PERSPECTIVE

nature genetics

Conceptualizing human variation

S O Y Keita[1,2], R A Kittles[1,3], C D M Royal[1], G E Bonney[1], P Furbert-Harris[1], G M Dunston[1] & C N Rotimi[1]

What is the relationship between the patterns of biological and sociocultural variation in extant humans? Is this relationship accurately described, or best explained, by the term 'race' and the schema of 'racial' classification? What is the relationship between 'race', genetics and the demographic groups of society? Can extant humans be categorized into units that can scientifically be called 'races'? These questions underlie the discussions that address the explanations for the observed differences in many domains between named demographic groups across societies. These domains include disease incidence and prevalence and other variables studied by biologists and social scientists. Here, we offer a perspective on understanding human variation by exploring the meaning and use of the term 'race' and its relationship to a range of data. The quest is for a more useful approach with which to understand human biological variation, one that may provide better research designs and inform public policy.

'Race': semantics and confusion

The term 'race' engenders much discussion, with little agreement between those who claim that 'races' are real (meaning natural) biological entities and those who maintain that they are socially constructed[1]. The former group sometimes stresses empirical evidence for the existence of biological 'racial' differences, and the latter stresses the role that human agency has had in creating distinctions between people (on any level). Biologists also disagree about the meaning of 'race', and whether it is applicable to human infraspecific (within-species) variation[2–5].

An examination of these discussions indicates that there is a problem with semantics. 'Race' is not being defined or used consistently; its referents are varied and shift depending on context. The term is often used colloquially to refer to a range of human groupings. Religious, cultural, social, national, ethnic, linguistic, genetic, geographical and anatomical groups have been and sometimes still are called 'races'[6,7]. In anthropology, the meaning of race became formalized for humans and restricted to units based on biological variation in keeping with general zoological practice[8,9]. Classifications were based on somatic traits.

[1]National Human Genome Center, College of Medicine, Howard University, Washington, DC 20060, USA. [2]Department of Anthropology, Smithsonian Institution, Washington, DC, USA. [3]Department of Molecular Virology, Immunology, and Medical Genetics, The Ohio State University, Columbus, Ohio 43210, USA. Correspondence should be addressed to R.A.K. (kittles-1@medctr.osu.edu).

Published online 26 October 2004; doi:10.1038/ng1455

'Race' is applied in formal taxonomy to variation below the species level. In traditional approaches, substantively morphologically distinct populations or collections of populations occupying a section of a species range are called subspecies and given a three-part Latin name[10]. In current systematic practice, the designation 'subspecies' is used to indicate an objective degree of microevolutionary divergence[11]. Do any of the human groups called 'races', including those from traditional anthropology, meet this latter criterion?

We argue that the correct use of the term 'race' is the most current taxonomic one, because it has been formalized. 'Race' gains its force from its natural science root. The term denotes 'natural' distinctions and connotes differences not susceptible to change. One is led to ask, therefore, whether everything that is called a 'racial' difference is actually natural. 'Racial' differences carry a different weight than cultural differences. In terms of taxonomic precision and best practice, is it scientifically correct to identify European Americans, Asians and Pacific Islanders, Han Chinese, Hispanics and African Americans of Middle Passage descent as different races? Although individuals may refer to themselves as belonging to a particular 'race', it is doubtful that this has been done with knowledge of, or concern for, zoological taxonomy, because the common use of the term has come from sociopolitical discourse. Individuals learned the 'race' to which they were assigned.

Although 'race' and subspecies are usually treated as equivalent, some zoological taxonomists reserve the word 'race' for local breeding populations, with subspecies being geographical collections of populations that are similar or the same in the defining traits. This causes no serious problem to this discussion, because the most commonly known anthropological classification of humans is said to consist of races. If 'Caucasoid' is a subspecies, however, then an endogamous village population or ethnic group becomes a 'race'. This illustrates an inconsistency even in biological usage not found in scientific or sociopolitical practice: for example, how often are the Old Order Amish referred to as a 'race' in recent scientific literature? This group of people is a breeding population, based on a particular behavioral pattern of mate choice, as opposed to being defined by an anatomical trait complex.

'Race' and subspecies

Although the subspecies level is formally recognized in taxonomy, it has been criticized. Subspecies were primarily delimited by differences in selected observable morphological traits within a restricted geographical range. In practice, divisions were made based on a few prominent traits, with subsequent variation interpreted in terms of established units.

BOX 1 SUMMARY POINTS

1. Modern human biological variation is not structured into phylogenetic subspecies ('races'), nor are the taxa of the standard anthropological 'racial' classifications breeding populations. The 'racial taxa' do not meet the phylogenetic criteria.
2. 'Race' denotes socially constructed units as a function of the incorrect usage of the term. US demographic units are not 'races'. But social units were politically constructed from the somatically defined 'races' of classical anthropology. In addition, rules of descent were created that delimited group membership, based on some notion of desirability by those who created the laws.
3. Human geographical and group variation in health and disease are real and require study to partition, as much as possible, environmental and genetic variance.
4. The absence of 'races' does not mean the absence of racism, or the structured inequality based on operationalized prejudice used to deprive people who are deemed to be fundamentally biologically different of social and economic justice. The 'no biological race' position does not exclude the idea that racism is a problem that needs to be addressed.
5. Group studies should obtain the specific ancestral histories of individuals. Ancestral histories are different from self-reported group membership, as some groups have multiple ancestral origins.
6. Many terms requiring definition for use describe demographic population groups better than the term 'race' because they invite examination of the criteria for classification. Population labels that may apply are 'ethnoancestral', 'bioethnic', 'ethnobiohistorical', 'ancestral-ethnic', 'social-designation', 'biocultural', 'biopopulation', 'ethnosocial', 'ancestral', 'ancestor-historical', 'origin group' and 'ethnogeographical'.

In the 1950s many zoological taxonomists became dissatisfied with the subspecies as a way to understand variation[10,12,13]. Criticisms included (i) the nonconcordance of traits, which made it possible to produce different classifications using the same individuals; (ii) the existence of polytopic populations, which are the product of parallel evolution; (iii) the existence of true breeding populations (demes) within previously delimited subspecies; and (iv) the arbitrariness of criteria used to recognized subspecies[10]. In addition, some traits were found to be clinally distributed, making the creation of divisions arbitrary.

Current systematic theory emphasizes that taxonomy at all levels should reflect evolutionary relationships[11]. For instance, the term 'Negro' was once a racial designation for numerous groups in tropical Africa and Pacific Oceania (Melanesians). These groups share a broadly similar external phenotype; this classification illustrates 'race' as type, defined by anatomical complexes. Although the actual relationship between African 'Negroes' and Oceanic 'Negroes' was sometimes questioned, these groups were placed in the same taxon. Molecular and genetic studies later showed that the Oceanic 'Negroes' were more closely related to mainland Asians.

Molecular systematics makes it possible to explore infraspecific variation to detect patterns that would reflect phylogenetic substructuring. Avise and Ball suggest a definition of 'subspecies' that is consistent with the goals of evolutionary taxonomy[11]: "Subspecies are groups of actually or potentially interbreeding populations phylogenetically distinguishable from, but reproductively compatible with, other such groups. Importantly the evidence for phylogenetic distinction must normally come from the concordant distributions of multiple, independent, genetically based traits."

This definition is different from the previous one in that it emphasizes phylogenetics. It is, in theory, more objective and consistent with neodarwinian evolutionary theory and can be used as the basis for determining whether or not modern *Homo sapiens* can be structured into populations divergent enough to be called 'races'. We know that there is human geographical variation, but does this infraspecific diversity reach a threshold that merits the designation 'subspecies', as is true with chimpanzees[14]?

'Race' and social construction

'Race' is 'socially constructed' when the word is incorrectly used as the covering term for social or demographic groups. Broadly designated groups, such as 'Hispanic' or 'European American' do not meet the classical or phylogenetic criteria for subspecies or the criterion for a breeding population. Furthermore, some of the 'racial' taxa of earlier European science used by law and politics were converted into social identities[2]. For example, the self-defined identities of enslaved Africans were replaced with the singular 'Negro' or 'black', and Europeans became 'Caucasian', thus creating identities based on physical traits rather than on history and cultural tradition. Another example of social construction is seen in the laws of various countries that assigned 'race' (actually social group or position) based on the proportion of particular ancestries held by an individual. The entities resulting from these political machinations have nothing to do with the substructuring of the species by evolutionary mechanisms.

Human races as human variation

Arguments against the existence of human races (the taxa 'Mongoloid', 'Caucasoid' and 'Negroid' and those from other classifications) include those stated for subspecies[10] and several others[15]. The within- to between-group variation is very high for genetic polymorphisms (~85%; refs. 16,17). This means that individuals from one 'race' may be overall more similar to individuals in one of the other 'races' than to other individuals in the same 'race'. This observation is perhaps insufficient[18], although it still is convincing because it illustrates the lack of a boundary. Coalescence times[19,20] calculated from various genes suggest that the differentiation of modern humans began in Africa in populations whose morphological traits are unknown; it cannot be assumed from an evolutionary perspective that the traits used to define 'races' emerged simultaneously with this divergence[15]. There was no demonstrable 'racial' divergence.

Y-chromosome and mitochondrial DNA genealogies are especially interesting because they demonstrate the lack of concordance of lineages with morphology[15] and facilitate a phylogenetic analysis. Individuals with the same morphology do not necessarily cluster with each other by lineage, and a given lineage does not include only individuals with the same trait complex (or 'racial type'). Y-chromosome DNA from Africa alone suffices to make this point. Africa contains populations whose members have a range of external phenotypes. This variation has usually been described in terms of 'race' (Caucasoids, Pygmoids, Congoids, Khoisanoids). But the Y-chromosome clade defined by the PN2 transition (PN2/M35, PN2/M2) shatters the

boundaries of phenotypically defined races and true breeding populations across a great geographical expanse[21]. African peoples with a range of skin colors, hair forms and physiognomies have substantial percentages of males whose Y chromosomes form closely related clades with each other, but not with others who are phenotypically similar. The individuals in the morphologically or geographically defined 'races' are not characterized by 'private' distinct lineages restricted to each of them.

Human genome variation, demographic groups and disease

'Race' is a legitimate taxonomic concept that works for chimpanzees but does not apply to humans (at this time). The nonexistence of 'races' or subspecies in modern humans does not preclude substantial genetic variation that may be localized to regions or populations. More than 10 million single-nucleotide polymorphisms (SNPs) probably exist in the human genome[22]. More than 5 million of these SNPs are expected to be common (minor allele frequency >10%)[23]. Most of these SNPs vary in frequency across human populations, and a large fraction of them are private or common in only a single population. Other genetic variants are also asymmetrically distributed. This makes forensic distinctions possible even within restricted regions such as Scandinavia[24]. Anonymous human DNA samples will structure into groups that correspond to the divisions of the sampled populations or regions when large numbers of genetic markers are used. This has been demonstrated with autosomal microsatellites, which are the most rapidly evolving genetic variants[25]. The DNA of an unknown individual from one of the sampled populations would probably be correctly linked to a population. Because this identification is possible does not mean that there is a level of differentiation equal to 'races'. The genetics of *Homo sapiens* shows gradients of differentiation[15,26].

Because substantial genetic variation may be localized to regions or populations, attention has been focused on how geographic origins may contribute to differential distribution of disease and mortality or 'health disparities'. In January of 2000, the US Department of Health and Human Services launched "Healthy People 2010," a program committed to eliminating 'ethnic' and 'racial' health disparities. Although there is considerable debate regarding the definition, measurement and causes of health disparities, there is increased focus on the potential role of the distribution of DNA sequence variation in contributing to observed differences in disease status among groups.

Several competing, but not necessarily exclusive, hypotheses exist to describe the genetic contribution to complex disease, including the common disease–common variant (CDCV) hypothesis and the multiple rare variants (MRV) hypothesis[27–32]. If it turns out to be that much of the genetic variation contributing to disease is old and shared by most human populations, as implied by the CDCV hypothesis, then differences in the health status of population groups (health disparities) will be largely due to differences in exposure to cumulative environmental insults. If the MRV hypothesis turns out to be true, however, then more comprehensive sampling of multiple human populations will be necessary to adequately describe the extent to which a differential distribution of genes underlies the pathophysiology of disease susceptibility or resistance. Under this hypothesis, a substantial proportion of genetic polymorphisms will be rare and will probably be specific to groups that experienced similar evolutionary forces of selection or drift. In the end, both the CDCV and the MRV hypotheses may apply, depending on the phenotype under consideration. The etiologies of diseases such as lupus, diabetes and Alzheimer disease are examples that may require strategies derived from both hypotheses.

An important implication of the MRV model is that no one map of polymorphic markers (*e.g.*, a SNP map such as that generated by the HapMap project) will probably be sufficient for understanding the complex interplay between multiple genetic variants and multiple environmental factors in the etiology of human diseases across all global populations. Therefore, it may be premature at this time to completely disregard all population (or group) identifiers in biomedical research, as some propose. Group identifiers are important for seeing group patterns in disparities. For example, African Americans have a higher prevalence of some chronic and degenerative diseases. African American males have a 60% greater risk of developing prostate cancer, twice the risk of developing its aggressive form and twice the mortality relative to European Americans[33]. Study designs should reflect efforts to partition the genetic, environmental and geographic variance for the diseases that contribute most to group disparity statistics, such as obesity and related disorders.

The finding that the demographic group called 'African American' has a higher prevalence of prostate cancer, obesity and hypertension is not to be denied. This does not mean, however, that this is a 'racial' phenomenon, as disease is probably due to gene-environment interaction and not linked to the physical traits assumed to covary with this population. This group has heterogeneous ancestral continental origins, predominantly West African and West Central African. They are heterogeneous in their African origins also. Continental African immigrants to the US, including some suprasaharan Africans (*e.g.*, Tunisians and Egyptians) sometimes call themselves 'African Americans', which is true as an epithet but false as a marker of the bioethnic history of those whose ancestors share the experiences of the Middle Passage and slavery. It is this history, and its constituent elements, that are specific to the group. The Middle Passage African descendants, whether in North America or South America, do have a particular biocultural history[34]. It may be necessary to craft specific group identifiers to facilitate good research design[2]. 'Racial' approaches to identity, as found in Office of Management and Budget directive 15, operate from the Platonic mold that groups so defined would necessarily be genetically the same, and this is false. The New World descendants of Middle Passage Africans, whether found in specifically labeled communities (e.g., African Argentinian, African Mexican, African Venezuelan or African Canadian) or in the 'majority' populations ('mestizos' or 'whites') cannot be lumped with newcomers from the continent under the label 'black' or 'African American'. Designations like 'Arab' are also fraught with biohistorical complexity because they often designate peoples who became acculturated. For example, Syrian and Shuwa Arabs illustrate the great biological and cultural variation that may be found under a single ethnolinguistic label.

The causes of health disparities among groups are not well understood, but genetic explanations are frequently the default position for a variety of reasons, including a tradition of biological determinism[4]. Although genes probably have a role, we must realize that some environmental influences can be so subtle and occur so early in life as to be missed, thereby facilitating acceptance of a genetic explanation that is probably false. The fetal programming and early childhood insult hypotheses for the origins of adult disease may have a role in explaining health disparities[35,36].

'Race' and research

Modern human genetic variation does not structure into phylogenetic subspecies (geographical 'races'), nor do the taxa from the most common racial classifications of classical anthropology qualify as 'races' (**Box 1**). The social or ethnoancestral groups of the US and Latin America are not 'races', and it has not been demonstrated that any human breeding population is sufficiently divergent to be taxonomically recognized by the standards of modern molecular systematics.

These observations are not to be taken as statements against doing research on demographic groups or populations. They only support a brief for linguistic precision and careful descriptions of groups under study. Terms and labels have qualitative implications.

Detailed description of study populations and their specific histories is advocated. The study of well-defined local populations of demographic groups of the same name should be carried out in order to understand possible gene-environment effects. Likewise, data from nationwide studies on particular demographic groups should always be disaggregated by locale. Local names should replace macrodesignations in studies in order to reflect specific populations. Generalizations that invoke 'genetic' explanations are to be avoided unless they are warranted. All of these have policy implications for health studies.

'Racial' thinking can still be found in scientific literature[15]. Evolutionary and other biohistorical studies should be model-based and should acknowledge the ongoing legacy of 'racial' thinking. Collaborations with experts in appropriate fields such as historical linguistics, archaeology, ethnology and recent history would improve the quality of multidisciplinary studies.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Received 9 September; accepted 23 September 2004
Published online at http://www.nature.com/naturegenetics/

1. Andreasen, R.O. Race: Biological reality or social construct. *Philos. Sci. (Proc.)* **67**, S653–S666 (2000).
2. Keita, S.O.Y. & Boyce, A.J. "Race": Confusion about zoological and social taxonomies, and their places in science. *Am. J. Hum. Biol.* **13**, 569–575 (2001).
3. Andreason, R.O. A new perspective on the race debate. *Brit. J. Philos. Sci.* **49**, 199–225 (1998).
4. Lewontin, R. *Not In Our Genes* (Pantheon, New York, 1984).
5. Livingstone, F. On the non-existence of human races. *Curr. Anthropol.* **3**, 279–281 (1962).
6. Gordon, H. Genetics and race. *S. Afr. Med. J.* **39**, 533–536 (1965).
7. Stepan, N. The Idea of Race in Science: Great Britain 1800-1960 (London and Basingstoke, 1982).
8. Deniker, J. *The Races of Man* (Walter Scott, London, 1900).
9. Garn, S. *Human Races.* (McGraw Hill, Springfield, 1961).
10. Mayr, E. & Ashlock, P. *Principles of Systematic Zoology* 2nd edn. (McGraw Hill, New York, 1991).
11. Avise, J.C. & Ball, R.M. Principles of genealogical concordance in species concepts and biological taxonomy. *Oxf. Surv. Evol. Biol.* **7**, 45–67 (1990).
12. Wilson, E.O. & Brown, W.L. The subspecies concept and its taxonomic application. *Syst. Zool.* **2**, 97–111 (1953).
13. Brown, W.L. & Wilson, E.O. The case against the Trinomen. *Syst. Zool.* **3**, 174–176 (1953).
14. Ruvolo, M. Genetic diversity in hominoid primates. *Annu. Rev. Anthropol.* **26**, 515–540 (1997).
15. Keita, S.O.Y. & Kittles, R.A. The persistence of racial thinking and the myth of racial divergence. *Am. Anthropol.* **99**, 534–544 (1997).
16. Latter, B.D. Genetic differences within and between populations of the major human groups. *Am. Nat.* **116**, 220 (1980)
17. Lewontin, R.C. The apportionment of human diversity. *Evol. Biol.* **6**, 381–398 (1972).
18. Long, J.C. & Kittles, R.A. Human genetic diversity and the non-existence of biological races. *Hum. Biol.* **75**, 449–471 (2003).
19. Nei, M. & Roychoudhury, A.K. Evolutionary relationships of human populations on a global scale. *Mol. Biol. Evol.* **10**, 927–943 (1993).
20. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes.* (Princeton University Press, Princeton, New Jersey, 1994).
21. Underhill, P.A. *et al.* The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65**, 43–62 (2001).
22. Kruglyak, L. & Nickerson, D. Variation is the spice of life. *Nat. Genet.* **27**, 234–236 (2001).
23. Carlson, C.S. *et al.* Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat. Genet.* **33**, 518–521 (2003).
24. Allen, M., Salden, T., Patterson, U. & Gyllensten, U. Genetic typing of HLA class II genes in Swedish populations: applications in forensic analyses. *J. Forensic Sci.* **38**, 554–570 (1993).
25. Rosenberg, N. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
26. Serre, D. & Paabo, S. Evidence of gradients of human genetic diversity within and among continents. *Genome Res.* **14**, 1679–1685 (2004).
27. Collins, F.S., Brooks, L.D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**, 1229–1231 (1998).
28. Reich, D. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
29. Weiss, K.M. & Clark, A.G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**, 19–24 (2002).
30. 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).
31. Carlson, C.S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2004).
32. Neale, B. & Shan, P. The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet.* **75**, 353–362 (2004).
33. Stanford, J.L. *et al. Prostate Cancer Trends 1973-1995.* SEER Program, National Cancer Institute NIH Pub No 99-4543. (National Institutes of Health, Bethesda, Maryland, 1999).
34. Rout, L. The African Experience in Spanish America, 1502 to the Present Day (Cambridge University Press, Cambridge, 1976).
35. Barker, D.J.P. In utero programming of chronic disease. *Clin. Sci.* **95**, 115–128 (1998).
36. Sallout, B. & Walker, M. The fetal origin of adult diseases. *J. Obstet. Gynaecol.* **23**, 555–560 (2003).