Original Investigation

Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer

Andrew E. Teschendorff, PhD; Zhen Yang, PhD; Andrew Wong, PhD; Christodoulos P. Pipinikas, PhD; Yinming Jiao, MSc; Allison Jones, BSc; Shahzia Anjum, PhD; Rebecca Hardy, PhD; Helga B. Salvesen, MD; Christina Thirlwell, PhD; Samuel M. Janes, PhD; Diana Kuh, PhD; Martin Widschwendter, MD

IMPORTANCE The utility of buccal cells as an epithelial source tissue for epigenome-wide association studies (EWASs) remains to be demonstrated. Given the direct exposure of buccal cells to potent carcinogens such as smoke, epigenetic changes in these cells may provide insights into the development of smoke-related cancers.

OBJECTIVE To perform an EWAS in buccal and blood cells to assess the relative effect of smoking on the DNA methylation (DNAme) patterns in these cell types and to test whether these DNAme changes are also seen in epithelial cancer.

DESIGN, SETTING, AND PARTICIPANTS In 2013, we measured DNAme at more than 480 000 CpG sites in buccal samples provided in 1999 by 790 women (all aged 53 years in 1999) from the United Kingdom Medical Research Council National Survey of Health and Development. This included matched blood samples from 152 women. We constructed a DNAme-based smoking index and tested its sensitivity and specificity to discriminate normal from cancer tissue in more than 5000 samples.

MAIN OUTCOMES AND MEASURES CpG sites whose DNAme level correlates with smoking pack-years, and construction of an associated sample-specific smoking index, which measures the mean deviation of DNAme at smoking-associated CpG sites from a normal reference.

RESULTS In a discovery set of 400 women, we identified 1501 smoking-associated CpG sites at a genome-wide significance level of $P < 10^{-7}$, which were validated in an independent set of 390 women. This represented a 40-fold increase of differentially methylated sites in buccal cells compared with matched blood samples. Hypermethylated sites were enriched for bivalently marked genes and binding sites of transcription factors implicated in DNA repair and chromatin architecture ($P < 10^{-10}$). A smoking index constructed from the DNAme changes in buccal cells was able to discriminate normal tissue from cancer tissue with a mean receiver operating characteristic area under the curve of 0.99 (range, 0.99-1.00) for lung cancers and of 0.91 (range, 0.71-1.00) for 13 other organs. The corresponding area under the curve of a smoking signature derived from blood cells was lower than that derived from buccal cells in 14 of 15 cancer types (Wilcoxon signed rank test, $P = .001$).

CONCLUSIONS AND RELEVANCE These data point toward buccal cells as being a more appropriate source of tissue than blood to conduct EWASs for smoking-related epithelial cancers.

JAMA Oncol. 2015;1(4):476-485. doi:10.1001/jamaoncol.2015.1053
Published online May 14, 2015. Corrected on May 21, 2015.

Supplemental content at jamaoncology.com

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Authors: Martin Widschwendter, MD, Department of Women's Cancer, EGA Institute for Women's Health, University College London, 74 Huntley St, Rm 340, London WC1E 6AU, England (m.widschwendter@ucl.ac.uk) and Andrew E. Teschendorff, PhD, Statistical Cancer Genomics, UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, England (a.teschendorff@ucl.ac.uk).

Exposure to tobacco smoke is one of the best-known and potent risk factors for many diseases, including epithelial cancers and notably lung cancer.[1,2] Mortality rates are set to increase, with more than 1 billion expected tobacco-related deaths during this century.[3,4] There is, therefore, an urgent need to advance our understanding of tobacco-related cancer etiology.[5,6]

Recent work has demonstrated a role for epigenetic changes, especially changes in DNA methylation (DNAme), during the earliest stages of carcinogenesis.[7-14] It is therefore important to study the potential impact of tobacco smoke exposure on the epigenome. Although several epigenome-wide association studies (EWASs) conducted in whole blood have been performed that have identified smoking-related DNAme changes centered around specific genes, eg, *AHRR*,[15-19] these associations are relatively few in number and the role of these alterations in cancer etiology is unclear.

We hypothesized that a more natural and relevant tissue in which to perform an EWAS for smoking would be buccal cells because this constitutes an easily accessible source of epithelial cells with direct exposure to tobacco smoke. In particular, we posited that buccal cells would provide us with a more relevant tissue than blood to conduct an EWAS aimed at understanding epigenetic misprogramming in epithelial cancer. Specifically, we set out to explore the hypothesis that DNAme marks measured in buccal cells, and which correlate with a measure of the cumulative exposure to smoke, may exhibit similar changes in epithelial cancers, especially in those cancers strongly linked to smoking.

To test these hypotheses, we conducted an EWAS in 790 buccal samples collected from women within the Medical Research Council National Survey of Health and Development (NSHD),[20,21] a longitudinal birth cohort with extensive epidemiological data, including detailed information on smoking history and smoking status at sample collection.[22] To assess the relevance of smoking-associated DNAme changes in cancer, we further analyzed genome-wide DNAme data from more than 5000 samples encompassing normal, preneoplastic, and cancer tissue from 15 different epithelial cancer types.

Methods

The overall analysis strategy is summarized in eFigure 1 in the Supplement.

Data Sets and Ethics Approval

In 2013, we analyzed buccal samples that had been provided in 1999 by 790 women enrolled in the NSHD, a birth cohort study of men and women all born in Britain in March 1946.[22,23] All women gave written informed consent for their samples to be used in genetic studies of health, and the Central Manchester Ethics Committee approved the use of these samples for epigenetic studies of health in 2012. Women were selected from those who provided a buccal and blood sample at age 53 years in 1999, who had not previously developed any cancer, and who had complete information on epidemiological variables of interest and follow-up (**Table 1**). For 152 of the 790 women, we also analyzed a matched whole blood sample.

Samples from preinvasive lung lesions were taken from a cohort described recently.[24,25] A subset of 24 laser-microdissected samples, consisting of lesions that did (n = 19) and did not (n = 5) progress to invasive lung cancer (all assessed by means of bronchoscopy) and that were matched for smoking pack-years (SPY), was used. In addition, 21 normal lung samples (bronchial brushings) from individuals at high risk of developing lung cancer were taken from anatomical sites with no documented history of preinvasive lesions. See eMethods in the Supplement for details regarding the data sets used.

At a Glance

- The purpose of the research was to assess the suitability of buccal cells as an epithelial source of tissue to examine the effects of smoking on the epigenome, and to test whether these effects are also seen in smoke-related epithelial cancers.
- Smoking is associated with widespread changes in the DNA methylation landscape of buccal cells.
- Some smoking-associated DNA methylation changes are common to buccal and blood tissue, but buccal cells exhibit significantly more changes than blood cells.
- Smoking-associated DNA methylation changes in buccal cells correlate with DNA methylation changes in epithelial cancers and do so most strongly in smoke-related epithelial cancer.

DNAme Analysis

Methylation analysis was performed on DNA from 790 buccal and 152 blood samples using the Illumina Infinium Human Methylation450 BeadChip array.[26,27] The methylation status of a specific CpG site was calculated from the intensity of the methylated (M) and unmethylated (U) alleles, as the ratio of fluorescent signals $\beta = \max(M,0)/[\max(M,0) + \max(U,0) + 100]$. On this scale, $0 < \beta < 1$, with β values close to 1 (0) indicating 100% methylation (no methylation). Data were processed and normalized, correcting for type 2 probe bias,[28,29] and using a quality control pipeline that assesses the nature of the largest components of variation[30] (see eMethods in the Supplement). DNA from preinvasive lung lesions and normal adjacent tissue was extracted from fresh frozen laser capture microdissected sections (or bronchial brushings from controls), and genome-wide DNAme profiles were obtained using the Methylation450 BeadChip.

Statistical Analyses

A discovery set was defined by randomly selecting 400 of the 790 buccal samples, and linear regression analyses adjusted for bisulfite conversion efficiency were used to identify CpGs correlating significantly with SPY. The Bonferroni threshold was subsequently used to define a 1501-CpG smoking-associated DNAme signature. Gene set enrichment analysis (GSEA) of this signature was done using the Molecular Signatures Database,[31] using 1-tailed Fisher exact tests, as done previously.[30] The 1501 smoking-associated CpGs were then used to construct a sample-specific smoking index, which measures the deviation of DNAme in a given sample from a normal reference, with

Table 1. Characteristics of the Participants in the Study Sets

Characteristic	Discovery Set (n = 400)	Replication Set (n = 390)	Subset[a] (n = 152)
Age	53	53	53
Sex	Female	Female	Female
Sample statistics for smoking history, No. (%)			
Never-smoker	191 (47.8)	193 (49.5)	72 (47.4)
Ex-smoker	127 (31.8)	112 (28.7)	49 (32.2)
Current smoker	82 (20.5)	85 (21.8)	31 (20.4)
Sample statistics for SPY			
≥10	110 (27.5)	109 (27.9)	35 (23.0)
<10	215 (53.8)	213 (54.6)	87 (57.2)
NA	75 (18.8)	68 (17.4)	30 (19.7)
BMI			
At 15 y, mean (SD)	20.7 (2.7)	20.8 (2.9)	20.9 (2.4)
SCC for 15-y BMI ~ SPY[b]	0.01	0.15	0.24
P value	.89	.01	.01
At 53 y, mean (SD)	27.5 (5.3)	27.9 (5.9)	28.1 (6.4)
SCC for 53-y BMI ~ SPY[b]	0.01	0.03	−0.03
P value	.82	.65	.77
Current physical activity, No. (%)[c]			
None	209 (52)	183 (47)	73 (48)
Moderate, 1-4 times/wk	55 (14)	64 (16)	25 (16)
High, >5 times/wk	136 (34)	143 (37)	54 (36)
Kruskal-Wallis *P* value for physical activity ~ SPY[d]	.01	.05	.31
Parity[e]			
SCC for parity ~ SPY[f]	0.01	0.02	−0.14
P value	.73	.76	.13

Abbreviations: BMI, body mass index (calculated as weight in kilograms divided by height in meters squared); NA, data not available; SCC, Spearman correlation coefficient; SPY, smoking pack-years.

[a] The subset consisted of a set of samples from the discovery and replication set for which we also analyzed blood samples.

[b] 15-y or 53-y BMI ~ SPY indicates correlation between BMI at this age and SPY.

[c] Estimate based on 4 weeks prior to sample collection at age 53 years.

[d] Physical activity ~ SPY indicates whether SPY is associated with physical activity.

[e] Defined as number of children before age 53 years.

[f] Parity ~ SPY indicates correlation between parity and SPY.

the mean taken over the 1501 CpGs. In more detail, given a set of normal reference DNAme profiles, we computed, for each of the 1501 CpGs, the mean β-value, μ_c, and standard deviation, σ_c, across the reference samples. For any given sample, *s*, we then defined the smoking index score, SI(*s*), as

$$\mathrm{SI}(s) = \tfrac{1}{n}\sum_{c}^{n} w_c \frac{\beta_{cs} - \mu_c}{\sigma_c},$$

where w_c is +1 (−1) if the smoking-associated CpG, *c*, is hypermethylated (hypomethylated) in smokers and where β_{cs} is the β-methylation value of the CpG *c* in sample *s*. In the formula, *n* is the number of the 1501 CpGs that have β-values in the given samples, and the summation is over all *n* of these smoking-associated CpGs. In the case of the buccal set cohort, the normal reference samples were those of the nonsmokers. When computing the smoking index in the cancer samples from a given cancer type, the normal reference was chosen as the normal samples (from nonsmokers if this information was available) from the corresponding tissue type. This is key, because this avoids confounding of the smoking index by methylation differences between tissue types (see eMethods in the Supplement for full details).

Rationale for Using Smoking Pack-Years

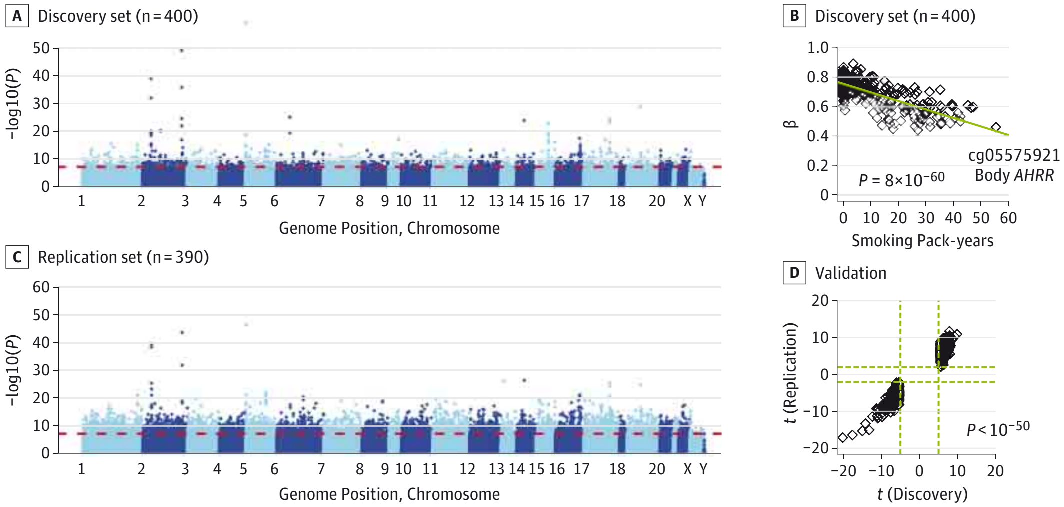
It is important to justify the use of SPY as the outcome of interest, and not the smoking status at sample collection. The latter would have provided a biased measure of the cumulative risk exposure of an individual, especially for ex-smokers. In fact, SPY anticorrelated significantly with the time between quitting and sample collection in ex-smokers (eFigure 2 in the Supplement). Moreover, we observed that specific genes such as *AHRR* exhibited a reversal of DNAme changes in ex-smokers who quit smoking at least 10 years before sample collection (eFigure 3 in the Supplement). Regression analysis using smoking status as a graded response variable (never-smokers, ex-smokers, smokers at sample draw) would have resulted in only 406 CpGs with $P < 10^{-7}$ compared with 1501 when using SPY.

Results

Association of Smoking With Alteration of the DNAme Landscape in Buccal Cells

To assess whether smoking history affects the DNA methylome of buccal cells, we performed genome-wide DNAme analysis[26] in a discovery set of 400 buccal samples from women all aged 53 years (Table 1), thus eliminating chronological age and sex as potential confounding sources of data variation.[32] Singular value decomposition of the DNAme data matrix, encompassing 479 491 probes, revealed that the largest component of variation, accounting for approximately 55% of data variation, correlated with SPY, an epidemiological indicator of an individual's smoking history (eFigure 4 in the Supplement). Other epidemiological variables, such as parity or body mass index, were not significantly correlated with smoking history (Table 1). None of these other variables showed a stronger association with the top principal component than SPY, supporting the view that most of the data variation is linked to smoking (eFigure 4 in the Supplement).

Figure 1. Smoking Epigenome-Wide Association Study in Buccal Cells



A, Manhattan plot in the discovery buccal set (n = 400) samples. A total of 1501 smoking-associated CpGs passed a Bonferroni threshold of approximately 10^{-7} (indicated by red dotted line). B, The methylation β value (y-axis) of the top-ranked CpG, mapping to the gene body of *AHRR*, as a function of smoking pack-years (x-axis). *P* value from a linear regression is given. C, Manhattan plot in the replication set of 390 buccal samples. D, Scatterplot of the linear regression DNA methylation *t* statistics of the 1501 smoking-associated CpGs in the discovery set (x-axis) against those in the replication set (n = 390). *P* value of agreement is from a Fisher exact test. Vertical green dashed lines indicate the level of significance as given by the Bonferroni threshold in the discovery set. Horizontal green dashed lines indicate a level of significance of *P* = .05 in the replication set.

Using linear regression models, adjusted for bisulfite conversion efficiency, we identified 1501 CpGs whose DNAme β-values correlated with SPY, all passing a Bonferroni threshold of $P < .05/479\,491$ (approximately 10^{-7}) (**Figure 1**A and eTables 1 and 2 and eFigure 5 in the Supplement). Top-ranked CpGs were mostly hypomethylated in smokers (eTable 2 in the Supplement). Changes in DNAme were modest, with the top-ranked CpG (mapping to the gene *AHRR*) showing a 6% decrease for every 10 SPY (Figure 1B and **Table 2**). More than 73% of the 1501 smoking-associated CpGs were validated at the same Bonferroni significance level of 10^{-7} in an independent replication set of 390 buccal samples (Figure 1C and eTable 2 in the Supplement), with all sites exhibiting the same directional changes as in the discovery set (Figure 1D).

Comparison of DNAme Changes Between Buccal and Blood Tissue

The top-ranked CpGs were all hypomethylated with increasing SPY and many mapped to genes previously identified in smoking EWASs conducted in whole-blood tissue (eg, *AHRR, CYP1A1, F2RL3, PTK2, GNG12, GFI1*)[15,18,33-35] (Table 2), suggesting that much of the DNA hypomethylation is common to both tissue types. To investigate this further, we conducted a detailed comparison on a matched subset of 152 women for whom both a blood and buccal sample had been collected at age 53 years. Using only the matched samples from these 152 women, we derived smoking-associated DNAme signatures from the buccal and blood cells. Focusing on the top-ranked CpGs, this revealed consistency and broad agreement between the 2 tissue types, driven mainly by the commonly hypomethylated sites (Table 2 and eFigure 6 in the Supplement).

However, the analysis also revealed significantly many more associations in buccal cells (eFigure 6A in the Supplement). In the case of whole blood, only 38 CpGs passed a false discovery rate threshold of <.05, in contrast to more than 1500 (ie, a 40-fold increase) passing this same threshold in the 152 matched buccal samples. Thus, although the top-ranked CpGs, which were generally hypomethylated, agreed between the 2 tissue types, smoking was associated with a greater proportion of altered CpG sites in buccal cells.

Biological Significance

To assess biological significance, we performed GSEA separately on the 912 hypermethylated and 589 hypomethylated sites of the 1501 differentially methylated CpGs (**Figure 2**). For the hypermethylated sites, the strongest enrichment was attained for genes bivalently marked in human embryonic stem cells, for binding sites of transcription factors implicated in chromatin organization and specification of stem cell identity (RAD21, CTCF, and EZH2),[36-40] and finally also for genes hypermethylated in lung cancer,[41] a cancer strongly linked to tobacco smoke exposure (Figure 2 and eTables 3-5 in the Supplement). The results of GSEA on the hypomethylated sites did not reveal a strong enrichment of bivalently marked genes but instead showed an enrichment of genes overexpressed in a poorly differentiated human papillomavirus–negative subtype of head and neck cancer,[42] a cancer for which smoking is a main risk factor (Figure 2 and eTable 4 in the Supplement). Thus, the fact that the top-ranked enriched biological terms point toward smoking-related cancers strongly supports the biological relevance of our smoking DNAme-based signature.

DNAme-Based Smoking Index Derived From Buccal Cells

Given the GSEA results, we reasoned that smoking-associated DNAme changes in buccal cells might be seen in epithelial can-

Table 2. Change in DNA Methylation β Value Per 10 Smoking Pack-years for Top-Ranked Probes in the Discovery Set[a]

CpG ID	Gene Symbol	Region	Discovery Set, Buccal (n = 400)		Replication Set, Buccal (n = 390)		Whole Blood Subset (n = 152)	
			Δβ	*P* Value[b]	Δβ	*P* Value[b]	Δβ	*P* Value[b]
cg05575921	*AHRR*	Body	−0.06	8×10^{-60}	−0.05	7×10^{-47}	−0.05	1×10^{-27}
cg12101586	*CYP1A1*	TSS1500	−0.03	5×10^{-19}	−0.03	6×10^{-14}	−0.02	.02
cg03636183	*F2RL3*	Body	−0.02	4×10^{-29}	−0.02	2×10^{-25}	−0.02	5×10^{-21}
cg25189904	*GNG12*	TSS1500	−0.04	3×10^{-16}	−0.04	1×10^{-16}	−0.03	2×10^{-7}
cg09935388	*GFI1*	Body	−0.05	7×10^{-15}	−0.04	5×10^{-15}	−0.04	4×10^{-13}
cg13899718	*LASS2*	TSS1500	0.03	9×10^{-9}	0.03	4×10^{-14}	0.01	4×10^{-3}
cg10801607	*SLC3A1*	Body	0.01	4×10^{-13}	0.01	2×10^{-10}	0.01	.03
cg00393487	*RTP1*	Body	−0.01	5×10^{-9}	−0.01	.004	−0.01	9×10^{-3}
cg23485307	*TGFBR2*	Body	0.04	2×10^{-9}	0.04	3×10^{-13}	0.003	.02
cg17619755	*VARS*	Body	0.02	1×10^{-11}	0.02	2×10^{-14}	0.01	1×10^{-3}
cg12075928	*PTK2*	Body	−0.01	7×10^{-10}	−0.01	8×10^{-6}	−0.02	2×10^{-5}
cg24126592	*DGKZ*	Body	0.01	7×10^{-11}	0.01	1×10^{-15}	0.01	.02
cg22782986	*ODZ4*	TSS1500	0.04	2×10^{-10}	0.04	1×10^{-14}	0.01	.02
cg04583842	*BANP*	Body	0.02	1×10^{-11}	0.02	6×10^{-10}	0.02	7×10^{-5}
cg07251887	*RECQL5*	Body	−0.03	6×10^{-13}	−0.03	1×10^{-14}	−0.01	2×10^{-5}
cg15393221	*PRX*	TSS1500	0.02	9×10^{-10}	0.02	2×10^{-14}	0.01	.01
cg20244340	*SLC24A3*	Body	−0.02	4×10^{-9}	−0.01	5×10^{-7}	−0.02	1×10^{-4}
cg02532700	*NCF4*	Body	−0.02	4×10^{-10}	−0.01	1×10^{-7}	−0.02	6×10^{-4}
cg00788739	*TCN2*	TSS1500	0.03	5×10^{-11}	0.03	3×10^{-15}	0.01	.03

[a] Effect sizes are shown for the discovery and replication set, as well as for the matched subset of 152 whole blood samples.

[b] *P* values are from the linear regression between DNA methylation and smoking pack-years using bisulfite conversion efficiency as a covariate.

cers for which smoking is a potent risk factor. To investigate this and to further assess whether the changes are specific to smoke-related cancers, we collected DNAme data from 15 epithelial cancer types, profiled as part of the Cancer Genome Atlas, encompassing more than 5000 samples, some strongly linked to smoking (lung squamous cell carcinoma [LSCC] and lung adenocarcinoma [LUAD]), others for which smoking is a moderate risk factor (esophageal, head and neck, bladder), and others unrelated to smoking (endometrial and breast cancer).[43,44]

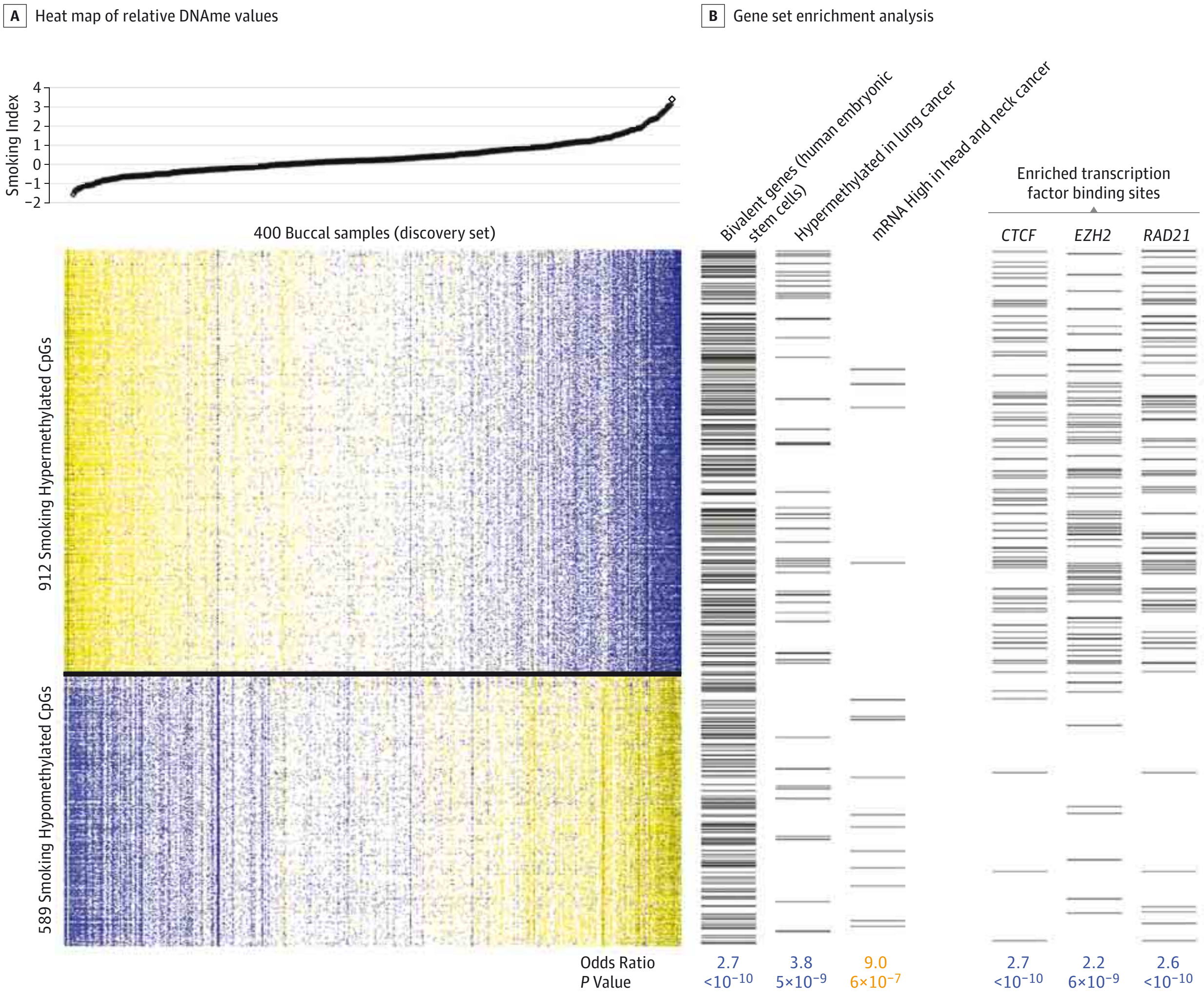
To be able to quantify the similarity of smoking-associated DNAme changes in buccal cells to those in cancer, we constructed a DNAme-based "smoking index," computable for any given independent sample, from the 1501 smoking-associated CpGs of the discovery buccal set. To validate the smoking index, we verified that it correlated significantly ($P < 10^{-10}$) with SPY in the independent replication buccal set (eFigure 7 in the Supplement).

Next, we computed the smoking index in the normal vs cancer data sets. This revealed higher values in cancer compared with its corresponding normal tissue, independently of tissue type (**Figure 3**A and eFigure 8 in the Supplement). Notably, the smoking index was highest for LSCC, a lung cancer with the highest proportion of smokers,[45] followed by LUAD (Figure 3A). Importantly, the smoking index values were similar for 2 independent LUAD data sets, demonstrating the reproducibility of these scores (Figure 3A). All other cancer types (13 in total) exhibited significantly lower index values than the lung cancers (Wilcoxon $P < .01$), in line with smoking being a weaker risk factor for these other cancers (Figure 3A and eFigure 8 in the Supplement). Detailed receiver operating characteristic analyses showed that the smoking index was highly discriminative of normal vs cancer status in every tissue type (Figure 3B and eFigure 8 in the Supplement). Thus, these results indicate that smoking-associated DNAme changes in normal cells are not only present in smoking-associated cancers but also in those cancers for which smoking is not a risk factor, suggesting that other cancer risk factors may be causing similar epigenetic aberrations in normal cells. Similar results were obtained for the smoking signature derived from the 152 buccal samples (Figure 3B and eFigure 8 in the Supplement). In contrast, the signature derived from the matched 152 blood samples and a signature constructed by randomly selecting 1501 CpGs were significantly less accurate (Wilcoxon $P = .001$ [blood] and $P < .001$ [random]) (Figure 3B).

Association of Hypermethylation With Cancer

The GSEA results also suggested to us that the power of the smoking signature to discriminate cancer from normal tissue could be due to the hypermethylated component. Thus, to dissect the relative contribution of hypermethylation and hypomethylation to the smoking index, we recomputed the smoking index in the normal vs cancer sets using 4 different subsets of the 1501 differentially methylated CpGs: (1) all 912 hypermethylated CpGs, (2) all 589 hypomethylated CpGs, (3) the top 50 hypermethylated CpGs, and (4) the top 50 hypomethylated sites. Including a gene set based on the top 50 CpGs allowed us to assess the significance of the CpGs exhibiting the strongest effect sizes. By restricting to the hypermethylated subsets, we observed that the smoking index nearly always increased in cancer, providing high discriminatory power (eFigures 9 and 10 in the Supplement). In contrast,

Figure 2. Gene Set Enrichment Analysis of the Smoking DNA Methylation (DNAme) Signature



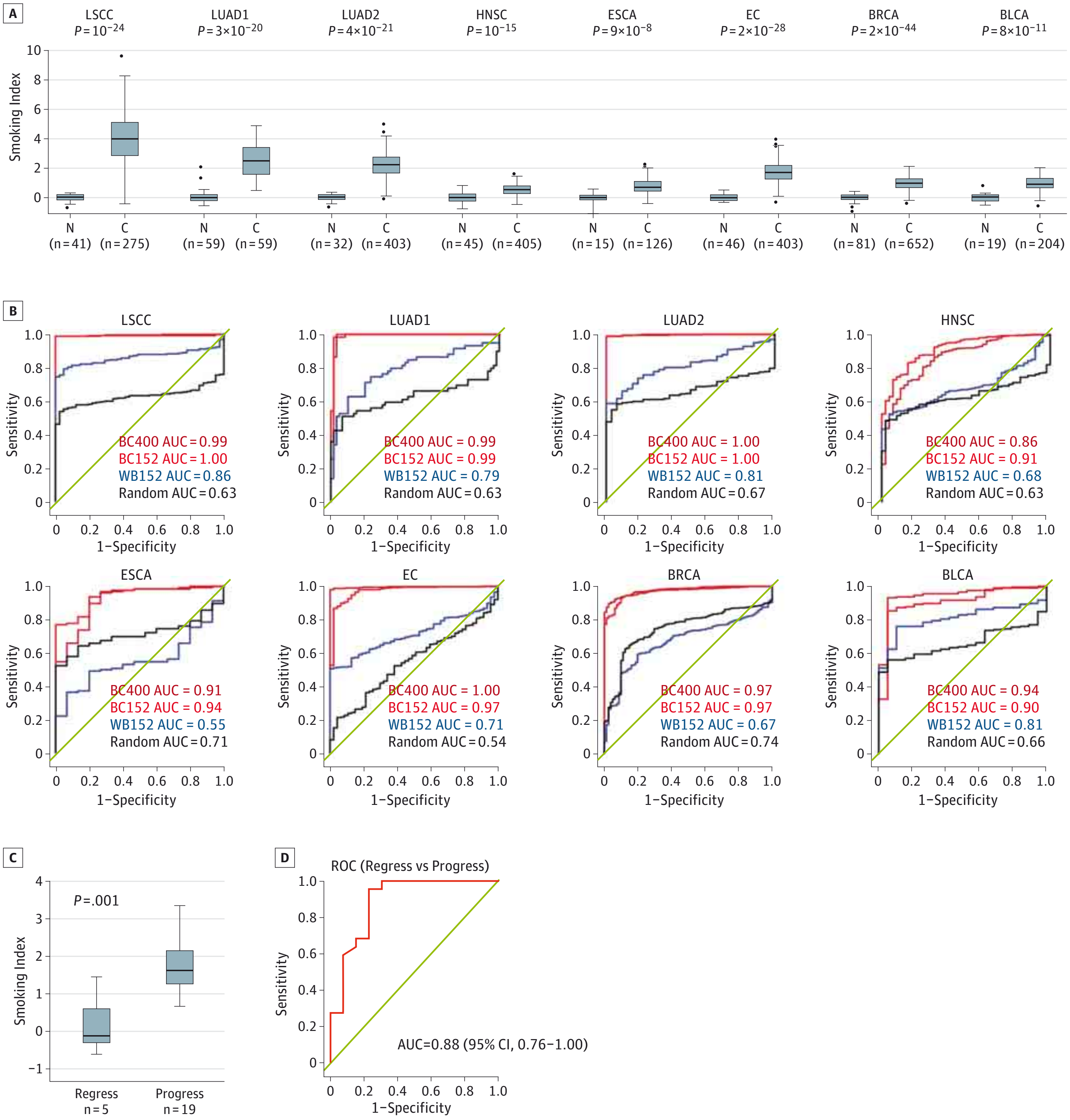
A, Heat map of relative DNAme values (CpG methylation β values were standardized to mean zero and unit variance across the 400 buccal samples) of the 1501 smoking-associated CpGs separated according to hypermethylation and hypomethylation as indicated. The 400 buccal samples were ordered according to the smoking index, a measure of the deviation in DNAme from a normal reference (here buccal cells from nonsmokers). B, Results of the gene set enrichment analysis assessing enrichment of genes or transcription factor binding sites among the hypermethylated and hypomethylated CpG categories. We give the odds ratios and enrichment *P* values (1-tailed Fisher test) of some of the enriched categories. Enrichment was assessed for hypermethylated CpGs and hypomethylated CpGs separately. Blue color indicates enrichment in the hypermethylated class compared with CpGs not associated with smoking; orange indicates enrichment in the hypomethylated class compared with CpGs not associated with smoking.

by restricting to the hypomethylated subsets, we found that the smoking index was less frequently discriminative of normal vs cancer status (eFigures 9 and 10 in the Supplement). Focusing on the top 50 hypomethylated CpGs (which included genes such as *AHRR*, *CYP1A1*), we found that the changes seen in cancer were often random, with some cancer tissues exhibiting directional changes exactly opposite to those seen in smoking (eFigures 9 and 10 in the Supplement).

Smoking Index in Preneoplastic Lesions

The observed correlation between smoking-associated DNAme changes in buccal cells with those seen in epithelial cancers prompted us to explore whether these specific DNAme changes are a consequence of cancer or whether they represent an early event in carcinogenesis. Indeed, the observed overlap between smoking-associated and cancer-associated DNAme changes could be the result of the widespread DNAme changes caused by uncontrolled proliferation of cancer cells. To address this, we asked whether the smoking index is increased in preneoplastic lesions. To this end, we generated Illumina 450k DNAme data for a series of 8 endometrial hyperplasias and 33 endometrial cancers[46] and analyzed these data jointly with the 46 normal endometrial tissue samples and 403 endometrial cancers from the Cancer Genome Atlas.[47] We observed a significant increase in the smoking index between normal tissues and hyperplasia, with a further increase between hyperplasia and cancer (eFigure 11 in the Supplement). We also computed the smoking index in a series of 152 cytologically normal cervical

Figure 3. Smoking Index in Cancer Compared With Normal Tissue



A, Box-and-whisker plots comparing the smoking index of cancers (C) to their respective normal tissue (N) for 8 independent data sets encompassing the following cancers: LSCC (lung squamous cell carcinoma), LUAD1 and LUAD2 (lung adenocarcinoma data sets 1 and 2), HNSC (head and neck squamous carcinoma), ESCA (esophageal carcinoma), EC (endometrial cancer), BRCA (breast cancer), and BLCA (bladder cancer). The number of samples in each category is below the x-axis. *P* values are from a Wilcoxon rank sum test. The smoking index for each sample was computed using the 1501 smoking-associated CpGs derived from the discovery set of 400 buccal samples. The horizontal line in the middle of each box indicates the median, while the top and bottom borders of the box mark the 75th and 25th percentiles, respectively. The horizontal lines above and below the box mark deviations from the median given by 1.5 times the interquartile range. The points beyond these horizontal lines define outliers. B, Corresponding receiver operating curve (ROC) and area under the curve (AUC) analysis for each of the 8 data sets and for the smoking indices derived from the original 400 buccal samples (BC) (brown), the 152 matched buccal samples (orange), the 152 matched blood samples (WB) (blue), and a random 1501-CpG signature (black). C, Comparison of the smoking index (as calculated using the smoking DNA methylation signature derived from buccal cells) of preinvasive lung lesions that regress with those that progress to lung cancer. Wilcoxon rank sum test *P* value is given. D, The ability of the smoking index to discriminate regressors from progressors in an ROC and AUC analysis. The AUC plus 95% confidence interval is given.

smear samples at different risk of neoplastic transformation,[8] revealing an increased smoking index in cells at higher risk of neoplastic transformation (eFigure 12 in the Supplement). Thus, the DNAme changes seen in normal buccal cells may indeed represent early events in carcinogenesis.

Association With Smoking-Associated Gene Expression Changes in Nontumor Lung Tissue

Given the particularly strong association of the smoking index with lung cancers, we next asked whether the 1501 CpGs from our smoking signature are informative of smoking-associated gene expression changes in the normal tissue that gives rise to lung cancer. Specifically, we analyzed gene expression data of 344 nontumor lung tissue samples from smokers and nonsmokers, all of whom developed lung cancer (Bossé et al).[48] For this analysis, we focused on the subset of the 1501 smoking-associated CpGs, which mapped to within 200 bp of the transcription start sites of genes profiled in Bossé et al.[48] We observed that CpGs hypermethylated in buccal cells of smokers exhibited lower levels of expression in the normal lung tissue of patients who were smokers compared with those of patients who did not smoke, whereas hypomethylated CpGs showed significantly higher expression (eFigure 13 and eTable 6 in the Supplement). Importantly, these results were validated in 2 independent but equivalent cohorts totaling 509 (285 + 224) samples (eFigure 13 and eTable 6 in the Supplement). This supports the view that part of the smoking-associated DNAme changes identified in buccal cells, if also present in nontumor lung tissue of smokers, may represent functional changes in a tissue relevant to the development of lung cancer.

Smoking Index and Progression of Preinvasive Lung Lesions

Given the fact that the smoking index is able to discriminate between normal tissue and cancer irrespective of the tissue analyzed, we wanted to test whether this signature is also able to predict the fate of lesions that originate from the same organ. Despite only analyzing 24 samples, we found that the smoking index was able to identify preinvasive lung lesions that subsequently progress to an invasive lung cancer with a high sensitivity and specificity (area under the curve, 0.88 [95% CI, 0.76-1.00]; *P* = .001) (Figure 3C and D).

Discussion

The EWAS presented here has demonstrated the suitability of using buccal cells to examine the effects of smoking on the epigenome. Specifically, our key novel findings are as follows: (1) Smoking is associated with widespread changes in the DNAme landscape of buccal cells, in contrast to blood cells, supporting the view that buccal tissue is a more appropriate source to examine the effects of smoking. (2) Nevertheless, the top-ranked CpG sites, which were overwhelmingly hypomethylated in smokers and which mapped to genes such as *AHRR*, *CYP1A1*, and *CYP1B1* (all involved in toxin response pathways), were common to buccal and blood tissue. (3) Smoking-associated DNAme changes in buccal cells, in particular, hypermethylation of bivalent marks and binding sites of transcription factors implicated in DNA repair (RAD21) and chromatin architecture (CTCF), correlated with DNAme changes in epithelial cancers and did so most strongly in smoke-related cancer, notably lung cancer (eFigures 14 and 15 in the Supplement). This suggests that smoking is associated with disruption of RAD21 binding and hence DNA repair deficiency, which in turn has been associated with increased lung cancer risk.[49] (4) The smoking DNAme signature correlated with an increased risk that a preinvasive lung lesion will progress to lung cancer. This not only opens a new window of opportunities in personalized medicine but also provides additional evidence that alteration of the epigenome is an important early step in cancer development.

One of the most intriguing observations of our analysis is the increased smoking index observed in all epithelial cancers relative to their normal tissue, although in line with our expectations, this index was highest for lung cancers. We note, however, that the correlation with cancer was driven by hypermethylation of sites that are strongly enriched for bivalently marked genes in human embryonic stem cells, which in turn is a common feature of DNAme signatures associated with other cancer risk factors (including aging) and with cancer itself.[10,12,14,32,50-52] We stress that hypomethylated CpG sites could not consistently discriminate normal from cancer tissue, unless we specifically focused on genes previously implicated in cancer, for instance those observed to be overexpressed in human papillomavirus–negative head and neck cancer (a cancer for which smoking is a risk factor). It would thus appear that most of the DNA hypomethylation seen in the buccal and blood cell epigenome of smokers is unrelated to cancer etiology. This is not inconsistent with the observation that global DNA hypomethylation is seen in cancer and preneoplastic lesions,[10] because this widespread hypomethylation has so far only been seen in cells that have already undergone morphological transformation. In contrast, our buccal DNAme signature was derived from entirely normal cells exposed to different levels of a carcinogen, so the DNA hypomethylation that we observe in these normal cells could reflect a different underlying mechanism such as a specific response to smoke toxins.

We end by discussing 2 potential limitations of this study. First, the signature was derived from women only. Thus, whether the smoking DNAme signature would change substantially had we used a male population is unclear. Previous EWASs in blood suggest, however, that most smoking-related changes are independent of sex.[15,18] A second limitation is that we did not analyze any functional or expression data in the same buccal samples. However, we did analyze gene expression data from nontumor lung tissue, comparing expression levels of genes implicated by our DNAme signature between smokers and nonsmokers, providing evidence that specific DNAme changes in a relevant cell of origin may indeed be functional.

Conclusions

This study has demonstrated that smoking is associated with a widespread alteration of the DNAme landscape in buccal cells but not so in blood tissue. The DNAme alterations seen in buccal tissue may be important for the etiology of specific epithelial cancers and the fate of preinvasive lesions.

ARTICLE INFORMATION

Accepted for Publication: March 23, 2015.

Published Online: May 14, 2015. doi:10.1001/jamaoncol.2015.1053.

Author Affiliations: Statistical Genomics Group at the University College London Cancer Institute, London, United Kingdom (Teschendorff); CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai Institute for Biological Sciences, Shanghai, China (Teschendorff, Yang, Jiao); Medical Research Council Unit for Lifelong Health and Ageing at University College London, London, United Kingdom (Wong, Hardy, Kuh); University College London Cancer Institute, University College London, London, United Kingdom (Pipinikas, Thirlwell); Department of Women's Cancer, University College London Elizabeth Garrett Anderson Institute for Women's Health, London, United Kingdom (Jones, Anjum, Widschwendter); Department of Obstetrics and Gynaecology, Haukeland University Hospital, Bergen, Norway (Salvesen); Lungs for Living Research Centre, Division of Medicine, University College London, London, United Kingdom (Janes).

Author Contributions: Drs Teschendorff and Wong had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Teschendorff, Yang, and Wong served as co–first authors, each with equal contribution to the manuscript.
Study concept and design: Teschendorff, Widschwendter.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: Teschendorff, Widschwendter.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Teschendorff, Yang, Pipinikas, Jiao, Anjum.
Obtained funding: Teschendorff, Kuh, Janes, Widschwendter.
Administrative, technical, or material support: Wong, Pipinikas, Jones, Salvesen, Janes, Widschwendter.
Study supervision: Teschendorff, Thirlwell, Janes, Widschwendter.

Conflict of Interest Disclosures: None reported.

Funding/Support: This work was funded by the Eve Appeal (https://www.eveappeal.org.uk/) and the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement 305428 (Project EpiFemCare) and was done at University College London Hospital/University College London, which received a proportion of its funding from the Department of Health National Institute for Health Research Biomedical Research Centres funding scheme. Drs Teschendorff and Yang and Mr Jiao are supported by the Chinese Academy of Sciences, the Shanghai Institute for Biological Sciences, and the Max-Planck-Gesellschaft. Drs Wong, Hardy, and Kuh are supported by the Medical Research Council (MC_UU_12019/1). Dr Janes is a Wellcome Trust Senior Fellow in Clinical Science (WT091730MA) and Drs Pipinikas and Janes were supported by the Rosetrees Trust, Roy Castle Lung Cancer Foundation, and University College London Hospital Charitable Foundation. Dr Thirlwell is a Cancer Research UK funded Clinician Scientist. Dr Widschwendter was supported by the Eve Appeal and the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement 305428 (Project EpiFemCare).

Role of the Funder/Sponsor: The funding bodies had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Correction: This article was corrected online May 21, 2015, for an error in the abstract Results.

REFERENCES

1. Thun MJ, Carter BD, Feskanich D, et al. 50-year trends in smoking-related mortality in the United States. *N Engl J Med*. 2013;368(4):351-364.

2. Jha P, Ramasundarahettige C, Landsman V, et al. 21st-century hazards of smoking and benefits of cessation in the United States. *N Engl J Med*. 2013;368(4):341-350.

3. Jha P, Peto R. Global effects of smoking, of quitting, and of taxing tobacco. *N Engl J Med*. 2014;370(1):60-68.

4. Giovino GA, Mirza SA, Samet JM, et al; GATS Collaborative Group. Tobacco use in 3 billion individuals from 16 countries: an analysis of nationally representative cross-sectional household surveys. *Lancet*. 2012;380(9842):668-679.

5. Been JV, Nurmatov UB, Cox B, Nawrot TS, van Schayck CP, Sheikh A. Effect of smoke-free legislation on perinatal and child health: a systematic review and meta-analysis. *Lancet*. 2014;383(9928):1549-1560.

6. Sitas F, Egger S, Bradshaw D, et al. Differences among the coloured, white, black, and other South African populations in smoking-attributed mortality at ages 35-74 years: a case-control study of 481,640 deaths. *Lancet*. 2013;382(9893):685-693.

7. Zhuang J, Jones A, Lee SH, et al. The dynamics and prognostic potential of DNA methylation changes at stem cell gene loci in women's cancer. *PLoS Genet*. 2012;8(2):e1002517.

8. 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*. 2012;4(3):24.

9. Jones A, Teschendorff AE, Li Q, et al. Role of DNA methylation and epigenetic silencing of HAND2 in endometrial cancer development. *PLoS Med*. 2013;10(11):e1001551.

10. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. *Nat Rev Genet*. 2006;7(1):21-33.

11. Hansen KD, Sabunciyan S, Langmead B, et al. Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization. *Genome Res*. 2014;24(2):177-184.

12. Issa JP, Ahuja N, Toyota M, Bronner MP, Brentnall TA. Accelerated age-related CpG island methylation in ulcerative colitis. *Cancer Res*. 2001;61(9):3573-3577.

13. Ahuja N, Li Q, Mohan AL, Baylin SB, Issa JP. Aging and DNA methylation in colorectal mucosa and cancer. *Cancer Res*. 1998;58(23):5489-5494.

14. Ahuja N, Issa JP. Aging, methylation and cancer. *Histol Histopathol*. 2000;15(3):835-842.

15. 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*. 2013;22(5):843-851.

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

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

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

19. 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(4):450-457.

20. Medical Research Council National Survey of Health and Development website. http://www.nshd.mrc.ac.uk. Accessed April 9, 2015.

21. Kuh D, Pierce M, Adams J, et al; NSHD Scientific and Data Collection Team. Cohort profile: updating the cohort profile for the MRC National Survey of Health and Development: a new clinic-based data collection for ageing research. *Int J Epidemiol*. 2011;40(1):e1-e9.

22. Wadsworth M, Kuh D, Richards M, Hardy R. Cohort profile: the 1946 National Birth Cohort (MRC National Survey of Health and Development). *Int J Epidemiol*. 2006;35(1):49-54.

23. Rousseau K, Vinall LE, Butterworth SL, et al. MUC7 haplotype analysis: results from a longitudinal birth cohort support protective effect of the MUC7*5 allele on respiratory function. *Ann Hum Genet*. 2006;70(pt 4):417-427.

24. Banerjee AK, Rabbitts PH, George PJ. Preinvasive bronchial lesions: surveillance or intervention? *Chest*. 2004;125(5)(suppl):95S-96S.

25. McCaughan F, Pipinikas CP, Janes SM, George PJ, Rabbitts PH, Dear PH. Genomic evidence of pre-invasive clonal expansion, dispersal and progression in bronchial dysplasia. *J Pathol*. 2011;224(2):153-159.

26. Sandoval J, Heyn H, Moran S, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*. 2011;6(6):692-702.

27. Ma X, Wang YW, Zhang MQ, Gazdar AF. DNA methylation data analysis and its application to cancer research. *Epigenomics*. 2013;5(3):301-316.

28. Maksimovic J, Gordon L, Oshlack A. SWAN: subset-quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol*. 2012;13(6):R44.

29. 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*. 2013;29(2):189-196.

30. Teschendorff AE, Menon U, Gentry-Maharaj A, et al. An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS One*. 2009;4(12):e8274.

31. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550.

32. Teschendorff AE, Menon U, Gentry-Maharaj A, et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res*. 2010;20(4):440-446.

33. Zeilinger S, Kühnel B, Klopp N, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS One*. 2013;8(5):e63812.

34. Shenker NS, Ueland PM, Polidoro S, et al. DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology*. 2013;24(5):712-716.

35. Zhang Y, Yang R, Burwinkel B, et al. F2RL3 methylation in blood DNA is a strong predictor of mortality. *Int J Epidemiol*. 2014;43(4):1215-1225.

36. Nitzsche A, Paszkowski-Rogacz M, Matarese F, et al. RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. *PLoS One*. 2011;6(5):e19470.

37. Merkenschlager M, Odom DT. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell*. 2013;152(6):1285-1297.

38. van Vlerken LE, Kiefer CM, Morehouse C, et al. EZH2 is required for breast and pancreatic cancer stem cell maintenance and can be used as a functional cancer stem cell reporter. *Stem Cells Transl Med*. 2013;2(1):43-52.

39. Suvà ML, Riggi N, Janiszewska M, et al. EZH2 is essential for glioblastoma cancer stem cell maintenance. *Cancer Res*. 2009;69(24):9211-9218.

40. Kamminga LM, Bystrykh LV, de Boer A, et al. The polycomb group gene Ezh2 prevents hematopoietic stem cell exhaustion. *Blood*. 2006;107(5):2170-2179.

41. Fukasawa M, Kimura M, Morita S, et al. Microarray analysis of promoter methylation in lung cancers. *J Hum Genet*. 2006;51(4):368-374.

42. Rickman DS, Millon R, De Reynies A, et al. Prediction of future metastasis and molecular characterization of head and neck squamous-cell carcinoma based on transcriptome and genome analysis by microarrays. *Oncogene*. 2008;27(51):6607-6622.

43. Viswanathan AN, Feskanich D, De Vivo I, et al. Smoking and the risk of endometrial cancer: results from the Nurses' Health Study. *Int J Cancer*. 2005;114(6):996-1001.

44. Gaudet MM, Gapstur SM, Sun J, Diver WR, Hannan LM, Thun MJ. Active smoking and breast cancer risk: original cohort data and meta-analysis. *J Natl Cancer Inst*. 2013;105(8):515-525.

45. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers [published correction appears in *Nature*. 2012 Nov 8;491(7423):288]. *Nature*. 2012;489(7417):519-525.

46. Hoivik EA, Kusonmano K, Halle MK, et al. Hypomethylation of the CTCFL/BORIS promoter and aberrant expression during endometrial cancer progression suggests a role as an Epi-driver gene. *Oncotarget*. 2014;5(4):1052-1061.

47. Kandoth C, Schultz N, Cherniack AD, et al; Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature*. 2013;497(7447):67-73.

48. Bossé Y, Postma DS, Sin DD, et al. Molecular signature of smoking in human lung tissues. *Cancer Res*. 2012;72(15):3753-3763.

49. Luo S, Sun M, Jiang R, Wang G, Zhang X. Establishment of primary mouse lung adenocarcinoma cell culture. *Oncol Lett*. 2011;2(4):629-632.

50. Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? *Hum Mol Genet*. 2013;22(R1):R7-R15.

51. Widschwendter M, Fiegl H, Egle D, et al. Epigenetic stem cell signature in cancer. *Nat Genet*. 2007;39(2):157-158.

52. Maegawa S, Hinkal G, Kim HS, et al. Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Res*. 2010;20(3):332-340.