Title

Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort.

Journal

Genetics, 200(4)

ISSN

0016-6731

Authors

Banda, Yambazi
Kvale, Mark N
Hoffmann, Thomas J
et al.

Publication Date

2015-08-01

DOI

10.1534/genetics.115.178616

Peer reviewed

Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort

Yambazi Banda,*,[1] Mark N. Kvale,* Thomas J. Hoffmann,*,[†] Stephanie E. Hesselson,* Dilrini Ranatunga,[‡] Hua Tang,[§] Chiara Sabatti, Lisa A. Croen,[‡] Brad P. Dispensa,* Mary Henderson,[‡] Carlos Iribarren,[‡] Eric Jorgenson,[‡] Lawrence H. Kushi,[‡] Dana Ludwig,[‡] Diane Olberg,[‡] Charles P. Quesenberry Jr.,[‡] Sarah Rowell,[‡] Marianne Sadler,[‡] Lori C. Sakoda,[‡] Stanley Sciortino,[‡] Ling Shen,[‡] David Smethurst,[‡] Carol P. Somkin,[‡] Stephen K. Van Den Eeden,[‡] Lawrence Walter,[‡] Rachel A. Whitmer,[‡] Pui-Yan Kwok,* Catherine Schaefer,[‡,1,2] and Neil Risch*,[†,‡,1,2]**

*Institute for Human Genetics, University of California, San Francisco, California 94143-0794, [†]Department of Epidemiology and Biostatistics, University of California, San Francisco, California 94158-2549, [‡]Kaiser Permanente Northern California Division of Research, Oakland, California 94612-2304, [§]Department of Genetics, Stanford University, Stanford, California 94305-5120, and **Department of Health Research and Policy, Stanford University, Stanford, California 94305-5405

ABSTRACT Using genome-wide genotypes, we characterized the genetic structure of 103,006 participants in the Kaiser Permanente Northern California multi-ethnic Genetic Epidemiology Research on Adult Health and Aging Cohort and analyzed the relationship to self-reported race/ethnicity. Participants endorsed any of 23 race/ethnicity/nationality categories, which were collapsed into seven major race/ethnicity groups. By self-report the cohort is 80.8% white and 19.2% minority; 93.8% endorsed a single race/ethnicity group, while 6.2% endorsed two or more. Principal component (PC) and admixture analyses were generally consistent with prior studies. Approximately 17% of subjects had genetic ancestry from more than one continent, and 12% were genetically admixed, considering only nonadjacent geographical origins. Self-reported whites were spread on a continuum along the first two PCs, indicating extensive mixing among European nationalities. Self-identified East Asian nationalities correlated with genetic clustering, consistent with extensive endogamy. Individuals of mixed East Asian–European genetic ancestry were easily identified; we also observed a modest amount of European genetic ancestry in individuals self-identified as Filipinos. Self-reported African Americans and Latinos showed extensive European and African genetic ancestry, and Native American genetic ancestry for the latter. Among 3741 genetically identified parent–child pairs, 93% were concordant for self-reported race/ethnicity; among 2018 genetically identified full-sib pairs, 96% were concordant; the lower rate for parent–child pairs was largely due to intermarriage. The parent–child pairs revealed a trend toward increasing exogamy over time; the presence in the cohort of individuals endorsing multiple race/ethnicity categories creates interesting challenges and future opportunities for genetic epidemiologic studies.

doi: 10.1534/genetics.115.178616
Manuscript received January 16, 2015; accepted for publication June 2, 2015; published Early Online June 19, 2015.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.178616/-/DC1.
The accession number for the Research Program on Genes, Environment, and Health Parent Level/Study on the database of Genotypes and Phenotypes (dbGaP) is phs00078. The title of the dbGaP Parent study is "The Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH)." GO Project IRB: CN-09CScha-06-H.
[1]Corresponding authors: University of California San Francisco, Institute for Human Genetics, 513 Parnassus Ave., San Francisco, CA 94143-0794. E-mail: banday@humgen.ucsf.edu; University of California San Francisco, Institute for Human Genetics, 513 Parnassus Ave., San Francisco, CA 94143-0794. E-mail: rischn@humgen.ucsf.edu; Kaiser Permanente Northern California, Division of Research, 2000 Broadway, Oakland, CA 94612-2304. E-mail: cathy.schaefer@kp.org
[2]These authors contributed equally to this work.

KEYWORDS RPGEH GERA; population structure; principal components; admixture; race/ethnicity

POPULATION genetic structure analyses have recently increased in number due to improvements in capabilities to perform large-scale genomic investigations. Technological developments have improved our ability to address questions associated with phenotypic variation (Wellcome Trust Case Consortium 2007), human genetic variation (Jakobsson *et al.* 2008; Li *et al.* 2008), and evolution (Lohmueller *et al.* 2008). These studies play an important role in a variety of applied settings, including genome-wide association studies (GWAS), admixture analyses, and dissection of traits associated with ancestry. For example, in association studies, error rates due

to confounding by ancestry can be improved when population structure is taken into account (Tian *et al.* 2008a). At the same time, the relationship between self-identified race/ethnicity/nationality and genetic ancestry based on genetic marker data has become a topic of great interest (Risch *et al.* 2002; Burchard *et al.* 2003; Cooper *et al.* 2003).

Studies of human evolution have typically focused on indigenous population samples broadly distributed geographically across the globe. One such resource that has been highly exploited for this purpose is the Human Genome Diversity Project panel of 55 indigenous populations (Jakobsson *et al.* 2008; Li *et al.* 2008). On the other hand, GWAS utilizing U.S.-based samples often include more heterogeneous populations in terms of ancestry, although the number of ethnic groups included is typically limited.

In the present study we utilize the large, ethnically diverse Kaiser Permanente (KP) Research Program on Genes, Environment, and Health (RPGEH) Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort to examine the question of genetic ancestry in a representative northern California population and how it relates to racial/ethnic self-identification. The cohort consists of 103,006 adult members of Kaiser Permanente Northern California (KPNC), ranging in age from 18 to 100 years at enrollment. The cohort was created to enable studies of genetic and environmental influences on many different health conditions and traits by linking high-density genome-wide SNP data with comprehensive longitudinal clinical information from electronic health records (EHR) as well as self-reported data on demographic factors and health behaviors from a structured survey. The GERA cohort is one of the first very large multi-ethnic cohorts created for GWAS for a wide variety of health conditions. The cohort was genotyped using custom ancestry-specific SNP arrays to better capture rare variants specific to different ethnic groups and provide better genome-wide coverage, thus permitting investigation of potential associations that may differ between groups. Understanding and characterizing the genetic diversity within a sample is essential to GWAS, since population structure both within and between groups can lead to artifactual associations. The multi-ethnic GERA cohort thus provides an unprecedented opportunity to understand human genetic diversity in a U.S. population sample. This article presents the results of analyses of population genetic structure, confirming previous observations, but also adding further understanding of mixed genetic ancestry, including the extent of distant *vs.* recent admixture. We also provide estimates of principal components needed for adjustment of population structure in GWAS and examine the self-reported race/ethnicity distribution of first-degree relative (parent–child and full sib) and monozygotic (MZ) twin pairs. Finally, we examine how the identified genetic structure correlates with participants' self-identification in terms of race/ethnicity/nationality.

Materials and Methods

Participants

Individuals composing the GERA cohort are participants in the KPNC RPGEH. KPNC is an integrated health care delivery system with >3 million members in northern California. The membership is representative of the general population with respect to race/ethnicity and socioeconomic status, although extremes of income are under-represented (Krieger *et al.* 1993). The RPGEH was established as a resource for research on genetic and environmental influences on health and disease. The development of the RPGEH and GERA cohort are described elsewhere (dbGaP phs000674.v1.p1). Briefly, adult members of KPNC were asked to complete a mailed survey; survey respondents then completed a broad written consent and provided a saliva sample for extraction of DNA. Participants self-reported their race, ethnicity, and nationality on the survey by endorsing as many of 23 race, ethnicity, and nationality categories as applied (Table 1 provides a list of the choices). Participants were asked their religion, and this question in conjunction with a race/ethnicity/nationality question was used to identify Ashkenazi individuals (those who responded "Ashkenazi Jewish" to the nationality question or "Jewish" to the religion question).

To maximize the diversity of the sample, the GERA cohort was formed by including all racial and ethnic minority participants with saliva samples (19% of the total); the remaining participants were drawn randomly from white non-Hispanic participants (81% of the total). Among cohort members, the average length of KPNC health plan membership was >23 years, providing extensive longitudinal data on diagnoses and procedures, laboratory test results, pharmaceutical prescriptions, radiological findings, and other clinical information from EHR for use in GWAS of health conditions and traits.

The Human Genome Diversity Project (HGDP) (Cavalli-Sforza 2005; Li *et al.* 2008) subjects were used to facilitate geographic interpretation of the GERA principal components.

Self-reported race/ethnicity

Self-reported race/ethnicity for each individual was derived from responses to the survey question on race/ethnicity/nationality (Table 1). Nationalities within a single race/ethnicity group were collapsed. Specifically, all East Asian nationalities (codes 10–15) were collapsed into a single East Asian group; all Pacific Islander nationalities (codes 16–18) were collapsed into a single Pacific Islander group; all Latino nationalities (codes 4–8) were collapsed into a single Latino category; all African descent populations (codes 1–3) were collapsed into a single group; all white-European ethnicities (codes 20–22) were collapsed into a single category; the single categories of South Asians and Native Americans remained as such. A small number of individuals (<1%) had implausible race/ethnicity responses from the survey (*e.g.*, checked off every category) or specified "other." For

Table 1 Distribution of responses to survey question on race/ethnicity/nationality along with proportion female and average ages

	Category	No.	% female	Mean age (SE)
1	African American	3,117	0.57	60.66 (0.24)
2	African	129	0.43	52.90 (1.43)
3	Afro-Caribbean	119	0.68	56.24 (1.30)
4	Mexican	4,613	0.56	56.67 (0.22)
5	Central-South American	1,034	0.70	55.34 (0.46)
6	Puerto Rican	322	0.69	56.68 (0.83)
7	Cuban	106	0.71	55.41 (1.42)
8	Other Latino/Hispanic	1,545	0.70	57.41 (0.38)
9	South Asian–Indian/Pakistani	575	0.42	54.58 (0.60)
10	Chinese	3,433	0.58	56.75 (0.25)
11	Japanese	1,739	0.61	61.56 (0.34)
12	Korean	234	0.66	53.83 (1.04)
13	Filipino	1,708	0.59	55.59 (0.37)
14	Vietnamese	317	0.50	53.23 (0.82)
15	Other Southeast Asia	176	0.64	51.85 (1.10)
16	Native Hawaiian	144	0.65	58.41 (1.23)
17	Samoan	14	0.64	59.36 (3.44)
18	Other Pacific Islander	132	0.57	53.88 (1.35)
19	Native American Indian/Alaska Native	3,884	0.66	61.20 (0.22)
20	White European American	80,079	0.59	63.27 (0.05)
21	Middle Easterner	914	0.43	62.18 (0.48)
22	Ashkenazi Jewish	2,399	0.66	62.49 (0.28)
23	Other ethnicity	75	0.73	56.53 (1.64)

these individuals, we used KPNC administrative databases to assign race/ethnicity. For other individuals, a discrepancy was observed between their original and scanned survey responses. These subjects were also adjudicated to their original form results as described in *SI Methods* (File S1).

Genotyping and array assignment

To maximize genome-wide coverage of common and less common variants, four custom Affymetrix Axiom arrays (Hoffmann *et al.* 2011a,b) were designed for individuals of non-Hispanic white (EUR), East Asian (EAS), African American (AFR), and Latino (LAT) race/ethnicity. The number of SNPs varied among arrays, ranging from 674,518 on the EUR array to 893,631 on the AFR array (Hoffmann *et al.* 2011b). A total of 254,438 SNPs were common to all four arrays. Genotyping was performed at the University of California, San Francisco and is described elsewhere (Kvale *et al.* 2015).

The assignment of subjects to arrays was based on the race/ethnicity categories formed as described above (Table S2 and Table S3). Assignments were hierarchical to accommodate individuals reporting multiple racial/ethnic categories. Specifically, individuals reporting any Latino or Native American race/ethnicity/nationality (possibly in combination with other races/ethnicities/nationalities) were assigned to the LAT array, with the exception of individuals who reported African/African American race/ethnicity and Native American race/ethnicity, who were assigned to the AFR array, and individuals reporting East Asian race/ethnicity and Native American race/ethnicity, who were assigned to the EAS array. All other individuals reporting any African, African American, or Afro-Caribbean race/ethnicity but no Latino race/ethnicity were assigned to the AFR array. All those reporting any East Asian but not African, African American, Afro-Caribbean, or Latino race/ethnicity were assigned to the EAS array. Subjects reporting white-European American, South Asian, Middle Eastern, or Ashkenazi race/ethnicity, but none of the previously mentioned races/ethnicities, were assigned to the EUR array. Therefore, for example, individuals with European and East Asian race/ethnicity were assigned to the EAS array; individuals with African American and East Asian race/ethnicity/nationality were analyzed on the AFR array. The various arrays were designed to allow for the relevant admixture (Hoffmann *et al.* 2011b).

Quality control

High-quality genotype data for the GERA cohort was obtained by systematic examination and removal of SNP genotypes according to a specific protocol, as described in detail elsewhere (Kvale *et al.* 2015). For the genetic structure analyses, only SNPs that were common across all four arrays and that had a call rate >99.5% were considered. This set also excluded SNPs that showed extreme deviation from Hardy–Weinberg equilibrium ($P < 10^{-5}$). This resulted in a set of 144,799 high-performing SNPs used in further analyses of population structure and admixture.

Principal components analysis

Filtering: Principal components analysis (PCA) was performed using the *smartpca* program, which is part of the EIGENSOFT4.2 software package (Patterson *et al.* 2006). The initial PCA runs were performed separately for individuals genotyped on different arrays. The initial set of 144,799 high-performing SNPs (described above) that were common across all four array types was used in the preliminary

analyses. When the HGDP samples were included in subsequent runs and projected onto the GERA principal components (PCs) to facilitate geographic interpretation, 43,988 high-performing SNPs were used. Initial analyses revealed that a number of individuals appeared to be discordant between their genetic ancestry and the array to which they were assigned, and the PCA was re-run after reclassifying these individuals (see *SI Methods*, File S1).

PC projection approach: PCA requires the inversion of a data matrix, which for very large data sets may be computationally challenging. For the East Asian, African American, and Latino subgroups in the GERA data set, the sample sizes were small enough so that all subjects within each subgroup were run together. For example, all 7520 East Asian subjects were run together in one PCA. The white-European American sample, however, is very large and required inverting an ~80,000 by 80,000 (6.4 billion elements) matrix. Furthermore, the version of the *Smartpca* program used at the time of analyses was not able to analyze the entire European ancestry sample of >83,000 individuals. Therefore, our approach was to select a large but manageable number of subjects on which to perform an initial PCA and then use the resulting SNP loadings to project the remaining subjects.

Because we planned to select a random subset of 20,000 individuals for the initial PCA on which the remaining subjects would be projected, we examined the effect of using different subsets by calculating the correlations of the SNP loadings for three different random subsets (Supporting Information, Table S1). The numbers of subjects in the three subsets were the following: 18,677 for set 1; 20,121 for set 2; and 17,691 for set 3. For the first six PCs there was very good correlation of the SNP loadings for all three pairs of subsets, also suggesting that most of the signal regarding genetic structure is derived from the first six PCs. Given these results, we selected a random set of 20,000 European ancestry subjects and projected the remaining subjects onto the PCs obtained.

Since the SNPs used for the PCA and admixture estimation were common among all four genotyping arrays it was possible to produce "global" PCA scores for the GERA subjects. Subsets of individuals from the EUR (15,500) AFR (3100), EAS (5600), and LAT (3000) arrays were used for the initial PCA, and the remaining subjects were projected onto these PCs to obtain PC scores for each individual. Table S4 shows the number of SNPs remaining after LD and structural variation loci pruning for each of the eight different PCA runs (File S1).

Genetic ancestry/admixture estimation

To determine individual ancestral admixture proportions in admixed subjects such as African Americans and Latinos (and others), the full maximum-likelihood software package *frappe* (Tang *et al.* 2005) was used. In this analysis, individual ancestry proportions are estimated by calculating the probability of a set of genome-wide genotypes in an individual as a weighted average of allele frequencies of putative ancestors, where the weights represent the admixture proportions. In general, the same HGDP population samples described above were used to derive allele frequencies for the ancestral groups.

Relationship determination

Relationships were determined using the software KING_v1.4 (Manichaikul *et al.* 2010) with the robust version that allows for population substructure. KING provides standard thresholds for characterizing monozygotic twin, parent–child, and sibling relationships, which we followed. In our data, these relationships were clearly separated into distinct clusters. All subjects were included irrespective of the array type used for their analysis. This analysis was based on the 144,799 high-performing SNPs common across the four arrays described above.

Results

Distribution of race/ethnicity/nationality categories reported

This multi-ethnic cohort includes representation from a broad distribution of races/ethnicities/nationalities (Table 1). For individuals who reported more than one category, all categories are included; hence, the numbers in Table 1 sum to greater than 103,006, the total cohort size. All of the major continents are represented and many nationalities/ethnicities. Collapsing the selections into race/ethnicity categories (see *Materials and Methods*), of the 106,733 total selections, 3365 (3.2%) include an African/African American race/ethnicity; 7620 (7.1%) include a Latino race/ethnicity; 575 (0.5%) include South Asian race/ethnicity; 7607 (7.1%) include an East Asian race/ethnicity; 290 include a Pacific Islander race/ethnicity (0.3%); 3884 (3.6%) include Native American race/ethnicity; and 83,392 (78.1%) include a white-European race/ethnicity. The majority of those endorsing a Latino race/ethnicity are Mexican and Central American, while the largest groups endorsing an East Asian race/ethnicity are Chinese, Japanese, and Filipino. We also examined the sex and age distributions across the different categories (Table 1). Compared to those reporting white-European race/ethnicity, those endorsing African/Afro-Caribbean, Latino, East Asian, and Pacific Islander race/ethnicity are younger; with the exception of those reporting Mexican nationality, the Latino groups tend to have a higher proportion of females, as do those reporting Ashkenazi Jewish ethnicity; those reporting South Asian and Middle Eastern nationalities have a lower proportion of females.

Structure of individuals run on the EUR array

Individuals who self-reported Ashkenazi, Middle Eastern, and non-Hispanic white or European race/ethnicity but no other ethnicities were run on the EUR array and analyzed together. The initial analysis showed, as expected, a clear Ashkenazi cluster and a larger cluster depicting the

northwest–southeast European cline (Price *et al.* 2008; Tian *et al.* 2008c). Figure S1A shows those who self-reported a single ethnicity/nationality, while Figure S1B shows individuals who self-reported more than one. It is evident that endorsement of more than one ethnicity can imply mixed genetic ancestry but not automatically. Comparing Figure S1, A and B, we observe a higher proportion of individuals with mixed genetic ancestry among those who endorsed both Ashkenazi and European or Middle Eastern ethnicity; however, we still observe a large proportion of nonadmixed individuals, suggesting that endorsement of Ashkenazi and European may reflect a joint perception of ethnicity and continent of origin. By contrast, in Figure S1A we observe a substantial number of individuals who appear to have Ashkenazi and European admixture but self-reported a single category only (most often European).

A similar observation can be made about those endorsing Middle Eastern ethnicity, where those endorsing that as a sole response appear to have more Middle Eastern genetic ancestry, while those endorsing Middle Eastern and European ethnicity show more evidence of European genetic ancestry. However, in Figure S1A we also observe substantial numbers of individuals reporting only European ethnicity whose genetic ancestry appears to be Middle Eastern and vice versa. Again, these reports may reflect recent geographic origin as well as nationality/ethnicity.

We also repeated the PC analysis after removing the Ashkenazi and part-Ashkenazi subjects. The PC scores for the Ashkenazi subjects were then derived by projecting their genotypes onto the resulting PCs. Individuals reporting a single ethnicity/nationality are depicted in Figure S2A, while those endorsing more than one are displayed in Figure S2B. The first PC corresponds to a northwest–southeast cline through Europe and the Middle East and the second PC corresponds to a southwest–northeast cline within Europe, as has been observed in numerous previous studies (Menozzi *et al.* 1978; Sokal *et al.* 1991; Cavalli-Sforza *et al.* 1993; Cavalli-Sforza *et al.* 1996; Barbujani and Bertorelle 2001; Belle *et al.* 2006; Seldin *et al.* 2006; Bauchet *et al.* 2007; Novembre *et al.* 2008; Price *et al.* 2008; Tian *et al.* 2008c). The first and second PCs account for 31.9 and 13.4% of the total variance of the first 10 PCs, respectively.

Subjects who self-identified as South Asian were also run on the EUR array and subjected to a separate PCA. For these subjects, to characterize the observed PCs and the relationship to geographic ancestry, we employed onomastics. In particular, we analyzed surnames to characterize individuals based on surname geographic region of origin. These subjects are mainly of Indian origin, and the clusters formed in the PCA depict subgroups from different regions of India (Figure S3). The first PC accounts for 19.1% of the total variance of the first 10 PCs and the second PC accounts for 10.0%. The analysis also shows that northern Indians are genetically closer to Europeans (Reich *et al.* 2009) and eastern Indians are genetically more similar to East Asian populations. As expected, those reporting European as well as South Asian ethnicity are positioned closer in the diagram to the HGDP Europeans.

Structure of individuals run on the EAS array

Individuals run on the EAS array included subjects self-reporting European and East Asian race/ethnicity and those reporting solely East Asian race/ethnicity. The first PC for these individuals (Figure S4, A and B) is responsible for clustering of individuals with different East Asian–European ancestry proportions (mostly 50 or 75% European). Those with genetic ancestry that is both East Asian and European are most clearly observed in Figure S4B, among those self-reporting both races/ethnicities, and there are very few GERA individuals in this figure who do not have mixed genetic ancestry. Among individuals reporting only an East Asian nationality (Figure S4A), the large majority have only East Asian genetic ancestry; however, there are also individuals who appear to have mixed East Asian–European genetic ancestry who self-reported only their East Asian nationality. Of particular interest is the continuous nature of a modest amount of European genetic ancestry in self-identified Filipinos, consistent with older European admixture. The second PC corresponds to the north-to-south cline in East Asia (Su *et al.* 1999; Tian *et al.* 2008b; Hugo Pan-Asian Snp Consortium 2009), and the distinct clusters observed that represent different East Asian nationalities are consistent with extensive endogamy in these groups. The first and second PCs account for 59.71 and 20.39% of the total variance of the first 10 PCs, respectively.

Individuals endorsing a Pacific Islander ethnicity are displayed in Figure S5. Those also reporting an East Asian ethnicity appear to cluster more closely to the HGDP East Asians, while those also reporting European ethnicity appear to cluster more closely to the HGDP Europeans. While those reporting Hawaiian and Samoan ethnicity are reasonably well separated from both the HGDP Europeans and East Asians, some individuals who identified as "other Pacific Islander" appear to overlap quite closely with the HGDP East Asians. Also of interest, another subgroup of "other Pacific Islanders" appears to form its own cluster at the bottom of Figure S5. We note that a number of these individuals self-reported both Pacific Islander and South Asian ethnicity. Based on onomastics, these individuals have Indian surnames and are likely to be Indo Fijians. Approximately 37.5% of the population of Fiji is of Indian origin, according to the 2007 census (http://www.statsfiji.gov.fj). The observation that some Pacific Islanders cluster near to the East Asians is also an indication that clear separation of genetic ancestry for these groups is likely to be challenging.

Structure of individuals run on the AFR array

Subjects run on the AFR array revealed, as expected, extensive African and European genetic ancestry (Figure S6, A and B) (Parra *et al.* 1998; Fernandez *et al.* 2003; Tang *et al.* 2006; Tishkoff *et al.* 2009; Zakharia *et al.* 2009). The first PC, which accounts for 63.8% of the total variance of

the first 10 PCs, reflects African *vs.* European genetic ancestry, while the second PC denotes East Asian and/or Native American genetic ancestry. This is consistent with the array assignments, whereby individuals reporting both African/African American race/ethnicity and East Asian or Native American race/ethnicity were assigned to the AFR array. Individuals who self-reported African ancestry only were also subject to onomastics to determine likely countries of origin. We were able to identify subjects of Ethiopian, Eritrean, and Kenyan nationality. For the Kenyans, Figure S6A indicates a location consistent with 100% African genetic ancestry. By contrast, the Ethiopian/Eritrean subjects occupy an intermediate position on the PC1 axis, suggesting proximity to European/Middle Eastern populations. Also of note is the modest variation in their PC1 scores. This is likely due to ancient admixture with Middle Eastern populations (Hodgson *et al.* 2014). These results confirm that Ethiopians have a unique genetic structure among African populations.

Individuals self-reporting mixed African and East Asian race/ethnicity generally reflect that admixture from the genetic perspective as well (Figure S6B); however, a number of individuals who reported only African American ethnicity also appear to have similar levels of East Asian admixture (Figure S6A). Those reporting both African American and European ethnicity generally occupy a position on the PC1 axis closer to Europeans than those who do not (Figure S6B).

The mean African ancestry proportion in this sample is 73.6 ± 17.4%. There is a reasonably high level of variation in the African genetic ancestry proportion, ranging from 10.6 to 100%.

Structure of individuals run on the LAT array

Latinos may have ancestry deriving from multiple continents, including Europe, Africa, Asia, and the Americas (Bonilla *et al.* 2004; Tang *et al.* 2006, 2007). Figure S7A provides the PCA results for all those who endorsed Latino or Native American as their sole race/ethnicity. PC1 represents the European *vs.* Native American axis of genetic variation, and PC2 represents the African axis of genetic variation. PC1 and PC2 account for 70.95 and 11.57% of the total variance of the first 10 PCs, respectively. Nearly all Latinos show evidence of European/West Asian genetic ancestry, and a substantial subset also show evidence of African genetic ancestry. Similarly, all individuals self-reporting Native American race/ethnicity show some degree of European/West Asian genetic ancestry. Latinos of different nationalities exhibit varying proportions of European, African, and Native American ancestries (Figure S7B). Those reporting Mexican and Central-South American nationality have genetic ancestry that is primarily European and Native American, with slight but varying amounts of African ancestry. Those reporting Cuban nationality have primarily European genetic ancestry with a small number of individuals having primarily African genetic ancestry. Those reporting Puerto Rican nationality show some Native American genetic ancestry but are primarily admixed between European and African genetic ancestry. Individual ancestral admixture proportions were determined for these subjects and are provided in Table S5.

The LAT array also included a variety of individuals who self-reported more than one race/ethnicity. These individuals are represented in Figure S7C. Individuals who reported European as well as Latino race/ethnicity tend to have slightly more European genetic ancestry than those who did not; similarly, a number of individuals who reported African/African American race/ethnicity in addition to Latino race/ethnicity have substantial African genetic ancestry; however, many such individuals also appear to have the same modest degree of African genetic ancestry as those who reported only a Latino race/ethnicity. Those who reported Native American race/ethnicity in addition to Latino race/ethnicity also appear to have slightly increased Native American genetic ancestry. Those who reported European and Native American race/ethnicity appear to be similar to those who solely reported Native American race/ethnicity; all have European/West Asian genetic ancestry, and while some show evidence of Native American genetic ancestry, European/West Asian is the sole or primary genetic ancestry for the majority. For those with 100% European genetic ancestry and who self-reported only European and Native American race/ethnicity (n = 2155), we also calculated European PCs. Finally, those who reported East Asian in addition to Latino race/ethnicity generally have evidence of East Asian genetic ancestry (as observed in Figure S7C by proximity to the HGDP East Asians) ranging from 25 to 50 and 100%.

Global PCA for GERA subjects

Figure S8 shows that the first PC mainly separates Europeans from East Asians (and Native Americans) and PC2 separates Africans from all the other groups; PC3 seems to separate Native Americans from the other groups and PC4 also separates Native Americans from the other groups but also shows some separation among the Europeans; PC5 separates the different East Asian groups (mainly north *vs.* south) and also East Asians from Oceania, and PC6 separates Central–South Asians from the other groups; PC7 again separates the various East Asian regions, and PC8 separates the European groups (mainly north to south); PC9 and PC10 separate East Asians from Oceania, but also the Russians (not labeled) are separated from the other European groups.

Relationship between self-reported race/ethnicity and genetic ancestry

Table S6 displays the full relationship of self-reported race/ethnicity to genetic ancestry for the six continental genetic ancestries of Europe/West Asia, Africa, East Asia, Pacific Islands, South Asia, and the Americas. A genetic continental ancestry was assigned to an individual if her/his estimate for that ancestry was at least 5%. A total of 91,502 individuals (93.9%) reported a single race/ethnicity; 5475 individuals reported two races/ethnicities (5.9%); and 512

individuals (0.5%) reported three races/ethnicities (Table 2). As expected, all individuals who self-identified as European/West Asian had evidence of European/West Asian genetic ancestry. The next largest genetic ancestry component in this group was South Asian (4.3%), primarily attributable to individuals of West Asian ethnicity. Because there is a continuum of genetic ancestry from Europe to West Asia, Central-South Asia to East Asia, genetic overlap exists for individuals whose national origins are geographically between these divisions (Li *et al.* 2008). Nearly 1% of this group also had evidence of Native American genetic ancestry, while a smaller fraction had evidence of African or East Asian genetic ancestry (0.3 and 0.4%, respectively). Nearly all individuals (99.7%) self-reporting African/African American race/ethnicity had evidence of African genetic ancestry; 91% also had evidence of European genetic ancestry, consistent with broad European admixture among African Americans. Native American and East Asian genetic ancestry occurred in this group at a similar low level as observed in the Europeans/West Asians (1.3 and 0.5%, respectively). Among self-reported East Asians, all had evidence of East Asian genetic ancestry; a sizable proportion (21.7%) also had evidence of Pacific Islander genetic ancestry, but this likely represents difficulty in differentiating East Asian and Pacific Islander genetic ancestry. A modest subgroup (3.4%) had evidence of European/West Asian genetic ancestry (majority are self-reported Filipinos), while small proportions had evidence of African or Native American genetic ancestry (0.1 and 0.5%, respectively). Among the Latinos, nearly all had evidence of European/West Asian genetic ancestry; a similar high proportion (94.2%) had evidence of Native American genetic ancestry, and an additional 27.7% had evidence of African ancestry. A substantial number of self-reported Pacific Islanders had evidence of East Asian genetic ancestry (91.3%) in addition to Pacific Islander genetic ancestry (66.3%); these results are again likely due to close genetic similarity between East Asians and Pacific Islanders. There is also evidence of substantial European/West Asian and South Asian genetic ancestry in this group (57.6 and 26.1%, respectively). The former reflects a high rate of European admixture among some self-reported Pacific Islander groups, while the latter likely reflects Fijians of Indian origin. Most self-reported South Asians have evidence of South Asian genetic ancestry; a substantial proportion also has evidence of European or East Asian genetic ancestry, likely due to inability to cleanly separate South Asian genetic ancestry from West Asian or East Asian (Li *et al.* 2008). Among those reporting Native American race/ethnicity, 14.4% have evidence of Native American genetic ancestry, and all have evidence of European/West Asian genetic ancestry.

For those with missing or mis-scanned self-reported race/ethnicity, and whose race/ethnicity was derived from KP administrative databases (Table 3 and Table S7), results align closely with those in Table 2. For individuals self-reporting two or three races/ethnicities, the correspondence between the self-report and genetic ancestry is generally quite high (Table 2).

We also observed a decrease in average age and increasing proportion of females with the number of different race/ethnicity/ancestry groups reported (Table 2). While the different minority groups, and in particular the self-reported East Asians and Latinos, are younger on average, those reporting mixed race/ethnicity are even younger. These patterns likely reflect increasing exogamy over time. As expected, these patterns are also reflected in the genetic PC scores, where, for example, the proportion of mixed East Asian/European genetic ancestry increases with decreasing age. The excess of females among those reporting mixed race/ethnicity appears to reflect a reporting preference, as there was no significant difference in the proportion of individuals with mixed genetic ancestry by sex.

A more in-depth examination of the distribution of continental genetic ancestry for the various self-report race/ethnicity groups is provided in Table S8.

Relatives

We were able to clearly identify first-degree relative (parent–child and full sib) and MZ twin pairs and categorized them based on self-reported race/ethnicity (Figure S9 and Table S9). We also observed thousands of likely second- and third-degree relatives (Figure S9); however, the figure also indicates substantial overlap between these groups based on kinship estimates.

The 34 MZ pairs, who are perfectly concordant for genetic ancestry, are also perfectly concordant for self-reported race/ethnicity. Sib pairs are also (virtually) identical for genetic ancestry. We identified a total of 2018 sib pairs, 1936 (96%) of whom are concordant for self-reported race/ethnicity. Among the 82 discordant pairs, the majority (n = 66) involve pairs where one self-reports Native American or Latino race/ethnicity (solely or in combination with European/West Asian race/ethnicity) while the other reports only European/West Asian race ethnicity (Table S10); in most of these cases, the genetic ancestry is solely European/West Asian, although in some there is also evidence of Native American genetic ancestry. A modest number of pairs are also discordant in their reports of East Asian race/ethnicity, and, again for most of these, the genetic ancestry is solely European/West Asian. Similarly, a few pairs with mixed genetic ancestry including African are discordant in terms of self-reporting of African American race/ethnicity.

We identified 3741 parent–child pairs, of which 3478 (93%) were concordant for self-identified race/ethnicity. The lower rate of concordance compared to the sib pairs is not surprising as parent and child reports may differ if the child's parents are of different race/ethnicity. In 116 of 263 discordant pairs (Table S11), the child has genetic ancestry that her/his parent does not (Native American in 69 cases, East Asian in 41 cases, and African in 11 cases), and this difference is reflected in the self-report, where the child is self-reporting a race/ethnicity that the parent is not. By contrast, in only 9 cases did the parent have a genetic ancestry that the child did not, and in 8 of these 9 cases the parent has

Table 2 Proportion of individuals with genetic ancestry from each of six ancestral populations by self-reported race/ethnicity

		Genetic ancestry							
Race/ethnicity	*n*	EW	AF	EA	NA	PI	SA	% female	Mean age (SE)
One group	91,502							0.59	62.92 (0.04)
EW	76,401	1.000	0.003	0.004	0.009	0.000	0.043	0.59	63.71 (0.05)
AA	2,679	0.910	0.997	0.005	0.013	0.000	0.021	0.57	61.28 (0.25)
EA	6,389	0.034	0.001	1.000	0.005	0.217	0.008	0.58	58.51 (0.18)
NA	674	0.999	0.022	0.022	0.144	0.000	0.037	0.55	64.34 (0.51)
LT	4,807	0.999	0.277	0.008	0.942	0.000	0.024	0.58	57.92 (0.21)
PI	92	0.576	0.000	0.913	0.000	0.663	0.261	0.48	56.89 (1.49)
SA	460	0.307	0.007	0.109	0.004	0.050	0.961	0.39	54.29 (0.67)
Two groups	5,476							0.67	57.37 (0.19)
EW/AA	123	1.000	0.976	0.024	0.033	0.000	0.081	0.67	52.76 (1.50)
EW/EA	572	0.960	0.005	0.942	0.014	0.063	0.080	0.68	49.13 (0.65)
EW/NA	2,548	1.000	0.008	0.007	0.096	0.000	0.024	0.68	61.63 (0.26)
EW/LT	1,564	1.000	0.071	0.010	0.710	0.000	0.068	0.68	54.05 (0.38)
EW/PI	48	1.000	0.000	0.813	0.042	0.625	0.021	0.79	59.64 (2.00)
EW/SA	44	0.955	0.000	0.068	0.045	0.000	0.682	0.66	53.55 (2.26)
AA/EA	29	0.655	0.931	0.828	0.034	0.000	0.069	0.56	50.06 (2.46)
AA/NA	99	1.000	0.99	0.000	0.051	0.000	0.030	0.68	59.67 (1.30)
AA/LT	114	0.991	0.596	0.018	0.754	0.000	0.026	0.34	55.09 1.42)
AA/SA	13	0.167	0.167	0.167	0.083	0.250	0.833	0.17	54.33 (4.23)
EA/LT	95	0.789	0.042	0.926	0.642	0.063	0.000	0.67	56.07 (1.44)
EA/PI	40	0.275	0.025	1.000	0.000	0.475	0.025	0.60	56.93 (2.37)
EA/SA	17	0.059	0.000	0.765	0.000	0.059	0.235	0.47	62.06 (2.88)
NA/LT	129	1.000	0.140	0.031	0.953	0.000	0.047	0.68	58.22 (1.19)
LT/PI	12	1.000	0.417	0.250	0.917	0.000	0.167	0.64	53.93 (3.95)
LT/SA	10	0.600	0.000	0.400	0.600	0.200	0.500	0.63	61.50 (4.56)
Three groups	512							0.70	53.52 (0.75)
EW/AA/NA	115	0.991	0.991	0.000	0.043	0.000	0.017	0.74	59.71 (1.58)
EW/AA/LT	23	0.957	0.696	0.043	0.522	0.000	0.087	0.52	50.09 (4.11)
EW/EA/NA	32	0.969	0.000	0.875	0.250	0.000	0.125	0.69	46.06 (3.11)
EW/EA/LT	48	1.000	0.041	0.857	0.490	0.000	0.061	0.72	45.98 (2.49)
EW/EA/PI	35	0.943	0.000	1.000	0.029	0.486	0.000	0.67	51.92 (3.02)
EW/NA/LT	198	1.000	0.066	0.000	0.803	0.000	0.086	0.70	53.83 (0.99)

Only those with at most three self-reported race/ethnicities and three genetic ancestries are included; race/ethnicity categories with at least 10 members are shown. For individuals self-reporting two or three races/ethnicities, the correspondence between self-report and genetic ancestry is generally quite high. For example, for those reporting European/West Asian and East Asian race/ethnicity, 96 and 94% have evidence of European/West Asian and East Asian genetic ancestry, respectively; for those reporting African/African American and East Asian race/ethnicity, 93.1 and 82.8% have evidence of African and East Asian genetic ancestry, while 65.5% have evidence of European/West Asian genetic ancestry. Among those reporting European/West Asian and Native American race/ethnicity, 9.6% have evidence of Native American genetic ancestry; for those reporting African/African American and Native American race/ethnicity, 5.1% have evidence of Native American genetic ancestry. EW: European/West Asian; AA: African/African American/Afro-Caribbean; EA: East Asian; NA: Native American/Alaska Native; LT: Latino; PI: Pacific Islander; SA: South Asian. Genetic ancestry abbreviations are the same except for AF, which represents sub-Saharan African ancestry.

a low level of Native American ancestry (but >5%) whereas the child is below our 5% threshold. Interestingly, in 5 of these cases the parent self-reports as Latino race/ethnicity but the child does not, whereas the opposite is true in 3 of the 8 cases. In an additional 114 cases, the genetic information for parent and child matches but the self-reports for race/ethnicity are different. The largest subgroup (49) of these cases reflects differences in the reporting of Native American or Latino race/ethnicity; and in 47 of these, there is no evidence of Native American genetic ancestry in the parent or child; it is approximately equally split as to whether the parent or child reports the Native American race/ethnicity. Among 53 cases where parent and child are discordant for self-report of Latino race/ethnicity, in ~2/3 it is the child who self-reports Latino race/ethnicity whereas the parent does not. There are 11 cases of discordance for self-report of East Asian race/ethnicity, and in nearly all of them there is no evidence of continental East Asian genetic ancestry. In slightly more than half of these cases it is the parent who self-reports East Asian race/ethnicity.

Discussion

The RPGEH GERA cohort provides an excellent opportunity to characterize a large, representative northern California population from the perspectives of self-reported race/ethnicity/nationality and genetic ancestry. Overall, the cohort is 80.8% non-Hispanic white and 19.2% minority and includes a broad spectrum of races/ethnicities/nationalities. The results of our PC analyses to characterize genetic structure within each of the major race/ethnicity groups are largely consistent with prior reports.

For the non-Hispanic white individuals, we see a broad spectrum of genetic ancestry ranging from northern Europe

Table 3 Proportion of individuals with genetic ancestry from each of six ancestral populations by race/ethnicity as determined by KP administrative databases

		Genetic ancestry					
Race/ethnicity	*n*	EW	AF	EA	NA	PI	SA
White	4575	1	0.007	0.009	0.017	0.001	0.030
African American	102	0.941	0.990	0.000	0.020	0.000	0.020
Asian	311	0.106	0.003	0.952	0.006	0.167	0.074
Latino	255	0.988	0.192	0.043	0.816	0.000	0.035
Other/uncertain	84	0.929	0.131	0.357	0.167	0.071	0.083

Abbreviations are the same as in Table 2.

to southern Europe and the Middle East. Within that large group, with the exception of Ashkenazi Jews, we see little evidence of distinct clusters. This is consistent with considerable exogamy within this group. By comparison, we do see structure in the East Asian population, correlated with nationality, reflecting continuing endogamy for these nationalities and also recent immigration. On the other hand, we did observe a substantial number of individuals who are admixed between East Asian and European ancestry, reflecting ~10% of all those reporting East Asian race/ethnicity. The majority of these reflected individuals with one East Asian and one European parent or one East Asian and three European grandparents. In addition, we noted that for self-reported Filipinos, a substantial proportion have modest levels of European genetic ancestry reflecting older admixture.

As expected, most self-reported African Americans show some degree of European genetic ancestry, with an overall average of 26%. Among individuals self-reporting as African American and East Asian, all showed evidence of genetic ancestry from three continents: Africa, Europe/West Asia, and East Asia.

Latinos are the most complex from a genetic perspective, as they can possess genetic ancestry from essentially any of the major continents. Most of the Latinos in our study derive from Mexico and Central/South America, with smaller proportions from Puerto Rico and Cuba. These individuals have varying proportions of Native American, European, and African genetic ancestry. We also found evidence of East Asian genetic ancestry in some individuals, but these were primarily individuals who self-reported both East Asian and Latino nationalities.

Of note, ~17% of the cohort had evidence of genetic ancestry from more than one continent. However, this does not mean that all or even most of these individuals represent recent continental admixture. As has been true in other analyses (Li *et al.* 2008), genetic similarity between West Asians and South Asians (and to some degree South Asians and East Asians) did not allow for a clear distinction among these genetic ancestries. As such, while some individuals were estimated to have South Asian genetic ancestry, this more likely reflects the difficulty in demarking West Asian *vs.* South Asian genetic ancestry. A similar situation holds for Pacific Islanders and East Asians, where we and others have shown strong genetic similarity for some Pacific Islander groups with East Asians. Also, some individuals may have reported more than a single race/ethnicity that may reflect recent country of origin in addition to, or rather than, more distant ancestry, with Indo-Fijians as one example.

If we include only individuals with genetic admixture from nonadjacent continents, the proportion with continental admixture is ~12%. However, we also note that this fraction depends on our cutoff of 5% for defining genetic admixture as well as some imprecision in the admixture estimation. Of course, a lower threshold would increase the proportion of the cohort that is considered to be genetically admixed, while a higher threshold would do the opposite.

As expected, in a large cohort such as this, we were easily able to identify a substantial number of close relatives: specifically, 34 identical twins, 2018 full sibs, and 3741 parent-child pairs. We also had clear evidence of a large number of likely second-and third-degree relatives, but these kinship groups did not separate clearly from each other. More refined methods may be able to provide more precise kinship estimates.

A major goal was to examine the relationship between self-reported race/ethnicity and genetic ancestry. By and large, there was very high correspondence between the two, allowing for the broad range of genetic ancestry that exists among African Americans and Latinos. We were also able to compare the self-report data of identical twins, parent–child pairs, and sib pairs. All MZ twin pairs were concordant, as were most of the sib pairs. However, we did note that for some sib pairs the self-report data differed. For the majority of these, the discordance related to reporting of Native American or Latino race/ethnicity.

The results obtained here are important for the study of complex genetic disease in this large, population-based cohort through association studies, admixture analysis and admixture mapping and in particular for investigating observed ethnic variation in diseases and traits. As described previously (Risch *et al.* 2002; Tang *et al.* 2006), the strong correspondence, also observed here, between the social categories of race/ethnicity and genetic ancestry makes dissection of racial/ethnic differences challenging. The patterns that we observed reflect historical and recent mating practices and their impact on genetic variation. On a global level, geography continues to create strong local endogamy, which is also reflected among the recent U.S. migrant populations. However, the increasing frequency of interracial individuals that we observed in this cohort—a reflection of increasing exogamy—will enhance both the complexity of such analyses and the opportunities to investigate the genetic and environmental contributors to racial/ethnic differences. While the advent of myriad genetic markers can provide accurate estimates of individuals' genetic ancestry, the social aspects of race/ethnicity may be more challenging to characterize. For example, in our study, considering the various combinations of 7 race/ethnicity categories that an individual could

endorse, we observed 50 different combinations, and this does not include individuals who endorsed >3 (although they were few in number). While overall 6% of the cohort endorsed more than a single category, that number is likely to grow as mating patterns continue to evolve.

Acknowledgments

We thank the Kaiser Permanente Northern California members who have generously agreed to participate in the Kaiser Permanente Research Program on Genes, Environment and Health and Judith Millar for her assistance in preparing the manuscript for publication. This work was supported by grants RC2 AG036607 and R01 GM073059 from the National Institutes of Health and by a postdoctoral fellowship from the Lamond Family Foundation. The development of the Research Program on Genes, Environment, and Health, including enrollment and consent of participants and collection of surveys and saliva samples, was supported by grants from the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, the Ellison Medical Foundation, and Kaiser Permanente Community Benefit Programs. Information about data access can be obtained at http://www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs000674.v1.p1 and https://rpgehportal.kaiser.org.

Note added in proof: See Kvale *et al.* 2015 (pp. 1051–1060) and Lapham *et al*. 2015 (pp. 1061–1072) in this issue for related works.

Literature Cited

Barbujani, G., and G. Bertorelle, 2001 Genetics and the population history of Europe. Proc. Natl. Acad. Sci. USA 98: 22–25.

Bauchet, M., B. McEvoy, L. N. Pearson, E. E. Quillen, T. Sarkisian *et al.*, 2007 Measuring European population stratification with microarray genotype data. Am. J. Hum. Genet. 80: 948–956.

Belle, E. M., P. A. Landry, and G. Barbujani, 2006 Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. Proc. Biol. Sci. 273: 1595–1602.

Bonilla, C., E. J. Parra, C. L. Pfaff, S. Dios, J. A. Marshall *et al.*, 2004 Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. Ann. Hum. Genet. 68: 139–153.

Burchard, E. G., E. Ziv, N. Coyle, S. L. Gomez, H. Tang *et al.*, 2003 The importance of race and ethnic background in biomedical research and clinical practice. N. Engl. J. Med. 348: 1170–1175.

Cavalli-Sforza, L. L., 2005 The Human Genome Diversity Project: past, present and future. Nat. Rev. Genet. 6: 333–340.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza, 1993 Demic expansions and human evolution. Science 259: 639–646.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza, 1996 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ, pp xiii and 413.

Cooper, R. S., J. S. Kaufman, and R. Ward, 2003 Race and genomics. N. Engl. J. Med. 348: 1166–1170.

Fernandez, J. R., M. D. Shriver, T. M. Beasley, N. Rafla-Demetrious, E. Parra *et al.*, 2003 Association of African genetic admixture with resting metabolic rate and obesity among women. Obes. Res. 11: 904–911.

Hodgson, J. A., C. J. Mulligan, A. Al-Meeri, and R. L. Raaum, 2014 Early back-to-Africa migration into the Horn of Africa. PLoS Genet. 10: e1004393.

Hoffmann, T. J., M. N. Kvale, S. E. Hesselson, Y. Zhan, C. Aquino *et al.*, 2011a Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics 98: 79–89.

Hoffmann, T. J., Y. Zhan, M. N. Kvale, S. E. Hesselson, J. Gollub *et al.*, 2011b Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics 98: 422–430.

HUGO Pan-Asian SNP Consortium; M. A. Abdulla, I. Ahmed, A. Assawamakin, J. Bhak, S. K. Brahmachari *et al*., 2009 Mapping human genetic diversity in Asia. Science 326: 1541–1545.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451: 998–1003.

Krieger, N., D. L. Rowley, A. A. Herman, B. Avery, and M. T. Phillips, 1993 Racism, sexism, and social class: implications for studies of health, disease, and well-being. Am. J. Prev. Med. 9: 82–122.

Kvale, M. N., S. E. Hesselson, T. J. Hoffmann, Y. Cao, D. Chan *et al.*, 2015 Genotyping informatics and quality control for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. DOI: 10.1534/genetics.115.178905.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104.

Lohmueller, K. E., A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez *et al.*, 2008 Proportionally more deleterious genetic variation in European than in African populations. Nature 451: 994–997.

Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale *et al.*, 2010 Robust relationship inference in genome-wide association studies. Bioinformatics 26: 2867–2873.

Menozzi, P., A. Piazza, and L. Cavalli-Sforza, 1978 Synthetic maps of human gene frequencies in Europeans. Science 201: 786–792.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko *et al.*, 2008 Genes mirror geography within Europe. Nature 456: 98–101.

Parra, E. J., A. Marcini, J. Akey, J. Martinson, M. A. Batzer *et al.*, 1998 Estimating African American admixture proportions by use of population-specific alleles. Am. J. Hum. Genet. 63: 1839–1851.

Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet. 2: e190.

Price, A. L., J. Butler, N. Patterson, C. Capelli, V. L. Pascali *et al.*, 2008 Discerning the ancestry of European Americans in genetic association studies. PLoS Genet. 4: e236.

Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh, 2009 Reconstructing Indian population history. Nature 461: 489–494.

Risch, N., E. Burchard, E. Ziv, and H. Tang, 2002 Categorization of humans in biomedical research: genes, race and disease. Genome Biol. 3: comment2007.

Seldin, M. F., R. Shigeta, P. Villoslada, C. Selmi, J. Tuomilehto *et al.*, 2006 European population substructure: clustering of northern and southern populations. PLoS Genet. 2: e143.

Sokal, R. R., N. L. Oden, and C. Wilson, 1991 Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature 351: 143–145.

Su, B., J. Xiao, P. Underhill, R. Deka, W. Zhang *et al.*, 1999 Y-chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. Am. J. Hum. Genet. 65: 1718–1724.

Tang, H., J. Peng, P. Wang, and N. J. Risch, 2005 Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. 28: 289–301.

Tang, H., E. Jorgenson, M. Gadde, S. L. Kardia, D. C. Rao *et al.*, 2006 Racial admixture and its impact on BMI and blood pressure in African and Mexican Americans. Hum. Genet. 119: 624–633.

Tang, H., S. Choudhry, R. Mei, M. Morgan, W. Rodriguez-Cintron *et al.*, 2007 Recent genetic selection in the ancestral admixture of Puerto Ricans. Am. J. Hum. Genet. 81: 626–633.

Tian, C., P. K. Gregersen, and M. F. Seldin, 2008a Accounting for ancestry: population substructure and genome-wide association studies. Hum. Mol. Genet. 17: R143–R150.

Tian, C., R. Kosoy, A. Lee, M. Ransom, J. W. Belmont *et al.*, 2008b Analysis of East Asia genetic substructure using genome-wide SNP arrays. PLoS One 3: e3862.

Tian, C., R. M. Plenge, M. Ransom, A. Lee, P. Villoslada *et al.*, 2008c Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet. 4: e4.

Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro *et al.*, 2009 The genetic structure and history of Africans and African Americans. Science 324: 1035–1044.

Wellcome Trust Case Consortium, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.

Zakharia, F., A. Basu, D. Absher, T. L. Assimes, A. S. Go *et al.*, 2009 Characterizing the admixed African ancestry of African Americans. Genome Biol. 10: R141.

Communicating editor: N. R. Wray

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.178616/-/DC1

Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort

Yambazi Banda, Mark N. Kvale, Thomas J. Hoffmann, Stephanie E. Hesselson, Dilrini Ranatunga, Hua Tang, Chiara Sabatti, Lisa A. Croen, Brad P. Dispensa, Mary Henderson, Carlos Iribarren, Eric Jorgenson, Lawrence H. Kushi, Dana Ludwig, Diane Olberg, Charles P. Quesenberry,Jr. Sarah Rowell, Marianne Sadler, Lori C. Sakoda, Stanley Sciortino, Ling Shen, David Smethurst, Carol P. Somkin, Stephen K. Van Den Eeden, Lawrence Walter, Rachel A. Whitmer, Pui-Yan Kwok, Catherine Schaefer, and Neil Risch

DOI: 10.1534/genetics.115.178616

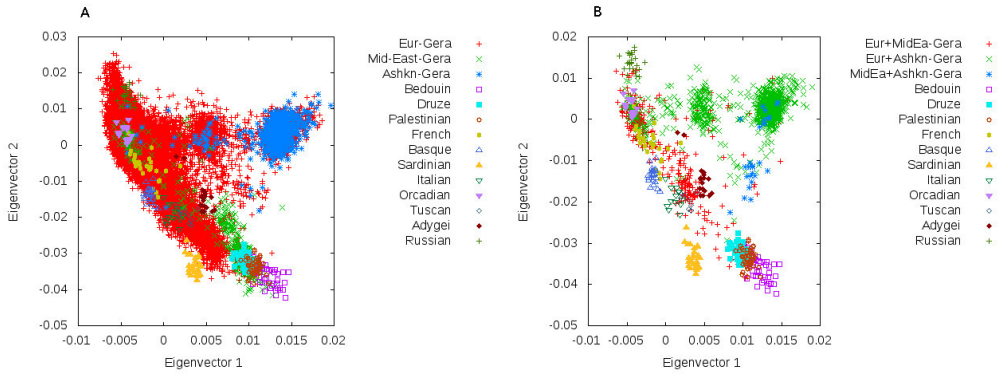

Figure S1 PCA of European and West Asian subjects on the EUR array. A clear Ashkenazi cluster is observed. The largest cluster depicts the northwest-southeast cline within Europe. **A** Those reporting a single ethnicity; **B** Those reporting multiple ethnicities.

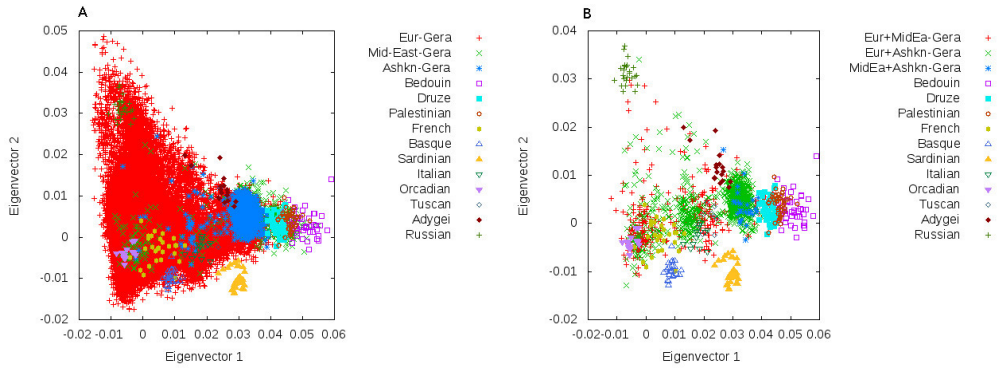

Figure S2 PCA of subjects who are neither South Asian nor Ashkenazi. The Ashkenazi subjects were later projected onto the PCs obtained. **A** Those reporting a single ethnicity; **B** Those reporting more than one ethnicity.

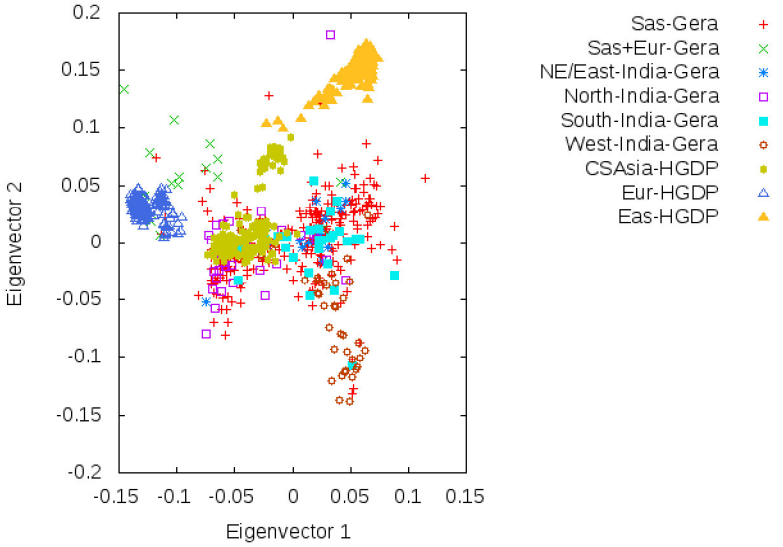

Figure S3 PCA of individuals reporting South Asian ethnicity, either alone or in combination with European ethnicity. Separate clusters of the Indian subgroups from different Indian regions identified by onomastics are also identified.

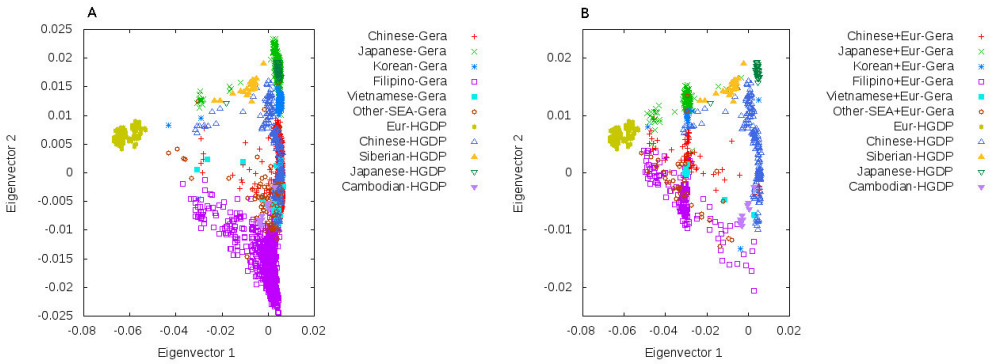

Figure S4 PCA of subjects on the EAS array. **A** Individuals reporting only East Asian ancestry; **B** Individuals reporting East Asian and European ancestry.

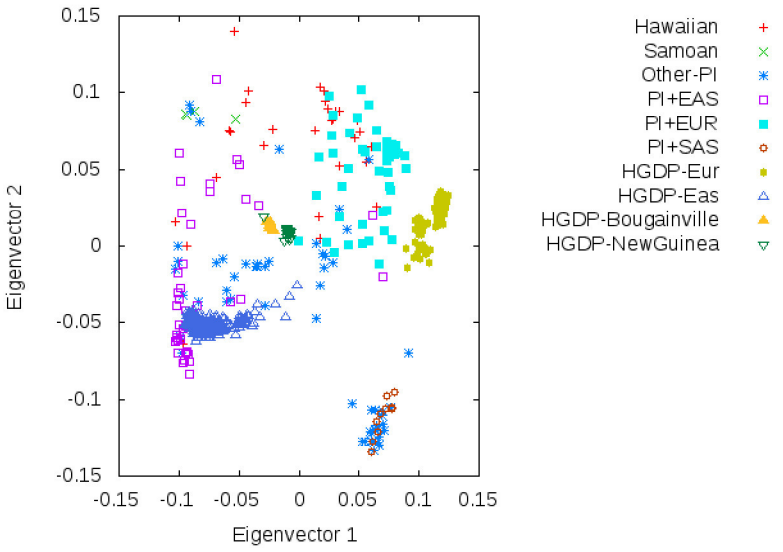

Figure S5 PCA of individuals run on the EAS array reporting Pacific Islander ethnicity, including those reporting another ethnicity.

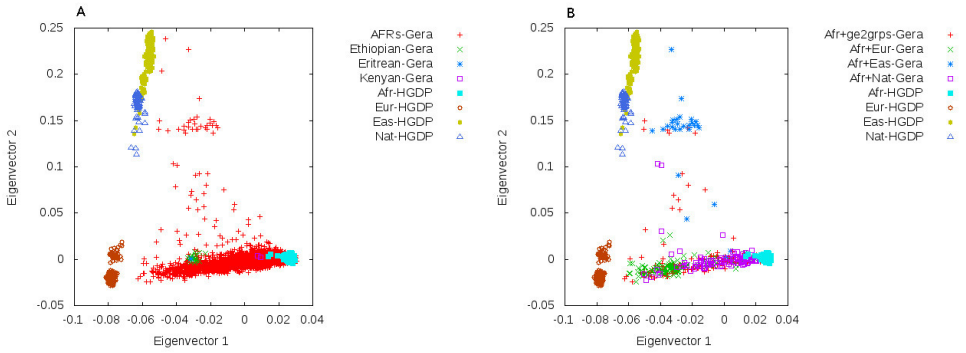

Figure S6 PCA of individuals run on the AFR array. **A** Individuals reporting only African or African American ethnicity; individuals identified by onomastics as Ethiopian, Eritrean and Kenyan depicted separately; **B** Individuals reporting African/African American ethnicity plus at least one additional ethnicity.

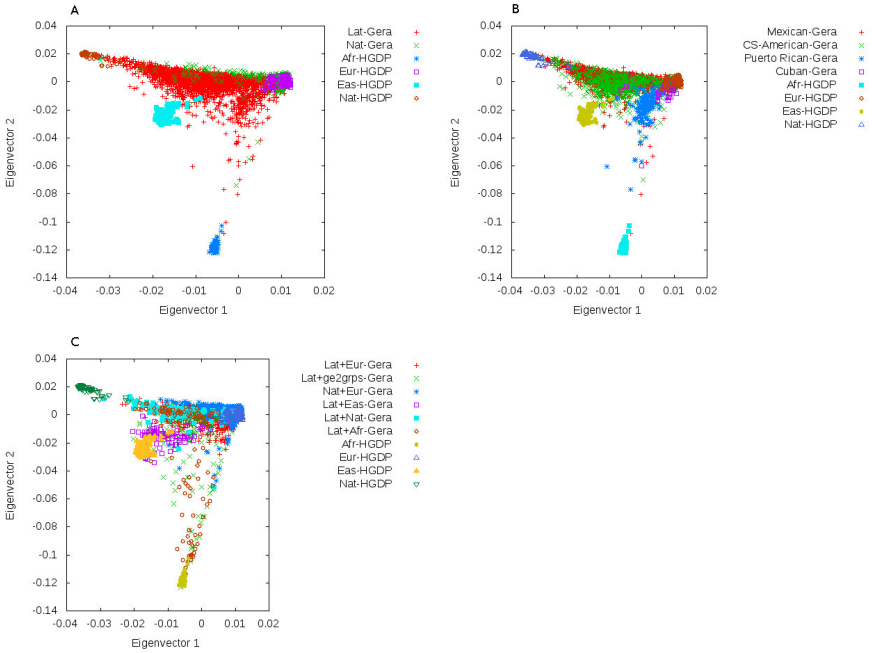

Figure S7 PCA of individuals run on the LAT array. **A** Individuals reporting Latino ethnicity only and Native American ethnicity only; **B** Individuals reporting Latino ethnicity by nationality; **C** Individuals reporting Latino ethnicity and at least one additional ethnicity, and also individuals reporting Native American and European ethnicities only.

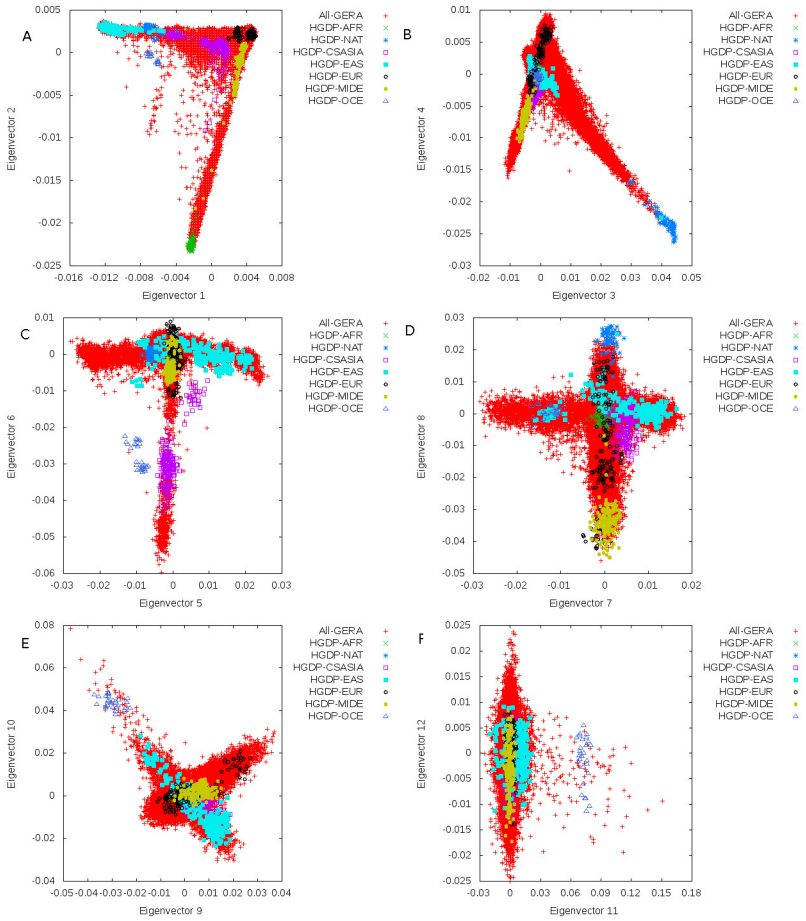

Figure S8 Global PCA of GERA subjects. **A-F** Individuals distributed according to continental differentiation. Admixed individuals also observed.

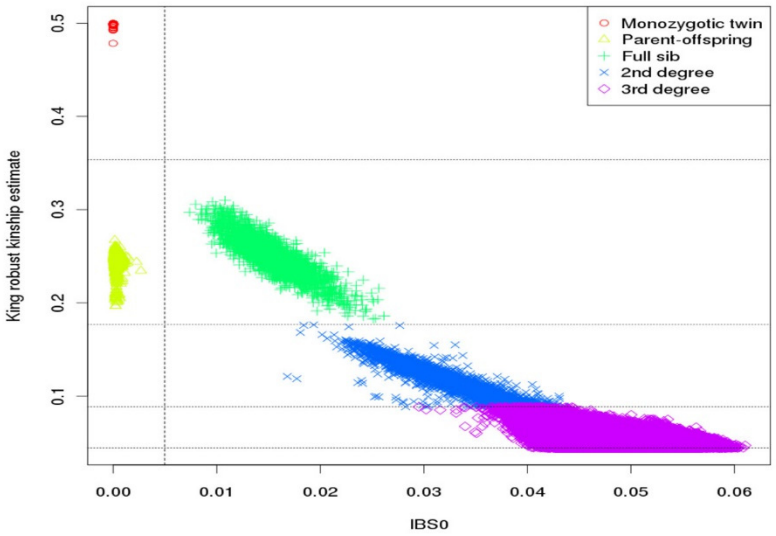

Figure S9 Identification of MZ twin and first-degree relative pairs by KING. The x-axis represents the proportion of SNPs with zero IBS (Identity by State) e.g. TT and CC.

File S1

Supplementary Methods and Results

Adjudication of Race/Ethnicity Information

As described in Methods, PC and admixture analyses were performed on the genotype data to characterize the genetic ancestry of the GERA cohort. In so doing, we discovered a large number of individuals (2,140) run on the AFR array who were estimated to have 100% European/West Asian genetic ancestry and a smaller number (123) to have 100% East Asian genetic ancestry. A small number (276) of individuals run on the EAS array were estimated to have 100% European ancestry. This led to further investigation of the self-report assignments for these individuals.

Direct examination of the original survey forms for these subjects revealed that the self-reported race/ethnicity information on the form was not consistent with the computerized information for some individuals. Further investigation revealed that an artifact had occurred when the forms were originally scanned due to a black mark on the glass scanning plate overlaying the box for African American, which led to a number of subjects being recorded as having African American race/ethnicity (in addition to another race/ethnicity) when in fact they did not (according to the original form). By our algorithm for array assignment, because these individuals were assumed to have some African ancestry, they were assigned to the AFR array. These individuals were then adjudicated to race/ethnicity categories based on the actual survey information, supplemented by other data on race/ethnicity in the KP administrative databases, as necessary, and principal component scores were recalculated. This error only affected individuals with originally recorded African American race/ethnicity from the computerized scanning. The individuals with 100% European/West Asian genetic ancestry that were run on the EAS array had no scanning errors and hence were not adjudicated. Reviewing the race/ethnicity self-report of these individuals, the large majority self-reported both East Asian and European race/ethnicity.

Table S2 displays the relationship between self-reported race/ethnicity and genotyping array for those with self-report race/ethnicity data that was not missing or adjudicated due to scanning errors. Because of the time element required for processing samples, the array assignments were made based on raw data from the race/ethnicity questions on the survey, prior to data cleaning. After data cleaning, we noticed that some individuals were assigned to arrays based on the raw information that was not consistent with their final race/ethnicity categorization and the assignment algorithm. As can be seen in Table S2, this primarily affected a modest number of individuals with a final assignment of mixed race/ethnicity who were run on the EUR array. Table S3 shows array assignments for individuals with scanning errors or missing self-report race/ethnicity data whose

final race/ethnicity was determined from KPNC administrative databases. In this group are the 2,140 Whites and 123 Asians with scanning errors who were run on the AFR array. We also note a moderate number of individuals classified as White (267) who were run on the EAS array.

Finally, we emphasize that no race/ethnicity assignments or re-assignments were based on genetic information. The genetic information was only used in the detection of the scanning problems for some individuals by comparing their computerized responses to those on the original forms. All self-report race/ethnicity data reported in Results are based on the final cleaned and adjudicated categories.

Numbers of pruned SNPs for various PCAs and Individual Admixture Estimation

To reduce the linkage disequilibrium (LD) between markers (e.g. those in the lactase and MHC regions on chromosomes 2 and 6, respectively), pairs of SNPs that had an r^2 greater than 0.5 and within 5 MB of each other were considered and one member of the pair removed. Also removed were SNPs located in regions with inversions such as chromosomes 8p23 and 17q21. These structural variations have previously been shown to influence PCA results in European ancestry samples. As reported in Results, various PCA runs were performed separately for individuals genotyped on different arrays. The numbers of SNPs remaining after LD and structural variation loci pruning, for each of the eight different PCA runs, are shown in Table S4.

For the initial individual admixture analyses, a set of 43,988 SNPs which were common between the HGDP dataset and our set of 144,799 high quality SNPs was used. A set of 38,301 SNPs (used for the 'global' PCA run), remaining after LD pruning and removal of SNPs located in regions with structural variations, was also used for admixture analyses. We found very minimal difference in admixture estimates obtained from the two types of analyses.

Distribution of continental genetic ancestry as a function of self-reported race/ethnicity.

Table S8 provides a more detailed examination of the distribution of continental genetic ancestry for the various self-report race/ethnicity groups. Among those reporting European/West Asian race/ethnicity, 5.6% had evidence of genetic ancestry from two continents; however, for the large majority the second continent was South Asia. Hence, this likely does not reflect recent admixture, but rather the genetic similarity of West Asians and South Asians. For a moderate proportion, the second genetic ancestry is Native American. By contrast, among the self-reported Africans/African Americans, 88.2% have evidence of genetic

ancestry from two continents. This represents the European/West Asian genetic admixture present in most African Americans that has occurred over 5 centuries. For those with a single genetic ancestry, that ancestry is African. As expected, nearly all self-reported Latinos have genetic ancestry from more than one continent; a substantial proportion (29.9%) have genetic ancestry from 3 continents – European/West Asian, Native American and African, while the majority (65.2%) have genetic ancestry from two continents, European/West Asian and Native American. The large majority of self-reported East Asians have genetic ancestry that is solely East Asian or East Asian and Pacific Islander. The latter combination primarily reflects the close genetic relatedness of East Asians to some Pacific Islander groups and not necessarily recent admixture, although that likely applies to some. We do note that for a modest number of individuals, the second continental ancestry is European/West Asian. Similarly, for self-reported South Asians, the large percentage (38.1%) corresponding to two continents primarily reflects genetic similarity of West Asians and South Asians in the admixture analysis. Among those self-reporting only Native American race/ethnicity, 82.3% have a single genetic ancestry which is European/West Asian, although 17.7% have genetic ancestry from two or more continents, which are European/West Asian and Native American.

Those who self-reported more than one race/ethnicity comprise multiple combinations. Among those reporting two, 51% have a single genetic ancestry which is nearly always European/West Asian. Similarly, for those with genetic ancestry from more than one continent, for nearly all European/West Asian is one of them, with the second continental group being Native American (60%), East Asian (22%) or African (13%). The pattern is similar for those with genetic ancestry from three continents. The pattern is also closely reproduced in those reporting 3 race/ethnicity categories; the large majority with a single continental genetic ancestry reflects European/West Asian genetic ancestry. For those with two or more continental genetic ancestries, European/West Asian is nearly always one of them; but in this case, African ancestry is more prominent than East Asian ancestry as the second continent.

Table S1 Magnitude of correlation of PC loadings for three 'supersets'.

PC	Set1-Set2 correlation	Set1-Set3 correlation	Set2-Set3 correlation
1	0.981	0.979	0.980
2	0.876	0.853	0.856
3	0.850	0.816	0.809
4	0.766	0.776	0.752
5	0.658	0.654	0.639
6	0.605	0.604	0.592
7	0.001	0.001	0.210
8	0.043	0.179	0.045
9	0.312	0.068	0.089
10	0.007	0.067	0.201

Table S2 Self-reported Race/Ethnicity versus Genotyping Array. Race/Ethnicity abbreviations: EW = European/West Asian; AA = African/African America/Afro-Caribbean; EA = East Asian; NA = Native American; LT = Latino; PI = Pacific Islander; SA = South Asian.

Race/Ethnicity	**Genotyping Array**			
	EUR	EAS	AFR	LAT
EW	76403	0	0	1
AA	0	0	2682	0
EA	0	6394	0	0
NA	0	0	0	674
LT	0	0	0	4855
PI	0	92	0	0
SA	461	0	0	0
EW,AA	0	0	123	0
EW,EA	38	538	0	0
EW,NA	91	0	0	2463
EW,LT	12	0	0	1561
EW,PI	8	45	0	0
EW,SA	44	0	0	0
AA,EA	0	0	32	0
AA,NA	0	0	100	0
AA,LT	0	0	3	113
AA.PI	0	0	4	0
AA,SA	0	0	12	0
EA,NA	0	7	0	0
EA,LT	0	2	0	108
EA,PI	0	40	0	0
EA,SA	2	15	0	0
NA,LT	0	0	0	130
NA,PI	0	2	0	0
LT,PI	0	0	0	14
LT,SA	0	0	0	8
PI,SA	0	10	0	0
EW,AA,EA	0	0	6	0
EW,AA,NA	0	0	115	0
EW,AA,LT	0	0	0	23
EW,AA,PI	0	0	1	0
EW,AA,SA	0	0	2	0
EW,EA,NA	2	0	0	30
EW,EA,LT	0	1	0	52
EW,EA,PI	0	36	0	0
EW,NA,LT	0	0	0	199
EW,NA,PI	0	0	0	2
EW,LT,PI	2	0	0	6
EW,LT,SA	0	0	0	4
EW,PI,SA	0	1	0	0
AA,EA,NA	0	0	8	0
AA,EA,LT	0	0	0	8
AA,EA,PI	0	0	1	0
AA,EA,SA	0	0	3	0
AA,NA,LT	0	0	1	1
AA,LT,SA	0	0	1	6
EA,NA,LT	0	0	0	9
EA,LT,PI	0	0	0	2
EA,PI,SA	0	1	0	0
NA,LT,PI	0	0	0	3
Total	77063	7184	3094	10272

Table S3 Race/Ethnicity derived from KP administrative databases versus genotyping array used for those with missing or mis-scanned self-report data

Race/Ethnicity	**Genotyping Array**			
	EUR	EAS	AFR	LAT
White	2115	267	2140	55
African American	0	0	99	3
Hispanic	5	0	0	255
Asian	5	183	123	1
Other/Uncertain	6	11	43	24
Total	2131	461	2405	338

Table S4 Numbers of SNPs remaining after LD and structural variation locus pruning, for each of the eight different PCAs.

PCA run	Number of SNPs after pruning
European/West Asian and Ashkenazi	38050
European only	37967
South Asian	38823
East Asian	38077
Pacific Islander	38833
African American	39849
Latino	38365
All GERA	38301

Table S5 Individual ancestral admixture proportions for subjects run on the LAT array.

Nationality	**Ancestral admixture proportion (%)**		
	African	European	Native American
Mexican	2.3 ± 5.0	67.0 ± 14.5	30.7 ± 13.5
Central-South American	5.5 ± 10.8	65.4 ± 17.9	29.1 ± 16.3
Puerto-Rican	12.3 ± 14.1	75.5 ± 14.8	12.4 ± 7.6
Cuban	12.7 ± 20.5	79.7 ± 20.0	7.6 ± 8.1
LAT mean	4.4 ± 7.9	66.5 ± 15.8	28.6 ± 14.8

Table S6 Distribution of genetic ancestry by self-reported race/ethnicity. A particular genetic ancestry was assigned to an individual if at least 5% of that individual's ancestry was estimated from that group. Race/Ethnicity abbreviations as in Table S2. Genetic ancestry abbreviations are the same except for AF which represents sub-Saharan African

Race-Ethnicity	Genetic Ancestry																										
	EW	AF	EA	NA	SA	EW AF	EW EA	EW NA	EW SA	AF EA	AF SA	EA NA	EA PI	EA SA	PI SA	EW AF EA	EW AF NA	EW AF SA	EW EA NA	EW EA PI	EW EA SA	EW NA SA	EW PI SA	AF EA PI	AF EA SA	AF PI SA	EA PI SA	All
EW	71992					191	243	646	3208							1	27	9	11	9	24	38	1				1	76401
AA		230				2349		5		1	7					9	28	46	1		1				1	1		2679
EA			4835		2		108					23	1267	21		4			8	93	4			1			23	6389
NA	555					5	3	67	11			1				1	7	2	10			12						674
LT	232	1	1	1		34		3094	4			1	1		1	2	1294	1	33			107						4807
PI			4		3		19		1				13	3	4					32	1						12	92
SA	1		15		256		1	1	132		1			24	16			2			3	1					7	460
EW/AA						107	2										3	10	1									123
EW/EA	32		16				431	1					7			3			7	29	46							572
EW/NA	2248					16	2	206	34							1	4		11		3	23						2548
EW/LT	404					25	5	929	17							1	84	1	9			89						1564
EW/PI	8						7	1											1	30	1							48
EW/SA	14								25					2							1	2						44
AA/EA		1			1	1	1			7	1					16	1											29
AA/NA						91		1									4	3										99
AA/LT						26		43			1						40	1	2			1						114

AA/PI															1	1								1			1	4			
AA/SA		1			6	1		1					1	1												1	1	13			
EA/NT			2				1											3										6			
EA/LT			15				10	6				5			3	1			54	1								95			
EA/PI			18				1					11			1					8	1							40			
EA/SA			12		3				1			1																17			
NA/LT		3					1	1	98	1						17			3			5						129			
NA/PI									1																			1			
LT/PI									3							5			2		1	1						12			
LT/SA			1		1				3										1			2					2	10			
PI/SA					5								1	2														8			
EW/AA/EA						1	2								2													5			
EW/AA/NA			1			107		1								4	2											115			
EW/AA/LT			1			9		5								5	1	1				1						23			
EW/AA/SA						1																						1			
EW/AA/PI															1													1			
EW/EA/NA		2					20				1								5		2	2						32			
EW/EA/LT		2					19	4							2				19		2							48			
EW/EA/PI							17					2							1	15								35			
EW/NA/LT		34				2		135	2							10	1					14						198			
EW/NA/PI							1													1								2			
EW/LT/PI		3						2								1				1	1							8			

EW/LT/SA	2							1										1											4
EW/PI/SA																												1	1
AA/EA/NA						1				1						4	1	1											8
AA/EA/LT						1		1											2										4
AA/EA/SA					2										1														3
AA/NA/LT						1																1							2
AA/LT/SA						1		3									2					1							7
EA/NA/LT																			8										8
EA/LT/PI			1														1												2
EA/PI/SA													1																1
NA/LT/PI								2											1										3
Total	75532	235	4920	1	279	2971	894	5260	3436	9	10	26	1308	52	25	52	1540	81	194	219	91	300	1	2	1	2	48	97489

Table S7 Distribution of genetic ancestry by race/ethnicity as reported in the KP electronic health records for those with missing or mis-scanned self-report race/ethnicity. Abbreviations as in Table S6.

Race-Ethnicity																						
	EW	AF	EA	SA	EW AF	EA NA	EA PI	EW EA	EW NA	EW SA	EA SA	SA PI	EW AF EA	EW AF NA	EW AF SA	EW EA NA	EW EA PI	EW EA SA	EW SA NA	EW SA PI	EA SA PI	Total
White	4302	0	0	0	25	0	0	33	68	131	0	0	0	3	2	3	2	3	2	1	0	4575
Afr. Am.	0	6	0	0	92	0	0	0	1	0	0	0	0	1	2	0	0	0	0	0	0	102
Asian	4	0	218	8	0	1	41	18	0	2	3	1	1	0	0	1	4	3	0	0	6	311
Latino	37	0	3	0	3	0	0	2	151	0	0	0	1	44	1	5	0	0	8	0	0	255
Other-uncertain	34	0	3	0	6	0	0	15	8	2	1	0	2	2	1	3	4	0	1	0	2	84

Table S8 Distribution of continental genetic ancestry as a function of self-reported race/ethnicity.

Race-Ethnicity		Genetic Ancestry									
		One Continent		Two Continents		Three Continents		All	Continental Distribution		
Number		Number	%	Number	%	Number	%		1 Continent	2 Continents	3 Continents
One	EW	71,992	94.2	4,288	5.6	121	0.2	76401	100% EW	75% EW,SA	
										15% EW,NA	
	AA	230	8.6	2,362	88.2	87	3.2	2679	100% AF	99% AF,EW	
	LT	235	4.9	3,135	65.2	1,437	29.9	4807	99% EW	99% EW,NA	90% EW,NA,AF
	EA	4,837	75.7	1,419	22.2	133	2.1	6389	100% EA	89% EA,PI	
										8% EA,EW	
	PI	7	7.6	40	43.5	45	48.9	92		48% PI,EW	71% PI,EW,EA
										33% PI,EA	
	SA	272	59.1	175	38.1	13	2.8	460	94% SA	75% SA,EW	
										14% SA,EA	

	NA	555	82.3	87	12.9	32	4.8	674	100% EW	77% EW,NA	
	All	78,128	85.4	11,506	12.6	1,868	2.0	91502			
	% of Total							93.9			
Two	All	2,791	51.0	2140	39.1	545	10.0	5476	97% EW	60% EW,NA	29% EW,NA,AF
										22% EW,EA	23% EW,SA,NA
										13% EW,AF	17% EW,EA,NA
	% of Total							5.6			
Three	All	48	9.4	345	67.5	118	23.1	511	90% EW	45% EW,NA	31% EW,EA,NA
										36% EW,AF	20% EW,AF,NA
										19% EW,EA	16% EW,SA,NA
	% of Total							0.5			
All	**All**	**80967**	**83.1**	**13,991**	**14.4**	**2,531**	**2.6**	**97,489**			

Table S9 First degree relatives organized by self-reported race/ethnicity

MZ pair

	White	African American	Latino	Asian	Other/Uncertain
White	28	0	0	0	0
African American	0	0	0	0	0
Latino	0	0	2	0	0
Asian	0	0	0	4	0
Other/Uncertain	0	0	0	0	0

Parent (column) – Offspring (row)

	White	African American	Latino	Asian	Other/Uncertain
White	3044	1	23	6	21
African American	8	54	0	1	1
Latino	122	6	175	3	0
Asian	35	2	0	205	1
Other/Uncertain	32	0	1	0	0

Full sibs

	White	African American	Latino	Asian	Other/Uncertain

White	1531	2	33	5	30
African American		45	7	0	0
Latino			155	2	2
Asian				205	1
Other/Uncertain					0

Table S10 Race/Ethnicity and Genetic Ancestry for Sib Pairs Discordant for Race/Ethnicity. Abbreviations as in Table S6.

Sib 1 Race/Ethnicity	Sib 2 Race/Ethnicity	Genetic Ancestry	Number
Both Sibs Self-Report			
EW	NA	EW	26
EW	LT	EW	6
EW	LT	EW,NA	1
EW	AA	EW,AF	1
EW	EW,LT	EW	15
EW	EW,LT	EW,NA	6
EW	EW,LT	EW,AF,NA	1
EW	EW,AA,LT	EW,AF	1
EW	EW,EA	EW	3
EW	EW,EA	EW,EA	1
EW	EW,SA	EW	1
EW,NA	NA	EW	1
EW,NA	NA	EW,NA	1
EW,NA	NA	EW,AF	1
EW,NA	EW,NA,LT	EW	1
EW,LT	NA	EW	1
EW,LT	AA,LT,SA	EW,AF	1
EW,LT	EW,AA,NA,LT	EW,AF	1
EW,EA	LT	EW,NA	1
EW,AA,NA,LT,SA	LT	EW,NA	1
EW,LT,PI	PI	EW,EA,PI	1
LT	AA,LT	EW,NA	3
LT	AA,LT	EW,AF,NA	1
One Sib Self-Report; One Sib EHR			
EW	Latino	EW	1
EW,LT	Other/Uncertain	EW	1
LT	White	EW	1
LT,EA	Other/Uncertain	EW,EA,PI	1
EA,PI	Other/Uncertain	EA	1
Both Sibs EHR			
White	Latino	EW,NA	1

Table S11 Race/Ethnicity and Genetic Ancestry for Parent-Child pairs Discordant for Race/Ethnicity. Abbreviations as in Table S6.

Parent Race/Ethnicity	Child Race/Ethnicity	Parent Genetic Ancestry	Child Genetic Ancestry	Number
Parent and Child Self-Report				
EW	EW,AA	EW	EW,AF	4
EW	EW,AA,NA	EW	EW,AF	1
EW	EW,EA	EW	EW	3
EW	EW,EA	EW	EW,EA	19
EW	EW,EA	EW	EW,EA,SA	3
EW	EW,EA	EW,SA	EW	1
EW	EW,EA	EW,SA	EW,EA	1
EW	EW,LT	EW	EW	18
EW	EW,LT	EW	EW,NA	41
EW	EW,LT	EW	EW,SA	2
EW	EW,LT	EW	EW,NA,SA	2
EW	EW,LT	EW,NA	EW	2
EW	EW,LT	EW,NA	EW,NA	5
EW	EW,LT	EW,NA	EW,NA	1
EW	EW,LT	EW,SA	EW,NA	1
EW	EW,LT	EW,SA	EW,NA,SA	1
EW	EW,LT	EW	EW,EA	1
EW	EW,LT	EW	EW,EA,NA	1
EW	EW,LT	EW,NA	EW,AF,EA	1
EW	EW,EA,LT	EW	EW,EA	1
EW	EW,EA,LT	EW	EW,EA,NA	1
EW	EW,EA,NA	EW	EW,EA	1
EW	EW,EA,NA,LT	EW	EW,EA	1
EW	EW,EA,PI	EW	EW,EA,PI	1
EW	EW,NA,LT	EW	EW,AF	1
EW	EW,NA,LT	EW	EW,NA	2
EW	EW,NA,LT	EW	EW,NA,SA	1
EW	EW,PI	EW	EW,EA	1
EW	EW,PI	EW	EW,EA,PI	1
EW	AA	EW	EW,AF	1
EW	AA,LT	EW	EW,NA	1
EW	EA	EW	EW,EA	1
EW	LT	EW	EW	6
EW	LT	EW	EW,NA	9
EW	LT	EW	EW,NA,SA	2
EW	LT	EW,NA	EW,NA	3
EW	LT	EW	EW,AF,NA	1
EW	NA	EW	EW	24
EW	NA	EW,SA	EW	1
EW	NA	EW,SA	EW,SA	1
EW,AA	EW,AA	EW	EW,AF	1
EW,EA	EW	EW	EW	3
EW,LT	EW	EW	EW	6
EW,LT	EW	EW	EW,NA	1
EW,LT	EW	EW,EA	EW,EA	1
EW,LT	EW	EW,NA	EW	3
EW,LT	EW	EW,NA	EW,NA	1
EW,EA	EW,LT	EW	EW	1
EW,EA	EW	EW,EA	EW,EA	1

EW,EA	EW,AA,PI	EW,EA,PI	EW,AF,EA	1
EW,NA	NA	EW,NA	EW,NA	1
EW,NA	EW,LT	EW	EW,NA	2
EW,NA	NA,LT	EW	EW,NA	1
EW,NA	EW,EA,LT	EW	EW,EA	1
EW,NA	EW,EA,NA	EW,NA,SA	EW,EA,NA	2
EW,AA,NA	EW,NA	EW,AF,SA	EW,AF,EA	1
EW,EA,NA,PI	EW,EA,LT	EW,EA	EW,EA	1
EW,EA,PI	EW	EW,EA,PI	EW,EA	1
EW,NA,LT	EW	EW	EW	1
EW,NA,LT	EW	EW	EW,SA	1
LT	EW	EW	EW	3
LT	EW	EW,NA	EW,NA	1
LT	EW	EW,NA	EW,NA,SA	1
LT	EW	EW,AF,NA	EW,NA	1
LT	EW,NA	EW,AF,SA	EW,AF,NA,SA	1
NA	EW	EW	EW	18
NA,LT	EW	EW,NA	EW	1
NA	EW,NA	EW	EW	1
AA,LT	LT	EW,AF,NA	EW,AF,NA	2
AA,LT	LT	EW,NA	EW,NA,SA	1
AA,SA	EW,LT	EA,SA	EW,SA	1
AA,EA,LT	EW,LT	EW,NA	EW,NA	1
AA,EA,SA	SA	PI,SA	PI,SA	1
AA,LT,SA	LT	EW,NA	EW,NA	1
EA	EW	EW,EA	EW,EA	1
EA	EA,LT	EW,NA	EW,EA	1
EA	EW,EA	EW	EW,EA	1
EA	EW,EA,LT	EW	EW,EA	1
Parent Self-Report; Child EHR				
EW	Other/Uncertain	EW	EW	1
EW	Other/Uncertain	EW	EW,NA,SA	1
EW	Latino	EW	EW,NA	3
EW,LT	Other/Uncertain	EW	EW	1
AA	Asian	EW,AF	EW,AF,EA	1
EA	Latino	EA	EA	1
NA	Asian	EW,NA	EW,EA,NA	1
Parent EHR; Child Self-Report				
Latino	EW	EW	EW	1
Other/Uncertain	EW	EW	EW	2
White	EW,LT	EW	EW	2
White	EW,LT	EW	EW,NA	3
White	EW,LT	EW,NA	EW	1
White	EW,NA,LT	EW	EW	1
White	EW,NA,LT	EW	EW,NA	1
White	NA	EW	EW	3
White	EW,EA	EW	EW,EA	1
Other/Uncertain	EW,AA,LT	EW,AF	EW,AF	1
White	EW,EA,LT	EW	EW,EA	1