Epigenetic Signatures of Cigarette Smoking

Roby Joehanes[1,2*], Allan C. Just[3*], Riccardo E. Marioni[4,5,6*], Luke C. Pilling[7*], Lindsay M. Reynolds[8*], Pooja R. Mandaviya[9,10*], Weihua Guan[11*], Tao Xu[12*], Cathy E. Elks[13*], Stella Aslibekyan[14*], Hortensia Moreno-Macias[15,16*], Jennifer A. Smith[17*], Jennifer A Brody[18*], Radhika Dhingra[19*], Paul Yousefi[20], James S. Pankow[21], Sonja Kunze[12], Sonia Shah[6,22], Allan F. McRae[6,22], Kurt Lohman[23], Jin Sha[14], Devin M. Absher[24], Luigi Ferrucci[25], Wei Zhao[17], Ellen W. Demerath[20], Jan Bressler[26], Megan L. Grove[26], Tianxiao Huan[2], Chunyu Liu[2], Michael M. Mendelson[2,27], Chen Yao[2], Douglas P. Kiel[1], Annette Peters[12], Rui Wang-Sattler[12], Peter M. Visscher[4,6,22], Naomi R. Wray[6], John M. Starr[4,28], Jingzhong Ding[29], Carlos J. Rodriguez[8], Nicholas J. Wareham[13], Marguerite R. Irvin[14], Degui Zhi[30], Myrto Barrdahl[31], Paolo Vineis[32,33], Srikant Ambatipudi[16] , André G. Uitterlinden[9], Albert Hofman[34], Joel Schwartz[35], Elena Colicino[35], Lifang Hou[36], Pantel S. Vokonas[37], Dena G. Hernandez[38], Andrew B. Singleton[38], Stefania Bandinelli[39], Stephen T. Turner[40], Erin B. Ware[17,41], Alicia K. Smith[42], Torsten Klengel[43,44], Elisabeth B. Binder[43,45], Bruce M. Psaty[18,47], Kent D. Taylor[47,48,49], Sina A. Gharib[50], Brenton R. Swenson[18], Liming Liang[51], Dawn L. DeMeo[52], George T. O'Connor[53], Zdenko Herceg[16], Kerry J. Ressler[42,44,54], Karen N. Conneely[55#], Nona Sotoodehnia[17#], Sharon L. R. Kardia[17#], David Melzer[7#], Andrea A. Baccarelli[35,56#], Joyce B. J. van Meurs[9#], Isabelle Romieu[16#], Donna K. Arnett[14#], Ken K. Ong[13#], Yongmei Liu[8#], Melanie Waldenberger[12#], Ian J. Deary[4,57#], Myriam Fornage[26,58#], Daniel Levy[2#], Stephanie J. London[59#]

*These authors contributed equally as first authors

#These authors contributed equally as senior authors

Correspondence is to be sent to Stephanie J. London (london2@niehs.nih.gov; T: 919-541-5772; F: 301-480-3290)

Affiliations

[1]Institute for Aging Research, Hebrew SeniorLife, Department of Medicine Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA

[2]Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA, and the Framingham Heart Study, Framingham, MA, USA

[3]Department of Preventive Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[4]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, UK

[5]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, UK

[6]Queensland Brain Institute, University of Queensland, Australia

[7]Epidemiology and Public Health Group, Institute of Biomedical and Clinical Science, University of Exeter Medical School, Exeter, UK

[8]Department of Epidemiology & Prevention, Public Health Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA

[9]Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands

[10]Department of Clinical Chemistry, Erasmus University Medical Center, Rotterdam, The Netherlands

[11]Division of Biostatistics, Schoold of Public Health, Univerisity of Minnesota, Minneapolis, MN, USA

[12]Research Unit of Molecular Epidemiology, Institute of Epidemiology II, Munich, Germany

[13]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK

[14]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA

[15]Autonomous Metropolitan University-Iztapalapa, Mexico City, Mexico

[16]International Agency for Research on Cancer (IARC)

[17]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA

[18]Cardiovascular Health Research Unit, Department of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, WA, USA

[19]Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA

[20]School of Public Health, University of California, Berkeley, CA, USA

[21]Division of Epidemiology & Community Health, School of Public Health, University of Minnesota, Minneapolis, USA

[22]University of Queensland Diamantina Institute, Translational Research Institute, The University of Queensland, Brisbane 4072, QLD, Australia

[23]Department of Biostatistical Sciences, Division of Public Health Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA

[24]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA

[25]Clinical Research Branch, National Institute on Aging, Baltimore, MD, USA

[26]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, USA
[27]Children's Hospital, Boston, MA, USA
[28]Alzheimer Scotland Dementia Research Centre, University of Edinburgh, UK
[29]Department of Internal Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA
[30]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, USA
[31]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ) Heidelberg, Im Neuenheimer Feld 581, D-69120, Heidelberg, Germany
[32]MRC/PHE Centre for Environment and Health, School of Public Health, Imperial College London, UK
[33]HuGeF Foundation, Torino, Italy
[34]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands
[35]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[36]Department of Preventive Medicine and the Robert H. Lurie Comprehensive Cancer Center, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
[37]VA Normative Aging Study, VA Boston Healthcare System and Department of Medicine, Boston University School of Medicine, Boston, MA, USA
[38]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA
[39]Geriatric Unit, Azienda Sanitaria di Firenze, Florence, Italy
[40]Division of Nephrology and Hypertension, Mayo Clinic, Rochester, MN, USA
[41]Research Center for Group Dynamics, Institute for Social Research, University of Michigan, Ann Arbor, MI
[42]Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA, USA
[43]Department of Translational Research in Psychiatry, Max-Planck Institute of Psychiatry, Munich, Germany
[44]Division of Depression & Anxiety Disorders, McLean Hospital, Belmont, MA, USA
[45]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA, USA
[47]Group Health Research Institute, Group Health Cooperative, Seattle, WA, USA
[47]Institute for Translational Genomics and Population Sciences, Los Angeles BioMedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA.
[48]Division of Genomic Outcomes, Department of Pediatrics, Harbor-UCLA Medical Center, Torrance, CA, USA.
[49]Departments of Pediatrics, Medicine, and Human Genetics, UCLA, Los Angeles, CA, USA
[50]Center for Lung Biology, Division of Pulmonary & Critical Care Medicine, Department of Medicine, University of Washington, Seattle, WA
[51]Harvard School of Public Health, Boston, MA, USA
[52]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[53]Boston University School of Medicine, Boston, MA, USA
[54]Department of Psychiatry, Harvard Medical School, Boston MA
[55]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA
[56]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[57]Department of Psychology, University of Edinburgh, UK
[59]Institute of Molecular Medicine, The University of Texas Health Science Center McGovern Medical School, Houston, TX, USA
[60]Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, NC, USA

Abstract

Background: Cigarette smoking increases the risk of multiple diseases including cancers, osteoporosis, lung, and cardiovascular disorders. DNA methylation leaves a long-term signature of smoking exposure and is one potential mechanism by which tobacco exposure predisposes to these adverse health outcomes.

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 2,433 current, 6,518 former, and 6,956 never smokers). Comparing current versus never smokers, 2,623 CpG sites (CpGs), annotated to 1,405 genes, were statistically significantly differentially methylated at Bonferroni threshold of $p<1x10^{-7}$. 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 ($p<1x10^{-7}$), indicating a pattern of persistent altered methylation, with attenuation, after smoking cessation. Transcriptomic integration identified effects on gene expression at many differentially methylated CpGs.

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 biologic 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.

Keywords: Epigenomic, smoking, meta-analysis, epidemiology

Introduction

Cigarette smoking is a major causal risk factor for various diseases including cancers, cardiovascular disease (CVD), chronic obstructive pulmonary disease (COPD)[1], and osteoporosis[2]. 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 US, there are more former smokers than current smokers[3]. Despite the decline in the prevalence of smoking in many countries, it remains the leading preventable cause of death in the world, accounting for nearly 6 million deaths each year[4].

Even decades after cessation, cigarette smoking confers long-term risk of diseases including some cancers[5], chronic obstructive pulmonary disease, and stroke[6]. The mechanisms for these long-term effects are not well understood. DNA methylation changes have been proposed as one possible explanation.

DNA methylation appears to reflect exposure to a variety of lifestyle factors[7,8], including cigarette smoking. Several studies have shown reproducible associations between tobacco smoking and altered DNA methylation at multiple cytosine-phosphate-guanine (CpG) sites[9–20]. Some DNA methylation sites associated with tobacco smoking have also localized to genes related to coronary heart disease[10] and pulmonary disease[21]. Some studies have found different associated CpGs in smokers versus non-smokers[13,16]. 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 employed. 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[18] 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 4,750 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 which can identify signals that do not reach statistical significance in individual studies [22].

In the current study, we meta-analyzed association results between DNA methylation and cigarette smoking in 15,907 individuals from 16 cohorts in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium 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 to non-smokers 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 (Supplementary Table 1). 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. Of these, 12,161 are of European Ancestry (EA) and 3,746 are of African Ancestry (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 (Supplementary Table 1). 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 online Supplementary Materials.

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[23] and BMIQ[24], 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 online Supplementary Materials.

Smoking phenotype definition

Self-reported cigarette smoking status was divided into three categories. Current smokers were defined as those who have smoked at least one cigarette a day within 12 months prior to the blood draw, former smokers were defined as those who had ever smoked at least one cigarette a day, but had stopped at least 12 months prior to the blood draw, and never smokers reported never having smoked. Pack years was calculated based on self-report as the average number of cigarettes per day smoked 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 two 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[25]. 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 (*e.g.*, for FHS, see Supplementary Material 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 (BMI) because it is associated with methylation at some loci, making it a potential confounder[26]. Only three cohorts participated in model 2 analysis: FHS, KORA, and NAS. Model 3 substituted smoking phenotypes for pack years. Only three cohorts participated in model 3 analysis: FHS, Rotterdam, and Inchianti. The pack year analysis was performed only on two subsets: current *vs*. never smokers and former *vs*. never smokers. Combining all three categories would require accurate records of time of quitting, which among the three cohorts was available for only FHS. To investigate cell type differences, we removed blood counts from Model 1 and called it Model 4. Only three cohorts participated in this analysis: FHS, KORA, and NAS. All models were run with the lme4 package[27] in R[28], except for FHS (See Supplementary Materials for details).

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

$$E_i = \mu + s_i + e_i, \qquad (2)$$

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 (REML) criterion in R's *metafor*[29] package. Multiple-testing adjustment on the resulting p-values was performed using the False Discovery Rate (FDR) method of Benjamini and Hochberg[30]. 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 0-100%.

Literature review to identify genes previously associated with smoking and methylation

We used the same literature search strategy published previously[31]. A broad query of NCBI's 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 >1,000 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, which is found in Supplementary Table 4 of Joubert et al.

article[31]. CpG level results (P-values and gene annotations) for sites showing genome-wide statistically significant associations (FDR<0.05) were extracted and resulted in 1,185 genes previously associated with adult or maternal smoking. All CpGs annotated to these 1,185 genes were marked as "previously found."

Gene set enrichment analysis (GSEA)

Gene-Set Enrichment Analysis (GSEA)[32] 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 (GO) biological process (C5-BP) and collected all categories with FDR<0.05 (up to 100 categories).

Enrichment analysis for localization to different genomic features

Enrichment analysis on genomic features were performed using the annotation file supplied by the 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 (*i.e.*, body, first exon, 3' UTR, 5'UTR, within 200 base pairs of Transcriptional Start Site [TSS200], and TSS1500), the relation of CpG site to a CpG island (*i.e.*, 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 (DHS). Enrichment analysis was performed using one-sided Fisher's exact set for each feature, using R's `fisher.test`.

Genome-wide association study (GWAS) analysis

We intersected our results with SNPs having genome-wide association study (GWAS) p-values≤5 x 10^{-8} in the NHGRI GWAS catalog (accessed November 2, 2015)[33]. The catalog contained 9,777 SNPs annotated to 7,075 genes associated with 865 phenotypes at p≤5x10^{-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 one-sided Fisher's exact test.

For bone mineral phenotype enrichment, we included all SNPs containing terms "bone mineral density" or "osteoporosis". For cardiovascular disease (CVD), we included all SNPs containing terms "cardiovascular disease", "stroke", "coronary disease", "cardiomyopathy", or "myocardial infarction". For CVD 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 over 35 years. The analysis was performed on seven dichotomous variables, indicating cessation of smoking

for 5, 10, 15, 20, 25, and 30 years versus never smokers. For example, for five year cessation variable, those who quit smoking five years or more are marked as ones, while never smokers are marked as zeroes and current smokers are excluded. For this analysis, we used the *pedigreemm* package[34] with the same set of covariates as in the primary analysis. Sites with $p<0.002$ across all seven variables were deemed to be statistically significant compared to never-smoker levels.

Methylation by expression (MxE) analysis

To determine transcriptomic association of each significant CpG site, we interrogated such CpG sites in the FHS gene-level methylation by expression (MxE) database, at genome-wide false discovery rate (FDR)<0.05. The MxE database was constructed from 2,262 individuals from the FHS Offspring cohort attending examination cycle eight (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 one-sided Fisher's 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 African Ancestry (AA) and European Ancestry (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, CHS-EA) was performed separately from those of AA cohorts (ARIC, GTP, GENOA, CHS-AA). The meta-analysis procedure was identical to that discussed previously.

Analysis of samples types for DNA extraction

Meta-analysis of the results of 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) was performed identically to that discussed previously. 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 Supplementary Table 1.

Current versus Never Smokers

In the meta-analysis of current cigarette smokers (N=2,433) versus never smokers (N=6,956), 2,623 CpGs annotated to 1,405 genes met Bonferroni significance after correction for 485,381 tests ($P<1x10^{-7}$). Based on genome-wide false discovery rate (FDR)<0.05, 18,760 CpG sites (CpGs) annotated to 7,201 genes were differentially methylated. There was a moderate inflation factor[35] λ of 1.32 (Supplementary Figure 1), which is consistent with a large number of sites being impacted by smoking. Our results

lend support many previously reported loci[12,13,16,18], including CpGs annotated to *AHRR*, *RARA*, *F2RL3*, and *LRRN3* (Supplementary Table 2). Not surprisingly, cg05575921 annotated to *AHRR*, the top CpG identified in most prior studies of smoking, was highly significant in our meta-analysis ($P=4.6x10^{-26}$; ranked 36, Supplementary Table 2) 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 6,720 genes) have not been previously reported to be associated with cigarette smoking – these include 1,500 of the 2,623 CpGs that met Bonferroni significance. The 25 CpGs with lowest p-values for both overall and novel findings are shown in Table 2. Supplementary Table 2 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 (BMI) into the model did not appreciably alter the results (Supplementary Figure 2).

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 Supplementary Figure 3.

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 7,201 genes in regards to the length of

the gene or number of exons and found only three 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[9,12,14] as well as our own, and its key role in the AHR pathway which is crucial in the response to polyaromatic hydrocarbons, such as are produced by smoking[36], it seems very unlikely that the *AHRR* findings are false positives. Likewise there is strong support in the literature for *PRRT1*[37] and *TNF*[38]. The enrichment results for methylation platform coverage also yielded the same three genes.

In a subset of three cohorts (1,827 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 (Supplementary Table 3).

To investigate the pathways implicated by these genes, we performed a gene-set enrichment analysis[39] 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=2.8×10^{-79}), protein metabolic processes (FDR=1.2×10^{-43}), and transcription pathways (FDR=8.4×10^{-31}). The complete list of 99 enriched molecular processes can be found in Supplementary Table 4.

Former versus Never Smokers

Meta-analysis of former (N=6,518) versus never smokers (N=6,956) restricted to the 18,760 CpG sites that were differentially methylated in current versus never smokers

identified 2,568 CpGs annotated to 1,326 genes at FDR<0.05 (Supplementary Table 5). There were 285 CpGs (annotated to 149 genes) that also met Bonferroni correction (P< $0.05/18760 \approx 2.67 \times 10^{-6}$). There was no evidence of inflation[35] (λ=0.98) (Supplementary Figure 4). We also confirmed previously reported findings for CpGs annotated to *AHRR*, *RARA*, and *LRRN3*[12,13,16,18]. Effect sizes of these CpGs were all weaker than in the analysis of current versus never smokers [61.2% ±15.3% weaker] for the 2,568 CpGs that remained significantly differentially methylated in former *vs.* never smokers compared with current *vs*. never smokers. Results for the top 25 CpGs are displayed in Table 3. Adding BMI to the model did not appreciably alter the results (Supplementary Figure 5). A volcano plot can be found in Supplementary Figure 6. In a subset of three cohorts (3,349 subjects), analyses using pack-years confirmed a significant dose response for 1,804 of the 2,568 CpGs (70%) annotated to 942 genes at FDR<0.05 (Supplementary Table 6).

The gene-set enrichment analysis[32] in the former versus never smoker analyses on all 1,326 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}) (Supplementary Table 7). The gene-set enrichment analysis on the 942 genes for which the 1,804 CpGs exhibited dose responses with pack-years also revealed similar pathways to those summarized in Supplementary Table 7, except with weaker enrichment FDR values.

In 2,648 Framingham Heart Study participants with up to 30 years of prospectively collected smoking data, we examined the 2,568 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 five 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 1, Table 4).

The EPIC studies included cancer cases plus non-cancer 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 $\rho = 0.99$ 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 NHGRI-EBI GWAS Catalog[33] (accessed November 2, 2015). The catalog contained 9,777 SNPs annotated to 7,075 genes associated with 865 phenotypes at $p \leq 5 \times 10^{-8}$. Of the 7,201 genes (mapped by 18,760 CpG sites) significantly differentially methylated in current versus never smokers, we found overlap with 1,791 genes (4,187 CpGs are mapped to these) associated in GWAS with 700 phenotypes (enrichment $p = 2.4 \times 10^{-52}$). We identified smoking-related traits using the 2014 US Surgeon General's (USSG)

report[2]. Enrichment results for a selection of smoking-related phenotypes including coronary heart disease (CHD) 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 2,568 CpGs associated with former versus never smoking status. We identified enrichment for CHD, pulmonary traits, and some cancers (Table 5). More detailed results are available in Supplementary Tables 8 and 9. Differentially methylated genes in relation to smoking status that are associated in GWAS with CHD or CHD risk factors are available in Supplementary Table 10. 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 (Supplementary Table 11). Trends were similar for the two 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, 1,430 were significantly associated in *cis* with the expression of 924 genes at FDR<0.05 (enrichment $p=3.6 \times 10^{-215}$, Supplementary Table 12) using whole blood samples from 2,262 Framingham Heart Study participants. Of these, 424 CpGs associated with the expression of 285 genes were replicated at FDR<0.0001 in 1,264 CD14+ samples from the Multi-Ethnic Study of Atherosclerosis (MESA)[40]. These genes are associated with pathways similar to those described earlier (Supplementary Table 13).

Comparison between African ancestry and European ancestry

Meta-analysis of the current versus never smokers in 11 cohorts with participants of European ancestry (N=6,750 subjects) yielded 10,977 CpGs annotated to 4,940 genes at FDR<0.05. Meta-analysis the results of the smaller dataset of four cohorts with African ancestry participants (N=2,639) yielded 3,945 CpGs annotated to 2,088 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 $\rho=0.89$). The results by ancestry are shown in Supplementary Table 14.

We performed the same ancestry-stratified analysis on former versus never smokers (Supplementary Table 15). Meta-analysis of the results of European ancestry participants yielded 2,045 CpG sites annotated to 1,081 genes at FDR<0.05. Meta-analysis of the results of African ancestry 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 (2,234 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 European ancestry sample size, we performed a meta-analysis on subset of European ancestry 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 African ancestry 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 due to better power in the European ancestry 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[25]. 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 *vs*. never smoker analysis (Supplementary Figure 7). Similarly for the analysis of former versus never smokers the effect estimates were highly correlated before

and after adjustment (ρ=0.93; Supplementary Figure 8). In addition, in two 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; Supplementary Table 16).

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 Supplementary Figure 9. The profile across 18,760 CpG sites significantly associated with current *vs*. never smoking status can be found in Supplementary Figure 10. 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 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 over 7,000 genes, or roughly a 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 over 16,000 novel differentially methylated CpGs in response to cigarette smoking. Many of these genes have not been previously implicated in the biologic 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 lifespan[2], 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 (Supplementary Tables 2 and 5) because we accounted for inter-study variability by using random effect meta analyses, which is conservative when heterogeneity is present[41]. 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, following smoking cessation.

We found that genes differentially methylated in relation to smoking are enriched for variants associated in GWAS with smoking–related diseases[2] including, osteoporosis, colorectal cancers, chronic obstructive pulmonary disease, pulmonary function, cardiovascular disease (CVD) 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 [42]. 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[43]. 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 five 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. While speculative, it is possible that persistent methylation changes at some loci might contribute to risks of some conditions that remain elevated after smoking cessation.

In all but two 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 (*i.e.*, in *cis*) even though a single measurement of gene expression in blood is probably subject to considerably more within-subject variability than DNA methylation,[44] 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 DHS sites are more dynamic (*i.e.*, susceptible to methylation changes) than CpG islands[45], which may be more resistant

to abrupt changes in DNA methylation in response to environmental exposures[46]. Thus our results suggest that many of the smoking-associated CpG sites may have regulatory effects.

While 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 past smoking history[14,47]. Indeed, several studies have identified several of such markers[10,47,48]. The large number of persistently modified CpGs may be 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 as well as to validate self-reported never smoker status. Further, our analyses of both former and current smokers show dose-dependent effects at a number of CpGs (Supplementary Tables 3 and 7). Methylation based biomarkers could be informative for investigating dose response relationships with disease endpoints. This is useful because smokers often underreport 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 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 (*e.g.*, heart and lung) would be of additional interest. In addition, our analyses could not distinguish smoking's direct effects from its indirect effects due to 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 past smoking history. These findings may provide new insights into molecular mechanisms underlying the protean effects of smoking on human health and disease.

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 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.

Funding Sources

Infrastructure for the CHARGE Consortium is supported in part by the National Heart, Lung, and Blood Institute grant R01HL105756. Additional funding sources for each cohort can be found in the supplementary materials.

Disclosure

B. M. P. 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.E. is currently employed by Astra Zeneca, although the work was completed prior to the employment. All other authors declare no financial interests or potential conflicts of interest.

References

1. Adler I. Primary malignant growths of the lung and bronchi. *Longmans, Green*. 1912;3–12.

2. 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 [Internet]. Atlanta (GA): Centers for Disease Control and Prevention (US); 2014 [cited 2015 Nov 24]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK179276/

3. Centers for Disease Control and Prevention. Current Cigarette Smoking Among Adults — United States, 2011. *Morbidity and Mortality Weekly Report*. 2012;61:889–894.

4. World Health Organization. WHO global report on trends in prevalence of tobacco smoking [Internet]. 2015 [cited 2015 Mar 19];Available from: http://apps.who.int/iris/bitstream/10665/156262/1/9789241564922_eng.pdf?ua=1

5. Tabuchi T, Ito Y, Ioka A, Nakayama T, Miyashiro I, Tsukuma H. Tobacco smoking and the risk of subsequent primary cancer among cancer survivors: a retrospective cohort study. *Ann Oncol*. 2013;

6. Lu M, Ye W, Adami H-O, Weiderpass E. Stroke Incidence in Women under 60 Years of Age Related to Alcohol Intake and Smoking Habit. *Cerebrovascular Diseases*. 2008;25:517–525.

7. 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.

8. Barua S, Junaid MA. Lifestyle, pregnancy and epigenetic effects. *Epigenomics*. 2015;7:85–102.

9. 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.

10. Breitling LP, Salzmann K, Rothenbacher D, Burwinkel B, Brenner H. Smoking, F2RL3 methylation, and prognosis in stable coronary heart disease. *European heart journal* [Internet]. 2012 [cited 2012 May 31];Available from: http://www.ncbi.nlm.nih.gov/pubmed/22511653

11. 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. *Human molecular genetics* [Internet]. 2012 [cited 2012 May 31];Available from: http://www.ncbi.nlm.nih.gov/pubmed/22492999

12. 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*. 2014;

13. 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.

14. 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.

15. 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.

16. 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;

17. 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.

18. 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.

19. Shah S, Bonder MJ, Marioni RE, Zhu Z, McRae AF, Zhernakova A, et al. Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. *Am J Hum Genet*. 2015;97:75–85.

20. 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 Biology*. 2007;8:R201.

21. 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;

22. Garg AX, Hackam D, Tonelli M. Systematic Review and Meta-analysis: When One Study Is Just not Enough. *Clinical Journal of the American Society of Nephrology*. 2008;3:253–260.

23. 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.

24. 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.

25. 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.

26. 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.

27. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* [Internet]. 2015 [cited 2015 Dec 9];67. Available from: http://www.jstatsoft.org/v67/i01/

28. R Development Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: 2010. Available from: http://www.R-project.org

29. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*. 2010;36.

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

31. 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.

32. 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.

33. 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.

34. 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.

35. Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999;55:997–1004.

36. 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:L391-399.

37. 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 Oncology*. 2015;1:476.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

44. Suderman M, Pappas JP, 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;

45. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature*. 2013;500:477–481.

46. Ivanova E, Chen J-H, 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.

47. 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*. 2015;

48. Zhang Y, Yang R, Burwinkel B, Breitling LP, Brenner H. F2RL3 Methylation as a Biomarker of Current and Lifetime Smoking Exposures. *Environmental Health Perspectives* [Internet]. 2013 [cited 2016 Apr 5];Available from: http://ehp.niehs.nih.gov/1306937/

Tables

Table 1. Participant characteristics

Characteristics	Current Smokers, N=2,433	Former Smokers N=6,518	Never Smokers N=6,956
Sex (% Male)	46.3%	55.6%	31.7%
Age (years)*	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

*weighted mean ± pooled standard deviation across cohorts

1

2 Table 2. Most statistically significant CpG sites that were associated with current *vs.* never

3 smoker status

Probe ID	**Chr**	**Location**	**Gene Symbol***	**Coef†**	**S.E.**	**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}
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 sites without gene names are intergenic. These are all included in all the analyses.

†Coef stands for regression coefficients

‡Not previously discovered by other studies

Table 3. Twenty-five most statistically significant CpG sites that were associated with former *vs.* never smoker status

Probe ID	Chr	Location	Gene Symbol*	Coef†	S.E.	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 sites without gene names are intergenic. These are all included in all the analyses.

†Coef stands for regression coefficients

Table 4. The top 36 most statistically significant CpG sites that did not return to never-smoker levels 30 years after smoking cessation in the Framingham Heart Study (N=2,648)

Probe ID	Chr	Location	Gene Symbol	P
cg05951221	2	233284402		3.2×10^{-15}
cg06644428	2	233284112		1.2×10^{-14}
cg05575921	5	373378	*AHRR*	6.5×10^{-14}
cg21566642	2	233284661		8.6×10^{-10}
cg03636183	19	17000585	*F2RL3*	5.7×10^{-7}
cg06126421	6	30720080		1.3×10^{-6}
cg01940273	2	233284934		1.9×10^{-6}
cg23771366	11	86510998	*PRSS23*	3.1×10^{-6}
cg17272563	6	32116548	*PRRT1*	4.4×10^{-6}
cg23916896	5	368804	*AHRR*	1.3×10^{-5}
cg11660018	11	86510915	*PRSS23*	1.3×10^{-5}
cg08118908	16	15787920	*NDE1*	3.0×10^{-5}
cg13937905	12	53612551	*RARG*	1.5×10^{-4}
cg24172324	2	232258363		1.7×10^{-4}
cg10780313	6	33501379		2.0×10^{-4}
cg14027333	6	32116317	*PRRT1*	2.1×10^{-4}
cg11245297	19	8117898	*CCL25*	2.1×10^{-4}
cg01692968	9	108005349		3.1×10^{-4}
cg00706683	2	233251030	*ECEL1P2*	3.4×10^{-4}
cg25317941	2	233351153	*ECEL1*	4.0×10^{-4}
cg25189904	1	68299493	*GNG12*	4.0×10^{-4}
cg14179389	1	92947961	*GFI1*	4.7×10^{-4}
cg13641317	3	127255552		4.9×10^{-4}
cg19847577	15	29213748	*APBA2*	5.1×10^{-4}
cg14239618	7	110281356		5.8×10^{-4}
cg25955180	6	32116538	*PRRT1*	6.3×10^{-4}

cg00774149	3	52255721	*TLR9*	6.4×10^{-4}
cg21351392	6	161607487	*AGPAT4*	7.1×10^{-4}
cg11902777	5	368843	*AHRR*	7.6×10^{-4}
cg07251887	17	73641809	*LOC100130933; RECQL5*	7.7×10^{-4}
cg19382157	7	2124566	*MAD1L1*	8.9×10^{-4}
cg19925780	1	101509557		1.1×10^{-3}
cg03679544	6	155537972	*TIAM2*	1.1×10^{-3}
cg08559712	20	16030674	*MACROD2*	1.3×10^{-3}
cg09837977	7	110731201	*LRRN3; IMMP2L*	1.3×10^{-3}
cg00931843	6	155442993	*TIAM2*	1.4×10^{-3}

1 *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 (GWAS) phenotypes that are regarded as causally related to cigarette smoking[2]

GWAS Phenotype	**Enrichment p-value**
Current vs. never smoking	
Coronary heart disease (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 (HDL)	0.0009
Type 2 diabetes	0.0106
Rheumatoid arthritis (RA)	2.9×10^{-5}
Bone mineral density (BMD) and osteoporosis	0.0467
All pulmonary traits	2.8×10^{-6}
All chronic obstructive pulmonary disease (COPD)	0.0295
Moderate to severe COPD	0.0156
Pulmonary function	0.0044
Crohn's 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 (RA)	6.3×10^{-5}
All pulmonary traits	0.0217
Inflammation bowel disease	5.2×10^{-6}
Crohn's Disease	0.0064
All cancer	7.8×10^{-6}

1

2

Figure

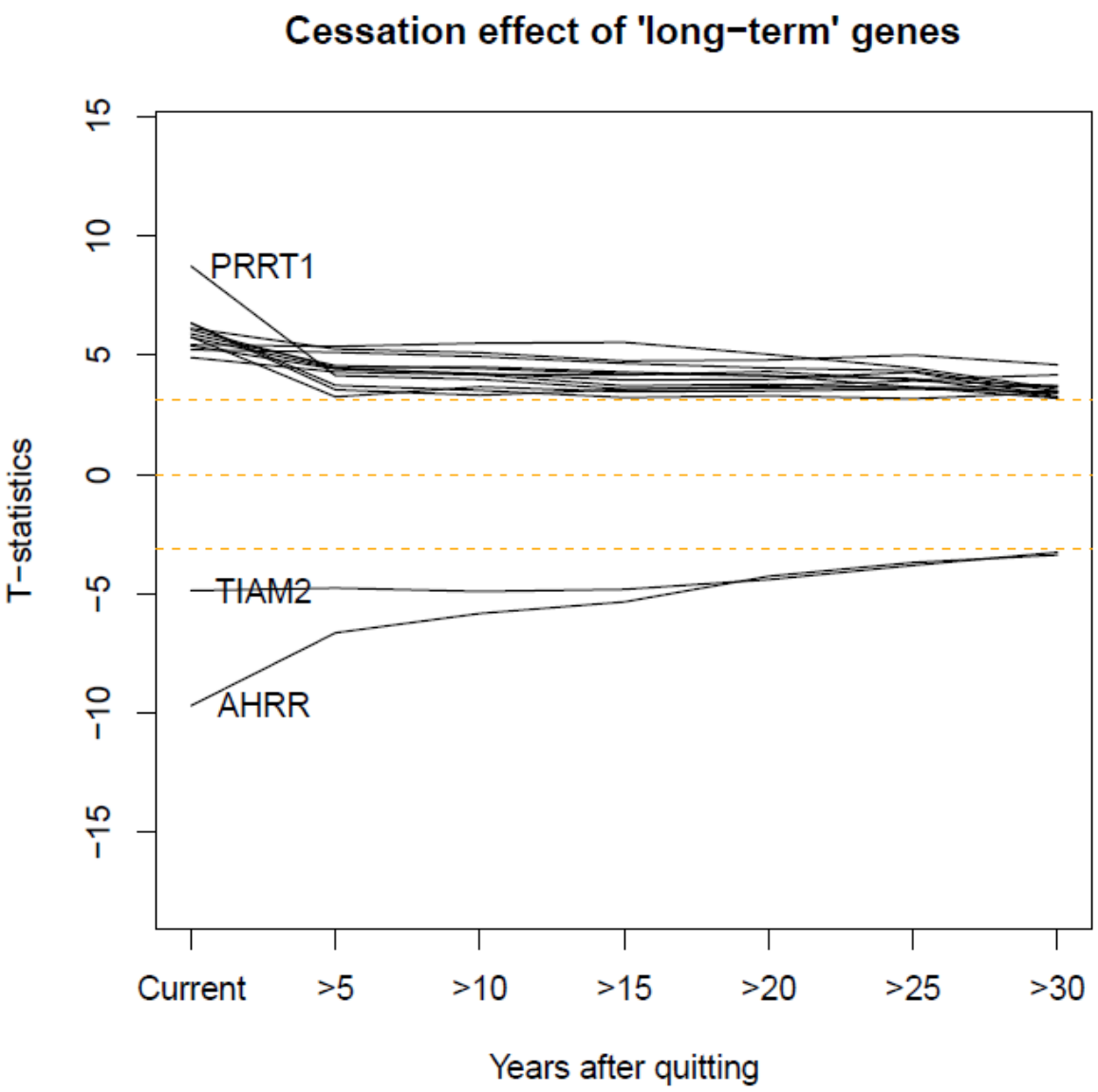

Figure 1. Trajectories of CpG sites that did not return to never-smoker levels within 30 years after cessation.