgements

The InCHIANTI study was supported in part by the Intramural Research Program, National Institute on Aging, NIH, Baltimore MD USA. D.M. is generously supported by a Wellcome Trust Institutional Strategic Support Award (WT097835MF).

Cooperative health research in the Region of Augsburg (KORA)

Description

Cooperative health research in the Region of Augsburg (KORA) is a population-based cohort study conducted in the region of Augsburg, Southern Germany[13,14]. The study has been conducted according to the principles expressed in the Declaration of Helsinki. Written informed consent has been given by each participant. The study was reviewed and approved by the local ethical committee (Bayerische Landesärztekammer). The baseline survey 4 (KORA S4) consists of 4,261 individuals (aged 25-74 years) examined between 1999 and 2001. During the years of 2006 to 2008, 3,080 participants took part in the follow-up survey 4 (KORA F4). Phenotypic data were retrieved from self-reports and medical records.

DNA methylation sample, measurement, normalization, and quality control

In KORA F4, the analysis was performed using whole blood DNA of fasting participants (n=1776). Genomic DNA (1 µg) was bisulfite converted using the EZ-96 DNA Methylation Kit (Zymo Research, Orange, CA, USA) according to the manufacturer's procedure, with the alternative incubation conditions recommended when using the Illumina Infinium Methylation Assay. Genome-wide DNA methylation was assessed using the Illumina HumanMethylation450 BeadChip, following the Illumina Infinium HD Methylation protocol. This consisted of a whole genome amplification step using 4µl of each bisulfite converted sample, followed by enzymatic fragmentation and application of the samples to BeadChips (Illumina). The arrays were fluorescently stained and scanned with the Illumina HiScan SQ scanner. Raw methylation data were extracted with Illumina Genome Studio (version 2011.1) with methylation module (version 1.9.0). The percentage of methylation at a given cytosine is reported as a beta-value. Low-confidence probes, which has less than three functional beads or has a detection p-value larger than 0.01, were excluded. Sites representing or being located in a 50 bp proximity to SNPs with a minor allele frequency of at least 5% were also excluded from the data set. β-mixture quantile normalization[15] was applied to the DNA methylation data using the R package wateRmelon[2], version 1.0.3.8. KORA F4 samples were processed on 20/7 96-well plates in 9/4 batches, a plate effect representing 4.8% of the total variance of the methylation level was observed. Additionally, plate was included as a random effect in the analyses. Detailed quality control process was described the previous publication[16].

Smoking phenotype

The smoking phenotype was defined based on self-reports, see main text for details.

Analysis

The data was analyzed using two linear mixed effect models as described in the main text. In KORA F4, the blood count comprises the fractions of CD4+ T-cells, CD8+ T-cells, NK cells, and monocyte estimated using the Houseman et al. method. The chip number, the row and column number of the samples on the plates were included as the technical covariates with random effect in the model. No significant population stratification was found in the KORA F4 data, familial relationship was not adjusted in the model.

Acknowledgements

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. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 261433 (BioSHaRE-EU), n°603288 (SysVasc) and HEALTH-F2-2013-602736 (PAIN-OMICS). Furthermore, this study was supported in part by a grant from the BMBF to the German Center for Diabetes Research (DZD). In addition, part of this work was financed by the German National Genome Research Network (NGFNplus, project number 01GS0834) and through additional funds from the University of Ulm.

Grady Trauma Project (GTP)

Description

The Grady Trauma Project (GTP) is a population-based, prospective study of demographic characteristics, trauma exposure, and prevalence of post-traumatic stress disorder and major depressive disorder in an urban, predominantly African-American population[17]. Subjects were recruited prospectively from the waiting rooms of primary care and obstetrics-gynecology clinics of Grady Memorial Hospital in Atlanta, GA. Exclusion criteria included mental retardation, active psychosis, or the inability to give informed consent. Written and verbal informed consent was obtained for all participating subjects. All procedures in this study were approved by the Institutional Review Boards of Emory University School of Medicine and Grady Memorial Hospital. Since its inception in 2005, over 5000 subjects have been interviewed for the study.

DNA methylation sample, measurement, normalization, and quality control

We extracted DNA from whole blood at the Max Planck Institute in Munich for 425 GTP participants using the Gentra Puregene Kit (Qiagen); for this study, we focus on 286 participants who are African American and have complete information for the smoking phenotype (described below). Genomic DNA was then bisulfite converted using the Zymo EZ-96 DNA Methylation Kit (Zymo Research). We assessed DNA methylation at >480,000 CpG sites using Illumina HumanMethylation450 BeadChip arrays, with hybridization and processing performed according to the instructions of the manufacturer. For each CpG site and individual, we collected two data points: M (the total methylated signal), and U (the total unmethylated signal). We set to missing data points with 1) a detection p-value greater than 0.001 or 2) a combined signal less than 25% of the total median signal and less than both the median unmethylated and median methylated signal. We removed individual samples from analysis if they were outliers in a hierarchical clustering analysis or had 1) a mean total signal less than half of the median of the overall mean signal or 2000 arbitrary units, or 2) a missingness rate above 5%. Similarly, we removed from analysis CpG sites with a missingness rate above 10%. We then computed β-values for each individual at each CpG site as the total methylated signal divided by the total signal: $\beta = M/(M+U)$. For quantile normalized data, the M and U signals were quantile normalized together prior to computation of β-values.

Smoking phenotype

Smoking information was collected from GTP participants using an adapted KMSK questionnaire tool. This tool (originally described in Kellogg et al. 2003[18]) records, using a numerical scale, the current frequency of smoking, the duration of time that this frequency has been maintained and the amount of cigarettes smoked during this period. The adapted tool used in GTP recorded this information for both the 30 days prior and the time period where participant smoking was greatest for 425 individuals. Frequency (coded on a 0-5 point scale, where 5 = smoking at regular intervals most/all days; 4 = smoking at specific times of day most/all days; 3 = once a day most/all days; 2 = 20-100 times in lifetime; 1 = less than 20 times in lifetime; 0 = never smoked) for both time periods (hereafter referred to as '30-day' and 'maximum') was used to create a variable describing whether the individual is a current, former or never smoker (CFN).

The CFN scale was determined as follows:

- An individual was classified as a current smoker (N = 94), if their 30-day frequency was coded as a 3, 4 or 5 and their maximum frequency was coded as a 3, 4, 5 or missing.
- An individual was classified as a former smoker (N = 64), if 30-day frequency was coded as a 0, 1, or 2, and maximum frequency was coded as a 3, 4 or 5.
- An individual was classified as a never smoker (N = 128), if their maximum frequency was coded as a 0, 1, or 2 and their 30-day frequency was coded as a 0, 1, 2 or missing.

If an individual did not meet the above criteria (N = 45) or did not supply any smoking information (N = 61), their score on the CFN scale was recorded as missing.

Analysis

Data were analyzed with two linear mixed effects models, as described in the main paper.

Acknowledgements

This work was primarily supported by the National Institute of Mental Health (MH096764 and MH071537 to K.J.R.). Support was also received from Emory and Grady Memorial Hospital General Clinical Research Center, the Max Planck Society, the Behrens-Weise Foundation, NIH National Centers for Research Resources (M01RR00039), and the National Center for Advancing Translational Sciences of the NIH (UL1TR000454).

Lothian Birth Cohorts of 1921 and 1936 (LBC1921 and LBC1936)

Description

The Lothian Birth Cohorts of 1921 and 1936 are two longitudinal studies of ageing[19–21]. They derive from the Scottish Mental Surveys of 1932 and 1947, respectively, when nearly all 11 year old children in Scotland completed a test of general cognitive ability[21]. Survivors living in the Lothian area of Scotland were recruited in late-life at mean age 79 for LBC1921 (n=550) and mean age 70 for LBC1936 (n=1,091). Follow-up has taken place at ages 70, 73, and 76 in LBC1936 and ages 79, 83, 87, and 90 in LBC1921. Collected data include genetic information, longitudinal epigenetic information, longitudinal brain imaging (LBC1936), and numerous blood biomarkers, anthropomorphic and lifestyle measures. Post QC, DNA methylation data were available for 920 LBC1936 participants at age 70, and for 446 LBC1921 participants at age 79.

DNA methylation sample, measurement, normalization, and quality control

Detailed information about the collection and QC steps undertaken on the LBC methylation data have been reported previously[22]. Briefly, the Infinium HumanMethylation450 BeadChip (Illumina Inc, San Diego, CA) was used to measure DNA methylation in whole blood of consenting participants. Background correction was performed and QC was used to remove probes with a low detection rate (<95% at $P < 0.01$), low quality (manual inspection), low call rate (below 450,000 probes at $P < 0.01$), and samples with a poor match between genotypes and SNP control probes, and incorrect predicted sex. Background correction and internal normalisation were performed; the betas were modified such that the minimum was 0.001 and the maximum was 0.999.

Smoking phenotype

Smoking was measured via self-response. Participants were asked if they were current smokers, never smokers, or former smokers.

Analysis

Linear mixed effects models were used to analyze the data in both cohorts. Measured white blood cell counts (eosinophils, neutrophils, basophils, monocytes, and lymphocytes) were included as fixed effects along with age and sex; technical covariates (sample plate, BeadChip, position on BeadChip, and hybridisation date) were included as random effects.

Acknowledgements

We thank the cohort participants and team members who contributed to these studies. This work was supported by numerous funding bodies. Phenotype collection in the Lothian Birth Cohort 1921 was supported by the UK’s Biotechnology and Biological Sciences Research Council (BBSRC), The Royal Society and The Chief Scientist Office of the Scottish Government. Phenotype collection in the Lothian Birth Cohort 1936 was supported by Age UK (The Disconnected Mind project). Methylation typing was supported by the Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, The Wellcome Trust Institutional Strategic Support Fund, The University of Edinburgh, and The University of Queensland. REM, PMV, and IJD are members of the University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology (CCACE). CCACE is supported by funding from the BBSRC, the Medical

Research Council (MRC), and the University of Edinburgh as part of the cross-council Lifelong Health and Wellbeing initiative (MR/K026992/1). Research reported in this publication was supported by National Health and Medical Research Council (NHMRC) project grants 613608, 1010374 and 1046880. NHMRC Fellowships to PMV (1078037) and NRW (1078901) and Australia Research Council (ARC) Future Fellowship to NRW (FT0991360). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NHMRC or ARC.

The Multi Ethnic Study of Atherosclerosis (MESA)

Description

The Multi-Ethnic Study of Atherosclerosis (MESA) was designed to investigate the prevalence, correlates, and progression of subclinical cardiovascular disease in a population cohort of 6,814 participants. Since its inception in 2000, five clinic visits collected extensive clinical, socio-demographic, lifestyle, behavior, laboratory, nutrition, and medication data[23]. DNA methylation and gene expression were measured in purified (CD14+) monocyte samples from the April 2010 – February 2012 examination (exam 5) of 1,264 randomly selected MESA participants from four MESA field centers (Baltimore, MD; Forsyth County, NC; New York, NY; and St. Paul, MN) as previously described[24]. The study protocol was approved by the Institutional Review Board at each site. All participants signed informed consent.

DNA methylation sample, measurement, normalization, and quality control

As previously described[24], blood was initially collected in sodium heparin-containing Vacutainer CPT™ cell separation tubes (Becton Dickinson, Rutherford, NJ, USA) to separate peripheral blood mononuclear cells from other elements within 2 h from blood draw. Subsequently, monocytes were isolated with the anti-CD14-coated magnetic beads, using AutoMACs automated magnetic separation unit (Miltenyi Biotec, Bergisch Gladbach, Germany). Based on flow cytometry analysis of 18 specimens, monocyte samples were consistently >90% pure.   DNA and RNA were isolated from samples simultaneously using the AllPrep DNA/RNA Mini Kit (Qiagen, Inc., Hilden, Germany). DNA and RNA QC metrics included optical density measurements, using a NanoDrop spectrophotometer and evaluation of the integrity of 18s and 28s ribosomal RNA.

Illumina HumanMethylation450 BeadChips and HiScan reader were used to perform the epigenome-wide methylation analysis. Bead-level methylation data were summarized in GenomeStudio. Because a two-channel system and both Infinium I and II assays were used, normalization was performed in several steps using the lumi package. “Smooth quantile normalization” was used to adjust for color bias. Next, the data were background adjusted by subtracting the median intensity value of the negative control probes. Lastly, data were normalized across all samples by standard quantile normalization applied to the bead-type intensities and combined across Infinium I and II assays and both colors. QC measures included checks for sex and race/ethnicity mismatches, and outlier identification by multidimensional scaling plots. To estimate residual sample contamination for data analysis, we generated separate enrichment scores for neutrophils, B cells, T cells, monocytes, and natural killer cells. We implemented a Gene Set Enrichment Analysis[25] as previously described[24] to calculate the enrichment scores using the gene signature of each blood cell type from previously defined lists[26]. To remove technical error in methylation levels associated with batch effects across the multiple chips, positional effects of the sample on the chip, and residual sample contamination with non-monocyte cell types, we adjusted methylation values for chip, sample position on the chip, and estimated residual sample contamination with neutrophils, B cells, T cells, monocytes, and natural killer cells. The final methylation value for each methylation probe was computed as the beta-value, essentially the proportion of the methylated to the total intensity.

Smoking phenotype

Smoking status was ascertained longitudinally (Exams 1-5). Current smokers reported to be current smokers at Exam 5, the time of the blood draw. Former smokers reported to be former smokers at any exam (1-5) or reported to ever smoke at least 100 cigarettes in their lifetime at exam 1. Never smokers reported never smoking at all exams.

Analysis

Data were analyzed with two linear mixed effects models, as described in the main paper. A look-up for significant (FDR<0.001) methylation by expression (MxE) associations was performed using data previously reported in the same 1,264 CD14+ samples[24], including genes located within 1 MB of smoking-associated methylation identified in current vs. never and former vs. never analyses.

Acknowledgements

MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-001079, UL1-TR-000040, and DK063491. The MESA Epigenomics & Transcriptomics Study was funded by NHLBI grant R01HL101250 to Wake Forest University Health Sciences. Analysis of MESA data reported in this publication was also supported by the NHLBI under Award Number

P50HL120163. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

European Prospective Investigation into Cancer (EPIC)

Description

The EPIC study is an on-going multi-center prospective cohort study designed to investigate the relation between nutrition and cancer occurrence. The cohort consists of 23 centers in 10 European countries (*i.e.*, Denmark, France, Germany, Greece, Italy, Netherlands, Norway, Spain, Sweden and United Kingdom). From 1992–2000, more than 500,000 individuals aged between 25 and 70 years were recruited. All participants gave written or oral informed consent. The study was approved by the International Agency for Research on Cancer (IARC) ethical review committee and by local ethical committees at the participating centers. The design of EPIC is described in detail elsewhere[27] DNA methylation was measured on 450 breast cancer cases and 450 matched controls among women using a nested case-control approach (2005-2008). Of these, 898 had both measurements on methylation and smoking status (196 current, 190 former, and 512 never smokers). All participants provided written informed consent for genetic research.

DNA methylation sample, measurement, normalization, and quality control

DNA was isolated from white blood cells as per the standard DNA extraction procedure (Autopure LS, Qiagen). DNA methylome profiling was carried out using the Illumina Infinium HumanMethylation450 (HM450) BeadChip assay, which interrogates more than 480,000 methylation sites, essentially as described previously[28]. Briefly, 500 ng of extracted DNA was bisulfite-modified using the EZ DNA Methylation kit (Zymo Research, D5004), following the manufacturer's instructions for the HM450 BeadChip assay. The conversion was confirmed by performing PCR for *GAPDH* primers specific for modified/unmodified DNA samples. The hybridization and scanning of the arrays were performed as per the manufacturer's instructions.

Data pre-processing and analysis were performed using R (version 3.2.2) /Bioconductor packages. To avoid spurious associations, we excluded the cross-reactive probes and probes overlapping with a known single nucleotide polymorphism (SNP) with an allele frequency of ≥5% in the overall population (European ancestry[29]), leaving 423,066 probes. In any given sample, a probe with a detection *P*-value (a measure of an individual probe's performance) of greater than or equal to 0.05 was assigned missing status. If a probe was missing in more than 5% of samples, it was excluded from all samples. Thus, 1,625 probes were excluded on this basis. Finally, 421,441 probes were available for the analyses, which were corrected for probe colour bias, inter-sample quantile normalization followed by beta-mixture quantile normalization (BMIQ) to align Type I and Type II probe distributions[15]. The array annotations from FDb.InfiniumMethylation.hg19 were used to assign probes to their corresponding genes.

Smoking phenotype

Each center participating in EPIC cohort had their own questionnaire, thus questions regarding smoking habits were slightly different. Responses were harmonized in order to classified participants as never, former and current smoker according to the responses to their respective questionnaires with the rationale of clearly distinguish between groups. Some examples of questionnaire information are: "Do you currently smoke", "Have you ever smoked for over 3 month?"," Did you smoked 1 cigarette per day or more per day in the past?”,” Have you ever smoked as much as one cigarette a day for as long as a year?”,” Do you smoke cigarettes regularly? , etc. Women were classified as never, former and current smoker according their responses to their respective questionnaire with the rationale of clearly distinguish between groups.

Analysis

First the proportion of methylation (beta values) for each CpG site were explored and then the range of beta values were checked. The vector of raw betas and the vector of normalized betas were used as the outcome in separate linear mixed models with center and pool ID as random effects. Models were adjusted for age, BMI, cancer status (case or control) and proportion of CD8 T lymphocytes, CD4 T lymphocytes, B cells, monocytes and natural killer cells. The genomic inflation factor (lambda) was calculated and the QQ-plot was generated for each model. The Benjamini & Hochberg[30] procedure was used for controlling False Discovery Rate because multiple testing. R software v.3.1.3 was used.

Acknowledgements

This work was supported by the International Agency for Research on Cancer.The coordination of EPIC is financially supported by the European Commission (DG-SANCO) and the International Agency for Research on Cancer. The national cohorts are supported by DanishCancer Society (Denmark);Deutsche Krebshilfe, Deutsches Krebsforschungszentrum and Federal Ministry of Education and Research (Germany); Hellenic Health Foundation, the Stavros Niarchos Foundation, and the Hellenic Ministry of Health and Social Solidarity (Greece); Italian Association for Research on Cancer (AIRC) and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS),Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (ZorgOnderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (TheNetherlands); ERC-2009-AdG 232997 and Nordforsk, Health Research Fund (FIS), Regional Governments of Andalucía, Asturias, Basque Country, Murcia (project number 6236) and Navarra, ISCIII RETIC(RD06/0020) (Spain); Cancer Research UK, Medical Research Council, StrokeAssociation, British Heart Foundation, Department of Health, Food Standards Agency, and WellcomeTrust (United Kingdom). We would like to acknowledge technical help from Cyrille Cuenin and Geoffrey Durand. The work carried out by the Epigenetics group at IARC was supported by grants from the Institut National du Cancer (INCa, France) to I.R. and Z.H. and the European Commission (EC) Seventh Framework Programme (FP7) Translational Cancer Research (TRANSCAN) Framework (EpiMARK project) to Z.H.. Z.H. was also supported by the Fondation ARC pour la Recherche sur le Cancer (France) and the EC FP7 EurocanPlatform: A European Platform for Translational Cancer Research (grant number: 260791). The work reported in this paper was undertaken during the tenure of a Postdoctoral Fellowship (to SA) from the International Agency for Research on Cancer, partially supported by the EC FP7 Marie Curie Actions – People – Co-funding of regional, national and international programmes (COFUND). This study depended on the participation of the women in the EPIC cohort, to whom we are grateful.

The Atherosclerosis Risk in Communities (ARIC) Study

Description

The Atherosclerosis Risk in Communities (ARIC) Study is a prospective cohort study of cardiovascular disease risk in four U.S. communities. Between 1987 and 1989, 7,082 men and 8,710 women aged 45–64 years were recruited from Forsyth County, North Carolina; Jackson, Mississippi (African Americans only); suburban Minneapolis, Minnesota; and Washington County, Maryland. The ARIC Study protocol was approved by the institutional review board of each participating institution. After written informed consent was obtained, including that for genetic studies, participants underwent a baseline clinical examination (Visit 1) and four subsequent follow-up clinical exams (Visits 2 – 5).

DNA methylation sample, measurement, normalization, and quality control

At this time, DNA methylation data are available for African American members of the cohort from two centers (Forsyth County and Jackson). The present study comprises a cross-sectional analysis of smoking and methylation measured in samples collected at visit 2 and 3, with covariates obtained at the same visit.

Genomic DNA was extracted from peripheral blood leukocyte samples using the Gentra Puregene Blood Kit (Qiagen; Valencia, CA, USA) according to the manufacturer's instructions (www.qiagen.com). Bisulfite conversion of 1 ug genomic DNA was performed using the EZ-96 DNA Methylation Kit (Deep Well Format) (Zymo Research; Irvine, CA, USA) according to the manufacturer's instructions (www.zymoresearch.com). Bisulfite conversion efficiency was determined by PCR amplification of the converted DNA before proceeding with methylation analyses on the Illumina platform using Zymo Research's Universal Methylated Human DNA Standard and Control Primers.

Bisulfite-converted DNA was used for hybridization on the Illumina Infinium HumanMethylation450 (HM450) BeadChip, following the Illumina Infinium HD Methylation protocol (www.illumina.com). This consisted of a whole genome amplification step followed by enzymatic end-point fragmentation, precipitation and re-suspension. The re-suspended samples were hybridized to the complete set of bead-bound probes, followed by ligation and single-base extension during which a fluorescently-labeled nucleotide is incorporated, and scanned. The degree of methylation is determined for each CpG cytosine by measuring the amount of incorporated label for each probe. The intensities of the images were extracted using Illumina GenomeStudio 2011.1, Methylation module 1.9.0 software. The methylation score for each CpG was represented as a beta (β) value according to the fluorescent intensity ratio. Background subtraction was conducted with the GenomeStudio software using built-in negative control bead types on the array.

Positive and negative controls and sample replicates were included on each 96-well plate assayed. After exclusion of controls, replicates, and 22 samples that failed bisulfite conversion, a total of 2,905 study participants had HM450 data available for further quality control analyses. We removed poor-quality samples with pass rate <99% (N=32). At the target level, we flagged poor-quality CpG sites with average detection p-value > 0.01, and calculated the percentage of samples having detection p-value > 0.01 for each autosomal and X chromosome CpG site. There were 5,174 autosomal and X chromosomal markers where >1% of samples showed detection p-value > 0.01, and these sites were excluded.

Methylation values were normalized using the Beta MIxture Quantile dilation (BMIQ) method[15].

Smoking phenotype

Smoking phenotype is as described in the main paper.

Analysis

Since white blood cell proportions were not directly measured in most participants in ARIC, they were imputed from the methylation data using the Houseman method. Specifically, the proportions of neutrophils, lymphocytes, monocytes, eosinophils, and basophils were estimated based on the measured differential cell counts available for a subset of ARIC participants at Visit 2 (n = 175). All association analyses were performed in R using linear mixed models with DNA methylation beta values as the outcome, as described in the main paper.

Acknowledgements

The Atherosclerosis Risk in Communities (ARIC) study is carried out as a collaborative study supported by the National Heart, Lung, and Blood Institute (NHLBI) contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). The authors thank the staff and participants of the ARIC study for their important contributions. Funding support for “Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium” was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419). The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper. Stephanie London is supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

Genetic Epidemiology Network of Arteriopathy (GENOA)

Description

The Genetic Epidemiology Network of Arteriopathy (GENOA) study is a community-based study of hypertensive sibships that was designed to investigate the genetics of hypertension and target organ damage in African Americans from Jackson, Mississippi and non-Hispanic whites from Rochester, Minnesota[31]. In the initial phase of GENOA (Phase I: 1996-2001), all members of sibships containing ≥ 2 individuals with essential hypertension clinically diagnosed before age 60 were invited to participate, including both hypertensive and normotensive siblings. DNA methylation was measured on the peripheral blood leukocytes of 422 unrelated African American participants using stored blood samples collected during the Phase I examination. Participants were excluded if they were identified as an outlier in principal component plots generated during the methylation data cleaning process. A total of 420 African American GENOA participants were included in this analysis. All participants provided written informed consent for genetic research.

DNA methylation sample, measurement, normalization, and quality control

Genomic DNA of 422 participants was extracted from stored peripheral blood leukocytes collected at the Phase I GENOA examination. The EZ DNA Methylation Gold Kit (Zymo Research, Irvine CA) was used for bisulfite conversion, and methylation was measured with the Illumina Infinium HumanMethylation450 BeadChip. The *minfi* R package was used to preprocess, normalize (SWAN), and calculate beta values. Principal components analysis was performed using the SWAN method to identify and exclude sample outliers (>6sd from the mean of the top 10 PCs). The proportion of each cell type was estimated using Houseman's method. Detection p-values were calculated for each sample at each CpG site, and values were set as missing when detection P-value was >0.01. CpG sites were excluded if >10% of samples had a detection P-value of >0.01. All samples had a call rate >90%.

Smoking phenotype

Participants were categorized as being a current smoker (smoker within the past 1 year), former smoker (not having smoked in the past 1 year), or never smoker. A person was considered a never smoker if they answered "No" to the following question: "Have you ever smoked more than 100 cigarettes in your entire life?". A person was considered a former smoker if they answered "Yes" to "Have you ever smoked more than 100 cigarettes in your entire life?", answered "No" to "Do you now smoke cigarettes?", and there was greater than 1 year between their current age/date of exam and their answer to the question, "In what year or how old were you when you last quit smoking?" A person was considered a current smoker if they answered "Yes" to "Have you ever smoked more than 100 cigarettes in your entire life?" and answered "Yes" to "Do you now smoke cigarettes?". A person was also considered a current smoker if they answered "Yes" to "Have you ever smoked more than 100 cigarettes in your entire life?", answered "No" to "Do you now smoke cigarettes?", and there was less than 1 year between their current age/date of exam and their answer to the question, "In what year or how old were you when you last quit smoking?".

Analysis

GENOA data were analyzed with linear mixed effect models using the R software, as described in the main paper. DNA methylation beta values were used as the outcome variables.

Acknowledgements

Support for the Genetic Epidemiology Network of Arteriopathy was provided by the National Heart, Lung and Blood Institute (HL054457, HL100185, and HL119443) of the National Institutes of Health. We appreciate technical assistance from Jodie L. Van de Rostyne, Pamela I. Hammond, Julie M. Cunningham, and the Mayo Clinic Advanced Genomics Technology Center. We would also like to thank the families that participated in the GENOA study.

Cardiovascular Health Study (CHS)

Description

The CHS is a population-based cohort study of risk factors for coronary heart disease and stroke in adults ≥65 years conducted across four field centers[32]. The original predominantly European ancestry cohort of 5,201 persons was recruited in 1989-1990 from random samples of the Medicare eligibility lists; subsequently, an additional predominantly African-American cohort of 687 persons was enrolled for a total sample of 5,888.

DNA methylation was measured on 200 European ancestry and 200 African-American ancestry participants. The samples were randomly selected among participants without presence of coronary heart disease, congestive heart failure, peripheral vascular disease, valvular heart disease, stroke or transient ischemic attack at study baseline or lack of available DNA at study year 5.

CHS was approved by institutional review committees at each field center and individuals in the present analysis had available DNA and gave informed consent including consent to use of genetic information for the study of cardiovascular disease.

DNA methylation sample, measurement, normalization, and quality control

Methylation measurements were performed at the Institute for Translational Genomics and Population Sciences at the Harbor-UCLA Medical Center Institute for Translational Genomics and Population Sciences using the Infinium HumanMethylation450 BeadChip (Illumina Inc, San Diego, CA).

Quality control was performed in in the minfi R package[33–35] (version 1.12.0, http://www.bioconductor.org/packages/release/bioc/html/minfi.html). Samples with low median intensities of below 10.5 ($\log_2$) across the methylated and unmethylated channels, samples with a proportion of probes falling detection of greater than 0.5%, samples with QC probes falling greater than 3 standard deviation from the mean, sex-check mismatches, or failed concordance with prior genotyping were removed. In total, 11 samples were removed for sample QC resulting in a sample of 191 European-ancestry and 198 African-American samples. Methylation values were normalized using the SWAN quantile normalization method[34]. Since white blood cell proportions were not directly measured in CHS they were estimated from the methylation data using the Houseman method[36].

Smoking phenotype

Smoking phenotype is as described in the main paper.

Analysis

All association analyses were performed in R using linear models with DNA methylation beta values as the outcome. Analyses were stratified by race and all analyses were adjusted for age, gender, total white blood cell count, study clinic and estimated white blood cell proportions, as well as chip number and position on the chip.

Acknowledgements

CHS: Infrastructure for the CHARGE Consortium is supported in part by the National Heart, Lung, and Blood Institute grant R01HL105756. The CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086; and NHLBI grants U01HL080295, R01HL087652, R01HL092111, R01HL105756, R01HL103612, R01HL111089, and R01HL120393 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through R01AG023629 from the National Institute on Aging (NIA) as well as Laughlin Family, Alpha Phi Foundation, and Locke Charitable Foundation. A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

European Prospective Investigation into Cancer and Nutrition-Norfolk (EPIC-Norfolk)

Description

The European Prospective Investigation of Cancer (EPIC)-Norfolk study enrolled more than 25,000 community-based men and women at baseline (1993-1997), who were aged 40-79 years old and registered with a participating general practitioner in and around the city of Norwich (Norfolk, UK). The full details of the study design and follow up of participants has been reported previously[37]. Written informed consent was obtained from all participants. The study complies with the principles of the Declaration of Helsinki and ethical approval was given by the Norfolk Local Research Ethics Committee and the East Norfolk and Waveney NHS Research Governance Committee.

DNA methylation sample, measurement, normalization, and quality control

DNA was isolated from white blood cells as per the standard DNA extraction procedure (Autopure LS, Qiagen). DNA methylome profiling was carried out using the Illumina Infinium HumanMethylation450 (HM450) BeadChip assay. 500 ng of extracted DNA was bisulfite-modified using the EZ DNA Methylation kit (Zymo Research, D5004) following the manufacturer’s instructions. The minfi R package was used to preprocess, normalize (SWAN), and calculate beta values.

Smoking phenotype

Personal medical history was assessed using the question in the Health and Lifestyle Questionnaire. Yes/no responses to the questions “Have you ever smoked as much as one cigarette a day for as long as a year?” and “Do you smoke cigarettes now?” were used to derive smoking history

Analysis

All association analyses were performed in R using linear models with DNA methylation beta values as the outcome. Analyses were analyses were adjusted for age, gender, and estimated white blood cell proportions, as well as plate number and position.

Acknowledgements

EPIC-Norfolk is supported by program grants from the Medical Research Council (MRC) [G9502233; G0401527] and Cancer Research UK [C864/A8257]. The generation and management of the Illumina 450K methylation array data in this cohort is supported through the MRC Cambridge initiative in metabolomic science [MR/L00002/1]. CEE, KKO and NJW are supported by MRC program grants [MC_UU_12015/1 and MC_UU_12015/2].

Normative Aging Study (NAS)

Description

The US Department of Veterans Affairs (VA) Normative Aging Study (NAS) is an ongoing longitudinal cohort established in 1963, which included men who were aged 21 to 80 years and free of known chronic medical conditions at entry[38]. Participants were subsequently invited to medical examinations every 3 to 5 years. At each visit, participants provided information on medical history, lifestyle, and demographic factors, and underwent a physical examination and laboratory tests. DNA samples were collected from 1999 to 2007 from the active participants and used for DNA methylation analysis.

DNA was extracted from buffy coat using the QIAamp DNA Blood Kit (QIAGEN, Valencia, CA, USA). A total of 500 ng of DNA was used to perform bisulfite conversion using the EZ-96 DNA Methylation Kit (Zymo Research, Orange, CA, USA). To limit chip and plate effects, a two-stage age-stratified algorithm was used to randomize samples and ensure similar age distributions across chips and plates; we randomized 12 samples - which were sampled across all the age quartiles - to each chip, then chips were randomized to plates (each housing eight chips). Quality control analysis was performed to remove samples where >1% of probes had a detection P value >0.05 and probes where >1% of passing samples had a detection P value >0.05. The passing samples were preprocessed using out-of-band background correction[39], dye bias adjustment, and probe type adjustment using the Beta MIxture Quantile dilation (BMIQ) method[15].

Smoking phenotype

At each in-person examination visit, participants completed a questionnaire that included their smoking status that was classified as in the main paper.

Analysis

Data were analyzed with linear mixed effects models, as in the main paper with main models adjusted with a fixed effect for age and an indicator for sentrix column (position on chip) and random effects for sentrix row (position on chip) and chip number. As the NAS does not include females there was no adjustment for sex.

Acknowledgements

The present work on the US Department of Veterans Affairs (VA) Normative Aging Study has been supported by funding from the U.S. National Institute of Environmental Health Sciences (NIEHS) (R01ES015172, R01ES021733). The VA Normative Aging Study is supported by the Cooperative Studies Program/ERIC, US Department of Veterans Affairs, and is a research component of the Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC). Additional support to the VA Normative Aging Study was provided by the US Department of Agriculture, Agricultural Research Service (contract 53-K06-510). The views expressed in this paper are those of the authors and do not necessarily represent the views of the US Department of Veterans Affairs.

References

1. Feinleib M, Kannel WB, Garrison RJ, McNamara PM, Castelli WP. The Framingham Offspring Study. Design and preliminary data. *Prev Med*. 1975;4:518–525.

2. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*. 2013;14:293.

3. Vazquez AI, Bates DM, Rosa GJM, Gianola D, Weigel KA. Technical note: an R package for fitting generalized linear mixed models in animal breeding. *J Anim Sci*. 2010;88:497–504.

4. Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study [Internet]. Available from: http://www.biostat.wustl.edu/goldn/

5. Absher DM, Li X, Waite LL, Gibson A, Roberts K, Edberg J, Chatham WW, Kimberly RP. Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations. *PLoS Genet*. 2013;9:e1003678.

6. Hidalgo B, Irvin MR, Sha J, Zhi D, Aslibekyan S, Absher D, Tiwari HK, Kabagambe EK, Ordovas JM, Arnett DK. Epigenome-Wide Association Study of Fasting Measures of Glucose, Insulin, and HOMA-IR in the Genetics of Lipid Lowering Drugs and Diet Network Study. *Diabetes*. 2014;63:801–807.

7. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2006;8:118–127.

8. Hofman A, Brusselle GGO, Darwish Murad S, van Duijn CM, Franco OH, Goedegebure A, Ikram MA, Klaver CCW, Nijsten TEC, Peeters RP, Stricker BHC, Tiemeier HW, Uitterlinden AG, Vernooij MW. The Rotterdam Study: 2016 objectives and design update. *Eur J Epidemiol*. 2015;30:661–708.

9. 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;4:325–341.

10. Ferrucci L, Bandinelli S, Benvenuti E, Di Iorio A, Macchi C, Harris TB, Guralnik JM. Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. *J Am Geriatr Soc*. 2000;48:1618–1625.

11. Debey-Pascher S, Eggle D, Schultze JL. RNA stabilization of peripheral blood and profiling by bead chip analysis. *Methods Mol Biol*. 2009;496:175–210.

12. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai S-L, Arepalli S, Dillman A, Rafferty IP, Troncoso J, Johnson R, Zielke HR, Ferrucci L, Longo DL, Cookson MR, Singleton AB. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genetics*. 2010;6:e1000952.

13. Holle R, Happich M, Löwel H, Wichmann HE, MONICA/KORA Study Group. KORA--a research platform for population based health research. *Gesundheitswesen*. 2005;67 Suppl 1:S19-25.

14. Wichmann H-E, Gieger C, Illig T, MONICA/KORA Study Group. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*. 2005;67 Suppl 1:S26-30.

15. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*. 2013;29:189–196.

16. Pfeiffer L, Wahl S, Pilling LC, Reischl E, Sandling JK, Kunze S, Holdt LM, Kretschmer A, Schramm K, Adamski J, Klopp N, Illig T, Hedman ÅK, Roden M, Hernandez DG, Singleton AB, Thasler WE, Grallert H, Gieger C, Herder C, Teupser D, Meisinger C, Spector TD, Kronenberg F, Prokisch H, Melzer D, Peters A, Deloukas P, Ferrucci L, Waldenberger M. DNA methylation of lipid-related genes affects blood lipid levels. *Circ Cardiovasc Genet*. 2015;8:334–342.

17. Gillespie CF, Bradley B, Mercer K, Smith AK, Conneely K, Gapen M, Weiss T, Schwartz AC, Cubells JF, Ressler KJ. Trauma exposure and stress-related disorders in inner city primary care patients. *Gen Hosp Psychiatry*. 2009;31:505–514.

18. Kellogg SH, McHugh PF, Bell K, Schluger JH, Schluger RP, LaForge KS, Ho A, Kreek MJ. The Kreek-McHugh-Schluger-Kellogg scale: a new, rapid method for quantifying substance abuse and its possible applications. *Drug Alcohol Depend*. 2003;69:137–150.

19. Deary IJ, Gow AJ, Pattie A, Starr JM. Cohort profile: the Lothian Birth Cohorts of 1921 and 1936. *Int J Epidemiol*. 2012;41:1576–1584.

20. Deary IJ, Gow AJ, Taylor MD, Corley J, Brett C, Wilson V, Campbell H, Whalley LJ, Visscher PM, Porteous DJ, Starr JM. The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. *BMC Geriatr*. 2007;7:28.

21. Deary IJ, Whiteman MC, Starr JM, Whalley LJ, Fox HC. The impact of childhood intelligence on later life: following up the Scottish mental surveys of 1932 and 1947. *J Pers Soc Psychol*. 2004;86:130–147.

22. Shah S, McRae AF, Marioni RE, Harris SE, Gibson J, Henders AK, Redmond P, Cox SR, Pattie A, Corley J, Murphy L, Martin NG, Montgomery GW, Starr JM, Wray NR, Deary IJ, Visscher PM. Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res*. 2014;24:1725–1733.

23. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, Jacob DR, Kronmal R, Liu K, Nelson JC, O’Leary D, Saad MF, Shea S, Szklo M, Tracy RP. Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am J Epidemiol*. 2002;156:871–881.

24. Liu Y, Ding J, Reynolds LM, Lohman K, Register TC, De La Fuente A, Howard TD, Hawkins GA, Cui W, Morris J, Smith SG, Barr RG, Kaufman JD, Burke GL, Post W, Shea S, McCall CE, Siscovick D, Jacobs DR, Tracy RP, Herrington DM, Hoeschele I. Methylomics of gene expression in human monocytes. *Hum Mol Genet*. 2013;22:5065–5074.

25. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005;102:15545–15550.

26. Abbas AR, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F, Fong S, van Lookeren Campagne M, Godowski P, Williams PM, Chan AC, Clark HF. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes Immun*. 2005;6:319–331.

27. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, Charrondière UR, Hémon B, Casagrande C, Vignat J, Overvad K, Tjønneland A, Clavel-Chapelon F, Thiébaut A, Wahrendorf J, Boeing H, Trichopoulos D, Trichopoulou A, Vineis P, Palli D, Bueno-De-Mesquita HB, Peeters PHM, Lund E, Engeset D, González CA, Barricarte A, Berglund G, Hallmans G, Day NE, Key TJ, Kaaks R, Saracci R. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr*. 2002;5:1113–1124.

28. Hernandez-Vargas H, Castelino J, Silver MJ, Dominguez-Salas P, Cros M-P, Durand G, Calvez-Kelm FL, Prentice AM, Wild CP, Moore SE, Hennig BJ, Herceg Z, Gong YY, Routledge MN. Exposure to aflatoxin B1 in utero is associated with DNA methylation in white blood cells of infants in The Gambia. *Int J Epidemiol*. 2015;44:1238–1248.

29. Chen Y, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*. 2013;8:203–209.

30. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JSSRB*. 1995;57:289–300.

31. Daniels PR, Kardia SLR, Hanis CL, Brown CA, Hutchinson R, Boerwinkle E, Turner ST, Genetic Epidemiology Network of Arteriopathy study. Familial aggregation of hypertension treatment and control in the Genetic Epidemiology Network of Arteriopathy (GENOA) study. *Am J Med*. 2004;116:676–681.

32. Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A. The Cardiovascular Health Study: design and rationale. *Ann Epidemiol*. 1991;1:263–276.

33. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30:1363–1369.

34. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol*. 2012;13:R44.

35. Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, Greenwood CM, Hansen KD. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol*. 2014;15:503.

36. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13:86.

37. Day N, Oakes S, Luben R, Khaw KT, Bingham S, Welch A, Wareham N. EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br J Cancer*. 1999;80 Suppl 1:95–103.

38. Bell B, Rose CL, Damon A. The Veterans Administration longitudinal study of healthy aging. *Gerontologist*. 1966;6:179–184.

39. Triche TJ, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res*. 2013;41:e90.