Human Molecular Genetics, 2015, Vol. 24, No. 8 2349–2359

doi: 10.1093/hmg/ddu751
Advance Access Publication Date: 2 January 2015
Original Article

ORIGINAL ARTICLE

Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation

Florence Guida[1,†], Torkjel M. Sandanger[2,†], Raphaële Castagné[1,†], Gianluca Campanella[1], Silvia Polidoro[3], Domenico Palli[4], Vittorio Krogh[5], Rosario Tumino[6], Carlotta Sacerdote[3], Salvatore Panico[7], Gianluca Severi[3,8,9], Soterios A. Kyrtopoulos[10], Panagiotis Georgiadis[10], Roel C.H. Vermeulen[1,11,12], Eiliv Lund[2], Paolo Vineis[1,3,‡] and Marc Chadeau-Hyam[1,11,‡,*]

[1]MRC-PHE Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, Norfolk Place, London W2 1PG, UK, [2]Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway, [3]HuGeF, Human Genetics Foundation, Torino, Italy, [4]Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Institute—ISPO, Florence, Italy, [5]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy, [6]Ragusa Cancer Registry Azienda Ospedaliera "Civile M.P. Arezzo," Ragusa, Italy, [7]Department of Clinical Medicine and Surgery, Federico II University, Naples, Italy, [8]Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Australia, [9]Centre for Epidemiology and Biostatistics, University of Melbourne, Melbourne, Australia, [10]Institute of Biology, Pharmaceutical Chemistry and Biotechnology, National Hellenic Research Foundation, Athens, Greece, [11]Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands and [12]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

*To whom correspondence should be addressed at: MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, UK. Tel: +44 2075941637; Fax: +44 2074022150; Email: m.chadeau@imperial.ac.uk

Abstract

Several studies have recently identified strong epigenetic signals related to tobacco smoking. However, an aspect that did not receive much attention is the evolution of epigenetic changes with time since smoking cessation. We conducted a series of epigenome-wide association studies to capture the dynamics of smoking-induced epigenetic changes after smoking cessation, using genome-wide methylation profiles obtained from blood samples in 745 women from 2 European populations. Two distinct classes of CpG sites were identified: sites whose methylation reverts to levels typical of never smokers within decades after smoking cessation, and sites remaining differentially methylated, even more than 35 years after smoking cessation. Our results suggest that the dynamics of methylation changes following smoking cessation are driven by a differential and site-specific magnitude of the smoking-induced alterations (with persistent sites being most affected) irrespective of the intensity and duration of smoking. Analyses of the link between methylation and expression levels revealed that methylation predominantly and remotely down-regulates gene expression. Among genes whose expression was associated with our candidate CpG sites, LRRN3 appeared to be particularly interesting as it was one of the few genes whose methylation and

† The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.
‡ The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.
Received: October 23, 2014. **Revised:** December 23, 2014. **Accepted:** December 30, 2014

expression were directly associated, and the only gene in which both methylation and gene expression were found associated with smoking. Our study highlights persistent epigenetic markers of smoking, which can potentially be detected decades after cessation. Such historical signatures are promising biomarkers to refine individual risk profiling of smoking-induced chronic disease such as lung cancer.

Introduction

Smoking is a leading cause of death worldwide (1,2) and has been identified as a major risk factor for several diseases including cancer (3,4), cardiovascular (5,6) and respiratory diseases (7,8). Tobacco is one of the most potent carcinogenic mixtures, and the carcinogenic effect of tobacco smoking can persist for decades after smoking cessation. The risk of developing lung cancer remains much higher in former smokers compared with never smokers, even 11–20 years after smoking cessation (9), and the duration of smoking has been found to have a greater effect than intensity on health outcomes (10,11).

To better characterize the dynamics of smoking-related biological effects, and to assess their impact on subsequent disease risk (12), the development of biologically relevant long-term markers of tobacco smoke exposure is crucial. This includes the exploration of the dynamics of the biomarker changes over time after exposure cessation and specifically the estimation of the time period during which biomarker levels remain altered.

While some biomarkers of exposure to tobacco smoke have been well-established (e.g. cotinine levels in blood, urine or saliva), they have so far failed to identify the effects of past exposures. The identification of a long-term biomarker, such as changes in DNA methylation, that measures exposures decades prior to biosample collection, constitutes a great leap forward in the study of exposure-induced risk of chronic diseases.

Over the recent years, several studies have investigated epigenetic changes relating to exposure to tobacco smoke. These studies developed an Epigenome-Wide Association Study (EpWAS) approach to identify CpG sites associated with smoking status (13–19). Few studies have, however, reported on the dynamics following cessation, with one study reporting the potential effect modification of time since smoking cessation (20), showing that DNA methylation levels in former smokers approached those of never smokers several years after smoking cessation. More recently, one investigation formally assessed the relationship between methylation levels in five preselected smoking-related CpG loci on *F2RL3* and smoking cessation (21). One EpWAS reported three differentially methylated sites between first and last quartiles of time since smoking cessation (22).

The present study represents one of the largest EpWAS for smoking-related methylation alterations including ($N = 745$) blood samples from two independent European populations (the Italian EPIC-Italy cohort and the Norwegian NOWAC cohort), using the Illumina HumanMethylation450 BeadChips array. In addition to investigating epigenetic signatures reflecting smoking status, our study also constitutes the first agnostic investigation of the dynamics of methylation changes after smoking cessation. To preserve power, we propose a novel strategy relying on a binary recoding of the smoking status as a function of time since smoking cessation, to identify the time taken for potential epigenetic smoking signatures to disappear. In addition, we use genome-wide gene expression data, measured in the Norwegian component of our study to identify transcripts potentially associated with our CpG sites of interest.

Results

Epigenome-wide markers of smoking status

The EpWAS comparing methylation levels in current smokers to those in never smokers revealed 461 significant associations (Fig. 1 and Supplementary Material, Table S1). Of these, the vast majority (448 CpG sites) were hypomethylated, in particular 3 different loci on the *AHRR* gene (chromosome 5) (*P*-values from 10^{-106} to 10^{-14}), 4 CpG sites in a non-annotated region on chromosome 2 (2q37.1), 5 sites on chromosome 1 (including 4 sites on *GFI1* and 1 on *GNG12*), 1 CpG site on chromosome 19 (*F2RL3*), 1 CpG site in a non-annotated region on chromosome 6 (6p21.33) and 3 CpG sites on chromosome 11 (*KCNQ1OT1*), 12 (*RARG*) and 16 (*ADCY9*), respectively. The two strongest hypermethylated CpG sites were located on the *MYO1G* (chromosome 7).

The three differentially hypomethylated CpG sites that were found significant in the former-to-never smokers comparisons were all located on the same non-annotated region of chromosome 2 (Table 1) and were also identified in the current-to-never smoker analyses. The distribution of their methylation level as a function of time since smoking cessation (Supplementary Material, Fig. S1) shows strong hypomethylation close to smoking cessation, and subsequent reversion to methylation levels that are typical of never smokers.

Dynamics of DNA methylation following smoking cessation

To investigate further methylation changes after smoking cessation and characterize, at the whole methylome level, their dynamics, we ran a set of additional EpWAS using a recoded binary indicator for smoking status. As detailed in the Materials and Methods section, for each value of t (time since smoking cessation) investigated (from 0 to 45 years), we compared the genome-wide methylation profiles in recoded 'smokers' (current and former smokers having quit $< t$ years ago) and in 'non-smokers' (never smokers and former smokers having quit $\geq t$). Sample sizes and corresponding number of significant associations found at a Bonferroni 5% level for each value of t are summarized in Supplementary Material, Figures S2 and S3, respectively. Supplementary Material, Figure S3 highlights two distinct phases: (i) a decrease in the number of associations for values of t ranging from 0 to 35 years and (ii) a levelling-off of the number of associations (for $t \geq 35$). We found 751 CpG sites associated with the recoded smoking status at least once across the 46 values of t, and we report in Figure 2 the evolution of their *P*-values as a function of t. Two classes of CpG sites clearly emerge: (i) reversible sites losing statistical significance after a certain time since smoking cessation and (ii) persistent sites which remain differentially methylated even >35 years after smoking cessation.

Classification of the markers

To obtain an objective classification of our 751 candidate CpG sites with respect to the value of t at which they lose statistical

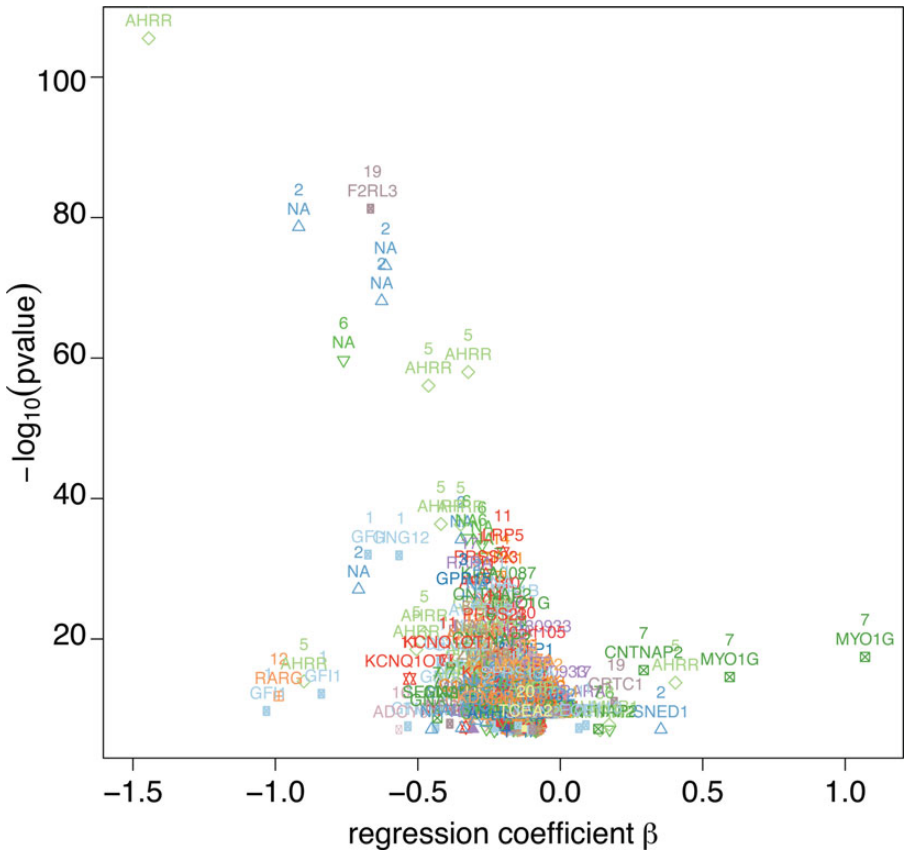

Figure 1. Description of CpG sites significantly associated with smoking status. Summary of the significant associations found at a Bonferroni 5% level (significant threshold is $0.05/432\,414 = 1.16 \times 10^{-7}$). Each significant association is represented by its strength as measured by its *P*-value (*Y*-axis on the $-\log_{10}$ scale), and its effect size (*X*-axis, linear regression coefficient β). Associations are colour-coded with reference to the chromosome the CpG sites are located on. Results are presented for the analyses comparing current with never smokers.

Table 1. List of CpG sites found differentially methylated in former smokers when compared with never smokers

Probe name	Chromosome	Position	Gene name	Gene region	CGI region	Former versus never	
						β	*P*-value
cg06644428	2	233 284 112	NA	NA	Island	−0.40	1.63×10^{-10}
cg05951221	2	233 284 402	NA	NA	Island	−0.22	1.12×10^{-11}
cg21566642	2	233 284 661	NA	NA	Island	−0.28	1.09×10^{-10}

significance, we ran an automatic clustering procedure allocating CpG sites to any of the ($N = 2$) *a priori* defined clusters. As detailed in the Materials and Methods section, our procedure was not only based on a cut-off value for the time since quitting but accounted for the full significance history across the 46 values of t investigated. As indicated in Supplementary Material, Table S2, one class combined ($N = 602$) reversible CpG sites whose methylation reverts back to that of never smokers from up to 35 years after smoking cessation. The second cluster comprised ($N = 149$, Supplementary Material, Table S3) persistent CpG sites remaining differentially methylated >35 years after smoking. Visual inspection of the average methylation levels in never and current smokers at these 751 CpG sites (Fig. 3) highlights a homogeneous distribution of the methylation levels across both classes of probes. Distribution of the methylation levels at our smoking-related CpG sites, irrespective of their reversible or persistent nature, is different from the typical bimodal distribution obtained over the whole epigenome, with a single mode at mid-to-high levels of methylation. In addition, as depicted in Supplementary Material, Figure S4, we find that the largest absolute (Supplementary Material, Fig. S4A) and relative (Supplementary Material, Fig. S4B) current-to-never smokers' methylation differences are observed for persistent CpG sites, which exhibits a right-shifted distribution compared with that obtained from reversible sites. This suggests that irrespective of the methylation level in never smokers, the magnitude of the changes in methylation levels following exposure to tobacco smoke is site specific.

The potential role of dose in the magnitude of the per-site methylation changes was investigated in current smokers ($N = 177$) by running an EpWAS for two smoking exposure metrics. In Figure 4A and B, we report, for our 751 reversible and persistent CpG sites, the funnel plots for both smoking intensity

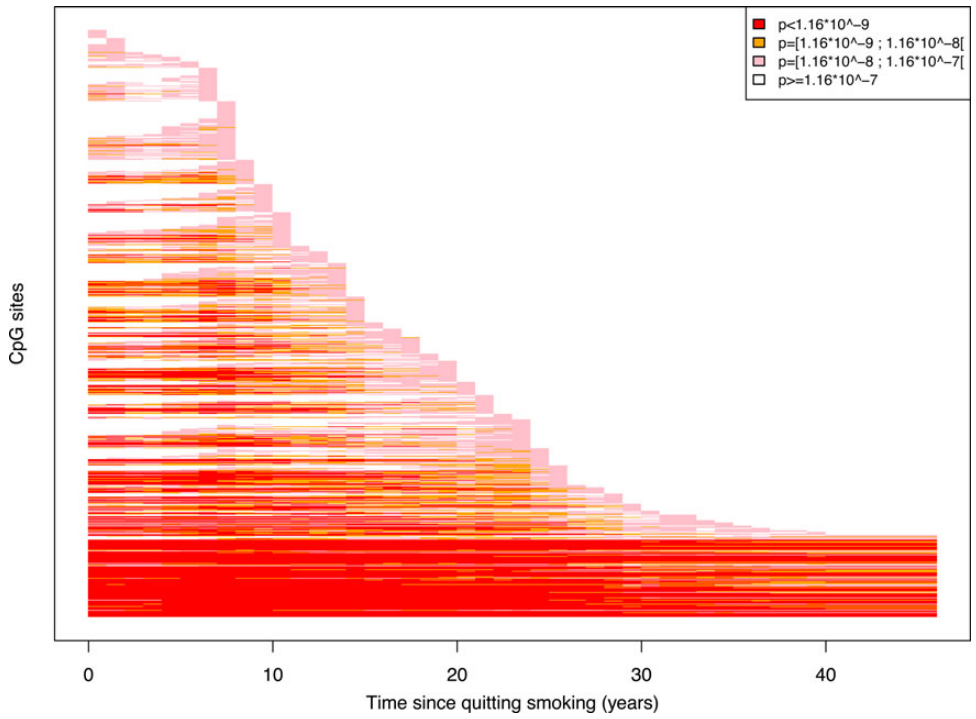

Figure 2. Evolution of the strength of association between methylation level and dichotomized smoking status. Only *P*-value for probes found at least once differentially methylated between smokers and non-smokers across the >40 models (one for each *t*) are plotted ($N = 751$ probes, represented in lines). The colour of each segment indicates the strength of association between methylation level and binary smoking status found for the dataset corresponding to the given *t*. For clarity, *P*-values greater than the Bonferroni-corrected threshold are omitted from the plot.

and smoking duration at the time of blood collection, respectively. Our analyses identified a single genome-wide significant association linking methylation at cg05575921 (*AHRR*, $\beta = 5.93 \times 10^{-2}$, *P*-value $= 1.64 \times 10^{-8}$) and smoking intensity, and none for smoking duration (minimal *P*-value $= 4.6 \times 10^{-4}$). These analyses across our 751 CpG sites also suggest an over-representation of inverse relations (negative regression coefficients) between methylation and smoking intensity (94%), and to a lesser extent, for smoking duration (56%). As expected, irrespective of the smoking metric considered, the strongest effect size estimates correspond to the strongest associations, and overall weaker associations are observed with smoking duration. While the strongest associations corresponded to persistent CpG sites, there is no clear discrimination between the two classes of sites with respect to their association with either smoking intensity or smoking duration.

Genomic location of candidate CpG sites

To attempt describing the two classes of CpG sites, we first investigated their distribution across the genomic regions and position on the CpG island (Supplementary Material, Figs S5 and S6, respectively). We observed a slight under-representation of persistent sites in transcription start sites (TSS regions) compared with the proportion observed over all sites assayed in the array, and a corresponding over-representation of persistent sites in gene bodies. Reversible sites were found under-represented in intergenic regions and over-represented both in gene bodies and in the 5′UTR region. Physical repartition within the island (Supplementary Material, Fig. S6) shows an over-representation of CpG sites on shores for both classes (39 and 47% for persistent and reversible sites versus 23% over all probes assayed in the chip) and a corresponding lower proportion of sites in CpG islands.

Investigation of the CpG-transcript pairs

To explore the relationship between our 751 candidate CpG sites and transcriptomic profiles ($N = 8952$ transcripts, see Materials and Methods) available in NOWAC, we assessed the association of the $751 \times 8952 = 6\,722\,952$ CpG–transcript pairs. We identified 5636 significant CpG–transcript pairs involving 426 unique transcripts and 265 unique CpG sites.

Supplementary Material, Figure S7A and B shows that the expression of the vast majority (88%) of genes was inversely associated with methylation levels. Only 41 and 10 CpG sites were found to up-regulate the expression of at least 1 gene in the reversible and persistent groups, respectively. As a result, the distribution of the gene expression across the 426 transcripts involved in the candidate pairs mirrors trends observed for methylation levels at the candidate CpG sites (Fig. 5): gene expression is found clearly up-regulated (especially for transcripts associated with persistent CpG sites) in current smokers, and this is gradually attenuated in former and never smokers.

Detailed investigation of our candidate CpG–transcript pairs (Supplementary Material, Table S4) showed that 23 of the 265 CpG–transcript associations involved *LRRN3* as the most associated transcript. Only five CpG–transcript pairs involved CpG sites and transcripts relating to the same gene: one pair for *PP1R15A*, *AMICA1*, *RUNX3* genes and two pairs for *LRRN3* (2.01×10^{-9}, 2.7×10^{-11}). *LRRN3* was also found significantly overexpressed in current smokers, when compared with never smokers (fold change: 2.85, *P*-value: 2.1×10^{-24}). The expression of only one additional gene, *FOXO3*, was found up-regulated in current smokers (fold change: 1.27, *P*-value: 4.3×10^{-6}), and no CpG–transcript pairs involving *FOXO3* were identified.

As summarized in Table 2, across the full set ($N = 26$) of CpG–transcript pairs involving *LRRN3* transcript, 20 pairs involved persistent CpG sites.

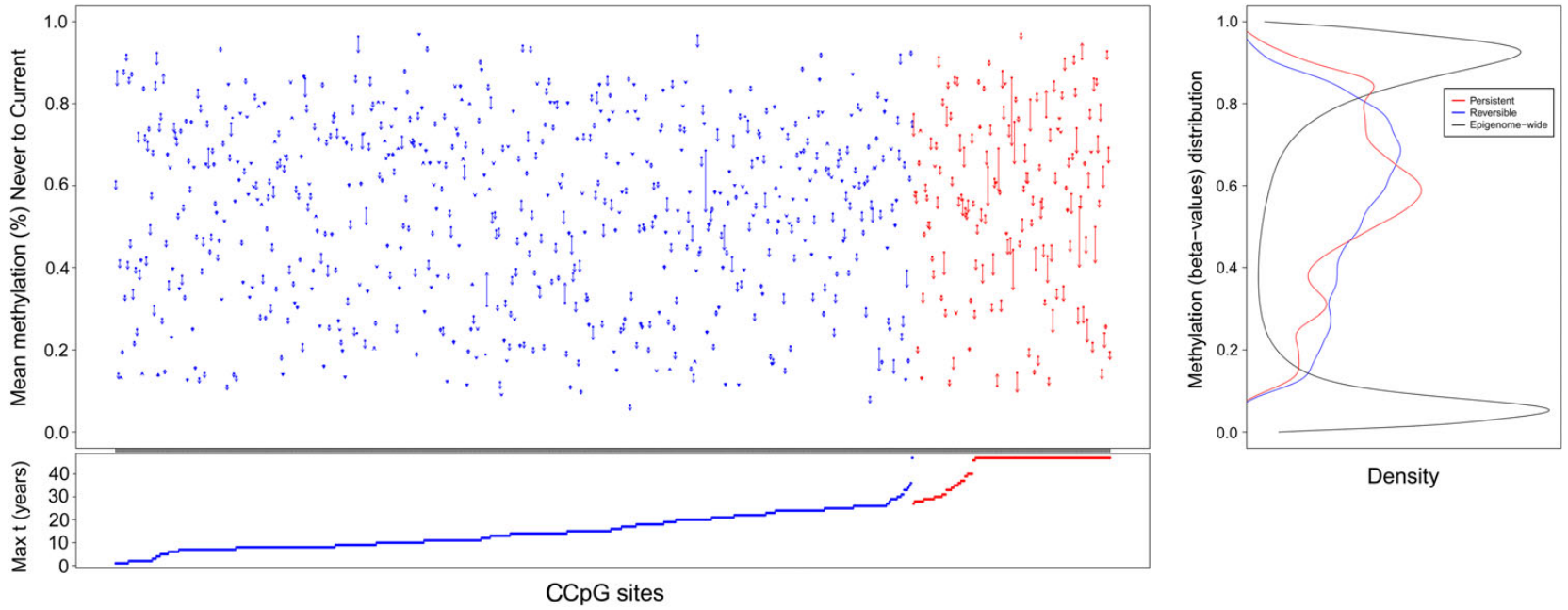

Figure 3. Average methylation levels from never smokers to current smokers. Average methylation levels (%) in never smokers (bullet point) and current smokers (arrowhead) for each of the 751 probes found at least once differentially methylated between smokers and non-smokers using the dichotomized smoking status. Reversible probes are presented in blue and persistent probes in red. For clarity, probes (*X*-axis) are ordered, within each classes, with respect to the largest time since quitting smoking (Max *t*, bottom panel) at which they were found significant. The top right hand plot depicts the methylation level (β-values) across (i) all probes assayed in the Illumina HumanMethylation450 BeadChips array (black line), (ii) all reversible probes (blue line) and (iii) all persistent probes (red line).

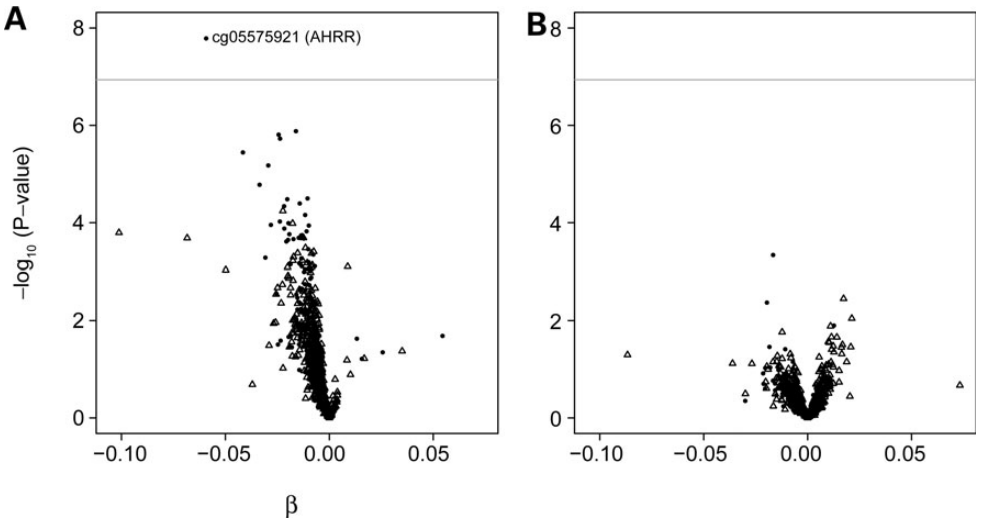

Figure 4. Funnel plots summarizing the associations between methylation level and smoking intensity (**A**) and smoking duration (**B**) in current smokers. Results are presented for 751 CpG sites classified either as reversible (triangles points) or persistent (bullet points) CpG loci. Each plot represents the effect size estimates (*X*-axis) as a function of the $-\log_{10}$ (*P*-values) measuring the strength of association between per-site methylation level and the smoking exposure metric.

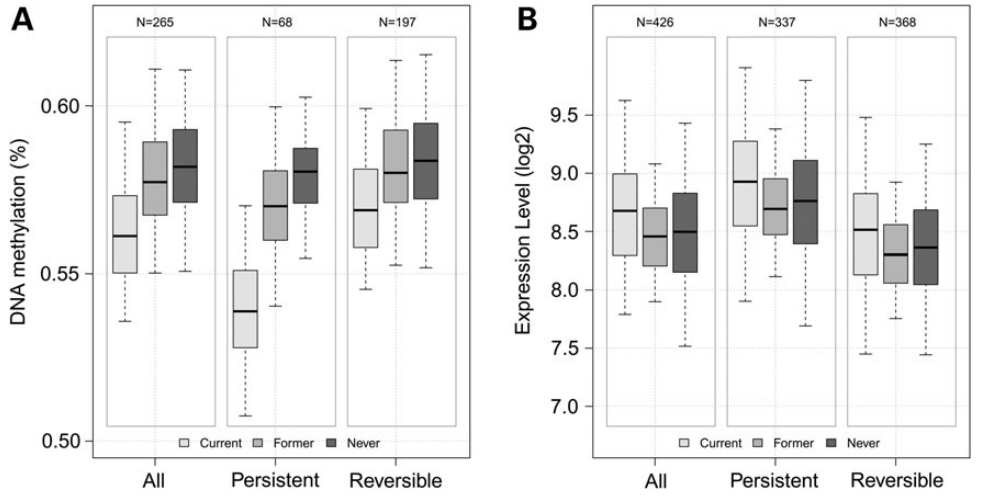

Figure 5. Distribution of the methylation fraction of the CpG site/transcripts pairs associated with smoking exposure. The distribution of the methylation fraction in the 751 CpG sites involved in these pairs is presented by smoking status for the full set of probes and for reversible and persistent probes separately (**A**). The distribution of the $\log_2$-transformed gene expression is presented for the 474 unique transcripts involved in these pairs by smoking status (**B**).

Discussion

Our primary EpWAS identified numerous CpG sites differentially methylated (mostly hypomethylated) in current smokers, when compared with never smokers. The strongest associations involved several CpG sites previously reported: on the *AHRR* gene (14,15,18–20), on the region 2q37.1 of chromosome 2 (14,15,18–20), on the region 6p21.33 on chromosome 6 (14,15,19,20), on the *F2RL3* gene (13–15,17–22) and on the *GFI1* gene (14,19,20). Hypomethylation related to maternal smoking at CpG sites on AHRR (23,24) and GFI1 (24) was recently confirmed in new-born cord blood. We found several hypermethylated CpG sites, including three sites on *MYO1G*, as already reported in several studies (14,19,20,24), and two sites on *CNTNAP2*, which lacks consistent evidence in the literature (18,20–22,24).

Few significant associations were found in the former-to-never analyses, and all of them were also found in the current-versus-never comparisons. To characterize the dynamics of the methylation changes after smoking cessation, we ran series of epigenome-wide analyses using a recoded binary indicator for smoking status. We identified two homogeneous classes of smoking biomarkers: the reversible markers whose methylation reverts back to the levels of non-smokers several years (from 0 to 35) after quitting smoking and the persistent markers whose methylation level remains altered >35 years after quitting smoking. In contrast to what was previously reported (20), the exploration of the main features of these two classes showed that the magnitude of the methylation changes in persistent CpG sites were much higher (~2-fold) than in reversible CpG sites. This indicates that the permanence of methylation alterations for persistent CpG sites is more likely to be attributed to a higher smoking-related change in methylation rather than to a slower reversion rate (towards levels characteristic of never smokers). This interpretation is further supported by an additional regression model linking, in former smokers, methylation levels and time since smoking cessation. Results show a marked

Table 2. Description of the CpG–transcript pairs involving LRRN3 transcript (Chr 7, NM_001099658)

CpG name	CHR	Position	Gene	β	P-value	Cluster
cg25189904	1	68 299 493	*GNG12*	−0.78	2.61×10^{-9}	Persistent
cg08884752	1	2 162 001	*SKI*	−1.18	2.46×10^{-9}	Reversible
cg05951221	2	233 284 402	*NA*	−1.18	1.82×10^{-15}	Persistent
cg21566642	2	233 284 661	*NA*	−0.94	1.30×10^{-18}	Persistent
cg01940273	2	233 284 934	*NA*	−1.23	3.62×10^{-15}	Persistent
cg00295485	2	106 755 721	*UXS1*	−1.44	8.32×10^{-11}	Persistent
cg00501876	3	39 193 251	*CSRNP1*	−1.69	1.06×10^{-9}	Persistent
cg05575921	5	373 378	*AHRR*	−0.64	6.69×10^{-21}	Persistent
cg26703534	5	377 358	*AHRR*	−1.53	4.84×10^{-10}	Persistent
cg14817490	5	392 920	*AHRR*	−1.09	1.95×10^{-11}	Persistent
cg17287155	5	393 347	*AHRR*	−1.20	7.85×10^{-23}	Persistent
cg04551776	5	393 366	*AHRR*	−1.85	1.64×10^{-19}	Persistent
cg21161138	5	399 360	*AHRR*	−1.57	1.19×10^{-15}	Persistent
cg06126421	6	30 720 080	*NA*	−0.77	2.57×10^{-9}	Persistent
cg24859433	6	30 720 203	*NA*	−1.41	1.76×10^{-9}	Persistent
cg19798735	7	110 730 805	*IMMP2L*	−1.45	5.32×10^{-14}	Reversible
cg11556164	7	110 738 315	*LRRN3*	−1.04	2.01×10^{-9}	Reversible
cg05221370	7	110 738 836	*LRRN3*	−1.08	2.70×10^{-11}	Persistent
cg09084200	11	134 095 863	*VPS26B*	−1.98	3.44×10^{-9}	Persistent
cg13937905	12	53 612 551	*RARG*	−0.64	3.58×10^{-15}	Persistent
cg10592478	12	53 612 641	*RARG*	−0.78	9.97×10^{-13}	Reversible
cg20124610	13	111 357 885	*CARS2*	−2.11	8.76×10^{-11}	Reversible
cg07756788	13	30 532 829	*NA*	−1.18	3.33×10^{-10}	Reversible
cg19572487	17	38 476 024	*RARA*	−1.46	2.16×10^{-14}	Persistent
cg03636183	19	17 000 585	*F2RL3*	−1.18	1.36×10^{-16}	Persistent
cg07381806	19	2 094 327	*MOBKL2A*	−0.80	8.01×10^{-11}	Persistent

consistency of the effect size estimates in both classes of CpG sites, hence suggesting a comparable methylation reversion rate in both reversible and persistent sites (Supplementary Material, Fig. S8). Persistent loci could therefore be interpreted as those being mostly altered by exposure to tobacco smoking, which would therefore take longer to revert back to methylation levels observed in never smokers.

Additional analyses in current smokers showed that the magnitude of methylation changes in the CpG sites classified as reversible or persistent were neither associated with smoking intensity (with the exception of one site on *AHRR*) nor with smoking duration. In addition, the classification of these CpG sites across the two classes appeared to be, at most, weakly driven by the exposure dose (i.e. intensity or duration). However, the two classes of CpG sites showed some differences in terms of their physical distribution on the gene with an under-representation of reversible sites in intergenic regions compared with persistent sites. One possible explanation, deserving more investigation, would be that, irrespective of the dose, CpG sites in the intergenic regions are more accessible and therefore more likely to be demethylated by exogenous compounds.

Our recoding strategy splits former smokers into two categories with respect to the time since smoking cessation t and pools the earliest quitters ($<t$) with current smokers and later quitters ($\geq t$) with never smokers. Each value of t investigated gave rise to a different population split. For instance moving from t to $t+1$ resulted in shifting former smokers who quit smoking $t+1$ years ago from the 'non-smoker' to the 'smoker' subpopulation. For each of the resulting datasets, we ran an EpWAS and identified across all EpWAS's the CpG sites that were at least once associated with the binary (recoded) smoking status. For these CpG sites of interest, we investigated the trajectory of the strength of the methylation–smoking association as measured by their P-value and specifically identified the value of t at which the association lost significance. That value varied across CpG sites and measured the time since smoking cessation after which pooling (i) current smokers and earlier quitters ($<t$) and (ii) never smokers and later quitters ($\geq t$) yielded too heterogeneous populations to ensure the identification of the effect of smoking. As such, it indicates, on a site-by-site basis, the time after which the smoking-induced methylation alterations become non-detectable. By construction, the reference population (non-smokers) varies across values of t, which precludes the direct comparison of the effect size estimates and limits our analyses to P-values. These measure the heterogeneity induced by pooling current and former ($<t$) smokers on the one hand, and never and former ($\geq t$) on the other hand. Hence, probes remaining statistically significant up to t can be interpreted as those whose methylation remains altered up to t years after smoking cessation. A further limitation of our approach is that some associations may be so strong that they are only marginally affected by the dilution effect induced by our pooling strategy. As depicted in Supplementary Material, Figure S9, these correspond to persistent CpG sites with high relative methylation changes and high P-values for the ever-versus-never smoker comparisons (e.g. 2q37.1, *AHRR*, *F2RL3*). For these probes, it may be difficult to rule out whether their clustering in the persistent group is solely due to the large effect of smoking or to actual long-lasting effect of smoking. However, Supplementary Material, Figure S9 also shows persistent probes with moderate-to-low relative methylation changes, and reciprocally reversible sites with high level of methylation changes. Despite interpretation issues for the (over-represented) strongly affected sites in the persistent class and less affected for reversible sites, our approach is able to unambiguously identify persistent and reversible CpG sites. The specificity of our classification procedure is further

highlighted by the fact that some of the CpG sites showing the largest relative methylation changes (absolute change between 0.4 and 0.5, e.g. cg21070864; chr 17; *BIRC5*) are neither classified as persistent nor reversible site in our analyses.

Our study highlights the existence of methylation alterations in white blood cells (WBC) decades after exposure cessation. One potential explanation is that we may be observing smoking-induced alterations in haematopoietic stem cells of the bone marrow. One can speculate that such methylation changes can also be considered markers of events occurring in other stem cells of the body, including target organs such as the lung.

We also identified >5500 CpG–transcript pairs that were significantly associated, showing mainly inverse associations between methylation and gene expression. Only five pairs comprised CpG sites and transcripts relating to the same genes (among which *LRRN3*). In addition, 334 pairs involved genes located on the same chromosome, suggesting either complex regulatory cascades linking methylation and gene expression, or the possibility of a remote regulation. The strongest CpG–transcript associations involved gene expression of *LRRN3* and CpG sites on *AHRR*, *F2RL3* and importantly on *LRRN3* itself. We also found, consistently with two recent studies (25,26), that the expression of *LRRN3* was strongly up-regulated by smoking exposure (almost a 3-fold higher expression in current smokers than in never smokers, P-value $< 10^{-23}$).

Among our candidate genes, three have been shown to have important biological implications. *GFI1* is involved in the regulation of haematopoietic stem cells (27); different stem/progenitor populations are characterized by distinctive transcription factor expression states, including relationships between the genes *Gata2*, *Gfi1* and *Gfi1b*. Therefore, the finding of hypomethylation of this gene in WBC of smokers and ex-smokers is consistent with increased activity of haematopoietic stem cells. *AHRR* is the repressor of the aryl hydrocarbon receptor, a key regulator of the relationships between the cell and the external environment, and of the effects of stressors such as dioxin and Polycyclic Aromatic Hydrocarbons (that are contained in tobacco smoke) (28,29). *AHRR*, notably, is expressed in all tissues (30). Finally, the role of *LRRN3* methylation is unclear. This gene encodes a leucine-rich repeat protein and has been related mainly to neurological/psychiatric conditions: polymorphisms in leucine-rich repeat genes are associated with autism spectrum disorder susceptibility in populations of European ancestry (31).

Our dataset comprised prospectively collected biological samples of participants from three case–control studies (on breast and colon cancers) nested in two cohorts. Although those two cancers are not highly associated with exposure to tobacco smoking, we performed a similar set of EpWAS restricting the study population to controls only. Reassuringly, these sensitivity analyses confirmed the 100 strongest associations of the EpWAS but lacked power to confirm the weaker ones. Additional sensitivity analyses were carried out by stratifying our analyses by cohort and despite the resulting loss of power, results from NOWAC and EPIC-Italy separately showed strong consistency and confirmed 83 and 94% of sites identified in the pooled analysis, respectively.

Overall, our study confirms the existence of already reported methylation markers of current smoking exposure (the main being *AHRR*, *F2RL3*, 2q37.1 and *GFI1*) and extends these findings with several previously undescribed methylation markers. Novel analyses on the dynamics of these methylation markers revealed two broad classes of methylation changes: CpG sites whose methylation reverts back to normal within the first three decades after smoking cessation and some persistent CpG markers remaining differentially methylated even three decades after smoking cessation. With reference to previous gene expression analyses (32), the truly irreversible nature of our persistent markers is yet to be confirmed, for instance in former smokers surviving decades after smoking cessation.

The dynamics of the reversible methylation sites mimic risk profiles observed after smoking cessation for some chronic diseases (e.g. lung cancer). Further research should link the dynamics in these markers to subsequent individual risk profiles. Our results also hold promise for the idea that exposure-related methylation changes can be detected years after actual exposure happened. Such historical fingerprints, if specific, could then potentially be used in epidemiological investigations on environmentally related chronic diseases.

Materials and Methods

Study population

Our study includes participants from three case–control studies on breast and colon cancer nested in the Italian component of the European Prospective Investigation into Cancer and Nutrition (33) (EPIC-Italy, $N = 47\,749$ volunteers aged 35–70) and the Norwegian Women and Cancer Study cohort (NOWAC, $N = 50\,000$ healthy women aged 46–63) (34,35).

For all EPIC participants, anthropometric measurements and lifestyle variables including detailed information on smoking and smoking history were collected at recruitment (1993–1998) through standardized questionnaires, together with a blood sample that was subsequently transported to local laboratories for processing and aliquot preparation. Blood was separated into 0.5 mL fractions and stored in liquid nitrogen at –196°C. All EPIC participants signed an informed consent form, and the ethical review boards of the International Agency for Research on Cancer and of local participating centres approved the study protocol.

Women enrolled in NOWAC (from 1991 to 2006) completed an eight-page questionnaire with information on current use of hormonal treatments and other pharmaceutical therapies, dietary supplements, smoking and height and weight. At the time of blood sampling (2003–2006), a complementary questionnaire including detailed information about smoking habits was distributed. Blood samples were sent by overnight mail to the Department of Community Medicine at the University of Tromsø, Norway. Upon arrival, the citrate glass tube was centrifuged and buffy-coat and plasma (two tubes) were separated. Both plasma and the PAXgene™ tubes were frozen immediately at –80°C. All participants gave written informed consent. The study was approved by the Regional Committee for Medical and Health Research Ethics and the Norwegian Data Inspectorate.

After exclusion of cases diagnosed <1 year after enrolment ($N = 30$) or individuals with a blood-related cancer ($N = 1$), the 451 samples from EPIC women comprised 129 and 70 prospective breast and colon cancer cases, respectively, and 252 healthy controls (Supplementary Material, Table S5A). Samples from the NOWAC study included 333 women from which ($N = 39$) participants diagnosed <1 year after recruitment were excluded leaving us with 294 women (129 breast cancer cases and 165 healthy controls, Supplementary Material, Table S5B). All retained subjects from both cohorts were cancer free at enrolment, and clinical onset in cases occurred >1 year after enrolment (on average after 5.8 years after enrolment in EPIC-Italy and after 2.6 years in NOWAC). Therefore, we considered all subjects as healthy at blood drawing, and we additionally adjusted our analyses for case–control status.

DNA methylation measurement, data pre-processing and quality control

Genome-wide DNA methylation analyses were performed on samples from both cohorts using the Illumina Infinium HumanMethylation450 platform. All laboratory procedures were carried out at the Human Genetics Foundation (Turin, Italy) according to manufacturers' protocols. Buffy coats stored in liquid nitrogen were thawed, and genomic DNA was extracted using the QIAGEN QIAsymphony DNA Midi Kit. DNA (500 ng) was bisulphite-converted using the Zymo Research EZ-96 DNA Methylation-Gold™ Kit, and hybridized to Illumina Infinium HumanMethylation450 BeadChips. These were subsequently scanned using the Illumina HiScanSQ system, and sample quality was assessed using control probes present on the micro-arrays. Finally, raw intensity data were exported from Illumina GenomeStudio (version 2011.1).

Data pre-processing was carried out using in-house software written for the R statistical computing environment. For each sample and each probe, measurements were set to missing if obtained by averaging intensities over less than three beads, or if averaged intensities were below detection thresholds estimated from negative control probes. Background subtraction and dye bias correction (for probes using the Infinium II design) were also performed. The resulting subset of 473 929 probes targeting autosomal CpG loci was selected for further analyses, of which probes detected in <20% of the samples were excluded from the analyses, leaving us with 432 414 probes.

Methylation levels at each locus were expressed as $\log_2$-transformed ratios of intensities arising from methylated cytosines over those arising from unmethylated cytosines (*M*-values).

Genome-wide gene expression profiles from NOWAC samples

Blood samples from NOWAC participants underwent transcriptomic profiling as already reported (34). Samples were analysed in three runs: one on an Illumina HumanWG-6 chip V3 and two on Illumina HT-12 chips at NTNU (Norwegian University of Science and Technology). Original probe values were background corrected. Probes reported to have poor quality from Illumina, no annotation or expressed in <1% of the total samples processed at the same time were removed. Among the remaining probes, the probe showing the highest signal per gene was kept. As each run was pre-processed separately, depending on the results of the filtering procedure, across the different runs, transcriptomic profiles can contain different probes. Only the set of probes commonly assayed across all runs was retained for subsequent analysis ($N = 8952$ transcripts, and an equal number genes).

EpWAS statistical models

Linear models were used for all analyses, with DNA methylation levels as dependent variable. To account for residual technical confounding, all models were adjusted for micro-array ($N = 89$) and position of the sample on the micro-array ($N = 12$). All analyses were additionally adjusted for blood cell composition estimated using the algorithm developed by Houseman *et al.* (36) by including in the model the estimated blood cell composition. The Houseman prediction model was calibrated using DNA methylation profiles of purified human leukocytes from six healthy male blood donors (37), and predictions were obtained using the subset of 89 490 probes found to be significantly differentially methylated across cell types at a stringent Bonferroni-corrected significance threshold ensuring a family-wise error rate lower than. The associations between smoking status and (estimated) blood cell composition are summarized in Supplementary Material, Table S6, and show, irrespective of the cell subtype, strong associations, especially in the current-to-never analyses.

Further adjustment covariates included age at blood collection (continuous), case–control status (binary) and centre (categorical, 6 classes).

Multiple testing was accounted for by using, as a stringent strategy, Bonferroni correction ensuring a strong control of the family-wise error rate at a 0.05 level.

Dynamics of methylation changes after smoking cessation

To investigate the dynamics of methylation changes after smoking cessation, we ran a set of additional EpWAS, as defined earlier, using a recoded binary indicator for smoking status. For a given time since smoking cessation t (ranging from 0 to 45), non-smokers included never and former smokers who quit more than t years ago; and smokers included current and former smokers having quit less than t years ago. For $t = 0$, our model compares methylation in current smokers to that of never and former smokers, and for $t = 45$, we compare methylation profiles from ever smokers against those of never smokers.

We investigated the classification of probes with respect to their time since smoking cessation after which they lose significance by running an unsupervised k-means clustering procedure (38) on the vectors (one per probe) each containing one binary variable per value of t indicating if the probe was significant for that value of t.

Targeted integration of gene expression data, pathway analyses

Log_2-transformed expression levels of the 8952 transcripts/genes assayed in NOWAC samples were regressed against methylation levels in a subset of CpG sites found to be associated with smoking. In these analyses, technical variation in the transcriptomic profiles was accounted for by adjusting our results for the analytical run (categorical variable, three classes). We declared CpG–transcript pairs as significant based on Bonferroni 5% significance level (per-test significance level $\alpha' = 0.05/(n \times 8952)$, where n denotes the number of CpG sites under investigation) (39–41).

Supplementary Material

Supplementary Material is available at *HMG* online.

Conflict of Interest statement. None declared.

Funding

This study was supported by the Human Genetics Foundation (HuGeF) and has been carried out within the Transdisciplinary Research in Cancer of the Lung (TRICL) project, supported by the National Cancer Institute, Grant U19 CA148127 02 to Christopher I. Amos. M.C.-H., S.A.K., R.C.H.V. and P.V. acknowledge the 7th European Framework Programme (Exposomics—308610—to P.V.). F.G. was supported by the Fondation de France (Grant 2012-00031604) and R.C. by the Fondation pour la Recherche Medicale (FRM, Grant number SPE20120523823). Gene expression and DNA methylation analyses in NOWAC samples were funded by ERC advanced grant Transcriptomics in cancer Epidemiology—ERC-2008-AdG-232997 and Tromsø research foundation.

References

1. Ezzati, M. and Lopez, A.D. (2003) Estimates of global mortality attributable to smoking in 2000. *Lancet*, **362**, 847–852.
2. Mathers, C.D. and Loncar, D. (2006) Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med.*, **3**, e442.
3. Newcomb, P.A. and Carbone, P.P. (1992) The health consequences of smoking—cancer. *Med. Clin. North Am.*, **76**, 305–331.
4. Vineis, P., Alavanja, M., Buffler, P., Fontham, E., Franceschi, S., Gao, Y.T., Gupta, P.C., Hackshaw, A., Matos, E., Samet, J. *et al.* (2004) Tobacco and cancer: recent epidemiological evidence. *J. Natl Cancer Inst.*, **96**, 99–106.
5. Conen, D., Everett, B.M., Kurth, T., Creager, M.A., Buring, J.E., Ridker, P.M. and Pradhan, A.D. (2011) Smoking, smoking cessation, [corrected] and risk for symptomatic peripheral artery disease in women: a cohort study. *Ann. Intern. Med.*, **154**, 719–726.
6. Kawachi, I., Colditz, G.A., Stampfer, M.J., Willett, W.C., Manson, J.E., Rosner, B., Speizer, F.E. and Hennekens, C.H. (1993) Smoking cessation and decreased risk of stroke in women. *JAMA*, **269**, 232–236.
7. Vernooy, J.H., Kucukaycan, M., Jacobs, J.A., Chavannes, N.H., Buurman, W.A., Dentener, M.A. and Wouters, E.F. (2002) Local and systemic inflammation in patients with chronic obstructive pulmonary disease: soluble tumor necrosis factor receptors are increased in sputum. *Am. J. Respir. Crit. Care Med.*, **166**, 1218–1224.
8. Willemse, B.W., ten Hacken, N.H., Rutgers, B., Lesman-Leegte, I.G., Postma, D.S. and Timens, W. (2005) Effect of 1-year smoking cessation on airway inflammation in COPD and asymptomatic smokers. *Eur. Respir. J.*, **26**, 835–845.
9. Ebbert, J.O., Yang, P., Vachon, C.M., Vierkant, R.A., Cerhan, J.R., Folsom, A.R. and Sellers, T.A. (2003) Lung cancer risk reduction after smoking cessation: observations from a prospective cohort of women. *J. Clin. Oncol.*, **21**, 921–926.
10. Vermeulen, R. and Chadeau-Hyam, M. (2012) Dynamic aspects of exposure history-do they matter? *Epidemiology*, **23**, 900–901.
11. Vlaanderen, J., Portengen, L., Schuz, J., Olsson, A., Pesch, B., Kendzia, B., Stucker, I., Guida, F., Bruske, I., Wichmann, H.E. *et al.* (2014) Effect modification of the association of cumulative exposure and cancer risk by intensity of exposure and time since exposure cessation: a flexible method applied to cigarette smoking and lung cancer in the SYNERGY Study. *Am. J. Epidemiol.*, **179**, 290–298.
12. Chadeau-Hyam, M., Tubert-Bitter, P., Guihenneuc-Jouyaux, C., Campanella, G., Richardson, S., Vermeulen, R., De Iorio, M., Galea, S. and Vineis, P. (2014) Dynamics of the risk of smoking-induced lung cancer: a compartmental hidden Markov model for longitudinal analysis. *Epidemiology*, **25**, 28–34.
13. Breitling, L.P., Yang, R., Korn, B., Burwinkel, B. and Brenner, H. (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am. J. Hum. Genet.*, **88**, 450–457.
14. Besingi, W. and Johansson, A. (2014) Smoke-related DNA methylation changes in the etiology of human disease. *Hum. Mol. Genet.*, **23**, 2290–2297.
15. Shenker, N.S., Polidoro, S., van Veldhoven, K., Sacerdote, C., Ricceri, F., Birrell, M.A., Belvisi, M.G., Brown, R., Vineis, P. and Flanagan, J.M. (2013) 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**, 843–851.
16. Dogan, M.V., Shields, B., Cutrona, C., Gao, L., Gibbons, F.X., Simons, R., Monick, M., Brody, G.H., Tan, K., Beach, S.R. *et al.* (2014) The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. *BMC Genomics*, **15**, 151.
17. Sun, Y.V., Smith, A.K., Conneely, K.N., Chang, Q., Li, W., Lazarus, A., Smith, J.A., Almli, L.M., Binder, E.B., Klengel, T. *et al.* (2013) Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. *Hum. Genet.*, **132**, 1027–1037.
18. Harlid, S., Xu, Z., Panduri, V., Sandler, D.P. and Taylor, J.A. (2014) CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the sister study. *Environ. Health Perspect.*, **122**, 673–678.
19. Elliott, H.R., Tillin, T., McArdle, W.L., Ho, K., Duggirala, A., Frayling, T.M., Davey Smith, G., Hughes, A.D., Chaturvedi, N. and Relton, C.L. (2014) Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin. Epigenetics*, **6**.
20. Zeilinger, S., Kühnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., Weidinger, S., Lattka, E., Adamski, J., Peters, A. *et al.* (2013) Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS ONE*, **8**.
21. Zhang, Y., Yang, R., Burwinkel, B., Breitling, L.P. and Brenner, H. (2014) F2RL3 methylation as a biomarker of current and lifetime smoking exposures. *Environ. Health Perspect.*, **122**, 131–137.
22. Wan, E.S., Qiu, W., Baccarelli, A., Carey, V.J., Bacherman, H., Rennard, S.I., Agusti, A., Anderson, W., Lomas, D.A. and Demeo, D.L. (2012) Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum. Mol. Genet.*, **21**, 3073–3082.
23. Novakovic, B., Ryan, J., Pereira, N., Boughton, B., Craig, J.M. and Saffery, R. (2014) Postnatal stability, tissue, and time specific effects of AHRR methylation change in response to maternal smoking in pregnancy. *Epigenetics*, **9**, 377–386.
24. Joubert, B.R., Haberg, S.E., Nilsen, R.M., Wang, X., Vollset, S.E., Murphy, S.K., Huang, Z., Hoyo, C., Midttun, O., Cupul-Uicab, L.A. *et al.* (2012) 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.*, **120**, 1425–1431.
25. Charlesworth, J.C., Curran, J.E., Johnson, M.P., Goring, H.H., Dyer, T.D., Diego, V.P., Kent, J.W. Jr., Mahaney, M.C., Almasy, L., MacCluer, J.W. *et al.* (2010) Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. *BMC Med. Genomics*, **3**, 29.
26. Beineke, P., Fitch, K., Tao, H., Elashoff, M.R., Rosenberg, S., Kraus, W.E., Wingrove, J.A. and Investigators, P. (2012) A whole blood gene expression-based signature for smoking status. *BMC Med. Genomics*, **5**, 58.
27. Moignard, V., Macaulay, I.C., Swiers, G., Buettner, F., Schutte, J., Calero-Nieto, F.J., Kinston, S., Joshi, A., Hannah, R., Theis, F.J. *et al.* (2013) Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol.*, **15**, 363–372.
28. Philibert, R.A., Beach, S.R. and Brody, G.H. (2012) Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. *Epigenetics*, **7**, 1331–1338.
29. Zudaire, E., Cuesta, N., Murty, V., Woodson, K., Adams, L., Gonzalez, N., Martinez, A., Narayan, G., Kirsch, I., Franklin, W. *et al.* (2008) The aryl hydrocarbon receptor repressor is a

putative tumor suppressor gene in multiple human cancers. *J. Clin. Invest.*, **118**, 640–650.
30. Yamamoto, J., Ihara, K., Nakayama, H., Hikino, S., Satoh, K., Kubo, N., Iida, T., Fujii, Y. and Hara, T. (2004) Characteristic expression of aryl hydrocarbon receptor repressor gene in human tissues: organ-specific distribution and variable induction patterns in mononuclear cells. *Life Sci.*, **74**, 1039–1049.
31. Sousa, I., Clark, T.G., Holt, R., Pagnamenta, A.T., Mulder, E.J., Minderaa, R.B., Bailey, A.J., Battaglia, A., Klauck, S.M., Poustka, F. *et al.* (2010) Polymorphisms in leucine-rich repeat genes are associated with autism spectrum disorder susceptibility in populations of European ancestry. *Mol. Autism*, **1**, 7.
32. Beane, J., Sebastiani, P., Liu, G., Brody, J.S., Lenburg, M.E. and Spira, A. (2007) Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol.*, **8**, R201.
33. Palli, D., Berrino, F., Vineis, P., Tumino, R., Panico, S., Masala, G., Saieva, C., Salvini, S., Ceroti, M., Pala, V. *et al.* (2003) A molecular epidemiology project on diet and cancer: the EPIC-Italy prospective study. Design and baseline characteristics of participants. *Tumori*, **89**, 586–593.
34. Dumeaux, V., Borresen-Dale, A.L., Frantzen, J.O., Kumle, M., Kristensen, V.N. and Lund, E. (2008) Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer postgenome cohort study. *Breast Cancer Res.*, **10**, R13.
35. Lund, E., Dumeaux, V., Braaten, T., Hjartåker, A., Engeset, D., Skeie, G. and Kumle, M. (2008) Cohort profile: the Norwegian Women and Cancer Study—NOWAC—Kvinner og kreft. *Int. J. Epidemiol.*, **37**, 36–41.
36. Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K. and Kelsey, K.T. (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, **13**.
37. Reinius, L.E., Acevedo, N., Joerink, M., Pershagen, G., Dahlén, S.-E., Greco, D., Söderhäll, C., Scheynius, A. and Kere, J. (2012) Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS ONE*, **7**.
38. Hartigan, J.A. and Wong, M.A. (1979) Algorithm AS 136: a K-means clustering algorithm. *J. Roy. Stat. Soc. C-App.*, **28**, 100–108.
39. Bird, A.P. and Wolffe, A.P. (1999) Methylation-induced repression—belts, braces, and chromatin. *Cell*, **99**, 451–454.
40. Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Gene Dev.*, **16**, 6–21.
41. Klose, R.J. and Bird, A.P. (2006) Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.*, **31**, 89–97.