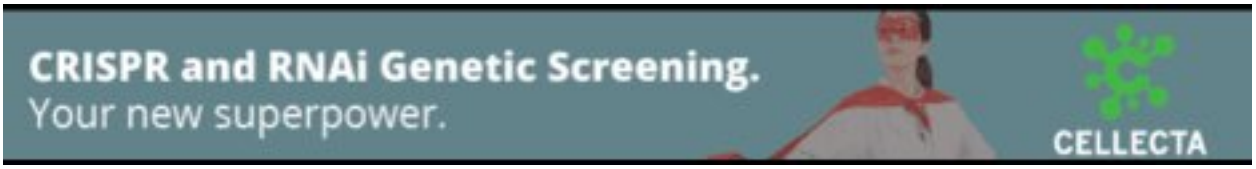
Race, Ethnicity, and Genomics: Social Classifications as Proxies of Biological Heterogeneity

Morris W. Foster and Richard R. Sharp

Genome Res. 2002 12: 844-850
Access the most recent version at doi:10.1101/gr.99202



Commentary

Race, Ethnicity, and Genomics: Social Classifications as Proxies of Biological Heterogeneity

Morris W. Foster[1,3] and Richard R. Sharp[2]

[1]Department of Anthropology, University of Oklahoma, Norman, Oklahoma 73019, USA; Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma 73104, USA; [2]National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina 27709-2233

Over the past century, genetics has experienced a tension between the view that racial and ethnic categories are biologically meaningful and the view that these social classifications have little or no biological significance. That tension continues to inform genomics and is evident in the assembly of biological collections and sequence databases that seek to approximate the genetic variation found in human populations. Although social identities can be useful and convenient proxies of some biological features, for example, in ensuring that genomic resources capture a range of genetic variants found in most human populations, the ways in which geneticists conceptualize the relationship between racial and ethnic identities and genetic variation can be problematic. Inclusion of racial and ethnic identifiers in genomic resources can create risks for all members of those identified populations and influence lay perceptions of the nature of racial and ethnic groups. Thus, the burden of showing the scientific utility of racial and ethic identities in the construction and analysis of genomic resources falls on researchers. This requires that genetic researchers pay as much attention to the social constitution of human populations as presently is paid to their genetic composition.

Concepts of race and ethnicity are among the most controversial, contentious, and misunderstood classifications in our social and scientific landscapes. Among geneticists, there are two dominant historical perspectives on race and ethnicity. On the one hand, there is a perspective rooted in the eugenics movement and early population genetics that treats racial and ethnic categories as biological classifications (Kevles 1995), attempting to use scientific analysis to specify the precise nature of presumed biological differences between those socially labeled populations (Huxley 1951). On the other hand, there is a perspective with its roots in the work of physical anthropologists and social scientists who argue that race and ethnicity are primarily cultural and historical constructions with little biological significance (Boas 1942). Although there has been much debate among geneticists regarding these two competing perspectives throughout the 20th century, geneticists at the beginning of the 21st century have yet to reconcile these divergent views on race and ethnicity.

In this regard, a mid-century celebration of the first 50 years of modern genetics by Laurence Snyder (1951), one of the first presidents of the American Society of Human Genetics, bears a noteworthy resemblance to the present state of affairs:

"Human populations differ one from another almost entirely in the varying *proportions* of the allelic genes of the various sets of hereditary factors, and not in the *kinds* of genes they contain. The extreme positions held by those who on the one hand maintain that there are no significant genetic differences between human races, and those who on the other hand hold that certain races are 'superior' and others 'inferior,' require drastic modification in the light of the accumulated data on the gene frequency dynamics of human populations."

Some 50 years later, researchers have accumulated a far more extensive collection of data on allelic distributions within and between human populations. Despite this ever-increasing pool of information, however, geneticists have yet to resolve that basic tension between social and biological conceptions of race and ethnicity.

Debates about race and ethnicity have changed in one important respect—today nearly all geneticists reject the idea that biological differences belie racial and ethnic distinctions. Geneticists have abandoned the search for "Indian" or "African" genes, for example, and few if any accept racial typologies. Even so, although simplistic biological interpretations of race and ethnicity have been discredited for decades, studies in clinical and population genetics continue to associate biological findings with the social identities of research participants. In some cases, scientists use genetic features to reconstruct population histories, applying biological criteria to define and redefine commonalities among research participants recruited on the basis of shared linguistic, racial, ethnic, or geographical identities. In these and other cases, researchers name the racial or ethnic communities being studied, thereby implicitly indicating that genetic features can be used to characterize contemporary social populations. Thus, although the simplistic biological understanding of race and ethnicity associated with the eugenics movement may be dead, the far more subtle presumption that racial and ethnic distinctions nonetheless capture "some" meaningful biological differences is alive and flourishing (Kaufman and Cooper 2001).

It was hoped by some that the sequencing of the human genome would undermine the view that racial and ethnic

[3]Corresponding author.
E-MAIL Morris.W.Foster-1@ou.edu; FAX (405) 325-7386.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.99202.

12:844–850 ©2002 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

classifications have biological significance (Gilbert 1992). This position was based on the prospect that by showing that there are numerous genetic similarities across all social classifications and no genetic features that are entirely unique to any particular racial or ethnic population, genomics would provide definitive evidence that race and ethnicity are social, not biological, classifications. Ironically, the sequencing of the human genome has instead renewed and strengthened interest in biological differences between racial and ethnic populations, as genetic variants associated with disease susceptibility (Collins and McKusick 2001), environmental response (Olden and Guthrie 2001), and drug metabolism (Nebert and Menon 2001) are identified, and frequencies of these variants in different populations are reported.

This renewed interest in identifying interindividual and interpopulation differences has prompted the construction of several types of genomic resources designed to facilitate the study of human genetic variation. Existing genomic resources include (1) biological materials for use in coordinated sequencing projects, (2) electronic databases of sequence information, and (3) databases describing genotype-phenotype associations. These resources seek to describe, as best as possible with existing technologies, the broad range of genetic variation that exists across human populations. For example, anthropologists and evolutionary biologists are interested in creating a collection of human cell lines for the study of genetic diversity among culturally defined populations (http://www.nsf.gov/pubs/2001/nsf01120/nsf01120.html). Other projects include a cell line collection for the discovery of single-nucleotide polymorphisms (SNPs; Collins et al. 1998), a publicly funded electronic database for clinical investigations of genetic influences on drug response (http://www.pharmgkb.org), a research database for investigators examining how genetic variation influences response to adverse environmental agents (Olden and Wilson 2000), and, more recently, a federally supported collection of cell lines for use in assembling a comprehensive human haplotype map (Helmuth 2001).

In this paper, we discuss the use of race and ethnicity in assembling genomic resources and examine how this practice may affect how research questions are framed, how findings are translated into clinical applications, how race and ethnicity will influence access to medical services, and how race and ethnicity are perceived in nonmedical contexts. Although it has been argued that "race" is an irrelevant analytic concept for genomics (Rothstein and Epps 2001), the alternative use of "ethnicity" often functions as a racialized euphemism rather than a conceptual break from treating social identities as biological categories (Lee et al. 2001a; Oppenheimer 2001). Consequently, we consider both race and ethnicity in our discussion.

Variation among and within Social Populations

To capture the vast number of genetic differences that exist and may be of biological importance, genomic resources must themselves be genetically diverse. A critical question facing geneticists, however, is how best to capture this genetic diversity in the construction of those resources (Przeworski et al. 2000). Except for genetic features related directly to a particular phenotype, there are no self-evident biological criteria for identifying donors who represent a range of human variation. In the absence of such criteria, social categories—such as race, ethnicity, nationality, and geographic locality—often are used to approximate the range of genetic variation that might exist across populations (Weber 1999; Stephens et al. 2001). In other words, genomic researchers frequently rely on the "social" identities of sample donors to ensure the "biological" heterogeneity of the genetic materials they collect and analyze.

A difficulty with this approach is that social classifications may correspond more or less precisely with the biological variation of interest to geneticists (Marks 1995). This makes the use of racial and ethnic classifications as a proxy for genetic heterogeneity paradoxical: On the one hand, the use of these social identities can be critical for assembling biologically diverse genomic resources; on the other hand, using these social categories in the construction of genomic resources indicates a substantive biological significance that racial and ethnic classifications do not necessarily possess. How best to resolve this tension, and how to help the lay public understand the manner in which race and ethnicity are being conceived by scientists, is critical to the success of the genomic sciences. Nonetheless, both the empirical and conceptual relationships between social populations and biological features can be subtle (Juengst 1998).

With regard to empirical relationships, for example, some geneticists have argued that most common complex diseases will have genetic contributors that can be found in most social populations (Chakravarti 1999; Daly et al. 2001). That claim is the basis for population-specific genomic projects, including the deCODE project in Iceland and a similar undertaking in Estonia. In those projects, phenotype-genotype linkages within a single social population are being investigated with the goal of identifying specific alleles associated with common diseases across populations (Gulcher and Stefannson 1998; Frank 2000). This approach has been supported by recent haplotype studies that indicate ancestral haplotypes are indeed common across populations with contrasting social histories (Wilson et al. 2001). These findings are very preliminary, however. It may still be the case that variation in genomic structures related to disease susceptibility, drug metabolism, and/or environmental response is often specific to socially defined populations. For instance, current evidence for genetic contributors to type 2 diabetes indicates that variants associated with that disease in some European populations (e.g., localized Finnish and French populations) are not associated with the disease in Mexican American and Native American populations (Ghosh et al. 2000; Horikawa et al. 2000; Vionnet et al. 2000; Wolford et al. 2001;). If this finding is typical of many other common diseases as well, then genomic resources and sequencing projects that focus on variation across populations may not provide sufficient resolution to discover intrapopulation variants (Weiss and Clark 2002). Thus, there currently is insufficient data to determine the precise relationships that exist between socially defined populations and noteworthy genetic features.

Other difficulties in how we think about the relationships between social and biological classifications can be seen in the controversy surrounding the Human Genome Diversity Project (HGDP), an unsuccessful effort to establish a collection of DNA samples from socially, culturally, and linguistically defined populations from around the world (Greely 2001). Proponents of the HGDP were interested in sampling from small, geographically isolated populations thought to have experienced relatively little or relatively recent admixture (Cavalli-Sforza et al. 1991). Their hope was to correlate specific genetic features with particular language families, cul-

tural affiliations, and racial and ethnic identities (Kidd et al. 1993). Members of some of those populations to be sampled, however, argued that the HGDP effort amounted to reducing distinctive histories of social populations to analyses of shared biological ancestry, in effect giving priority to evidence for genetic homogeneity in the face of compelling records of considerable linguistic, cultural, and social interaction and heterogeneity (Macilwain 1996). The historical processes that have produced that heterogeneity include intermarriage; cultural preferences for claiming only one's father's or mother's identity; fictive and adoptive kinship; instrumental and situational choices in asserting alternative identities; colonialist, racist, and nationalist ideologies that impose new identities on subjugated populations; voluntary and forced migration; and economic, religious, and other barriers to reproduction within larger social categories. These processes (and others) undermine the view that small "isolated" populations with distinctive racial, ethnic, linguistic, or cultural identities are better proxies of biological relatedness or homogeneity than are larger populations comprised of members with multiple such identities.

Despite these problems, social categories cannot be completely dismissed as irrelevant to genomic analysis. The lines that exist between social categories can constitute partial barriers to interaction, reproduction, and migration. Consequently, some members of a particular social population may share similar SNPs, similar conserved or ancestral haplotypes, or similar variants of disease-susceptibility, drug metabolism, and environmental-response genes in frequencies and patterns that are not found in other social populations. Those differences often are owing to distinctive demographic structures that result from historical events such as population bottlenecks and founder effects that affect the ways in which genetic features are distributed in subsequent generations. These historical events also influence the techniques that geneticists use to discover genetic features. For example, some of the techniques that have been most successful in identifying disease genes, such as linkage disequilibrium analysis, have been developed explicitly for use in the context of discrete, socially identifiable populations with unique demographic histories (Jorde 2000). Haplotype mapping techniques also depend on the analysis of discrete populations with differing kinds of demographic histories and structures (Goldstein and Weale 2001).

Clearly, deciphering the relationships that may exist between social classifications and biological categories is not a simple matter. The biological significance that a social distinction may have for one purpose can dissolve when those same social categories are used to answer other biological questions. Thus, it may be appropriate to use social categories as a proxy for biological relatedness (or unrelatedness) in some circumstances but not in others. Precisely because social categories like race and ethnicity are problematic as proxies for genomic features, the burden for showing their scientific utility in any specific resource or study should fall on researchers.

Efforts to meet this burden of establishing that social identities should be used in the construction of genomic resources can be divided into two general strategies. First, it has been claimed that the use of social identities helps increase the likelihood of representing a more complete range of variation in human populations, particularly if the social identities used correspond to relatively disparate geographical localities. For example, genomic resources developed using biological samples from members of social groups from different continents (and from different regions within each continent) can serve as a proxy for human genetic variation in general, even though they may not include samples from all racial or ethnic groups within each continent. Although imperfect because of subsequent population migrations and intermarriages, this proxy strategy remains powerful because of the initial but still relatively recent ancestral movement of human populations out of Africa into other continents. Second, the use of social identities in genomic research has been defended on the grounds that recruiting individuals with shared social identities may increase the likelihood that the collected samples reflect a common demographic history (such as a population bottleneck) that may be useful in identifying discrete health-related genomic structures or features (such as specific polymorphisms, genes, or haplotypes).

Notably, both of these two strategies use social categories as loose proxies of genetic features. In the first instance, differing social identities are used to approximate genetic differences in the human population as a whole, whereas in the second strategy, shared social identities are used to approximate common genealogical relationships. In contrast to these two applications, the use of social identity is most problematic, both scientifically and ethically, when used to associate all members of a population with a particular genetic feature or features. When social identities are used in this manner in the construction of genomic resources, the complex histories of social populations are reduced to naive biological characterizations.

Social Identifiability

The foregoing considerations indicate that there can be clear benefits (and potential harms) in using social classifications as a proxy of biological features in the construction of genomic resources. In addition to ensuring that DNA collections represent some approximation of human genetic variation among populations or a common demographic history within a population, socially identified samples also assist researchers in building a common literature that takes advantage of the carefully delimited biological significance of social populations and ensures the critical examination and replicability of findings. Nonetheless, social identities also can be used to assemble genomic resources without necessarily linking specific samples to particular racial or ethnic identifiers.

An important genomic resource for the discovery of polymorphic variation is the Polymorphism Discovery Resource (PDR). The PDR was assembled to approximate genetic variation among all human populations, consisting of samples from socially diverse populations within the United States (Collins et al. 1998). The creation of this resource caused considerable disagreement regarding the number of human cell lines needed to establish a standard reference set, which populations should be included in the collection, and what proportion of samples should come from each population (Marshall 1997). Much of this debate focused on scientific questions of inter- and intrapopulation variation. An additional consideration, however, was the potential social risks posed by the association of genetic features with particular racial or ethnic groups represented in the resource. To address these concerns, the PDR removed all social identifiers once the collection was established, thereby removing the possibility of using the PDR to estimate allelic distributions in racial or ethnic populations. As a result of this decision, the PDR has

been described as less useful to researchers than it might otherwise have been were this information retained (Davidson 2000).

Biological collections such as the PDR illustrate a difficult challenge in the construction of genomic resources, namely, how to decide whether the social identities of sample donors ought to be made available to users of the resource. Naming the socially defined populations of which individual donors are members means that *all* members of those populations may be affected by research findings, including those who did not consent to or take part in the resource. Some genomic resources, such as the CEPH (Centre du Etude Polymorphisme Humain) family samples and the proposed haplotype map repository, will be used by many more researchers and may take on greater significance than other resources (such as disease-specific DNA repositories) will. That higher profile may entail greater risks for both participating individuals (whose entire genomes may become publicly available on the Internet) and the social populations of which they are members (about which many more genomic findings may be generated than for other populations).

Certainly, research that does not explicitly link socially defined populations with genetic findings has fewer ethical risks. At the same time, the distribution of benefits resulting from genomic resources may depend on identifying populations from which study participants are recruited. For common alleles that cross social boundaries, identification is less crucial. However, for less-common alleles, or for rare variants of common alleles that tend to occur most frequently in the social populations in which they originated ancestrally, identification may be more important for clinical diagnosis, targeted delivery of appropriate health services, and public health education.

Choosing how to socially identify donors as members of social populations is not a trivial or self-evident matter. An individual donor, for instance, may be known simultaneously as a resident of a particular Indian village in Arizona, a member of the Hopi tribe, a descendant of a Laguna family (through a paternal ancestor who is not explicitly noted in matrilineal Hopi society), a Native American, and as someone of Spanish ancestry (owing to 18th-century intermarriages between Lagunas and Spaniards) in addition to being a member of the general U.S. population. Each of these identities, or several in combination, would have different social implications (as well as resulting in somewhat different scientific findings about genetic variation in relation to specific social categories), depending on which label or labels are placed on an individually anonymous DNA sample.

This potential for scientific inaccuracy in defining social categories and in recording them for specific donors could reduce the value of their association with specific genetic findings, while creating additional risks for members of the populations that happen to be named. Particularly in a larger socially defined population in which there is considerable difference in members' haplotypes and SNPs, a small number of donor participants may not represent the full range of genetic variation present in the population. Ascertaining the intrapopulation range of variation will thus be important for any specific downstream clinical benefits that basic genetic research may have for persons with that social categorization (apart from any general benefits from the discovery of common genes that occur across social boundaries).

Many genetic studies rely solely or primarily on participant self-reporting to assess membership in a socially defined population. Although this standard is appropriate for studying race and ethnicity as social phenomena, it does not allow for the critical scientific investigation of their biological accuracy. Moreover, genetic studies often use social labels such as "Chinese," "Nigerian," and "African American" that conceal a great deal of cultural, linguistic, and biological variation (Cooper et al. 2000). Even a more specific ethnic label such as "Yoruba," which refers to a population of >10 million people distributed across a large multinational region of West Africa, can mask significant intragroup variation (Reich et al. 2001). Consequently, the accuracy (or lack thereof) in the choice of social label for a group of study participants may affect the scientific validity of genetic findings for that larger social category.

Among members of smaller populations that are, in effect, a small number of extended families, there will be fewer differences in conserved haplotypes and SNPs than those among members of larger social categories (Tarazona-Santos et al. 2001). Those smaller population studies will, essentially, be pedigree studies and may require more stringent human subjects protections to preserve family and individual privacy (Botkin et al. 2001). Moreover, many smaller populations already experience the disadvantages of minority status within larger polities and, for that reason as well as their size, may be more vulnerable to social stigmatization. These observations indicate that smaller social populations (particularly those that already are economically or politically disadvantaged) should not be identified in genomic resources or publications unless there is the potential for direct benefits to those populations such as identifying genetic variants that predispose members to disease and that are less common in other populations.

In the case of larger social populations, identifying donors' racial or ethnic identities frequently involves a trade-off between gaining additional information that may be useful in designing genetic studies and interpreting their results (a process that may produce downstream benefits to members of donor populations) while exposing all members of those populations to greater risks of discrimination and stigmatization. The evaluation of this risk-benefit calculus can be complex and should include members of the populations in question (Sharp and Foster 2000). When faced with the concerns of members of populations that may be identified, many of which are informed by historical experiences of mistreatment and exploitation by outside researchers, the goal of minimizing the risks of social identifiability can become a central element in the scientific design of the resource or project, as in the case of the PDR (Lee et al. 2001a).

An argument can be made however, that racially and ethnically neutral genomic resources such as the PDR tend to preserve the illusion that genetic variation research can be performed in a social vacuum. Racial and ethnic identities often feature heavily in addressing health disparities, including access to care, choice of treatment, and differential health outcomes. Although those disparities are primarily social in origin, determining whether there are genetic predispositions to common diseases that are similar across racial and ethnic groups or that some social identities appear to be associated with higher frequencies of particular disease-related genetic structures can help place greater emphasis on the need to address the underlying social contributors that lead to health disparities. In addition, the process of selecting and involving members of different racial and ethnic groups in the formation of genomic resources and projects can cause researchers

to engage with issues such as race and social difference that inevitably will shape the applications of their studies outside the laboratory. In the final analysis, however, genomic researchers must justify both the importance of using social identities in the construction of genomic resources and their choices with regard to the release of those identities to users of genomic resources.

Risk and Reification

Among the most significant implications of population-specific genomic research is the potential for reifying racial and ethnic identities as biological constructs. Using broad, socially defined populations to structure participant recruitment for a genomic resource that retains social identifiers may foster an unintended genetic essentialism in the way the public understands race and ethnicity (Davis 2000). That essentialism could obscure the important fact that the "boundaries" between social groups are highly fluid and that most genetic variation exists *within* all social groups—not *between* them. Even the smallest socially defined population will have multiple haplotypes, alleles, polymorphisms, and other genetic characteristics that will be shared among different socially defined groups.

Nonetheless, public perceptions of genetic information tend to collapse categories and distinctions that scientists use to maintain distance between social and genetic definitions of populations (Hubbard and Wald 1993; Nelkin 2001). At the same time, genetic information tends to have greater public authority than social information (Nelkin and Lindee 1995). Hence, there is a likelihood that genetic information associated with race and ethnicity will result in the reductionary reconfiguration of those categories along simplified biological lines. Genetic linkage studies, for instance, emphasize one particular lineage through which donors are related to one another and ignore all the other ancestral lineages they may have that are unrelated to one another. Thus, a single genomic feature (such as a disease-related haplotype in linkage disequilibrium or a particular mitochondrial chromosome) mistakenly may be perceived by the lay public as defining a social population. In fact, donors for these kinds of feature-specific studies often are selected for the likelihood that they share the genetic feature in question (e.g., as evidenced by expression of a disease phenotype) rather than that they are representative of the social population as a whole.

Genetic interpretations of existing social distinctions and population histories are not without risk, having the potential for unintended uses in political and legal settings, as well as nationalist and racist reinterpretations by nonscientists. For example, the United States historically has used a limited set of phenotypic and biological characteristics as a statutory basis for determining which individuals are legally African American or Native American (Gossett 1963). Because of this historical practice, a haplotype map that associates racial, ethnic, and other socially constructed identities with specific ancestral haplotypes could potentially be used in legislation or lawsuits to determine which racial or ethnic identity an individual can or must use in certain circumstances. That propensity also may be found in forensic DNA profiling (Lowe et al. 2001). Indigenous and colonized populations are most vulnerable to these misuses, but minority populations elsewhere (e.g., those in Europe) also could be at risk.

Treating social populations as though they are essentially genetic in nature can lead to a number of harms. We already know that the association of racial, ethnic, or other social identities with genetic findings about disease susceptibility may be exclusionary and stigmatizing (Andrews et al. 1994). Other times, disease diagnoses may be missed or delayed when patients are not identified as being members of higher risk racial or ethnic populations, for example, in cases of sickle cell disease or cystic fibrosis (Spencer et al. 1994; Shafer et al. 1996). Being a member of what is publicly labeled as a higher risk population also can lead to a shared stigmatization (as happened historically in the case of sickle cell trait), even though one is neither a carrier nor an affected. In addition to how it might have an effect on existing social categories, perceptions of genetic difference could be used as the basis for the construction of new social categories, such as communities of persons who share genetically based disease susceptibilities or genetically based drug or environmental responses (or lack of response). These new social categories also could be subjected to discrimination and stigma. The experiences of patients diagnosed with leprosy and HIV are historical examples of social categories that arose from medical diagnoses (Rabinow 1996).

Researchers who assemble and use socially identified DNA sample collections have little control over how the lay public understands the relationship between race and ethnicity and the genetic findings associated with those identities "after" the composition of resources and findings based on them are made public. Clearly, researchers should attempt to correct inaccurate portrayals of their work or inappropriate depictions of the relationships between genetics and race and/or ethnicity. Once public mischaracterizations occur, however, it may be difficult to undo the damage caused. In contrast, "before" genomic resources or findings go public, researchers can take actions that would better frame subsequent public uses of race and ethnicity in relation to genetic features. In that regard, scientists can do a better job of explicitly noting the limitations of the genomic resources they assemble and the genetic studies they perform for biologically characterizing or defining social populations. Although researchers have tended to be relatively free in their uses of social identities in labeling samples and publishing population-specific findings, they have been remarkably restrained in explicitly noting the scientific limits within which those uses can or should be understood.

The initial social labeling of samples in commonly used genomic resources may be most crucial in this regard. Secondary users of those samples often are not in a position to be aware of the issues surrounding the choice of a particular social label for a given biological sample. More explicit framing of the scientific limitations of using race and ethnicity in labeling those samples can better shape lay perceptions of the relationship between social populations and biological features. For example, collections of DNA samples from members of identifiable communities (e.g., Yoruba donors) can be labeled more clearly as not necessarily being representative of the genomic structures of members of those communities (e.g., all Yorubas).

Providing this contextual background also should include discussions with members of the social populations from which samples are donated about how they identify themselves. The names that we use for other peoples often are more products of the ways in which European Americans historically have defined those others than of how they have defined themselves. Although both imposed and asserted identities are social constructions, population-specific genetic resources and findings may have implications for each. In

some cases, genetic findings may be taken by some as supporting or contesting external social classifications. However, genomic resources and publications that include clear concise accounts of how the social populations represented by donors have been defined historically by their members (as well as by outsiders) will make more evident the dynamic nature of social constructions of racial and ethnic identities, thereby reducing the likelihood that those classifications will be naively construed as markers of underlying genetic features.

Conclusion

The need for diversity in genomic resources is compelling, but the use of social classifications as a proxy for biological heterogeneity can be problematic. A biologically diverse resource will identify more genetic variants and increase the generalizability of findings. A resource that draws donors from multiple populations also recognizes that genetic variants may be more or less common among individuals with particular social identities. Without studying different socially defined populations, we cannot assume that all will benefit equally. Moreover, a basic genomic resource (such as a haplotype map) may provide important "spillover" benefit to members of sampled populations—if only because it increases the probability that the genetic structures of greatest relevance to their health will be found expeditiously. Correspondingly, the decision to "exclude" certain groups could have the effect of "depriving" them of these potential (albeit indirect) benefits. Thus, choices made today about the ways in which genomic research resources are established and how the social identities of sample donors are viewed by researchers will have significant implications.

Inclusion in a genomic resource is not just a question of potential benefits. It also is a question of perceived risks. This point is clear when one recalls the many ethical controversies surrounding the HGDP (National Research Council 1997). In addition to its problematic scientific assumption that many social populations are equivalent to biological demes, the project also floundered because of an inability to adequately address concerns from members of indigenous communities (Lone Dog 1999). These included concerns about the commodification of indigenous persons and their bodies, worries about potential uses that might conflict with cultural beliefs and practices, harmful legal and political implications resulting from genetic interpretations of population histories and memberships, and lack of provision for both short-term and long-term benefit for participating communities.

The implications of population-specific genomic research are not always self-evident, as both risk and benefit tend to be culturally defined (Douglas and Wildavsky 1982). In the cases of many nonmajority communities, researchers and institutional review boards will be unable to anticipate all culturally specific risks (e.g., those associated with the creation of immortalized cell lines and public databases; Foster and Sharp 2000). Outsiders often are unable to fully appreciate community-specific risks of these kinds even once they have been identified for them, in part because minority community members' perceptions of these risks may have been heightened by their historical experiences of being economically and politically disadvantaged with respect to the majority society (of which biomedical research is a tangible manifestation; Foster et al. 1999). The difference in power and privilege between researchers and socially defined populations lacking in significant economic and political resources may affect the ability of the latter to fully conceptualize and negotiate the conditions for research participation and to take effective action on any subsequent concerns about sample misuse or adverse interpretations of genetic findings. Moreover, differences in economic power may prevent donor populations from taking advantage of genetic tests and therapies developed from genomic research using members' samples. For these reasons, community involvement and consultation are essential in planning genomic resources that include and/or identify socially defined populations (Sharp and Foster 2000).

Categories such as race and ethnicity can be useful as heuristic starting-points for the investigation of genetic variation across populations. As such, social categories used to recruit participants should themselves be investigated in the course of assembling genomic resources, with the goals of discovering smaller population subgroups that are more predictive of biological relatedness and of better understanding the social factors that confound the relationship between socially defined populations and human genetic variation, factors that also will affect how genomic findings and technologies will be perceived and used by the general public. Only in that way will we be able to improve our understanding of genetic variation within social populations while simultaneously minimizing the risks of identifiability and reification and maximizing the potential benefits of genomics. In the end, genomic research with socially defined populations requires as much attention to the social organization of populations as it does to the genetic organization of chromosomes.

ACKNOWLEDGMENTS

This paper was made possible by grant No. ES11174 from the National Institute of Environmental Health Sciences (NIEHS) and the National Human Genome Research Institute (NHGRI) and by grant No. AI24717-S2 from the National Institute of Allergy and Infectious Diseases (NIAID). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS, the NHGRI, the NIAID, or the National Institutes of Health. We thank C. Aston, L. Brooks, A. Chakravarti, J. McEwen, A. Stone, and M. Yudell for comments on earlier versions of this paper.

REFERENCES

Andrews, L., et al. 1994. *Assessing genetic risks: Implications for health and social policy*. (eds. L.B. Andrews, J.E. Fullarton, N.A. Holtzman, and A.G. Motulsky) National Academy Press, Washington, DC.

Boas, F. 1942. *Race, language, and culture*. University of Chicago Press, Chicago.

Botkin, J.R. 2001. Protecting the privacy of family members in survey and pedigree research. *JAMA* **285:** 207–211.

Cavalli-Sforza, L.L., Wilson, A.C., Cantor, C.R., Cook-Deegan, R.M., and King, M.C. 1991. Call for a world-wide survey of human genetic diversity: A vanishing opportunity for the human genome project. *Genomics* **11:** 480–491.

Chakravarti, A. 1999. Population genetics: Making sense out of sequence. *Nat. Genet.* **21:** 56–60

Collins, F.S. and McKusick, V.A. 2001. Implications of the Human Genome Project for medical science. *JAMA* **285:** 540–544

Collins, F.S., Brooks, L.D., and Chakravarti, A. 1998. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8:** 1229–1231.

Cooper, R.S., Guo, X., Rotimi, C.N., Luke, A., Ward, R., Adeyemo, A., and Danilov, S.M. 2000. Heritability of angiotensin-converting enzyme and angiotensinogen: A comparison of US blacks and Nigerians. *Hypertension* **35:** 1141–1147.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., and Lander, E.S.

2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29:** 229–232.
Davidson, S. 2000. Research suggests importance of haplotypes over SNPs. *Nat. Biotech.* **18:** 1134–1135.
Davis, D. 2000. Groups, communities, and contested identities in genetic research. *Hastings Cent. Rpt.* **30(Nov/Dec):** 38–45.
Douglas, M. and Wildavsky, A. 1982. *Risk and culture: An essay on the relation of technical and environmental dangers*. University of California Press, Berkeley, CA.
Foster, M.W. and Sharp, R.R. 2000. Genetic research and culturally specific risks: One size does not fit all. *Trends Genet.* **16:** 93–95.
Foster, M.W., Sharp, RR., Freeman, W.L., Chino, M., Bernsten, D., and Carter, T.H. 1999. The role of community review in evaluating the risks of human genetic variation research. *Am. J. Hum. Genet.* **64:** 1719–1727.
Frank, L. 2000. Population genetics: Estonia prepares for national DNA database. *Science* **290:** 31.
Ghosh, S., Watanabe, R.M., Valle, T.T., Hauser, E.R., Magnuson, V.L., Langefeld, C.D., Ally, D.S., Mohlke, K.L., Silander, K., Kohtamaki, K., et al. 2000. The Finland–United States investigation of non–insulin-dependent diabetes mellitus genetics (FUSION) study, I: An autosomal genome scan for genes that predispose to type 2 diabetes. *Am. J. Hum. Genet.* **67:** 1174–1185.
Gilbert, W. 1992. A vision of the grail. In *The code of codes: Scientific and social issues in the human genome project* (eds. D.J. Kevles and L. Hood), pp. 83–97. Harvard University Press, Cambridge, MA.
Goldstein, D.B. and Weale, M.E. 2001. Population genomics: Linkage disequilibrium holds the key. *Curr. Bio.* **11:** R576–R579.
Gossett, T. 1963. *Race: The history of an idea in America*. Schocken Books, New York.
Greely, H.T. 2001. Human genome diversity: What about the other human genome project? *Nat. Rev. Genet.* **2:** 222–227.
Gulcher, J. and Stefannson, K. 1998. Population genomics: Laying the groundwork for genetic disease modeling and targeting. *Clin. Chem. Lab. Med.* **36:** 523–527.
Helmuth, L. 2001. Genome research: Map of the human genome 3.0. *Science* **293:** 583–585.
Horikawa, Y., Oda, N., Cox, N.J., Li, X., Orho-Melander, M., Hara, M., Hinokio, Y., Lindner, T.H., Mashima, H., Schwarz, P.E., et al. 2000. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat. Genet.* **26:** 163–175.
Hubbard, R. and Wald, E. 1993. *Exploding the gene myth*. Beacon Press, Boston.
Huxley, J. 1951. Genetics, evolution, and human destiny. In *Genetics in the twentieth century: Essays on the progress of genetics during its first 50 years* (ed. J.L. Dunn), pp. 591–621. Macmillan, New York.
Jorde, L.B. 2000. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10:** 1435–1444.
Juengst, E.T. 1998. Groups as gatekeepers to genomic research: Conceptually confusing, morally hazardous, and practically useless. *Kennedy Inst. Ethics J.* **8:** 183.
Kaufman, J.S. and Cooper, R.S. 2001. Commentary: Considerations for use of racial/ethnic classification in etiologic research. *Am. J. Epidemiol.* **154:** 291–298.
Kevles, D. 1995. *In the name of eugenics: Genetics and the uses of human heredity*. Harvard University Press, Cambridge, MA.
Kidd, J.R., Kidd, K.K., and Weiss, K.M. 1993. Human genome diversity initiative. *Hum. Bio.* **65:** 1–6.
Lee, S., Mountain, J., and Koenig, B.A. 2001. The meanings of "race" in the new genomics: Implications for health disparities research. *Yale J. Health Policy Law Ethics* **1:** 33–75.
Lone Dog, L. 1999. Whose genes are they? The Human Genome Diversity Project. *J. Health Soc. Policy* **10:** 51–66.
Lowe, A.L., Urquhart, A., Foreman, L.A., and Evett, I.W. 2001. Inferring ethnic origin by means of an STR profile. *Forensic Sci. Int.* **119:** 17–22.
Macilwain, C. 1996. Tribal groups attack ethics of genome diversity project. *Nature* **383:** 208.
Marks, J. 1995. *Human biodiversity: Genes, race, and history*. Aldine de Gruyter, New York.
Marshall, E. 1997. "Playing chicken" over gene markers. *Science* **278:** 2046.
National Research Council. 1997. *Evaluating human genetic diversity*. National Academy Press, Washington, DC.
Nebert, D.W. and Menon, A.G. 2001. Pharmacogenomics, ethnicity, and susceptibility genes. *Pharmacogenomics J.* **1:** 19–22.
Nelkin, D. 2001. Molecular metaphors: The gene in popular discourse. *Nat. Genet. Rev.* **2:** 555–559.
Nelkin, D. and Lindee, S.M. 1995. *The DNA mystique: The gene as a cultural icon*. WH Freeman, New York.
Olden, K. and Wilson, S. 2000. Environmental health and genomics: Visions and implications. *Nat. Genet. Rev.* **1:** 149–153.
Olden, K. and Guthrie, J. 2001. Genomics: Implications for toxicology. *Mutat. Res.* **473:** 3–10.
Oppenheimer, G.M. 2001. Paradigm lost: Race, ethnicity, and the search for a new population taxonomy. *Am. J. Pub. Health* **91:** 1049–1055.
Przeworski, M., Hudson, R.R., and DiRienzo, A. 2000. Adjusting the focus on human variation. *Trends Genet.* **16:** 296–302.
Rabinow, P. 1996. *Essays on the anthropology of reason*. Princeton University Press, Princeton, NJ.
Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumijian, R., Farhadian, S.F., et al. 2001. Linkage disequilibrium in the human genome. *Nature* **411:** 199–204.
Rothstein, M.A. and Epps, P.G. 2001. Pharmacogenomics and the (ir)relevance of race. *Pharmacogenomics J.* **1:** 104–108.
Shafer, F.E., Lorey, F., Cunningham,G.C., Klumpp, C., Vichinsky, E., and Lubin, B. 1996. Newborn screening for sickle cell disease: Four years of experience from California's newborn screening program. *J. Pediatr. Hematol. Oncol.* **18:** 36–41.
Sharp, R.R. and Foster, M.W. 2000. Involving study populations in the review of genetic research. *J. Law Med. Ethics* **28:** 41–51.
Snyder, L.H. 1951. Old and new pathways in human genetics. In *Genetics in the twentieth century: Essays on the progress of genetics during its first 50 years* (ed. L.C. Dunn), pp. 369–392. New Macmillan, New York.
Spencer, D.A., Venkatataman, M., Higgins, S., Stephenson, K., and Weller, P.H. 1994. Cystic fibrosis in children from ethnic minorities in the West Midlands. *Respir. Med.* **88:** 671–675.
Stephens, J.C., Schneider, J.A., Tanguay, D.A., Choi, J., Acharya, T., Stanley, S.E., Jiang, R., Messer, C.J., Chew, A., Han, J.H., et al. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293:** 489–493.
Tarazona-Santos, E., Carvalho-Silva, D.R., Pettener, D., Luiselli, D., De Stefano, G.F., Labarga, C.M., Rickards, O., Tyler-Smith, C., Pena, S.D., and Santos, F.R. 2001. Genetic differentiation in South Amerindians is related to environmental and cultural diversity: Evidence from the Y chromosome. *Am. J. Hum. Genet.* **68:** 1485–1496.
Vionnet, N., Hani, E.H., Dupont, S., Gallina, S., Francke, S., Dotte, S., De Matos, F., Durand, E., Lepretre, F., Lecoeur, C., et al. 2000. Genomewide search for type 2 diabetes–susceptibility genes in French whites: Evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. *Am. J. Hum. Genet.* **67:** 1470–1480.
Weber, W.W. 1999. Populations and genetic polymorphisms. *Mol. Diagn.* **4:** 299–307.
Weiss, K.M. and Clark, A.G. 2002. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18:** 19–24.
Wilson, J.F., Weale, M.E., Smith, A.C., Gratrix, F., Fletcher, B., Thomas, M.G., Bradman, N., and Goldstein, D.B. 2001. Population genetic structure of variable drug response. *Nat. Genet.* **29:** 265–269.
Wolford, J.K., Hanson, R.L., Kobes, S., Bogardus, C., and Prochazka, M. 2001. Analysis of linkage disequilibrium between polymorphisms in the KCNJ9 gene with type 2 diabetes mellitus in Pima indians. *Mol. Genet. Metab.* **73:** 97–103.

WEB SITE REFERENCES

http://www.nsf.gov/pubs/2001/nsf01120/nsf01120.html; Collection of human cell lines for the study of genetic diversity among culturally defined populations.
http://www.pharmgkb.org; Electronic database for clinical investigations of genetic influences on drug response.