EBioMedicine 37 (2018) 214–220

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

EBioMedicine
Published by THE LANCET

Epigenetic signatures of starting and stopping smoking

Daniel L. McCartney [a], Anna J. Stevenson [a], Robert F. Hillary [a], Rosie M. Walker [a,d], Mairead L. Bermingham [a], Stewart W. Morris [a], Toni-Kim Clarke [b], Archie Campbell [a], Alison D. Murray [c], Heather C. Whalley [b], David J. Porteous [a,d], Peter M. Visscher [d,e], Andrew M. McIntosh [a,b,d], Kathryn L. Evans [a,d], Ian J. Deary [d,f], Riccardo E. Marioni [a,d,*]

[a] Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, Scotland, United Kingdom
[b] Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, Scotland, United Kingdom
[c] Aberdeen Biomedical Imaging Centre, University of Aberdeen, Aberdeen, Scotland, United Kingdom
[d] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, Scotland, United Kingdom
[e] Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, Australia
[f] Department of Psychology, University of Edinburgh, Edinburgh, Scotland, United Kingdom

ARTICLE INFO

Article history:
Received 21 September 2018
Received in revised form 16 October 2018
Accepted 18 October 2018
Available online 30 October 2018

Keywords:
DNA methylation
Epigenetics
Smoking
Epidemiology

ABSTRACT

Background: Multiple studies have made robust associations between differential DNA methylation and exposure to cigarette smoke. But whether a DNA methylation phenotype is established immediately upon exposure, or only after prolonged exposure is less well–established. Here, we assess DNA methylation patterns from peripheral blood samples in current smokers in response to dose and duration of exposure, along with the effects of smoking cessation on DNA methylation in former smokers.

Methods: Dimensionality reduction was applied to DNA methylation data at 90 previously identified smoking–associated CpG sites for over 4900 individuals in the Generation Scotland cohort. K–means clustering was performed to identify clusters associated with current and never smoker status based on these methylation patterns. Cluster assignments were assessed with respect to duration of exposure in current smokers (years as a smoker), time since smoking cessation in former smokers (years), and dose (cigarettes per day).

Findings: Two clusters were specified, corresponding to never smokers (97·5% of whom were assigned to Cluster 1) and current smokers (81·1% of whom were assigned to Cluster 2). The exposure time point from which >50% of current smokers were assigned to the smoker–enriched cluster varied between 5 and 9 years in heavier smokers and between 15 and 19 years in lighter smokers. Low–dose former smokers were more likely to be assigned to the never smoker–enriched cluster in the first year following cessation. In contrast, a period of at least two years was required before the majority of former high–dose smokers were assigned to the never smoker–enriched cluster.

Interpretation: Our findings suggest that smoking–associated DNA methylation changes are a result of prolonged exposure to cigarette smoke, and can be reversed following cessation. The length of time in which these signatures are established and recovered is dose dependent. Should DNA methylation–based signatures of smoking status be predictive of smoking–related health outcomes, our findings may provide an additional criterion on which to stratify risk.

1. Background

Cigarette smoking is among the leading causes of illness and premature death worldwide [1]. In addition to multiple cancers [2], it is a major risk factor for cardiovascular and respiratory disorders [3,4]. Recent studies suggest that altered DNA methylation may play an important role in the biological pathways linking smoking to adverse health outcomes [5–9].

DNA methylation is an epigenetic modification, typically characterised by the addition of a methyl group to a cytosine–guanine dinucleotide (CpG). Both genetic and environmental factors can modulate DNA methylation levels, which in turn can regulate gene expression [10]. To date, the most informative environmental correlate of DNA methylation has been cigarette smoking. Multiple epigenome–wide association studies (EWAS) have been performed on smoking, using

* Corresponding author at: Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, Scotland, United Kingdom.
E-mail address: riccardo.marioni@ed.ac.uk (R.E. Marioni).

https://doi.org/10.1016/j.ebiom.2018.10.051

Research in context

Evidence before this study

The effects of cigarette smoking on DNA methylation have been well established. However, fewer studies have investigated: (1) how long these effects persist upon cessation; (2) the extent to which they can be reversed; and (3) how long it takes for such smoking-based methylation patterns to appear.

Added value of this study

We show the extent to which smoking-associated DNA methylation profiles are time- and dose-dependent in current smokers. In addition, we demonstrate the reversibility of these changes in former smokers is dependent on time since cessation and dose prior to quitting. To our knowledge, this is currently the largest study of DNA methylation in former smokers. Furthermore, the broad age range of our cohort has permitted us to investigate DNA methylation in both recent and long-term smokers (from <1 to >50 years as a smoker), and recent and long-term quitters (from <1 to >35 years since quitting).

Implications of all the available evidence

The establishment of smoking-associated DNA alterations provides an important public health message as a deterrent from smoking initiation. Furthermore, our reports on the dose-dependency and reversibility of these changes may encourage a reduction in the cigarette intake of current smokers (if not cessation), and an incentive against relapse in former smokers.

either status (e.g. current smoker, former smoker, never smoker) or intake (e.g. pack years) as the trait of interest [5,7,9], identifying thousands of smoking–associated loci. Moreover, cohort studies have reported altered DNA methylation in the offspring of women who smoked during pregnancy [11–13]. These analyses have identified a large number of loci where methylation is altered by exposure to cigarette smoke, with the cg05572921 locus in the aryl hydrocarbon receptor (AHR) repressor (*AHRR*) gene being among the most robustly implicated.[6–8,13] The relationship between exposure to cigarette smoke and DNA methylation changes has been widely reported. However, when these effects are established in smokers and whether they can be recovered by cessation is not well understood. Studying the mechanics of smoking–associated DNA methylation changes may provide a novel means of identifying risk of smoking–related morbidities.

We investigated the extent to which smoking–associated DNA methylation changes were associated with duration of exposure in current smokers and time since cessation in former smokers. We examined the relationship between DNA methylation from peripheral blood samples and smoking in a cohort of over 4900 individuals, incorporating self–reported years as a smoker and cigarettes per day as metrics for duration of exposure and dose, respectively.

2. Methods

2.1. The Generation Scotland cohort

Details of the Generation Scotland: Scottish Family Health Study (GS:SFHS) have been described previously [14,15]. DNA samples were collected for genotype– and DNA methylation–profiling along with detailed clinical, lifestyle, and sociodemographic data. The current study comprised 4905 individuals from the cohort for whom both DNA methylation and smoking data were available. A summary of variables assessed in this analysis is presented in Table 1.

GS:SFHS smoking data were collected using two different questionnaires. The first version of the questionnaire was answered by 2158/4905 (44·0%) of the participants and collected data on absolute values with respect to number of cigarettes smoked, and age started/stopped smoking. The second version of the questionnaire, which was answered by the remaining 2747 individuals in the analysis sample (56·0%), collected data using binned intervals. In order to harmonise the two sets of measurements, mid–point interval estimates were calculated for the ordinal data from the second version of the questionnaire (e.g., 17 years of exposure was assigned to individuals who had reported smoking between 15 and 19 years). Second–hand smoking status was assigned based on whether participants reported exposure to cigarette smoke at home, work or elsewhere, or whether they reported cohabiting with a smoker. Both questionnaires can be accessed from the GS:SFHS website (www.generationscotland.co.uk).

In the current study, exposure data were placed into ten five–year bins from 0 to 4 years to 45–49 years (N ≥ 32 per bin), with the longest exposure defined as ≥50 years (N = 23). Data on time since cessation were placed into five–year bins from 10 to 14 years to 30–34 years (N ≥ 48 per bin). The longest time since cessation was defined as ≥35 years (N = 53), whereas the most recent cessation time points (0–9 years) were presented as yearly intervals (N ≥ 26). Sample counts at each exposure and cessation time point are presented in Supplementary Tables 1–2.

2.2. Ethics

All components of GS:SFHS received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). GS:SFHS has also been granted Research Tissue Bank status by the Tayside Committee on Medical Research Ethics (REC Reference Number: 10/S1402/20), providing generic ethical approval for a wide range of uses within medical research.

2.3. GS:SFHS DNA methylation

Genome–wide DNA methylation was profiled from peripheral blood samples in 5200 individuals using the Illumina HumanMethylationEPIC

Table 1
Summary of the Generation Scotland cohort and variables assessed. Sample numbers are presented for each variable (N) along with mean and standard deviation (SD) values, where applicable.

Variable	N	Mean	SD
Sex			
Males	1872	–	–
Females	3033	–	–
Age (years)	4905	48·5	13·9
Smoking status			
Current smoker	917	–	–
Former smoker	1466	–	–
Never smoker	2522	–	–
Smoking variables[a]			
Cigarettes per day (current and former smokers)	2177	11·1	9·8
Cigarettes per day (current smokers)	859	15·2	9·8
Cigarettes per day (former smokers)	1318	8·5	8·8
Pack years (current and former smokers)	2037	15·9	16·8
Pack years (current smokers)	854	23·3	19·7
Pack years (former smokers)	1183	10·6	11·8
Years as a smoker (current and former smokers)	2221	27·7	13·4
Years as a smoker (current smokers)	917	28·7	13·2
Years as a smoker (former smokers)	1304	27·0	13·5
Years since cessation (former smokers only)	1324	9·0	10·0

[a] Information relating to dose and time since cessation/duration of exposure was not available for all current and former smokers.

BeadChip. Quality control was conducted in R [16]. ShinyMethyl was used to plot the log median intensity of methylated versus unmethylated signal per array, with outliers excluded upon visual inspection [17]. WateRmelon was used to remove (1) samples where ≥1% of CpGs had a detection p–value in excess of 0·05 (2) probes with a beadcount of less than three in more than five samples, and (3) probes where ≥0·5% of samples had a detection p–value in excess of 0·05 [18]. Methylation β-values were calculated using the dasen() normalisation method. Briefly, the dasen method performs background adjustment and quantile normalises Type I and Type II probes separately. From these, M-values were calculated using the Beta2M() function in wateRmelon [18]. ShinyMethyl was used to exclude samples where predicted sex did not match recorded sex. Ten saliva–derived samples and three samples from individuals who had answered "yes" for all self–reported conditions were also excluded (e.g. stroke, Alzheimer's disease, depression. Further details on these conditions are available in the GS:SFHS questionnaire, accessible from the GS:SFHS website: http://www.generationscotland.co.uk). This left a sample of 5088 participants with blood–derived samples available for analysis, of whom 4905 had smoking data available.

2.4. Statistical analysis

All analyses were performed in R [16].

Data–driven cluster analysis was performed on the top 100 p–value–ranked methylation sites from a recent, large meta–analysis EWAS of current versus never smoking (Joehanes et al. Supplementary Table 1, Sheet 02) [5,19]. Ninety of the top 100 probes were present in the GS:SFHS DNA methylation dataset following quality control (Supplementary Table 3). Clusters were visualised by plotting the first two principal coordinates, identified via data reduction analysis (multi dimensional scaling), using the cmdscale() function in the stats package [16,19]. K–means clustering was performed to partition the data, using the kmeans() function in the stats package [16]. As the probe set under consideration was associated with current/never smoker status, two clusters were specified.

Logistic regression was performed to assess the relationship between a genetic variant in the *CHRNA5–A3–B4* gene cluster that is associated with heaviness of smoking (rs1051730) and cluster assignment in current smokers, adjusting for sex [20]. The relationships between cluster assignment and batch, sex, and passive smoking were assessed using Chi–Squared Tests. The relationship between cluster assignment and alcohol consumption (current, former, and never drinker) was assessed using a Fisher's Exact Test. The relationship between cluster assignment and time since cessation (former smokers) and duration of exposure (current smokers) was assessed using logistic regression, adjusting for sex, age and dose.

Data were visualised using "broken stick" regression lines using the default parameters for the segmented() function in the segmented package in R [21]. Comprehensive smoking index (CSI) values were calculated for former smokers using the method described by Dietrich and Hoffman, using a half-life estimate of 1.5 [22].

3. Results

Descriptive data for the 917 current–, 1466 ex–, and 2522 never–smokers are summarised in Table 1. On average, current smokers had a greater duration of exposure compared to former smokers (28·7 years vs 27·0 years), and a greater cumulative dose (23·3 pack years vs 10·6 pack years).

3.1. Clustering of current smokers depends on dose and duration smoked

Data reduction was performed on DNA methylation data for 90 smoking-associated sites (multidimensional scaling; Fig. 1). Of the 2522 never smokers, 2459 (97·5%) were assigned to a *never smoker–enriched* cluster whereas, of the 917 current smokers, 744 (81·1%) were assigned to a *smoker–enriched* cluster (K-means clustering with two clusters specified). There was no association between misclassification of current smokers to the *never smoker–enriched* cluster and sex, alcohol consumption, batch, or genotype at the well–established nicotine addiction genetic variant rs1051730, ($P \geq 0{\cdot}103$; Supplementary Table 4) [20]. Similarly, there was no association between misclassification of never smokers ($N = 63$) to the *smoker–enriched* cluster and exposure to second–hand smoke, sex, alcohol consumption, or plate processing batch ($P \geq 0{\cdot}179$; Supplementary Table 4). The proportion of current smokers assigned to the *smoker–enriched* cluster increased with years as a smoker (Fig. 2 and Supplementary Table 1; $OR_{\text{smoker-enriched cluster}} = 1.07$ per year of smoking; 95% CI = 1.03–1.12; $P = 2.4 \times 10^{-4}$). A significant association was also present between cluster assignment and cigarettes per day ($OR_{\text{smoker-enriched cluster}} = 1.12$ per cigarette smoked per day; 95% CI = 1.09–1.15; $P = 3.9 \times 10^{-14}$). Of the 32 individuals who reported smoking for 0–4 years prior to DNA methylation sampling, seven (21·9%) were assigned to the *smoker–enriched* cluster; for the 76 individuals who reported smoking for 5–9 years prior to sampling, 34 (44·7%) were assigned to the *smoker–enriched* cluster. The proportion of assignments to the *smoker–enriched* cluster increased to 87·3% for current smokers at 20–24 years of exposure, remaining stable thereafter. Of the 670 current smokers reporting at least 20 years of exposure, 605 (90·3%) were assigned to the *smoker–enriched* cluster.

There was a significant association between dose (cigarettes per day) and duration of exposure (years as a smoker) in current smokers. Individuals who had smoked for a longer duration were more likely to be heavier smokers (age– and sex–adjusted linear regression Beta = 0·38 cigarettes per day for each year as a smoker; $P < 0.0001$). To minimise confounding between dose and duration of exposure, data for current smokers were split based on the median dose to generate time point–specific subsets of heavy and light smokers. The proportion of *smoker–enriched* cluster assignments increased with duration of exposure in both dose groups, stabilising at 15–19 years of exposure in heavy smokers, and 25–29 years in lighter smokers (Fig. 2 and Supplementary Table 1). The proportion of individuals assigned to the *smoker–enriched* cluster over time in heavy smokers was significantly greater than that in light smokers (Wilcoxon signed rank test P = 0·002).

3.2. Clustering of former smokers depends on dose and time since cessation

Of the 1466 former smokers assessed, 359 (24·5%) were assigned to the *smoker–enriched* cluster. The proportion of *smoker–enriched* cluster assignments decreased as time since smoking cessation increased (Fig. 3 and Supplementary Table 2; $OR_{\text{smoker-enriched cluster}} = 0.86$ per year since cessation; 95% CI = 0.83–0.89; $P < 2.0 \times 10^{-16}$). A significant association was also present between cluster assignment and cigarettes per day in former smokers ($OR_{\text{smoker-enriched cluster}} = 1.08$ per cigarette smoked per day; 95% CI = 1.06–1.10; $P = 1.67 \times 10^{-14}$). The highest proportion of *smoker–enriched* cluster assignments (64·4%) was observed in individuals who had quit smoking within a year prior to sampling. The proportion of *smoker–enriched* cluster assignments fell below 50% by 1 year following cessation. Contrary to the findings in current smokers, there was a significant negative relationship between dose and duration of exposure in former smokers (age– and sex–adjusted linear regression Beta = −0·18 cigarettes smoked per day for each year as a smoker $P < 0.0001$). Samples were next split on the median dose at each cessation time point to obtain a high–dose and low–dose group. The proportion of *smoker–enriched* cluster assignments was significantly lower in the low–dose group relative to the high–dose group (Wilcoxon signed rank test $P = 1{\cdot}5 \times 10^{-4}$). The proportion of *smoker–enriched* cluster assignments in the low–dose group was consistently below 50% (Fig. 3 and Supplementary Table 2). The proportion of *smoker–enriched* cluster assignments for former smokers exposed to a

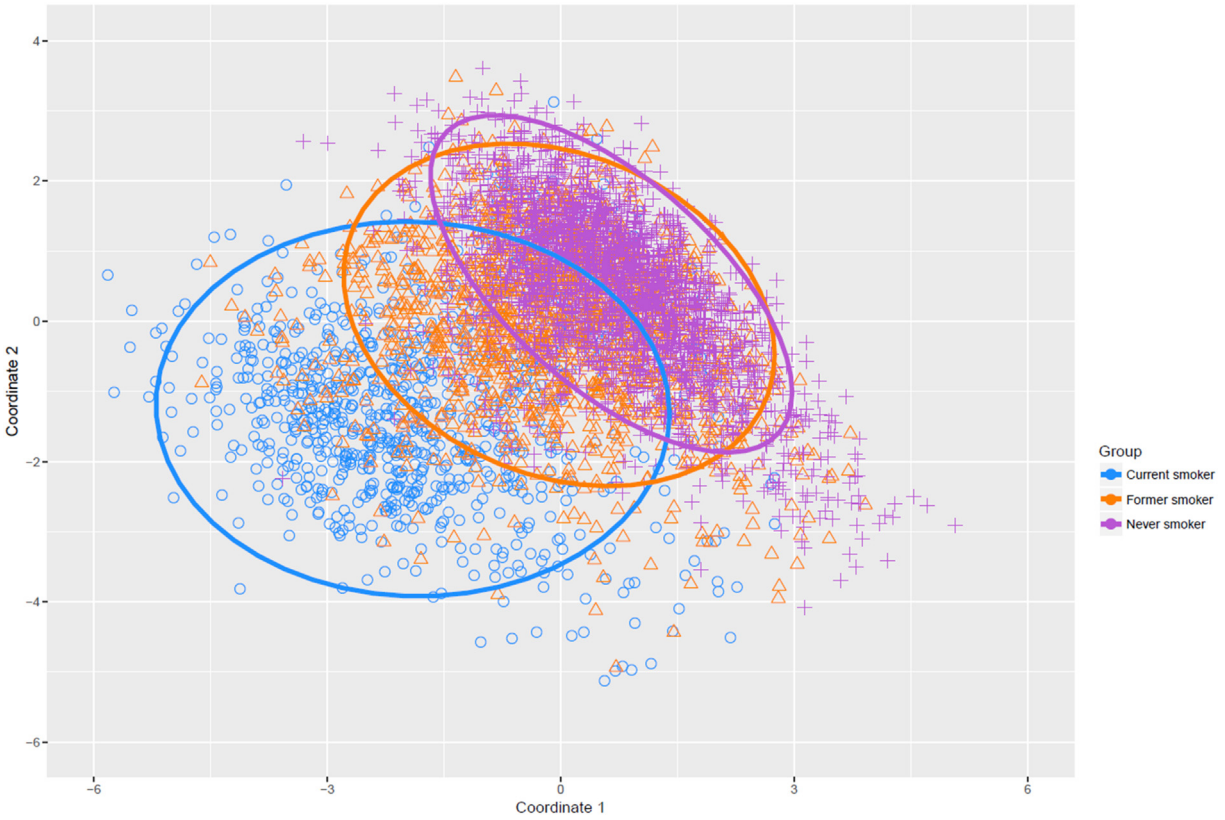

Fig. 1. Principal coordinate vectors 1 and 2 from a multidimensional scaling analysis of 90 smoking-associated probes. Points and ellipses are coloured by smoking status (blue circles = current smokers, orange triangles = former smokers, purple crosses = never smokers). Ellipses represent normal confidence ellipses.

high dose fell below 50% two years following cessation. From five years following smoking cessation, the proportion of *smoker–enriched* cluster assignments stabilised in high– and low–dose groups. Of the 760 individuals who had quit at least 5 years prior to sampling, 84 (11·1%) were assigned to the *smoker–enriched* cluster. As duration of exposure was not considered here, the analysis was repeated substituting years since cessation with pack years (years as a smoker × packs smoked per day), revealing a similar trend (Supplementary Table 5, Supplementary Fig. 1). Using pack years as a metric, the proportion of *smoker–enriched* cluster assignments in the low–dose group stabilised from two years following smoking cessation, compared to five years following cessation in the high–dose group.

The comprehensive smoking index (CSI) was calculated as an additional metric to incorporate duration, intensity and recency of exposure in former smokers, and its relationship with cluster assignment was assessed [22]. In a five-year period from cessation, individuals with lower CSI scores were less likely to be assigned to the *smoker-enriched* cluster relative to those with a higher CSI score. Cluster assignments between high- and low-CSI individuals stabilised beyond 5 years following cessation (Supplementary Fig. 2).

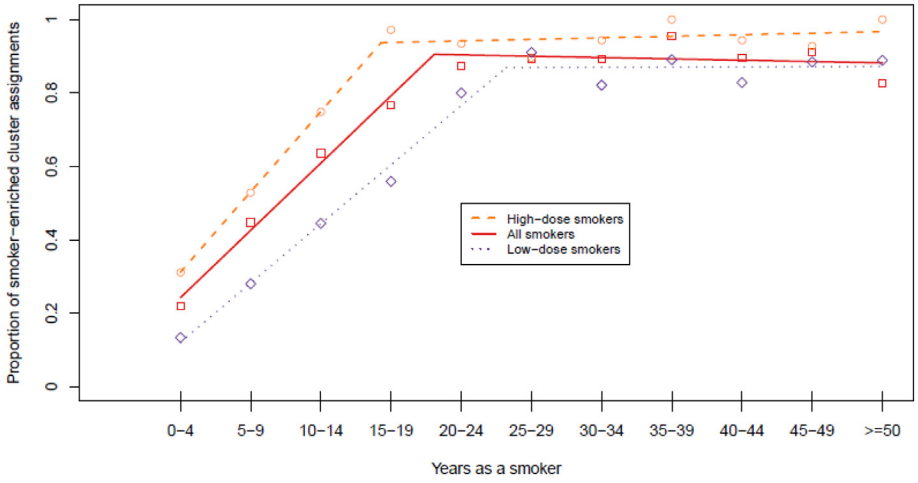

Fig. 2. Proportion of current smokers assigned to Cluster 2 (smoker-enriched cluster) by duration of exposure. "Broken stick" regression lines are presented for all current smokers (red solid line, square points), high-dose current smokers (orange dashed line, circular points) and low-dose current smokers (purple dotted line, diamond points).

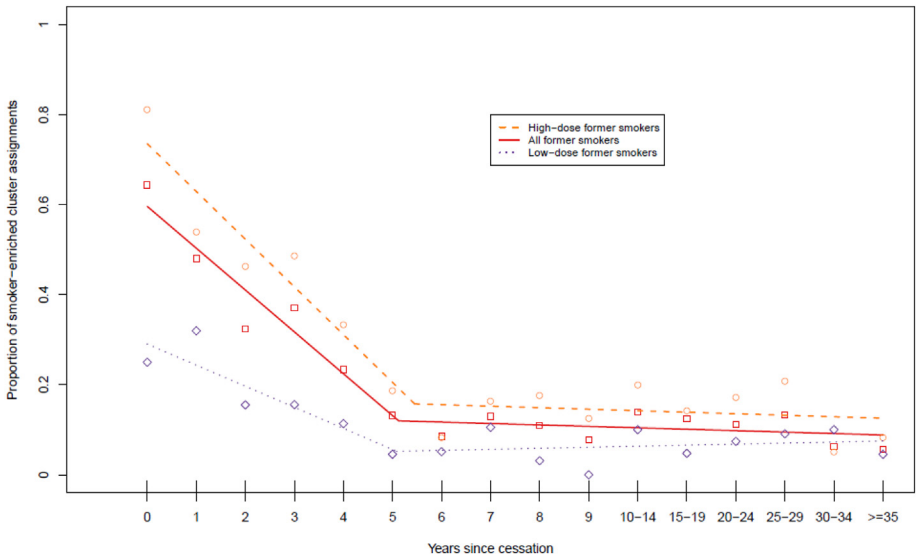

Fig. 3. Proportion of former smokers assigned to Cluster 2 (smoker-enriched cluster) by years since smoking cessation. "Broken stick" regression lines are presented for all former smokers (red solid line, square points), high-dose former smokers (orange dashed line, circular points) and low-dose former smokers (purple dotted line, diamond points).

Finally, we investigated the trajectories of DNA methylation at probes where smoking-associated modifications were reported to persist up to 30 years following cessation [5]. Of 36 probes reported by Joehanes et al., 30 were present in the GS:SFHS DNA methylation data [5]. Absolute t-statistics for DNA methylation in former smokers versus never smokers decreased with increasing years since cessation (Supplementary Fig. 3; Supplementary Tables 6–7). Four probes (cg05575921, cg21566642, cg01940273 and cg00706683) remained significantly differentially methylated in former smokers relative to never smokers up to 30 years following cessation.

3.3. Sensitivity analysis

To check the robustness of the predictions, three sensitivity analyses were considered. In the first analysis, a parsimonious predictor was developed by selecting CpG sites that discriminated smokers from non–smokers with an AUC > 0.9 (five out of the 18,760 genome–wide EWAS sites identified by Joehanes et al. – cg05575921, cg21566642, cg01940273, cg03636183 and cg21161138) [5]. There was a slight improvement in the prediction of current versus never smokers using this score (Supplementary Table 8; Supplementary Fig. 4). However, the proportion of current smoker assignments in high–dose former smokers was consistently higher in the five–CpG predictor compared with the cluster–based predictor. Moreover, low–dose former smokers displayed a consistent proportion of current smoker assignments over time in comparison to the cluster–based assignments (Supplementary Fig. 5 comparison with Fig. 3).

In the second analysis, two predictors were developed based on polygenic scores for a subset of the most significant smoking–associated CpG sites (N = 90), and all smoking–associated sites (N = 17,529) [5]. The 90–probe polygenic predictor yielded similar results to the cluster– and AUC–based predictors (Supplementary Table 9, Supplementary Figs. 6–7). In contrast, the polygenic score derived from the larger probe set displayed poorer predictions (Supplementary Table 10, Supplementary Figs. 8–9).

In the final analysis, DNA methylation-based smoking scores were generated for former smokers, based on a signature developed from current and never smokers in the GS:SFHS dataset [23]. Average DNA methylation scores in former smokers decreased within the first 2–3 years of quitting, remaining stable thereafter (Supplementary Fig. 10).

4. Discussion

In this study, we showed that smoking–based DNA methylation patterns are time– and dose–dependent. We identified two clusters from DNA methylation data in over 4900 individuals – one enriched for current smokers and another enriched for never–smokers. It took 15–19 years for the majority of low–dose smokers to display a methylation profile that assigned them to the smoker–enriched cluster. It took <1 year for the majority of low–dose ex–smokers to be assigned to the never smoker–enriched cluster. By contrast, it took 5–9 years for the majority of heavy–dose smokers to display DNA methylation profiles corresponding to the smoker–enriched cluster, and up to 2 years since quitting before the majority of heavy–dose ex–smokers had methylation patterns that more strongly resembled those of never smokers. Furthermore, there is little impact of smoking dose on methylation–based clustering of smoking for those who had smoked for >25 years or for those who had stopped smoking for at least 6 years.

These findings suggest that a prolonged period of exposure to cigarette smoke is required before a smoking–related signature can be reliably identified using DNA methylation data. This is supported by evidence from multiple studies, which have reported an association between duration of exposure to cigarette smoke and an increased risk of oesophageal, lung, and bladder cancers [24–26]. Moreover, a longer duration of exposure has been linked to an increased risk of chronic obstructive pulmonary disorder (COPD) and respiratory symptoms [27]. It is therefore worth considering our findings in the context of molecular pathological epidemiology (MPE), an approach that implicates exogenous factors such as lifestyle and the environment on both disease pathogenesis and omics measures such as DNA methylation and gene expression [28,29]. In the current study, we examined DNA methylation from blood and not from tumour or more likely disease-targeted tissues such as lung. Nonetheless, there may still be precision medicine applications of blood-based DNA methylation smoking signatures. Should the DNA methylation profile of smokers be associated with an increased risk of smoking–related pathologies, the current findings suggest there is a dose–dependent period of exposure within which this risk is comparable to that of never smokers.

Others have reported reversion of smoking–associated DNA methylation changes in former smokers persisting beyond 30 years from cessation, with the most rapid reversion rates occurring in the within the first 14 years [30]. Moreover, increased methylation levels at *AHRR* has been reported in smokers undergoing cessation therapy [31]. Examination of

cluster assignments in former smokers revealed to some degree the reversible nature of smoking–associated DNA methylation changes. Former light smokers were more likely to be assigned to the never smoker–enriched cluster, regardless of time since cessation. In contrast, a period of two years was required before the rate of never–smoker cluster assignments for former heavy smokers reached >50%.

A small proportion of never smokers were assigned to the smoker–enriched cluster. Such misclassifications may be a result of passive smoking, or other lifestyle–related correlates of smoking status. Although we did not observe an association between alcohol consumption and assignment of never smokers to the *smoker–enriched* cluster, it is possible that additional smoking–associated factors contribute to their misclassification. The effects of passive (i.e. second–hand) smoking on DNA methylation have been well established, with differential DNA methylation reported to persist up to decades following exposure to cigarette smoke in–utero [11–13]. It was not possible to determine whether the individuals profiled in the current study were exposed to cigarette smoke in utero as maternal smoking data were unavailable. Second–hand smoke exposure has also been linked to differential DNA methylation in adults. Similar DNA methylation patterns have been observed in lung tumours of smokers and second–hand smokers [32]. Hypomethylation of *AHRR* at cg05575921 has been linked to recent exposure to second–hand smoke [33], while others have reported significant associations between second–hand smoke exposure and differential DNA methylation in bladder cancer [34]. There was no association between misclassification of never smokers and co–habitation with, or other exposure to, smokers. However, information on the duration of co–habitation with smokers was not available, and there was no information regarding co–habitants and exposures prior to sampling.

We showed the use of AUC–based prediction of current/never smoking status using five probes is more accurate than the cluster–based prediction. However, smoking–associated DNA methylation changes at four of these five probes have been reported to persist decades following cessation [5]. Moreover, as the prediction thresholds for the five CpGs were selected to discriminate smoking status in the current sample, this generates a biased predictor when applied to the same data. While the predictive performance of the cluster–based predictions is less accurate for current/never smokers, its application to former smokers may be more suitable due to the inclusion of sites with reversible smoking–associated DNA methylation changes. This is reflected by the consistently higher proportion of current smoker assignments over time from the AUC–based predictor relative to the cluster–based predictor.

In a further sensitivity analysis, Z–score based polygenic methylation scores were built from all 18,760 genome–wide significant CpGs (N = 17,529 present in the GS:SFHS dataset) and also from the top 100 CpGs (N = 90 present in the GS:SFHS dataset). In the primary analysis, clusters were defined in relation to the methylation values in the Generation Scotland cohort, which may have introduced ascertainment bias. The polygenic analysis for the 90 CpGs yielded very similar results to the primary models. In contrast, the polygenic analysis derived from 17,529 probes did not perform as well. This was possibly due to the introduction of noise from many features of small effects. Conversely, predictive performance was improved by the inclusion of fewer features of larger effects.

In addition to the lack of information regarding maternal and past exposures to second–hand smoke, a further limitation to the current analysis is the presence of confounding between cigarette dose and duration of exposure to cigarette smoke. In order to minimise this association and to focus primarily on duration of exposure, the sample was stratified on the median dose at each time point assessed. A strength of this study is the use of a large and homogeneous analysis cohort. The Generation Scotland cohort comprises participants across a broad age range (18–99 years) which has permitted the analysis of smoking exposure in a large number of both recently–started and long–term smokers, as well as recently–quit and long–term former–smokers. Moreover, future analysis of smoking phenotypes and related health outcomes are possible, as a result of data linkage capabilities and sample collection for longitudinal DNA methylation profiling.

In conclusion, our findings suggest there is a dose–dependent interval within which smoking–associated DNA methylation are established. Furthermore, we have demonstrated a degree of reversibility of these changes in former smokers, whereby the interval of reversion is dependent on dose prior to smoking cessation. Consideration of duration of exposure in current smokers, and years since cessation in former smokers, coupled with dose, all measured via DNA methylation patterns, may assist in determining and stratifying risk of smoking–associated morbidities. Highlighting the establishment of smoking-associated DNA alterations provides an important public health message as a deterrent from smoking initiation. Furthermore, our reports on the dose-dependency and reversibility of these changes may encourage a reduction in the cigarette intake of current smokers (if not cessation), and an incentive against relapse in former smokers.

Author contributions

Conception and design: DLM, REM, IJD, DJP, PMV.
Data analysis: DLM.
Drafting the article: DLM, REM.
Data preparation: DLM, REM, RMW, MLB, SWM, AC.
Data collection: DJP, AMM, KLE, ADM.
Revision of the article: all authors.

Declaration of interests

Dr. McIntosh reports grants from The Sackler Trust, grants from Eli Lilly, and grants from Janssen outside the submitted work. The remaining authors have nothing to disclose.

Acknowledgements

This work was supported by Alzheimer's Research UK Major Project Grant [ARUK–PG2017B–10]. Generation Scotland received core funding from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006]. We are grateful to all the families who took part, the general practitioners and the Scottish School of Primary Care for their help in recruiting them, and the whole Generation Scotland team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, healthcare assistants, and nurses. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award "STratifying Resilience and Depression Longitudinally" (STRADL) [104036/Z/14/Z]. DNA methylation data collection was funded by the Wellcome Trust Strategic Award [10436/Z/14/Z]. The research was conducted in The University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology (CCACE), part of the cross–council Lifelong Health and Wellbeing Initiative [MR/K026992/1]; funding from the Biotechnology and Biological Sciences Research Council (BBSRC) and Medical Research Council (MRC) is gratefully acknowledged. CCACE supports Ian Deary, with some additional support from Dementias Platform UK [MR/L015382/1]. HCW is supported by a JMAS SIM fellowship from the Royal College of Physicians of Edinburgh. AMM and HCW have received support from the Sackler Institute.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2018.10.051.

References

[1] World Health Organization. WHO global report on trends in prevalence of tobacco smoking 2015. WHO Mag 2015:1–359.
[2] Blakely T, Barendregt JJ, Foster RH, et al. The association of active smoking with multiple cancers: National census–cancer registry cohorts with quantitative bias analysis. Cancer Causes Control 2013;24:1243–55. https://doi.org/10.1007/s10552-013-0204-2.
[3] Berry JD, Dyer A, Cai X, et al. Lifetime risks of cardiovascular disease. N Engl J Med 2012;366:321–9. https://doi.org/10.1056/NEJMoa1012848.
[4] Jayes L, Haslam PL, Gratziou CG, et al. SmokeHaz: Systematic reviews and meta–analyses of the effects of smoking on respiratory health. Chest; 2016. p. 164–79.
[5] Joehanes R, Just AC, Marioni RE, et al. Epigenetic signatures of cigarette smoking. Circ Cardiovasc Genet 2016;9:436–47. https://doi.org/10.1161/CIRCGENETICS.116.001506.
[6] Shenker NS, Ueland PM, Polidoro S, et al. DNA methylation as a long–term biomarker of exposure to tobacco smoke. Epidemiology 2013;24:712–6. https://doi.org/10.1097/EDE.0b013e31829d5cb3.
[7] Zeilinger S, Kühnel B, Klopp N, et al. Tobacco smoking leads to extensive genome–wide changes in DNA methylation. PLoS One 2013;8. https://doi.org/10.1371/journal.pone.0063812.
[8] Guida F, Sandanger TM, Castagné R, et al. Dynamics of smoking–induced genome–wide methylation changes with time since smoking cessation. Hum Mol Genet 2015;24:2349–59. https://doi.org/10.1093/hmg/ddu751.
[9] Freeman JR, Chu S, Hsu T, Huang Y–T. Epigenome–wide association study of smoking and DNA methylation in non–small cell lung neoplasms. Oncotarget 2016;7:69579–91.
[10] Jaenisch R, Bird A. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. Nat Genet 2003;33:245–54. https://doi.org/10.1038/ng1089.
[11] Richmond RC, Simpkin AJ, Woodward G, et al. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: Findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). Hum Mol Genet 2015;24:2201–17. https://doi.org/10.1093/hmg/ddu739.
[12] Richmond RC, Suderman M, Langdon R, Relton CL, Davey Smith G. DNA methylation as a marker for prenatal smoke exposure in adults. Int J Epidemiol 2018. https://doi.org/10.1093/ije/dyy091.
[13] Joubert BR, Felix JF, Yousefi P, et al. DNA methylation in newborns and maternal smoking in pregnancy: Genome–wide consortium meta–analysis. Am J Hum Genet 2016;98:680–96. https://doi.org/10.1016/j.ajhg.2016.02.019.
[14] Smith BH, Campbell H, Blackwood D, et al. Generation Scotland: The Scottish family health study; a new resource for researching genes and heritability. BMC Med Genet 2006;7. https://doi.org/10.1186/1471–2350–7–74.
[15] Smith BH, Campbell A, Linksted P, et al. Cohort profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. Int J Epidemiol 2013;42:689–700. https://doi.org/10.1093/ije/dys084.
[16] R Core Team, R Development Core Team. R A Lang. Environ. Stat. Comput, 55; 2017; 275–86.
[17] Fortin J–P, Fertig E, Hansen K. shinyMethyl: Interactive quality control of Illumina 450k DNA methylation arrays in R. F1000Res 2014. https://doi.org/10.12688/f1000research.4680.2.
[18] Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data–driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics 2013;14. https://doi.org/10.1186/1471–2164–14–293.
[19] Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 1966;53:325. https://doi.org/10.1093/biomet/53.3-4.325.
[20] Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 2008;452:638–42. https://doi.org/10.1038/nature06846.
[21] Muggeo VMR. Segmented: An R package to fit regression models with broken–line relationships. R News 2008;8:20–5.
[22] Dietrich T, Hoffmann K. A comprehensive index for the modeling of smoking history in periodontol research. J Dent Res 2004;83(12):966.
[23] McCartney DL, Hillary RF, Stevenson AJ, et al. Epigenetic prediction of complex traits and death. Genome Biol 2018;19(1):136.
[24] Pandeya N, Williams GM, Sadhegi S, Green AC, Webb PM, Whiteman DC. Associations of duration, intensity, and quantity of smoking with adenocarcinoma and squamous cell carcinoma of the esophagus. Am J Epidemiol 2008;168:105–14. https://doi.org/10.1093/aje/kwn091.
[25] Flanders WD, C A Lally, Zhu B–P, Henley SJ, Thun MJ. Lung cancer mortality in relation to age, duration of smoking, and daily cigarette consumption: Results from Cancer Prevention Study II. Cancer Res 2003;63:6556–62.
[26] Baris D, Karagas MR, Verrill C, et al. A case–control study of smoking and bladder cancer risk: Emergent patterns over time. J Natl Cancer Inst 2009;101:1553–61.
[27] Liu Y, Pleasants RA, Croft JB, et al. Smoking duration, respiratory symptoms, and COPD in adults aged ≥45 years with a smoking history. Int J COPD 2015;10:1409–16.
[28] Ogino S, Lochhead P, Chan AT, et al. Molecular pathological epidemiology of epigenetics: Emerging integrative science to analyze environment, host, and disease. Mod Pathol 2013. https://doi.org/10.1038/modpathol.2012.214.
[29] Ogino S, Nowak JA, Hamada T, Milner DA, Nishihara R. Insights into pathogenic interactions among environment, host, and tumor at the crossroads of molecular pathology and epidemiology. Annu Rev Pathol 2019;14.
[30] Wilson R, Wahl S, Pfeiffer L, et al. The dynamics of smoking–related disturbed methylation: a two time–point study of methylation change in smokers, non–smokers and former smokers. BMC Genomics 2017;18. https://doi.org/10.1186/s12864–017–4198–0.
[31] Philibert R, Hollenbeck N, Andersen E, et al. Reversion of AHRR demethylation is a quantitative biomarker of smoking cessation. Front Psych 2016. https://doi.org/10.3389/fpsyt.2016.00055.
[32] Scesnaite A, Jarmalaite S, Mutanen P, et al. Similar DNA methylation pattern in lung tumours from smokers and never–smokers with second–hand tobacco smoke exposure. Mutagenesis 2012;27:423–9. https://doi.org/10.1093/mutage/ger092.
[33] Reynolds LM, Magid HS, Chi GC, et al. Secondhand tobacco smoke exposure associations with DNA methylation of the aryl hydrocarbon receptor repressor. Nicotine Tob Res 2017;19:442–51.
[34] CS Wilhelm–Benartzi, Christensen BC, Koestler DC, et al. Association of secondhand smoke exposures with DNA methylation in bladder carcinomas. Cancer Causes Control 2011;22:1205–13.