Blood- and brain-based genome-wide association studies of smoking

Aleksandra D. Chybowska[1], Elena Bernabeu[1], Paul Yousefi[2,3,4], Matthew Suderman[2,3,4], Robert F. Hillary[1], Louise MacGillivray[5], Lee Murphy[5], Sarah E. Harris[6], Janie Corley[6], Archie Campbell[1], Tara L. Spires-Jones[7,8], Daniel L. McCartney[1], Simon R. Cox [6,9], Jackie F. Price[10], Kathryn L. Evans[1], Riccardo E. Marioni[1*]

[1.] Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, EH4 2XU, UK

[2.] Medical Research Council Integrative Epidemiology Unit at the University of Bristol, University of Bristol, Bristol, UK

[3.] NIHR Bristol Biomedical Research Centre, University Hospitals Bristol and Weston NHS Foundation Trust and University of Bristol

[4.] Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK

[5.] Edinburgh Clinical Research Facility, University of Edinburgh, Western General Hospital, Edinburgh, EH4 2XU, UK

[6.] Lothian Birth Cohorts, Department of Psychology, The University of Edinburgh, UK

[7.] Centre for Discovery Brain Sciences, University of Edinburgh, Edinburgh, UK

[8.] UK Dementia Research Institute, University of Edinburgh, Edinburgh, UK

[9.] Scottish Imaging Network, A Platform for Scientific Excellence (SINAPSE) Collaboration, Edinburgh, UK

[10.] Usher Institute, University of Edinburgh, Old Medical School, Teviot Place, Edinburgh, EH8 9AG, UK

*Corresponding author:
Name: Riccardo Marioni
Contact Details: riccardo.marioni@ed.ac.uk

Abstract

Background: Self-reported smoking is often incorporated into disease prediction tools but suffers from recall bias and does not capture passive exposure. Blood-based DNA methylation (DNAm) is an objective way to assess smoking. However, studies have not fully explored tissue-specificity or epigenome-wide coverage beyond array data. Here, we update the existing biomarkers of smoking and conduct a detailed analysis of the associations between blood DNAm and self-reported smoking.

Methods and Findings: A blood-based Bayesian epigenome-wide association study (EWAS) of smoking was carried out in 17,865 Generation Scotland individuals at ~850k CpG sites (Illumina EPIC array). For 24 pairs of smokers and non-smokers a high-resolution approach was implemented (~4 million sites, TWIST methylome panel). A DNAm-derived biomarker of smoking (mCigarette) was tested in the independent Lothian Birth Cohort 1936 (n=882, Illumina 450k array) and in the ALSPAC parents and offspring at four time points (range n=496–1,207). To explore tissue specific signals, EWASs of smoking were run across five brain regions for 14 individuals using DNAm from the EPIC array. Lastly, genome-wide association studies (GWASs) of smoking pack years and an epigenetic score for smoking (GrimAge DNAm pack years) were conducted (n=17,105). The primary EWAS analyses identified two novel genome-wide significant loci, mapping to genes related to addiction and carcinogenesis. Associations with CpG sites which are currently absent from methylation arrays were identified by the high resolution EWAS of smoking (n=48). The mCigarette pack years biomarker showed excellent discrimination across all smoking categories (current, former, never), and outperformed existing predictors in associations with pack years in an external test dataset (Pearson r=0.75). Several CpGs showed near-perfect discrimination of smoking status in both blood and brain, but these loci did not overlap across tissues. The GWAS of DNAm (but not self-reported) pack years identified novel and established smoking-related loci. However, the self-reported phenotype GWAS had a higher genetic correlation with a large meta-analysis GWAS of self-reported pack years. Among the study shortcomings are its potential lack of generalizability to non-Europeans and the absence of serum cotinine data.

Conclusion: A multi-tissue, multi-cohort analysis of the relationship between smoking, DNA and DNAm (assessed via arrays and targeted sequencing) has improved our understanding of the biological consequences of smoking.

1. Introduction

Cigarette smoking remains a leading cause of preventable death and disease, accounting for approximately 8 million global deaths annually (1). It is a major risk factor of more than 50 diseases including cardiovascular disease, lung cancer and dementia (2). As smoking history is often used in clinical risk stratification assessments, enhancing the accuracy of cumulative tobacco consumption measurements has the potential to improve prevention and treatment of smoking-related diseases.

Traditionally, tobacco use has often been quantified using self-report questionnaire-response data, such as pack years or indication of current smoking status (i.e. current, former, never smoker), which are prone to recall bias and do not account for passive smoke exposure (3). A more objective approach to assess smoking is by measuring the concentration of tobacco-related chemicals. However, the commonly used nicotine biomarker, cotinine, has an average half-life of 15–20 hours (4). Consequently, the concentration of serum cotinine does not inform about time since cessation in recent quitters and it cannot help to distinguish former smokers from never smokers. This limitation is especially pertinent when estimating the risk of diseases that take many years to develop, such as cardiovascular disease.

Blood-based DNA methylation patterns show great promise as a long-term biomarker of smoking (5). DNA methylation (DNAm) is a cell- and tissue-specific epigenetic modification of DNA molecule that does not change the DNA sequence itself. It involves the addition of a methyl group to cytosine residues and occurs predominantly at cytosine-phosphate-guanine dinucleotides, also known as CpG sites. CpG methylation levels reflect not only smoking status but also cumulative tobacco exposure and can be informative of time since quitting in former smokers (6–8). In the majority of CpG sites, smoking-related DNAm changes are dose-dependent and reversible after cessation (9).

In previous studies, blood-based DNAm biomarkers of smoking almost perfectly discriminated smokers from never smokers (7,10). However, their ability to differentiate former smokers from never smokers was relatively modest. Additionally, most of these studies relied on whole-blood DNAm assessments using arrays, which only measure a pre-selected subset of CpG sites present in the epigenome. For example, the current largest epigenome wide association study (EWAS) of smoking in adults (n=15,907) was a meta-analysis conducted using the Illumina

450k BeadChip array (n~450,000 CpG sites) (6). The largest EWAS of smoking, in which DNAm was measured with Illumina EPIC array (n~850,000 CpG sites, approximately 5% of all sites on the epigenome), considered blood samples of 15,014 individuals (11).

Here, we update the existing biomarkers of smoking by analysing whole-blood DNAm levels measured with Illumina EPIC array in 17,865 individuals. For a subset of 48 individuals, we implement a high-resolution approach (~4 million CpG sites, TWIST human methylome panel) aimed at measuring methylation levels at CpG sites which are currently absent from arrays. Furthermore, we develop a smoking biomarker (mCigarette) using the EPIC dataset and investigate its associations with self-reported smoking in two external cohorts (Lothian Birth Cohort – LBC1936 and The Avon Longitudinal Study of Parents and Children - ALSPAC). We also investigate variations in methylation patterns in both blood and brain, with DNA methylation measured across five *post-mortem* brain regions in 14 individuals. Finally, we compare the epigenetic proxy of smoking to self-reported smoking by considering the genetic loci associated with these phenotypes in genome-wide association studies (GWASs).

2. Methods

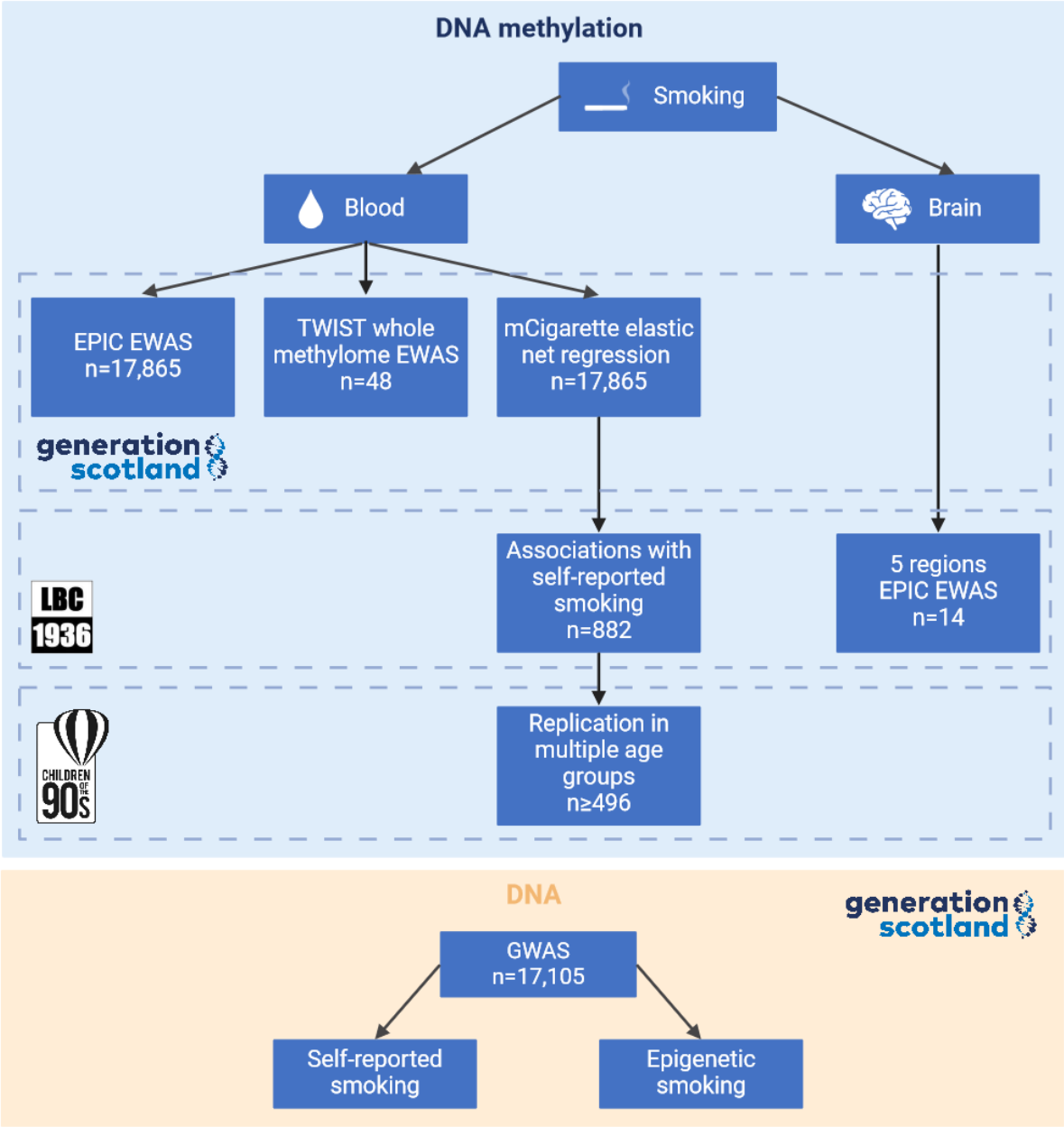

Figure 1. Project overview. *The analysis was conducted using data from three Scottish cohorts: Generation Scotland (GS), the Lothian Birth Cohort of 1936 (LBC1936) and the Avon Longitudinal Study of Parents and Children (ALSPAC). While blood-based DNAm data were available in all three cohorts, LBC1936 also contained information about the DNAm levels in post-mortem brain tissue, covering five brain regions from 14 individuals. A Bayesian epigenome-wide association study (EWAS) of smoking was performed in GS (~850k sites, Illumina EPIC array). For 24 pairs of age- and sex-matched smokers and non-smokers from GS, a high-resolution methylation measurement approach was implemented (~4 million sites, TWIST human methylome panel), followed by an EWAS analysis. An epigenetic biomarker of smoking, mCigarette, was developed in GS and tested as a predictor of self-reported smoking*

in LBC1936. The association between mCigarette and self-reported smoking was replicated in multiple age groups present in ALSPAC. Next, EWASs of smoking were run across five brain regions for 14 individuals using EPIC DNAm from LBC1936. Finally, genome-wide association studies (GWASs) were run in GS to compare genetic signal of self-reported and epigenetic smoking (GrimAge DNAm pack years score). Image created with BioRender.

2.1. Generation Scotland

Volunteer recruitment to Generation Scotland (GS) has been detailed in a previous publication (12). Between 2006 and 2011, patients of collaborating general medical practices in Scotland between 35 and 65 years of age were selected at random and invited to take part in the study. These individuals were then encouraged to recruit family members to volunteer to join the cohort. A total of 24,088 individuals between 18-99 years of age completed a health questionnaire. All individuals were asked to report their smoking status (current smoker, former smoker or never smoker). In addition, current and former smokers provided information about the age they started smoking, the age they quit smoking (former smokers only), and the number of cigarettes they smoked per day. Pack years were computed by multiplying years of smoking by the number of cigarettes smoked per day divided by 20 (number of cigarettes in a pack), and assigning a value of zero to those who never smoked. Further information about the distribution of phenotypic data is available in **Supplemental Table 1**. DNA extracted from whole blood collected at the baseline visit was genotyped using the Illumina HumanOmniExpressExome array (8v1-2 and 8v1) for 19,992 individuals. Following quality control (QC), imputation to the Haplotype Reference Consortium (HRC) panel (13) and post imputation QC, 7,626,922 single nucleotide polymorphisms (SNPs) remained for downstream analyses. Methylation levels were measured with Illumina EPIC850k array (18,413 individuals, 752,722 sites after QC). GWAS and EWAS quality control steps have been described before (10) and are also documented in **Supplemental Tables 2 and 3**. Phenotype pre-processing steps are detailed in **Supplemental Figure 1**.

2.2. Lothian Birth Cohort 1936

The Lothian Birth Cohort of 1936 (n=1,091, LBC1936) comprises community-dwelling older adults in Scotland, most of whom completed an intelligence test aged around 11 years in 1947. Later in life, those living in the Edinburgh and Lothians region were recruited to the cohort at

a mean age of ~70 years and then followed up at 3-yearly intervals. Data collected at each wave comprised cognitive test scores as well as biological measures obtained from blood samples. During a baseline interview at age 70, the participants' self-reported smoking status (never smoker, past smoker, current smoker) and smoking behaviour (age at starting, age at stopping, average number of cigarettes smoked per day) were determined. Pack years were calculated as in GS. A brain tissue bank was established at wave 3 (from age ~76 years). Detailed information about the cohort, brain imaging and post-mortem brain samples can be found in a cohort update and brain protocol papers (14,15). DNAm from whole blood and 5 post-mortem brain tissues has been measured using Illumina Infinium HumanMethylation450 BeadChip array (16). Quality control and processing details are provided in **Supplemental Table 2**. Phenotype pre-processing steps are detailed in **Supplemental Figure 2**.

2.3. ALSPAC

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a cohort study conducted among pregnant women residing in Avon, UK, with expected delivery dates falling between 1st April 1991 and 31st December 1992 (17,18). Out of the 20,248 eligible pregnancies, 14,541 were enrolled to the study, resulting in 14,062 live births, of which 13,988 children survived to age one. During pregnancy, mothers invited fathers to take part in the study. A total of 12,113 fathers completed questionnaires, with 3,807 currently formally enrolled. For a subset of ALSPAC participants (mothers, fathers and children) DNAm was assayed as part of the Accessible Resource for Integrated Epigenomic Studies (ARIES) initiative (19,20). DNA was extracted from blood samples collected at various time intervals between birth and death, and methylation levels were measured using the Illumina Infinium HumanMethylation450 or MethylationEPIC BeadChip arrays. 450,838 CpG sites passed quality control and were common to these methylation arrays. This study used four collections of DNAm data. Antenatal collection includes data from the ALSPAC mothers only, the FOM/FOF collection corresponds to the mothers/fathers at midlife (~50 years) (21), and the F17 and F24 (22) collections contain ALSPAC children at ages 15-17 (time-point '15up') and 24 (time-point 'F24'), respectively. Smoking status for F17 was based on three questionnaires administered ages 14-17. Former smokers reported having quit at the age 17 questionnaire. Current smokers reported that they had smoked weekly in one or more questionnaires but had not quit. Never smokers reported having never smoked at least one time and never reported having smoked. Smoking status for F24 was assessed using six questionnaires administered ages 14-24. Former

smokers reported having smoked regularly at some point prior to age 24, but at age 24 reported not having smoked in the previous 30 days. Current smokers reported at age 24 having smoked in the last 30 days and having smoked at least 50 cigarettes in their lifetime. Never smokers reported never having smoked at age 24. Maternal antenatal smoking was assessed by questionnaires administered at 18 and 32 weeks gestation. Former smokers reported having smoked previously but have stopped smoking for the pregnancy. Current smokers reported smoking regularly in the first trimester. Never smokers reported having never smoked before or during the pregnancy. Smoking status of FOM mothers was assessed using 15 questionnaires administered at study child ages up to age 12 and a final questionnaire at age 18. Former smokers reported having smoked on at least one questionnaire but reported not smoking on the 18y questionnaire. Current smokers reported being current regular smokers on the 18y questionnaire. Never smokers consistently reported never having smoked on questionnaires. Smoking status of FOF fathers was assessed using 11 questionnaires administered at study child ages up to age 12 and a final questionnaire at age 20. Former smokers reported having smoked on at least one questionnaire but reported not smoking on the 20y questionnaire. Current smokers reported being current regular smokers on the 20y questionnaire. Never smokers consistently reported never having smoked on questionnaires.

Data for the study were gathered and administered utilizing REDCap electronic data capture tools, which are hosted at the University of Bristol. REDCap (Research Electronic Data Capture) is a secure web application specifically designed to facilitate data capture for research (23). The study website provides comprehensive details on all available data, accessible through a fully searchable data dictionary located at http://www.bristol.ac.uk/alspac/researchers/our-data/.

2.4. Sequencing-based approach

For 48 unrelated smokers and non-smokers from GS, a whole methylome approach was implemented (TWIST methylome panel, ~4 million CpG sites). To ensure a robust methylation signal would be present when comparing cases against controls, only heavy smokers (12 males and 12 females with between 53.9 and 87.7 pack years of tobacco use) were selected as cases. Controls were matched by age and sex using R package called Matchit (24). Sequencing was performed by the Edinburgh Clinical Research Facility using the TWIST Human Methylome Panel and according to TWIST Targeted Methylation Sequencing Protocol (25). Pre-

processing of raw FASTQ files, read aligning to human reference genome (GRCh38, n=29,401,795 total reference CpGs) with bwa-meth and quality-control of the results was performed using MethylSeq bioinformatics analysis pipeline, version 2.2.0 (26,27). This analysis yielded information about the methylation level and depth of coverage (DoC) at 18,248,472 covered CpG sites, which were subsequently processed in R version 4.2.2 (28) using Methrix package (29). As part of post-processing, loci a) of extremely low (minimal DoC = 2) and high coverage (beyond 0.99 quantile), and b) overlapping with known cytosine to thymine polymorphisms were removed from the methylation dataset. Finally, a coverage filter was applied, retaining only the loci that were covered in at least 40 samples by 10 or more reads. This left an analysis sample of 3,391,718 CpGs, which were annotated with Annotatr R package (30).

2.5. Blood-based DNAm EWAS

A blood-based epigenome-wide association study (EWAS) of smoking was carried out in 17,865 GS individuals using BayesR (31). As BayesR implicitly corrects for confounding effects without requiring a full characterization of all covariates, our models were only adjusted for measured variables i.e., estimated white blood cell proportions were not included. Before running the EWAS, each CpG was corrected for the effects of age, sex, and batch using linear regression (saving the residuals from each model as the new variable). The smoking phenotype (measured in pack years, with a pack defined as 20 cigarettes) was natural log+1 transformed and adjusted for age and sex using linear regression (again, the residuals were saved and used for the downstream analyses). Both the smoking and the CpG variables were scaled to have mean of 0 and variance of 1. The data served as inputs of a Bayesian penalised linear regression model. Four Gaussian priors were specified to model CpGs with varying effect sizes (mixture variances of 0.1%, 1%, 10% and 100%) along with a discrete spike at the origin to model CpGs with no effect. A Gibbs sampler was used to sample over the posterior distribution, conditioning on the input data. A burn-in of 5,000 samples was used, after which every fifth sample was retained across 10,000 iterations. A CpG with a posterior inclusion probability of greater than 0.95 was considered as epigenome-wide significant. Previously unpublished associations were identified by searching the literature and the EWAS catalog (32).

2.6. Comparison between TWIST and EPIC850k

DNAm of 24 pairs of smokers and non-smokers from GS was profiled using both the EPIC array and the TWIST platform (see **Methods - Sequencing-based approach**). For each DNAm profiling method, an EWAS of smoking (pack years) was conducted. The association between DNAm level at each CpG site (outcome) and binary smoking status, age and sex was modelled using linear regression. Given the small sample size (n=48), the significance threshold was set at $P<1x10^{-5}$. The results were compared on a Manhattan plot generated with the CMplot R package (33).

2.7. Biomarker of cumulative smoking

An elastic net biomarker of pack years (mCigarette) was trained in 17,865 GS individuals using the glmnet library in R (34). As part of data pre-processing, CpG sites were filtered to 18,760 loci associated with smoking at False Discovery Rate (FDR)<0.05 in a previous meta-analysis EWAS (n=18,760) of tobacco use (Joehanes *et al.*, 2016). Alpha was fixed at 0.5 and the lambda value that minimised the mean prediction error was selected via 10-fold cross validation. The selected model assigned non-zero coefficients to 1,255 CpGs.

Next, mCigarette was tested in wave 1 of LBC1936 (n=882, mean age 70 years) and ALSPAC (range n=496–1,207, four time points). In both cohorts, the biomarker was projected to blood-based DNAm and benchmarked against three epigenetic scores for smoking: EpiSmokEr score (7,35,36), and two scores derived from previous GS analyses on smaller subsets of the dataset - one based on BayesR weights (31), the other via lasso penalised regression by McCartney *et al.* (10). In LBC1931, the predictive performance of mCigarette was additionally compared to that of GrimAge DNAm pack years (37). Pearson's r was calculated to estimate the degree of correlation between self-reported pack years of smoking and the studied score. The amount of variance in pack years explained by the studied score was assessed by comparing R^2 estimates of null and full models. While the null model was adjusted for age and sex, the full model also included the studied score. Incremental R^2 was calculated as the difference between variance explained by null and full models.

The ability of the scores to distinguish between current, former, and never smokers was assessed by the area under the receiver operating characteristic curve (AUC). Receiver operating characteristic (ROC) curves were produced using pROC R package (38). Additional prediction performance metrics such as the area under precision-recall curve (PRAUC) were obtained using MLmetrics R package (39).

2.8. Tissue specificity analyses in LBC1936

DNAm was measured in 5 brain regions (hippocampus - BA35, dorsolateral prefrontal cortex – BA46, primary visual cortex - BA17, anterior cingulate cortex - BA24, ventral/lateral inferior temporal cortex - BA20/21) from post-mortem brain samples of 14 LBC1936 individuals, with one sample missing from hippocampus. Tissue acquisition and processing details are detailed in Stevenson *et al.* (16). Using blood-DNAm measured at wave 1 and brain-DNAm data, exploratory EWAS analyses were performed (CpG ~ age + sex + smoking). In these analyses, smoking was treated as a categorical variable (0 = never smoked, 1 = former smoker, 2 = current smoker). Given the small sample size, nominally significant CpG-smoking associations were defined as having $P<1x10^{-5}$ and were displayed using a ggplot2 boxplot. Due to the same constraints, an enrichment analysis for significant associations was not carried out.

2.9. GWAS of smoking

Associations between genetic variants and smoking were examined using genome-wide association studies (GWASs). Two phenotypes were considered: natural log(transformed pack years of smoking + 1) and an epigenetic score for smoking pack years generated by an online calculator that uses the algorithm derived for the GrimAge epigenetic clock (37). After initial filtering (see **Supplemental Table 3**), there were 17,105 GS individuals with data available at 7,626,922 SNPs. Both GWASs were conducted using the GCTA software (40), with a Genetic Relationship Matrix fitted into a fastGWA-lmm model to account for relatedness. Each trait was adjusted for age, sex, and 20 genetic principal components. The results of these analyses, along with the findings of previous GWASs of tobacco use, were visualised using CMplot (33). Lead and methylation quantitative loci among the results of GrimAge DNAm pack years GWAS were identified using the default settings in FUMA (41) and goDMC (42), respectively. GWAS catalog was accessed via FUMA website. Genetic correlations were calculated with LDSC (43).

3. Results

3.1. EWAS of smoking

First, we ran a Bayesian EWAS of smoking. There were 17,865 GS volunteers with a measure of pack years of smoking and Illumina EPIC DNAm data. The average age of the sample was 47.6 years (standard deviation [SD] = 14.9), and 59.1% of the participants were female (**Supplemental Table 1**). In the BayesR EWAS, DNAm explained 50.0% (95% Credible Interval 46.0 – 53.9) of the variance in the pack years phenotype. Forty-two independent CpGs were associated with smoking at posterior inclusion probability (PIP)>80% and 26 associations at PIP>95% (**Supplemental Table 4**).

Of the 26 associations significant at PIP>95%, 24 were previously reported at $P<1x10^{-4}$ (with 23 having $P<1x10^{-7}$) in the EWAS catalogue (32), a resource that curates findings from published EWAS studies. Two associations with PIP>95% were novel: cg02517189 and cg00562553. The former CpG site was annotated to *GRIK5*, which encodes glutamate receptor, while the latter was annotated to a homeobox family gene called *HOXA4*.

3.2. Comparison of Illumina EPIC850k and TWIST human methylome panel

Next, we extended this analysis by running high resolution EWASs of smoking on a subset of 24 pairs (n=48) of current vs never smokers with both Illumina EPIC array (~850k CpG sites) and TWIST human methylation panel (4 million CpG sites) data. In the EPIC-based analysis, 44 CpG sites showed associations with smoking status, whereas TWIST-based analysis revealed 97 associations at $P<1x10^{-5}$. Of the 97 CpG significant sites identified in the TWIST-based analysis, 10 were available on the EPIC array and overlapped between the two methods. **Figure 2** presents a comparison of the results obtained from the TWIST- and EPIC-based EWAS.

The results of TWIST EWAS revealed associations that had been previously reported in the EWAS catalog (based on DNAm profiled with array technologies), as well as novel loci. Some of the previously reported associations were *ME3* (chr11-86632814-86632815, beta=-0.12, $P=3.1x10^{-6}$) and *SEMA4B* (chr15-90185008-90185009, beta=-0.08, $P=4.0x10^{-6}$). These loci were not identified in the n=48 EPIC analysis. Novel TWIST associations included *TSPAN5*

(chr4-98472405-98472406, beta=-0.06, SE=0.01, P=1.8x10^{-6}), *SST* (chr3-187670342-187670343, beta=-0.08, SE=0.01, P=5.6x10^{-7}) and *USP42* (chr7-6126706-6126707, beta=0.05, SE=0.01, P=1.0x10^{-7}). Further details are given in **Supplemental Table 5**.

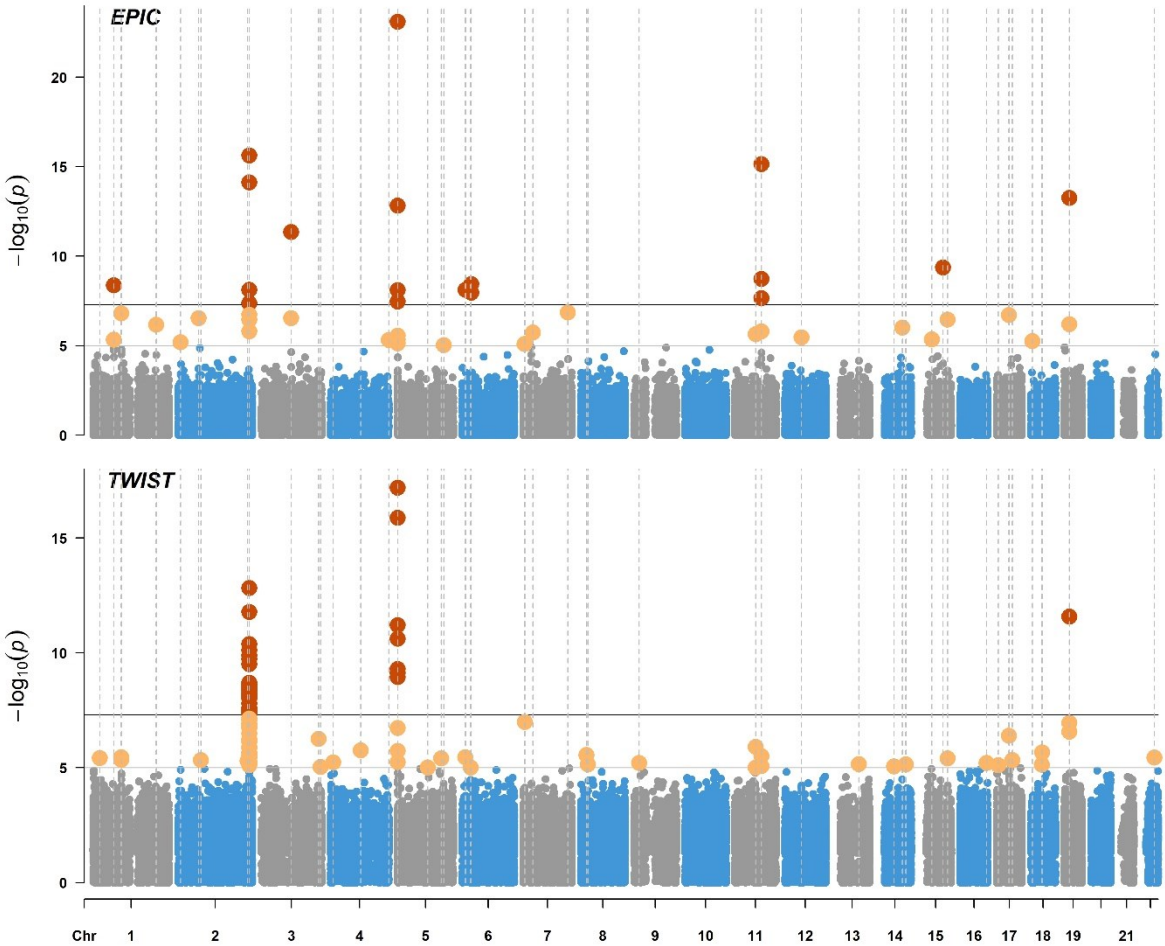

Figure 2. *EWAS of current versus never smokers in Generation Scotland (n=24 pairs). Analyses were performed using DNAm data obtained using the Illumina EPIC array (~850k CpG sites, top panel) and the TWIST human methylation panel (4 million CpG sites, bottom panel). While yellow dots correspond to CpGs significant at 1x10^{-5}, red dots represent CpGs significant at 3.6x10^{-8}. Dotted vertical lines indicate loci associated with smoking status at P<1x10^{-5}.*

3.3. DNAm biomarker of cigarette consumption - mCigarette

A DNAm biomarker of cigarette consumption, mCigarette, was then developed. The biomarker was trained in GS (n=17,865) using elastic net regression with 10-fold cross-validation. Prior to training, CpG sites were prefiltered to those associated with tobacco use at FDR<0.05

(n_{CpG}=18,760) in the previous largest EWAS of smoking (6). This meta-analysis did not include GS and was conducted using Illumina 450k BeadChip array. Following 10-fold cross validation, an optimal lambda value that minimised the mean prediction error was selected (lambda=0.012577) and fed into an elastic net model of smoking pack years. As a result, non-zero coefficients were assigned to 1,255 CpG sites (**Supplemental Table 6**).

The biomarker was tested in the external LBC1936 study (n=882, mean age 69.6 years, SD=0.8). Assessing the predictive performance using Area Under the Curve (AUC) (**Figure 3**) revealed very good (AUC=0.85) to near perfect (AUC=0.98) ability to distinguish between current (n=101), former (n=368) and never smokers (n=413). Similar results were obtained when Area Under the Precision-Recall Curve was used as a performance metric ($PRAUC_{current\ vs\ never}$=0.96, $PRAUC_{former\ vs\ never}$ = 0.85, PRAUC current vs former = 0.71).

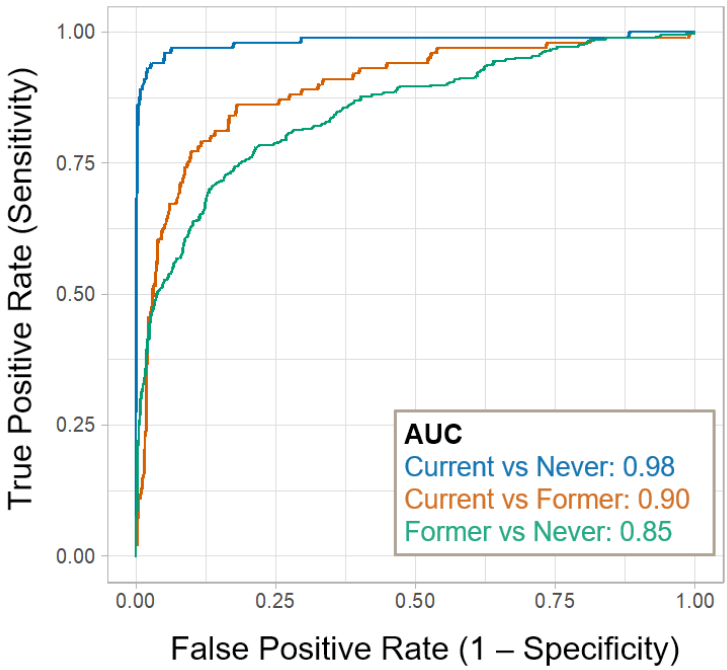

***Figure 3**. Predictive performance of the epigenetic biomarker of smoking. The areas under receiver operating characteristic curves represent the ability of the model to distinguish between the following smoking categories: blue for current smokers vs. never smokers, orange for current smokers vs. former smokers, and green for former smokers vs. never smokers.*

The predictive performance of mCigarette in was benchmarked against four previously developed epigenetic scores for smoking in wave 1 of LBC1936 (n=882). The results of this comparison can be found in **Table 1.** mCigarette yielded improved incremental R^2 estimates compared to EpiSmokEr, the score developed by McCartney *et al.* and GrimAge DNAm pack years.

Metric	*EpiSmokEr*	*BayesR*	*McCartney et al*	*GrimAge*	*mCigarette*
N training	1,793	9,448	5,087	1,731	17,865
a) Variance explained in measured pack years (R^2) and correlation (r) metrics					
Incremental R^2	0.351	0.514	0.330	0.419	0.534
r	0.610	0.735	0.581	0.664	0.750
b) Binary classification performance (AUC)					
Current vs Never	0.985	0.977	0.982	0.983	0.984
Current vs Former	0.916	0.853	0.921	0.939	0.897
Former vs Never	0.755	0.846	0.725	0.815	0.852

Table 1. *Benchmarking of mCigarette against five biomarkers of smoking: EpiSmokEr* (7)*, GrimAge DNAm pack years* (37)*, and three scores developed using GS data: BayesR score* (31)*, a score developed by McCartney et al.* (10) *and a score developed in this study - mCigarette. The BayesR score and the score developed by McCartney et al. were trained on smaller subsets of the GS DNAm dataset. Table a) compares the variance in self-reported pack years explained by null and full models. While the null model was adjusted for age and sex, the full model also included the studied score. The difference between variance explained by null and full models is denoted as the incremental R2. Pearson's correlation coefficient is referred to as r. b) Performance of the studied scores in distinguishing between different smoking categories. Binary classification performance of the scores was measured using Areas Under the Curve (AUC).*

Subsequently, the weights used to construct the studied scores (apart from the GrimAge DNAm pack years, as weights are not publicly available) were applied to methylation data in the ALSPAC cohort. Within this dataset, no single methylation score consistently outperformed others in distinguishing between current, former, and never smokers across studied age groups. While in young adults (mean age 18 and 24 years old) EpiSmokEr and the score constructed by McCartney *et al*. achieved the highest AUCs (median AUC=0.720 [IQR:0.642-0.763]), mCigarette and EpiSmokEr showed excellent performance in older adults (mean age 29 and 50 years old, median AUC=0.890 [IQR:0.771-0.931]). Full results of the replication study are available in **Supplemental Table 7**.

3.4. Tissue specificity

To assess whether the findings translated across different tissue types, the association between tobacco use and DNAm levels across five brain regions were explored using *post-mortem* samples from LBC1936 (n=14, $n_{hippocampus}$=13). At significance threshold $P<1x10^{-5}$, five loci in the hippocampus (BA35), one locus in dorsolateral prefrontal cortex (BA46), four loci in primary visual cortex (BA17), nine loci in anterior cingulate cortex (BA24) and three loci in ventral/lateral inferior temporal cortex (BA20/21) were associated with smoking status (**Supplemental Table 8**). There was no overlap between the significant loci across the studied brain regions.

Some loci demonstrated nearly perfect discrimination of smoking status in blood and brain; however, these loci did not overlap (**Figure 4**). For instance, the methylation status at cg0557592, annotated to the *AHRR* gene, is a well-established marker of smoking status in whole-blood DNAm. However, this marker did not discriminate smoking category in hippocampal DNAm. On the other hand, cg26381592, annotated to the *PMS1* gene, did not effectively distinguish smokers in blood samples, but it exhibited a strong correlation with smoking status in brain tissue samples.

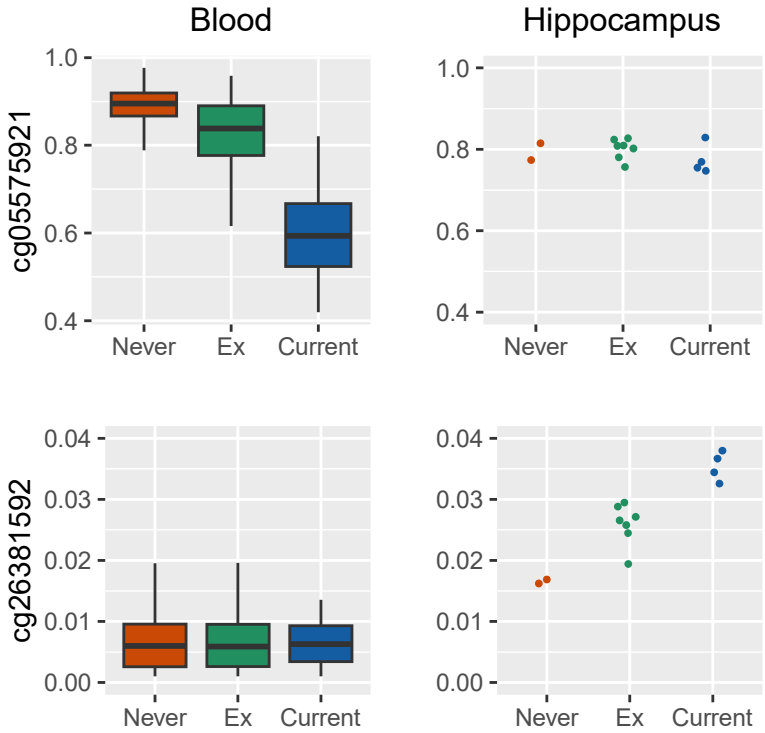

Figure 4. *CpG methylation levels in blood (LBC1936 baseline, n=882) and brain across different self-reported smoking categories (LBC1936, n=13).*

Subsequently, we tested the performance of a blood-derived DNAm biomarker of smoking (mCigarette) in the brain tissue. When applied to brain DNAm data, mCigarette did not distinguish between smoking categories in any of the studied brain regions (**Supplemental Figure 3**).

3.5. GWAS of self-reported and epigenetic smoking

Finally, GWASs of self-reported and epigenetic smoking were conducted. To avoid overfitting, we used the GrimAge DNAm pack years estimator (trained externally in 1,731 individuals from the Framingham Heart Cohort study (37)) instead of mCigarette, which was trained in GS.

In 17,105 GS individuals, there was a SNP-based heritability of 27.3% ($P=7.0x10^{-46}$) for self-reported pack years of smoking compared to 41.0% ($P=8.2x10^{-98}$) for GrimAge DNAm pack years (**Figure 5**). At the genome-wide significance level of $P<5x10^{-8}$, only one locus

(rs117836409 annotated to *GDPD1*) was associated with self-reported pack years, in contrast to 39 SNPs (three lead SNPs at two genomic risk loci) associated with GrimAge DNAm pack years (detailed in **Supplemental Tables 9** and **10**). Two lead loci (rs1800440 and rs6495309, annotated to *CYP1B1* and *CHRNA3 - CHRNB4,* respectively) have been previously documented in the GWAS catalogue, which aggregates data from published GWAS studies (44) (**Supplemental Table 11**). They have been associated with carcinogenesis and nicotine dependence, respectively (45,46). The SNP not previously annotated in the GWAS Catalog (rs114342890) maps to *RMDN2:RMDN2-AS1*, a long non-coding RNA previously studied in relation to eosinophil counts and melanoma (47,48). According to GeneHancer, an online database of enhancers, promoters and their inferred targets, all 27 regulatory elements which target *RMDN2:RMDN2-AS1* also regulate the expression of *CYP1B1* (49).

The three lead SNPs have been characterised as methylation quantitative trait loci (mQTLs). They all act in *cis* on 43 CpGs (**Supplemental Table 12**). Only one of these CpGs (cg06264984) is featured in mCigarette, while none are included in EpiSmokEr – the list of CpGs included for the GrimAge DNAm pack years is not publicly available.

Next, the GrimAge DNAm pack years GWAS results were compared to previously published GWAS studies of tobacco use (**Supplemental Table 13**). At a significance level of $P<5x10^{-8}$, seven SNPs annotated to *CHRNA3* and *CHRNA5* aligned with the findings of the largest pack years GWAS to date (n=131,892) (50). Thirty-seven SNPs, mapping to *CHRNA3, CHRNA5,* and *CHRNB4,* overlapped at $P<5x10^{-8}$ with the results of a related phenotype, cigarettes per day (n= 618,489) (51).

The genetic correlation (r_g) between GrimAge DNAm pack years and pack years from Erzurumluoglu et al. (50) was 0.51 (SE=0.12, $P=8.3x10^{-6}$). By contrast, the r_g between self-reported pack years in the 17,105 GS and 131,892 meta-analysis individuals was 0.67 (SE=0.19, $P=5.0x10^{-4}$). Additional information on the genetic correlation between epigenetic smoking and previously studied self-reported smoking behaviours (ranging from -0.62 to 0.72) (51) can be found in **Supplemental Figure 4.**

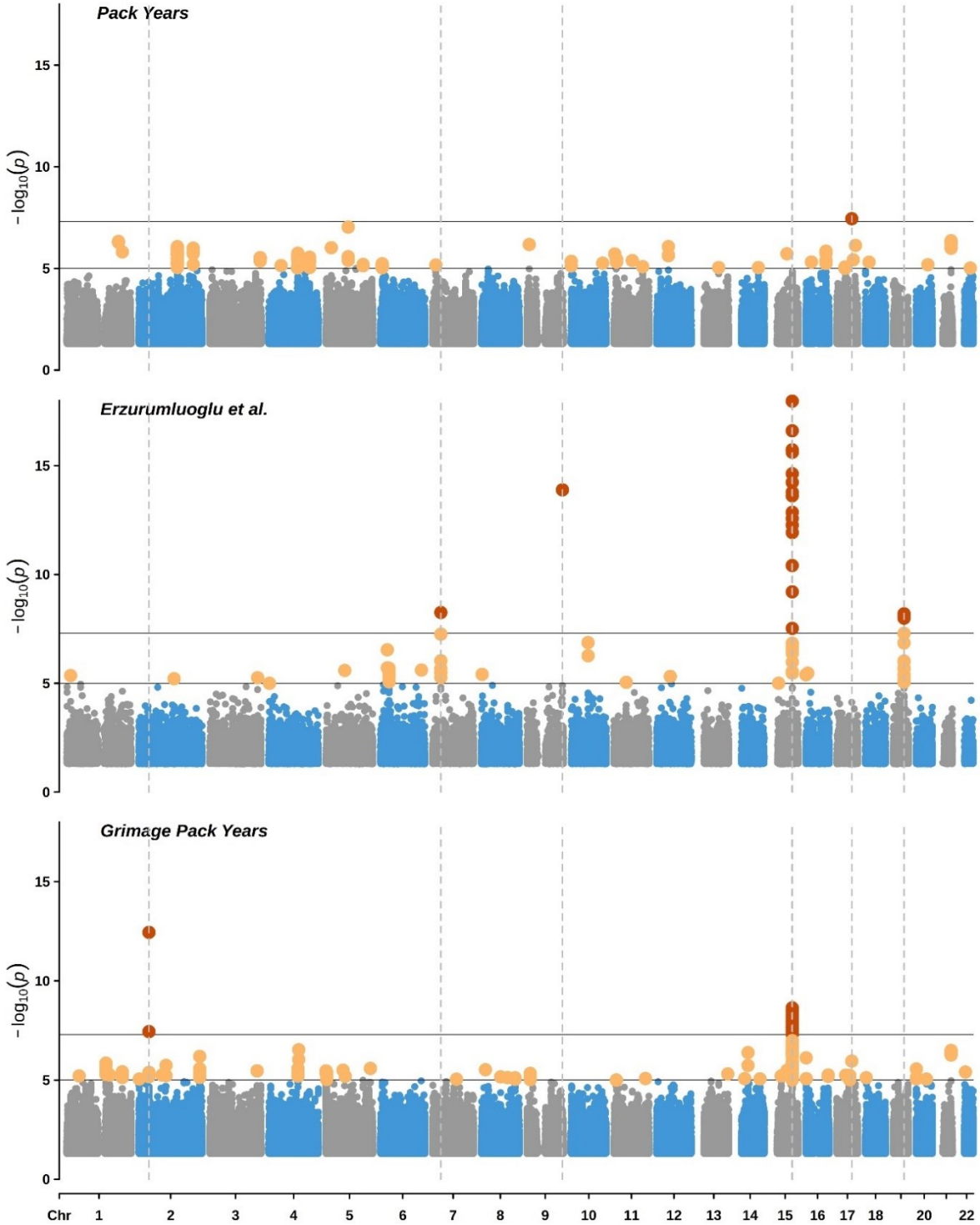

Figure 5. *Manhattan plots visualising overlap between various GWASs of smoking. Studied phenotypes included a) pack years (Generation Scotland, n=17,105), b) pack years (largest meta-analysis to date, n=131,892), c) GrimAge-based DNAm estimator of smoking pack years (n=17,105). Yellow dots represent SNPs that are significant at* 1×10^{-5}*, while red dots correspond to SNPs that are significant at* $P<5x10^{-8}$*. Dotted vertical lines mark genomic positions of significant associations at* $P<5x10^{-8}$*.*

4. Discussion

This multi-tissue, multi-cohort analysis of the relationship between smoking and DNAm (assessed via arrays and targeted sequencing) has improved both our understanding of the biological consequences of smoking and our ability to measure it objectively. The array-based study, which identified two novel loci, represents the largest single cohort EWAS of smoking and the largest EPIC array EWAS of smoking, to date. The updated epigenetic biomarker of tobacco-use, mCigarette, reliably predicted smoking status and was strongly correlated with self-reported pack years of tobacco use. The analysis of sites differentially methylated in the brains of smokers and non-smokers revealed evidence of tissue-specific signals. There was a partial overlap between the results of the GrimAge DNAm pack years GWAS conducted in GS (n=17,105) and the most extensive GWAS of self-reported smoking to date (n=131,892).

Of the novel loci identified in the EWASs, the two array-based CpGs are annotated to *GRIK5* and *HOXA4*. *GRIK5* encodes a receptor for glutamate (52), a neurotransmitter of the central nervous system that plays an important role in addiction (53). *HOXA4*, in turn, is a homeobox domain gene that is normally involved in embryonic development. Homeobox genes are abnormally expressed in cancer cells and changes in the expression of *HOXA4* has been specifically associated with colorectal, ovarian and lung cancer (54).

The twenty loci (at a threshold of $P<1x10^{-5}$) from the substantially less well-powered targeted sequencing analysis included CpGs that mapped to *TSPAN5, USP42* and *SST* – genes linked to carcinogenesis. *TSPAN5* regulates expression of tumour suppressor genes and is required for the growth of hepatocellular carcinoma (55). *USP42,* belongs to the ubiquitin C-terminal hydrolases family that is involved in the pathogenesis of various human malignancies including head and neck cancer (56). Lastly, *SST*, which encodes somatostatin, a growth hormone-inhibiting hormone, implicated in the development of pancreatic cancer (57).

When compared to previously published epigenetic biomarkers of tobacco use, mCigarette was a better predictor of smoking pack years. It also showed an excellent performance in discriminating smoking status (current, former, never) in two external cohorts. While individual CpG sites offered excellent discrimination of cases and controls, these loci varied

by tissue. This is consistent with the low correlations in methylation patterns between blood and brain tissue reported previously (58). Future work should explore if tissue-specific signals identify pathways and mechanisms by which smoking influences brain health.

In GS, the GrimAge DNAm pack years GWAS results did not align with self-reported smoking findings but did show partial overlap of lead loci with the most extensive GWAS of self-reported smoking to date. This may suggest an increased power to detect significant loci when the epigenetic score is analysed as a phenotype. However, the genetic correlation between GrimAge DNAm pack years and the meta-analysis smoking pack years was considerably lower than the latter and self-reported pack years in GS. The shared loci between the GrimAge DNAm pack years GWAS and previous self-reported smoking GWAS included *CHRNA3, CHRNA5,* and *CHRNB4*. These genes encode subunits of the nicotinic acetylcholine receptor, responsible for neurotransmission and binding of nicotine in the brain. Variations in these genes can affect nicotine dependence and may be associated with neurological conditions as well as lung cancer.

The key strengths of this study include high resolution of epigenetic data used in the main EWAS analysis (both sample size and the number of measured CpG sites), diverse DNAm profiling techniques and the availability of multi-tissue DNAm. Using targeted sequencing allowed for identifying associations between smoking and loci not included on the EPIC chip. Given the mean age of GS volunteers, mCigarette represents many years of exposure to cigarette smoke, and is likely to capture long term DNAm changes associated with smoking.

Limitations of this study include its potential lack of generalizability to non-Europeans and the small number of brain- and TWIST-DNAm samples. Nevertheless, given that smoking is associated with major epigenetic alternations, the effects of tobacco use are detectable, even in small datasets. The absence of serum cotinine concentrations prevented us from comparing mCigarette against a clinically established smoking biomarker. We were also unable to verify self-reported smoking status. To mitigate potential bias, we cross-referenced the reported smoking status with other variables (such as self-reported smoking initiation and cessation) and eliminated records with conflicting responses.

In conclusion, this study explored methylation patterns associated with smoking in blood and brain. Novel loci were identified from the blood-based analyses (through both sequencing- and

array-based approaches), in addition to the development of a highly accurate blood-based DNAm biomarker for smoking, and gaining insights into differential effects of smoking in the blood and brain. Together, these enhance our understanding of the epigenetic architecture of smoking and shed light on the molecular mechanisms by which tobacco use influences health.

5. References

1. GBD 2019 Tobacco Collaborators. Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019. Lancet. 2021 Jun 19;397(10292):2337–60.

2. West R. Tobacco smoking: Health impact, prevalence, correlates and interventions. Psychol Health. 2017 Aug 3;32(8):1018–36.

3. Connor Gorber S, Schofield-Hurwitz S, Hardt J, Levasseur G, Tremblay M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. Nicotine Tob Res. 2009 Jan;11(1):12–24.

4. Murphy SE, Wickham KM, Lindgren BR, Spector LG, Joseph A. Cotinine and trans 3'-hydroxycotinine in dried blood spots as biomarkers of tobacco exposure and nicotine metabolism. J Expo Sci Environ Epidemiol. 2013;23(5):513–8.

5. Breitling LP. Current Genetics and Epigenetics of Smoking/Tobacco-Related Cardiovascular Disease. Arteriosclerosis, Thrombosis, and Vascular Biology. 2013 Jul;33(7):1468–72.

6. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic Signatures of Cigarette Smoking. Circulation: Cardiovascular Genetics. 2016 Oct;9(5):436–47.

7. Bollepalli S, Korhonen T, Kaprio J, Anders S, Ollikainen M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. Epigenomics. 2019 Oct;11(13):1469–86.

8. Christiansen C, Castillo-Fernandez JE, Domingo-Relloso A, Zhao W, El-Sayed Moustafa JS, Tsai PC, et al. Novel DNA methylation signatures of tobacco smoking with trans-ethnic effects. Clin Epigenetics. 2021 Feb 16;13(1):36.

9. McCartney DL, Stevenson AJ, Hillary RF, Walker RM, Bermingham ML, Morris SW, et al. Epigenetic signatures of starting and stopping smoking. EBioMedicine. 2018 Nov;37:214–20.

10. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. Genome Biology. 2018 Sep 27;19(1):136.

11. Hoang TT, Lee Y, McCartney DL, Kersten ETG, Page CM, Hulls PM, et al. Comprehensive evaluation of smoking exposures and their interactions on DNA methylation. eBioMedicine [Internet]. 2024 Feb 1 [cited 2024 Apr 16];100. Available from: https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(23)00522-4/fulltext

12. Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ, et al. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. BMC Med Genet. 2006 Dec;7(1):74.

13. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016 Oct;48(10):1279–83.

14. Henstridge CM, Jackson RJ, Kim JM, Herrmann AG, Wright AK, Harris SE, et al. Post-mortem brain analyses of the Lothian Birth Cohort 1936: extending lifetime cognitive and brain phenotyping to the level of the synapse. Acta Neuropathologica Communications. 2015 Sep 4;3(1):53.

15. Taylor AM, Pattie A, Deary IJ. Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. International Journal of Epidemiology. 2018 Aug 1;47(4):1042–1042r.

16. Stevenson AJ, McCartney DL, Gadd DA, Shireby G, Hillary RF, King D, et al. A comparison of blood and brain-derived ageing and inflammation-related DNA methylation signatures and their association with microglial burdens. European Journal of Neuroscience. 2022;56(9):5637–49.

17. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2013 Feb;42(1):111–27.

18. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol. 2013 Feb;42(1):97–110.

19. Relton CL, Gaunt T, McArdle W, Ho K, Duggirala A, Shihab H, et al. Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). International Journal of Epidemiology. 2015 Aug 1;44(4):1181–90.

20. Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinformatics. 2018 Dec 1;34(23):3983–9.

21. Northstone K, Suarez-Perez A, Matthews S, Crawford M, Timpson N. The Avon Longitudinal Study of Parents and Children - a resource for COVID-19 research: questionnaire data capture July 2021 to December 2021, with a focus on long COVID [version 2; peer review: 1 approved]. Wellcome Open Research. 2023;8(292).

22. Northstone K, Lewcock M, Groom A, Boyd A, Macleod J, Timpson N, et al. The Avon Longitudinal Study of Parents and Children (ALSPAC): an update on the enrolled sample of index children in 2019. Wellcome Open Res. 2019;4:51.

23. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr;42(2):377–81.

24. Ho D, Imai K, King G, Stuart EA. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software. 2011 Jun 14;42:1–28.

25. Twist Targeted Methylation Sequencing Protocol [Internet]. 2023 [cited 2023 Jun 5]. Available from: https://www.twistbioscience.com/resources/protocol/twist-targeted-methylation-sequencing-protocol

26. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017 Apr;35(4):316–9.

27. Ewels P, Hüther P, Hammarén R, mashehu, Peltzer A, bot nf core, et al. nf-core/methylseq: nf-core/methylseq version 1.6.1 [Nauseous Serpent] [Internet]. Zenodo; 2021. Available from: https://doi.org/10.5281/zenodo.4744708

28. R: The R Project for Statistical Computing [Internet]. 2023 [cited 2023 Jun 5]. Available from: https://www.r-project.org/

29. Mayakonda A, Schönung M, Hey J, Batra RN, Feuerstein-Akgoz C, Köhler K, et al. Methrix: an R/Bioconductor package for systematic aggregation and analysis of bisulfite sequencing data. Bioinformatics. 2020 Dec 1;36(22–23):5524–5.

30. Cavalcante RG, Sartor MA. annotatr: genomic regions in context. Bioinformatics. 2017 Aug 1;33(15):2381–3.

31. Trejo Banos D, McCartney DL, Patxot M, Anchieri L, Battram T, Christiansen C, et al. Bayesian reassessment of the epigenetic architecture of complex traits. Nat Commun. 2020 Jun 8;11:2865.

32. Battram T, Yousefi P, Crawford G, Prince C, Sheikhali Babaei M, Sharp G, et al. The EWAS Catalog: a database of epigenome-wide association studies. Wellcome Open Res. 2022;7:41.

33. Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, et al. rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool for Genome-wide Association Study. Genomics, Proteomics & Bioinformatics. 2021 Aug 1;19(4):619–28.

34. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010 Feb 2;33:1–22.

35. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. PLOS ONE. 2013 May 17;8(5):e63812.

36. Elliott HR, Tillin T, McArdle WL, Ho K, Duggirala A, Frayling TM, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clinical Epigenetics. 2014 Feb 3;6(1):4.

37. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging (Albany NY). 2019 Jan 21;11(2):303–27.

38. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011 Mar 17;12(1):77.

39. Yan Y. CRAN - Package MLmetrics [Internet]. 2024 [cited 2024 Apr 26]. Available from: https://cran.r-project.org/web/packages/MLmetrics/index.html

40. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011 Jan 7;88(1):76–82.

41. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017 Nov 28;8(1):1826.

42. Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. Nat Genet. 2021 Sep;53(9):1311–21.

43. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015 Mar;47(3):291–5.

44. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 2022 Nov 9;51(D1):D977–85.

45. Flora AV, Zambrano CA, Gallego X, Miyamoto JH, Johnson KA, Cowan KA, et al. Functional characterization of SNPs in *CHRNA3/B4* intergenic region associated with drug behaviors. Brain Research. 2013 Sep 5;1529:1–15.

46. Choquet H, Jiang C, Yin J, Kim Y, Hoffmann TJ, 23andMe Research Team, et al. Multi-ancestry genome-wide meta-analysis identifies novel basal cell carcinoma loci and shared genetic effects with squamous cell carcinoma. Commun Biol. 2024 Jan 5;7(1):33.

47. Ransohoff KJ, Wu W, Cho HG, Chahal HC, Lin Y, Dai HJ, et al. Two-stage genome-wide association study identifies a novel susceptibility locus associated with melanoma. Oncotarget. 2017 Mar 14;8(11):17586–92.

48. Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am J Hum Genet. 2019 Jan 3;104(1):65–75.

49. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford). 2017 Apr 17;2017:bax028.

50. Erzurumluoglu AM, Liu M, Jackson VE, Barnes DR, Datta G, Melbourne CA, et al. Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. Mol Psychiatry. 2020 Oct;25(10):2392–409.

51. Saunders GRB, Wang X, Chen F, Jang SK, Liu M, Wang C, et al. Genetic diversity fuels gene discovery for tobacco and alcohol use. Nature. 2022 Dec;612(7941):720–4.

52. Unlu G, Gamazon ER, Qi X, Levic DS, Bastarache L, Denny JC, et al. GRIK5 Genetically Regulated Expression Associated with Eye and Vascular Phenomes:

Discovery through Iteration among Biobanks, Electronic Health Records, and Zebrafish. Am J Hum Genet. 2019 Mar 7;104(3):503–19.

53. D'Souza MS, Markou A. The "Stop" and "Go" of Nicotine Dependence: Role of GABA and Glutamate. Cold Spring Harb Perspect Med. 2013 Jun;3(6):a012146.

54. Cheng S, Qian F, Huang Q, Wei L, Fu Y, Du Y. HOXA4, down-regulated in lung cancer, inhibits the growth, motility and invasion of lung cancer cells. Cell Death Dis. 2018 Apr 27;9(5):1–13.

55. Schreyer L, Mittermeier C, Franz MJ, Meier MA, Martin DE, Maier KC, et al. Tetraspanin 5 (TSPAN5), a Novel Gatekeeper of the Tumor Suppressor DLC1 and Myocardin-Related Transcription Factors (MRTFs), Controls HCC Growth and Senescence. Cancers (Basel). 2021 Oct 26;13(21):5373.

56. Rong C, Zhou R, Wan S, Su D, Wang SL, Hess J. Ubiquitin Carboxyl-Terminal Hydrolases and Human Malignancies: The Novel Prognostic and Therapeutic Implications for Head and Neck Cancer. Frontiers in Oncology [Internet]. 2021 [cited 2023 Jun 1];10. Available from: https://www.frontiersin.org/articles/10.3389/fonc.2020.592501

57. Manoochehri M, Wu Y, Giese NA, Strobel O, Kutschmann S, Haller F, et al. SST gene hypermethylation acts as a pan-cancer marker for pancreatic ductal adenocarcinoma and multiple other tumors: toward its use for blood-based diagnosis. Molecular Oncology. 2020;14(6):1252–67.

58. Walton E, Hass J, Liu J, Roffman JL, Bernardoni F, Roessner V, et al. Correspondence of DNA Methylation Between Blood and Brain Tissue and Its Application to Schizophrenia Research. Schizophr Bull. 2016 Mar;42(2):406–14.

6. Ethics approval and consent to participate

All components of Generation Scotland received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). All participants provided broad and enduring written informed consent for biomedical research. Generation Scotland has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 15/0040/ES), providing generic ethical approval for a wide range of uses within medical research. This study was performed in accordance with the Helsinki declaration.

Ethical approval for the LBC1936 study was obtained from the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) and the Lothian Research Ethics committee (LREC/1998/4/183; LREC/2003/2/29). All participants provided written informed consent. These studies were performed in accordance with the Helsinki declaration.

Ethical approval for the ALSPAC study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Consent for biological samples has been collected in accordance with the Human Tissue Act (2004). Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time.

7. Availability of data and materials

According to the terms of consent for Generation Scotland participants, access to data must be reviewed by the Generation Scotland Access Committee. Applications should be made to access@generationscotland.org.

Lothian Birth Cohort data are available on request from the Lothian Birth Cohort Study, University of Edinburgh (https://www.ed.ac.uk/lothian-birth-cohorts/data-access-collaboration). Lothian Birth Cohort data are not publicly available due to them containing information that could compromise participant consent and confidentiality.

ALSPAC data are available on request from bona fide researchers. The study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool (http://www.bristol.ac.uk/alspac/researchers/our-data/).

All custom R (version 4.3.1), Python (version 3.9.7), and bash code is available with open access at the following GitHub repository: https://github.com/aleksandra-chybowska/Smoking_EpiScore/

GWAS and EWAS summary statistics will be made available on Edinburgh DataShare on publication.

8. Competing interests

R.E.M has received a speaker fee from Illumina and is an advisor to the Epigenetic Clock Development Foundation. R.F.H. has received consultant fees from Illumina. R.E.M and R.F.H. have received consultant fees from Optima partners. All other authors declare no competing interests.

9. Funding

Generation Scotland: Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006). Genotyping and DNA methylation profiling of the Generation Scotland samples was carried out by the Genetics Core Laboratory at the Edinburgh 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; Reference 104036/Z/14/Z). The DNA methylation data assayed for Generation Scotland was partially funded by a 2018 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (Ref: 27404; awardee: Dr David M Howard) and by a JMAS SIM fellowship from the Royal College of Physicians of Edinburgh (Awardee: Dr Heather C Whalley).

LBC1936: The LBC1936 is supported by the BBSRC, and the Economic and Social Research Council [BB/W008793/1] (which supports S.E.H.), Age UK (Disconnected Mind project), the Milton Damerel Trust, the Medical Research Council (MR/M01311/1), and the University of Edinburgh. Methylation typing of LBC1936 was supported by the Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, The Wellcome Trust Institutional Strategic Support Fund, The University of Edinburgh, and The University of Queensland. Genotyping was funded by the BBSRC (BB/F019394/1). S.R.C. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 221890/Z/20/Z).

ALSPAC: The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and they will serve as guarantors for the contents of this paper. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). Funding for ALSPAC DNAm measurements were supported by the Wellcome (102215/2/13/2); the University of Bristol; the UK Economic and Social Research Council (ES/N000498/1); the UK Medical Research Council (MC_UU_12013/1, MC_UU_12013/2); the Biotechnology and Biological Sciences Research Council (BBI025751/1 and BB/I025263/1); and the John Templeton Foundation (60828). P.Y. and M.S. work is supported by the National Institute for Health and Care Research Bristol Biomedical Research Centre, the Medical Research Council Integrative Epidemiology Unit at the University of Bristol (MC_UU_00032/3, MC_UU_00032/4, MC_UU_00032/6), and Cancer Research UK [C18281/A29019, EDDISA-Jan22\100003].

A.D.C. is supported by a Medical Research Council PhD Studentship in Precision Medicine with funding from the Medical Research Council Doctoral Training Program and the University of Edinburgh College of Medicine and Veterinary Medicine. R.F.H is supported by an MRC IEU Fellowship. E.B. and R.E.M. are supported by Alzheimer's Society major project grant AS-PG-19b-010.

This research was funded in whole, or in part, by the Wellcome Trust (104036/Z/14/Z, 108890/Z/15/Z, 220857/Z/20/Z, and 221890/Z/20/Z). For the purpose of open access, the

author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

10. Authors' contributions

A.D.C. analysed the data. E.B. and R.F.H developed the Bayesian EWAS pipeline. P.Y. and M.S. replicated results in the ALSPAC cohort. D.L.M., R.F.H., L.McG., L.M.., S.E.H., J.C., A.C., T.L.S, S.R.C., and K.L.E. were involved in the data generation. E.B., R.E.M., and C.A.V. drafted the initial manuscript. A.D.C., J.F.P., K.L.E., and R.E.M. designed the study. All authors read and approved the final manuscript.

11. Acknowledgements

Generation Scotland: 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 GS team that includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, healthcare assistants, and nurses.

LBC1936: The authors thank all LBC1936 study participants and research team members who have contributed, and continue to contribute, to ongoing studies.

ALSPAC: We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.