The Structure of Common Genetic Variation in United States Populations

Stephen L. Guthery, Benjamin A. Salisbury, Manish S. Pungliya, J. Claiborne Stephens, and Michael Bamshad

The common-variant/common-disease model predicts that most risk alleles underlying complex health-related traits are common and, therefore, old and found in multiple populations, rather than being rare or population specific. Accordingly, there is widespread interest in assessing the population structure of common alleles. However, such assessments have been confounded by analysis of data sets with bias toward ascertainment of common alleles (e.g., HapMap and Perlegen) or in which a relatively small number of genes and/or populations were sampled. The aim of this study was to examine the structure of common variation ascertained in major U.S. populations, by resequencing the exons and flanking regions of 3,873 genes in 154 chromosomes from European, Latino/Hispanic, Asian, and African Americans generated by the Genaissance Resequencing Project. The frequency distributions of private and common single-nucleotide polymorphisms (SNPs) were measured, and the extent to which common SNPs were shared across populations was analyzed using several different estimators of population structure. Most SNPs that were common in one population were present in multiple populations, but SNPs common in one population were frequently not common in other populations. Moreover, SNPs that were common in two or more populations often differed significantly in frequency from one population to another, particularly in comparisons of African Americans versus other U.S. populations. These findings indicate that, even if the bulk of alleles underlying complex health-related traits are common SNPs, geographic ancestry might well be an important predictor of whether a person carries a risk allele.

Health is primarily determined by conditions that are both common and have a complex pattern of inheritance (i.e., risk is influenced by a combination of several different genetic and environmental factors). A popular model of the genetic architecture of common disease posits that the minor-allele frequencies (MAFs) of genetic variants influencing susceptibility are often also common (i.e., $\geq$5%) and that such alleles are therefore old and found in multiple populations, rather than being rare and population specific. This model is known as the common-variant/common-disease (CV/CD) hypothesis.[1–4]

To facilitate testing of whether common variants influence susceptibility to common diseases, substantial efforts have been made to characterize the distribution of common alleles, particularly SNPs, among populations. This is important, because the extent to which common alleles explain risk of common disease across populations depends, in part, on how often alleles common in one population are common, or at least shared, in other populations.[5] Although only a relatively small number of alleles associated with complex disease have been reported, some alleles putatively associated with complex disease are common and are found at similar frequencies among populations,[6] whereas others, such as those that influence risk for atherosclerosis,[7] hypertension,[8] and acquired immunodeficiency syndrome[9] and some drug responses,[10] either are common in only a single population or differ significantly in frequency among groups. The extent to which such differences explain overall variation in heritable disease risk across populations remains to be determined.

A frequent claim about human population structure is that most common variation is shared among all populations.[11–13] This, of course, depends on how population boundaries are defined, but often cited to support such comments are the comparisons of SNP frequencies in pairs of populations in the HapMap data and the Perlegen data. Analyses of these data indicated that common SNPs were frequently both shared and common among populations of predominately African, Asian, and European ancestry.[14,15] However, population-genetics analysis was not the intended goal of either the HapMap or the Perlegen projects, and common, shared SNPs were oversampled by the ascertainment strategies used for each project.[16,17]

Other projects avoided this ascertainment bias by resequencing the entire sample from which SNP frequencies were estimated. Examples of these projects include the Environmental Genome Project (EGP),[18] the Seattle SNP project,[19,20] the Applera SNP project,[21] and the ENCyclopedia of DNA Elements (ENCODE) project.[22] Yet, comparison of common coding-SNP variation across U.S. populations was limited by the design of each of these studies as well (table 1). For example, the EGP used the Poly-

From the Department of Pediatrics, University of Utah, Salt Lake City (S.L.G.); Genaissance Pharmaceuticals, New Haven, CT (B.A.S.; M.S.P.); Motif BioSciences, New York (J.C.S.); and Departments of Pediatrics and Genome Sciences, University of Washington (M.B.), and Children's Hospital and Regional Medical Center (M.B.), Seattle

Received June 1, 2007; accepted for publication August 3, 2007; electronically published October 16, 2007.

Address for correspondence and reprints: Dr. Michael Bamshad, Department of Pediatrics, Division of Genetics and Developmental Medicine, University of Washington School of Medicine, 1959 NE Pacific Street, HSB RR349, Seattle, WA 98195. E-mail: mbamshad@u.washington.edu

Am. J. Hum. Genet. 2007;81:1221–1231. 0002-9297/2007/8106-0009$15.00
DOI: 10.1086/522239

Table 1. Comparison of Samples among Different Resequencing Projects

Project Name	No. of Chromosomes	No. of Individuals in Population Sample[a]				No. of Genes Resequenced
		EA	AfA	AsA	HA	
Seattle SNP	94	46	48	…	…	100
EGP	180	…	…	…	…	213
Applera SNP	78	40	38	…	…	11,624
ENCODE	128	32	32[b]	32[b]	…	10 × 500–kb regions
GRP	152	40	40	38	34	3,873

[a] EA = European American; Afa = African American; AsA = Asian American; HA = Latino/Hispanic American.

[b] Samples were ascertained from native populations, not U.S. populations.

morphism Discovery Resource, in which the sample identities are unknown, precluding comparisons across populations. The Seattle SNP and the Applera SNP projects resequenced samples only from self-identified African Americans and European Americans; Asian Americans and Latino/Hispanic Americans were not included. Furthermore, with the exception of Applera, all these projects resequenced a relatively modest number of genes, and several projects concentrated on genes with similar functional properties (e.g., genes involved in inflammation, immune defense, etc.).

To estimate how frequently common SNPs ascertained by resequencing are shared among major U.S. populations, we analyzed the Genaissance Resequencing Project (GRP) SNP frequency data from 3,873 genes on 152 chromosomes (~14 Mb of DNA sequence per individual) from self-identified African, Asian, Latino/Hispanic, and European Americans.[23–25] These population labels were used despite the controversy surrounding the correspondence between notions of race and population structure inferred from explicit genetic data, because they are the labels used by the National Institutes of Health (NIH), the U.S. Food and Drug Administration, and many, if not most, biomedical researchers. Insofar as these labels capture information about genetic ancestry, it is of substantial biomedical interest to understand the distribution of common variation across populations such defined.

Subjects and Methods

Laboratory Methods

The data set used herein consisted of genotypes ascertained by resequencing each exon (including the coding regions, 5′ UTR, and 3′ UTR), up to 100 bp upstream and downstream of each exon, up to 1,000 bp upstream of the transcription start site, and 100 bp downstream of the termination codon of 3,873 genes in 76 unrelated individuals (152 chromosomes), including 20 European Americans, 17 Latino/Hispanic Americans, 19 East Asian Americans, and 20 African Americans. All samples were obtained with institutional review board approval from individuals of self-identified group membership who participated in the GRP.[23–25] Individuals were sampled from two locations in the United States—Anaheim, CA, and Miami, FL. Sampling 40 chromosomes in a population provides a 95% probability of detecting a SNP with a true population MAF ⩾5% (i.e., the common polymorphisms in which we are most interested). These data were provided as anonymous genotypes, so the identities of the genes that were resequenced and the location of each SNP were unknown to M.B. and S.L.G, the two authors responsible for the analysis. Accordingly, this precluded the performance of analyses that require such information (e.g., stratifying estimates of SNP sharing on the basis of functional and/or structural similarities among genes).

For each individual, a blood sample was obtained, and lymphocytes were immortalized as Epstein-Barr virus–transformed cell lines. Genomic DNA was extracted using standard techniques and was used as the template for all subsequent PCRs. Sequencing reactions were performed using Applied Biosystems Big Dye Terminator chemistry, essentially in accordance with the manufacturer's protocol, and results were analyzed on ABI Prism 3700/3730 DNA Analyzers. The presence of a polymorphism was confirmed by sequencing both strands of DNA. After initial data processing with the ABI instruments, sequence trace files were reanalyzed with the Phred program, which adds a quantitative base-quality value. This base-quality value provides a probabilistic estimate of the correctness of the base call. The quality values are the log of the probability that the base call is correct, such that a Phred value of 20 corresponds to a 99% probability that the base call is accurate, whereas a Phred value of 30 corresponds to a 99.9% probability that the base call is accurate. A minimum Phred value of 20 was used as a threshold. The sequence was assembled with consensus sequence with use of the Phrap program, and potential polymorphisms were identified using the Polyphred program. All sequence assemblies (i.e., reads plus consensus sequence and tagged polymorphisms) were then compiled into one Consed project for review. Potential polymorphisms were catalogued and underwent human review of original trace files. This final list of verified polymorphisms was loaded into a database, where they could be further reviewed.

Sample mix-ups were controlled in three ways, by (1) genotyping several triads (i.e., parents and offspring) that were included on each sequencing plate, (2) confirming the identity of each sample by use of a subset of the Combined DNA Index System microsatellite markers each time a new master plate of DNA was generated, and (3) positioning a "null sample" in the same well on each sequencing plate to ensure that each plate was oriented properly. Hardy-Weinberg equilibrium (HWE) for each SNP within each population was calculated on the basis of a comparison of observed and expected heterozygosities and significance, tested against a χ^2 distribution. For <5% of SNPs, the genotype frequencies differed significantly (i.e., $P < .05$) from HWE. This result suggests there were no gross systematic errors in base calls and/or sample mix-ups.

Statistical Analysis

Genotypes were available for 96.5%–99.9% (mean 98.4%) of sites per individual. For each SNP, the minor allele was defined as the allele with the lowest frequency in the total chromosome sample. To assess the degree of allele sharing among populations, we first determined the proportion of SNPs that were common in each population, defined as the number of SNPs with a MAF ⩾5% or ⩾10% in each population. We used both Spearman rank and Pearson correlation coefficients to calculate the pairwise corre-

lations between the MAFs in each population. Simulations to generate expected values between populations of similar sample size for the proportion of SNPs shared, minor-SNP frequency differences, Spearman rank correlation coefficients for minor-SNP frequencies, and pairwise F_{ST} values were performed by randomly sampling 40 individuals without replacement from the total sample of 76 individuals, with the use of Floyd's ordered hash table algorithm implemented in the surveyselect procedure of SAS 9.1.3., then by randomly allocating them into two groups to be used for analysis. The sample size of 20 individuals for each population matches the maximum sample size of each population from which empirical data were available. Reported values and SDs were generated from 1,000 such simulated data sets. Contour plots were constructed using the kernel-density estimation procedure of SAS 9.1.3, with use of a bandwidth multiplier equal to 1 and grid points of 60×60.

We performed a principal-components analysis and distance-based cluster analysis, using the number of sequence differences between two individuals for all pairwise comparisons of individuals as the distance metric. Eigenvalues for the principal-components analysis are shown in figure 1*A*. We used the unweighted pair group method with arithmetic mean (UPGMA), implemented in SAS and PHYLIP[26] for cluster analysis, and estimated the number of clusters where the pseudo *F*-test statistics were maximized (fig. 1*B*).[27] The distance matrix, principal-components analysis, and pseudo *F*-test statistics were generated in SAS 9.1.3. A radial tree depicting the relationships between individuals was drawn in TREEVIEW.[28] The estimated log-likelihood of the probability of the data over the range of *K* is demonstrated in figure 1*C*.

For the model-based cluster analysis, we used STRUCTURE 2.0,[29] using the correlated allele-frequency model.[30] Among the 63,127 SNPs, we selected those in the top 10th percentile for expected heterozygosity, since this is a readily available measure and since data from Pritchard et al. suggest that markers with high expected heterozygosity are informative when used to infer population structure.[31] These selected markers had low pairwise linkage disequilibrium. We used the following settings for the STRUCTURE run: admixture model, correlated markers, $K = 1$–6, a length of 100,000 for the burn-in period, and 100,000 repetitions following the burn-in period. The estimated log-likelihood of the probability of the data over the range of *K* is demonstrated in figure 1*C*.

For each biallelic locus, Wright's locus-by-locus fixation index, F_{ST}, was estimated using

$$F_{ST} = 1 - \frac{\sum_j 2p_j\frac{1-p_j}{j}}{2\bar{p}(1-\bar{p})},$$

where p_j is the MAF in population j and $\bar{p}$ is the MAF in all j populations.[32] Total F_{ST} is expressed as an average over all alleles.

Results

A total of 63,127 SNPs were identified in 3,873 genes (data available at the Bamshad Lab Web site). Of these SNPs, 24,982 (39.6%) were singletons, meaning that the minor allele was observed on only one chromosome (fig. 2). Of all singletons, 45% (11,244) were observed in African Americans, and the lowest number of singletons was found in Asian Americans (table 2). More than half of all SNPs (35,385, or 56%) were private—that is, observed in only one population (table 2). The majority of private SNPs were rare; 70.6% were singletons and 99% were observed at a frequency of <5% (table 2). The percentage of all nonsingleton SNPs (i.e., the number of SNPs in a population divided by the total number of SNPs identified in all populations combined) found in any single group ranged between 50% in Asian Americans and 83% in African Americans.

The absolute number of SNPs with an MAF of either ≥5% or ≥10% (i.e., common SNPs) was highest in African

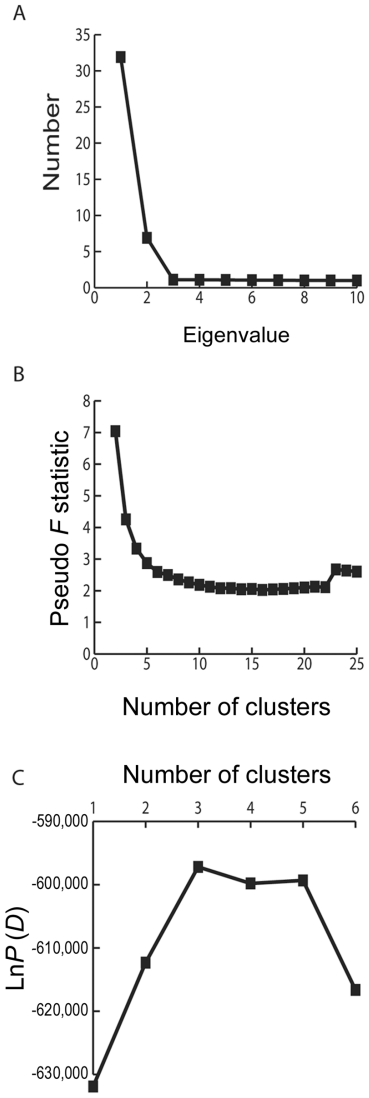

Figure 1. Summary statistics for data-reduction methods used to estimate population structure. *A,* Eigenvalues versus the number of principal components. *B,* Number of clusters plotted as a function of the pseudo *F* statistic obtained from the UPGMA algorithm. *C,* Number of clusters plotted as a function of Ln*P*(*D*) from STRUCTURE.

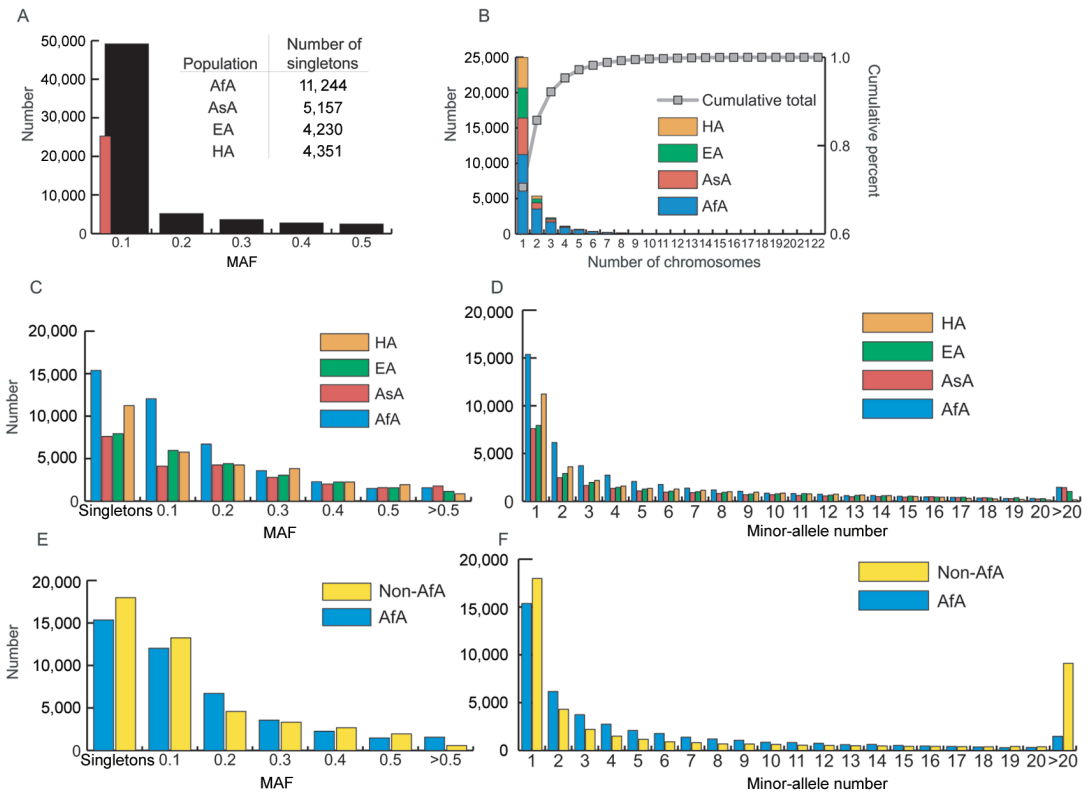

Figure 2. Site-frequency distributions for SNP data from 3,873 genes. *A,* SNP site–frequency distribution for the total sample. Of a total 63,127 SNPs (*black bars*) in the data set, 39% (n = 24,982) were singletons (*red bar*). *B,* Number and distribution of private SNPs in each population determined from the site-frequency distribution of the total sample. The majority of private SNPs were observed in seven or fewer chromosomes, illustrated by cumulative frequency (*gray line*). *C* and *D,* SNP site–frequency distribution for each population. *E* and *F,* SNP site–frequency distributions for African Americans versus non–African Americans.

Americans, and, for every three common SNPs in European Americans, there were four common SNPs in African Americans. For SNPs with an MAF ⩾10%, pairwise population comparisons showed that 67%–96% of SNPs common in one population were at least present in both populations (fig. 3*A*). However, only 44%–72% of such SNPs were common in both populations (fig. 3*C*). These findings were similar when SNPs with an MAF ⩾5% in at least one population were compared (fig. 3*B* and 3*D*). Of the 23,220 SNPs with MAF ⩾10% in at least one population, 7,436 (32%) were common in all four populations, and 13,285 (57%) were present in all four populations. Additionally, common SNPs were often not shared among the African American population and other U.S. populations. These results indicate that, in this sample of U.S. populations, slightly more than half of all common SNPs are shared among populations, but two-thirds of them are not common in all populations.

Even when common alleles are common in two or more groups, it is important to know whether the frequencies of such alleles differ substantially between groups. Indeed, there is predicted to be greater variation among the frequencies of common alleles than among those of rare alleles. To what extent the frequencies of common SNPs were similar among populations was assessed by estimating the pairwise correlation coefficient between frequencies of SNPs with an MAF ⩾10% in both populations. The MAFs of common SNPs varied widely between groups (table 3). They were most highly correlated between Latino/Hispanic Americans and European Americans (ρ = 0.84) and were least correlated between African Americans and Asian Americans (ρ = 0.26). Contour plots demonstrated that pairwise correlation coefficients were consistently lower between African Americans and non–African American populations (fig. 4); results were similar when SNPs ⩾5% were compared between populations. This is due, in part, to the presence of more high-frequency SNPs in African Americans, which leads to greater differences among MAFs between populations (fig. 5). Therefore, whereas rare alleles contributing to common disease might be less likely to be found in multiple populations, common alleles influencing risk of common disease are likely to vary more in frequency among groups.

To examine the relationship between the MAF and the sharing of SNPs among groups, we estimated the proportion of SNPs shared (i.e., present in both populations) be-

Table 2. Summary of Private Genetic Variation in U.S. Populations

SNP Frequency	No. of SNPs (% in Population)[a]				
	AfA	EA	AsA	HA	Total
Singletons	11,244	5,157	4,230	4,351	24,982
MAF ⩾5%	7,498 (40)	1,579 (23)	820 (16)	506 (10)	10,403 (29)
MAF ⩾10%	2,297 (12)	712 (11)	263 (5)	99 (2)	3,371 (10)
Total	18,742	6,736	5,050	4,857	35,385

[a] AfA = African American; EA = European American; AsA = Asian American; HA = Latino/Hispanic American.

tween pairs of populations for SNPs with frequencies ranging from all "nonsingletons" to ⩾40% in either population. For all pairwise comparisons, the proportion of SNPs shared between groups was substantially less than that shared among individuals who were randomly sorted into two populations, and, as the frequency of a SNP increased, it was more likely to be shared between populations (fig. 5*A*). Indeed, >95% of SNPs with an MAF ⩾20% and 99% of SNPs with an MAF ⩾30% were shared between pairs of populations. However, despite being more frequently shared between groups, SNPs with a higher frequency were also associated with a greater mean difference in frequency between populations (fig. 5*B*). Accordingly, the correlation between MAFs between groups decreased as SNP frequency increased (fig. 5*C*). One outcome of this phenomenon is that the effect of population structure among pairs of groups, as estimated by Wright's fixation index, or F_{ST}, is more pronounced when calculated using SNPs with a higher allele frequency (fig. 5*D*). Therefore, the difference between the frequencies of a given SNP in

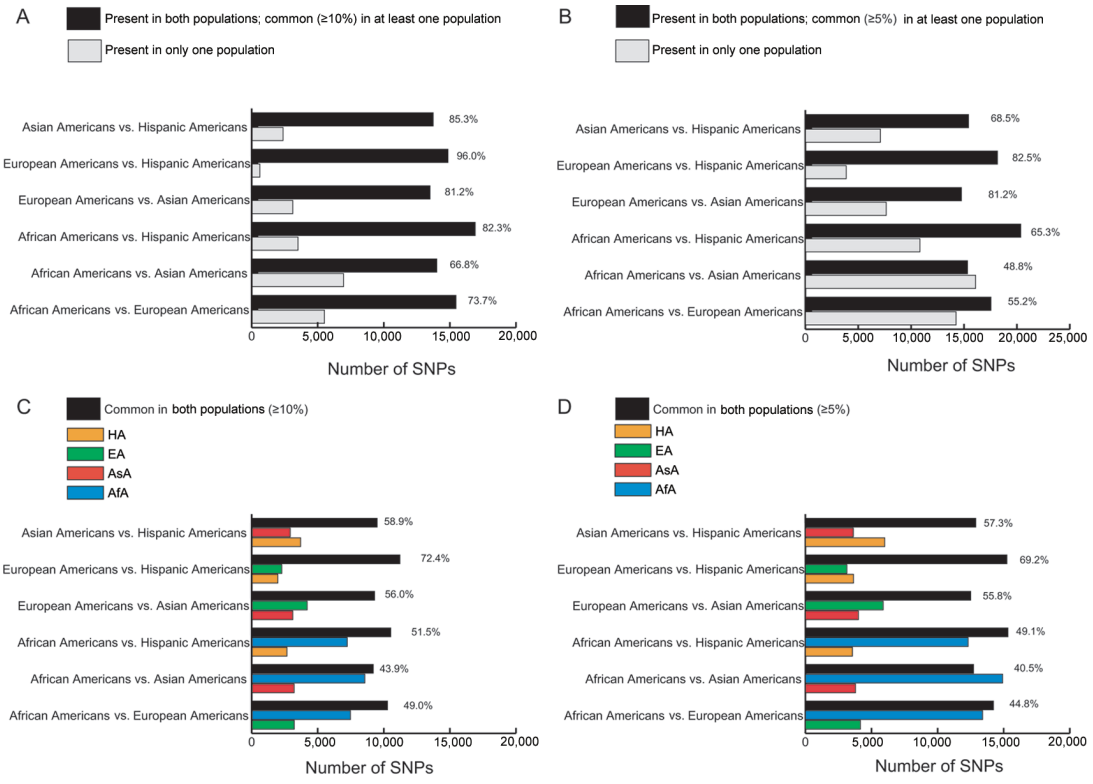

Figure 3. Distribution of common SNPs among Latino/Hispanic, African, Asian, and European Americans. *A* and *B,* The percentage of SNPs that are common (i.e., ⩾10%) in at least one population but are found in both populations (*black bars*) is high overall but varies from ~74% to 96%. A modest percentage of common SNPs that are common in at least one population are absent in the other populations (*gray bars*). *C* and *D,* The percentage of common SNPs common in both populations (*black bars*) compared with SNPs common in only each population compared: African Americans (AfA) (*blue bars*), Asian Americans (AsA) (*red bars*), European Americans (EA) (*green bars*), and Latino/Hispanic Americans (HA) (*orange bars*). Overall, only a modest percentage (44%–72%) of SNPs common in at least one population are common in both populations. A substantial proportion of common SNPs in African Americans are common only in African Americans.

Table 3. MAF Correlation of Common SNPs between Populations

Population	AfA	EA	AsA	HA
AfA	...	.36[a]	.26	.45
EA	.19[b]	...	.58	.84
AsA	.084	.21	...	.62
HA	.28	.65	.25	...

NOTE.—AfA = African American; AsA = Asian American; EA = European American; HA = Latino/Hispanic American.

[a] Correlation between SNPs with an MAF ⩾10%, where MAF is defined for the whole sample (n = 152 chromosomes).

[b] Spearman rank correlation between SNPs with an MAF ⩾10% in both populations where MAF is defined in each subpopulation (below the diagonal).

two different populations was positively correlated with MAF, and the magnitude of the correlation varied among pairwise population comparisons.

For each comparison (i.e., fig. 5*A*–5*D*), it was of interest to assess the departure from expectations under a simple model in which individuals mated at random (i.e., no population structure). Accordingly, we created a thousand data sets in which individuals were randomly allocated into two groups composed of 20 individuals each and repeated each of these analyses (see the "Subject and Methods" section for details). These simulations demonstrated that pairwise sharing of SNPs, differences in MAF, the correlation of SNP frequencies among populations, and pairwise F_{ST} estimates differed more significantly between all pairs of populations than expected by chance. The departure from expectations was consistently greatest for comparisons between African Americans and Asian Americans and was least for comparisons between European Americans and Latino/Hispanic Americans.

Each SNP in the GRP data set does not have an equal a priori probability of being a functional SNP and, therefore, influencing a phenotype, and yet SNPs influencing variation in risk for a health-related trait must be functionally distinct from other alleles. Therefore, it is of particular interest to know to what extent functional SNPs are shared among populations. Data from direct testing for functionality for each SNP are unavailable. However, SNPs that result in nonsynonymous substitutions or nonsense mutations are more likely to have functional consequences than are SNPs that cause synonymous substitutions. Accordingly, we examined a subset of 28,810 SNPs from 1,928 genes for which information was available about whether a coding SNP resulted in a synonymous substitution versus a nonsynonymous substitution or nonsense mutation.

In the total sample, the number of SNPs predicted to

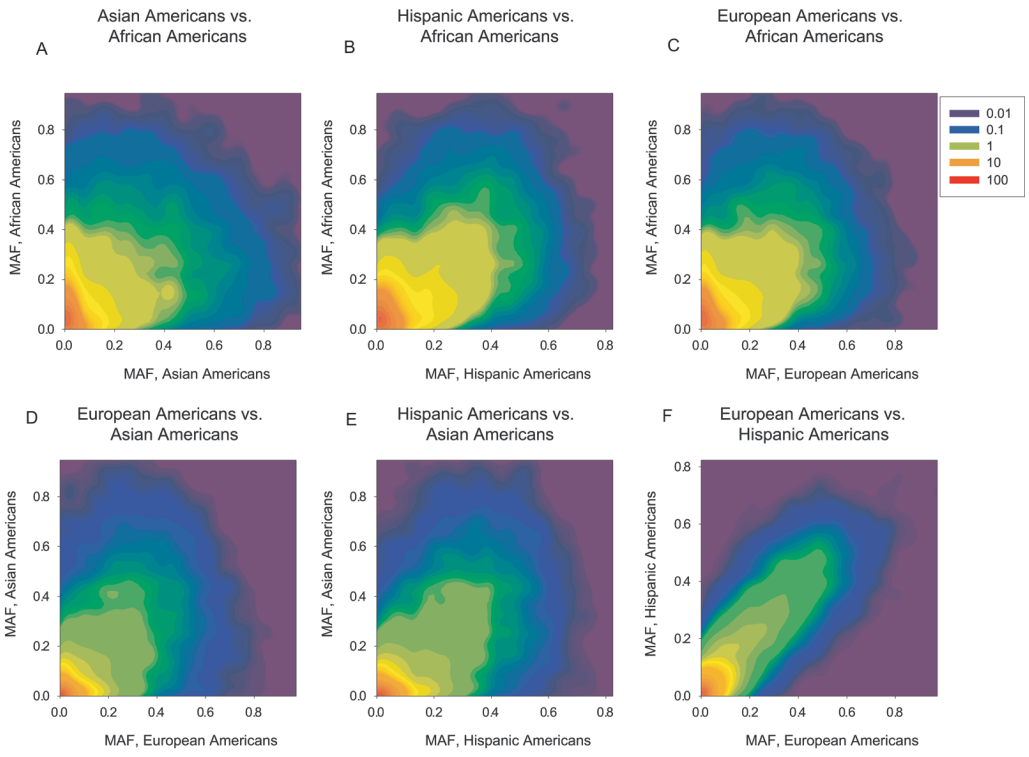

Figure 4. Contour plot of minor-SNP frequencies between pairs of populations. Plots compare frequencies of SNPs (n = 38,145) excluding singletons. Each plot represents a scatterplot with minor-SNP frequency from a given population on each axis. Plots are divided into 3,600 grids (60 × 60 grids), and the number of data points within each grid is color coded. For example, purple represents 0.01 data points per grid, and red represents 100 data points per grid (see legend in the upper right-hand corner).

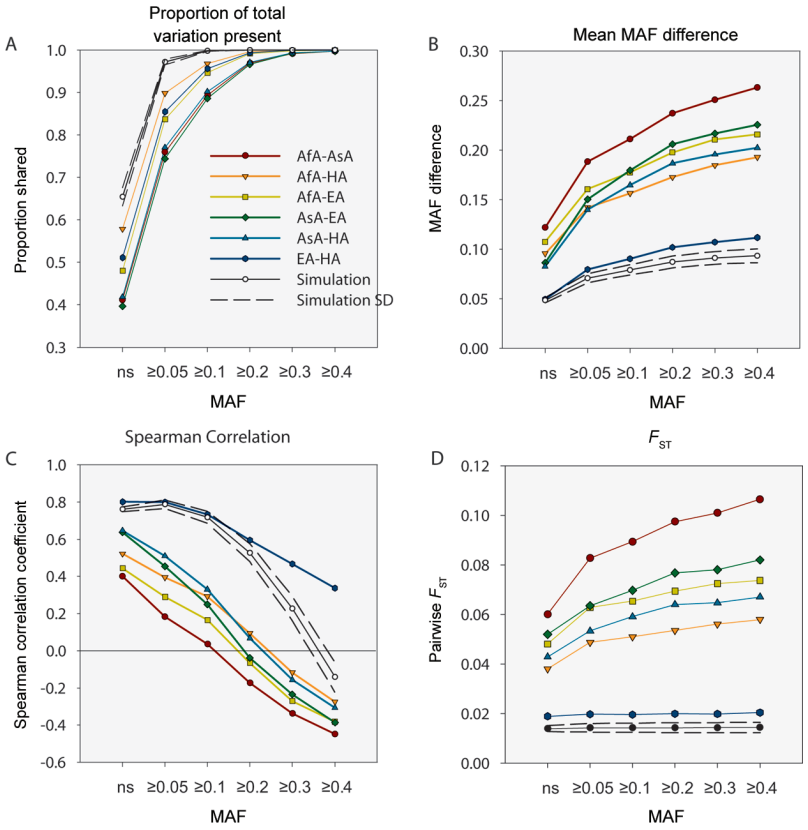

Figure 5. Measures of SNP sharing among Latino/Hispanic (HA), African (AfA), Asian (AsA), and European (EA) Americans. For all figures, the *X*-axis represents overlapping bins (i.e., >0.05 represents all SNPs with MAF >0.05), and MAF is calculated across all 152 chromosomes. When two populations are compared, MAF is calculated separately for each population. *A,* Pairwise comparisons of the proportion of SNPs shared between populations. *B,* Mean differences of pairwise comparisons of MAF between SNPs. *C,* Spearman rank correlation coefficients among pairwise comparisons of MAF between SNPs. *D,* Pairwise F_{ST} estimates between SNPs. The solid black line in each figure represents the mean value, and the dotted lines indicate the CI of values estimated from 1,000 data sets in which individuals were randomly distributed into pairs of populations (see text for details). ns = nonsingletons.

cause nonsynonymous substitutions (3,810) versus synonymous substitutions (3,776) was nearly identical. This averages to ~2 synonymous or nonsynonymous SNPs per gene and is similar to the 1.5–3 coding SNPs per gene found in other analyses of resequencing efforts.[21,33] Fifty-eight SNPs predicted to cause nonsense mutations were found in the GRP data set. The remaining 21,166 SNPs were not located in coding regions. The site-frequency distributions for the total sample, African Americans, and non–African Americans were similar for both synonymous versus nonsynonymous and nonsense mutations combined (fig. 6). Compared with the frequency with which all SNPs ⩾10% were shared between groups, coding SNPs ⩾10% were shared between populations about as often (table 4), although common SNPs predicted to cause either nonsynonymous substitutions or nonsense mutations were shared slightly less frequently than were all SNPs combined.

Estimates of the extent to which common alleles are shared among populations are potentially confounded by population admixture. In individual self-identified African Americans, ~80% of their ancestry is of West African origin, but this contribution ranges from ~20% to 100%.[34] Similarly, the genetic composition of self-identified Latinos/Hispanics in the United States varies depending on

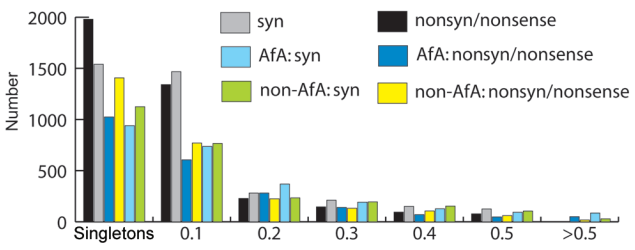

Figure 6. Site-frequency distribution of synonymous (syn) (*gray bars*) and nonsynonymous (nonsyn) (*black bars*) SNPs for the total sample and for African Americans (AfA) versus non–African Americans.

Table 4. Proportion of Coding SNPs Common in at Least One Population That Are Common in Both

	Percentage (%)				
Population	AfA	EA	AsA	HA	Non-AfA
AfA	...	46.0[a]	39.6	48.2	48.6
EA	48.0[b]	...	51.0	72.3	...
AsA	42.0	56.5	...	54.0	...
HA	50.9	72.6	59.8	...	...
Non-AfA	49.7	...	...	...	...

NOTE.—AfA = African American; AsA = Asian American; EA = European American; HA = Latino/Hispanic American.

[a] Sharing of nonsynonymous or nonsense SNPs between populations (above the diagonal).

[b] Sharing of synonymous SNPs between populations (below the diagonal).

the geographical locale where individuals were sampled—that is, more African American admixture in Latino/Hispanic Americans from Puerto Rico is more prevalent in the Southeast, whereas more admixture with Native Americans in Latino/Hispanic Americans from Mexico is more common in the Southwest.[35] We assessed population structure and individual ancestry proportions in the total sample, to verify that our SNP data supported assignment of the 76 genotyped individuals to four populations and to assess the impact of admixture on our estimates of sharing of common SNPs among U.S. populations. The posterior probabilities that $k = 3$ versus $k = 4$ were nearly identical. All individuals who identified themselves as African American or Asian American were sorted into separate groups. However, individuals who identified themselves as European American or Latino/Hispanic American were sorted into the same cluster when $k = 3$ and even when $k = 4$. Hispanic Americans frequently were misclassified with European Americans.

For both self-identified European Americans and Asian Americans, the average fraction of ancestry shared with other groups was low and varied little among individuals. The mean "African" contribution to individual African American ancestry was ~83% but varied from 62% to 100% (fig. 7). Similarly, the African American and European American contribution to individual Latino/Hispanic American ancestry varied from 2% to 41% and from 45% to 98%, respectively, when analyzed with $k = 4$. These admixture estimates are comparable to those obtained in other studies[36,37] and suggest that it is reasonable, given the limitations imposed by the small sample size, to extrapolate estimates of SNP sharing among these sample populations to the U.S. population in general. It also means that these estimates of SNP sharing are conservative in contrast to pairwise comparisons of native sub-Saharan Africans, Europeans, and East Asians. Finally, it underscores the observation that inference of whether an individual African American or Latino/Hispanic American shares an allele common in other populations is most appropriately based on explicit genetic data or ancestry information rather than a racial label.

Discussion

The primary aim of this study was to understand the structure of common variation ascertained by resequencing versus genotyping a large fraction of human genes in four major U.S. populations. Our concentration on genetic variation and SNP ascertainment via resequencing is motivated by several concerns. First, the ascertainment schemes for both the Perlegen and HapMap data sets involved resequencing in a small discovery panel, followed by genotyping of a larger sample to estimate allele frequencies. This strategy biased estimates of allele sharing among populations because common SNPs were both more likely to be found and more likely to be shared among groups. Fully sampling variation in the genes studied via resequencing provides an opportunity for a more robust population-genetics analysis. Second, functional SNPs that influence common phenotypes might be more likely to be found in the flanking and coding regions of genes rather than in intergenic regions, so estimates of sharing of coding SNPs between populations might be more valuable for evaluating genotype-phenotype relationships in different groups. Finally, the use of common SNPs in disease-association studies is experimentally more tractable and more valuable for the development of diagnostics and therapeutics, because, for the same phenotypic effect, a common allele explains a larger fraction of the population-attributable risk than does a rare allele.

A simple comparison of data from the GRP with other large SNP data sets generated from a comparable number of chromosomes by Perlegen, HapMap, and the National Institute of Environmental Health Sciences is instructive. In the total population and in each separate population, a smaller proportion of SNPs in the GRP data have a MAF $\geqslant$10%, and a smaller proportion of private SNPs are common. For example, in the Perlegen data set, an average of ~63% of all SNPs and 37% of private SNPs were reported to have an MAF $\geqslant$10%, whereas an average of 45% of all SNPs and 9.5% of private SNPs had an MAF $\geqslant$10% in the GRP data set. As demonstrated elsewhere, the higher fraction of SNPs with an MAF $\geqslant$10% in both the Perlegen data and HapMap data sets is due in part to an ascertainment bias that resulted in the oversampling of common SNPs.[16] Likewise, a larger proportion of SNPs with an MAF $\geqslant$5% were shared among populations in the Perlegen data versus the GRP data (table 5).

The structure of common-SNP variation differed substantially in African Americans compared with all other U.S. populations studied. The largest absolute numbers of SNPs, common SNPs, and private SNPs were found in African Americans. African Americans exhibited the highest percentage of rare SNPs (64%) and the lowest percentage of common SNPs (36%), and nearly half of all SNPs (44%) in African Americans were private. Correlation of the fre-

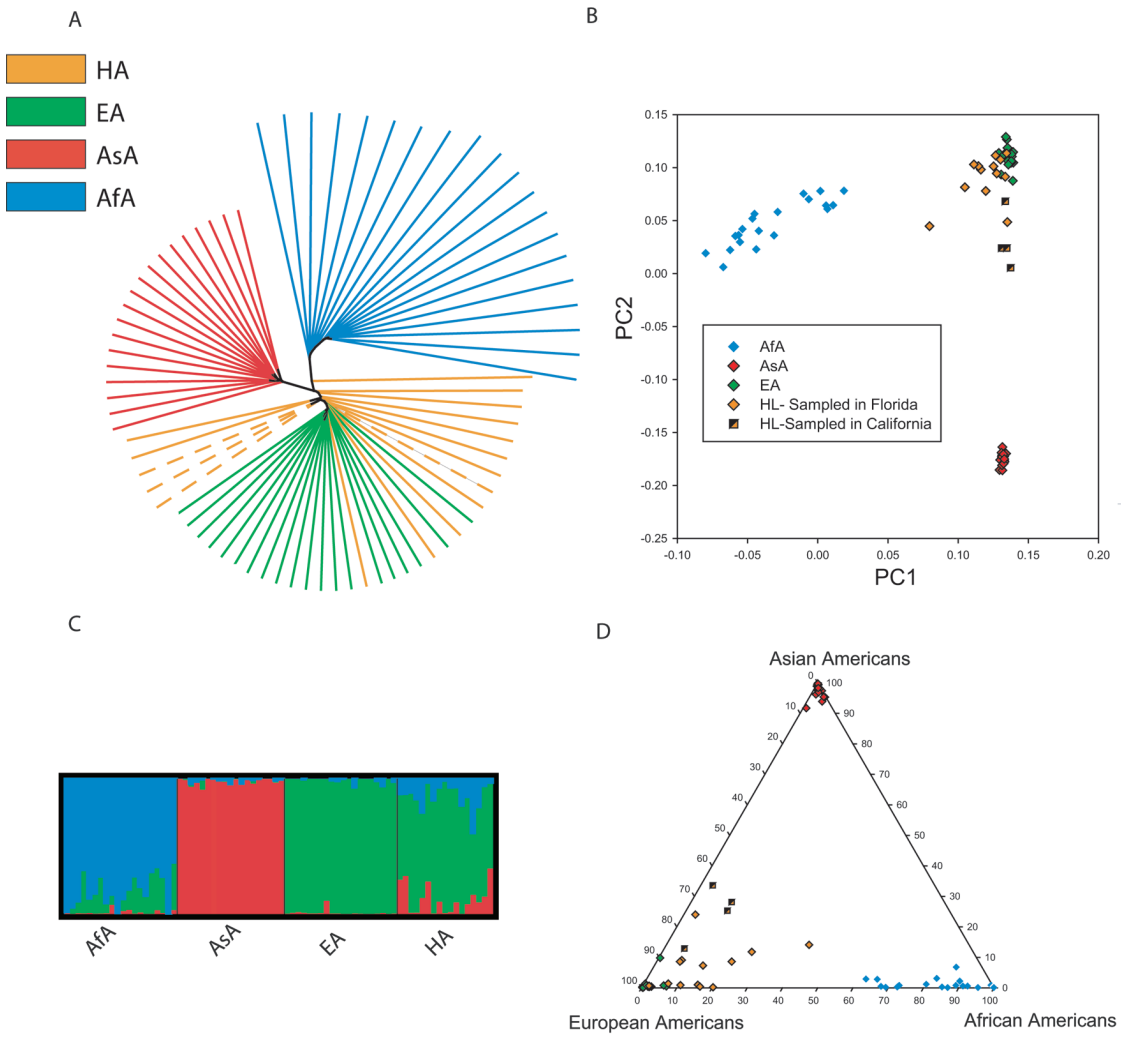

Figure 7. Estimation of population structure in GRP samples. AfA = African American; AsA = Asian American; EA = European American; HA = Latino/Hispanic American. *A,* Phylogenetic network based on genetic distances with the use of UPGMA. *B,* Plot of principal components (PCs) estimated from a genetic-distance matrix. *C,* Stacked bar chart with inferences from results of a model-based cluster analysis with the use of STRUCTURE 2.0. Each bar represents an individual, and each bar is divided according to the fraction of cluster membership. *D,* Triangle plot illustrating the percentage of African, Asian, and European American ancestry of each individual (indicated by colored shapes, as given in panel B) estimated from STRUCTURE 2.0.

quency of common SNPs between populations was consistently the lowest in comparisons with African Americans. Common SNPs predicted to cause protein truncation or nonsynonymous substitutions were also shared less frequently between African Americans and other U.S. populations than were common SNPs predicted to cause synonymous substitutions. Collectively, these observations suggest that a sizable fraction of genetic factors that influence common health-related traits might vary between African Americans and other U.S. populations. Therefore, it would be prudent to consider oversampling African Americans or to develop new initiatives to characterize common variation in African Americans.

Alleles that are subject to negative selection and are therefore deleterious might contribute substantially to genetic variation underlying complex health-related traits.[38] Negative selection appears to be widespread among SNPs predicted to cause nonsynonymous substitutions,[21,33] but it appears to be weak enough that such SNPs can reach modest frequencies via genetic drift and form a substantial fraction of extant genetic variation. Accordingly, it is of particular interest to understand the population structure of SNPs predicted to cause nonsynonymous substitutions. In the GRP data set, there was an excess of low-frequency nonsynonymous or nonsense SNPs compared with synonymous SNPs. Similar differences in the site-frequency distributions among different functional classes of SNPs have been observed in other data sets.[21,33,39,40] Among SNPs $\geq$10%, there were more synonymous SNPs than nonsynonymous or nonsense SNPs, and synonymous SNPs

Table 5. Comparison of Common SNPs between GRP and Perlegen Data

	MAF Correlation Coefficient		Proportion Shared		Mean (Median) MAF Difference	
Population	GRP	Perlegen	GRP	Perlegen	GRP	Perlegen
AfA-AsA	.18	.176	.753	.854	.19 (.16)	.2 (.17)
AfA-EA	.29	.304	.8365	.91	.16 (.13)	.17 (.14)
AsA-EA	.45	.4	.744	.839	.15 (.15)	.16 (.14)

NOTE.—AfA = African American; AsA = Asian American; EA = European American.

were shared between populations more often than were nonsynonymous or nonsense SNPs. Among nonsynonymous or nonsense SNPs that were common in at least one population, 55% were present in all four populations, and 28% were common in all populations. Therefore, on average, about one of every two to three common SNPs, whether coding or noncoding, are shared among populations.

The extent to which SNP frequencies differ across populations is also influenced by the effect of local positive selection and balancing selection, the former increasing differentiation among groups and the latter decreasing it.[41] Although the overall effect of these forces on the structure of common variation is difficult to predict, it highlights one potential limitation of the extrapolation of our results to the entire genome. That is, if many of the genes in the GRP data set have been affected by selection in a similar fashion, the cumulative effect could be to skew the distribution of common SNPs among groups. Such a scenario is unlikely. Even so, the genes resequenced in the GRP (e.g., transcriptional regulators, signal transducers, and drug-metabolism enzymes) were chosen because they were considered candidates for complex diseases and, therefore, genes for which there is substantial interest in empirical estimates of the structure of common variation.

The extent to which rare versus common SNPs that are either private or shared among populations contribute to the ways in which humans differ from one another is an empirical question for which there are too few data to answer comfortably, much less definitively. Nevertheless, the distribution of genetic variation is, for major U.S. populations at least, becoming clearer. Analysis of the GRP data shows that the effects of population structure on common-SNP variation in U.S. populations is greater than has previously been reported, particularly for common SNPs that might be deleterious (i.e., coding SNPs). These findings underscore the necessity of considering ancestry and the effects of admixture in the design and execution of association-mapping studies.[42] Furthermore, studies of the effects of common variation on health-related traits have concentrated on European Americans, with the expectation that the findings could be extended to other U.S. populations. There is increasing recognition that some alleles associated with disease risk in European Americans are found at substantially different frequencies or virtually not at all in other populations, particularly African Americans, which makes it more difficult to study their effects.[6–10,43] Greater emphasis is therefore needed on understanding patterns of variation and the relationship of risk alleles to health-related traits in U.S. populations other than European Americans.

To date, all studies of the structure of common variation in U.S. populations, including this one, have been limited to surveys of a small number of individuals collected from a few geographic regions. Coupled with the heterogeneity of U.S. populations distinguished by commonly used racial labels, many scientists are likely to perceive our knowledge of the distribution of common variation in the U.S. to be incomplete. We agree and suggest that a more comprehensive understanding will be contingent on sampling a much larger spectrum of the diverse groups with varied geographical ancestries who now reside in the U.S.

Acknowledgments

We thank L. Jorde, D. Nickerson, H. Tang, D. Witherspoon, and all the anonymous reviewers, for comments and suggestions on the manuscript. This work was supported by NIH grants AI065357 (to M.B.) and DK069513 (to S.G.).

Web Resource

The URL for data presented herein is as follows:

Bamshad Lab, http://depts.washington.edu/bamshad/data/

References

1. Chakravarti A (1999) Population genetics—making sense out of sequence. Nat Genet 21:56–60
2. Lander ES (1996) The new genomics: global views of biology. Science 274:536–539
3. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137
4. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17:502–510
5. Ioannidis JP, Ntzani EE, Trikalinos TA (2004) 'Racial' differences in genetic effects for complex diseases. Nat Genet 36: 1312–1318
6. Lohmueller KE, Mauney MM, Reich D, Braverman JM (2006) Variants associated with common disease are not unusually differentiated in frequency across populations. Am J Hum Genet 78:130–136
7. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH (2005) Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in *PCSK9*. Nat Genet 37:161–165
8. Rosskopf D, Manthey I, Siffert W (2002) Identification and ethnic distribution of major haplotypes in the gene *GNB3* encoding the G-protein β3 subunit. Pharmacogenetics 12: 209–220
9. Gonzalez E, Dhanda R, Bamshad M, Mummidi S, Geevarghese R, Catano G, Anderson SA, Walter EA, Stephan KT, Hammer MF, et al (2001) Global survey of genetic variation in *CCR5,* RANTES, and MIP-1α: impact on the epidemiology of the HIV-1 pandemic. Proc Natl Acad Sci USA 98:5199–5204

10. Tate SK, Goldstein DB (2004) Will tomorrow's medicines work for everyone? Nat Genet 36:S34–S42
11. Rebbeck TR, Halbert CH, Sankar P (2006) Genetics, epidemiology, and cancer disparities: is it black and white? J Clin Oncol 24:2164–2169
12. Weigmann K (2006) Racial medicine: here to stay? The success of the International HapMap Project and other initiatives may help to overcome racial profiling in medicine, but old habits die hard. EMBO Rep 7:246–249
13. Abecasis G, Tam PK, Bustamante CD, Ostrander EA, Scherer SW, Chanock SJ, Kwok PY, Brookes AJ (2007) Human Genome Variation 2006: emerging views on structural variation and large-scale SNP analysis. Nat Genet 39:153–155
14. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
15. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079
16. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15:1496–1502
17. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. Genome Res 15:1468–1476
18. Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA (2004) Pattern of sequence variation across 213 environmental response genes. Genome Res 14:1821–1831
19. Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat Genet 33:518–521
20. Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD, Eberle MA, Kruglyak L, Nickerson DA (2004) Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. Am J Hum Genet 74:610–622
21. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al (2005) Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157
22. ENCODE (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636–640
23. Salisbury BA, Pungliya M, Choi JY, Jiang R, Sun XJ, Stephens JC (2003) SNP and haplotype variation in the human genome. Mutat Res 526:53–61
24. Schneider JA, Pungliya MS, Choi JY, Jiang R, Sun XJ, Salisbury BA, Stephens JC (2003) DNA variability of human genes. Mech Ageing Dev 124:17–25
25. Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, et al (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489–493
26. Felsenstein J (2004) PHYLIP (Phylogeny Inference Package) release 3.6. Department of Genome Sciences, University of Washington, Seattle
27. Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Statistics 3:1–27
28. Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12: 357–358
29. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
30. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587
31. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181
32. Hartl DL, Clark AG (1997) Principles of population genetics. Sinauer Associates, Sunderland, MA
33. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci USA 102:7882–7887
34. Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N, et al (2003) Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 112:387–399
35. Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, Tsai HJ, Naqvi M, Phong A, Ung N, et al (2006) Population stratification confounds genetic association studies among Latinos. Hum Genet 118:652–664
36. Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB (2003) Human population genetic structure and inference of group membership. Am J Hum Genet 72:578–589
37. Tsai HJ, Shaikh N, Kho JY, Battle N, Naqvi M, Navarro D, Matallana H, Lilly CM, Eng CS, Kumar G, et al (2006) β2-adrenergic receptor polymorphisms: pharmacogenetic response to bronchodilator among African American asthmatics. Hum Genet 119:547–557
38. Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M (2003) Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc Natl Acad Sci USA 100: 15754–15757
39. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, et al (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231–238
40. Sunyaev S, Ramensky V, Bork P (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 16:198–200
41. Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4:99–111
42. Reiner AP, Ziv E, Lind DL, Nievergelt CM, Schork NJ, Cummings SR, Phong A, Burchard EG, Harris TB, Psaty BM, et al (2005) Population structure, admixture, and aging-related phenotypes in African American adults: the Cardiovascular Health Study. Am J Hum Genet 76:463–477
43. Kugathasan S, Loizides A, Babusukumar U, McGuire E, Wang T, Hooper P, Nebel J, Kofman G, Noel R, Broeckel U, et al (2005) Comparative phenotypic and *CARD15* mutational analysis among African American, Hispanic, and white children with Crohn's disease. Inflamm Bowel Dis 11:631–638