bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Methylation Analysis Reveals Fundamental Differences Between Ethnicity and Genetic Ancestry

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

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

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

17. Department of Health Sciences, Graduate Program in Public Health, City University of New York, Bronx, NY

* These authors contributed equally to this work

Please address correspondence to:

Esteban G. Burchard, MD, MPH

University of California, San Francisco

Departments of Bioengineering & Therapeutic Sciences and Medicine

UCSF Box 2911

San Francisco, CA 94143-2911

Ph: (415) 514-9677

Fax: (415) 514-4365

e-mail: esteban.burchard@ucsf.edu

or

Noah Zaitlen, PhD

University of California, San Francisco

Department of Medicine

UCSF Box 2552

San Francisco, CA 94143-2552

Ph: (415) 502-2027

e-mail: noah.zaitlan@ucsf.edu

or

Joshua Galanter, MD, MAS

University of California, San Francisco

Departments of Medicine, Bioengineering & Therapeutic Sciences and Epidemiology & Biostatistics

UCSF Box 2911

San Francisco, CA 94143-2552

Ph: (415) 514-9931

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

e-mail: galanter@gmail.com

Sources of Funding:

This research was supported in part by National Institutes of Health (R01 ES015794, R01 HL088133, M01 RR000083, R01 HL078885, R01 HL104608, P60 MD006902, U19 AI077439, M01 RR00188); ARRA grant RC2 HL101651; EGB was supported in part through grants from the Flight Attendant Medical Research Institute (FAMRI), the Sandler Foundation, the American Asthma Foundation and NIH (K23 HL004464); NZ was supported in part by an NIH career development award from the NHLBI (K25HL121295). JMG was supported in part by NIH Training Grant T32 (GM007546) and career development awards from the NHLBI K23 (K23HL111636) and NCATS KL2 (KL2TR000143) as well as the Hewett Fellowship; N.T. was supported in part by a institutional training grant from the NIGMS (T32-GM007546) and career development awards from the NHLBI (K12-HL119997), Parker B. Francis Fellowship Program, and the American Thoracic Society; CRG was supported in part by NIH Training Grant T32 (GM007175) and the UCSF Chancellor's Research Fellowship and Dissertation Year Fellowship; RK was supported with a career development award from the NHLBI (K23HL093023); HJF was supported in part by the GCRC (RR00188); PCA was supported in part by the Ernest S. Bazley Grant; MAS was supported in part by 1R01HL128439-01. This publication was supported by various institutes within the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

In clinical practice and biomedical research populations are often divided categorically into distinct racial and ethnic groups. In reality, these categories comprise diverse groups with highly heterogeneous histories, cultures, traditions, religions, as well as social and environmental exposures. While the factors captured by these categories contribute to clinical practice and biomedical research, the use of race/ethnicity is widely debated. As a response to this debate, genetic ancestry has been suggested as a complement or alternative to this categorization. However, few studies have examined the effect of genetic ancestry, racial/ethnic identity, and environmental exposures on biological processes. Herein, we examine the contribution of self-identification within ethnicity, genetic ancestry, and environmental exposures on epigenetic modification of DNA methylation, a phenomenon affected by both genetic and environmental factors. We typed over 450,000 variably methylated CpG sites in primary whole blood of 573 individuals of Mexican and Puerto Rican descent who also had high-density genotype data. We found that methylation levels at a large number of CpG sites were significantly associated with ethnicity even when adjusting for genetic ancestry. In addition, we found an enrichment of ethnicity-associated sites amongst loci previously associated with environmental and social exposures. Interestingly, one of the strongest associated sites is driven by the Duffy Null blood type variant, demonstrating a new function of the locus in lymphocytes. Overall, the methylation changes associated with race/ethnicity, driven by both genes and environment, highlight the importance of measuring and accounting for both self-identified race/ethnicity and genetic ancestry in clinical and biomedical studies and the benefits of studying diverse populations.

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

Race, ethnicity, and genetic ancestry have had a complex and often controversial history within biomedical research and clinical practice[1,2]. For example, race- and ethnicity-specific clinical reference standards are based on an average derived from statistical modeling applied to population-based sampling on a given physical trait such as pulmonary function[3,4]. However, because race and ethnicity are social constructs, they ignore the heterogeneity within the categories[5]. To account for these heterogeneities and avoid social and political controversies, the genetics community has integrated the use of genetic ancestry as a proxy for race and ethnicity because genetic sequence is not altered by environmental or social factors, such as those related to racial or ethnic identity. Indeed, recent work from our group and others have demonstrated that genetic ancestry improves diagnostic precision compared to crude categorizations of racial/ethnic assignment for specific medical conditions and clinical decisions[6-8]. Herein, we propose that ethnicity, above and beyond genome-wide ancestry, could be correlated to variation in methylation, a fundamental biological process.

Epigenetic modification of the genome through methylation plays a key role in the regulation of diverse cellular processes[9]. Changes in DNA methylation patterns have been associated with complex diseases, including various cancers[10], cardiovascular disease[11,12], obesity[13], diabetes[14], autoimmune and inflammatory diseases[15], and neurodegenerative diseases[16]. Epigenetic changes are thought to reflect influences of both genetic[17] and environmental factors[18]. 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[17], with meQTL's explaining between 22% and 63% of the variance in methylation levels. Multiple environmental

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

factors have also been shown to affect methylation levels, including endocrine disruptors, tobacco smoke[19], polycyclic aromatic hydrocarbons, infectious pathogens, particulate matter, diesel exhaust particles[20], allergens, heavy metals, and other indoor and outdoor pollutants[21]. Psychosocial factors, including measures of traumatic experiences[22-24], socioeconomic status[25,26], and general perceived stress[27], also affect methylation levels.

Given the roles of both genetic and environmental influences upon methylation, we leveraged genome-wide methylation data as an intermediate phenotype to examine the degree to which self-identified ethnicity and genetic ancestry are reflected in differences in methylation. We hypothesized that while genetic ancestry can explain many of the differences in methylation between these groups, some ethnic-specific methylation differences reflecting social and environmental differences between groups, would remain. We further examined the relationship between genome-wide (global) estimates of ancestry and locus-specific (local) ancestry to determine the extent to which associations between global ancestry and methylation are reflective of genetic factors acting in -cis. Finally, by using dense genotyping arrays, we queried whether methylation differences associated with ancestry can be traced back to meQTLs whose allele frequencies differ by ancestry. To address these aims, we analyzed data from 573 Latino children according to their national origin identity or ethnic subgroup (such as Puerto Rican and Mexican), enrolled in the Genes-Environments and Admixture in Latino Americans (GALA II) study of childhood asthma[28].

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

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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). Among the 524 participants with genomic ancestry estimates, European ancestry represented slightly over 50% of the average participant's ancestry, while Native American ancestry had the largest inter-quartile range. There were significant differences in ancestry between ethnic subgroups; Mexicans as a whole had a greater proportion of Native American Ancestry while Puerto Ricans had a significantly greater proportion European and African ancestry [Table 1 and Supplementary Figure 1].

Global patterns of methylation

We first examined whether differences in ethnicity and ancestry resulted in discernible patterns in the global methylation profile by performing multidimensional scaling analysis (Supplementary Figure 2A). We tested for association of each of the first ten principal coordinates (PC) with ethnicity, ancestry, and other confounding factors (Supplementary Figure 2B). As expected[25,29], the first few principal coordinates are strongly correlated to imputed cell composition (Supplementary Figure 2B and C). There are also significant associations of self-identified sub-ethnicity with PC2 (p-ANOVA = 0.003), PC3 (p-ANOVA = 0.004), PC6 (p-ANOVA = 0.0001), PC7 (p-ANOVA = 0.0003) [Supplementary Figure 3A], and PC8 (p-ANOVA = 0.0003), after adjusting for age, sex, cell components, and technical factors (plate and position). Genetic ancestry was associated with PC3 (p-ANOVA = 0.001), PC7 (p-ANOVA = 0.0003) [Supplementary Figure 3B] and PC8 (p-ANOVA = 0.001) in a two degree of freedom ANOVA test, adjusting for age, sex, cell components, technical factors, and ethnicity.

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

Differences in methylation by ethnicity

We next investigated associations between ethnicity and individual loci by performing an epigenome-wide association study of self-identified ethnicity (see methods for details of ascertainment of ethnicity) and methylation. We identified a significant difference in methylation M-values between ethnic groups at 916 CpG sites at a Bonferroni-corrected significance level of less than 1.5×10^{-7} [Figure 1A and Supplementary Table 1]. 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 1B]. A two degree of

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

freedom ANOVA test for genomic ancestry was also significantly associated with methylation level at this site ($p = 2.3 \times 10^{-5}$), 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 ethnic differences beyond geography and ancestry are driving the association.

When we repeat the analysis, adjusting for ancestry, a significant association remained in 314 of the 834 CpG sites associated with ethnicity [Supplementary Figure 4A] (82 sites were excluded because correlations among predictors rendered the models unstable). Genomic ancestry explained a median of 4.2% (IQR 1.8% to 8.3%) of the variance in methylation at these loci and accounts for a median of 75.7% (IQR 45.8% to 92%) of the total variance in methylation explained jointly by ethnicity and ancestry [Supplementary Figure 4B].

Therefore, genetic ancestry explains much of the association between ethnicity and methylation, but other non-genetic factors associated with ethnicity could explain the ethnicity-associated methylation changes that cannot be accounted for by genomic ancestry alone. 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

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 Puerto Ricans was not due to the loss of power from the smaller sample size, we repeated our analysis of the association between ethnicity and ancestry randomly down-sampling to 276 participants to match the sample size in the analysis of geography in Mexicans. While down-sampling the study to this degree resulted in a loss of power, 128 methylation sites were still associated with ancestry. We conclude that recruitment site was unlikely to be a significant confounder of our associations between ethnicity and methylation and was not a significant independent predictor of methylation.

Ethnic differences in environmentally-associated methylation sites

Differences in environmental exposures between ethnic subgroups may explain some of the observed differences in methylation. To investigate this possibility, we identified CpG loci that had previously been reported to be associated with environmental exposures and whose exposure prevalence differs between ethnic groups. We then tested whether methylation at these loci was associated with ethnicity in this study. We have reported that maternal smoking during pregnancy varies significantly by ethnicity[28]. Maternal smoking during pregnancy has also been associated with statistically significant differences in methylation at 26 CpG loci in Norwegian newborns[19]. Of these 26 loci, 19 passed quality control (QC) in our own analysis, and the association between methylation and ethnicity was found to be nominally significant at 6 CpG loci. At a more stringent Bonferroni correction adjusting for 19 tests, cg23067299 in the aryl

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

hydrocarbon receptor repressor (*AHRR*) gene on chromosome 5 remained statistically significant [Table 2]. The protein encoded by *AHRR* participates in the aryl hydrocarbon receptor (AhR) signaling cascade, which mediates dioxin toxicity, and is involved in regulation of cell growth and differentiation. 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[20] were significantly enriched among the set of loci whose methylation levels varied between ethnic groups. Specifically, of the 101 CpG sites that were significantly associated with exposure to DEP and passed QC in our dataset, 31 were nominally associated with ethnicity ($p < 0.05$), and 5 were associated with ethnicity after adjusting for 101 comparisons ($p < 0.005$) [Table 2]. 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[22] and with heavy trauma exposure in adults[23], was significantly associated with ethnicity ($p = 0.02$).

Differences in methylation by ancestry

Our epigenome-wide association study found 194 loci with a significant association between global genetic ancestry and methylation levels at a Bonferroni corrected association p-value of less than 1.6×10^{-7} [Figure 2A and Supplementary Table 2]. Of these significant associations, 55 were driven primarily by differences in African ancestry, 94 by differences in European ancestry, and 45 by differences in Native American ancestry. The most significant association between methylation and ancestry

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

occurred at cg04922029 in the Duffy antigen receptor chemokine gene (DARC) on chromosome 1 (ANOVA p-value 3.1×10^{-24}; Bonferroni corrected p-value 9.9×10^{-19}) [Figure 2B]. 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-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-adjusted = 0.4), demonstrating that ethnic differences in methylation at this site are due to differences in African ancestry. A substantial proportion of the effect of global ancestry on local methylation levels is due to local ancestry acting in -cis. Among the 194 CpG sites associated with global ancestry, local ancestry at the CpG site explained a median of 10.4% (IQR 3.0% to 19.4%) of the variance in methylation at these sites, accounting for a median of 52.8% (IQR 20.3% to 84.9%) of the total variance explained jointly by local and global ancestry [Supplementary Figure 5].

Admixture mapping of methylation

We next performed an admixture mapping study, examining the association between methylation levels at each CpG site and ancestry at the same locus. Of the 321,503 CpG's examined, methylation at 3,694 (1.1%) was significantly associated with ancestry at the CpG site at a Bonferroni corrected association p-value of less than 1.6×10^{-7}. [Figure 3A

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

and Supplementary Table 3] This included 118 of the 194 loci identified above (61%), where global ancestry was associated with methylation. The most significant CpG site was again cg04922029, which was almost perfectly correlated with African ancestry at the locus ($p = 6 \times 10^{-162}$) [Figure 3B]. Each African haplotype at the CpG site was associated with an increase in methylation M-value of 2.7, corresponding to a 6.5-fold increase in the ratio of methylated to unmethylated DNA per African haplotype at that locus. The second most significant association occurred at cg06957310 on chromosome 17; each increase in African ancestry at the locus was associated with a decrease in M-value of 1.7 (a 3.2-fold decrease in the ratio of methylated to unmethylated DNA; $p = 3.7 \times 10^{-75}$).

Finally, we explored whether our admixture mapping results were indicative of the presence of a meQTL. For each of the admixture mapping loci, we tested whether a single nucleotide polymorphism (SNP) within 1 Mb from the CpG was associated with methylation. We found 3637 loci out of the 3694 (98.5%) admixture mapping findings with at least one SNP within 1 Mb that was significantly associated with methylation levels (after adjustment of the number of SNPs in cis-). The SNP/CpG pair were separated by a median distance of 10.9 kb (interquartile range 2.9 kb to 35.1 kb). The furthest SNP/CpG pair were 998 kb apart. The most significant SNP/CpG pair was cg25134647/rs4963867, on chromosome 12, which are separated by 412 base pairs. Each copy of the T allele was associated with a decrease in M-value of 3.58, corresponding to a nearly 12-fold decrease in the ratio of methylated to unmethylated DNA at the site. We found that CpG cg04922029 (our top admixture mapping association) was significantly correlated with SNP rs2814778 [Figure 3C], the Duffy null mutation, 212 base pairs away; each copy of the C allele was associated with an increase

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 3D].

Discussion

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

We conclude that systematic environmental differences between ethnic subgroups likely play an important role in shaping the methylome for both individuals and populations. Loci previously associated with diverse environmental exposures such as *in utero* exposure to tobacco smoke[19], diesel exhaust particles[20], and psychosocial stress[22] were enriched in our set of loci, where methylation was associated with ethnicity. 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.

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[30]. The strongest patterns of methylation are associated with cell composition in whole blood[25].

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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[31]. Also, our models of genetic ancestry assumed a linear effect of ancestry on methylation. Although we would expect the effects to be small, 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. To rule out any residual confounding due to recruitment sites, we conducted an additional analysis on the effect of recruitment site on methylation both for the overall study and for the Mexican participants (the largest study population in this analysis). We observed no significant independent effect of recruitment site suggesting that confounding due to recruitment region was limited, at least within the United States. We were unable to test for the effect of geographic differences between the United States and Puerto Rico because our study included relatively few Puerto Ricans recruited outside of Puerto Rico.

The presence of a strong association between genetic ancestry and methylation raises the possibility that epigenetic studies can be confounded by population stratification, similar to genetic association studies, and that adjustment for either genetic ancestry or selected principal components is warranted. This possibility was first demonstrated in a previous analysis of the association between self-described race and methylation[32]. However, the study only evaluated two distinct racial groups (African Americans and Whites), while the present study demonstrates the possibility of population

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 that do not adjust for ethnic heterogeneity is likely to result in spurious associations even if genomic ancestry is included as a covariate.

Our analysis of local genetic ancestry and methylation demonstrates that loci associated with genome-wide ancestry are driven primarily by allele frequency differences between ancestral populations in 118 out of 194 loci, suggesting that in most cases global ancestry is acting in -cis. In addition, methylation-QTLs whose allele frequencies differ between ethnic groups are found in 95% (3637/3694) of loci associated with local ancestry. Of particular interest, the most significant ancestry-associated locus, the *DARC* gene, harbors an association between ancestry and methylation at cg04922029, which can be entirely explained by the genotype at rs2814778, the Duffy null mutation. This mutation, which confers resistance to *P. vivax* malaria, has an allele frequency of 100% in individuals in the five 1000 genomes populations[33-35] from Africa (Esan in Nigeria, Gambian in the Western Division of Gambia, Luhya in Webuye, Kenya, Mende in Sierra Leone, and Yoruban in Ibadan, Nigeria), and 80% to 90% in admixed populations of African origin in the Americas (89% in Afro Caribbeans in Barbados, and 80% in African Americans in the Southwest U.S.). In contrast, the allele frequency is 1% in the five European populations and 0% in the five Asian populations [Figure 3E], consistent with the known high level of ancestry information at the locus. Latinos, being admixed,

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

have intermediate and more varied minor allele frequencies, ranging from 3% in Mexicans to 14% in Puerto Ricans.

In summary, the present 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 some of the ethnic differences in methylation in this study, ethnic identity is associated with methylation independent of genetic ancestry. This may be 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 profound implications for the independent and joint effects of race, ethnicity, and genetic ancestry in biomedical research and clinical practice, especially in studies conducted in diverse or admixed populations. Our conclusions may be generalizable to any population that is racially mixed such as those from South Africa, India, and Brazil, though we would encourage further study in diverse populations. As the National Institutes of Health (NIH) embarks on a precision medicine initiative, this

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

research underscores the importance of including diverse populations and studying factors capturing the influence of social, cultural, and environmental factors, in addition to genetic ones, upon disparities in disease and drug response.

Methods

Participants

Institutional review boards at University of California, San Francisco and recruitment sites approved the study, 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[28]. A total of 4,702 children (2,374 participants with asthma and 2,328 healthy controls) were recruited from five centers (Chicago, Bronx, Houston, San Francisco Bay Area, and Puerto Rico) using a combination of community- and clinic-based recruitment. Participants were eligible if they were 8-21 years of age and self-identified as a specific Latino ethnicity and had four Latino grandparents. Asthma cases were defined as participants with a history of physician diagnosed asthma and the presence of two or more symptoms of coughing, wheezing, or shortness of breath in the 2 years preceding enrollment. Participants were excluded if they reported any of the following: (1) 10 or more pack-years of smoking; (2) any smoking within 1 year of recruitment date; (3) history of lung diseases other than asthma (cases) or chronic illness (cases and controls); or (4) pregnancy in the third trimester. Further details of recruitment are described elsewhere[28]. Latino sub-ethnicity was determined by self-identification and the ethnicity of the participants' 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.

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

Failed probes were identified using detection p-values using Illumina's recommendations. Probes on sex chromosomes and those known to contain genetic polymorphisms in the probe sequence were also excluded, leaving 321,503 probes for analysis. Raw data were normalized using Illumina's control probe scaling procedure. Beta values of methylation (ranging from 0 to 1) were converted to M-values via a logit transformation[36].

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Genotyping

Details of genotyping and quality control procedures for single nucleotide polymorphisms (SNPs) and individuals have been described elsewhere[37]. Briefly, participants were genotyped at 818,154 SNPs on the Axiom® Genome-Wide LAT 1, World Array 4 (Affymetrix, Santa Clara, CA)[38]. We removed SNPs with >5% missing data and failing platform-specific SNP quality criteria (n=63,328), along with those out of Hardy-Weinberg equilibrium (n=1845; $p<10^{-6}$) within their respective populations (Puerto Rican, Mexican, and other Latino), as well as non-autosomal SNPs. Subjects were filtered based on 95% call rates and sex discrepancies, identity by descent and standard Affymetrix Axiom metrics. The total number of participants passing QC was 3,804 (1,902 asthmatic cases, 1,902 healthy controls), and the total number of SNPs passing QC was 747,129. The number of participants with both methylation and genotyping data was 524.

Ancestry estimation

GALA II participants were combined with ancestral data from 1000 Genomes European (CEU) and African (YRI) populations and 71 Native American (NAM) samples genotyped on the Axiom® Genome-Wide LAT 1 array. A final sample of 568,037 autosomal SNPs with relevant ancestral data was used to estimate local and global ancestry. Global ancestry was estimated using the program ADMIXTURE[39], with a three population model. Local ancestry at all positions across the genome was estimated using the program LAMP-LD[40], assuming three ancestral populations.

Statistical Analysis

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

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the cell type makeup of whole blood we inferred white cell counts using the method by Houseman et al[29]. 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, n-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 n-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)[41] and the Bioconductor package version 2.13.

Global patterns of methylation

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 [Supplementary Figure 2A]. We performed a simple linear regression 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 Supplementary Figure 2B].

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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"[42] for methylation principal coordinates, which demonstrated a significant association with ethnicity.

Local patterns of methylation

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

We also examined whether differences in methylation patterns by ethnicity could be associated with known loci that had previously been reported to vary based on common environmental exposures, including maternal smoking during pregnancy[19], diesel exhaust particles (DEP)[20], and exposure to violence[22]. We have previously shown that exposure to these common environmental exposures or similar exposures varied by ethnicity within our own GALA II study populations[28,43,44].

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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) and local ancestry (γ) in a joint model by calculating the proportion of variance explained as above.

cis-Admixture mapping

To determine whether ancestry associations with methylation were due to variation in local ancestry, we performed a cis-admixture mapping study, comparing estimates of local ancestry at each CpG site with methylation at the site. Because ancestry LD is much stronger than genotypic LD, it is possible to accurately interpolate ancestry at each CpG site based on the ancestry estimated at the nearest SNPs[37,45]. 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.

Allelic associations

In order to determine the extent to which admixture mapping results could be explained by allelic associations, we performed a meQTL analysis at all Bonferroni-corrected significant admixture mapping associations ($p < 1.6\times10^{-7}$), by comparing methylation at a given locus with the genotype of SNPs within 1 MB of the CpG site using an additive genotypic model, adjusted for both global and local genomic ancestry, demographic

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

variables including ethnicity, estimated cell proportions, case status, and technical factors. The significance threshold was based on Bonferroni correction for the number of SNPs within 1 MB of the CpG site.

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.

Tables

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

	Mexican	Puerto Rican	Mixed Latino	Other Latino
n	276	220	16	61
Males (%)	125 (45.3%)	127 (57.7%)	6 (37.5%)	28 (45.9%)
Age	11.4 [9.3: 14.7]	12.3 [10.4: 14.2]	11.8 [10.7: 14.9]	11.8 [10: 15.7]
Asthma cases (%)	124 (44.9%)	147 (66.8%)	9 (56.3%)	31 (50.8%)
Ancestry (n = 524)				
African	4.3% [2.9%: 6.0%]	22.8% [16.6%: 29.4%]	8.5% [5.6%: 19.2%]	12.3% [6.3%: 25.8%]
Native American	55.4% [44.5%: 65.7%]	11.2% [9.8%: 13%]	31.5% [20.9%: 45.6%]	32.8% [10.4%: 49.3%]
European	40.5% [29.9%: 50.2%]	65.7% [59.2%: 71%]	50.5% [44.6%: 57.6%]	48.9% [40%: 58.5%]
Recruitment Site				
Chicago	140 (50.7%)	15 (6.8%)	11 (68.9%)	15 (24.6%)
New York	18 (6.5%)	10 (4.5%)	1 (6.3%)	23 (37.7%)

Puerto Rico	0	193 (87.7%)	0	0
San Francisco	78 (28.3%)	0	2 (12.5%)	23 (37.7%)
Houston	40 (14.5%)	2 (0.9%)	2 (12.5%)	5 (8.2%)
Cell Counts (estimated)				
Granulocytes	51.2% [46.0%: 55.7%]	51.6% [46.8%: 57%]	51% [43.6%: 57.2%]	49.1% [43.8%: 55.8%]
Lymphocytes	41.9% [36.9%: 46.6%]	41.8% [36.9%: 46.5%]	41.9% [36.1%: 51.6%]	43.9% [36.8%: 49.6%]
Monocytes	7.1% [5.8%: 8.3%]	6.74% [5.74%: 8.24%]	6.6% [5.7%: 7.6%]	7.4% [6.2%: 8.6%]

For continuous variables, the mean and interquartile range are displayed. For categorical variables, the number and proportion of subjects in each category are shown.

TABLE 2: Significant associations between ethnicity and methylation loci previously associated with environmental exposures.

Exposure	CpG	Gene	Chr	Position	p	p (adj)
Diesel Exhaust Particles	cg05084827	*RPS27A*	2	55402999	0.0002	0.02
Diesel Exhaust Particles	cg27457191	*PHTF2*	7	77429766	0.0004	0.04
Diesel Exhaust Particles	cg06106484	*TRNA (Pseudo)*	8	32985289	0.0004	0.04
Diesel Exhaust Particles	cg07462448	*CASP7*	10	115441840	0.00005	0.005
Diesel Exhaust Particles	cg08131547	*ZNF121*	19	9691569	0.0002	0.02
Exposure to Violence	cg11218385	*ADCYAP1R1*	7	31092854	0.03	0.03
Prenatal Smoke	cg23067299	*AHRR*	5	323907	0.0002	0.004

Unadjusted p-values correspond to the ANOVA p-value between ethnicity and methylation; adjusted p-values are Bonferroni corrected.

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure Legends

FIGURE 1: [A] Manhattan plot showing the associations between ethnicity and methylation at individual CpG loci. [B] Violin plot showing one such locus, cg19145607. Mexicans are relatively hypermethylated compared to Puerto Ricans ($p = 1.4 \times 10^{-19}$). [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.

FIGURE 2: [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.

FIGURE 3: [A] Manhattan plot showing the association between local ancestry and methylation at individual CpG loci. [B] Association between cg04922029 on the *DARC* locus and African ancestry, color coded by ethnic group. There is near perfect correlation between the two. [C] Association between SNPs located within 1Mb of cg04922029 and methylation levels at that CpG. [D] Association between rs2814778 (Duffy null) genotype and methylation at cg04922029, color coded by the number of African alleles present. There is near perfect correlation between genotype, ancestry and methylation at the locus. [E] Allele frequency of rs2814778 by 1000 Genomes population. The C allele is nearly ubiquitous in African populations and nearly absent outside of African populations and their descendants.

SUPPLEMENTAL 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.

SUPPLEMENTARY FIGURE 2: [A] Distribution of the first 10 principal coordinates of the methylation data. Plots in the diagonal show the univariate distribution; those in the lower left triangle show bivariate relationship between each pair of PCs, while those in the upper right show the bivariate density. [B] Bivariate or ANOVA associations between principal coordinates and technical factors (chip, position), cell counts, genetic

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

ancestry (European, Native American, African), recruitment site (New York, NY, San Francisco, CA, Chicago, IL, Houston, TX, and Puerto Rico), demographic factors (ethnicity, age, sex), and case status. [C] Correlation coefficients between the various factors and principal coordinates.

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

SUPPLEMENTARY FIGURE 4: Relationship between genomic ancestry and the association between ethnicity and methylation. [A] Venn diagram showing the effect of adjustment for ancestry on the association between ethnicity and methylation. The components of the diagram represent the number of CpG's that remained associated with ethnicity after adjustment for ancestry and the number of CpG's that were associated with ancestry. [B] Relative proportion of variance in methylation explained by ethnicity and genomic ancestry across loci significantly associated with ethnicity. Mediation analysis of associations between ethnicity and methylation M-values for [C] Native American ancestry and [D] African ancestry. For simplicity, only significant mediation effects are shown.

SUPPLEMENTAL FIGURE 5: Relative proportion of variance in methylation explained by global and local ancestry across loci significantly associated with global ancestry.

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

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

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

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

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

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

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

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

8. Nalls, M. A. *et al.* Admixture Mapping of White Cell Count: Genetic Locus Responsible for Lower White Blood Cell Count in the Health ABC and Jackson Heart Studies. *The American Journal of Human Genetics* **82,** 81–87 (2008).

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

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

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

12. Kato, N. *et al.* Trans-ancestry genome-wide association study identifies 12 genetic

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

loci influencing blood pressure and implicates a role for DNA methylation. *Nat. Genet.* (2015). doi:10.1038/ng.3405

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

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

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

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

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

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

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

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

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

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

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

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

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

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

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

33. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

34. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

35. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

genomes. *Nature* **526,** 75–81 (2015).

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

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

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

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

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

41. Team, R. C. R: A language and environment for statistical computing. at <http://www.R-project.org/>

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

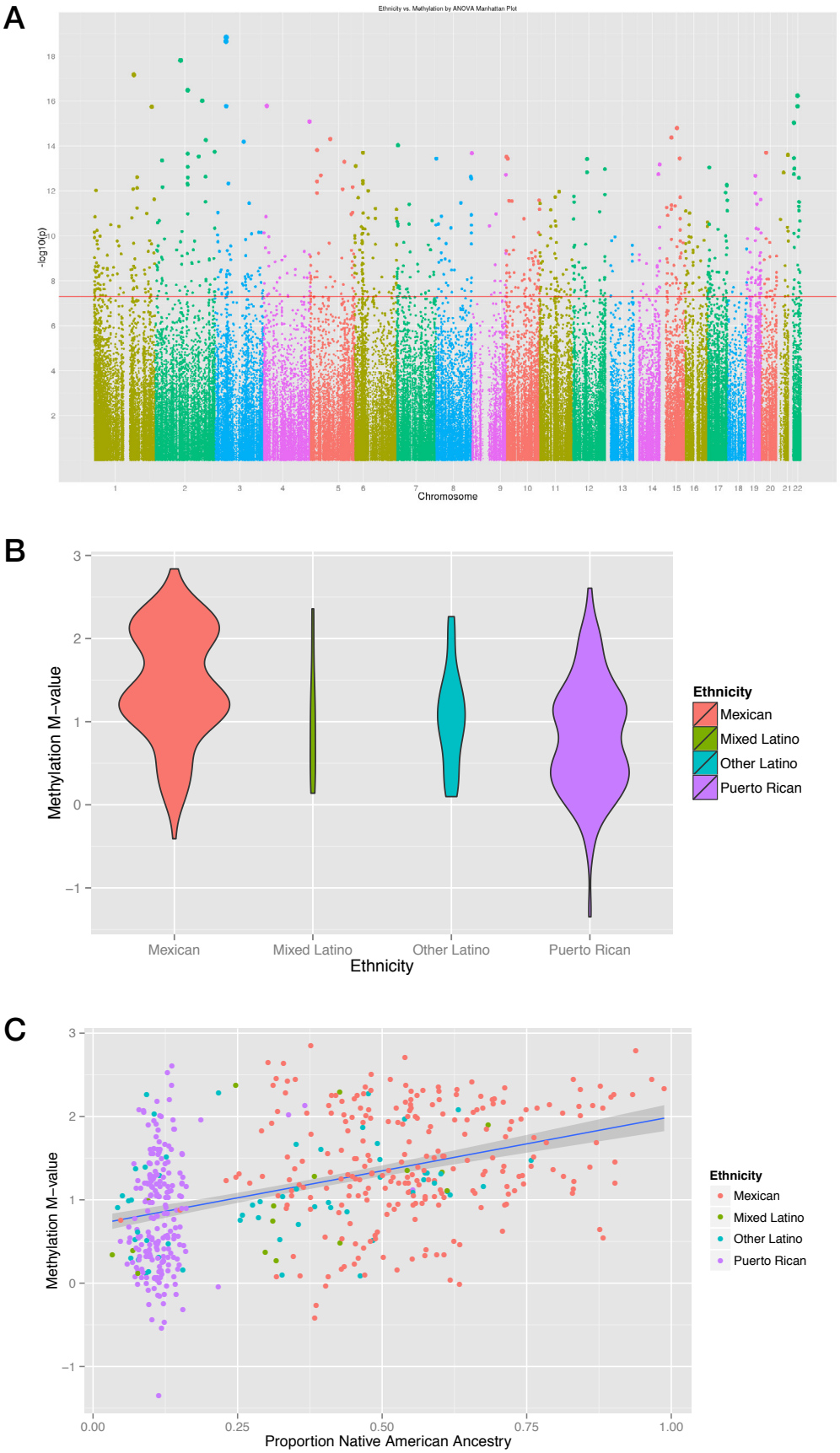
43. Nishimura, K. K. *et al.* Early-life air pollution and asthma risk in minority children. The GALA II and SAGE II studies. *Am. J. Respir. Crit. Care Med.* **188,** 309–318 (2013).

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

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

bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

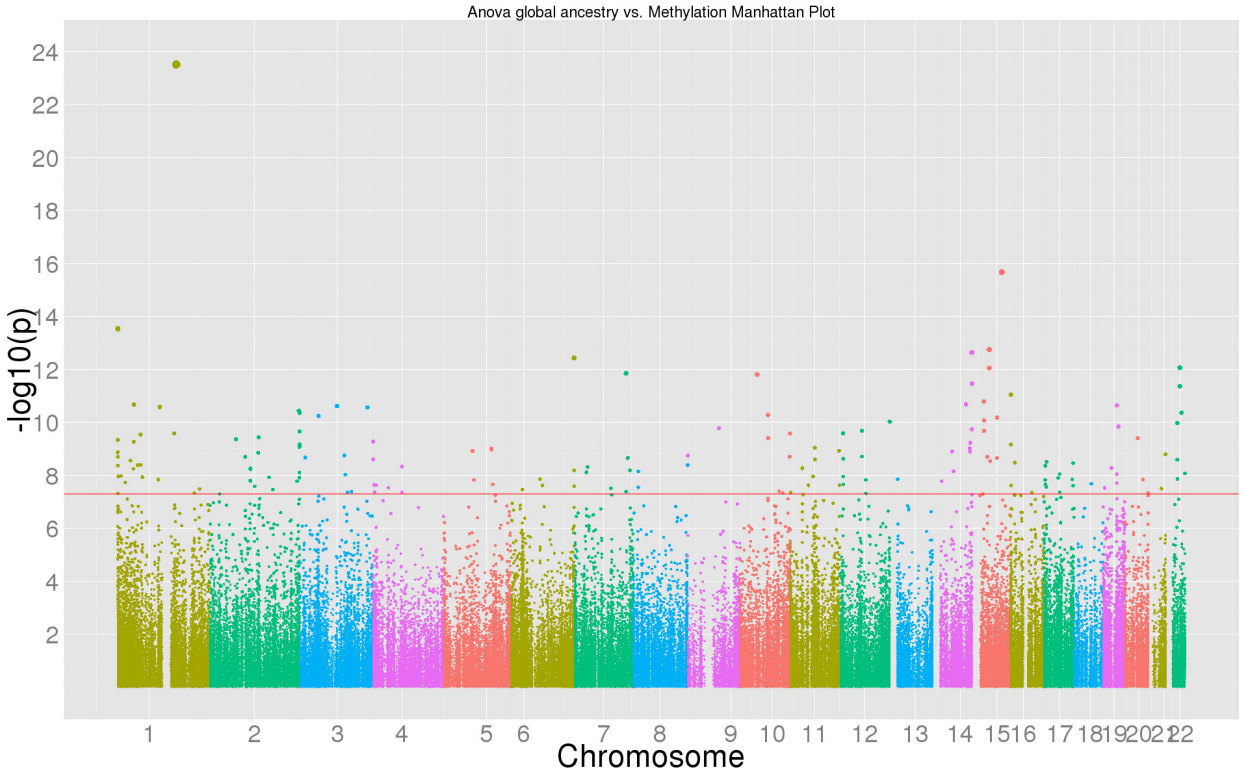
Figure 1



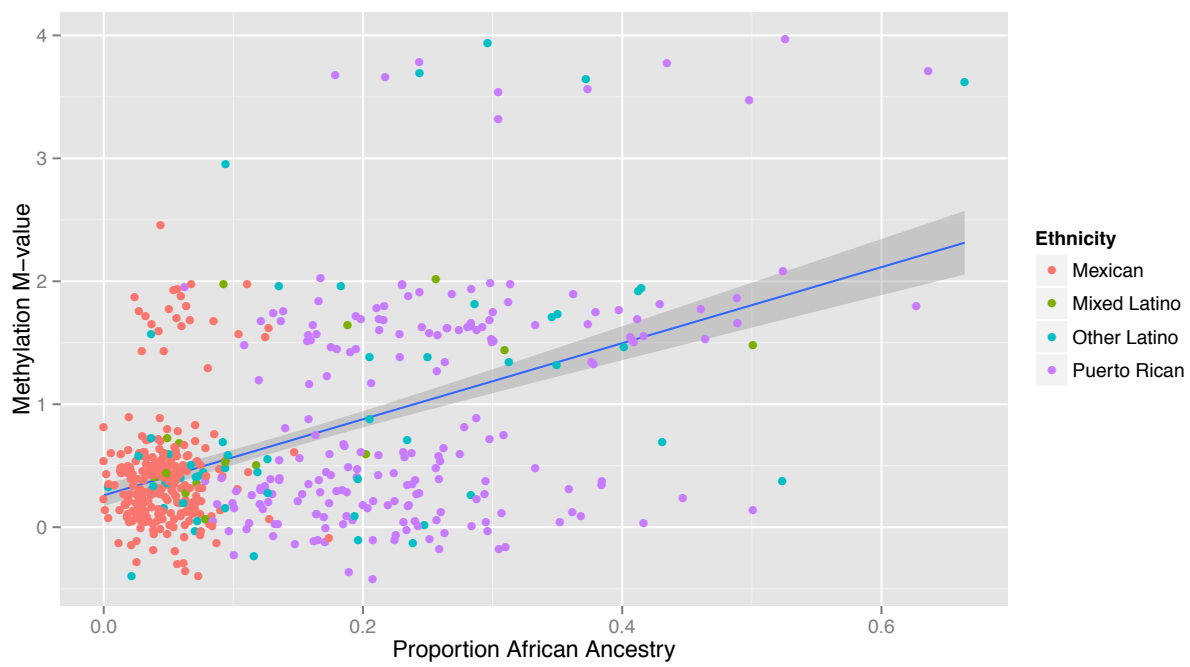
bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 2

A



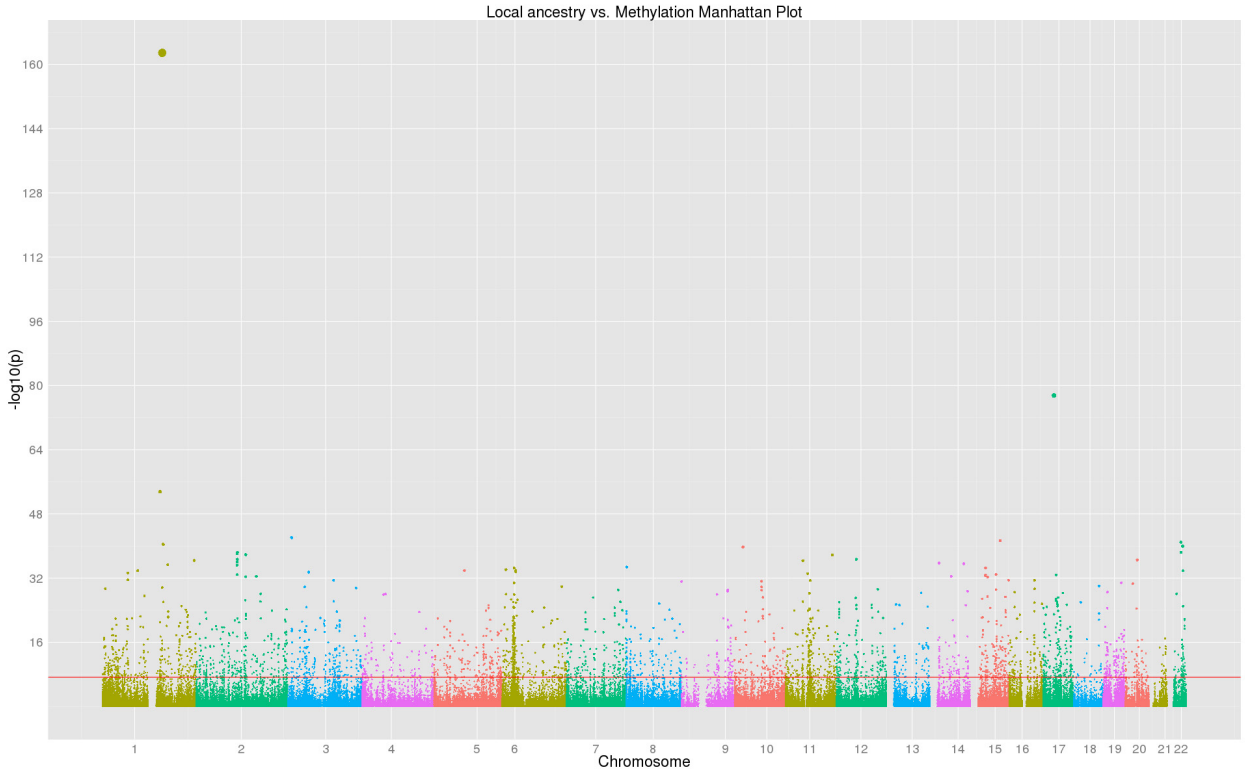
B



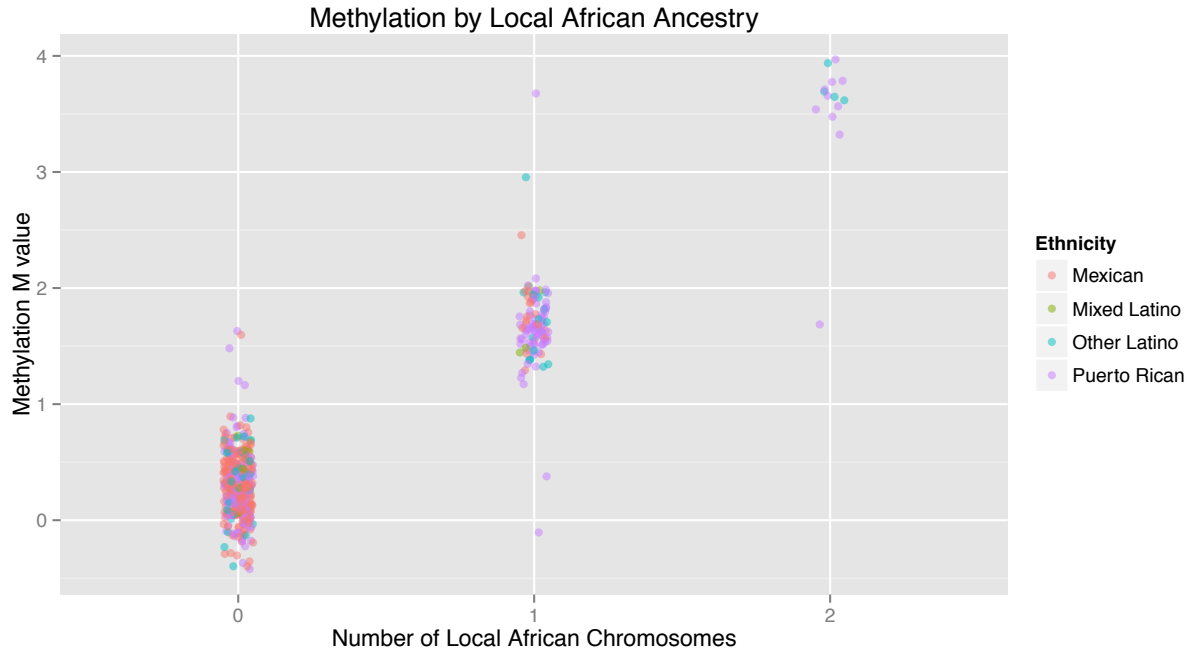
bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 3

A

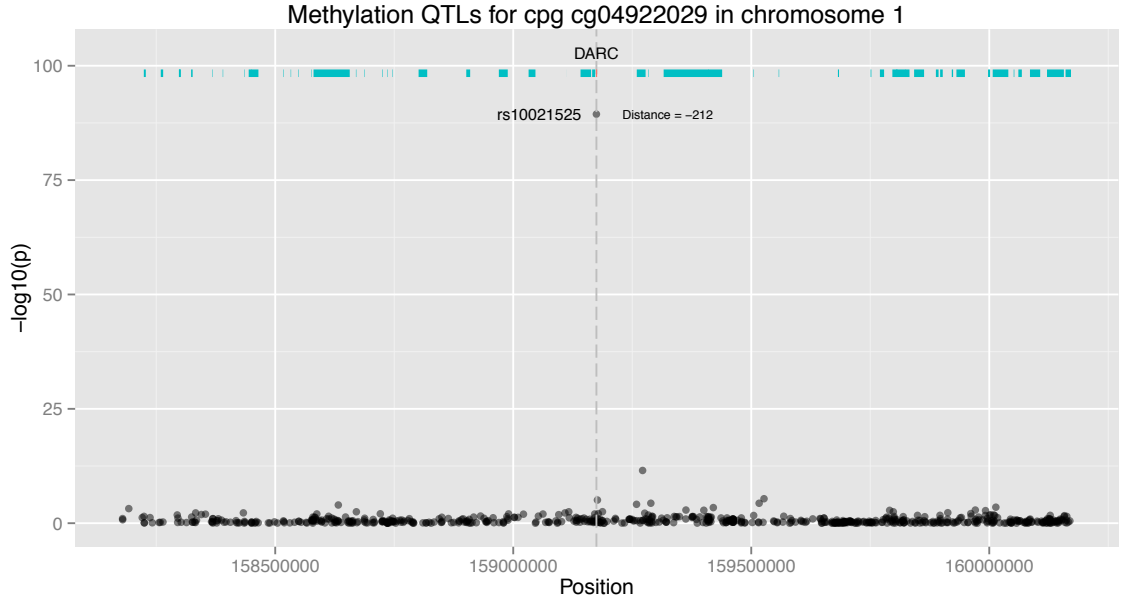


B



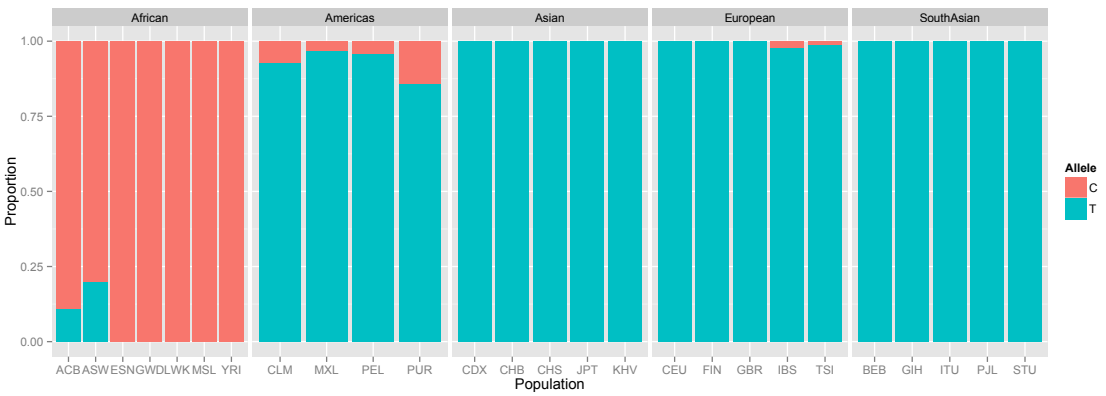
bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

C



D

E

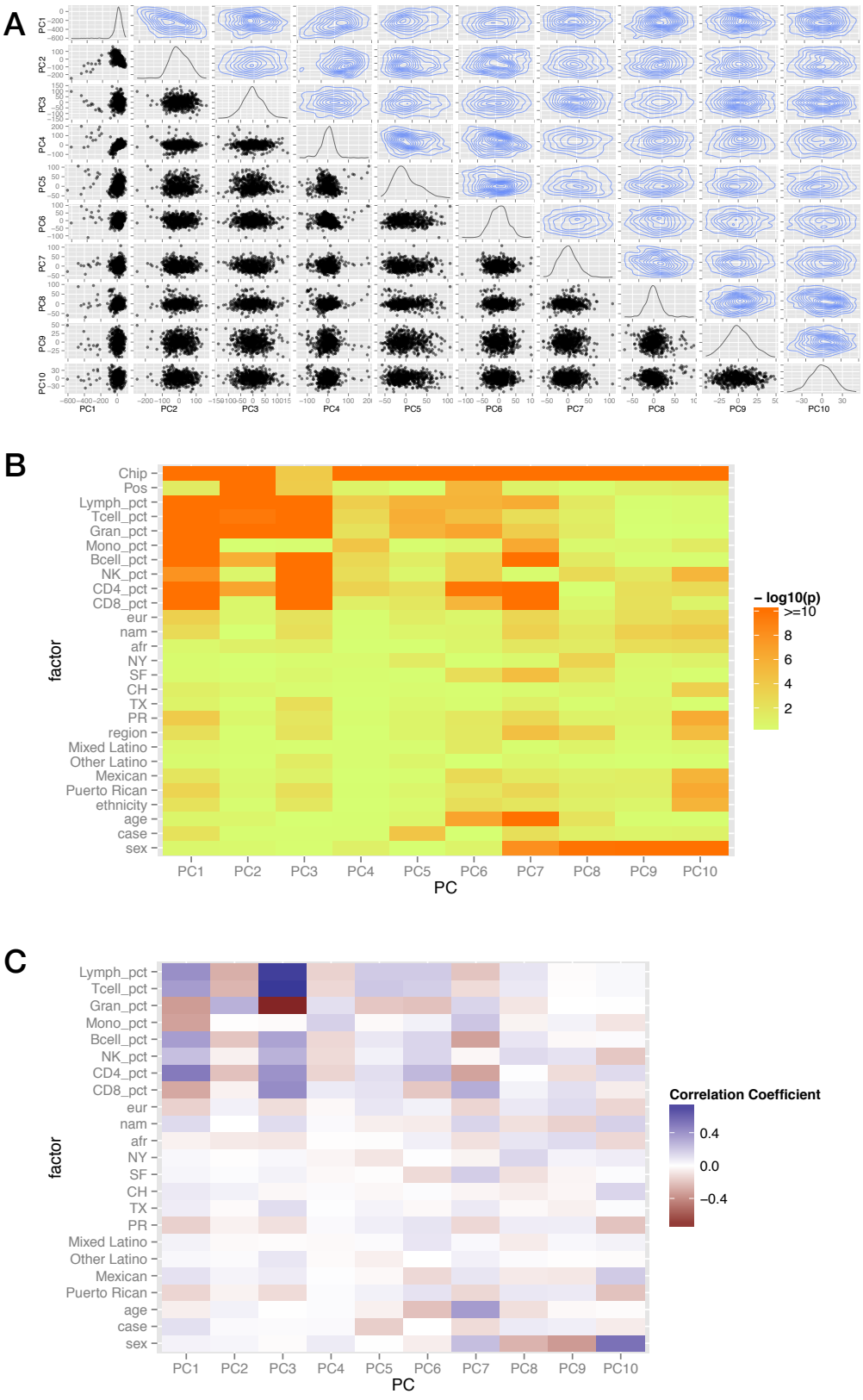


bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Figure 1

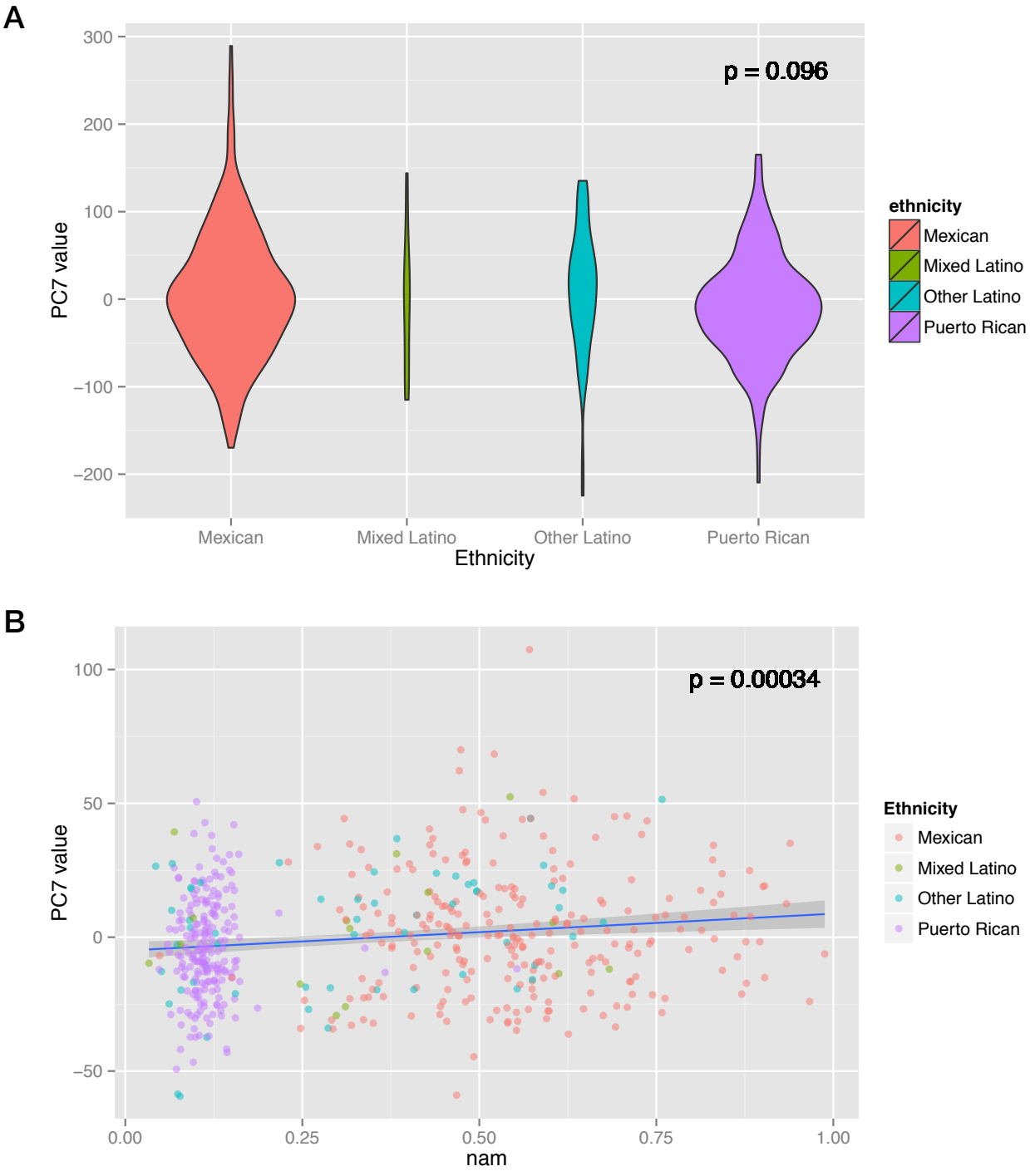
bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Figure 2



bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

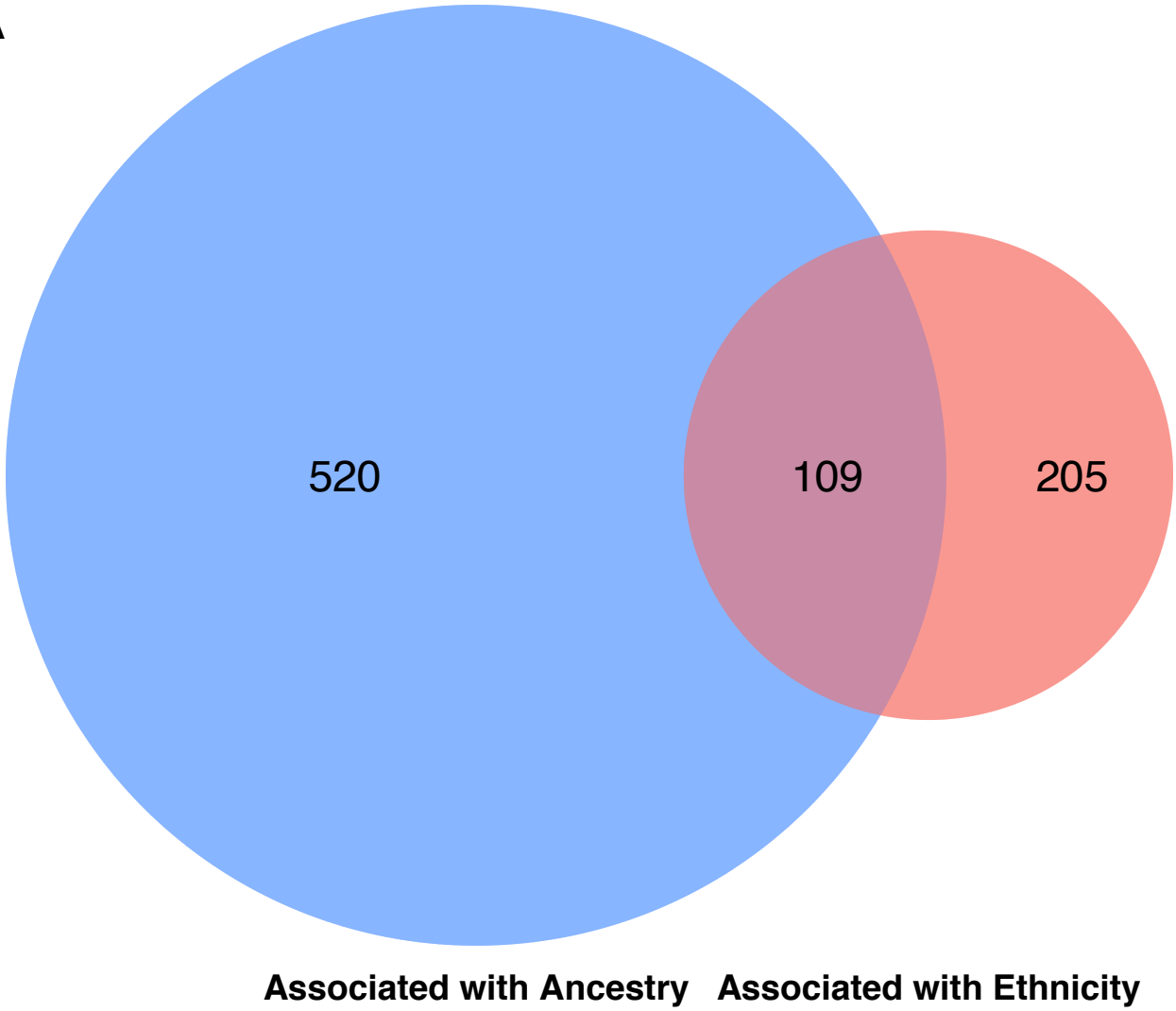
Supplementary Figure 3



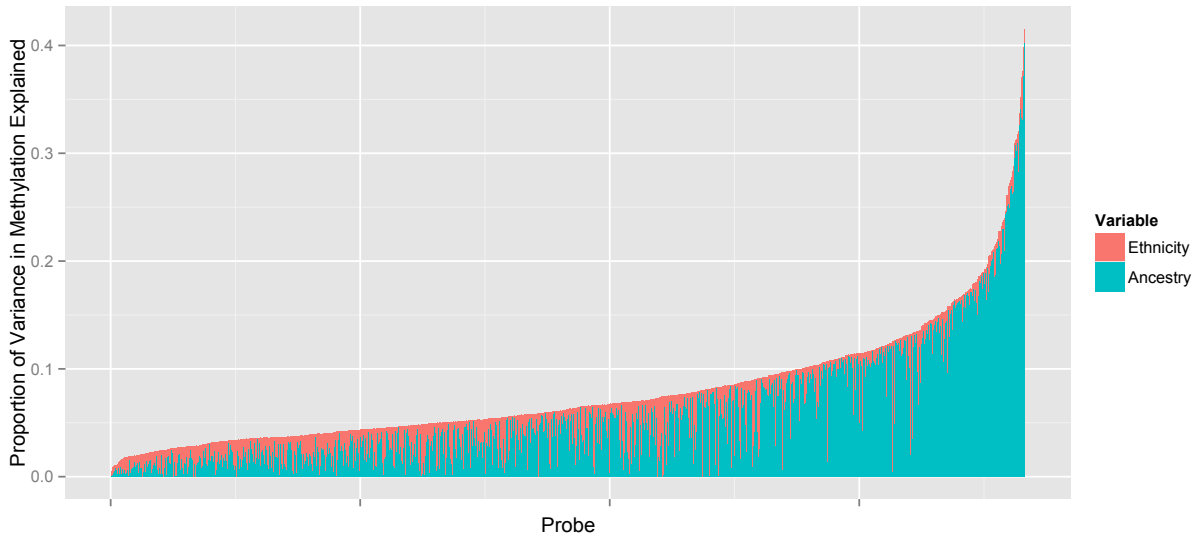
bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

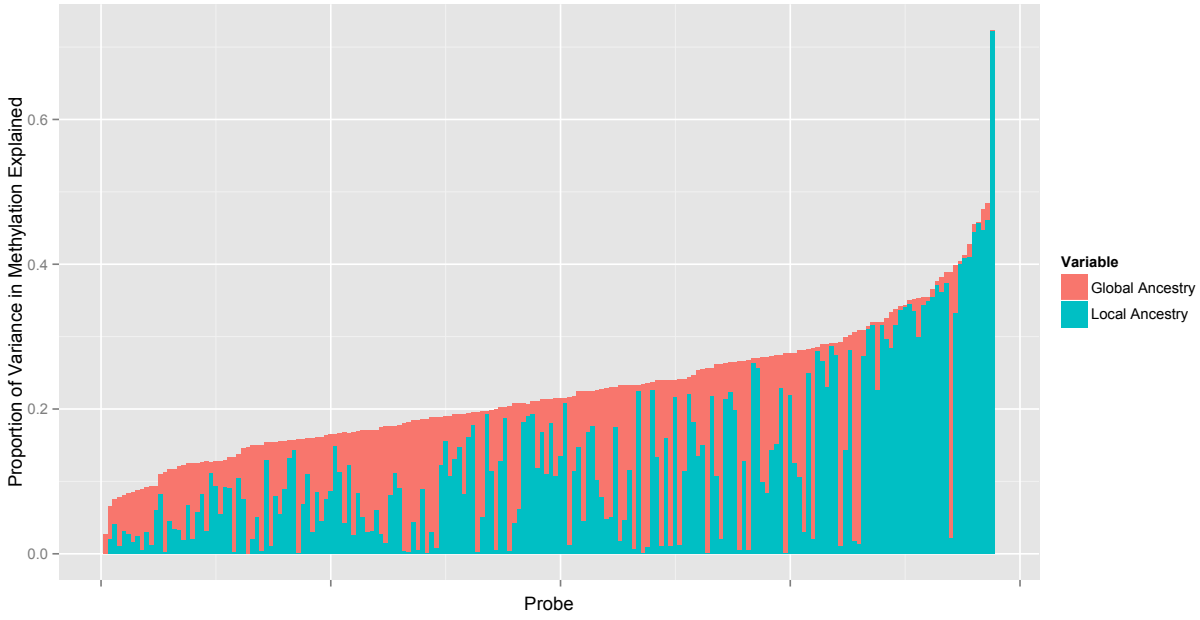
Supplementary Figure 4

A



B



bioRxiv preprint doi: https://doi.org/10.1101/036822; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Figure 5