Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures

Joshua M. Galanter, MD, MAS[1,2,3], Christopher R. Gignoux, PhD[4], Sam S. Oh, PhD[1], Dara Torgerson, PhD[2], Maria Pino-Yanes, PhD[5,6], Neeta Thakur, MD, MPH[1], Celeste Eng, BS[1], Donglei Hu, PhD[1], Scott Huntsman, MS[1], Harold J. Farber, MD[7], Pedro C Avila, MD[8], Emerita Brigino-Buenaventura, MD[9], Michael A LeNoir, MD[10], Kelly Meade, MD[11], Denise Serebrisky, MD[12], William Rodríguez-Cintrón, MD[13], Rajesh Kumar, MD[14], Jose R Rodríguez-Santana, MD[15], Max A. Seibold, PhD[16], Luisa N. Borrell, DDS, PhD[17], Esteban G. Burchard, MD, MPH[1,2*], Noah Zaitlen, PhD[1*]

1. Department of Medicine, University of California, San Francisco, CA
2. Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA
3. Department of Epidemiology and Biostatistics, University of California, San Francisco, CA
4. Department of Genetics, Stanford University, Stanford, CA
5. Hospital Universitario Nuestra Señora de Candelaria, Tenerife, Spain
6. CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain
7. Department of Pediatrics, Baylor College of Medicine and Texas Children's Hospital, Houston, Texas
8. Division of Allergy and Immunology, Feinberg School of Medicine, Northwestern University, Chicago, IL
9. Kaiser Permanente-Vallejo Medical Center, Vallejo, CA
10. Bay Area Pediatrics, Oakland, CA
11. Department of Pediatrics, Children's Hospital and Research Center, Oakland, CA
12. Jacobi Medical Center, Bronx, NY
13. Veterans Caribbean Health System, San Juan, Puerto Rico
14. Division of Allergy and Immunology, The Ann and Robert H. Lurie Children's Hospital of Chicago, Chicago, IL
15. Centro de Neumología Pediátrica, San Juan, Puerto Rico
16. Center for Genes, Environment, and Health, Department of Pediatrics, National Jewish Health, Denver, CO
17. Graduate School of Public Health and Health Policy, City University of New York, New York, NY

* These authors contributed equally to this work

Please address correspondence to:

Esteban G. Burchard, MD, MPH

University of California, San Francisco

Departments of Bioengineering & Therapeutic Sciences and Medicine

UCSF Box 2911

San Francisco, CA 94143-2911

Ph: (415) 514-9677

Fax: (415) 514-4365

e-mail: esteban.burchard@ucsf.edu

or

Noah Zaitlen, PhD

University of California, San Francisco

Department of Medicine

UCSF Box 2552

San Francisco, CA 94143-2552

Ph: (415) 502-2027

e-mail: noah.zaitlan@ucsf.edu

or

Joshua Galanter, MD, MAS

University of California, San Francisco

Departments of Medicine and Epidemiology & Biostatistics

UCSF Box 2911

San Francisco, CA 94143-2552

Ph: (415) 514-9931

e-mail: galanter@gmail.com

Sources of Funding:

This research was supported in part by the Sandler Family Foundation, the American Asthma Foundation, National Institutes of Health (P60 MD006902, R01 HL117004, R21ES24844, U54MD009523, R01 ES015794, R01 HL088133, M01 RR000083, R01 HL078885, R01 HL104608, U19 AI077439, M01 RR00188), ARRA grant RC2 HL101651, and TRDRP 24RT-0025; EGB was supported in part through grants from the Flight Attendant Medical Research Institute (FAMRI), and NIH (K23 HL004464); NZ was supported in part by an NIH career development award from the NHLBI (K25HL121295). JMG was supported in part by NIH Training Grant T32 (T32GM007546) and career development awards from the NHLBI (K23HL111636) and NCATS (KL2TR000143) as well as the Hewett Fellowship; N.T. was supported in part by an institutional training grant from the NIGMS (T32-GM007546) and career development awards from the NHLBI (K12-HL119997 and K23-HL125551), Parker B. Francis Fellowship Program, and the American Thoracic Society; CRG was supported in part by NIH Training Grant T32 (GM007175) and the UCSF Chancellor's Research Fellowship and Dissertation Year Fellowship; RK was supported with a career development award from the NHLBI (K23HL093023); HJF was supported in part by the GCRC (RR00188); PCA was supported in part by the Ernest S. Bazley Grant; MAS was supported in part by 1R01HL128439-01. This publication was supported by various institutes within the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Abstract

In clinical practice and biomedical research populations are often divided categorically into distinct racial and ethnic groups. In reality, these categories, which are based on social rather than biological constructs, comprise diverse groups with highly heterogeneous histories, cultures, traditions, religions, as well as social and environmental exposures. While the factors captured by these categories contribute to clinical practice and biomedical research, the use of race/ethnicity is widely debated. As a response to this debate, genetic ancestry has been suggested as a complement or alternative to this categorization. However, few studies have examined the effect of genetic ancestry, racial/ethnic identity, and environmental exposures on biological processes. Herein, we examine the associations between self-identified ethnicity and epigenetic modification of DNA methylation, a phenomenon affected by both genetic and environmental factors. We also assess the relative contributions of genetic ancestry and environmental factors on these associations. We typed over 450,000 variably methylated CpG sites in primary whole blood of 573 individuals of Mexican and Puerto Rican descent who also had high-density genotype data. We found that both genetically determined ancestry and self-identified ethnicity were significantly associated with methylation levels at a large number of CpG sites. In addition, we found an enrichment of ethnicity-associated sites amongst loci previously associated with environmental and social exposures, particularly maternal smoking during pregnancy. This suggests that ethnic self-identification may function as a proxy for environmental exposures. Overall we conclude that race and ethnicity provide important and relevant clinical and biomedical information above and beyond and individual's genetic ancestry.

Significance Statement

Race, ethnicity, and genetic ancestry have had a complex and often controversial history within biomedical research and clinical practice. For example, race- and ethnicity-specific clinical reference standards are based on an average derived from statistical modeling applied to population-based sampling on a given physical trait such as pulmonary function. However, because race and ethnicity are social constructs, they ignore the heterogeneity within the categories. To account for these heterogeneities and avoid social and political controversies, the genetics community has integrated the use of genetic ancestry as a proxy because genetic sequence is not altered by environmental or social factors. We explore the relative contributions of ancestry, ethnicity and environment to variation in methylation, a fundamental biological process.

Introduction

Race, ethnicity, and genetic ancestry have had a complex and often controversial history within biomedical research and clinical practice[1-3]. For example, race- and ethnicity-specific clinical reference standards are based on an average derived from statistical modeling applied to population-based sampling on a given physical trait such as pulmonary function[4,5]. However, because race and ethnicity are social constructs and poor markers for genetic diversity, they ignore the heterogeneity within the categories[6]. To account for these heterogeneities and avoid social and political controversies, the genetics community has integrated the use of genetic ancestry instead of race and ethnicity[3] because genetic sequence is not altered by environmental or social factors, such as those related to racial or ethnic self-identity. Indeed, recent work from our

47 group and others have demonstrated that genetic ancestry improves diagnostic
48 precision compared to crude racial/ethnic categorizations for specific medical
49 conditions and clinical decisions[7-9].

50 Epigenetic modification of the genome through methylation plays a key role in the
51 regulation of diverse cellular processes[10]. Changes in DNA methylation patterns have
52 been associated with complex diseases, including various cancers[11], cardiovascular
53 disease[12,13], obesity[14], diabetes[15], autoimmune and inflammatory diseases[16], and
54 neurodegenerative diseases[17]. Epigenetic changes are thought to reflect influences of
55 both genetic[18] and environmental factors[19]. The discovery of methylation quantitative
56 trait loci (meQTL's) across populations by Bell *et al.* established the influence of genetic
57 factors on methylation levels in a variety of tissue types[18], with meQTL's explaining
58 between 22% and 63% of the variance in methylation levels. Multiple environmental
59 factors have also been shown to affect methylation levels, including endocrine
60 disruptors, tobacco smoke[20,21], polycyclic aromatic hydrocarbons, infectious pathogens,
61 particulate matter, diesel exhaust particles[22], allergens, heavy metals, and other indoor
62 and outdoor pollutants[23]. Psychosocial factors, including measures of traumatic
63 experiences[24-26], socioeconomic status[27,28], and general perceived stress[29], also affect
64 methylation levels. Racial and ethnic categories reflect the shared experiences and
65 exposures to known risk factors for disease, such as air pollution and tobacco smoke,
66 poverty, and inadequate access to medical services, which have all contributed to worse
67 disease outcomes in certain populations[30,31]. Since environmental exposures affect
68 methylation, they may be reflected in systematic differences in methylation levels

69 between racial or ethnic groups that cannot be captured by using genetically defined
70 ancestry alone.

71 Given the roles of both genetic and environmental influences upon methylation, we
72 leveraged genome-wide methylation data as an intermediate phenotype to examine the
73 degree to which self-identified ethnicity and genetic ancestry are reflected in differences
74 in methylation. We hypothesized that while genetic ancestry can explain many of the
75 differences in methylation between these groups, some ethnic-specific methylation
76 differences reflecting social and environmental differences between groups, might
77 remain. If our hypothesis is correct, our findings would have important implications for
78 both the use of ancestry to capture the biological changes and of race/ethnicity to
79 account for social and environmental exposures. Epigenome-wide association studies in
80 diverse populations may be susceptible to confounding due to environmental exposures
81 in addition to confounding due to population stratification[32].

82 We also examined the relationship between genome-wide (global) estimates of ancestry
83 and locus-specific (local) ancestry to determine the extent to which associations between
84 global ancestry and methylation are reflective of genetic factors acting in -cis. Finally, by
85 using dense genotyping arrays, we queried whether methylation differences associated
86 with ancestry can be traced back to meQTLs whose allele frequencies differ by ancestry.
87 To address these aims, we analyzed data from 573 Latino children according to their
88 national origin identity or ethnic subgroup (such as Puerto Rican and Mexican),
89 enrolled in the Genes-Environments and Admixture in Latino Americans (GALA II)
90 study of childhood asthma[33].

91 **Results**

92 The study included 573 participants, the majority of whom self-identified as being either
93 of Puerto Rican (n = 220) or Mexican origin (n = 276). Table 1 displays baseline
94 characteristics of the GALA II study participants with methylation data included in this
95 study, stratified by ethnic subgroups (Puerto Rican, Mexican, Other Latino, and Mixed
96 Latinos who had grandparents of more than one national origin). SI Appendix Figure 1
97 shows the distribution of African, European, and Native American ancestry among the
98 524 participants with genomic ancestry estimates.

99 Differences in ethnicity and ancestry resulted in discernible patterns in the global
100 methylation profile as demonstrated in a multidimensional scaling analysis (SI
101 Appendix Figure 2A). As expected[27,34], the first few principal coordinates are strongly
102 correlated to imputed cell composition (SI Appendix Figure 2B and C). There are also
103 significant associations of self-identified sub-ethnicity with PC2 (p-ANOVA = 0.003),
104 PC3 (p-ANOVA = 0.004), PC6 (p-ANOVA = 0.0001), PC7 (p-ANOVA = 0.0003) [SI
105 Appendix Figure 3A], and PC8 (p-ANOVA = 0.0003), after adjusting for age, sex,
106 disease status, cell components, and technical factors (plate and position). Genetic
107 ancestry was associated with PC3 (p-ANOVA = 0.002), PC7 (p-ANOVA = 0.0004) [SI
108 Appendix Figure 3B] and PC8 (p-ANOVA = 0.001) in a two degree of freedom ANOVA
109 test, adjusting for age, sex, disease status, cell components, technical factors, and
110 ethnicity. SI Appendix Table 1 summarizes the results of the simple correlation analysis
111 of methylation with ethnicity and ancestry, as well as the adjusted nested ANOVA
112 models described above and the mediation results described below.

113 A mediation analysis[35] revealed that the associations between ethnicity and PCs 3, 7,
114 and 8 were significantly mediated by Native American ancestry (mediation $p = 0.01$,
115 <0.001, and <0.001, respectively) and inclusion of Native American ancestry in the
116 regression model of PCs 3, 7, and 8 caused the ethnicity associations to be non-
117 significant. However, the associations of ethnicity with PCs 2 and 6 were not explained
118 by Native American, African or European ancestry (mediation $p > 0.05$), suggesting that
119 ethnic differences are associated with global methylation patterns not captured by
120 genetic differences. When genetic ancestry was regressed on the methylation data with
121 the principal coordinates recalculated using the residuals of the regression between
122 methylation and ancestry, there was an association between ethnicity and PC6 (p-
123 ANOVA = 0.003). However, there was no association with any of the other principal
124 coordinates. These observations suggest that while genetic ancestry can explain some of
125 the association between ethnicity and global methylation patterns, other non-genetic
126 factors, such as environmental and social exposure differences associated with ethnicity,
127 influence methylation independent of genetic ancestry.

128 An epigenome-wide association study of self-identified ethnicity (see methods for
129 details of ascertainment of ethnicity) and methylation identified a significant difference
130 in methylation M-values between ethnic groups at 916 CpG sites at a Bonferroni-
131 corrected significance level of less than 1.6×10^{-7} [Figure 1A and SI Appendix Table 2].
132 The most significant association with ethnicity occurred at cg12321355 in the ABO blood
133 group gene (*ABO*) on chromosome 3 (p-ANOVA 6.7×10^{-22}) [Figure 1B]. A two degree of
134 freedom ANOVA test for genomic ancestry was also significantly associated with
135 methylation level at this site ($p = 2.3\times10^{-5}$) [Figure 1C], and when the analysis was

136 stratified by ethnic sub-group, showed an association in both Puerto Ricans and
137 Mexicans ($p = 0.001$ for Puerto Ricans, $p = 0.003$ for Mexicans). Although adjusting for
138 genomic ancestry attenuated the effect of ethnicity, a significant association between
139 ethnicity and methylation remained ($p = 0.04$). Recruitment site, an environmental
140 exposure proxy, was not significantly associated with methylation at this locus ($p = 0.5$),
141 suggesting that environmental differences associated with ethnicity beyond geography
142 and ancestry are driving the association.

143 When we repeat the analysis adjusting for ancestry, a significant association remained
144 in 314 of the 834 (37.8%, $p < 2.2\times10^{-16}$ for enrichment) CpG sites associated with
145 ethnicity [SI Appendix Figure 4A and SI Appendix Table 2] (82 sites were excluded
146 because they demonstrated unstable coefficient estimates and inflated standard errors
147 due to strong correlations between ethnicity and ancestry, especially Native American
148 ancestry [see SI Appendix Figure 1]). Genomic ancestry explained a median of 4.2%
149 (IQR 1.8% to 8.3%) of the variance in methylation at these loci and accounts for a
150 median of 75.7% (IQR 45.8% to 92%) of the total variance in methylation explained
151 jointly by ethnicity and ancestry [SI Appendix Figure 4B]. Sensitivity tests for
152 departures from linearity, fine scale population substructure and the exclusion of the 16
153 participants who self-identified as “Mixed Latino” sub-ethnicity, did not meaningfully
154 affect our results [See SI Appendix Results and SI Appendix Tables 2-6]. We conclude
155 that genetic ancestry explains much but not all of the association between ethnicity and
156 methylation. We hypothesize that other non-genetic factors associated with ethnicity
157 could explain the ethnicity-associated methylation changes that cannot be accounted for
158 by genomic ancestry alone.

159 Methylation at CpG loci that had previously been reported to be associated with
160 environmental exposures whose exposure prevalence differs between ethnic groups
161 were tested for association with ethnicity in this study. A recent meta-analysis of
162 maternal smoking during pregnancy, an exposure that varies significantly by ethnicity[33],
163 identified associations with methylation at over 6,000 CpG loci[21]. We found 1341 of
164 4404 that passed QC in our own study (30.4%) were nominally associated with ethnicity
165 ($p < .05$), which represented a highly significant ($p < 2\times10^{-16}$) enrichment. Using a
166 Bonferroni correction for the 4404 loci tested, 126 maternal-smoking related loci were
167 associated with ethnicity ($p < 1.1\times10^{-5}$), and 27 loci were among the 916 CpG's reported
168 above as associated with ethnicity [SI Appendix Table 7]. We also examined methylation
169 loci from an earlier study of maternal smoking in Norwegian newborns[20] as well as
170 studies of diesel exhaust particles[22] and exposure to violence[24]. These results are
171 supportive of our hypothesis that environmental exposures may be responsible for the
172 observed differences in methylation between ethnic groups and are presented in the SI
173 Appendix Text and in SI Appendix Table 8.

174 We also found 194 loci with a significant association between global genetic ancestry
175 and methylation levels (after adjusting for ethnicity) at a Bonferroni corrected
176 association p-value of less than 1.6×10^{-7} [SI Appendix Figure 5 and SI Appendix Table
177 9], including 48 that were associated with ethnicity in our earlier analysis. Of these
178 significant associations, 55 were driven primarily by differences in African ancestry, 94
179 by differences in Native American ancestry, and 45 by differences in European ancestry.
180 The most significant association between methylation and ancestry occurred at
181 cg04922029 in the Duffy antigen receptor chemokine gene (DARC) on chromosome 1

182 (ANOVA p-value 3.1×10^{-24}) [SI Appendix Figure 5B]. This finding was driven by a strong
183 association between methylation level and global African ancestry; each 25 percentage
184 point increase in African ancestry was associated with an increase in M-value of 0.98,
185 which corresponds to an almost doubling in the ratio of methylated to unmethylated
186 DNA at the site (95% CI 0.72 to 1.06 per 25% increase in African ancestry, $p = 1.1\times10^{-21}$).
187 The distribution of methylation M-values at cg04922029 is tri-modal, raising the
188 possibility that a SNP whose allele frequency differs between African and non-African
189 populations may be driving the association. There was no significant heterogeneity in
190 the association between genetic ancestry and methylation between Puerto Ricans and
191 Mexicans (p-het = 0.5). Mexicans have a mean unadjusted methylation M-value 0.48
192 units lower than Puerto Ricans (95% CI 0.35 to 0.62 units, $p = 1.1\times10^{-11}$). However,
193 adjusting for African ancestry accounts for the differences in methylation level between
194 the two sub-groups (p-adjusted = 0.4), demonstrating that ethnic differences in
195 methylation at this site are due to differences in African ancestry.

196 A substantial proportion of the effect of global ancestry on local methylation levels is
197 due to local ancestry acting in -cis. Among the 194 CpG sites associated with global
198 ancestry, local ancestry at the CpG site explained a median of 10.4% (IQR 3.0% to
199 19.4%) of the variance in methylation, accounting for a median of 52.8% (IQR 20.3% to
200 84.9%) of the total variance explained jointly by local and global ancestry [SI Appendix
201 Figure 5].

202 In an admixture mapping study, we find that methylation at 3,694 of 321,503 CpG's
203 (1.1%), was significantly associated with ancestry at the CpG site at a Bonferroni
204 corrected association p-value of less than 1.6×10^{-7} [Figure 2A and SI Appendix Table

205 10]. This included 118 of the 194 loci identified above (61%), where global ancestry was
206 associated with methylation. The most significant CpG site was again cg04922029,
207 which was almost perfectly correlated with African ancestry at the locus ($p = 6\times10^{-162}$)
208 [Figure 2B]. Each African haplotype at the CpG site was associated with an increase in
209 methylation M-value of 2.7, corresponding to a 6.5-fold increase in the ratio of
210 methylated to unmethylated DNA per African haplotype at that locus. The second most
211 significant association occurred at cg06957310 on chromosome 17; each increase in
212 African ancestry at the locus was associated with a decrease in M-value of 1.7 (a 3.2-fold
213 decrease in the ratio of methylated to unmethylated DNA; $p = 3.7\times10^{-75}$). We obtained
214 similar results when performing the analysis using methylation ß values instead of M-
215 values [See SI Appendix text and SI Appendix Table 11]

216 For each of the admixture mapping loci, we tested whether a single nucleotide
217 polymorphism (SNP) within 1 Mb of the CpG was associated with methylation. We
218 found 3637 loci out of the 3694 (98.5%) admixture mapping findings with at least one
219 SNP within 1 Mb that was significantly associated with methylation levels (after
220 adjustment of the number of SNPs in cis-) [SI Appendix Table 12]

221 The SNP/CpG pair were separated by a median distance of 10.9 kb (interquartile range
222 2.9 kb to 35.1 kb). The furthest SNP/CpG pair were 998 kb apart. The most significant
223 SNP/CpG pair were cg25134647/rs4963867, on chromosome 12, which are separated by
224 412 base pairs [SI Appendix Figure 6A/B]. Each copy of the T allele was associated with
225 a decrease in M-value of 3.58, corresponding to a nearly 12-fold decrease in the ratio of
226 methylated to unmethylated DNA at the site [SI Appendix Figure 6C]
227 ($p < 10^{-370}$). We found that CpG cg04922029 (our top admixture mapping association)

228 was significantly correlated with SNP rs2814778 [Figure 2C], the Duffy null mutation,
229 212 base pairs away; each copy of the C allele was associated with an increase in M-value
230 of 1.5, or a 2.9-fold increase in the ratio of methylated to unmethylated DNA ($p = 3.8\times10^{-90}$) [Figure 2D]. We obtained similar results when performing the analysis using
232 methylation ß values instead of M-values [See SI Appendix text and SI Appendix Table
233 13]

234 Discussion

235 We have shown that both genomic ancestry and self-identified ethnicity are
236 independently associated with methylation levels throughout the genome. While
237 genomic ancestry can explain a portion of the association between ethnicity and
238 methylation, genomic ancestry inadequately accounts for the association between
239 ethnicity and methylation at 34% (314/916) of loci. Our results suggest ethnic self-
240 identification is unlikely to have a direct causal relationship with methylation but rather,
241 that other non-genetic factors associated with self-identified ethnicity may affect
242 differences in methylation patterns between Latino subgroups. These factors may
243 include social, economic, cultural, and environmental exposures.

244 We conclude that systematic environmental differences between ethnic subgroups likely
245 play an important role in shaping the methylome for both individuals and populations.
246 Loci previously associated with diverse environmental exposures such as *in utero*
247 exposure to tobacco smoke[20,21], as well as diesel exhaust particles[22] and psychosocial
248 stress[24] were enriched in our set of loci where methylation was associated with ethnicity.
249 Twenty-seven of the loci associated with maternal smoking during pregnancy in a large
250 consortium meta-analysis[21] were associated with methylation at a genome-wide

251 significance threshold of 1.6×10^{-7}. Thus, inclusion of relevant social and environmental
252 exposures in studies of methylation may help elucidate racial/ethnic disparities in
253 disease prevalence, health outcomes and therapeutic response. However, in many cases,
254 a detailed environmental exposure history is unknown, unmeasurable or poorly
255 quantifiable, and race/ethnicity becomes a useful, albeit imperfect proxy.

256 Our comprehensive analysis of high-density methyl- and genotyping from genomic DNA
257 allowed us to investigate the genetic control of methylation in great detail and without
258 the potential destabilizing effects of EBV transformation and culture in cell lines[36]. The
259 strongest patterns of methylation are associated with cell composition in whole blood[27].
260 However, the specific type of Latino ethnic-subgroups (Puerto Rican, Mexican, other, or
261 mixed) is also associated with principal coordinates of genome-wide methylation.

262 Our approach has some potential limitations. It is possible that fine-scale population
263 structure (sub-continental ancestry) within European, African, and Native American
264 populations may contribute to ethnic differences in methylation, as we had previously
265 reported in the case of lung function[37]. However, despite the presence of additional
266 substructure among the GALA II participants, PC's 3-10 explained the association
267 between ethnicity and ancestry at only 51 loci. PCs from chip-based genotypes will not
268 capture all forms of genetic variation. Clusters of ethnicity specific rare variants of large
269 effect or strong ethnicity-specific selective sweeps in the last 8-12 generations[38] could
270 also give rise to methylation differences, but these are inconsistent with existing rare
271 variant and selection analyses[39,40]. Our models of genetic ancestry assumed a linear
272 effect of ancestry on methylation, whereas a nonlinear association or other model
273 misspecification could have led to incomplete adjustment for genetic ancestry, and thus,

274 led to a residual association between ethnicity and methylation. However, when we
275 added second and third order polynomials or cubic splines to our models, we found
276 evidence for a nonlinear association between ancestry and methylation at only 25 and
277 26 loci, respectively, and it did not affect the association between ethnicity and
278 methylation. Although it is impossible to account for all types of non-linearity and non-
279 additivity (such as gene by gene or gene by environment interaction), our analysis
280 suggests that non-linear effects are unlikely to be significant. To rule out any residual
281 confounding due to recruitment sites, we conducted an additional analysis on the effect
282 of recruitment site on methylation both for the overall study and for the Mexican
283 participants (the largest study population in this analysis). We observed no significant
284 independent effect of recruitment site suggesting that confounding due to recruitment
285 region was limited, at least within the United States. We were unable to test for the
286 effect of geographic differences between the United States and Puerto Rico because our
287 study included relatively few Puerto Ricans recruited outside of Puerto Rico.

288 The presence of a strong association between genetic ancestry and methylation raises
289 the possibility that epigenetic studies can be confounded by population stratification,
290 similar to genetic association studies, and that adjustment for either genetic ancestry or
291 selected principal components is warranted. This possibility was first demonstrated in a
292 previous analysis of the association between self-described race and methylation[41].
293 However, the study only evaluated two distinct racial groups (African Americans and
294 Whites), while the present study demonstrates the possibility of population
295 stratification in an admixed and heterogeneous population with participants from
296 diverse Latino national origins. The tendency to consider Latinos as a homogenous or

297 monolithic ethnic group makes any analysis of this population particularly challenging.
298 Our finding of loci whose methylation patterns differed between Latino ethnic
299 subgroups, even after adjusting for genetic ancestry, suggests that any analysis of these
300 populations in disease-association studies without adjusting for ethnic heterogeneity is
301 likely to result in spurious associations even after controlling for genomic ancestry.

302 Our analysis of local genetic ancestry and methylation demonstrates that loci associated
303 with genome-wide ancestry are driven primarily by allele frequency differences between
304 ancestral populations in 118 out of 194 loci, suggesting that in most cases global ancestry
305 is acting in -cis. In addition, methylation-QTLs whose allele frequencies differ between
306 ethnic groups are found in 95% (3637/3694) of loci associated with local ancestry. Of
307 particular interest, the most significant ancestry-associated locus, the *DARC* gene,
308 harbors an association between ancestry and methylation at cg04922029, which can be
309 entirely explained by the genotype at rs2814778, the Duffy null mutation. This mutation,
310 which confers resistance to *P. vivax* malaria, has an allele frequency of 100% in
311 individuals in the five 1000 genomes populations[42-44] from Africa (Esan in Nigeria,
312 Gambian in the Western Division of Gambia, Luhya in Webuye, Kenya, Mende in Sierra
313 Leone, and Yoruban in Ibadan, Nigeria), and 80% to 90% in admixed populations of
314 African origin in the Americas (89% in Afro Caribbeans in Barbados, and 80% in
315 African Americans in the Southwest U.S.). In contrast, the allele frequency is 1% in the
316 five European populations and 0% in the five Asian populations [Figure 2E], consistent
317 with the known high level of ancestry information at the locus. Latinos, being admixed,
318 have intermediate and more varied minor allele frequencies, ranging from 3% in
319 Mexicans to 14% in Puerto Ricans.

320 In summary, this study provides a framework for understanding how genetic, social and
321 environmental factors can contribute to systematic differences in methylation patterns
322 between ethnic subgroups, even between presumably closely related populations such as
323 Puerto Ricans and Mexicans. Methylation QTL's whose allele frequency varies by
324 ancestry lead to an association between local ancestry and methylation level. This, in
325 turn, leads to systematic variation in methylation patterns by ancestry, which then
326 contributes to ethnic differences in genome-wide patterns of methylation. However,
327 although genetic ancestry has been used to adjust for confounding in genetic studies,
328 and can account for much of the ethnic differences in methylation in this study, ethnic
329 identity is associated with methylation independent of genetic ancestry. This is likely
330 due to social and environmental effects captured by ethnicity. Indeed, we find that CpG
331 sites known to be influenced by social and environmental exposures are also
332 differentially methylated between ethnic subgroups. These findings called attention to a
333 more complete understanding of the effect of social and environmental variables on
334 methylation in the context of race and ethnicity to fully understanding this complex
335 process.

336 Our findings have important implications for the independent and joint effects of race,
337 ethnicity, and genetic ancestry in biomedical research and clinical practice, especially in
338 studies conducted in diverse or admixed populations. Our conclusions may be
339 generalizable to any population that is racially mixed such as those from South Africa,
340 India, and Brazil, though we would encourage further study in diverse populations. As
341 the National Institutes of Health (NIH) embarks on a precision medicine initiative, this
342 research underscores the importance of including diverse populations and studying

343 factors capturing the influence of social, cultural, and environmental factors, in addition
344 to genetic ones, upon disparities in disease and drug response.

345 Methods

346 All research on human subjects was approved by the Institutional Review Board at the
347 University of California and each of the recruitment sites (Kaiser Permanente Northern
348 California, Children's Hospital Oakland, Northwestern University, Children's Memorial
349 Hospital Chicago, Baylor College of Medicine on behalf of the Texas Children's Hospital,
350 VA Medical Center in Puerto Rico, the Albert Einstein College of Medicine on behalf of
351 the Jacobi Medical Center in New York and the Western Review Board on behalf of the
352 Centro de Neumologia Pediatrica), and all participants/parents provided age-
353 appropriate written assent/consent. Latino children were enrolled as a part of the
354 ongoing GALA II case-control study, where details of recruitment can be found[33] as well
355 as in the SI Appendix text.

356 Genomic DNA (gDNA) was extracted from whole blood using Wizard® Genomic DNA
357 Purification Kits (Promega, Fitchburg, WI). A subset of 573 participants (311 cases with
358 asthma and 262 healthy controls) was selected for methylation. Methylation was
359 measured using the Infinium HumanMethylation450 BeadChip (Illumina, Inc., San
360 Diego, CA) following the manufacturer's instructions. Details of methylation measures
361 and quality control are described in the SI Appendix text.

362 Details of genotyping and quality control procedures for single nucleotide
363 polymorphisms (SNPs) and individuals have been described elsewhere[45] and
364 summarized in the SI Appendix Text.

365 Unless otherwise noted, all regression models were adjusted for case status, age, sex,
366 estimated cell counts, and plate and position. To account for possible heterogeneity in
367 the cell type makeup of whole blood we inferred white cell counts using the method by
368 Houseman et al[34]. Indicator variables were used to code categorical variables with more
369 than two categories, such as ethnicity. In these cases, a nested analysis of variance
370 (ANOVA) was used to compare models with and without the variables to obtain an
371 omnibus p-value for the association between the categorical variable and the outcome.
372 For analyses of dependent beta-distributed variables (such as African, European, and
373 Native American ancestries), or cell proportion, k-1 variables were included in the
374 analysis, and a nested analysis of variance (ANOVA) was used to compare models with
375 and without the variables to obtain an k-1 degree of freedom omnibus p-value for the
376 association between predictor (such as ancestry) and the outcome variable.

377 The Bonferroni method was used to adjust for multiple comparisons. For methylome-
378 wide associations, the significance threshold was adjusted for 321,503 probes, resulting
379 in a Bonferroni threshold of 1.6×10^{-7}. Analyses were performed using R version 3.2.1
380 (The R Foundation for Statistical Computing)[46] and the Bioconductor package version
381 2.13. Further details on the models used for specific statistical analyses is in the SI
382 Appendix Text.

383 Acknowledgements

384 The authors acknowledge the families and patients for their participation and thank the
385 numerous health care providers and community clinics for their support and
386 participation in GALA II. In particular, the authors thank study coordinator Sandra
387 Salazar; the recruiters who obtained the data: Duanny Alva, MD, Gaby Ayala-Rodriguez,

388 Lisa Caine, Elizabeth Castellanos, Jaime Colon, Denise DeJesus, Blanca Lopez , Brenda
389 Lopez, MD, Louis Martos, Vivian Medina, Juana Olivo, Mario Peralta, Esther Pomares,
390 MD, Jihan Quraishi, Johanna Rodriguez, Shahdad Saeedi, Dean Soto, Ana Taveras. We
391 also thank Sasha Gusev for helpful discussion. Computations in this manuscript were
392 performed using the UCSF Biostatistics High Performance Computing System.

393 ## References

394 1. Risch, N., Burchard, E., Ziv, E. & Tang, H. Categorization of humans in biomedical
395 research: genes, race and disease. *Genome Biology* **3,** comment2007 (2002).

396 2. Cooper, R. S., Kaufman, J. S. & Ward, R. Race and genomics. *N. Engl. J. Med.*
397 **348,** 1166–1170 (2003).

398 3. Yudell, M., Roberts, D., Desalle, R. & Tishkoff, S. SCIENCE AND SOCIETY.
399 Taking race out of human genetics. *Science* **351,** 564–565 (2016).

400 4. Hankinson, J. L., Odencrantz, J. R. & Fedan, K. B. Spirometric reference values
401 from a sample of the general U.S. population. *Am. J. Respir. Crit. Care Med.* **159,**
402 179–187 (1999).

403 5. Quanjer, P. H. *et al.* Multi-ethnic reference values for spirometry for the 3-95-yr
404 age range: the global lung function 2012 equations. *European Respiratory*
405 *Journal* **40,** 1324–1343 (2012).

406 6. Borrell, L. N. Racial identity among Hispanics: implications for health and well-
407 being. *Am J Public Health* **95,** 379–381 (2005).

408 7. Kumar, R. *et al.* Genetic Ancestry in Lung-Function Predictions. *N. Engl. J. Med.*
409 **363,** 321–330 (2010).

410 8. Udler, M. S. *et al.* Effect of Genetic African Ancestry on eGFR and Kidney Disease.
411 *J. Am. Soc. Nephrol.* **26,** 1682–1692 (2015).

412 9. Nalls, M. A. *et al.* Admixture Mapping of White Cell Count: Genetic Locus
413 Responsible for Lower White Blood Cell Count in the Health ABC and Jackson

414 Heart Studies. *The American Journal of Human Genetics* **82,** 81–87 (2008).

415 10. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development.
416 *Nat Rev Genet* **14,** 204–220 (2013).

417 11. Kulis, M. & Esteller, M. DNA methylation and cancer. *Adv. Genet.* **70,** 27–56
418 (2010).

419 12. Udali, S., Guarini, P., Moruzzi, S., Choi, S.-W. & Friso, S. Cardiovascular
420 epigenetics: from DNA methylation to microRNAs. *Mol. Aspects Med.* **34,** 883–
421 901 (2013).

422 13. Kato, N. *et al.* Trans-ancestry genome-wide association study identifies 12 genetic
423 loci influencing blood pressure and implicates a role for DNA methylation. *Nat.*
424 *Genet.* (2015). doi:10.1038/ng.3405

425 14. Bell, C. G. *et al.* Integrated genetic and epigenetic analysis identifies haplotype-
426 specific methylation in the FTO type 2 diabetes and obesity susceptibility locus.
427 *PLoS ONE* **5,** e14040 (2010).

428 15. Chambers, J. C. *et al.* Epigenome-wide association of DNA methylation markers
429 in peripheral blood from Indian Asians and Europeans with incident type 2
430 diabetes: a nested case-control study. *Lancet Diabetes Endocrinol* **3,** 526–534
431 (2015).

432 16. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an
433 intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31,** 142–147
434 (2013).

435 17. Lardenoije, R. *et al.* The epigenetics of aging and neurodegeneration. *Prog.*
436 *Neurobiol.* **131,** 21–64 (2015).

437 18. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene
438 expression variation in HapMap cell lines. *Genome Biology* **12,** R10 (2011).

439 19. Feil, R. & Fraga, M. F. Epigenetics and the environment: emerging patterns and
440 implications. *Nat Rev Genet* **13,** 97–109 (2011).

441 20. Joubert, B. R. *et al.* 450K epigenome-wide scan identifies differential DNA

442 methylation in newborns related to maternal smoking during pregnancy. *Environ.*
443 *Health Perspect.* **120,** 1425–1431 (2012).

444 21. Joubert, B. R. *et al.* DNA Methylation in Newborns and Maternal Smoking in
445 Pregnancy: Genome-wide Consortium Meta-analysis. *Am. J. Hum. Genet.* **98,**
446 680–696 (2016).

447 22. Jiang, R., Jones, M. J., Sava, F., Kobor, M. S. & Carlsten, C. Short-term diesel
448 exhaust inhalation in a controlled human crossover study is associated with
449 changes in DNA methylation of circulating mononuclear cells in asthmatics. *Part*
450 *Fibre Toxicol* **11,** 71 (2014).

451 23. Ho, S.-M. *et al.* Environmental epigenetics and its implication on disease risk and
452 health outcomes. *ILAR J* **53,** 289–305 (2012).

453 24. Chen, W. *et al.* ADCYAP1R1 and asthma in Puerto Rican children. *Am. J. Respir.*
454 *Crit. Care Med.* **187,** 584–588 (2013).

455 25. Ressler, K. J. *et al.* Post-traumatic stress disorder is associated with PACAP and
456 the PAC1 receptor. *Nature* **470,** 492–497 (2011).

457 26. van der Knaap, L. J. *et al.* Glucocorticoid receptor gene (NR3C1) methylation
458 following stressful events between birth and adolescence. The TRAILS study.
459 *Transl Psychiatry* **4,** e381 (2014).

460 27. Lam, L. L. *et al.* Factors underlying variable DNA methylation in a human
461 community cohort. *Proceedings of the National Academy of Sciences* **109 Suppl**
462 **2,** 17253–17260 (2012).

463 28. Borghol, N. *et al.* Associations with early-life socio-economic position in adult
464 DNA methylation. *International Journal of Epidemiology* **41,** 62–74 (2012).

465 29. Vidal, A. C. *et al.* Maternal stress, preterm birth, and DNA methylation at imprint
466 regulatory sequences in humans. *Genet Epigenet* **6,** 37–44 (2014).

467 30. Nguyen, A. B., Moser, R. & Chou, W. Y. Race and health profiles in the United
468 States: an examination of the social gradient through the 2009 CHIS adult survey.
469 *Public Health* **128,** 1076–1086 (2014).

470 31. Evans, G. W. & Kantrowitz, E. Socioeconomic Status and Health: The Potential
471 Role of Environmental Risk Exposure. *Annual Review of Public Health* **23,** 303–
472 331 (2002).

473 32. Michels, K. B. *et al.* Recommendations for the design and analysis of epigenome-
474 wide association studies. *Nat. Methods* **10,** 949–955 (2013).

475 33. Oh, S. S. *et al.* Effect of secondhand smoke on asthma control among black and
476 Latino children. *J. Allergy Clin. Immunol.* **129,** 1478–83.e7 (2012).

477 34. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell
478 mixture distribution. *BMC Bioinformatics* **13,** 86 (2012).

479 35. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. mediation: R package
480 for causal mediation analysis. *UCLA Statistics/American Statistical Association*
481 (2014).

482 36. Grafodatskaya, D. *et al.* EBV transformation and cell culturing destabilizes DNA
483 methylation in human lymphoblastoid cell lines. *Genomics* **95,** 73–83 (2010).

484 37. Moreno-Estrada, A. *et al.* Human genetics. The genetics of Mexico recapitulates
485 Native American substructure and affects biomedical traits. *Science* **344,** 1280–
486 1285 (2014).

487 38. Galanter, J. M. *et al.* Development of a panel of genome-wide ancestry informative
488 markers to study admixture throughout the Americas. *PLoS Genet.* **8,** e1002554
489 (2012).

490 39. Hernandez, R. D. *et al.* Classic selective sweeps were rare in recent human
491 evolution. *Science* **331,** 920–924 (2011).

492 40. Tang, H. *et al.* Recent Genetic Selection in the Ancestral Admixture of Puerto
493 Ricans. *The American Journal of Human Genetics* **81,** 626–633 (2007).

494 41. Barfield, R. T. *et al.* Accounting for population stratification in DNA methylation
495 studies. *Genet. Epidemiol.* **38,** 231–241 (2014).

496 42. 1000 Genomes Project Consortium *et al.* A map of human genome variation from
497 population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

498 43. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation
499 from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

500 44. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human
501 genomes. *Nature* **526,** 75–81 (2015).

502 45. Galanter, J. M. *et al.* Genome-wide association study and admixture mapping
503 identify different asthma-associated loci in Latinos: the Genes-environments &
504 Admixture in Latino Americans study. *J. Allergy Clin. Immunol.* **134,** 295–305
505 (2014).

506 46. Team, R. C. R: A language and environment for statistical computing.

507 Figure Legends

508 **Figure 1**: [A] Manhattan plot showing the associations between ethnicity and
509 methylation at individual CpG loci. [B] Violin plot showing one such locus, cg19145607.
510 Mexicans are relatively hypermethylated compared to Puerto Ricans ($p = 1.4\times10^{-19}$). [C]
511 Plot showing the association between Native American ancestry at the locus and
512 methylation levels at the locus colored by ethnicity; Native American ancestry accounts
513 for 58% of the association between ethnicity and methylation at the locus.

514 **Figure 2:** [A] Manhattan plot showing the association between local ancestry and
515 methylation at individual CpG loci. [B] Association between cg04922029 on the *DARC*
516 locus and African ancestry, color coded by ethnic group. There is near perfect
517 correlation between the two. [C] Association between SNPs located within 1Mb of
518 cg04922029 and methylation levels at that CpG. [D] Association between rs2814778
519 (Duffy null) genotype and methylation at cg04922029, color coded by the number of
520 African alleles present. There is near perfect correlation between genotype, ancestry and
521 methylation at the locus. [E] Allele frequency of rs2814778 by 1000 Genomes
522 population. The C allele is nearly ubiquitous in African populations and nearly absent
523 outside of African populations and their descendants.

TABLE 1: Baseline characteristics of GALA II participants with methylation data, stratified by ethnicity.

	Mexican	Puerto Rican	Mixed Latino	Other Latino
n	276	220	16	61
Males (%)	125 (45.3%)	127 (57.7%)	6 (37.5%)	28 (45.9%)
Age	11.4 [9.3: 14.7]	12.3 [10.4: 14.2]	11.8 [10.7: 14.9]	11.8 [10: 15.7]
Asthma cases (%)	124 (44.9%)	147 (66.8%)	9 (56.3%)	31 (50.8%)
	Ancestry (n = 524)			
African	4.3% [2.9%: 6.0%]	22.8% [16.6%: 29.4%]	8.5% [5.6%: 19.2%]	12.3% [6.3%: 25.8%]
Native American	55.4% [44.5%: 65.7%]	11.2% [9.8%: 13%]	31.5% [20.9%: 45.6%]	32.8% [10.4%: 49.3%]
European	40.5% [29.9%: 50.2%]	65.7% [59.2%: 71%]	50.5% [44.6%: 57.6%]	48.9% [40%: 58.5%]
	Recruitment Site			
Chicago	140 (50.7%)	15 (6.8%)	11 (68.9%)	15 (24.6%)
New York	18 (6.5%)	10 (4.5%)	1 (6.3%)	23 (37.7%)
Puerto Rico	0	193 (87.7%)	0	0
San Francisco	78 (28.3%)	0	2 (12.5%)	23 (37.7%)
Houston	40 (14.5%)	2 (0.9%)	2 (12.5%)	5 (8.2%)
	Cell Counts (estimated)			
Granulocytes	51.2% [46.0%: 55.7%]	51.6% [46.8%: 57%]	51% [43.6%: 57.2%]	49.1% [43.8%: 55.8%]
Lymphocytes	41.9% [36.9%: 46.6%]	41.8% [36.9%: 46.5%]	41.9% [36.1%: 51.6%]	43.9% [36.8%: 49.6%]
Monocytes	7.1% [5.8%: 8.3%]	6.74% [5.74%: 8.24%]	6.6% [5.7%: 7.6%]	7.4% [6.2%: 8.6%]

Figure 1

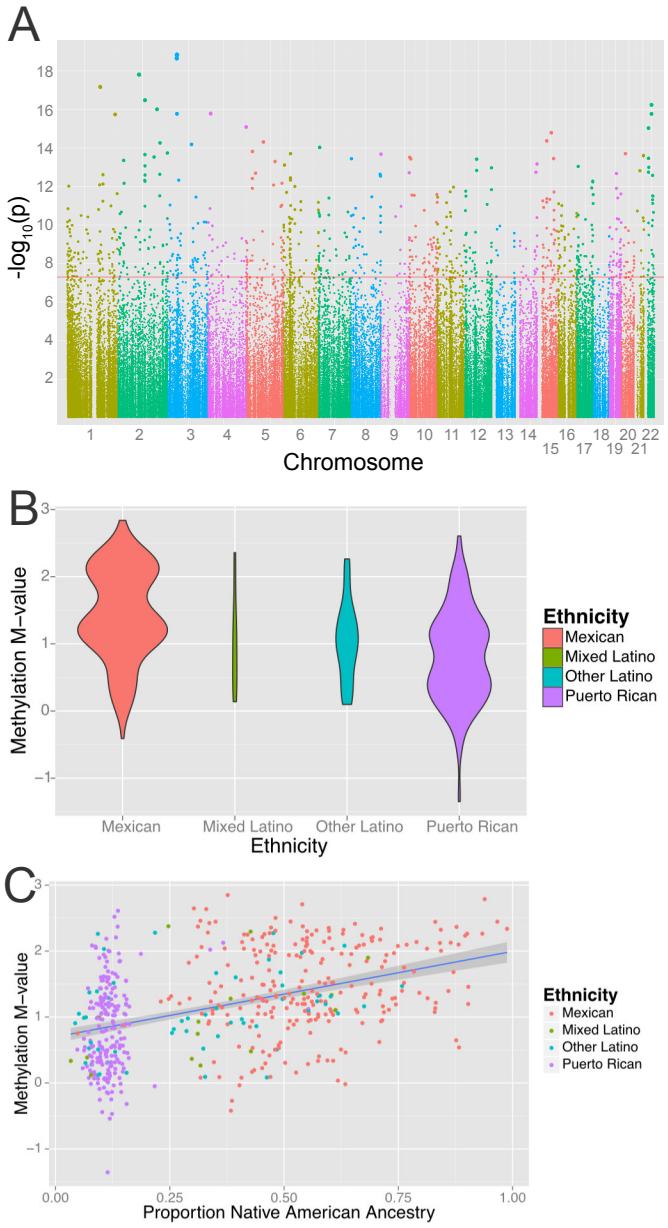
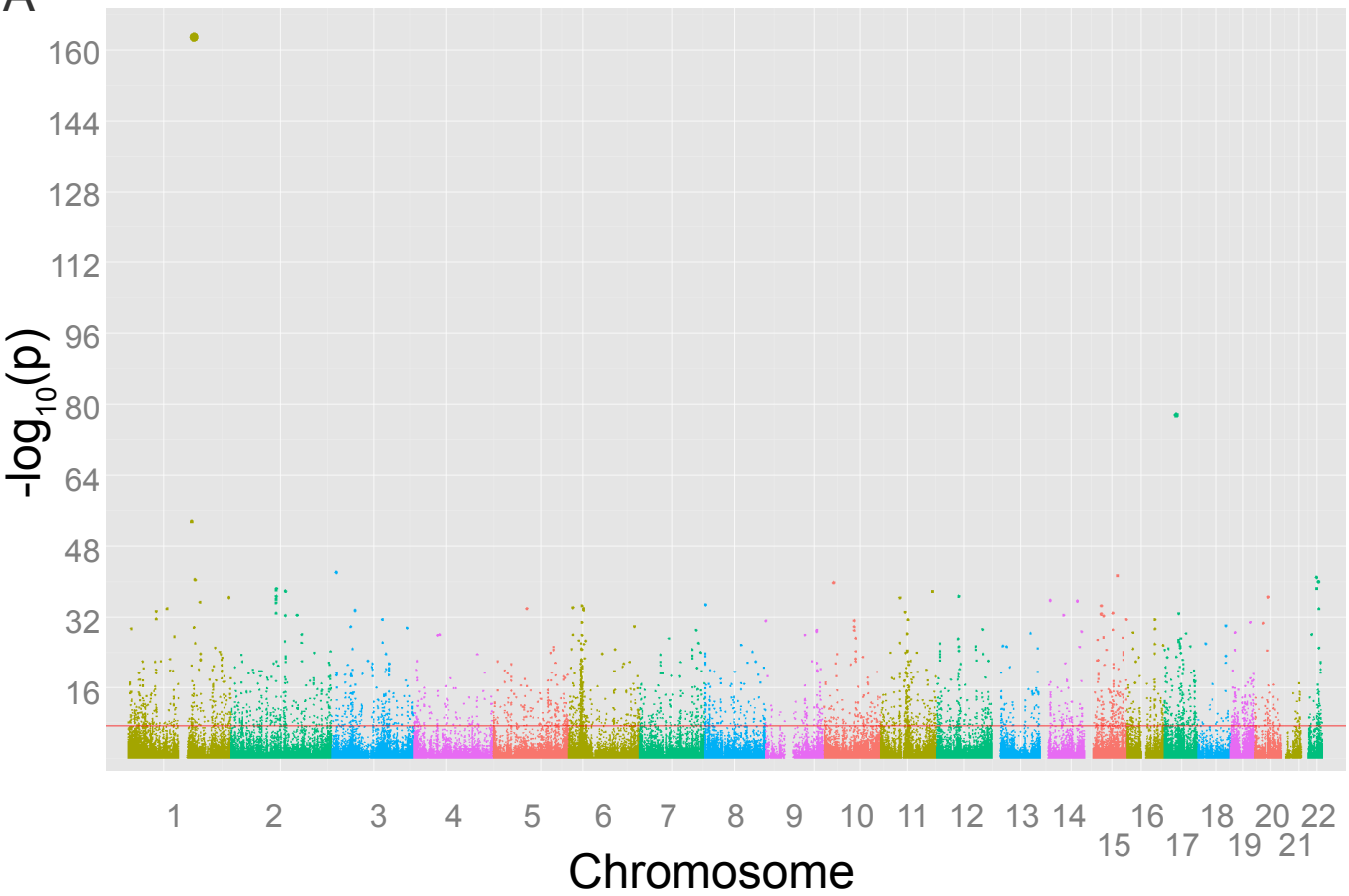
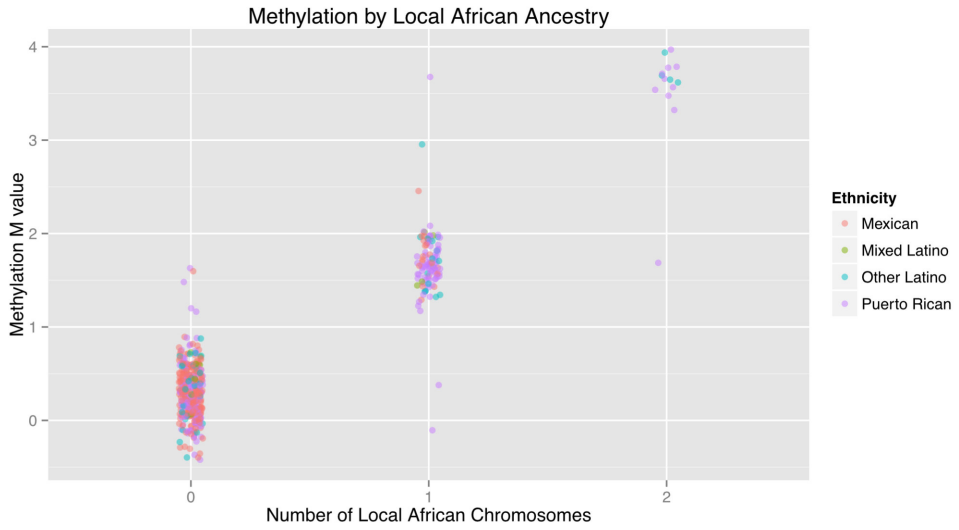

A
-log10(p)
Chromosome
B
Methylation M−value
Mexican
Mixed Latino
Other Latino
Puerto Rican
Ethnicity
C
Methylation M−value
Proportion Native American Ancestry
Ethnicity
Mexican
Mixed Latino
Other Latino
Puerto Rican

Figure 2

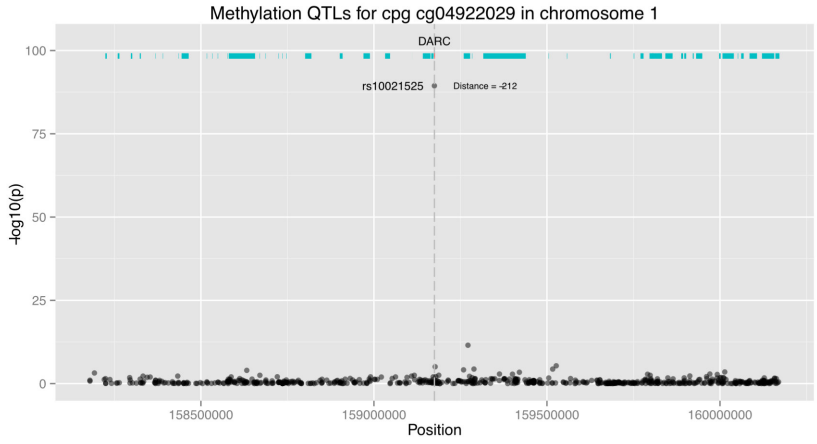
A


-log10(p)
160
144
128
112
96
80
64
48
32
16
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome

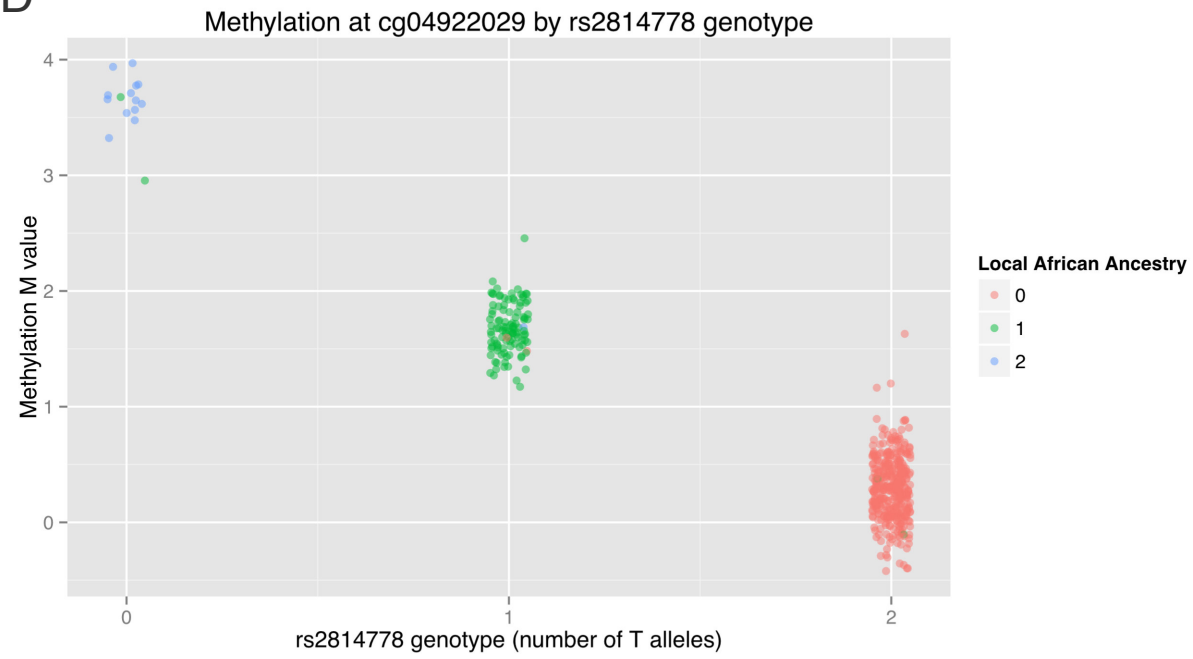
B


Methylation by Local African Ancestry
Methylation M value
4
3
2
1
0
0
1
2
Number of Local African Chromosomes
Ethnicity
Mexican
Mixed Latino
Other Latino
Puerto Rican

C


Methylation QTLs for cpg cg04922029 in chromosome 1
DARC
rs10021525
Distance = -212
-log10(p)
100
75
50
25
0
158500000
159000000
159500000
160000000
Position

D


Methylation at cg04922029 by rs2814778 genotype
Methylation M value
4
3
2
1
0
0
1
2
rs2814778 genotype (number of T alleles)
Local African Ancestry
0
1
2

E

African
Americas
Asian
European
SouthAsian
Proportion
1.00
0.75
0.50
0.25
0.00
ACB ASW ESN GWD LWK MSL YRI
CLM MXL PEL PUR
CDX CHB CHS JPT KHV
CEU FIN GBR IBS TSI
BEB GIH ITU PJL STU
Population
Allele
C
T