nature genetics

Implications of biogeography of human populations for 'race' and medicine

Sarah A Tishkoff[1] & Kenneth K Kidd[2]

In this review, we focus on the biogeographical distribution of genetic variation and address whether or not populations cluster according to the popular concept of 'race'. We show that racial classifications are inadequate descriptors of the distribution of genetic variation in our species. Although populations do cluster by broad geographic regions, which generally correspond to socially recognized races, the distribution of genetic variation is quasicontinuous in clinal patterns related to geography. The broad global pattern reflects the accumulation of genetic drift associated with a recent African origin of modern humans, followed by expansion out of Africa and across the rest of the globe. Because disease genes may be geographically restricted due to mutation, genetic drift, migration and natural selection, knowledge of individual ancestry will be important for biomedical studies. Identifiers based on race will often be insufficient.

One of the grand challenges of the post–human-genome-sequence era is to "develop a detailed understanding of the heritable variation in the human genome"[1]. By characterizing genetic variation among individuals and populations, we may gain a better understanding of differential susceptibility to disease, differential response to pharmacological agents and the complex interaction of genetic and environmental factors in producing phenotypes[1]. Informed policy decisions must be made about which populations to include in genome projects such as HapMap[2] and whether or not to include racial and ethnic identifiers in biomedical research. Such decisions require understanding of the structure of genetic variation across human populations, its correlation with human demographic and evolutionary history and its implications for the study of differential risk to disease.

History of racial classification

The topic of race, genetics and biomedical research continues to be of considerable interest and debate[3–8]. Historically, biological classification of races has been associated with hierarchical ranking of races, biological determinism, eugenics and justification for genocide (*e.g.*, the Nazi-led holocaust), colonialism, slavery, and other social inequities[9]. Given this tainted history of biological studies of race, it is no wonder that there has been an understandable fear of using biological markers to make racial classifications.

Few of the early classifications of race were disassociated from the social and political views of the time. In 1758, Linnaeus proposed what he considered to be natural taxonomic categories of the human species[10]. He distinguished between *Homo sapiens afer* and *Homo sapiens europaeus* and later added four geographical subdivisions of humans: white Europeans, red Americans, yellow Asians and black Africans[9,10]. Although Linnaeus intended an objective classification, he used both biological and cultural data in his subdivision descriptions[10]. In 1775, Blumenbach categorized humans into five 'races', which largely corresponded with Linnaeus's classifications except for the addition of Oceanians (whom he called 'Malay')[10]. In 1962, the physical anthropologist Carlton Coon further refined this classification of five races on the basis of phenotypic physical features; he called the races Caucasoid, Mongoloid, Australoid, Negroid and Capoid[10]. Despite disagreement among anthropologists, this classification remains in use by many researchers, as well as lay persons.

One of the problems with using 'race' as an identifier is the lack of a clear definition of race[7]. Historically, 'race' has been classified based on both sociocultural and biological characteristics including morphology, skin color, language, culture, religion, ethnicity and geographic origin. Morphology and skin color are not always good indicators of race because they probably result from adaptation to environmental conditions and may have been subject to convergent evolution (*e.g.*, people with dark skin are found in New Guinea, Southern India and Africa, and even within these regions, there can be tremendous variation in skin color). Culture, language, religion and ethnicity have strong sociocultural components and may not always be a good indicator of shared ancestry (*e.g.*, 'Hispanics' in the US include individuals of European, Native American and African ancestry in all possible combinations). Nor is geographic origin always adequate for defining 'race' because of recent, historical and prehistorical migrations of peoples.

Some argue that there is no such thing as 'race' or that it is biologically meaningless[11]. Yet the lay person will ridicule that position as nonsense, because people from different parts of the world look different, whereas people from the same part of the world tend to look similar. The popular concept of five races corresponds well to both geographic regions (Africa, Europe, East Asia, Oceania and the Americas) and bureaucratic definitions (*e.g.*, the US census bureau; http://www.census.gov/population/www/socdemo/race/racefactcb.html).

[1]Department of Biology, Building #144, University of Maryland, College Park, Maryland 20742, USA. [2]Department of Genetics, Yale University School of Medicine, PO Box 208005, New Haven, Connecticut 06520-8005, USA. Correspondence should be addressed to S.A.T. (tishkoff@umd.edu) or K.K.K. (kenneth.kidd@yale.edu).

Published online 26 October 2004; doi:10.1038/ng1438

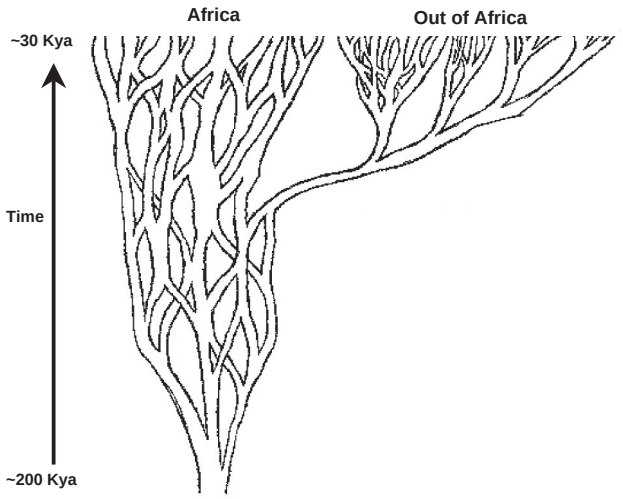

Figure 1 This pencil sketch of large-scale events abstractly illustrates the RAO model of human evolutionary history from ~200–30 Kya. The expansion of modern humans out of Africa within the past 100 Kya sampled only a subset of the variation in the African gene pool. As humans migrated farther from Africa and then expanded locally to occupy all of Eurasia, even more drift accumulated because the subsets of the local gene pools carried forward by migrants became successively more homogeneous. The result is a general clinal pattern of decreasing heterozygosity and increasing linkage disequilibrium with increasing distance from Africa, but with a marked founder effect associated with the expansion out of Africa. Not shown are the ever-present small-scale migrations among groups or the recent migrations between regions.

In this review, we focus on the biogeographical distribution of genetic variation and we address the question of whether or not populations cluster according to this popular concept of 'race'. We show that racial classifications are inadequate descriptors of the distribution of genetic variation in our species.

Models of human evolution

Early theories of modern human origins proposed that human races were distinct biological species that originated independently with little or no gene flow between them (*e.g.*, polygenism)[10]. More recent models based on the fossil record (*e.g.*, the Multiregional Origin model) propose that after the migration of *Homo erectus* out of Africa ~800,000 to 1.8 million years ago, there has been parallel evolution from *H. erectus* to *Homo sapiens* among geographically dispersed populations, with limited gene flow between populations[12].

In contrast to these models, which predict that populations from distinct geographic regions have been differentiating over long periods of time, the genetic data accumulated over the past two decades overwhelmingly support the Recent African Origin (RAO) model (also called the Out of Africa model)[13]. According to the RAO model, all non-African populations descend from an anatomically modern *H. sapiens* ancestor that evolved in Africa ~200 thousand years ago (Kya) and then spread and diversified throughout the rest of the world starting ~50–100 Kya, supplanting any archaic *Homo* populations still present outside of Africa, such as Neanderthals[14] (although low levels of admixture between these groups cannot be ruled out[13]). Studies of variation in autosomal[15,16], mitochondrial DNA (mtDNA)[17] and Y-chromosome[18] haplotypes indicate that the migration(s) out of Africa originated from an East African gene pool. The RAO model predicts a recent common African ancestor with subsequent recent expansions after the initial migration(s) out of Africa ~100 Kya (summarized in ref. 19). This history of expansions into Australo-Melanesia (~60 Kya), Europe (~40 Kya), Asia (~35 Kya), the New World (~30 Kya) and the Pacific (~3 Kya) is supported by patterns of allele frequency variation[19]. **Figure 1** summarizes the RAO model of human evolution up to ~30 Kya. Before reviewing some of the data supporting this model, we must place human genetic variation in a quantitative context.

Amounts of genetic variation

Humans are ~98.8% similar to chimpanzees at the nucleotide level[20] and are considerably more similar to each other, differing on average at only 1 of every 500–1,000 nucleotides between chromosomes[21,22]. This degree of diversity is less than what typically exists among chimpanzees[23–25]. Current estimates of how much variation occurs species-wide indicates that all *H. sapiens* are ~99.6–99.8% identical at the nucleotide sequence level. The other 0.2–0.4% of 3 billion nucleotides comprises ~10 million DNA variants that can potentially occur in all different combinations (these numbers may be underestimates, as Build 121 of dbSNP already contains nearly 10 million single-nucleotide polymorphisms (SNPs), of which 4.5 million are validated; http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi). This is vastly more than enough variation to ensure individual uniqueness at the DNA level, but still represents a very small fraction of the total genome. Most of this variation occurs in DNA of no known function, but common variants also occur in coding regions of genes, altering amino acid sequences of proteins, and in regulatory regions that affect gene expression.

Global patterns of genetic variation

Sequencing of the human genome, and recent advances in identifying and genotyping genetic variation at hundreds of loci in hundreds of individuals, is providing a more detailed understanding of global patterns of genetic variation. Most studies of genetic variation in autosomes, the X chromosome and mtDNA, using many types of markers, show higher levels of genetic variation in African populations than in non-African populations[13,26]. Exceptions to this pattern are studies of restriction fragment length polymorphisms and SNPs, which show higher variability in Europeans but are biased because the polymorphisms were first identified in non-African populations[13,16,26]. Additionally, studies of autosomal[15,27–31] and X-chromosomal haplotype variation[32–34], as well as mtDNA variation[17], indicate that Africans have the largest number of population-specific alleles and that non-African populations carry only a fraction of the genetic diversity that is present in Africa. This would be expected if there were a genetic bottleneck at the time of migration of modern humans out of Africa. For example, data from 94 short tandem repeat polymorphism (STRP) loci genotyped in ten populations[35,36] indicate that there are more heterozygosity and more private (population-specific) alleles in Africans and a clinal pattern of less heterozygosity and fewer population-specific alleles with increasing distance from Africa (**Fig. 2**).

Populations also differ with respect to the organization of variants along a chromosome (haplotypes). The nonrandom association of alleles at different sites is referred to as linkage disequilibrium (LD). Levels and patterns of LD depend on gene-specific factors, such as selection and rates of mutation and recombination, as well as demographic factors that have a genome-wide effect, such as population size, population structure, founder effect and admixture[13,26]. Numerous studies of LD between SNPs and microsatellites show greater LD in Eurasians than in Africans and still greater LD in Native Americans[15,16,27–31,37] (**Fig. 3**). Additionally, levels of LD may vary within geographic regions (**Fig. 3**). This pattern of LD is consistent

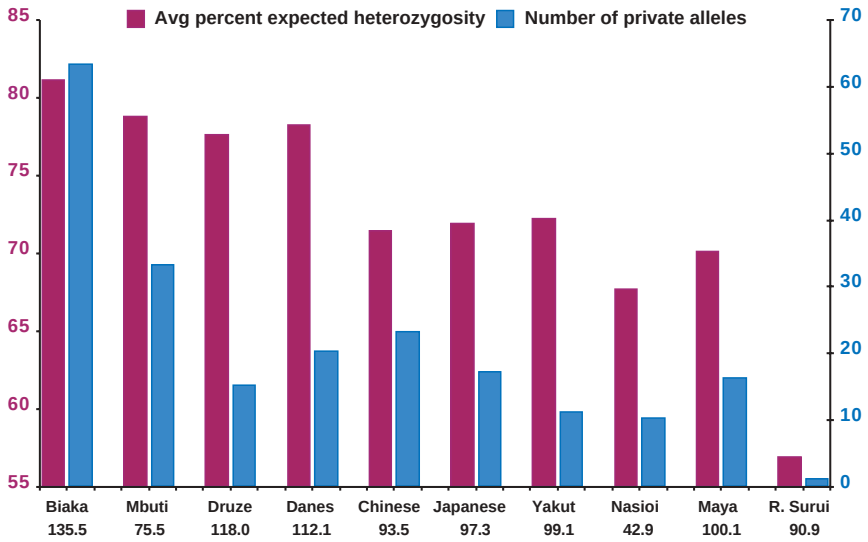

Figure 2 Diversity within populations for 94 dinucleotide STRPs. For each population, the left bar (left axis) gives the average percent expected heterozygosity calculated as the mean of $1-\Sigma p_i^2$ values for the 94 loci. The right bar (right axis) gives the number of private alleles across all 94 loci, where a private allele is present only in that population (considering only those populations in this study). The average sample size (2N, number of chromosomes typed) is given below each population name. Modified from ref. 36 including data in ref. 35; data are in ALFRED[69] (http://alfred.med.yale.edu). R, Rondonian.

with human demographic history; ancestral African populations have maintained a larger effective population size (N_e) and have had more time for recombination and mutation to reduce LD. The bottleneck associated with the expansion of modern humans out of Africa resulted in many of the African haplotypes being lost, leading to greater LD in non-African populations. Another bottleneck, associated with the expansion into the Americas, is reflected in the even higher amounts of LD in this region.

Genetic substructure

Given that humans, as a species, are extremely similar at the genetic level, what is the structure of the genetic variability that does exist and does it correlate with commonly used racial classifications? Isolation by distance is the norm in human populations because humans do not mate at random; individuals living in the same geographic region and sharing a language are more likely to mate with each other than with individuals from more distant regions. Therefore, due to the process of genetic drift, populations have differentiated over time. Genetic isolation of ethnic groups is reinforced by assortative mating (the tendency for individuals with phenotypic resemblance to mate with each other) and localized endogamy based on sociocultural factors such as language and ethnicity. Because we have greater mobility today, distance is not as great a barrier, and as sociocultural barriers against interethnic marriages are decreasing, admixture is becoming increasingly common.

Most studies of population structure focus on selectively neutral variation, which is most likely to reflect human demographic history. The classic measure for partitioning genetic variance within populations relative to between populations is Sewall Wright's F_{ST}, a statistic

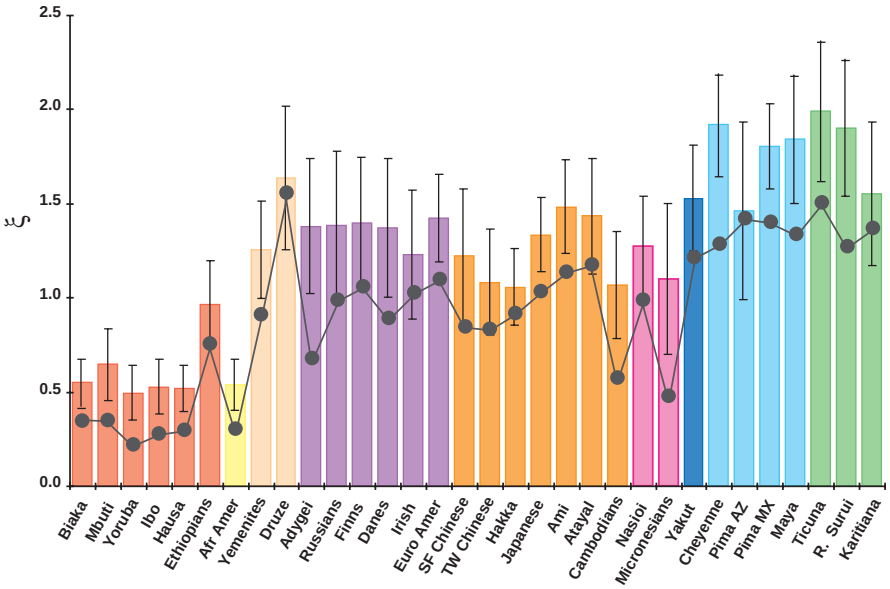

Figure 3 The average LD for 83 SNPs across 21 haplotypes for 32 populations. LD is measured as the ξ coefficient, a standardized measure of overall nonrandomness of alleles at the sites in the haplotypes[70]. The bars are the mean values of ξ across the same 21 independent haplotype systems in all populations. The standard errors of the means are given as the error bars and the median values are plotted as dots connected by the line. Bars are color-coded by geographic region of origin of the populations, from left to right as sub-Saharan Africa, African Americans, Southwest Asia, Europe, East Asia, Pacific, Siberia, North America and South America. Population and sample descriptions are in ALFRED[69]. Different samples of populations with the same name are distinguished by initials: SF, San Francisco; TW, Taiwan; AZ, Arizona; MX, Mexico; R, Rondonian. The haplotyped loci were chosen with no prior knowledge of LD values at the locus. The number of sites per haplotyped locus varied from 2 to 7 for a total of 83 SNPs. The graph is based on published data on *CD4*, *DM1*, *DRD2*, *DRD4*, *PAH* and *COMT* plus unpublished data. These results show less LD in African populations than elsewhere and greater LD in the Native American populations than in other regions, as well as variation in LD within geographic regions.

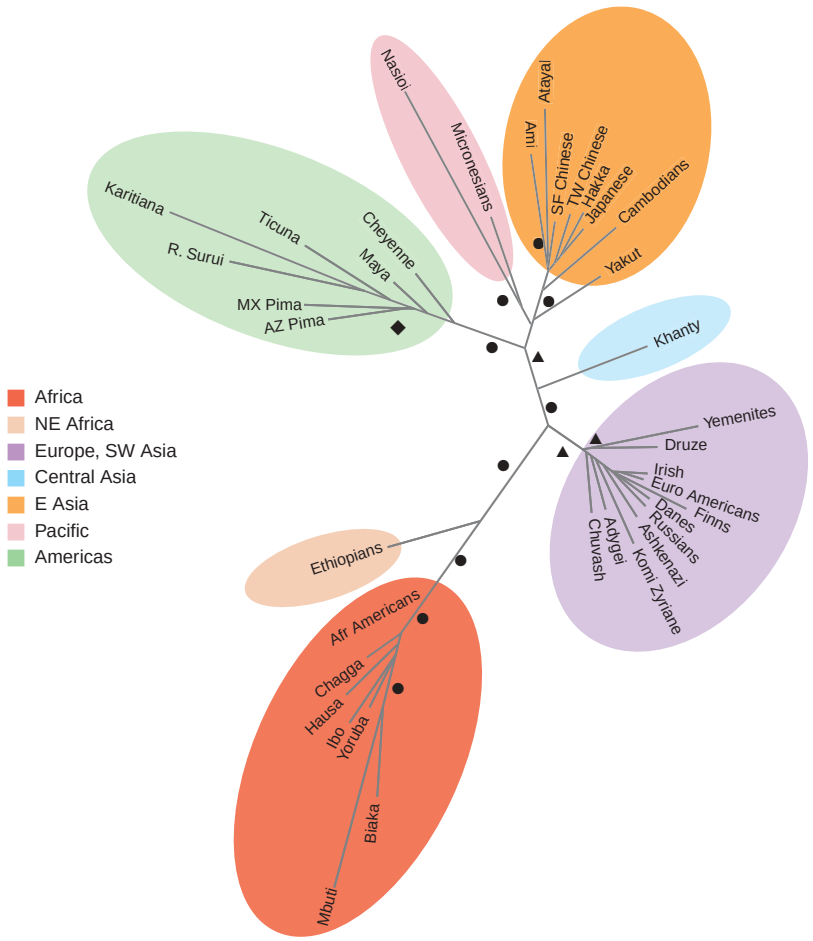

Figure 4 A least-squares tree for 37 populations based on 80 independent loci (41 haplotyped loci, 36 biallelic loci and 3 STRPs) with ~620 statistically independent alleles. This is the best fit found among several exact least-squares evaluations of the similar trees found by a heuristic search algorithm, improving on the neighbor-joining tree. Under the assumption of random genetic drift and no migration or selection (clearly not applicable to African Americans), the branch lengths are proportional to $t/2N_e$ (ref. 16). Because the time from the root (in Africa) to all terminals (modern populations) is the same, the increasing distance from populations in Africa to those in East Asia and the Americas represents increasing drift caused by decreasing effective population sizes. Within a geographic region, the genetic clustering of populations often parallels the linguistic clustering because stochastic factors affect both as populations diverge[19]. SF, San Francisco; TW, Taiwan; AZ, Arizona; MX, Mexico; R, Rondonian. The symbols represent bootstrap values (based on 1,000 replicates): circles, >95%; diamonds, 90–95%; triangles, 85–90%.

ranging from a value of zero (no differentiation) to a value of one (no shared genetic variation)[38]. Under a model of random genetic drift, F_{ST} increases with the amount of time that populations are separated at rates that are inversely related to N_e. At equilibrium between gene flow among populations and genetic drift within populations, the value of F_{ST} will depend on the number of migrants (N_m) exchanged between populations each generation. Thus, a low F_{ST} value could reflect either recent common ancestry or high levels of migration.

Estimates of F_{ST} (or equivalent measures) within and between main geographic regions (Africa, Europe and Asia) typically range from 0.11 to 0.23 for protein polymorphisms, blood groups, RFLPs, SNPs and autosomal microsatellites, indicating that only 11–23% of observed variation is due to differences among populations[13,16,26] (**Table 1**). F_{ST} estimates based on variation in mtDNA (F_{ST} = 0.24–0.27) or the Y chromosome (F_{ST} = 0.23–0.64) are higher than estimates from autosomal DNA, possibly due to the smaller effective population size for mtDNA and the Y chromosome, which results in more genetic drift[13,26].

F_{ST} estimates for individual polymorphisms showed variation around the mean in several different studies of SNPs[3,16,39,40]. F_{ST} values that are exceptionally high or low could reflect differential selection acting at particular loci rather than genetic drift and migration[39,41,42]. Additionally, F_{ST} values may differ depending on the type of polymorphism studied, from 3–5% based on STRPs[43] to 14% based on SNPs[16]. These differences illustrate that the F_{ST} statistic is affected by the nature of the polymorphisms studied (*e.g.*, the high heterozygosity of multiallelic microsatellites within populations results in lower F_{ST} values[7,31,44]). Additionally, Long and Kittles[7] argue that the true level of population differentiation may be underestimated due to other violations of assumptions of the F_{ST} model (*e.g.*, divergence between all pairs of populations is not always equal and independent). Nonetheless, all studies are concordant in showing that the amount of genetic variation between populations is a small fraction of the total variation in the human species.

Pairwise F_{ST} values can be represented as a principal-components or multidimensional-scaling plot or as a tree diagram. The tree in **Figure 4** indicates that populations cluster by geographic region (Africa, Europe/Middle East, East Asia, Oceania, New World) and that African populations are most divergent. Much of that clustering is the result of the nonrandom geographic sampling of these specific populations[16]. Those studies that include populations from geographically intermediate regions place those populations in an intermediate position (*e.g.*, the central Asian Khanty and northeast African Ethiopians; **Fig. 4**). The simplest explanation of this pattern is genetic drift resulting from isolation by distance after the initial expansion out of Africa. There is a bottleneck and considerable drift associated with the initial expansion out of Africa, as shown by the large genetic distances between African and non-African populations (**Fig. 4**). Additional drift is associated with the founding of geographic regions such as Europe and the Middle East, Asia, Oceania and the Americas.

There is also considerable substructure within geographic regions[13,26] (**Fig. 4** and **Table 1**). Several studies of mtDNA and Y-chromosome variation, in addition to autosomal RFLPs, microsatellites and Alu elements studied singly and as haplotypes, have shown more divergent genetic lineages and higher levels of subdivision in

Table 1 F_{ST} comparisons based on 369 SNPs for pairs of populations within and across global regions

	F_{ST} across regions						F_{ST} within regions			
	Africa Europe	Africa E. Asia	Africa Americas	Europe E. Asia	Europe Americas	E. Asia Americas	Africa	Europe	E. Asia	Americas
n	48	42	42	56	56	49	15	28	21	21
Mean	0.152	0.228	0.232	0.122	0.156	0.159	0.051	0.019	0.041	0.091
Median	0.147	0.223	0.229	0.115	0.154	0.161	0.051	0.020	0.040	0.084
Minimum	0.112	0.183	0.171	0.089	0.084	0.086	0.013	0.021	0.016	0.033
Maximum	0.202	0.283	0.308	0.182	0.223	0.229	0.097	0.031	0.079	0.148

n, average number of pairwise comparisons. Africa includes Biaka, Mbuti, Yoruba, Ibo, Hausa and Chagga. Europe includes Adygei, Chuvash, Russians, Ashkenazi, Finns, Danes, Irish and European Americans. East Asia includes San Francisco Chinese, Taiwan Chinese, Hakka, Japanese, Ami, Atayal and Cambodians. Americas includes Cheyenne, Arizona Pima, Mexico Pima, Maya, Ticuna, Rondonian Surui and Karitiana. The SNPs are primarily noncoding SNPs but include a few silent substitutions.

African populations than in those from other regions, as expected under a RAO model[26]. But haplotype studies suggest there has been sufficient gene flow among African populations such that common haplotypes are present in most African populations, though often at very different frequencies[15,16,30,31,33,34]. This African heterogeneity means that descendents of the African slave trade, who originated from diverse West Africa ethnic groups and have varying levels of European and Native American admixture, are genetically heterogeneous. Similarly, the considerable substructure that exists in all other regions means that 'racial' classifications refer to heterogeneous groups. For example, 'Asian American' refers to a heterogeneous population with possible ancestry from Japan, China, southeast Asia or elsewhere. Knowledge of substructure could be important in designing and interpreting biomedical studies.

Determining individual ancestry

Although the amount of genetic diversity between populations is relatively small compared with the amount of genetic diversity within populations, populations usually cluster by geographic region based on genetic distance (**Fig. 4**). Rosenberg *et al.*[43] analyzed 377 microsatellites genotyped in 52 global populations using a clustering algorithm (STRUCTURE[45]) to assign individuals to subgroups (clusters) that have distinctive allele frequencies. They could distinguish five main clusters of individuals that corresponded to broad geographic regions (Africa, Middle East and Europe, Asia, Oceania, Americas). They identified a sixth cluster specific to a Pakistani population, which probably reflects high levels of inbreeding and genetic drift in that group. Without reference to sampling location, individuals from the same predefined population nearly always shared membership in one of the five main clusters.

There were some exceptions, however, for populations from geographically intermediate regions (*e.g.*, Central Asia, the Middle East), in which individuals had partial membership in multiple clusters, especially those of flanking geographic regions, indicating a continuous gradient of variation among some regions. Thus, although the main clusters correlate with the common concept of 'races' (as expected, because populations from different parts of the world have larger differences in allele frequencies than populations from the same region of the world), the analyses by STRUCTURE do not support discrete boundaries between races. Had there been a more geographically continuous sampling (*e.g.*, from regions such as Ethiopia), there would probably be an even more continuous gradient of genetic variation across all geographic regions. This and other studies[3,8,44] indicate that one can assign individual ancestry to continent of origin with high accuracy using a large enough number of polymorphic markers (>60) and that "self-reported population ancestry likely provides a suitable proxy for genetic ancestry"[43]. Accuracy of assigning ancestry can be even higher using ancestry-informative markers, markers with very different allele frequencies in populations from different regions[46–49], or functional variants that may be under differential selection[50]. The accuracy of assigning ancestry decreases for populations from intermediate geographic regions such as Central Asia[43], the Middle East[43], Ethiopia[50] or South Asia[3,44] and for individuals of mixed ancestry.

The role of selection

We currently know little about the genetic basis of the phenotypic traits that people often associate with racial classification (*e.g.*, hair texture, skeletal morphology, skin color). These are typically quantitative traits that result from interaction of multiple genes and environment and are probably influenced by natural selection. Mutations that are involved in disease may also differ in frequency across ethnic groups due to historical selection. The best-known examples are for genes that have a role in resistance to infectious disease (*e.g.*, sickle cell anemia, G6PD deficiency and possibly cystic fibrosis). Many common diseases, however, such as hypertension, diabetes and obesity, which differ in prevalence across ethnic groups, may also be influenced by genes that have been under natural selection (*e.g.*, the thrifty gene hypothesis[51]). Thus, characterization of signatures of natural selection could enable identification of functional mutations that influence susceptibility to genetic disease and differential drug response[13,41].

Examples of genes that have a signature for selection include those that are involved in skin color (*MC1R*[52]), resistance to malaria (*G6PD*[33,53–55], Duffy (*FY*)[56] and *HBE*1; ref. 57), resistance to human immunodeficiency virus (*CCR*5; ref. 58), lactase persistence (*LCT*[42]), drug metabolism (*CYP1A2*; ref. 59), color vision (*OPN1LW*[34]) and alcohol metabolism (*ADH1B*[60] and *ALDH2*; ref. 61). Many functional variants at these genes are geographically restricted owing to geographically restricted parasites (*G6PD*, Duffy and *HBE1*), nutritional and cultural factors (*LCT*) or unknown factors (*ADH1B*, *ALDH2*, *CCR5* and *CYP1A2*). Depending on the geographic distribution of the selective forces, such loci may reinforce or obscure the overall pattern of population relationships. But traits that typify *H. sapiens* (*e.g.*, language capacity, a large brain and intelligence) are shared among populations from all regions owing to recent common ancestry and shared selective pressure during human speciation.

Biomedical implications

The processes of mutation, migration, genetic drift and selection have resulted in the differential distribution of normal genetic variation and of genetic variation affecting disease. Therefore, knowledge of ancestry

can be important clinically and in design of biomedical research studies. The common disease–common variant hypothesis states that common genetic diseases are affected by common disease-susceptibility alleles at a few loci that exist at high frequency across ethnically diverse populations[21,62]. These alleles probably arose before population differentiation and are common across populations. But complex diseases may also be influenced by geographically restricted rare susceptibility alleles[63,64]. Because LD is believed to be useful in mapping genes for complex disease, and given the divergent pattern of haplotype frequencies and LD across global populations as well as high levels of substructure in regions such as Africa, there is increased need to characterize haplotypes and LD across ethnically diverse populations[13,26,65]. The HapMap project, which proposes to characterize LD in a small subset of ethnic groups[2], may be insufficient.

Additionally, undetected population structure in case-control association studies can result in false positive association[66,67]. Thus, knowledge of ethnicity (not just broad geographic ancestry) and statistical tests of substructure are important for proper design of case-control association studies and for identifying disease predisposing alleles that may differ across ethnic groups. The particular identifier used (language, ethnicity, geographic origin, religion) will depend on the particular study and the hypotheses being tested. For example, use of religion as a descriptor will be important if discussing diseases prevalent in Jewish populations, which may result from genetic drift due to founder effect[68] (*e.g.*, Tay Sachs, Torsion dystonia, breast cancer and Gaucher disease).

Information about individual ancestry could also provide important medical information for diagnosis and treatment. It is not desirable to treat individuals on the basis of their ethnic identity; the goal is individualized medicine—identifying individual risk factors and treating for the specific etiology in the individual. But many different disorders have similar symptoms, and the process of differential diagnosis can use ethnicity to prioritize tests according to the most likely etiology. Whether genetic, infectious or environmental, causes of disorders vary among ethnic groups. Economics and common sense argue that one would attempt to confirm (or reject) the most likely cause before attempting to confirm a very remote etiology. Taking 'ethnicity' (genetic ancestry and sociocultural characterization) into account can be good medical practice[4,8], and if one is interested only in broad geographic ancestry, self-reported ancestry will probably suffice[43,44]. At the same time, one must be wary of racial profiling and ignorance of the continuous nature of genetic variation and high levels of admixture in modern populations, which can result in misclassification and misdiagnosis[5,11]. Although information about ethnicity can be informative for biomedical research, it is imperative to move away from describing populations according to racial classifications such as 'black', 'white' or 'Asian', unless the aims of the study are to distinguish sociocultural and environmental risk factors or to distinguish broad geographic ancestry. Because there can be considerable genetic heterogeneity within a region, it is most useful to be as specific as possible about geographic origins, ethnicity or tribal affiliation.

Conclusions

The emerging picture is that populations do, generally, cluster by broad geographic regions that correspond with common racial classification (Africa, Europe, Asia, Oceania, Americas). This is not surprising as the distribution of variation seen today is primarily the result of the history of human expansion out of Africa, the pathways of expansion through Eurasia, subsequent demographic expansions of populations into Oceania and the Americas and local and long-range migrations. A general pattern of isolation by distance has allowed drift to accumulate in spite of some damping due to local migrations. The pattern laid down by the initial expansion of modern humans out of Africa is detectable using Y-chromosome, mtDNA and autosomal markers. Selection in response to region-specific factors has enhanced the differences at some loci, and recent migrations and demic expansions have added complexity to the pattern. But 'races' are neither homogeneous nor distinct for most genetic variation.

Understanding the global distribution of genetic variation is biomedically important, but we emphasize that existence of differences, however small, should not be a basis for discrimination. Statements like "We hold these truths to be self-evident, that all men are created equal…" (US Declaration of Independence, 1776) reflect morality, not science. One can accept this moral imperative and still recognize that all individuals, independently conceived, are genetically unique.

ACKNOWLEDGMENTS

We thank A. Pakstis for help with the analyses and graphic representations in the figures; B. Verrelli, F. Reed and J. Kidd for critical review of the manuscript; and the many hundreds of individuals who volunteered to give DNA samples for studies such as those reviewed here. This work was supported in part by grants from the US National Institutes of Health (to K.K.K.), by a contract from the National Institute of Diabetes, Digestive and Kidney Diseases (to K.K.K.), by a grant from the Alfred P. Sloan Foundation (to K.K.K.), by a grant from the National Science Foundation (to S.A.T.) and by the Burroughs Wellcome Fund and David and Lucile Packard Career Awards (to S.A.T.).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 10 August; accepted 9 September 2004
Published online at http://www.nature.com/naturegenetics/

1. Collins, F.S., Green, E.D., Guttmacher, A.E. & Guyer, M.S. A vision for the future of genomics research. *Nature* **422**, 835–847 (2003).
2. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
3. Bamshad, M., Wooding, S., Salisbury, B.A. & Stephens, J.C. Deconstructing the relationship between genetics and race. *Nat. Rev. Genet.* **5**, 598–609 (2004).
4. Burchard, E.G. *et al.* The importance of race and ethnic background in biomedical research and clinical practice. *N. Engl. J. Med.* **348**, 1170–1175 (2003).
5. Cooper, R.S., Kaufman, J.S. & Ward, R. Race and genomics. *N. Engl. J. Med.* **348**, 1166–1170 (2003).
6. Kittles, R.A. & Weiss, K.M. Race, ancestry, and genes: implications for defining disease risk. *Annu. Rev. Genomics Hum. Genet.* **4**, 33–67 (2003).
7. Long, J.C. & Kittles, R.A. Human genetic diversity and the nonexistence of biological races. *Hum. Biol.* **75**, 449–471 (2003).
8. Risch, N., Burchard, E., Ziv, E. & Tang, H. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol.* **12**, 602–612 (2002).
9. Gould, S.J. *The Mismeasure of Man* (Norton & Company, New York, London, 1981).
10. Marks, J. *Human biodiversity: genes, race, and history.* (Aldine de Gruyter, New York, 1995).
11. Schwartz, R.S. Racial profiling in medical research. *N. Engl. J. Med.* **344**, 1392–1393 (2001).
12. Wolpoff, M.H. Interpretations of multiregional evolution. *Science* **274**, 704–706 (1996).
13. Tishkoff, S.A. & Verrelli, B.C. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.* **4**, 293–340 (2003).
14. Stringer, C.B. & Andrews, P. Genetic and fossil evidence for the origin of modern humans. *Science* **239**, 1263–1268 (1988).
15. Tishkoff, S.A. *et al.* Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387 (1996).
16. Kidd, K.K., Pakstis, A.J., Speed, W.C. & Kidd, J.R. Understanding human DNA sequence variation. *J. Heredity* **95**, 406–420 (2004).
17. Quintana-Murci, L. *et al.* Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat. Genet.* **23**, 437–441 (1999).
18. Underhill, P.A. *et al.* Y chromosome sequence variation and the history of human populations. *Nat. Genet.* **26**, 358–361 (2000).
19. Cavalli-Sforza, L.L., Piazza, A. & Menozzi, P. *History and Geography of Human Genes.* (Princeton University Press, Princeton, 1994).
20. Britten, R.J. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl. Acad. Sci. USA* **99**, 13633–13635 (2002).
21. Chakravarti, A. Single nucleotide polymorphisms...to a future of genetic medicine. *Nature* **409**, 822–823 (2001).
22. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).

23. Fischer, A., Wiebe, V., Paabo, S. & Przeworski, M. Evidence for a complex demographic history of chimpanzees. *Mol. Biol. Evol.* **21**, 799–808 (2004).
24. Li, W.H. & Sadler, L.A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
25. Yu, N. *et al*. Low nucleotide diversity in chimpanzees and bonobos. *Genetics* **164**, 1511–1518 (2003).
26. Tishkoff, S.A. & Williams, S.M. Genetic analysis of African populations: Human evolution and complex disease. *Nat. Rev. Genet.* **3**, 611–621 (2002).
27. Gabriel, S.B. *et al*. The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
28. Kidd, K.K. *et al*. A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Hum. Genet.* **103**, 211–227 (1998).
29. Kidd, J.R. *et al*. Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus (*PAH*) in a global representation of populations. *Am. J. Hum. Genet.* **66**, 1882–1899 (2000).
30. Tishkoff, S.A. *et al*. A global haplotype analysis of the myotonic dystrophy locus: Implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *Am. J. Hum. Genet.* **62**, 1389–1402 (1998).
31. Tishkoff, S.A. *et al*. Short tandem-repeat polymorphism/Alu haplotype variation at the PLAT locus: Implications for modern human origins. *Am. J. Hum. Genet.* **67**, 901–925 (2000).
32. Kaessmann, H., Wiebe, V., Weiss, G. & Paabo, S. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat. Genet.* **27**, 155–156 (2001).
33. Verrelli, B.C. *et al*. Evidence for balancing selection from nucleotide sequence analyses of human G6PD. *Am. J. Hum. Genet.* **71**, 1112–1128 (2002)
34. Verrelli, B.C. & Tishkoff, S.A. Signatures of selection and gene conversion associated with human color vision variation. *Am. J. Hum. Genet.* **75**, 363–375 (2004).
35. Calafell, F., Shuster, A., Speed, W.C., Kidd, J.R. & Kidd, K.K. Short tandem repeat polymorphism evolution in humans. *Eur. J. Hum. Genet.* **6**, 38–49 (1998).
36. Kidd, K. Race, Human Genes & Human Origins: How Genetically Diverse Are We? in *New Dimensions in Bioethics: Science, Ethics and the Formulation of Public Policy* (A.W. Galston, E. Shurr & M.A. Norwell, eds.) 11–24 (Kluwer Academic Press, 2001).
37. Reich, D.E., Cargill, M., Bolk, S., Ireland, J. & Sabeti, P.C. Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
38. Wright, S. Evolution and the Genetics of Populations. Volume 2: The Theory of Gene Frequencies. (University of Chicago Press, Chicago, 1969).
39. Akey, J.M., Zhang, G., Zhang, K., Jin, L. & Shriver, M.D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805–1814 (2002).
40. Stephens, J.C. *et al*. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
41. Bamshad, M. & Wooding, S.P. Signatures of natural selection in the human genome. *Nat. Rev. Genet.* **4**, 99–111 (2003).
42. Bersaglieri, T. *et al*. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
43. Rosenberg, N.A. *et al*. Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
44. Bamshad, M.J. *et al*. Human population genetic structure and inference of group membership. *Am. J. Hum. Genet.* **72**, 578–589 (2003).
45. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
46. Rosenberg, N.A., Li, L.M., Ward, R. & Pritchard, J.K. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **73**, 1402–1422 (2003).
47. Parra, E.J., Marcini, A., Akey, J., Martinson, J. & Batzer, M.A. Estimating African American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.* **63**, 1839–1851 (1998).
48. Shriver, M.D. *et al*. Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum. Genet.* **112**, 387–399 (2003).
49. Smith, M.W. *et al*. A high-density admixture map for disease gene discovery in african americans. *Am. J. Hum. Genet.* **74**, 1001–1013 (2004).
50. Wilson, J.F. *et al*. Population genetic structure of variable drug response. *Nat. Genet.* **29**, 265–269 (2001).
51. Neel, J.V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *Am. J. Hum. Genet.* **14**, 353–362 (1962).
52. Harding, R.M. *et al*. Evidence for variable selective pressures at MC1R. *Am. J. Hum. Genet.* **66**, 1351–1361 (2000).
53. Sabeti, P.C. *et al*. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
54. Saunders, M.A., Hammer, M.F. & Nachman, M.W. Nucleotide variability at G6pd and the signature of malarial selection in humans. *Genetics* **162**, 1849–1861 (2002).
55. Tishkoff, S.A. *et al*. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293**, 455–462 (2001).
56. Hamblin, M.T., Thompson, E.E. & Di Rienzo, A. Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* **70**, 369–383 (2002).
57. Ohashi, J. *et al*. Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. *Am. J. Hum. Genet.* **74**, 1198–1208 (2004).
58. Bamshad, M.J. *et al*. A strong signature of balancing selection in the 5′ cis-regulatory region of CCR5. *Proc. Natl. Acad. Sci. USA* **99**, 10539–10544 (2002).
59. Wooding, S.P. *et al*. DNA sequence variation in a 3.7-kb noncoding sequence 5 ' of the CYP1A2 gene: Implications for human population history and natural selection. *Am. J. Hum. Genet.* **71**, 528–542 (2002).
60. Osier, M.V. *et al*. A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. *Am. J. Hum. Genet.* **71**, 84–99 (2002).
61. Oota, H. *et al*. The evolution and population genetics of the ALDH2 locus: random genetic drift, selection, and low levels of recombination. *Ann. Hum. Genet.* **68**, 93–109 (2004).
62. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
63. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
64. Pritchard, J.K. & Cox, N.J. The allelic architecture of human disease genes: common disease - common variant...or not? *Hum. Mol. Genet.* **11**, 2417–2423 (2002).
65. Tishkoff, S.A.& Verrelli, B.C. Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping. *Curr. Opin. Genet. Dev.* **13**, 569–575 (2003).
66. Pritchard, J.K. & Rosenberg, N.A. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65**, 220–228 (1999).
67. Kang, A.M., Palmatier, M.A. & Kidd, K.K. Global variation of a 40-bp VNTR in the 3′-untranslated region of the dopamine transporter gene (SLC6A3). *Biol. Psychiatry* **46**, 151–160 (1999).
68. Risch, N., Tang, H., Katzenstein, H. & Ekstein, J. Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. *Am. J. Hum. Genet.* **72**, 812–822 (2003).
69. Osier, M.V. *et al*. ALFRED: an allele frequency database for anthropology. *Am. J. Phys. Anthropol.* **119**, 77–83 (2002).
70. Zhao, H., Pakstis, A.J., Kidd, J.R. & Kidd, K.K. Assessing linkage disequilibrium in a complex genetic system I. Overall deviation from random association. *Ann. Hum. Genet.* **63**, 167–179 (1999).