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

Epigenetic modification of the genome through methylation plays a key role in the regulation of diverse cellular processes[10]. Changes in DNA methylation patterns have been associated with complex diseases, including various cancers[11], cardiovascular disease[12,13], obesity[14], diabetes[15], autoimmune and inflammatory diseases[16], and neurodegenerative diseases[17]. Epigenetic changes are thought to reflect influences of both genetic[18] and environmental factors[19]. The discovery of methylation quantitative trait loci (meQTL's) across populations by Bell *et al.* established the influence of genetic factors on methylation levels in a variety of tissue types[18], with meQTL's explaining between 22% and 63% of the variance in methylation levels. Multiple environmental factors have also been shown to affect methylation levels, including endocrine disruptors, tobacco smoke[20,21], polycyclic aromatic hydrocarbons, infectious pathogens, particulate matter, diesel exhaust particles[22], allergens, heavy metals, and other indoor and outdoor pollutants[23]. Psychosocial factors, including measures of traumatic experiences[24-26], socioeconomic status[27,28], and general perceived stress[29], also affect methylation levels. Racial and ethnic categories reflect the shared experiences and exposures to known risk factors for disease, such as air pollution and tobacco smoke, poverty, and inadequate access to medical services, which have all contributed to worse disease outcomes in certain populations[30,31]. Since environmental exposures affect methylation, they may be reflected in systematic differences in methylation levels between racial or ethnic groups that cannot be captured by using genetically defined ancestry alone.

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

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

## Results

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

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

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

An epigenome-wide association study of self-identified ethnicity (see methods for details of ascertainment of ethnicity) and methylation identified a significant difference in methylation M-values between ethnic groups at 916 CpG sites at a Bonferroni-corrected significance level of less than $1.6\times10^{-7}$ [Figure 1A and SI Appendix Table 2]. The most significant association with ethnicity occurred at cg12321355 in the ABO blood group gene (*ABO*) on chromosome 3 (p-ANOVA $6.7\times10^{-22}$) [Figure 1B]. A two degree of freedom ANOVA test for genomic ancestry was also significantly associated with methylation level at this site (p = $2.3\times10^{-5}$) [Figure 1C], and when the analysis was stratified by ethnic sub-group, showed an association in both Puerto Ricans and Mexicans ($p = 0.001$ for Puerto Ricans, $p = 0.003$ for Mexicans). Although adjusting for genomic ancestry attenuated the effect of ethnicity, a significant association between ethnicity and methylation remained ($p = 0.04$). Recruitment site, an environmental exposure proxy, was not significantly associated with methylation at this locus ($p = 0.5$), suggesting that environmental differences associated with ethnicity beyond geography and ancestry are driving the association.

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

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

We also found 194 loci with a significant association between global genetic ancestry and methylation levels (after adjusting for ethnicity) at a Bonferroni corrected association p-value of less than $1.6\times10^{-7}$ [SI Appendix Figure 5 and SI Appendix Table 9], including 48 that were associated with ethnicity in our earlier analysis. Of these significant associations, 55 were driven primarily by differences in African ancestry, 94 by differences in Native American ancestry, and 45 by differences in European ancestry. The most significant association between methylation and ancestry occurred at cg04922029 in the Duffy antigen receptor chemokine gene (DARC) on chromosome 1

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

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

In an admixture mapping study, we find that methylation at 3,694 of 321,503 CpG's (1.1%), was significantly associated with ancestry at the CpG site at a Bonferroni corrected association p-value of less than $1.6\times10^{-7}$ [Figure 2A and SI Appendix Table

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

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

The SNP/CpG pair were separated by a median distance of 10.9 kb (interquartile range 2.9 kb to 35.1 kb). The furthest SNP/CpG pair were 998 kb apart. The most significant SNP/CpG pair were cg25134647/rs4963867, on chromosome 12, which are separated by 412 base pairs [SI Appendix Figure 6A/B]. Each copy of the T allele was associated with a decrease in M-value of 3.58, corresponding to a nearly 12-fold decrease in the ratio of methylated to unmethylated DNA at the site [SI Appendix Figure 6C] ($p < 10^{-370}$). We found that CpG cg04922029 (our top admixture mapping association) was significantly correlated with SNP rs2814778 [Figure 2C], the Duffy null mutation, 212 base pairs away; each copy of the C allele was associated with an increase in M-value 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 methylation ß values instead of M-values [See SI Appendix text and SI Appendix Table 13]

## Discussion

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

We conclude that systematic environmental differences between ethnic subgroups likely play an important role in shaping the methylome for both individuals and populations. Loci previously associated with diverse environmental exposures such as *in utero* exposure to tobacco smoke[20,21], as well as diesel exhaust particles[22] and psychosocial stress[24] were enriched in our set of loci where methylation was associated with ethnicity. Twenty-seven of the loci associated with maternal smoking during pregnancy in a large consortium meta-analysis[21] were associated with methylation at a genome-wide significance threshold of $1.6\times10^{-7}$. Thus, inclusion of relevant social and environmental exposures in studies of methylation may help elucidate racial/ethnic disparities in disease prevalence, health outcomes and therapeutic response. However, in many cases, a detailed environmental exposure history is unknown, unmeasurable or poorly quantifiable, and race/ethnicity becomes a useful, albeit imperfect proxy.

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

Our approach has some potential limitations. It is possible that fine-scale population structure (sub-continental ancestry) within European, African, and Native American populations may contribute to ethnic differences in methylation, as we had previously reported in the case of lung function[37]. However, despite the presence of additional substructure among the GALA II participants, PC's 3-10 explained the association between ethnicity and ancestry at only 51 loci. PCs from chip-based genotypes will not capture all forms of genetic variation. Clusters of ethnicity specific rare variants of large effect or strong ethnicity-specific selective sweeps in the last 8-12 generations[38] could also give rise to methylation differences, but these are inconsistent with existing rare variant and selection analyses[39,40]. Our models of genetic ancestry assumed a linear effect of ancestry on methylation, whereas a nonlinear association or other model misspecification could have led to incomplete adjustment for genetic ancestry, and thus, led to a residual association between ethnicity and methylation. However, when we added second and third order polynomials or cubic splines to our models, we found evidence for a nonlinear association between ancestry and methylation at only 25 and 26 loci, respectively, and it did not affect the association between ethnicity and methylation. Although it is impossible to account for all types of non-linearity and non-additivity (such as gene by gene or gene by environment interaction), our analysis suggests that non-linear effects are unlikely to be significant. To rule out any residual confounding due to recruitment sites, we conducted an additional analysis on the effect of recruitment site on methylation both for the overall study and for the Mexican participants (the largest study population in this analysis). We observed no significant independent effect of recruitment site suggesting that confounding due to recruitment region was limited, at least within the United States. We were unable to test for the effect of geographic differences between the United States and Puerto Rico because our study included relatively few Puerto Ricans recruited outside of Puerto Rico.

The presence of a strong association between genetic ancestry and methylation raises the possibility that epigenetic studies can be confounded by population stratification, similar to genetic association studies, and that adjustment for either genetic ancestry or selected principal components is warranted. This possibility was first demonstrated in a previous analysis of the association between self-described race and methylation[41]. However, the study only evaluated two distinct racial groups (African Americans and Whites), while the present study demonstrates the possibility of population stratification in an admixed and heterogeneous population with participants from diverse Latino national origins. The tendency to consider Latinos as a homogenous or monolithic ethnic group makes any analysis of this population particularly challenging. Our finding of loci whose methylation patterns differed between Latino ethnic subgroups, even after adjusting for genetic ancestry, suggests that any analysis of these populations in disease-association studies without adjusting for ethnic heterogeneity is likely to result in spurious associations even after controlling for genomic ancestry.

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

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

Our findings have important implications for the independent and joint effects of race, ethnicity, and genetic ancestry in biomedical research and clinical practice, especially in studies conducted in diverse or admixed populations. Our conclusions may be generalizable to any population that is racially mixed such as those from South Africa, India, and Brazil, though we would encourage further study in diverse populations. As the National Institutes of Health (NIH) embarks on a precision medicine initiative, this research underscores the importance of including diverse populations and studying factors capturing the influence of social, cultural, and environmental factors, in addition to genetic ones, upon disparities in disease and drug response.

## Methods

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

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

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

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

The Bonferroni method was used to adjust for multiple comparisons. For methylome-wide associations, the significance threshold was adjusted for 321,503 probes, resulting in a Bonferroni threshold of $1.6\times10^{-7}$. Analyses were performed using R version 3.2.1 (The R Foundation for Statistical Computing)[46] and the Bioconductor package version 2.13. Further details on the models used for specific statistical analyses is in the SI Appendix Text.

## Acknowledgements

The authors acknowledge the families and patients for their participation and thank the numerous health care providers and community clinics for their support and participation in GALA II. In particular, the authors thank study coordinator Sandra Salazar; the recruiters who obtained the data: Duanny Alva, MD, Gaby Ayala-Rodriguez, Lisa Caine, Elizabeth Castellanos, Jaime Colon, Denise DeJesus, Blanca Lopez , Brenda Lopez, MD, Louis Martos, Vivian Medina, Juana Olivo, Mario Peralta, Esther Pomares, MD, Jihan Quraishi, Johanna Rodriguez, Shahdad Saeedi, Dean Soto, Ana Taveras. We also thank Sasha Gusev for helpful discussion. Computations in this manuscript were performed using the UCSF Biostatistics High Performance Computing System.

## References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

20. 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).

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

22. 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).

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

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

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

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

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

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

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

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

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

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

33. 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).

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

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

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

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

38. 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).

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

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

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

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

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

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

45. 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).

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

## Figure Legends

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

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

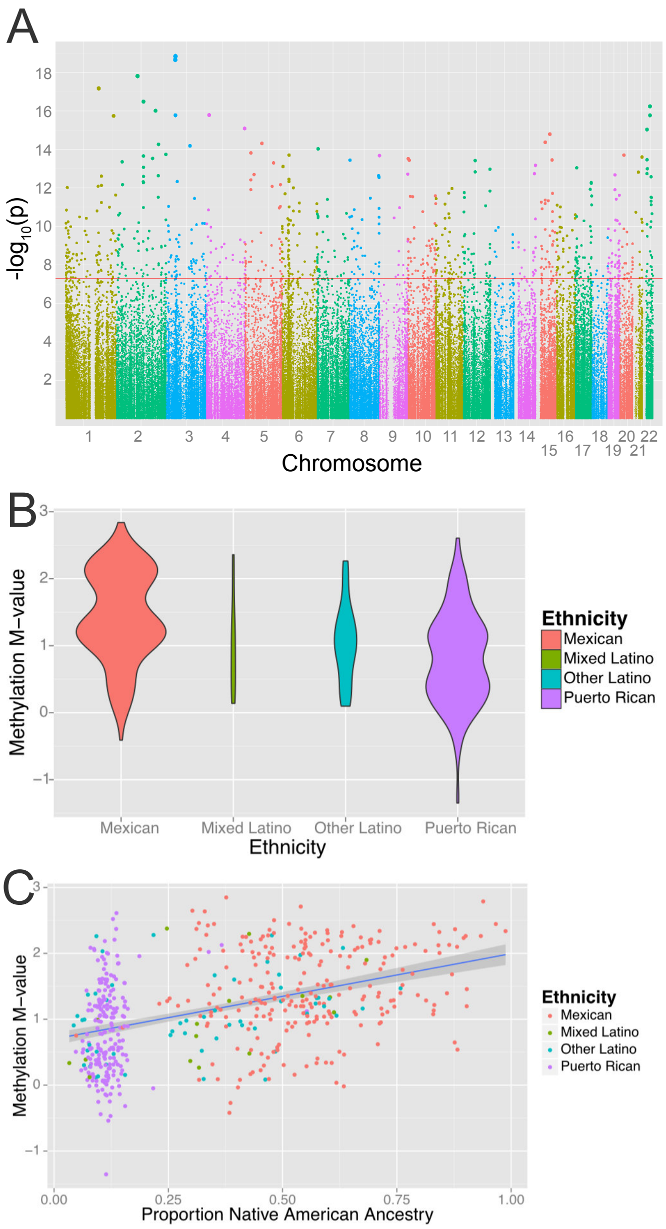
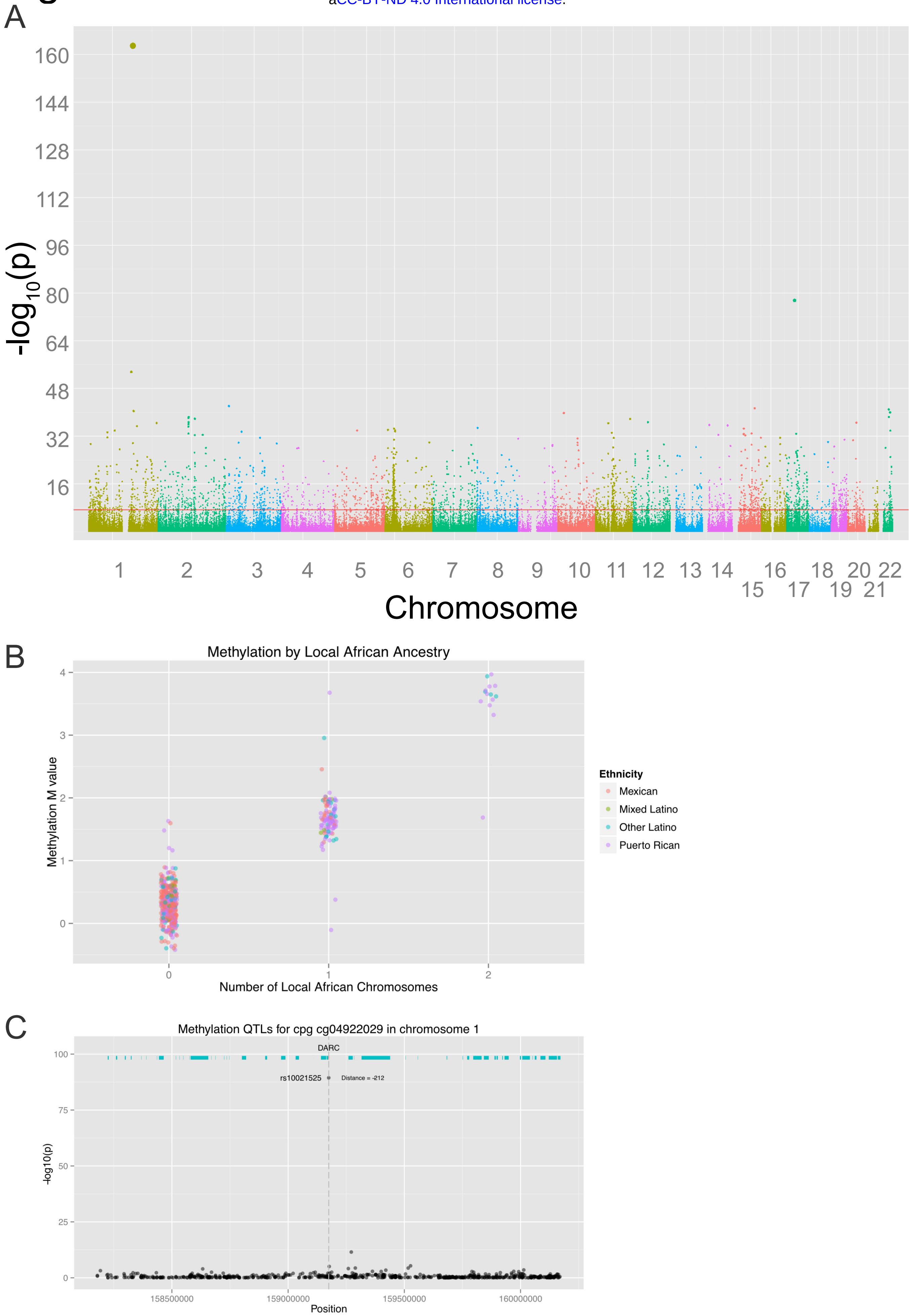
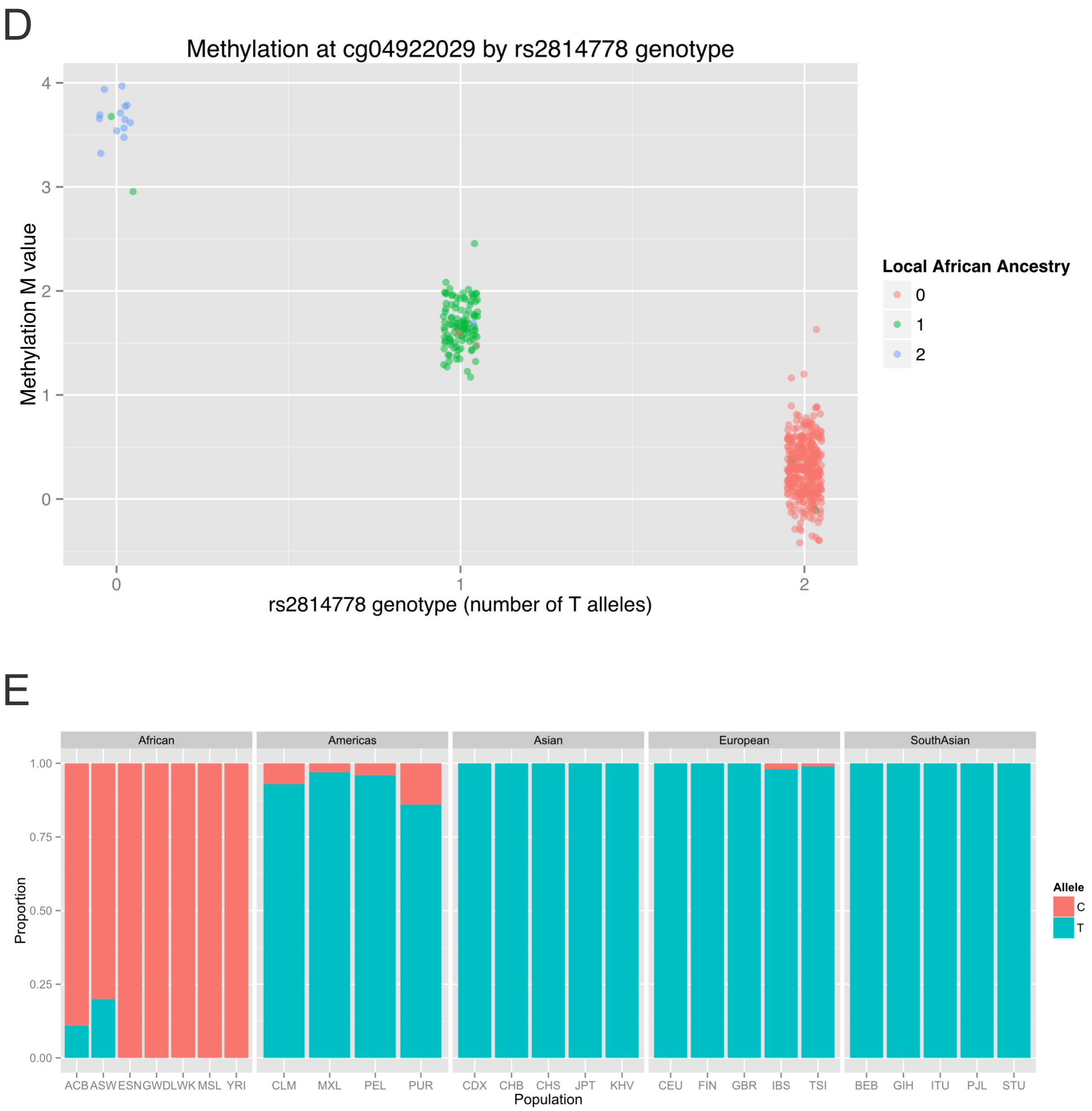

A
-log10(p)
18
16
14
12
10
8
6
4
2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome
B
Methylation M−value
3
2
1
0
−1
Mexican
Mixed Latino
Other Latino
Puerto Rican
Ethnicity
Ethnicity
🟥 Mexican
🟩 Mixed Latino
🟦 Other Latino
🟪 Puerto Rican
C
Methylation M−value
3
2
1
0
−1
0.00
0.25
0.50
0.75
1.00
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