Supplementary material for "Longitudinal data reveal strong genetic and weak non-genetic components of ethnicity-dependent blood DNA methylation levels"

Chris McKennan[1*], Katherine Naughton[2], Catherine Stanhope[2], Meyer Kattan[3], George T. O'Connor[4], Megan T. Sandel[4], Cynthia M. Visness[5], Robert A. Wood[6], Leonard B. Bacharier[7], Avraham Beigelman[7], Stephanie Lovinsky-Desir[3], Alkis Togias[8], James E. Gern[9], Dan Nicolae[1,2¶] Carole Ober[2¶]

June 29, 2020

This supplement contains a description of additional analyses, additional statistical methods and all supplemental figures referenced in the main text of the manuscript.

Detailed sample description and additional analyses

A complete description of the available covariate information

Maternal questionnaires, including those on smoking, stress, and depression, were administered prenatally and annually after the child's birth. Postnatal child health questionnaires were administered to a parent every three months through age 7 years. Annual visits of child and parent to the study center, starting at one year of age, included questionnaires, anthropomorphic measurements, and phlebotomy. Questionnaires included the Perceived Stress Scale [1], the Edinburgh Perinatal/Postnatal Depression Scale [2], and additional questionnaires to assess stress related to neighborhood factors, violence, and economic hardship [3]. Parent-reported colds were ascertained by telephone questionnaire every three months throughout the first three years of life. Gestational age at birth, maternal infections during pregnancy, and obstetric history were obtained from medical records. Bedroom allergens were measured in vacuumed dust from the child's bedroom, and cotinine levels were measured in cord blood plasma. Pet ownership, number of smokers in the household, daycare attendance, number of siblings, and maternal asthma were ascertained by interview with the mother. Allergic sensitization was determined by prick skin testing for 14 common aeroallergens at 3, 5, or 7 years of age. Aeroallergen sensitization was defined as a wheal $\geq$ 3mm larger than the saline control on prick skin testing or specific IgE $\geq$ 0.35kU/L.

Differences between the inferred genetic ancestry and reported race analyses are robust to differences in power

We showed in "Inferred genetic ancestry is more correlated with DNA methylation than is self-reported race" that we identified 8,597 conserved inferred genetic ancestry-associated CpGs (IGA-CpGs), but only 2,162 conserved self-reported race-associated CpGs (RR-CpGs). To show

that these results were robust to any differences between the powers of these analyses, we used Model (1) (see Methods) to determine

$$\delta_{RR} = \mathbb{P}\left\{\beta^{(0)}_{g,RR} \neq 0 \text{ or } \beta^{(7)}_{g,RR} \neq 0\right\}$$
$$\delta_{IGA} = \mathbb{P}\left\{\beta^{(0)}_{g,IGA} \neq 0 \text{ or } \beta^{(7)}_{g,IGA} \neq 0\right\},$$

which are the fraction of CpGs with non-zero self-reported race (RR) or inferred genetic ancestry (IGA) effects at birth or age 7. Since these quantities do not depend on the observed data, their values are invariant of our power to identify RR-CpGs and IGA-CpGs. We estimated that δ_{IGA} was 11% larger than δ_{RR}, indicating that IGA-CpGs outnumber RR-CpGs, which is consistent with our results in the main text.

Additional analyses to identify exposure-associated CpG sites

We performed additional analyses on every measured direct or indirect measures of environmental exposures to attempt to identify exposures that correlated with DNA methylation (DNAm) levels at either birth or age 7. These exposures were maternal asthma, maternal infections during pregnancy, pet ownership, bedroom allergens, number of smokers in the household, number of siblings, number of previous live births, daycare attendance, number of colds at age 2 or 3, and allergic sensitization or asthma in the child. Unlike maternal cotinine levels measured during pregnancy, there were no CpGs associated with DNAm levels at birth or age 7 at a 5% FDR.

We also tested the association between DNAm at birth and scores derived from the Perceived Stress Scale (PSS) and the Edinburgh Perinatal/Postnatal Depression Scale (EPDS), which are maternal stress-related phenotypes that were measured during the second or third trimester of pregnancy. There is ample evidence that suggests children exposed to adverse maternal mental well-being have compromised hypothalamus–pituitary–adrenal (HPA) axis signaling, possibly due to heightened exposure to maternal cortisol levels *in utero* [4]. Many observational studies have therefore tried to understand the relationship between prenatal stress and cord blood DNAm

on the promoters of the HPA-related genes NR3C1, HSD11B2, FKBP5, CRH, CRHBP, SLC6A4, and OXTR [5]. Despite there being evidence from studies with smaller sample sizes than ours that the promoter region DNAm in the cord blood of these genes (especially NR3C1) is related to the aforementioned mental-health variables [6–8], we found little to no evidence supporting these associations in our study (Figures S6 and S7). These null results are recapitulated in Mansell et al. [4], which, as far as we are aware (and stated by the authors of the Mansell et al. study), is the largest study which targets the cord blood DNAm on the promoters of HPA-related genes. Interestingly, they show that their ostensibly contradictory results can be attributed to erroneous statistical techniques used in previous studies [4].

Unfortunately, the existing epigenetic literature devoted to studying the effect of exposure histories on the blood DNAm in infants, children and adults is replete with studies whose conclusions are based on faulty statistics. Some notable examples include:

(i) Kippler et al. [9] (106 citations): The authors study the effect of maternal exposure to cadmium on the cord blood DNAm of 127 infants using the HumanMethylation450K BeadChip. While there were no significant associations after adjusting for multiple testing, the authors used the non-uniformity of their CpG P value histograms in males and females to claim cadmium affected global DNAm patterns, and had a sex-specific effect. However, since the authors failed to adjust for latent confounding factors (including cellular composition) that plague high throughput DNAm studies [10], their conclusions can easily be explained away by the almost sure systematic correlation between their $> 450,000$ P values [11, 12].

(ii) Koestler et al. [13] (64 citations): The authors used the HumanMethylation450K BeadChip to measure the cord blood DNAm of 138 newborns to understand the effect of maternal mercury and arsenic exposure on DNAm. The authors failed to account for multiple testing in all of their analyses:

- The authors reported that 9 out of the 348,569 CpGs in their study were significantly associated with maternal mercury at a P value threshold of 10^{-4} (Figure 1 in Koestler

et al. [13]). This is 26 fewer than expected by chance.

- The authors reported that 37 out of the 348,569 CpGs in their study were significantly associated with the multiplicative interaction between maternal mercury and arsenic at a P value threshold of 10^{-4} (Figure 2 in Koestler et al. [13]). This is just 2 more than expected by chance (binomial P value = 0.38).

(iii) Hernandez-Vargas et al. [14] (50 citations): The authors determined the concentration of aflatoxin B1 (AFB1) in the plasma of 115 pregnant mothers and subsequently measured the blood DNAm of their children at age 2-8months. They reported that the blood DNAm M-values at 71 CpGs were significantly linearly related to AFB1 concentration (5% FDR). However, we have reason to believe that most, if not all, of their reported associations are spurious:

- We downloaded their data from GEO and used their analysis pipeline to replicate their results. However, we found that a single outlier was driving most, if not all, of their associations (Figure S8).
- We removed the apparent outlier and re-ran their analysis pipeline. Only 1 CpG was associated with AFB1 concentration at a 5% FDR (Figure S9).
- Given that the long right hand tail of AFB1 concentrations is likely driving spurious associations (Figures S8 and S9), we log-transformed AFB1 concentration and repeated the original analysis. Unsurprisingly, there were no significant AFB1-associated CpGs at 25% FDR (the smallest q-value was 0.99). This is congruent with the author's observations that there were no AFB1-associated CpGs when AFB1 was treated as a categorical covariate (low AFB1 vs. high AFB1).

(iv) Geraghty et al. [15] (5 citations): The authors performed a randomized clinical trial from 60 neotates to study the global cord blood DNAm response to a low glycaemic index diet intervention in mothers. The major flaws include:

- Even though cord blood cellular composition differed significantly between cases (mothers on a low glycaemic index diet) and controls (no intervention) (see Table 4 in Geraghty et al.), they ignored cell type in all analyses comparing cases and controls. This almost certainly explains why the intervention correlated with the DNAm data matrix's second principal component (PC).
- As the authors note, none of their results replicated in direction or magnitude in independent data derived from the same cohort. This is congruent with the results of Amarasekera et al. [16], which also found no significant associations between maternal PUFA intervention and cord blood DNAm profiles.

(v) Peters et al. [17] (DMRcate; 203 citations): DMRcate is a method to infer differentially methylated regions (DMRs), which utilizes a kernel smoother to smooth marginal test statistics from neighboring CpGs. Although it is widely used, inference with DMRcate is fallacious:

- DMRcate's region-based P values rely on the critical assumption that DNAm from neighboring CpGs (CpGs < 1000bp apart) is independent (see the second equation in "Model for smoothed data" in Peters et al.). However, it is well known that neighboring CpGs (CpGs < 1000bp apart) are highly dependent [18, 19], which will lead to anti-conservative inference [20].
- Anti-conservative inference with DMRcate could be driving the results from Hibler et al. [21], which used blood DNAm measured on the EPIC 850K chip from 68 adults to infer > 400 genomic regions whose DNAm was associated with a nutritional and exercise intervention. For example, the correlation coefficients between DNAm from neighboring CpGs in one of the two regions Hibler et al. discussed in their results (chr4:185369135–185370076) were as high as 0.86 in our data.

(vi) Jiang et al. [22] (48 citations): The authors performed a double-blind crossover study in 16 young adults to understand the effect of diesel exhaust on the DNAm in PBMCs using the

HumanMethylation450K BeadChip. Some serious flaws include:

- They failed to incorporate cell composition into any of their regression models.
- Their reported 2,827 diesel exhaust-associated CpGs were not actually diesel exhaust-associated CpGs. Instead, they found that principal component (PC) 22 (out of 95) was correlated with one of their diesel exhaust phenotypes, and defined a diesel exhaust-associated CpG to be a CpG whose loading on PC 22 lied in the "tails" of the distribution of loadings ("tail" was defined as ±3 standard deviations around the mean loading). Besides "tail" being arbitrary, the reported 2,827 significant CpGs were actually PC 22-associated CpGs, and not diesel exhaust-associated CpGs.

These observations imply that besides maternal cotinine levels, there are likely few to no exposures considered in the existing literature that would be expected to be associated with blood DNAm at birth or age 7 in our data. This sentiment is echoed in Hannon et al. [23], which in addition to also showing that environmental exposures are responsible for a relatively small fraction of the variation in blood DNAm levels in young adults, states the following:

> "Social-science and health researchers in search of evidence for environmental effects on the genome should not assume that 'epigenetic' equates to 'environmental'. Importantly, DNA methylation at sites robustly associated with extrinsic factors such as smoking and BMI can also be under strong genetic control." [23]

Identifying sample collection site-associated CpG sites

We next attempted to identify CpGs whose DNAm levels at birth or age 7 were associated with sample collection site, which was used to argue that sample collection site might be confounding the relationship between ethnicity and DNAm levels in Galanter et al. [24]. We restricted the analysis to samples from self-reported black and Hispanic children, and regressed methylation at birth or age 7 onto sample collection site (a factor variable with four levels), while accounting for self-reported race (black or Hispanic), sex (male or female), gestational age, inferred genetic

ancestry, methylation plate number (a factor variable with five levels) and nine latent factors estimated with the method proposed in McKennan et al. [10]. We identified 865 CpGs whose DNAm levels at birth or age 7 were significantly associated with sample collection site at a 5% FDR. To test for differences in the effect due to sample collection site at birth and age 7, we regressed the difference in methylation at birth and age 7 onto sample collection site while accounting for the aforementioned covariates. There were no CpGs at which the effect due to sample collection site differed at birth and age 7 at a 5% FDR.

Additional statistical methods

The benefits of our Bayesian model for longitudinal data

The methodology described in "Joint modelling of DNA methylation at birth and age 7" of Methods is novel, and therefore warrants some discussion. As we do in Model (1), let $\beta_g^{(a)}$ be the expected difference in DNAm at CpG $g = 1, \ldots, p$ between self-reported Black and Hispanic children at age $a = 0, 7$. Recall that our goals were:

(i) Determine the strength of the association between DNAm and self-reported race and birth and age 7. Mathematically speaking, this amounts to estimating $\beta_g^{(0)}$ and $\beta_g^{(7)}$ for each CpG site g.

(ii) Identify CpGs whose self-reported race-associated DNAm is conserved from birth to age 7. This is equivalent to identifying the CpGs g at which $\beta_g^{(0)}$ and $\beta_g^{(7)}$ are similar and non-zero.

(iii) Identify CpGs whose self-reported race-associated DNAm changes from birth to age 7. This is equivalent to identifying the CpGs g at which $\beta_g^{(0)}$ and $\beta_g^{(7)}$ are different.

Standard frequentist inference that regresses DNAm onto self-reported race at birth and age 7 separately can easily be used to solve (i). However, standard statistical techniques are not appropriate for carrying out (ii) or (iii).

To see this, consider the usual hypothesis testing framework to solve (ii) and (iii), which attempts to assess the evidence for or against the hypotheses $H_{0,g} : \beta_g^{(0)} = \beta_g^{(7)}$ for each CpG $g = 1, \ldots, p$. Testing these hypotheses is straightforward, and can be done by regressing the difference in DNAm at birth and age 7 onto self-reported race. However, since accepting $H_{0,g}$ is forbidden in standard frequentist inference [25] and (ii) requires us to identify CpGs for which we accept $H_{0,g}$, it is impossible to solve (ii) using this framework. Additionally, it is not clear that $H_{0,g}$ is the hypothesis that best matches our inferential goals. Consider, for example, a situation in which differences in CBMC and PBMC cellular composition cause $\beta_g^{(7)} = \beta_g^{(0)} + \epsilon$ for some small, but non-zero, ϵ [26]. While studies with a large enough sample size will almost always reject $H_{0,g}$ in this case, the difference between $\beta_g^{(0)}$ and $\beta_g^{(7)}$ is of little biological consequence. This simple example shows that rejecting $H_{0,g}$ does not imply (iii) is true.

Our Bayesian estimator circumvents the abovementioned issues by allowing us to infer the relationship between the signs of $\beta_g^{(0)}$ and $\beta_g^{(7)}$. We therefore define "similar" in (ii) to mean $\beta_g^{(0)}$ and $\beta_g^{(7)}$ have the same non-zero sign, and "different" in (iii) if at least one of $\beta_g^{(0)}$ and $\beta_g^{(7)}$ is non-zero and the other is zero or in the opposite direction. The latter also provides a notion of a biologically relevant difference.

Model (1) in Methods also allows us to estimate the fraction of CpGs whose self-reported race effects are present at one age and not present at another ($\sum_{k=1}^{K} \pi_{(1,0)}^{(k)}$ and $\sum_{k=1}^{K} \pi_{(0,1)}^{(k)}$), the fraction that are independent of one another ($\sum_{k=1}^{K} \pi_{(1,1)}^{(k,s)}$ such that $\rho_s = 0$), the fraction that are moderately similar ($\sum_{k=1}^{K} \pi_{(1,1)}^{(k,s)}$ such that $\rho_s = 1/3$), the fraction that are very similar ($\sum_{k=1}^{K} \pi_{(1,1)}^{(k,s)}$ such that $\rho_s = 2/3$), and the fraction that are identical ($\sum_{k=1}^{K} \pi_{(1,1)}^{(k,s)}$ such that $\rho_s = 1$). Such quantities are impossible to estimate using standard methodology.

A lower bound on fraction of reported race-associated CpGs mediated through local genotype

Here we describe how we conservatively estimated the fraction of reported race-associated CpGs (RR-CpGs) with a SNP in a 10kB window that were mediated by a neighboring SNP. Fix

some network composed of a CpG with methylation M and the SNP whose genotype G was most correlated with M according to the meQTL discovery procedure described in Methods. Let $\{\mathrm{RR} \to \mathrm{M}\}$ be the event the CpG is an RR-CpG, $\{\mathrm{RR} \to \mathrm{G}\}$ the event RR affects genotype, and $\{\mathrm{G} \to \mathrm{M}\}$ the event G affects M independently of RR. We would like to estimate

$$\begin{aligned}\mathbb{P}(\mathrm{RR} \to \mathrm{G}, \mathrm{G} \to \mathrm{M} \mid \mathrm{RR} \to \mathrm{M}) &= \frac{\mathbb{P}(\mathrm{RR} \to \mathrm{G}, \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M})}{\mathbb{P}(\mathrm{RR} \to \mathrm{M})} \\ &= \frac{\mathbb{P}(\mathrm{RR} \to \mathrm{M} \mid \mathrm{RR} \to \mathrm{G}, \mathrm{G} \to \mathrm{M})\,\mathbb{P}(\mathrm{RR} \to \mathrm{G}, \mathrm{G} \to \mathrm{M})}{\mathbb{P}(\mathrm{RR} \to \mathrm{M})} \quad \text{(S1)}\end{aligned}$$

Define $H_0 = \{\mathrm{RR} \not\to \mathrm{G}\}$. For each SNP we computed a P value for the null hypothesis H_0 using the logistic regression model $\mathrm{G} \sim \mathrm{RR}$, where RR was either Black or Hispanic (we assumed a Hardy-Weinberg equilibrium model for the genotypes of all SNPs considered). Let t be the test statistic from the regression and $t^*_\alpha > 0$ be some threshold with significance level α. Then because G and RR are independent under H_0 (regardless of whether or not $\{\mathrm{G} \to \mathrm{M}\}$ or $\{\mathrm{RR} \to \mathrm{M}\}$ hold),

$$\begin{aligned}q = \mathbb{P}(H_0 \mid |t| \geq t^*_\alpha, \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M}) &= \frac{\mathbb{P}(|t| \geq t^*_\alpha \mid H_0, \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M})\,\mathbb{P}(H_0 \mid \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M})}{\mathbb{P}(|t| \geq t^*_\alpha \mid \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M})} \\ &= \frac{\mathbb{P}(|t| \geq t^*_\alpha \mid H_0)\,\mathbb{P}(H_0 \mid \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M})}{\mathbb{P}(|t| \geq t^*_\alpha \mid \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M})} \\ &\leq \frac{\alpha}{\mathbb{P}(|t| \geq t^*_\alpha \mid \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M})}\end{aligned}$$

where the equality in the second line comes from the fact that under the null hypothesis and given the rest of the graph, the behavior of G and RR are independent. We therefore upper-bounded q by estimating $\mathbb{P}(|t| \geq t^*_\alpha \mid \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M})$ using the RR-meQTL logistic regression test statistics (this is just the Benjamini-Hochberg procedure interpreted in a Bayesian framework). We finally established an estimated lower bound for (S1) by using the following:

$$\begin{aligned}
\mathbb{P}\left(\mathrm{RR} \to \mathrm{G}, \mathrm{G} \to \mathrm{M} \mid \mathrm{RR} \to \mathrm{M}\right) &= \frac{\#\text{ networks with RR} \to \mathrm{G}, \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M}}{\#\text{ networks with RR} \to \mathrm{M}} \\
&\geq \frac{\#\text{ networks with RR} \to \mathrm{G}, \mathrm{G} \to \mathrm{M}, \mathrm{RR} \to \mathrm{M} \text{ and } q \leq 0.2}{\#\text{ of networks with RR} \to \mathrm{M}} \\
&\gtrsim (1 - 0.2)\frac{\#\text{ of networks with } q \leq 0.2 \text{ among RR-meQTLs}}{\#\text{ RR-CpGs}} \\
&= 0.26.
\end{aligned}$$

Calculating the *P* value for the overlap between gestational age- and chronological age-associated CpGs with the same effect sign

Define $\boldsymbol{y}_g^{(a)}$ to be the DNAm for CpG $g = 1, \ldots, 784,484$ at age $a = 0, 7$. We estimated the effect of gestational age $\boldsymbol{X} \in \mathbb{R}^n$ on $\boldsymbol{y}_g^{(0)}$ in the model

$$\boldsymbol{y}_g^{(0)} = \boldsymbol{X}\beta_g^{\mathrm{GA}} + \boldsymbol{Z}_0\boldsymbol{\gamma}_g + \boldsymbol{C}_0\boldsymbol{\ell}_g + \boldsymbol{e}_g^{(0)}, \quad \boldsymbol{e}_g^{(0)} \sim N_n\left(0, \left(\sigma_g^2 + \delta_g^2\right) I_n\right) \quad (g = 1, \ldots, p)$$

and the effect of age, $\beta_g^{(0\to7)}$, in the model

$$\boldsymbol{y}_g^{(7)} - \boldsymbol{y}_g^{(0)} = \mathbf{1}_n\beta_g^{(0\to7)} + \boldsymbol{Z}_{\mathrm{diff}}\boldsymbol{\gamma}_g + \boldsymbol{C}_{\mathrm{diff}}\boldsymbol{\ell}_g + \boldsymbol{e}_g^{(\mathrm{diff})}, \quad \boldsymbol{e}_g \sim N_n\left(0, \sigma_g^2 I_n\right) \quad (g = 1, \ldots, p)$$

using ordinary least squares, where the nuisance covariates in $\boldsymbol{Z}_0, \boldsymbol{Z}_{\mathrm{diff}}$ are given in Methods and $\boldsymbol{C}_0, \boldsymbol{C}_{\mathrm{diff}}$ were estimated using McKennan et al. [10]. Define the estimated gestational age and age effects to be $\hat{\beta}_g^{\mathrm{GA}}$ and $\hat{\beta}_g^{(0\to7)}$, respectively. We use the output of these two regressions to get an approximate upper bound for the expected number of pairs $\left(\beta_g^{\mathrm{GA}}, \beta_g^{(0\to7)}\right)$ out of all 16,172 age-related CpG sites that had the same sign, under the null hypothesis that the effects due to gestational age and chronological age were generated independently (see the Results section).

Assume the variance model for the data at birth and age seven is given by (1c) and let $r_g =$

$\frac{\delta_g^2}{\delta_g^2+\sigma_g^2}$. Then using the estimates for $\boldsymbol{C}_0, \boldsymbol{C}_{\text{diff}}$, along with the observed nuisance covariates $\boldsymbol{Z}_0, \boldsymbol{Z}_{\text{diff}}$,

$$\hat{\text{Corr}}\left(\hat{\beta}_g^{\text{GA}}, \hat{\beta}_g^{(0\to 7)}\right) = 0.66\left(1-\hat{r}_g\right).$$

We then have that conditional on the true effects $\left(\beta_g^{\text{GA}}, \beta_g^{(0\to 7)}\right)^T$,

$$\left(\hat{\beta}_g^{\text{GA}}, \hat{\beta}_g^{(0\to 7)}\right)^T \approx N_2\left\{\left(\beta_g^{\text{GA}}, \beta_g^{(0\to 7)}\right)^T, \text{diag}\left(c_g, d_g\right)\begin{pmatrix} 1 & 0.66\left(1-\hat{r}_g\right) \\ 0.66\left(1-\hat{r}_g\right) & 1 \end{pmatrix}\text{diag}\left(c_g, d_g\right)\right\} \tag{S2}$$

for each $g \in [p]$, where c_g, d_g are positive constants. Let A_g be the event that CpG g is an age-CpG at a 5% FDR. We assume that $A_g = \left\{|z|_g \geq t\right\}$, where $z_g = \hat{\beta}_g^{0\to 7}/d_g$ is the z-score corresponding to $\hat{\beta}_g^{0\to 7}$ and t can be estimated as the smallest z-score with a q-value less than 0.05. The empirical distributions of $\left\{\hat{\beta}_g^{(0\to 7)}\right\}_{g\in\{5\%\text{ FDR age CpGs}\}}$ and $\left\{\hat{\beta}_g^{GA}\right\}_{g\in\{5\%\text{ FDR gestational age CpGs}\}}$ were approximately symmetric around 0, which we took to imply $\left\{\beta_g^{(0\to 7)}\right\}_{g\in[p]}$ and $\left\{\beta_g^{GA}\right\}_{g\in[p]}$ were symmetric around 0. For simplicity, we assume for density functions

$$h_{GA}(\cdot) = \sum_{r=1}^{R}\pi_r^{(GA)}N_1\left(\cdot; 0, \phi_r^{(GA)}\right) \qquad h_{(0\to 7)}(\cdot) = \sum_{j=1}^{J}\pi_j^{(0\to 7)}N_1\left(\cdot; 0, \phi_j^{(0\to 7)}\right),$$

$\beta_g^{\text{GA}} \overset{i.i.d}{\sim} h_{GA}(\cdot)$ and $\beta_g^{(0\to 7)} \overset{i.i.d}{\sim} h_{(0\to 7)}(\cdot)$. Such mixture normal densities can approximate a large class of parametric and non-parametric distributions [27]. Define $X_g, Y_g \in \mathbb{R}$ to be such that

$$\left(X_g, Y_g\right)^T \sim N_2\left\{0, \begin{pmatrix} 1 & 0.66\left(1-\hat{r}_g\right) \\ 0.66\left(1-\hat{r}_g\right) & 1 \end{pmatrix}\right\}.$$

Then under the null hypothesis that $\beta_g^{(0\to 7)}$ and β_g^{GA} are independent and assuming (S2) is correct,

$$\mathbb{P}\left\{\hat{\beta}_g^{\text{GA}}\hat{\beta}_g^{(0\to 7)} > 0 \mid A_g\right\} \leq \frac{\mathbb{P}\left(\left\{X_gY_g > 0\right\} \cap \left\{\left|Y_g\right| \geq t\right\}\right)}{\mathbb{P}\left(\left|Y_g\right| \geq t\right)}.$$

We can easily estimate the above upper bound. Therefore, conditional on knowing whether or not

each CpG is an age-associated CpG,

$$\mu = \mathbb{E}\left\{ \sum_{g \in \{5\% \text{ FDR age CpGs}\}} 1\left(\hat{\beta}_g^{\text{GA}} \hat{\beta}_g^{(0\to 7)} > 0\right) \right\} = \sum_{g \in \{5\% \text{ FDR age CpGs}\}} \mathbb{P}\left\{ \hat{\beta}_g^{\text{GA}} \hat{\beta}_g^{(0\to 7)} > 0 \mid A_g \right\}$$

$$\leq 14,236$$

under the null hypothesis. Since the maximum variance for a Bernoulli random variable is $1/4$, an approximate lower bound for the test-statistic is

$$\frac{0.97 \times 16,172 - 14,236}{\sqrt{16,172/4}} = 23.3,$$

which has a corresponding P value $\leq 10^{-119}$ under the normal approximation.

Determining the fraction of the variance in DNA methylation levels explained by maternal cotinine levels during pregnancy

Here we discuss our method for determining the fraction of the variance explained by maternal cotinine levels during pregnancy, which accounts for potential differences in the standard errors (i.e. sample sizes) of the maternal cotinine and genotype analyses.

The phenotype for maternal smoking was taken to be a factor variable with two levels, where the levels were smoker (cord blood plasma cotinine levels $\geq$ 10ng/mL) and non-smoker (cord blood plasma cotinine levels $<$ 10ng/mL). The 10ng/mL threshold was chosen because it was the same cutoff used to define sustained maternal smoking in Joubert et al. [28]. We remark that 98% of the non-smoking mothers had cotinine levels below 2ng/mL, the limit of detection of the assay. We report results for DNAm levels at birth, although the results at age 7 are identical.

Let $c_i \in \{0, 1\}$ and $s_{gi} \in \{0, 1, 2\}$ be maternal smoking status and the genotype for the SNP closest to CpG g for individual i, respectively. We assume that DNAm levels at birth at CpG g in

individual i (y_{gi}) could be modeled as

$$y_{gi} = \mu_g + \beta_g^{(s)} s_{gi} + \beta_g^{(c)} c_i + \boldsymbol{\gamma}_g^T \boldsymbol{z}_i + \epsilon_{gi},\ \epsilon_{gi} \sim N\left(0, \sigma_g^2\right), \quad g = 1, \ldots, p; i = 1, \ldots, n, \tag{S3}$$

where $\boldsymbol{z}_i$ contain the nuisance covariates inferred genetic ancestry, sex, gestational age and methylation plate number. Using this model, we defined the fraction of the variance in DNAm levels at CpG g explainable by maternal smoking and genotype to be

$$\pi_g^{(c)} = \frac{\left\{\beta_g^{(c)}\right\}^2 \mathrm{Var}\left(c_i\right)}{\mathrm{Var}\left(y_{gi}\right)}, \quad \pi_g^{(s)} = \frac{\left\{\beta_g^{(s)}\right\}^2 \mathrm{Var}\left(s_{gi}\right)}{\mathrm{Var}\left(y_{gi}\right)}, \tag{S4}$$

respectively. We set $\mathrm{Var}\left(c_i\right) = p^{(c)}\left\{1 - p^{(c)}\right\}$ and $\mathrm{Var}\left(s_{gi}\right) = 2p_g^{(s)}\left\{1 - p_g^{(s)}\right\}$, where $p^{(c)}$ is the fraction of maternal smokers in our dataset and $p_g^{(s)}$ is the minor allele frequency for the SNP adjacent to CpG g. We remark that $p^{(c)} = 0.17$ was 2.8 times the national smoking during pregnancy (SDP) rate for non-Hispanic black mothers, 1.7 times the national SDP rate for non-Hispanic white mothers and 10.0 times the national SDP rate for Hispanic mothers [29], indicating $\pi_c^{(g)}$, depending on the exact population of interest, is likely an overestimate for the fraction of variance explained by maternal smoking. Since our goal was to determine the relative proportion of the variance explained by maternal smoking and genotype, i.e. $\pi_g^{(c)} / \left\{\pi_g^{(c)} + \pi_g^{(s)}\right\}$ and $\pi_g^{(s)} / \left\{\pi_g^{(c)} + \pi_g^{(s)}\right\}$, we need only estimate $\left\{\beta_g^{(c)}\right\}^2$ and $\left\{\beta_g^{(s)}\right\}^2$.

Since there was little detectable correlation between maternal smoking and the genotypes of SNPs adjacent to maternal smoking CpGs identified in Joubert et al. [28] (ms-CpGs), we ignored genotype when estimating $\beta_g^{(c)}$ and used McKennan et al. [10] to determine $\hat{\beta}_g^{(c)}$, an estimate for $\beta_g^{(c)}$, and subsequently used the method proposed in Stephens et al. [30] with only ms-CpGs as input to determine $\mathbb{E}\left[\left\{\beta_g^{(c)}\right\}^2 \mid \hat{\beta}_g^{(c)}\right]$, our estimate for $\left\{\beta_g^{(c)}\right\}^2$. To estimate $\left\{\beta_g^{(s)}\right\}^2$, we first computed $\hat{\beta}_g^{(s)}$, the ordinary least squares estimate for $\beta_g^{(s)}$ in Model (S3), using a random subset of 56% of the self-reported black children, and subsequently used Stephens et al. [30] with only ms-CpGs as input to determine $\mathbb{E}\left[\left\{\beta_g^{(s)}\right\}^2 \mid \hat{\beta}_g^{(s)}\right]$, our estimate for $\left\{\beta_g^{(s)}\right\}^2$. We used self-reported black children to avoid

heterogeneous genetic effect sizes (due to population stratification), and only used 56% of those samples to determine $\hat{\beta}_g^{(s)}$ to ensure that the standard errors of $\hat{\beta}_g^{(c)}$ and $\hat{\beta}_g^{(s)}$ were approximately the same (Figure S5). This sub-sampling helped guarantee that the precision of the estimates $\mathbb{E}\left[\left\{\beta_g^{(c)}\right\}^2 \mid \hat{\beta}_g^{(c)}\right]$ and $\mathbb{E}\left[\left\{\beta_g^{(s)}\right\}^2 \mid \hat{\beta}_g^{(s)}\right]$ was the same, meaning that any difference in those estimates could not be attributed to the relatively small number of smokers in our study. We lastly plugged-in $\mathbb{E}\left[\left\{\beta_g^{(c)}\right\}^2 \mid \hat{\beta}_g^{(c)}\right]$ and $\mathbb{E}\left[\left\{\beta_g^{(s)}\right\}^2 \mid \hat{\beta}_g^{(s)}\right]$ for $\left\{\beta_g^{(c)}\right\}^2$ and $\left\{\beta_g^{(s)}\right\}^2$ into (S4) to estimate $\pi_g^{(c)} / \left\{\pi_g^{(c)} + \pi_g^{(s)}\right\}$ and $\pi_g^{(s)} / \left\{\pi_g^{(c)} + \pi_g^{(s)}\right\}$.

The relationship between inferred genetic ancestry and meQTLs

Since inferred genetic ancestry (IGA) is derived from genoytpe data, it could be the case that the SNPs used to estimate IGA would *a priori* be expected to be cis-meQTL, which might inflate the correlated between IGA and DNAm. However, we found that the genotype at only 3% of the 39,670 identified mSNPs (SNPs that are cis-meQTLs for at least 1 CpG at a 5% FDR) was significantly correlated with IGA at a 20% FDR, which was 17% fewer than what was expected by chance.

Supplemental figures

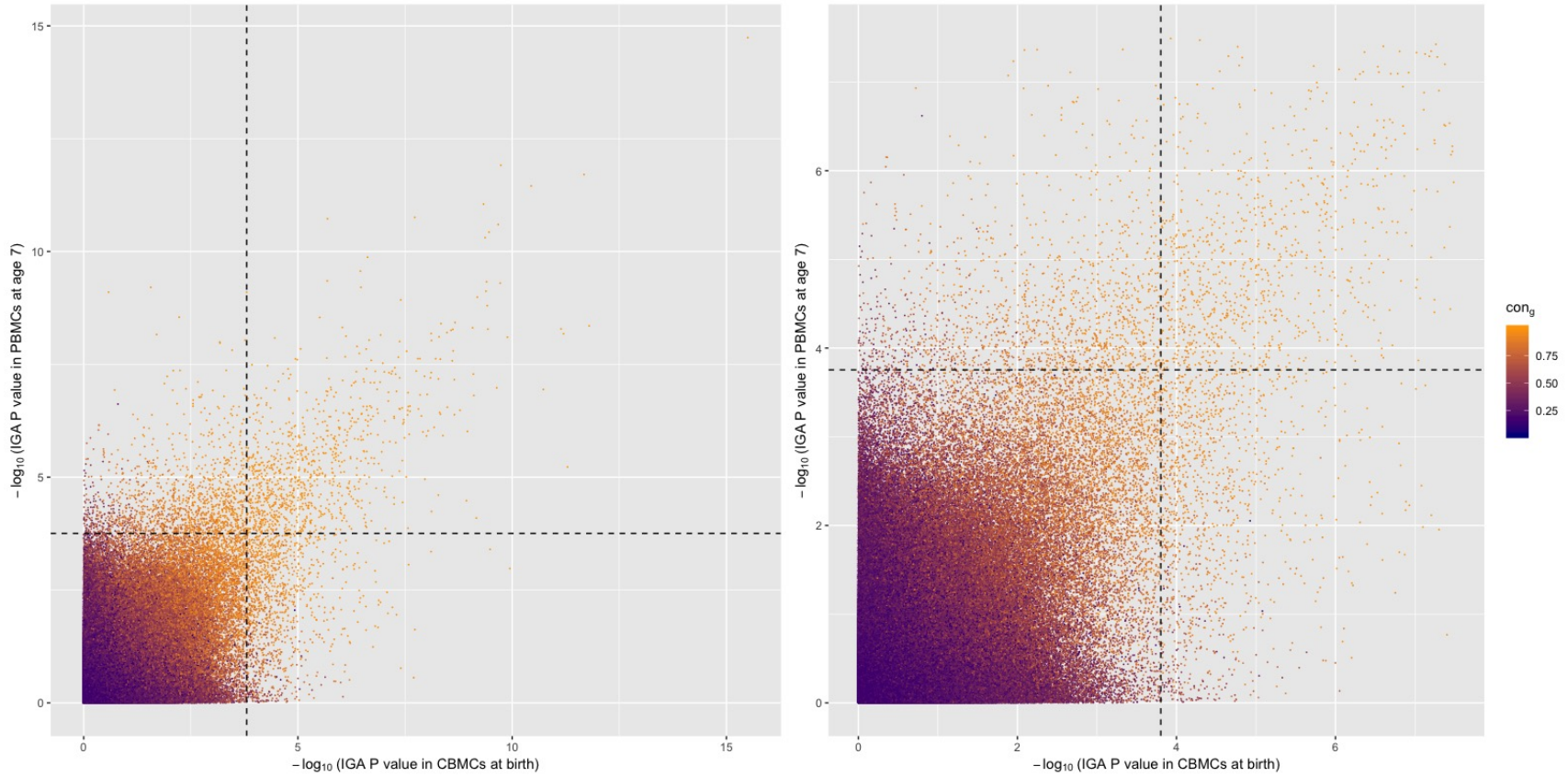


Figure S1: Relationship between inferred genetic ancestry (IGA) *P* values at birth and age 7 and estimated conserved sign rates (con_g). The *P* values were estimated using a standard regression model and the dashed lines indicate the 5% FDR threshold. The plot on the right is a zoomed-in version of the plot on the left.

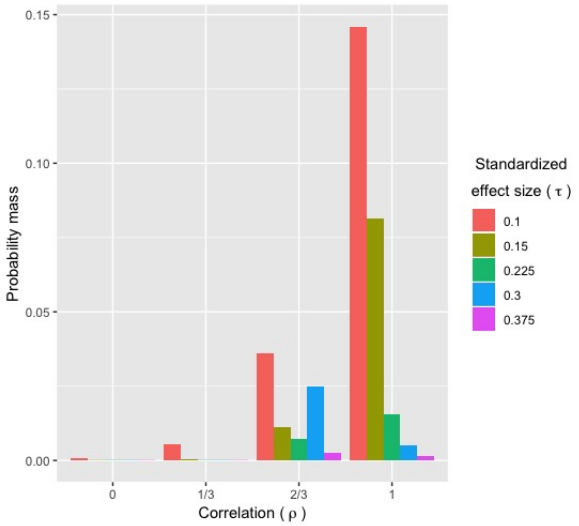


Figure S2: Probability mass of the components of $\hat{\boldsymbol{\pi}}_{(1,1)}$, where ρ is the correlation between the reported race effect at birth and age 7 and τ is proportional to the expected magnitude of the effect sizes (see Model (1)). This trend was echoed in the inferred genetic ancestry analysis.

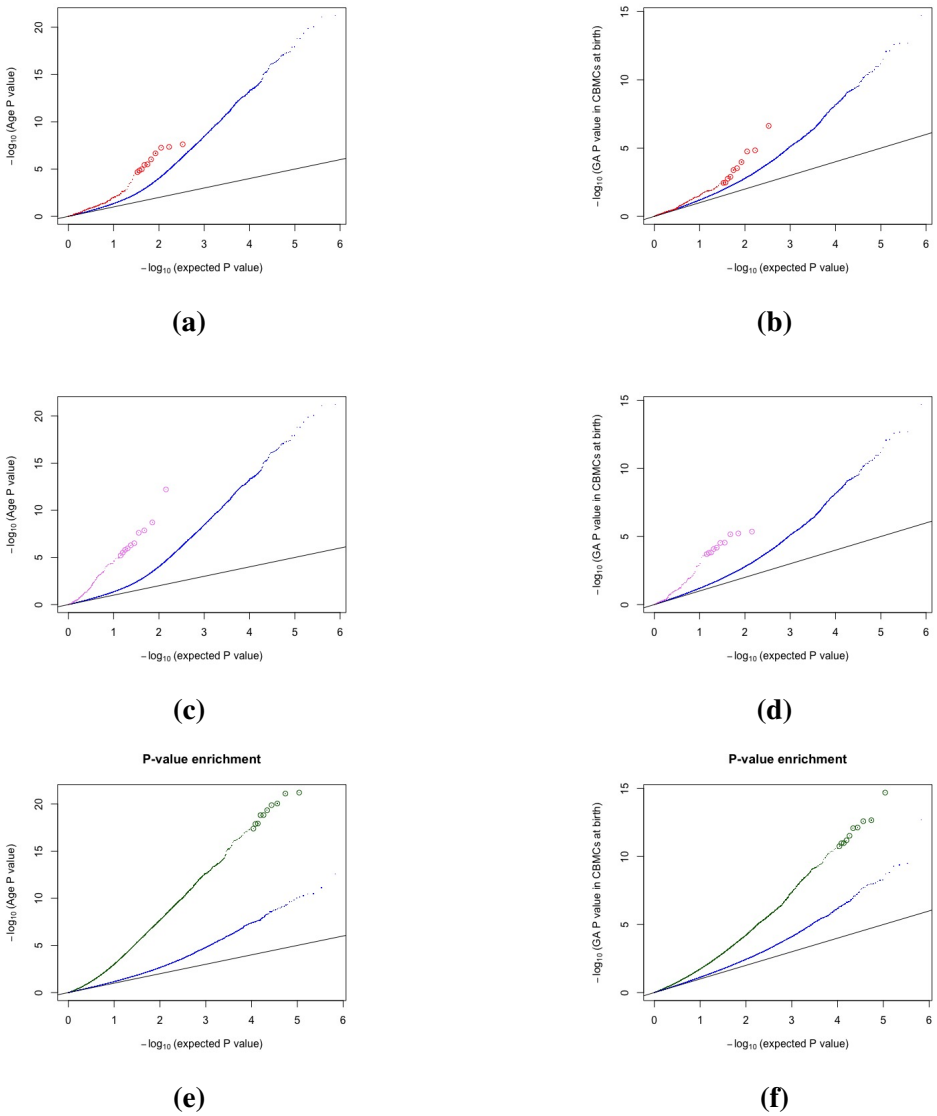
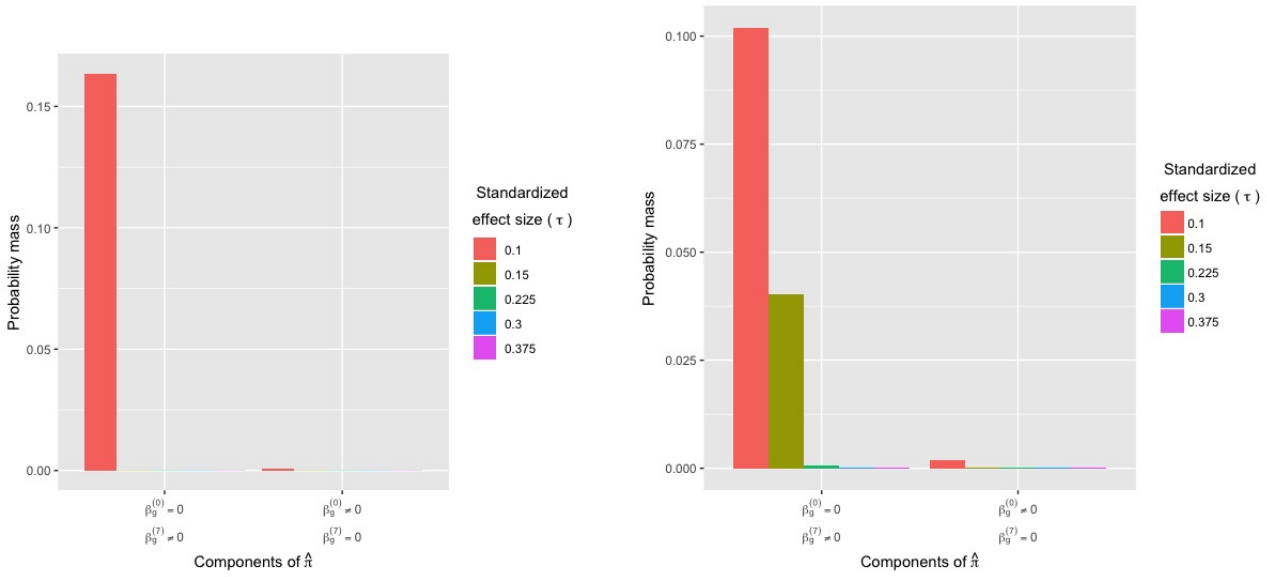


Figure S3: Distribution of *P* values for age (birth to age 7) (panels (a), (c) and (e)) and gestational age (GA) (panels (b), (d) and (f)). The red, violet and dark green dots in the upper, middle and lower panels are the 353 CpGs used to build the linear predictor of age in Horvath [31], the 148 CpGs used to build the linear predictor of gestational age in Knight et al. [32] and the 109,597 age (birth to age 5) CpGs discovered in Pérez et al. [33]. The blue dots are all of the other CpGs and the 10 enlarged circles are for visual aid.



(a) **(b)**

Figure S4: Probability mass of components of $\hat{\boldsymbol{\pi}}_{(1,0)}$ ($\beta_g^{(0)} \neq 0, \beta_g^{(7)} = 0$) and $\hat{\boldsymbol{\pi}}_{(0,1)}$ ($\beta_g^{(0)} = 0, \beta_g^{(7)} \neq 0$) in the inferred genetic ancestry (a) and reported race (b) analyses.

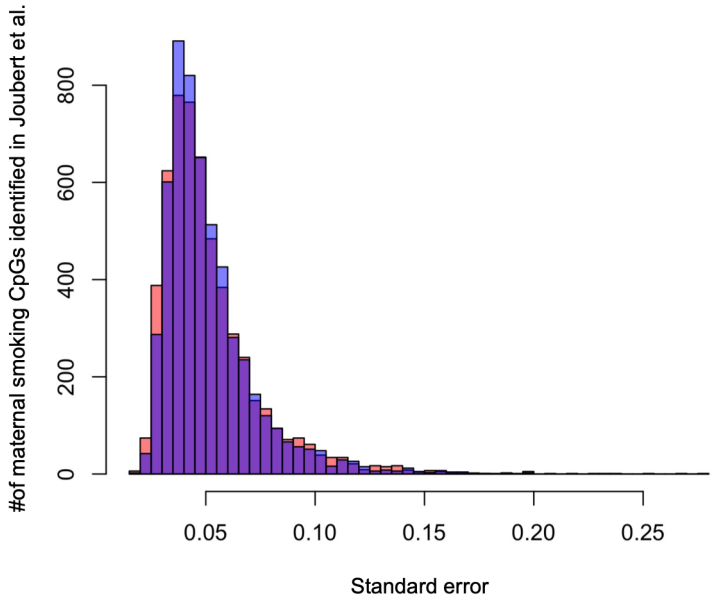


Figure S5: Histograms of standard errors for $\hat{\beta}_g^{(c)}$ (blue) and $\hat{\beta}_g^{(s)}$ (red) for all CpGs $g = 1, \ldots, 784,484$ that are also maternal smoking-associated CpGs identified in Joubert et al. [28].

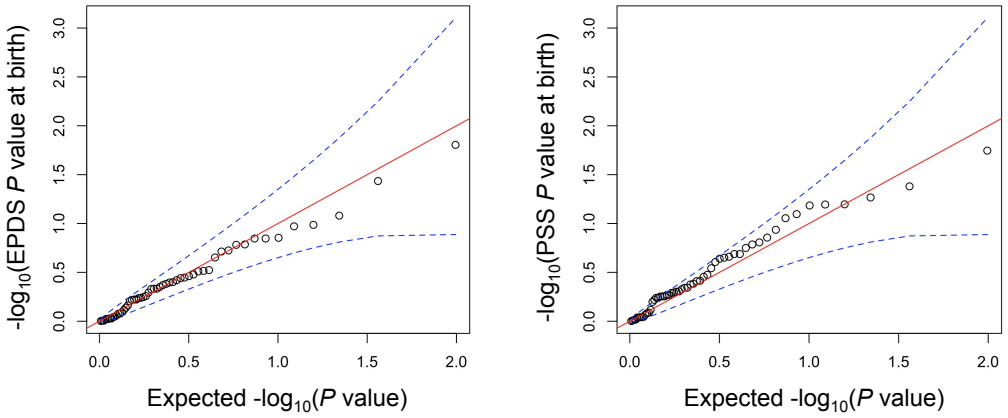


Figure S6: *P* values for CpGs on the promoter region of NR3C1 whose cord blood DNAm was measured in our study. Consistent with existing work, PSS was treated as a continuous variable and EPDS was treated as categorical variables in our regressions, where mothers with an EPDS score ≥ 10 (< 10) were classified as depressed (not-depressed) [4, 6]. The dashed blue error lines are drawn ± 2 standard deviations around the solid red 45° line.

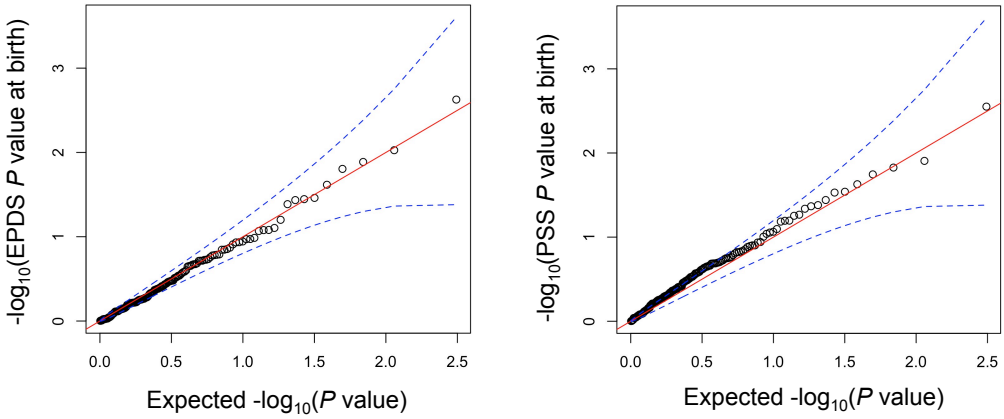
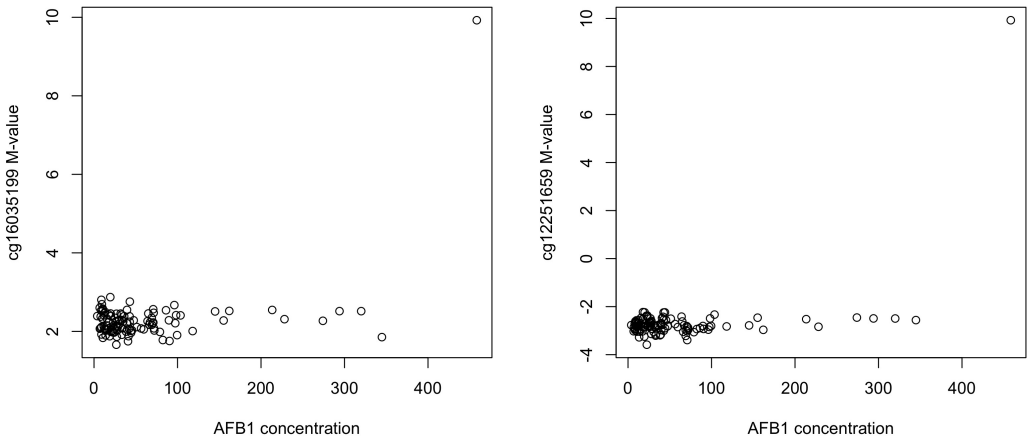


Figure S7: The same as Figure S6, except for all HPA-related genes considered in Sosnowski et al. [5] (NR3C1, HSD11B2, FKBP5, CRH, CRHBP, SLC6A4, and OXTR).

277

278



279

Figure S8: M-values of the two most significant aflatoxin B1- (AFB1-) associated CpGs identified in Hernandez-Vargas et al. [14] as a function of maternal plasma AFB1 concentration.

280

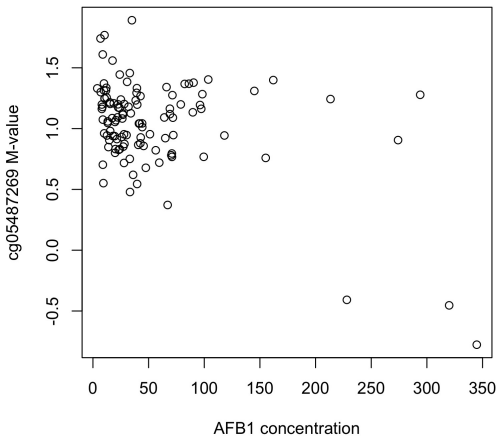


Figure S9: M-values of the only significant AFB1-associated CpG (5% FDR) after removing the outlier in Figure S8.

References

1. S. Cohen, T. Kamarck, and R. Mermelstein. "A Global Measure of Perceived Stress". In: *Journal of Health and Social Behavior* 24.4 (1983), pp. 385–396. ISSN: 00221465.
2. J. L. Cox, J. M. Holden, and R. Sagovsky. "Detection of Postnatal Depression: Development of the 10-item Edinburgh Postnatal Depression Scale". In: *The British Journal of Psychiatry* 150.6 (1987), pp. 782–786.
3. R. J. Wright et al. "Prenatal Maternal Stress and Cord Blood Innate and Adaptive Cytokine Responses in an Inner-City Cohort". In: *American Journal of Respiratory and Critical Care Medicine* 182.1 (July 2010), pp. 25–33.
4. T. Mansell, P. Vuillermin, A.-L. Ponsonby, F. Collier, R. Saffery, and J. Ryan. "Maternal mental well-being during pregnancy and glucocorticoid receptor gene promoter methylation in the neonate". In: *Development and Psychopathology* 28.4pt2 (2016), pp. 1421–1430.
5. D. W. Sosnowski, C. Booth, T. P. York, A. B. Amstadter, and W. Kliewer. "Maternal prenatal stress and infant DNA methylation: A systematic review". In: *Developmental Psychobiology* 60.2 (2018), pp. 127–139.
6. T. Hompes et al. "Investigating the influence of maternal cortisol and emotional state during pregnancy on the DNA methylation status of the glucocorticoid receptor gene (NR3C1) promoter region in cord blood". In: *Journal of Psychiatric Research* 47.7 (2013), pp. 880–891.
7. A. L. Non, A. M. Binder, L. D. Kubzansky, and K. B. Michels. "Genome-wide DNA methylation in neonates exposed to maternal depression, anxiety, or SSRI medication during pregnancy". In: *Epigenetics* 9.7 (July 2014), pp. 964–972.
8. T. F. Oberlander, J. Weinberg, M. Papsdorf, R. Grunau, S. Misri, and A. M. Devlin. "Prenatal exposure to maternal depression, neonatal methylation of human glucocorticoid receptor gene (NR3C1) and infant cortisol stress responses". In: *Epigenetics* 3.2 (Mar. 2008), pp. 97–106.
9. M. Kippler, K. Engström, S. JurkovićMlakar, M. Bottai, S. Ahmed, M. Hossain, R. Raqib, M. Vahter, and K. Broberg. "Sex-specific effects of early life cadmium exposure on DNA methylation and implications for birth weight". In: *Epigenetics* 8 (May 2013), pp. 494–503.
10. C. McKennan and D. Nicolae. "Accounting for unobserved covariates with varying degrees of estimability in high-dimensional biological data". In: *Biometrika* 106.4 (Sept. 2019), pp. 823–840. ISSN: 0006-3444.
11. B. Efron. "Correlated z-values and the accuracy of large-scale statistical estimates". In: *Journal of the American Statistical Association* 105.491 (Sept. 2010), pp. 1042–1055.
12. D. Gerard and M. Stephens. "Empirical Bayes shrinkage and false discovery rate estimation, allowing for unwanted variation". In: *Biostatistics* 21.1 (July 2018), pp. 15–32. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxy029.
13. D. C. Koestler, M. Avissar-Whiting, E. A. Houseman, M. R. Karagas, and C. J. Marsit. "Differential DNA Methylation in Umbilical Cord Blood of Infants Exposed to Low Levels of Arsenic in Utero". In: *Environmental Health Perspectives* 121.8 (Aug. 2013), pp. 971–977.

14. H. Hernandez-Vargas et al. “Exposure to aflatoxin B1 in utero is associated with DNA methylation in white blood cells of infants in The Gambia”. In: *International journal of epidemiology* 44.4 (Aug. 2015), pp. 1238–1248.

15. A. A. Geraghty, A. Sexton-Oates, E. C. O’Brien, G. Alberdi, P. Fransquet, R. Saffery, and F. M. McAuliffe. “A Low Glycaemic Index Diet in Pregnancy Induces DNA Methylation Variation in Blood of Newborns: Results from the ROLO Randomised Controlled Trial”. In: *Nutrients* 10.4 (Apr. 2018), p. 455.

16. M. Amarasekera, P. Noakes, D. Strickland, R. Saffery, D. J. Martino, and S. L. Prescott. “Epigenome-wide analysis of neonatal CD4(+) T-cell DNA methylation sites potentially affected by maternal fish oil supplementation”. In: *Epigenetics* 9.12 (Dec. 2014), pp. 1570–1576.

17. T. J. Peters, M. J. Buckley, A. L. Statham, R. Pidsley, K. Samaras, R. V Lord, S. J. Clark, and P. L. Molloy. “De novo identification of differentially methylated regions in the human genome”. In: *Epigenetics & Chromatin* 8.1 (2015), p. 6.

18. F. Eckhardt et al. “DNA methylation profiling of human chromosomes 6, 20 and 22”. In: *Nature genetics* 38.12 (Dec. 2006), pp. 1378–1385.

19. A. E. Jaffe, P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg, and R. A. Irizarry. “Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies”. In: *International journal of epidemiology* 41.1 (Feb. 2012), pp. 200–209.

20. T. R. Makin and J.-J. Orban de Xivry. “Ten common statistical mistakes to watch out for when writing or reviewing a manuscript”. In: *eLife* 8 (Oct. 2019), e48175.

21. E. Hibler, L. Huang, J. Andrade, and B. Spring. “Impact of a diet and activity health promotion intervention on regional patterns of DNA methylation”. In: *Clinical Epigenetics* 11.1 (2019), p. 133.

22. R. Jiang, M. J. Jones, F. Sava, M. S. Kobor, and C. Carlsten. “Short-term diesel exhaust inhalation in a controlled human crossover study is associated with changes in DNA methylation of circulating mononuclear cells in asthmatics”. In: *Particle and Fibre Toxicology* 11.1 (2014), p. 71.

23. E. Hannon et al. “Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins”. In: *PLOS Genetics* 14.8 (Aug. 2018), e1007544–.

24. J. M. Galanter et al. “Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures”. In: *eLife* 6 (Jan. 2017), e20532. ISSN: 2050-084X.

25. J. Leppink, P. O’Sullivan, and K. Winston. “Evidence against vs. in favour of a null hypothesis”. In: *Perspectives on Medical Education* 6.2 (2017), pp. 115–118.

26. J. Fu et al. “Unraveling the Regulatory Mechanisms Underlying Tissue-Dependent Genetic Variation of Gene Expression”. In: *PLOS Genetics* 8.1 (Jan. 2012), e1002431–.

27. A. Norets. “Approximation of conditional densities by smooth mixtures of regressions”. In: *The Annals of Statistics* 38.3 (2010), pp. 1733–1766.

28. B. R. Joubert et al. “DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis”. In: *The American Journal of Human Genetics* 98.4 (2016), pp. 680–696.

29. P. Drake, A. K. Driscoll, and T. Mathews. “Cigarette smoking during pregnancy: United States, 2016”. In: *National Center for Health Statistics Data Brief* 305 (Feb. 2018).

30. M. Stephens, P. Carbonetto, D. Gerard, M. Lu, L. Sun, J. Willwerscheid, and N. Xiao. *ashr: Methods for Adaptive Shrinkage, using Empirical Bayes*. R package version 2.2-39. 2019. URL: `https://CRAN.R-project.org/package=ashr`.

31. S. Horvath. “DNA methylation age of human tissues and cell types”. In: *Genome Biology* 14.10 (2013), p. 3156.

32. A. K. Knight et al. “An epigenetic clock for gestational age at birth based on blood methylation data”. In: *Genome Biology* 17.1 (2016), p. 206.

33. R. F. Pérez, P. Santamarina, J. R. Tejedor, R. G. Urdinguio, J. Álvarez-Pitti, P. Redon, A. F. Fernández, M. F. Fraga, and E. Lurbe. “Longitudinal genome-wide DNA methylation analysis uncovers persistent early-life DNA methylation changes”. In: *Journal of translational medicine* 17.1 (Jan. 2019), pp. 15, 15–15.