An epigenome-wide association study meta-analysis of educational attainment

"Keeping an open mind is a virtue—but, as the space engineer James Oberg once said, not so open that your brains fall out."

Carl Sagan

"Study hard what interests you the most in the most undisciplined, irreverent and original manner possible."

Richard Feynman

4

Based on Karlsson Linnér et al. (2017). *Molecular Psychiatry.*

Abstract

The epigenome is associated with biological factors, such as disease status, and environmental factors, such as smoking, alcohol consumption, and body mass index. Although there is a widespread perception that environmental influences on the epigenome are pervasive and profound, there has been little evidence to date in humans with respect to environmental factors that are biologically distal. Here, we provide evidence on the associations between epigenetic modifications—in our case, CpG methylation—and educational attainment (EA), a biologically distal environmental factor that is arguably among the most important life-shaping experiences for individuals. Specifically, we report the results of an epigenome-wide association study meta-analysis of EA based on data from 27 cohort studies with a total of 10,767 individuals. We find nine CpG probes significantly associated with EA. However, robustness analyses show that all nine probes have previously been found to be associated with smoking. Only two associations remain when we perform a sensitivity analysis in the subset of never-smokers, and these two probes are known to be strongly associated with maternal smoking during pregnancy, and thus their association with EA could be due to correlation between EA and maternal smoking. Moreover, the effect sizes of the associations with EA are far smaller than the known associations with the biologically proximal environmental factors alcohol consumption, BMI, smoking and maternal smoking during pregnancy. Follow-up analyses that combine the effects of many probes also point to small methylation associations with EA that are highly correlated with the combined effects of smoking. If our findings regarding EA can be generalized to other biologically distal environmental factors, then they cast doubt on the hypothesis that such factors have large effects on the epigenome.

Introduction

The epigenome has been shown to be associated with biological factors such as disease status[1,2]. While there is a widespread perception in the social sciences that a variety of social environmental factors have an effect on the epigenome[3–10], virtually all of the replicated evidence to date in humans relates to environmental factors that have a fairly direct biological impact, such as smoking[11–13], alcohol consumption[14,15], and excess energy intake resulting in increased body mass index (BMI)[16,17]. Here, we study the associations between epigenetic modifications—specifically, the methylation of cytosine-guanine pairs connected by a phosphate link (CpG methylation)—and educational attainment (EA). EA is biologically distal, and yet it is arguably among the most important life-shaping experiences for individuals in modern societies. EA therefore provides a useful test case for whether and to what extent biologically distal environmental factors may affect the epigenome.

In this paper, we report the results of a large-scale epigenome-wide association study (EWAS) meta-analysis of EA. By meta-analyzing harmonized EWAS results across 27 cohort studies, we were able to attain an overall sample size of 10,767 individuals of recent European ancestry, making this study one of the largest EWAS to date[13,15,18]. A large sample size is important because little is known about plausible EWAS effect sizes for complex phenotypes such as EA, and an underpowered analysis would run a high risk of both false negatives and false positives[19,20].

As is standard in EWAS, we used data on CpG DNA methylation. This is the most widely studied epigenetic mark in large cohort studies[1]. Methylation level was measured by the beta value, which is the proportion of methylated molecules at each CpG locus, a continuous variable ranging between 0 and 1[21]. The Illumina 450k Bead Chip measures methylation levels at over 480,000 loci in human DNA and has been used in many cohort studies[1].

We report results from two common methods for the analysis of such methylation datasets. The first main analysis is an EWAS, which considers regression models for each CpG loci with EA. Using the EWAS results we then performed a series of follow-up analyses: enrichment analyses, prediction analyses, correlation with tissue-specific methylation, and gene-expression analysis (**Supplementary Note**). The second main analysis uses the 'epigenetic clock'[22,23] method, which employs a weighted linear combination of a subset of probes (i.e., measured CpG methylation loci) to predict an individual's so-called 'biological age.' The resulting variable can then be linked to phenotypes and health outcomes.

EWAS studies to date have found associations between DNA methylation and, for example, smoking[11,12], body mass index (BMI)[16,24], traumatic stress[25], alcohol consumption[14,26], and cancer[2,27]. In prior work, an age-accelerated epigenetic clock (i.e. an older biological than chronological age) has been linked to increased mortality risk[28], poorer cognitive and physical health[29], greater Alzheimer's disease pathology[30], Down's syndrome[31], high lifetime stress[32], and lower income[33].

Methods

Participating cohorts

We obtained summary-level association statistics from 27 independent cohort studies across 15 cohorts located in Europe, the US, and Australia (**Supplementary Table S1.1**). The total sample size comprised 10,767 individuals of recent European ancestry. All participants provided written informed consent, and all contributing cohorts confirmed compliance with their Local Research Ethics Committees or Institutional Review Boards.

Educational attainment measures

Following earlier work of the Social Science Genetic Association Consortium (SSGAC)[34,35], EA was harmonized across cohorts. The EA variable is defined in accordance with the ISCED 1997 classification (UNESCO), leading to seven categories of EA that are internationally comparable. The categories are translated into US years-of-schooling equivalents, which have a quantitative interpretation (**Supplementary Table S1.2–3**).

Participant inclusion criteria

To be included in the current analysis, participants had to satisfy six criteria: 1) participants were assessed for educational attainment at or after 25 years of age; 2) participants were of European ancestry; 3) all relevant covariate data were available for each participant; 4) participants passed the cohort-level methylation quality control; 5) participants passed cohort-specific standard quality controls (for example, genetic outliers were excluded); and 6) participants were not disease cases from a case/control study.

DNA methylation measurement and cohort-level quality control

Whole-blood DNA CpG methylation was measured genome-wide in all cohorts using the Illumina 450k Human Methylation chip. We standardized the cohort-level quality control and pre-processing of the methylation data as much as possible, while ensuring some degree of flexibility to keep the implementation feasible for all cohorts (leading to slight variation in preprocessing across cohorts, as is common[13,15,17]). Cohort-specific information regarding technical treatment of the data, such as background-correction[36], normalisation[37], and quality control, is reported in **Supplementary Table S1.4**.

Epigenome-wide association study (EWAS)

Our analyses were performed in accordance with a pre-registered analysis plan archived at Open Science Framework (OSF) in September 2015 (available at: https://osf.io/9v3nk/). We first performed a meta-analysis of the EWAS of EA to investigate associations with individual methylation markers (**Supplementary Note 2**). As is standard, the EWAS was performed as a set of linear regressions in each cohort, one methylation marker at a time, with the methylation beta value (0–1) as the dependent variable. The key independent variable was EA. We estimated two regression models that differ in the set of covariates included. In the *basic model*, the covariates were age, sex, imputed or measured white blood-cell counts, technical covariates from the methylation array, and four genetic principal components to account for population stratification. In the *adjusted model*, we additionally controlled for body mass index (BMI, kg/m^2), smoker status (three categories: current, previous, or never smoker), an interaction term between age and sex, and a quadratic term for age. Since BMI and smoking are correlated with EA[38,39] and known to be associated with methylation[13,17], the basic model may identify associations with EA that are actually due to BMI or smoking. While the adjusted model reduces that risk, it may also reduce power to identify true associations with EA (by controlling for factors that are correlated with EA). While we present the results for both models, we focus on the adjusted model because it is more conservative. Details of cohort-specific covariates are presented in **Supplementary Table S1.4**.

EWAS quality control and meta-analysis

Each participating cohort uploaded EWAS summary statistics to a central secure server for quality control (QC) and meta-analysis. The number of CpG probes filtered at each step of the QC is presented in **Supplementary Table S1.5.** We removed: (a) probes with missing *P* value, standard error, or coefficient estimate; (b) probes with a call rate less than 95%; (c) probes with a combined sample size less than 1,000; (d) probes not available in the probe-annotation reference by Price et al. (2013)[40]; (e) CpH probes (H = A/C/T); (f) probes on the sex chromosomes; and (g) cross-reactive probes highlighted in a recent paper by Chen et al.[41]. We performed a sample-size-weighted meta-analysis of the cleaned results using METAL[42]. We used single genomic control, as is common in genome-wide association studies (GWAS)[43], to stringently correct the meta-analysis *P* values for possible unaccounted-for population stratification[44]. Probes with a *P* value less than 1×10^{-7}, a commonly-used threshold in EWAS[1] that we pre-specified in the analysis plan, were considered epigenome-wide significant associations.

Epigenetic clock analyses

To construct our epigenetic clock variables (**Supplementary Note 3**), the cohort-level raw beta-value data was entered into the online Horvath calculator[23], as per our pre-registered analysis plan. The "normalize data" and "advanced analysis for Blood Data" options were selected. The following variables were selected from the calculator's output for subsequent analysis:

- Clock 1. Horvath age acceleration residuals, which are the residuals from the regression of chronological age on Horvath age.
- Clock 2. White blood cell count adjusted Horvath age acceleration, which is the residual from Clock 1 after additional covariate adjustment for imputed white blood cell counts.
- Clock 3. White blood cell count adjusted Hannum age acceleration, which is the same as Clock 2 but with the Hannum age prediction in place of the Horvath prediction.
- Clock 4. Cell-count enriched Hannum age acceleration, which is the basic Hannum predictor plus a weighted average of aging-associated cell counts. This index has been found to have the strongest association with mortality[45].

These Clock measures are annotated in the Horvath software as follows: 'AgeAccelerationResidual', 'AAHOAdjCellCounts', 'AAHAAdjCellCounts', and 'BioAge4HAAdjAge'. We analyzed two regression models for each clock variable, both with EA as an independent variable and a clock variable as the dependent variable. In the *basic age acceleration model*, we control for chronological age, and in the *adjusted age acceleration model*, we additionally control for BMI and smoker status (current, previous, or never smoker). In total, in each adult cohort, we estimated eight regressions: each of the two models with each of the four clock variables as a dependent variable. For each of the eight regressions, we performed a sample-size-weighted meta-analysis of the cohort-level results.

Polygenic predictions with polygenic methylation score

We performed a prediction analysis with polygenic methylation scores (PGMS), analogous to polygenic-score prediction in the GWAS literature (**Supplementary Note 6**). We tested the predictive power in three independent adult cohort studies: Lothian Birth Cohort 1936 (LBC1936, *n* = 918), RS-BIOS (Rotterdam Study – BIOS, *n* = 671), and RS3 (Rotterdam Study 3, *n* = 728). We re-ran the EWAS meta-analysis for each prediction cohort to obtain the weights for the PGMS, while holding out the prediction cohort to avoid overfitting. We constructed the PGMS for each individual as a weighed sum of the individual's methylation markers' beta values and the EWAS effect-sizes, using the Z statistics from the EWAS as weights. (The Z statistics were used instead of the EWAS coefficients because CpG methylation is the dependent variable in the EWAS regression.) We constructed PGMS using two different thresholds for probe inclusion, $P < 1\times10^{-5}$ and $P < 1\times10^{-7}$, with weights from the basic and adjusted EWAS models, for a total of four PGMS in each prediction cohort.

To shed light on the direction of causation of epigenetic associations, we used a fourth prediction cohort study, a sample of children in the ALSPAC ARIES cohort[46]. We constructed the PGMS using the same approach as described above, in this case using data from cord-blood-based DNA methylation at birth. The outcome variables in this cohort were average educational achievement test scores (Key Stages 1–4[47]) from age 7 up to age 16 years.

To examine the relationship between epigenetic and genetic associations, we also constructed a single-nucleotide polymorphism polygenic score (SNP PGS) for EA. We used SNP genotype data available in the three adult prediction cohort studies (LBC1936, RS-BIOS, and RS3). We constructed the SNP PGS in each cohort as a weighted sum of the individual genotypes from all available genotyped SNPs, with GWAS meta-analysis coefficients as weights. We obtained the coefficients by re-running the largest GWAS meta-analysis to date of EA[34] after excluding our prediction cohorts (LBC1936, RS-BIOS, and RS3).

We evaluated the predictive power of the PGMS by examining the incremental coefficient of determination (incremental R^2) for predicting EA (or test scores in ALSPAC ARIES). The incremental R^2 is the difference in R^2 between the regression model with only covariates, and the same regression model that additionally includes the PGMS as a predictor. The covariate-only models in the LBC1936, RS-BIOS, and RS3 cohorts controlled for age, sex, and the SNP PGS. In the ALSPAC ARIES cohort we controlled for age at assessment and sex. In the ALSPAC ARIES cohort, when we investigate maternal smoking as a potential confound for our EA associations, we add maternal smoking to the set of covariates. We finally restricted the ALSPAC ARIES cohort to children with non-smoking mothers. To investigate a possible interaction effect between the PGMS and SNP PGS, we re-estimated the regression model after adding an interaction term between the PGMS and the SNP PGS, and the incremental R^2 was calculated as the difference in R^2 relative to the model that included the PGMS and the SNP PGS as additive main effects.

Results

Descriptive statistics

Summary statistics from the 27 independent cohort studies from the 15 contributing cohorts are shown in **Supplementary Table S1.1**. The mean age at reporting ranges from 26.6 to 79.1 years, and the sample size ranges from 48 to 1,658, with a mean of 399 individuals per cohort. The mean cohort EA ranges from 8.6 to 18.3 years of education, and the sample-size-weighted mean is 13.6 (SD = 3.62). The meta-analysis sample is 54.1% female.

EWAS

The QC filtering is reported in **Supplementary Table S1.5**. We inspected the quantile-quantile (QQ) plot of the filtered EWAS results from each contributing cohort as part of the QC procedure before meta-analysis. The genomic inflation factor (λ_{GC}), defined as the ratio of the median of the empirically observed chi-square test statistics to the expected median under the null, had a mean across the cohorts of 1.02 for the adjusted model (SD = 0.18). We report the cohort-level genomic inflation factor after probe filtering in **Supplementary Table S1.5**. The variation in λ_{GC} across cohorts was comparable to that from EWAS performed in cohorts of similar sample size[12]. We applied genomic control at the cohort level, which is a conservative method of controlling for residual population stratification that may remain even despite the regression controls for principal components[44]. The meta-analysis λ_{GC} was 1.19 for the basic model and 1.06 for the adjusted model.

Figure 4.1 shows the Manhattan plot for the meta-analysis results of the adjusted model. The Manhattan plot for the basic model is reported in the **Supplementary Note**, together with the QQ plots for the basic and adjusted model. In the basic model there were 37 CpG probes associated with EA at our preregistered epigenome-wide P value threshold ($P < 10^{-7}$); these results are reported in **Supplementary Table S1.6a**. In the adjusted model there were 9 associated probes, listed in **Table 4.1** (with additional details in **Supplementary Table S1.7a**), all of which were also associated in the basic model. We hereafter refer to the adjusted model's 9 associated probes as the "lead probes." In **Supplementary Note 2.4.1** we present the association results with false discovery rate (FDR) less than 0.05, but since this threshold was not pre-specified in the analysis plan we do not present these results as main findings.

To investigate how the EWAS results look at a regional level, we analyzed the distribution of the EWAS associations across the genome by performing enrichment tests for methylation density regions[40] (the so-called "HIL" categories; **Supplementary Note 5.2**). We found that the number of probes with $P < 10^{-7}$ is more or less proportional to the total number of probes in every region and that there is enrichment for association in all four methylation density categories: high-density CpG islands (HC), intermediate-density CpG islands (IC), intermediate-density CpG islands bordering HCs (ICshore), and non-islands (LC).

The effect sizes of the associations for the nine lead probes are shown in **Table 4.1**. The coefficients of determination (R^2's) range from 0.3% to 0.7%. To put these effect sizes in perspective, **Figure 4.2** and **Supplementary Table S1.8** compare the R^2's for the top 50 probes in our adjusted model with the top 50 probes from recent large-scale EWAS of smoking[13], maternal smoking[12], alcohol consumption[15], and BMI[17], as well as the top 50 GWAS SNP associations with EA[34]. The EA EWAS associations of our study are an order of magnitude larger than the largest EA SNP effect sizes. However, our EWAS associations are small in magnitude relative to the EWAS associations reported for more biologically proximal environmental factors. BMI is the most similar to EA, with R^2's of associated probes approximately 20–50% larger than those for EA. Relative to the largest R^2 for an EA-associated probe, the largest effect for probes associated with smoking and maternal smoking are greater by factors of roughly three and 17, respectively.

Lookup of lead probes in published EWAS of smoking

Since our smoker-status control variable is coarse and discrete (current, former, or never smoker), we were concerned that the adjusted EWAS model might not have adequately controlled for exposure to smoking (i.e., amount and duration of smoking and exposure to second-hand smoke). Therefore, we performed a lookup of our lead probes in the published EWAS on smoking (**Supplementary Note 4** and **Supplementary Table S1.10**). We found that all nine lead probes have previously been associated with smoking. The results of this lookup motivated our analysis of the never-smoker subsample, discussed next.

Robustness of EWAS results in the never-smoker subsample

To minimize the possible confounding effect of smoking on the association between EA and CpG methylation, we conducted a set of analyzes that we did not anticipate when we preregistered our analysis plan. Specifically, we went back to the cohorts and asked them to re-conduct their EWAS, this time restricting the analysis to individuals that self-reported as never smokers. After following the same QC steps as above, we performed a new meta-analysis of these results (n = 5,175).

In this subsample, the effect-size estimates were smaller by at least 60% for seven of the nine lead probes (see **Table 4.1** and panel A in **Figure 4.3**), whereas two probes (cg12803068 and cg22132788) had similar effect-size estimates as in the full sample (statistically distinguishable from zero with $P = 1.48\times10^{-4}$ and $P = 4.35\times10^{-4}$, respectively). These two probes, however—both in proximity to the gene *MYO1G*—have been found to be associated with maternal smoking during pregnancy, and the effects on the methylation of this gene are persistent when measured at age 17 in the offspring[12,48]. This influence has been shown to continue through to middle age[49]. We cannot distinguish between the hypothesis that these probes have some true association with EA and the hypothesis that their apparent association with EA is entirely driven by more maternal smoking during pregnancy among lower-EA individuals. We also cannot rule out that the probes' association with EA is driven by second-hand smoke exposure, which could also be correlated with EA.

To assess the how widely such confounding may affect the EA results, in Panel B of **Figure 4.3** we compare the effect sizes of all the probes associated with EA at $P < 10^{-5}$ in the adjusted EWAS model to the effect sizes found for the same probes in EWAS meta-analyses of smoking[13] and maternal smoking[12] (see also **Supplementary Table S.1.11**). Many of the EA-associated probes are also associated with smoking or maternal smoking, strongly suggesting that residual smoking exposure (i.e., the misclassification of amount and duration of smoking, and second-hand smoke that is not captured by the smoking covariate) and maternal smoking remain potential confounding factors for the probe associations with EA, even in the subsample of individuals who are self-reported never-smokers.

Epigenetic clock associations with EA

Two cohorts, FINRISK and MCCS, did not contribute to the epigenetic clock analyses. Therefore, the sample sizes for these analyses were smaller than for the EWAS meta-analysis: 8,173 for the basic age acceleration model and 7,691 for the adjusted age acceleration model (the difference being due to the lack of covariates for some individuals). The effect-size estimates are presented in **Figure 4.4** and **Supplementary Table S1.9**. There was no evidence for an association between EA and Clocks 1, 2, or 3, but the association between EA and Clock 4 was strong ($P = 3.51\times10^{-6}$ and $P = 4.51\times10^{-4}$ in the basic and adjusted age acceleration model, respectively). The point estimates were small, however: using Clock 4, each year of EA was associated with a 0.071-year (i.e., 26-day) reduction in age acceleration in the basic model and a 0.055-year (i.e., 20-day) reduction in the adjusted model. Overall then, higher educational attainment was associated with slightly younger biological age when compared with chronological age. We note that the epigenetic clock that was found to be associated with EA, Clock 4, has previously been found to be the most predictive epigenetic-clock measure of mortality[45].

Prediction using polygenic methylation scores

The incremental R^2's from the prediction of EA with polygenic methylation scores (PGMS) in our adult prediction cohort studies, the LBC1936, RS-BIOS, and RS3, are reported in **Supplementary Table S1.13a** and **Figure 4.5.** Across the four PGMS constructed with weights from the basic and adjusted model, and with the two probe-inclusion thresholds ($P < 10^{-5}$ and $P < 10^{-7}$), the incremental R^2's ranged from 1.4% to 2.0% ($P = 3.28\times10^{-8}$ and lower). There was also weak evidence for an interaction between the PGMS and the SNP PGS in predicting EA, with the R^2's for the interaction term ranging from 0.1% to 0.3% (P values ranged from 0.01 to 0.12).

In the subsample of never-smokers the PGMS (constructed with weights derived from the full EWAS sample), the PGMS is far less predictive, with incremental R^2's ranging from 0.3% to 0.9% (**Figure 4.5** and **Supplementary Table S1.13b**). The two PGMS constructed from probes with $P < 1\times10^{-5}$ in the EWAS were associated with EA at $P < 0.05$, while the two PGMS constructed only from the lead probes with $P < 1\times10^{-7}$ were not ($P > 0.05$). No interaction effect was found between the PGMS and the SNP PGS in the never-smoker subsample.

To further investigate confounding by smoking in the prediction analysis, we examined the correlations between our PGMS constructed from the lead probes (i.e., those associated with EA at significance threshold $P < 1\times10^{-7}$) in either our basic or adjusted model and a PGMS for smoking (see **Supplementary Note 6.2.2** for details). For the smoking PGMS, we use the 187 probes that were identified at epigenome-wide significance ($P < 1\times10^{-7}$) and then successfully replicated in a recent EWAS of smoking[50]. We examine the PGMS correlations in our full prediction samples, not restricted to never-smokers. For the EA PGMS from our basic model, we find a correlation with the smoking PGMS of –0.96 in RS3, –0.94 in RS-BIOS, and –0.93 in LBC1936. For the EA PGMS from our adjusted model, the correlations are –0.90, –0.89, and –0.91, respectively. In all cases, the nearly perfect correlation between the smoking and EA methylation scores strongly suggests that smoking status confounds the EWAS associations with EA.

Turning to the child sample in the ALSPAC ARIES cohort[46,48], we examined whether a PGMS constructed from methylation assessed in cord blood samples at birth was predictive of four prospective measures of educational achievement test scores (Key Stage 1–4[47]), collected between ages 7 and 16 (**Supplementary Note 6.1.1**). The results are reported in **Supplementary Table S1.13c**. The largest incremental R^2 was 0.73% (P = 0.0094), and it was attained in the model predicting school performance at age 14–16 (i.e., the Key Stage 4 test scores). However, once maternal smoking status was added as a control variable, the predictive power of the PGMS became essentially zero (incremental $R^2 = 0.05\%$, P = 0.234). This suggests that the confounding effects of maternal smoking strongly influenced the predictive power of the PGMS for EA. We draw two conclusions from these results from the child sample. First, they reinforce the concern that maternal smoking was a major confound for any probe associations with EA. Second, they suggest that any true methylation-EA associations were unlikely to be driven by a causal effect of methylation on EA.

Overlap between EWAS probes and published GWAS associations

To supplement our polygenic-score analyses of the overlap between epigenetic and genetic associations, we next investigated whether our lead probes are located at loci that contain SNPs previously identified in GWAS of EA and smoking **(Supplementary Note 5**). Considering jointly the 141 approximately independent EWAS probes with $P < 10^{-4}$, we did not find evidence of enrichment for either EA-linked SNPs ($P = 0.206$) or smoking-linked SNPs ($P =$

0.504). Considering the probes individually, one probe (cg17939805) was found to be in the same genomic region as a SNP (rs9956387) associated with EA (with a genomic distance of 607 bp), whereas no probes were close to SNPs previously identified as linked to smoking.

Correlation of EWAS results with tissue-specific methylation

To answer the question of whether our EWAS associations are correlated with any tissue-specific DNA methylation, we utilized the tissue-specific methylation data made available by the Epigenomic Roadmap Consortium[51]. That data was used to calculate tissue-specific deviations from average cross-tissue methylation at the loci corresponding to the EWAS CpG probes associated with EA at a P value less than $1{\times}10^{-4}$ (**Supplementary Note 7**). We examined the correlation between these tissue-specific deviations and the EWAS association test statistics (Z statistics) of the probes, using the results from the adjusted EWAS model. We report the results in **Figure 4.6** and in **Supplementary Table S1.14**. The strongest correlations were found for primary haematopoietic stem cells G-CSF-mobilized female and IMR90 fetal lung fibroblast. Intermediate-strength correlations were found across multiple, seemingly unrelated tissues, while no correlations of relevant magnitudes were found with the brain tissues available in the Roadmap. We interpret the lack of correlation with tissues plausibly related to EA (such as brain tissues) as supporting the conclusion that the EWAS results are driven by confounding factors rather than by a true association with EA.

Pathway analysis with gene-expression data

Using the GTEx[52] expression data and the webtool 'Functional mapping and annotation of genetic associations' (FUMA)[53] we performed a pathway analysis. The analysis used the GTEx gene-expression levels to cluster the 29 genes physically closest to the EA-associated (at $P < 1{\times}10^{-5}$) CpG probes of the adjusted model (**Supplementary Note 8**). The results of the expression analysis are displayed in **Supplementary Figure 4**. We find that the genes closest to the EA-associated probes are expressed across multiple tissues that have no clear relationship to EA (such as blood tissues, among many other); for further discussion, see **Supplementary Note**. Overall, these results are consistent with the hypothesis that the EWAS results are driven by confounding factors.

Discussion

This study provides one of the first large-scale investigations in humans of epigenetic changes linked to a biologically distal environmental factor. In our EWAS meta-analysis—one of the largest EWAS conducted to date—we found nine CpG probes associated with EA. Each of these probes explains 0.3% to 0.7% of the variance in EA—effect sizes somewhat smaller than the largest EWAS effects that have been observed for BMI and many times smaller than those observed for alcohol consumption, smoking, and especially maternal smoking during pregnancy. When we restrict our analysis to the subsample of never-smokers, the effect sizes of seven out of the nine lead probes are substantially attenuated. Moreover, the other two lead probes have been found in previous work to be strongly associated with maternal smoking during pregnancy[12]. More generally, comparing our own results to those from previous EWAS highlights a variety of factors correlated with EA, including not only maternal smoking but also alcohol consumption and BMI, as potentially major confounding factors for the EA associations we detect. We also cannot rule out that other factors correlated with EA, such as exposure to second-hand smoke, could confound the EA associations. This should be taken into account in future endeavors of associating methylation with biologically distal factors that are known to correlate with environmental factors that have a fairly direct biological impact, such as smoking.

Convincingly establishing a causal effect of EA would require analyzing a sample with quasi-random variation in EA, such as a sample in which some individuals were educated after an increase in the number of years of compulsory schooling and other individuals were educated before the law change[54]. We are not aware of large EWAS samples with quasi-random variation at present, but we anticipate that such samples will become available as methylation becomes more widely measured.

Although the EWAS we report here is among the largest conducted to date, our sample size of 10,767 individuals is only large enough to identify nine probes associated with EA at the conventional epigenome-wide significance threshold. Subsequent EWAS conducted in larger samples that have sufficient statistical power to identify a much larger number of EA-associated probes will enable more extensive investigations of overlap with probes associated with other phenotypes than were possible from our results, as well as analyses of the biological functions of the probes. Besides limited statistical power, other limitations of our study, common to EWAS research designs, are that we study methylation cross-sectionally and not longitudinally and that we only investigate CpG methylation and no other types of epigenetic modifications. Also, our study focuses on single CpG sites; future studies could consider additional analytical approaches to assess regions of differential methylation (e.g., genes). Once suitable methods have been developed, it would also be of interest to estimate the overall proportion of variance in EA that can be attributed to individual differences in DNA methylation patterns.

Conclusion

One plausible hypothesis is that environmental influences on the epigenome—even those due to everyday, social environmental factors—are pervasive and profound[3]. According to the logic of this view, a major life experience that occurs over many years, such as EA, should leave a powerful imprint on the epigenome. Motivated by this view and by the evidence of large EWAS effects in studies of lifestyle factors, when we embarked on this project we entertained the hypothesis that we might find large associations between EA and methylation. We also entertained the alternative hypothesis that EA, because it is so biologically distal, may exhibit much weaker associations with methylation.

While our results do not allow us to distinguish how much of the effects we find are due to true associations with EA and how much are due to confounding factors, they strongly suggest that the effect sizes we estimate are an upper bound on the effect sizes of any true methylation associations with EA. These upper-bound effect sizes are far smaller than associations with more biologically proximal environmental factors that have been studied. If our results can be generalized beyond EA to other biologically distal environmental factors, then they cast doubt on the hypothesis that such factors have large effects on the epigenome.

Authors – Chapter 4

Richard Karlsson Linnér, Riccardo E Marioni, Cornelius A Rietveld, Andrew J Simpkin, Neil M Davies, Kyoko Watanabe, Nicola J Armstrong, Kirsi Auro, Clemens Baumbach, Marc Jan Bonder, Jadwiga Buchwald, Giovanni Fiorito, Khadeeja Ismail, Stella Iurato, Anni Joensuu, Pauliina Karell, Silva Kasela, Jari Lahti, Allan F McRae, Pooja R Mandaviya, Ilkka Seppälä, Yunzhang Wang, Laura Baglietto, Elisabeth B Binder, Sarah E Harris, Allison M Hodge, Steve Horvath, Mikko Hurme, Magnus Johannesson, Antti Latvala, Karen A Mather, Sarah E Medland, Andres Metspalu, Lili Milani, Roger L Milne, Alison Pattie, Nancy L Pedersen, Annette Peters, Silvia Polidoro, Katri Räikkönen, Gianluca Severi, John M Starr, Lisette Stolk, Melanie Waldenberger, Johan G Eriksson, Tõnu Esko, Lude Franke, Christian Gieger, Graham G Giles, Sara Hägg, Pekka Jousilahti, Jaakko Kaprio, Mika Kähönen, Terho Lehtimäki, Nicholas G Martin, Joyce B C van Meurs, Miina Ollikainen, Markus Perola, Danielle Posthuma, Olli T Raitakari, Perminder S Sachdev, Erdogan Taskesen, André G Uitterlinden, Paolo Vineis, Cisca Wijmenga, Margaret J Wright, Caroline Relton, George Davey Smith, Ian J Deary, Philipp D Koellinger, Daniel J Benjamin

References – Chapter 4

1. Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* **12,** 529–541 (2012).
2. van Veldhoven, K. *et al.* Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis. *Clin. Epigenetics* **7,** 1–12 (2015).
3. Champagne, F. A. & Curley, J. P. Epigenetic Influence of the Social Environment. in *Brain, Behavior, and Epigenetics* (eds. Petronis, A. & Mill, J.) 23–40 (Springer-Verlag Berlin Heidelberg, 2011). doi:10.1007/978-3-642-17426-1
4. Beach, S. R. H. *et al.* Parenting, Socioeconomic Status Risk, and Later Young Adult Health: Exploration of Opposing Indirect Effects via DNA Methylation. *Child Dev.* **87,** 111–121 (2016).
5. Borghol, N. *et al.* Associations with early-life socio-economic position in adult DNA methylation. *Int. J. Epidemiol.* **41,** 62–74 (2012).
6. Cunliffe, V. T. The epigenetic impacts of social stress: how does social adversity become biologically embedded? *Epigenomics* **8,** 1653–1669 (2016).
7. Jones-Mason, K., Allen, I. E., Bush, N. & Hamilton, S. Epigenetic marks as the link between environment and development: examination of the associations between attachment, socioeconomic status, and methylation of the SLC6A4 gene. *Brain Behav.* **6,** 1–18 (2016).
8. Stringhini, S. *et al.* Life-course socioeconomic status and DNA methylation of genes regulating inflammation. *Int. J. Epidemiol.* **44,** 1320–1330 (2015).
9. Szyf, M. DNA methylation, the early-life social environment and behavioral disorders. *J. Neurodev. Disord.* **3,** 238–249 (2011).
10. Szyf, M., McGowan, P. & Meany, M. J. The Social Environment and the Epigenome. *Environ. Mol. Mutagen.* **49,** 46–60 (2008).
11. Gao, X., Jia, M., Zhang, Y., Breitling, L. P. & Brenner, H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clin. Epigenetics* **7,** 113 (2015).
12. Joubert, B. R. *et al.* DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am. J. Hum. Genet.* 1–17 (2016). doi:10.1016/j.ajhg.2016.02.019
13. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circ. Cardiovasc. Genet.* CIRCGENETICS.116.001506 (2016). doi:10.1161/CIRCGENETICS.116.001506
14. Ungerer, M., Knezovich, J. & Ramsay, M. In utero alcohol exposure, epigenetic changes, and their consequences. *Alcohol Res* **35,** 37–46 (2013).
15. Liu, C. *et al.* A DNA methylation biomarker of alcohol consumption. *Mol. Psychiatry* 1–12 (2016). doi:10.1038/mp.2016.192
16. Demerath, E. W. *et al.* Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple

replicated loci. *Hum. Mol. Genet.* **24,** 4464–79 (2015).

17. Mendelson, M. M. *et al.* Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease : A Mendelian Randomization Approach. 1–30 (2017). doi:10.1371/journal.pmed.1002215

18. Ligthart, S. *et al.* DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biol.* 1–15 (2016). doi:10.1186/s13059-016-1119-5

19. Ioannidis, J. P. a. Why most published research findings are false. *PLoS Med.* **2,** e124 (2005).

20. Benjamin, D. J. *et al.* The promises and pitfalls of genoeconomics. *Annu. Rev. Econom.* **4,** 627–662 (2012).

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

22. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol. Cell* **49,** 359–367 (2013).

23. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14,** R115 (2013).

24. Rönn, T. *et al.* Impact of age, BMI and HbA1c levels on the genome-wide DNA methylation and mRNA expression patterns in human adipose tissue and identification of epigenetic biomarkers in blood. *Hum. Mol. Genet.* **24,** 3792–3813 (2015).

25. Houtepen, L. C. *et al.* Genome-wide DNA methylation levels and altered cortisol stress reactivity following childhood trauma in humans. *Nat. Commun.* **7,** 10967 (2016).

26. Masemola, M. L., van der Merwe, L., Lombard, Z., Viljoen, D. & Ramsay, M. Reduced DNA methylation at the PEG3 DMR and KvDMR1 loci in children exposed to alcohol in utero: A South African Fetal Alcohol Syndrome cohort study. *Front. Genet.* **5,** 1–12 (2015).

27. Barrow, T. M. & Michels, K. B. Epigenetic epidemiology of cancer. *Biochem. Biophys. Res. Commun.* **455,** 70–83 (2014).

28. Marioni, R. E. *et al.* The epigenetic clock and telomere length are independently associated with chronological age and mortality. *Int. J. Epidemiol.* 1–9 (2016). doi:10.1093/ije/dyw041

29. Marioni, R. E. *et al.* The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *Int. J. Epidemiol.* **44,** 1388–1396 (2015).

30. Levine, M. E., Lu, A. T., Bennett, D. A. & Horvath, S. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. *Aging (Albany. NY).* **7,** 1198–1211 (2015).

31. Horvath, S. *et al.* Accelerated epigenetic aging in Down syndrome. *Aging Cell* **14,** 491–495 (2015).

32. Zannas, A. S. *et al.* Lifetime stress accelerates epigenetic aging in an urban, African American cohort: relevance of glucocorticoid signaling. *Genome Biol.* **16,** 266 (2015).

33. Simons, R. L. *et al.* Economic hardship and biological weathering: The epigenetics of aging in a U.S. sample of black women. *Soc. Sci. Med.* **150,** 192–200 (2016).

34. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533,** 539–542 (2016).

35. Rietveld, C. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340,** 1467–71 (2013).

36. Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* **41,** 1–11 (2013).

37. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15,** 503 (2014).

38. Johnson, W. *et al.* Does education confer a culture of healthy behavior? Smoking and drinking patterns in Danish twins. *Am. J. Epidemiol.* **173,** 55–63 (2011).

39. Hermann, S. *et al.* The Association of Education with Body Mass Index and Waist Circumference in the EPIC-PANACEA Study. *BMC Public Health* **11,** 169 (2011).

40. Price, M. E. *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* **6,** 4 (2013).

41. Chen, Y. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8,** 203–9 (2013).

42. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).

43. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).

44. van Iterson, M., van Zwet, E., the BIOS Consortium & Heijmans, B. T. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.* **18,** 1–13 (2017).

45. Chen, B. H. *et al.* DNA methylation-based measures of biological age: Meta-analysis predicting time to death. *Aging (Albany. NY).* **8,** 1844–1865 (2016).

46. Relton, C. L. *et al.* Data resource profile: Accessible resource for integrated epigenomic studies (ARIES). *Int. J. Epidemiol.* **44,** 1181–1190 (2015).

47. GOV.UK. National Curriculum. *GOV.UK* (2016).

48. Richmond, R. C. *et al.* Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: Findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Hum. Mol. Genet.* **24,** 2201–2217 (2015).

49. Richmond, R. C., Suderman, M., Langdon, R., Relton, C. L. & Davey Smith, G. DNA methylation as a marker for prenatal smoke exposure in adults. *bioRxiv* (2017).

50. Zeilinger, S. *et al.* Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. *PLoS One* **8,** (2013).

51. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

52. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–5 (2013).

53. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. FUMA: Functional

mapping and annotation of genetic associations. *Manuscr. Submitt. Publl.* (2017).

54. Lager, A. C. J. & Torssander, J. Causal effect of education on mortality in a quasi-experiment on 1.2 million Swedes. *Proc. Natl. Acad. Sci.* **109,** 8461–8466 (2012).

Figures and tables – Chapter 4

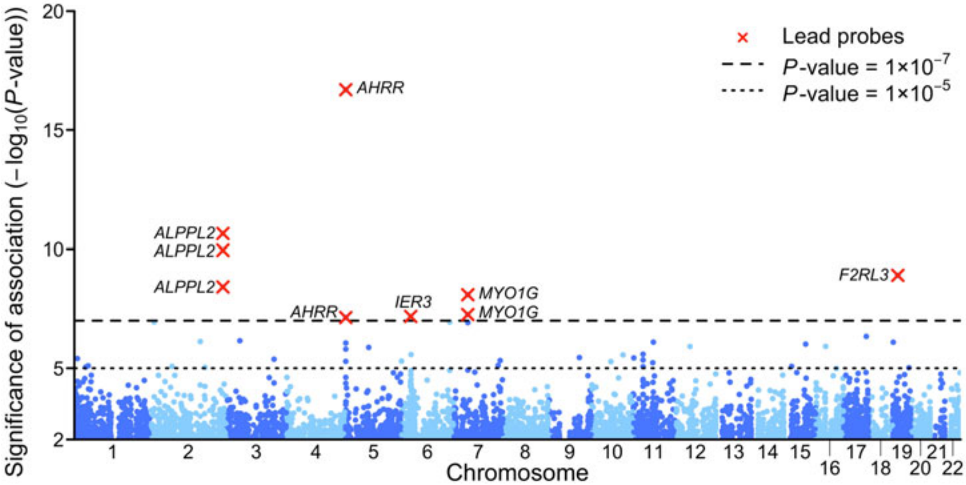

Figure 4.1. Manhattan plot of the adjusted EWAS model. The figure displays the Manhattan plot of the meta-analysis of the adjusted EWAS model (the Manhattan plot of the basic model is reported in **Supplementary Note**). The *x*-axis is chromosomal position, and the *y*-axis is the significance on a $-\log_{10}$ scale. The dashed lines mark the threshold for epigenome-wide significance ($P = 1\times10^{-7}$) and for suggestive significance ($P = 1\times10^{-5}$). Each epigenome-wide associated probe is marked with a red ×, and the symbol of the closest gene based on physical position.

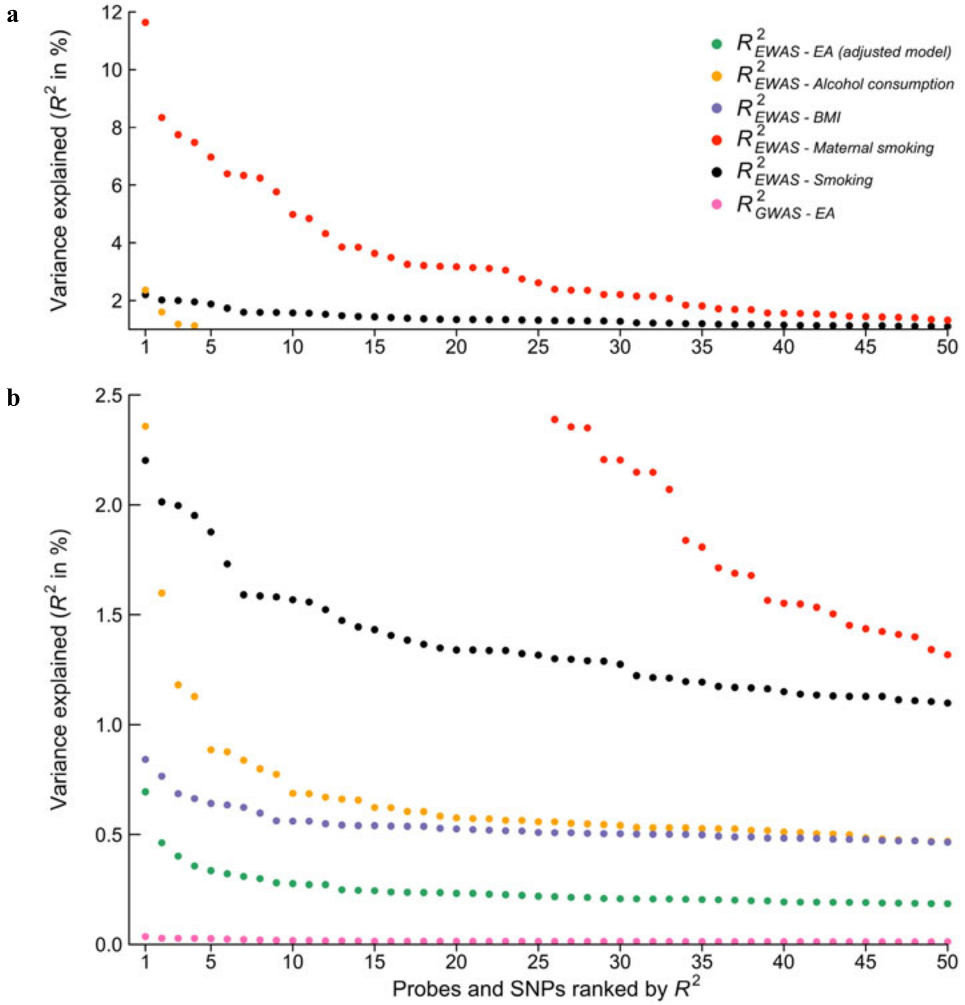

Figure 4.2. EWAS effect sizes (in terms of variance explained) across traits and with GWAS. The figure displays the effect size estimates in terms of R^2, in descending order, for the 50 top probes of the adjusted EWAS model. For comparison we present the 50 top probes from recent EWAS on alcohol consumption (n = 9,643, Liu et al., 2016), BMI (n = 7,798, Mendelson et al., 2017), smoking (n = 9,389, Joehanes et al., 2016), and maternal smoking (n = 6,685, Joubert et al., 2016). For comparison with GWAS effect sizes we contrast the EWAS probes with the effect sizes of the 50 top approximately independent SNPs from a recent GWAS on educational attainment (n = 405,073, Okbay et al., 2016). Panel **(a)** and **(b)** display the same results but with a different scaling of the *y*-axis in order for the smaller effect sizes to be visible.

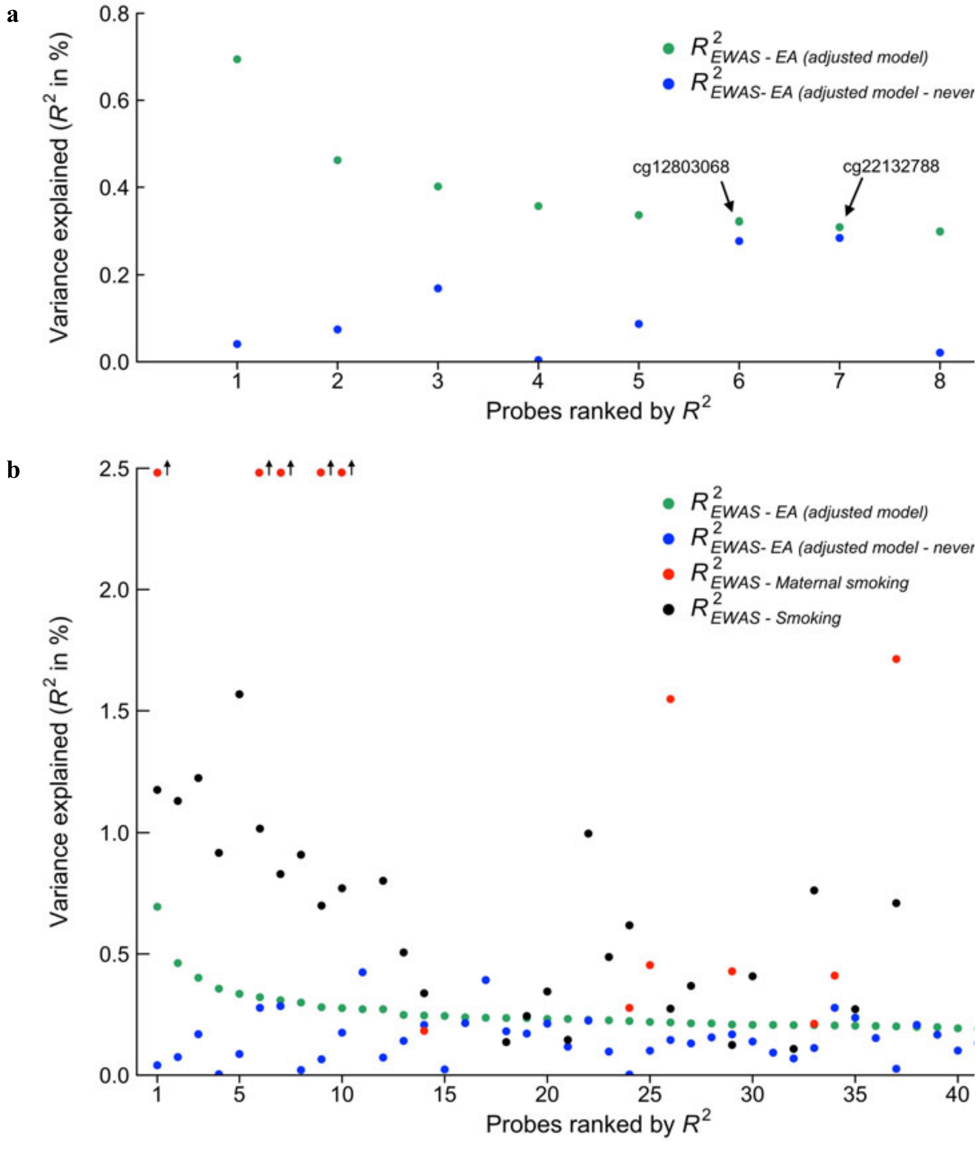

Figure 4.3. Comparison of EA EWAS effect sizes with the effect sizes in the never-smoker subsample and in smoking EWAS results. Panel **(a)** displays the effect-size estimates in terms of R^2 for the 9 lead probes, in descending order, and the lead probe's corresponding effect size when re-estimated in the subsample of never smokers. Panel **(b)** displays the same information for the probes of the adjusted model with $P < 1{\times}10^{-5}$ (including the 9 lead probes), as well as the same probes' effect-size estimates from two recent EWAS of smoking ($n = 9{,}389$, Joehanes et al., 2016), and maternal smoking ($n = 6{,}685$, Joubert et al., 2016). The smoking and maternal smoking estimates are only publicly available for probes associated at FDR < 0.05 in the respective EWAS.

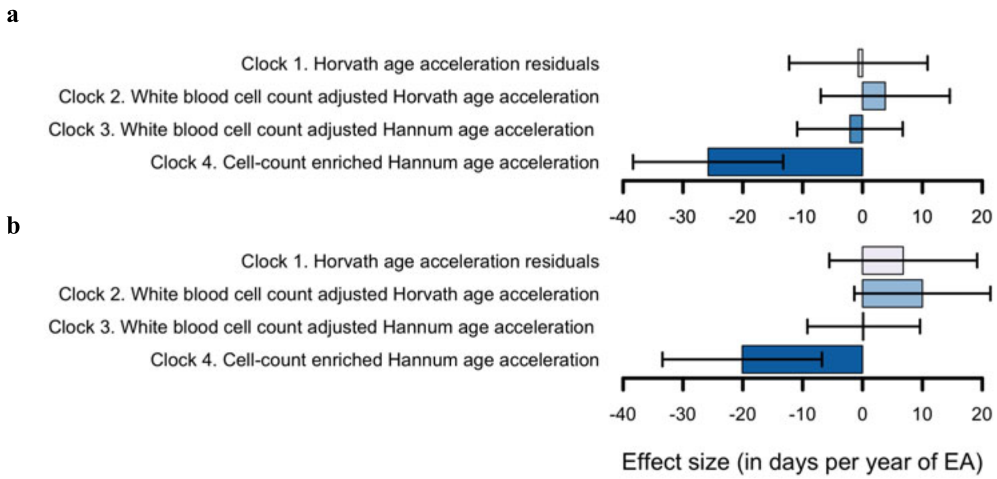

Figure 4.4. Effect size estimates (in days) of the epigenetic clock analyses with 95% confidence intervals. Panel **(a)** displays the effect size estimates from the basic age acceleration model, and panel **(b)** displays the effect size estimates from the adjusted age acceleration model. The effect size is denoted in days of age acceleration per year of EA, and error bars represent 95% confidence intervals.

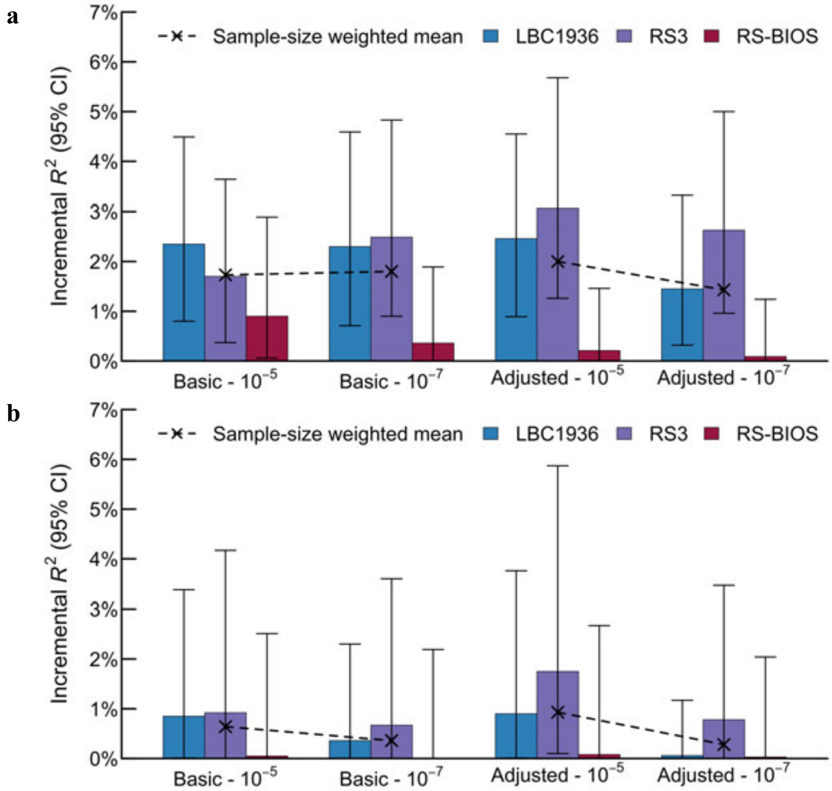

Figure 4.5. Methylation score prediction of educational attainment in independent holdout samples. Panel **(a)** displays the prediction in all individuals, and panel **(b)** displays the prediction in the subsample of never smokers. Four methylation scores were constructed: using coefficient estimates from the basic model versus adjusted model, crossed with a *P*-value threshold of 10^{-5} and 10^{-7}. The sample sizes of the LBC1936, the RS3, and the RS-BIOS cohorts are 918, 728, and 671 individuals, respectively. We performed sample-size-weighted meta-analysis across the cohorts for each of the four methylation-score prediction analyses. From left to right, the respective *P*-values testing the null hypothesis of zero predictive power are 4.42×10^{-11}, 7.76×10^{-11}, 2.02×10^{-11}, and 3.28×10^{-8} for the full sample and 0.0183, 0.0898, 0.0051, and 0.1818 for the never-smokers, respectively. The full prediction results are presented in **Supplementary Table S1.9a and S1.9b**.

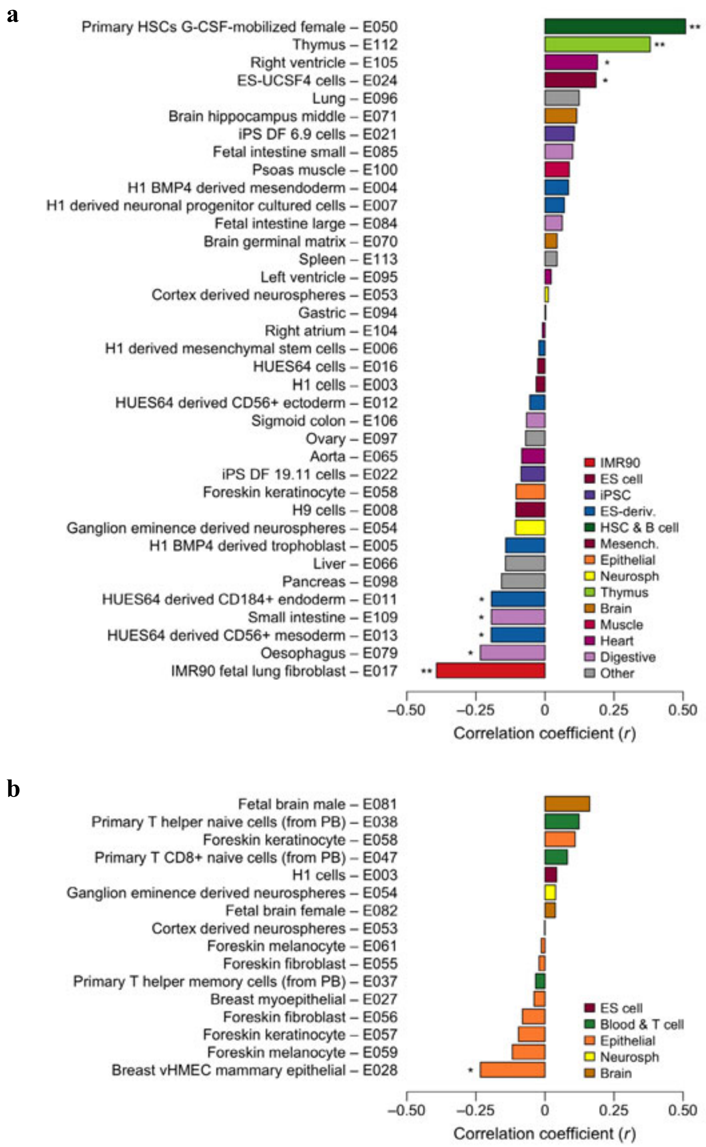

Figure 4.6. Correlations between tissue-specific methylation and the EWAS association results (adjusted model). Panel **(a)** displays the correlation estimates based on the whole-genome bisulfite sequencing (WGBS) methylation measurement, and **(b)** displays results based on the mCRF methylation measurement. (The mCRF measurement combines sequencing data from the MeDIP-seq and MRE-seq methods.) The method is described in **Supplementary Note 7**. Correlations that are significant after Bonferroni correction are marked with two asterisks (**), and marginal significance ($P < 0.05$) is marked with one asterisk (*). The tissue-specific methylation data is from the Roadmap Epigenomics Consortium, and we used their categorization and color code for simplicity of comparison[51].

Table 4.1 | EWAS association results - adjusted model

EWAS association results - adjusted model (9 probes with P-value < 1×10^{-7})										
Probe	ChrPosID	n	P-value	R^2	Closest gene	Distance closest gene TSS	Expected power in never-smokers (n = 5,175)	n (never-smokers)	P-value (never-smokers)	R^2 (never-smokers)
cg05575921	5:373379	10,315	2.03×10^{-17}	0.70%	*AHRR*	-47,656	75.1%	5,174	1.46×10^{-1}	0.04%
cg21566642	2:233284662	9,633	2.26×10^{-11}	0.46%	*ALPPL2*	13,110	33.3%	4,627	6.36×10^{-2}	0.07%
cg05951221	2:233284403	10,313	1.12×10^{-10}	0.40%	*ALPPL2*	12,851	22.2%	5,174	3.10×10^{-3}	0.17%
cg03636183	19:17000586	10,313	1.24×10^{-9}	0.36%	*F2RL3*	760	15.2%	5,172	6.55×10^{-1}	0.00%
cg01940273	2:233284935	10,316	3.84×10^{-9}	0.34%	*ALPPL2*	13,383	12.3%	5,175	3.37×10^{-2}	0.09%
cg12803068	7:45002920	10,316	8.09×10^{-9}	0.32%	*MYO1G*	6,067	10.6%	5,174	1.48×10^{-4}	0.28%
cg22132788	7:45002487	9,531	5.52×10^{-8}	0.31%	*MYO1G*	6,500	9.2%	4,334	4.35×10^{-4}	0.28%
cg06126421	6:30720081	9,718	6.63×10^{-8}	0.30%	*IER3*	-7,753	8.2%	5,174	2.98×10^{-1}	0.02%
cg21161138	5:399361	10,309	7.39×10^{-8}	0.28%	*AHRR*	-21,674	6.4%	5,170	6.59×10^{-2}	0.07%

Note: "Distance closest gene TSS" measured in base pairs. An extended version of this table is available in **Supplementary Table S1.7a**.

Chapter 4 Supplementary methods

Study overview

This meta-analysis study of educational attainment (EA) and cytosine-phosphate-guanine (CpG) methylation in human DNA was performed according to a pre-specified analysis plan. The analysis plan was publicly archived on Open Science Framework (OSF) in September 2015 (available at https://osf.io/9v3nk/), and it specifies two main analyses to be conducted at the cohort level – an *epigenome-wide association study* (EWAS)[1], and an *epigenetic clock analysis*[2]. The EWAS is considered hypothesis-free as it is performed genome-wide without an expected direction of effect for individual CpG loci (often referred to as CpG probes), while the epigenetic clock analysis is hypothesis-driven since we *a priori* expect lower EA to be associated with a faster running epigenetic clock (a higher DNA methylation age).

When we designed the study, we aimed to achieve a sample size of at least 10,000 individuals, which would lead to 80% power to detect an effect as small as $R^2 = 0.38\%$. This effect size is much smaller than those found in EWAS of other traits[3–5], but smaller effect sizes could possibly be expected as EA is a biologically distal environmental factor. To achieve the desired sample size, we were required to pool data across cohort studies and perform a meta-analysis. Due to privacy and data sharing restrictions the EWAS and epigenetic clock analysis could only be performed locally, and subsequently meta-analyzed. Cohorts could join this study by providing a signed collaboration agreement and sample descriptives during the fall of 2015. The Principal Investigator (PI) of each cohort affirmed that the results contributed to the study were based on analyses approved by the local Research Ethics Committee and/or Institutional Review Board responsible for overseeing research. All participants provided written informed consent. An overview of the 27 independent cohort studies from the 15 participating cohorts is reported in **Supplementary Table S1.1**.

In summary, the EWAS was performed with a *basic model* with covariates for age, sex and white blood-cell counts, and an *adjusted model* with additional covariates to correct for the two lifestyle factors body mass index (BMI) and smoking, which were available in all cohorts. For a full description of the control variables, see below. Since lifestyle factors such as BMI and smoking are known to be strongly associated with both methylation and EA[6–8], we focus the presentation and follow-up analyses on the results of the adjusted model as we consider this to be more conservative.

Sample inclusion criteria

The analysis plan specified the individual inclusion criteria, limiting the analysis to individuals of European ancestry, who were at least 25 years of age at the time of assessment of EA. Individuals were to be excluded if they were cases from cohorts with a case-control study design; if they did not have successfully measured methylation or did not pass other cohort-specific standard quality control (QC) filters. Cohort-level individual and CpG probe filtering is explained below. A second quality control was implemented after meta-analysis, and this procedure is described below.

The total meta-analysis sample sizes of the basic and adjusted EWAS models are respectively 10,767, and 10,317. Some individuals were excluded from the adjusted model due to missing covariates, which caused the difference in sample size.

Cohorts performed the analyses with code pre-specified in the analysis plan[9], and this code can be accessed at Open Science Framework[q]. For the EWAS the cohorts provided the following summary statistics for each probe to the meta-level analysis - CpG probe ID (“markername”), estimated regression coefficient (beta), standard error of the coefficient estimate, *P* value of the coefficient estimate, and sample size. For the epigenetic clock analyses the cohorts provided the parameter estimates for the four different age acceleration measures, standard error of the parameter estimates, *P* value of the parameter estimate, and sample size.

Phenotype definition and sample descriptive statistics

Educational attainment (EA) was harmonized across cohorts in accordance with earlier work of the Social Science Genetics Association Consortium (SSGAC)[10,11], i.e., in accordance with the ISCED 1997 classification (UNESCO)[12]. The classification consists of seven internationally comparable categories, which were subsequently translated into US years-of-schooling equivalents, which have a quantitative interpretation (**Supplementary Table S1.2**). The translated measure maintains a high level of variance in the phenotype and the cohort-specific translation is reported in **Supplementary Table S1.3,** together with the within-cohort means and standard deviations.

A summary of the 27 independent cohort studies from the 15 contributing cohorts is reported in **Supplementary Table S1.1**. The within-cohort mean age at reporting ranges from 26.6 to 79.1 years, and the minimum and maximum age is respectively 18[r] and 94 years. The sample size ranges from 48 to 1,658, with an average of 399 individuals. The average EA within the cohorts ranges from 8.6 to 18.3 years of education, and the sample-size weighted average is 13.6 (SD = 3.62). Females comprise 54.1% of the meta-analysis sample.

DNA methylation measurement and cohort-level quality control (QC)

Genome-wide CpG methylation was measured in whole blood with the Illumina 450k Human Methylation chip. Since background correction and normalization of methylation data is time-consuming and dependent on sample-specific properties, and as no method is considered overall superior, we encouraged the cohorts to perform background correction[13] and normalization[14] according to their standard QC protocols to prepare the chip data for analyses. We report the cohort-specific technical details in **Supplementary Table S1.4.**

Cohorts were recommended to implement exclusions of CpG probes and samples of individual participants using the following thresholds: probes should be excluded if the probe-detection *P* value was greater than 0.01 in more than 5% of the individuals, and individuals should be excluded if the detection *P* value was greater than 0.01 for more than 5% of the probes within an individual.

White blood cell counts, which were included as covariates to avoid confounding due to differential leukocyte cell composition across individuals, could either be measured directly or imputed based on the Houseman algorithm[15].

To summarize, we standardized the analysis protocol as much as possible, while ensuring some degree of flexibility to keep the implementation feasible for all samples.

[q] To access the code for follow-up analyses please contact the corresponding author.

[r] The FTC and LLD cohorts were allowed to additionally include individuals slightly below 25 years of age because the FTC cohort includes on-going education in the EA measurement, and the cohorts has validated the phenotype with later cohort waves (Cohen’s κ = 0.82). The LLD cohort measures EA with high enough precision to capture most of the variation even for individuals that had not yet begun higher education.

Epigenome-wide association study (EWAS)

The dependent variable in the epigenome-wide association study (EWAS) was the methylation beta-value, i.e., the proportion of methylation at a CpG locus across the measured cells within an individual. The beta-value hence lies in the interval [0, 1], and has a biologically meaningful interpretation. An alternative approach is to regress the methylation as the so-called M-value, which is the $\log_2$ transformed beta-value[16]. For comparability, we used the beta-values following the methodology of the largest EWAS studies to our knowledge[3,5,6,17]. The relation between CpG methylation and the technical covariates motivates having the methylation beta-value as the dependent variable since the technical covariates reduce the error variance leading to greater statistical power to find associations with the phenotype of interest.

Each cohort study estimated the following regression for each CpG probe passing cohort-level quality control:

$$CpG_i = \beta_0 + \beta_1 EA_i + \boldsymbol{PC_i}\boldsymbol{\gamma} + \boldsymbol{X_i}\boldsymbol{\alpha} + \boldsymbol{C_i}\boldsymbol{\theta} + \epsilon_i, \tag{6}$$

where CpG_i is the methylation beta-value for individual i, EA_i is the harmonized continuous measure of EA, $\boldsymbol{PC_i}$ is a vector of the first four principal components of the genetic relatedness matrix, and $\boldsymbol{X_i}$ is a vector of control variables further explained below. $\boldsymbol{C_i}$ is a vector containing study-specific controls and technical covariates (such as dummy variables for plates, hybridization date, and batches) that were encouraged in the analysis plan, which is discussed below.

In accordance with the pre-specified analysis plan, probes with *P* value less than 1×10^{-7} (the commonly-used threshold in epigenome-wide association studies[1]) were considered as epigenome-wide significant associations. We report two-tailed *P* values throughout the paper unless otherwise specified.

Control variables in the adjusted model

Two epigenome-wide association analyses were performed in each cohort, a *basic model* and an *adjusted model*. The two models differ in the included covariates ($\boldsymbol{X_i}$) in accordance with the pre-specified analysis plan, where the basic model controls for age, sex, and white blood-cell counts. The adjusted model additionally includes BMI (kg/m^2), and smoking measured as a categorical variable (measured either as Ever/Never smoker or Current/Former/Never smoker, depending on data availability), together with a squared age term to account for non-linear age effects, and an interaction term between age and sex. Since lifestyle factors such as BMI and smoking are known to be strongly associated with both methylation and EA[6–8], we focus the article on the results of the adjusted model as we consider this to be more conservative.

The quadratic age term and the interaction between age and sex were added to the adjusted model, because of the known non-linear relationship between age and CpG methylation (see Bollati et al., 2009; Langevin et al., 2011; Horvath et al., 2012; Bell et al., 2013; Florath et al., 2014)[18–22], as well as the relationship between sex and CpG methylation (e.g. Boks et al., 2009; Liu et al., 2010; Zhang et al. 2011)[23–25]. Moreover, levels of educational attainment (our main explanatory variable) vary by age (birthyear) and sex (see e.g. Barro and Lee, 2001)[26]. Hence, the inclusion of these control variables is both biologically and statistically warranted.

Cohort-specific control variables

After discussion with the meta-analysts, a few additional cohort-specific control variables were encouraged if they were deemed necessary to control for confounding relationships between

EA and CpG methylation, and these cohort-specific control variables were included in $\boldsymbol{C_i}$. The EPIC and MCCS cohorts did not have genetic data available to allow inclusion of genetic PCs, so the first principal component of the methylation matrix was included instead as a stringent alternative to control for possible inflation of the test statistic due to subtle population stratification. The HBCS cohort included a dummy for childhood separation exposure, KORA F4 a dummy for World War 2, and MCSS included country of birth control variables, all of which could plausibly be correlated with both EA and methylation.

Moreover, in ALSPAC, EPIC-Breast Cancer, MCCS-Breast Cancer, and MCCS-Prostate Cancer, sex was not included because these samples consist of solely females or males. In EGCUT1 smoking was not included because all participants were non-smokers. The EPIC and MCSS samples did not control for genetic PCs because of unavailability of these variables but controlled instead for the first PC of the methylation data as a stringent alternative. The age^2 term was not included in the LBC models due to the very narrow age-range in these birth cohort samples.

Description of major steps in meta-level quality-control (QC) analyses

A stringent quality-control (QC) protocol was performed to ensure that only high-quality CpG probes were meta-analyzed. All cohorts were asked to supply descriptive statistics and phenotype definitions according to the pre-specified analysis plan, and the completeness of these documents was assessed as the first step of quality control, together with examination of the uploaded EWAS summary statistics. Thereafter we applied the following probe filters.

Filters were applied to remove (a) probes with missing *P* value, standard error or coefficient estimate; (b) probes not available in the probe annotation reference by Price et al. (2013)[27]; (c) CpH probes (H = A/C/T); (d) probes on the sex chromosomes; (e) cross-reactive probes highlighted in a recent paper by Chen et al.[28]; and (f) probes with a cohort-level call rate less than 95%.

Probes were annotated if they were so-called *SNP-probes*, i.e., if the probe is located on a single-nucleotide polymorphism (SNP). This was done so that any positive results could be interpreted with caution if the associated probe would be located on a SNP, as this is known to affect the methylation status of the CpG probe[27]. We however chose to keep all SNP-probes in the results, rather than removing all of them.

The output from the quality control was examined to see if any filters removed an unusual or unexpected number of probes, and two analysts independently performed and crosschecked the QC. After probe filtering, the distributions of the coefficient estimates (betas) were compared across the cohorts to identify possible outliers and birth-year effects.

Meta-analysis

Due to the differences in the mean and standard deviation of CpG methylation across cohorts we decided to perform sample-size weighted fixed-effect meta-analysis of the cohort-level EWAS summary statistics using the METAL[29] software, as fixed effect meta-analysis is robust to differences in units of measurement. Due to the variability of λ_{GC} across cohorts we applied cohort-level genomic control to deflate the association test statistic prior to meta-analysis, equivalent to its GWAS analogue[30], to stringently control for possible population stratification that could remain even after controlling for genetic principal components in regression. In the final meta-analysis results, only probes with a meta-level sample size greater than 1,000 were considered.

Investigation of possible confounding with tobacco smoking

The adjusted EWAS model included a categorical control variable for tobacco smoking (categorized as current, former or never smoker). Due to the discrete nature of the variable and the measurement error that follows by not measuring tobacco smoking as a continuous variable, we believe that the control variable might not have completely controlled for the exact smoking exposure. Unfortunately, controlling for smoking with a categorical variable is the common approach due to the lack of a more precise smoking measure in most cohort studies. As a result, smoking could potentially bias the regression coefficient of EA because of the known negative correlation between smoking and EA[36], and the strong association between smoking and methylation found in a large number of studies[4,6]. Therefore, we performed a literature review using PubMed to see if any of the lead probes from the adjusted model have been associated with smoking in previous studies.

The literature review was performed February 24, 2016 by searching for the term "smoking" together with each of the adjusted model's nine lead probes separately. The search resulted in 30 eligible studies or smaller meta-analyses ($n < 1{,}800$) on CpG methylation and different forms of smoking exposure, one systematic review by Gao et al. (2015)[37], and two larger meta-analyses by Joehanes et al. (2016, $n = 9{,}389$)[6] and Joubert et al. (2016, $n = 6{,}685$)[4] that include many of the individual studies. The EWAS meta-analysis by Joehanes et al., 2016[6], comprised of several of the 30 individual studies, was published after the literature review had been performed and we added this study to the literature review post-hoc due to its relevance and large sample size. The result of the literature review is presented in **Supplementary Table S.1.10**.

The studies could largely be categorized as an investigation of either the relationship between a person's own smoking and methylation, or the association between newborn's methylation and maternal smoking, both at birth and at later stages in the life of the offspring. Most of the individual studies have relatively small sample sizes (smallest sample size consists of only 21 monozygotic twin pairs, and the average sample size is 471), and we note that many of the individual studies on personal smoking are meta-analyzed in Joehanes et al. (2016), and most of the individual studies on maternal smoking are meta-analyzed in Joubert et al. (2016).

We found that all of the adjusted model's nine lead probes have been associated with smoking in at least one previous study, and in **Supplementary Note 4.3** we contrast the EWAS effect sizes estimates for EA with the effect size estimates from the two meta-analyses studies. However, not all of these previous studies controlled for EA, and the associations between smoking and methylation could therefore be biased due to the omission of EA as a control variable, which makes it complicated to draw definite conclusions of the magnitude of confounding. Also, the sample sizes of many of the individual studies are small and likely to contain a higher number of false positives than the meta-analyses.

In light of this finding we interpret our EWAS findings carefully, and the literature review motivated the *post-hoc* sensitivity analyses in the subsample of people that never smoked presented in the next section.

Robustness of EWAS results in the never-smoker subsample

To further investigate the possible confounding effect of smoking on the association between EA and CpG methylation, all cohorts reran the EWAS in the subsample of individuals that reported as 'never smoker'. We refer to this subsample as the *never smokers*. The full EWAS sample consists of the *ever smokers* (i.e., the individuals who answered that they were ever, current, or former smokers) and the never smokers. This approach would give us confidence in the associations of the adjusted model if the lead probes were found to be associated with EA in the never-smoker subsample. We note that this sample could be biased in other ways when selected on this specific variable, and the statistical power will be lower due to the decrease in sample size, and we therefore do not perform any further analysis of any other probes than the nine lead probes in this subset.

The meta-analysis sample size of the never-smoker subsample was 5,175, and with the effect-size estimates from the adjusted model we calculated the power to detect the effects of the lead probes in the never-smoker subsample. The expected power ranged from 6.4% to 74% (**Supplementary Table S1.7a**), at P value < 10^{-7}, and the expected number of replications was 1.93 given the expected power. If a more liberal Bonferroni-corrected threshold would be used, such as 0.05/9, then the power in the never smokers ranged from 81% to 100%. Hence, we expected considerable statistical power to replicate the lead probes of the adjusted model in the never-smoker subsample assuming the full sample effect size estimates.

After quality control, we performed a meta-analysis of the re-estimation in the never-smoker subsample, and the results are displayed in **Supplementary Table S1.7a**. None of the 9 lead probes were significant at the stringent epigenome-wide threshold of $P < 10^{-7}$, while 2 probes, cg12803068 and cg22132788, were estimated with strong associations even though the sample size was halved (with respective P values of 1.48×10^{-4} and 4.35×10^{-4}).

Joint test of no association in the never-smoker subsample

Next, we performed a joint test of the null hypothesis that the regression coefficients of the lead probes are all equal to zero, with the sample-size weighted Z statistic estimated in the never smokers. Since we have no a priori hypothesis of a direction of effect for different CpG probes, and to avoid that positive and negative Z-statistics cancel out, we performed this test on the absolute value of the Z-statistic with the following weighting

$$Z_{Combined} = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}} \qquad (2)$$

where

$$w_i = \sqrt{N_i}$$

The right-tailed P value of the joint test of no association was 2.18×10^{-11}. As a robustness check, we also performed this test while pruning the lead probes so that only the strongest association within 250kb was kept. In that case, we observed a right-tailed P value of 4.91×10^{-6}, based on five probes. We interpret this as evidence that at least one probe has an estimated effect different from zero in the meta-analysis of the never smokers (however see **Supplementary Note 4.4** for more discussion of maternal smoking as a confounding factor).

Effect size comparison ever versus never smokers

Next, we investigated whether the effect-size estimates changed between estimation performed in the ever and never smokers. To be able to compare the differences in effect sizes and the precision of the estimates (*SE*), we used inverse-variance weighted meta-analysis to retain meta-analytical betas and standard errors for the nine lead probes. We thereafter derived the effect sizes and standard errors of the ever smokers by assuming that the effect size estimates in the full sample is a weighted average of the effect size estimates in the ever smokers and never smokers, as the two groups are mutually exclusive and the union between the two is the full EWAS sample.

The result of the effect size comparison is presented in **Supplementary Figure 4.3**. We aligned the effect sizes to the first quadrant by taking their absolute values so that all effect sizes were positive for the comparison. If the effect sizes would be similar across the subsamples we would expect them to align along the 45-degree line, and if confounding would be driving the effect-size estimates then we would expect a cluster above the 45-degree line as the effect-size estimates in the ever smokers would be greater than those in the never smokers.

By visual inspection we observed a cluster of probes with larger effect size estimates in the ever smokers than in the never smokers (i.e., those above the 45-degree line). We also observed a second cluster of effect size estimates with similar estimates in both subsamples (i.e., those along the 45-degree line). The two lead probes that we found suggestively associated in the never smokers i.e., cg12803068 and cg22132788, have slightly larger effect size estimates in the ever smokers than in the never smokers, however these larger estimates lie within the confidence intervals of the estimates in the ever smokers. The effects of the other seven probes were estimated to be more than 50% smaller in the never smokers than in the ever smokers, and all of these effect size estimates are outside the confidence intervals of the estimates among ever smokers.

Lookup of probes in EWAS meta-analysis on smoking and maternal smoking during pregnancy

In **Figure 4.3** and **Supplementary Table S1.11** we present the effect-size estimates, in terms of R^2, of the 44 probes of the adjusted model with P value less than 1×10^{-5} (including the nine lead probes). The probes are listed in descending order based on their effects, for comparison, we display their effects when re-estimated in the never-smoker subsample, as well as their effects as reported in the recent EWAS meta-analyses on smoking by Joehanes et al. (2016)[6] and maternal smoking by Joubert et al. (2016)[4]. We report the effect sizes for smoking and maternal smoking if a probe was significantly associated in these studies at FDR < 0.05, as those results were publicly available.

Notably, the effect sizes of smoking and maternal smoking are many times larger for most probes compared to the effect of EA. Further, the suggestively associated probes cg12803068 and cg22132788, and the closest gene; *MYO1G*, have all been reported as affected by maternal smoking in newborn's, and persistently later in life[4,38]. The effect of maternal smoking on these two probes is extreme; their R^2 are 8.34% and 6.33%, respectively. In our study, we do not have access to individual-level information on maternal smoking. Thus, we cannot distinguish between the hypothesis that these probes are truly associated with EA and the hypothesis that the observed association with EA is entirely driven by a greater exposure to maternal smoking during pregnancy among lower-EA individuals. The extreme effects of smoking and maternal smoking substantially weaken the support of the premise that these probes would be directly affected by EA, rather than indirectly via smoking. Also, we cannot rule out other potential confounders, such as exposure to second-hand smoke. The lookup strongly suggests that smoking and maternal smoking are worrying confounding factors for the probe associations with EA, even in the subsample of individuals who self-report as never smokers.

Enrichment analyses

Enrichment analysis of GWAS of smoking and EA

We performed an enrichment analysis to investigate whether the genetic loci, in proximity with the probes found to be associated in the adjusted EWAS model (at $P < 1\times10^{-4}$), also contain SNPs enriched for EA and smoking. For EA, we used the GWAS summary-statistics from a large ($n = 405{,}073$), recent meta-analysis by Okbay et al. (2016)[34]. For smoking, we used a meta-analysis combining GWAS results from the UK Biobank together with the publicly available summary statistics from the Tobacco, Alcohol, and Genetics Consortium (TAG, total $n = 186{,}102$)[39]. The latter meta-analysis is described in detail in Karlsson Linnér et al. (2017)[40].

The enrichment analysis was performed with the 179 EWAS probes from the adjusted model with P value less than 1×10^{-4}. We pruned the probes so that the probe with the lowest P value in a locus was selected as the lead probe, and all probes within 250kb from the lead probe was clumped with the strongest association. This resulted in 141 "approximately independent" lead probes (k). For each of the phenotypes, we extracted the closest SNP available in the GWAS summary statistics based on the physical position of the k lead probes. Using the reference panel 1000 Genomes phase 3 (October 2014 haplotype release version 5), we extracted the SNPs in strong LD ($r^2 >= 0.8$) with the k SNPs. Next, we averaged the absolute value of their GWWAS Z statistics, as well as the minor allele frequency (MAF) of the SNPs in strong LD. This leads to a distribution of k average Z statistics and average MAF.

From the summary statistics we extracted a set of 1,000 random SNPs for each of the k SNPs, matched on the average MAF (+/- 1pp), and for the matched SNPs we also averaged the absolute value of the Z statistics across the SNPs in strong LD, just as for the 141 "first-stage" SNPs. This leads to a distribution of $k\times1{,}000$ average absolute Z statistics. Next, we ordered the Z statistics and performed a test of joint enrichment with the non-parametric Mann-Whitney test of the null hypothesis that the association test statistic of the k SNPs are drawn from the same distribution as those of the 141,000 MAF-matched SNPs.

Prediction of EA with polygenic methylation scores

To test the out-of-sample predictive power of our EWAS findings, we constructed polygenic methylation scores (PGMS) to perform prediction of EA. The prediction was performed in three independent cohort studies, the Lothian Birth Cohort 1936 (LBC1936, $n = 917$), Rotterdam Study BIOS (RS-BIOS, $n = 671$), and Rotterdam Study 3 (RS3, $n = 729$). For each of the prediction cohorts we created a new EWAS meta-analysis withholding the respective prediction cohort study to avoid overfitting[42]. The effect sizes (in terms of Z statistics) from the respective holdout meta-analysis were used as weights when constructing the PGMS. The Z statistic were used instead of the EWAS coefficients because CpG methylation is the dependent variable in the EWAS regression.

For each prediction sample we created four PGMS using two P-value thresholds $P < 10^{-5}$ and $P < 10^{-7}$, and the estimates from the basic and adjusted models. The PGMS was defined as the weighted sum of the Q CpG probes:

$$\widehat{M}_i = \sum_{j=1}^{q} \hat{Z}_j \, \mathrm{B}_{ij} \, , \qquad (5)$$

where $\widehat{M}_i$ denotes the methylation score of individual i, $\hat{Z}_j$ is the estimated Z-statistic for CpG probe j, and B_{ij} is the methylation beta-value for individual i at CpG probe j.

A recent study showed that the phenotypic prediction can be improved if the PGMS is combined with a single-nucleotide polymorphism polygenic score (SNP PGS)[43]. We therefore included a SNP PGS constructed with the effect size estimates from a recent large-scale GWAS on EA (n = 405,073)[11], and the prediction cohorts were excluded from the GWAS meta-analysis to avoid overfitting. The SNP PGS was defined as the weighted sum of the T directly genotyped SNPs:

$$\hat{S}_i = \sum_{j=1}^{t} \hat{\beta}_j \, g_{ij} \, , \quad (6)$$

where $\hat{S}_i$ denotes the polygenic score of individual i, $\hat{\beta}_j$ is the estimated additive effect size of the effect-coded allele at SNP j, and g_{ij} is the genotype of individual i at SNP j (coded as having 0, 1 or 2 instances of the effect-coded allele)[44].

For each prediction cohort study we performed an OLS regression with EA as the dependent variable, and age and sex as control variables. The predictive power of the PGMS was evaluated as the incremental R^2 of adding the PGMS to the regression model, i.e., the difference in R^2 between the regression model with only the SNP PGS and the covariates, and the regression model with the PGMS together with the SNP PGS and the covariates. The incremental R^2 of the interaction term (between the PGMS and SNP PGS) was evaluated as the difference in R^2 between the regression model with the PGMS and the SNP PGS as additive main effects, together with the covariates, and the same regression model with the interaction term added.

To estimate 95% confidence intervals for the incremental R^2 we performed bootstraps with 1,000 samples with the percentile interval method, and this was done with the 'boot' package in R[9,45,46]. Finally, the incremental R^2 was meta-analysed across the three prediction cohorts (LBC, RS-BIOS, and RS3) as the sample-size weighted incremental R^2.

ALSPAC prediction of educational achievement

A further prediction analysis, using the same PGMS prediction method as described in **Supplementary Note 6**, was carried out in the ARIES substudy (n = 678) of the ALSPAC cohort[47,38]. Cord-blood methylation was assessed in newborn children and EA was measured using four sets of standardised Key Stage[48] school grades from primary level through to high school (educational achievement was evaluated at ages 7-16 years), and the Key Stage educational achievement test scores are part of the state education system in the UK[48]. The cord blood signatures were collected prior to any educational exposure and may help identify if methylation differences might be a cause or consequence of EA. For each of the Key Stages 1 to 4, an average score across all school subjects was derived for each child. The predictive accuracy of the PGMS, built using the association Z-statistic from a meta-analysis excluding the ALSPAC cohort, was assessed as the incremental adjusted R^2 of adding the score to the OLS regression model, with sex and age at assessment as control variables. Secondly, we tested if these associations were robust to the inclusion of maternal smoking as an additional control variable. Finally, we performed the prediction in a restricted sample using data from children of non-smokers only.

Correlation of EWAS associations with tissue-specific methylation

The NIH Roadmap Epigenomics Consortium recently made an impressive analysis and categorization of a multitude of different epigenetic marks, across 111 different tissues[51]. The data is publicly archived, and we harnessed their tissue-specific methylation data to answer the question of whether our EWAS associations are correlated with any tissue-specific DNA methylation. We hypothesize that EWAS associations for a given phenotype would be more likely to be located at loci that are differentially methylated in tissues relevant for the phenotype or endophenotypes. E.g., if our EWAS associations would be correlated with the tissue-specific methylation of brain tissues, that would increase the credibility of the associations, and improve the biological interpretation.

Genome-wide methylation data was available for three kinds of CpG methylation measurements: whole-genome bisulfite sequencing (WGBS), reduced representation bisulfite sequencing (RRBS), and mCRF (a method combining sequencing data from the MeDIP-seq and MRE-seq methods). Methylation was measured as the beta value, ranging from 0 to 1, and it was available for 37 tissues measured with WGBS, 49 tissues with RRBS, and for 16 tissues with mCRF. Only eight tissues were available for two or more methylation measurements.

We pruned the probes of the adjusted EWAS model with *P*-value less than 1×10^{-4} using a window of 250kb, as described in **5.1 Enrichment analysis of GWAS of smoking and EA**, which resulted in 141 "approximately independent" probes. Based on the physical location of the 141 pruned probes we extracted the beta-value for WGBS, RRBS, and mCRF. The beta-values were first converted to M-values[s], and thereafter we calculated standardised tissue-specific deviations ($Z_{CpG,t}$) from the cross-tissue average methylation ($\overline{Mval}_{CpG}$) for each of the 141 CpG loci as:

$$Z_{CpG,t} = \frac{Mval_{CpG,t} - \overline{Mval}_{CpG}}{s.d(Mval_{CpG})} \tag{7}$$

where $Z_{CpG,t}$ is the tissue-specific deviation for tissue *t,* at locus *CpG.* This procedure was performed within each methylation measurement. Hence, in the case that a tissue had no missing CpG loci, there were 141 tissue-specific deviations; each corresponding to one of the loci identified in the association results of the adjusted EWAS model. For RRBS, there were many missing values across all 49 tissues, and the maximum number of loci available was 38 out of 141. We therefore excluded the RRBS measurement from further analysis, while both WGBS and mCRF had close to 141 overlapping loci for all tissues (**Supplementary Table 1.14**).

For WGBS and mCRF we calculated correlations with the tissue-specific *Z*-statistics and the EWAS association *Z*-statistics of the pruned probes (from the adjusted EWAS model). Bonferroni correction for the number of tissues within each methylation measurement was performed.

methQTL analysis

It has been shown that genetic variants also explain variation in CpG levels[1] in addition to environmental influences on DNA methylation. EWAS probes under genetic influence may help us understand the direction of association between CpGs and outcomes, but it can also be a confounding factor. SNPs affecting the level of methylation, usually referred to as methylation quantitative trait loci (methQTL)[1], sometimes have effect sizes that can be found in samples of less than a thousand individuals[54,55], which is much less than what is necessary for most GWAS of e.g. behavioral phenotypes[34]. Therefore, as is customary in the EWAS literature, we performed a GWAS for each of the 9 lead probes to investigate if any SNPs were associated with the level of methylation. The genome-wide association analysis was performed in the

[s] As the consortium methylation data was only available with two decimals precision, we imputed 0's and 1's with 0.005 and 0.995 to avoid infinite M-values.

LBC1936 (n = 918) and RS3 (n = 731) cohorts that estimated the following GWAS regression equation for each autosomal bi-allelic SNP and lead probe:

$$Y_i = \beta_0 + \beta_1 SNP_i + \boldsymbol{PC_i\gamma} + \boldsymbol{X_i\alpha} + \boldsymbol{C_i\theta} + \epsilon_i, \tag{8}$$

where Y_i is the methylation beta-value for individual i, SNP_i is the number of reference alleles of the SNP, $\boldsymbol{PC_i}$ is a vector of the first four principal components of the genetic relatedness matrix, and $\boldsymbol{X_i}$ is a vector of the control variables age and sex, as well as an interaction term between age and sex. $\boldsymbol{C_i}$ is a vector containing study-specific controls and technical covariates, such as dummy variables to control for genotyping array and batch.

Quality control of methQTL analyses

The GWAS results of the methQTL analyses were quality controlled according to a stringent protocol by the GIANT consortium[56], and the protocol was implemented with the EasyQC software. In summary the following major filters were applied; removal of monomorphic and multi-allelic SNPs, and structural variants such as INDELs; removal of SNPs with an IMPUTE imputation quality < 0.7; and of SNPs with a minor allele frequency (MAF) < 0.05. The quality control procedure ensures that all SNPs have alleles aligned to the 1000G phase 3, version 5 (October 2014 haplotype release)[57], and SNPs that could not be aligned with the reference were removed.

References – Chapter 4 Supplementary methods

1. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* 2012; **12,** 529–541.

2. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol.* 2013; **14**, 1-19 (2013).

3. Mendelson MM, Marioni RE, Joehanes R, Liu C, Hedman ÅK, Aslibekyan *et al.* Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease : A Mendelian Randomization Approach. *PLoS Med.* 2017; **14**: 1–30.

4. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C *et al.* DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am. J. Hum. Genet.* 2016; **98**: 680–696.

5. Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T *et al.* DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biol.* 2016; **17**: 1–15.

6. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR *et al.* Epigenetic Signatures of Cigarette Smoking. *Circ. Cardiovasc. Genet.* 2016; **9**: 436-447.

7. Demerath EW, Guan W, Grove ML, Aslibekyan S, Mendelson MM, Zhou Y-H *et al.* Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Hum. Mol. Genet.* 2015; **24**: 4464–4479.

8. Johnson W, Kyvik KO, Mortensen EL, Batty GD, Deary IJ. Does education confer a culture of healthy behavior? Smoking and drinking patterns in Danish twins. *Am. J. Epidemiol.* 2011; **173**: 55–63.

9. Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. [Internet]. 2015 [cited 4 April 2016]. Available from: <https://cran.r-project.org/web/packages/boot/>

10. Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 2013; **340**: 1467–1471.

11. Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 2016; **533**: 539–542.

12. United Nations Educational, Scientific and Cultural Organization. International Standard Classification of Education [Internet]. 2006 [cited 18 September 2015]. Available from: <http://www.uis.unesco.org/Library/Documents/isced97-en.pdf>

13. Triche TJ, Weisenberger DJ, Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* 2013; **41**: 1–11.

14. Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* 2014; **15**: 503-520.

15. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH

et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 2012; **13**: 86-102.

16. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Lifang H *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 2010; **11**: 587-596.

17. Liu C, Marioni RE, Hedman ÅK, Pfeiffer L, Tsai P-C, Reynolds LM *et al.* A DNA methylation biomarker of alcohol consumption. *Mol. Psychiatry* 2016; e-pub ahead of print 15 November 2016; doi:10.1038/mp.2016.192

18. Bollati V, Schwartz J, Wright R, Litonjua A, Tarantini L, Suh H *et al.* Decline in genomic DNA methylation through aging in a cohort of elderly subjects. *Mech. Ageing Dev.* 2009; **130**: 234–239.

19. Langevin SM, Houseman EA, Christensen BC, Wiencke JK, Nelson HH, Karagas MR *et al.* The influence of aging, environmental exposures and local sequence features on the variation of DNA methylation in blood. *Epigenetics* 2011; **6**: 908–919.

20. Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MPM, Eijk, K *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* 2012; **13**: R97.

21. Bell JT, Yang TP-C, Pidsley R, Nisbet J, Glass D, Mangino M *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet.* 2012; **8**: e1002629.

22. Florath I, Butterbach K, Müller H, Bewerunge-Hudler M, Brenner H. Cross-sectional and longitudinal changes in DNA methylation with age: An epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Hum. Mol. Genet.* 2012; **23**: 1186–1201.

23. Boks MP, Derks EM, Weisenberger, DJ, Strengman E, Janson E, Sommer IE *et al.* The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS One* 2009; **4**: 21–23.

24. Liu J, Morgan M, Hutchison K, Calhoun VD. A study of the influence of sex on genome wide methylation. *PLoS One* 2010; **5**: e10028.

25. Zhang FF, Cardarelli R, Carroll J, Fulda KG, Kaur M, Gonzalez K *et al.* Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. *Epigenetics* 2011; **6**: 623–629.

26. Barro RJ, Lee J-W. International data on educational attainment: updates and implications. *Oxf. Econ. Pap.* 2001; **53**: 541–563.

27. Price ME, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* 2013; **6**: 4-19.

28. Chen Y-A, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013; **8**: 203–209.

29. Willer CJ, Li Y, Abecasis GR. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26**: 2190–2191.

30. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ *et al.* Genomic

inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* 2011; **19**: 807–812.

31. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.

32. Iterson M, Zwet E, the BIOS Consortium, Heijmans BT. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.* 2017; **18**: 1–13.

33. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. *Stat. Med.* 2014; **33**: 1946–1978.

34. Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 2016; **533**: 539–542.

35. Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai P-C *et al.* DNA methylation-based measures of biological age: Meta-analysis predicting time to death. *Aging* 2016; **8**: 1844–1865.

36. Johnson W, Kyvik KO, Mortensen EL, Batty GD, Deary IJ. Does education confer a culture of healthy behavior? Smoking and drinking patterns in Danish twins. *Am. J. Epidemiol.* 2011; **173**: 55–63.

37. Gao X, Jia M, Zhang Y, Breitling LP, Brenner, H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clin. Epigenetics* 2015; **7**, 113-123.

38. Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL *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**: 2201–2217.

39. The Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* 2010; **42**: 441–447.

40. Karlsson Linnér R, Beauchamp JP. Large-scale genetic study of risk tolerance and risky behaviors identifies new loci and reveals shared genetic influences. *Manuscript in preparation for publication* 2017.

41. Rietveld CA, Esko T, Davies G, Pers TH, Turley P, Benyamin B *et al.* Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proc. Natl. Acad. Sci.* 2014; **111**: 13790–13794.

42. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* 2013; **14**: 507–515.

43. Shah S, Bonder MJ, Marioni, RE, Zhu Z, McRae AF, Zhernakova A *et al.* Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. *Am. J. Hum. Genet.* 2015; **97**: 75–85.

44. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* 2013; **9**: e1003348.

45. Davison AC, Hinkley DV. *Bootstrap Methods and Their Application.* Cambridge University Press: Cambridge, UK, 1997.

46. The R Core Team. R: A language and environment for statistical computing [Internet]. 2013 [cited 4 April 2016]. Available from: <https://cran.r-project.org/doc/manuals/r-

release/fullrefman.pdf>.

47. Relton CL, Gaunt T, McArdle W, Ho K, Duggirala A, Shihab H *et al.* Data Resource Profile : Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int. J. Epidemiol.* 2015; **44**: 1181–1190.

48. GOV.UK. National Curriculum [Internet]. 2016 [cited 14 January 2017]. Available from: <https://www.gov.uk/national-curriculum>

49. 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; **8**: e63812.50. Fraser, A. *et al.* Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort. *Int. J. Epidemiol.* **42,** 97–110 (2012).

51. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Erst J, Bilenky M, Yen A *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015; **518**: 317–330.

52. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 2013; **45**: 580–585.

53. Watanabe K, Taskesen E, Bochoven A, Posthuma D. FUMA: Functional mapping and annotation of genetic associations. *Manuscript submitted for publication* 2017.

54. Quon, G., Lippert, C., Heckerman, D. & Listgarten, J. Patterns of methylation heritability in a genome-wide analysis of four brain regions. *Nucleic Acids Res.* **41,** 2095–2104 (2013).

55. Lemire M, Zaidi SHE, Ban M, Ge B, Aïssi D, Germain M *et al.* Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat. Commun.* 2015; **6**: 6326.

56. Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* 2014; **9**: 1192–1212.

57. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56–65.

58. Gibbs JR, Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai S-L *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in Human Brain. *PLoS Genet.* 2010; **6**: e1000952.