Genome Biology IMPACT FACTOR 6.63

Opinion

Highly accessed

Categorization of humans in biomedical research: genes, race and disease

Neil Risch[1,2*], **Esteban Burchard**[3], **Elad Ziv**[3] and **Hua Tang**[4]

* Corresponding author: Neil Risch risch@lahmed.stanford.edu

Author Affiliations

1 Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305-5120, USA
2 Division of Research, Kaiser Permanente, Oakland, CA 94611-5714, USA
3 Department of Medicine, University of California, San Francisco, CA 94143, USA
4 Department of Statistics, Stanford University, Stanford, CA 94305, USA

For all author emails, please log on.

Genome Biology 2002, **3**:comment2007-comment2007.12 doi:10.1186/gb-2002-3-7-comment2007

The electronic version of this article is the complete one and can be found online at:
http://genomebiology.com/2002/3/7/comment/2007

Published: 1 July 2002

Abstract

A debate has arisen regarding the validity of racial/ethnic categories for biomedical and genetic research. Some claim 'no biological basis for race' while others advocate a 'race-neutral' approach, using genetic clustering rather than self-identified ethnicity for human genetic categorization. We provide an epidemiologic perspective on the issue of human categorization in biomedical and genetic research that strongly supports the continued use of self-identified race and ethnicity.

A major discussion has arisen recently regarding optimal strategies for categorizing humans, especially in the United States, for the purpose of biomedical research, both etiologic and pharmaceutical. Clearly it is important to know whether particular individuals within the population are more susceptible to particular diseases or most likely to benefit from certain therapeutic interventions. The focus of the dialogue has been the relative merit of the concept of 'race' or 'ethnicity', especially from the genetic perspective. For example, a recent editorial in the *New England Journal of Medicine* [1] claimed that "race is biologically meaningless" and warned that "instruction in medical genetics should emphasize the fallacy of race as a scientific concept and the dangers inherent in practicing race-based medicine." In support of this perspective, a recent article in *Nature Genetics* [2] purported to find that "commonly used ethnic labels are both insufficient and inaccurate representations of inferred genetic clusters." Furthermore, a supporting editorial in the same issue [3] concluded that "population clusters identified by genotype analysis seem to be more informative than those identified by skin color or self-declaration of 'race'." These conclusions seem consistent with the claim that "there is no biological basis for 'race'" [3] and that "the myth of major genetic differences across 'races' is nonetheless worth dismissing with genetic evidence" [4]. Of course, the use of the term "major" leaves the door open for possible differences but *a priori* limits any potential significance of such differences.

In our view, much of this discussion does not derive from an objective scientific perspective. This is understandable, given both historic and current inequities based on perceived racial or ethnic identities, both in the US and around the world, and the resulting sensitivities in such debates. Nonetheless, we demonstrate here that from both an objective and scientific (genetic and epidemiologic) perspective there is great validity in racial/ethnic self-categorizations, both from the research and public policy points of view.

Definition of risk factors: human categorization

The human population is not homogeneous in terms of risk of disease. Indeed, it is probably the case that every human being has a uniquely defined risk, based on his/her inherited (genetic) constitution plus non-genetic or environmental characteristics acquired during life. It is the goal of etiological epidemiological research to characterize such risks, both on an individual as well as population level, for the effective planning of prevention and/or treatment strategies. This public health perspective applies not only to disease, but to variation in normal traits (for example, for quantitative variables that are risk factors for disease such as blood pressure) as well as treatment response and adverse effects of pharmacologic agents.

The term 'risk factor' is widely used in epidemiology to define a characteristic associated either directly or indirectly with risk of disease. Some risk factors are fixed at birth (for example, sex or ethnicity) while others are acquired during life (for example, exposure to tobacco smoke or other environmental toxins). It is often assumed that risk factors fixed at birth are non-modifiable while those acquired after birth are modifiable and thus amenable to intervention strategies. But a

multifactorial model of risk requires the interplay of multiple inherited and non-inherited factors in producing a particular risk profile. Identification of inherited factors can both aid in the discovery of their environmental counterparts as well as provide a rational strategy for identifying, *a priori,* the most vulnerable members of our population, on whom prevention strategies can be focused. These concepts apply not only to disease prevention but to disease treatment as well, given that providing timely and efficacious treatment to individuals benefits both patients and health-care providers.

The ultimate goal of characterizing each individual's unique risk would require knowledge of every causal factor and the quantitative relationship of all possible combinations of such factors. In most cases, causal variables are not known, however, so epidemiologists resort to other means of categorizing people by the use of surrogate variables. Examples of such variables include gender, occupation, geographic location, socioeconomic status and dietary intake of a given food. It is understood that these variables do not themselves reflect a direct causal relationship with disease but rather are correlated with such a causal variable or variables.

Each of these classification systems masks within it inherent risk heterogeneity that could be further resolved if the specific agent(s) were identified. For example, geographic gradients in disease rates are well known, but it is not the geographic location, *per se,* that is causally related but rather some underlying correlated causal factor(s) such as temperature, humidity, rainfall, sunlight or presence of ground toxins. Even geographic categorizations, for example those based on latitude, mask heterogeneity within strata. Vancouver has a similar latitude to Winnipeg, but individuals born and raised in these two locations have very different climatologic experiences.

Although risk factor associations do not usually imply direct causal links, they do provide a starting point for further investigation. For example, there are sex differences in the rate of a variety of disorders. Sometimes these differences are related to endogenous (and hence non- or poorly-modifiable) differences between men and women (for example, the differential rates of breast cancer in men and women), while other examples are due to behavioral (presumably modifiable) differences (for example, differential lung cancer rates in men and women due to different smoking experiences). When direct causal factors are identified, risk estimates on both an individual and population basis can be made much more precise. Before such identification, however, the use of cruder surrogate factors can still provide valuable input for prevention and treatment decisions, even while acknowledging the latent heterogeneity within strata defined by such variables.

Rationale for the genetic categorization of humans

The discussion above provides an objective perspective from which to examine the question of genetic categorization of humans for biomedical research purposes. As for non-genetic risk factors, the ultimate goal of genetic research is to identify those specific genes and gene variants that influence the risk of disease, a quantitative outcome of interest, or response to a particular drug. Once all such genes are identified, every individual's unique risk can be assessed according to some quantitative model (if the number of genetic factors involved is large, however, problems will arise from the large number of possible combinations). In this case, categorization can occur at the level of the individual, irrespective of his/her racial, ethnic or geographic origins, moving us "closer to the ultimate goal of individualized therapy" [3]. But unless financial considerations in genetic testing become moot, there may still be practical issues concerning which genes to test in which individuals, and whether all genes should be tested in everyone.

To date, few genes underlying susceptibility to common diseases or influencing drug response have been identified. A question then arises, as it does when considering non-genetic factors, as to whether humans can or should be categorized genetically according to a surrogate scheme in the absence of known, specific gene effects. Wilson *et al.* [2] argue forcefully that genetic structure exists in the human population, that it can easily be identified even with a relatively modest number of marker loci, and that such structure is highly predictive of drug response: "We conclude that it is not only feasible but a clinical priority to assess genetic structure as a routine part of drug evaluation." [2]. This conclusion was based on their identification of four genetic "clusters" within a diverse sample of humans, and significant differences in the frequencies across these clusters for functional allelic variants in drug metabolizing enzymes.

Human evolution

Probably the best way to examine the issue of genetic subgrouping is through the lens of human evolution. If the human population mated at random, there would be no issue of genetic subgrouping because the chance of any individual carrying a specific gene variant would be evenly distributed around the world. For a variety of reasons, however, including geography, sociology and culture, humans have not and do not currently mate randomly, either on a global level or within countries such as the US. A clearer picture of human evolution has emerged from numerous studies over the past decade using a variety of genetic markers and involving indigenous populations from around the world. In summary, populations outside Africa derive from one or more migration events out of Africa within the last 100,000 years [5,6,7,8,9,10,11]. The greatest genetic variation occurs within Africans, with variation outside Africa representing either a subset of African diversity or newly arisen variants. Genetic differentiation between individuals depends on the degree and duration of separation of their ancestors. Geographic isolation and in-breeding (endogamy) due to social and/or cultural forces over extended time periods create and enhance genetic differentiation, while migration and inter-mating reduce it.

With this as background, it is not surprising that numerous human population genetic studies have come to the identical conclusion - that genetic differentiation is greatest when defined on a continental basis. The results are the same irrespective of the type of genetic markers employed, be they classical systems [5], restriction fragment length polymorphisms (RFLPs) [6], microsatellites [7,8,9,10,11], or single nucleotide polymorphisms (SNPs) [12]. For example, studying 14 indigenous populations from 5 continents with 30 microsatellite loci, Bowcock *et al.* [7] observed that the 14 populations clustered into the five continental groups, as depicted in Figure 1. The African branch included three sub-Saharan populations, CAR pygmies, Zaire pygmies, and the Lisongo; the Caucasian branch included Northern Europeans and Northern Italians; the Pacific Islander branch included Melanesians, New Guineans and Australians; the East Asian branch included Chinese, Japanese and Cambodians; and the Native American branch included Mayans from Mexico and the Surui and Karitiana from the Amazon basin. The identical diagram has since been derived by others, using a similar or greater number of microsatellite markers and individuals [8,9]. More recently, a survey of 3,899 SNPs in 313 genes based on US populations (Caucasians, African-Americans, Asians and Hispanics) once again provided distinct and non-overlapping clustering of the Caucasian, African-American and Asian samples [12]: "The results confirmed the integrity of the self-described ancestry of these individuals". Hispanics, who represent a recently admixed group between Native American, Caucasian and African, did not form a distinct subgroup, but clustered variously with the other groups. A previous cluster analysis based on a much smaller number of SNPs led to a similar conclusion: "A tree relating 144 individuals from 12 human groups of Africa, Asia, Europe and Oceania, inferred from an average of 75 DNA polymorphisms/individual, is

remarkable in that most individuals cluster with other members of their regional group" [13]. Effectively, these population genetic studies have recapitulated the classical definition of races based on continental ancestry - namely African, Caucasian (Europe and Middle East), Asian, Pacific Islander (for example, Australian, New Guinean and Melanesian), and Native American.

Figure 1. The evolutionary tree of human races. Population genetic studies of world populations support the categorization into five major groups, as shown. See text for further details.

The terms race, ethnicity and ancestry are often used interchangeably, but some have also drawn distinctions. For the purpose of this article, we define racial groups on the basis of the primary continent of origin, as discussed above (with some modifications described below). Ethnicity is a self-defined construct that may be based on geographic, social, cultural and religious grounds. It has potential meaning from the genetic perspective, provided it defines an endogamous group that can be differentiated from other such groups. Ancestry refers to the race/ethnicity of an individual's ancestors, whatever the individual's current affiliation. From the genetic perspective, the important concept is mating patterns, and the degree to which racially or ethnically defined groups remain endogamous.

The continental definitions of race and ancestry need some modification, because it is clear that migrations have blurred the strict continental boundaries. For example, individuals currently living in South Africa, although currently Africans, have very different ancestry, race and ethnicity depending on the ancestry of their forbears (for example from Europe or Asia) and the degree to which they have remained endogamous. For our purposes here, on the basis of numerous population genetic surveys, we categorize Africans as those with primary ancestry in sub-Saharan Africa; this group includes African Americans and Afro-Caribbeans. Caucasians include those with ancestry in Europe and West Asia, including the Indian subcontinent and Middle East; North Africans typically also are included in this group as their ancestry derives largely from the Middle East rather than sub-Saharan Africa. 'Asians' are those from eastern Asia including China, Indochina, Japan, the Philippines and Siberia. By contrast, Pacific Islanders are those with indigenous ancestry from Australia, Papua New Guinea, Melanesia and Micronesia, as well as other Pacific Island groups further east. Native Americans are those that have indigenous ancestry in North and South America. Populations that exist at the boundaries of these continental divisions are sometimes the most difficult to categorize simply. For example, east African groups, such as Ethiopians and Somalis, have great genetic resemblance to Caucasians and are clearly intermediate between sub-Saharan Africans and Caucasians [5]. The existence of such intermediate groups should not, however, overshadow the fact that the greatest genetic structure that exists in the human population occurs at the racial level.

Most recently, Wilson *et al.* [2] studied 354 individuals from 8 populations deriving from Africa (Bantus, Afro-Caribbeans and Ethiopians), Europe/Mideast (Norwegians, Ashkenazi Jews and Armenians), Asia (Chinese) and Pacific Islands (Papua New Guineans). Their study was based on cluster analysis using 39 microsatellite loci. Consistent with previous studies, they obtained evidence of four clusters representing the major continental (racial) divisions described above as African, Caucasian, Asian, and Pacific Islander. The one population in their analysis that was seemingly not clearly classified on continental grounds was the Ethiopians, who clustered more into the Caucasian group. But it is known that African populations with close contact with Middle East populations, including Ethiopians and North Africans, have had significant admixture from Middle Eastern (Caucasian) groups, and are thus more closely related to Caucasians [14]. Furthermore, the analysis by Wilson *et al.* [2] did not detect subgroups within the four major racial clusters (for example, it did not separate the Norwegians, Ashkenazi Jews and Armenians among the Caucasian cluster), despite known genetic differences among them. The reason is clearly that these differences are not as great as those between races and are insufficient, with the amount of data provided, to distinguish these subgroups.

Are racial differences merely cosmetic?

Two arguments against racial categorization as defined above are firstly that race has no biological basis [1,3], and secondly that there are racial differences but they are merely cosmetic, reflecting superficial characteristics such as skin color and facial features that involve a very small number of genetic loci that were selected historically; these superficial differences do not reflect any additional genetic distinctiveness [2]. A response to the first of these points depends on the definition of 'biological'. If biological is defined as genetic then, as detailed above, a decade or more of population genetics research has documented genetic, and therefore biological, differentiation among the races. This conclusion was most recently reinforced by the analysis of Wilson *et al.* [2]. If biological is defined by susceptibility to, and natural history of, a chronic disease, then again numerous studies over past decades have documented biological differences among the races. In this context, it is difficult to imagine that such differences are not meaningful. Indeed, it is difficult to conceive of a definition of 'biological' that does not lead to racial differentiation, except perhaps one as extreme as speciation.

A forceful presentation of the second point - that racial differences are merely cosmetic - was given recently in an editorial in the *New England Journal of Medicine* [1]: "Such research mistakenly assumes an inherent biological difference between black-skinned and white-skinned people. It falls into error by attributing a complex physiological or clinical phenomenon to arbitrary aspects of external appearance. It is implausible that the few genes that account for such outward characteristics could be meaningfully linked to multigenic diseases such as diabetes mellitus or to the intricacies of the therapeutic effect of a drug." The logical flaw in this argument is the assumption that the blacks and whites in the referenced study differ only in skin pigment. Racial categorizations have never been based on skin pigment, but on indigenous continent of origin. For example, none of the population genetic studies cited above, including the study of Wilson *et al.* [2], used skin pigment of the study subjects, or genetic loci related to skin pigment, as predictive variables. Yet the various racial groups were easily distinguishable on the basis of even a modest number of random genetic markers; furthermore, categorization is extremely resistant to variation according to the type of markers used (for example, RFLPs, microsatellites or SNPs).

Genetic differentiation among the races has also led to some variation in pigmentation across races, but considerable variation within races remains, and there is substantial overlap for this feature. For example, it would be difficult to distinguish most Caucasians and Asians on the basis of skin pigment alone, yet they are easily distinguished by genetic markers. The author of the above statement [1] is in error to assume that the only genetic differences between races, which may differ on average in pigmentation, are for the genes that determine pigmentation.

Common versus rare alleles

Despite the evidence for genetic differentiation among the five major races, as defined above, numerous studies have shown that local populations retain a great deal of genetic variation. Analysis of variance has led to estimates of 10% for the proportion of variance due to average differences between races, and 75% of the variance due to genetic variation within populations. Comparable estimates have been obtained for classical blood markers [15,16], microsatellites [17], and SNPs [12]. Unfortunately, these analysis of variance estimates have also led to misunderstandings or misinterpretations. Because of the large amount of variation observed within races versus between races, some commentators have denied genetic differentiation between the races; for example, "Genetic data ... show that any two individuals within a particular population are as different genetically as any two people selected from any two populations in the world." [18]. This assertion is both counter-intuitive and factually incorrect [12,13]. If it were true, it would be impossible to create discrete clusters of humans (that end up corresponding to the major races), for example as was done by Wilson *et al.* [2], with even as few as 20 randomly chosen genetic markers. Two Caucasians are more similar to each other genetically than a Caucasian and an Asian.

In these variance assessments, it is also important to consider the frequency of the allelic variants examined. These studies are based primarily on common alleles, and may not reflect the degree of differentiation between races for rare alleles. This is an important concern because alleles underlying disease susceptibility, especially deleterious diseases, may be less frequent than randomly selected alleles. Similarly, it has also been shown that among different classes of SNPs, those that lead to non-conservative amino-acid substitutions (which most frequently are associated with clinical outcomes) occur least often, and when they do occur they tend to have lower allele frequencies than non-coding or synonymous coding changes [12,19,20].

It is likely that genetic differentiation among races is enhanced for disease-predisposing alleles because such alleles tend to be in the lower frequency range. It is well known that rarer alleles are subject to greater fluctuation in frequency due to genetic drift than common alleles. Indeed, for most Mendelian diseases, even higher frequency alleles are found only in specific races (for example, cystic fibrosis and hemochromatosis in Caucasians). Furthermore, recent SNP surveys of the different races have shown that lower frequency variants are much more likely to be specific to a single race or shared by only two races than are common variants [12,19,20]. In one study, only 21% of 3,899 SNPs were found to be pan-ethnic, and some race-specific SNPs were found to have a frequency greater than 25% [12].

Admixture and genetic categorization in the United States

Most population genetic studies that focus on human evolution and the relatedness of people, including the ones cited above, utilize indigenous groups from the various continents. These groups would not necessarily adequately depict the US population, for example, where admixture between races has occurred over many centuries. Nonetheless, during the same period of time, as well as currently, mating patterns are far from random. The tendency toward endogamy is reflected within the 2000 US Census [21], which allowed individuals to report themselves to be of a single race or of mixed race. Six racial categories were provided (White; Black or African American; American Indian and Alaska Native; Asian; Native Hawaiian and other Pacific Islander; Some other race). In response to this question, 97.6% of subjects reported themselves to be of one race, while 2.4% reported themselves to be of more than one race; 75% reported themselves as White, 12.3% as Black or African American, 3.6% as Asian, 1% as American Indian or Alaska Native, 0.1% as Hawaiian or Pacific Islander; and 5.5% of other race. Of the 5.5% who reported themselves as 'other race', most (97%) also reported themselves to be Hispanic. According to these numbers, if mating were at random with respect to these racial categories, 42% of individuals would result from 'mixed' matings and hence derive from more than one race, as opposed to the 2.4% reported. These figures highlight the strong deviation from random mating in the US.

What are the implications of these census results and the admixture that has occurred in the US population for genetic categorization in biomedical research studies in the US? Gene flow from non-Caucasians into the US Caucasian population has been modest. On the other hand, gene flow from Caucasians into African Americans has been greater; several studies have estimated the proportion of Caucasian admixture in African Americans to be approximately 17%, ranging regionally from about 12% to 23% [22]. Thus, despite the admixture, African Americans remain a largely African group, reflecting primarily their African origins from a genetic perspective. Asians and Pacific Islanders have been less influenced by admixture and again closely represent their indigenous origins. The same is true for Native Americans, although some degree of Caucasian admixture has occurred in this group as well [23].

The most complex group is made up of those who self-identify as Hispanic/Latino. The US Census did not consider this group as a separate race, although 42% of respondents who considered themselves Hispanic checked the category 'other race' for the racial question, while 48% checked 'White'. Hispanics are typically a mix of Native American, Caucasian and African/African American, with the relative proportions varying regionally. Southwest Hispanics, who are primarily Mexican-American, appear to be largely Caucasian and Native American; recent admixture estimates are 39% Native American, 58% Caucasian and 3% African [24]. By contrast, East Coast Hispanics are largely Caribbean in origin, and have a greater proportion African admixture [25]. Thus, depending on geography, self-identified Hispanics could aggregate genetically with Caucasians, Native Americans, African Americans or form their own cluster.

The persistence of genetic differentiation among these US racial groups (as defined by the US Census) has also been verified recently in a study of nearly 4,000 SNPs in 313 genes [12]. These authors found distinct clusters for Caucasian Americans, African Americans and Asian Americans; the Hispanic Americans did not form a separate cluster but were either grouped with Caucasians or not easily classified. Although the US Census results suggest the large majority of individuals can be categorized into a single ancestral group, there remain individuals of mixed ancestry who will not be easily categorized by any simple system of finite, discrete categories. On the other hand, such individuals can be particularly informative in epidemiologic studies focused on differentiating genetic versus environmental sources for racial/ethnic difference, as we describe further below.

Genetic clustering versus self-reported ancestry

A major conclusion from the study of Wilson *et al.* [2], reiterated in accompanying editorials, is that "Clusters identified by genotyping... are far more robust than those identified using geographic and ethnic labels" [26]. But closer examination of the study and other data actually leads to the opposite conclusion: namely, that self-defined race, ethnicity or ancestry are actually more genetically informative than clusters based on analysis of random genetic markers.

In their analysis, Wilson *et al.* [2] found greater variation in allele frequencies for drug metabolizing enzymes based on four "genetically" defined clusters than in three "ethnically" defined clusters. The ethnic clusters included Caucasians (Norwegians, Ashkenazi Jews and Armenians), Africans (Bantus, Afro-Caribbeans and Ethiopians), and Asians (Chinese and

Papua New Guineans). The inclusion of Papua New Guineans as Asians would be considered highly controversial by most population geneticists, as all prior studies of this group show them to cluster with Pacific Islanders [7,8], and as we discussed above, population genetic studies have shown Pacific Islanders to be distinct from Asians [6,7,8,9]. Futhermore, the racial categories of the US census would also not merge Chinese with New Guineans. Nonetheless, examination of the variance in allele frequencies for the six drug-metabolizing enzymes across the four "genetically defined" clusters versus the three "racial" groups does not reveal greater differentiation of the former (Table 1). In fact, for 5 of the 6 loci, the variance is greater among the three "racial" groups, although for most loci the variance is very similar for the two categorization schemes. This is not surprising, as the racial categories aligned nearly perfectly with genotype clusters. The only exception was for the Ethiopians, who (as we discussed above) are known to genetically resemble Caucasians, probably as a result of considerable Caucasian admixture [14]. On the other hand, neither of these ethnic categorizations (of New Guineans and Ethiopians) would have much impact on studies in the US, as these groups represent only a tiny fraction of the US population, even among Pacific Islanders and African Americans, respectively.

Table 1. Allele frequency differentiation of drug metabolizing enzymes on the basis of "genetic clusters" versus "racial groups," from the data of Wilson *et al.* [2]

Indeed, in another respect, the results of Wilson *et al.* [2] demonstrate the superiority of ethnic labels over genetic clustering. Consider the group they labeled Caucasian, consisting of Norwegians, Ashkenazi Jews and Armenians. Their genetic cluster analysis lumped these three populations together into a single (Caucasian) cluster. Yet numerous genetic studies of these groups have shown them to differ in allele frequencies for a variety of loci. For example, the hemochromatosis gene mutation C282Y has a frequency of less than 1% in Armenians and Ashkenazi Jews but of 8% in Norwegians [27]. Thus, in this case, self-defined ethnicity provides greater discriminatory power than the single genotype cluster obtained by Wilson *et al.* [2].

This conclusion obtains not just for the ethnically homogeneous but for admixed subjects as well. A study by Williams *et al.* [28] considered the degree of Caucasian admixture in a Pima Indian population. These authors compared self-report (the number of Caucasian and Pima grandparents) with a genetic estimate of admixture based on 18 conventional blood markers. Using type 2 diabetes mellitus (high frequency in Pimas, low frequency in Caucasians) as the outcome, disease frequency correlated more strongly with self-reported admixture than with genetically estimated admixture.

As seen in the study of Wilson *et al.* [2], genetic cluster analysis is only powerful in separating out individuals whose ancestors diverged many millennia ago, leading to substantial genetic differences. It is much less capable of differentiating more recently separated groups, whose genetic differentiation is smaller. This conclusion may also be a result of the small number of genetic markers employed, however, and finer resolution might be possible with a much larger number of loci.

How many loci are needed for clustering?

A natural question arises as to the number of loci required to categorize individuals into ancestrally defined clusters. The answer depends on the degree of genetic differentiation of the populations in question. Two groups with ancient separation and no migration will require far fewer markers than groups that have separated more recently or have been influenced by recent migrations and/or admixture.

A simple quantification of this question is possible, as described in Box 1. The number of biallelic loci necessary for given misclassification rates is given in Table 2; when δ_{av} is high, 20 or fewer loci are adequate for accurate classification, but even for low δ_{av} values 200 markers are more than adequate. How do these numbers relate to the ability to differentiate genetically various racial/ethnic groups? Recent large-scale surveys of 257 SNPs [29] and 744 short tandem repeat polymorphisms (STRPs, or microsatellites) [30] provide an answer to this question, based on the distribution of δ values observed in these surveys for various population comparisons. Both studies included Caucasian Americans (CA), Asian Americans (AS) and African Americans (AA); Dean *et al.* [29] also included Native Americans (NA), while Smith *et al.* [30] included Hispanic Americans (HA). Table 3 provides the median δ values for all markers, for the top 50th percentile of δ values, and for the top 20th percentile of δ values.

Table 2. The number of markers required for clustering as a function of the misclassification rate (calculated as shown in Box 1)

Table 3. Median δ values for different racial/ethnic group comparisons, from data of Dean *et al.* [29] and Smith *et al.* [30]

In conjunction with Table 2, we can estimate that about 120 unselected SNPs or 20 highly selected SNPs can distinguish group CA from NA, AA from AS and AA from NA. A few hundred random SNPs are required to separate CA from AA, CA from AS and AS from NA, or about 40 highly selected loci. STRP loci are more powerful and have higher effective δ values because they have multiple alleles. Table 3 reveals that fewer than 100 random STRPs, or about 30 highly selected loci, can distinguish the major racial groups. As expected, differentiating Caucasians and Hispanic Americans, who are admixed but mostly of Caucasian ancestry, is more difficult and requires a few hundred random STRPs or about 50 highly selected loci. These results also indicate that many hundreds of markers or more would be required to accurately differentiate more closely related groups, for example populations within the same racial category.

Gene-environment correlation and confounding - the real problem

From an epidemiologic perspective, the use of 'genetic' clusters, as suggested by Wilson *et al.* [2], instead of self-reported ethnicity will not alleviate but rather will actually create and/or exacerbate problems associated with genetic inferences based on racial differences. The true complication is due to the fact that racial and ethnic groups differ from each other on a variety of social, cultural, behavioral and environmental variables as well as gene frequencies, leading to confounding between genetic and environmental risk factors in an ethnically heterogeneous study. For example, with respect to treatment response, "An individual's response to a drug depends on a host of factors, including overall health, lifestyle, support system, education and socioeconomic status - all of which are difficult to control for and likely to be affected, at least in the United States, by a person's 'race'" [3].

Specifically, let us consider the practical implications of the "race-neutral approach" [3] advocated by Wilson *et al.* [2]. As

an example, we revisit a recent study of the efficacy of inhibitors of angiotensin-converting enzyme (ACE) in 1,200 white versus 800 black patients with congestive heart failure [31] that generated a great deal of controversy [1,32]. In that study, the authors showed that black patients on the ACE inhibitor Enalapril showed no reduction in hospitalization compared with those on placebo, whereas white patients showed a strong, statistically significant difference between treatment versus placebo arms. Let us suppose that instead of using racial labels, the authors had performed genotype cluster analysis on their combined sample. They would have obtained two clusters - cluster A containing approximately 1,200 subjects, and cluster B, containing approximately 800 subjects. They would then demonstrate that cluster A treated subjects show a dramatic response to Enalapril compared to placebo subjects, while cluster B subjects show no such response. The direct inference from this analysis would be that the difference in responsiveness between individuals in cluster A and cluster B is genetic - that is, due to a frequency difference in one or more alleles between the two groups. But the problem should be obvious: cluster A is composed of the Caucasian subjects and cluster B the African Americans. Although a genetic difference in treatment responsiveness between these two groups is inferred, the conclusion is completely confounded with the myriad other ways these two groups might differ from each other; hence the culprit may not be genetic at all.

A racial difference in the frequency of some phenotype of interest (disease, or drug response) or quantitative trait is but a first clue in the search for etiologic causal factors. As we have illustrated, without such racial/ethnic labels, these underlying factors cannot be adequately investigated. Although some investigators might quickly jump to a genetic explanation for an ethnic difference, this is rarely the case with epidemiologists, who have a broad view of the complex nature of most human traits [33]. Indeed, epidemiologists employ several different approaches to disentangling genetic from environmental causes of ethnic differences, including migrant studies and stratified analyses.

The rationale underlying migrant studies is to compare the frequency of a trait (such as disease) between members of the same ethnic group (who are assumed to be genetically homogeneous) who are residing in different environments. For example, breast cancer rates in Asian (Chinese and Japanese) women are vastly lower than the rates among US Caucasians. However, the breast cancer rates of Chinese and Japanese women living in Hawaii and the San Francisco Bay Area are comparable to those of US Caucasians [34]. These results suggest an environmental source of the racial difference. Asians are also known to have much lower rates of multiple sclerosis than European Caucasians [35]. But within a single country, namely Canada, this racial difference persists [36], increasing support for (but not proving) a genetic explanation.

The best approach to resolve confounding is through matched, adjusted or stratified analyses, but this depends on having the confounding variables (or their surrogates). Such analyses can be performed in a racially heterogeneous sample, but it is potentially more powerful when performed within a single group. The reason is that the correlation between confounding variables (such as genes and environment) may be stronger in a heterogeneous study population than in a more homogeneous one. The ability to disentangle the effects of confounding variables is greater when their sample correlation is low.

A simple example is provided in Figure 2. Here we assume two populations (for example, races), groups A and B. As shown in the figure, where both environmental and genetic factors differ between the populations, it is impossible to determine which is the functional cause of the racial difference if the genetic and environmental effects are completely correlated in frequency within the two groups. More importantly, if the relative frequency in the two groups of the environmental factor was not measured, analysis stratified on the genetic differences yields the correct interpretation that the genetic difference does not contribute to the racial difference if the genetic and environmental factors are not correlated. But if they are fully correlated, analysis stratified on the genetic factor alone would lead to the incorrect conclusion that it is the cause of the racial difference.

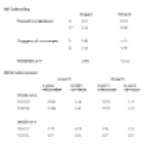

Figure 2. An example of confounding and a stratified analysis of environmental and genetic factors. Here we assume two populations (for example, races), groups A and B. G1 and G2 represent dichotomous genotype classes at a candidate gene locus (here one of the classes represents two genotypes for simplification, as would be the case for a dominant model), and E1 and E2 represent two strata of an environmental factor. **(a)** We assume that the probability (P) of trait D depends only on E, so that the risk of D given E1 is 10%, versus 1% given E2. In group A, the frequency of G1, G2, E1 and E2 are each 50%, whereas in group B, the frequency of G1 and E1 are each 10% and the frequency of G2 and E2 are each 90% Then, within group A, the prevalence of D is 5.5% whereas in group B the prevalence is 1.9%; hence, a racial difference exists in the prevalence of D. **(b)** We next consider the prevalence of D within strata defined by G and E. First, we assume G and E are frequency-independent within each group. In this case, the frequency difference in D between groups A and B persists within strata defined by G, but not within strata defined by E. Thus, the environmental factor E can completely explain the racial difference between groups A and B, but the genetic factor does not. Next consider the case where G and E are completely correlated in frequency within groups. In this case, analysis stratified on G or E eliminates the prevalence difference between groups A and B, and it is impossible to determine which is the functional cause of the racial difference. More important, consider the situation where factor E was not measured. Then for the first scenario (G and E independent within group), analysis stratified on G yields the correct interpretation that G does not contribute to the racial difference; for the second scenario (G and E fully correlated), however, analysis stratified on G would lead to the incorrect conclusion that G is the cause of the racial difference. P(D|G1) denotes the probability of disease given an individual has genotype G1, and similarly for G2, E1 and E2.

Box 1.

Epidemiologists often perform analyses of racial differences stratified on numerous environmental variables, such as socio-economic status, access to health care, education, and so on. The persistence of racial differences after accounting for these covariates raises the index of suspicion that genetic differences may be involved. For example, Karter *et al.* [37] recently demonstrated persistence of racial differences in diabetes complications in a health maintenance organization after

controlling for numerous potential confounders including measures of socioeconomic status, education, and health-care access and utilization. Such evidence is indirect, however, as other unmeasured factors may still be responsible [38]. Ultimate proof depends on identifying a specific gene effect within each population, with an allele frequency difference between populations. One such example involves the lower risk of type 1 diabetes in US Hispanic versus Caucasian children. The HLA allele DR3 is predisposing to type 1 diabetes in both populations but has a lower frequency in Hispanics than Caucasians [39].

Another approach often taken by genetic epidemiologists is to consider the prevalence of disease or drug response (D) in individuals who are admixed between groups A and B - for example, in individuals who are 100%A, 75%A-25%B, 50%A-50%B, 25%A-75%B and 100%B (corresponding to 4, 3, 2, 1 and 0 grandparents that are group A, respectively). A continuous cline in the frequency of D with genome proportion that is group A is taken as suggestive evidence of genetic factors explaining the prevalence difference between groups A and B. An example of this type of analysis is the decreasing trend in type 2 diabetes in Pima Indians with degree of Caucasian admixture [23]. Analyses stratified on environmental factors can again strengthen the argument. But the same caveat applies here as described above. If an unmeasured environmental variable (such as socio-economic status) covaries in the same fashion as the proportion group A, the racial difference could be due to this unmeasured variable. At best, one could argue that the racial difference is not explained by any of the measured covariates.

Identical treatment is not equal treatment

Both for genetic and non-genetic reasons, we believe that racial and ethnic groups should not be assumed to be equivalent, either in terms of disease risk or drug response. A 'race-neutral' or 'color-blind' approach to biomedical research is neither equitable nor advantageous, and would not lead to a reduction of disparities in disease risk or treatment efficacy between groups. Whether African Americans, Hispanics, Native Americans, Pacific Islanders or Asians respond equally to a particular drug is an empirical question that can only be addressed by studying these groups individually. Differences in treatment response or disease prevalence between racial/ethnic groups need to be studied carefully; naive inferences about genetic causation without evidence should be avoided. At the same time, gratuitous dismissal of a genetic interpretation without evidence for doing so is also unjustified.

We strongly support the search for candidate genes that contribute both to disease susceptibility and treatment response, both within and across racial/ethnic groups. Identification of such genes can help provide more precise individualized risk estimates. Environmental variables that influence risk and interact with genetic variables also require identification. Only if consideration of all these variables leaves no residual difference in risk between racial/ethnic groups is it justified to ignore race and ethnicity.

Is there any advantage to using random genetic markers and genetically defined clusters over self-reported ancestry in attempting to assess risk? In the case where a sample was collected without ancestry information, or for individuals on whom ancestral background is missing - for example, adoptees - such data could substitute for self-identified ancestry. On the other hand, we see considerable disadvantage in avoiding self-reported ancestry in favor of a 'color-blind' approach of genetically defined clusters. Any study in the US that randomly samples subjects without regard to ancestry will obtain, on average, 75% Caucasians, 12% African Americans, 4% Asians, 12% Hispanics (broadly defined) and few Pacific Islanders, although these frequencies would vary regionally. Thus, results from such studies would be largely derived from the Caucasian majority, with obtained parameter estimates that might not apply to the groups with minority representation. It might be possible in such studies to subsequently identify racial/ethnic differences, either based on self-reported ancestry of the study subjects or by genetic cluster analysis. However, the low frequency of the non-Caucasian groups in the total sample would lead to reduced power to detect and investigate any racial/ethnic differences that might be present. The best way to avoid this power issue is to specifically over-sample the lower frequency racial/ethnic groups to obtain larger sample sizes. Obviously, the only way to effectively and economically do so is based on self-identified ancestry, as genotype information from the population at large is not available and not likely to become so. Fortunately, the National Institutes of Health have instituted policies to encourage the inclusion of US ethnic minorities, and we strongly support continuation and expansion of this policy.

Finally, we believe that identifying genetic differences between races and ethnic groups, be they for random genetic markers, genes that lead to disease susceptibility or variation in drug response, is scientifically appropriate. What is not scientific is a value system attached to any such findings. Great abuse has occurred in the past with notions of 'genetic superiority' of one particular group over another. The notion of superiority is not scientific, only political, and can only be used for political purposes.

As we enter this new millennium with an advancing arsenal of molecular genetic tools and strategies, the view of genes as immutable is too simplistic. Every race and even ethnic group within the races has its own collection of clinical priorities based on differing prevalence of diseases. It is a reflection of the diversity of our species - genetic, cultural and sociological. Taking advantage of this diversity in the scientific study of disease to gain understanding helps all of those afflicted. We need to value our diversity rather than fear it. Ignoring our differences, even if with the best of intentions, will ultimately lead to the disservice of those who are in the minority.

Acknowledgements

We are grateful to Andrew Karter and Catherine Schaefer for many helpful comments and discussion on an earlier version of this manuscript. N.R. was supported by NIH grant GM057672, E.B. and E.Z. by the Sandler Family Supporting Foundation, and H.T. by a Howard Hughes Medical Institute fellowship.

References

Schwartz RS: **Racial profiling in medical research.**
N Engl J Med 2001, **344:**1392-1393. PubMed Abstract | Publisher Full Text

Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MF, Bradman N, Goldstein DB: **Population genetic structure of variable drug response.**
Nat Genet 2001, **29:**265-269. PubMed Abstract | Publisher Full Text

Editorial: **Genes, drugs and race.**
Nat Genet 2001, **29:**239-240. PubMed Abstract | Publisher Full Text

Owens K, King M-C: **Genomic views of human history.**
Science 1999, **286:**451-453. PubMed Abstract | Publisher Full Text

Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J: **Reconstruction of human evolution; bringing together genetic, archaeological, and linguistic data.**
Proc Natl Acad Sci USA 1988, **85:**6002-6006. PubMed Abstract

Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, Kidd KK, Cavalli-Sforza LL: **Drift, admixture, and selection in human evolution: a study with DNA polymorphisms.**
Proc Natl Acad Sci USA 1991, **88:**839-843. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL: **High resolution of human evolutionary trees with polymorphic microsatellites.**
Nature 1994, **368:**455-457. PubMed Abstract | Publisher Full Text

Perez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R, Bertranpetit J: **Microsatellite variation with the differentiation of modern humans.**
Hum Genet 1997, **99:**1-7. PubMed Abstract | Publisher Full Text

Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK: **Short tandem repeat polymorphism evolution in humans.**
Eur J Hum Genet 1998, **6:**38-49. PubMed Abstract | Publisher Full Text

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M, *et al*.: **Global patterns of linkage disequilibrium at the CD4 locus and modern human origins.**
Science 1996, **271:**1380-1387. PubMed Abstract

Jorde LB, Rogers AR, Bamshad M, Watkins WS, Krakowiak P, Sung S, Kere J, Harpending HC: **Microsatellite diversity and the demographic history of modern humans.**
Proc Natl Acad Sci USA 1997, **94:**3100-3103. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, *et al*.: **Haplotype variation and linkage disequilibrium in 313 human genes.**
Science 2001, **293:**489-493. PubMed Abstract | Publisher Full Text

Mountain JL, Cavalli-Sforza LL: **Multilocus genotypes, a tree of individuals, and human evolutionary history.**
Am J Hum Genet 1997, **61:**705-718. PubMed Abstract

Cavalli-Sforza LL, Menozzi P, Piazza A:
The History and Geography of Human Genes. Princeton University Press: Princeton New Jersey;. 1994.

Lewontin RC: **The apportionment of human diversity.**
Evol Biol 1972, **6:**381-398.

Latter BDH: **Genetic differences within and between populations of the major human subgroups.**
Amer Nat 1980, **116:**220-237. Publisher Full Text

Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL: **An apportionment of human DNA diversity.**
Proc Natl Acad Sci USA 1997, **94:**4516-4519. PubMed Abstract | Publisher Full Text | PubMed Central Full Text

Editorial: **Census, race and science.**
Nat Genet 2000, **24:**97-98. PubMed Abstract | Publisher Full Text

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, *et al*.: **Characterization of single-nucleotide polymorphisms in coding regions of human genes.**
Nat Genet 1999, **22:**231-238. PubMed Abstract | Publisher Full Text

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A: **Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis.**
Nat Genet 1999, **22:**239-247. PubMed Abstract | Publisher Full Text

[http://www.census.gov/population/www/cen2000/briefs.html] *webcite*
Overview of Race and Hispanic Origin: Census 2000 Brief. United States Census US Census Bureau, US Department of Commerce.. 2000.

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD: **Estimating African American admixture proportions by use of population-specific alleles.**
Am J Hum Genet 1998, **63:**1839-1851. PubMed Abstract | Publisher Full Text

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG: **Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture.**
Am J Hum Genet 1988, **43:**520-526. PubMed Abstract

Tseng M, Williams RC, Maurer KR, Schanfield MS, Knowler WC, Everhart JE: **Genetic admixture and gallbladder disease in Mexican Americans.**
Am J Phys Anthropol 1998, **106:**361-371. PubMed Abstract | Publisher Full Text

Hanis CL, Newett-Emmett D, Bertin TK, Schull WJ: **Origins of U.S. Hispanics. Implications for diabetes.**
Diabetes Care 1991, **14:**618-627. PubMed Abstract

McLeod HL: **Pharmacogenetics: more than skin deep.**
Nat Genet 2001, **29:**247-248. PubMed Abstract | Publisher Full Text

Merryweather-Clarke AT, Pointon JJ, Jouanolle AM, Rochette J, Robson KJ: **Geography of HFE C282Y and H63D mutations.**
Genet Test 2000, **42:**183-198. Publisher Full Text

Williams RC, Long JC, Hanson RL, Sievers ML, Knowler WC: **Individual estimates of European genetic admixture associated with lower body-mass index, plasma glucose, and prevalence of type 2 diabetes in Pima Indians.**
Am J Hum Genet 2000, **66:**527-538. PubMed Abstract | Publisher Full Text

Dean M, Stephens JC, Winkler C, Lomb DA, Ramsburg M, Boaze R, Stewart C, Charbonneau L, Goldman D, Albaugh BJ, *et al*.: **Polymorphic admixture typing in human ethnic populations.**
Am J Hum Genet 1994, **55:**788-808. PubMed Abstract

Smith MW, Lautenberger JA, Shin HD, Chretien J-P, Shrestha S, Gilbert DA, O'Brien SJ: **Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations.**
Am J Hum Genet 2001, **69:**1080-1094. PubMed Abstract | Publisher Full Text

Exner DV, Dries DL, Domanski MJ, Cohn JN: **Lesser response to angiotensin-converting-enzyme inhibitor therapy in black as compared with white patients with left ventricular dysfunction.**
N Engl J Med 2001, **344:**1351-1357. PubMed Abstract | Publisher Full Text

Wood AJJ: **Racial differences in the response to drugs - pointers to genetic differences.**
N Engl J Med 2001, **344:**1393-1396. PubMed Abstract | Publisher Full Text

Lin SS, Kelsey JL: **Use of race and ethnicity in epidemiologic research: concepts, methodological issues, and suggestions for research.**
Epidem Revs 2000, **22:**187-202. PubMed Abstract

Ziegler RG, Hoover RN, Pike MC, Hildesheim A, Nomura AM, West DW, Wu-Williams AH, Koloner LN, Horn-Ross PL, Rosenthal JF, *et al*.: **Migration patterns and breast cancer risk in Asian-American women.**
J Natl Cancer Inst 1993, **85:**1819-1827. PubMed Abstract

Compston A: **Distribution of multiple sclerosis.**
In McAlpine's Multiple Sclerosis. Edited by Compston A, Ebers G, Lassman H, McDonald I, Matthews B, Wekerle H. Churchill Livingston: London; 1991, 63-100.

Ebers GC, Sadovnick AD: **Epidemiology.**
In Multiple Sclerosis. Edited by Paty DW, Ebers GC. Philadelphia: FA Davis Company; 1998, 5-28.

Karter AJ, Ferrara A, Liu JY, Moffet HH, Ackerson LM, Selby JV: **Ethnic disparities in diabetic complications in an insured population.**
JAMA 2002, **287:**2519-2527. PubMed Abstract | Publisher Full Text

Kaufman JS, Cooper RS: **Commentary: considerations for use of racial/ethnic classificaiton in etiologic research.**
Am J Epidemiol 2001, **154:**291-298. PubMed Abstract | Publisher Full Text

Cruickshanks KJ, Jobim LF, Lawler-Heavner J, Neville TG, Gay EC, Chase HP, Klingensmith G, Todd JA, Hamman RF: **Ethnic differences in human leukocyte antigen markers of susceptibility to IDDM.**
Diabetes Care 1994, **17:**132-137. PubMed Abstract