Longitudinal studies at birth and age 7 reveal strong effects of genetic variation on ancestry-associated DNA methylation patterns in blood cells from ethnically admixed children

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 Lovisky-Desir[3], Alkis Togias[8], James E. Gern[9], Dan Nicolae[1,2¶], Carole Ober[2¶] for the NIAID-sponsored Inner-City Asthma Consortium

[1]Department of Statistics, University of Chicago, Chicago, IL
[2]Department of Human Genetics, University of Chicago, Chicago, IL
[3]Department of Pediatrics, Columbia University Medical Center, New York, NY
[4]Department of Medicine, Boston University School of Medicine, Boston, MA
[5]Rho Federal Systems Division, Chapel Hill, NC
[6]Department of Pediatrics, Johns Hopkins University Medical Center, Baltimore, MD
[7]Department of Pediatrics, Washington University School of Medicine and St Louis Children's Hospital, St. Louis, MO
[8]National Institute of Allergy and Infectious Disease, Bethesda, MD
[9]Departments of Pediatrics and Medicine, University of Wisconsin School of Medicine and Public Health, Madison WI

¶ Equal contributions

*Corresponding Author:
Email: cgm29@uchicago.edu

Abstract

The epigenetic architecture in humans is influenced by genetic factors, exposure histories and biological factors such as age, but little is known about their relative contribution or their longitudinal dynamics. Here, we studied DNA methylation levels at over 750,000 CpG sites in mononuclear blood cells collected at birth and age 7 from 196 children of primarily self-reported Black and Hispanic ethnicities to study age- and ancestry-related patterns in DNA methylation. We developed a novel Bayesian inference method for longitudinal data and showed that even though average methylation levels changed from birth to age 7, the vast majority of the ancestry-associated methylation patterns present at birth are also present at age 7. A large proportion of ancestry-associated CpGs (59%) had a nearby methylation quantitative trait locus (meQTL) and we show that at least 13% of the ancestry-associated methylation patterns were mediated through local genotype. These combined results indicate that ancestry-associated methylation patterns in blood are in large part genetically determined. Our results further suggest that DNA methylation patterns in blood cells are robust to many environmental exposures, at least during the first 7 years of life.

43 ## Introduction

44 Epigenetic patterning in human genomes reflects the contributions of genetic variation [1, 2]
45 exposure histories [3-8], and biological factors, such as age [9-18], ethnicity [19-24] and disease
46 status [25-28], among others. However, little work has been done to elucidate the relative
47 contributions or longitudinal dynamics of each on epigenetic patterning.

48 To directly examine the relationship between age, ethnicity, genetic variation, early life
49 exposures and allergic phenotypes and an epigenetic mark, we studied global DNA methylation
50 patterns at over 750,000 CpG sites on the EPIC array in cord blood mononuclear cells (CBMCs)
51 collected at birth and in peripheral blood mononuclear cells (PBMCs) collected at 7 years of age
52 from 196 children participating in the Urban Environment and Childhood Asthma (URECA)
53 birth cohort study[29, 30]. This cohort is part of the NIAID-funded Inner-City Asthma
54 Consortium and is comprised of children primarily of Black and Hispanic self-reported ethnicity,
55 with a mother and/or father with a history of at least one allergic disease living in poor urban
56 areas (see Gern et al. [30] for details of enrollment criteria). Mothers of children in the URECA
57 study were enrolled during pregnancy and children were followed from birth through at least 7
58 years of age.

59 The longitudinal design of the URECA study provided us with the resolution to partition
60 genetic from non-genetic effects on ancestry-associated DNA methylation patterns, and yielded
61 new insight into the factors affecting DNA methylation patterns at CpG sites in mononuclear
62 (immune) cells during early life in ethnically admixed children. Using a novel statistical
63 inference method that provides a general framework for analyzing longitudinal genetic and
64 epigenetic data, we show that ancestry-dependent methylation patterns are conserved over the
65 first 7 years of life and that these patterns are strongly influenced, and often mediated, by local

genotype. Further, chronological age, but not measured exposures during pre- or post-natal periods or disease status by age 7, was associated with methylation patterns in this sample of children. Considering the results of our study and those of a recently published comprehensive review on environmental epigenetics research [31], we suggest that methylation levels in blood are not as responsive to environmental exposures as previously suggested [20], at least during the first 7 years of life.

Results

Our study included 196 children participants in the Urban Environment and Childhood Asthma (URECA) cohort who had stored cord blood mononuclear cells (CBMCs) and peripheral blood mononuclear cells (PBMCs) collected at birth and age 7, respectively [29], and passed quality control (QC) checks as described in Methods. The URECA children were classified by parent- or guardian-reported race into one of the following categories: Black, n =147; Hispanic, n = 39; White, n = 1; Mixed race n = 7, and Other, n = 2. A description of the study population is shown in Table 1 and in Supplementary Materials. Ancestry, assessed using ancestral PCs, revealed varying proportions of African and European ancestry along PC1. Becuase there is little separation along PC2 (Figure 1) and no genome-wide significant correlation between PC2 through PC10 and methylation levels at either age, we defined PC1 as inferred genetic ancestry (IGA). The reported races (RR) of the children are also shown in Figure 1. The means and ranges of gestational age stratified by reported race are shown in Table 1; the distribution of gestational age at birth is shown in Figure S1 in the supplement.

90 **Table 1**: Covariates for the 196 URECA children in our study, stratified by self-reported race.
91
92

	Black	Hispanic	White	Mixed	Other
Sample Size (N)	147	39	1	7	2
Males (%)	71 (48%)	25 (64%)	0 (0%)	4 (57%)	0 (0%)
Asthma diagnosis at age 7 (%)	38 (26%)	12 (31%)	0 (0%)	2 (29%)	0 (0%)
Gestational age at birth, in weeks (mean [range])	39.0 [34,42]	38.9 [35,41]	36.0	39.1 [37,40]	39.0 [38,40]
Sample Collection Site					
Baltimore (%)	64 (44%)	1 (3%)	1 (100%)	3 (43%)	2 (100%)
Boston (%)	17 (12%)	5 (13%)	0 (0%)	3 (29%)	0 (0%)
New York (%)	23 (16%)	32 (82%)	0 (0%)	1 (14%)	0 (0%)
St. Louis (%)	43 (29%)	1 (3%)	0 (0%)	1 (14%)	0 (0%)

93
94

95 **Inferred genetic ancestry effects on DNA methylation patterns are conserved in magnitude**
96 **and direction between birth and age 7**

97 Previous cross-sectional studies have revealed associations between ancestry and DNA
98 methylation at birth [19, 23] and later in life [20-22, 24, 25]. These correlations were generally
99 attributed to the combined effects of genetic variation and environmental exposures [20-23].
100 However, because of the cross-sectional nature of these studies, it is not known if the association
101 between ancestry and methylation patterns present at birth persist (or change) in childhood.
102 Moreover, because ancestry is typically confounded with environmental exposures [41], it has
103 been proposed that ancestry effects on methylation levels may reflect the effects of exposure

104 histories, which also may vary by race or ethnicity [20]. Alternatively, ancestry effects on DNA
105 methylation patterns could also be due to genetic differences. In these cases, we would expect
106 ancestry-associated methylation patterns to be conserved from birth to later childhood. Using the
107 longitudinal data from the URECA cohort, we tested this hypothesis by addressing three
108 questions. What is the correlation between ancestry and methylation levels at individual CpG
109 sites at birth and age 7? Is the direction and magnitude of the correlation between ancestry and
110 methylation levels conserved between birth and age 7? Are there any CpGs for which the
111 correlation between methylation and ancestry at birth is significantly different from the
112 correlation between methylation and ancestry at age 7?

113 Standard hypothesis testing can be used to answer the first question but is not appropriate
114 for answering the second or third because failure to reject the null hypothesis that the effects are
115 equal at birth and age 7 does not imply the null hypothesis is true. Additionally, because our
116 studies were conducted in cord blood cells at birth and peripheral blood cells at age 7, effects at
117 birth and age 7 may differ slightly due to differences in cell composition [42]. To circumvent
118 these issues, we built a Bayesian model (see model (S3) in the Supplement) and let the data
119 determine both the strength of the correlation between inferred genetic ancestry or reported race
120 and methylation, and how similar the correlations were at birth and age 7. We then answered the
121 first, second and third questions by defining and estimating the correct (*cor*), conserved (*con*)
122 and discordant (*dis*) sign rates for each CpG $g = 1,\ldots,784{,}484$:

123

124 $cor_g^{(0)}$ = Posterior probability that the estimate for the direction of CpG g's ancestry effect at
125 birth was correct.

126 $con_g^{(0,7)}$ = Posterior probability that the directions of CpG *g*'s ancestry effects at birth and
127 age 7 were the same AND the directions were estimated correctly.
128 $dis_g^{(0)}$ = Posterior probability that the ancestry effect for CpG *g* at birth was non-zero and
129 was 0 or in the opposite direction at age 7 AND both directions were estimated
130 correctly.
131
132 The correct and discordant sign rates at age 7 ($cor_g^{(7)}$, $dis_g^{(7)}$) were defined analogously. Because
133 the correct sign rate at birth and age 7 is always at least as large as the conserved sign rate, we
134 say that the birth and age 7 effects for CpG *g* overlap if its conserved sign rate ($con_g^{(0,7)}$) is above
135 a designated threshold. Supplemental Figures S2 and S3 provide insight into how the correct and
136 conserved sign rates compare with standard univariate *P* values. We refer the reader to "Joint
137 modeling of methylation at birth and seven" in the Supplement for a detailed description of our
138 model and estimation procedure.
139 After fitting the relevant parameters in the model to the data, we were able to estimate the
140 fraction of CpGs with non-zero effects at both ages that fell into one of four possible bins: the
141 two effects were completely unrelated ($\rho = 0$), moderately similar ($\rho = 1/3$), very similar ($\rho =$
142 $2/3$), or identical ($\rho = 1$). Note that if a non-trivial fraction of CpG sites had effects that were
143 negatively related, they would be assigned to the first bin ($\rho = 0$). In fact, less than 1% of the
144 CpGs with non-zero inferred genetic ancestry effects at both ages had unrelated or moderately
145 similar IGA effects ($\rho = 0$ or $1/3$), whereas 34% fell in the very similar bin and 66% had
146 identical inferred genetic ancestry effects at birth and age 7 (Supplemental Figure S4). This
147 indicates that when inferred genetic ancestry effects on methylation are non-zero at both birth

148 and age 7, they tend to be very similar or exactly the same with respect to both direction and
149 magnitude.

150 We then estimated the correct and conserved sign rates for all 784,484 probes, and
151 identified 2,873 inferred genetic ancestry-associated CpGs (IGA-CpGs) in CBMCs ($cor_g^{(0)} \geq$
152 0.95), 3,834 in PBMCs at age 7 ($cor_g^{(7)} \geq 0.95$), and 2,659 whose effects were conserved in sign
153 ($con_g^{(0,7)} \geq 0.95$). Methylation tended to increase with increasing African ancestry at 1,494 of the
154 2,659 conserved CpGs (56%), suggesting that individuals with more African ancestry tend to
155 have more methylation ($P < 10^{-10}$). This is consistent with the study of Moen et al. [22], which
156 used the Illumina 450K array to quantify the differences in methylation between European and
157 African populations. Supplemental Figure S5 shows the IGA-CpG locations in the genome and
158 Figure 2a illustrates the overlap between IGA-CpGs at birth and age 7. This strong overlap
159 corroborates our above observations and answers the second question in the affirmative: if
160 methylation is strongly correlated with inferred genetic ancestry at birth, the magnitude and
161 direction of the correlation is conserved at age 7.

162

163 **Inferred genetic ancestry is more correlated with methylation than is self-reported race**

164 The observed correlations between ancestry and methylation levels may reflect differences in
165 environmental exposures [20, 22], due to associations between race or ethnicity with socio-
166 cultural, nutritional, and geographic exposures, among others [41]. In fact, Galanter et al. [20]
167 showed in a cross-sectional study that self-reported ethnicity explained a substantial portion of
168 the variability in whole blood DNA methylation patterns from Latino children of diverse
169 ethnicities, even more so than inferred genetic ancestry. They concluded that ethnicity captures
170 genetic, as well as the socio-cultural and environmental differences that influence methylation

171 levels. If this were the case in the URECA children, we should observe just as large, if not larger,
172 an effect as we did for inferred genetic ancestry when we substitute reported race for inferred
173 genetic ancestry in the analyses presented in the previous section. However, using reported race,
174 we identified only 457 CpGs at birth and 709 CpGs at age 7 whose correct sign rate was at least
175 0.95, and 424 whose reported effects were conserved from birth to age 7 ($con_g^{(0,7)} \geq 0.95$), far
176 fewer than the 3,991 CpGs whose methylation was significantly correlated with inferred genetic
177 ancestry at birth or at age 7.

178 To explore this further, we examined the overlap between IGA-CpGs and reported race-
179 associated CpGs (RR-CpGs) in CBMCs at birth and in PBMCs at 7 (Figures 2c-d). Because
180 reported race is an estimate of inferred genetic ancestry, there is a still substantial overlap
181 between IGA-CpGs and RR-CpGs. In fact, almost all of the RR-CpGs are among the IGA-CpGs,
182 but the opposite is not true. This indicates that while IGA-CpGs include most RR-CpGs, reported
183 race does not capture most of the variation in methylation attributable to inferred genetic
184 ancestry in these children.

185

186 **The observed correlations between DNA methylation and ancestry are primarily genetic**

187 To further address the question of whether ancestry effects on methylation at either birth or age 7
188 were due to genetic variation or to environmental exposures, we used local genetic variation
189 (within 5kb of a CpG site) and DNA methylation data at birth and age 7 in the 147 self-reported
190 Black children in our study. Of the 519,622 CpGs within 5kb of a SNP, 65,068 and 70,898 had at
191 least one meQTL in CBMCs at birth and in PBMCs at age 7, respectively, at an FDR of 5%. In
192 addition, 59% of IGA-CpGs with at least one SNP in the ±5kb window had at least one meQTL

193 at birth or age 7 at an FDR of 5%, indicating IGA-CpGs were enriched for CpGs with meQTLs
194 (Figure 3a-b).

195 To provide additional evidence that local genotype mediates the effect of inferred genetic
196 ancestry on methylation, we used logistic regression to regress the genotype of each of the
197 269,622 SNPs in our study set onto inferred genetic ancestry. The goal was to determine the
198 fraction of IGA-CpGs that were mediated through local genotype, i.e. IGA-CpGs with both
199 edges a and c in Figure 3a. Our analysis first revealed that the genotypes at meQTLs whose
200 target CpGs were IGA-CpGs in either CBMCs at birth or PBMCs at age 7 (IGA-meQTLs) were
201 significantly more correlated with inferred genetic ancestry than the genotype of non-IGA-
202 meQTLs (Figure 3c). Moreover, approximately 13% of the IGA-CpGs with at least one SNP in
203 their ±5kb windows had an inferred genetic ancestry effect that was mediated through local
204 genotype (i.e. had edges a and c, see Supplement for calculation details), which is likely an
205 underestimate of the true number IGA-CpGs mediated through local genotype because our
206 sample size was relatively small [43]. Nonetheless, this is striking compared to the 0.1% of non-
207 IGA-CpGs whose corresponding SNP had edge c at a 20% FDR.

208 Lastly, we used DNA methylation data on 573 ethnically diverse U.S. Latino children
209 ages 9 to 16 years old from the Galanter et al. study [20] to further explore the effect of ancestry
210 on DNA methylation in whole blood. Children and teenagers in the Galanter study were
211 classified as Mexican, Puerto Rican, mixed Latino, or other Latino. They reported 916 CpG sites
212 whose methylation was significantly associated with reported race, 773 of which were also
213 significantly associated with estimated percent European, Native American and African ancestry
214 at a Bonferroni *P* value threshold of 1.6×10^{-7}. A total of 726 of their 916 ethnic-associated
215 CpGs were also in the set of 784,484 probes CpG sites in our study. Our set of IGA-CpGs at

216 birth or age 7 contained a significant fraction of the 726 ethnic-associated CpGs from the cross-
217 sectional study, but there was considerably less overlap with our RR-CpGs (Figure 4). If the
218 correlation we observed between ancestry and methylation was largely due to responses to
219 environmental exposures, as suggested in the Galanter study, then the overlap with reported race
220 should have been at least as large as the overlap with inferred genetic ancestry. That was not the
221 case, further suggesting that the inferred genetic ancestry effects on methylation in the URECA
222 cohort are primarily genetic in origin.
223
224 **Non-genetic factors influence the observed ancestry-methylation correlation more at age 7**
225 **than at birth**
226 Although most of the variation in methylation levels at ancestry-associated CpGs can be
227 attributed to genetic variation in the URECA children, a small proportion may be due to non-
228 genetic (environmental) factors. To explore this possibility, we further hypothesized that non-
229 genetic effects on methylation levels at ancestry-associated CpGs would be greater at age 7 than
230 at birth, due to accumulated exposures over the first 7 years of life. We note that none of the
231 direct or indirect measures of exposures that were available in this cohort were associated with
232 methylation levels at either age, including maternal asthma, maternal infections during
233 pregnancy, pet ownership, bedroom allergens, mother stress, anxiety and depression metrics,
234 maternal cotinine levels during pregnancy, number of smokers in the household, number of
235 siblings, number of previous live births, daycare attendance, number of colds at age 2 or 3, and
236 allergic sensitization or asthma in the child (see Supplement for details). We did, however,
237 identify 16,172 age-related CpGs (CpGs whose methylation changed from birth to age 7).
238 Besides being strongly enriched for CpGs used to predict gestational age in Knight et al. [13] and

239 chronological age in Horvath [10] (see Figure S7 in the Supplement), estimates of the age effects
240 among the CpGs that changed from birth to age 7 showed the same direction of change as their
241 corresponding estimated gestational age effects at birth in 97% of the 16,172 CpGs, which
242 included 14,186 gestational age-associated effects that were not significant at the 5% FDR
243 threshold. This concordance in direction of effect is unlikely to occur by chance ($P < 10^{-10}$, see
244 Supplement for calculation), and indicates that the majority of the change in mean methylation
245 levels from birth to age 7 was due to aging-related mechanisms rather than age-dependent
246 environmental exposures.

247 To directly test the hypothesis that non-genetic factors tend to have larger effects on
248 methylation levels at age 7 than at birth, we used our Bayesian model to estimate the proportion
249 of CpGs in our study whose methylation was not associated with ancestry at birth but associated
250 at age seven and the proportion that were associated with ancestry at birth but not at age 7. The
251 former was greater than 14% while the latter was less than 1.5% using either inferred genetic
252 ancestry or reported race as a measure of ancestry. Even though over 14% of all CpGs in our
253 study had ancestry effects present at age 7 but not at birth, we were only able to identify 18
254 discordant IGA-CpGs and 4 discordant RR-CpGs at age 7 using a liberal threshold of $dis_g^{(7)} \geq$
255 0.80. That is, for almost all CpGs that are associated with ancestry at age 7 but not at birth, the
256 expected ancestry effect sizes were quite small relative to the statistical error, making it
257 impossible to assign the direction of effect on methylation changes with confidence (Figure S6).
258 Therefore, while some CpG sites may be influenced by exposures that are correlated with
259 ancestry at age 7 but not at birth, their effects were far too small to estimate in this sample size.
260

261 **Discussion**

262 The relationships between DNA methylation, chronological age, and ancestry have the potential
263 to shed light on disease etiology and may help determine the relative genetic and environmental
264 contributions to the observed inter-individual variability of the epigenome [9-15, 19-24]. While it
265 has previously been shown that ancestry is related to DNA methylation in cross-sectional studies
266 [19-24], and that statistically significant meQTLs are conserved as one ages [44], it has yet to be
267 shown whether or not *ancestry-dependent* methylation marks are conserved as children age.

268 Even though there was substantial change in blood methylation levels over time among
269 children in this cohort, inferred genetic ancestry and self-reported race effects on methylation
270 were overwhelmingly conserved in both direction and magnitude from birth to age 7. This result
271 is interesting in and of itself because it provides an example of perinatal epigenetic variation that
272 persists later in life, and more generally an example of a persistent effect on DNA methylation
273 levels, which has been cited as a critical area of future epigenetic research [31, 45]. The
274 consistency of our estimates for the effect due to ancestry also demonstrates the fidelity of our
275 processing step to account for unobserved factors like cell composition, since failure to account
276 for latent covariates often leads to biased and irreproducible estimates [46, 47]. Furthermore, the
277 novel statistical framework we used to infer effects that are conserved versus those that vary over
278 time can be easily applied to other longitudinal DNA methylation data, as a way to avoid the
279 spurious logic often used in applications of frequentist hypothesis testing that failing to reject the
280 null hypothesis implies the null is true.

281 While the observation that inferred genetic ancestry and reported race effects are
282 conserved from birth to age 7 gives credence to the hypothesis that the effects are genetic in
283 nature, it does not rule out the possibility of environmental components or gene-environment
284 interactions that could determine ancestry-related methylation prior to birth and persist as the

285 child ages. To further explore this, we showed that the IGA-CpGs were enriched among CpGs
286 with meQTLs, and that methylation levels at many of the IGA-CpGs are mediated by local
287 genotype, indicating that much of the ancestry-methylation correlation could be attributed to
288 genetic variation. Moreover, the RR-CpGs were only a small subset of IGA-CpGs in our study.
289 This is opposite to the findings of Galanter et al. [20], who argued that ancestry-dependent
290 methylation patterns in admixed populations are in large part determined by differences in
291 exposure histories. Because their data were cross-sectional they could not evaluate whether the
292 observed patterns arose during childhood or were also present at birth. Our results provide
293 evidence for genetics accounting for most of the correlation between methylation and ancestry,
294 and implies that the genetic contribution to variability in blood methylation is substantial.
295 Our observations in support of strong genetically – and weak environmentally –
296 determined ancestry-associated methylation patterns in blood may seem paradoxical to the
297 plethora of studies showing that DNA methylation levels in blood cells are associated with
298 environmental exposures, such as cadmium, arsenic and smoking, to name a few [5-8, 20, 48-
299 52]. Whereas the estimated genetic effect sizes in our study are substantially larger than many of
300 the environmentally-associated effects on methylation patterns previously reported, the effects of
301 environmental exposures on methylation in blood are probably too small to estimate with even
302 moderate to large sample sizes [31]. For example, it was only by performing a meta-analysis in
303 6,685 individuals that Joubert et al. [5] were able to identify 6,000 CpGs whose DNA
304 methylation levels in blood from infants and adolescents were associated maternal smoking
305 exposure. In one sense, we were able to corroborate previous observations of small non-genetic
306 effects on methylation in blood by showing that while methylation patterns at an estimated 14%
307 of all CpGs in our study were not correlated with ancestry at birth but correlated with ancestry at

308 age 7, the correlation at individual CpGs at age 7 was too small to be identified as statistically
309 significant. We were also not able to find any statistically significant correlations between
310 methylation at birth or at age 7 and any of the environmental exposure variables measured in this
311 cohort. We note that cord blood cotinine levels, a measure of *in utero* tobacco smoke exposure,
312 were above the level of detection in only 34 of the 196 mothers in our study.

313 An unsurprising feature of these longitudinal data is that average methylation levels of
314 over 16,000 CpGs changed significantly from birth to age 7. However, what was quite
315 remarkable was that the direction of the change in 97% of those CpGs matched the direction of
316 their corresponding estimated gestational age effect at birth, which included over 14,000
317 gestational age-associated effects that were not genome-wide significant. Not only does this fit
318 with the above narrative and suggest that methylation levels of the vast majority of the 16,172
319 age-related CpGs were in fact changing due to age-related mechanisms and not because of
320 differences in environmental exposures at birth and age 7, it also indicates that the "epigenetic
321 clock" present at birth may be the same as that present later in life. While we do not have the
322 data to explore this further, this remains an important avenue of future research.

323 The results of our study suggest that DNA methylation levels in blood cells are fairly
324 robust to environmental exposures, including those that are correlated with self reported race. A
325 better understanding of tissue-specific methylation responses to environmental exposures could
326 inform the design of future studies and provide insights into the mechanisms through which
327 exposures and gene-environment interactions influence health and disease.

328

329 Materials and methods

330 Sample composition

331 URECA is a birth cohort study initiated in 2005 in Baltimore, Boston, New York City and St.
332 Louis under the NIAID-funded Inner City Asthma Consortium [29]. Pregnant women were
333 recruited. Either they or the father of their unborn child had a history of asthma, allergic rhinitis,
334 or eczema, and deliveries prior to 34 weeks gestation were excluded (see Gern et al. [29] for full
335 entry criteria). Informed consent was obtained from the women at enrollment and from the
336 parent or legal guardian of the infant after birth.
337 Maternal questionnaires were administered prenatally and child health questionnaires
338 administered to a parent or caregiver every 3 months through age 7 years. Gestational age at
339 birth and obstetric history were obtained from medical records. Additional details on study
340 design are described in Gern et al. [29] and in the Supplement. Frozen paired cord blood
341 mononuclear cells (CBMCs) and peripheral blood mononuclear cells (PBMCs) at age 7, were
342 available for 196 of the 560 URECA children after completing other studies. After QC (see
343 below), DNA methylation data were available for 194 children at birth, 195 children at age 7,
344 and 193 children at both time points; genotype data were available in 193 children (194 at birth;
345 195 at age 7) (Supplementary Table 1).
346
347 **DNA Methylation**
348 DNA for methylation studies was extracted from thawed CBMCs and PBMCs using the Qiagen
349 AllPrep kit (QIAGEN, Valencia, CA). Genome-wide DNA methylation was assessed using the
350 Illumina Infinium MethylationEPIC BeadChip (Illumina, San Diego, CA) at the University of
351 Chicago Functional Genomics Facility (UC-FGF). Birth and 7-year samples from the same child
352 were assayed on the same chip and the data were processed using Minfi [32]; Infinium type I and
353 type II probe bias were corrected using SWAN [33]. Raw probe values were corrected for color

354 imbalance and background by control normalization. Three out of the 392 samples (two at birth
355 and one at age 7) were removed as outliers following normalization. We removed 82,352 probes
356 that mapped either to the sex chromosomes or to more than one location in a bisulfite-converted
357 genome, had detection *P* values greater than 0.01% in 25% or more of the samples, or
358 overlapped with known SNPs with minor allele frequency of at least 5% in African, American or
359 European populations. After processing, 784,484 probes were retained and M-values were used
360 for all downstream analyses, which were computed as $\log_2$ (methylated intensity + 100) - $\log_2$
361 (unmethylated intensity + 100). The offset of 100 was recommended in Du et al. [34].
362

363 **Genotyping**

364 DNA from the 196 URECA children was genotyped with the Illumina Infinium
365 CoreExome+Custom array. Of the 532,992 autosomal SNPs on the array, 531,755 passed
366 Quality control (QC) (excluding SNPs with call rate <95%, Hardy-Weinberg *P* values $<10^{-5}$, and
367 heterozygosity outliers). We conducted all analyses in 293,696 autosomal SNPs with a minor
368 allele frequency ≥5%. Genotypes for three children failed QC and were excluded from
369 subsequent analysis that involved genotypes, including methylation quantitative locus (meQTL)
370 mapping, inferred genetic ancestry, or used genetic ancestry PC1 as a covariate (see below).
371 These three children were included in all other analyses.
372

373 **Estimating inferred genetic ancestry**

374 Ancestral principal component analysis (PCA) was performed using a set of 801 ancestry
375 informative markers (AIMs) from Tandon et al. [35] that were genotyped in both the URECA
376 children and in HapMap [36] release 23. Because PC1 captured the majority of variation in

377 genetic ancestry (Figure 1), we refer to PC1 as inferred genetic ancestry and consider it as a
378 surrogate measure for percent African ancestry.
379
380 **Statistical analysis**
381 To determine the effect of gestational age on methylation in CBMCs, we used standard linear
382 regression models with the child's gender, sample collection site, inferred genetic ancestry and
383 methylation plate number as covariates in our model. We also estimated cell composition and
384 other unobserved confounding factors using a method described in McKennan et al. [37]. We
385 then computed a gestational age *P* value for each CpG site and used q-values [38] to control the
386 false discovery rate at a nominal level. We took the same approach to determine CpGs whose
387 methylation changed from birth to age 7, except the response was measured as the difference in
388 methylation at birth and age 7. In this analysis, we included the child's gender, gestational age at
389 birth, inferred genetic ancestry and sample collection site as covariates. Because all paired
390 samples were on the same plate, we did not include plate number as a covariate in this analysis.
391 We also estimated unobserved factors that influence differences in methylation at birth and age 7
392 using McKennan et al. [37] and included these latent factors in our linear model. See models
393 (S1) and (S2) in the Supplement for more detail.
394 We used data from the self-reported Hispanic and Black individuals with methylation
395 measured at both time points to analyze the effect of ancestry (either inferred genetic ancestry or
396 self-reported race) on methylation at birth and age 7 jointly using a Bayesian model. We did not
397 include the 10 individuals of other reported races in this analysis because we did not want our
398 estimates to be influenced by the groups with small samples sizes. We included age (birth or age
399 7), sample collection site, gestational age at birth, gender and methylation plate number as

400 covariates in our model, and estimated additional unobserved covariates (including cell
401 composition) using a method specifically designed for correlated data [39]. Once we estimated
402 the relevant hyper-parameters, we extended the sign rate paradigm developed in Stephens [40] to
403 perform inference in longitudinal data. This is discussed in more detail in the context of the
404 specific questions we present in the results section. We encourage the reader to review the
405 Supplement for a more detailed presentation of this model and previously discussed models.
406 We performed meQTL mapping in the 145 genotyped, self-reported Black children using
407 the set of 269,622 SNPs with 100% genotype call rate in this subset. We restricted ourselves to
408 this subset of samples to minimize heterogeneity in effect sizes. To identify CpG-SNP pairs, we
409 considered SNPs within 5kb of each CpG, as this region has been previously shown to contain
410 the majority of genetic variability in DNA methylation [1] and is small enough to mitigate the
411 multiple testing burden, and computed a *P* value for the effect of the genotype at a single SNP on
412 methylation at the corresponding CpG with ordinary least squares. We then defined the meQTL
413 for each CpG site as the SNP with the lowest *P* value. In addition to genotype, we included
414 inferred genetic ancestry (i.e., ancestry PC1), gestational age at birth, gender, sample collection
415 site and methylation plate number in the linear model, along with the first nine principal
416 components of the residual methylation data matrix after regressing out the intercept and the five
417 additional covariates. We then tested the null hypothesis that a CpG did not have an meQTL in
418 the 10kb region by using the minimum marginal *P* value in the region as the test statistic and
419 computed its significance via bootstrap. Lastly, we used q-values to control the false discovery
420 rate.
421
422

Acknowledgements

This work was supported in part by NIH grants U19 AI106683, R01 HL129735, R01 HL122712, and P01 HL070831. The URECA study has been funded by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, under Contract numbers NO1-AI-25496, NO1-AI-25482, HHSN272200900052C, HHSN272201000052I, 1UM1AI114271-01 and UM2AI117870. Additional support was provided by the National Center for Research Resources, National Institutes of Health, under grants RR00052, M01RR00533, 1UL1RR025771, M01RR00071, 1UL1RR024156, UL1TR000040, UL1TR001079 and 5UL1RR024992-02.

432 **References**

433 1. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA
434 methylation patterns associate with genetic and gene expression variation in HapMap cell lines.
435 Genome Biol. 2011;12(1):R10. Epub 2011/01/22. doi: 10.1186/gb-2011-12-1-r10. PubMed
436 PMID: 21251332; PubMed Central PMCID: PMCPMC3091299.

437 2. Smith AK, Kilaru V, Kocak M, Almli LM, Mercer KB, Ressler KJ, et al. Methylation
438 quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage,
439 and tissue type. BMC Genomics. 2014;15:145. Epub 2014/02/22. doi: 10.1186/1471-2164-15-
440 145. PubMed PMID: 24555763; PubMed Central PMCID: PMCPMC4028873.

441 3. Chatterton Z, Hartley BJ, Seok MH, Mendelev N, Chen S, Milekic M, et al. In utero
442 exposure to maternal smoking is associated with DNA methylation alterations and reduced
443 neuronal content in the developing fetal brain. Epigenetics Chromatin. 2017;10:4. Epub
444 2017/02/06. doi: 10.1186/s13072-017-0111-y. PubMed PMID: 28149327; PubMed Central
445 PMCID: PMCPMC5270321.

446 4. Goodrich JM, Dolinoy DC, Sanchez BN, Zhang Z, Meeker JD, Mercado-Garcia A, et al.
447 Adolescent epigenetic profiles and environmental exposures from early life through peri-
448 adolescence. Environ Epigenet. 2016;2(3):dvw018. Epub 2016/08/14. doi: 10.1093/eep/dvw018.
449 PubMed PMID: 29492298; PubMed Central PMCID: PMCPMC5804533.

450 5. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. DNA
451 Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-
452 analysis. Am J Hum Genet. 2016;98(4):680-96. Epub 2016/04/05. doi:
453 10.1016/j.ajhg.2016.02.019. PubMed PMID: 27040690; PubMed Central PMCID:
454 PMCPMC4833289.

455 6. Kippler M, Engstrom K, Mlakar SJ, Bottai M, Ahmed S, Hossain MB, et al. Sex-specific
456 effects of early life cadmium exposure on DNA methylation and implications for birth weight.
457 Epigenetics. 2013;8(5):494-503. Epub 2013/05/07. doi: 10.4161/epi.24401. PubMed PMID:
458 23644563; PubMed Central PMCID: PMCPMC3741219.

459 7. Koestler DC, Avissar-Whiting M, Houseman EA, Karagas MR, Marsit CJ. Differential
460 DNA methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero.
461 Environ Health Perspect. 2013;121(8):971-7. Epub 2013/06/13. doi: 10.1289/ehp.1205925.
462 PubMed PMID: 23757598; PubMed Central PMCID: PMCPMC3733676.

463 8. Rzehak P, Saffery R, Reischl E, Covic M, Wahl S, Grote V, et al. Maternal Smoking
464 during Pregnancy and DNA-Methylation in Children at Age 5.5 Years: Epigenome-Wide-
465 Analysis in the European Childhood Obesity Project (CHOP)-Study. PLoS One.
466 2016;11(5):e0155554. Epub 2016/05/14. doi: 10.1371/journal.pone.0155554. PubMed PMID:
467 27171005; PubMed Central PMCID: PMCPMC4865176.

468 9. Bocklandt S, Lin W, Sehl ME, Sanchez FJ, Sinsheimer JS, Horvath S, et al. Epigenetic
469 predictor of age. PLoS One. 2011;6(6):e14821. Epub 2011/07/07. doi:
470 10.1371/journal.pone.0014821. PubMed PMID: 21731603; PubMed Central PMCID:
471 PMCPMC3120753.

472 10. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol.
473 2013;14(10):R115. Epub 2013/10/22. doi: 10.1186/gb-2013-14-10-r115. PubMed PMID:
474 24138928; PubMed Central PMCID: PMCPMC4015143.

475 11. Horvath S, Erhart W, Brosch M, Ammerpohl O, von Schonfels W, Ahrens M, et al.
476 Obesity accelerates epigenetic aging of human liver. Proc Natl Acad Sci U S A.
477 2014;111(43):15538-43. Epub 2014/10/15. doi: 10.1073/pnas.1412759111. PubMed PMID:
478 25313081; PubMed Central PMCID: PMCPMC4217403.

479 12. Johnson AA, Akman K, Calimport SR, Wuttke D, Stolzing A, de Magalhaes JP. The role
480 of DNA methylation in aging, rejuvenation, and age-related disease. Rejuvenation Res.
481 2012;15(5):483-94. Epub 2012/10/27. doi: 10.1089/rej.2012.1324. PubMed PMID: 23098078;
482 PubMed Central PMCID: PMCPMC3482848.

483 13. Knight AK, Craig JM, Theda C, Baekvad-Hansen M, Bybjerg-Grauholm J, Hansen CS,
484 et al. An epigenetic clock for gestational age at birth based on blood methylation data. Genome
485 Biol. 2016;17(1):206. Epub 2016/10/09. doi: 10.1186/s13059-016-1068-z. PubMed PMID:
486 27717399; PubMed Central PMCID: PMCPMC5054584.

487 14. Levine ME, Crimmins EM. Evidence of accelerated aging among African Americans and
488 its implications for mortality. Soc Sci Med. 2014;118:27-32. Epub 2014/08/03. doi:
489 10.1016/j.socscimed.2014.07.022. PubMed PMID: 25086423; PubMed Central PMCID:
490 PMCPMC4197001.

491 15. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, et al. DNA
492 methylation age of blood predicts all-cause mortality in later life. Genome Biol. 2015;16:25.
493 Epub 2015/01/31. doi: 10.1186/s13059-015-0584-6. PubMed PMID: 25633388; PubMed Central
494 PMCID: PMCPMC4350614.

495 16. Parets SE, Conneely KN, Kilaru V, Fortunato SJ, Syed TA, Saade G, et al. Fetal DNA
496 Methylation Associates with Early Spontaneous Preterm Birth and Gestational Age. PLoS One.
497 2013;8(6):e67489. Epub 2013/07/05. doi: 10.1371/journal.pone.0067489. PubMed PMID:
498 23826308; PubMed Central PMCID: PMCPMC3694903.

499 17. Schroeder JW, Conneely KN, Cubells JC, Kilaru V, Newport DJ, Knight BT, et al.
500 Neonatal DNA methylation patterns associate with gestational age. Epigenetics.
501 2011;6(12):1498-504. Epub 2011/12/06. doi: 10.4161/epi.6.12.18296. PubMed PMID:
502 22139580; PubMed Central PMCID: PMCPMC3256334.

503 18. Simpkin AJ, Suderman M, Gaunt TR, Lyttleton O, McArdle WL, Ring SM, et al.
504 Longitudinal analysis of DNA methylation associated with birth weight and gestational age.
505 Hum Mol Genet. 2015;24(13):3752-63. Epub 2015/04/15. doi: 10.1093/hmg/ddv119. PubMed
506 PMID: 25869828; PubMed Central PMCID: PMCPMC4459393.

507 19. Adkins RM, Krushkal J, Tylavsky FA, Thomas F. Racial differences in gene-specific
508 DNA methylation levels are present at birth. Birth Defects Res A Clin Mol Teratol.
509 2011;91(8):728-36. Epub 2011/02/11. doi: 10.1002/bdra.20770. PubMed PMID: 21308978;
510 PubMed Central PMCID: PMCPMC3429933.

511 20. Galanter JM, Gignoux CR, Oh SS, Torgerson D, Pino-Yanes M, Thakur N, et al.
512 Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and
513 environmental exposures. Elife. 2017;6. Epub 2017/01/04. doi: 10.7554/eLife.20532. PubMed
514 PMID: 28044981; PubMed Central PMCID: PMCPMC5207770.

515 21. Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, Sandoval J, et al. DNA
516 methylation contributes to natural human variation. Genome Res. 2013;23(9):1363-72. Epub
517 2013/08/03. doi: 10.1101/gr.154187.112. PubMed PMID: 23908385; PubMed Central PMCID:
518 PMCPMC3759714.

519 22. Moen EL, Zhang X, Mu W, Delaney SM, Wing C, McQuade J, et al. Genome-wide
520 variation of cytosine modifications between European and African populations and the
521 implications for complex traits. Genetics. 2013;194(4):987-96. Epub 2013/06/25. doi:
522 10.1534/genetics.113.151381. PubMed PMID: 23792949; PubMed Central PMCID:
523 PMCPMC3730924.

524 23. Mozhui K, Smith AK, Tylavsky FA. Ancestry dependent DNA methylation and influence
525 of maternal nutrition. PLoS One. 2015;10(3):e0118466. Epub 2015/03/06. doi:
526 10.1371/journal.pone.0118466. PubMed PMID: 25742137; PubMed Central PMCID:
527 PMCPMC4350920.

528 24. Rahmani E, Shenhav L, Schweiger R, Yousefi P, Huen K, Eskenazi B, et al. Genome-
529 wide methylation data mirror ancestry information. Epigenetics Chromatin. 2017;10:1. Epub
530 2017/02/06. doi: 10.1186/s13072-016-0108-y. PubMed PMID: 28149326; PubMed Central
531 PMCID: PMCPMC5267476.

532 25. Chan MA, Ciaccio CE, Gigliotti NM, Rezaiekhaligh M, Siedlik JA, Kennedy K, et al.
533 DNA methylation levels associated with race and childhood asthma severity. J Asthma.
534 2017;54(8):825-32. Epub 2016/12/09. doi: 10.1080/02770903.2016.1265126. PubMed PMID:
535 27929694.

536 26. Ladd-Acosta C, Hansen KD, Briem E, Fallin MD, Kaufmann WE, Feinberg AP.
537 Common DNA methylation alterations in multiple brain regions in autism. Mol Psychiatry.

538 2014;19(8):862-71. Epub 2013/09/04. doi: 10.1038/mp.2013.114. PubMed PMID: 23999529;
539 PubMed Central PMCID: PMCPMC4184909.

540 27. Nicodemus-Johnson J, Myers RA, Sakabe NJ, Sobreira DR, Hogarth DK, Naureckas ET,
541 et al. DNA methylation in lung cells is associated with asthma endotypes and genetic risk. JCI
542 Insight. 2016;1(20):e90151. Epub 2016/12/13. doi: 10.1172/jci.insight.90151. PubMed PMID:
543 27942592; PubMed Central PMCID: PMCPMC5139904 PulmOne Advanced Medical Devices
544 Ltd., Israel, and received reimbursement for expenses. He served on the Respiratory Therapy
545 Clinical Advisory Board for Hollister Inc., and for this received honoraria and was reimbursed
546 for travel and meal expenses incurred during meetings. He has received a research grant from
547 AstraZeneca Inc., from 2006 to 2014, that was administered through the University of Chicago.
548 He has multiple patents concerning a smooth muscle gene promoter and one pending concerning
549 a method to determine respiratory physiological parameters (6090618; 6114311; 6284743;
550 6291211; 6297221; 6331527; 7169764). He has consulted for Novartis Institute for Biomedical
551 Research, for which he received an honorarium and travel reimbursement. He was a member of
552 the scientific advisory board for Cytokinetics Inc., for which he received honoraria and travel
553 reimbursement.

554 28. Yokoyama AS, Rutledge JC, Medici V. DNA methylation alterations in Alzheimer's
555 disease. Environ Epigenet. 2017;3(2):dvx008. Epub 2018/03/02. doi: 10.1093/eep/dvx008.
556 PubMed PMID: 29492310; PubMed Central PMCID: PMCPMC5804548.

557 29. Gern JE, Visness CM, Gergen PJ, Wood RA, Bloomberg GR, O'Connor GT, et al. The
558 Urban Environment and Childhood Asthma (URECA) birth cohort study: design, methods, and
559 study population. BMC Pulm Med. 2009;9:17. Epub 2009/05/12. doi: 10.1186/1471-2466-9-17.
560 PubMed PMID: 19426496; PubMed Central PMCID: PMCPMC2689166.

561 30. O'Connor GT, Lynch SV, Bloomberg GR, Kattan M, Wood RA, Gergen PJ, et al. Early-
562 life home environment and risk of asthma among inner-city children. J Allergy Clin Immunol.
563 2017. Epub 2017/09/25. doi: 10.1016/j.jaci.2017.06.040. PubMed PMID: 28939248.

564 31. Breton CV, Marsit CJ, Faustman E, Nadeau K, Goodrich JM, Dolinoy DC, et al. Small-
565 Magnitude Effect Sizes in Epigenetic End Points are Important in Children's Environmental
566 Health Studies: The Children's Environmental Health and Disease Prevention Research Center's
567 Epigenetics Working Group. Environ Health Perspect. 2017;125(4):511-26. Epub 2017/04/01.
568 doi: 10.1289/EHP595. PubMed PMID: 28362264; PubMed Central PMCID: PMCPMC5382002.

569 32. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al.
570 Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA
571 methylation microarrays. Bioinformatics. 2014;30(10):1363-9. Epub 2014/01/31. doi:
572 10.1093/bioinformatics/btu049. PubMed PMID: 24478339; PubMed Central PMCID:
573 PMCPMC4016708.

574 33. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile within array normalization
575 for illumina infinium HumanMethylation450 BeadChips. Genome Biol. 2012;13(6):R44. Epub
576 2012/06/19. doi: 10.1186/gb-2012-13-6-r44. PubMed PMID: 22703947; PubMed Central
577 PMCID: PMCPMC3446316.

578 34. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value
579 and M-value methods for quantifying methylation levels by microarray analysis. BMC
580 Bioinformatics. 2010;11:587. Epub 2010/12/02. doi: 10.1186/1471-2105-11-587. PubMed
581 PMID: 21118553; PubMed Central PMCID: PMCPMC3012676.

582 35. Tandon A, Patterson N, Reich D. Ancestry informative marker panels for African
583 Americans based on subsets of commercially available SNP arrays. Genet Epidemiol.
584 2011;35(1):80-3. Epub 2010/12/25. doi: 10.1002/gepi.20550. PubMed PMID: 21181899;
585 PubMed Central PMCID: PMCPMC4386999.

586 36. International HapMap C. The International HapMap Project. Nature.
587 2003;426(6968):789-96. Epub 2003/12/20. doi: 10.1038/nature02168. PubMed PMID:
588 14685227.

589 37. McKennan C, Nicolae DL. Accounting for unobserved covariates with varying degrees
590 of estimability in high dimensional biological data; 2018. Preprint. Available from: arXiv:
591 arXiv:1801.00865v2. Cited 22 August 2018

592 38. Storey JD. A direct approach to false discovery rates. J Royal Stat Soc B. 2002;64:479-
593 98.

594 39. McKennan C, Nicolae DL. Estimating and accounting for unobserved covariates in high
595 dimensional correlated data; 2018. Preprint. Available from: arXiv:1808.05895v1. Cited 22
596 August 2018

597 40. Stephens M. False discovery rates: a new deal. Biostatistics. 2017;18(2):275-94. Epub
598 2016/10/21. doi: 10.1093/biostatistics/kxw041. PubMed PMID: 27756721; PubMed Central
599 PMCID: PMCPMC5379932.

600 41. Nguyen AB, Moser R, Chou WY. Race and health profiles in the United States: an
601 examination of the social gradient through the 2009 CHIS adult survey. Public Health.
602 2014;128(12):1076-86. Epub 2014/12/03. doi: 10.1016/j.puhe.2014.10.003. PubMed PMID:
603 25457801.

604 42. Fu J, Wolfs MG, Deelen P, Westra HJ, Fehrmann RS, Te Meerman GJ, et al. Unraveling
605 the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression.
606 PLoS Genet. 2012;8(1):e1002431. Epub 2012/01/26. doi: 10.1371/journal.pgen.1002431.
607 PubMed PMID: 22275870; PubMed Central PMCID: PMCPMC3261927.

608 43. Fritz MS, Mackinnon DP. Required sample size to detect the mediated effect. Psychol
609 Sci. 2007;18(3):233-9. Epub 2007/04/21. doi: 10.1111/j.1467-9280.2007.01882.x. PubMed
610 PMID: 17444920; PubMed Central PMCID: PMCPMC2843527.

611 44. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic
612 identification of genetic influences on methylation across the human life course. Genome Biol.
613 2016;17:61. Epub 2016/04/03. doi: 10.1186/s13059-016-0926-z. PubMed PMID: 27036880;
614 PubMed Central PMCID: PMCPMC4818469.

615 45. Martin EM, Fry RC. Environmental Influences on the Epigenome: Exposure-Associated
616 DNA Methylation in Human Populations. Annu Rev Public Health. 2018. Epub 2018/01/13. doi:
617 10.1146/annurev-publhealth-040617-014629. PubMed PMID: 29328878.

618 46. Peixoto L, Risso D, Poplawski SG, Wimmer ME, Speed TP, Wood MA, et al. How data
619 analysis affects power, reproducibility and biological insight of RNA-seq studies in complex
620 datasets. Nucleic Acids Res. 2015;43(16):7664-74. Epub 2015/07/24. doi: 10.1093/nar/gkv736.
621 PubMed PMID: 26202970; PubMed Central PMCID: PMCPMC4652761.

622 47. Yao C, Li H, Shen X, He Z, He L, Guo Z. Reproducibility and concordance of
623 differential DNA methylation and gene expression in cancer. PLoS One. 2012;7(1):e29686.
624 Epub 2012/01/12. doi: 10.1371/journal.pone.0029686. PubMed PMID: 22235325; PubMed
625 Central PMCID: PMCPMC3250460.

626 48. Joubert BR, Haberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 450K
627 epigenome-wide scan identifies differential DNA methylation in newborns related to maternal
628 smoking during pregnancy. Environ Health Perspect. 2012;120(10):1425-31. Epub 2012/08/02.
629 doi: 10.1289/ehp.1205412. PubMed PMID: 22851337; PubMed Central PMCID:
630 PMCPMC3491949.

631 49. Lee KW, Richmond R, Hu P, French L, Shin J, Bourdon C, et al. Prenatal exposure to
632 maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery
633 sample of adolescents and replication in an independent cohort at birth through 17 years of age.
634 Environ Health Perspect. 2015;123(2):193-9. Epub 2014/10/18. doi: 10.1289/ehp.1408614.
635 PubMed PMID: 25325234; PubMed Central PMCID: PMCPMC4314251.

636 50. Markunas CA, Xu Z, Harlid S, Wade PA, Lie RT, Taylor JA, et al. Identification of DNA
637 methylation changes in newborns related to maternal smoking during pregnancy. Environ Health
638 Perspect. 2014;122(10):1147-53. Epub 2014/06/07. doi: 10.1289/ehp.1307892. PubMed PMID:
639 24906187; PubMed Central PMCID: PMCPMC4181928.

640 51. Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL, et al.
641 Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse:
642 findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). Hum Mol

643 Genet. 2015;24(8):2201-17. Epub 2015/01/02. doi: 10.1093/hmg/ddu739. PubMed PMID:
644 25552657; PubMed Central PMCID: PMCPMC4380069.

645 52. Wu D, Yang H, Winham SJ, Natanzon Y, Koestler DC, Luo T, et al. Mediation analysis
646 of alcohol consumption, DNA methylation, and epithelial ovarian cancer. J Hum Genet.
647 2018;63(3):339-48. Epub 2018/01/13. doi: 10.1038/s10038-017-0385-8. PubMed PMID:
648 29321518.
649

650 **Figure Legends**

651

652 **Figure** 1: Estimated ancestry principal components (PCs) 1 and 2. Nearly all the variation in
653 ancestry separates along PC1 in the URECA sample. Filled triangles represent the 196
654 URECA children in this study, with their self-reported race shown in different colors.
655 Open circles are reference control samples from HapMap; red = Utah residents from
656 northern and western Europe (CEU), yellow = east Asian (Chinese and Japanese); dark
657 blue = Africans from Nigeria (Yoruban).

658 **Figure** 2: Overlapping ancestry CpGs at birth and at age 7. (a): IGA-CpGs in CBMCs at birth
659 ($\hat{cor}^{(0)} \geq 0.95$) and PBMCs at age 7 ($\hat{cor}^{(7)} \geq 0.95$). Overlapping CpGs (violet) have
660 $\hat{con}^{(0,7)} \geq 0.95$. (b): RR-CpGs at in CBMCs at birth ($\hat{cor}^{(0)} \geq 0.95$) and PBMCs at age 7 (
661 $\hat{cor}^{(7)} \geq 0.95$). Overlapping CpGs (violet) have $\hat{con}^{(0,7)} \geq 0.95$. (c): IGA-CpGs and RR-
662 CpGs ($\hat{cor}^{(0)} \geq 0.95$) in CBMCs at birth. (d): IGA-CpGs and RR-CpGs ($\hat{cor}^{(0)} \geq 0.95$) in
663 PBMCs at age 7.

664 **Figure** 3: IGA-CpGs are enriched for CpGs with meQTLs. (a) Illustration of the causal
665 relationship between the methylation (M) at a CpG site, the genotype (G) at the SNP
666 within ±5kb of the CpG that had the smallest meQTL *P* value and inferred genetic
667 ancestry (IGA). Each graph corresponds to a unique CpG. (b) Plots of the meQTL *P* value
668 for edge a in CBMCs at birth, where CpGs were stratified by whether or not it was an
669 IGA-CpG at birth or age 7 ($\max\left(\hat{cor}_g^{(0)}, \hat{cor}_g^{(7)}\right) \geq 0.95$). The ten enlarged red circles are
670 just for visual aid. (c) Plots of the logistic regression *P* value for edge c (Genotype $\sim$ IGA
671 $+$ Ancestry PC2 (see Figure 1)), stratified by whether or not the SNP was an IGA-meQTL.

672 **Figure** 4: Inferred genetic ancestry effects on methylation are primarily genetic in origin.
673 Histograms of $\max\left(\hat{cor}^{(0)}, \hat{cor}^{(7)}\right)$ in the IGA (a) and the RR (b) analyses. If

674 $\max\left(\hat{cor}_g^{(0)}, \hat{cor}_g^{(7)}\right)$ is close to 1 in the IGA or RR analysis, CpG g is an IGA-CpG or RR-

675 CpG, respectively, in CBMCs at birth or PBMCs at 7. The red histogram is created with

676 the set of 726 ethnicity-associated CpGs identified in Galanter et al. [20] that were also

677 among the 784,484 CpGs in our study and the blue are the remaining 783,758 CpGs.

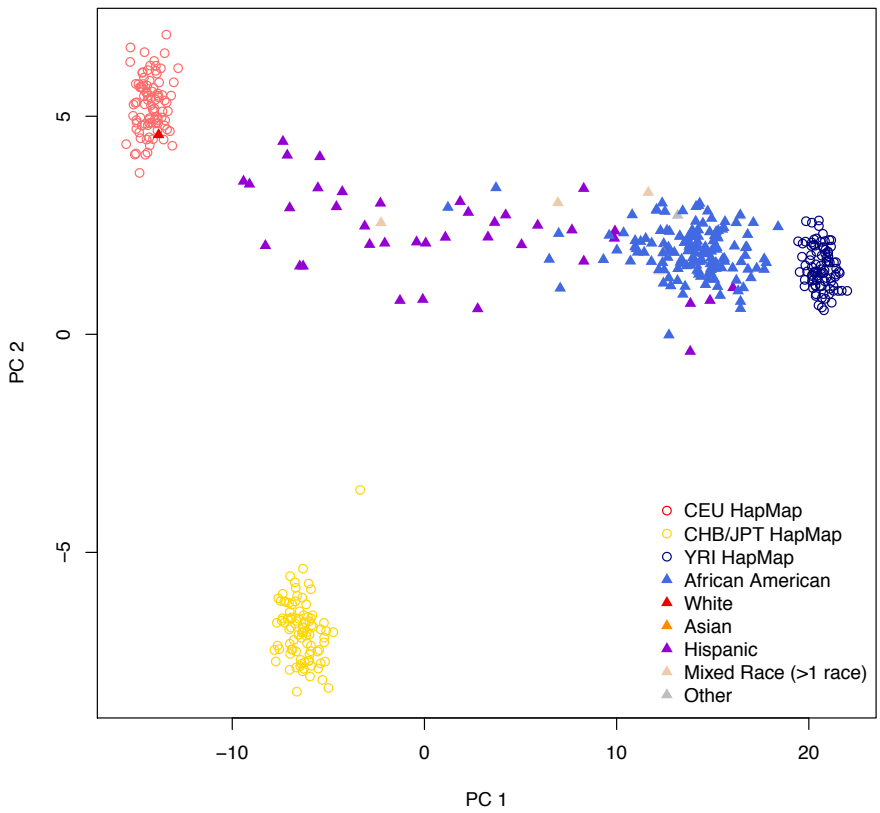
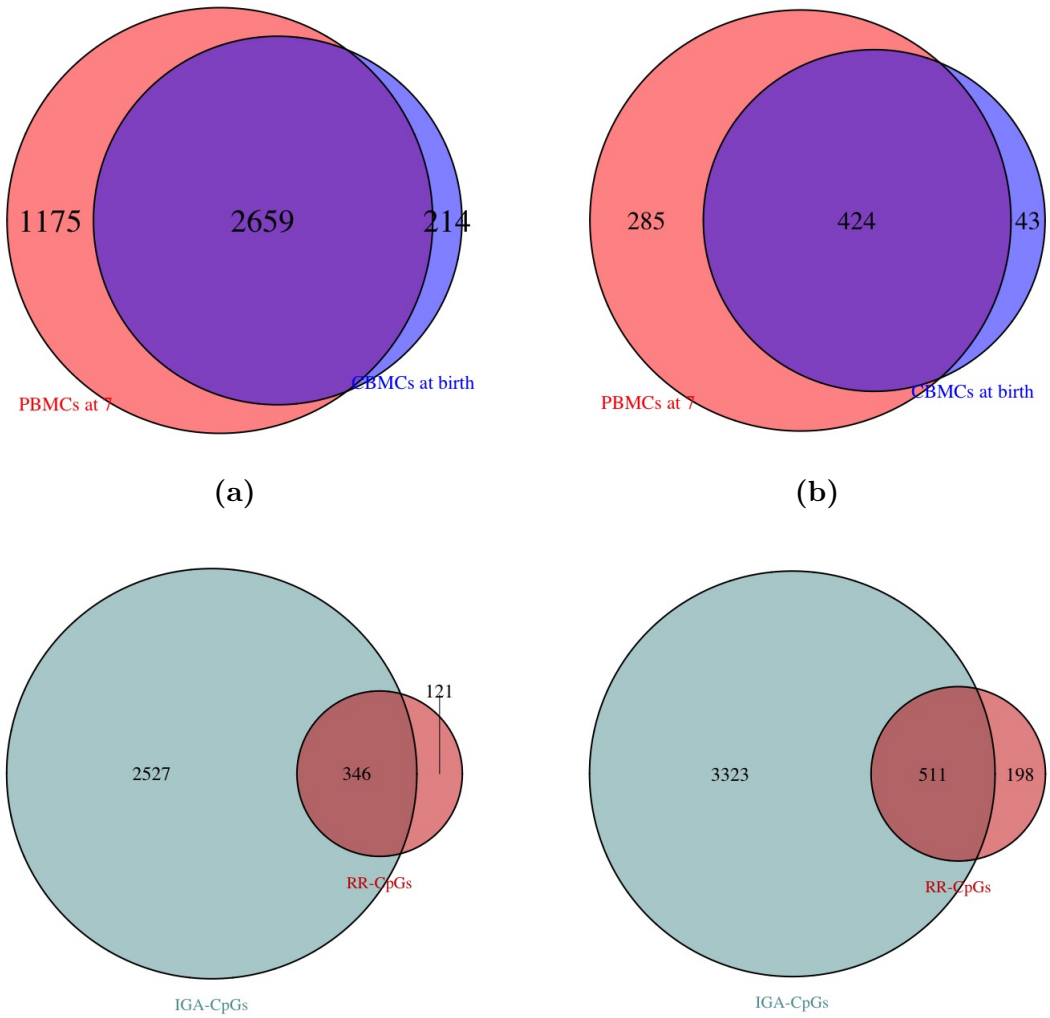

PC 2
5
0
−5
−10
0
10
20
PC 1
CEU HapMap
CHB/JPT HapMap
YRI HapMap
African American
White
Asian
Hispanic
Mixed Race (>1 race)
Other

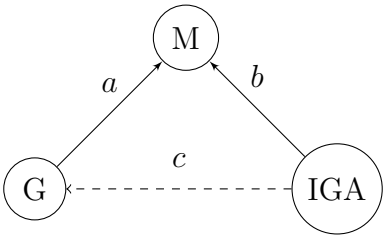
Figure 1



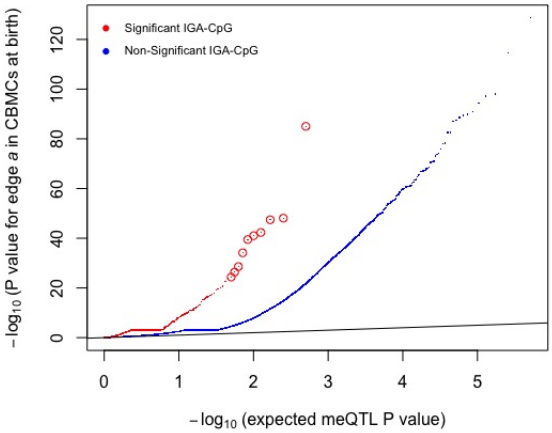
(a) **(b)**

(c) **(d)**

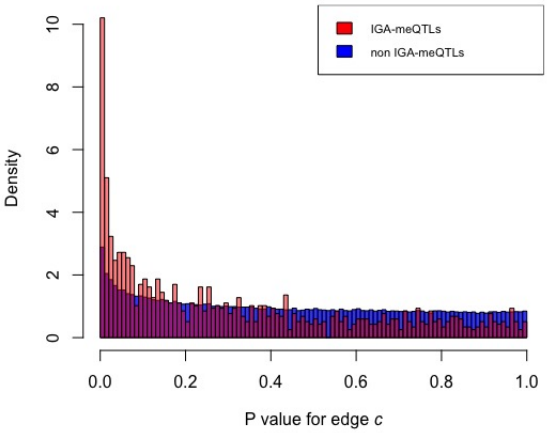
Figure 2


M
a
b
c
G
IGA

(a)

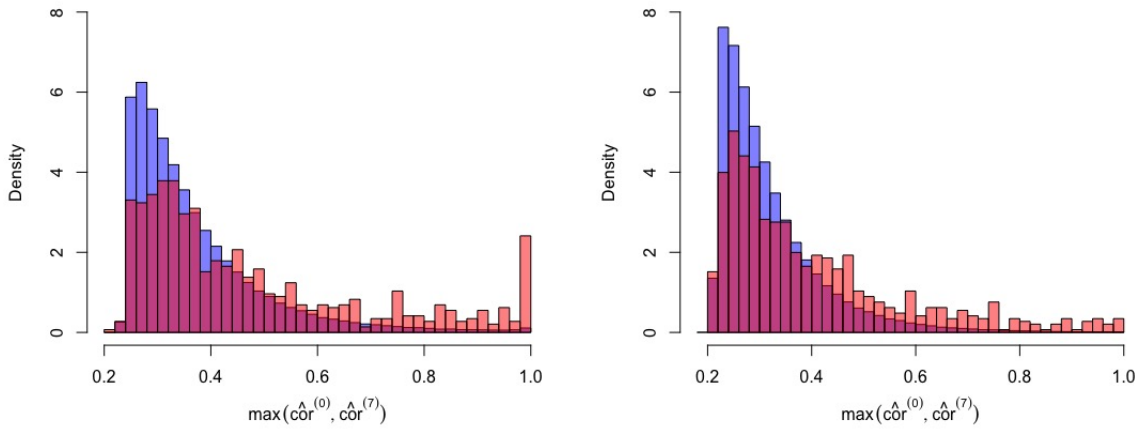

Significant IGA-CpG
Non-Significant IGA-CpG
−log10 (P value for edge a in CBMCs at birth)
−log10 (expected meQTL P value)
0
20
40
60
80
100
120
0
1
2
3
4
5

(b)


IGA-meQTLs
non IGA-meQTLs
Density
P value for edge c
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0

(c)

Figure 3


Density
0
2
4
6
8
0.2
0.4
0.6
0.8
1.0
$\max(\hat{cor}^{(0)}, \hat{cor}^{(7)})$
Density
0
2
4
6
8
0.2
0.4
0.6
0.8
1.0
$\max(\hat{cor}^{(0)}, \hat{cor}^{(7)})$

(a) **(b)**

Figure 4