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

Joshua M. Galanter, MD, MAS, Christopher R. Gignoux, PhD, Sam S. Oh, PhD, Dara Torgerson, PhD, Maria Pino-Yanes, PhD, Neeta Thakur, MD, MPH, Celeste Eng, BS, Donglei Hu, PhD, Scott Huntsman, MS, Harold J. Farber, MD, Pedro C Avila, MD, Emerita Brigino-Buenaventura, MD, Michael A LeNoir, MD, Kelly Meade, MD, Denise Serebrisky, MD, William Rodríguez-Cintrón, MD, Rajesh Kumar, MD, Jose R Rodríguez-Santana, MD, Max A. Seibold, PhD, Luisa N. Borrell, DDS, PhD, Esteban G. Burchard, MD, MPH, Noah Zaitlen, PhD

SI Appendix Text

Results

Differences in methylation by ethnicity

To explore the effect of departures from a linear association between ancestry and methylation, we incorporated both higher order polynomials and cubic splines of ancestry into our models. We observed a significant departure from linearity ($p < 0.05$) in only 26 (for splines) and 25 (for polynomials) of the 314 CpG's where an association between ethnicity and methylation remained after adjusting for ancestry; however, the association between ethnicity and methylation remained even after adjusting for non-linearity at all sites [SI Appendix Tables 3 and 4].

While most population substructure in Latinos would be expected to arise from differences in continental ancestry[1,2], there is evidence of finer scale (sub-continental) ancestry in Latino populations[3]. We tested for the effect of fine scale substructure by

calculating principal components for all participants with genotyping data using Eigensoft[4]. We found significant associations between principal components 3-10 (PC's 1 and 2 were almost perfectly collinear with ancestry, with an adjusted $R^2 > 0.998$ for all three ancestry proportions, and were therefore excluded) and ethnicity. We therefore added these 8 PC's to models of ethnicity and methylation, and found an association between these genetic PC's and methylation in 63/314 CpG's that had remained associated with ethnicity after adjusting for ancestry. Adjusting for higher order substructure in these CpG's explained the association between ethnicity and methylation in 51 additional loci. This left 263 loci associated with ethnicity after adjustment for ancestry where there was either no association between PC's 3-10 and methylation or the inclusion of these PC's did not affect the association between ethnicity and methylation.[SI Appendix Table 5]

As only 16 participants self-identified as "Mixed Latino", we performed a sensitivity analysis to test the effect of excluding these participants from the analysis and only examining Puerto Ricans, Mexicans, and "Other Latinos". We found that excluding self-identified "Mixed Latino" participants from the analysis did not significantly alter the results in most cases [SI Appendix Table 6]. Of the 916 CpG's associated with ethnicity at a genome-wide scale ($p < 1.6 \times 10\text{-}7$) in models including individuals self-identified as "Mixed Ethnicity", 894 (97.5%) were still significant at a genome-wide scale when "Mixed Latinos" were excluded. All but two of the CpG's that did not meet genome-wide significance were significant when correcting for 916 tests ($p < 5\times10\text{-}5$). In addition, an additional 290 CpG loci that did not meet genome-wide significance in the original analysis were significant at a genome-wide scale when self-identified "Mixed Latinos" were excluded. While these loci did not meet genome-wide significance in the original

analysis that included Mixed Latinos, they all had p-values lower than 2×10^{-6}. Thus we conclude that a sensitivity test excluding individuals of mixed Latino ethnicity did not significantly alter the conclusions.

Environmental differences between geographic locations or recruitment sites are a potential non-genetic explanation for ethnic differences in methylation. We investigated the independent effect of recruitment site on methylation by analyzing the associations between recruitment site and individual methylation loci after adjusting for ethnicity. We did not find any loci significantly associated with recruitment site at a significance threshold of 1.6×10^{-7}. We then performed an analysis to assess the effect of recruitment sites on methylation stratified by ethnicity. We did not find any loci significantly associated with recruitment site and methylation among Mexican participants. We were underpowered to perform a similar analysis for Puerto Ricans because there were only 27 Puerto Rican participants recruited outside of Puerto Rico. To ensure that the absence of association in Mexicans was not due to the loss of power from the smaller sample size, we repeated our analysis of the association between ethnicity and ancestry randomly down-sampling to 276 participants to match the sample size in the analysis of geography in Mexicans. While down-sampling the study to this degree resulted in a loss of power, 128 methylation sites were still associated with ancestry. We conclude that recruitment site was unlikely to be a significant confounder of our associations between ethnicity and methylation and was not a significant independent predictor of methylation.

Ethnic differences in environmentally-associated methylation sites

In an earlier study of maternal smoking in Norwegian newborns[5] that identified 26 loci associated with maternal smoking during pregnancy, 19 passed quality control (QC) in

our own analysis, and the association between methylation and ethnicity was found to be nominally significant ($p < 0.05$)at 6 (31.6%) CpG loci. Adjusting for 19 tests ($p < .0026$), cg23067299 in the aryl hydrocarbon receptor repressor (*AHRR*) gene on chromosome 5 remained statistically significant [SI Appendix Table 8]. These results suggest that ethnic differences in methylation at loci known to be responsive to tobacco smoke exposure *in utero* may be explained in part by ethnic-specific differences in the prevalence of maternal smoking during pregnancy.

We also found that CpG loci previously reported to be associated with diesel-exhaust particle (DEP) exposure[6] were significantly enriched among the set of loci whose methylation levels varied between ethnic groups. Specifically, of the 101 CpG sites that were significantly associated with exposure to DEP and passed QC in our dataset, 31 were nominally associated with ethnicity ($p < 0.05$), and 5 were associated with ethnicity after adjusting for 101 comparisons ($p < 0.005$) . Finally, we found that methylation levels at cg11218385 in the pituitary adenylate cyclase-activating polypeptide type I receptor gene (*ADCYAP1R1*), which had been associated with exposure to violence in Puerto Ricans[7] and with heavy trauma exposure in adults[8], was significantly associated with ethnicity ($p = 0.02$).

Admixture mapping of methylation

We repeated the admixture mapping analysis using methylation beta values [methylated / (methylated + unmethylated)] instead of methylation M-values [log2(ß / (1 - ß)]. We report these results in SI Appendix Table 11 and note that they did not significantly alter our findings; 3695 loci were associated with local ancestry, compared to 3694 when the analysis was done on the methylation M scale. The most significantly associated CpG in the admixture mapping analysis remained cp04922029 on the Duffy Locus, with a p-

value of 4×10^{-152}, only slightly less significant than the 6×10^{-162} significance level found for that locus using methylation M-value. Each increase in African ancestry was associated with an increase in methylation ß of 0.37.

We also repeated the mQTL analysis using methylation ß values instead of M-values in SI Appendix Table 12. Of the 3694 loci significantly associated with local ancestry, 3631 (98.3%) have at least one SNP within 1 Mb that is significantly associated with methylation levels (after adjustment of the number of SNPs in cis-), compared to 3637 loci when the analysis was performed with M-values. The most significant SNP-CpG pair was cg17857094/KG_6_31014327 (rs56366011), which has a p-value of 10^{-354}; each copy of the C allele was associated with a decrease in methylation ß of 0.31. The cp04922029/rs2814778 was also highly significant, but not as significant as in the original analysis; the p-value was 2×10^{-65}; each copy of the T allele was associated with an increase of methylation ß of 0.20.

SI Appendix Methods

Recruitment

A total of 4,702 children (2,374 participants with asthma and 2,328 healthy controls) were recruited from five centers (Chicago, Bronx, Houston, San Francisco Bay Area, and Puerto Rico) using a combination of community- and clinic-based recruitment. Participants were eligible if they were 8-21 years of age and self-identified as a specific Latino ethnicity and had four Latino grandparents. Asthma cases were defined as participants with a history of physician diagnosed asthma and the presence of two or more symptoms of coughing, wheezing, or shortness of breath in the 2 years preceding enrollment. Participants were excluded if they reported any of the following: (1) 10 or

more pack-years of smoking; (2) any smoking within 1 year of recruitment date; (3) history of lung diseases other than asthma (cases) or chronic illness (cases and controls); or (4) pregnancy in the third trimester. Further details of recruitment are described elsewhere[9]. Latino sub-ethnicity was determined by self-identification and the ethnicity of the their four grandparents. Due to small numbers, ethnicities other than Puerto Rican and Mexican were collapsed into a single category, "other Latino". Participants whose four grandparents were of discordant ethnicity were considered to be of "mixed Latino" ethnicity.

Trained interviewers, proficient in both English and Spanish, administered questionnaires to gather baseline demographic data, as well as information on general health, asthma status, acculturation, social, and environmental exposures.

Methylation

1 μg of gDNA was bisulfite-converted using the Zymo EZ DNA Methylation Kit™ (Zymo research, Irvine, CA) according to the manufacturer's instructions. Bisulfite converted DNA was isothermally amplified overnight, enzymatically fragmented, precipitated, and re-suspended in hybridization buffer. The fragmented, re-suspended DNA samples were dispensed onto Infinitum HumanMethylation450 BeadChips and incubated overnight in an Illumina hybridization oven. Following hybridization, free DNA was washed away, and the BeadChips were extended through single nucleotide extensions with fluorescent labels. The BeadChips were imaged using an Illumina iScan system, and processed using the Illumina GenomeStudio Software.

Failed probes were identified using detection p-values using Illumina's recommendations. Probes on sex chromosomes and those known to contain genetic

polymorphisms in the probe sequence were also excluded, leaving 321,503 probes for analysis. Raw data were normalized using Illumina's control probe scaling procedure. Beta values of methylation (ranging from 0 to 1) were converted to M-values via a logit transformation[10].

Genotyping

Participants were genotyped at 818,154 SNPs on the Axiom® Genome-Wide LAT 1, World Array 4 (Affymetrix, Santa Clara, CA)[11]. We removed SNPs with >5% missing data and failing platform-specific SNP quality criteria (n=63,328), along with those out of Hardy-Weinberg equilibrium (n=1845; p<10-6) within their respective populations (Puerto Rican, Mexican, and other Latino), as well as non-autosomal SNPs. Subjects were filtered based on 95% call rates and sex discrepancies, identity by descent and standard Affymetrix Axiom metrics. The total number of participants passing QC was 3,804 (1,902 asthmatic cases, 1,902 healthy controls), and the total number of SNPs passing QC was 747,129. The number of participants with both methylation and genotyping data was 524.

Ancestry and PCA calculations

GALA II participants were combined with ancestral data from 1000 Genomes European (CEU) and African (YRI) populations and 71 Native American (NAM) samples genotyped on the Axiom® Genome-Wide LAT 1 array. A final sample of 568,037 autosomal SNPs with relevant ancestral data was used to estimate local and global ancestry. Global ancestry was estimated using the program ADMIXTURE[12], with a three population model. Local ancestry at all positions across the genome was estimated using the program LAMP-LD[13], assuming three ancestral populations.

Principal components for the genetic data were determined using the program EIGENSTRAT[4].

Statistical Analyses

Multidimensional scaling of the logit transformed methylation data (M-values) was performed by first calculating the Euclidian distance matrix between each pair of individuals and then calculating the first 10 principal coordinates of the data [SI Appendix Figure 2A]. We performed both a simple correlation analysis of these principal coordinates to demographic factors (age, sex, ethnicity), estimated cell counts and technical factors (batch, plate, and position) to identify factors that correlated with global methylation patterns [see SI Appendix Figure 2B]. In addition, we performed a multiple regression analysis of methylation principal coordinates by ethnicity and ancestry, adjusting for case status, age, sex, estimated cell counts, and plate and position [SI Appendix Table 1].

We also sought to establish the extent to which global differences in methylation between Puerto Ricans and Mexicans could be explained by differences in ancestry between the two groups. We estimated the proportion of the ethnicity association that was mediated by genomic ancestry using the R package "mediation"[14] for methylation principal coordinates, which demonstrated a significant association with ethnicity.

We also sought to correlate ethnicity and methylation at a locus-specific level. We thus performed a linear regression between methylation at each CpG site and self-reported ethnicity (Mexican, Puerto Rican, Mixed Latino, and Other Latino), followed by a three degree of freedom analysis of variance to determine the overall effect of ethnicity on methylation We repeated the analysis excluding the 16 participants that were self-

described as “Mixed Latino”, and tested for non-linearity in two ways: by adding second and third order polynomials to the model, and by adding a 3-degree of freedom cubic spline and comparing models with the non-linear terms to those without using a nested ANOVA. At loci where there was evidence for non-linearity, we tested whether ethnicity remained associated with methylation after adjusting for ancestry as well as the deviations from linearity. Finally, we tested for the presence of population sub-structure beyond that conveyed through ancestry by adding the genetic principal components 3-10 (PCs 1 and 2 were co-linear with ancestry with a correlation coefficient $R2 > 0.998$) and comparing models with those PCs to those without. At loci where there was evidence for association between PC’s 3-10 and methylation, we tested whether ethnicity remained associated with methylation after adjusting for ancestry as well as the PC’s 3-10.

We calculated the proportion of variance in methylation explained by ethnicity and genomic ancestry at each site where ethnicity was significantly associated with methylation. To do this, we fit a model that included both ethnicity and global ancestry as well as the confounders described above and calculated the proportion of variance explained by multiplying the ratio of the variance between predictors (ethnicity and genomic ancestry) and outcome (methylation) by the square of the effect magnitude (ß).

We also examined whether differences in methylation patterns by ethnicity could be associated with known loci that had previously been reported to vary based on common environmental exposures, including maternal smoking during pregnancy[5], diesel exhaust particles (DEP)[6], and exposure to violence[7]. We have previously shown that

exposure to these common environmental exposures or similar exposures varied by ethnicity within our own GALA II study populations[9,15,16].

In addition, we examined the association between global ancestry and methylation across all CpG loci using a two-degree of freedom likelihood ratio test as well as by examining the association between individual ancestral components (African, European, and Native American) and methylation at each CpG site. At each site where methylation was significantly associated with genomic ancestry proportions, we determined the relative effect of global ancestry (θ, theta) and local ancestry (γ, gamma) in a joint model by calculating the proportion of variance explained as above.

To determine whether ancestry associations with methylation were due to variation in local ancestry, we performed a cis-admixture mapping study, comparing estimates of local ancestry at each CpG site with methylation at the site. Because ancestry LD is much stronger than genotypic LD, it is possible to accurately interpolate ancestry at each CpG site based on the ancestry estimated at the nearest SNPs[17,18]. Measures of locus-specific ancestry were correlated with local methylation using linear regression. We performed a two-degree of freedom analysis of variance test evaluating the overall effect of all three ancestries as well as single-ancestry associations comparing methylation at a given locus with the number of African, European and Native American chromosomes at that CpG site.

In order to determine the extent to which admixture mapping results could be explained by allelic associations, we performed a meQTL analysis at all Bonferroni-corrected significant admixture mapping associations ($p < 1.6 \times 10\text{-}7$), by comparing methylation at a given locus with the genotype of SNPs within 1 MB of the CpG site using an additive

genotypic model, adjusted for both global and local genomic ancestry, demographic variables including ethnicity, estimated cell proportions, case status, and technical factors. The significance threshold was based on Bonferroni correction for the number of SNPs within 1 MB of the CpG site.

SI Appendix Figures

SI Appendix Figure 1: Ancestry estimates for GALA II participants, by ethnic group. Mexicans, on average, had a greater proportion of Native American ancestry than Puerto Ricans; Puerto Ricans had a greater proportion of European and African ancestry. Mixed and other Latinos were intermediate.

SI Appendix Figure 2: [A] Distribution of the first 10 principal coordinates of the methylation data. Plots in the diagonal show the univariate distribution; those in the lower left triangle show bivariate relationship between each pair of PCs, while those in the upper right show the bivariate density. [B] Bivariate or ANOVA associations between principal coordinates and technical factors (chip, position), cell counts, genetic ancestry (European, Native American, African), recruitment site (New York, NY, San Francisco, CA, Chicago, IL, Houston, TX, and Puerto Rico), demographic factors (ethnicity, age, sex), and case status. [C] Correlation coefficients between the various factors and principal coordinates.

SI Appendix Figure 3: [A] Association between ethnicity and principal coordinate 7. [B] Association between Native American ancestry proportion and PC7, colored by ethnicity. Native American ancestry explains approximately 81% of the association between PC7 and ethnicity.

SI Appendix Figure 4: Relationship between genomic ancestry and the association between ethnicity and methylation. [A] Venn diagram showing the effect of adjustment for ancestry on the association between ethnicity and methylation. The components of the diagram represent the number of CpG's that remained associated with ethnicity after adjustment for ancestry and the number of CpG's that were associated with ancestry. [B] Relative proportion of variance in methylation explained by ethnicity and

genomic ancestry across loci significantly associated with ethnicity. Mediation analysis of associations between ethnicity and methylation M-values for [C] Native American ancestry and [D] African ancestry. For simplicity, only significant mediation effects are shown.

SI Appendix Figure 5: [A] Manhattan plot showing the associations between genomic ancestry and methylation at individual CpG loci. [B] Plot showing one such locus, cg04922029, and genomic African ancestry, showing a strong correlation between African ancestry and hypermethylation at that site.

SI Appendix Figure 6: Relative proportion of variance in methylation explained by global and local ancestry across loci significantly associated with global ancestry.

SI Appendix Figure 7: [A] Violin plot showing the association between cg25134647 on chromosome 12 and European ancestry at the locus. [B] Association between SNPs located within 1Mb of cg25134647 and methylation levels at that CpG. [C] Association between rs4963867 genotype and methylation at cg25134647, color coded by the number of European alleles present. There is near perfect correlation between genotype and methylation at the locus. [D] Allele frequency of cg25134647by 1000 Genomes population. The C allele is more common in African populations than in other populations.

SI Appendix Tables

SI Appendix Table 1: Correlation between methylation principal components and both ethnicity and ancestry, association between ethnicity and methylation adjusted for ancestry, and mediation of the association between ethnicity and methylation by ancestry.

SI Appendix Table 2: Significant associations between ethnicity and methylation ($p < 1.6 \times 10^{-7}$), and effect of adjustment for ancestry on the association of ethnicity and methylation.

SI Appendix Table 3: Effect of adding cubic spline ancestry terms to the association between ethnicity and methylation.

SI Appendix Table 4: Effect of adding quadratic and cubic ancestry terms to the

association between ethnicity and methylation.

SI Appendix Table 5: Effect of adding genetic principal components 3-10 to the association between ethnicity and methylation.

SI Appendix Table 6: Significant associations between ethnicity and methylation ($p < 1.6 \times 10^{-7}$), and effect of adjustment for ancestry on the association of ethnicity and methylation, excluding participants of "Mixed Latino" ethnicity.

SI Appendix Table 7: Association of ethnicity and methylation in loci previously associated with maternal smoking during pregnancy.

SI Appendix Table 8: Significant associations between ethnicity and methylation loci previously associated with environmental exposures.

SI Appendix Table 9: Significant associations between global ancestry and methylation, and effect of adjustment for local ancestry on the association between global ancestry and methylation.

SI Appendix Table 10: Significant associations between local ancestry and methylation (cis- admixture mapping), and effect of adjustment for local ancestry on the association between global ancestry and methylation.

SI Appendix Table 11: Significant associations between local ancestry and methylation (cis- admixture mapping) using the methylation ß scale (proportion of DNA that is methylated), and effect of adjustment for local ancestry on the association between global ancestry and methylation.

SI Appendix Table 12: mQTLs within 1 Mb of admixture mapping loci.

SI Appendix Table 13: mQTLs within 1 Mb of admixture mapping loci on the ß scale and.

SI Appendix References

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

2. Bryc, K. *et al.* Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences* **107,** 8954–8961 (2010).

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

4. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLoS Genet.* **2,** e190 (2006).

5. Joubert, B. R. *et al.* 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.* **120,** 1425–1431 (2012).

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

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

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

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

10. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11,** 587 (2010).

11. Hoffmann, T. J. *et al.* Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* 1–9 (2011). doi:10.1016/j.ygeno.2011.08.007

12. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19,** 1655–1664 (2009).

13. Baran, Y. *et al.* Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **28,** 1359–1367 (2012).

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

15. Nishimura, K. K. *et al.* Early-life air pollution and asthma risk in minority children. The GALA II and SAGE II studies. *Am. J. Respir. Crit. Care Med.* **188,** 309–318 (2013).

16. Thakur, N. *et al.* Socioeconomic status and childhood asthma in urban minority youths. The GALA II and SAGE II studies. *Am. J. Respir. Crit. Care Med.* **188,** 1202–1209 (2013).

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

18. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat Rev Genet* **11,** 356–366 (2010).