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:
Carole Ober, PhD`
Department of Human Genetics
University of Chicago
TEL: 773-834-0735
FAX: 773-834-0505
Email: c-ober@genetics.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 first identified a large set of age-related CpGs at birth and age 7 and showed that gestational age related CpGs at birth are the same as the age-related CpGs at age 7. We then developed a novel Bayesian inference method for longitudinal data and showed that 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 environmental exposures, at least during the first 7 years of life.

48 **Introduction**

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

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

63 The longitudinal design of the URECA study provided us with the resolution to partition
64 genetic from non-genetic effects on ancestry-associated DNA methylation patterns, and yielded
65 new insight into the factors affecting DNA methylation patterns at CpG sites in mononuclear
66 (immune) cells during early life in ethnically admixed children. Using a novel statistical
67 inference method that provides a general framework for analyzing longitudinal genetic and
68 epigenetic data, we show that ancestry-dependent methylation patterns are conserved over the
69 first 7 years of life and that these patterns are strongly influenced, and often mediated, by local
70 genotype. Further, chronological age, but not measured exposures during pre- or post-natal

71 periods or disease status by age 7, was associated with methylation patterns in this sample of
72 children. Finally, we show that the "epigenetic clock"[10] present during pregnancy is the same
73 clock that is present after birth. Together with published data, we suggest that genetic ancestry
74 contributes to the epigenetic prediction of age at all stages of the life cycle, beginning *in utero*.
75

76 ## Materials and methods

77 **Sample composition**

78 URECA is a birth cohort study initiated in 2005 in Baltimore, Boston, New York City and St.
79 Louis under the NIAID-funded Inner City Asthma Consortium[29]. Pregnant women were
80 recruited. Either they or the father of their unborn child had a history of asthma, allergic rhinitis,
81 or eczema, and deliveries prior to 34 weeks gestation were excluded (see Gern et al.[29] for full
82 entry criteria). Informed consent was obtained from the women at enrollment and from the
83 parent or legal guardian of the infant after birth.

84 Maternal questionnaires were administered prenatally and child health questionnaires
85 administered to a parent or caregiver every 3 months through age 7 years. Gestational age at
86 birth and obstetric history were obtained from medical records. Additional details on study
87 design are described in Gern et al.[29] and in the Supplement. Frozen paired cord blood
88 mononuclear cells (CBMCs) and peripheral blood mononuclear cells (PBMCs) at age 7, were
89 available for 196 of the 560 URECA children after completing other studies. After QC (see
90 below), DNA methylation data were available for 194 children at birth, 195 children at age 7,
91 and 193 children at both time points; genotype data were available in 193 children (194 at birth;
92 195 at age 7) (Supplementary Table 1).
93

94 **DNA Methylation**

95 DNA for methylation studies was extracted from thawed CBMCs and PBMCs using the Qiagen
96 AllPrep kit (QIAGEN, Valencia, CA). Genome-wide DNA methylation was assessed using the
97 Illumina Infinium MethylationEPIC BeadChip (Illumina, San Diego, CA) at the University of
98 Chicago Functional Genomics Facility (UC-FGF). Birth and 7-year samples from the same child
99 were assayed on the same chip and the data were processed using Minfi[31]; Infinium type I and
100 type II probe bias were corrected using SWAN[32]. Raw probe values were corrected for color
101 imbalance and background by control normalization. Three out of the 392 samples (two at birth
102 and one at age 7) were removed as outliers following normalization. We removed 82,352 probes
103 that mapped either to the sex chromosomes or to more than one location in a bisulfite-converted
104 genome, had detection *P* values greater than 0.01% in 25% or more of the samples, or
105 overlapped with known SNPs with minor allele frequency of at least 5% in African, American or
106 European populations. After processing, 784,484 probes were retained and M-values were used
107 for all downstream analyses, which were computed as $\log_2$ (methylated intensity + 100) - $\log_2$
108 (unmethylated intensity + 100). The offset of 100 was recommended in Du et al.[33].

109

110 **Genotyping**

111 DNA from the 196 URECA children was genotyped with the Illumina Infinium
112 CoreExome+Custom array. Of the 532,992 autosomal SNPs on the array, 531,755 passed
113 Quality control (QC) (excluding SNPs with call rate <95%, Hardy-Weinberg *P* values $<10^{-5}$, and
114 heterozygosity outliers). We conducted all analyses in 293,696 autosomal SNPs with a minor
115 allele frequency ≥5%. Genotypes for three children failed QC and were excluded from
116 subsequent analysis that involved genotypes, including methylation quantitative locus (meQTL)

117 mapping, inferred genetic ancestry, or used genetic ancestry PC1 as a covariate (see below).
118 These three children were included in all other analyses.
119
120 **Estimating inferred genetic ancestry**
121 Ancestral principal component analysis (PCA) was performed using a set of 801 ancestry
122 informative markers (AIMs) from Tandon et al.[34] that were genotyped in both the URECA
123 children and in HapMap[35] release 23. Because PC1 captured the majority of variation in genetic
124 ancestry (Figure 1), we refer to PC1 as inferred genetic ancestry and consider it as a surrogate
125 measure for percent African ancestry.
126
127 **Statistical analysis**
128 To determine the effect of gestational age on methylation in CBMCs, we used standard linear
129 regression models with the child's gender, recruitment city, inferred genetic ancestry and
130 methylation plate number as covariates in our model. We also estimated cell composition and
131 other unobserved confounding factors using a method described in McKennan et al.[36]. We then
132 computed a gestational age *P* value for each CpG site and used q-values[37] to control the false
133 discovery rate at a nominal level. We took the same approach to determine CpGs whose
134 methylation changed from birth to age 7, except the response was measured as the difference in
135 methylation at birth and age 7. In this analysis, we included the child's gender, gestational age at
136 birth, inferred genetic ancestry and recruitment site as covariates. Because all paired samples
137 were on the same plate, we did not include plate number as a covariate in this analysis. We also
138 estimated unobserved factors that influence differences in methylation at birth and age 7 using

139 McKennan et al.[36] and included these latent factors in our linear model. See models (S1) and
140 (S2) in the Supplement for more detail.

141 We used data from the self-reported Hispanic and Black individuals with methylation
142 measured at both time points to analyze the effect of ancestry (either inferred genetic ancestry or
143 self-reported race) on methylation at birth and age 7 jointly using a Bayesian model. We did not
144 include the 10 individuals of other reported races in this analysis because we did not want our
145 estimates to be influenced by the groups with small samples sizes. We included age (birth or age
146 7), recruitment site, gestational age at birth, gender and methylation plate number as covariates
147 in our model, and estimated additional unobserved covariates (including cell composition) using
148 a method specifically designed for correlated data[36]. Once we estimated the relevant hyper-
149 parameters, we extended the sign rate paradigm developed in Stephens[38] to perform inference in
150 longitudinal data. This is discussed in more detail in the context of the specific questions we
151 present in the results section; the mathematical formulation is presented in the Supplement.

152 We performed meQTL mapping in the 145 genotyped, self-reported Black children using
153 the set of 269,622 SNPs with 100% genotype call rate in this subset. We restricted ourselves to
154 this subset of samples to minimize heterogeneity in effect sizes. To identify CpG-SNP pairs, we
155 considered SNPs within 5kb of each CpG, as this region has been previously shown to contain
156 the majority of genetic variability in DNA methylation[1], and computed a *P* value for the effect of
157 the genotype at a single SNP on methylation at the corresponding CpG with ordinary least
158 squares. We then defined the meQTL for each CpG site as the SNP with the lowest *P* value. In
159 addition to genotype, we included inferred genetic ancestry (i.e., ancestry PC1), gestational age
160 at birth, gender, recruitment city and methylation plate number in the linear model, along with
161 the first nine principal components of the residual methylation data matrix after regressing out

162 the intercept and the five additional covariates. We then tested the null hypothesis that a CpG did
163 not have an meQTL in the 10kb region by using the minimum marginal *P* value in the region as
164 the test statistic and computed its significance via bootstrap. Lastly, we used q-values to control
165 the false discovery rate.
166
167 **Results**
168 Our study included 196 children participants in the Urban Environment and Childhood Asthma
169 (URECA) cohort who had stored CBMCs and PBMCs collected at birth and age 7,
170 respectively[29], and passed quality control (QC) checks as described in Methods. The URECA
171 children were classified by parent- or guardian-reported race into one of the following
172 categories: Black, n =147; Hispanic, n = 39; White, n = 1; Mixed race n = 7, and Other, n = 2. A
173 description of the study population is shown in Table 1 and in Supplementary Materials.
174 Ancestry, assessed using ancestral PCs, revealed varying proportions of African and European
175 ancestry along PC1, which we refer to as inferred genetic ancestry (IGA); there is little
176 separation along PC2 (Figure 1). The reported races (RR) of the children are also shown in
177 Figure 1. The means and ranges of gestational age stratified by reported race are shown in Table
178 1; the distribution of gestational age at birth is shown in Figure S1 in the supplement.
179
180 **The "epigenetic clock" *in utero* is the same "epigenetic clock" after birth**
181 DNA methylation patterns change at many CpG sites across the genome as humans age[9; 13; 14; 39]
182 and an individual's epigenetic age may be more associated with their health and disease states
183 than chronological age[11; 15]. Two epigenetic age predictors have been proposed: one after birth[10]
184 and one prior to birth[13]. However, it is not known how the "epigenetic clock" present *in utero*

185 (measured at birth) is related to the epigenetic clock present later in life, and we therefore sought
186 to determine whether or not the same CpGs determine one's epigenetic age *in utero* and at age 7.
187 Among the 784,484 CpG sites in CBMCs that passed QC, the methylation levels at 2,836 CpGs
188 were significantly correlated with gestational age at an FDR of 5%. Methylation at 62% of these
189 sites decreased with increasing gestational age (P value for trend $< 10^{-10}$). We did not find any
190 significant associations between gestational age and methylation at age 7, indicating that
191 gestational age within the ranges available has little to no effect on methylation patterns later in
192 life.

193 To determine whether methylation levels at gestational age-associated CpG sites
194 continued to change between birth and age 7, we next identified a set of 16,172 CpG sites whose
195 methylation levels changed between these ages at an FDR of 5% (see model (S2) in the
196 Supplement for more detail). The 16,172 age-related CpGs included 1,986 (70%) of the 2,836
197 gestational age-associated CpGs. Thus, the gestational age-associated CpGs are largely a subset
198 of the CpGs that changed from birth to age 7. Furthermore, the estimates of age effects among
199 the CpGs that changed from birth to age 7 showed the same direction of change as the estimated
200 gestational age effects in 97% of the 16,172 CpGs, which included 14,186 gestational age-
201 associated effects that were not significant at the 5% FDR threshold. This concordance in
202 direction of effect is unlikely to occur by chance ($P < 10^{-10}$, see Supplement for calculation) and
203 provides strong evidence that CpGs whose methylation decreases with increasing gestational
204 ages *in utero* continue to decrease after birth (and vice-versa if the methylation increases *in*
205 *utero*).

206 To further validate that the same CpGs determine one's epigenetic age both *in utero* and
207 after birth, we investigated whether or not our gestational age- and age-related CpGs were

208 enriched for the 353 CpGs reported by Horvath[10] to predict age after birth using DNA
209 methylation patterns across many human tissues, as well as for the 148 CpGs used by Knight et
210 al.[13] to predict gestational age from cord blood DNA methylation. Even though only 6 CpGs
211 overlapped between the sets of 353 and 148 CpGs in each study, our gestational age and age-
212 related CpGs are highly enriched for both sets of CpGs from the Knight et al. and Horvath
213 studies (Figure 2).
214
215 **Inferred genetic ancestry effects on DNA methylation patterns are conserved in magnitude**
216 **and direction between birth and age 7**
217 Previous cross-sectional studies have revealed associations between ancestry and DNA
218 methylation at birth[19; 23] and later in life[20-22; 24; 25]. These correlations were generally attributed to
219 the combined effects of genetic variation and environmental exposures[20-23]. However, because of
220 the cross-sectional nature of these studies, it is not known if the association between ancestry and
221 methylation patterns present at birth persist (or change) in childhood. Moreover, because
222 ancestry is typically confounded with environmental exposures[40], it has been proposed that
223 ancestry effects on methylation levels may reflect the effects of exposure histories, which also
224 may vary by race or ethnicity[20]. Alternatively, ancestry effects on DNA methylation patterns
225 could also be due to genetic differences. In these cases, we would expect ancestry-associated
226 methylation patterns to be conserved from birth to later childhood. Using the longitudinal data
227 from the URECA cohort, we tested this hypothesis by addressing three questions. What is the
228 correlation between ancestry and methylation levels at individual CpG sites at birth and age 7? Is
229 the direction and magnitude of the correlation between ancestry and methylation levels
230 conserved between birth and age 7? Are there any CpGs for which the correlation between

231 methylation and ancestry at birth is significantly different from the correlation between
232 methylation and ancestry at age 7?
233 Standard hypothesis testing can be used to answer the first question but is not appropriate
234 for answering the second or third because failure to reject the null hypothesis that the effects are
235 equal at birth and age 7 does not imply the null hypothesis is true. Additionally, because our
236 studies were conducted in cord blood cells at birth and peripheral blood cells at age 7, effects at
237 birth and age 7 may differ due to differences in cell composition[41]. To circumvent these issues,
238 we built a Bayesian model (see model (S3) in the Supplement) and let the data determine both
239 the strength of the correlation between inferred genetic ancestry or reported race and
240 methylation, and how similar the correlations were at birth and age 7. We then answered the
241 first, second and third questions by defining and estimating the correct (*cor*), conserved (*con*)
242 and discordant (*dis*) sign rates for each CpG $g = 1,\ldots, 784{,}484$:
243
244 $cor_g^{(0)}$ = Posterior probability that the estimate for the direction of CpG *g*'s ancestry effect
245 at birth was correct
246 $cor_g^{(0,7)}$ = Posterior probability that the directions of CpG *g*'s ancestry effects at birth and
247 age 7 were the same AND the directions were estimated correctly
248 $dis_g^{(0)}$ = Posterior probability that the ancestry effect for CpG *g* at birth was non-zero and
249 was 0 or in the opposite direction at age 7 AND both directions were estimated
250 correctly
251
252 The correct and discordant sign rates at age 7 ($cor_g^{(7)}$, $dis_g^{(7)}$) were defined as follows. The
253 correct sign rate at birth and age 7 is always at least as large as the conserved sign rate. That is, if

254 a CpG has a significant conserved sign rate, the correct sign rates at birth and age 7 will also be
255 significant. We therefore say that the birth and age 7 effects for CpG *g* overlap if its conserved
256 sign rate ($con_g^{(0,7)}$) is above a designated threshold. Supplemental Figures S2 and S3 provide
257 insight into how the correct and conserved sign rates compare with standard univariate *P* values.
258 We refer the reader to "Joint modeling of methylation at birth and seven" in the Supplement for a
259 detailed description of our model and estimation procedure.

260 After fitting the relevant parameters in the model to the data, we were able to estimate the
261 fraction of CpGs with non-zero effects at both ages that fell into one of four possible bins: the
262 two effects were completely unrelated ($\rho = 0$), moderately similar ($\rho = 1/3$), very similar ($\rho =$
263 $2/3$), or identical ($\rho = 1$). Note that if a non-trivial fraction of CpG sites had effects that were
264 negatively related, they would be assigned to the first bin ($\rho = 0$). In fact, less than 1% of the
265 CpGs with non-zero inferred genetic ancestry effects at both ages had unrelated or moderately
266 similar IGA effects ($\rho = 0$ or $1/3$), whereas 34% fell in the very similar bin and 66% had
267 identical inferred genetic ancestry effects at birth and age 7 (Supplemental Figure S4). This
268 indicates that when inferred genetic ancestry effects on methylation are non-zero at both birth
269 and age 7, they tend to be very similar or exactly the same with respect to both direction and
270 magnitude.

271 We then estimated the correct and conserved sign rates for all 784,484 probes, and
272 identified 2,873 inferred genetic ancestry-associated CpGs (IGA-CpGs) in CBMCs ($cor_g^{(0)} \geq$
273 0.95), 3,834 in PBMCs at age 7 ($cor_g^{(7)} \geq 0.95$), and 2,659 whose effects were conserved in sign
274 ($cor_g^{(0,7)} \geq 0.95$). Methylation tended to increase with increasing African ancestry at 1,494 of the
275 2,659 conserved CpGs (56%), suggesting that individuals with more African ancestry tend to
276 have more methylation ($P < 10^{-10}$). This is consistent with the study of Moen et al.[22], which used

277 the Illumina 450K array to quantify the differences in methylation between European and
278 African populations. Supplemental Figure S5 shows the IGA-CpG locations in the genome and
279 Figure 3a illustrates the overlap between IGA-CpGs at birth and age 7. This strong overlap
280 corroborates our above observations and answers the second question in the affirmative: if
281 methylation is strongly correlated with inferred genetic ancestry at birth, the magnitude and
282 direction of the correlation is conserved at age 7.
283
284 **Inferred genetic ancestry is more correlated with methylation than is self-reported race**
285 The observed correlations between ancestry and methylation levels may reflect differences in
286 environmental exposures[20; 22], due to associations between race or ethnicity with socio-cultural,
287 nutritional, and geographic exposures, among others[40]. In fact, Galanter et al.[20] showed in a
288 cross-sectional study that self-reported ethnicity explained a substantial portion of the variability
289 in whole blood DNA methylation patterns from Latino children of diverse ethnicities, even more
290 so than inferred genetic ancestry. They concluded that ethnicity captures genetic, as well as the
291 socio-cultural and environmental differences that influence methylation levels. If this were the
292 case in the URECA children, we should observe just as large, if not larger, an effect as we did for
293 inferred genetic ancestry when we substitute reported race for inferred genetic ancestry in the
294 analyses presented in the previous section. However, using reported race, we identified only 457
295 CpGs at birth and 709 CpGs at age 7 whose correct sign rate was at least 0.95, and 424 whose
296 reported effects were conserved from birth to age 7 ($cor_g^{(0,7)} \geq 0.95$), far fewer than the 3,991
297 CpGs whose methylation was significantly correlated with inferred genetic ancestry at birth or at
298 age 7.

299 To explore this further, we examined the overlap between IGA-CpGs and reported race-
300 associated CpGs (RR-CpGs) in CBMCs at birth and in PBMCs at 7 (Figures 3c-d). Because
301 reported race is an estimate of inferred genetic ancestry, there is a still substantial overlap
302 between IGA-CpGs and RR-CpGs. In fact, almost all of the RR-CpGs are among the IGA-CpGs,
303 but the opposite is not true. This indicates that while IGA-CpGs include most RR-CpGs, reported
304 race does not capture most of the variation in methylation attributable to genetic ancestry in these
305 children.
306
307 **The observed correlations between DNA methylation and ancestry are primarily genetic**
308 Both age and ancestry effects on methylation could be due to genetic variation or to
309 environmental exposures that differ throughout life or by ethnicity. To address this question, we
310 used local genetic variation (within 5kb of a CpG site) and DNA methylation data at birth and
311 age 7 in the 147 self-reported Black children in our study. Of the 519,622 CpGs within 5kb of a
312 SNP, 65,068 and 70,898 had at least one meQTL in CBMCs at birth and in PBMCs at age 7,
313 respectively, at an FDR of 5%. In addition, 59% of IGA-CpGs with at least one SNP in the ±5kb
314 window had at least one meQTL at birth or age 7 at an FDR of 5%, indicating IGA-CpGs were
315 enriched for CpGs with meQTLs (Figure 4a-b).
316 To provide further evidence that local genotype mediates the effect of inferred genetic
317 ancestry on methylation, we used logistic regression to regress the genotype of each of the
318 269,622 SNPs in our study set onto inferred genetic ancestry. The goal was to determine the
319 fraction of IGA-CpGs that were mediated through local genotype, i.e. IGA-CpGs with both
320 edges a and c in Figure 4a. Our analysis first revealed that the genotypes at meQTLs whose
321 target CpGs were IGA-CpGs in either CBMCs at birth or PBMCs at age 7 (IGA-meQTLs) were

322 significantly more correlated with inferred genetic ancestry than the genotype of non-IGA-
323 meQTLs (Figure 4c). Moreover, approximately 13% of the IGA-CpGs with at least one SNP in
324 their ±5kb windows had an inferred genetic ancestry effect that was mediated through local
325 genotype (i.e. had edges a and c, see Supplement for calculation details), which is likely an
326 underestimate of the true number IGA-CpGs mediated through local genotype because our
327 sample size was relatively small[42]. Nonetheless, this is striking compared to the 0.1% of non-
328 IGA-CpGs whose corresponding SNP had edge c at a 20% FDR.

329 Lastly, we used DNA methylation data on 573 ethnically diverse U.S. Latino children
330 ages 9 to 16 years old from the Galanter et al. study[20] to further explore the effect of ancestry on
331 DNA methylation in whole blood. Children and teenagers in the Galanter study were classified
332 as Mexican, Puerto Rican, mixed Latino, or other Latino. They reported 916 CpG sites whose
333 methylation was significantly associated with reported race, 773 of which were also significantly
334 associated with estimated percent European, Native American and African ancestry at a
335 Bonferroni *P* value threshold of 1.6×10^{-7}. A total of 726 of their 916 ethnic-associated CpGs
336 were also in the set of 784,484 probes CpG sites in our study. Our set of IGA-CpGs at birth or
337 age 7 contained a significant fraction of the 726 ethnic-associated CpGs from the cross-sectional
338 study, but there was considerably less overlap with our RR-CpGs (Figure 5a-b). If the correlation
339 we observed between ancestry and methylation was largely due to responses to environmental
340 exposures, as suggested in the Galanter study, then the overlap with reported race should have
341 been at least as large as the overlap with inferred genetic ancestry. That was not the case, further
342 suggesting that the inferred genetic ancestry effects on methylation in the URECA cohort are
343 primarily genetic in origin.
344

345 **Age-related CpGs overlap with ancestry-associated CpGs that are also genetically**
346 **regulated**
347 Recent work has suggested that race and ethnicity may play a role in determining one's
348 epigenetic age[13; 14]. However, it has not been determined if there is a genetic basis for this
349 phenomenon or if differences arise solely due different environmental exposures that are
350 confounded with ethnicity. To investigate this, we examined the overlap between age-related
351 CpGs identified in our study and the 263 ethnic-associated CpGs identified in the Galanter study
352 that also had an meQTL at a 5% FDR. Figure 5c-d illustrates that these 263 ethnic-associated
353 meQTLs are enriched for age-related CpGs in our study, suggesting that the previously observed
354 discrepancy in epigenetic age between ethnic groups may in part be due to genetic differences
355 between the groups.
356
357 **Non-genetic factors influence the observed ancestry-methylation correlation more at age 7**
358 **than at birth**
359 Although most of the variation in methylation levels at ancestry-associated CpGs can be
360 attributed to genetic variation in the URECA children, a small proportion may be due to non-
361 genetic (environmental) factors. To explore this possibility, we further hypothesized that non-
362 genetic effects on methylation levels at ancestry-associated CpGs would be greater at age 7 than
363 at birth, due to accumulated exposures over the first 7 years of life. We note that none of the
364 direct or indirect measures of exposures that were available in this cohort were associated with
365 methylation levels at either age, including maternal asthma, maternal infections during
366 pregnancy, pet ownership, bedroom allergens, mother stress, anxiety and depression metrics,
367 maternal cotinine levels during pregnancy, number of smokers in the household, number of

368 siblings, number of previous live births, daycare attendance, number of colds at age 2 or 3, and
369 allergic sensitization or asthma in the child (see Supplement for details). Therefore, if exposure
370 histories influenced methylation levels in these children, it is not likely due to any of these
371 measured variables.

372 To directly test the hypothesis that non-genetic factors tend to have larger effects on
373 methylation levels at age 7 than at birth, we used our Bayesian model to estimate the proportion
374 of CpGs in our study whose methylation was not associated with ancestry at birth but associated
375 at age seven and the proportion that were associated with ancestry at birth but not at age 7. The
376 former was greater than 14% while the latter was less than 1.5% using either inferred genetic
377 ancestry or reported race as a measure of ancestry. Even though over 14% of all CpGs in our
378 study had ancestry effects present at age 7 but not at birth, we were only able to identify 18
379 discordant IGA-CpGs and 4 discordant RR-CpGs at age 7 using a liberal threshold of $dis_g^{(7)} \geq$
380 0.80. That is, for almost all CpGs that are associated with ancestry at age 7 but not at birth, the
381 expected ancestry effect sizes were quite small relative to the statistical error, making it
382 impossible to assign the direction of effect on methylation changes with confidence (Figure S6).
383 Therefore, while some CpG sites may be influenced by exposures that are correlated with
384 ancestry at age 7 but not at birth, their effects were far too small to estimate in this sample size.
385

386 **Discussion**

387 The relationships between DNA methylation, chronological age, and ancestry have the potential
388 to shed light on disease etiology and may help determine the relative genetic and environmental
389 contributions to the observed inter-individual variability of the epigenome[9-15; 19-24]. For example,
390 it is well known that the methylome undergoes global changes as humans age[9; 10; 12], and

391 accelerated aging (the difference in epigenetic age and chronological age) is related to a variety
392 of disease phenotypes in adults[11; 15; 39], and with pregnancy complications, as well as with
393 developmental and stress-related phenotypes and birthweight in infants[42-44]. However, it has not
394 previously been shown that the same CpGs influence methylation age both prior to and after
395 birth. Our results suggest that not only is the collection of age-related CpGs similar pre- and
396 post-natally, but that the direction of age-related change in methylation levels is consistent before
397 and after birth.

398 Our discovery of the strict monotonicity of the change in methylation levels from birth to
399 age 7 is particularly interesting because it suggests the biological mechanism(s) involved with
400 advancing methylation age begin(s) *in utero* and remain(s) active at least to age 7. This provides
401 further support for and mechanistic insights into hypotheses on the fetal origins of adult
402 diseases[45; 46]. Moreover, because methylation age is an aggregate measure of methylation levels
403 across the genome, using this metric could reveal associations with phenotypes of interest even
404 when correlations with individual CpG sites are weak. Our results also suggest that including
405 CpGs that also predict epigenetic age *in utero* may provide more accurate measurements of
406 accelerated aging in adults.

407 While there was substantial change in blood methylation levels over time among children
408 in this cohort, inferred genetic ancestry and self-reported race effects on methylation are
409 overwhelmingly conserved in both direction and magnitude from birth to age 7, indicating that
410 variation in methylation patterns in blood cells are robust to changes in environment and are
411 consistent through early childhood. This result is interesting in and of itself because it provides
412 an example of fetal epigenetic variation that persists later in life, and more generally an example
413 of a persistent effect on DNA methylation levels, which has been cited as a critical area of future

414 epigenetic research[47; 48]. Furthermore, the novel statistical framework we used to infer effects
415 that are conserved versus those that vary over time can be easily applied to other longitudinal
416 DNA methylation data, as a way to avoid the spurious logic often used in applications of
417 frequentist hypothesis testing that failing to reject the null hypothesis implies the null is true.

418 The observation that inferred genetic ancestry and reported race effects are conserved
419 from birth to age 7 does not necessarily mean that the observed effects are primarily genetic in
420 nature because genetic and environmental components, or their interactions, could determine
421 ancestry-related methylation *in utero* and those patterns could persist as the child ages. To
422 further explore this, we showed that the IGA-CpGs were enriched among CpGs with meQTLs,
423 and that methylation levels at many of the IGA-CpGs are mediated by local genotype, indicating
424 that much of the ancestry-methylation correlation could be attributed to genetic variation.
425 Moreover, the RR-CpGs were only a small subset of IGA-CpGs in our study. This is opposite to
426 the findings of Galanter et al. [20], who argued that ancestry-dependent methylation patterns in
427 admixed populations are in large part determined by differences in exposure histories. Because
428 their data were cross-sectional they could not evaluate whether the observed patterns arose
429 during childhood or were also present at birth. While our results only provide evidence for
430 genetics accounting for most of the correlation between methylation and ancestry, they imply
431 that the genetic contribution to variability in blood methylation is substantial.

432 Our observations in support of strong genetically – and weak environmentally –
433 determined ancestry-associated methylation patterns in blood may seem paradoxical to the
434 plethora of studies showing that DNA methylation levels in blood cells are associated with
435 environmental exposures, such as cadmium, arsenic and smoking, to name a few[5-8; 20; 49-53].
436 Whereas the estimated genetic effect sizes in our study are substantially larger than many of the

437 environmentally-associated effects on methylation patterns previously reported, the effects of
438 environmental exposures on methylation in blood is probably too small to estimate with even
439 moderate to large sample sizes[45]. For example, it was only by performing a meta-analysis in
440 6,685 individuals that Joubert et al.[5] were able to identify 6,000 CpGs whose DNA methylation
441 levels in blood from infants and adolescents were associated maternal smoking exposure. In one
442 sense, we were able to corroborate previous observations of small non-genetic effects on
443 methylation in blood by showing that while methylation patterns at an estimated 14% of all
444 CpGs in our study were not correlated with ancestry at birth but correlated with ancestry at age 7,
445 the correlation at individual CpGs at age 7 was too small to be identified as statistically
446 significant. However, we were not able to find any statistically significant correlations between
447 methylation at birth or at age 7 and any of the environmental exposure variables measured in
448 blood cells from this cohort. We note that cord blood cotinine levels, a measure of in-utero
449 tobacco smoke exposure, were above the level of detection in only 34 of the 196 mothers in our
450 study.

451 The combined results of our study and those cited above suggest that DNA methylation
452 levels in blood cells are fairly robust to environmental perturbations. A better understanding of
453 tissue-specific methylation responses to environmental exposures could inform the design of
454 future studies and provide insights into the mechanisms through which exposures and gene-
455 environment interactions influence health and disease.

456

457 **Acknowledgements**

458 This work was supported in part by NIH grants U19 AI106683, R01 HL129735, R01 HL122712,
459 and P01 HL070831. The URECA study has been funded by the National Institute of Allergy and

460 Infectious Diseases, National Institutes of Health, under Contract numbers NO1-AI-25496, NO1-
461 AI-25482, HHSN272200900052C, HHSN272201000052I, 1UM1AI114271-01 and
462 UM2AI117870. Additional support was provided by the National Center for Research
463 Resources, National Institutes of Health, under grants RR00052, M01RR00533,
464 1UL1RR025771, M01RR00071, 1UL1RR024156, UL1TR000040, UL1TR001079 and
465 5UL1RR024992-02.

466 **References**

467 1. Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y., and
468 Pritchard, J.K. (2011). DNA methylation patterns associate with genetic and gene expression
469 variation in HapMap cell lines. Genome Biol 12, R10.

470 2. Smith, A.K., Kilaru, V., Kocak, M., Almli, L.M., Mercer, K.B., Ressler, K.J., Tylavsky, F.A.,
471 and Conneely, K.N. (2014). Methylation quantitative trait loci (meQTLs) are consistently
472 detected across ancestry, developmental stage, and tissue type. BMC Genomics 15, 145.

473 3. Chatterton, Z., Hartley, B.J., Seok, M.H., Mendelev, N., Chen, S., Milekic, M., Rosoklija, G.,
474 Stankov, A., Trencevsja-Ivanovska, I., Brennand, K., et al. (2017). In utero exposure to
475 maternal smoking is associated with DNA methylation alterations and reduced neuronal
476 content in the developing fetal brain. Epigenetics Chromatin 10, 4.

477 4. Goodrich, J.M., Dolinoy, D.C., Sanchez, B.N., Zhang, Z., Meeker, J.D., Mercado-Garcia, A.,
478 Solano-Gonzalez, M., Hu, H., Tellez-Rojo, M.M., and Peterson, K.E. (2016). Adolescent
479 epigenetic profiles and environmental exposures from early life through peri-adolescence.
480 Environ Epigenet 2, dvw018.

481 5. Joubert, B.R., Felix, J.F., Yousefi, P., Bakulski, K.M., Just, A.C., Breton, C., Reese, S.E.,
482 Markunas, C.A., Richmond, R.C., Xu, C.J., et al. (2016). DNA Methylation in Newborns and
483 Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. Am J Hum
484 Genet 98, 680-696.

485 6. Kippler, M., Engstrom, K., Mlakar, S.J., Bottai, M., Ahmed, S., Hossain, M.B., Raqib, R.,
486 Vahter, M., and Broberg, K. (2013). Sex-specific effects of early life cadmium exposure on
487 DNA methylation and implications for birth weight. Epigenetics 8, 494-503.

488 7. Koestler, D.C., Avissar-Whiting, M., Houseman, E.A., Karagas, M.R., and Marsit, C.J.
489 (2013). Differential DNA methylation in umbilical cord blood of infants exposed to low
490 levels of arsenic in utero. Environ Health Perspect 121, 971-977.

491 8. Rzehak, P., Saffery, R., Reischl, E., Covic, M., Wahl, S., Grote, V., Xhonneux, A.,
492 Langhendries, J.P., Ferre, N., Closa-Monasterolo, R., et al. (2016). Maternal Smoking during
493 Pregnancy and DNA-Methylation in Children at Age 5.5 Years: Epigenome-Wide-Analysis
494 in the European Childhood Obesity Project (CHOP)-Study. PLoS One 11, e0155554.

495 9. Bocklandt, S., Lin, W., Sehl, M.E., Sanchez, F.J., Sinsheimer, J.S., Horvath, S., and Vilain, E.
496 (2011). Epigenetic predictor of age. PLoS One 6, e14821.

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

499 11. Horvath, S., Erhart, W., Brosch, M., Ammerpohl, O., von Schonfels, W., Ahrens, M., Heits,
500 N., Bell, J.T., Tsai, P.C., Spector, T.D., et al. (2014). Obesity accelerates epigenetic aging of
501 human liver. Proc Natl Acad Sci U S A 111, 15538-15543.

502 12. Johnson, A.A., Akman, K., Calimport, S.R., Wuttke, D., Stolzing, A., and de Magalhaes, J.P.
503 (2012). The role of DNA methylation in aging, rejuvenation, and age-related disease.
504 Rejuvenation Res 15, 483-494.

505 13. Knight, A.K., Craig, J.M., Theda, C., Baekvad-Hansen, M., Bybjerg-Grauholm, J., Hansen,
506 C.S., Hollegaard, M.V., Hougaard, D.M., Mortensen, P.B., Weinsheimer, S.M., et al. (2016).
507 An epigenetic clock for gestational age at birth based on blood methylation data. Genome
508 Biol 17, 206.

509 14. Levine, M.E., and Crimmins, E.M. (2014). Evidence of accelerated aging among African
510 Americans and its implications for mortality. Soc Sci Med 118, 27-32.

511 15. Marioni, R.E., Shah, S., McRae, A.F., Chen, B.H., Colicino, E., Harris, S.E., Gibson, J.,
512 Henders, A.K., Redmond, P., Cox, S.R., et al. (2015). DNA methylation age of blood
513 predicts all-cause mortality in later life. Genome Biol 16, 25.

514 16. Parets, S.E., Conneely, K.N., Kilaru, V., Fortunato, S.J., Syed, T.A., Saade, G., Smith, A.K.,
515 and Menon, R. (2013). Fetal DNA Methylation Associates with Early Spontaneous Preterm
516 Birth and Gestational Age. PLoS One 8, e67489.

517 17. Schroeder, J.W., Conneely, K.N., Cubells, J.C., Kilaru, V., Newport, D.J., Knight, B.T.,
518 Stowe, Z.N., Brennan, P.A., Krushkal, J., Tylavsky, F.A., et al. (2011). Neonatal DNA
519 methylation patterns associate with gestational age. Epigenetics 6, 1498-1504.

520 18. Simpkin, A.J., Suderman, M., Gaunt, T.R., Lyttleton, O., McArdle, W.L., Ring, S.M.,
521 Tilling, K., Davey Smith, G., and Relton, C.L. (2015). Longitudinal analysis of DNA
522 methylation associated with birth weight and gestational age. Hum Mol Genet 24, 3752-
523 3763.

524 19. Adkins, R.M., Krushkal, J., Tylavsky, F.A., and Thomas, F. (2011). Racial differences in
525 gene-specific DNA methylation levels are present at birth. Birth Defects Res A Clin Mol
526 Teratol 91, 728-736.

527 20. Galanter, J.M., Gignoux, C.R., Oh, S.S., Torgerson, D., Pino-Yanes, M., Thakur, N., Eng, C.,
528 Hu, D., Huntsman, S., Farber, H.J., et al. (2017). Differential methylation between ethnic
529 sub-groups reflects the effect of genetic ancestry and environmental exposures. Elife 6.

530 21. Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., Monk, D.,
531 Hata, K., Marques-Bonet, T., Wang, L., et al. (2013). DNA methylation contributes to natural
532 human variation. Genome Res 23, 1363-1372.

533 22. Moen, E.L., Zhang, X., Mu, W., Delaney, S.M., Wing, C., McQuade, J., Myers, J., Godley,
534 L.A., Dolan, M.E., and Zhang, W. (2013). Genome-wide variation of cytosine modifications
535 between European and African populations and the implications for complex traits. Genetics
536 194, 987-996.

537 23. Mozhui, K., Smith, A.K., and Tylavsky, F.A. (2015). Ancestry dependent DNA methylation
538 and influence of maternal nutrition. PLoS One 10, e0118466.

539 24. Rahmani, E., Shenhav, L., Schweiger, R., Yousefi, P., Huen, K., Eskenazi, B., Eng, C.,
540 Huntsman, S., Hu, D., Galanter, J., et al. (2017). Genome-wide methylation data mirror
541 ancestry information. Epigenetics Chromatin 10, 1.

542 25. Chan, M.A., Ciaccio, C.E., Gigliotti, N.M., Rezaiekhaligh, M., Siedlik, J.A., Kennedy, K.,
543 and Barnes, C.S. (2017). DNA methylation levels associated with race and childhood asthma
544 severity. J Asthma 54, 825-832.

545 26. Ladd-Acosta, C., Hansen, K.D., Briem, E., Fallin, M.D., Kaufmann, W.E., and Feinberg,
546 A.P. (2014). Common DNA methylation alterations in multiple brain regions in autism. Mol
547 Psychiatry 19, 862-871.

548 27. Nicodemus-Johnson, J., Myers, R.A., Sakabe, N.J., Sobreira, D.R., Hogarth, D.K.,
549 Naureckas, E.T., Sperling, A.I., Solway, J., White, S.R., Nobrega, M.A., et al. (2016). DNA
550 methylation in lung cells is associated with asthma endotypes and genetic risk. JCI Insight 1,
551 e90151.

552 28. Yokoyama, A.S., Rutledge, J.C., and Medici, V. (2017). DNA methylation alterations in
553 Alzheimer's disease. Environ Epigenet 3, dvx008.

554 29. Gern, J.E., Visness, C.M., Gergen, P.J., Wood, R.A., Bloomberg, G.R., O'Connor, G.T.,
555 Kattan, M., Sampson, H.A., Witter, F.R., Sandel, M.T., et al. (2009). The Urban
556 Environment and Childhood Asthma (URECA) birth cohort study: design, methods, and
557 study population. BMC Pulm Med 9, 17.

558 30. O'Connor, G.T., Lynch, S.V., Bloomberg, G.R., Kattan, M., Wood, R.A., Gergen, P.J.,
559 Jaffee, K.F., Calatroni, A., Bacharier, L.B., Beigelman, A., et al. (2017). Early-life home
560 environment and risk of asthma among inner-city children. J Allergy Clin Immunol.

561 31. Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D.,
562 and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the
563 analysis of Infinium DNA methylation microarrays. Bioinformatics 30, 1363-1369.

564 32. Maksimovic, J., Gordon, L., and Oshlack, A. (2012). SWAN: Subset-quantile within array
565 normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol 13,
566 R44.

567 33. Du, P., Zhang, X., Huang, C.C., Jafari, N., Kibbe, W.A., Hou, L., and Lin, S.M. (2010).
568 Comparison of Beta-value and M-value methods for quantifying methylation levels by
569 microarray analysis. BMC Bioinformatics 11, 587.

570 34. Tandon, A., Patterson, N., and Reich, D. (2011). Ancestry informative marker panels for
571 African Americans based on subsets of commercially available SNP arrays. Genet Epidemiol
572 35, 80-83.

573 35. International HapMap, C. (2003). The International HapMap Project. Nature 426, 789-796.

574 36. McKennan, C., and Nicolae, D.L. (2018). Accounting for unobserved covariates with varying
575 degrees of estimability in high dimensional biological data. arXiv 1801.00865y3.

576 37. Storey, J.D. (2002). A direct approach to false discovery rates. J Royal Stat Soc B 64, 479-
577 498.

578 38. Stephens, M. (2017). False discovery rates: a new deal. Biostatistics 18, 275-294.

579 39. Knight, A.K., Conneely, K.N., and Smith, A.K. (2017). Gestational age predicted by DNA
580 methylation: potential clinical and research utility. Epigenomics.

581 40. Nguyen, A.B., Moser, R., and Chou, W.Y. (2014). Race and health profiles in the United
582 States: an examination of the social gradient through the 2009 CHIS adult survey. Public
583 Health 128, 1076-1086.

584 41. Fu, J., Wolfs, M.G., Deelen, P., Westra, H.J., Fehrmann, R.S., Te Meerman, G.J., Buurman,
585 W.A., Rensen, S.S., Groen, H.J., Weersma, R.K., et al. (2012). Unraveling the regulatory
586 mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet
587 8, e1002431.

588 42. Fritz, M.S., and Mackinnon, D.P. (2007). Required sample size to detect the mediated effect.
589 Psychol Sci 18, 233-239.

590 43. Davis, E.G., Humphreys, K.L., McEwen, L.M., Sacchet, M.D., Camacho, M.C., MacIsaac,
591 J.L., Lin, D.T.S., Kobor, M.S., and Gotlib, I.H. (2017). Accelerated DNA methylation age in

592 adolescent girls: associations with elevated diurnal cortisol and reduced hippocampal
593 volume. Transl Psychiatry 7, e1223.

594 44. Simpkin, A.J., Howe, L.D., Tilling, K., Gaunt, T.R., Lyttleton, O., McArdle, W.L., Ring,
595 S.M., Horvath, S., Smith, G.D., and Relton, C.L. (2017). The epigenetic clock and physical
596 development during childhood and adolescence: longitudinal analysis from a UK birth
597 cohort. Int J Epidemiol 46, 549-558.

598 45. Almond, D., and Currie, J. (2011). Killing Me Softly: The Fetal Origins Hypothesis. J Econ
599 Perspect 25, 153-172.

600 46. Smith, C.J., and Ryckman, K.K. (2015). Epigenetic and developmental influences on the risk
601 of obesity, diabetes, and metabolic syndrome. Diabetes Metab Syndr Obes 8, 295-302.

602 47. Breton, C.V., Marsit, C.J., Faustman, E., Nadeau, K., Goodrich, J.M., Dolinoy, D.C.,
603 Herbstman, J., Holland, N., LaSalle, J.M., Schmidt, R., et al. (2017). Small-Magnitude Effect
604 Sizes in Epigenetic End Points are Important in Children's Environmental Health Studies:
605 The Children's Environmental Health and Disease Prevention Research Center's Epigenetics
606 Working Group. Environ Health Perspect 125, 511-526.

607 48. Martin, E.M., and Fry, R.C. (2018). Environmental Influences on the Epigenome: Exposure-
608 Associated DNA Methylation in Human Populations. Annu Rev Public Health.

609 49. Joubert, B.R., Haberg, S.E., Nilsen, R.M., Wang, X., Vollset, S.E., Murphy, S.K., Huang, Z.,
610 Hoyo, C., Midttun, O., Cupul-Uicab, L.A., et al. (2012). 450K epigenome-wide scan
611 identifies differential DNA methylation in newborns related to maternal smoking during
612 pregnancy. Environ Health Perspect 120, 1425-1431.

613 50. Lee, K.W., Richmond, R., Hu, P., French, L., Shin, J., Bourdon, C., Reischl, E.,
614 Waldenberger, M., Zeilinger, S., Gaunt, T., et al. (2015). Prenatal exposure to maternal
615 cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample
616 of adolescents and replication in an independent cohort at birth through 17 years of age.
617 Environ Health Perspect 123, 193-199.

618 51. Markunas, C.A., Xu, Z., Harlid, S., Wade, P.A., Lie, R.T., Taylor, J.A., and Wilcox, A.J.
619 (2014). Identification of DNA methylation changes in newborns related to maternal smoking
620 during pregnancy. Environ Health Perspect 122, 1147-1153.

621 52. Richmond, R.C., Simpkin, A.J., Woodward, G., Gaunt, T.R., Lyttleton, O., McArdle, W.L.,
622 Ring, S.M., Smith, A.D., Timpson, N.J., Tilling, K., et al. (2015). Prenatal exposure to
623 maternal smoking and offspring DNA methylation across the lifecourse: findings from the
624 Avon Longitudinal Study of Parents and Children (ALSPAC). Hum Mol Genet 24, 2201-
625 2217.

626 53. Wu, D., Yang, H., Winham, S.J., Natanzon, Y., Koestler, D.C., Luo, T., Fridley, B.L.,
627 Goode, E.L., Zhang, Y., and Cui, Y. (2018). Mediation analysis of alcohol consumption,
628 DNA methylation, and epithelial ovarian cancer. J Hum Genet 63, 339-348.
629

630

631 **Table 1**: Covariates for the 196 URECA children in our study, stratified by self-reported race.
632
633

	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]
Recruitment city					
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%)

634
635

636 **Figure Legends**

637

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

644 **Figure** 2: Distribution of *P* values for age (birth to age 7) (panels a and c) and gestational age
645 (GA) (panels b and d). Upper panels, stratified by whether or not the CpG was one of the
646 353 CpGs used to build the linear predictor of age in Horvath 2013[10] (red dots). The blue
647 dots are all other CpGs; the 10 enlarged red circles are for visual aid. Lower panels,
648 stratified by whether or not the CpG was one of the 148 CpGs used to build the linear
649 predictor of age in Knight et al.[13] (violet dots). The blue dots are all of the other CpGs; the
650 ten enlarged violet circles are just for visual aid.

651 **Figure** 3: Overlapping ancestry CpGs at birth and at age 7. (a): IGA-CpGs in CBMCs at birth
652 ($\widehat{cor}^{(0)} \geq 0.95$) and PBMCs at age 7 ($\widehat{cor}^{(7)} \geq 0.95$). Overlapping CpGs (violet) have
653 $\widehat{con}^{(0,7)} \geq 0.95$. (b): RR-CpGs at in CBMCs at birth ($\widehat{cor}^{(0)} \geq 0.95$) and PBMCs at age 7 (
654 $\widehat{cor}^{(7)} \geq 0.95$). Overlapping CpGs (violet) have $\widehat{con}^{(0,7)} \geq 0.95$. (c): IGA-CpGs and RR-
655 CpGs ($\widehat{cor}^{(0)} \geq 0.95$) in CBMCs at birth. (d): IGA-CpGs and RR-CpGs ($\widehat{cor}^{(0)} \geq 0.95$) in
656 PBMCs at age 7.

657 **Figure** 4: IGA-CpGs are enriched for CpGs with meQTLs. (a) Illustration of the causal
658 relationship between the methylation (M) at a CpG site, the genotype (G) at the SNP
659 within ±5kb of the CpG that had the smallest meQTL *P* value and inferred genetic

660 ancestry (IGA). Each graph corresponds to a unique CpG. (b) Plots of the meQTL *P* value
661 for edge a in CBMCs at birth, where CpGs were stratified by whether or not it was an
662 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
663 just for visual aid. (c) Plots of the logistic regression *P* value for edge c (Genotype $\sim$ IGA
664 $+$ Ancestry PC2 (see Figure 1)), stratified by whether or not the SNP was an IGA-meQTL.
665 **Figure** 5: Inferred genetic ancestry effects on methylation are primarily genetic in origin.
666 Upper panel. Histograms of $\max\left(\hat{cor}^{(0)}, \hat{cor}^{(7)}\right)$ in the IGA (a) and the RR (b) analyses. If
667 $\max\left(\hat{cor}_g^{(0)}, \hat{cor}_g^{(7)}\right)$ is close to 1 in the IGA or RR analysis, CpG g is an IGA-CpG or RR-
668 CpG, respectively, in CBMCs at birth or PBMCs at 7. The red histogram is created with
669 the set of 726 ethnicity-associated CpGs identified in Galanter et al.[20] that were also
670 among the 784,484 CpGs in our study and the blue are the remaining 783,758 CpGs.
671 Lower panel. *P* values for age (0 to 7) and gestational age (GA). The red points are the 263
672 ethnic-associated CpGs identified in Galanter et al., 2017[20] that had an meQTL at a 5%
673 FDR identified in our study. The blue points are the remaining 179,385 CpGs that were
674 within 5kB of a SNP in our data and among the CpGs studied in Galanter et al.[20]. The
675 large circles are shown for visual aid.

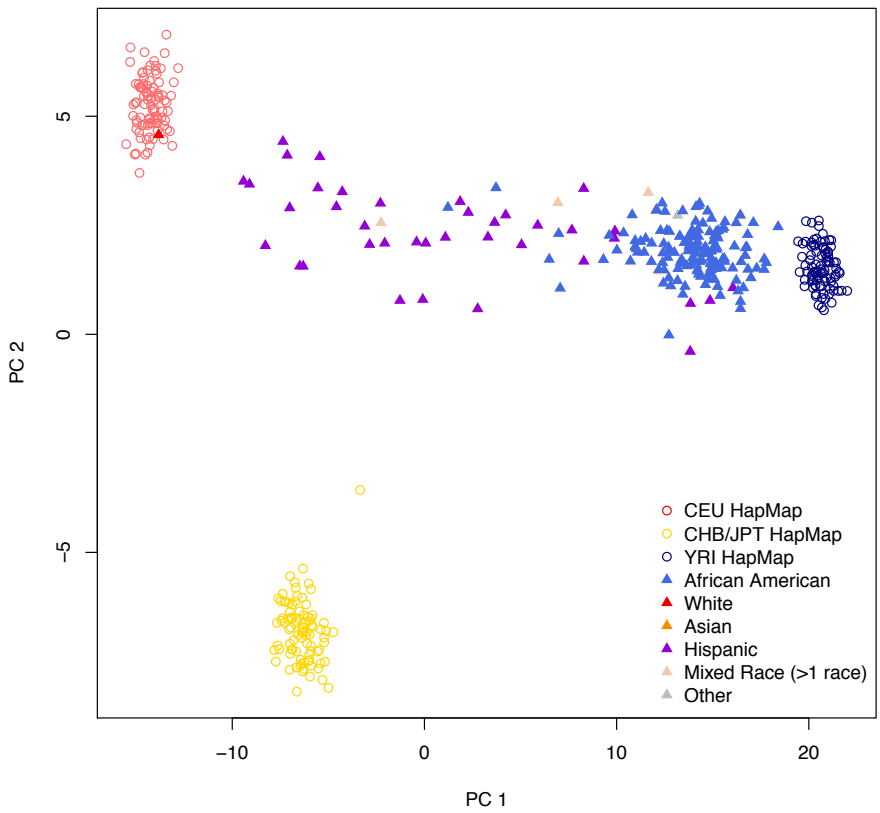
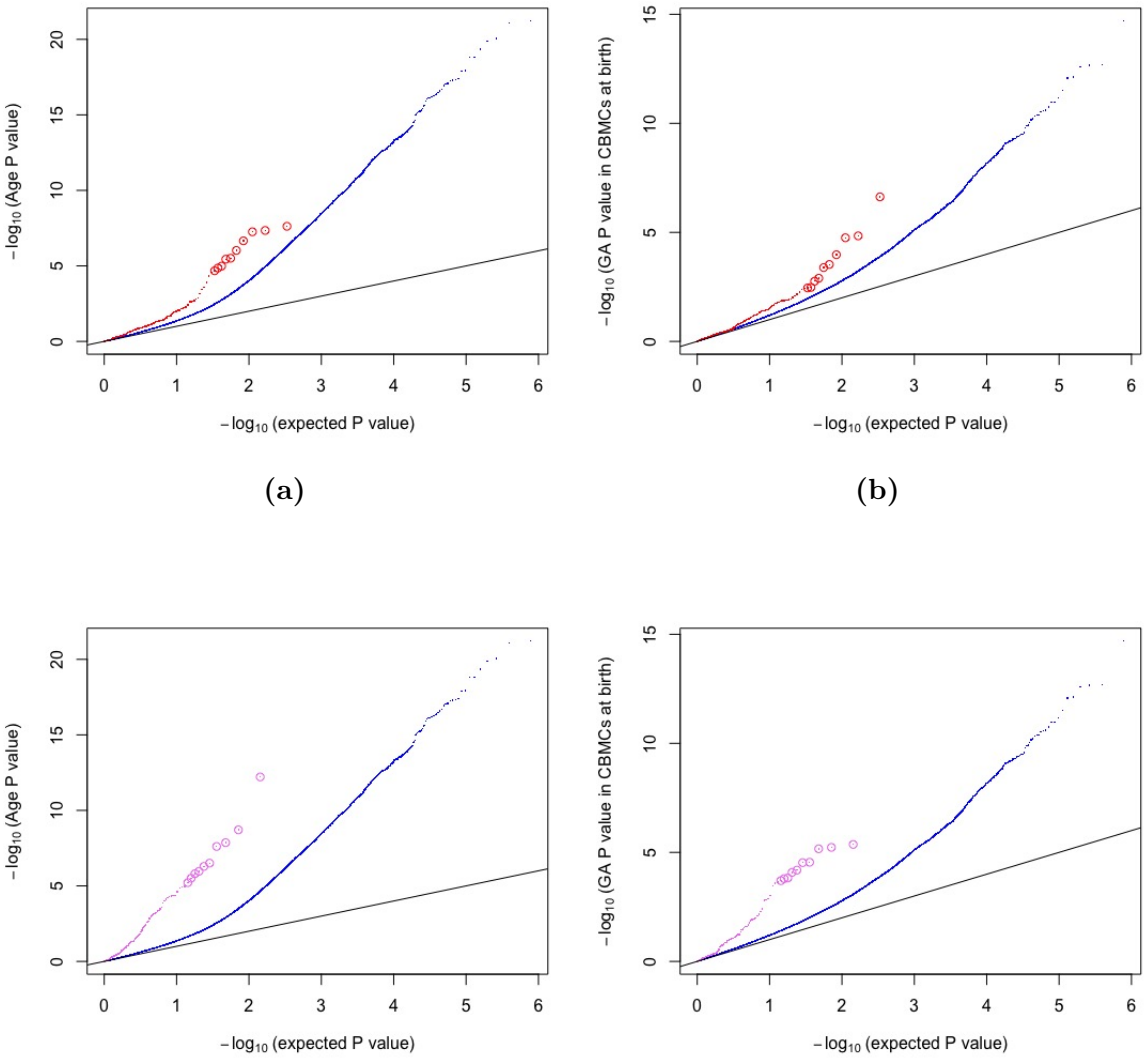

PC 2
5
0
−5
−10
0
10
20
PC 1
CEU HapMap
CHB/JPT HapMap
YRI HapMap
African American
White
Asian
Hispanic
Mixed Race (>1 race)
Other

Figure 1



(a) **(b)**

(c) **(d)**

Figure 2

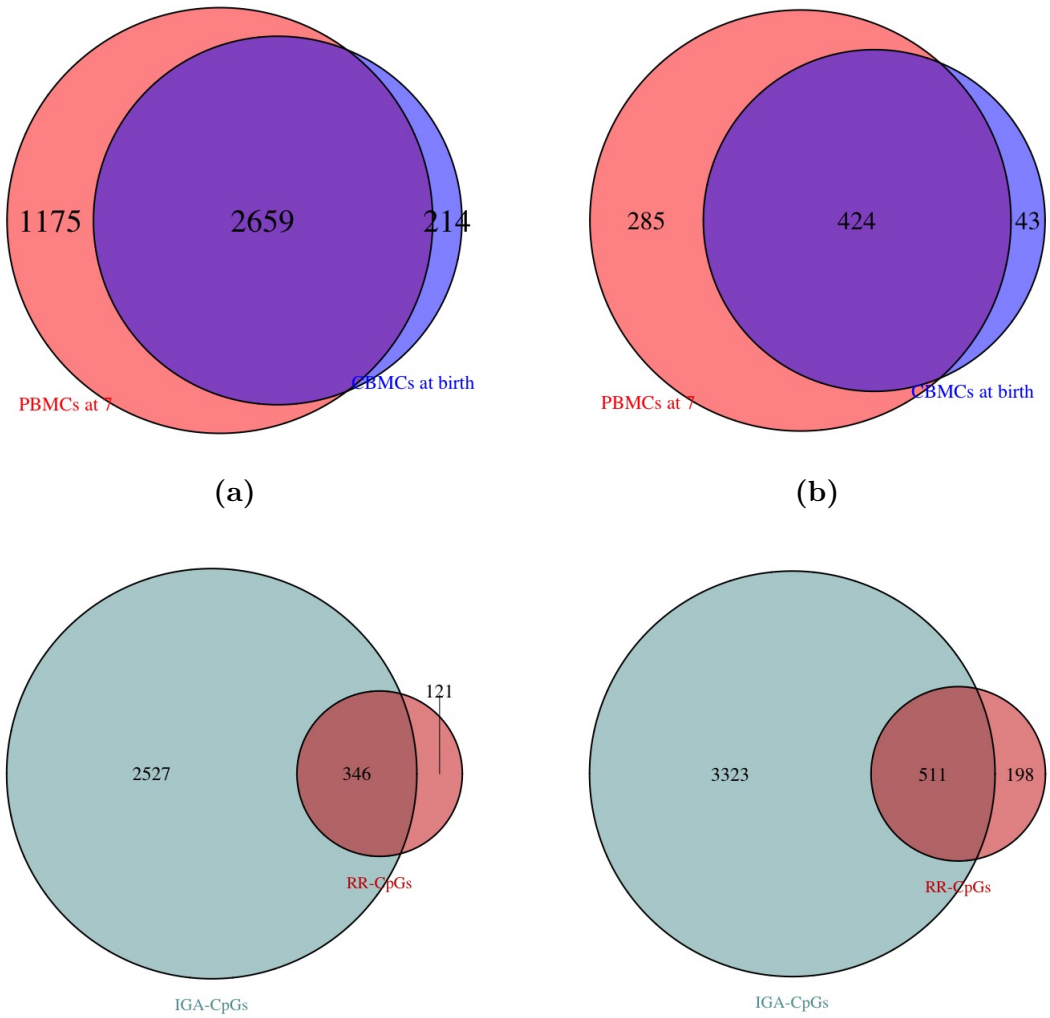
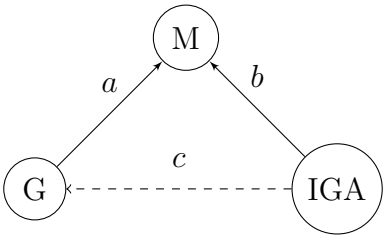
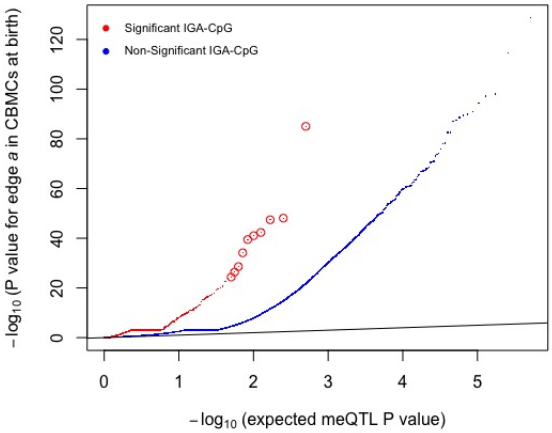

1175
2659
214
PBMCs at 7
CBMCs at birth
(a)
285
424
43
PBMCs at 7
CBMCs at birth
(b)
121
2527
346
RR-CpGs
IGA-CpGs
(c)
3323
511
198
RR-CpGs
IGA-CpGs
(d)

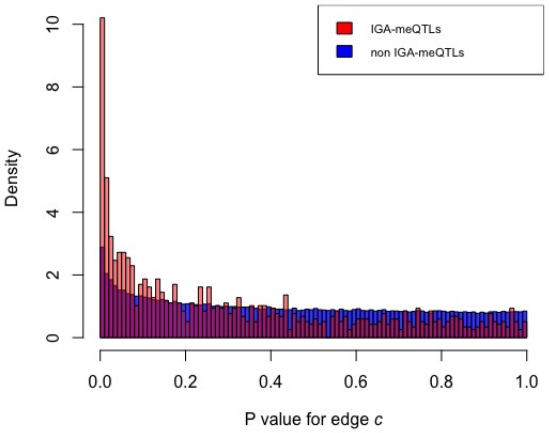
Figure 3


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 4

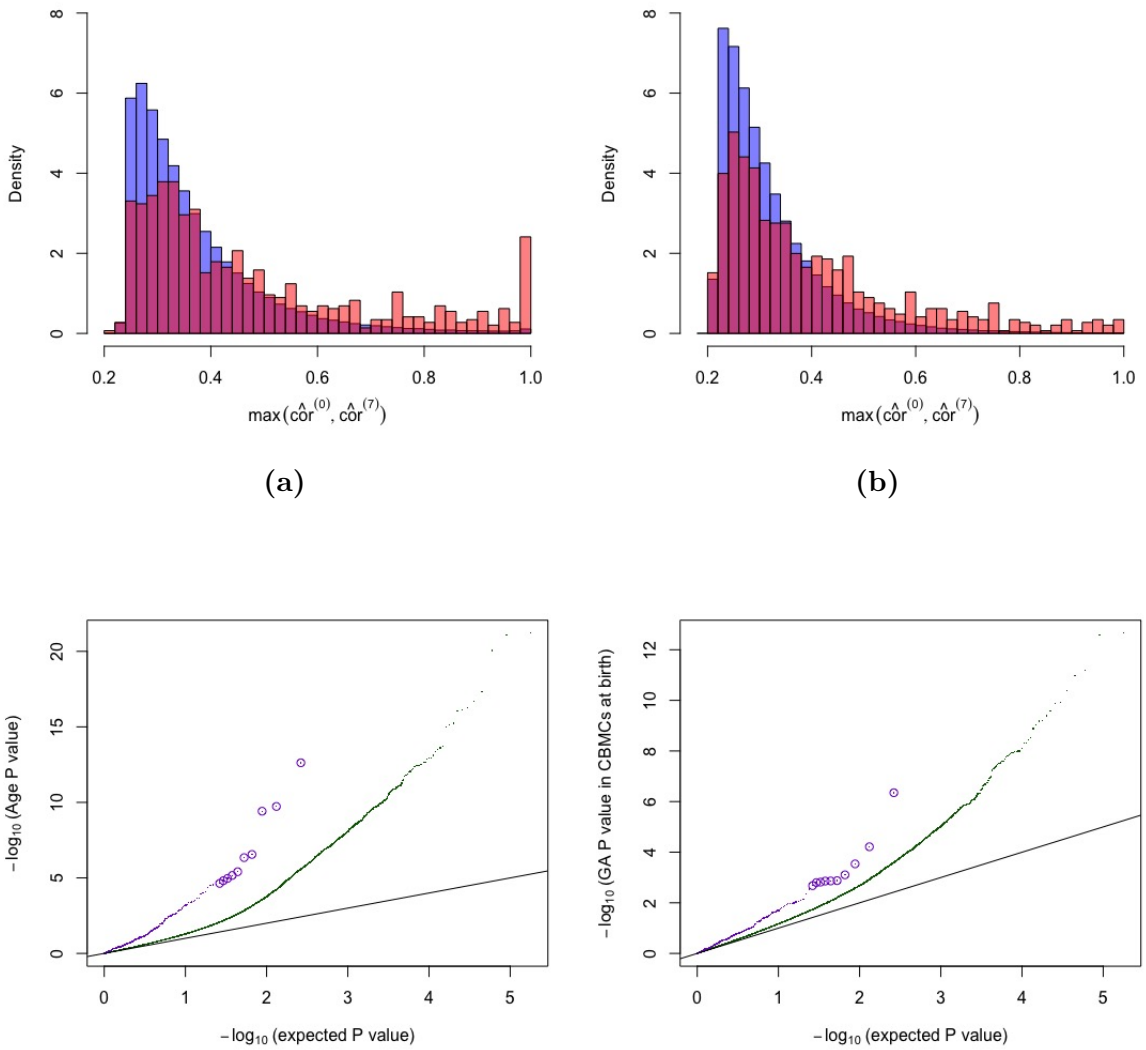

Density
0
2
4
6
8
0.2
0.4
0.6
0.8
1.0
$\max(\hat{cor}^{(0)}, \hat{cor}^{(7)})$
(a)
Density
0
2
4
6
8
0.2
0.4
0.6
0.8
1.0
$\max(\hat{cor}^{(0)}, \hat{cor}^{(7)})$
(b)
$-\log_{10}$ (Age P value)
0
5
10
15
20
0
1
2
3
4
5
$-\log_{10}$ (expected P value)
(c)
$-\log_{10}$ (GA P value in CBMCs at birth)
0
2
4
6
8
10
12
0
1
2
3
4
5
$-\log_{10}$ (expected P value)
(d)

Figure 5