

City University of New York (CUNY)

CUNY Academic Works

Publications and Research

CUNY Graduate Center

2017

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

Joshua M. Galanter
University of California, San Francisco

Christopher R. Gignoux
Stanford University

Sam S. Oh
University of California, San Francisco

Dara Torgerson
University of California, San Francisco

Maria Pino-Yanes
Hospital Universitario Nuestra Senora de Candelaria

See next page for additional authors

[How does access to this work benefit you? Let us know!](#)

More information about this work at: https://academicworks.cuny.edu/gc_pubs/500

Discover additional works at: <https://academicworks.cuny.edu>

This work is made publicly available by the City University of New York (CUNY).
Contact: AcademicWorks@cuny.edu

Authors

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

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

Joshua M Galanter^{1,2,3*†}, Christopher R Gignoux⁴, Sam S Oh^{1,2,3}, Dara Torgerson², Maria Pino-Yanes^{5,6}, Neeta Thakur¹, Celeste Eng¹, Donglei Hu¹, Scott Huntsman¹, Harold J Farber⁷, Pedro C Avila⁸, Emerita Brigino-Buenaventura⁹, Michael A LeNoir¹⁰, Kelly Meade¹¹, Denise Serebrisky¹², William Rodríguez-Cintrón¹³, Rajesh Kumar¹⁴, Jose R Rodríguez-Santana¹⁵, Max A Seibold¹⁶, Luisa N Borrell¹⁷, Esteban G Burchard^{1,2*†}, Noah Zaitlen^{1*†}

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

*For correspondence: galanter@protonmail.com (JMG); esteban.burchard@ucsf.edu (EGB); noah.zaitlen@ucsf.edu (NZ)

†These authors contributed equally to this work

Present address: [†]Genentech, South San Francisco, United States

Competing interests: The authors declare that no competing interests exist.

Funding: See page 19

Received: 10 August 2016

Accepted: 23 November 2016

Published: 03 January 2017

Reviewing editor: Magnus Nordborg, Austrian Academy of Sciences, Austria

© Copyright Galanter et al. This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

Abstract Populations are often divided categorically into distinct racial/ethnic groups based on social rather than biological constructs. Genetic ancestry has been suggested as an alternative to this categorization. Herein, we typed over 450,000 CpG sites in whole blood of 573 individuals of diverse Hispanic origin who also had high-density genotype data. We found that both self-identified ethnicity and genetically determined ancestry were each significantly associated with methylation levels at 916 and 194 CpGs, respectively, and that shared genomic ancestry accounted for a median of 75.7% (IQR 45.8% to 92%) of the variance in methylation associated with ethnicity. There was a significant enrichment ($p=4.2 \times 10^{-64}$) of ethnicity-associated sites amongst loci previously associated environmental exposures, particularly maternal smoking during pregnancy. We conclude that differential methylation between ethnic groups is partially explained by the shared genetic ancestry but that environmental factors not captured by ancestry significantly contribute to variation in methylation.

DOI: [10.7554/eLife.20532.001](https://doi.org/10.7554/eLife.20532.001)

eLife digest Whether a person develops a particular disease can depend on both genetic and environmental factors. Many studies have found that people of different races and ethnicities have different likelihoods of acquiring certain diseases. Race and ethnicity are social constructs; that is, they are not necessarily defined biologically. However, shared ancestry will produce genetic links between members of a group. In addition, members of an ethnic group often share a culture or environment that may influence their risk of disease. For example, the 'Mediterranean diet' inspired by the dietary habits of Southern Italians has been shown to reduce the risk of heart disease, diabetes and cancer.

The addition of chemical groups – such as methyl groups – to DNA strands can affect the activity of nearby genes. Methylation is controlled by both genetic and environmental factors, and altered patterns of DNA methylation are seen in some diseases. It is therefore an ideal biological process to study to determine how race/ethnicity and ancestry contribute to a person's susceptibility to disease.

Galanter et al. have now studied the patterns of methylation found in the blood of 573 people from diverse Latino ethnic sub-groups. The different groups displayed significantly different patterns of methylation at hundreds of locations across the genome. Genetic ancestry explained approximately 75% of the variation in methylation between the sub-groups. In addition, the methylation patterns at DNA locations known to be affected by environmental exposures – for example, by exposure to tobacco while in the womb – were disproportionately likely to be methylated differently in different sub-groups.

Now that more is known about the relative effects of race/ethnicity and genetic ancestry on methylation, the next step is to apply this knowledge to disease processes. This will help us to better understand the source of health disparities across different groups of people.

DOI: [10.7554/eLife.20532.002](https://doi.org/10.7554/eLife.20532.002)

Introduction

Race, ethnicity, and genetic ancestry have had a complex and often controversial history within biomedical research and clinical practice (*Risch et al., 2002; Cooper et al., 2003; Yudell et al., 2016; Burchard et al., 2003; Phimister, 2003*). For example, race- and ethnicity-specific clinical reference standards are based on population-based sampling on a given physical trait such as pulmonary function (*Hankinson et al., 1999; Quanjer et al., 2012*). However, because race and ethnicity are social constructs and poor markers for genetic diversity, they fail to capture the heterogeneity present within racial/ethnic groups and in admixed populations (*Borrell, 2005*). To account for these heterogeneities and to avoid social and political controversies, the genetics community has grouped individuals by genetic ancestry instead of race and ethnicity (*Yudell et al., 2016*). Indeed, recent work from our group and others have demonstrated that genetic ancestry improves diagnostic precision compared to racial/ethnic categorizations for specific medical conditions and clinical decisions (*Kumar et al., 2010; Udler et al., 2015; Nalls et al., 2008*).

However, racial and ethnic categories also 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 (*Nguyen et al., 2014; Evans and Kantrowitz, 2002*). Thus, it is unclear whether defining groups through genetic ancestry can capture these shared exposures. In this work we seek to explore the contributions of genetically defined ancestry and social, cultural and environmental factors to understanding differential methylation between ethnic groups.

Epigenetic modification of the genome through methylation plays a key role in the regulation of diverse cellular processes (*Smith and Meissner, 2013*). Changes in DNA methylation patterns have been associated with complex diseases, including various cancers (*Kulis and Esteller, 2010*), cardiovascular disease (*Udali et al., 2013; Kato et al., 2015*), obesity (*Bell et al., 2010*), diabetes (*Chambers et al., 2015*), autoimmune and inflammatory diseases (*Liu et al., 2013*), and

neurodegenerative diseases (Lardenoije et al., 2015). Epigenetic changes are thought to reflect influences of both genetic (Bell et al., 2011) and environmental factors (Feil and Fraga, 2011), and have been shown to vary between racial groups (Barfield et al., 2014). 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 (Bell et al., 2011), 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 (Joubert et al., 2012, 2016), polycyclic aromatic hydrocarbons, infectious pathogens, particulate matter, diesel exhaust particles (Jiang et al., 2014), allergens, heavy metals, and other indoor and outdoor pollutants (Ho et al., 2012). Psychosocial factors, including measures of traumatic experiences (Chen et al., 2013; Ressler et al., 2011; van der Knaap et al., 2014), socioeconomic status (Lam et al., 2012; Borghol et al., 2012), and general perceived stress (Vidal et al., 2014), also affect methylation levels. Since both genetic and environmental exposures affect methylation, this represents an ideal phenotype to explore the contributions of these two factors on differential methylation between ethnic groups.

In this work, we leveraged genome-wide methylation data in 573 Latino children of diverse Latino sub-ethnicities enrolled in the Genes-Environment and Admixture in Latino Americans (GALA II) study (Oh et al., 2012) whose genetic ancestry had been determined from dense genotyping arrays. This allowed us to explore the extent to which the differences in methylation between Latino sub-groups could be explained by their shared genetic ancestry. We found that many of the methylation differences associated with ethnicity could be explained by shared genetic ancestry. However, even after adjusting for ancestry, significant differences in methylation remained between the groups at multiple loci, reflecting social and environmental influences upon methylation.

Our findings have important implications for both the use of ancestry to capture 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 (Michels et al., 2013). The findings also have implications for the common practice of considering individuals of Latino descent, regardless of origin as a single ethnic group.

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). **Figure 1** shows the distribution of African, European, and Native American ancestry among the 524 participants with genomic ancestry estimates.

Methylation data used in this study has been previously made publicly available at the Gene Expression Omnibus at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77716> (Rahmani et al., 2016). Genotyping data has been deposited in dbGaP; link will be activated when the data becomes publicly available (Burchard, <http://www.ncbi.nlm.nih.gov/gap/?term=phs001180>).

Global patterns of methylation

Differences in ethnicity and ancestry resulted in discernible patterns in the global methylation profile as demonstrated in a multidimensional scaling analysis (**Figure 2A**). As expected (Houseman et al., 2012; Lam et al., 2012), the first few principal coordinates are strongly correlated to imputed cell composition (**Figure 2B–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) (**Figure 3A**), and PC8 (p -ANOVA = 0.0003), after adjusting for age, sex, disease status, cell components, and technical laboratory factors (plate and position). Genetic ancestry was associated with PC3 ($p=0.002$), PC7 ($p=0.0004$) (**Figure 3B**) and PC8 ($p=0.001$) in a two degree of freedom ANOVA test, adjusting for age, sex, disease status, cell components, technical factors, and ethnicity. **Supplementary file 1A** summarizes the results of the simple correlation analysis of

Table 1. Baseline characteristics of GALA II participants with methylation data, stratified by ethnicity. Continuous variables are reported with inter-quartile range in brackets.

	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)				
Granulo cytes	51.2% [46.0%: 55.7%]	51.6% [46.8%: 57%]	51% [43.6%: 57.2%]	49.1% [43.8%: 55.8%]
Lympho cytes	41.9% [36.9%: 46.6%]	41.8% [36.9%: 46.5%]	41.9% [36.1%: 51.6%]	43.9% [36.8%: 49.6%]
Mono cytes	7.1% [5.8%: 8.3%]	6.74% [5.74%: 8.24%]	6.6% [5.7%: 7.6%]	7.4% [6.2%: 8.6%]

DOI: [10.7554/eLife.20532.004](https://doi.org/10.7554/eLife.20532.004)

methylation with ethnicity and ancestry, as well as the adjusted nested ANOVA models described above and the mediation results described below.

A mediation analysis (Tingley *et al.*, 2014) revealed that the associations between ethnicity and PCs 3, 7, and eight were significantly mediated by Native American ancestry, which explained ~100% (95% CI: 37–100%, $p=0.01$) of PC3, 83% (95% CI 37–100%, $p<0.001$) of PC7 and 66% (95% CI: 25% to 100%, $p<0.001$) of PC8. Inclusion of Native American ancestry in the regression model of PCs 3, 7, and eight 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 the ethnic differences in these principal components are associated with global methylation patterns not captured by the shared genetic ancestry of each ethnic group. 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 shared genetic ancestry can explain over 50% of the association between ethnicity and global methylation patterns in three PC's, other non-genetic factors, such as environmental and social exposure differences associated with ethnicity influence methylation and are not captured by measures of genetic ancestry in two others.

Epigenome-wide association of self-identified ethnicity

An epigenome-wide association study of self-identified ethnicity (see Materials and 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

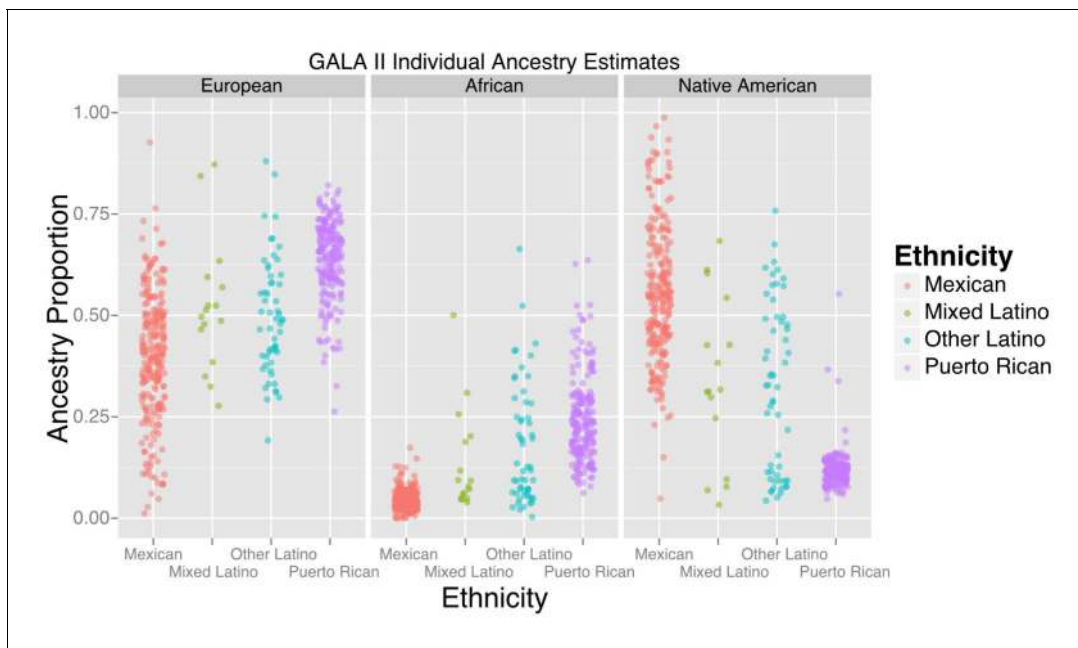


Figure 1. Ancestry estimates for GALA II participants, by ethnic group. Mexicans, on average, had a greater proportion of Native American ancestry than Puerto Ricans; Puerto Ricans had a greater proportion of European and African ancestry. Mixed and other Latinos were intermediate.

DOI: [10.7554/eLife.20532.003](https://doi.org/10.7554/eLife.20532.003)

less than 1.6×10^{-7} (Figure 4A and Supplementary file 1B). The most significant association with ethnicity occurred at cg12321355 in the ABO blood group gene (*ABO*) on chromosome 3 (p -ANOVA 6.7×10^{-22}) (Figure 4B). A two degree of freedom ANOVA test for genomic ancestry was also significantly associated with methylation level at this site ($p=2.3 \times 10^{-5}$) (Figure 4C), 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.

To determine the contribution of shared genetic ancestry and other factors associated with ethnicity, we repeated the analysis adjusting for ancestry. A significant association remained in 314 of the 834 (37.8%, $p=1.7 \times 10^{-183}$ for enrichment) CpG sites associated with ethnicity (Figure 5A and Supplementary file 1B) (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 Figure 1]).

Table 2 and Figure 5b show the proportion of variance explained by ethnicity, genomic ancestry, and their joint effect in the 916 CpG's associated with ethnicity, as well as the 314 CpG's that remained associated with ethnicity after adjustment for ancestry and the 520 CpG's whose association with ethnicity was no longer significant when ancestry terms were introduced into the model. Even after adjusting for genomic ancestry, ethnicity explained 1.7% (IQR 0.785% to 3.0%) but as much as 13.4% of the variance in methylation across these loci. Genomic ancestry explained a median of 4.2% (IQR 1.8% to 8.3%) of the variance in methylation at all loci associated with ethnicity 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 (median of 6.8%, IQR 4.5% to 10.0%) (Figure 5B).

Ethnicity and ancestry jointly explained as much as 38.5% of the variance in methylation in one CpG (cg0966827) and there were 17 CpG's where ethnicity and ancestry jointly explain more 25% of the variance. Among the 314 CpG's that remained associated with ethnicity after adjustment for ancestry, ethnicity accounted for a larger share of the joint variance than genomic ancestry (3.5%, IQR 2.2% to 5.1% versus 1.8%, IQR 0.8% to 4.0%). We saw a moderate amount of correlation

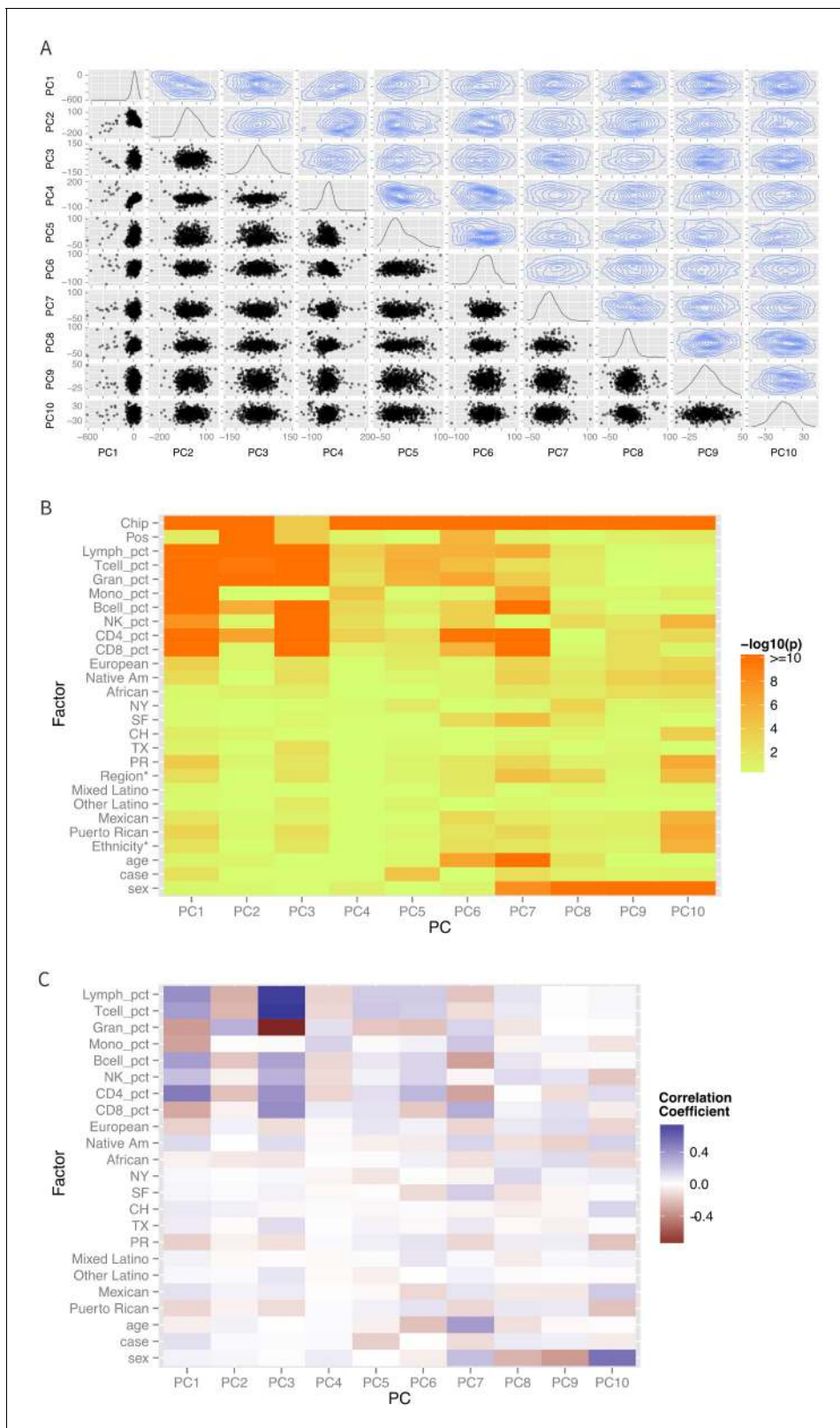


Figure 2. Patterns of global methylation. (A) Distribution of the first 10 principal coordinates of the methylation data. Plots in the diagonal show the univariate distribution; those in the lower left triangle show bivariate relationship between each pair of PCs, while those in the upper right show the bivariate density. (B) Bivariate or ANOVA associations between principal coordinates and technical factors (chip, position), cell counts, genetic ancestry
Figure 2 continued on next page

Figure 2 continued

(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.

DOI: [10.7554/eLife.20532.005](https://doi.org/10.7554/eLife.20532.005)

between the 314 methylation sites associated with ethnicity after adjusting for ancestry (median R^2 of 0.044, IQR 0.01 to 0.13).

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

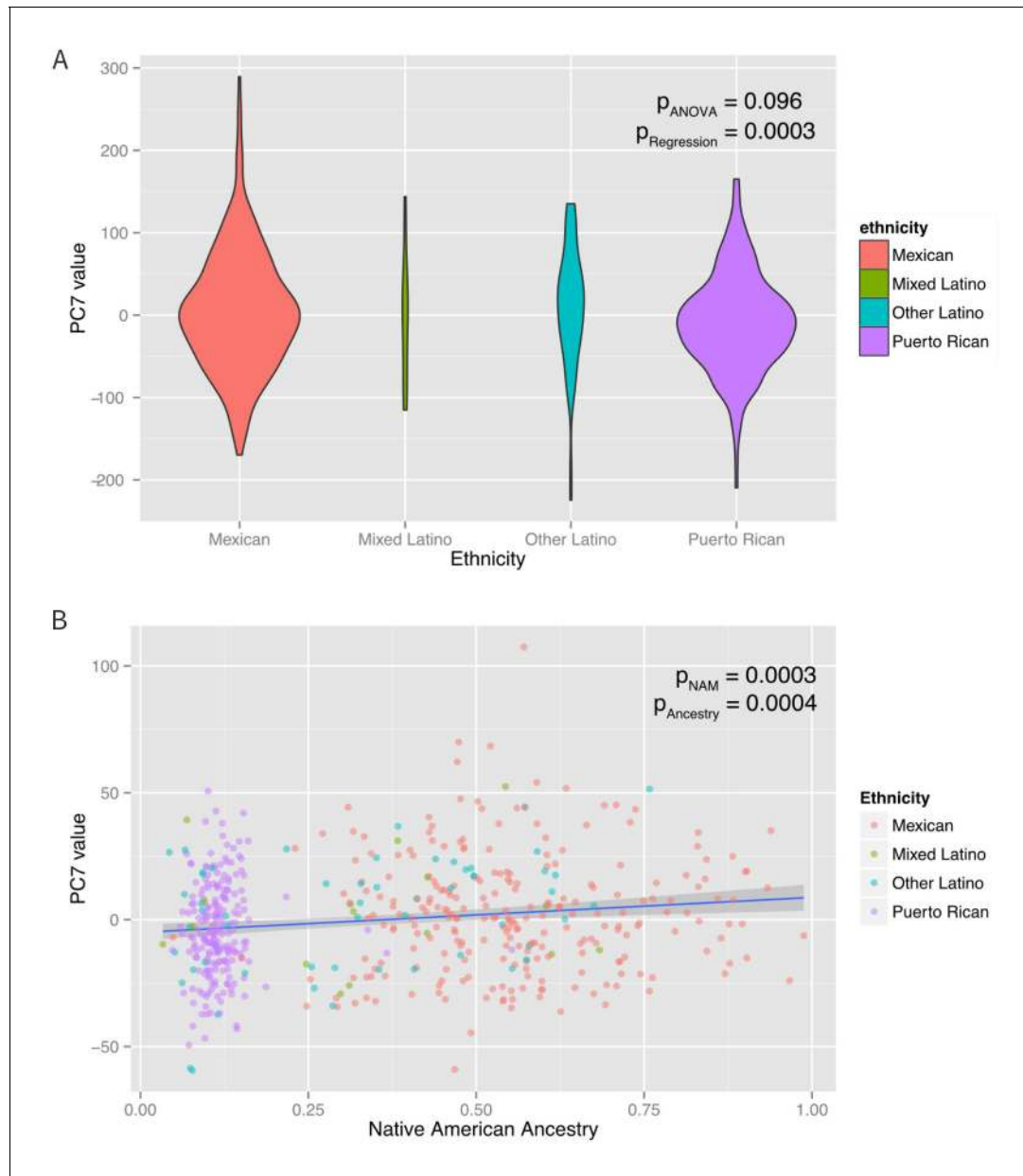


Figure 3. Associations between ethnicity, ancestry and global methylation. (A) Association between ethnicity and principal coordinate 7. (B) Association between Native American ancestry proportion and PC7, colored by ethnicity. Native American ancestry explains approximately 81% of the association between PC7 and ethnicity.

DOI: [10.7554/eLife.20532.006](https://doi.org/10.7554/eLife.20532.006)

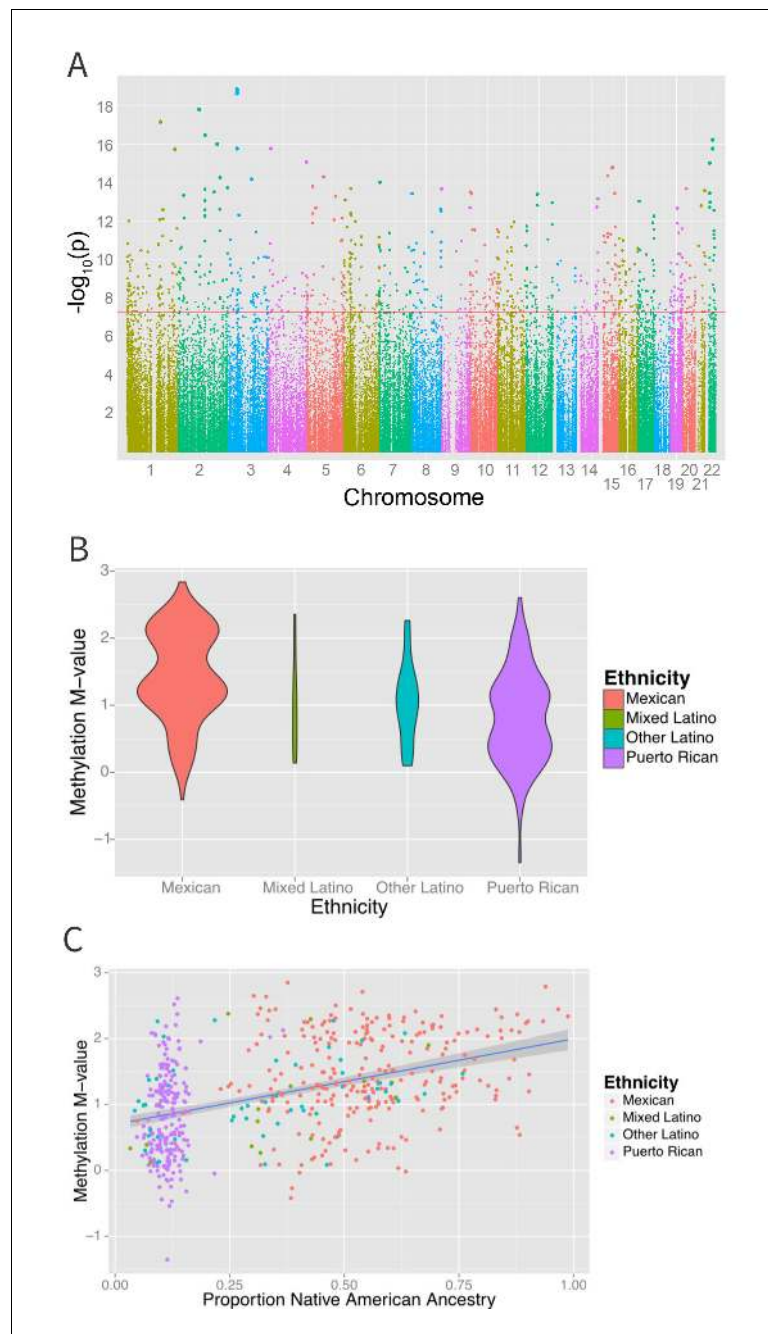


Figure 4. Associations between ethnicity and methylation (A) Manhattan plot showing the associations between ethnicity and methylation at individual CpG loci. (B) Violin plot showing one such locus, cg19145607. Mexicans are relatively hypermethylated compared to Puerto Ricans ($p=1.4 \times 10^{-19}$). (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.
DOI: [10.7554/eLife.20532.007](https://doi.org/10.7554/eLife.20532.007)

our results (See *Supplementary file 1B–F*). 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.

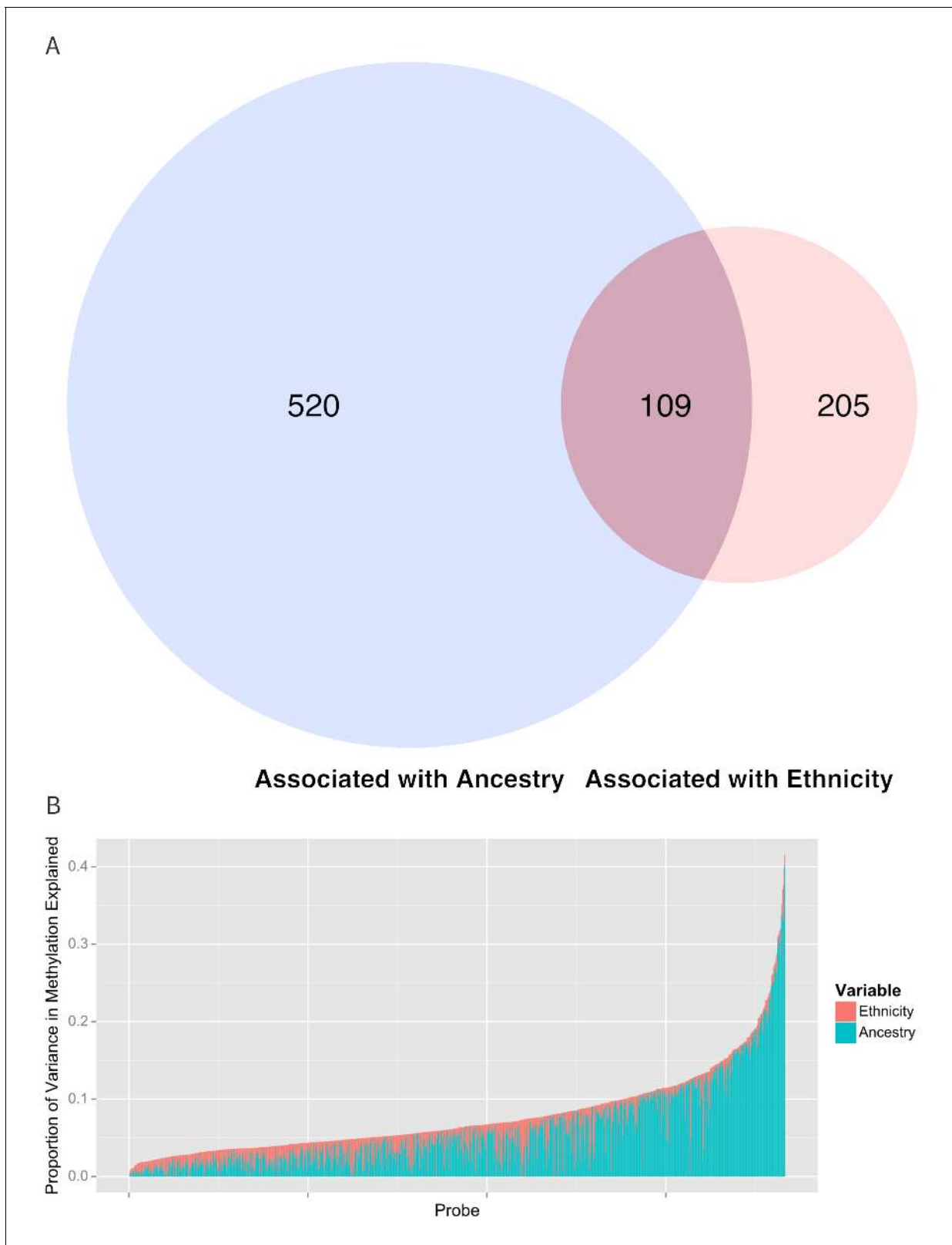


Figure 5. Relationship between genomic ancestry and the association between ethnicity and methylation. (A) Venn diagram showing the effect of adjustment for ancestry on the association between ethnicity and methylation. The components of the diagram represent the number of CpG's that remained associated 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
Figure 5 continued on next page

Figure 5 continued

between ethnicity and methylation M-values for (C) Native American ancestry and (D) African ancestry. For simplicity, only significant mediation effects are shown.

DOI: [10.7554/eLife.20532.008](https://doi.org/10.7554/eLife.20532.008)

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 (**Supplementary file 1C,D**).

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.

While most population substructure in Latinos would be expected to arise from differences in continental ancestry (*Galanter, 2012; Bryc et al., 2010*), there is evidence of finer scale (sub-continental) ancestry in Latino populations (*Moreno-Estrada et al., 2014*). We tested for the effect of fine scale substructure by calculating principal components for all participants with genotyping data using Eigensoft (*Patterson et al., 2006*). 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. (**Supplementary file 1E**)

At these 314 loci, the median total variance accounted for by ethnicity, ancestry, and fine-scale substructure was 10.4% (IQR 6.6% to 16.1%), of which ethnicity explained a median of 1.7% (IQR 0.8% to 3.8%), ancestry explained a median of 2.9% (IQR 1.0 to 4.6%) and fine scale substructure explained a median of 3.4% (IQR 2.0% to 4.2%). Among the 263 CpG's whose association with

Table 2. Proportion of variance in methylation explained by ethnicity and ancestry. Numbers represent the median and interquartile range.

Component	All CpG's associated with ethnicity (n = 916)	CpG's associated with ethnicity after adjusting for ancestry (n = 314)	CpG's whose association with ethnicity is explained by ancestry (n = 520)
Joint	6.8% (4.5% to 10%)	6.2% (4.4% to 8.8%)	7.8% (5.3% to 11.1%)
Ethnicity	1.7% (0.78% to 3.0%)	3.5% (2.2% to 5.1%)	<1%
Ancestry	4.2% (1.8% to 8.3%)	1.8% (0.8% to 4.0%)	6.6% (4.0% to 10.2%)

DOI: [10.7554/eLife.20532.009](https://doi.org/10.7554/eLife.20532.009)

ethnicity could not be explained by fine-scale substructure, ethnicity explained a median of 1.9% (IQR 1.0% to 4.0%; max 26.7%), ancestry explained 2.8% (IQR 1.0% to 6.2%), and fine scale ancestry explained 3.2% (IQR 1.9% to 4.7%).

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 (**Supplementary file 1F**). Of the 916 CpG's associated with ethnicity at a genome-wide scale ($p < 1.6 \times 10^{-7}$) 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 ($p < 5 \times 10^{-5}$). 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.

We conclude that shared genetic ancestry explains much but not all of the association between ethnicity and methylation. Other, non-genetic factors associated with ethnicity likely explain the ethnicity-associated methylation changes that cannot be accounted for by genomic ancestry alone.

Ethnic differences in environmentally-associated methylation sites

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 (**Oh et al., 2012**), identified associations with methylation at over 6000 CpG loci (**Joubert et al., 2016**). We found 1341 of 4404 that passed QC in our own study (30.4%) were nominally associated with ethnicity ($p < 0.05$), which represented a highly significant ($p < 2 \times 10^{-16}$) enrichment. Using a Bonferroni correction for the 4404 loci tested, 126 maternal-smoking related loci were associated with ethnicity ($p < 1.1 \times 10^{-5}$), and 27 loci were among the 916 CpG's reported above as associated with ethnicity (**Supplementary file 1G**). Of these, 14 were among the 314 CpG's whose association with ethnicity could not be explained by ancestry and 12 were among the 263 CpG's whose association with ethnicity could not be explained by ancestry or fine-scale substructure. We also examined methylation loci from an earlier study of maternal smoking in Norwegian newborns (**Joubert et al., 2012**) as well as studies of diesel exhaust particles (**Jiang et al., 2014**) and exposure to violence (**Chen et al., 2013**). 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 **Supplementary file 1H**.

In an earlier study of maternal smoking in Norwegian newborns (**Joubert et al., 2012**) 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 ($p < 0.0026$), cg23067299 in the aryl hydrocarbon receptor repressor (*AHRR*) gene on chromosome five remained statistically significant (**Supplementary file 1H**). 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 (**Jiang et al., 2014**) 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 five 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 (**Chen et al., 2013**) and with heavy trauma exposure in adults (**Ressler et al., 2011**), was significantly associated with ethnicity ($p = 0.02$).

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

1.6×10^{-7} (Figure 6 and Supplementary file 1I), 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 3.1×10^{-24}) (Figure 6B). 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, $p=1.1 \times 10^{-21}$). There was no significant heterogeneity in the association between genetic ancestry and methylation between Puerto Ricans and Mexicans ($p\text{-het} = 0.5$). Mexicans have a mean unadjusted methylation M-value 0.48 units lower than Puerto Ricans (95% CI 0.35 to 0.62 units, $p=1.1 \times 10^{-11}$). However, adjusting for African ancestry accounts for the differences in methylation level between the two sub-groups ($p\text{-adjusted} = 0.4$), demonstrating that ethnic differences in methylation at this site are due to differences in African ancestry.

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. We therefore looked at the association between methylation at cg04922029 and ancestry at that locus. We found almost perfect correlation between methylation and African ancestry at the locus ($p=6 \times 10^{-162}$) (Figure 7A). 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. We then looked for SNPs within 10,000 base pairs of the CpG site that explained the admixture mapping association. We found that methylation at cg04922029 was significantly correlated with SNP rs2814778 (Figure 7B), the Duffy

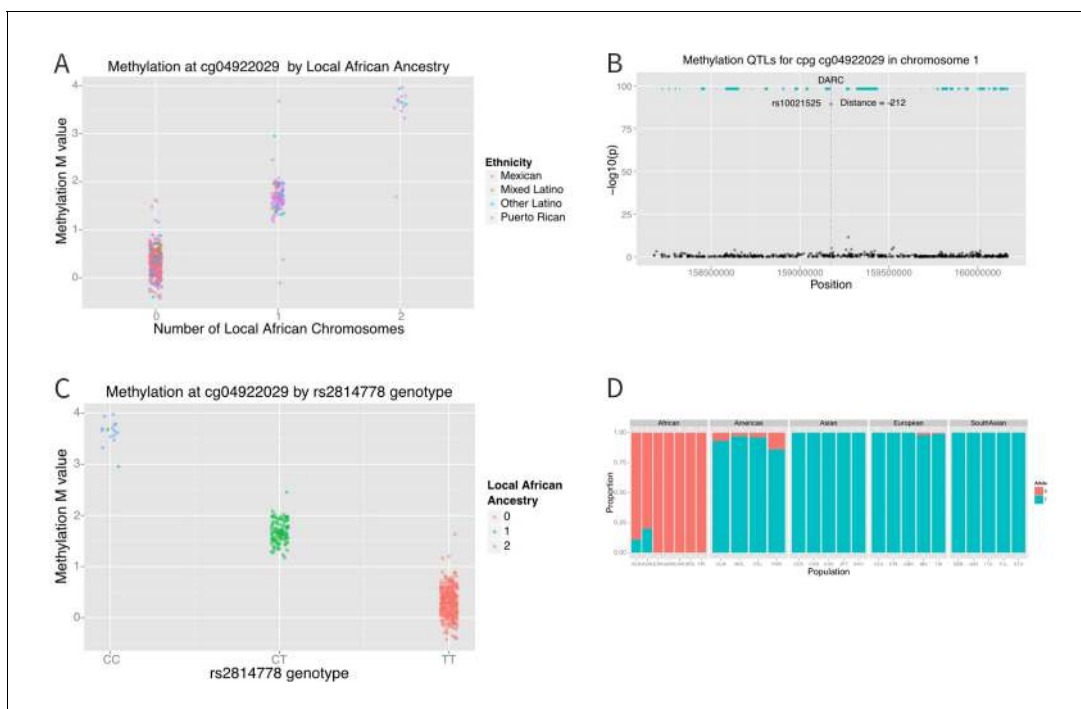


Figure 7. Association between local ancestry and methylation. (A) Association between cg04922029 on the DARC locus and African ancestry, color coded by ethnic group. There is near perfect correlation between the two. (B) Association between SNPs located within 1 Mb of cg04922029 and methylation levels at that CpG. (C) 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. (D) 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.

DOI: 10.7554/eLife.20532.011

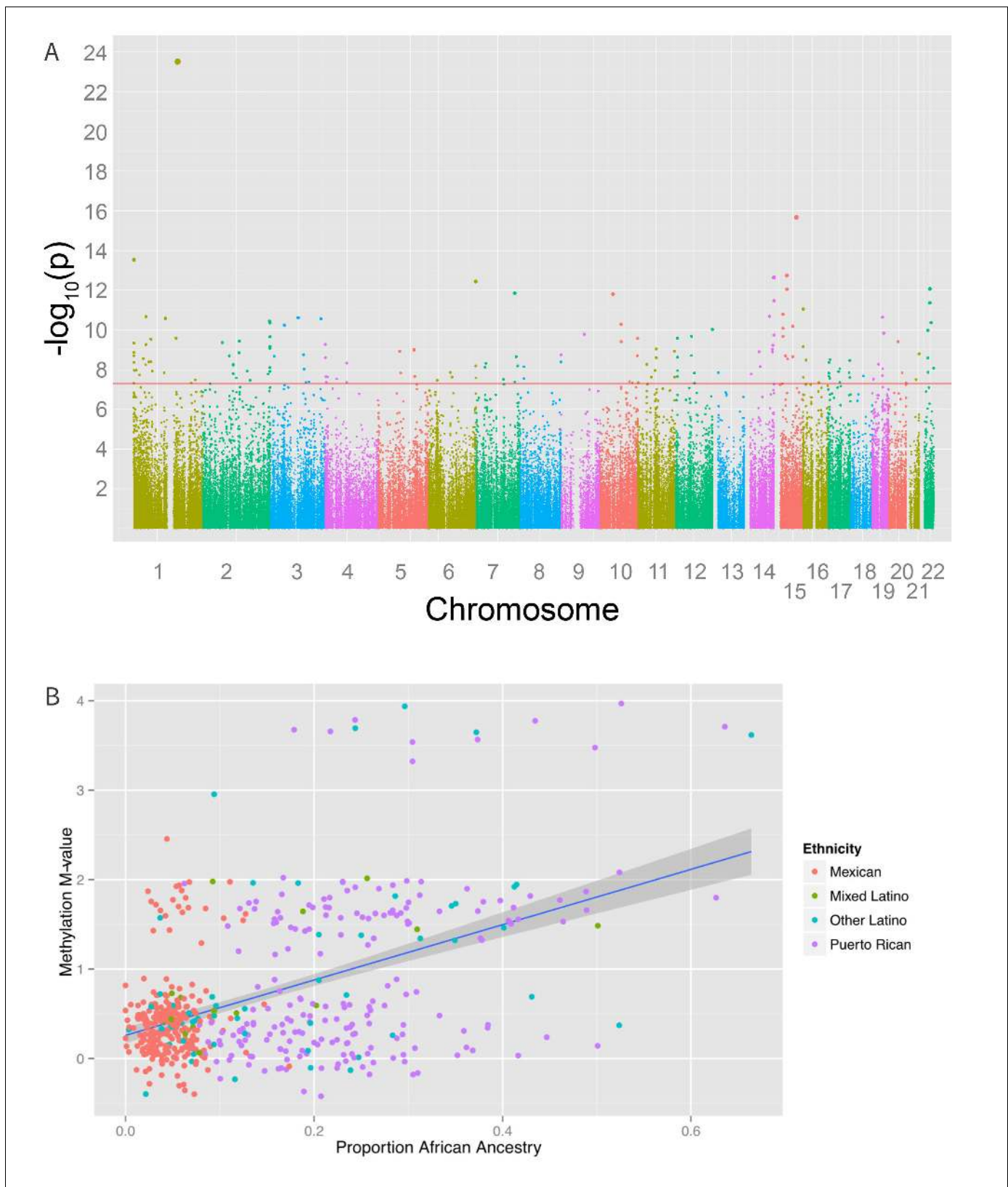


Figure 6. Associations between genomic ancestry and individual methylation loci. (A) Manhattan plot showing the associations between genomic ancestry and methylation at individual CpG loci. (B) Plot showing one such locus, cg04922029, and genomic African ancestry, showing a strong correlation between African ancestry and hypermethylation at that site.

DOI: [10.7554/eLife.20532.010](https://doi.org/10.7554/eLife.20532.010)

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 ($p=3.8 \times 10^{-90}$) (**Figure 7C**).

When we examined the effect of local ancestry at the other 194 CpG's we find that 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 (**Figure 8**).

Discussion

In a diverse population of Latinos, we have shown that a substantial number of loci are differentially methylated between ethnic sub-groups. While genomic ancestry can explain the association between ethnicity and methylation at 66% of the 916 loci associated with ethnicity, factors other than shared ancestry that correlate with ethnicity, such as social, economic, cultural and environmental exposures account for the association between ethnicity and methylation at 34% (314/916) of loci.

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 (**Joubert et al., 2012, 2016**), as well as diesel exhaust particles (**Jiang et al., 2014**) and psychosocial stress (**Chen et al., 2013**) 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 (**Joubert et al., 2016**) were differentially methylated between Latino subgroups at a genome-wide significance threshold of 1.6×10^{-7} . Interestingly, this included both loci whose association persisted after adjustment for ancestry and fine-scale population substructure and are thus presumed to be due to environmental differences between ethnic groups and loci in which the association between ethnicity and methylation could be fully explained by genetically defined ancestry.

There are a number of plausible reasons for overlap between CpG's associated with ancestry and those associated with environmental exposure. It is possible that this represents a gene-environment interaction, and that individuals with certain genetic backgrounds are more susceptible to the effects

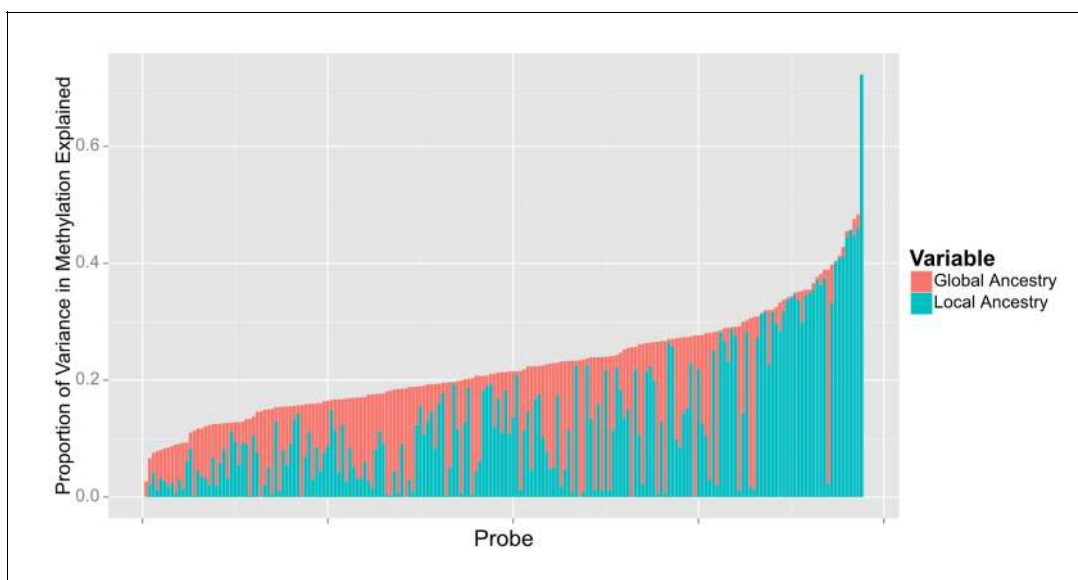


Figure 8. Relative proportion of variance in methylation explained by global and local ancestry across loci significantly associated with global ancestry. DOI: 10.7554/eLife.20532.012

of environmental exposures such as *in utero* tobacco smoke than those of other genetic backgrounds. It has been previously reported that Hispanic smokers with high Native American ancestry had reduced risk of methylation across 12 genes, suggesting an ancestry by smoking interaction (**Leng et al., 2013**). Because the majority of studies that comprised the consortium that identified differentially methylated regions enrolled participants of European descent, such interactions might not have been evident in their study.

It is also possible that environmental exposures correlate with ancestry and that participants with certain ancestral backgrounds may have been more exposed to *in utero* tobacco smoke than those of other backgrounds. Several studies have shown correlations between genetic ancestry and environmental exposures, including socioeconomic status (**Florez et al., 2011**), overweight and obesity (**Ziv et al., 2006**), and birth site and country of residence (**González Burchard et al., 2005**).

Though our analysis of global ancestry showed that a majority of the variance explained jointly by local and global ancestry can be traced to specific loci in the genome acting in *-cis*, a substantial proportion cannot. Some of the residual association between global ancestry and methylation may be due to genetic effects acting in *-trans*; however, the possibility that some of it may be due to environmental exposures correlating with global ancestry cannot be excluded. Thus, it is plausible that genomic ancestry is acting as a proxy for both genetic and environmental effects in our study. If this is the case, our study likely underestimates the degree to which environmental factors explain differential methylation between ethnic groups.

Finally, it is possible that our analysis identified DMRs that are independently modifiable by both genetic and environmental exposures. Thus, regions of the genome that are differentially methylated due to genetic polymorphisms may also be more susceptible to differential methylation due to environmental exposures.

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 may be a useful, albeit imperfect proxy. However, if a comprehensive catalog of the effects of exposures can be compiled, it may be possible to use genome-wide methylation analysis as a biomarker of exposure long after the exposure has passed and can no longer be measured.

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 (**Grafodatskaya et al., 2010**). The strongest patterns of methylation are associated with cell composition in whole blood (**Lam et al., 2012**). 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 (**Moreno-Estrada et al., 2014**). 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 (**Galanter et al., 2012**) could also give rise to methylation differences, but these are inconsistent with existing rare variant and selection analyses (**Hernandez et al., 2011**; **Tang et al., 2007**). 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. Since our study was geographically diverse, recruiting participants at five recruitment sites in the United States and Puerto Rico, it is possible that systematic differences associated with site of recruitment might have influenced observed methylation differences between ethnic

groups. However, when we included recruitment site as a covariate, we found no significant effect on methylation independent of ethnicity.

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 ([Barfield et al., 2014](#)). 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. However, the methylation loci identified in this study, as well as studies of environmental exposures, could be particularly interesting loci for the study of biomedical outcomes, particularly those with disparate prevalence between racial/ethnic groups, such as asthma ([Barr et al., 2016](#)). If methylation loci associated with ethnicity or ancestry were shown to be associated with a biomedical outcome, it could help explain racial/ethnic disparities in disease.

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 beyond the effects of shared 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, and likely has implications for all studies of 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.

Materials and methods

Participant recruitment

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 Pediátrica), 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 ([Oh et al., 2012](#)).

A total of 4702 children (2374 participants with asthma and 2328 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 two 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 (*Oh et al., 2012*). Latino sub-ethnicity was determined by self-identification and the ethnicity of 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

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.

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 Infinium 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 (*Du et al., 2010*).

Genotyping

Details of genotyping and quality control procedures for single nucleotide polymorphisms (SNPs) and individuals have been described elsewhere (*Galanter et al., 2014*). Briefly, participants were genotyped at 818,154 SNPs on the Axiom Genome-Wide LAT 1, World Array 4 (Affymetrix, Santa Clara, CA) (*Hoffmann et al., 2011*). 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 3804 (1902 asthmatic cases, 1902 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 analysis

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 one 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 (*Alexander et al., 2009*), with a three population model. Local ancestry at all

positions across the genome was estimated using the program LAMP-LD (Baran *et al.*, 2012), assuming three ancestral populations.

Principal components for the genetic data were determined using the program EIGENSTRAT (Patterson *et al.*, 2006).

Statistical analysis

Using a variance in methylation m -value of 0.2 units, which corresponded to approximately the 90th percentile of the variance in m -value in our pilot data, we determined that in order to have an 80% power to detect a difference in mean methylation between the two major ethnic groups of 0.25 units, using a Bonferroni significance threshold of 1.6×10^{-7} a sample, a sample size of 251 participants in each group was required. That total sample size of 502 participants gave us 80% power to detect correlations between ancestry and methylation of medium (Pearson $r > 0.25$) effect, meaning that we had 80% power to detect loci where ancestry accounted for at least 6.25% of the variance in methylation.

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.* (Houseman *et al.*, 2012). 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 1.6×10^{-7} . Analyses were performed using R version 3.2.1 (The R Foundation for Statistical Computing)(R Core Team) and the Bioconductor package version 2.13.

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 (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 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 (Supplementary file 1A).

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' (Tingley *et al.*, 2014) 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 coefficient of determination $R^2 > 0.998$) 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 (β).

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 ([Joubert et al., 2012](#)), diesel exhaust particles (DEP) ([Jiang et al., 2014](#)), and exposure to violence ([Chen et al., 2013](#)). We have previously shown that exposure to these common environmental exposures or similar exposures varied by ethnicity within our own GALA II study populations ([Oh et al., 2012](#); [Nishimura et al., 2013](#); [Thakur et al., 2013](#)).

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 correlated 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 ([Galanter et al., 2014](#); [Rosenberg et al., 2010](#)). 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.

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.

Additional information

Funding

Funder	Grant reference number	Author
National Institutes of Health	multiple	Joshua M Galanter Christopher R Gignoux Neeta Thakur Harold J Farber Rajesh Kumar Max A Seibold Esteban G Burchard Noah Zaitlen
American Asthma Foundation		Esteban G Burchard
Sandler Family Foundation		Esteban G Burchard

Tobacco-Related Disease Research Program	24RT-0025	Esteban G Burchard
Flight Attendant Medical Research Institute		Esteban G Burchard
Hewett Fellowship		Joshua M Galanter
Parker B. Francis Fellowship Program		Neeta Thakur
American Thoracic Society		Neeta Thakur
University of California, San Francisco	Chancellor's Research Fellowship	Christopher R Gignoux
University of California, San Francisco	Dissertation of the Year Fellowship	Christopher R Gignoux
Ernest S. Bazley Grant		Pedro C Avila

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

JMG, CRG, EGB, NZ, Conception and design, Analysis and interpretation of data, Drafting or revising the article; SSO, MP-Y, Analysis and interpretation of data, Drafting or revising the article; DT, Analysis and interpretation of data; NT, MAS, Drafting or revising the article; CE, Acquisition of data; DH, SH, Acquisition of data, Analysis and interpretation of data; HJF, RK, LNB, Recruited participants, Acquisition of data, Drafting or revising the article; PCA, EB-B, MAL, KM, DS, WR-C, JRR-S, Recruited participants, Acquisition of data

Author ORCIDs

Joshua M Galanter, <http://orcid.org/0000-0002-2561-6384>

Christopher R Gignoux, <http://orcid.org/0000-0001-9728-6567>

Sam S Oh, <http://orcid.org/0000-0002-2815-6037>

Maria Pino-Yanes, <http://orcid.org/0000-0003-0332-437X>

Neeta Thakur, <http://orcid.org/0000-0001-6126-6601>

Esteban G Burchard, <http://orcid.org/0000-0001-7475-2035>

Ethics

Human subjects: 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.

Additional files

Supplementary files

- Supplementary file 1. (A) Methylation principal components and ethnicity and ancestry. (B) Significant associations between ethnicity and methylation. (C) Effect of adding cubic spline ancestry terms to the association between ethnicity and methylation. (D) Effect of adding quadratic and cubic ancestry terms to the association between ethnicity and methylation. (E) Effect of adding genetic principal components 3–10 to the association between ethnicity and methylation. (F) 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. (G) Association of ethnicity and methylation in loci previously associated with maternal smoking during pregnancy. (H) Significant associations between ethnicity and methylation loci previously

associated with environmental exposures. (I) Significant associations between global ancestry and methylation.

DOI: [10.7554/eLife.20532.013](https://doi.org/10.7554/eLife.20532.013)

Major datasets

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset URL	Database, license, and accessibility information
Burchard EG	2016	Differential DNA methylation in Latino population	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77716	Publicly available at the NCBI Gene Expression Omnibus (accession no. GSE77716)
Burchard EG	2016	Genes-Environment and Ancestry in Latino Americans	https://www.ncbi.nlm.nih.gov/gap/?term=phs001180	Publicly available at NCBI dbGaP (accession no. phs001180.v1.p1)

References

- Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**:1655–1664. doi: [10.1101/gr.094052.109](https://doi.org/10.1101/gr.094052.109), PMID: [19648217](https://pubmed.ncbi.nlm.nih.gov/19648217/)
- Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, Rodriguez-Cintron W, Chapela R, Ford JG, Avila PC, Rodriguez-Santana J, Burchard EG, Halperin E. 2012. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **28**:1359–1367. doi: [10.1093/bioinformatics/bts144](https://doi.org/10.1093/bioinformatics/bts144), PMID: [22495753](https://pubmed.ncbi.nlm.nih.gov/22495753/)
- Barfield RT, Almli LM, Kilaru V, Smith AK, Mercer KB, Duncan R, Klengel T, Mehta D, Binder EB, Epstein MP, Ressler KJ, Conneely KN. 2014. Accounting for population stratification in DNA methylation studies. *Genetic Epidemiology* **38**:231–241. doi: [10.1002/gepi.21789](https://doi.org/10.1002/gepi.21789), PMID: [24478250](https://pubmed.ncbi.nlm.nih.gov/24478250/)
- Barr RG, Avilés-Santa L, Davis SM, Aldrich TK, Gonzalez F, Henderson AG, Kaplan RC, LaVange L, Liu K, Loredó JS, Mendes ES, Ni A, Ries A, Salathe M, Smith LJ. 2016. Pulmonary disease and age at immigration among hispanics. results from the hispanic community health study study of latinos. *American Journal of Respiratory and Critical Care Medicine* **193**:386–395. doi: [10.1164/rccm.201506-1211OC](https://doi.org/10.1164/rccm.201506-1211OC), PMID: [26451874](https://pubmed.ncbi.nlm.nih.gov/26451874/)
- Bell CG, Finer S, Lindgren CM, Wilson GA, Rakyán VK, Teschendorff AE, Akan P, Stupka E, Down TA, Prokopenko I, Morison IM, Mill J, Pidsley R, Deloukas P, Frayling TM, Hattersley AT, McCarthy MI, Beck S, Hitman GA, International Type 2 Diabetes 1q Consortium. 2010. Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. *PLoS One* **5**:e14040. doi: [10.1371/journal.pone.0014040](https://doi.org/10.1371/journal.pone.0014040), PMID: [21124985](https://pubmed.ncbi.nlm.nih.gov/21124985/)
- Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK. 2011. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology* **12**:R10. doi: [10.1186/gb-2011-12-1-r10](https://doi.org/10.1186/gb-2011-12-1-r10), PMID: [21251332](https://pubmed.ncbi.nlm.nih.gov/21251332/)
- Borghol N, Suderman M, McArdle W, Racine A, Hallett M, Pembrey M, Hertzman C, Power C, Szyf M. 2012. Associations with early-life socio-economic position in adult DNA methylation. *International Journal of Epidemiology* **41**:62–74. doi: [10.1093/ije/dyr147](https://doi.org/10.1093/ije/dyr147), PMID: [22422449](https://pubmed.ncbi.nlm.nih.gov/22422449/)
- Borrell LN. 2005. Racial identity among Hispanics: implications for health and well-being. *American Journal of Public Health* **95**:379–381. doi: [10.2105/AJPH.2004.058172](https://doi.org/10.2105/AJPH.2004.058172), PMID: [15727961](https://pubmed.ncbi.nlm.nih.gov/15727961/)
- Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, Hammer M, Bustamante CD, Ostrer H. 2010. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *PNAS* **107** Suppl 2:8954–8961. doi: [10.1073/pnas.0914618107](https://doi.org/10.1073/pnas.0914618107), PMID: [20445096](https://pubmed.ncbi.nlm.nih.gov/20445096/)
- Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, Mountain JL, Pérez-Stable EJ, Sheppard D, Risch N. 2003. The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine* **348**:1170–1175. doi: [10.1056/NEJMs025007](https://doi.org/10.1056/NEJMs025007), PMID: [12646676](https://pubmed.ncbi.nlm.nih.gov/12646676/)
- Chambers JC, Loh M, Lehne B, Drong A, Kriebel J, Motta V, Wahl S, Elliott HR, Rota F, Scott WR, Zhang W, Tan ST, Campanella G, Chadeau-Hyam M, Yengo L, Richmond RC, Adamowicz-Brice M, Afzal U, Bozaoglu K, Mok ZY, et al. 2015. 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. *The Lancet Diabetes & Endocrinology* **3**:526–534. doi: [10.1016/S2213-8587\(15\)00127-8](https://doi.org/10.1016/S2213-8587(15)00127-8), PMID: [26095709](https://pubmed.ncbi.nlm.nih.gov/26095709/)
- Chen W, Boutaoui N, Brehm JM, Han YY, Schmitz C, Cressley A, Acosta-Pérez E, Alvarez M, Colón-Semidey A, Baccarelli AA, Weeks DE, Kolls JK, Canino G, Celedón JC. 2013. ADCYAP1R1 and asthma in Puerto Rican children. *American Journal of Respiratory and Critical Care Medicine* **187**:584–588. doi: [10.1164/rccm.201210-1789OC](https://doi.org/10.1164/rccm.201210-1789OC), PMID: [23328528](https://pubmed.ncbi.nlm.nih.gov/23328528/)
- Cooper RS, Kaufman JS, Ward R. 2003. Race and genomics. *New England Journal of Medicine* **348**:1166–1170. doi: [10.1056/NEJMs022863](https://doi.org/10.1056/NEJMs022863), PMID: [12646675](https://pubmed.ncbi.nlm.nih.gov/12646675/)

- Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**:587. doi: [10.1186/1471-2105-11-587](https://doi.org/10.1186/1471-2105-11-587), PMID: [21118553](https://pubmed.ncbi.nlm.nih.gov/21118553/)
- Evans GW, Kantrowitz E. 2002. Socioeconomic status and health: the potential role of environmental risk exposure. *Annual Review of Public Health* **23**:303–331. doi: [10.1146/annurev.publhealth.23.112001.112349](https://doi.org/10.1146/annurev.publhealth.23.112001.112349), PMID: [11910065](https://pubmed.ncbi.nlm.nih.gov/11910065/)
- Feil R, Fraga MF. 2012. Epigenetics and the environment: emerging patterns and implications. *Nature Reviews Genetics* **13**:97–109. doi: [10.1038/nrg3142](https://doi.org/10.1038/nrg3142), PMID: [22215131](https://pubmed.ncbi.nlm.nih.gov/22215131/)
- Florez JC, Price AL, Campbell D, Riba L, Parra MV, Yu F, Duque C, Saxena R, Gallego N, Tello-Ruiz M, Franco L, Rodríguez-Torres M, Villegas A, Bedoya G, Aguilar-Salinas CA, Tusié-Luna MT, Ruiz-Linares A. 2011. Strong Association of Socioeconomic Status and Genetic Ancestry in Latinos: Implications for Admixture Studies of Type 2 Diabetes. *Racial Identities, Genetic Ancestry, and Health in South America*:137–153. doi: [10.1057/9781137001702_7](https://doi.org/10.1057/9781137001702_7)
- Galanter JM, Fernandez-Lopez JC, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, Via M, Hidalgo-Miranda A, Contreras AV, Figueroa LU, Raska P, Jimenez-Sanchez G, Zolezzi IS, Torres M, Ponte CR, Ruiz Y, Salas A, Nguyen E, Eng C, Borjas L, Zabala W, et al. 2012. Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genetics* **8**:e1002554. doi: [10.1371/journal.pgen.1002554](https://doi.org/10.1371/journal.pgen.1002554), PMID: [22412386](https://pubmed.ncbi.nlm.nih.gov/22412386/)
- Galanter JM, Gignoux CR, Torgerson DG, Roth LA, Eng C, Oh SS, Nguyen EA, Drake KA, Huntsman S, Hu D, Sen S, Davis A, Farber HJ, Avila PC, Brigino-Buenaventura E, LeNoir MA, Meade K, Serebrisky D, Borrell LN, Rodríguez-Cintrón W, et al. 2014. Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: the Genes-environments & Admixture in Latino Americans study. *Journal of Allergy and Clinical Immunology* **134**:295–305. doi: [10.1016/j.jaci.2013.08.055](https://doi.org/10.1016/j.jaci.2013.08.055), PMID: [24406073](https://pubmed.ncbi.nlm.nih.gov/24406073/)
- González Burchard E, Borrell LN, Choudhry S, Naqvi M, Tsai HJ, Rodríguez-Santana JR, Chapela R, Rogers SD, Mei R, Rodríguez-Cintrón W, Arena JF, Kittles R, Perez-Stable EJ, Ziv E, Risch N. 2005. Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. *American Journal of Public Health* **95**:2161–2168. doi: [10.2105/AJPH.2005.068668](https://doi.org/10.2105/AJPH.2005.068668), PMID: [16257940](https://pubmed.ncbi.nlm.nih.gov/16257940/)
- Grafodatskaya D, Choufani S, Ferreira JC, Butcher DT, Lou Y, Zhao C, Scherer SW, Weksberg R. 2010. EBV transformation and cell culturing destabilizes DNA methylation in human lymphoblastoid cell lines. *Genomics* **95**:73–83. doi: [10.1016/j.ygeno.2009.12.001](https://doi.org/10.1016/j.ygeno.2009.12.001), PMID: [20005943](https://pubmed.ncbi.nlm.nih.gov/20005943/)
- Hankinson JL, Odenchantz JR, Fedan KB. 1999. Spirometric reference values from a sample of the general U.S. population. *American Journal of Respiratory and Critical Care Medicine* **159**:179–187. doi: [10.1164/ajrccm.159.1.9712108](https://doi.org/10.1164/ajrccm.159.1.9712108), PMID: [9872837](https://pubmed.ncbi.nlm.nih.gov/9872837/)
- Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M, 1000 Genomes Project. 2011. Classic selective sweeps were rare in recent human evolution. *Science* **331**:920–924. doi: [10.1126/science.1198878](https://doi.org/10.1126/science.1198878), PMID: [21330547](https://pubmed.ncbi.nlm.nih.gov/21330547/)
- Ho SM, Johnson A, Tarapore P, Janakiram V, Zhang X, Leung YK. 2012. Environmental epigenetics and its implication on disease risk and health outcomes. *ILAR Journal* **53**:289–305. doi: [10.1093/ilar.53.3-4.289](https://doi.org/10.1093/ilar.53.3-4.289), PMID: [23744968](https://pubmed.ncbi.nlm.nih.gov/23744968/)
- Hoffmann TJ, Zhan Y, Kvale MN, Hesselson SE, Gollub J, Iribarren C, Lu Y, Mei G, Purdy MM, Quesenberry C, Rowell S, Shapero MH, Smethurst D, Somkin CP, Van den Eeden SK, Walter L, Webster T, Whitmer RA, Finn A, Schaefer C, et al. 2011. 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* **98**:422–430. doi: [10.1016/j.ygeno.2011.08.007](https://doi.org/10.1016/j.ygeno.2011.08.007), PMID: [21903159](https://pubmed.ncbi.nlm.nih.gov/21903159/)
- Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**:86. doi: [10.1186/1471-2105-13-86](https://doi.org/10.1186/1471-2105-13-86), PMID: [22568884](https://pubmed.ncbi.nlm.nih.gov/22568884/)
- Jiang R, Jones MJ, Sava F, Kobor MS, Carlsten C. 2014. Short-term diesel exhaust inhalation in a controlled human crossover study is associated with changes in DNA methylation of circulating mononuclear cells in asthmatics. *Particle and Fibre Toxicology* **11**:71. doi: [10.1186/s12989-014-0071-3](https://doi.org/10.1186/s12989-014-0071-3), PMID: [25487561](https://pubmed.ncbi.nlm.nih.gov/25487561/)
- Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, Reese SE, Markunas CA, Richmond RC, Xu CJ, Küpers LK, Oh SS, Hoyo C, Gruziova O, Söderhäll C, Salas LA, Baiz N, Zhang H, Lepeule J, Ruiz C, et al. 2016. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *The American Journal of Human Genetics* **98**:680–696. doi: [10.1016/j.ajhg.2016.02.019](https://doi.org/10.1016/j.ajhg.2016.02.019), PMID: [27040690](https://pubmed.ncbi.nlm.nih.gov/27040690/)
- Joubert BR, Håberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, Huang Z, Hoyo C, Midttun Ø, Cupul-Uicab LA, Ueland PM, Wu MC, Nystad W, Bell DA, Peddada SD, London SJ. 2012. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environmental Health Perspectives* **120**:1425–1431. doi: [10.1289/ehp.1205412](https://doi.org/10.1289/ehp.1205412), PMID: [22851337](https://pubmed.ncbi.nlm.nih.gov/22851337/)
- Kato N, Loh M, Takeuchi F, Verweij N, Wang X, Zhang W, Kelly TN, Saleheen D, Lehne B, Mateo Leach I, Drong AW, Abbott J, Wahl S, Tan ST, Scott WR, Campanella G, Chadeau-Hyam M, Afzal U, Ahluwalia TS, Bonder MJ, et al. 2015. Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nature Genetics* **47**:1282–1293. doi: [10.1038/ng.3405](https://doi.org/10.1038/ng.3405), PMID: [26390057](https://pubmed.ncbi.nlm.nih.gov/26390057/)
- Kulis M, Esteller M. 2010. DNA methylation and cancer. *Advances in Genetics* **70**:27–56. doi: [10.1016/B978-0-12-380866-0.60002-2](https://doi.org/10.1016/B978-0-12-380866-0.60002-2), PMID: [20920744](https://pubmed.ncbi.nlm.nih.gov/20920744/)
- Kumar R, Seibold MA, Aldrich MC, Williams LK, Reiner AP, Colangelo L, Galanter J, Gignoux C, Hu D, Sen S, Choudhry S, Peterson EL, Rodríguez-Santana J, Rodríguez-Cintrón W, Nalls MA, Leak TS, O'Meara E, Meibohm

- B, Kritchevsky SB, Li R, et al. 2010. Genetic ancestry in lung-function predictions. *New England Journal of Medicine* **363**:321–330. doi: [10.1056/NEJMoa0907897](https://doi.org/10.1056/NEJMoa0907897), PMID: [20647190](https://pubmed.ncbi.nlm.nih.gov/20647190/)
- Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, Miller GE, Kobor MS. 2012. Factors underlying variable DNA methylation in a human community cohort. *PNAS* **109** Suppl 2:17253–17260. doi: [10.1073/pnas.1121249109](https://doi.org/10.1073/pnas.1121249109), PMID: [23045638](https://pubmed.ncbi.nlm.nih.gov/23045638/)
- Lardenoije R, Iatrou A, Kenis G, Kompotis K, Steinbusch HW, Mastroeni D, Coleman P, Lemere CA, Hof PR, van den Hove DL, Rutten BP. 2015. The epigenetics of aging and neurodegeneration. *Progress in Neurobiology* **131**:21–64. doi: [10.1016/j.pneurobio.2015.05.002](https://doi.org/10.1016/j.pneurobio.2015.05.002), PMID: [26072273](https://pubmed.ncbi.nlm.nih.gov/26072273/)
- Leng S, Liu Y, Thomas CL, Gauderman WJ, Picchi MA, Bruse SE, Zhang X, Flores KG, Van Den Berg D, Stidley CA, Gilliland FD, Belinsky SA. 2013. Native American ancestry affects the risk for gene methylation in the lungs of Hispanic smokers from New Mexico. *American Journal of Respiratory and Critical Care Medicine* **188**:1110–1116. doi: [10.1164/rccm.201305-0925OC](https://doi.org/10.1164/rccm.201305-0925OC), PMID: [24032348](https://pubmed.ncbi.nlm.nih.gov/24032348/)
- Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, Shchetynsky K, Scheynius A, Kere J, Alfredsson L, Klareskog L, Ekström TJ, Feinberg AP. 2013. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology* **31**:142–147. doi: [10.1038/nbt.2487](https://doi.org/10.1038/nbt.2487), PMID: [23334450](https://pubmed.ncbi.nlm.nih.gov/23334450/)
- Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Grealley JM, Gut I, Houseman EA, Izzi B, Kelsey KT, Meissner A, Milosavljevic A, Siegmund KD, Bock C, Irizarry RA. 2013. Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods* **10**:949–955. doi: [10.1038/nmeth.2632](https://doi.org/10.1038/nmeth.2632), PMID: [24076989](https://pubmed.ncbi.nlm.nih.gov/24076989/)
- Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M, Contreras AV, Acuña-Alonso V, Sandoval K, Eng C, Romero-Hidalgo S, Ortiz-Tello P, Robles V, Kenny EE, Nuño-Arana I, Barquera-Lozano R, Macín-Pérez G, Granados-Arriola J, Huntsman S, Galanter JM, Via M, et al. 2014. Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* **344**:1280–1285. doi: [10.1126/science.1251688](https://doi.org/10.1126/science.1251688), PMID: [24926019](https://pubmed.ncbi.nlm.nih.gov/24926019/)
- Nalls MA, Wilson JG, Patterson NJ, Tandon A, Zmuda JM, Huntsman S, Garcia M, Hu D, Li R, Beamer BA, Patel KV, Akyzbekova EL, Files JC, Hardy CL, Buxbaum SG, Taylor HA, Reich D, Harris TB, Ziv E. 2008. 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. doi: [10.1016/j.ajhg.2007.09.003](https://doi.org/10.1016/j.ajhg.2007.09.003), PMID: [18179887](https://pubmed.ncbi.nlm.nih.gov/18179887/)
- Nguyen AB, Moser R, Chou WY. 2014. 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. doi: [10.1016/j.puhe.2014.10.003](https://doi.org/10.1016/j.puhe.2014.10.003), PMID: [25457801](https://pubmed.ncbi.nlm.nih.gov/25457801/)
- Nishimura KK, Galanter JM, Roth LA, Oh SS, Thakur N, Nguyen EA, Thyne S, Farber HJ, Serebrisky D, Kumar R, Brigino-Buenaventura E, Davis A, LeNoir MA, Meade K, Rodriguez-Cintron W, Avila PC, Borrell LN, Bibbins-Domingo K, Rodriguez-Santana JR, Sen S, et al. 2013. Early-life air pollution and asthma risk in minority children. The GALA II and SAGE II studies. *American Journal of Respiratory and Critical Care Medicine* **188**:309–318. doi: [10.1164/rccm.201302-0264OC](https://doi.org/10.1164/rccm.201302-0264OC), PMID: [23750510](https://pubmed.ncbi.nlm.nih.gov/23750510/)
- Oh SS, Tcheurekdjian H, Roth LA, Nguyen EA, Sen S, Galanter JM, Davis A, Farber HJ, Gilliland FD, Kumar R, Avila PC, Brigino-Buenaventura E, Chapela R, Ford JG, LeNoir MA, Lurmann F, Meade K, Serebrisky D, Thyne S, Rodriguez-Cintron W, et al. 2012. Effect of secondhand smoke on asthma control among black and Latino children. *Journal of Allergy and Clinical Immunology* **129**:1478–1483. doi: [10.1016/j.jaci.2012.03.017](https://doi.org/10.1016/j.jaci.2012.03.017), PMID: [22552109](https://pubmed.ncbi.nlm.nih.gov/22552109/)
- Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genetics* **2**:e190. doi: [10.1371/journal.pgen.0020190](https://doi.org/10.1371/journal.pgen.0020190), PMID: [17194218](https://pubmed.ncbi.nlm.nih.gov/17194218/)
- Phimister EG. 2003. Medicine and the racial divide. *New England Journal of Medicine* **348**:1081–1082. doi: [10.1056/NEJMp030028](https://doi.org/10.1056/NEJMp030028), PMID: [12646663](https://pubmed.ncbi.nlm.nih.gov/12646663/)
- Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, Enright PL, Hankinson JL, Ip MS, Zheng J, Stocks J, ERS Global Lung Function Initiative. 2012. 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. doi: [10.1183/09031936.00080312](https://doi.org/10.1183/09031936.00080312), PMID: [22743675](https://pubmed.ncbi.nlm.nih.gov/22743675/)
- R Core Team. 2016. R: a language and environment for statistical computing. <http://www.R-project.org/>.
- Rahmani E, Zaitlen N, Baran Y, Eng C, Hu D, Galanter J, Oh S, Burchard EG, Eskin E, Zou J, Halperin E. 2016. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods* **13**:443–445. doi: [10.1038/nmeth.3809](https://doi.org/10.1038/nmeth.3809), PMID: [27018579](https://pubmed.ncbi.nlm.nih.gov/27018579/)
- Ressler KJ, Mercer KB, Bradley B, Jovanovic T, Mahan A, Kerley K, Norrholm SD, Kilaru V, Smith AK, Myers AJ, Ramirez M, Engel A, Hammack SE, Toufexis D, Braas KM, Binder EB, May V. 2011. Post-traumatic stress disorder is associated with PACAP and the PAC1 receptor. *Nature* **470**:492–497. doi: [10.1038/nature09856](https://doi.org/10.1038/nature09856), PMID: [21350482](https://pubmed.ncbi.nlm.nih.gov/21350482/)
- Risch N, Burchard E, Ziv E, Tang H. 2002. Categorization of humans in biomedical research: genes, race and disease. *Genome Biology* **3**:comment2007. PMID: [12184798](https://pubmed.ncbi.nlm.nih.gov/12184798/)
- Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. 2010. Genome-wide association studies in diverse populations. *Nature Reviews Genetics* **11**:356–366. doi: [10.1038/nrg2760](https://doi.org/10.1038/nrg2760), PMID: [20395969](https://pubmed.ncbi.nlm.nih.gov/20395969/)
- Smith ZD, Meissner A. 2013. DNA methylation: roles in mammalian development. *Nature Reviews Genetics* **14**:204–220. doi: [10.1038/nrg3354](https://doi.org/10.1038/nrg3354), PMID: [23400093](https://pubmed.ncbi.nlm.nih.gov/23400093/)

- Tang H**, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ. 2007. Recent genetic selection in the ancestral admixture of Puerto Ricans. *The American Journal of Human Genetics* **81**:626–633. doi: [10.1086/520769](https://doi.org/10.1086/520769), PMID: [17701908](https://pubmed.ncbi.nlm.nih.gov/17701908/)
- Thakur N**, Oh SS, Nguyen EA, Martin M, Roth LA, Galanter J, Gignoux CR, Eng C, Davis A, Meade K, LeNoir MA, Avila PC, Farber HJ, Serebrisky D, Brigino-Buenaventura E, Rodriguez-Cintron W, Kumar R, Williams LK, Bibbins-Domingo K, Thyne S, et al. 2013. Socioeconomic status and childhood asthma in urban minority youths. The GALA II and SAGE II studies. *American Journal of Respiratory and Critical Care Medicine* **188**:1202–1209. doi: [10.1164/rccm.201306-1016OC](https://doi.org/10.1164/rccm.201306-1016OC), PMID: [24050698](https://pubmed.ncbi.nlm.nih.gov/24050698/)
- Tingley D**, Yamamoto T, Hirose K, Keele L, Imai K. 2014. mediation: R package for causal mediation analysis. *Journal of Statistical Software* **59**. doi: [10.18637/jss.v059.i05](https://doi.org/10.18637/jss.v059.i05)
- Udali S**, Guarini P, Moruzzi S, Choi SW, Friso S. 2013. Cardiovascular epigenetics: from DNA methylation to microRNAs. *Molecular Aspects of Medicine* **34**:883–901. doi: [10.1016/j.mam.2012.08.001](https://doi.org/10.1016/j.mam.2012.08.001), PMID: [22981780](https://pubmed.ncbi.nlm.nih.gov/22981780/)
- Udler MS**, Nadkarni GN, Belbin G, Lotay V, Wyatt C, Gottesman O, Bottinger EP, Kenny EE, Peter I. 2015. Effect of Genetic African Ancestry on eGFR and Kidney Disease. *Journal of the American Society of Nephrology* **26**:1682–1692. doi: [10.1681/ASN.2014050474](https://doi.org/10.1681/ASN.2014050474), PMID: [25349204](https://pubmed.ncbi.nlm.nih.gov/25349204/)
- van der Knaap LJ**, Riese H, Hudziak JJ, Verbiest MM, Verhulst FC, Oldehinkel AJ, van Oort FV. 2014. Glucocorticoid receptor gene (NR3C1) methylation following stressful events between birth and adolescence. The TRAILS study. *Translational Psychiatry* **4**:e381. doi: [10.1038/tp.2014.22](https://doi.org/10.1038/tp.2014.22), PMID: [24713862](https://pubmed.ncbi.nlm.nih.gov/24713862/)
- Vidal AC**, Benjamin Neelon SE, Liu Y, Tuli AM, Fuemmeler BF, Hoyo C, Murtha AP, Huang Z, Schildkraut J, Overcash F, Kurtzberg J, Jirtle RL, Iversen ES, Murphy SK. 2014. Maternal stress, preterm birth, and DNA methylation at imprint regulatory sequences in humans. *Genetics & Epigenetics* **6**:37–44. doi: [10.4137/GEG.S18067](https://doi.org/10.4137/GEG.S18067), PMID: [25512713](https://pubmed.ncbi.nlm.nih.gov/25512713/)
- Yudell M**, Roberts D, DeSalle R, Tishkoff S. 2016. Science and society. Taking race out of human genetics. *Science* **351**:564–565. doi: [10.1126/science.aac4951](https://doi.org/10.1126/science.aac4951), PMID: [26912690](https://pubmed.ncbi.nlm.nih.gov/26912690/)
- Ziv E**, John EM, Choudhry S, Kho J, Lorizio W, Perez-Stable EJ, Burchard EG. 2006. Genetic ancestry and risk factors for breast cancer among Latinas in the San Francisco Bay Area. *Cancer Epidemiology Biomarkers & Prevention* **15**:1878–1885. doi: [10.1158/1055-9965.EPI-06-0092](https://doi.org/10.1158/1055-9965.EPI-06-0092), PMID: [17035394](https://pubmed.ncbi.nlm.nih.gov/17035394/)