

Commentary

Ethnicity, Ancestry, and Race in Molecular Epidemiologic Research

Timothy R. Rebbeck¹ and Pamela Sankar²

¹Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics, Abramson Cancer Center and ²Department of Medical Ethics and the Center for Bioethics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania

Why the Controversy?

The concepts of ethnicity, ancestry, and race are widely used in molecular epidemiologic research, often based on the assumption that these correlate (however roughly) with increased genetic homogeneity among people claiming a similar identity. However, some have questioned the implications of research on ethnic or racial differences, particularly in the area of health disparities research (1, 2). This emphasis may lead to an overly simplistic attribution of poor health to biological rather than social or political factors (3, 4). Thus, a danger of the use of genetics as a major determinant of health and health disparities is that stereotypes of health and disease in certain groups may be perpetuated and alternative solutions to these problems will be ignored. Similarly, others have postulated that self-identified race or ethnicity (SIRE) is primarily sociocultural rather than biological and their use in genetic research is invalid (5, 6). Some argue that ethnicity is built of social, legal, and historical factors and bears no necessary or predictable relationship to genetic factors (7, 8).

To theorists who eschew SIRE as a biological concept, its function as a variable is to capture predictable dimensions of a person's or a group's daily experience that are tied to shared beliefs and practices. Beliefs (and practices based on those beliefs) might include what people hold true about themselves as members of a recognized group. For example, researchers may capture an individual respondent's notion of belonging to a group, such as "As a member of group A, I prefer these foods." Those who define SIRE as a social construct grant that there might be biological consequences of membership in these groups, such as result from diet or family size, but the consequences, without basis in genetics, are mutable and transient. According to this thinking, membership and boundaries of SIRE groups evolve over time, reflecting and influencing political and cultural events. This notion is supported by a study reporting that one third of individuals asked to assign themselves to an ethnic group in 2 consecutive years chose a different ethnic group in the second year (9). Thus, to social constructionists, ethnicity labeled "self-identified" would have to be spontaneously named by a subject and could not be assumed to have genetic correlates.

In contrast, population and evolutionary genetics research has examined genetic variability within and between members of SIRE groups. For example, the F_{ST} statistic of Wright (10) has been used as a measure of subdivision within species, where $F_{ST} = 100\%$ suggests that genomic variability is completely between groups (i.e., the groups are genetically distinct) and $F_{ST} = 0\%$ suggests the absence of genetically distinct groups with all of the observed genomic variability occurring within populations. Using large-scale genomic information, the average value of F_{ST} in humans has been estimated to be $\sim 13\%$ (11). Although there is significant variability in the value of F_{ST} across the human genome, $F_{ST} = 13\%$ supports the existence of between-group genomic differences, but does not indicate that SIRE groups are genomically distinct. Nonetheless, biological correlates of self-identified race or ethnicity exist. Numerous large-scale genomic studies have reported that the observed distribution of allelic variation in the human genome differs among groups. Although most genetic variation is common to all populations, genetic variants exist that are unique to specific SIRE groups (12). Using a genetic-cluster analysis algorithm to group individuals based on their genetic marker information alone, Tang et al. (13) reported that among 3,636 participants from four multicenter studies of blood pressure using 326 microsatellite linkage markers, only 17 individuals (0.5%) did not cluster according to SIRE. Although these results are striking, the analyses did not consider admixture within populations and thus overemphasized the distinctions between SIRE groups. Nonetheless, these and other data support the notion that correlations between genomic variability and SIRE exist, and thus refute those who reject any genetic basis to race or ethnicity.

These constructs are apparently at odds with one another and emphasize the need to define carefully and use appropriately the concepts of race, ancestry, or ethnicity in molecular epidemiologic research. We propose that it is not productive to continue the debate about whether racial or ethnic groups do or do not have genomic correlates. Rather, the need is to recognize that multiple, distinct concepts of race or ethnicity exist, and to move molecular epidemiologic research toward effectively characterizing and using these.

Ancestry, Culture, Environment, and Phenotype

Because SIRE can be correlated with both biological and sociocultural phenomena, models that consider only one of these explanations are ineffective for molecular epidemiologic research. Thus, we consider group membership as a complex, multifactorial trait similar to height, blood pressure, intelligence, or common disease phenotypes. Figure 1 suggests a framework for considering SIRE that includes both biological phenomena as well as environmental and sociocultural differences that modify the relationship between genotype

Cancer Epidemiol Biomarkers Prev 2005;14(11):2467–71

Received 8/23/05; revised 8/29/05; accepted 9/7/05.

Grant support: Public Health Service grants R21-ES11658, R01-CA85074, and P50-CA105641 (T.R. Rebbeck), and R01-HG03191 to PS, and University of Pennsylvania Abramson Cancer Center.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Timothy R. Rebbeck, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 904 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021. Phone: 215-898-1793; Fax: 215-898-2265. E-mail: trebbeck@ceceb.med.upenn.edu

Copyright © 2005 American Association for Cancer Research.

doi:10.1158/1055-9965.EPI-05-0649

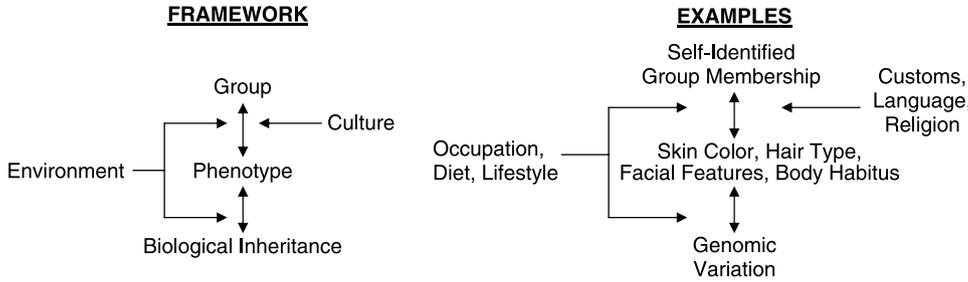


Figure 1. Framework for defining the influence of biological inheritance, phenotype, environment, and culture on group membership identity.

and phenotypes. These categories are observational and relational. They are observational in the sense that their component features are those that interest molecular epidemiology rather than those that group members might cite. These component features might overlap with features that group members cite but the premise for emphasizing these features is their relevance to research rather than to social identity. These categories are relational in that they are meaningful only in the context of specific comparisons and are not discrete entities that stand alone outside a particular research context. For example, genomically similar groups that are culturally dissimilar may represent major or minor comparisons between ethnic groups, depending on the context of the research question.

We use the generic term “group” to denote self-identified group membership. “Biological inheritance” denotes those innate factors that capture information about genomic variability and historical events, such as migration or selection, that have influenced the current patterns of genomic variation. Biological inheritance is correlated with phenotypes that in part define the physical features associated with SIRE. These phenotypes include skin color, hair type, facial features such as eye shape or nasal bridge structure, and body habitus. These phenotypes can be further correlated with group membership. The strength of the correlation between biological inheritance and phenotype and between phenotype and group can also be influenced by an individual’s environment, including exposures such as diet, lifestyle, occupation, or residence. Finally, the relationship between phenotype and group is further influenced by cultural factors such as religion, language, and social customs.

The use of ethnicity, ancestry, or race in molecular epidemiologic research is to account for differences that reflect etiologic or prognostic differences or to ameliorate study biases. Assuming that we can identify features that allow us to define group differences in epidemiologic research (i.e., ancestry, culture, environment, and phenotype; Fig. 1), we attempt to construct concepts of ethnicity, ancestry, or race that can serve research effectively and which recognize and benefit

from the complexity inherent to these concepts. To this end, we define four general terms that may be used when comparing groups of individuals. These do not represent discrete categorizations but are landmarks on a continuum (Fig. 2).

“Minor ethnicity” refers to locally constituted (i.e., physically proximal) groups who think of their members as similar to one another and distinct from neighboring groups (other minor ethnicities), although these different groups share many genomic, cultural, and environmental factors in common. The actual cultural or environmental differences between groups distinguished as different minor ethnic groups may be negligible and arbitrary, particularly if ethnic labels have been assigned by outsiders without reference to differences that might be recognizable to members of those groups. Alternatively, these differences may be apparent to both group members and outsiders (7). The Nuer and the Dinka of East Africa are examples of minor ethnicities. Although minor ethnic groups are similar in many ways, differences in diet, reproductive patterns, or other behaviors may influence disease.

“Major ethnicity” can be used to contrast groups that share some degree of common ancestry, but have diverged to some degree in terms of culture (e.g., language, religion) and environment. Natives of various European countries or the peoples of Northeast Asia are examples of major ethnicities. There is generally no presumed major genomic basis for differences among these groups.

“Ancestry” can be used to define comparisons of groups that are genomically divergent, but share cultural or environmental similarities. An example of ancestral groupings includes African Americans and European Americans, who share many common cultural and environmental characteristics, but diverge in terms of their genomic (geographic) ancestry.

“Race” can be used to characterize comparisons of groups that diverge in most respects. For example, Native Australian Aboriginals and East Asians share relatively fewer genomic, cultural, or environmental characteristics.

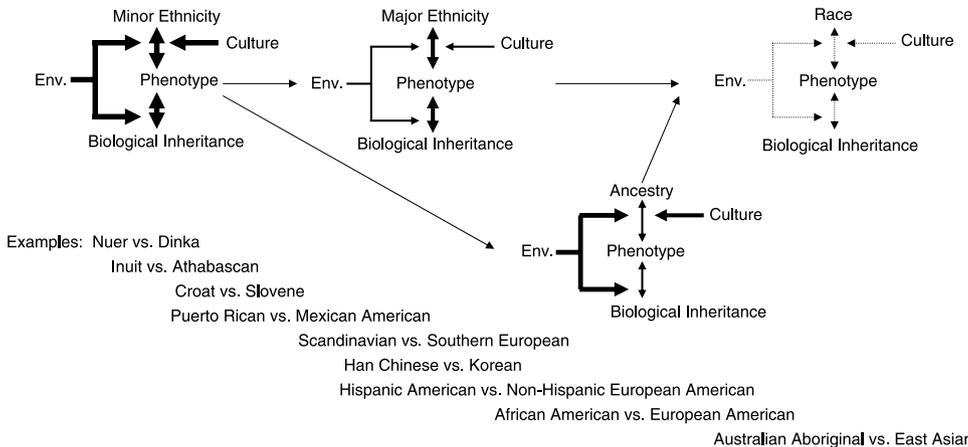


Figure 2. Relationships among ethnicity, ancestry, and race. Weight of the arrows reflects the strength of correlation between groups being compared.

Downloaded from <http://aacrjournals.org/cebp/article-pdf/14/11/2467/2263054/2467.pdf> by guest on 20 May 2022

This proposed nosology is meant to help to organize thinking rather than to represent a fixed underlying axis of differentiation among human groups. As in any attempt to define such complex concepts, numerous caveats must be considered. The components of the frameworks in Figs. 1 and 2 are potential, but not required, influences in determining group boundaries. To emphasize the point that these definitions are arbitrary and meant to depict a continuum to be used to conduct research, we point out that narrowly defined groups (minor ethnicities) may still constitute "racial" comparisons if they are sufficiently distinct in terms of genomic, cultural, and environmental characteristics (e.g., African Nuer versus Italian Calabrians). The same subjects could fall into different groups in different research projects. Cosenzans compared with other Calabrians may represent minor ethnicity comparisons (14). If Calabrians and Cosenzans were contrasted with Scandinavians, the internal distinctions disappear and both Calabrians and Cosenzans would become Southern Italian, members of a major ethnicity. In yet another project, distinctions among these populations might be submerged and the group as a whole is treated as European and contrasted with African as a racial comparison. An additional caution is to remain aware that, for historical and social reasons, some populations will fit more easily into certain categories. For example, the differences between Sicilians and Swedes seem relatively easy to characterize. However, comparisons of Puerto Ricans and Mexican Americans, or between Hispanic Americans and Non-Hispanic European Americans are complex, with the former not clearly a minor ethnicity nor a major ethnicity and the latter arguably definable as either major ethnicity or ancestry. Using this framework, the ease or difficulty with which labels can be assigned to research populations or populations can be assigned to categories is understood as an artifact of a particular research question rather than a feature of the group.

Why Use Ethnicity, Ancestry, or Race?

Given the many concerns about the meaning and use of ethnicity, ancestry, or race in epidemiologic research, why would a molecular epidemiologist choose to include these concepts in their research? As outlined below, study bias or inefficiencies may result if these concepts are ignored.

High-Risk Groups. Ethnicity, ancestry, or race can serve as surrogate measures to identify high-risk groups. Groups that have a particularly high incidence or strong familial aggregation of disease may represent an optimal resource in which to identify or characterize disease genes. For some diseases or traits, increased incidence or aggregation may identify exposed-predisposed groups. This is likely to be the case in diseases with a complex, multifactorial etiology caused by the interaction of inherited genotypes and exposures. Similarly, those prone to poor treatment response or increased toxicities also represent "high-risk" groups that may be, in part, determined by their genome. Although it has been proposed that treatment may be based on ethnicity alone (15), it is likely that ethnic-specific differences in treatment response are in fact determined by specific metabolic genotypes, the frequency of which may vary by ethnicity. Thus, the field of pharmacogenetics is likely to contribute to improved individual-specific, rather than race-specific, treatment.

Genetic and Etiologic Heterogeneity. Because geography and migration histories of populations share sufficient overlap with socially constructed categories of race or ethnicity, socially constructed categories have been widely used as an index of genetic homogeneity. Because exposures are funda-

mentally shaped by social environments, ethnicity, ancestry, and race may be used to operationalize different experiences of exposure. Most common diseases and phenotypes are under the influence of numerous etiologic agents, including both genes and environmental exposures. To the extent that socially defined ethnicity and allele frequencies are similarly patterned, choosing subjects by ethnicity may affect the genetic homogeneity of a study population. Gene identification studies have been proposed that specifically use admixed populations (e.g., mapping by admixture linkage disequilibrium; ref. 16). Similarly, founder mutations that occur in specific (usually culturally or geographically isolated) populations provide an opportunity to study homogeneous subgroups (e.g., French Canadians, Icelanders, and Ashkenazi Jews). For example, common founder mutations in the *BRCA1* and *BRCA2* genes in Jewish populations simplify genetic testing for clinical risk assessment and identify a relatively common limited set of background mutations in which to better understand *BRCA1*- and *BRCA2*-associated carcinogenesis.

Study Efficiency. Appropriate study sample size and power are critical aspects of all epidemiologic studies. Power is dependent on the frequency of the exposure or genotype being studied. Because genotype frequencies tend to vary by ethnicity, ancestry, or race, studies may have inadequate power or may be inefficient (i.e., using larger sample sizes than may be necessary) if race- or ethnicity-specific estimates of genotype or exposure frequencies are not considered. Similarly, genome scans that use concepts of linkage and linkage disequilibrium may suffer if ethnicity is not properly considered. Genome-wide association or linkage methods are dependent on the linkage disequilibrium among genetic variants on a chromosome. Differences in the pattern of linkage disequilibrium by race have been reported (17); this could affect the success of gene discovery efforts.

Study Biases. Numerous authors have examined the implications of unrecognized population structure to induce bias, false-positive associations, and lack of replication in association studies (18-28). Population stratification (i.e., confounding by ethnicity) can occur if both baseline disease risks and risk-conferring allele frequencies differ across the groups being studied (e.g., races or ethnicities). If either of these criteria is not fulfilled, this bias cannot occur. Although there are numerous examples of genotype frequency differences by ethnicity and disease risk differences by ethnicity, the evidence to date suggests that potential biases are small if baseline disease or allele frequency differences between ethnicities are small, diminish as admixture increases (e.g., if the number of component ethnicities is large, as may be the case in African Americans; refs. 24, 29), and may be more pronounced in recently admixed populations (30). However, it is also clear that population structure can be observed even in apparently homogeneous populations (31). Thus, a variety of methods have been proposed to correct for the biases that may result from these ethnic or ethnic differences, including the use of multiple unlinked markers to correct for or quantify potential biases (19, 32) or study designs that address potential bias by using relatives or matching strategies (33, 34).

The literature to date indicates that potential biases can be corrected for by the usual methods of statistical adjustment in study samples comprised of widely divergent races or may be unnecessary in genomically homogeneous populations (e.g., Northern Europeans; refs. 29, 30). However, studies of admixed populations (e.g., African Americans) represent a more complex situation. Millikan (35) reported that a minimal bias in point estimates was observed in African American populations. Ardlie et al. (30) observed the potential for population structure to exist in African American populations,

but also that this structure was reduced by removing recent African or Caribbean immigrants. Wang et al. (24) used simulation data that infer that odds ratios in association studies of unrelated African Americans would not be markedly distorted even under conditions of extreme differences in baseline disease risk and genotype frequency in the component admixed populations. Nonetheless, despite the limited evidence that population stratification causes biases to estimates in most common situations, it is clear that ignoring ethnicity in molecular epidemiologic studies can lead to some distortion of estimates of association. Therefore, all studies should carefully consider the potential for confounding by ethnicity, ancestry, or race, and respond with appropriate study design or analytic methods (28).

What is a Researcher to Do?

Although there are concerns about the use of ethnicity, ancestry, or race in molecular epidemiologic research, this information may prove valuable or necessary to achieve meaningful study results. Proper consideration of these concepts should be made in the early stages of research. Thus, molecular epidemiologic researchers must consider the proper use and context when applying ethnicity, ancestry, or race to ensure that these concepts enhance the value of research and do not undermine the translation of this research to improved human health.

Despite the widespread use of ethnicity, ancestry, or race in certain kinds of genetic research, details on how these concepts are defined or used are often sparse in the literature: Authors sometimes neglect to define the terms or to explain how labels were chosen or assigned to subjects. Similarly, different approaches are used in creating these definitions. For example, Tang et al. (13) relied on self-identified ethnicity; however, two study sites asked participants to choose one SIRE from among the four to seven provided by researchers; a third allowed participants to self-describe without a list of choices but then recorded all responses as "other" that did not correspond to Caucasian/White or African American. In certain analyses, those who identified themselves as "other" were excluded. The fourth site required participants to self-describe as either Chinese or Japanese and to identify four grandparents as either Chinese or Japanese. Although all of these procedures allowed subjects to choose their preferred SIRE, they also limited the categories with which subjects could identify. Operationalizing SIRE in this way is standard and, in itself, poses no threat to conclusions. However, these points emphasize how differently researchers may conceptualize ethnicity.

Most studies make use of a fairly crude measure of ethnicity, ancestry, or race that may not fully reflect group membership as a complex trait. Thus, although self-identified race or ethnicity may capture some of the relevant genetic, cultural, environmental, or phenotypic influences that determine group membership, it is likely to be an inherently misclassified measure of the true quantity of interest.

The effect of variable misclassification in epidemiologic research is well known and may lead to biases of various types. In the context of a (nested) case-control study design, a nondifferentially misclassified variable can lead to bias toward the null hypothesis. If used as a confounder, a misclassified variable will be less likely to account for confounding in the relationship between the risk factor of interest and disease risk. If the misclassification is differential, as might be expected if the relationship of a genotype and disease differed between ethnic groups and between cases and controls, the direction and magnitude of bias are unpredictable and could include bias of estimates either toward or away from the null hypothesis. Furthermore, coding of a self-identified ethnicity variable may lead to additional modeling errors in addition to

misclassification. For example, if self-identified ethnicity is coded for analysis as an ordered discrete variable (e.g., 0 = European, 1 = African, 2 = Asian), this ordering of an otherwise unordered variable may cause inferential errors.

Like many other variables, the analytic variable describing ethnicity, ancestry, or race should be constructed with regard to the specific research setting and hypotheses. As described above, self-identified group membership can be thought of as a complex continuously distributed variable under the influence of multiple factors, including genes, culture, and environment. By analogy, studies of hypertension may consider blood pressure as a continuous trait or may use hypertension as a discrete outcome. Continuous blood pressure measurements capture the complex distribution of this trait and may better reflect multiple exposures and genetic influences at a given point in time or under a specific circumstance. However, this measure may also be biased if, for example, some study subjects were taking antihypertensive medication. If so, the continuous blood pressure measurement may not reflect a relevant biological entity, and a binary variable such as "hypertensive" versus "normotensive" may be more valid. Similarly, the use of ethnicity, ancestry, or race may be best defined by a more complex composite (and possibly semi-continuous or continuous) variable in some research settings but may be more appropriate as a discrete variable with relatively few levels for other research questions. Recent studies of genomic markers have opened the door to purely genomic definitions of these concepts (36).

The absence of detail and consistency about how ethnicity, ancestry, or race are defined may affect replicability or comparability of studies and leaves the field vulnerable to complaints that its use of the terms meant to capture ethnicity, ancestry, or race is not scientifically adequate (37-39). Several journals request detailed information and standardization in papers that use concepts of ethnicity, ancestry, or race in genetics research (40, 41). To those, we add the following considerations:

Assess the Need for Using Ethnicity, Ancestry, or Race. Researchers should consider the use of these concepts before analysis. For example, is race being included out of habit as a baseline demographic variable? Is this information needed to test proposed hypotheses? Alternatively, if the concepts are being used as proxies for something else, can improved measures be developed? For example, would socioeconomic status provide a better measure of potential confounding than self-identified race?

Decide How Ethnicity, Ancestry, or Race Will be Used and What these Terms or Terms Will Mean in this Particular Study. Much of the persistent controversy over the use of the terms "ethnicity," "ancestry," or "race" may be attributable to the imprecision of their use. Ideally, the definition and use of these terms in different settings will share common features. It behooves users to explain their definitions of the meaning of these variables, not as general concepts, but making clear the particular way they are using them. For example, an appropriate concept for some genetics studies may be to consider the term "ancestry" (Figs. 1 and 2), which can capture concepts of genomic variation, biology, or geographic history. Studies may also use ethnicity to refer to regulatory or bureaucratic categories or to social identity as in aspects of access to health care that might relate to discrimination. These specific situations should be described to clarify meaning and use.

Choose the Term or Terms that Best Fits this Use or Meaning. Several attempts to address the debate about ethnicity, ancestry, or race in genetics have been made. Some disciplines have proposed new terminology as a solution, thus there are many terms from which to choose. These include many questionable neologisms such as macroethnicities or

“race/ethnicity” (42). The enormous variation in terms used today suggests that standardized terminology is not helpful and that a careful explanation of how and why one is using a specific term is likely to be the most appropriate approach. Criteria used to decide what term(s) best fit the proposed use might include prior practice in the literature, labels preferred by study participants, or categories that match important databases.

Be Consistent. Adhere to the chosen term or terms and vary terms only when there is different meaning that needs to be conveyed. If there is a different meaning to convey, consider a brief explanation in the text as to why was required.

Epidemiologic studies attempting to identify or characterize disease genes almost always use socially constructed measures of ethnicity, ancestry, or race. These measures may have utility in increasing study efficiency or reducing confounding. However, an important future direction for research will be to develop new measures that correlate with SIRE that may better reflect the complex nature of this variable.

References

- Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 2002;12:1805–14.
- Ardlie KG, Lunetta KL, Seielstad M. Testing for population subdivision and association in four case-control studies. *Am J Hum Genet* 2002;71:304–11.
- Braun L. Race, ethnicity, and health: can genetics explain disparities? *Perspect Biol Med* 2002;45:159–74.
- Briscoe D, Stephens JC, O'Brien SJ. Linkage disequilibrium in admixed populations: applications in gene mapping. *J Hered* 1994;85:59–63.
- British Medical Journal. Style matters: ethnicity, race, and culture: guidelines for research, audit, and publication. *BMJ* 1996;312:1094.
- Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003;361:598–604.
- Cooper PS, Kaufman JS, Ward R. Race and genomics. *N Engl J Med* 2003;348:1166–70.
- Barbaro A, Cormaci P, Falcone G, Barbaro A, Rizzo M. Genetic study of 11 Y-STRs in the populations of Reggio Calabria, Catanzaro, Cosenza (Calabria-South of Italy). *Forensic Sci Int* 2004;146 Suppl:S129–31.
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;55:997–1004.
- Freedman ML, Reich D, Penney KL, et al. Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004;36:388–93.
- Gauderman WJ, Witte JS, Thomas DC. Family-based association studies. *J Natl Cancer Inst Monograph* 1999;26:31–7.
- Goodman AH. Why genes don't count (for racial differences in health). *Am J Public Health* 2000;90:1699–702.
- Hamer D, Sirota L. Beware the chopsticks gene. *Mol Psychiatry* 2000;5:11–3.
- Hartl DL, Clark AG. Principles of population genetics. 3rd ed. Sunderland (MA): Sinauer Associates; 1997.
- Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K. An Icelandic example of the impact of population structure on association studies. *Nat Genet* 2005;37:90–5.
- Hinds DA, Stokowki RP, Patil N, et al. Matching strategies for genetic association studies in structured populations. *Am J Hum Genet* 2004;74:317–25.
- Hirschhorn J, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 2002;4:45–61.
- Instructions for authors. *JAMA* 2005;294:119–24.
- Khlat M, Cazes M-H, Genine E, Guiguet M. Robustness of case-control studies of genetic factors to population stratification: magnitude of bias and type I error. *Cancer Epidemiol Biomarkers Prev* 2004;13:1660–4.
- Kittles RA, Chen W, Panguluri RK, et al. Cyp3A4-V and prostate cancer in African Americans: causal or confounding association because of population stratification? *Hum Genet* 2002;110:553–60.
- Leech K. A Question in dispute: the debate about an ethnic question in the census. London: Runnymede Research Reports; 1989. p. 40.
- Macbeth HM. What is an ethnic group? A view from biological anthropology. In: Clarke A, editor. Culture, kinship and gene. London: Macmillan; 1997.
- Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet* 2004;36:512–7.
- Millikan RC. Re: Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst* 2001;93:156–8.
- Nebert D. Suggestions for the nomenclature of human alleles: relevance to ecogenetics, pharmacogenetics and molecular epidemiology. *Pharmacogenetics* 2000;10:279–90.
- Oppenheimer GM. Paradigm lost: race, ethnicity, and the search for a new population taxonomy. *Am J Public Health* 2001;91:1049–55.
- Parra EJ, Kittles RA, Argyropoulos G, et al. Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. *Am J Phys Anthropol* 2001;114:18–29.
- Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999;65:220–8.
- Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* 2001;20:4–16.
- Reich DE, Schaffner SF, Daly MJ, et al. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 2002;32:135–42.
- Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic structure of human populations. *Science* 2002;298:2381–5.
- Sankar P, Cho MK, Condit CM, et al. Genetic Research and Health Disparities. *JAMA* 2004;297:2985–9.
- Schwartz RS. Racial profiling in medical research. *N Engl J Med* 2001;344:1392–3.
- Sankar P, Kahn J. BiDil: race medicine or race marketing? *Health Aff. In press* 2005.
- Santelli JS, Lowry R, Brener ND, Robin L. The association of sexual behaviors with socioeconomic status, family structure, and race/ethnicity among US adolescents. *Am J Public Health* 2000;90:1582–8.
- Sheilds A, Fortun M, Hammonds E, et al. The use of race variables in genetic studies of complex traits and the goal of reducing health disparities: a transdisciplinary perspective. *Am Psychol* 2005;60:77–103.
- Tang H, Quertermous T, Rodriguez B, et al. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am J Hum Genet* 2005;76:268–75.
- Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev* 2002;11:505–12.
- Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst* 2000;14:1151–8.
- Wang Y, Localio R, Rebbeck TR. Evaluating bias due to population stratification in case-control association studies of admixed populations. *Genet Epidemiol* 2004;27:14–20.
- Wang Y, Localio R, Rebbeck TR. Consideration of the minimal number of unlinked markers required to correct for bias due to population stratification in candidate gene association studies. *Hum Hered* 2005;59:165–75.
- Witzig R. The medicalization of race: scientific legitimization of a flawed social construct. *Ann Intern Med* 1996;125:675–9.