Genome-wide methylation data mirror ancestry information

Elior Rahmani[1], Liat Shenhav[2], Regev Schweiger[1], Paul Yousefi[3], Karen Huen[3], Brenda Eskenazi[3], Celeste Eng[4], Scott Huntsman[4], Donglei Hu[4], Joshua Galanter[4,5], Sam Oh[4], Melanie Waldenberger[6,7], Konstantin Strauch[8,9], Harald Grallert[6,7,10], Thomas Meitinger[11,12], Christian Gieger[6,7,10], Nina Holland[3], Esteban Burchard[4,5], Noah Zaitlen[4], and Eran Halperin[1,13,14]

[1]Blavatnik School of Computer Science, Tel-Aviv University, Tel Aviv, Israel

[2]Department of Statistics, Tel Aviv University, Tel Aviv, Israel

[3]School of Public Health, CERCH, University of California, Berkeley, CA, USA

[4]Department of Medicine, University of California San Francisco, San Francisco, California

[5]Department of Bioengineering and Therapeutic Science, University of California San Francisco, San Francisco, California

[6]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München Research Center for Environmental Health, Neuherberg, Germany

[7]Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

[8]Institute of Genetic Epidemiology, Helmholtz Zentrum München Research Center for Environmental Health, Neuherberg, Germany

[9]Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany

[10]German Center for Diabetes Research (DZD), Neuherberg, Germany

[11]Institute of Human Genetics, Helmholtz Zentrum München, Munich, Germany

[12]Institute of Human Genetics, Technische Universität München, Munich, Germany

[13]International Computer Science Institute, Berkeley, California

[14]The Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel Aviv, Israel

July 5, 2016

Abstract

Genetic data are known to harbor information about human demographics, and genotyping data are commonly used for capturing ancestry information by leveraging genome-wide differences between populations. In contrast, it is not clear to what extent population structure is captured by whole-genome DNA methylation data. We demonstrate, using three large cohort 450K methylation array data sets, that ancestry information signal is mirrored in genome-wide DNA methylation data, and that it can be further isolated more effectively by leveraging the correlation structure of CpGs with cis-located SNPs. Based on these insights, we propose a method, Epistructure, for the inference of ancestry from methylation data, without the need for genotype data. Epistructure can be used to correct epignome-wide association studies (EWAS) for confounding due to population structure.

1 Introduction

The relation between ancestry and genetic variation has been repeatedly established over the last decade [1,2]. Several methods now provide accurate estimates of ancestry information by leveraging genome-wide systematic difference in allele frequencies between subpopulations, commonly referred to as population structure [3–7]. These methods often apply Principal Component Analysis (PCA) or variants of PCA.

Inferring population structure across individuals provides a powerful source of information for various fields, including genetic epidemiology, pharmacogenomics and anthropology. For instance, in genetic and molecular epidemiology, in which identifying genetic associations with disease or exposure is of primary interest, it is essential to have ancestry information in order to distinguish effects of demographic processes from biological or environmental effects. Specifically, the importance of controlling for population structure in genome-wide association studies (GWAS) is now well appreciated. Unless appropriately accounted for, population structure in GWAS can lead to numerous spurious associations and might obscure true signal [4, 8].

Emerging epigenome-wide association studies (EWAS) revealed thousands of CpG methylation sites correlated with genetics and with ancestry [9–20]. Not surprisingly, due to the genetic signal present in many CpGs, several studies have shown that the first several principal components (PCs) of methylation data capture population structure in some instances of data sets composed of European and African individuals [15, 21]. However, unlike the case of genotyping data in which global ancestry information is robustly captured by the top PCs, the first several PCs of methylation data were also shown to capture other factors in different scenarios, mainly cell type composition in case of data collected from heterogeneous source [22, 23], but also other factors, including technical variables, age and sex [15, 21]. Moreover, it is now appreciated that collecting methylation using probes with polymorphic CpGs is affected by hybridization sensitivity and thus not necessarily reflecting methylation variability but rather genetic variability [24], and therefore it is not clear to what extent global whole-genome DNA methylation states are affected by population structure when these artifacts are removed.

We hereby introduce Epistructure, a method for capturing ancestry information from DNA methylation. Epistructure is based on the observation that PCA computed from a set of methylation CpG sites that are highly correlated with SNPs efficiently captures population structure. Thus, we use a large reference data set that includes both genotypes and methylation in order to find correlations of CpGs with cis-located SNPs, and compile a reference list of genetically-informative CpGs. Then, given new methylation data we compute the principal components (PCs) of the methylation levels from the same sites included in the reference list. We evaluate the correlation between the methylation-inferred ancestry and the genetically inferred ancestry on two additional large methylation data sets.

In order to shed additional light on the relation between genetic ancestry and methylation-based ancestry, we further explore the unsupervised detection of ancestry from methylation data. We show that genome-wide methylation mirrors ancestry information in admixed populations after properly adjusting for known variability in genome-wide methylation, and after properly removing artificial artifacts, particularly probes that include SNPs that may confound the results. Thus, unlike previous studies that were potentially confounded by these artifacts, here we show that ancestry is indeed robustly mirrored by methylation data as one of the main principal variance components.

2 Results

Inferring ancestry from methylation data using Epistructure. Ancestry information signal in methylation is mostly expected to exist due to the large number of correlations between methylation sites and genetics [9–20]. We developed Epistructure, a method for the inference of ancestry from methylation data, which relies on reference data in which both genotype and methylation data are available. In a nutshell, Epistructure selects a set of CpGs that are highly correlated with genotype information in the reference data, and then, given new methylation data, performs principal component analysis on these sites while taking into account the cell type composition effects. More specifically, we use the KORA cohort of European adults (n=1,799), for which both whole-blood methylation and genotyping data are available [25] (see Methods). We fitted a regularized linear regression model for each CpG from SNPs in cis, and evaluated it based on a cross-validated linear correlation (see Methods). Since the vast majority of reported CpG-SNP associations are between CpGs and cis-located SNPs [9–11], we only considered cis-located SNPs in capturing the genetic component of each CpG. We observed that for most CpGs only a small fraction of the variance can be explained by cis-SNPs ($R^2 < 0.1$ for 92.9% of the CpGs tested; Supplementary Figure S1), thus motivating the use of only a relatively small subset of the CpGs for inferring ancestry information.

In order to test the performance of Epistructure we applied it on the GALA II data set ($n = 479$), a pediatric Latino population study with Mexican (MX) and Puerto-Rican (PR) individuals [26], for which both genotypes and 450K methylation array data (whole-blood) were available (see Methods). First, we computed the largest (first) two PCs of the genotypes (genotype-based PCs), known to capture population structure [4]. We observed that the first PC of Epistructure captured the top genotype-based PC well ($R^2 = 0.82$), as compared to the first PC of the methylation data (methylation-based PC; $R^2 = 0.01$) and as compared to the methylation-based PC computed only from CpGs residing in close proximity to nearby SNPs ($R^2 = 0.01$), as was suggested in a recent study for capturing ancestry information in methylation data [21]. More generally, we observed that Epistructure provides substantially improved correlation with the first two genotype-based PCs as compared with the alternatives (Figure 1).

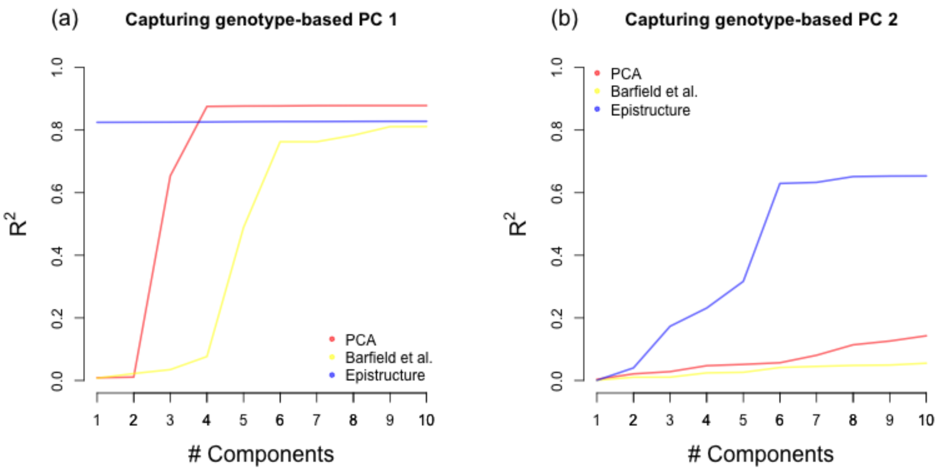

Figure 1: The fraction of variance explained in the first two genotype-based PCs of the GALA II data using several methods. Presented are linear predictors using increasing number of Epistructure PCs (in blue), methylation-based PCs (in red) and methylation-based PCs after feature selection based on a previous study [21] (in yellow) for capturing (a) the first genotype-based PC and (b) the second genotype-based PC.

Next, as an alternative measure of population structure, we used the ADMIXTURE software [5] to estimate for each individual ancestry proportions of the three ancestries known to compose the MX and PR populations: European (EU), Native-American (NA) and African (AF). In this case, the top two principal components of Epistructure capture very well both the Native American ancestry and the African Ancestry ($R^2 = 0.81$ and $R^2 = 0.56$ respectively), while the European ancestry was captured to a lesser extent ($R^2 = 0.32$, see Supplementary Figure S2).

We further tested whether ancestry information can be captured using Epistructure in case there is a weaker population structure in the data. We observed that the first two PCs of Epistructure could capture ancestry information well in both subpopulations of the GALA II data ($R^2 = 0.33$ in the PR group and $R^2 = 0.76$ in the MX group; Supplementary Figure S3). These results suggest that Epistructure can be used as an easy and efficient method for capturing ancestry information in methylation, even in data sets with relatively modest population structure.

Unsupervised Ancestry Inference from Methylation Data Epistructure is a supervised approach since it uses a reference data set in which both methylation and genotype data are available. In order to shed light on the extent to which ancestry is reflected by methylation data, we also explored unsupervised approaches for the inference of ancestry from methylation data. Consistent with a previous study of individuals from the same population [27], the first two genotype-based PCs of the GALA II data clustered the samples into two groups, generally corresponding to MX and PR subpopulations (Figure 2a). Since PCA has been shown to mirror ancestry very accurately in the case of genetic data [1], we first computed the top two methylation-based PCs while accounting for known technical factors as well as for age and sex, which are known to affect methylation genome-wide [28, 29]. Considering the population structure captured by the first two genotype-based PCs as the ground truth, the first two methylation-based PCs could not capture the population structure in the data (Figure 2b).

We then applied a few more sophisticated procedures, as follows. First, as before, we applied a feature selection step prior to calculating the methylation-based PCs according to a recent study, suggesting to consider only CpGs residing in close proximity to nearby SNPs in order to capture ancestry information in the first few PCs of the data [21]. We found that this procedure could not capture population structure in methylation data (Figure 2c). Next, since the first several PCs in methylation data coming from heterogeneous source such as blood are known to be dominated by cell type composition [22, 23], we adjusted the data for cell type composition using the ReFACTor software [23] and recalculated the first two PCs. This approach effectively reconstructed most of the separation determined by the genotype-based PCs (Figure 2d).

The above results show that methylation data indeed capture genotype data. Specifically, after accounting properly for known confounders, the top methylation-based PCs capture the genotype-based PCs. However, these results can potentially be driven by artifacts. Specifically, it is now acknowledged that many probes in the 450K methylation array contain single nucleotide polymorphisms (SNPs) in their target CpGs. Such polymorphic CpGs were shown to bias measured methylation levels as a function of the individual's genotypes, apparently due to changes in probe binding specificity [24]. Thus, our results so far might be biased by these probes. We therefore recalculated the first two methylation-based PCs after excluding 70,889 CpGs that are known to be polymorphic (see Methods). We found that the new methylation-based PCs could still capture well the first genotype-based PC ($R^2 = 0.77$ as opposed to $R^2 = 0.83$ before removing the polymorphic CpGs), accounting for the separation found using the first two genotype-based PCs (Figure 2e). In addition, we performed a more conservative analysis by repeating the last step after further excluding 167,738 probes containing at least one common SNP anywhere in the probe (i.e. not only in the target CpG; see Methods). We found that in this case as well the reconstruction using the top two methylation-based PCs provided almost the same separation determined by the genotype-based PCs (Figure 2f; $R^2 = 0.70$ with the first genotype-based PC).

We note that repeating the last two experiments while accounting for estimated cell proportions computed using a commonly applied reference-based method [30] as an alternative approach for correction of cell composition effects in methylation could not achieve the same results ($R^2 = 0.23$ and $R^2 = 0.14$ in the experiments without the polymorphic sites and in the experiment removing all probes with common SNPs, respectively).

We also compared the different approaches using the ancestry estimates of the ADMIXTURE software [5]. The results were consistent with our previous experiment - while the first two methylation-based PCs could not capture the ancestry estimates ($R^2 = 0.02$ with the EU fraction, $R^2 = 0.01$ with NA and $R^2 = 0.02$ with AF), we found the first two methylation-based PCs after adjusting for cell composition to capture the ancestry estimates well, even after excluding from the data all probes containing common SNPs ($R^2 = 0.28$ with the EU fraction, $R^2 = 0.69$ with NA and $R^2 = 0.47$ with AF; Supplementary Figure S4).

We further tested whether ancestry information can be captured in the same manner when applied to each of the two subpopulations in the data (MX and PR) separately. We found the methylation-based PCs to capture well only the first genotype-based PC of the Mexican group when not excluding probes containing common SNPs ($R^2 = 0.08$ for the PR cluster and $R^2 = 0.74$ for the MX cluster). After excluding the 167,738 probes containing at least one common SNP from the data the methylation-based PCs could not capture a substantial fraction of the first genotype-based PC in either clusters ($R^2 = 0.05$ for the PR cluster and $R^2 = 0.05$ also for the MX cluster). Thus, we conclude that under weak population structure unsupervised approaches currently do not mirror well the ancestry, and therefore supervised approaches such as Epistructure are needed.

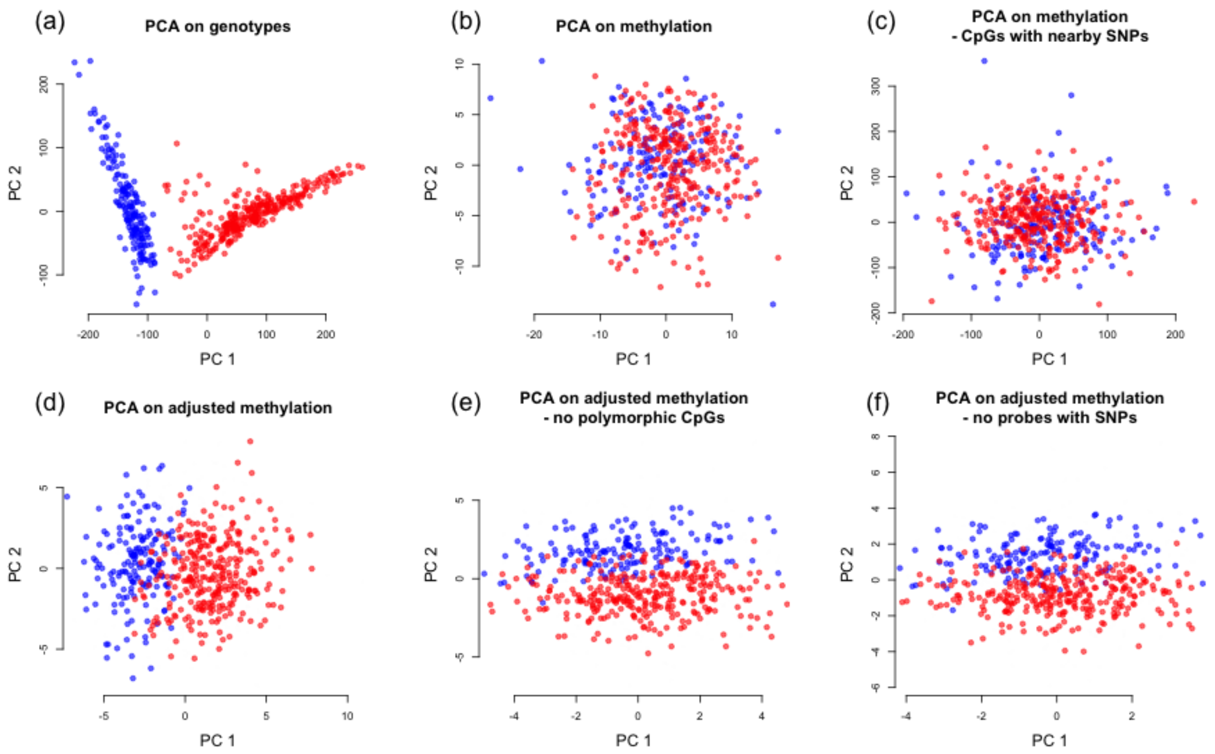

Figure 2: Capturing population structure in the GALA II data. (a) The first two PCs of the genotypes, considered as the ground truth, separates the samples into two subpopulations: Puerto-Ricans (in blue) and Mexicans (in red). (b) The first two PCs of the methylation levels (methylation PCs) cannot reconstruct the separation found in the genotypes. (c) Recalculating the first two PCs after applying a feature selection based on proximity of CpGs to nearby SNPs as was proposed by Barfield et al. [21] (d) The first two PCs of the methylation after adjusting the data for cell type composition (adjusted methylation PCs) can reconstruct most of the separation found in the genotypes either. (e) Using adjusted methylation PCs after excluding the 70,889 polymorphic sites from data. (f) Using adjusted methylation PCs after excluding the 167,738 probes containing at least one common SNP.

Validation using the CHAMACOS study data We further validated the effectiveness of Epistructure and the unsupervised approaches using data from the primarily Mexican-American CHAMACOS cohort [31, 32]. We used whole-blood methylation levels from nine years old participants (n=227) for which we had 106 ancestry informative markers (AIMs) [33], previously shown to approximate ancestry information well in another Hispanic admixed population [34].

We computed the first two PCs of the available AIMs (genotype-based PCs) in order to capture the ancestry information of the samples. Since the CHAMACOS cohort primarily consists of Mexican-American individuals, we observed no separation into distinct subpopulations in the first several genotype-based PCs. We then computed the first two methylation-based PCs, before and after adjusting the data for cell composition. In addition, we computed the first two Epistructure PCs of the data, and measured how much of the variance of the first genotype-based PC can be explained by each of the approaches. As shown in Figure 3, the first two methylation-based PCs could capture only a small portion of the first genotype-based PC ($R^2 = 0.04$ before adjusting for cell composition and $R^2 = 0.16$ after adjusting for cell composition), as opposed with the first two Epistructure PCs which could capture the first genotype-based PC substantially better ($R^2 = 0.38$). As in the GALA II data, applying feature selection based on proximity of CpGs to SNPs [21] could capture only a small portion of the ancestry information ($R^2 = 0.05$).

As before, we used the ADMIXTURE software [5] as an alternative measure of population structure. For each individual we estimated the ancestry proportions of the three ancestries known to compose Mexican individuals: European (EU), Native-American (NA) and African (AF). The first two Epistructure PCs were found to explain a large portion of the EU and NA fraction estimates ($R^2 = 0.46$ for EU and $R^2 = 0.6$ for NA ancestry), as opposed with the first two methylation-based PCs ($R^2 = 0.11$ for EU and $R^2 = 0.14$ for NA ancestry, after adjusting for cell type composition; Supplementary Figure S5). The estimates of African proportions, however, were not captured well by neither approach. This result was expected due to the low average proportion of African ancestry in Mexican samples (less than 10%) [35]. All the results are summarized in Table 1.

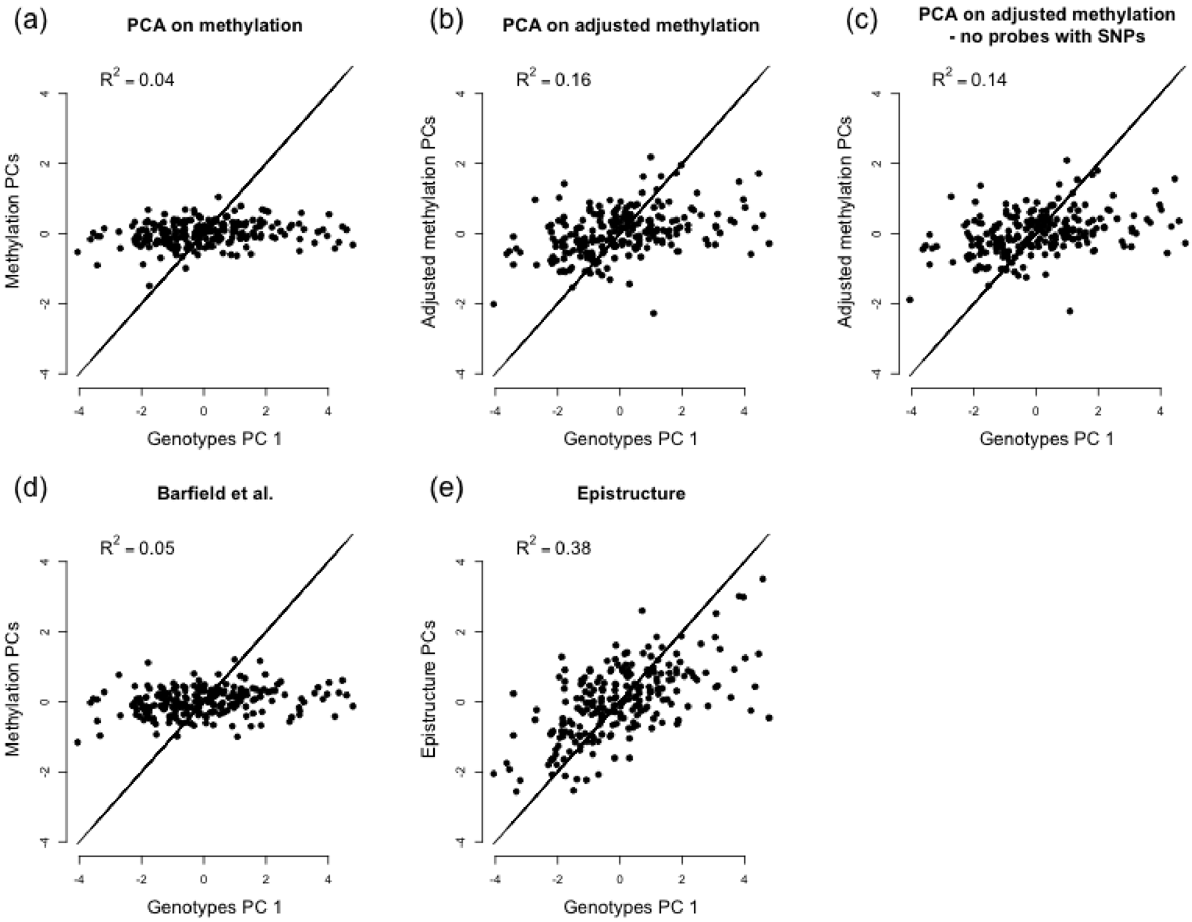

Figure 3: Capturing population structure in the CHAMACOS data. Presented are linear predictors of the first genotype-based PC using (a) the first two methylation PCs of the data, (b) the first two PCs after adjusting the data for cell type composition (adjusted methylation PCs), (c) the first two adjusted methylation PCs after excluding 167,738 probes containing SNPs from the data, (d) the first two PCs calculated based on a previously suggested feature selection [21] and (e) the using the first two Epistructure PCs.

Results summary						
data set	Meth PCs	Adj PCs	Adj PCs II	Barfield et al.	Epistructure	Measurement
GALA II	$R^2 = 0.01$	$R^2 = 0.83$	$R^2 = 0.70$	$R^2 = 0.02$	$R^2 = 0.83$	Genotypes PC 1
	$R^2 = 0.02$	$R^2 = 0.32$	$R^2 = 0.27$	$R^2 = 0.03$	$R^2 = 0.32$	EU fraction
	$R^2 = 0.01$	$R^2 = 0.81$	$R^2 = 0.69$	$R^2 = 0.03$	$R^2 = 0.81$	NA fraction
	$R^2 < 0.01$	$R^2 = 0.79$	$R^2 = 0.67$	$R^2 < 0.01$	$R^2 = 0.78$	AF fraction
CHAMACOS	$R^2 = 0.04$	$R^2 = 0.15$	$R^2 = 0.14$	$R^2 = 0.05$	$R^2 = 0.38$	Genotypes PC 1
	$R^2 = 0.03$	$R^2 = 0.11$	$R^2 = 0.08$	$R^2 = 0.01$	$R^2 = 0.46$	EU fraction
	$R^2 = 0.04$	$R^2 = 0.14$	$R^2 = 0.11$	$R^2 = 0.01$	$R^2 = 0.60$	NA fraction
	$R^2 = 0.01$	$R^2 = 0.01$	$R^2 = 0.01$	$R^2 = 0.01$	$R^2 = 0.06$	AF fraction

Table 1: Summary of the results in the GALA II data set and in the CHAMACOS data set. Linear squared correlations were measured between several measurements of ancestry information and linear predictors using the first two PCs of the data (Meth PCs), the first two PCs after adjusting the data for cell type composition (Adj PCs), the first two PCs after adjusting the data for cell type composition and excluding probes containing SNPs from the data (Adj PCs II), the first two PCs when considering only CpGs in close proximity to SNPs (Barfield et al.) and the first two Epistructure PCs.

3 Methods

Data and quality control The longitudinal KORA study (Cooperative health research in the Region of Augsburg) consists of independent population-based subjects from the general population living in the region of Augsburg, southern Germany [25]. Whole blood samples of the KORA F4 study were used ($n = 1,799$) as described elsewhere [36]. Briefly, DNA methylation levels were collected using the Infinium HumanMethylation450K BeadChip array (Illumina). Beta Mixture Quantile (BMIQ) [37] normalization was applied to the methylation levels using the R package wateRmelon, version 1.0.3 [38]. In total 431,360 probes were available for the analysis. As described elsewhere [39], genotyping was performed with the Affymetrix 6.0 SNP Array (534,174 SNP markers after quality control), with further imputation using HapMap2 as a reference panel. A total of 657,103 probes remained for the analysis.

We used whole-genome DNA methylation levels and genotyping data from the Genes-environments & Admixture in Latino Americans (GALA II) data set, a pediatric Latino population study. Details of genotyping data including quality control procedures for single nucleotide polymorphisms (SNPs) and individuals have been described elsewhere [35]. Briefly, participants were genotyped at 818,154 SNPs on the Axiom® Genome-Wide LAT 1, World Array 4 (Affymetrix, Santa Clara, CA) [40]. Non-autosomal SNPs and SNPs with missing data (> 0.05) and/or failing platform-specific SNP quality criteria ($n = 63,328$) were excluded as well as SNPs not in Hardy-Weinberg equilibrium ($n = 1,845$; $p < 10^{-6}$) within their respective populations (Puerto Rican, Mexican, and other Latino). Study participants were filtered based on 0.95 call rates and sex discrepancies, identity by descent and standard Affymetrix Axiom metrics. Finally, SNPs with low MAF (< 0.05; $n = 334,975$) were excluded. The total number of SNPs passing QC was 411,787.

Whole-blood methylation data for a subset of the GALA II participants ($n = 573$) are publicly available in the Gene Expression Omnibus (GEO) database (accession number GSE77716) and have been described elsewhere [13, 23]. Briefly, methylation levels were measured using the Infinium HumanMethylation450K BeadChip array and raw methylation data were processed using the R minfi package [41] and assessed for basic quality control metrics, including determination of poorly performing probes with insignificant detection P-values above background control probes and exclusion of probes on X and Y chromosomes. Finally, beta-normalized values of the the data were SWAN normalized [42], corrected for batch using COMBAT [43] and adjusted for age, gender and chip assignment information using linear regression. The number of participants with both methylation and genotyping data was 525. We further excluded 46 individuals collected in a separate batch since they were all Puerto-Ricans. A total of 479 individuals and 473,838 probes remained for the analysis.

In order to further evaluate and validate the performance of Epistructure we used data from the CHAMACOS longitudinal birth cohort study [31]. For this analysis, we had a subset of subjects that had Infinium HumanMethylation450K BeadChip array data available at 9 years of age. Briefly, samples were retained only if 95% of the sites assayed had insignificant detection P-value and samples demonstrating extremes level in the first two PCs of the data were removed. Probes where 95% of the samples had insignificant detection P-value (> 0.01; $n = 460$) as well as cross-reactive probes ($n = 29,233$) identified by Chen et al. [24] were dropped. At total, 227 samples and 455,590 probes remained for the analysis. Color channel bias, batch effects and difference in Infinium chemistry were minimized by application of All Sample Mean Normalization (ASMN) algorithm [44], followed by BMIQ normalization [37]. The data were adjusted for gender and technical batch information using linear regression.

In line with a previous study showing that a panel of small size is sufficient to approximate genetic ancestry in Latino populations well [45], 106 SNPs were collected and used as AIMs for estimating genetic ancestry of the CHAMACOS individuals [33]. The panel of AIMs was selected according to previously reported studies of Latino populations [12, 33, 46, 47]. Briefly, only SNPs with large differences in allele frequencies between ancestries were selected.

450K Human Methylation array This state of the art methodology allows for examination of $> 450,000$ CpG sites across the genome, representing 99% of RefSeq genes. Sites include promoters, gene bodies, and 96% of UCSC database CpG islands (dense concentrations of CpGs), many of which are known to be associated with transcriptional control [48–53]. This platform has been especially amenable to population studies because of its relative cost effectiveness and low sample requirements. Several studies have identified CpG sites differentially methylated by environmental exposures [54,55] (e.g. arsenic and tobacco smoke) and health outcomes including obesity [56], rheumatoid arthritis [57], and Crohns disease [58] demonstrating its utility in environmental and molecular epidemiology studies. The relative methylation (beta-normalized values) for each CpG site is calculated as the ratio of methylated-probe signal to total (methylated + unmethylated) fluorescent signal intensity. The Infinium pipeline is streamlined with excellent reproducibility [59].

Model and algorithm Previous studies reported a large number of correlations between DNA methylation and genetics, mainly cis-correlations between CpGs and nearby SNPs [9–12]. We therefore assume that cis-located SNPs can capture the genetic variability accounting for the methylation levels of a given CpGs. For a given CpG m we assume the following linear model:

$$m = \beta_0 + \sum_{s_i \in S_m} \beta_i s_i + \epsilon$$

where S_m is a group of w SNPs, cis-located with respect to m, β_i values are their corresponding effects on the methylation levels and ϵ represents an error term, assumed to be independent between different samples.

Given reference data of methylation levels and genotypes for the same individuals we fit the above linear model for each CpG. We regard the CpGs for which the model fits well as linear combinations of SNPs. We define the set of these genetically-informative CpGs as the reference list. Given methylation data for new individuals, we can estimate the population structure in the data by applying a standard PCA on the sites in the reference list. PCA is well-known to efficiently capture ancestry when applied to genotypes data [4], therefore applying PCA on CpGs that are linear combinations of SNPs is expected to capture population structure as well (see Appendix A for more details).

Given reference methylation and genotypes data, our suggested algorithm can be summarized as follows:

1. For each CpG m fit a linear model using w SNPs that are closest to m.

2. Define a reference list G of all the CpGs for which the linear model fits well. Evaluate model fit based on cross-validated squared linear correlation.

3. Given a new methylation data set, apply PCA on the sites defined by G and consider the first k principal components as the estimate of the population structure.

Note that creating a reference list, described in the first two steps of the algorithm, needs to be performed only once. Population structure can be then inferred in future data sets using this list of genetically-informative CpGs. In practice, an appropriate w may be relatively large (i.e. large number of predictors), while the sample size is typically limited. We therefore apply a regularized regression with ℓ_1 penalty, also known as LASSO regression [60]. For the same reason, we define a parameter p to limit the maximal number of predictors in each model. Furthermore, in order to avoid over-fitting of the model, we perform a k-fold cross-validation procedure for each CpG. The score of a CpG is defined as the median squared linear correlation of its predicted values with the real values across the k folds. Finally, a reference list of the CpGs is defined as the set of sites with highest scores.

In principle one could use the same approach taken here in order to create and use a reference list of CpGs which explain SNPs well rather than CpGs which are captured well by SNPs. However, modeling methylation levels as a function of SNPs is more natural with respect to the causality relations assumed between SNPs and methylation. Moreover, many methylation sites are known to be affected by several factors (e.g. age [28], gender [29] and smoking [61]), and therefore considering a group of methylation sites explaining a given SNP may introduce into the data more, potentially unknown, variance in addition to genetic variance. This potential problem is expected to be less severe in the opposite direction of modeling methylation using SNPs. In this case, methylation sites that are well explained by genetics are less likely to be highly explained by more factors. In particular, by including a CpG in the reference list only if at least 50% of its variance is explained by SNPs, we can ensure that the signal coming from genetics in each of the sites in the reference is the most dominant one.

Compiling a reference list from the KORA cohort The reference list of genetically-informative CpGs was created using the KORA cohort for which whole-blood methylation data as well as genotypes data were available for 1,799 European individuals. Following the algorithm described above, a score was computed for each CpG using k-fold cross-validation with $k = 10$ and using the parameters $w = 50$ and $p = 10$. A reference list was then compiled from CpGs with median correlation of $R^2 > 0.5$ in the cross-validated prediction procedure, resulting in a total of 4,913 CpGs, out of which 2,436 are polymorphic CpGs and additional 801 CpGs have at least on common SNP in their probe outside the CpG target. The number of these reference CpGs available in the GALA II data set and in the CHAMACOS data set were 4,912 and 4,450, respectively. Removing probes with polymorphic CpGs results in 2,476 and 2,229 CpGs, and further removing probes with common SNPs results in 1,676 and 1,554 CpGs for GALA II and CHAMACOS, respectively. We note that the Epistructure PCs in our experiments throughout the paper were computed using the entire reference list, including the polymorphic sites and CpGs with common SNPs in their probes, for capturing population structure better.

Detecting 450K probes containing SNPs Probes with a SNP in their CpG target (polymorphic CpGs) were shown to be biased with underlying genetic polymorphisms rather than capture methylation signals solely [24]. The authors reported a list of 70,889 such polymorphic CpGs in the 450k DNA methylation array, as well as a list of common SNPs residing in probes of the 450K array outside the CpG target (MAF > 0.01 according to at least one of the major continental groups in the 1000 Genome database [62]). The total number of probes containing SNPs reported is 167,738.

Estimating ancestry information Proportions of European, Native-American and African ancestries were estimated for each individual in both the GALA II and the CHAMACOS cohorts using the software ADMIXTURE [5] and the default reference data provided by the software. For the GALA II individuals we used the 411,787 SNPs remained after QC as an input, and for the CHAMACOS individuals we used the 106 available AIMs. The genotype based PCs were computed by applying PCA on the standardized values of the available genotypes in each data set. For the CHAMACOS data set, prior to computing PCA, we excluded sites with more than 5% missing values and completed the remaining missing values by assigning the mean. This resulted in a total of 99 SNPs.

Adjusting methylation levels for tissue heterogeneity Methylation levels of the GALA II and CHAMACOS data sets were adjusted for cell type composition using ReFACTor, a reference-free method for the correction of cell type heterogeneity in EWAS [23]. Each data set was adjusted for cell composition by regressing out the first six ReFACTor components computed using the default parameters and $K = 6$. For one of the experiments in the GALA II data we used an alternative approach for cell type composition correction - we obtained estimates of blood cell proportions using the default implementation available in the minfi package [41], defined and assembled for the 450K array [63] based on the approach suggested by Houseman et al [30] and a 450K reference data set [64].

Feature selection based on proximity to SNPs For evaluating our suggested method, we calculated alternative methylation-based PCs after applying a feature selection that was previously suggested as a method for capturing population structure [21]. Following the authors' recommendation, we considered a list of the CpGs residing within 50 bp from SNPs, as provided by the authors.

4 Discussion

We demonstrated that genome-wide DNA methylation data can capture population structure in admixed populations. In particular, we observed that in the presence of a relatively strong population structure (GALA II) the dominant genome-wide signal of ancestry information could be revealed once appropriately correcting for tissue heterogeneity. In contrast, we observed that in the presence of weaker population structure in the data (CHAMACOS) the genome-wide signal of ancestry methylation is only moderately reflected by the dominant axes of variance in the data after accounting for tissue heterogeneity.

Using a large data set for which both methylation levels and genotypes were available, we generated a reference list of genetically-informative CpGs and successfully used it to estimate ancestry information in new data sets by applying PCA on the reference sites. As we showed, by taking this approach, Epistructure was able to effectively isolate and capture ancestry information in the data. While we observed strong correlations of the Epistructure PCs with the genotype-based population structure estimates of the GALA II individuals, only moderate correlations were found in the CHAMACOS data set (though substantially better than unsupervised approaches, in which only negligible correlations to the true ancestry were found). These results can be explained by the fact that only 106 AIMs were available for us in the CHAMACOS for capturing ancestry information, as opposed with the dense genotype array information used in the GALA II analysis. Therefore, it is likely that our inference of population structure by methylation data is in fact more accurate than reflected in the experiments conducted on the CHAMACOS samples.

The reference-list of CpGs was generated using methylation states and genotypes collected from European individuals, therefore it may not be optimized for capturing ancestry information in non-European populations. However, since we successfully used this list for the inference of ancestry in the Latino GALA II and CHAMACOS individuals, we expect it to prove useful for some other non European populations as well. We further note that all of our experiments were conducted on methylation data collected from whole-blood specimens, and cell type specific correlations of methylation with genetics can potentially affect the selection of reference CpGs. However, to the best of our knowledge, there is currently no evidence in the literature for a dramatic impact on a reference list of thousands of CpGs due to interactions of cell type composition with genetics.

Finally, in line with previous works showing many associations of methylation with genetic variation and ancestry, our results further emphasize the importance of accounting for ancestry information in methylation studies of diverse populations, and therefore in the absence of genotyping data we suggest that Epistructure can be used in EWAS for adjusting methylation data for population structure.

References

[1] Novembre, J. *et al.* Genes mirror geography within europe. *Nature* **456**, 98–101 (2008).

[2] Price, A. L. *et al.* Discerning the ancestry of european americans in genetic association studies. *PLoS Genet* **4**, e236 (2008).

[3] Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

[4] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904–909 (2006).

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

[6] Yang, W.-Y., Novembre, J., Eskin, E. & Halperin, E. A model-based approach for analysis of spatial structure in genetic data. *Nature genetics* **44**, 725–731 (2012).

[7] Elhaik, E. *et al.* Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nature communications* **5** (2014).

[8] Pasaniuc, B. *et al.* Enhanced statistical tests for gwas in admixed populations: assessment using african americans from care and a breast cancer consortium. *PLoS Genet* **7**, e1001371 (2011).

[9] Gibbs, J. R. *et al.* Abundant quantitative trait loci exist for dna methylation and gene expression in human brain. *PLoS Genet* **6**, e1000952 (2010).

[10] Zhang, D. *et al.* Genetic control of individual differences in gene-specific methylation in human brain. *The American Journal of Human Genetics* **86**, 411–419 (2010).

[11] Bell, J. T. *et al.* Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome Biol* **12**, R10 (2011).

[12] Zhi, D. *et al.* Snps located at cpg sites modulate genome-epigenome interaction. *Epigenetics* **8**, 802–806 (2013).

[13] Galanter, J. M. *et al.* Methylation analysis reveals fundamental differences between ethnicity and genetic ancestry. *bioRxiv* 036822 (2016).

[14] Fraser, H. B., Lam, L. L., Neumann, S. M. & Kobor, M. S. Population-specificity of human dna methylation. *Genome Biol* **13**, R8 (2012).

[15] Moen, E. L. *et al.* Genome-wide variation of cytosine modifications between european and african populations and the implications for complex traits. *Genetics* **194**, 987–996 (2013).

[16] Smith, A. K. *et al.* Methylation quantitative trait loci (meqtls) are consistently detected across ancestry, developmental stage, and tissue type. *BMC genomics* **15**, 145 (2014).

[17] Boks, M. P. *et al.* The relationship of dna methylation with age, gender and genotype in twins and healthy controls. *PloS one* **4**, e6767 (2009).

[18] Kerkel, K. *et al.* Genomic surveys by methylation-sensitive snp analysis identify sequence-dependent allele-specific dna methylation. *Nature genetics* **40**, 904–908 (2008).

[19] Schalkwyk, L. C. *et al.* Allelic skewing of dna methylation is widespread across the genome. *The American Journal of Human Genetics* **86**, 196–212 (2010).

[20] Banovich, N. E. *et al.* Methylation qtls are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet* **10**, e1004663 (2014).

[21] Barfield, R. T. *et al.* Accounting for population stratification in dna methylation studies. *Genetic epidemiology* **38**, 231–241 (2014).

[22] Koestler, D. C. *et al.* Blood-based profiles of dna methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics* **8**, 816–826 (2013).

[23] Rahmani, E. *et al.* Sparse pca corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods* (2016).

[24] Chen, Y.-a. *et al.* Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).

[25] Wichmann, H., Gieger, C., Illig, T. *et al.* Kora-gen-resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* **67**, S26 (2005).

[26] Pino-Yanes, M. *et al.* Genetic ancestry influences asthma susceptibility and lung function among latinos. *Journal of Allergy and Clinical Immunology* **135**, 228–235 (2015).

[27] Galanter, J. M. *et al.* Cosmopolitan and ethnic-specific replication of genetic risk factors for asthma in 2 latino populations. *Journal of Allergy and Clinical Immunology* **128**, 37–43 (2011).

[28] Horvath, S. Dna methylation age of human tissues and cell types. *Genome biology* **14**, 1–20 (2013).

[29] Singmann, P. *et al.* Characterization of whole-genome autosomal differences of dna methylation between men and women. *Epigenetics & chromatin* **8**, 1 (2015).

[30] Houseman, E. A. *et al.* Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13**, 1 (2012).

[31] Eskenazi, B. *et al.* Chamacos, a longitudinal birth cohort study: lessons from the fields. *Journal of Children's Health* **1**, 3–27 (2003).

[32] Eskenazi, B. *et al.* Organophosphate pesticide exposure, pon1, and neurodevelopment in school-age children from the chamacos study. *Environmental research* **134**, 149–157 (2014).

[33] Huen, K., Harley, K., Beckman, K., Eskenazi, B. & Holland, N. Associations of pon1 and genetic ancestry with obesity in early childhood. *PloS one* **8**, e62565 (2013).

[34] Fejerman, L. *et al.* Genetic ancestry and risk of breast cancer among us latinas. *Cancer research* **68**, 9723–9728 (2008).

[35] 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. *Journal of Allergy and Clinical Immunology* **134**, 295–305 (2014).

[36] Pfeifferm, L. *et al.* Dna methylation of lipid-related genes affects blood lipid levels. *Circulation: Cardiovascular Genetics* CIRCGENETICS–114 (2015).

[37] Teschendorff, A. E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics* **29**, 189–196 (2013).

[38] Pidsley, R. *et al.* A data-driven approach to preprocessing illumina 450k methylation array data. *BMC genomics* **14**, 293 (2013).

[39] Kolz, M. *et al.* Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS Genet* **5**, e1000504 (2009).

[40] 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* **98**, 422–430 (2011).

[41] Aryee, M. J. *et al.* Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics* btu049 (2014).

[42] Maksimovic, J., Gordon, L., Oshlack, A. *et al.* Swan: Subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips. *Genome Biol* **13**, R44 (2012).

[43] Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8**, 118–127 (2007).

[44] Yousefi, P. *et al.* Considerations for normalization of dna methylation data by illumina 450k beadchip assay in population studies. *Epigenetics* **8**, 1141–1152 (2013).

[45] Tsai, H.-J. *et al.* Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations. *Human genetics* **118**, 424–433 (2005).

[46] Choudhry, S. *et al.* Genome-wide screen for asthma in puerto ricans: evidence for association with 5q23 region. *Human genetics* **123**, 455–468 (2008).

[47] Fejerman, L. *et al.* European ancestry is positively associated with breast cancer risk in mexican women. *Cancer Epidemiology Biomarkers & Prevention* **19**, 1074–1082 (2010).

[48] Bibikova, M. *et al.* High density dna methylation array with single cpg site resolution. *Genomics* **98**, 288–295 (2011).

[49] Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, cpg-rich dna. *Cell* **40**, 91–99 (1985).

[50] Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific cpg island shores. *Nature genetics* **41**, 178–186 (2009).

[51] Laurent, L. *et al.* Dynamic changes in the human methylome during differentiation. *Genome research* **20**, 320–331 (2010).

[52] Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences* **103**, 1412–1417 (2006).

[53] Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature biotechnology* **27**, 361–368 (2009).

[54] Joubert, B. R. *et al.* 450k epigenome-wide scan identifies differential dna methylation in newborns related to maternal smoking during pregnancy. *Environmental health perspectives* **120**, 1425 (2012).

[55] Seow, W. J. *et al.* Epigenome-wide dna methylation changes with development of arsenic-induced skin lesions in bangladesh: A case–control follow-up study. *Environmental and molecular mutagenesis* **55**, 449–456 (2014).

[56] Dick, K. J. *et al.* Dna methylation and body-mass index: a genome-wide analysis. *The Lancet* **383**, 1990–1998 (2014).

[57] Liu, Y. *et al.* Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology* **31**, 142–147 (2013).

[58] Adams, A. T. *et al.* Two-stage genome-wide methylation profiling in childhood-onset crohn's disease implicates epigenetic alterations at the vmp1/mir21 and hla loci. *Inflammatory bowel diseases* **20**, 1784–1793 (2014).

[59] Fan, J.-B. *et al.* [3] illumina universal bead arrays. *Methods in enzymology* **410**, 57–73 (2006).

[60] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288 (1996).

[61] Zeilinger, S. *et al.* Tobacco smoking leads to extensive genome-wide changes in dna methylation. *PloS one* **8**, e63812 (2013).

[62] Consortium, . G. P. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

[63] Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* **15**, R31 (2014).

[64] Reinius, L. E. *et al.* Differential dna methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS one* **7**, e41361 (2012).

Appendix A

PCA is well-known to efficiently capture ancestry information when applied to genotypes data [4]. In this appendix we show why applying PCA on CpGs that are linear combinations of SNPs is expected to capture population structure as well. The algorithm of Epistructure can be divided into two main steps. First, a reference list of genetically-informative methylation sites is complied from a group of CpGs, each found to be well approximated by its cis-located SNPs. Second, given a new methylation data set, the first several PCs of the data are calculated only from the same sites that were included in the reference list. The reason for applying PCA in the second part of the algorithm is motivated by the success of PCA to capture ancestry information in genotyping data. In the case of genotyping data coming from different populations, the first several PCs capture population structure by highlighting groups of individuals differing at the level of allele frequencies. Given an $n \times s$ centered genotyping data matrix G of s SNPs collected from n individuals, the generative model underlying PCA assumes:

$$\begin{aligned} G &= ZW + \Sigma \qquad (1) \\ \Sigma_j &\sim MVN\left(0, \tau^2 I_n\right) \end{aligned}$$

where Z is an $n \times k$ matrix representing k-dimensional latent structure of the ancestry information for each individual and W is a $k \times s$ matrix representing ancestry-specific differences in allele frequencies for each SNP. Σ is an $n \times s$ error term, typically assumed to have independent entries (that is, no relatedness between the n individuals and independence between the SNPs).

Any methylation site can be modeled as a linear function of SNPs and additional error term, and therefore the methylation level of this site in a given individual can be approximated to some extent using merely the individual's SNPs. Formally, given an $n \times m$ centered methylation data matrix O of m methylation sites coming from the same n individuals in G, we can describe O_j, the j-th column of O as:

$$\begin{aligned} O_j &= GB_j + E_j \qquad (2) \\ E_j &\sim MVN\left(0, \sigma_j^2\right) \end{aligned}$$

where B_j is an $s \times 1$ coefficients vector of the linear model and E_j is an $n \times 1$ error term. In particular, methylation site j that cannot be even partially explained by SNPs will have a corresponding B_j vector of only zeros. In the first step of the Epistructure algorithm we find a group of methylation sites which can be well explained by their cis-located SNPs. Restricting the data matrix O to be consisted only of such increases the signal-to-noise ratio in the data.

Plugging (1) into (2) we get

$$\begin{aligned} O_j &= (ZW + \Sigma)B_j + E_j \\ &= ZWBj + \Sigma B_j + E_j \end{aligned}$$

where ΣB_j and E_j are normally distributed as before. This model can be equivalently described as follows:

$$O_j \sim MVN\left(ZWB_j, \left(B_j^t B_j \tau^2 + \sigma_j^2\right) I_n\right)$$

Under this formulation there is a dependency between every two methylation sites. However, based on previous reports showing clear predominance of associations between CpGs and cis-located SNPs over trans-located SNPs [10–12], we assume that only cis-located SNPs are informative for explaining a given methylation site. As a result, B is expected to be very sparse with values concentrated around the diagonal, assuming the SNPs and CpGs are ordered by physical position. In particular, every two distant methylation sites are independent. In our case the matrix B was estimated from the KORA data for which both genotyping and methylation levels were available. We observed that the vast majority of the rows in the estimated matrix are sparse and only rarely have more than one non-zero entry (Supplementary Figure S6). The main reason for this is the fact that we consider only a sparse set of methylation sites from the genome, resulting from the first step of the algorithm in which only sites that can be well approximated by SNPs are selected. Therefore, we neglect the theoretical dependency between close sites and assume no dependency between any of the columns in B. Now, the model can be summarized as:

$$O_j \sim MVN(Z\tilde{W}_j, \psi_j^2 I_n) \tag{3}$$

where $\tilde{W}_j = WB_j$ and we are interested in extracting Z, the latent ancestry information structure of the individuals in the data. The maximum likelihood solution to the model in (3) is given by factor analysis, and the first k factors can be used as estimates of the latent population structure Z. In practice, factor analysis iteratively scales each site and the first iteration is equivalent to PCA after standardization of each of the sites. Empirically, applying more than one iteration did not improve the performance, therefore, in the second step of the Epistructure algorithm we suggest to perform a standardized PCA and to consider the first k PCs as the estimate of the population structure.