Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos

Matthew P. Conomos,[1,14,*] Cecelia A. Laurie,[1,14] Adrienne M. Stilp,[1,14] Stephanie M. Gogarten,[1,14] Caitlin P. McHugh,[1] Sarah C. Nelson,[1] Tamar Sofer,[1] Lindsay Fernández-Rhodes,[2] Anne E. Justice,[2] Mariaelisa Graff,[2] Kristin L. Young,[2] Amanda A. Seyerle,[2] Christy L. Avery,[2] Kent D. Taylor,[3] Jerome I. Rotter,[3] Gregory A. Talavera,[4] Martha L. Daviglus,[5] Sylvia Wassertheil-Smoller,[6] Neil Schneiderman,[7] Gerardo Heiss,[2] Robert C. Kaplan,[6] Nora Franceschini,[2] Alex P. Reiner,[8] John R. Shaffer,[9] R. Graham Barr,[10] Kathleen F. Kerr,[1] Sharon R. Browning,[1] Brian L. Browning,[11] Bruce S. Weir,[1] M. Larissa Avilés-Santa,[12] George J. Papanicolaou,[12] Thomas Lumley,[13] Adam A. Szpiro,[1] Kari E. North,[2] Ken Rice,[1] Timothy A. Thornton,[1] and Cathy C. Laurie[1,*]

US Hispanic/Latino individuals are diverse in genetic ancestry, culture, and environmental exposures. Here, we characterized and controlled for this diversity in genome-wide association studies (GWASs) for the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). We simultaneously estimated population-structure principal components (PCs) robust to familial relatedness and pairwise kinship coefficients (KCs) robust to population structure, admixture, and Hardy-Weinberg departures. The PCs revealed substantial genetic differentiation within and among six self-identified background groups (Cuban, Dominican, Puerto Rican, Mexican, and Central and South American). To control for variation among groups, we developed a multi-dimensional clustering method to define a "genetic-analysis group" variable that retains many properties of self-identified background while achieving substantially greater genetic homogeneity within groups and including participants with non-specific self-identification. In GWASs of 22 biomedical traits, we used a linear mixed model (LMM) including pairwise empirical KCs to account for familial relatedness, PCs for ancestry, and genetic-analysis groups for additional group-associated effects. Including the genetic-analysis group as a covariate accounted for significant trait variation in 8 of 22 traits, even after we fit 20 PCs. Additionally, genetic-analysis groups had significant heterogeneity of residual variance for 20 of 22 traits, and modeling this heteroscedasticity within the LMM reduced genomic inflation for 19 traits. Furthermore, fitting an LMM that utilized a genetic-analysis group rather than a self-identified background group achieved higher power to detect previously reported associations. We expect that the methods applied here will be useful in other studies with multiple ethnic groups, admixture, and relatedness.

Introduction

Individuals who identify as Hispanic and/or Latino (Hispanic/Latino) in the US are diverse in culture, environmental exposures, nativity, socioeconomic status, and disease burden.[1] They are also genetically diverse as a result of widespread geographic origins within the Americas, as well as variation in patterns of immigration from other continents. Many Hispanic/Latino individuals have admixed genomes consisting of three predominant continental ancestries: indigenous American (primarily of South and Central America, Mexico, and the Caribbean islands, hereafter referred to as "Amerindian"), European as a result of colonization, and African as a result of slave transport from West Africa. The proportion of genetic ancestry derived from each of these three continental regions varies substantially among and within ethnic groups from different countries in Latin America and in US Hispanic/Latino populations.[2–4] Furthermore, Amerindian genomic segments have additional genetic heterogeneity associated with geographic locations in Latin America.[2,4–6] To protect against confounding, it is important to take this complex admixture and genetic diversity into account in association studies that aim to identify the genetic basis of phenotypic variation.[7,8] In addition, accounting for the cultural and environmental diversity of US Hispanic/Latino groups might also improve the precision of detecting genetic risk factors.

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL) is a community-based cohort study

[1]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; [2]Department of Epidemiology, University of North Carolina, Chapel Hill, NC 27514, USA; [3]Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute and Department of Pediatrics, Harbor-UCLA Medical Center, Torrance, CA 90502, USA; [4]Graduate School of Public Health, San Diego State University, San Diego, CA 92182 USA; [5]Institute for Minority Health, University of Illinois at Chicago, Chicago, IL 60612, USA; [6]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA; [7]Department of Psychology and Behavioral Medicine, University of Miami, Miami, FL 33124, USA; [8]Department of Epidemiology, University of Washington, Seattle, WA 98195, USA; [9]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA; [10]Departments of Medicine and Epidemiology, Columbia University Medical Center, New York, NY 10032, USA; [11]Department of Medicine, University of Washington, Seattle, WA 98077, USA; [12]Division of Cardiovascular Sciences, NHLBI, NIH, Bethesda, MD 20892, USA; [13]Department of Statistics, University of Auckland, Auckland 1010, New Zealand
[14]These authors contributed equally to this work
*Correspondence: mconomos@uw.edu (M.P.C.), cclaurie@uw.edu (C.C.L.)
http://dx.doi.org/10.1016/j.ajhg.2015.12.001.

of self-identified Hispanic/Latino individuals from four US metropolitan areas.[9] General goals of this study are to identify risk and protective factors for various medical conditions, including cardiovascular disease, diabetes, pulmonary disease, and sleep disorders. As the largest Hispanic/Latino cohort study to date, it includes a baseline total of 16,415 participants, among whom 12,803 consented to genetic studies and were successfully genotyped on a genome-wide SNP array.

HCHS/SOL includes participants who self-identified as being of Cuban, Dominican, Puerto Rican, Mexican, Central American, or South American background. These specific group names were given as possible responses to a question regarding "ascendencia hispana o latina," which translates to Hispanic or Latino heritage or background. Participants could also respond as having "more than one" or "other" background. In this paper, we refer to these self-identified categories as "background groups." Previous studies of the HCHS/SOL cohort have shown that the prevalence of asthma[10] and cardiovascular risk factors,[11] including diet[12] and smoking,[13] differs substantially among background groups. Among all ethnic groups in the US, the highest prevalence of asthma occurs in individuals of Puerto Rican background, whereas the lowest prevalence occurs in those of Mexican background, and there is evidence that genetic risk factors for asthma might have group-specific effects.[14]

A categorical variable for ethnic group, such as Hispanic/Latino background, might have several useful applications in genome-wide association studies (GWASs). (1) Including it as a covariate in association tests can increase precision by controlling for complex cultural and environmental differences that might otherwise require numerous relevant measurements, which could be unknown or unavailable. (2) Including it as a covariate can also help to control confounding by ancestry if the ethnic group captures genetic differences not represented by the genetically inferred principal components (PCs) standardly used in accounting for population structure in association tests. (3) It can serve to aid detection of group-specific genetic effects in either stratified analyses or combined analyses with gene-by-group interactions. (4) It can also be used in accounting for trait-variance heterogeneity among groups, which could reduce genomic inflation or other artifacts. In HCHS/SOL and other studies, self-identified ethnic groups could be used for these purposes, but this can be problematic for the following reasons. First, self-identified groups can have genetic outliers, which could have undue influence on the PCs used for ancestry adjustment in group-stratified analyses. Such outliers could affect many of the PCs because of their orthogonality property, potentially hindering the detection of and adjustment for other important population structure.[15,16] Second, some individuals might not have self-reported membership with a specific ethnic group (e.g., 428 in HCHS/SOL), even though they are genetically similar to members of one or more groups in the study, leading to decreased sample size and a loss of power as a result of missing data.

Here, we address the potential problems associated with self-identified ethnic groups by using a multi-dimensional clustering method to construct HCHS/SOL "genetic-analysis groups" with the same categorical values as the self-identified background groups. These genetic-analysis groups are similar to self-identified background groups in that they share cultural and environmental characteristics, but they are more genetically homogeneous and include all study participants. For each background group, the minimum covariant determinant (MCD) method of Rousseeuw[17] is used for defining in PC space a hyper-ellipsoid that contains the majority of points representing individuals in that group. Once the hyper-ellipsoids for each group have been defined, all individuals, including those with missing or non-specific background-group membership, are assigned to a genetic-analysis group according to their distance to each hyper-ellipsoid in PC space in a manner that aims to preserve concordance with their self-identified background-group membership when reasonable. This clustering approach is versatile in that the degree of concordance between the self-identified background group and the genetic-analysis group is easily adjustable, allowing for a balance between maintaining self-identification and ensuring genetic similarity within groups.

Not only does genetic-ancestry variation (i.e., population structure) result from continental and sub-continental geographic differentiation, but the HCHS/SOL sample also has further genetic structure as a result of the presence of numerous familial relatives, as expected from the community- and household-based sampling design.[18] In this study, pedigree information was not collected from participants, so relatedness was inferred from the genotypic data. Identifying relatedness in the presence of population structure and vice versa is difficult in admixed populations, but robust approaches for identifying each of these structures in the presence of the other have recently been developed. PC-AiR[19] estimates PCs that reflect more distant ancestry and are robust to the presence of recent pedigree structure. PC-Relate[20] provides accurate estimates of recent genetic-relatedness measures, such as kinship coefficients (KCs), in structured populations, including those with ancestry admixture and departures from Hardy-Weinberg equilibrium. In this study, we estimated PCs and KCs simultaneously by using an iterative procedure combining both PC-AiR and PC-Relate. We used these estimates to characterize genetic diversity in the HCHS/SOL and to control for genomic inflation due to confounding in association studies.

Most previous GWASs using Hispanic/Latino samples have either (1) removed relatives from association tests[21–25] or (2) implemented a linear mixed model (LMM) that used a priori pedigree information to account for relatedness and PCs to account for population structure.[26–32] These approaches are not practical in

HCHS/SOL because removing inferred relatives would entail a decrease of approximately 20% in sample size and because pedigree information is not available. In the absence of pedigree information, an empirical genetic-relationship matrix (GRM)[33,34] estimated from SNP genotypes has often been used instead of PCs and a pedigree-based KC matrix in LMM analyses of primarily European populations.[8,35–38] The empirical GRM represents genetic similarity because of a combination of shared ancestry and relatedness, but it might not adequately account for population stratification at all SNPs of interest, whereas additional adjustment for fixed PC effects might do so.[8,19,39] However, double fitting the same structure as both fixed and random effects can lead to over-correction and a loss of power.[38,40,41] Therefore, in HCHS/SOL, we used a LMM that partitioned the overall genetic structure of samples into two separate components. We included PCs as fixed effects to adjust for population stratification due to ancestry variation, and we used a matrix of pairwise empirical KC estimates, calculated conditionally on the PCs, to account for familial relatedness.[41] Simulations that further support this approach will be presented elsewhere.

This paper describes genetic diversity in HCHS/SOL and how to account for it in genetic association studies, particularly with respect to controlling confounding due to population structure, admixture, and familial relatedness. We demonstrate that association testing with a LMM using robust PC and KC estimates effectively controls genomic inflation in GWASs of 22 biomedical traits in HCHS/SOL. We also demonstrate the utility of genetic-analysis groups in GWAS applications, including analyses stratified by group without substantial loss of sample size or PC outliers, adjustment for possible non-genetic-group effects, and accounting for variance heterogeneity among groups. We expect that the methods applied here will be useful in other studies with multiple ethnicities, admixture, and relatedness.

Subjects and Methods

Subjects and Study Design

The HCHS/SOL sample survey design was described previously.[18] It consisted of a two-stage probability sample of households at each of four recruitment centers: Chicago, Miami, the Bronx, and San Diego. Census block groups were selected in defined communities near each center, and households were sampled within block groups. Households with Hispanic/Latino surnames and individuals were oversampled as a means of increasing representation of the Hispanic/Latino target population; likewise, households with residents over 45 years of age were oversampled so a more uniform age distribution could be achieved. Sampling weights were calculated for each individual to reflect the probability of sampling. Baseline examination methods were described by Sorlie et al.[9] The traits analyzed here were measured at baseline. Statistics reported in this paper are based only on the 12,803 successfully genotyped participants. The HCHS/SOL study was approved by institutional review boards at participating institutions, and written informed consent was obtained from all participants.

Genotyping and Quality Control

DNA extracted from blood was genotyped on an Illumina custom array, SOL HCHS Custom 15041502 B3, consisting of the Illumina Omni 2.5M array (HumanOmni2.5-8v1-1) and ~150,000 custom SNPs selected to include ancestry-informative markers, variants characteristic of Amerindian populations, previously identified GWAS hits, and other candidate-gene polymorphisms (G.J.P., K.D.T., and J.I.R., unpublished data). Samples were checked for annotated sex or genetically determined sex, gross chromosomal anomalies,[42] unexpected duplicates, missing call rates, contamination, and batch effects.[43] Portions of the genome with large chromosomal anomalies were filtered out in 71 samples. A total of 12,803 samples passed quality control with a missing call rate < 1%. Quality metrics used to filter SNPs for the imputation basis and association testing included missing call rate (>2%), Mendelian errors (>3 in 1,343 trios or duos), duplicate-sample discordance (>2 in 291 sample pairs), and deviation from Hardy-Weinberg equilibrium ($p < 10^{-5}$ in a meta-analysis of nine groups within which individuals had both parents from the same country of origin). SNPs were regarded as "informative" if they had no positional duplicate on the array and were polymorphic in the sample. A total of 2,232,944 SNPs passed quality metrics and were informative.

In addition to the HCHS/SOL study participants, 401 control individuals were genotyped simultaneously on the same platform. These comprise Amerindian samples from Mexico and South America (NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research) and samples from five HapMap populations:[44] Utah residents with ancestry from northern and western Europe from the CEPH collection (CEU); Han Chinese in Beijing, China (CHB); Japanese in Tokyo, Japan (JPT); Mexican Ancestry in Los Angeles, California (MXL); and Yoruba in Ibadan, Nigeria (YRI).

Genotype Imputation

Genotype imputation was performed with the 1000 Genomes Project phase 1 reference panel.[45] The 12,803 samples were imputed together with genotyped SNPs that passed quality filters and represented unique positions on the autosomes and non-pseudo-autosomal parts of the X chromosome. SHAPEIT2 (v.2.r644)[46] was used for pre-phasing, followed by imputation with IMPUTE2[47,48] (v.2.3.0) software. Variants with at least two copies of the minor allele and present in any of the four 1000 Genomes continental panels were imputed (a total of 25,568,744 imputed variants). Quality control included examination of the "info score," masked SNP r^2, and "oevar," the ratio of observed variance of imputed dosages to the expected binomial variance.[49] Results of the association analysis were filtered according to an "effective minor allele count," $N_{eff} = 2p(1 - p)Nv$, where p is the estimated minor allele frequency, N is the sample size, and v is "oevar." Expected allelic dosages were used for imputed SNPs in association studies.

Continental-Ancestry Proportions

Continental-ancestry proportions were estimated with a model-based analysis using ADMIXTURE software[50] under the assumption of three or four ancestral populations (West African, European, and Amerindian for k = 3, plus East Asian for k = 4).

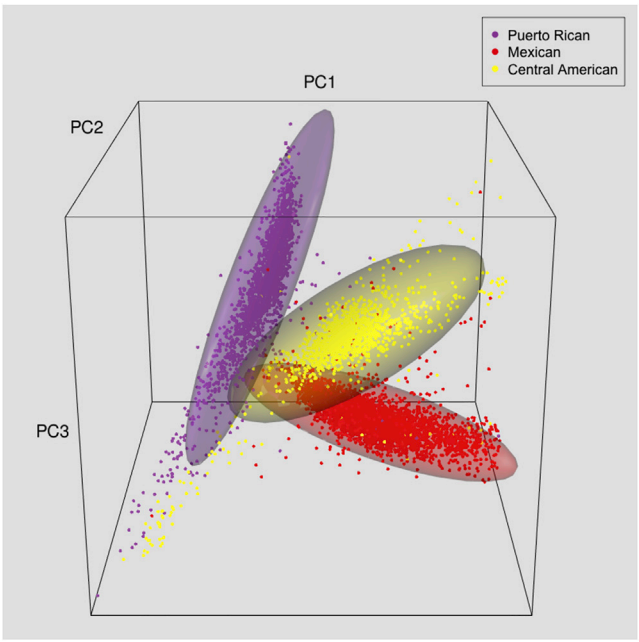

Figure 1. Hyper-ellipsoids Capturing Most of the Genetic Variation within Each Self-Identified Background Group
To illustrate the definition of genetic-analysis groups, here we use only PCs 1–3, although the full definition used PCs 1–5. These PCs are from PCA of all individuals except those with high East Asian ancestry. 3D hyper-ellipsoids, containing the highest density of points (defined by MCD), are shown for self-identified Mexican, Central American, and Puerto Rican background groups (Cuban, Dominican, and South American background groups are omitted for visual clarity).

Initially, an unsupervised analysis was performed on potential reference-population samples. These references were from the Human Genome Diversity Project (HGDP)[51] and the controls genotyped with the HCHS/SOL samples. These two datasets had 440,908 genotyped SNPs in common (including 431,143 autosomal and 9,694 X chromosome SNPs). KCs were estimated with KING-robust,[52] and the algorithm from PC-AiR was used to select an ancestrally diverse and mutually "unrelated" set of 574 samples at a second-degree threshold (i.e., KC < 0.088 for all pairs). An unsupervised ADMIXTURE analysis was performed on this "unrelated" set with $k = 4$ ancestral populations and 92,992 autosomal SNPs selected by linkage-disequilibrium (LD) pruning, described below. Relatively homogeneous reference samples were identified as those with >90% ancestry estimated from one group; in this way, we identified 101 African, 49 Amerindian, 176 European, and 161 East Asian reference samples. These samples were used as the reference samples for a supervised ADMIXTURE analysis performed on 10,642 mutually "unrelated" HCHS/SOL subjects, also identified with the PC-AiR algorithm with a third-degree relatedness threshold (i.e., KC < 0.044 for all pairs). With $k = 4$ ancestral populations, the East Asian component was very low in nearly all samples, so results presented here are for $k = 3$ (i.e., excluding East Asian reference samples). We also performed an X chromosome analysis with $k = 3$, the same set of reference samples, and a set of 2,233 SNPs selected with LD pruning and excluding any SNPs that fell into pseudo-autosomal regions. The male genotypes (allelic dosages) were coded as 0 or 2 for all X chromosome SNPs.

Inferring Population Structure and Estimating KCs

We used an iterative procedure to simultaneously estimate PCs reflecting population structure and KCs measuring familial relatedness. For each iteration of this procedure, we selected directly genotyped autosomal SNPs by LD pruning such that all pairs had $r^2 < 0.1$ in a sliding 10 Mb window in a set of individuals estimated to be more distant than third-degree relatives. SNPs with a missing call rate > 0.05, with a MAF < 0.05, or within lactase (*LCT*), human leukocyte antigen (HLA) genes, or polymorphic-inversion regions on chromosomes 8 and 17 were excluded from the initial pool. For each iteration, this selection procedure resulted in approximately 150,000 SNPs.

We applied the iterative procedure according to the following steps. We (1) obtained initial KC estimates with KING-robust, which is robust to discrete population structure but provides biased estimates in admixed samples;[20,53] (2) used PC-AiR to perform a PC analysis (PCA) that was robust to the relatedness identified in the sample; and (3) found updated KC estimates with PC-Relate, which uses PCs to account for population structure and provide accurate estimates, even in the presence of admixture. To protect against the potential misidentification of relatives from step (1) above, we followed this method with a second iteration of (4) PC-AiR and (5) PC-Relate analyses.

At this point, we determined that the sixth PC mainly separated 19 individuals with high levels of East Asian ancestry (39%–100%) from the remaining subjects (Figure S1A). Because these individuals are so different genetically from the rest of the sample, we removed them and repeated (6) PC-AiR and (7) PC-Relate analyses. We determined that this was the final iteration of PC-AiR and PC-Relate because (a) the pairwise correlations between the top five PCs from steps (4) and (6) were all >0.996, (b) the correlation between the estimated KCs from steps (5) and (7) was >0.999 for all pairs of individuals inferred to be fourth-degree relatives or closer (i.e., KC > 0.022) in either iteration, and (c) there were no extreme PC outliers in the final sample set (Figure S1B). In the rest of this paper, the full sample set of 12,784 refers to all HCHS/SOL genotyped participants excluding the 19 outliers with high proportions of East Asian ancestry, and the PCs and KCs that were used in all reported analyses are from the final iteration of this procedure.

Definition of the Genetic-Analysis Group

We constructed the categorical variable "genetic-analysis group," defined as having the same six values as the self-identified background groups (i.e., Cuban, Dominican, Puerto Rican, Mexican, Central American, and South American), by using the MCD method of Rousseeuw.[17] For each category separately, we first defined a hyper-ellipsoid in the five-dimensional space of the first five genetic PCs by following two steps. (1) From the set of points representing unrelated individuals whose self-identified background matched the category, we found the subset consisting of 99% of the points for which the covariance matrix had the minimum determinant (MCD points). Heuristically, the subset of points selected in this way represents the volume of highest point density. (2) We then defined a 99% tolerance hyper-ellipsoid for which (a) the center was the mean of the subset of points and (b) the boundary was set to include points within a fixed number of SDs of the center (i.e., the Mahalanobis distance) in any direction in PC space (see Figure 1). The fixed number of SDs was chosen so that, under multivariate normality, 99% of the data would be included in the ellipsoid. When multivariate normality

did not hold, the percentage might have differed from 99%, but such deviation does not invalidate the procedure because the hyper-ellipsoid was used as a classification tool rather than for hypothesis testing. The hyper-ellipsoids for the six groups actually contained 90%–96% of the unrelated individuals with a given self-identified background.

The ellipsoid-defining procedure was subsequently modified for the self-identified Cuban background group because it had a highly skewed distribution of values of PC 2, such that the skewed tail corresponded to individuals with more African and less European ancestry (Figure S2). This tail substantially overlapped the self-identified Dominican distribution. To include the self-identified Cubans in the tail of the distribution within the Cuban (rather than the Dominican) genetic-analysis group, we replaced the original hyper-ellipsoid with two hyper-ellipsoids defined sequentially as follows. First, we used 95% of all self-identified Cubans to construct a 95% tolerance hyper-ellipsoid; this hyper-ellipsoid represents the majority of Cubans. Second, we used 50% of self-identified Cubans with points not in the first Cuban hyper-ellipsoid to construct a 99% tolerance hyper-ellipsoid, which captured most of the self-identified Cubans in the tail of the distribution.

Assignments to genetic-analysis groups were based on the hyper-ellipsoids. Each individual was assigned to the genetic-analysis group with the same category as his or her self-identified background when his or her point in PC space was within the hyper-ellipsoid for that category. All other individuals were assigned to the "closest" (defined as the minimum Mahalanobis distance from his or her point in PC space to the center of a hyper-ellipsoid) genetic-analysis group. Individuals assigned to the two Cuban hyper-ellipsoids were combined into a single Cuban genetic-analysis group.

A group of 37 individuals, consisting mainly of Central Americans with unusually high proportions of African ancestry, formed a small PC cluster that was well separated from the main clusters of individuals with Central American, Mexican, and South American background. These 37 individuals were excluded from the definition of genetic-analysis groups and from all analyses that involved this variable.

Genetic Association Testing

We used LMMs to test for genetic associations with quantitative traits. Unless specified otherwise, for each trait analyzed, we included the top five PCs as fixed effects to account for population stratification. To protect against potential bias in effect-size estimates due to the survey sampling procedure implemented in HCHS/SOL, we also included individual sampling weights in the model as fixed effects, as advocated by Pfeffermann.[54] Additional fixed effects included sex, age, recruitment center, and other trait-specific covariates. Polygenic effects due to recent genetic relatedness (represented by a matrix of the pairwise empirical KCs estimated from autosomal SNPs with PC-Relate), household membership, and membership in a census block group were included as random effects. For each trait, we used the Average Information Restricted Maximum Likelihood (AI-REML),[55] applied to the null model (i.e., no genotype effect), to estimate variance components for each of these random effects, as well as the residual variance component. In analyses that allowed for heteroscedasticity in the error variances, we fit the model by using six separate residual variance components, one for each genetic analysis or self-identified group. Using the overall trait-specific covariance structure estimated from the null model,[35] we estimated individual SNP effects and SEs with a generalized least-squares estimator. Wald tests provided p values for tests of association.

Software

All analyses were performed with the R statistical computing environment and Bioconductor (Web Resources), including the following packages: SNPRelate[56] (v.1.1.3) for KING-robust and PCA, GWASTools[57] (v.1.14.0) for genotype quality control, GENESIS[41] for PC-AiR and PC-Relate (v.1.1.3), robustbase (v.0.92-3) for MCD,[58] and ggplot2[59] (v.1.0.1), ggmap[60] (v.2.5.2), GGally (v.0.5.0),[61] maptools[62] (v.0.8-34), rgdal[63] (v.0.9-2), and rgl[64] (v.0.95.1201) for graphics. The code used for fitting LMMs is available upon request.

Results

Self-Identified Ancestry

Nearly all HCHS/SOL participants self-identified as Hispanic/Latino, and most also self-identified as one of six different background groups. Among the 12,803 genotyped individuals, 16.1% identified their background as Cuban, 9.4% as Dominican, 17.1% as Puerto Rican, 37.1% as Mexican, 10.6% as Central American, 6.6% as South American, and 3.1% as multiple, other, or missing background. The proportions of these background groups varied greatly among the four recruitment centers (Figure S3). For example, 97% of individuals with a Cuban background were sampled from Miami, 93% of Dominican individuals from the Bronx, 67% of Puerto Rican individuals from the Bronx, and 59% of Mexican individuals from San Diego. A minority (18%) were born in the US (excluding Puerto Rico). The HCHS/SOL participants also provided the countries of origin of their parents and grandparents (Figure S4). Among the genotyped participants, 90.3% of grandparents were reported to be from Latin America, 4.5% from Europe, 1.7% from the US, and 3.6% from other countries such as China, Japan, and India. The European grandparents are predominantly from Spain (77% of 2,281). These observations illustrate the diverse origins of the US Hispanic/Latino populations sampled by HCHS/SOL.

Continental-Ancestry Admixture

Continental-ancestry proportions in HCHS/SOL (estimated under the assumption of three ancestral populations) vary substantially both within and among self-identified background groups (Figure 2). Participants who self-identified with the mainland backgrounds (Mexican, Central, and South American) have more Amerindian and less African ancestry than those from the Dominican and Puerto Rican groups, whereas those of Cuban background have more European ancestry than the other groups. These patterns are consistent with previous reports for Hispanic/Latino samples from the US[2,65] and in small samples from Latin American countries.[3,4] Nevertheless, we note that the HCHS/SOL participants were sampled from four urban areas in the US, so their ancestry might not be representative of the

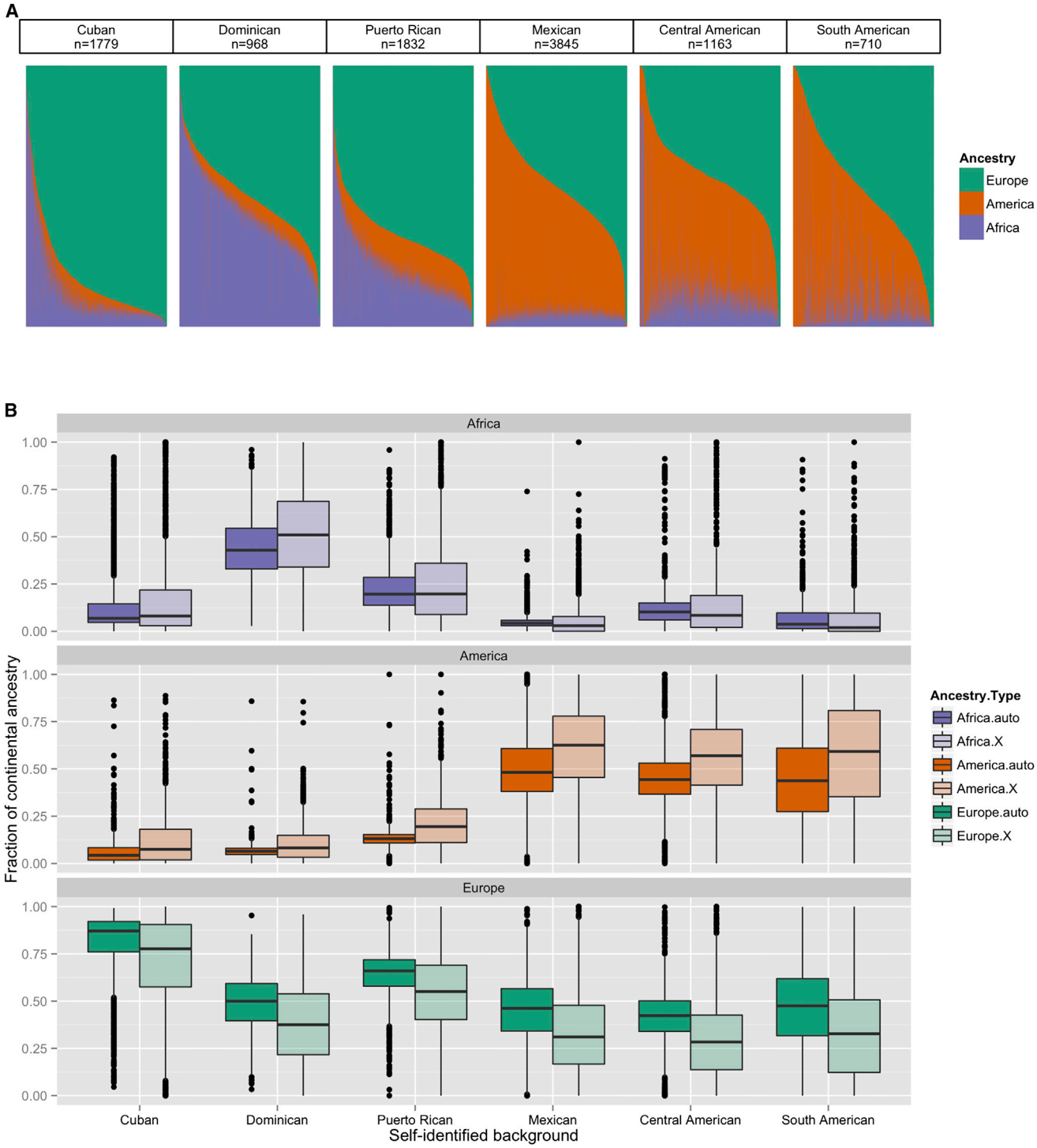

Figure 2. Continental-Ancestry Proportions for the Autosomes and X Chromosome

(A) Estimates of continental-ancestry proportions on the autosomes for an unrelated set of HCHS/SOL individuals are grouped by self-identified background, and the number of unrelated individuals is shown for each group. Each vertical bar represents a single individual, and the three color-coded segments represent the three ancestry fractions.

(B) Boxplots show distributions of estimates of continental-ancestry proportions within each self-identified background group for the autosomes (from A) and the X chromosome. The same individuals (excluding 15 individuals with X chromosome anomalies) were used for calculating X chromosome estimates and autosome estimates.

populations in their countries of origin, nor in the US as a whole. Consistent with the geographic variation in continental ancestry among Hispanic/Latino samples across the US as reported by Bryc et al.,[65] we observed variation in continental-ancestry proportions among HCHS/SOL recruitment centers within a given background group

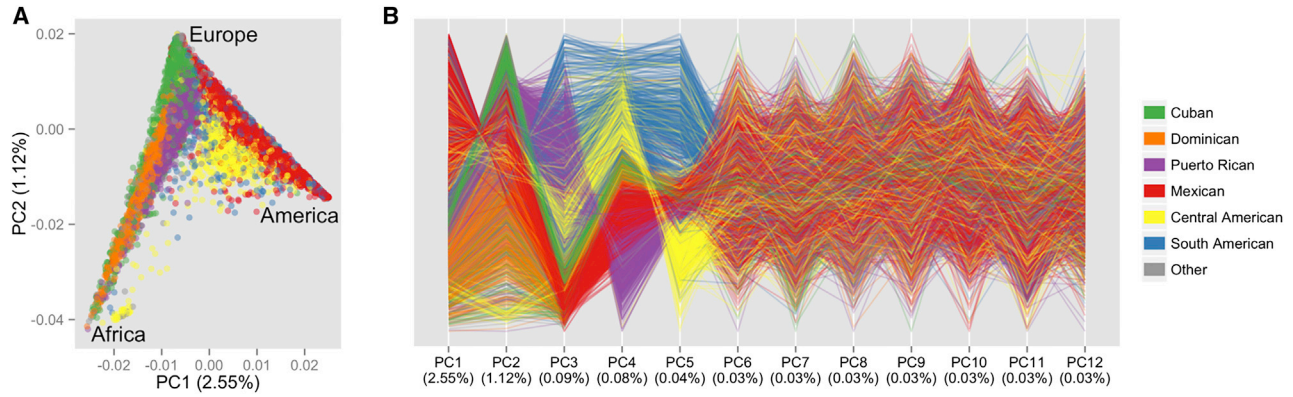

Figure 3. PCs of HCHS/SOL Participants

PCs were calculated with all individuals except for outliers with high East Asian ancestry. Color coding is by self-identified background. "Other" includes subjects who self-identified as having multiple or other backgrounds or had missing values.

(A) Scatter plot of PCs 1 and 2. Each point represents one individual. Regions representing high proportions of one continental ancestry are labeled. The three vertices of the triangle represent a high fraction of ancestry from each of the three continents, as determined by projecting control samples genotyped with the study samples (data not shown).

(B) Plot of parallel coordinates for the first 12 PCs. The 12 parallel vertical lines of equal length correspond to the first 12 PCs. Each individual is represented by a set of line segments connecting his or her PC values. The percentage of variance accounted for is given for each PC in the abscissa labels.

(Figure S5). For example, the fraction of Amerindian ancestry in self-identified Mexicans varied significantly among the recruitment centers ($p < 1 \times 10^{-16}$).

Figure 2B shows boxplots of the distributions of X and autosomal ancestry-proportion estimates for each of the six self-identified background groups. These plots show that the median value of Amerindian ancestry on the X chromosome is consistently higher than on the autosomes, whereas the corresponding European ancestry is consistently lower. The median value of African ancestry is also higher on the X in the Caribbean groups but is lower in the mainland groups. Mean values of ancestry proportions show the same pattern as the medians, except that the mainland groups have higher mean proportions of African ancestry on the X chromosomes (data not shown). All differences between the autosomes and X chromosome in the mean values are significant according to paired t tests ($p < 0.003$). Similarity due to relatedness is also expected to differ between the X and autosomes because of the different patterns of inheritance. Therefore, these chromosome types should be handled differently in association testing. The following results deal only with autosomal variation. X chromosome variation will be discussed in a separate communication.

Population Structure and Ancestry

Figure 3A shows a plot of the first two PCs for all individuals, color coded by self-identified background group. The points form a triangle, whose three vertices correspond to high proportions of the three major continental ancestries (European, African, and Amerindian). The left side contains mainly individuals reporting Caribbean backgrounds and represents an admixture gradient between European and African ancestry, whereas the right side contains predominantly individuals reporting mainland backgrounds and represents a gradient between European and Amerindian ancestries. This pattern has been observed previously in other studies of Hispanic/Latino individuals.[2,4]

A plot of parallel coordinates[66] for the first 12 PCs (Figure 3B) shows differentiation among the background groups for each of the first five PCs (see also pairwise PC plots in Figure S2), whereas PCs 6–12 show no clear separation. Even more differentiation among the six background groups is evident in three-dimensional (3D) plots of the first three PCs (Figure 4A). These figures show distinct differences among all six background groups, and they increase as the proportion of European ancestry decreases.

Genetic differentiation among individuals within a background group is associated with the geography of their countries of grandparental origin. Figures 4B and 4C show that individuals with origins in Colombia and Venezuela cluster closer to individuals with Central American origins than to those with origins in other South American countries. Additionally, plots of PCs from analyses using individuals for whom all four grandparents were born in a specific country in Central (Figures 5A and 5B) or South (Figures 5C and 5D) America show geographic structure (see also Figures S6 and S7). These results confirm the expected genetic diversity of the Central and South American background groups, which each represent multiple countries of origin. We defined these groups as self-identified background choices in the original HCHS/SOL questionnaire to avoid having many categories with small numbers of participants.

Relatedness

The HCHS/SOL sample contains substantial familial relatedness, as expected from the community-based, household sampling design. Figure S8 shows KC estimates from the final iteration of PC-Relate for all pairs of individuals

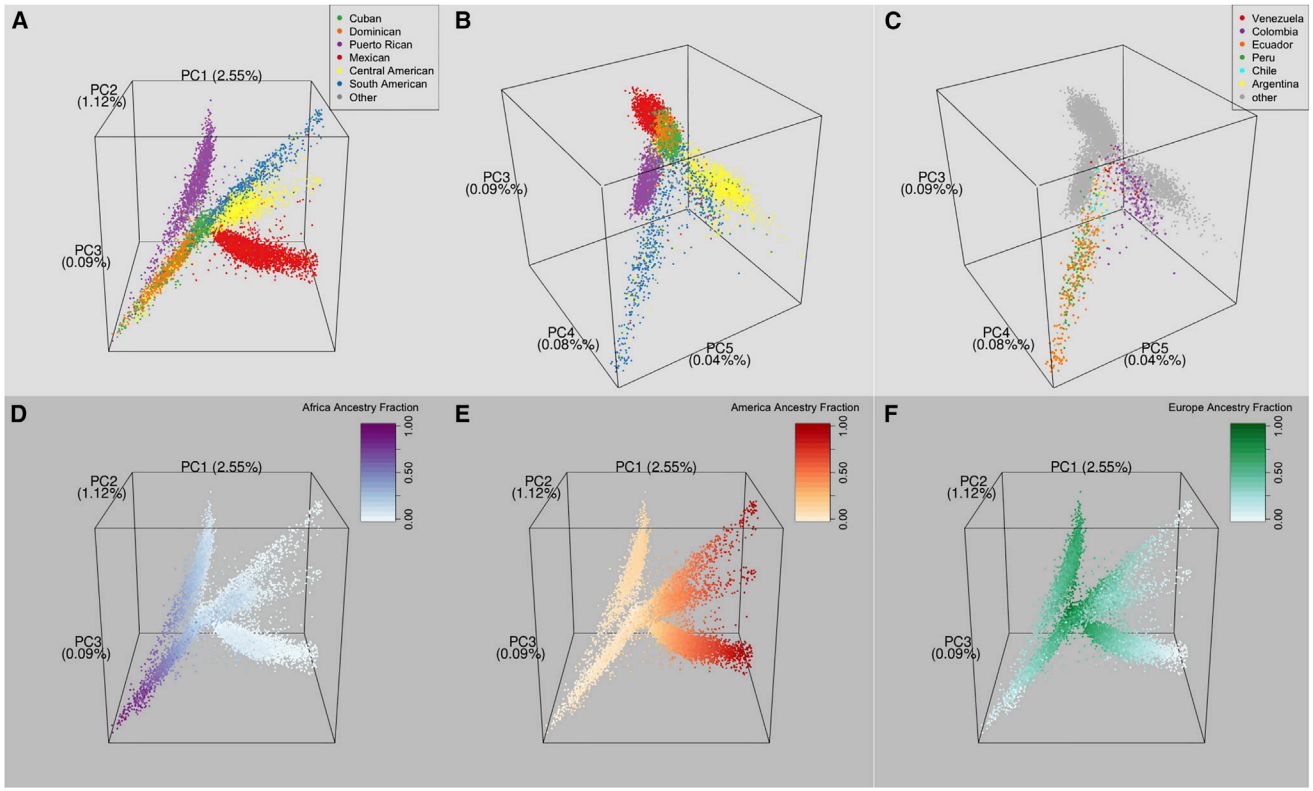

Figure 4. 3D Plots of PCs of HCHS/SOL Participants
The PCA included all study participants except for outliers with high East Asian ancestry.
(A) PCs 1–3 with color coding by self-identified background.
(B) PCs 3–5 with color coding as in (A).
(C) PCs 3–5 with color coding by country of grandparent origin for individuals who self-identified as having a South American background.
(D–F) PCs 1, 2, and 3 with color coding based on the ADMIXTURE estimates of continental-ancestry proportions from Africa (D), America (E), and Europe (F), respectively.

inferred to be related (and a subset of pairs inferred to be unrelated). The inferred relatives include 204 parent-offspring trios, an additional 1,042 parent-offspring duos, 699 full-sibling pairs, and numerous second-, third-, and fourth-degree relatives. The Figure S8 histogram showing the distribution of estimated KC values illustrates that first- and second-degree relative pairs form fairly distinct clusters, although more distant relationship types might have overlapping values and are difficult to distinguish. These results are expected given that the stochastic nature of segregation and recombination leads to variation in the realized relationship for a pair of individuals.[67,68] Among the genotyped participants, we identified a mutually "unrelated" set of 10,625 (83%) individuals by using a KC threshold of 0.044 (i.e., less than third-degree relatedness).[52] The remaining 2,159 individuals were inferred to be third-degree relatives or closer with at least one individual in the mutually unrelated set (and, in some cases, with each other).

Controlling for Confounding Due to Population Structure and Relatedness

Although only PCs 1–5 in the full sample set showed differentiation among the six main background groups, it is possible that higher-order PCs differentiate among geographic origins and/or immigration patterns within the six groups. To determine whether higher-order PCs are needed for adequate control of ancestry confounding in GWASs, we compared the genomic-control inflation factor[69] (λ_{GC}) for GWASs of 22 different traits by using LMMs in which no PCs, the first five PCs, or the first 20 PCs were included as covariates. Note that each of these LMMs used an empirical KC matrix that only measured genetic similarity due to familial relatedness, so inflated λ_{GC} values were expected if the PCs included in the model did not adequately control for population structure and ancestry. (This inflation would not be expected from LMM approaches that use an empirical GRM to measure genetic similarity due to both familial relatedness and shared ancestry.) Table S1 shows that the model with no PCs led to high inflation for all traits (λ_{GC} = 1.08–9.29; mean = 2.62), consistent with notable ancestral confounding. However, the models with either 5 or 20 PCs had essentially the same relatively low inflation (λ_{GC} = 1.00–1.07; mean = 1.03) for all 22 traits, suggesting that for many traits, higher-order PCs offer no further benefit in control of ancestry confounding beyond that achieved with the

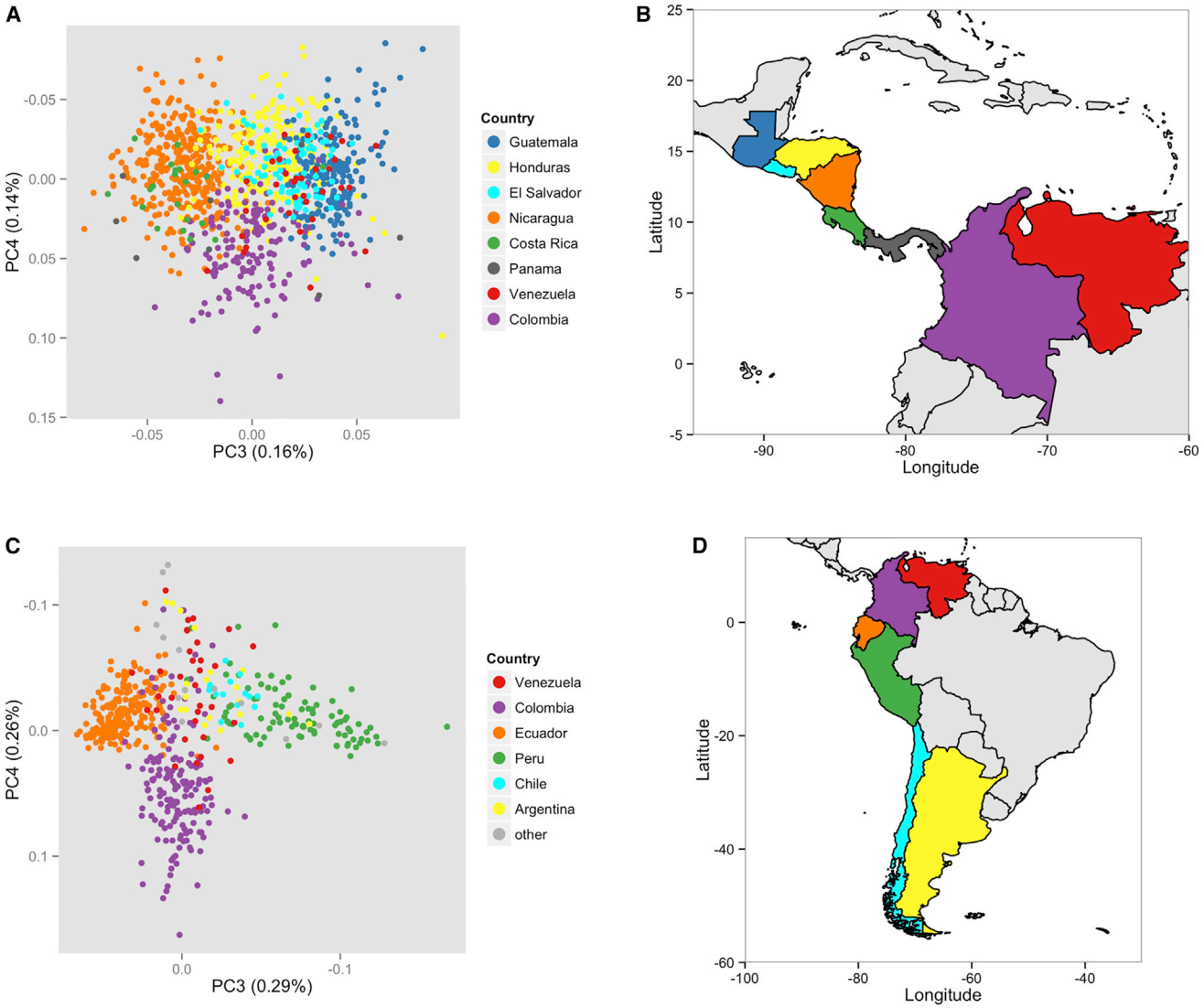

Figure 5. PCs by Country of Grandparental Origin

PCs were calculated with unrelated individuals whose four grandparents originated from the same country (according to participant reporting). The SNP set used for PCA was identical to that used for the overall PCA that excluded individuals with high East Asian ancestry.
(A) PCA of individuals with grandparents from Central America, Colombia, or Venezuela. A group of 37 outliers with high proportions of African ancestry (see Subjects and Methods) were excluded.
(B) Map showing geographical location of countries with grandparent origins of the individuals in (A).
(C) PCA of individuals with grandparents from the given South American countries. Two outliers were excluded.
(D) Map showing geographical location of countries with grandparent origins of the individuals in (C). The two PCs that most clearly separate the countries are shown in (A) and (C). Pairwise plots of other PCs are shown in Figures S6 and S7.

first five PCs. Additionally, we fit the linear-regression model that included the first five PCs as covariates but had no random effects to account for familial relatedness or shared environment. Ignoring the covariance structures due to kinship, household membership, and membership in a census block group led to higher genomic inflation for all 22 traits (λ_{GC} = 1.03–1.15; mean = 1.09).

Characteristics of the Genetic-Analysis Group

We used the categorical variable "genetic-analysis group," defined with both genetic variation and self-identified background, as a covariate in pooled GWASs and as a basis for stratified association studies. By design, the genetic-analysis groups have the same six values as the self-identified background groups (Cuban, Dominican, Puerto Rican, Mexican, Central American, or South American), and the two variables are highly concordant for these categories (95.6% overall; Figure S9). However, genetic-analysis groups were defined as having greater within-group genetic homogeneity, lacking within-group genetic outliers, and including all genotyped study participants (whereas self-identified background groups are missing or non-specific for 425 participants).

Within-group genetic outliers could have excessive influence in adjustment for ancestry in group-stratified analyses, and they could obscure the detection of more subtle

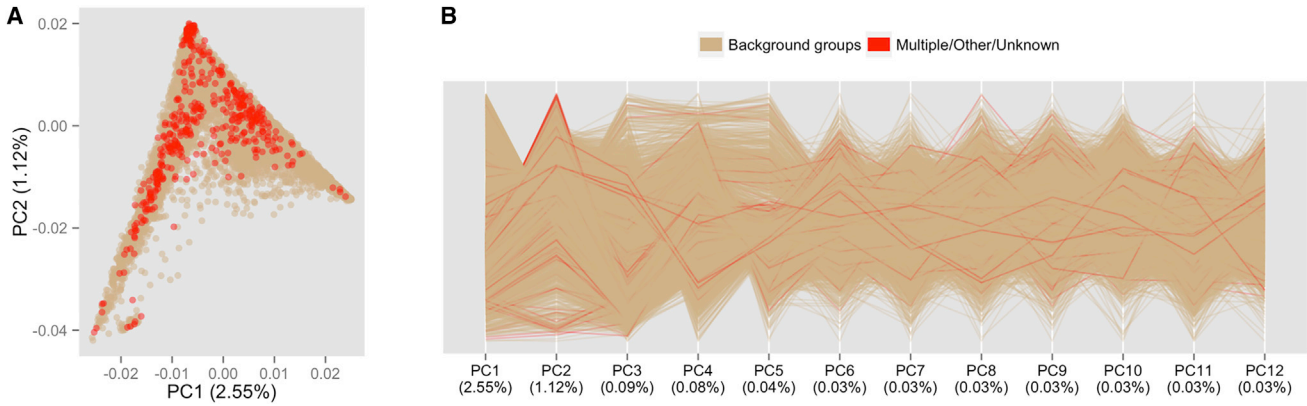

Figure 6. PCs Showing Genetic Similarity between Individuals with and without a Specific Self-Identified Background Group
The same PCs were calculated with all individuals except for outliers with high East Asian ancestry. Individuals self-identifying as one of the six specific background groups are tan, and individuals in the "other" group are red.
(A) PCs 1 and 2 and all of the "other" individuals (i.e., those with multiple, other, or missing responses) are plotted on top of those who self-identified with one of the six specific backgrounds.
(B) Plot of parallel coordinates for the first 12 PCs, including transparent colors and a random plot order. The percentage of variance accounted for is given for each PC in the abscissa labels.

population structure in group-specific PCA.[15] Figure 1 illustrates genetic outliers in a 3D plot of PCs from analysis of all HCHS/SOL participants, although only three of the six self-identified background groups are shown for clarity. The figure shows three background groups and their 99% tolerance ellipsoids, which capture most of the genetic variation within each background group. Genetic outliers for a given background group fall outside of that group's ellipsoid but are often within the ellipsoid of another background group. For example, some individuals with a Mexican background (red) fall within the Central American ellipsoid (yellow) and vice versa. These individuals would have discordant classifications for self-identified background and genetic-analysis groups. The example in Figure 1 illustrates ellipsoids defined with only three PCs, although five PCs were used for defining the actual genetic-analysis groups.

Initially, we used two sources of information to justify classification of individuals lacking a specific self-identified background into one of the six genetic-analysis groups. First, we examined the geographic origins of their grandparents (Figure S10), and these show that 67% were born in Latin American countries, 13% were born in Europe (mainly Spain), and only 20% were born elsewhere. Second, we examined the genetic homogeneity of the "other" self-identified group (including those who identified as being in more than one group or had no self-identification) in relation to the remaining participants in the PC space. Figure 6 shows that "other" individuals are not outliers for any of the first 12 PCs. Some of these individuals have high values for PC 2 but are not markedly different from other individuals with high European ancestry.

By definition, genetic-analysis groups are more homogeneous in the PC space than are the self-identified background groups from which they are derived. However, it is useful to characterize the efficacy of the multi-dimensional clustering method by visualizing group membership with the first five PCs and Mahalanobis distances within this 5-dimensional PC space. Figure 7A shows the distribution of Mahalanobis distances between individual points and the center of the hyper-ellipsoid for the Mexican group. This figure shows that the individuals who self-identify as Mexican but do not belong to the Mexican genetic-analysis group (the second boxplot) are further away from the hyper-ellipsoid center than are individuals who belong to the Mexican genetic-analysis group but have a different self-identified background (third boxplot) or no specific self-identification (fourth boxplot). These observations also hold for the other groups (Figure S11).

Figures 7B and 7C show plots of parallel coordinates from the PCA of the full sample set, but the plots include (on a common scale) only individuals in the self-identified Mexican background group (Figure 7B) or in the Mexican genetic-analysis group (Figure 7C). Figure 7B shows that self-identified Mexicans who do not belong to the Mexican genetic-analysis group appear as outliers for one or more of the first five PCs. Figure 7C shows that individuals who are assigned to the Mexican genetic-analysis group but do not self-identify as Mexican are not generally genetic outliers. We expect the genetic-analysis group to appear more homogeneous because of how it was defined, but this figure provides a visualization of the extent of improvement in homogeneity. This comparison is provided for the other groups in Figures S12 and S13.

Manichaikul et al.[2] previously described a variable similar to the genetic-analysis group in the Multi-Ethnic Study of Atherosclerosis (MESA). They defined four groups of Hispanic/Latino individuals by using k-means clustering ($k = 4$), initiated with the centers of four clusters in the PC 1–4 space, which corresponded approximately to four self-identification groups and gave high concordance for three of the four groups (98%, 93%, 90%, and 76%). We

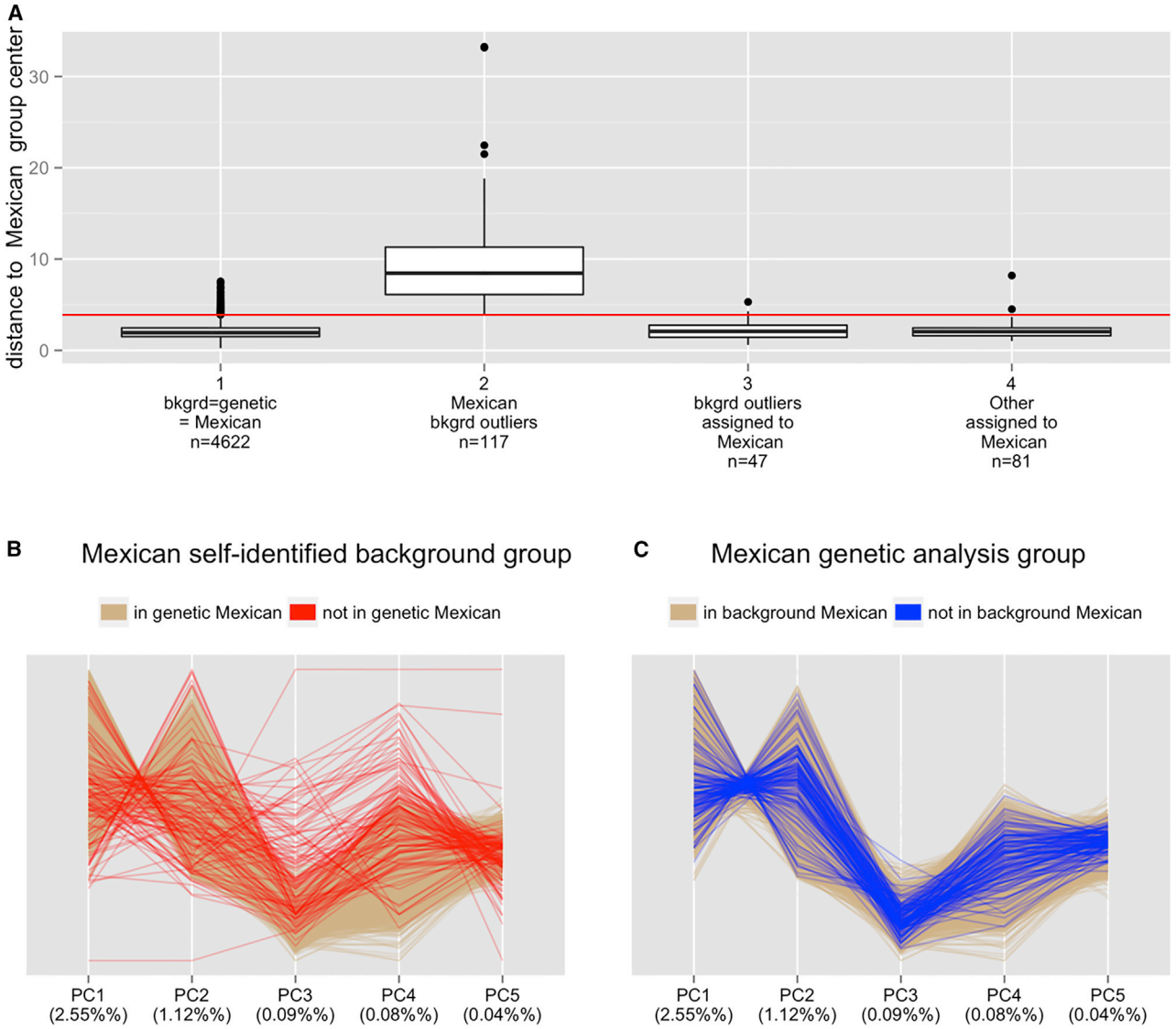

Figure 7. More Genetic Homogeneity in the Mexican Genetic-Analysis Group Than in the Mexican Self-Identified Background Group
(A) Distributions of the Mahalanobis distances between an individual in the five-dimensional PC space and the center of the Mexican hyper-ellipsoid. The four boxplots include individuals who belong to (1) both the Mexican self-identified background group and the Mexican genetic-analysis group, (2) the Mexican self-identified background group and another (not Mexican) genetic-analysis group, (3) one of the other (not Mexican) specific self-identified background groups and the Mexican genetic-analysis group, and (4) the "other" (i.e., multiple, other, or missing values) self-identified background group and the Mexican genetic-analysis group. The red line indicates the distance from the Mexican hyper-ellipsoid boundary to its center, which was one of the criteria for defining genetic-analysis groups. (B and C) Plots of parallel coordinates for individuals in either the Mexican self-identified background or the Mexican genetic-analysis group (PCs are from the PCA of all individuals except the outliers with high East Asian ancestry). The scaling of PCs is the same for both plots. Panel (B) shows only individuals in the Mexican self-identified background group and distinguishes those who are also in the Mexican genetic-analysis group from those who are not. Panel (C) shows only individuals who are in the Mexican genetic-analysis group and distinguishes those who are also in the Mexican self-identified background group from those who are not. Panel (B) shows that individuals who are in the self-identified Mexican background group but not the Mexican genetic-analysis group (red) tend to be outliers for one or a combination of PCs, whereas panel (C) shows that individuals who are in the self-identified non-Mexican background group and the Mexican genetic-analysis group are not outliers.

explored the use of k-means clustering in HCHS/SOL data by using the centers of the hyper-ellipsoids to initiate clustering, but this approach gave substantially lower concordance (82.8% overall, compared with 95.6% for the full hyper-ellipsoid method), and some groups performed particularly poorly (e.g., 45.6% for South Americans). The lower concordance in HCHS/SOL than in MESA might be related to larger sample size and higher diversity in HCHS/SOL. However, one of the main differences in the two approaches is the degree of supervision by self-identified background. The supervised k-means clustering that we applied uses only the centers of the background distributions, whereas the hyper-ellipsoid approach takes into account the multi-dimensional shape

of the distribution. Furthermore, in our approach, varying the parameters of the hyper-ellipsoid definition can modulate the degree of concordance between self-identified background and genetic-analysis group. Here, we aimed for concordance that was high enough to retain cultural and environmental information but not so high as to retain extreme genetic outliers.

Associations between Biomedical Traits and Genetic-Analysis Groups

For 22 quantitative traits in the HCHS/SOL study, we used Akaike's information criterion (AIC) to determine whether including genetic-analysis groups as a fixed effect in regression models contributes to model fit, after accounting for the increased model complexity that it entails. In a comparison of alternative models, lower AIC indicated better fit, although small AIC differences (less than about 3) might be expected because of chance alone.[70] The standard LMM described in Subjects and Methods was fit for each trait but included four model variations: PCs 1–5 with and without genetic-analysis groups and PCs 1–20 with and without genetic-analysis groups. AIC comparisons are summarized in Table S2. The addition of genetic-analysis groups to the model with PCs 1–5 resulted in an AIC that was lower by at least three units for eight traits, higher by at least three units for six traits, and within three units for eight traits. In addition, the six largest AIC differences (>10) favored the inclusion of genetic-analysis groups in the model. To determine whether the contribution of genetic-analysis groups depends on the number of PCs in the model, we repeated these comparisons by using models with the first 20 PCs; the results were essentially the same (Table S2). Therefore, genetic-analysis groups often improve the fit of a regression model that contains either 5 or 20 PCs. This observation indicates that genetic-analysis groups contain genetic information that is not captured by the first 20 PCs and/or that they capture non-genetic information associated with these traits, such as cultural and environmental diversity among the groups.

To compare the association between traits and either genetic-analysis groups or self-identified background groups, we evaluated models with PCs 1–5 plus either genetic-analysis groups or self-identified background groups (excluding individuals with a non-specific background group for a valid comparison of AIC values). Table S3 shows that using genetic-analysis groups rather than background groups resulted in an AIC that was lower by more than three units for three traits, higher by more than three units for four traits, and within three units for 15 traits. The number of traits analyzed (n = 22) is small, but it appears that genetic-analysis groups and self-identified background groups provide similar improvement to the fit of models that already have adjustment for genetic ancestry with PCs. However, genetic-analysis groups have the advantage of including individuals with missing or non-specific self-identified background groups.

Variance Heterogeneity among Genetic-Analysis Groups

Although results for most of the 22 traits analyzed have low genomic inflation according to the LMM with PC adjustment (Table S1), we investigated the possibility that heterogeneous residual variances among groups might contribute to the moderate inflation observed for a few traits (e.g., $\lambda_{GC} = 1.072$ for the FEV1/FVC ratio and 1.049 for log BMI). Using a pooled analysis for each trait (i.e., combined across the six genetic-analysis groups), we calculated the variance of the conditional residuals separately for each group and obtained the coefficient of variation (CV) among these six values as a measure of heterogeneity. This measure of heterogeneity ranged from 0.06 to 0.27 and was nearly identical when we used either genetic-analysis or self-identified background groups, after excluding individuals with a non-specific background (Table S4). Figure 8A shows a positive association between the CV of residual variance among genetic-analysis groups and genomic inflation across the 22 traits, suggesting that heteroscedasticity might contribute to genomic inflation.

To investigate further, we compared two LMMs for each trait: the original analysis assuming homoscedasticity and one that allowed for heteroscedasticity among groups. Both of these LMMs included PCs 1–5 and genetic-analysis groups as fixed-effect covariates, but the homoscedastic model fit one residual variance component for all observations, and the heteroscedastic model fit a separate residual variance component for each of the six genetic-analysis groups. Figure 9 compares the estimated residual variance component for the homoscedastic model with the six estimated residual variance components for the heteroscedastic model and provides a p value from a likelihood-ratio test of homoscedasticity. These p values are less than the Bonferroni-corrected threshold of 0.05/22 for 20 of 22 traits, indicating that heterogeneity of variance among groups is common. We found that genomic inflation was reduced in the heteroscedastic model for all traits, except for three that had λ_{GC} close to 1 in both homo- and heteroscedastic models (Figure 8B; Table S5). On average, allowing for heterogeneous residual variances in the LMM decreased λ_{GC} by 0.012. However, for the trait with the highest genomic inflation under the homoscedastic approach (FEV1/FVC ratio), allowing for heteroscedasticity led to a substantial reduction in λ_{GC} from 1.072 to 1.027.

Stratifying GWASs by genetic-analysis groups can also address heterogeneous variances among groups. Figure S14 shows that the LMM provides good control of genomic inflation in stratified analyses: $0.98 < \lambda_{GC} < 1.03$ in all groups for all 22 traits when we adjusted for ancestry confounding by using the first five PCs from either pooled or stratified PCA. Results of the stratified analysis can then be meta-analyzed, although this is complicated in HCHS/SOL by the fact that genetic-analysis (and self-identified background) groups lack independence because some individuals in different groups share census-block groups, households, and relatedness.

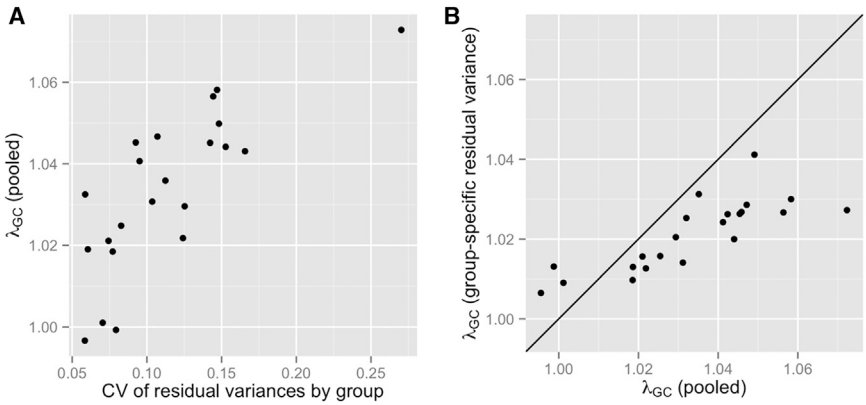

Figure 8. Relationship between Genomic Inflation Factor and Heteroscedasticity among Genetic-Analysis Groups in a Pooled GWAS

Each point in these plots is from a GWAS of 1 of 22 biomedical traits (see Table S1). (A) A measure of the degree of heteroscedasticity (the CV of residual variance by group) versus a measure of genomic inflation, λ_{GC}. Both were calculated from data pooled over the six genetic-analysis groups in an LMM analylsis assuming homoscedasticity.

(B) λ_{GC} is compared between a pooled LMM analysis that assumed homoscedasticity and a pooled LMM analysis that modeled the heterogeneous variances among the six genetic-analysis groups. (See Table S5 for the plotted data labeled by trait name.) All λ_{GC} values were calculated with autosomal SNPs filtered by an effective minor allele count, $N_{eff} > 120$, as described in Subjects and Methods. All models included age, sex, center, sampling weight, genetic-analysis group, and other trait-specific fixed effects, plus random effects for household, block group, and polygenic effects due to relatedness. The same sample set was used for both the homoscedastic and heteroscedastic models for each trait. Medians of 1,898,000 genotyped SNPs and 12,030,000 imputed SNPs were used in the λ_{GC} calculations.

Power to Detect Previously Known Associations

To further demonstrate the value of genetic-analysis groups, we compared the power of three LMM variations that differed in their use of a group variable. The first model used genetic-analysis groups as a covariate and as a stratification variable to fit separate residual variance components; this model excluded 37 individuals who had not been assigned to a genetic-analysis group. The second model was the same as the first but used self-identified background groups in place of genetic-analysis groups; this model excluded 425 subjects with a missing or non-specific background. The third model was the homoscedastic model that did not use a group variable and fit one residual variance component for all observations; this model had no sample exclusions.

We analyzed HCHS/SOL data with each of these three models and compared their Wald test statistics for the effects of SNPs with previously published associations for 12 of the 22 biomedical traits. Larger test statistics are indicative of higher power, given equivalent control of the type I error rate. However, whereas the models using genetic-analysis groups and self-identified background groups provided similar genomic inflation factors for all 12 traits, the model with no group variable typically provided larger genomic inflation factors, indicating an inflated type I error rate (Table S6). In order to provide a fair comparison among all three models, we divided the test statistics from each analysis by their respective genomic inflation factor. After this adjustment for genomic inflation, the test statistics from the model that used genetic-analysis groups were of similar magnitude on average to those from the model that used no groups, and they were systematically larger than those from the model that used self-identified background groups (Figure 10; Table S7). Most likely, smaller test statistics were observed from the model that used background groups because of the large number of samples that needed to be excluded from the analysis. The magnitude of the increase in the test statistics from the model using genetic-analysis groups is approximately what is expected from the difference in sample sizes (i.e., approximately equal to the expected increase in the test statistics, due to a decrease in the SEs, if the sample size in the analysis that used background groups was increased to match the same sample size in the analysis that used genetic-analysis groups). These results indicate that the LMM using genetic-analysis groups should be the preferred model, given that it provided the best control of genomic inflation and the most power among these models.

Discussion

Previous work has shown that a majority of US Hispanic/Latino individuals prefer to identify with their countries of origin (or background) rather than with either of the aggregate terms "Hispanic" and "Latino."[71] In the Americas, these countries span a vast geographic area that has diverse cultures, environments, colonization histories, and genetic ancestries. HCHS/SOL has participants who originate from many of these countries, and it has, to our knowledge, the largest sample size of any existing genetic study of US Hispanics/Latinos. Using self-reported grandparent origins, self-reported background, and genome-wide SNP data, we have demonstrated a high level of genetic diversity in the HCHS/SOL participants. We have also shown that this diversity is consistent with Latin American geography and the history of migration from other continents. The broad outlines of these genetic patterns have been reported previously— i.e., differences in continental-ancestry proportions between Caribbean and mainland populations (reflecting colonization history) and differentiation among small samples of background groups within these regions.[2–6,65] However, the HCHS/SOL analysis presented here reveals additional details of group differences, such as nearly complete separation of Puerto Rican

Figure 9. Residual Variance Components in LMM Regression with Samples Pooled over Genetic-Analysis Groups
For each of 22 biomedical traits, the "pooled" residual variance component was estimated with an LMM that included fixed effects for sex, age, center, sampling weight, genetic-analysis group, and PCs 1–5 and random effects for block group, household, and polygenic effects due to relatedness. In some cases, trait-specific fixed-effect covariates were also included. A second LMM that allowed for heterogeneous residual variance by background group was run. In each panel, the gray box shows the estimate for the residual variance component from the model that assumed homoscedasticity. The colored boxes show the estimated residual variance components by group in the heteroscedastic model in relation to the residual variance component from the homoscedastic model. The range of each box shows the 95% confidence interval. The p value from a likelihood-ratio test, with a null hypothesis of no heterogeneity, is also given for each trait.

from Cuban and Dominican background groups and of Mexican from Central and South American background groups in multi-dimensional genetic PC space. Self-identified grandparental origins of HCHS/SOL participants also document recent gene flow into US Hispanic/Latino populations from around the world. We note that these results are based on samples from communities in four US urban areas and might not represent the US as a whole or the Latin American populations from which they derive.

In HCHS/SOL, we found that continental-ancestry proportions differed substantially between the autosomes and the X chromosome. Across the background groups, the X chromosome consistently had higher Amerindian and lower European ancestry than the autosomes. These results are consistent with previous reports for US[65] and Latin American[4] samples. As suggested by Bryc and colleagues,[4] these differences between the X chromosome and autosomes might be due to sex-specific patterns of gene flow

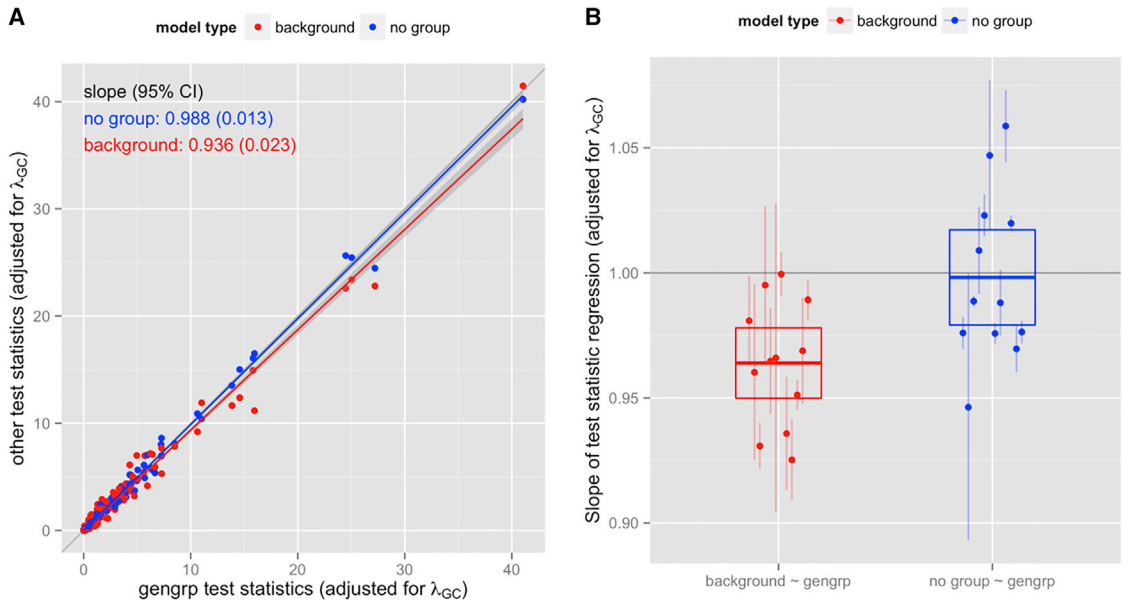

Figure 10. Using Models with Different Ethnic-Group Definitions to Compare Wald Test Statistics for the Effects of SNPs with Previously Published Trait Associations

For 12 biomedical traits, we performed association tests to assess power to detect known hits from the literature for models utilizing genetic-analysis groups, self-identified background groups, or no group variable. See Table S7 for SNP numbers and citations for each trait. All models included adjustment for sex, center, age, sampling weight, PCs 1–5, and trait-specific covariates. Random effects included block group, household, and genetic relatedness. The models also included genetic-analysis groups ("gengrp," up to 12,747 subjects), self-reported background groups ("background," up to 12,359 subjects), or no group variable ("no group," up to 12,784 subjects) as a covariate. For models using a group variable ("gengrp" and "background"), heterogeneous residual variance was fit by that group.

(A) For example, for log(BMI), the test statistics for 104 previously published SNP effects for the "no group" and "background" models are each plotted against the test statistics for the "gengrp" model. The test statistics for each model were divided by λ_{GC} from that model to provide genomic control. The solid lines show the linear regression (through the origin) of the "other" (i.e., "background" or "no group") test statistics on the "gengrp" test statistics, and the gray bands around the lines show the 95% confidence intervals. The thin gray line is $x = y$. The text on the plot gives the fitted slopes and their SEs. A slope less than 1 indicates a higher likelihood of detecting hits in the "gengrp" model than in the "other" model.

(B) Slopes from the linear regression of test statistics (as in A) for 12 different traits. The label "background ~ gengrp" refers to regression of the test statistic from the background-group model on that for the genetic-analysis- group model. Similarly, "no group ~ gengrp" refers to regression of the test statistic from the no-group model on that for the genetic-analysis-group model. Each point represents the slope for one trait and one model comparison, and its error bar represents the 95% confidence interval. The colored box for each model comparison shows the mean of the slopes from all 12 traits and its 95% confidence interval (see Table S7 for numeric values). The red points all have a slope less than 1.0 (mean 0.964), indicating that the background-group model has less power to detect previously detected GWAS hits than the genetic-analysis-group model. The blue points are scattered above and below a slope of 1.0 (mean 0.998), indicating no consistent difference in power between the no-group and genetic-analysis-group models.

in which European male colonists contributed more genetic material than did European females at the time of admixture. In addition to ancestry, similarity due to relatedness differs between the X and autosomes because of their different patterns of inheritance. These considerations indicate that adjustment for relatedness and ancestry should differ for the X and autosomes, and this could be one reason why the X chromosome has been neglected in GWASs.[72] We will describe the use of X-chromosome-specific PCs and KC estimates in X chromosome association studies in a separate report. In this article, the focus is on PCs and KC estimates based only on autosomal SNPs and their use in association testing of autosomal variants.

A primary goal of HCHS/SOL is to identify risk factors for disease, and this includes understanding the genetic architecture of variation in biomedical traits. The work presented here provides approaches to successfully account for genetic diversity in GWASs and other genetic studies involving Hispanic/Latino groups, and more generally in studies with multiple ethnicities, admixture, and relatedness. One of these tools is an approach to estimating KCs that are properly adjusted for ancestry and the multi-way admixture found in Hispanics/Latinos while simultaneously estimating PCs that represent ancestral variation without being influenced by familial relatedness. The basic methods that compose this procedure (PC-AiR and PC-Relate) have been described previously, but here we applied them in an iterative fashion to improve estimation. We used the resulting pairwise KC matrix and PCs to effectively control for the confounding effects of relatedness and ancestry in GWASs of many traits in HCHS/SOL, as determined by the genomic inflation factor (Table S1) and by examination of quantile-quantile and Manhattan plots (data not shown).

Although KCs and PCs are quite effective in controlling genomic inflation in the absence of knowledge of specific

Hispanic/Latino backgrounds, utilization of ethnic-group membership can increase the precision and versatility of genetic analyses. We have shown that many biomedical traits in HCHS/SOL have heterogeneous variances among ethnic groups, and modeling this heteroscedasticity typically further reduced genomic inflation. Although the reduction in inflation was modest for most traits (a mean decrease in λ_{GC} of 0.012), the effect was large for some traits, such as the FEV1/FVC ratio, for which λ_{GC} decreased from 1.072 to 1.027. We have also shown that including an ethnic-group variable (either genetic-analysis or self-identified background groups) as a fixed effect in the GWAS LMM improved the null model fit for many traits (as determined by the AIC), suggesting improved control of ancestry confounding and/or improved precision from modeling cultural and/or environmental effects. Furthermore, ethnic-group membership also can serve as a basis for stratified analyses when there is particular interest in certain groups. Examples include studying asthma in Puerto Rican individuals, among whom the prevalence is higher than in other groups,[10] and attempting to replicate discoveries made previously in a particular group, such as genetic associations with diabetes in Mexicans.[24]

Although either genetic-analysis groups or self-identified background groups can be used for controlling inflation, improving model fit, and stratifying analyses, we have shown two major advantages of genetic-analysis groups over self-identified background gruops. (1) Genetic-analysis groups are, by definition, more genetically homogeneous than self-identified background groups; they lack PC outliers in stratified analyses, which could have undue influence in ancestry adjustment and could hinder detection of and adjustment for important population structure. (2) The definition of genetic-analysis groups allows nearly all individuals to be classified into a specific group, whereas many individuals in HCHS/SOL have a missing or non-specific self-identified background. Therefore, analyses using genetic-analysis groups have a larger sample size than those using self-identified background groups, which contributes (solely or in conjunction with other factors) to increased power to detect previously documented associations with biomedical traits. The magnitude of the increase in power is, of course, related to the sample-size difference, which in HCHS/SOL was about 3%.

One can achieve the advantages of genetic-analysis groups while largely maintaining non-genetic effects associated with self-identified background groups. However, values of genetic-analysis groups are imputed (with the multi-dimensional PC method described here) for individuals who are PC outliers with respect to their self-identified background group and for individuals with a missing or non-specific self-identified background. There is a possibility that the imputed values create a mismatch between an individual's cultural and/or environmental characteristics and his or her group assignment, and this could potentially reduce power to detect associations. However, our investigation of power to detect previously known associations in the HCHS/SOL cohort indicates that the extra power obtained from being able to include nearly all participants in the analysis more than compensates for any potential power loss due to imputation inaccuracies. Our method for defining genetic-analysis groups is generally applicable to multi-ethnic populations in which genetic ancestry is associated with self-identified ethnicity. However, the power gained by imputing group membership for individuals with missing or non-specific self-identity will vary according to their number and similarity to other individuals in the study. Although the increase in power is dependent on the increased sample size achieved through imputation, the results presented here illustrate that genetic-analysis groups can provide an effective tool for imputing missing ethnic-group data.

In the presence of non-constant residual variance, which we observed for many traits in HCHS/SOL, alternatives to modeling group-specific residual variances include using robust SEs or generalized estimating equations (GEEs). However, in our experience, these approaches result in a loss of power to detect associations, and they require strong filtering because test statistics for SNPs with low minor allele frequencies are inflated (data not shown). For these reasons, we recommend using an LMM that allows for heteroscedasticity among groups. LMMs that fit heterogeneous residual variances have been used in agricultural statistics to account for heteroscedasticity due to environmental-interaction effects,[73,74] but to our knowledge, this approach has not previously been applied to human genetic data.

Meta-analysis of association analyses stratified by genetic-analysis group can test for overall effects and for heterogeneity of genetic effects among groups. Meta-analysis has been used widely in the GWAS field for combined analysis of summary-level data from different studies, for which pooled analysis is difficult in practice. When the data for different groups come from the same study (as in HCHS/SOL) and all individual-level data are readily available, both meta-analyses and combined analyses are practical. Lin and Zeng[75] have shown that mega- (i.e., pooled) and meta-analyses are equally efficient for detecting associations in many realistic settings. Furthermore, including a group-by-SNP interaction term in a pooled analysis can test for group-specific effects in essentially the same way as tests of heterogeneity in meta-analyses, such as Cochran's Q.[76] In HCHS/SOL, meta-analysis is complicated by the fact that genetic-analysis (and self-identified background) groups lack independence because some individuals in different groups share census-block groups, households, and relatedness. We are investigating methods to account for such correlations among groups, but at this time we recommend performing mega- rather than meta-analyses for detecting SNP associations and their possible differences across genetic-analysis groups in HCHS/SOL.

In this paper, we focused on ancestry adjustment based on global autosomal variation in GWASs and have not

addressed local-ancestry estimates, which are locus-specific estimates of the number of genomic segments derived from each putative ancestral population (e.g., zero, one, or two copies of each of three continental ancestries for Hispanic/Latino individuals). Previous studies have argued that it is important to adjust for local ancestry in admixed populations to avoid associations due to long-range admixture LD (e.g., Wang et al.[77]). However, adjusting for local ancestry can decrease power in discovery GWASs as a result of over-adjustment and might not be suitable.[78] Local-ancestry estimates are also used for admixture mapping, which relies on differences in causal allele frequencies between ancestral populations and has substantially lower resolution than the association studies discussed in this paper. We will pursue admixture mapping in a subsequent report.

Compared with other ethnic groups, Hispanic/Latino populations are under-represented in GWASs[79] (see GWAS Catalog in the Web Resources). Additional initiatives for participant recruitment and data collection in Hispanic/Latino studies are clearly needed to fill this gap, but there is also a need for genetic-analysis tools suited to the multi-way admixture and high diversity of Hispanic/Latino populations. In this paper, we have presented approaches for defining components of genetic diversity in Hispanic/Latino samples and for incorporating these components into association studies. These approaches include robustly estimating ancestry and relatedness, defining genetic-analysis groups that are more genetically homogeneous and more inclusive than self-identified background groups, and accounting for heterogeneous variances among groups. We expect that these approaches will help to narrow the gap in genetic discoveries in Hispanic/Latino populations.

Accession Numbers

The genotype and phenotype data reported in this paper are available under accession numbers dbGaP: phs000880.v1.p1 and phs000810.v1.p1.

Supplemental Data

Supplemental Data include 14 figures and 7 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.12.001.

Acknowledgments

We thank the participants and staff of the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) for their contributions to this study. The baseline examination of HCHS/SOL was carried out as a collaborative study supported by contracts from the NHLBI to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), and San Diego State University (N01-HC65237). The following institutes, centers, and offices contributed to the first phase of HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research (NIDCR), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Neurological Disorders and Stroke, and NIH Office of Dietary Supplements. The Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C AM03 and MOD03). Additional analysis support was provided by 1R01DK101855-01 and 13GRNT16490017. Genotyping efforts were supported by the NIH Department of Health and Human Services (HSN26220/20054C), National Center for Advancing Translational Science Clinical Translational Science Institute (UL1TR000124), and NIDDK Diabetes Research Center (DK063491). This manuscript has been reviewed by the HCHS/SOL Publications Committee for scientific content and consistency of data interpretation with previous HCHS/SOL publications.

Received: July 22, 2015
Accepted: December 2, 2015
Published: January 7, 2016

Web Resources

The URLs for data presented herein are as follows:

Bioconductor, http://www.bioconductor.org
GWAS Catalog, http://www.ebi.ac.uk/gwas/
R Statistical Computing Environment, http://www.r-project.org/

References

1. Rodriguez, C.J., Allison, M., Daviglus, M.L., Isasi, C.R., Keller, C., Leira, E.C., Palaniappan, L., Piña, I.L., Ramirez, S.M., Rodriguez, B., and Sims, M.; American Heart Association Council on Epidemiology and Prevention; American Heart Association Council on Clinical Cardiology; American Heart Association Council on Cardiovascular and Stroke Nursing (2014). Status of cardiovascular disease and stroke in Hispanics/Latinos in the United States: a science advisory from the American Heart Association. Circulation *130*, 593–625.
2. Manichaikul, A., Palmas, W., Rodriguez, C.J., Peralta, C.A., Divers, J., Guo, X., Chen, W.M., Wong, Q., Williams, K., Kerr, K.F., et al. (2012). Population structure of Hispanics in the United States: the multi-ethnic study of atherosclerosis. PLoS Genet. *8*, e1002640.
3. Gravel, S., Zakharia, F., Moreno-Estrada, A., Byrnes, J.K., Muzzio, M., Rodriguez-Flores, J.L., Kenny, E.E., Gignoux, C.R., Maples, B.K., Guiblet, W., et al.; 1000 Genomes Project (2013). Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS Genet. *9*, e1004023.
4. Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C.D., and Ostrer, H. (2010). Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proc. Natl. Acad. Sci. USA *107* (*Suppl 2*), 8954–8961.
5. Moreno-Estrada, A., Gignoux, C.R., Fernández-López, J.C., Zakharia, F., Sikora, M., Contreras, A.V., Acuña-Alonzo, V., Sandoval, K., Eng, C., Romero-Hidalgo, S., et al. (2014). Human genetics. The genetics of Mexico recapitulates Native

American substructure and affects biomedical traits. Science *344*, 1280–1285.
6. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., et al. (2013). Reconstructing the population genetic history of the Caribbean. PLoS Genet. *9*, e1003925.
7. Cardon, L.R., and Palmer, L.J. (2003). Population stratification and spurious allelic association. Lancet *361*, 598–604.
8. Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. *11*, 459–463.
9. Sorlie, P.D., Avilés-Santa, L.M., Wassertheil-Smoller, S., Kaplan, R.C., Daviglus, M.L., Giachello, A.L., Schneiderman, N., Raij, L., Talavera, G., Allison, M., et al. (2010). Design and implementation of the Hispanic Community Health Study/Study of Latinos. Ann. Epidemiol. *20*, 629–641.
10. Barr, R.G., Avilés-Santa, L., Davis, S.M., Aldrich, T., Gonzalez Ii, F., Henderson, A.G., Kaplan, R.C., LaVange, L., Liu, K., Loredo, J.S., et al. (2015). Pulmonary Disease and Age at Immigration Among Hispanics: Results from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Am. J. Respir. Crit. Care Med. Published online October 9, 2015.
11. Daviglus, M.L., Talavera, G.A., Avilés-Santa, M.L., Allison, M., Cai, J., Criqui, M.H., Gellman, M., Giachello, A.L., Gouskova, N., Kaplan, R.C., et al. (2012). Prevalence of major cardiovascular risk factors and cardiovascular diseases among Hispanic/Latino individuals of diverse backgrounds in the United States. JAMA *308*, 1775–1784.
12. Siega-Riz, A.M., Sotres-Alvarez, D., Ayala, G.X., Ginsberg, M., Himes, J.H., Liu, K., Loria, C.M., Mossavar-Rahmani, Y., Rock, C.L., Rodriguez, B., et al. (2014). Food-group and nutrient-density intakes by Hispanic and Latino backgrounds in the Hispanic Community Health Study/Study of Latinos. Am. J. Clin. Nutr. *99*, 1487–1498.
13. Kaplan, R.C., Bangdiwala, S.I., Barnhart, J.M., Castañeda, S.F., Gellman, M.D., Lee, D.J., Pérez-Stable, E.J., Talavera, G.A., Youngblood, M.E., and Giachello, A.L. (2014). Smoking among U.S. Hispanic/Latino adults: the Hispanic community health study/study of Latinos. Am. J. Prev. Med. *46*, 496–506.
14. Torgerson, D.G., Gignoux, C.R., Galanter, J.M., Drake, K.A., Roth, L.A., Eng, C., Huntsman, S., Torres, R., Avila, P.C., Chapela, R., et al. (2012). Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. J. Allergy Clin. Immunol. *130*, 76–82.e12.
15. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
16. Joliffe, I.T. (2004). Principal Component Analysis (Springer).
17. Rousseeuw, P.J., and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics *41*, 212–223.
18. Lavange, L.M., Kalsbeek, W.D., Sorlie, P.D., Avilés-Santa, L.M., Kaplan, R.C., Barnhart, J., Liu, K., Giachello, A., Lee, D.J., Ryan, J., et al. (2010). Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. Ann. Epidemiol. *20*, 642–649.
19. Conomos, M.P., Miller, M.B., and Thornton, T.A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet. Epidemiol. *39*, 276–293.
20. Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton, T.A. (2016). Model-free estimation of recent genetic relatedness. Am. J. Hum. Genet. *98*, this issue, 127–148.
21. Gao, X., Gauderman, W.J., Liu, Y., Marjoram, P., Torres, M., Haritunians, T., Kuo, J.Z., Chen, Y.D., Allingham, R.R., Hauser, M.A., et al. (2013). A genome-wide association study of central corneal thickness in Latinos. Invest. Ophthalmol. Vis. Sci. *54*, 2435–2443.
22. Cheng, I., Chen, G.K., Nakagawa, H., He, J., Wan, P., Laurie, C.C., Shen, J., Sheng, X., Pooler, L.C., Crenshaw, A.T., et al. (2012). Evaluating genetic risk for prostate cancer among Japanese and Latinos. Cancer Epidemiol. Biomarkers Prev. *21*, 2048–2058.
23. Drake, K.A., Torgerson, D.G., Gignoux, C.R., Galanter, J.M., Roth, L.A., Huntsman, S., Eng, C., Oh, S.S., Yee, S.W., Lin, L., et al. (2014). A genome-wide association study of bronchodilator response in Latinos implicates rare variants. J. Allergy Clin. Immunol. *133*, 370–378.
24. Williams, A.L., Jacobs, S.B., Moreno-Macías, H., Huerta-Chagoya, A., Churchhouse, C., Márquez-Luna, C., García-Ortíz, H., Gómez-Vázquez, M.J., Burtt, N.P., Aguilar-Salinas, C.A., et al.; SIGMA Type 2 Diabetes Consortium (2014). Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature *506*, 97–101.
25. Fejerman, L., Ahmadiyeh, N., Hu, D., Huntsman, S., Beckman, K.B., Caswell, J.L., Tsung, K., John, E.M., Torres-Mejia, G., Carvajal-Carmona, L., et al.; COLUMBUS Consortium (2014). Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25. Nat. Commun. *5*, 5260.
26. Palmer, N.D., Goodarzi, M.O., Langefeld, C.D., Wang, N., Guo, X., Taylor, K.D., Fingerlin, T.E., Norris, J.M., Buchanan, T.A., Xiang, A.H., et al. (2015). Genetic Variants Associated With Quantitative Glucose Homeostasis Traits Translate to Type 2 Diabetes in Mexican Americans: The GUARDIAN (Genetics Underlying Diabetes in Hispanics) Consortium. Diabetes *64*, 1853–1866.
27. Manichaikul, A., Hoffman, E.A., Smolonska, J., Gao, W., Cho, M.H., Baumhauer, H., Budoff, M., Austin, J.H., Washko, G.R., Carr, J.J., et al. (2014). Genome-wide study of percent emphysema on computed tomography in the general population. The Multi-Ethnic Study of Atherosclerosis Lung/SNP Health Association Resource Study. Am. J. Respir. Crit. Care Med. *189*, 408–418.
28. Comuzzie, A.G., Cole, S.A., Laston, S.L., Voruganti, V.S., Haack, K., Gibbs, R.A., and Butte, N.F. (2012). Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. PLoS ONE *7*, e51954.
29. Melton, P.E., Carless, M.A., Curran, J.E., Dyer, T.D., Göring, H.H., Kent, J.W., Jr., Drigalenko, E., Johnson, M.P., Maccluer, J.W., Moses, E.K., et al. (2013). Genetic architecture of carotid artery intima-media thickness in Mexican Americans. Circ. Cardiovasc. Genet. *6*, 211–221.
30. Norris, J.M., Langefeld, C.D., Talbert, M.E., Wing, M.R., Haritunians, T., Fingerlin, T.E., Hanley, A.J., Ziegler, J.T., Taylor, K.D., Haffner, S.M., et al. (2009). Genome-wide association study and follow-up analysis of adiposity traits in Hispanic Americans: the IRAS Family Study. Obesity (Silver Spring) *17*, 1932–1941.
31. Rich, S.S., Goodarzi, M.O., Palmer, N.D., Langefeld, C.D., Ziegler, J., Haffner, S.M., Bryer-Ash, M., Norris, J.M., Taylor, K.D., Haritunians, T., et al. (2009). A genome-wide association scan

for acute insulin response to glucose in Hispanic-Americans: the Insulin Resistance Atherosclerosis Family Study (IRAS FS). Diabetologia *52*, 1326–1333.
32. Palmer, N.D., Langefeld, C.D., Ziegler, J.T., Hsu, F., Haffner, S.M., Fingerlin, T., Norris, J.M., Chen, Y.I., Rich, S.S., Haritunians, T., et al. (2010). Candidate loci for insulin sensitivity and disposition index from a genome-wide association analysis of Hispanic participants in the Insulin Resistance Atherosclerosis (IRAS) Family Study. Diabetologia *53*, 281–289.
33. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
34. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. *43*, 519–525.
35. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354.
36. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods *8*, 833–835.
37. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.
38. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. *46*, 100–106.
39. Wu, C., DeWan, A., Hoh, J., and Wang, Z. (2011). A comparison of association methods correcting for population stratification in case-control studies. Ann. Hum. Genet. *75*, 418–427.
40. Listgarten, J., Lippert, C., Kadie, C.M., Davidson, R.I., Eskin, E., and Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. Nat. Methods *9*, 525–526.
41. Conomos, M.P. (2014). Inferring, estimating and accounting for population and pedigree structure in genetic analyses. PhD thesis (University of Washington).
42. Laurie, C.C., Laurie, C.A., Rice, K., Doheny, K.F., Zelnick, L.R., McHugh, C.P., Ling, H., Hetrick, K.N., Pugh, E.W., Amos, C., et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat. Genet. *44*, 642–650.
43. Laurie, C.C., Doheny, K.F., Mirel, D.B., Pugh, E.W., Bierut, L.J., Bhangale, T., Boehm, F., Caporaso, N.E., Cornelis, M.C., Edenberg, H.J., et al.; GENEVA Investigators (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. Genet. Epidemiol. *34*, 591–602.
44. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
45. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
46. Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.
47. Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. G3 (Bethesda) *1*, 457–470.
48. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.
49. Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. *34*, 816–834.
50. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.
51. Cavalli-Sforza, L.L. (2005). The Human Genome Diversity Project: past, present and future. Nat. Rev. Genet. *6*, 333–340.
52. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.
53. Thornton, T., Tang, H., Hoffmann, T.J., Ochs-Balcom, H.M., Caan, B.J., and Risch, N. (2012). Estimating kinship in admixed populations. Am. J. Hum. Genet. *91*, 122–138.
54. Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? Surv. Methodol. *37*, 115–136.
55. Gilmour, A.R., Thompson, R., and Cullis, B.R. (1995). Average Information REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics *51*, 1440–1450.
56. Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326–3328.
57. Gogarten, S.M., Bhangale, T., Conomos, M.P., Laurie, C.A., McHugh, C.P., Painter, I., Zheng, X., Crosslin, D.R., Levine, D., Lumley, T., et al. (2012). GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. Bioinformatics *28*, 3329–3331.
58. Rousseeuw, P.J., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., and Maechler, M. (2015). robustbase: Basic Robust Statistics. R package version 0.92-3. http://CRAN.R-project.org/package=robustbase.
59. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer).
60. Kahle, D., and Wickham, K. (2013). ggmap: spatial visualization with ggplot2. R J. *5*, 144–161. http://journal.r-project.org/archive/2013-2011/kahle-wickham.pdf.
61. Schloerke, B., Crowley, J., Cook, D., Hofmann, H., Wickham, H., Briatte, F., Marbach, M., and Thoen, E. (2014). GGally: Extension to ggplot. R package version 0.5.0, http://CRAN.R-project.org/package=GGally.
62. Bivand, R., and Lewin-Koh, N. (2015). maptools: Tools for reading and handling spatial objects. R package version 0.8-34. http://CRAN.R-project.org/package=maptools.
63. Bivand, R., Keitt, T., and Rowlingson, B. (2015). rgdal: bindings for the Geospatial Data Abstraction Library. R package version 0.9-2. http://CRAN.R-project.org/package=rgdal.
64. Adler, D., and Murdoch, D. (2014). rgl: 3D visualization using OpenGL. R package version 0.95.1201. http://CRAN.R-project/package=rgl.

65. Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D., and Mountain, J.L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am. J. Hum. Genet. *96*, 37–53.
66. Wegman, E. (1990). Hyperdimensional data analysis using parallel coordinates. J. Am. Stat. Assoc. *85*, 664–675.
67. Speed, D., and Balding, D.J. (2015). Relatedness in the post-genomic era: is it still useful? Nat. Rev. Genet. *16*, 33–44.
68. Hill, W.G., and Weir, B.S. (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. *93*, 47–64.
69. Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997–1004.
70. Burnham, K.P., and Anderson, D.R. (2002). Model Selection and Multimodel Inference (New York: Springer-Verlag).
71. Taylor, P., Lopez, M.H., Martinez, J.H., and Velasco, G. (2012). When labels don't fit: Hispanics and their views of identity (Pew Research Center).
72. Wise, A.L., Gyi, L., and Manolio, T.A. (2013). eXclusion: toward integrating the X chromosome in genome-wide association analyses. Am. J. Hum. Genet. *92*, 643–647.
73. Denis, J.B., Piepho, H.P., and Van Eeuwijk, F.A. (1997). Modelling expectation and variance for genotype by environment data. Heredity *79*, 162–171.
74. Piepho, H.P. (2000). A mixed-model approach to mapping quantitative trait loci in barley on the basis of multiple environment data. Genetics *156*, 2043–2050.
75. Lin, D.Y., and Zeng, D. (2010). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet. Epidemiol. *34*, 60–66.
76. Cochran, W.G. (1950). The comparison of percentages in matched samples. Biometrika *37*, 256–266.
77. Wang, X., Zhu, X., Qin, H., Cooper, R.S., Ewens, W.J., Li, C., and Li, M. (2011). Adjustment for local ancestry in genetic association analysis of admixed populations. Bioinformatics *27*, 670–677.
78. Liu, J., Lewinger, J.P., Gilliland, F.D., Gauderman, W.J., and Conti, D.V. (2013). Confounding and heterogeneity in genetic association studies with admixed populations. Am. J. Epidemiol. *177*, 351–360.
79. Bustamante, C.D., Burchard, E.G., and De la Vega, F.M. (2011). Genomics for the world. Nature *475*, 163–165.

The American Journal of Human Genetics

Supplemental Data

Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos

Matthew P. Conomos, Cecelia A. Laurie, Adrienne M. Stilp, Stephanie M. Gogarten, Caitlin P. McHugh, Sarah C. Nelson, Tamar Sofer, Lindsay Fernández-Rhodes, Anne E. Justice, Mariaelisa Graff, Kristin L. Young, Amanda A. Seyerle, Christy L. Avery, Kent D. Taylor, Jerome I. Rotter, Gregory A. Talavera, Martha L. Daviglus, Sylvia Wassertheil-Smoller, Neil Schneiderman, Gerardo Heiss, Robert C. Kaplan, Nora Franceschini, Alex P. Reiner, John R. Shaffer, R. Graham Barr, Kathleen F. Kerr, Sharon R. Browning, Brian L. Browning, Bruce S. Weir, M. Larissa Avilés-Santa, George J. Papanicolaou, Thomas Lumley, Adam A. Szpiro, Kari E. North, Ken Rice, Timothy A. Thornton, and Cathy C. Laurie

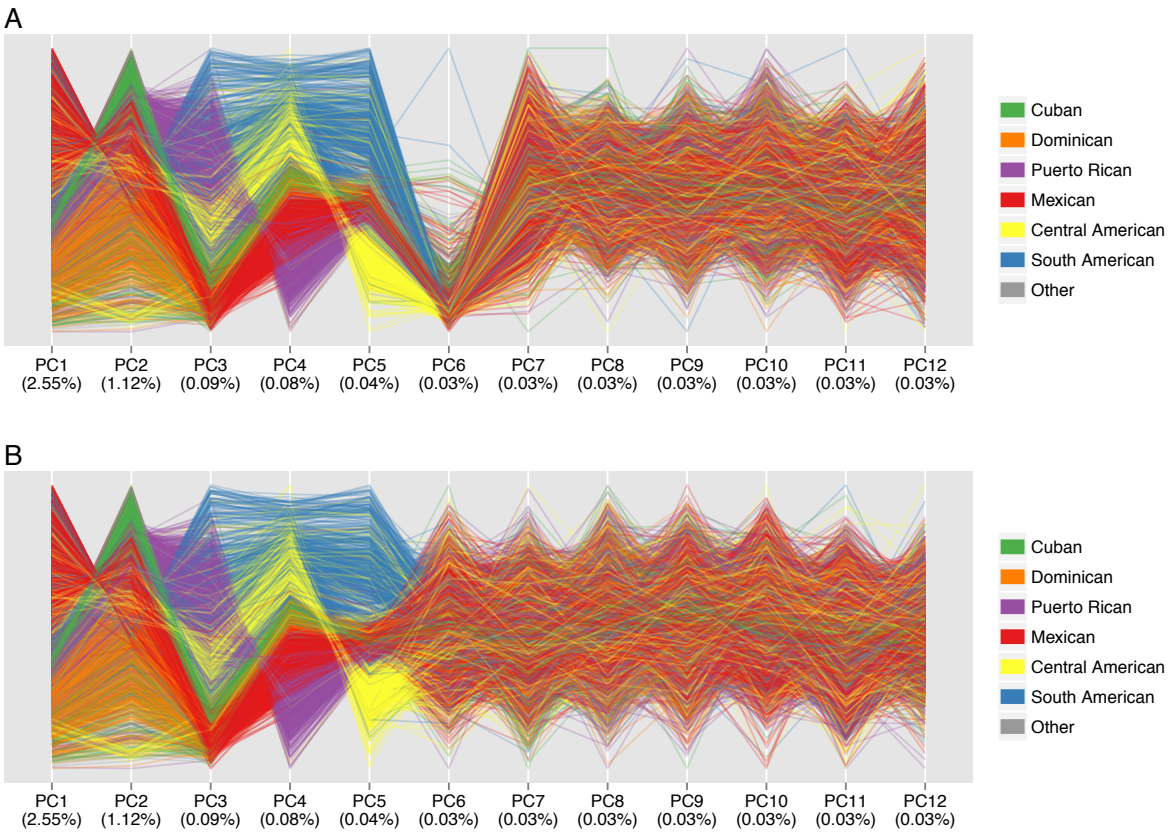

Figure S1: Principal components with and without 19 individuals having a high proportion of East Asian ancestry

(A) Parallel coordinates of the first 12 PCs using all HCHS/SOL individuals. PC6 separates a small number of individuals with predominantly East Asian ancestry.

(B) Parallel coordinates for the first 12 PCs for all individuals except for the 19 individuals with high East Asian ancestry. Percent variance accounted for is given for each PC in the horizontal axis labels. Color-coding is by self-identified background. "Other" includes subjects whose background is multiple, other or had a missing value.

The 12 parallel vertical lines of equal length correspond to the first 12 PCs. Each individual is represented by a set of line segments connecting their PC values.

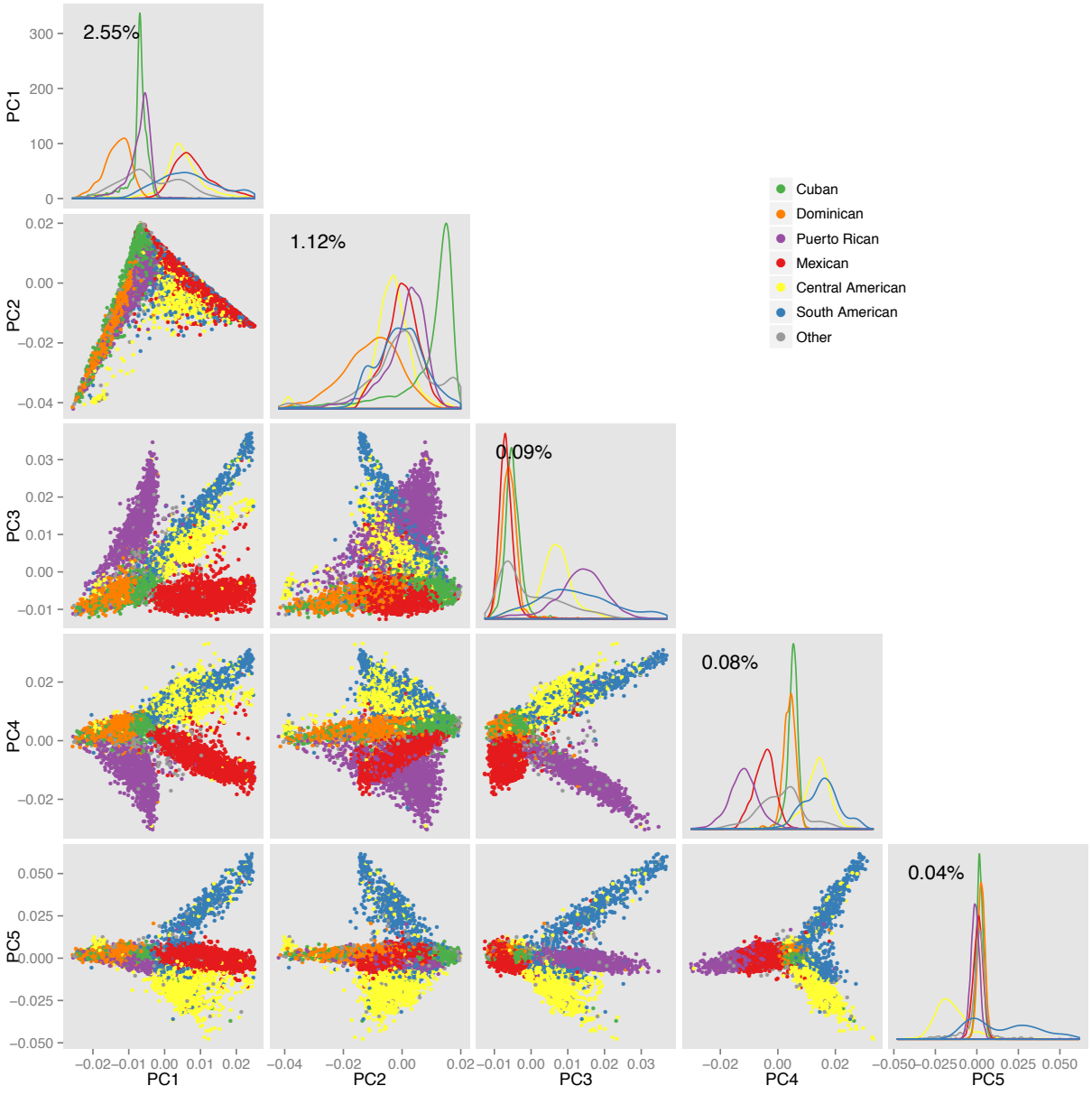

Figure S2: Pairwise PC plots for PCA of HCHS/SOL participants
All participants except the outliers with high East Asian ancestry were included in the PCA. Color-coding is by self-identified background. "Other" includes subjects whose background is multiple, other or had a missing value. The diagonal has density plots for the PC on the horizontal axis and the percent of variance accounted for by that PC.

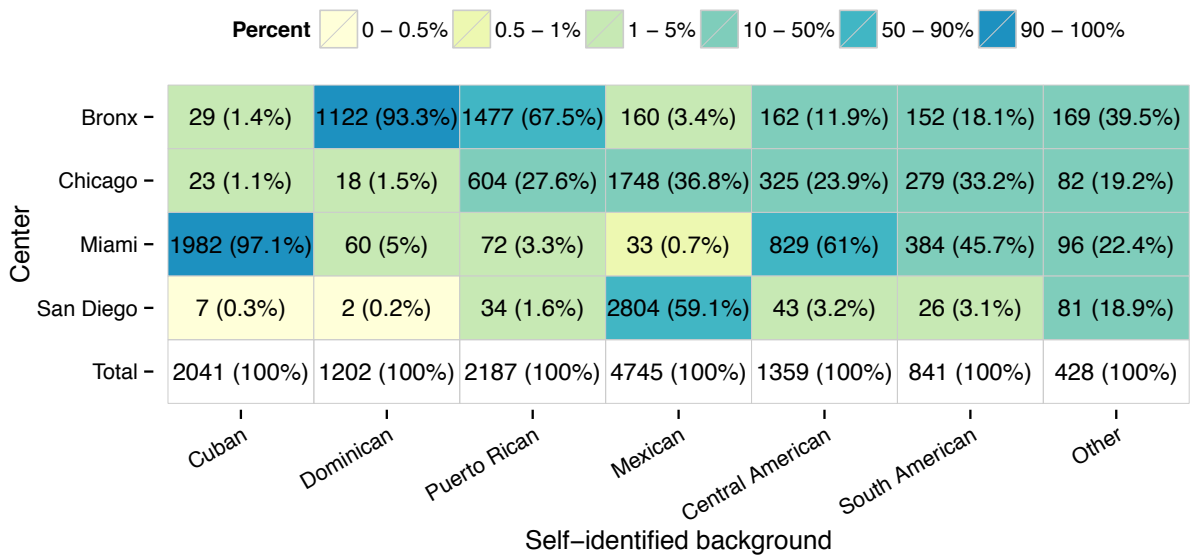

Figure S3: Cross-classification of recruitment center and self-identified background
The number in each box shows the number of genotyped participants from each recruitment center and self-identified background, while the color indicates the percentage of participants from each center for a given self-identified background. “Other” includes subjects whose background is multiple, other or had a missing value.

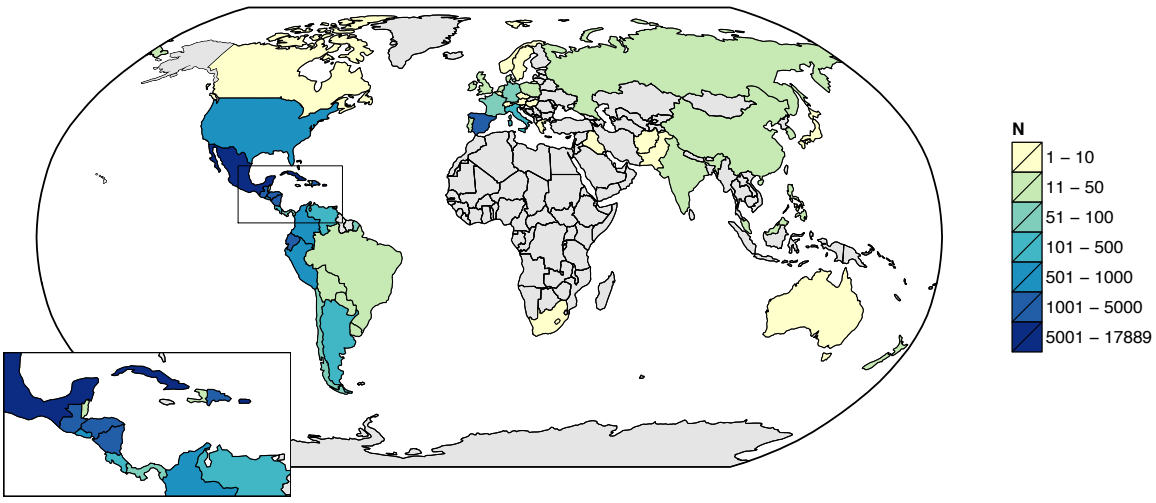

Figure S4: Grandparental origins for HCHS/SOL participants
World map with countries color-coded by the number of grandparents reported to originate from each country, with an inset that shows a larger version of Central America. Specific country origins for 49,626 grandparents were reported by 12,698 genotyped participants. The majority of grandparents come from the Americas. Grandparents from the U.S. are all placed in the lower 48 states (Alaska and Hawaii are colored grey even though it is possible that some U.S. grandparents could have come from those states). Grandparents from Puerto Rico are assigned to Puerto Rico, rather than the U.S.

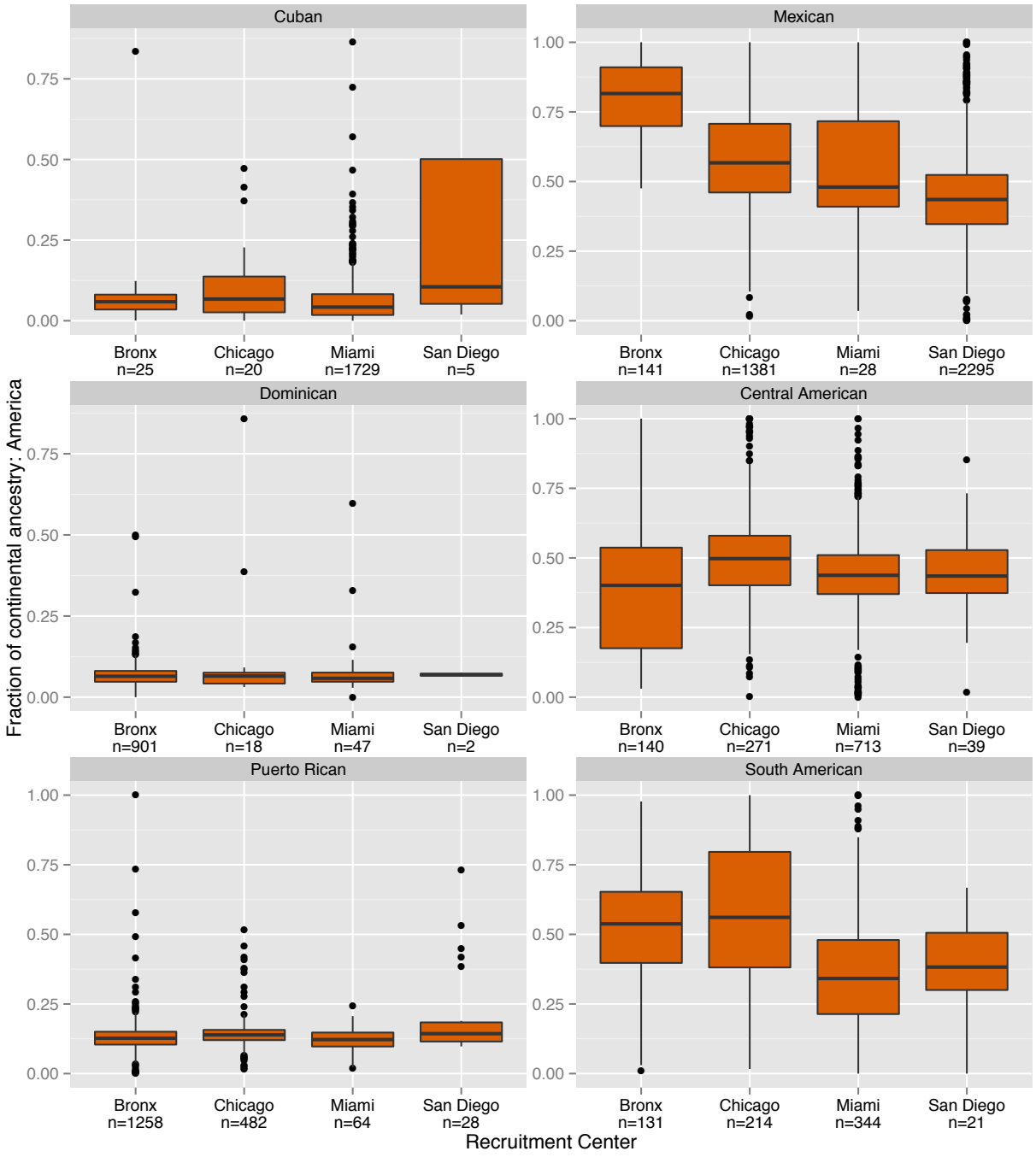

Figure S5: Distributions of the fraction of Amerindian ancestry by recruitment center and self-identified background

Boxplots showing the distributions of Amerindian ancestry proportions estimated for unrelated subjects only. Sample size for each center within a background group is indicated in the axis labels.

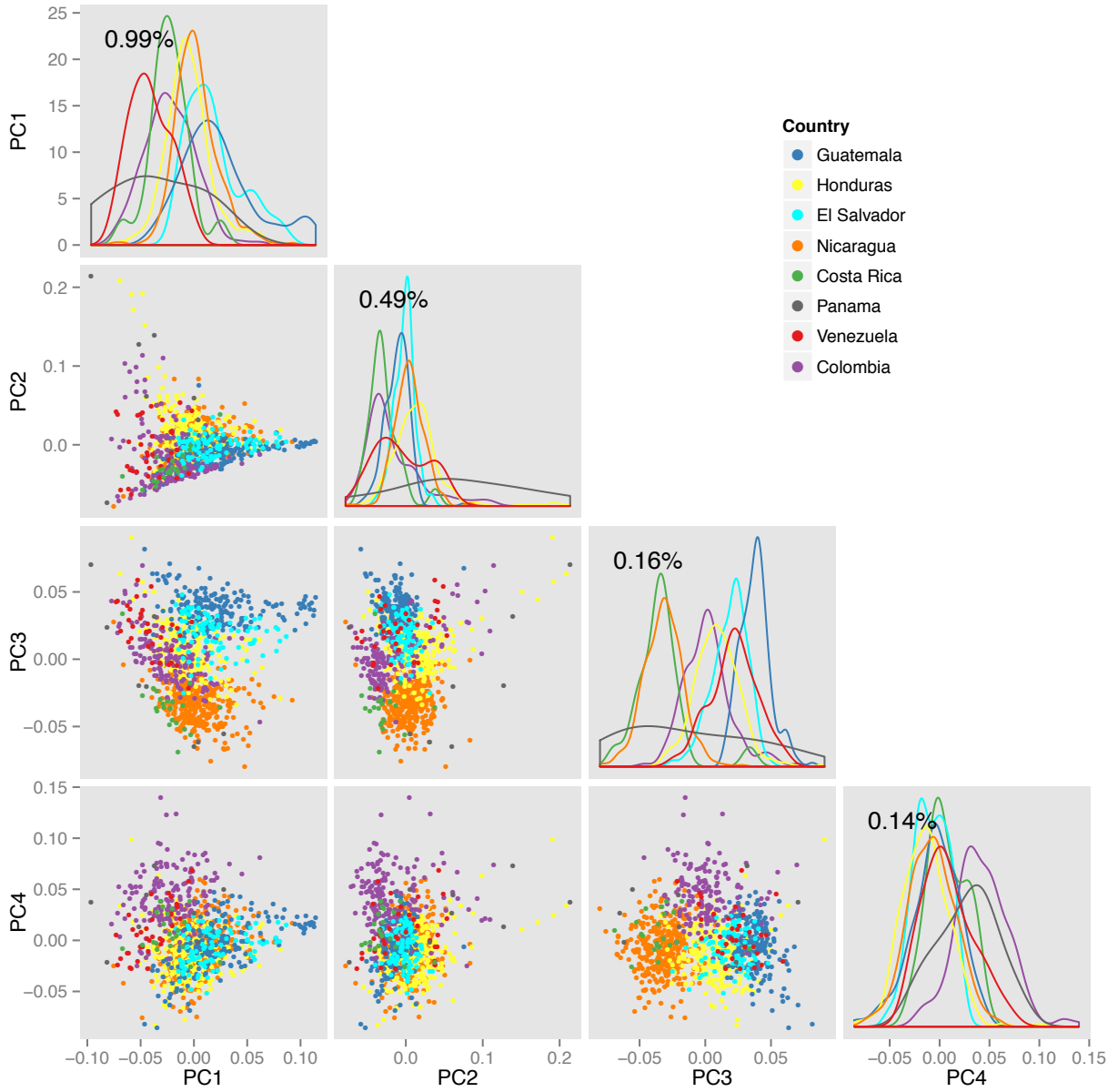

Figure S6: PCA of HCHS/SOL participants with all four grandparents from the same country for Central America, Colombia and Venezuela

Pairs plot showing PCs calculated using 1,094 unrelated subjects with all four grandparents from the countries given in the legend. The group of 37 outliers with high proportions of African ancestry (see Methods) were excluded. The diagonal has density plots for the PC on the horizontal axis and the percent of variance accounted for by that PC. The SNP set used was identical to the overall PCA excluding the outliers with high East Asian ancestry.

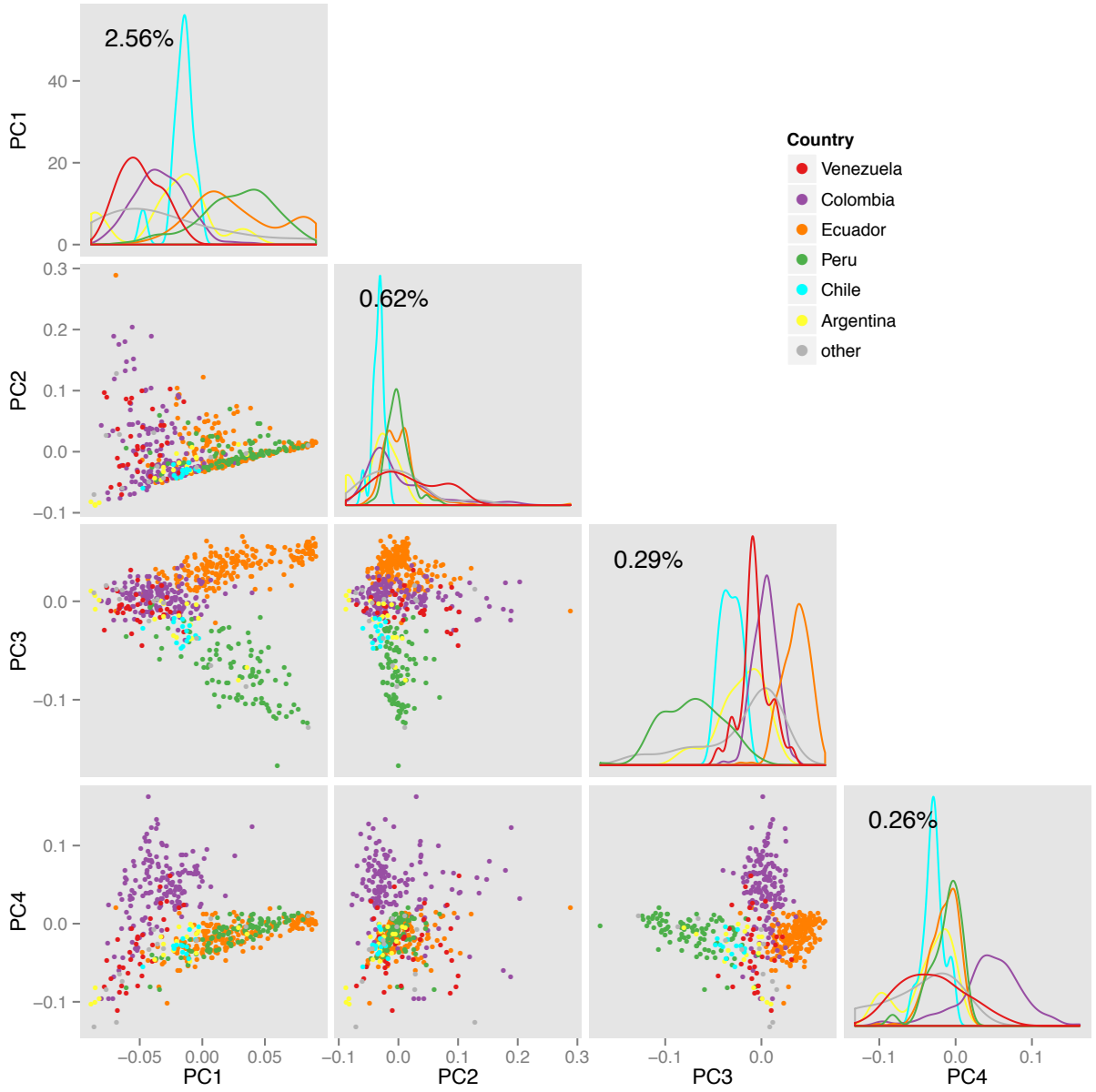

Figure S7: PCA of HCHS/SOL participants with all four grandparents from the same South American country

Pairs plots showing PCs calculated using 552 unrelated subjects with all four grandparents from the same South American country. Two outliers were excluded. The diagonals show density plots of each principal component. The SNP set used was identical to the overall PCA excluding the outliers with high East Asian ancestry.

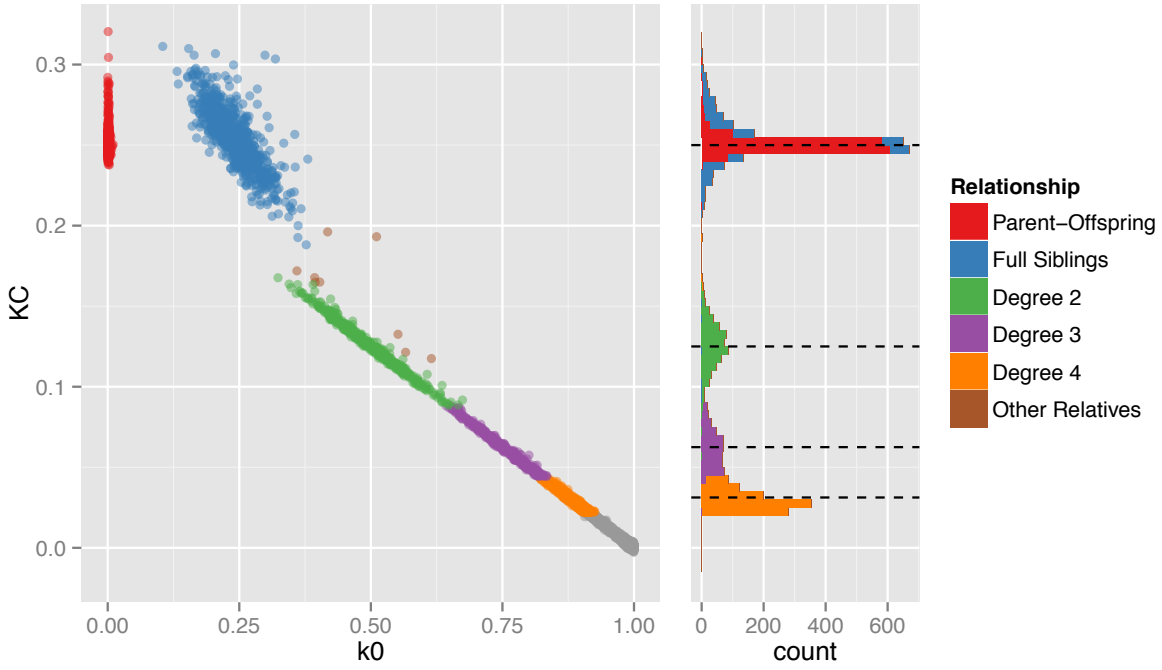

Figure S8: Relatedness estimates for pairs of HCHS/SOL participants
Relatedness was estimated with PC-Relate, excluding outliers with a high proportion of East Asian ancestry. Each point represents a pair of individuals. Because of the large number of unrelated pairs, only a random sample was included in the plot. The horizontal axis has estimates of k_0, which is the probably that, among the two alleles at a locus, zero are identical by descent. The vertical axis has estimates of the kinship coefficient, KC. A marginal histogram of KC color-coded by relationship (excluding unrelated subjects) is shown in the right panel. The black dashed lines represent the expected kinship coefficient for each degree of relatedness. A pair of individuals was assigned a relationship of degree d if $2^{-(d+3/2)} < KC \leq 2^{-(d+1/2)}$. Parent-Offspring pairs were distinguished from full siblings using a threshold of $k_0 = 2^{-11/2}$.

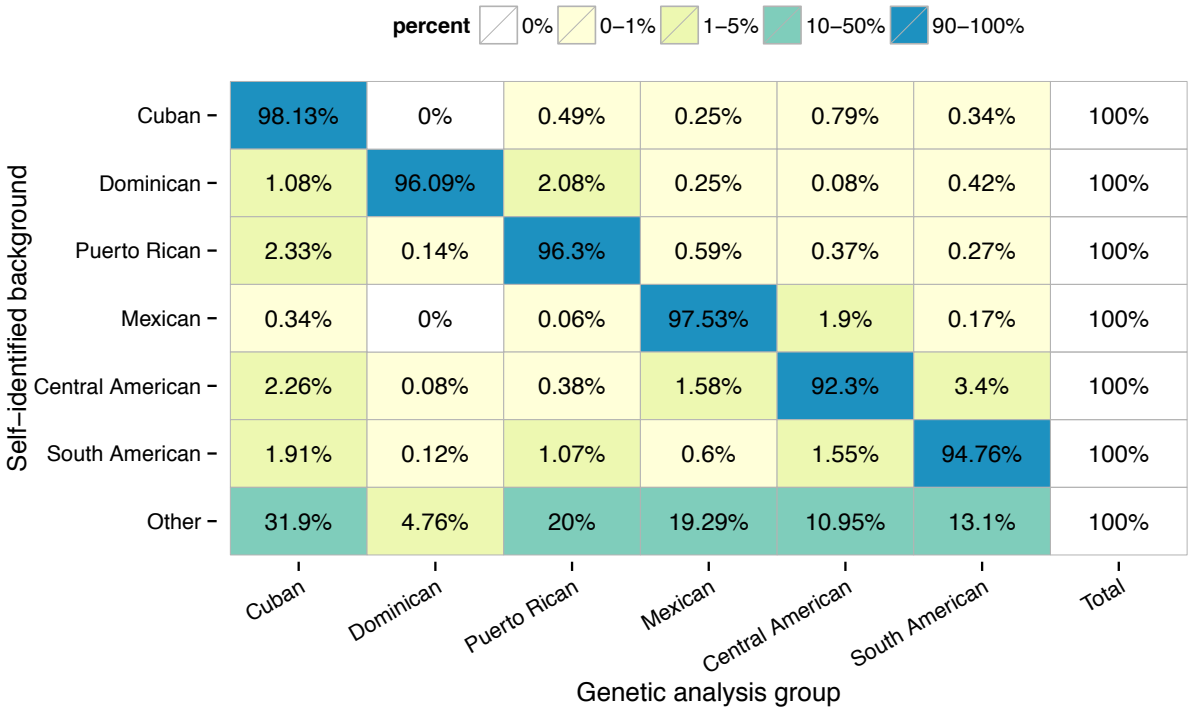

Figure S9: Cross-classification of genetic analysis group and self-identified background group Each cell has the percentage of individuals within a background group that are in the specified genetic analysis group (i.e. percentage of the row total). Color coding is based on this percentage. Individuals in the genetic analysis group having the same value appear on the diagonal. “Other” includes subjects whose background is multiple, other or had a missing value.

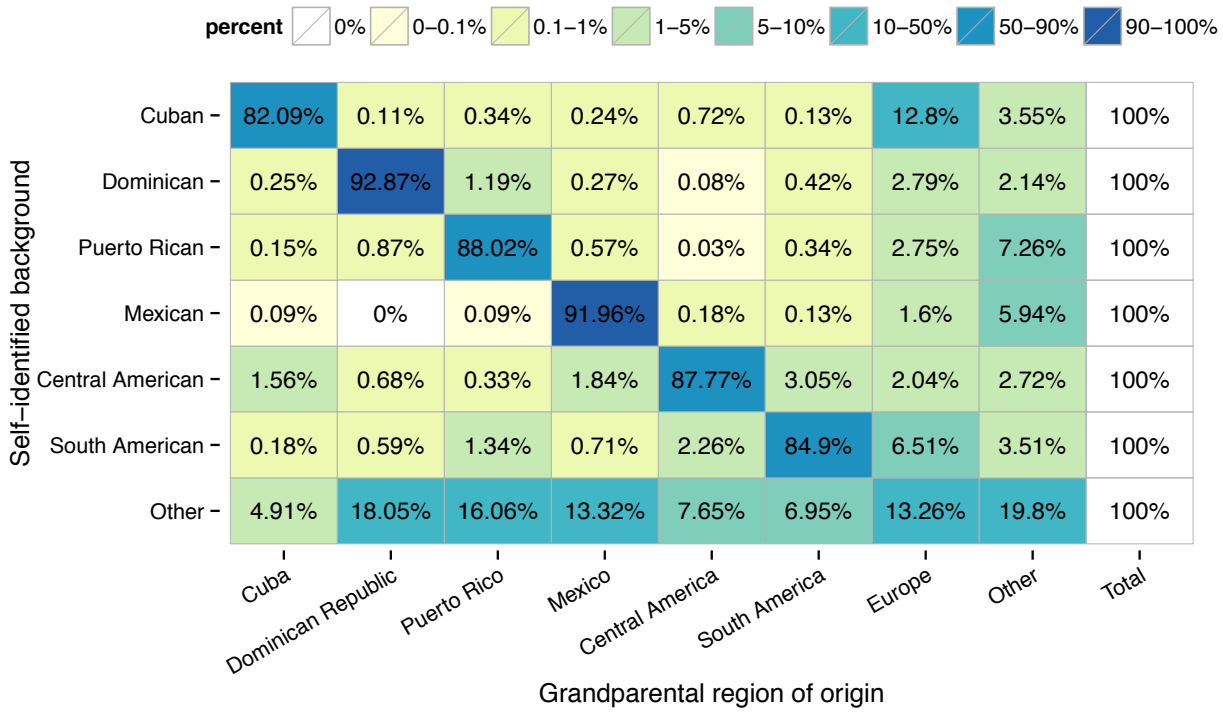

Figure S10: Cross-classification of grandparental region of origin and self-identified background group

Each cell (i,j) has the percentage of all grandparents for participants in the background group for the i^{th} row that originate from the region given in the j^{th} column. Color-coding is based on this percentage. "Other" background includes subjects whose background is multiple, other or had a missing value. "Other" grandparental region includes countries in regions other than the ones listed.

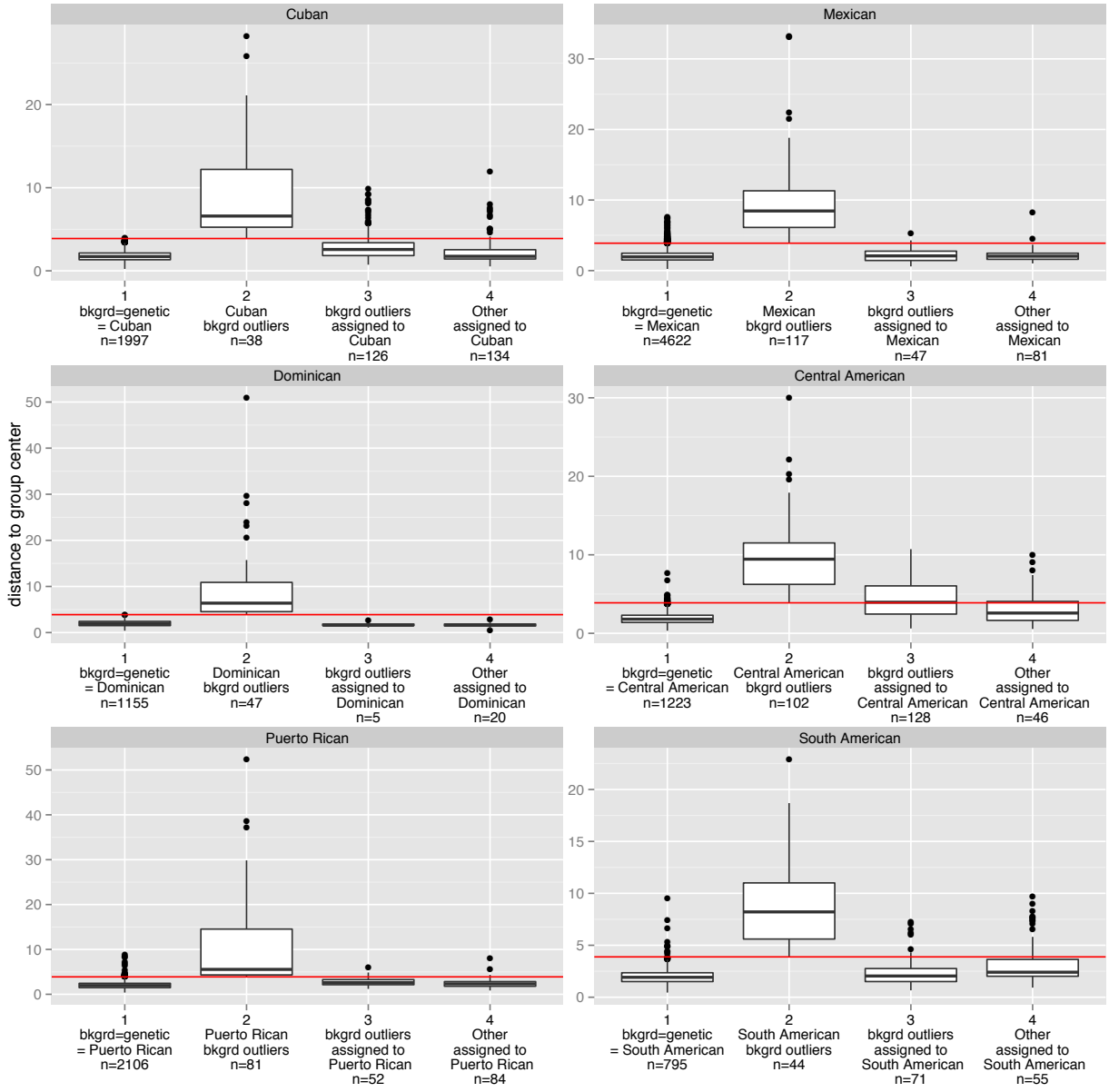

Figure S11: Genetic homogeneity of genetic analysis groups evaluated with Mahalanobis distances Distributions of the Mahalanobis distances between individual points in the five-dimensional PC space and the center of a given hyper-ellipsoid. For the Mexican hyper-ellipsoid, the four boxplots include individuals who belong to (1) both the Mexican self-identified background and Mexican genetic analysis groups, (2) the Mexican self-identified background group and another (not Mexican) genetic analysis group, (3) one of the other (not Mexican) specific self-identified background groups and the Mexican genetic analysis group, and (4) the "Other" (i.e. multiple, other or missing values) self-identified background group and the Mexican genetic analysis group. The red line indicates the distance from the Mexican hyper-ellipsoid boundary to its center, which was one of the criteria for defining genetic analysis group. These descriptions apply to each plot, substituting for "Mexican" the appropriate hyper-ellipsoid label at the top of the plot.

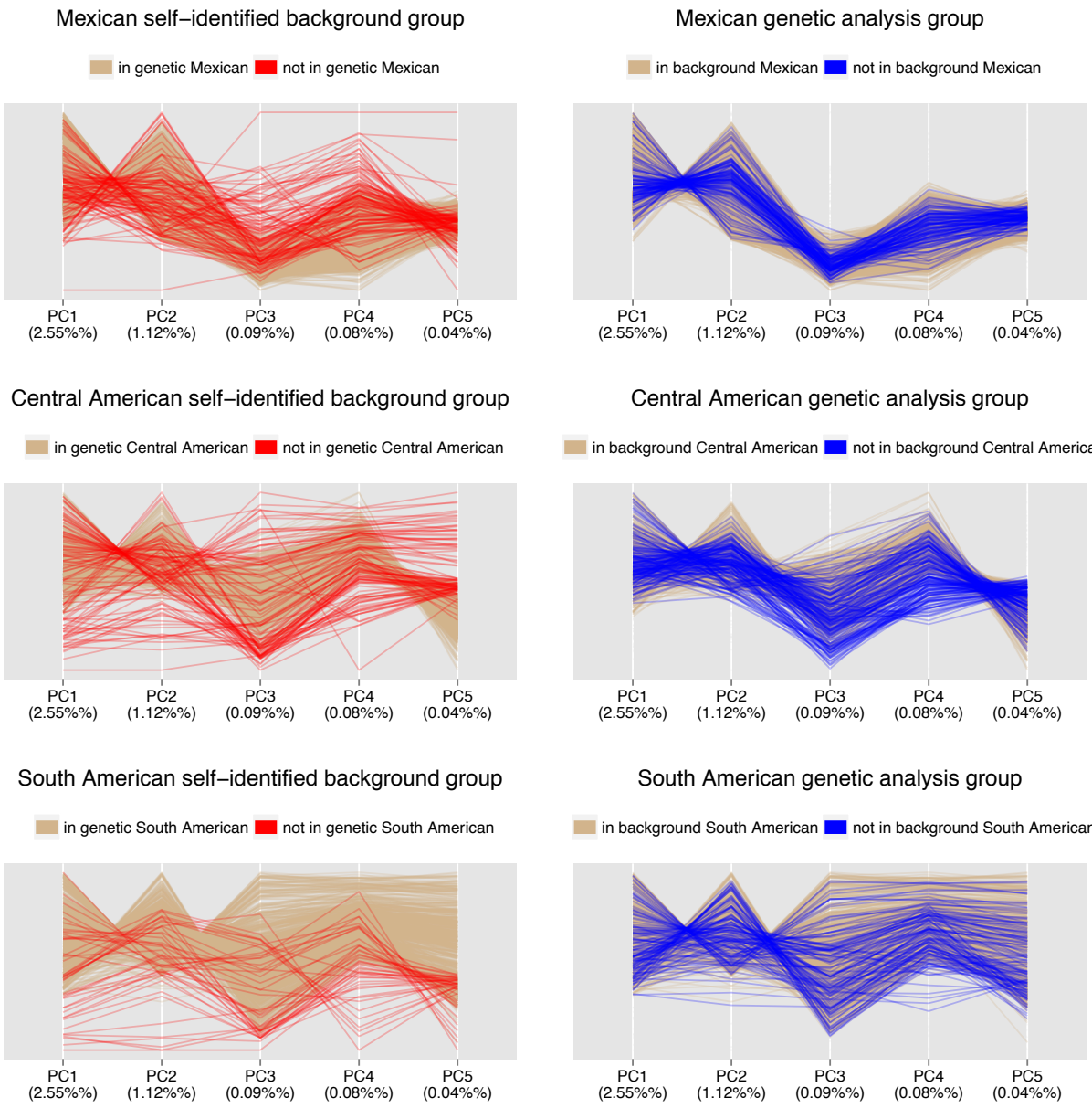

Figure S12: Genetic homogeneity of genetic analysis groups evaluated with PCs: Mainland groups The two Mexican group plots show parallel coordinates for individuals of either Mexican self-identified background (left) or Mexican genetic analysis group (right), from the PCA of all individuals except the outliers with high East Asian ancestry. The vertical scale is the same for both plots. The left plot shows only individuals in the Mexican self-identified background group, distinguishing those that are also in the Mexican genetic analysis group from those that are not. The right plot shows only individuals that are in the Mexican genetic analysis group, distinguishing those that are also in the Mexican self-identified background group from those that are not. The left plot shows that individuals with self-identified Mexican background that are not in the Mexican genetic analysis group (red) consist of outliers for one or more PCs. The right plot shows that self-identified non-Mexican background individuals who are in the Mexican genetic analysis group are not outliers. The same description applies to the other two groups shown here, substituting for “Mexican” the group name in the plot title.

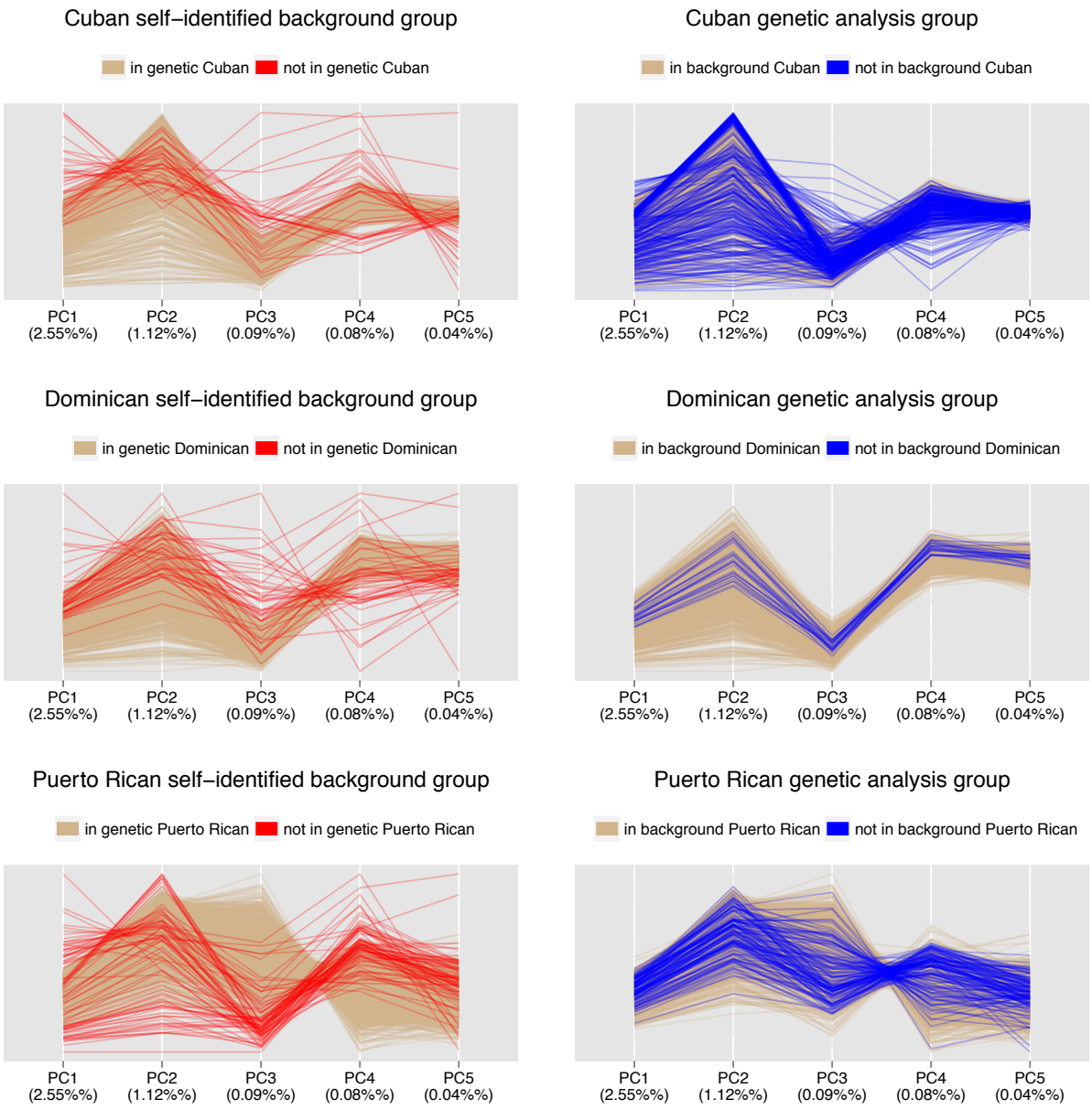

Figure S13: Genetic homogeneity of genetic analysis groups evaluated with PCs: Caribbean groups The two Cuban group plots show parallel coordinates for individuals of either Cuban self-identified background (left) or Cuban genetic analysis group (right), from the PCA of all individuals except the outliers with high East Asian ancestry. The vertical scale is the same for both plots. The left plot shows only individuals in the Cuban self-identified background group, distinguishing those that are also in the Cuban genetic analysis group from those that are not. The right plot shows only individuals that are in the Cuban genetic analysis group, distinguishing those that are also in the Cuban self-identified background group from those that are not. The left plot shows that individuals with self-identified Cuban background that are not in the Cuban genetic analysis group (red) consist of outliers for one or more PCs. The right plot shows that self-identified non-Cuban background individuals who are in the Cuban genetic analysis group are not outliers. The same description applies to the other two groups shown here, substituting for “Cuban” the group name in the plot title.

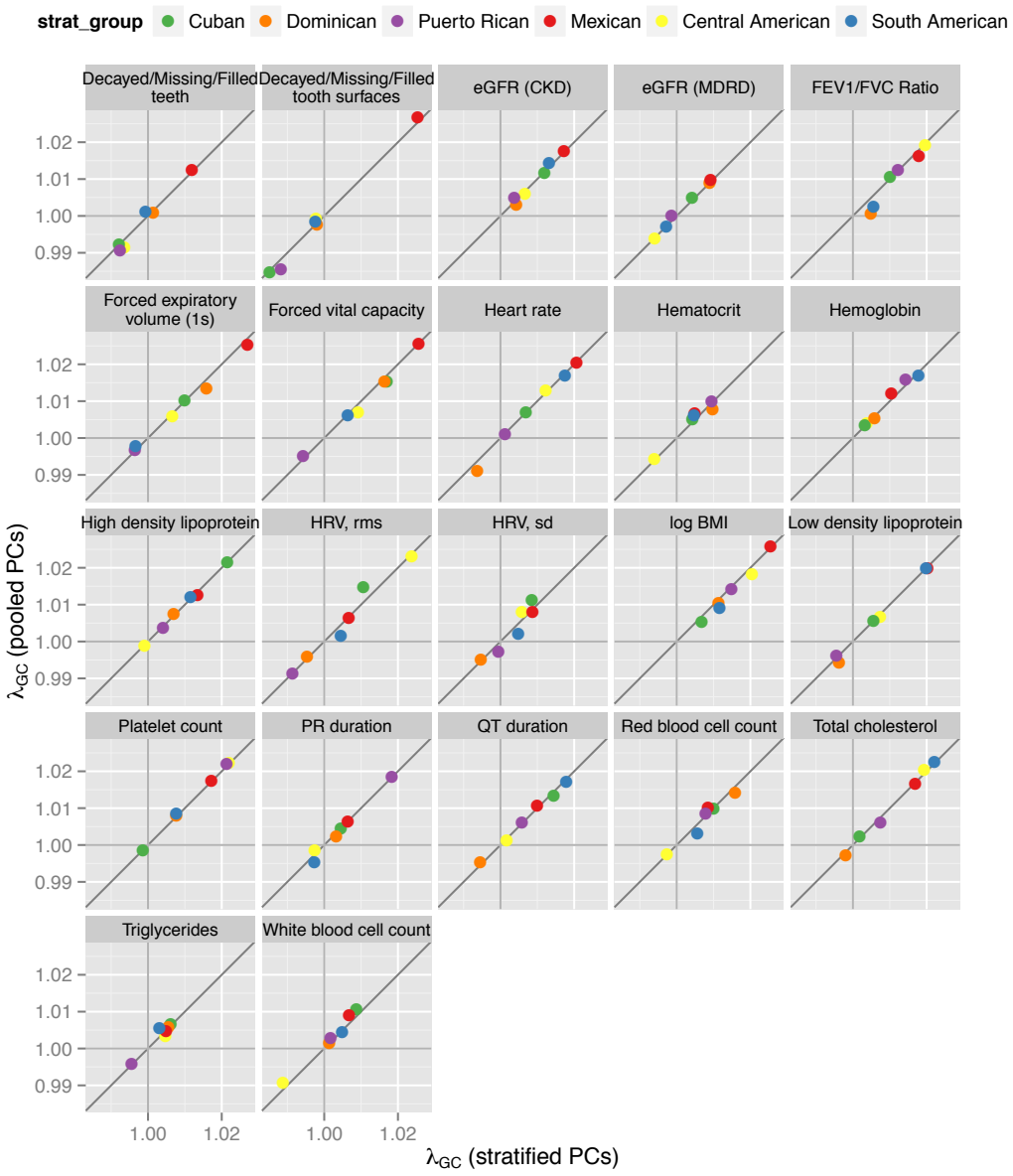

Figure S14: Genomic inflation factor (λ_{GC}) within each genetic analysis group for pooled PCs versus group-specific PCs

For each of 22 biomedical traits, λ_{GC} was calculated within each genetic analysis group for analyses with pooled principal components versus group-specific principal components. Five PCs were used in each case. SNPs were filtered on effective N $>$ 120 within each group. The median number of SNPs passing this filter for each group was: Cuban (8,206,000), Dominican (7,608,000), Puerto Rican (8,607,000), Mexican (8,660,000), Central American (6,736,000), and South American (5,817,000).

			λ_{GC}			
			LMM			LM
Trait group	Trait	N	no PCs	PC 1-5	PC 1-20	PC 1-5
Anthropometrics	log BMI	12,705	1.081	1.050	1.049	1.152
Blood cell count	Hematocrit	12,502	1.706	1.022	1.022	1.072
Blood cell count	Hemoglobin	12,502	1.901	1.030	1.030	1.073
Blood cell count	Platelet count	12,491	1.140	1.045	1.044	1.103
Blood cell count	Red blood cell count	12,502	1.152	1.058	1.059	1.124
Blood cell count	White blood cell count	11,809	2.168	1.057	1.057	1.121
Chronic kidney disease	eGFR (CKD)	12,739	5.672	1.047	1.045	1.111
Chronic kidney disease	eGFR (MDRD)	12,739	5.840	0.999	0.998	1.073
Dental	Decayed/Missing/Filled teeth	11,803	1.797	0.997	0.997	1.044
Dental	Decayed/Missing/Filled tooth surfaces	11,803	1.557	1.019	1.019	1.065
Electrocardiography	Heart rate	10,216	1.889	1.044	1.046	1.076
Electrocardiography	HRV, rms	10,798	1.253	1.025	1.025	1.051
Electrocardiography	HRV, sd	10,798	1.339	1.018	1.020	1.034
Electrocardiography	PR duration	11,686	1.645	1.041	1.041	1.099
Electrocardiography	QT duration	11,932	1.272	1.021	1.021	1.086
Lipids	High density lipoprotein	12,730	1.938	1.032	1.032	1.099
Lipids	Low density lipoprotein	12,467	1.090	1.045	1.046	1.115
Lipids	Total cholesterol	12,731	1.098	1.031	1.031	1.106
Lipids	Triglycerides	12,730	4.407	1.001	1.001	1.053
Pulmonary disease	FEV1/FVC Ratio	11,832	1.402	1.073	1.073	1.126
Pulmonary disease	Forced expiratory volume (1s)	11,833	6.972	1.043	1.043	1.101
Pulmonary disease	Forced vital capacity	11,832	9.287	1.036	1.036	1.097

Table S1: Genomic inflation factor (λ_{GC}) from GWAS for models that differ in accounting for relatedness and ancestry

λ_{GC} values for GWAS of 22 different biomedical traits run with 1) no PCs, 2) the first five or 3) the first twenty PCs to control for confounding by ancestry, and 4) without random effects to control for correlation structure. The model for each analysis included sex, center, age, sampling weight, and other trait-specific covariates. Models in the "LMM" category also included random effects for household, block group, and polygenic effects due to relatedness, and the "LM" model used a simple linear model that ignored correlations among subjects due to relatedness, household and block group. Genetic analysis group was not included. Sample size (N) varied by trait-specific exclusion criteria. All λ_{GC} values were calculated using autosomal SNPs filtered by an effective minor allele count, effN > 120, as described in Methods. The number of SNPs used to calculate λ_{GC} varied by analysis due to the varying sample size, but a median of 1,898,000 genotyped SNPs and 12,030,000 imputed SNPs were used in the calculations.

		AIC Difference	
Trait group	Trait	PC5 - PC5g	PC20 - PC20g
Anthropometrics	log BMI	18.4	16.8
Blood cell count	Hematocrit	-6.6	-6.9
Blood cell count	Hemoglobin	-6.6	-7.4
Blood cell count	Platelet count	-4.7	-4.3
Blood cell count	Red blood cell count	-6.6	-7.0
Blood cell count	White blood cell count	-0.5	-0.9
Chronic kidney disease	eGFR (CKD)	-1.9	-0.9
Chronic kidney disease	eGFR (MDRD)	-1.1	-0.2
Dental	Decayed/Missing/Filled teeth	19.3	18.9
Dental	Decayed/Missing/Filled tooth surfaces	21.8	21.6
Electrocardiography	Heart rate	1.8	1.1
Electrocardiography	HRV, rms	-0.4	-0.8
Electrocardiography	HRV, sd	-2.7	-2.9
Electrocardiography	PR duration	-0.8	-0.3
Electrocardiography	QT duration	3.3	2.2
Lipids	High density lipoprotein	10.1	8.7
Lipids	Low density lipoprotein	-5.4	-5.1
Lipids	Total cholesterol	-4.7	-4.2
Lipids	Triglycerides	3.7	2.8
Pulmonary disease	FEV1/FVC Ratio	0.1	-0.4
Pulmonary disease	Forced expiratory volume (1s)	10.9	11.9
Pulmonary disease	Forced vital capacity	15.8	15.6

Table S2: Evaluation of the contribution of genetic analysis group to regression models using AIC Each of 22 traits was analyzed using an LMM in which the model included fixed effects for sex, age, recruitment center, sampling weight, and random effects for census block group, household, and polygenic effects due to relatedness. Models also included trait-specific covariates in some cases. In addition, each model included either the first five or the first twenty PCs as fixed effects to adjust for ancestry, either with or without genetic analysis group. The column "PC5-PC5g" is PC5 (AIC for the model with PC1-5 but no genetic analysis group) minus PC5g (AIC for the model with PC1-5 and genetic analysis group); similarly for the column labeled as "PC20 - PC20g", except using PCs 1-20. A positive difference indicates that the model with genetic analysis group is a better fit than the model without genetic analysis group.

		AIC difference
Trait group	Trait	PC5b - PC5g
Anthropometrics	log BMI	3.0
Blood cell count	Hematocrit	-3.0
Blood cell count	Hemoglobin	-2.1
Blood cell count	Platelet count	-4.2
Blood cell count	Red blood cell count	-1.3
Blood cell count	White blood cell count	5.2
Chronic kidney disease	eGFR (CKD)	0.5
Chronic kidney disease	eGFR (MDRD)	3.2
Dental	Decayed/Missing/Filled teeth	-10.2
Dental	Decayed/Missing/Filled tooth surfaces	-14.4
Electrocardiography	Heart rate	4.3
Electrocardiography	HRV, rms	2.5
Electrocardiography	HRV, sd	1.6
Electrocardiography	PR duration	0.8
Electrocardiography	QT duration	-6.8
Lipids	High density lipoprotein	0.4
Lipids	Low density lipoprotein	0.0
Lipids	Total cholesterol	0.8
Lipids	Triglycerides	1.1
Pulmonary disease	FEV1/FVC Ratio	0.3
Pulmonary disease	Forced expiratory volume (1s)	1.8
Pulmonary disease	Forced vital capacity	-3.0

Table S3: Evaluation of the contribution of genetic analysis group versus self-identified background group to regression models using AIC

Each of 22 traits was analyzed using an LMM in which the model included fixed effects for sex, recruitment center, age, sampling weight, and the first five PCs, and random effects for census block group, household, and polygenic effects due to relatedness. The model labeled as "PC5g" also had genetic analysis group as a fixed effect, while "PC5b" had self-identified background as a fixed effect. Models also included trait-specific covariates in some cases. Only participants with one of the six specific self-identified background groups were included in both analyses. A positive AIC difference indicates that the model with genetic analysis group is a better fit than the model with background group.

		CV	
Trait group	Trait	gengrp	bkgrd
Lipids	High density lipoprotein	0.06	0.07
Dental	Decayed/Missing/Filled teeth	0.06	0.06
Dental	Decayed/Missing/Filled tooth surfaces	0.07	0.06
Lipids	Triglycerides	0.07	0.06
Electrocardiography	QT duration	0.07	0.06
Electrocardiography	HRV, sd	0.08	0.08
Chronic kidney disease	eGFR (MDRD)	0.08	0.09
Lipids	Low density lipoprotein	0.08	0.09
Electrocardiography	HRV, rms	0.09	0.08
Electrocardiography	PR duration	0.09	0.09
Lipids	Total cholesterol	0.10	0.10
Pulmonary disease	Forced vital capacity	0.12	0.11
Chronic kidney disease	eGFR (CKD)	0.12	0.12
Blood cell count	Hematocrit	0.13	0.12
Blood cell count	Hemoglobin	0.14	0.15
Anthropometrics	log BMI	0.15	0.14
Blood cell count	White blood cell count	0.15	0.15
Blood cell count	Platelet count	0.15	0.16
Blood cell count	Red blood cell count	0.16	0.14
Electrocardiography	Heart rate	0.16	0.16
Pulmonary disease	Forced expiratory volume (1s)	0.17	0.17
Pulmonary disease	FEV1/FVC Ratio	0.27	0.27

Table S4: Comparison of residual variance heterogeneity for genetic analysis versus self-identified background groups

Coefficient of variation for residual variance by group is given for each of 22 biomedical traits. These statistics derive from LMMs assuming homoscedasticity and using either genetic analysis group ("gengrp") or self-identified background group ("bkgrd") as a fixed effect covariate. The sample set was the same for both models, and it included only individuals who had one of the six specific values for both genetic analysis group and self-identified background. Both models included fixed effects for sex, age, center, sampling weight, and PCs 1-5, and random effects for block group, household, and polygenic effects due to relatedness. In some cases, trait-specific fixed effect covariates were also included in both models.

		λ_{GC}	
Trait group	Trait	homoscedastic	heteroscedastic
Dental	Decayed/Missing/Filled teeth	0.996	1.006
Chronic kidney disease	eGFR (MDRD)	0.999	1.013
Lipids	Triglycerides	1.001	1.009
Electrocardiography	HRV, sd	1.019	1.010
Dental	Decayed/Missing/Filled tooth surfaces	1.019	1.013
Electrocardiography	QT duration	1.021	1.016
Blood cell count	Hematocrit	1.022	1.013
Electrocardiography	HRV, rms	1.025	1.016
Blood cell count	Hemoglobin	1.029	1.020
Lipids	Total cholesterol	1.031	1.014
Lipids	High density lipoprotein	1.032	1.025
Pulmonary disease	Forced vital capacity	1.035	1.031
Electrocardiography	PR duration	1.041	1.024
Pulmonary disease	Forced expiratory volume (1s)	1.042	1.026
Electrocardiography	Heart rate	1.044	1.020
Blood cell count	Platelet count	1.045	1.026
Lipids	Low density lipoprotein	1.046	1.027
Chronic kidney disease	eGFR (CKD)	1.047	1.029
Anthropometrics	log BMI	1.049	1.041
Blood cell count	White blood cell count	1.056	1.027
Blood cell count	Red blood cell count	1.058	1.030
Pulmonary disease	FEV1/FVC Ratio	1.072	1.027

Table S5: Genomic inflation (λ_{GC}) for models with homoscedasticity and heteroscedasticity
Each of 22 traits was analyzed using an LMM in which the model included fixed effects for sex, recruitment center, age, sampling weight, genetic analysis group, and the first five PCs, and random effects for census block group, household, and polygenic effects due to relatedness. The model labeled "homoscedastic" fit one residual variance for all subjects, while the model labeled "heteroscedastic" fit a different residual variance for each genetic analysis group. Models also included trait-specific covariates in some cases. The data in this table are plotted in Figure 8B.

		λ_{GC}		
Trait group	Trait	gengrp	background	no group
Lipids	Triglycerides	1.009	1.012	1.001
Blood cell count	Hematocrit	1.013	1.018	1.022
Lipids	Total cholesterol	1.014	1.014	1.031
Electrocardiography	QT duration	1.016	1.016	1.022
Blood cell count	Hemoglobin	1.020	1.023	1.030
Lipids	High density lipoprotein	1.025	1.025	1.032
Blood cell count	Platelet count	1.026	1.026	1.045
Blood cell count	White blood cell count	1.027	1.028	1.057
Lipids	Low density lipoprotein	1.027	1.026	1.045
Pulmonary disease	FEV1/FVC Ratio	1.027	1.026	1.073
Blood cell count	Red blood cell count	1.030	1.035	1.058
Anthropometrics	log BMI	1.041	1.042	1.050

Table S6: Comparison of λ_{GC} for models using different group definitions
The table shows λ_{GC} for models with three different group definitions. All GWAS models were adjusted for sex, center, age, sampling weight, and PCs 1-5, and trait-specific covariates as fixed effects, and random effects for block group, household, and genetic relatedness were included. In addition, both models with group variables were run using heterogeneous residual variance. The "no group" model uses the full set of 12,784 subjects, minus any trait-specific exclusions. The "gengrp" model includes genetic analysis group as a fixed effect and excludes 37 subjects that have missing genetic analysis group. The "background" model includes self-reported background as a fixed effect, and excludes 425 subjects with missing self-reported background.

Trait group	Trait	N hits	slope [95% CI] background	slope [95% CI] no group	Refs
Blood cell count	White blood cell count	14	0.925 [0.909-0.941]	1.059 [1.044-1.073]	[1–4]
Lipids	High density lipoprotein	72	0.931 [0.922-0.940]	0.989 [0.987-0.991]	[5–7]
Anthropometrics	log BMI	104	0.936 [0.913-0.958]	0.988 [0.975-1.001]	[8–12]
Lipids	Triglycerides	40	0.951 [0.945-0.957]	1.020 [1.017-1.023]	[6,7]
Pulmonary disease	FEV1/FVC Ratio	21	0.960 [0.925-0.995]	0.946 [0.893-1.000]	[13–15]
Blood cell count	Hemoglobin	19	0.965 [0.944-0.986]	1.023 [1.015-1.031]	[2,16–18]
Blood cell count	Hematocrit	10	0.966 [0.904-1.028]	1.047 [1.017-1.077]	[2,17,19]
Blood cell count	Platelet count	59	0.969 [0.948-0.990]	0.970 [0.960-0.979]	[2,19–22]
Electrocardiography	QT duration	35	0.981 [0.963-0.999]	0.976 [0.970-0.982]	[23]
Lipids	Total cholesterol	74	0.989 [0.981-0.997]	0.976 [0.972-0.981]	[6,7]
Blood cell count	Red blood cell count	18	0.995 [0.963-1.027]	1.009 [0.992-1.026]	[2,17,18,24]
Lipids	Low density lipoprotein	58	0.999 [0.991-1.008]	0.976 [0.972-0.980]	[6,7]
	mean		0.964 [0.950-0.978]	0.998 [0.979-1.017]	

Table S7: Comparison of Wald test statistics for the effects of SNPs with previously published trait associations, using models with different ethnic group definitions.
For 12 biomedical traits, association tests were performed in HCHS/SOL to assess power to detect known hits from the literature for models utilizing genetic analysis group, self-identified background group, or no group variable. See Figure 10 for plots of these data. All models included adjustment for sex, center, age, sampling weight, PCs 1-5, and trait-specific covariates. Random effects included block group, household, and genetic relatedness. The models also included genetic analysis group ("gengrp", up to 12,747 subjects); self-reported background ("background", up to 12,359 subjects); or no group variable ("no group", up to 12,784 subjects) as a covariate. For models using a group variable ("gengrp" and "background"), heterogeneous residual variance was fit by that group. The slopes and 95% confidence intervals are from a linear regression (through the origin) of test statistics for a given trait from either the background or no-group model, regressed on the corresponding test statistics from the genetic analysis group model. See Figure 10A for an example using log BMI. All test statistics are from the Wald test for the effect of the published SNP hit ($\chi^2_{(1)}$) in HCHS/SOL data and were adjusted for λ_{GC} before fitting the linear model. The number of previously identified variants affecting each trait ("N hits") is shown. The "mean" row gives the mean of the slope estimates for each model type and its 95% confidence interval. References for papers used to determine known hits for each trait are shown in the "Refs" column.

References

[1] Crosslin, D. R., McDavid, A., Weston, N., Nelson, S. C., Zheng, X., Hart, E., de Andrade, M., Kullo, I. J., McCarty, C. A., Doheny, K. F., et al. (2012). Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. Hum Genet *131*, 639–652.

[2] Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y., and Kamatani, N. (2010). Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat Genet *42*, 210–215.

[3] Nalls, M. A., Couper, D. J., Tanaka, T., van Rooij, F. J. A., Chen, M.-H., Smith, A. V., Toniolo, D., Zakai, N. A., Yang, Q., Greinacher, A., et al. (2011). Multiple loci are associated with white blood cell phenotypes. PLoS Genet *7*, e1002113.

[4] Reiner, A. P., Lettre, G., Nalls, M. A., Ganesh, S. K., Mathias, R., Austin, M. A., Dean, E., Arepalli, S., Britton, A., Chen, Z., et al. (2011). Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). PLoS Genet *7*, e1002108.

[5] Elbers, C. C., Guo, Y., Tragante, V., Van Iperen, E. P., Lanktree, M. B., Castillo, B. A., Chen, F., Yanek, L. R., Wojczynski, M. K., Li, Y. R., et al. (2012). Gene-centric meta-analysis of lipid traits in African, East Asian and Hispanic populations. Plos One *7*.

[6] Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713.

[7] Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. Nat Genet *45*, 1274–1283.

[8] Berndt, S. I., Gustafsson, S., Magi, R., Ganna, A., Wheeler, E., Feitosa, M. F., Justice, A. E., Monda, K. L., Croteau-Chonka, D. C., Day, F. R., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat Genet *45*, 501–512.

[9] Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206.

[10] Monda, K. L., Chen, G. K., Taylor, K. C., Palmer, C., Edwards, T. L., Lange, L. A., Ng, M. C. Y., Adeyemo, A. A., Allison, M. A., Bielak, L. F., et al. (2013). A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. Nat Genet *45*, 690–696.

[11] Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Lango Allen, H., Lindgren, C. M., Luan, J., Magi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet *42*, 937–948.

[12] Wen, W., Cho, Y.-S., Zheng, W., Dorajoo, R., Kato, N., Qi, L., Chen, C.-H., Delahanty, R. J., Okada, Y., Tabara, Y., et al. (2012). Meta-analysis identifies common variants associated with body mass index in east Asians. Nat Genet *44*, 307–311.

[13] Hancock, D. B., Eijgelsheim, M., Wilk, J. B., Gharib, S. A., Loehr, L. R., Marciante, K. D., Franceschini, N., van Durme, Y. M. T. A., Chen, T.-H., Barr, R. G., et al. (2010). Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. Nat Genet *42*, 45–52.

[14] Repapi, E., Sayers, I., Wain, L. V., Burton, P. R., Johnson, T., Obeidat, M., Zhao, J. H., Ramasamy, A., Zhai, G., Vitart, V., et al. (2010). Genome-wide association study identifies five loci associated with lung function. Nat Genet *42*, 36–44.

[15] Soler Artigas, M., Loth, D. W., Wain, L. V., Gharib, S. A., Obeidat, M., Tang, W., Zhai, G., Zhao, J. H., Smith, A. V., Huffman, J. E., et al. (2011). Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nat Genet *43*, 1082–1090.

[16] Chambers, J. C., Zhang, W., Li, Y., Sehmi, J., Wass, M. N., Zabaneh, D., Hoggart, C., Bayele, H., McCarthy, M. I., Peltonen, L., et al. (2009). Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. Nat Genet *41*, 1170–1172.

[17] Ganesh, S. K., Zakai, N. A., van Rooij, F. J. A., Soranzo, N., Smith, A. V., Nalls, M. A., Chen, M.-H., Kottgen, A., Glazer, N. L., Dehghan, A., et al. (2009). Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat Genet *41*, 1191–1198.

[18] van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D. S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature *492*, 369–375.

[19] Li, J., Glessner, J. T., Zhang, H., Hou, C., Wei, Z., Bradfield, J. P., Mentch, F. D., Guo, Y., Kim, C., Xia, Q., et al. (2013). GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. Hum Mol Genet *22*, 1457–1464.

[20] Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A. H., Labrune, Y., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. Nature *480*, 201–208.

[21] Qayyum, R., Snively, B. M., Ziv, E., Nalls, M. A., Liu, Y., Tang, W., Yanek, L. R., Lange, L., Evans, M. K., Ganesh, S., et al. (2012). A meta-analysis and genome-wide association study of platelet count and mean platelet volume in African Americans. PLoS Genet *8*, e1002491.

[22] Shameer, K., Denny, J. C., Ding, K., Jouni, H., Crosslin, D. R., de Andrade, M., Chute, C. G., Peissig, P., Pacheco, J. A., Li, R., et al. (2014). A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. Hum Genet *133*, 95–109.

[23] Arking, D. E., Pulit, S. L., Crotti, L., van der Harst, P., Munroe, P. B., Koopmann, T. T., Sotoodehnia, N., Rossin, E. J., Morley, M., Wang, X., et al. (2014). Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. Nat Genet *46*, 826–836.

[24] Soranzo, N., Spector, T. D., Mangino, M., Kuhnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat Genet *41*, 1182–1190.