Research Article

*For reprint orders, please contact: reprints@futuremedicine.com*

Epigenomics

# Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study

**Aim**: Epigenetic changes may occur in response to environmental stressors, and an altered epigenome pattern may represent a stable signature of environmental exposure. **Materials & methods:** Here, we examined the potential of DNA methylation changes in 910 prediagnostic peripheral blood samples as a marker of exposure to tobacco smoke in a large multinational cohort. **Results**: We identified 748 CpG sites that were differentially methylated between smokers and nonsmokers, among which we identified novel regionally clustered CpGs associated with active smoking. Importantly, we found a marked reversibility of methylation changes after smoking cessation, although specific genes remained differentially methylated up to 22 years after cessation. **Conclusion**: Our study has comprehensively cataloged the smoking-associated DNA methylation alterations and showed that these alterations are reversible after smoking cessation.

First draft submitted: 4 January 2016; Accepted for publication: 26 January 2016; Published online: 11 February 2016

**Keywords:** DNA methylome • epigenetic signature • prospective cohort • tobacco smoking

Srikant Ambatipudi[1], Cyrille Cuenin[1], Hector Hernandez-Vargas[1], Akram Ghantous[1], Florence Le Calvez-Kelm[1], Rudolf Kaaks[2], Myrto Barrdahl[2], Heiner Boeing[3], Krasimira Aleksandrova[3], Antonia Trichopoulou[4,5], Pagona Lagiou[4,5], Androniki Naska[4,5], Domenico Palli[6], Vittorio Krogh[7], Silvia Polidoro[8], Rosario Tumino[9], Salvatore Panico[10], Bas Bueno-de-Mesquita[11,12,13,14], Petra HM Peeters[15,16], José Ramón Quirós[17], Carmen Navarro[18,19,20], Eva Ardanaz[19,21,22], Miren Dorronsoro[23], Tim Key[24], Paolo Vineis[25], Neil Murphy[25], Elio Riboli[25], Isabelle Romieu[1] & Zdenko Herceg*[,1]

*Author for correspondence:
Tel.: +33 4 72 73 83 98
Fax: +33 4 72 73 83 22
hercegz@iarc.fr

Author affiliations can be found at the end of this article.

Smoking is a leading cause of morbidity and mortality worldwide. Among middle-aged people, tobacco use is estimated to be the most important preventable risk factor for premature death in men and the second most important risk factor in women (after high blood pressure) [1]. Tobacco-related mortality is set to increase to almost 1 billion deaths during the 21st century, most of them in low-income countries [2]. This calls not only for addressing primary prevention by reducing tobacco consumption but also for active secondary prevention by actively following up current and former smokers. Smoking contributes toward disease development and progression through genetic and epigenetic mechanisms [3]. Epigenetic mechanisms broadly include DNA methylation, histone modification and noncoding RNAs, [4], and previous studies implicated exposure to tobacco smoking in deregulation of these mechanisms.

Due to their plastic nature, DNA methylation patterns are suspected to be under the influence of aging, environmental, lifestyle and demographic factors [4–7]. The malleable nature of DNA methylation has been increasingly exploited in biomarker discovery and mechanistic studies aimed at understanding constitutive and environmentally induced cancer risk [4,6,8–9]. A number of novel smoking-associated blood DNA methylation biomarkers have been identified using the Illumina 27K array [10,11] and more recently using the Illumina Infinium HumanMethylation 450K BeadChip array in cord blood and adult blood [12–16]. Among these markers, seven CpGs (*F2RL3* [cg03636183], *AHRR* [cg21161138 and cg05575921], 2q37.1 [cg21566642, cg01940273 and cg05951221] and 6p21.33 [cg06126421]) were common among most differentially methylated sites.

Previously published studies on smoking-associated methylation changes have been

Future Medicine part of fsg

limited by sample size and/or DNA methylome coverage. Moreover, all the studies focused on individual site-wise DNA methylation analysis associated with smoking, which provides useful information about the impact of smoking on specific CpG sites (differentially methylated probes [DMPs]). However, CpGs can be highly correlated by function and genomic density, unlike SNPS in genome-wide association studies (GWAS). Therefore, dimension reduction approaches that identify highly correlated CpG clusters have become crucial and have been applied in this study, leading to the identification of differentially methylated regional clusters (DMRs) [16]. Another way to measure the impact of smoking on the DNA methylome is to measure the methylome variability (most variable probe [MVP]), which has been shown to contribute to the observed differences in response to environmental agents and drugs [17].

None of the previous studies have documented dimension reduction or MVP analyses when studying alterations in DNA methylation due to smoking exposure. The present study was conducted in a case–control study on breast cancer nested within the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort [18]. Furthermore, this longitudinal analysis of former smokers has enabled us to identify CpG sites that remain differentially methylated for more than a decade after smoking cessation. Our results not only validated the previously reported findings in a large number of baseline blood samples from cohort participants but also identified novel markers as well as CpG clusters that respond to smoking exposure in a highly correlative manner.

## Materials & methods

### Study population

This study is based on the cohort of the European Prospective Investigation into Cancer and Nutrition (EPIC), a large prospective study conducted in 23 centers across ten European countries (Denmark, France, Germany, Greece, Italy, Norway, Spain, Sweden, the Netherlands and UK), aiming to investigate the relationship between diet, lifestyle, metabolism and cancer risk [19]. In brief, the EPIC cohort includes about 315,000 women and 200,000 men. At baseline recruitment, all study participants provided extensive questionnaire information about nutrition and other lifestyle factors. All study participants also provided a blood sample, which was processed, divided into aliquots of plasma, serum and buffy coat, and frozen at -196°C (in liquid nitrogen) for later use in specific research projects. All participants gave written informed consent. The study was approved by the local ethics committees and the Institutional Review Board of the International Agency for Research on Cancer (IARC, Lyon, France).

### Selection of participants

For the purpose of this study, we included 960 women from the EPIC-nested breast cancer case–control study. The matching criteria for the selection of the controls were: center, date of blood collection, age at blood collection, menopausal status, current use of the contraceptive pill and current use of hormone replacement therapy. The set of 960 samples included 20 technical replicates to compare inter- and intra-array batch variation. The technical replicates and 30 samples that did not pass the quality control filters were excluded from the final analyses, leaving 910 participants (460 controls and 450 cases). Smoking history included self-reported current smoking status, number of cigarettes smoked per day and time since smoking cessation (for former smokers). Current smokers, never-smokers and former smokers were defined as subjects self-reporting current smoking, lifelong nonsmokers and ex-smokers, respectively. Total lifetime dosage of tobacco smoke was measured in pack-years (PY) as per the following formula (assuming 20 cigarettes per pack):

$$PY = (\text{dose rate [cigarettes per day]}/20) \times \text{number of years smoked}$$

### Bisulfite conversion & genome-wide DNA methylation analysis

DNA was isolated from the white blood cells as per the standard DNA extraction procedure (Autopure LS, Qiagen). DNA methylome profiling was carried out using the Illumina Infinium HumanMethylation450K (HM450K) BeadChip assay, which interrogates more than 480,000 methylation sites [20], essentially as described previously [21,22].

### Bioinformatics analyses

Data preprocessing and analyses were performed using R (version 3.2.2)/Bioconductor packages. The DNA methylation level is described as the β-value, which is a continuous variable ranging between 0 (no methylation) and 1 (full methylation). To avoid spurious associations, we excluded the cross-reactive probes and probes overlapping with a known SNP with minor allele frequency of ≥5% in the overall population (European ancestry; [23]), leaving 423,066 probes. In any given sample, a probe with a detection p-value (a measure of an individual probe's performance) ≥0.05 was assigned 'missing' status. If a probe was missing in greater than 5% of samples, it was excluded from all samples. Thus, we excluded 1625 probes on this basis. Finally, 421,441 probes were available for the analyses,

which were corrected for probe color bias and intersample quantile normalization, followed by β-mixture quantile normalization (BMIQ) to align type I and type II probe distributions [24]. We used the array annotations from the Bioconductor package FDb.InfiniumMethylation.hg19 (version 2.2.0) to assign probes to their nearest corresponding genes.

### Estimates of white blood cell counts

Quantile-normalized data were used to infer blood cell proportions ($CD8^+$ T cells, $CD4^+$ T cells, natural killer [NK] cells, B cells, monocytes and granulocytes) using Houseman's estimation method [25,26], which is based on DNA methylation signatures from the purified leukocyte samples.

### Statistical analyses

#### Identification of DMPs

Logarithmically transformed methylation values [27] were batch-corrected using surrogate variable analysis (SVA) [28] and interrogated for association with smoking status (never vs current, former vs current and never vs former) by modeling the study variables and covariates (i.e., case–control status, age at blood collection, menopausal status, current use of the contraceptive pill and current use of hormone replacement therapy) together with latent surrogate variables by multivariable linear regression using the R/Bioconductor package limma [29]. Smoking status associated loci were selected based on a threshold for the adjusted p-value (false discovery rate [FDR]) of 0.05 [30].

#### Identification of DMRs

The bump hunting method [25] was used to identify predefined regional clusters of neighboring CpGs that are differentially methylated with smoking status using the recommended proximity-based criteria [31]. We fitted the linear model of methylation levels at each probe as a function of smoking status, adjusting for study variables and surrogate variables estimated by SVA. The family-wise error rate (FWER) for each DMR was estimated based on 1000 bootstraps under the null hypothesis.

#### Identification of MVPs

Sample-to-sample variability of DNA methylation at specific genomic locations captures the intersample variability, which distinguishes it from more commonly used mean methylation comparisons such as the DMP analysis. This intersample variability has previously been shown to be important for the identification of cancer-associated genes [32]. Differential variability between never-smokers and current smokers was identified using the DiffVar function implemented in the Bioconductor package missMethyl [33]. This function uses an empirical Bayes model framework to detect variability. The default criteria implemented in the package were used, except for the fact that latent variables identified through SVA were adjusted for in the model along with the covariates used for identifying differentially variable probes.

### Pathway analysis

For the pathway analysis, we used 748 CpG sites found to be differentially methylated between never-smokers and current smokers. Pathway analysis methods developed for gene expression studies give false results when used for analysis of 450k array data where the number of sites in each gene range from single digits to greater than 1000 [34]. In order to circumvent this problem, we first corrected the significant CpG sites for the number of probes present on 450k array followed by FDR correction for the number of gene symbols in Illumina 450k arrays. By doing this, we corrected for the identification of genes which have more probes on the 450k array and may potentially bias the pathway analysis. Thus, we had 538 genes for the pathway analysis, which was performed using Enrichr [35].

## Results

### Characteristics of the study population

To determine the differences in the DNA methylome between subjects who smoke and nonsmokers, we performed genome-wide DNA methylation profiling of baseline blood samples from cases and controls of the cohort-nested breast cancer case–control study. At baseline recruitment (time of blood collection), all subjects were aged 26.1–72.8 years, with an average age of 52.4 years. Based on self-reported smoking status, the samples were divided into never-smokers, former smokers and current smokers. The general characteristics of the participants are shown in Table 1. The proportion of premenopausal women was greater in current smokers compared with the other two categories. Based on our DNA methylation based estimates of the leukocyte subpopulations, current smokers had a slightly lower proportion of natural killer cells compared with the other two categories ($p < 0.001$) (Supplementary Figure 1).

### Differentially methylated CpG probes in response to smoking: DMP analysis

Differentially methylated probe (DMP) analysis identified a total of 748 CpG sites that were differentially methylated between current smokers and never-smokers (FDR ≤ 0.05). This included 450 hypomethylated and 298 hypermethylated CpG sites

Table 1. Clinical characteristics of participants who passed the quality control filters of the Illumina 450K array.

| Characteristic | All subjects (n = 910) | Never-smokers (n = 528) | Former smokers (n = 189) | Current smokers (n = 193) | p-value |
|---|---|---|---|---|---|
| Cases:controls | 450:460 | 263:265 | 96:93 | 91:102 | p = 0.75† |
| Age (years) | 52.4 ± 8.8 | 54.0 ± 8.5 | 51.6 ± 8.9 | 48.6 ± 8.8 | p < 0.001‡ |
| BMI (kg/m²) | 25.8 ± 4.4 | 26.2 ± 4.3 | 25.5 ± 4.7 | 24.8 ± 4.2 | p < 0.001‡ |
| Age at first menstrual cycle (years) | 13.1 ± 1.6 | 13.1 ± 1.6 | 13.1 ± 1.8 | 13.0 ± 1.4 | p = 0.51‡ |
| **Menopausal status** | | | | | |
| – Premenopausal | 370 (41%) | 179 (34%) | 85 (45%) | 106 (55%) | p < 0.001† |
| – Postmenopausal | 540 (59%) | 349 (66%) | 104 (55%) | 87 (45%) | |
| **Smoking related** | | | | | |
| – Pack-years | – | – | – | 19.3 ± 14.8 (190) | – |
| – PY missing | | | | 3 | |
| – Time since cessation (years) | – | – | 15.0 ± 10.2 (183) | – | – |
| – Time since cessation missing | | | 6 | | |
| Alcohol consumption (g/day) | 8.9 ± 12.3 | 7.0 ± 10.4 | 10.0 ± 12.2 | 12.9 ± 15.9 | p < 0.001‡ |

Continuous variables are mean ± SD.
†p-value from Pearson's $\chi^2$ test.
‡p-value from the Kruskal–Wallis test.

(Supplementary Table 1). The chromosome-wide distribution of significant CpG sites and their relevance is shown in Figure 1. Chromosome 5p and chromosome 2q showed a number of significant probes associated with smoking exposure. Including BMI, alcohol consumption and leukocyte subtypes in the model did not affect the main findings, and most of the CpG sites remained unchanged (Supplementary Figure 2). We therefore used the model without BMI, alcohol consumption and leukocyte subtypes for all future analyses.

The top five hypermethylated and the top five hypomethylated CpGs in current smokers versus never-smokers with ≥5% difference in the methylation between these two categories are shown in Table 2. Some of the differentially methylated sites (*AHRR, ALPPL2, F2RL3, GFI, GNG12, MYO1G, ZNF385D* and *CACNA1D*) were identical to those identified in previous studies [12–16], demonstrating the reliability and robustness of our analysis pipeline. In addition, we identified 12 novel CpG sites [cg22472290 (*ZNF577*), cg16071219 (*LPAR6*), cg00008629 (*PTBP3*), cg04224247 (*WWC3*), cg24874254 (*PRDM1*), cg24134897 (*TSPAN4*), cg02610360 (*TMEM136*), cg19925780 (*DPH5*), cg05156137 (*RCAN1*), cg04387347 (*MIR5189*), cg01899620 (*MCF2L*) and cg11028075 (*SORBS1*)] associated with smoking (Table 3). All of these CpG sites had ≥3% methylation difference between current smokers and never-smokers. Two of those sites, cg24134897 (*TSPAN4*) and cg05156137 (*RCAN1*), were also differentially methylated in current smokers relative to former smokers (≥2% difference). More than 90% of the DMPs in current smokers versus never-smokers were replicated in the comparison of current smokers versus former smokers. All the DMPs that were different between current and former smokers are shown in Supplementary Table 2. A similar DMP analysis comparing never and former smokers revealed four CpG sites shown in Supplementary Table 4.

An enrichment analysis for the functional distribution of significant smoking-associated CpG sites (n = 748) revealed an enrichment for open seas (regions containing isolated CpG sites in the genome that do not have a specific designation) and lower representation of the CpG islands, while the adjacent regions of CpG islands, such as the shores (0–2 kb from the promoter CpG islands) and shelves (2–4 kb from the promoter CpG islands) had an expected representation (data not shown) (Figure 2A). We found that there was a decrease in the promoter-related sites in the DMPs compared with the Illumina 450K array (Figure 2B). We observed a significant enrichment in the regulation of cell activation (GO: 0050865), while pathway analysis using the KEGG pathway revealed chronic myeloid leukemia, melanoma, hematopoietic cell lineage, regulation of actin cytoskeleton and glioma (Supplementary Figure 3).

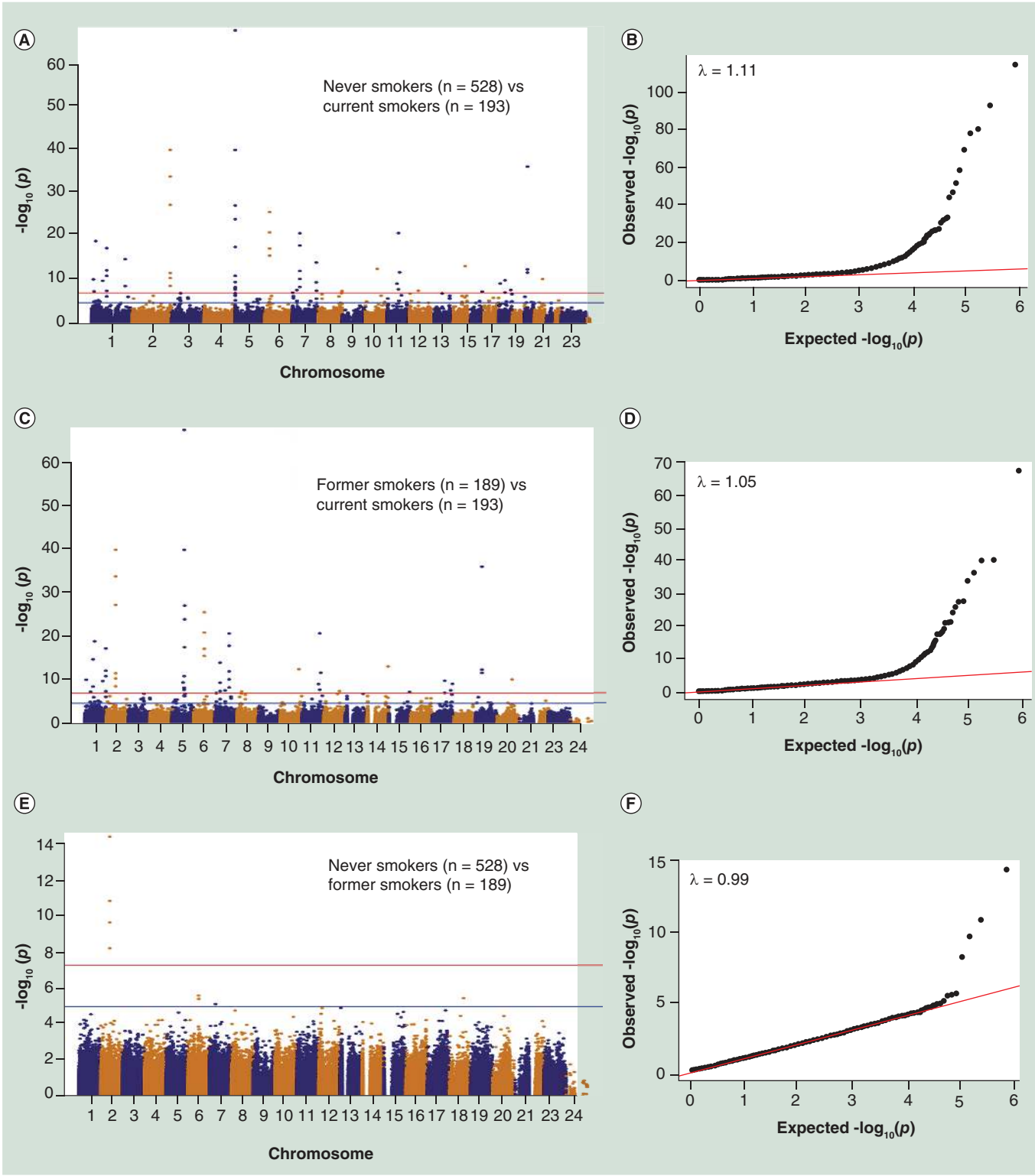


**Figure 1. Methylome-wide effect of tobacco smoking.** **(A, C, E)** Manhattan plots showing the distribution of p-values of CpG sites associated with smoking exposure across chromosomes: **(A)** never-smokers versus current smokers, **(C)** former smokers versus current smokers, **(E)** never-smokers versus former smokers. The red horizontal line indicates the genome-wide significance threshold of $p = 5 \times 10^{-8}$, and the blue line is a threshold for suggestive association ($p = 10^{-5}$). **(B, D, F)** Quantile-quantile (QQ) plots showing observed versus expected $-\log_{10}$(p-values) for the association between DNA methylation and smoking exposure (as a categorical variable): **(B)** never-smokers versus current smokers, **(D)** former smokers versus current smokers, **(F)** never-smokers versus former smokers.

Table 2. Top five differentially methylated probes associated with smoking (never smokers vs current smokers).

| CpG site | Gene | Region† | p-value‡ | Mean methylation (β values) in never smokers (SD) | Mean methylation (β values) in current smokers (SD) | Mean methylation difference (never-current)% |
|---|---|---|---|---|---|---|
| **Hypermethylated in never smokers** | | | | | | |
| cg05575921 | *AHRR* | GB_SHO | $3.40 \times 10^{-109}$ | 0.85 (0.05) | 0.67 (0.12) | 17.6 |
| cg23576855 | *AHRR* | GB_SHO | $1.06 \times 10^{-13}$ | 0.69 (0.19) | 0.52 (0.17) | 16.5 |
| cg21566642 | *ALPPL2* | IG_CGI | $5.47 \times 10^{-88}$ | 0.52 (0.06) | 0.38 (0.08) | 14.0 |
| cg03636183 | *F2RL3* | GB_SHO | $9.12 \times 10^{-65}$ | 0.70 (0.05) | 0.58 (0.10) | 12.0 |
| cg06126421 | *IER3* | DP_NC | $4.14 \times 10^{-54}$ | 0.76 (0.08) | 0.65 (0.10) | 11.5 |
| **Hypomethylated in never smokers** | | | | | | |
| cg03274391 | *ZNF385D* | IG_SHO | $3.26 \times 10^{-6}$ | 0.57 (0.14) | 0.68 (0.14) | -10.8 |
| cg23480021 | *ZNF385D* | IG_SHO | $1.54 \times 10^{-7}$ | 0.66 (0.14) | 0.76 (0.13) | -10.2 |
| cg12803068 | *MYO1G* | GB_SHO | $6.51 \times 10^{-22}$ | 0.75 (0.12) | 0.84 (0.11) | -9.2 |
| cg15693572 | *ZNF385D* | IG_SHO | $1.65 \times 10^{-6}$ | 0.56 (0.12) | 0.65 (0.11) | -9.1 |
| cg23126342 | *PCDH9* | GB_SHE | $2.75 \times 10^{-7}$ | 0.57 (0.09) | 0.64 (0.10) | -6.9 |

†Region denotes the relation of a CpG site to genes and CpG islands.
‡FDR corrected p-value.
CGI: CpG islands; DP: Distal promoter; GB:Gene body; IG: Intergenic; NC: Non-CpG islands; SD: Standard deviation; SHE: Shelves; SHO: Shores.

## Differentially methylated CpG regional clusters in response to smoking: DMR analysis

Next, we investigated potential regional clustering of differential methylation to check whether neighboring differentially methylated CpG sites are correlated with each other. For this, we ran the same model used for identifying DMPs adjusted for latent variables by SVA, but accounting for correlations among proximal CpGs. This analysis revealed regionally clustered CpGs that were differentially methylated with smoking status (Table 4). We identified eight DMRs associated with smoking status at a genome-wide family-wise empirical p-value [family-wise error rate (FWER)] <0.1, including six genes that were differentially methylated in the DMP analysis (*ALPPL2*, *GFI1*, *MYO1G*, *AHRR*, *ZNF385D* and *IER3*). The majority (6/8) of the DMRs showed lower methylation in current smokers compared with never-smokers, while two (*MYO1G* and *ZNF385D*) showed higher methylation in current smokers (Table 4 & Figure 3A & B).

The identification of *ALPPL2*, *GFI1*, *MYO1G*, *AHRR*, *ZNF385D* and *IER3* in both the DMP and DMR analyses substantiates their role in smoking-related exposure. Our DMR analysis revealed two novel candidates: *VTRNA2-1*, an imprinted small noncoding RNA, and *ZFAND2A*, also known as *AIRAP*. The effect of smoking-related methylation spans the entire *VTRNA2-1* DMR, and the current smokers (n = 193) exhibited hypomethylation at the *VTRNA2-1* DMR compared with the never-smokers (n = 528) (Figure 3C).

## Most variable CpG methylation in response to smoking: MVP analysis

Differentially variable CpG sites have been reported to show increased variance in normal cells from people predisposed to neoplasia [36]. Through this study, we tried to better understand the smoking exposure associated variability. We compared the differential variability between never-smokers and current smokers by DiffVar using a threshold of FDR ≤ 0.05. Our analysis revealed 14 differentially variable CpG sites associated with smoking exposure (Table 5). The 12 differentially variable sites with higher variability among current smokers compared with never-smokers are shown in Figure 4. Although we removed all potential CpG sites associated with SNPs in the initial quality control, we observed a site cg27126508 showing an SNP effect (Figure 4J). We then compared the DMP and MVP sets and found 50% (7/14) overlap, in which the top sites were common (Supplementary Figure 4).

## Effect of smoking dosage on DNA methylation

We studied the effect of smoking dosage measured in pack-years (PY) on site-specific DNA methylation. We restricted this analysis to current smokers (n = 190) as we believed that the time since smoking cessation in for-

Table 3. Novel differentially methylated probes associated with smoking (never smokers vs current smokers).

| CpG site | Gene | Region† | p-value‡ | Mean methylation (β values) in never smokers (SD) | Mean methylation (β values) in current smokers (SD) | Mean methylation difference (never-current)% |
|---|---|---|---|---|---|---|
| **Hypermethylated in never smokers** | | | | | | |
| cg22472290 | *ZNF577* | PP_SHO | 0.037257 | 0.72 (0.09) | 0.68 (0.10) | 3.8 |
| cg16071219 | *LPAR6* | GB_NC | 0.001598 | 0.50 (0.06) | 0.47 (0.06) | 3.6 |
| cg00008629 | *PTBP3* | GB_SHO | 0.006513 | 0.40 (0.09) | 0.37 (0.10) | 3.4 |
| cg04224247 | *WWC3* | NA | 0.04738 | 0.57 (0.08) | 0.54 (0.08) | 3.1 |
| cg24874254 | *PRDM1* | PP_NC | 0.009166 | 0.54 (0.08) | 0.51 (0.09) | 3.0 |
| **Hypomethylated in never smokers** | | | | | | |
| cg24134897 | *TSPAN4* | GB_SHO | $1.72 \times 10^{-5}$ | 0.75 (0.08) | 0.79 (0.07) | -3.9 |
| cg02610360 | *TMEM136* | GB_SHO | $8.18 \times 10^{-6}$ | 0.47 (0.07) | 0.51 (0.07) | -3.7 |
| cg19925780 | *DPH5* | IG_NC | $8.00 \times 10^{-5}$ | 0.78 (0.07) | 0.81 (0.07) | -3.5 |
| cg05156137 | *RCAN1* | PP_NC | 0.045051 | 0.24 (0.06) | 0.28 (0.06) | -3.4 |
| cg04387347 | *MIR5189* | GB_CGI | $2.51 \times 10^{-8}$ | 0.25 (0.06) | 0.28 (0.06) | -3.4 |
| cg01899620 | *MCF2L* | GB_SHO | 0.000526 | 0.51 (0.07) | 0.54 (0.07) | -3.3 |
| cg11028075 | *SORBS1* | PP_NC | $2.85 \times 10^{-8}$ | 0.55 (0.06) | 0.58 (0.06) | -3.2 |

†Region denotes the relation of a CpG site to genes and CpG islands.
‡FDR corrected p-value.
CGI: CpG islands; GB: Gene body; IG: Intergenic; NA: Not available; NC: Non-CpG islands; PP: Proximal promoter; SD: Standard deviation; SHO: Shores.

mer smokers would dilute the effect of smoking dosage. For the analysis, we selected all 748 CpG sites found to be differentially methylated between never-smokers and current smokers. We calculated the Pearson correlation coefficient between the PY values and the methylation values (β-values) of the CpG sites, and found that 153 CpG sites were significantly correlated with PY (Supplementary Table 5). A representative image of CpG sites with significant correlation to PY is shown in Figure 5A. CpGs hypomethylated in current smokers showed a decrease in methylation with increasing smoking dosage (*r* range = -0.39 to -0.14), whereas CpGs hypermethylated in current smokers showed an increase in methylation with increasing smoking dosage (*r* range = 0.14–0.30). We observed a smoking dosage-dependent change in the methylation of CpG sites (Figure 5B). The same trend was observed when we divided the current smoker PY into quartiles (Q1: 1.0–8.1; Q2: 8.2–16.4; Q3: 16.5–26.3 and Q4: 26.4–79.5) and compared them with never-smokers and former smokers. Figure 5C shows the effect of smoking dosage on one of the genes, *IER3* (four CpGs), in which the methylation values decrease with increasing smoking dosage.

To investigate the effect of duration of smoking before smoking cessation, we did analyses by correlating time since smoking cessation in former smokers to DNA methylation changes. We found that the genes differentially methylated in never-smokers relative to current smokers were also significantly associated with duration of smoking before cessation (Supplementary Figure 5).

## Effect of time since smoking cessation on DNA methylation

Next, we investigated the reversibility of CpG methylation changes after smoking cessation by carrying out a longitudinal analysis in former smokers (n = 183). This analysis was done only on the probes that were significantly altered by current smoking (current smokers vs former smokers, n = 163). We found that DMPs that were hypermethylated in current smokers relative to never-smokers showed decreased methylation with increasing time between smoking cessation and blood collection (Timq_smok) (negative correlation; *r* range = -0.16 to -0.27, n = 183; Figure 6A & B), sometimes returning to baseline levels (Figure 6C shows a representative gene, *IER3*). In contrast, DMPs that were hypomethylated in smokers showed increased methylation with increasing Timq_smok (positive correlation; *r* range = 0.15–0.44, n = 183; Figure 6A & B).

To better understand the association between smoking-associated DNA methylation changes and the quartiles of Timq_smok, we focused on 37 DMPs

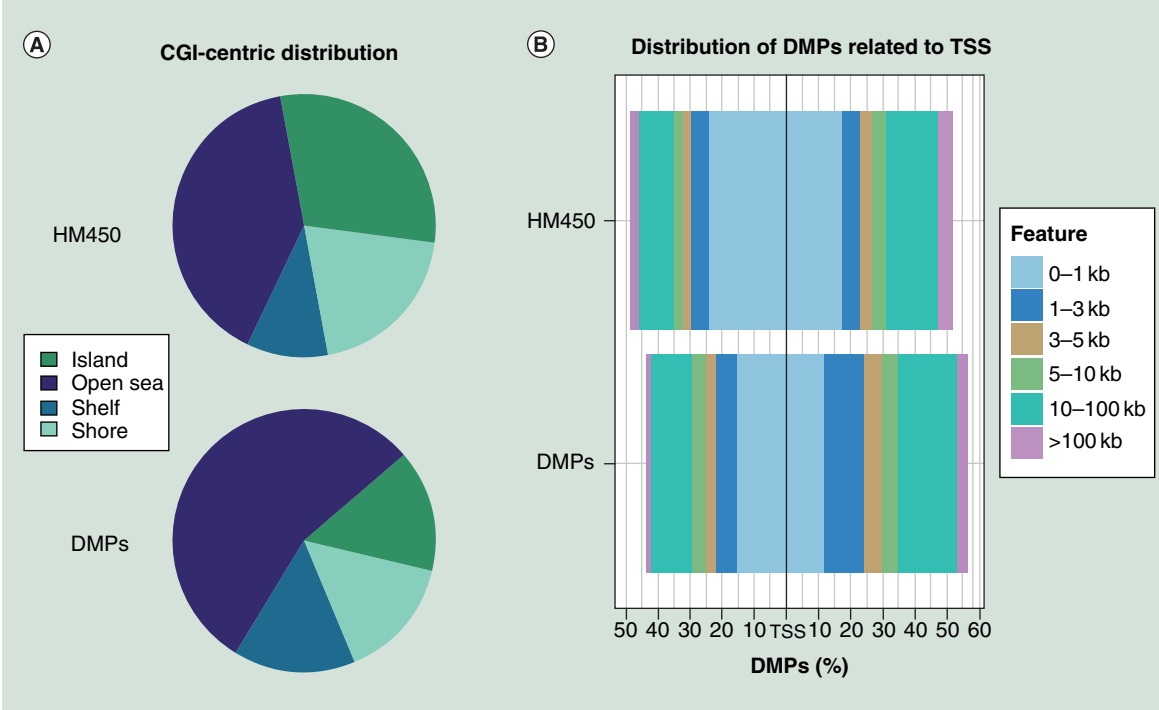


**Figure 2. Functional distribution of smoking-associated CpG sites. (A)** Distribution of smoking exposure-associated CpGs (never-smokers vs current smokers) relative to CGI (i.e., islands, shores, shelves). The distribution of all Illumina HM450K array probes is shown for comparison. **(B)** The distribution of differentially methylated CpGs based on distance from the TSS. The distribution of all Illumina HM450K array probes based on distance from the TSS is shown for comparison.
CGI: CpG islands; DMP: Differentially methylated probe; TSS: Transcription start site.

that were significantly correlated with Timq_smok (p-value < 0.05; Supplementary Table 6). We studied these sites (n = 37) in detail to discern the methylation dynamics and reversibility, and compared the methylation levels between never-smokers and former smokers in quartiles of time of cessation: Q1 (0.5–6.5 years), Q2 (6.6–14 years), Q3 (14.1–22 years) and Q4 (22.1–42.5 years). Compared with never-smokers, methylation levels of 15 CpG sites (Table 6 & Figure 6D), six CpG sites (cg00310412, cg03636183, cg05575921, cg06126421, cg11207515 and cg14817490), and four CpG sites (cg01940273, cg05951221, cg011554391 and cg21566642) were significantly different in Q1, Q2 and Q3 of former smokers, respectively. No DMPs were different between never-smokers and Q4 of former smokers, indicating that DNA methylation for those 37 DMPs almost reaches the baseline (levels in never-smokers) 22.1 years after smoking cessation.

## Discussion

We performed a genome-wide methylation study to investigate the effect of smoking exposure on DNA methylation using baseline blood samples collected in a prospective cohort. We identified 748 CpG sites that were differentially methylated between current smokers and never smokers (FDR <0.05). Most of the differentially methylated sites in our study were similar to those identified in previous studies on maternal smoking [12,16,37–40] and adult smoking [10,13–15,41–45], strengthening the view that alterations due to smoking exposure could be detected in the blood irrespective of the study population and sex. Thus, blood-based methylation markers are a robust measure of smoking exposure. However, unlike any of the previous studies, we carried out new analyses, which led to the identification of regional alterations and the variability of DNA methylation due to smoking exposure. Furthermore, 12 of the 748 differentially methylated sites (≥3% methylation difference) remain unreported to date: cg22472290 (*ZNF577*), cg16071219 (*LPAR6*), cg00008629 (*PTBP3*), cg04224247 (*WWC3*), cg24874254 (*PRDM1*), cg24134897 (*TSPAN4*), cg02610360 (*TMEM136*), cg19925780 (*DPH5*), cg05156137 (*RCAN1*), cg04387347 (*MIR5189*), cg01899620 (*MCF2L*) and cg11028075 (*SORBS1*).

Although the functional impact of differential methylation in the specific genes associated with smoking status (identified in our study and several recent studies by other groups) remains to be established, identification of AHHR and ALPPL2 makes sense biologically. The AHHR gene is part of the aryl hydro-

Table 4. The most significant smoking-associated differentially methylated regional clusters identified by the bump hunting analysis.

| Chr | DMR start | DMR end | Mean coefficient | Probes in DMR | Probes in cluster | p-value | Adj. p-value (FWER) | Gene | Distance from TSS | DMR on promoter |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 233284112 | 233285289 | -0.58998944 | 5 | 10 | 0 | 0 | *ALPPL2* | 12,560 | FALSE |
| 1 | 92946700 | 92947961 | -0.418801471 | 6 | 6 | 0 | 0 | *GFI1* | 1395 | FALSE |
| 7 | 45002287 | 45002919 | 0.540980799 | 4 | 6 | 0 | 0 | *MYO1G* | 11,907 | FALSE |
| 5 | 373299 | 373378 | -1.038399834 | 2 | 2 | 0 | 0 | *AHRR* | 50,597 | FALSE |
| 5 | 135415948 | 135416613 | -0.270404279 | 11 | 18 | $2.60 \times 10^{-6}$ | 0.01 | *VTRNA2-1* | 0 | TRUE |
| 3 | 2241124 | 22413232 | 0.406851616 | 4 | 11 | $5.20 \times 10^{-6}$ | 0.02 | *ZNF385D* | 891 | FALSE |
| 6 | 30720080 | 30720209 | -0.378740141 | 4 | 9 | $2.08 \times 10^{-5}$ | 0.08 | *IER3* | -7753 | FALSE |
| 7 | 1209495 | 1209742 | -0.393938267 | 3 | 3 | $2.34 \times 10^{-5}$ | 0.09 | *ZFAND2A* | -9640 | FALSE |

Analysis includes adjustment for case–control status, age at blood collection, menopausal status, current use of the contraceptive pill and current use of hormone replacement therapy together with latent surrogate variables.
FWER: Family-wise error rate; TSS: Transcription start site.

carbon pathway that metabolizes cigarette smoke components [46], and its hypomethylation was found to be associated with lung cancer risk [47]. The ALPPL2 gene is responsible for dephosphorylation of various molecules (including proteins, nucleotides or alkaloids) and ALPPL2 enzyme levels were found increased up to tenfold in cigarette smokers and patients with different cancers [48,49].

An interesting finding from our study was the hypermethylation of cg00008629 (*PTBP3*) in current smokers compared with never-smokers. *PTBP3* (also known as regulator of differentiation 1, or ROD1) encodes an RNA-binding protein that plays a role in the regulation of cell proliferation and differentiation and is presumably expressed in cells of the hematopoietic system [50]. The CpG site (cg00008629) is located on the gene body, which is generally associated with increased gene expression [51] and, considering this, we presume that methylation of *PTBP3* would lead to increased transcriptional activity in never-smokers. *PTBP3* has been reported to bind and post-transcriptionally regulate approximately 13,000 genes, including those related to the differentiation and proliferation of cells [52]. We believe that this homeostatic control of post-transcriptional regulation may be lost due to smoking exposure, although the mechanisms and underlying pathways remain to be elucidated.

Regional alterations in *VTRNA2-1* (nc886) and *ZFAND2A* related to active smoking exposure are novel. *VTRNA2-1* is an imprinted small noncoding RNA that is a putative tumor suppressor and a modulator of innate immunity. A previous report by Treppendahl *et al.* indicated that *VTRNA2-1* may be polymorphically imprinted [53], although our recent findings suggest that *VTRNA2-1* is not regulated by cis genetic variation but is affected by the maternal environment around the time of conception, occurs systemically, and is stable over many years [54]. Our results point to a regional loss of imprinting of *VTRNA2-1* in current smokers compared with never-smokers, showing that active smoking exposure could lead to the regional alteration of noncoding RNAs.

*ZFAND2A* behaves as a canonical heat shock gene, whose expression is strictly controlled by HSF1 in a temperature-dependent fashion [55]. It is an HSF1 target and is transcriptionally regulated by temperature [55]. To date, there are no reports linking methylation of *ZFAND2A* and active smoking exposure. More studies are needed to elucidate the functional relevance of the regional alterations mediated by smoking.

The fact that seven of the top DMPs (cg05575921, cg18146737, cg03636183, cg21566642, cg18316974, cg05951221 and cg14817490) were found within genes that show variable methylation in response to

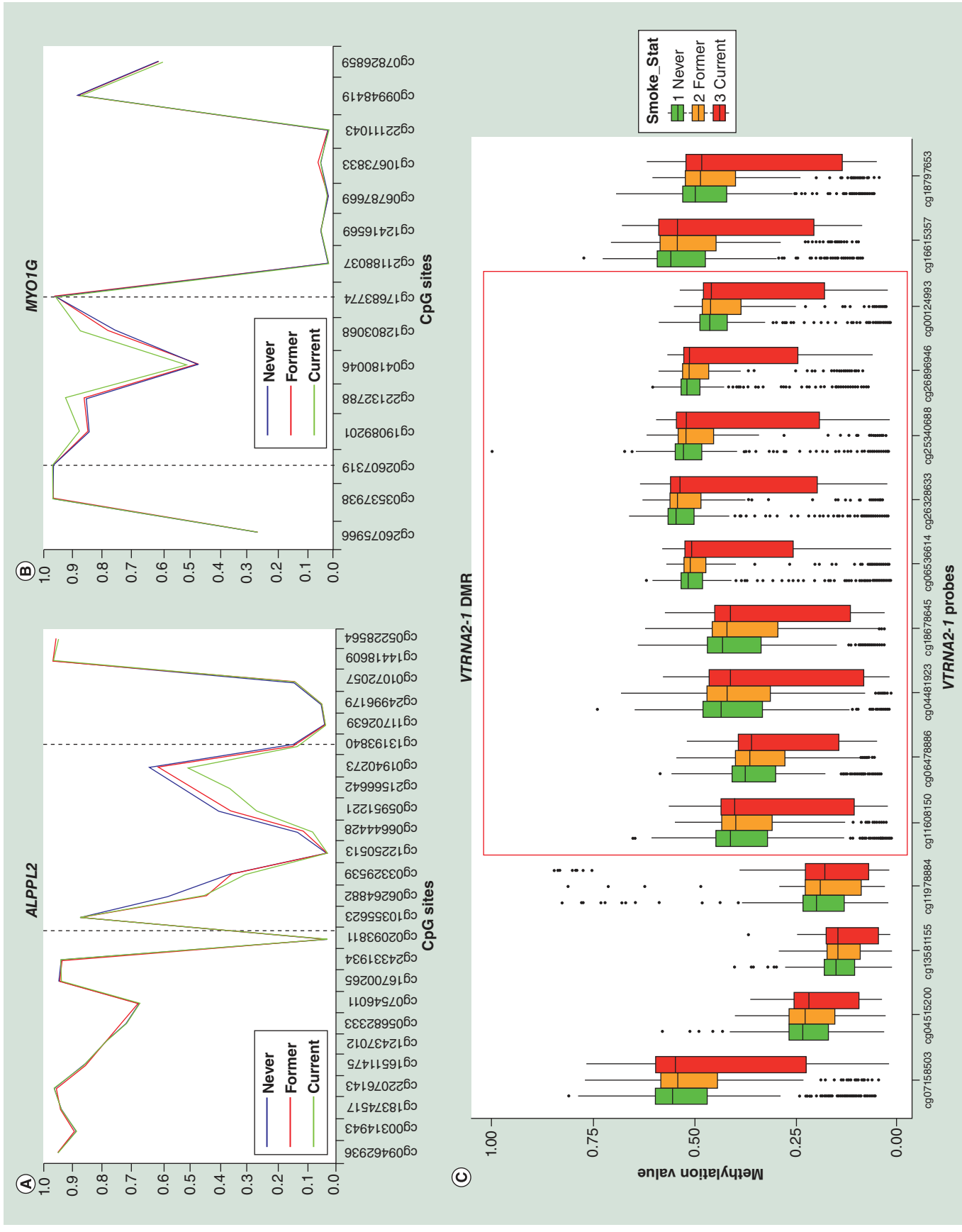
A
ALPPL2
Never
Former
Current
CpG sites
B
MYO1G
Never
Former
Current
CpG sites
C
VTRNA2-1 DMR
Smoke_Stat
1 Never
2 Former
3 Current
Methylation value
VTRNA2-1 probes

**Figure 3. Methylome-wide regional alterations (see facing page).** **(A)** Line plot showing *ALPPL2* hypomethylated in current smokers compared with never and former smokers. **(B)** Line plot showing *MYO1G* hypermethylated in current smokers compared with never and former smokers. X-axis shows the CpG sites represented by the gene and Y-axis shows the methylation value. **(C)** Regional alteration of *VTRNA2-1* in response to smoking exposure. Fifteen CpGs mapping to the *VTRNA2-1* locus are shown. The box highlights nine CpGs corresponding to the imprinted DMR. X-axis shows the CpG sites represented by *VTRNA2-1* and Y-axis shows the methylation value.

smoking exposure is consistent with the notion that an interplay exists between different epigenetic regulation mechanisms, which may exhibit different modes of alteration in response to the same exposure (smoking), and suggests that the combination of distinct differentially and variably methylated loci may have an important role in smoking-related exposures.

Seven of the most variable CpG sites have not been reported previously (cg17524265, cg27326062, cg27126508, cg12868738, cg04654261, cg00559054 and cg19160130). For instance, it is worth noting that one of the CpGs (cg17524265) on the gene *NAPRT* (nicotinic acid phosphoribosyltransferase) is more variable in never-smokers compared with current smokers. This variability may have some link to nicotinic acid metabolism in smokers, although further mechanistic and validation studies are needed.

With respect to the effect of smoking dosage and smoking duration as well as of the time since smoking cessation, we found that CpGs that are hypermethylated in current smokers showed decreased methylation with longer time since smoking cessation and increased methylation with increasing smoking dosage (PY). In contrast, CpGs that are hypomethylated in current smokers showed increased methylation with longer time since smoking cessation and decreased methylation with increasing smoking dosage. Our findings point to a strong influence of smoking dosage and time since smoking cessation in line with a recent study [14]. Interestingly, for four CpG sites [cg01940273 (*ALPPL2*), cg05951221 (*ALPPL2*), cg11554391 (*AHRR*) and cg21566642 (*ALPPL2*)], methylation levels did not decrease to the baseline levels of never-smokers even 14.1–22 years after smoking cessation, and these CpGs may serve as markers for follow-up of former smokers for secondary cancer prevention. Previous studies also identified elevated serum levels and differential methylation of *ALPPL2* in response to smoking [13–14,49], although no studies have looked at the response of *ALPPL2* to smoking cessation. All the CpG sites related to the *ALPPL2* gene are located in an intergenic CpG island, hinting at transcriptional initiation activity, as previously reported [56]. Based on our results and previously published literature, the methylation levels of the *ALPPL2* and *AHRR* genes could be used as markers of time since smoking cessation and of lifetime exposure to tobacco smoke.

Our study has many strengths, including a large sample size (n = 910), a comprehensive characterization of smoking-associated DNA methylation changes at site-specific, regional and variable positions, as well as an assessment of methylation markers for smoking dosage, smoking duration and reversibility after smoking cessation. A possible limitation of this study is the self-reported smoking information from the participants, which is sometimes under-reported [57]. Another possible limitation of our study is that we combined prediagnostic blood samples from prospective breast cancer cases and controls for the analysis. We addressed this issue by carrying out control-only and cases-only analyses to assess the effect of smoking on DNA methylation, revealing results similar to those obtained by combining both categories (Supplementary Figure 6).

It is noteworthy that we observed relatively small but significant differences in methylation of specific CpG sites and gene loci associated with smoking status. This is consistent with a number of recent studies that interrogated DNA methylation profiles in normal (nontumor) tissues (such as peripheral blood). Considering a robust platform applied and large number of samples analyzed, our study was sufficiently powered for detecting small differences in DNA methylation

**Table 5. Top differentially variable probes associated with smoking (never smokers vs current smokers).**

| CpG site | Gene | DiffLevene | p-value† |
|---|---|---|---|
| cg05575921 | *AHRR* | 0.408544 | $9.97 \times 10^{-27}$ |
| cg18146737 | *GFI1* | 0.529640 | $5.38 \times 10^{-10}$ |
| cg03636183 | *F2RL3* | 0.191574 | $1.73\times 10^{-8}$ |
| cg21566642 | *ALPPL2* | 0.152698 | $1.12 \times 10^{-6}$ |
| cg18316974 | *GFI1* | 0.282920 | 0.000268 |
| cg05951221 | *ALPPL2* | 0.13123 | 0.000463 |
| cg14817490 | *AHRR* | 0.150597 | 0.001486 |
| cg17524265 | *NAPRT* | -0.18824 | 0.005191 |
| cg27326062 | *LINC01205* | 1.265635 | 0.006142 |
| cg27126508 | *COL23A1* | 0.639640 | 0.007965 |
| cg12868738 | *ZNF212* | 0.467374 | 0.018975 |
| cg04654261 | *STK19* | 0.171126 | 0.034163 |
| cg00559054 | *SSR3* | 0.196784 | 0.043037 |
| cg19160130 | *RAB30* | -0.168180 | 0.043037 |

†FDR-corrected p-value.

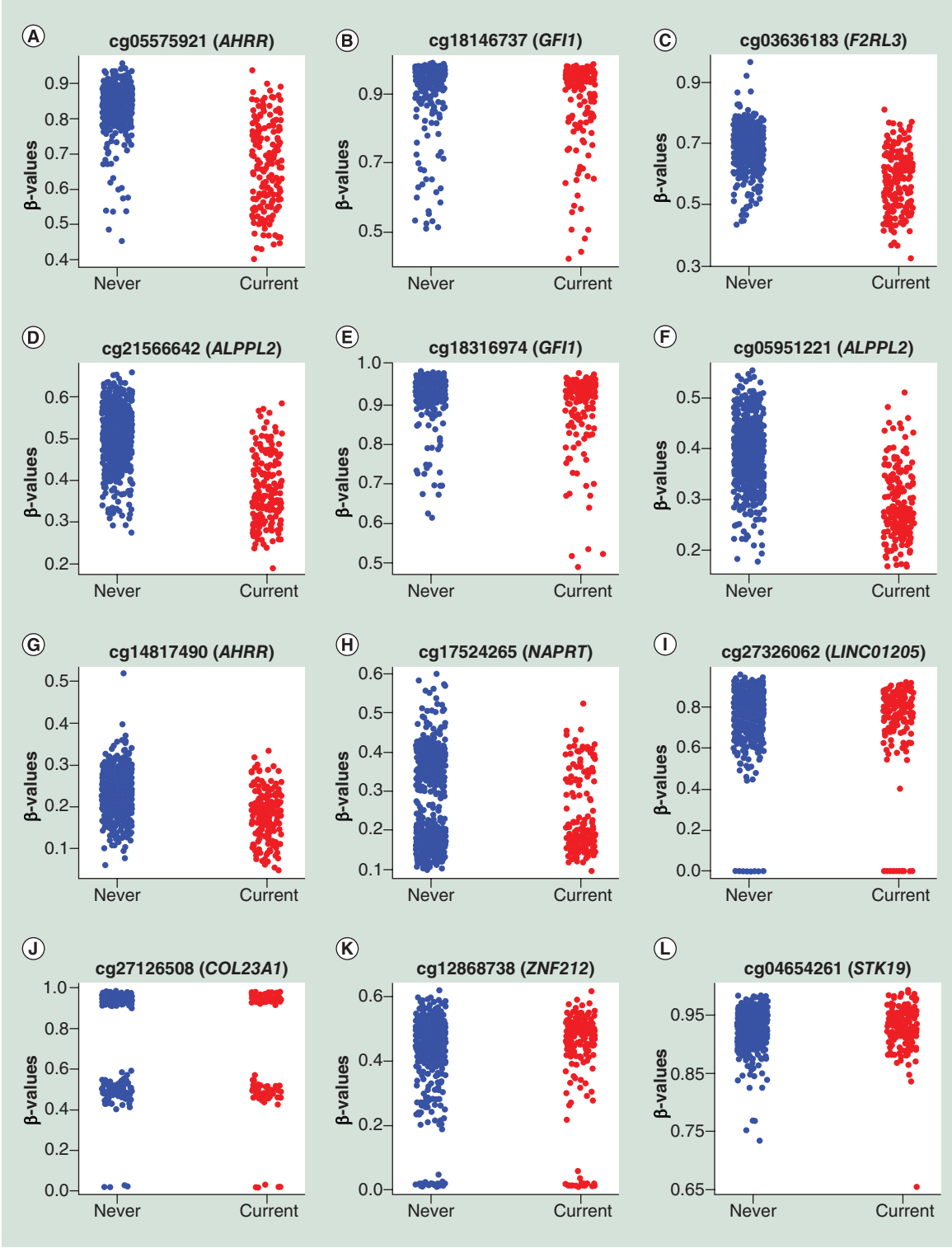


**Figure 4. Most variable probes in response to smoking exposure. (A–I)** Scatter plots showing the 12 most variable probes in the indicated genes in response to smoking exposure (never-smokers vs current smokers). The x axis shows the two categories, never-smokers and current smokers, and the y axis shows the methylation value (β-value). Note that current smokers show more variability compared with never-smokers, indicating that smoking may be responsible for the interindividual variability in DNA methylation patterns.

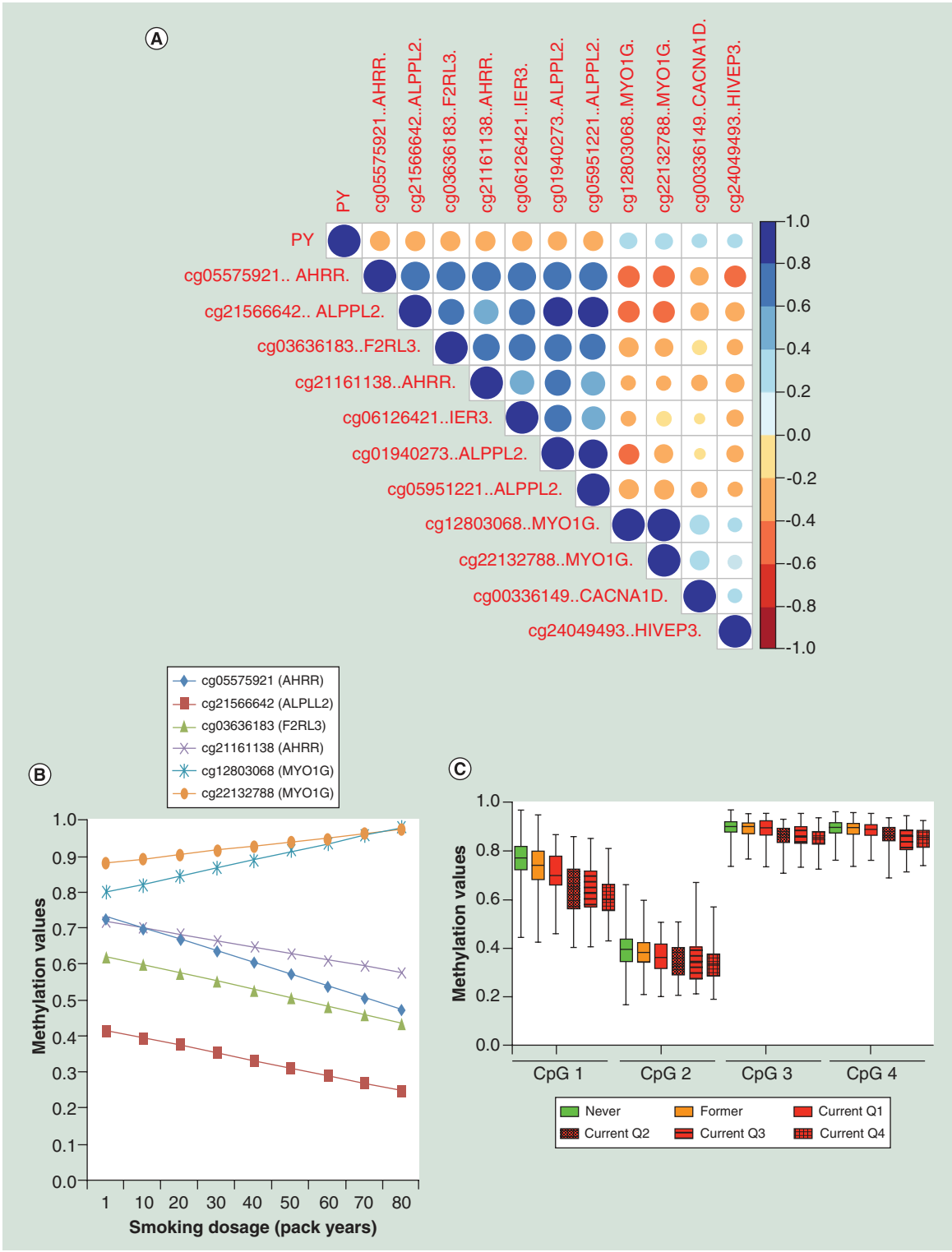


**Figure 5. Effect of smoking dosage on DNA methylation.** **(A)** Correlation plots showing the effect of smoking dosage measured in pack-years (PY) on DNA methylation across the indicated CpG sites. **(B)** Line plot showing the extent of methylation changes across six CpG sites in response to smoking dosage (PY). **(C)** An example shows alterations in methylation of the *IER3* gene in response to PY quartiles (Q1: 1.0–8.1; Q2: 8.2–16.4; Q3: 16.5–26.3 years; and Q4: 26.4–79.5 years). The x axis shows the groups of samples, and the y axis shows the methylation values. CpG1: cg06126421, CpG2: cg14753356, CpG3: cg15342087, CpG4: cg24859433.

Table 6. Dynamics of differentially methylated probes correlated to time since smoking cessation.

| CpGs | Mean methylation (β-values) in former smokers | | | | Means methylation (β-values) in never smokers | Genes | p-value vs Q1 | p-value vs Q2 | p-value vs Q3 | p-value vs Q4 |
|---|---|---|---|---|---|---|---|---|---|---|
| | Quartile 1 | Quartile 2 | Quartile 3 | Quartile 4 | | | | | | |
| cg01940273 | 0.57 | 0.61 | 0.61 | 0.63 | 0.63 | ***ALPPL2*** | **$5.5927 \times 10^{-9}$** | **0.00828874** | **0.02283992** | 0.881884 |
| cg05951221 | 0.33 | 0.36 | 0.37 | 0.40 | 0.40 | ***ALPPL2*** | **$6.7997 \times 10^{-11}$** | **0.00024** | **0.00295074** | 0.959675 |
| cg11554391 | 0.11 | 0.12 | 0.14 | 0.12 | 0.12 | ***AHRR*** | **0.01215898** | **0.03837115** | **0.04176547** | 0.84129 |
| cg21566642 | 0.43 | 0.47 | 0.49 | 0.52 | 0.52 | ***ALPPL2*** | **$1.8373 \times 10^{-11}$** | **0.00027905** | **0.01322902** | 0.520927 |
| cg00310412 | 0.48 | 0.48 | 0.49 | 0.50 | 0.50 | *SEMA7A* | 0.01688095 | 0.01333868 | 0.096259 | 0.473605 |
| cg03636183 | 0.63 | 0.66 | 0.69 | 0.71 | 0.70 | *F2RL3* | $2.9071 \times 10^{-06}$ | 0.00089251 | 0.4386763 | 0.045474 |
| cg05575921 | 0.78 | 0.82 | 0.84 | 0.86 | 0.85 | *AHRR* | $1.2797 \times 10^{-06}$ | 0.00787796 | 0.24663336 | 0.239024 |
| cg06126421 | 0.69 | 0.73 | 0.75 | 0.76 | 0.76 | *IER3* | $1.2063 \times 10^{-07}$ | 0.02072848 | 0.19120049 | 0.947284 |
| cg11207515 | 0.41 | 0.40 | 0.37 | 0.36 | 0.37 | *MIR548F4* | 0.00260903 | 0.01122727 | 0.89615034 | 0.206525 |
| cg14817490 | 0.21 | 0.22 | 0.24 | 0.24 | 0.23 | *AHRR* | 0.01120361 | 0.03547913 | 0.57060349 | 0.339304 |
| cg02583484 | 0.27 | 0.27 | 0.28 | 0.28 | 0.28 | *HNRNPA1* | 0.03747302 | 0.36482799 | 0.80813468 | 0.36306 |
| cg03329539 | 0.34 | 0.36 | 0.37 | 0.37 | 0.36 | *ALPPL2* | 0.0003429 | 0.65500009 | 0.3475772 | 0.435269 |
| cg07178945 | 0.35 | 0.34 | 0.32 | 0.33 | 0.34 | *FGF23* | 0.02734095 | 0.60297549 | 0.03766244 | 0.270658 |
| cg07251887 | 0.44 | 0.45 | 0.46 | 0.47 | 0.46 | *SMIM6* | 0.00239833 | 0.10115834 | 0.53153069 | 0.724767 |
| cg07826859 | 0.59 | 0.61 | 0.61 | 0.62 | 0.61 | *MYO1G* | 0.04224382 | 0.98981846 | 0.99753609 | 0.247555 |
| cg11660018 | 0.50 | 0.53 | 0.53 | 0.54 | 0.53 | *PRSS23* | 0.00070275 | 0.51864285 | 0.21720914 | 0.645713 |
| cg12678834 | 0.71 | 0.71 | 0.72 | 0.72 | 0.72 | *CXCR5* | 0.08283521 | 0.04989528 | 0.55835608 | 0.625891 |
| cg15342087 | 0.89 | 0.89 | 0.90 | 0.90 | 0.90 | *IER3* | 0.03029082 | 0.09337822 | 0.58724609 | 0.386983 |
| cg19254163 | 0.61 | 0.61 | 0.62 | 0.63 | 0.62 | *PTGDR2* | 0.01973658 | 0.06441623 | 0.66916667 | 0.245609 |
| cg21161138 | 0.74 | 0.76 | 0.76 | 0.77 | 0.76 | *AHRR* | 0.00702473 | 0.30804278 | 0.2701388 | 0.489021 |
| cg21611682 | 0.54 | 0.56 | 0.56 | 0.56 | 0.55 | *LRP5* | 0.03314418 | 0.5958874 | 0.50756132 | 0.618108 |
| cg23576855 | 0.60 | 0.68 | 0.67 | 0.72 | 0.69 | *AHRR* | 0.00184148 | 0.88022969 | 0.53958606 | 0.182808 |
| cg24859433 | 0.88 | 0.88 | 0.89 | 0.90 | 0.89 | *IER3* | 0.0062015 | 0.05360747 | 0.7276883 | 0.048765 |
| cg25189904 | 0.41 | 0.43 | 0.45 | 0.45 | 0.45 | *GNG12* | 0.00016015 | 0.09358426 | 0.56536275 | 0.70142 |
| cg26361535 | 0.71 | 0.73 | 0.72 | 0.75 | 0.73 | *ZC3H3* | 0.04332868 | 0.97194407 | 0.2789403 | 0.03573 |
| cg27241845 | 0.69 | 0.70 | 0.71 | 0.72 | 0.71 | *ECEL1P2* | 0.02366828 | 0.42788915 | 0.85639913 | 0.17679 |
| cg01901332 | 0.65 | 0.67 | 0.66 | 0.67 | 0.66 | *MIR326* | 0.05134314 | 0.87688668 | 0.68211387 | 0.4146 |

Quartiles of former smokers based on time since smoking cessation.
Q1 (0.5–6.5 years), Q2 (6.6–14 years), Q3 (14.1–22 years) and Q4 (22.1–42.5 years).
p-value calculated by comparing average values of the former smoker quartiles to never smokers (two-tailed *t*-test).
The genes and p-values found significant in three comparisons (never smokers vs Q1, never smokers vs Q2 and never smokers vs Q3) are highlighted in bold.

Table 6. Dynamics of differentially methylated probes correlated to time since smoking cessation (cont.).

| CpGs | Mean methylation (β-values) in former smokers | | | | Means methylation (β-values) in never smokers | Genes | p-value vs Q1 | p-value vs Q2 | p-value vs Q3 | p-value vs Q4 |
|---|---|---|---|---|---|---|---|---|---|---|
| | Quartile 1 | Quartile 2 | Quartile 3 | Quartile 4 | | | | | | |
| cg03991871 | 0.91 | 0.93 | 0.93 | 0.94 | 0.93 | *AHRR* | 0.10399607 | 0.98884032 | 0.72413396 | 0.047233 |
| cg04517079 | 0.60 | 0.60 | 0.61 | 0.61 | 0.60 | *MIR4641* | 0.54453422 | 0.30176138 | 0.44631373 | 0.119554 |
| cg05284742 | 0.73 | 0.73 | 0.75 | 0.74 | 0.74 | *ITPK1–AS1* | 0.18789432 | 0.28723319 | 0.4130127 | 0.72396 |
| cg10750182 | 0.54 | 0.54 | 0.55 | 0.56 | 0.55 | *C10orf105* | 0.07068565 | 0.33291944 | 0.59591936 | 0.162469 |
| cg12806681 | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 | *AHRR* | 0.14264054 | 0.8701124 | 0.71900069 | 0.028203 |
| cg18642234 | 0.45 | 0.45 | 0.47 | 0.47 | 0.46 | *GPX1* | 0.14499533 | 0.26460218 | 0.30863683 | 0.513251 |
| cg19859270 | 0.96 | 0.95 | 0.96 | 0.96 | 0.96 | *GPR15* | 0.8514209 | 0.22525174 | 0.12763399 | 0.046761 |
| cg23110422 | 0.90 | 0.91 | 0.91 | 0.92 | 0.91 | *ETS2* | 0.13075123 | 0.70350013 | 0.80569494 | 0.268912 |
| cg24908166 | 0.93 | 0.93 | 0.93 | 0.94 | 0.93 | *TERT* | 0.40839562 | 0.60392526 | 0.91739825 | 0.083909 |
| cg26963277 | 0.89 | 0.90 | 0.89 | 0.90 | 0.90 | *KCNQ1OT1* | 0.26523809 | 0.91639119 | 0.54018833 | 0.104414 |

Quartiles of former smokers based on time since smoking cessation.
Q1 (0.5–6.5 years), Q2 (6.6–14 years), Q3 (14.1–22 years) and Q4 (22.1–42.5 years).
p-value calculated by comparing average values of the former smoker quartiles to never smokers (two-tailed *t*-test).
The genes and p-values found significant in three comparisons (never smokers vs Q1, never smokers vs Q2 and never smokers vs Q3) are highlighted in bold.

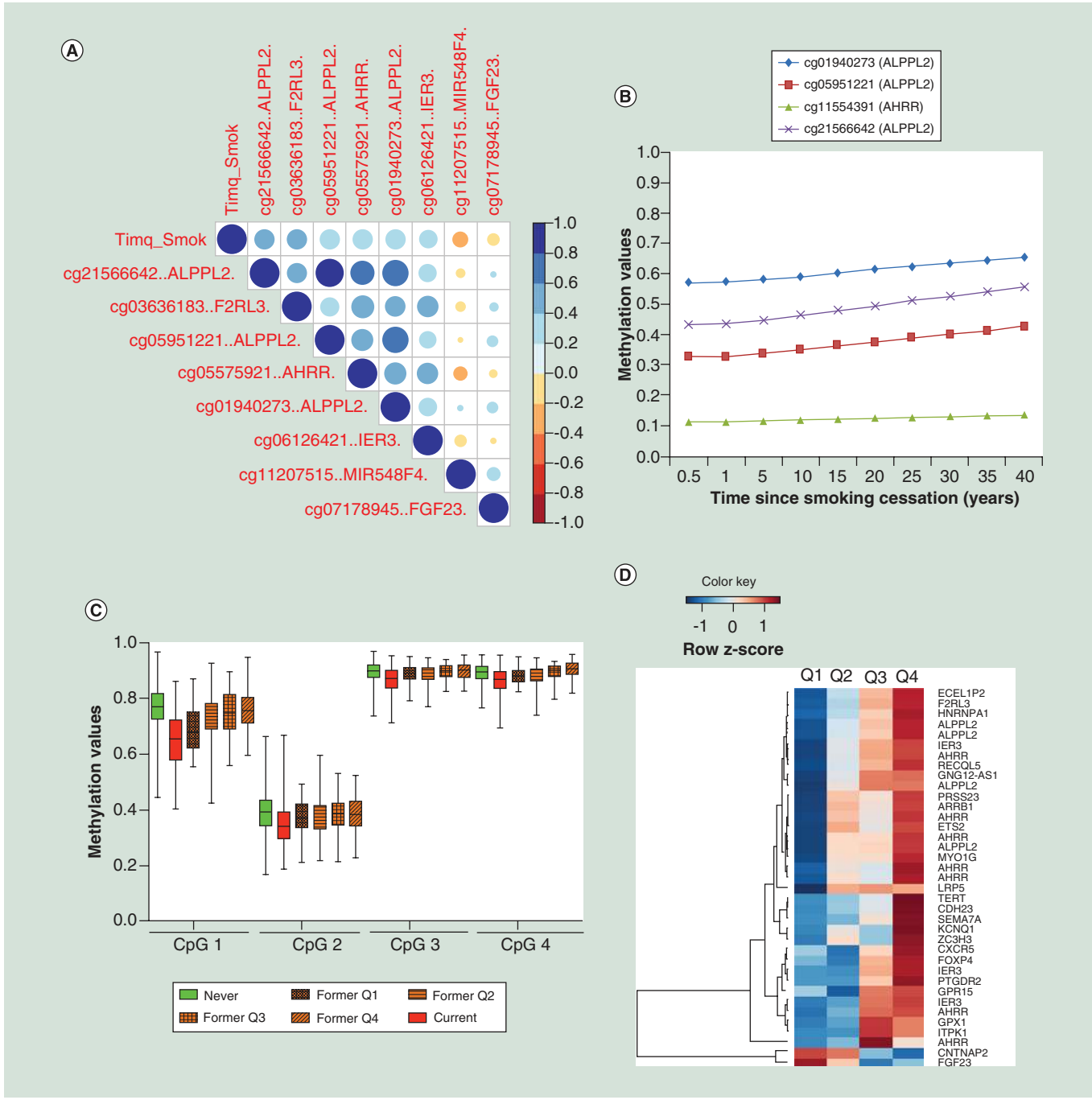


**Figure 6. Effect of time since smoking cessation on DNA methylation.** **(A)** Correlation plot showing the effect of time since smoking cessation on DNA methylation across the indicated CpG sites. **(B)** Line plot showing the extent of methylation changes across four CpG sites which remained differentially methylated between never smokers and former smokers up to 14.1–22 years after smoking cessation. **(C)** An example shows the effect of time since cessation on DNA methylation of the *IER3* gene based on quartiles of time since cessation; the x axis shows the sample groups, and the y axis shows the methylation values. CpG1: cg06126421, CpG2: cg14753356, CpG3: cg15342087, CpG4: cg24859433. **(D)** Heat map showing methylation dynamics and reversibility at 37 CpG sites that were significantly correlated with time between smoking cessation and blood collection (Timq_smok). We compared the methylation levels at these sites between never-smokers and Timq_smok in former smokers (Q1: 0.5–6.5 years; Q2: 6.6–14 years; Q3: 14.1–22 years; and Q4: 22.1–42.5 years).

and a high level of concordance between the smoking-associated methylation changes identified in our study and those in other groups, we think that our findings are robust and are highly unlikely to be a result of confounding and biases. Considering a binary type of methylation data (a given cytosine can be methylated

or not at the level of single cells) [58], small methylation changes in complex tissues, such as white blood cells, may reflect substantial changes in specific subpopulation. While purification of blood cell subtypes for epigenetic analyses in large molecular epidemiology studies is not feasible, a detailed follow-up study using *in vitro* and *in vivo* models should address functional implications of smoking-induced methylation changes.

Through the present study, we have comprehensively cataloged the smoking-associated DNA methylation alterations (DMPs, DMRs and MVPs) in a large prospective study. We have reported several novel epigenetic biomarkers of smoking, encompassing 12 novel DMPs, two coding and noncoding DMRs and seven MVPs. Furthermore, we have identified methylation biomarkers that may be used for follow-up of former smokers and to assess lifetime exposure to tobacco smoke. These findings may have important contributions to the understanding of the mechanistic biology and reversibility effects in response to smoking and may be coupled with existing biomarkers of smoking.

## Conclusion

The present study examined the potential of DNA methylation changes in prediagnostic peripheral blood samples as a marker of exposure to tobacco smoke in a large multinational cohort of the EPIC study, using the Illumina HumanMethylation 450K BeadChip array. We identified a total of 748 CpG sites that were differentially methylated between baseline smokers and nonsmokers. While many differentially methylated sites included CpG sites from genes identified in previous studies, we identified novel CpG sites associated with smoking. Dimension reduction approaches further revealed novel regionally clustered CpGs associated with active smoking exposure. Importantly, we found a marked reversibility of methylation changes after smoking cessation, although specific genes remained differentially methylated up to 22 years after cessation.

Thus, our study has comprehensively cataloged the smoking-associated DNA methylation alterations including newly identified coding and noncoding genes associated with active smoking. Our study also showed that prediagnostic smoking-related epigenetic alterations in human blood cells are reversible after smoking cessation, consistent with the known cancer risk reduction.

## Future perspective

We envisage that the markers of smoking-related methylation changes identified through our study could be used to assess the lifetime exposure to smoking in current smokers and more importantly to follow former smokers. Thus, our study provides interesting targets for smoking-related secondary cancer prevention. Future studies to understand the mechanistic link between our findings and validation of our findings in other study cohorts are warranted.

### Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/full/10.2217/epi-2016-0001

### Acknowledgements

The authors are grateful to K Müller for editing the manuscript. This study depended on the participation of the women in the EPIC cohort, to whom they are grateful.

### Financial & competing interests disclosure

This work was supported by grants from the Institut National du Cancer (INCa, France) to I Romieu and Z Herceg and the European Commission (EC) Seventh Framework Programme (FP7) Translational Cancer Research (TRANSCAN) Framework to Z Herceg. Z Herceg 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 S Ambatipudi and A Ghantous) 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). EPIC-Greece was supported by the Hellenic Health Foundation; DP was supported by a grant from Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy and JRQ was supported by the regional government of Asturias. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

### Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

## Author affiliations

[1]International Agency for Research on Cancer (IARC), Lyon, France
[2]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[3]Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbrücke, Nuthetal, Germany
[4]Hellenic Health Foundation, Athens, Greece
[5]WHO Collaborating Center for Nutrition & Health, Unit of Nutritional Epidemiology & Nutrition in Public Health, Department of Hygiene, Epidemiology & Medical Statistics, University of Athens Medical School, Athens, Greece
[6]Molecular & Nutritional Epidemiology Unit, Cancer Research & Prevention Institute–ISPO, Florence, Italy
[7]Epidemiology & Prevention Unit, Fondazione IRCCS Istituto Nazionale Tumori, Milano, Italy
[8]Human Genetic Foundation (HuGeF), Torino, Italy
[9]Cancer Registry & Histopathology Unit, 'Civic MP Arezzo' Hospital, ASP Ragusa, Italy
[10]Dipartimento di Medicina Clinica e Chirurgia, Federico II University, Naples, Italy
[11]Department of Determinants of Chronic Diseases (DCD), National Institute for Public Health & the Environment (RIVM), Bilthoven, The Netherlands
[12]Department of Gastroenterology & Hepatology, University Medical Centre, Utrecht, The Netherlands
[13]Department of Epidemiology & Biostatistics, The School of Public Health, Imperial College London, London, UK
[14]Department of Social & Preventive Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia
[15]Department of Epidemiology, Julius Center for Health Sciences & Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
[16]MRC-PHE Centre for Environment & Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London, UK
[17]Public Health Directorate, Asturias, Spain
[18]Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain
[19]CIBER Epidemiología y Salud Pública (CIBERESP), Spain
[20]Department of Health & Social Sciences, Universidad de Murcia, Spain
[21]Public Health Institute of Navarra, Pamplona, Spain
[22]IdiSNA, Navarra Institute for Health Research, Pamplona, Spain
[23]Public Health Direction and Biodonostia–Ciberesp, Basque Regional Health Department, San Sebastian, Spain
[24]Cancer Epidemiology Unit, University of Oxford, Oxford, UK
[25]School of Public Health, Imperial College London, London, UK


## Executive summary

- Using a large prospective study, our study has comprehensively cataloged the smoking-associated DNA methylation alterations including novel regionally altered coding and noncoding genes.
- We identified a total of 748 CpG sites that were differentially methylated between smokers and nonsmokers, including novel regionally clustered CpGs associated with active smoking exposure.
- We found a marked reversibility of methylation changes after smoking cessation and discrete genes that remained differentially methylated decades after cessation.
- Our study revealed that prediagnostic smoking-related epigenetic alterations in human blood cells are reversible after smoking cessation, consistent with the known cancer risk reduction.

## References


1 WHO. Burden: mortality, morbidity and risk factors. www.who.int/nmh/publications/ncd_report_chapter1.pdf

2 Jha P, Chaloupka FJ, Moore M *et al.* Tobacco Addiction. In: *Disease Control Priorities in Developing Countries*. Jamison DT, Breman JG, Measham AR (Eds). The International Bank for Reconstruction and Development, Washington DC, USA (2006).

3 Breitling LP. Current genetics and epigenetics of smoking/tobacco-related cardiovascular disease. *Arterioscler. Thromb. Vasc. Biol.* 33(7), 1468–1472 (2013).

4 Herceg Z, Vaissiere T. Epigenetic mechanisms and cancer: an interface between the environment and the genome. *Epigenetics* 6(7), 804–819 (2011).

5 Ghantous A, Hernandez-Vargas H, Byrnes G, Dwyer T, Herceg Z. Characterising the epigenome as a key component of the fetal exposome in evaluating in utero exposures and childhood cancer risk. *Mutagenesis* 30(6), 733–742 (2015).

6 Feil R, Fraga MF. Epigenetics and the environment: emerging patterns and implications. *Nat. Rev. Genet.* 13(2), 97–109 (2011).

7 Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* 12(8), 529–541 (2011).

8 Issa JP. Aging and epigenetic drift: a vicious cycle. *J. Clin. Invest.* 124(1), 24–29 (2014).

9 Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. *Nat. Rev. Genet.* 14(8), 585–594 (2013).

10 Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am. J. Hum. Genet.* 88(4), 450–457 (2011).

11 Wan ES, Qiu W, Baccarelli A *et al.* Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum. Mol. Genet.* 21(13), 3073–3082 (2012).

12 Joubert BR, Haberg SE, Nilsen RM *et al.* 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.* 120(10), 1425–1431 (2012).

13 Shenker NS, Polidoro S, Van Veldhoven K *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.* 22(5), 843–851 (2013).

14 Zeilinger S, Kuhnel B, Klopp N *et al.* Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS ONE* 8(5), e63812 (2013).

15 Harlid S, Xu Z, Panduri V, Sandler DP, Taylor JA. CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the Sister Study. *Environ. Health Perspect.* 122(7), 673–678 (2014).

16 Markunas CA, Xu Z, Harlid S *et al.* Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.* 122(10), 1147–1153 (2014).

17 Heyn H, Carmona FJ, Gomez A *et al.* DNA methylation profiling in breast cancer discordant identical twins identifies DOK7 as novel epigenetic biomarker. *Carcinogenesis* 34(1), 102–108 (2013).

18 Riboli E, Hunt KJ, Slimani N *et al.* European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.* 5(6B), 1113–1124 (2002).

19 Bingham S, Riboli E. Diet and cancer – the European Prospective Investigation into Cancer and Nutrition. *Nat. Rev. Cancer* 4(3), 206–215 (2004).

20 Bibikova M, Barnes B, Tsan C *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* 98(4), 288–295 (2011).

21 Hernandez-Vargas H, Castelino J, Silver MJ *et al.* Exposure to aflatoxin B1 in utero is associated with DNA methylation in white blood cells of infants in The Gambia. *Int. J. Epidemiol.* 44(4), 1238–1248 (2015).

22 Martin M, Ancey PB, Cros MP *et al.* Dynamic imbalance between cancer cell subpopulations induced by transforming growth factor beta (TGF-beta) is associated with a DNA methylome switch. *BMC Genomics* 15, 435 (2014).

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

24 Teschendorff AE, Marabita F, Lechner M *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 29(2), 189–196 (2013).

25 Aryee MJ, Jaffe AE, Corrada-Bravo H *et al.* Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30(10), 1363–1369 (2014).

26 Houseman EA, Accomando WP, Koestler DC *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 13, 86 (2012).

27 Du P, Zhang X, Huang CC *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 11, 587 (2010).

28 Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3(9), 1724–1735 (2007).

29 Ritchie ME, Phipson B, Wu D *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43(7), e47 (2015).

30 Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc.* 66(1), 187–205 (2004).

31 Jaffe AE, Murakami P, Lee H *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* 41(1), 200–209 (2012).

32 Hansen KD, Timp W, Bravo HC *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* 43(8), 768–775 (2011).

33 Phipson B, Oshlack A. DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biol.* 15(9), 465 (2014).

34 Harper KN, Peters BA, Gamble MV. Batch effects and pathway analysis: two potential perils in cancer studies involving DNA methylation array analysis. *Cancer Epidemiol. Biomarkers Prev.* 22(6), 1052–1060 (2013).

35 Chen EY, Tan CM, Kou Y *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14, 128 (2013).

36 Teschendorff AE, Jones A, Fiegl H *et al.* Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Med.* 4(3), 24 (2012).

37 Lee KW, Richmond R, Hu P *et al.* Prenatal exposure to maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. *Environ. Health Perspect.* 123(2), 193–199 (2015).

38 Richmond RC, Simpkin AJ, Woodward G *et al.* Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum. Mol. Genet.* 24(8), 2201–2217 (2014).

39 Breton CV, Siegmund KD, Joubert BR *et al.* Prenatal tobacco smoke exposure is associated with childhood DNA CpG methylation. *PLoS ONE* 9(6), e99716 (2014).

40 Joubert BR, Haberg SE, Bell DA *et al.* Maternal smoking and DNA methylation in newborns: in utero effect or epigenetic inheritance? *Cancer Epidemiol. Biomarkers Prev.* 23(6), 1007–1017 (2014).

41 Tsaprouni LG, Yang TP, Bell J *et al.* Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics* 9(10), 1382–1396 (2014).

42 Zaghlool SB, Al-Shafai M, Al Muftah WA, Kumar P, Falchi M, Suhre K. Association of DNA methylation with age, gender, and smoking in an Arab population. *Clin. Epigenet.* 7(1), 6 (2015).

43 Dogan MV, Shields B, Cutrona C *et al.* The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. *BMC Genomics* 15, 151 (2014).

44 Besingi W, Johansson A. Smoke-related DNA methylation changes in the etiology of human disease. *Hum. Mol. Genet.* 23(9), 2290–2297 (2014).

45 Wan ES, Qiu W, Baccarelli A *et al.* Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum. Mol. Genet.* 21(13), 3073–3082 (2012).

46 Nguyen LP, Bradfield CA. The search for endogenous activators of the aryl hydrocarbon receptor. *Chem. Res. Toxicol.* 21(1), 102–116 (2008).

47 Fasanelli F, Baglietto L, Ponzi E *et al.* Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat. Commun.* 6, 10192 (2015).

48 Tucker DF, Oliver RT, Travers P, Bodmer WF. Serum marker potential of placental alkaline phosphatase-like activity in testicular germ cell tumours evaluated by H17E2 monoclonal antibody assay. *Br. J. Cancer* 51(5), 631–639 (1985).

49 Koshida K, Stigbrand T, Munck-Wikland E, Hisazumi H, Wahren B. Analysis of serum placental alkaline phosphatase activity in testicular cancer and cigarette smokers. *Urol. Res.* 18(3), 169–173 (1990).

50 Tan LY, Whitfield P, Llorian M *et al.* Generation of functionally distinct isoforms of PTBP3 by alternative splicing and translation initiation. *Nucleic Acids Res.* 43(11), 5586–5600 (2015).

51 Maunakea AK, Nagarajan RP, Bilenky M *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466(7303), 253–257 (2010).

52 Brazao TF, Demmers J, Van IW *et al.* A new function of ROD1 in nonsense-mediated mRNA decay. *FEBS Lett.* 586(8), 1101–1110 (2012).

53 Treppendahl MB, Qiu X, Sogaard A *et al.* Allelic methylation levels of the noncoding VTRNA2-1 located on chromosome 5q31.1 predict outcome in AML. *Blood* 119(1), 206–216 (2012).

54 Silver MJ, Kessler NJ, Hennig BJ *et al.* Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment. *Genome Biol.* 16, 118 (2015).

55 Rossi A, Trotta E, Brandi R, Arisi I, Coccia M, Santoro MG. AIRAP, a new human heat shock gene regulated by heat shock factor 1. *J. Biol. Chem.* 285(18), 13607–13615 (2010).

56 Medvedeva YA, Fridman MV, Oparina NJ *et al.* Intergenic, gene terminal, and intragenic CpG islands in the human genome. *BMC Genomics* 11, 48 (2010).

57 Burke GL, Savage PJ, Manolio TA *et al.* Correlates of obesity in young black and white women: the CARDIA Study. *Am. J. Public Health* 82(12), 1621–1625 (1992).

58 Herceg Z, Hernandez-Vargas H. New concepts of old epigenetic phenomena and their implications for selecting specific cell populations for epigenomic research. *Epigenomics* 12(4), 383–386 (2011).