An epigenome-wide association study of educational attainment (*n* = 10,767)

Authors: All authors and their affiliations appear at the end of the paper*

Abstract: The epigenome has been shown to be influenced by 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 of 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. While we find that 9 CpG probes are significantly associated with EA, only two remain associated when we restrict the sample to never-smokers. These two 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, their effect sizes on EA are far smaller than the known associations between CpG probes and biologically proximal environmental factors. Two analyses that combine the effects of many probes—polygenic methylation score and epigenetic-clock analyses—both suggest small associations with EA. 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.

Key words: Epigenetics, EWAS, Educational Attainment, Epigenetic Clock, Smoking

Introduction:

The epigenome is known to be influenced by biological factors such as disease status[1,2]. There is a widespread perception that the epigenome is also affected by a variety of social environmental factors[3–10], but 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-analysing harmonised EWAS results across 27 cohort studies, we are able to attain an overall sample size of 10,767 individuals, 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 use data on CpG DNA methylation. This is the most widely studied epigenetic mark in large cohort studies[1]. Methylation level is measured by the beta value, which is the proportion of methylated molecules at each CpG site, a continuous variable ranging between 0 and 1[21]. The Illumina 450k Bead Chip measures methylation levels at over 480,000 sites in human DNA and has been used in many cohort studies[1].

We report results from both of the two common approaches for the analysis of such methylation datasets. Our main analysis is an EWAS, which considers regression models for each CpG methylation site with EA. We also present evidence from an 'epigenetic clock'[22,23] analysis, which uses a weighted linear combination of probes (i.e. measured CpG methylation sites) 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

Data for the EWAS and epigenetic clock analyses were available in 27 independent cohort studies from 15 cohorts from across Europe, the US, and Australia (**Supplementary Table S1.1**), and the total sample size comprised 10,767 individuals.

Ethics and preregistration

All contributing data were collected and analysed according to cohort-specific Local Research Ethics Committees or Institutional Review Boards, and the 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/). All participants provided written informed consent.

Educational attainment measures

Following earlier work of the Social Science Genetics 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-S1.3**).

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. 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**.

Statistical analysis

Epigenome-wide association study (EWAS)

We first ran an EWAS of EA to investigate associations with individual methylation markers (**Supplementary Note 1.4**). As is standard, the EWAS was performed as a set of linear regressions, one methylation marker at a time, with 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: probes with missing *P*-value, standard error, or coefficient estimate; probes with a call rate less than 95%; probes with a combined sample size less than 1,000; probes not available in the probe-annotation reference by Price et al. (2013)[40]; CpH probes (H = A/C/T); probes on the sex chromosomes; and 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 correct the meta-analysis P-values for possible unaccounted-for population stratification. Probes with a P-value less than 1×10^{-7}, a commonly-used threshold in EWAS[1], were considered epigenome-wide significant associations.

Polygenic predictions with polygenic methylation score

Analogous to polygenic-score prediction analyses from GWAS, a polygenic methylation prediction analysis was performed (**Supplementary Note 1.7**). We tested predictive power in three independent 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 three times, each time holding out one of the prediction samples. In each prediction sample, we constructed each polygenic methylation score (PGMS) as a weighed sum of the methylation markers' beta values, using the Z-statistics from the EWAS as weights. The Z-statistics are used instead of the EWAS coefficients because CpG methylation is the dependent variable in the EWAS regression. We constructed PGMSs using two different thresholds for probe inclusion, $P < 1 \times 10^{-5}$ and $P < 1 \times 10^{-7}$, and using the EWAS results from the basic and adjusted EWAS models, for a total of four PGMSs in each prediction sample.

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[44]. We constructed a 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 sample are average educational achievement test scores (Key Stages 1–4[45]) from age 7 up to age 16 years.

For both our adult and child samples, we measure the predictive power of the PGMS by examining the incremental coefficient of determination (incremental R^2) from adding the PGMS to the regression model for predicting EA (or test scores), i.e., the difference in R^2 between the regression model without the PGMS and the regression model with the PGMS added as a predictor.

To enable us to examine the relationship between epigenetic and genetic associations, we also constructed single-nucleotide polymorphism (SNP) polygenic scores for EA. To do so, we used SNP genotype data available in the three adult prediction cohort studies (LBC1936, RS-

5

BIOS, and RS3). We constructed a polygenic score in each sample as a weighted sum of 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[35] after excluding the LBC1936, RS-BIOS, and RS3 samples, respectively. We examine the incremental R^2 from adding the PGMS to a regression model for predicting EA that already includes the polygenic score. We then re-estimate the model after we add an interaction term between the PGMS and the polygenic score, and we assess the incremental R^2 compared to the model with the PGMS and the polygenic score as additive main effects.

Epigenetic clock analyses

To construct our epigenetic clock variables, we entered the raw beta-value data 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[46].

These Clock measures are annotated in the Horvath software as follows: 'AgeAccelerationResidual', 'AAHOAdjCellCounts', 'AAHAAdjCellCounts', and 'BioAge4HAAdjAge'. We analysed two regression models, both with EA as the dependent variable and a clock variable as an independent 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 an independent variable. For each of the eight regressions, we performed a sample-size weighted meta-analysis of the cohort-level results.

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

EWAS Quality Control (QC)

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, has 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 is 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[43]. The meta-analysis λ_{GC} is 1.19 for the basic model and 1.06 for the adjusted model.

EWAS results

Figure 1 shows the Manhattan plot for the meta-analysis results of the adjusted model. The Manhattan plot for the basic model, as well as the QQ plots for the basic and adjusted model, are in the **Supplementary Note**. In the basic model there are 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 are 9 associated probes, listed in **Table 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."

7

The effect sizes of the associations for the 9 lead probes are shown in **Table 1**. The coefficients of determination (R^2's) range from 0.3% to 0.7%. To put these effect sizes in perspective, **Figure 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[35]. The EA EWAS associations we find are an order of magnitude larger than the largest EA SNP effect sizes from the GWAS. However, our EWAS associations for EA are small in magnitude relative to the EWAS associations that have been 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 for probes associated with smoking and maternal smoking are greater by factors of roughly 3 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., the amount and duration of smoking. Therefore, we performed a lookup of our lead probes in published EWAS of smoking (**Supplementary Note 2.3**). The literature review is reported in **Supplementary Table S1.9**. We found that all 9 lead probes have previously been associated with smoking in at least one study. This evidence, however, is difficult to interpret because: (a) since not all studies included EA as a covariate, it is possible that these associations are truly due to a relationship with EA rather than with smoking, and (b) many of the published smoking EWAS are based on small samples and hence likely have an inflated rate of spurious associations that could include EA-associated probes just by chance. Nonetheless, the results of this literature review 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 analyses 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 ($n = 5,175$).

After re-estimation, none of the 9 lead probes are associated at the stringent epigenome-wide significance threshold ($P < 10^{-7}$). However, we strongly reject the joint null hypothesis that none of the 9 probes is associated with EA (one-tailed P-value of 2.18×10^{-11}; for more details, see **Supplementary Note 2.3.2**). Relative to their effect sizes in the full EWAS sample, in the never-smoker subsample 7 out of the 9 probes have effect sizes (in terms of R^2) that are reduced by more than 60% (see **Table 1** and panel A in **Figure 3**). In contrast, the effect sizes of two probes, cg12803068 and cg22132788, exhibit little or no attenuation. These are the probes for which the evidence of association with EA remains strongest ($P = 1.48 \times 10^{-4}$ and $P = 4.35 \times 10^{-4}$, 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[12,47]. 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 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.1.10**). Since effect size estimates for smoking and maternal smoking were not available to us for all the probes, and since all the effect sizes are estimated with error, we cannot draw firm conclusions. Nonetheless, the comparison strongly suggests 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 confounders for the probe associations with EA, even in the subsample of individuals who are self-reported never-smokers.

Prediction using polygenic methylation scores

The incremental R^2's from the prediction of EA with polygenic methylation scores (PGMSs) in our adult prediction cohort studies, the LBC1936, RS-BIOS, and RS3, are reported in **Supplementary Table S1.10a** and in **Figure 4.** Across the four PGMSs constructed from the

9

basic and adjusted model, constructed with estimates from the full EWAS sample, and from both probe-inclusion thresholds ($P < 10^{-5}$ and $P < 10^{-7}$), the incremental R^2's range from 1.4% to 2.0% ($P < 3.28 \times 10^{-8}$ and lower). There is also weak evidence for an interaction between the PGMS and the polygenic score in predicting EA, with R^2's for the interaction term ranging from 0.1% to 0.3% (P-values ranging from 0.01 to 0.12).

Relative to the prediction in the full sample, in the subsample of never-smokers the PGMSs (constructed with weights derived from the full EWAS sample) predict less of the variance in EA. As shown in **Figure 4** and **Supplementary Table S1.10b**, the incremental R^2's range from 0.3% to 0.9%. Two PGMSs constructed from probes reaching $P < 1 \times 10^{-5}$ in the EWAS are associated with EA at significance threshold $P < 0.05$, while the two PGMSs constructed only from the lead probes reaching $P < 1 \times 10^{-7}$ are not ($P > 0.05$). No interaction effect between the PGMS and the polygenic score is found in the never-smoker subsample.

Turning to the child sample in the ALSPAC ARIES cohort[44,47], we examine whether a PGMS constructed from methylation assessed in cord blood samples at birth is predictive of four prospective measures of educational achievement test scores (Key Stage 1–4[45]), collected between ages 7 and 16 (**Supplementary Note 1.7.1**). The results are reported in **Supplementary Table S1.10c**. The highest incremental R^2 is 0.73% ($P = 0.0094$), and it is attained in the model predicting school performance at age 14-16 (i.e. the Key Stage 4 test scores). However, once maternal smoking status is added as a control variable, the predictive power of the PGMS is essentially zero ($R^2 = 0.05\%$, $P = 0.234$). This suggests that the predictive power is driven by the confounding effects of maternal smoking. We draw two conclusions from these results from the child sample. First, they reinforce the concern that maternal smoking is a major confound for any probe associations with EA. Second, they suggest that any true methylation-EA associations are 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 in genomic regions that contain SNPs that have been identified in GWAS of EA or GWAS of smoking. Considering jointly the 141 approximately independent EWAS probes reaching $P < 10^{-4}$, we did not find evidence of enrichment for either EA-linked SNPs ($P = 0.206$) or smoking-linked SNPs ($P = 0.504$) (see **Supplementary Note 2.4**). Considering the probes individually, one

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

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 are 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 results are presented in **Supplementary Table S1.11**. There is no evidence for an association between EA and Clocks 1, 2, or 3, but the association between EA and Clock 4 is strong ($P = 3.51 \times 10^{-6}$ and $P = 4.51 \times 10^{-4}$ in the basic and adjusted model, respectively). The point estimates are small, however: using Clock 4, each year of EA is 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 is associated with slightly younger biological age when compared with chronological age. We note that the epigenetic clock that is found to be associated with EA, Clock 4, has previously been found to be the most predictive epigenetic clock measure of mortality[46].

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 9 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 7 out of the 9 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 confounds for

11

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.

Convincingly establishing a causal effect of EA would require analysing 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[48]. 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 not other types of epigenetic modifications.

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 confounds, 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.

References

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.   Veldhoven, K, V. *et al.* Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis. *Clin. Epigenetics* **7,** 67 (2015).

3.   Champagne, F. A., Curley, J. P. 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).

14

12.   Joubert, B. R. *et al.* DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am. J. Hum. Genet.* **98,** 680–696 (2016).

13.   Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circ. Cardiovasc. Genet.* **9,** 436-447 (2016).

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* **00**, 1–12 (2016).

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. *PLoS Med.* **14,** 1–30 (2017).

18.   Ligthart, S. *et al.* DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biol.* **17,** 1–15 (2016).

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* **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. Rietveld, C. A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340,** 1467–1471 (2013).

35. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with

educational attainment. *Nature* **533,** 539–542 (2016).

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. A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8,** 203–209 (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. Relton, C. L. *et al.* Data Resource Profile : Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int. J. Epidemiol.* **44,** 1181–1190 (2015).

45. GOV.UK. National Curriculum. (2016). Available at: <https://www.gov.uk/national-curriculum>

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

47. 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).

48.   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).

Richard Karlsson Linnér[*,1,2], Riccardo E Marioni[*,3,4,5], Cornelius A Rietveld[*,2,6,7], Andrew Simpkin[8], Neil M Davies[8], Kyoko Watanabe[1], Nicola J Armstrong[9], Kirsi Auro[10,11], Clemens Baumbach[12], Marc Jan Bonder[13], Jadwiga Buchwald[10], Giovanni Fiorito[14,15], Khadeeja Ismail[10], Stella Iurato[16], Anni Joensuu[10,11], Pauliina Karell[10], Silva Kasela[17,18], Jari Lahti[19,20], Allan F McRae[21], Pooja R Mandaviya[22,23], Ilkka Seppälä[24,25], Yunzhang Wang[26], Laura Baglietto[27], Elisabeth B Binder[16,28], Sarah E Harris[4,5], Allison M Hodge[29,30], Steve Horvath[31], Mikko Hurme[32,33,34], Magnus Johannesson[35], Antti Latvala[36], Karen A Mather[37], Sarah E Medland[38], Andres Metspalu[17,18], Lili Milani[17], Roger L Milne[29,30], Alison Pattie[39], Nancy L Pedersen[26], Annette Peters[12], Silvia Polidoro[14], Katri Räikkönen[19], Gianluca Severi[14,29,40], John M Starr[4,41], Lisette Stolk[22,42], Melanie Waldenberger[12], BIOS Consortium, Johan G Eriksson[43,44,45], Tõnu Esko[17,46], Lude Franke[13], Christian Gieger[12], Graham G Giles[29,30], Sara Hägg[26], Pekka Jousilahti[11], Jaakko Kaprio[10,36], Mika Kähönen[47,48], Terho Lehtimäki[24,25], Nicholas G Martin[49], Joyce B C van Meurs[23,42], Miina Ollikainen[10,36], Markus Perola[10,11], Danielle Posthuma[1], Olli T Raitakari[50,51], Perminder S Sachdev[37,52], Erdogan Taskesen[1,53], André G Uitterlinden[7,23,42], Paolo Vineis[14,54], Cisca Wijmenga[13], Margaret J Wright[55], Caroline Relton[8], George Davey Smith[8], Ian J Deary[§,4,39], Philipp D Koellinger[§,1,2], Daniel J Benjamin[§,56]

All affiliations appear at the end of the paper.

ACKNOWLEDGEMENTS

Figures

Figure 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 ($P = 1 \times 10^{-7}$) and for suggestive significance ($P = 1 \times 10^{-5}$). Each epigenome-wide associated probe is marked with a red ×, and the symbol of the closest gene based on physical position.
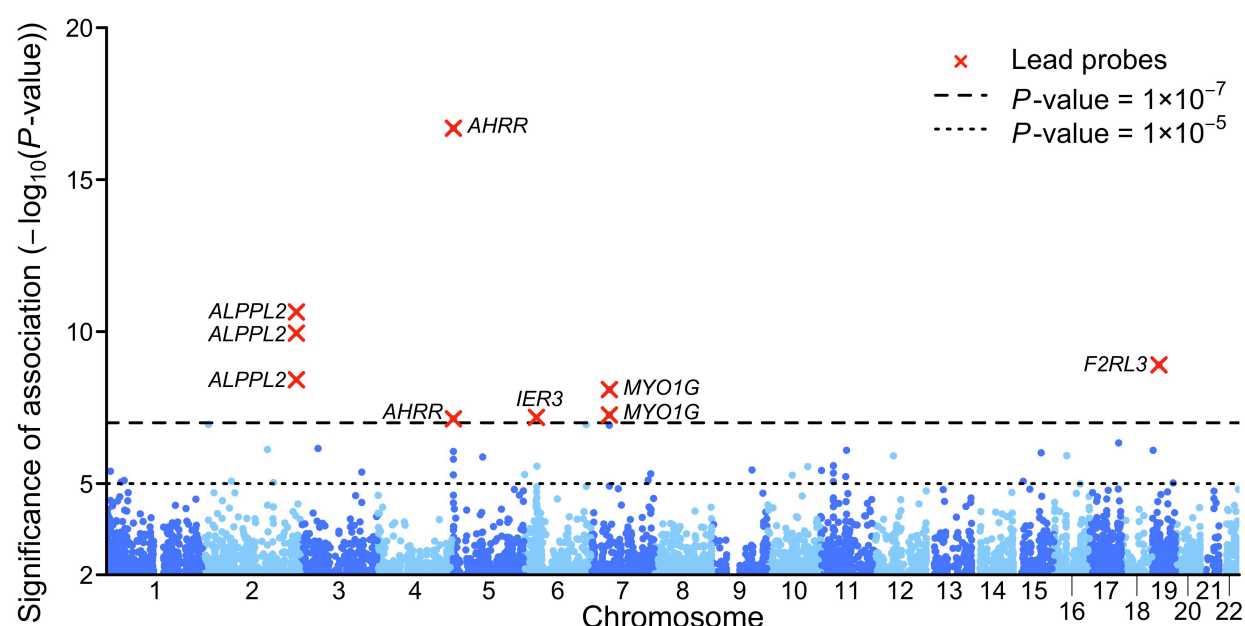
Figure 2 - Distribution of coefficient of determination (R^2's) across traits and methods.
The figure displays the effect size estimates in terms of R^2, in descending order, for the 50 top probes of the adjusted EWAS model. For comparison we present the 50 top probes from recent EWAS on alcohol consumption ($n = 9,643$, 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.
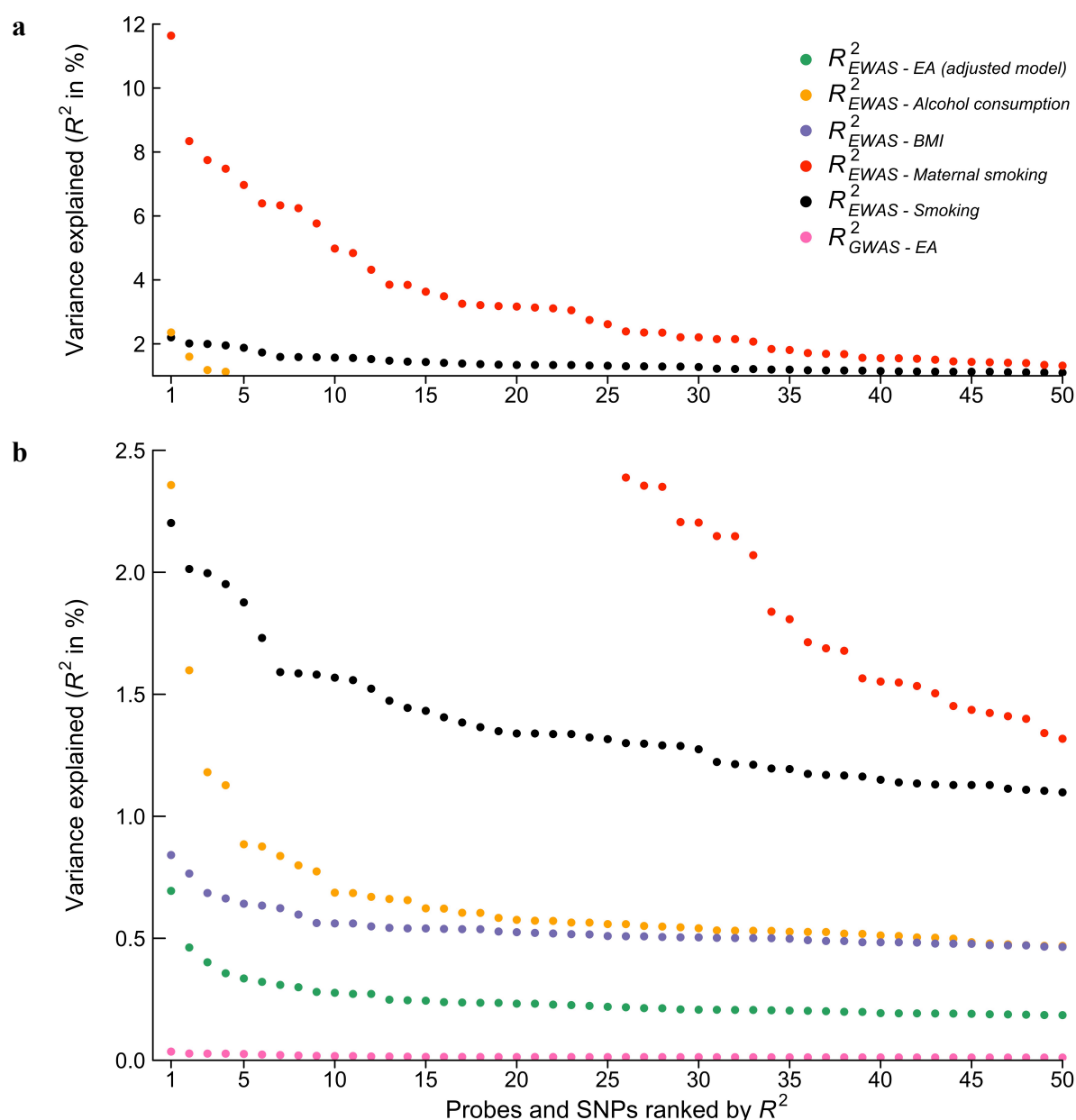
Figure 3 - Distribution of R^2 to compare the effect sizes across never smokers and smoking EWAS. Panel **(a)** displays the effect-size estimates in terms of R^2 for the 9 lead probes, in descending order, and the lead probe's corresponding effect size when re-estimated in the subsample of never smokers. 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.
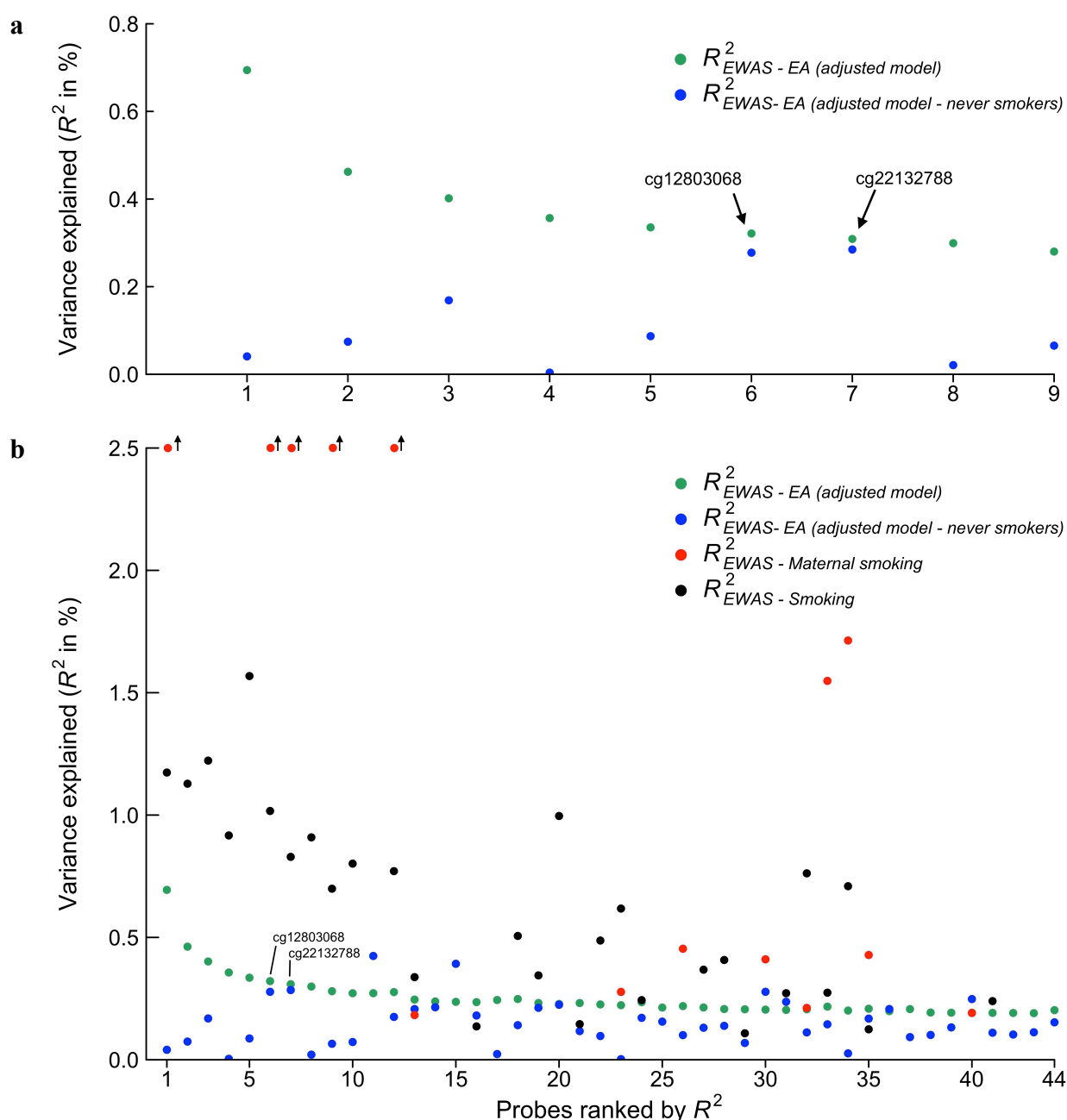
Figure 4 - Methylation score prediction of educational attainment in independent holdout samples. Panel **(a)** displays the prediction in all individuals, and panel **(b)** displays the prediction restricted to never smokers. The incremental R^2 of the methylation score was constructed with the coefficient estimates from meta-analysis of the basic model, as well as the adjusted model. The sample sizes of the LBC1936, the RS3, and the RS-BIOS cohorts are 918, 728, and 671 respectively. We performed sample size weighted meta-analysis across the cohorts for each of the 4 combinations of model estimates and P-value thresholds. The respective P-values are 4.42×10^{-11}, 7.76×10^{-11}, 2.02×10^{-11}, and 3.28×10^{-8} respectively for the full sample and 0.0183, 0.0898, 0.0051, and 0.1818 respectively for the never-smokers. The full prediction results are presented in **Supplementary Table S1.9a and S1.9b**.
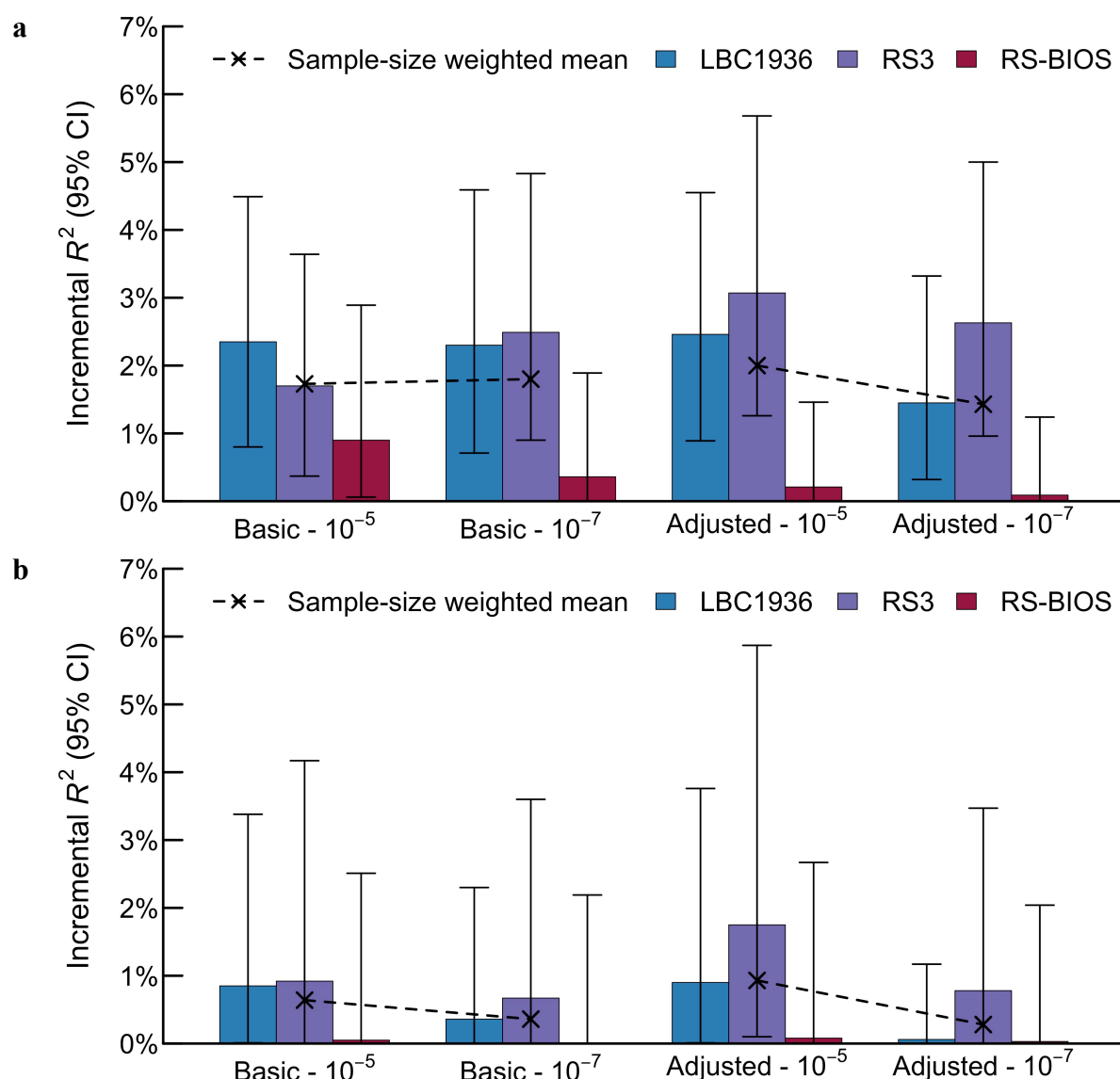
Tables

Table 1. EWAS association results - adjusted model

EWAS association results - adjusted model (9 probes with P-value $< 1\times10^{-7}$)										
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**.

[1] Department of Complex Trait Genetics, Vrije Universiteit, Center for Neurogenomics and Cognitive Research, De Boelelaan 1085, Amsterdam, 1081HV, the Netherlands

[2] Institute for Behavior and Biology, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, the Netherlands

[3] Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia

[4] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, 7 George Square, Edinburgh, EH8 9JZ, United Kingdom

[5] Medical Genetics Section, Centre for Genomic and Experimental Medicine, University of Edinburgh, Edinburgh, EH4 2XU, United Kingdom

[6] Department of Applied Economics, Erasmus School of Economics, Erasmus University Rotterdam, Burgemeester Oudlaan 50, Rotterdam, 3062 PA, the Netherlands

[7] Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands

[8] MRC Intergrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Barley House, Oakfield Grove, Bristol, BS28BN, United Kingdom

[9] Mathematics and Statistics, Murdoch University, 90 South St., Murdoch, 6150, WA, Australia

[10] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Tukholmankatu 8, 2B, Helsinki, FI-00014, Finland

[11] National Institute for Health and Welfare, Genomics and Biomarkers, PO Box 30, Helsinki, FI-00271, Finland

[12] Research Unit of Molecular Epidemiology (AME), Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstr. 1, Munich, Germany, Neuherberg, 85764, Germany

[13] Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

[14] Molecular and genetic epidemiology unit, Human Genetics Foundation Torino (HuGeF), Via Nizza 52, Turin, 10126, Italy

[15] Department of Medical Sciences, University of Torino, Corso Dogliotti 14, Turin, 10126, Italy

[16] Department Translational Research in Psychiatry, Max-Planck-Institute of Psychiatry, Munich, Germany, Kraepelinstr. 2-10, Munich, 80804, Germany

[17] Estonian Genome Center, University of Tartu, Riia 23B, Tartu, 51010, Estonia

[18] Institute of Molecular and Cell Biology, University of Tartu, Riia 23, Tartu, 51010, Estonia

[19] Institute of Behavioural Studies, Siltavuorenpenger 1A, University of Helsinki, Helsinki, FI-00014, Finland

[20] Collegium for Advanced Studies, University of Helsinki, Helsinki, FI-00014, Finland

[21] Institute for Molecular Bioscience, The University of Queensland, St. Lucia, QLD

[22] Department of Clinical Chemistry, Erasmus University Medical Center, Wytemaweg 80, Rotterdam, 3015 CN, The Netherlands

[23] Department of Internal Medicine, Erasmus University Medical Center, Wytemaweg 80, Rotterdam, 3015 CN, The Netherlands

[24] Department of Clinical Chemistry, Fimlab Laboratories, Tampere 33520, Finland

[25] Department of Clinical Chemistry, Faculty of Medicine and Life Sciences, University of Tampere, Tampere 33014, Finland

[26] Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12A, Stockholm, 17177, Sweden

[27] Centre for Research in Epidemiology and Population Health, Inserm (Institut National de la Santé et de la Recherche Médicale), 114 rue Edouard Vaillant, Villejuif, 94805, France

[28] Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, USA

[29] Cancer Epidemiology Centre, Cancer Council Victoria, 615 St Kilda Road, Melbourne, 3004, Victoria, Australia

[30] Centre for Epidemiology and Biostatistics, School of Population and Global Health, The University of Melbourne, Level 3, 207 Bouverie Street, Carlton, Melbourne, 3010, Victoria, Australia

[31] Human Genetics and Biostatistics, University of California Los Angeles, 695 Charles E. Young Drive South, Box 708822, Los Angeles, CA 90095-7088, USA

[32] Department of Microbiology and Immunology, Faculty of Medicine and Life Sciences, University of Tampere, Tampere 33014, Finland

[33] Gerontology Research Center, University of Tampere, Tampere 33014, Finland

[34] Fimlab Laboratories, Tampere 33520, Finland

[35] Department of Economics, Stockholm School of Economics, Box 6501, Stockholm, 11383, Sweden

[36] Department of Public Health, University of Helsinki, Tukholmankatu 8, 2B, Helsinki, FI-00014, Finland

[37] Centre for Healthy Brain Ageing, Psychiatry, UNSW Australia, High St., Sydney, NSW 2052, Australia

[38] Psychiatric Genetics, QIMR Berghofer Medical Research Institute, 300 Herston Rd., Herston, QLD 4006, Australia

[39] Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh, EH8 9JZ, United Kingdom

[40] Centre for Research in Epidemiology and Population Health (CESP), Inserm (Institut National de la Santé et de la Recherche Médicale), 28 Rue Laennec, Lyon, 69373, France

[41] Alzheimer Scotland Dementia Research Centre, University of Edinburgh, 7 George Square, Edinburgh, EH8 9JZ, United Kingdom

[42] Netherlands Consortium for Healthy Ageing, Erasmus University Medical Center, Wytemaweg 80, Rotterdam, 3015 CN, The Netherlands

[43] Department of General Practice and Primary Health Care, University of Helsinki and Helsinki University Hospital, Tukholmankatu 8 B, Helsinki, FI-00014, Finland

[44] National Institute for Health and Welfare, P.O. Box 30, Helsinki, FI-00271, Finland

[45] Folkhälsan Research Center, Helsinki, Finland

[46] Program in Medical and Population Genetics, Broad Institute, 415 Main St., Cambridge, MA 02142, USA

[47] Department of Clinical Physiology, Tampere University Hospital, Tampere 33521, Finland

[48] Department of Clinical Physiology, Faculty of Medicine and Life Sciences, University of Tampere, Tampere 33014, Finland

[49] Genetic Epidemiology, QIMR Berghofer Medical Research Institute, 300 Herston Rd, Herston, QLD 4006, Australia

[50] Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku 20521, Finland

[51] Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku 20014, Finland

[52] Neuropsychiatric Institute, Prince of Wales Hospital, Barker St. Randwick, Randiwick, NSW 2031, Australia

[53] VU University Medical Center (VUMC), Alzheimer Center, Department of Neurology, Amsterdam, the Netherlands

[54] MRC/PHE Centre for Environment and Health, School of Public Health, Imperial College London, St Mary's Campus, Norfolk Place, London, W2 1PG, United Kingdom

[55] Queensland Brain Institute and Centre for Advanced Imaging, The University of Queensland, Brisbane, QLD, Australia

[56] Center for Economic and Social Research, University of Southern California, Los Angeles, CA 90089-3332, USA

[*] These authors contributed equally to this work.

[§] These authors jointly supervised the work.