Individual ancestry inference and the reification of race as a biological phenomenon

Article · January 2008

CITATIONS
87

READS
1,046

1 author:

Deborah A Bolnick
University of Texas at Austin
81 PUBLICATIONS **1,630** CITATIONS

SEE PROFILE

Rutgers Series in Medical Anthropology

Edited by Mac Marshall

Revisiting Race in a Genomic Age

EDITED BY
BARBARA A. KOENIG,
SANDRA SOO-JIN LEE,
AND SARAH S. RICHARDSON

RUTGERS UNIVERSITY PRESS
NEW BRUNSWICK, NEW JERSEY, AND LONDON

4

Individual Ancestry Inference and the Reification of Race as a Biological Phenomenon

DEBORAH A. BOLNICK

Anthropological ideas about the pattern of human diversity shifted drastically during the 20th century. Prior to World War II, *Homo sapiens* was generally perceived as a polytypic species with biologically distinct subgroups, or races (Stepan, 1982; Marks, 1995). This biological differentiation was thought to be the result of long periods of independent evolution when each race was largely isolated from the others. Anthropologists gradually moved away from such typological thinking during the latter half of the 20th century, in part because new genetic data did not support this paradigm. Instead, genetic research suggested that humans could not be neatly divided into a few discrete, isolated races (Brown & Armelagos, 2001; Kittles & Weiss, 2003). Studies of human biological diversity therefore began to focus less on classification and more on the actual patterns of variation among populations, as well as on the evolutionary processes that shaped those patterns.

With this shift away from typological thought has come an increased interest in *individuals* and what genetics can tell us about the unique identity and history of each person. As part of this trend, anthropologists and geneticists have recently begun to explore how genomic data can be used to infer an individual's "ancestry." I will consider the meaning of this term in more detail later in this chapter, but "ancestry" is generally used to refer to the geographic region or regions where one's biological ancestors lived (Jorde & Wooding, 2004; Race, Ethnicity, and Genetics Working Group, 2005).

Several methods have been developed for inferring an individual's ancestry from genetic data (Rannala & Mountain, 1997; McKeigue, Carpenter, Parra, & Shriver, 2000; Pritchard, Stephens, & Donnelly, 2000), and these methods are starting to be used in a variety of contexts. For example, individual ancestry inference has important biomedical applications because

ancestry may influence disease susceptibility and drug response (Wilson et al., 2001; Risch, Burchard, Ziv, & Tang, 2002; Helgadottir et al., 2005; Tate & Goldstein, this volume). Individual ancestry inference can also aid forensic investigations by determining the genetic heritage of DNA left at a crime scene, which can then be used to narrow the pool of potential suspects (Frudakis et al., 2003; Shriver, Frudakis, & Budowle, 2005). Finally, these methods are also of great interest to members of the general public who want to reconstruct their personal genealogical histories (Elliott & Brodwin, 2002; Bolnick, 2003; TallBear, 2005; Greely, this volume; Shriver & Kittles, this volume; TallBear, this volume).

Although this body of work emphasizes the *individual* as the crucial unit of analysis, individual ancestry inference is closely tied to our understanding of human *groups* and the distribution of genetic variation among them. Inferring an individual's genetic ancestry entails deciding that his or her DNA was inherited from a certain group or groups, and that cannot be accomplished unless one first distinguishes groups that differ genetically in some way. Thus, even such individually oriented genetic research has implications for our understanding of race and the pattern of human biological diversity.

In this chapter, I begin with an overview of our current understanding of the pattern of human biodiversity. I then examine two widely cited studies that use the *structure* program (Pritchard et al., 2000) to infer individual ancestry (Rosenberg et al., 2002; Bamshad et al., 2003) and discuss what these studies imply about the relationship between human genetic structure and traditional notions of race.

The Distribution of Human Genetic Variation

Our current understanding of human genetic structure is based on hundreds of studies that have been conducted over the past few decades. Both mitochondrial DNA and nuclear loci have been surveyed using many different types of markers (Tishkoff & Verrelli, 2003). While the specific findings of each study have varied, two general patterns have consistently emerged.

First, African populations exhibit greater genetic diversity and less linkage disequilibrium than non-African populations (Tishkoff & Williams, 2002; Kittles & Weiss, 2003).[1] This pattern reflects the evolutionary and demographic history of our species. Because *Homo sapiens* evolved in Africa before dispersing throughout the rest of the world (Klein, 1999), African populations are older and have had more time to accumulate genetic differences through mutation. Similarly, the greater age of African populations helps to explain the lower levels of linkage disequilibrium in Africa since linkage disequilibrium decreases over time due to recombination (Kittles & Weiss,

2003). Differences between African and non-African populations also reflect a genetic bottleneck that occurred when humans dispersed out of Africa. The individuals who left Africa carried only a subset of the genetic variants found in the ancestral African population. Consequently, non-Africans are less genetically diverse and exhibit increased linkage disequilibrium compared to Africans (Bamshad, Wooding, Salisbury, & Stephens, 2004; Tishkoff & Kidd, 2004).

The second pattern that has emerged from many genetic studies is that human variation is clinally distributed (see fig. 4-1). Allele frequencies change gradually across geographic space, with few sharp discontinuities (Barbujani, 2005). Populations are most genetically similar to others that are found nearby, and genetic similarity is inversely correlated with geographic distance (Relethford, 2004; Ramachandran et al., 2005).

There are several reasons for this pattern. First, it reflects localized gene flow and isolation by distance (Cavalli-Sforza, Menozzi, & Piazza, 1994; Relethford, 2004). In other words, because geographic distance limits migration, individuals tend to mate with those who live nearby and geographically close populations tend to exchange more genes than geographically distant ones (Wright, 1943; Malécot, 1969). Restricted gene flow therefore

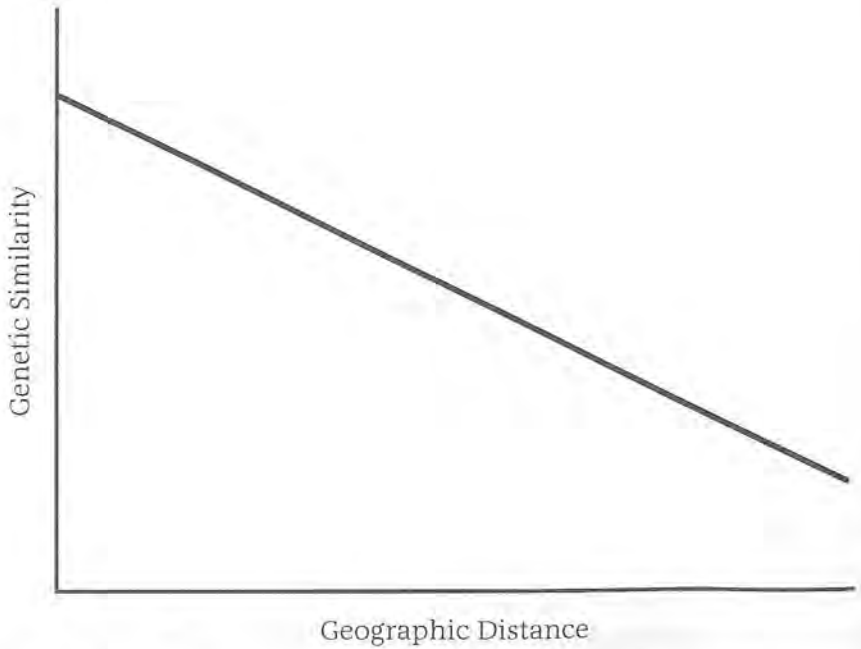

FIGURE 4-1. The relationship between genetic similarity and geographic distance under a pattern of clinal variation.

contributes to the observed pattern of decreasing genetic similarity with increasing geographic distance.

The clinal pattern of human genetic variation also reflects successive founder effects that occurred as humans migrated out of Africa to populate the rest of the world (Relethford, 2004; Prugnolle, Manica, & Balloux, 2005; Ramachandran et al., 2005). Prugnolle et al. (2005) and Ramachandran et al. (2005) suggest that this form of genetic drift played a particularly important role in shaping the patterns of variation among human populations. According to their analyses, serial founder effects explain 76–85% of the observed variation (Prugnolle et al., 2005; Ramachandran et al., 2005).

Finally, clinal variation at some loci reflects selection in response to environmental gradients. Clines due to selection vary from locus to locus (i.e., allele frequencies at one locus change faster than those at another locus over the same geographic distance). Since many loci show similar patterns of allele frequency change across human populations, the overall pattern of genetic variation in our species reflects selection less than serial founder effects and restricted gene flow with isolation by distance (Relethford, 2004).

Because of the patterns of human genetic variation described here, many anthropologists argue that traditional notions of race misrepresent human biological diversity and the evolutionary history of our species. While traditional notions of race are extremely variable—no consensus has ever been reached regarding the number or composition of human races, for example—most describe racial groups as equivalent, biologically distinct units (Barbujani, 2005). However, the patterns described above suggest that this is not the case. From a genetic perspective, non-Africans are essentially a subset of Africans (Quintana-Murci et al., 1999; Underhill et al., 2000; Kidd, Pakstis, Speed, & Kidd, 2004). No discrete boundaries separate humans into a few genetically distinct groups, and the members of each racial group are highly variable (Brown & Armelagos, 2001). Consequently, human racial groups do not appear to be distinct genetic groups.

Individual Ancestry Inference, Race, and Genetic Structure

Several recent studies of individual ancestry seem to challenge this understanding of the distribution of human genetic variation. These new studies instead suggest genetic differentiation among what are essentially races based on continental ancestry. For example, Rosenberg et al. "identified six main genetic clusters, five of which correspond to major geographic regions" (2002, p. 2381). Since the five "major geographic regions" comprise Africa, Eurasia, East Asia, Oceania, and America (Rosenberg et al., 2002), these results have been interpreted as showing that racial divisions based

on continental ancestry are biologically significant (Burchard et al., 2003; Mountain & Risch, 2004). Similarly, Bamshad et al. (2003) identified three genetic clusters that correspond to Africa, Europe, and Asia. These studies have been widely cited as verifying traditional ideas about race and the pattern of human biological diversity (Wade, 2002; Seebach, 2003).[2]

The conclusions of both the Rosenberg et al. (2002) and Bamshad et al. (2003) studies were based on the Bayesian computer program *structure* (Pritchard et al., 2000). To understand the results of these two studies and what they imply about the structure of the human gene pool, it is important to first understand how this computer program works.

The *structure* program implements a model-based clustering method to infer population structure from multilocus genotype data and then allocates individuals into populations (Pritchard et al., 2000). It can be used to estimate the number of genetic clusters or populations present in a given data set as well as the population of origin of each individual. The populations are expected to be in Hardy-Weinberg equilibrium. Pritchard et al. (2000) assume a model in which a number (K) of populations exist, each of which is characterized by a set of allele frequencies.

A data set of multilocus genotypes (X) is therefore viewed as being made up of individuals sampled from K separate populations (see fig. 4-2). When running the *structure* program, the user defines K in advance. *Structure* then assigns individuals probabilistically to K populations with the goal of maximizing Hardy-Weinberg equilibrium in each population. In other words, for any given value of K, *structure* searches for the most probable way to divide the sampled individuals into that pre-defined number of clusters based on their genotypes. If an individual's genotype suggests that he or she has ancestry from more than one population, *structure* can assign the individual jointly to two or more populations and estimate the proportion of ancestry from each. The analysis can (and should) be performed for multiple different values of K.

Thus, the fact that *structure* identifies a particular number of clusters is insignificant: it does so simply because the user told it to do so. What is more

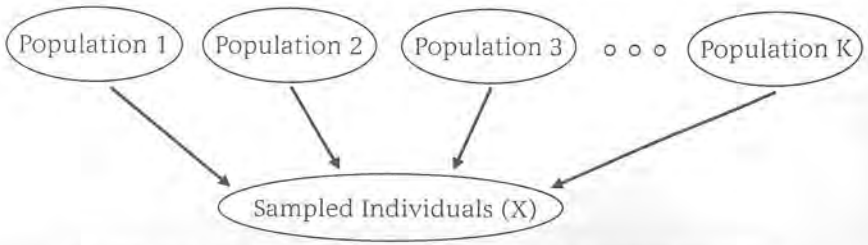

FIGURE 4-2. The population model assumed by the *structure* program.

important is that *structure* provides a way to determine the value of K that is most appropriate for the data set in question (i.e., the most likely number of clusters or populations represented). The "best" value of K is the one that maximizes the probability of observing that set of data. *Structure* calculates the probability of the data given each value of K submitted (i.e., Pr [X|K]), and the inferred value for K is the one associated with the highest Pr (X|K).

However, it is not entirely straightforward to determine the true number of genetic clusters in a given data set for three reasons. First, because it is computationally difficult to estimate Pr (X|K), *structure* provides only an approximation (Pritchard et al., 2000). Pritchard et al. note that "the assumptions underlying [this approximation] are dubious at best, and we do not claim (or believe) that our procedure provides a quantitatively accurate estimate of the posterior distribution of K. We see it merely as an *ad hoc* guide to which models are most consistent with the data, with the main justification being that it seems to give sensible answers in practice" (2000, p. 949). Thus, because *structure*'s estimates of Pr (X|K) may or may not be accurate, the value of K estimated as maximizing the probability of the data may not actually do so.

Second, if a data set is complex, different runs of *structure* may produce substantially different results. In these cases, the composition of genetic clusters varies among runs using the same pre-defined value for K. A simplistic illustration of this situation would be a case where the analysis of four individuals (A, B, C, and D) using $K = 2$ yielded clusters (A, B) and (C, D) in run 1, but clusters (A, C) and (B, D) in run 2. Pritchard and Wen (2004) suggest that this mostly occurs with data sets containing a large number of genetic clusters ($K > 5$) and is either because the program did not run long enough (i.e., *structure* did not have enough time to determine the optimal clustering of individuals) or because there are several highly probable ways to divide the sampled individuals into that number of clusters. If the latter is the case, it may not be possible to determine a single optimal clustering scheme. Furthermore, the different ways to divide individuals into a particular number of clusters may each yield a different Pr (X|K). For example, in the above illustration with $K = 2$, the clustering scheme in run 1 might be associated with a high Pr (X|K), whereas the clustering scheme in run 2 might be associated with a low Pr (X|K). Consequently, it is not always clear which of the probabilities associated with a given K should be used when determining the "best" value for K.

Third, the underlying model used in the *structure* program is not appropriate for all data sets. In particular, Pritchard and Wen (2004) note that the *structure* model is not well suited to data shaped by restricted gene flow with isolation by distance. If *structure* is used to analyze such data, they warn that "the inferred value of K . . . can be rather arbitrary" (2004, p. 14). Thus,

although the *structure* program can estimate the number and composition of genetic clusters present in a given data set, such estimates must be interpreted carefully.

Rosenberg et al.'s (2002) Study of Human Genetic Structure

Rosenberg et al. (2002) used the *structure* program to analyze genotypic variation at 377 autosomal microsatellite loci in 1,056 individuals from around the world (the HGDP-CEPH Human Genome Diversity Cell Line Panel). The abstract of their article mentioned the identification of 6 main genetic clusters (Africa, Eurasia, East Asia, Oceania, America, and the Kalash of Pakistan), but Rosenberg et al. (2002) actually presented results for multiple values of K (2–6) in the body of the paper. They also analyzed the data set using values of $K > 6$ (up to $K = 20$; N. Rosenberg, personal communication), but they did not publish those results because *structure* identified multiple ways to divide the sampled individuals into K clusters when $K > 6$ (Rosenberg et al., 2002). For example, in 10 replicates, *structure* found 9 different ways to divide the sampled individuals into 14 clusters and 10 different ways to divide them into 20 clusters (N. Rosenberg, personal communication). The different clustering schemes in these replicates were fairly similar, but they often yielded very different Pr (X|K), making it difficult to interpret the results for a given value of K when $K > 6$. Rosenberg et al. (2002) therefore published the results for $K < 7$ for the worldwide sample, as well as further analyses using regional subsets of the entire data set.

Thus, the fact that *structure* identified 6 genetic clusters is not significant in and of itself—the program also identified 2, 5, 10, and 20 genetic clusters using the same set of data. As noted above, *structure* will identify as many clusters as the user tells it to identify. While it may be interesting that 5 of the 6 clusters identified with $K = 6$ correspond to major geographic regions, such clustering does not necessarily provide a better representation of human genetic differentiation than the clustering observed when K is set to 4, 9, 12, or any other number. Only by evaluating the probability of the observed data given each value of K (i.e., Pr [X|K]) is it possible to determine the number of genetic clusters *most likely* represented in this data set.

Rosenberg et al. (2002) did not report the *most likely* number of genetic clusters, nor did they publish the probabilities of the observed data given each value of K. Since some of the larger values of K were associated with several different Pr (X|K) across runs, and since Rosenberg et al. wanted to present results that could be easily replicated, they felt that it was more informative to show the robust results for multiple small K than to focus on a larger value of K that was associated with variable clustering schemes and both high and low probabilities (N. Rosenberg, personal communication).

In other words, no single value of K clearly maximized the probability of the observed data. Probabilities increased sharply from $K = 1$ to $K = 4$ but were fairly similar for values of K ranging from 4 to 20 (N. Rosenberg, personal communication). The probability of the observed data was higher for $K = 6$ than for smaller values of K, but not as high as for some replicates with larger values of K (N. Rosenberg, personal communication). The highest Pr (X|K) was associated with a particular replicate of $K = 16$, but that value of K was also associated with very low probabilities when the individuals were grouped into 16 clusters in other ways (N. Rosenberg, personal communication). Consequently, it is uncertain what number of genetic clusters *best* fits this data set, but there is no clear evidence that $K = 6$ is the best estimate.

Thus, the Rosenberg et al. (2002) study does not challenge our current understanding of human genetic structure as much as some have suggested. Indeed, the fact that it was not possible to determine a single best value for K is exactly what we would expect given the clinal variation and pattern of isolation by distance found in our species. In addition, as Tishkoff and Kidd (2004) have noted, individuals from areas near the borders of the five "major geographic regions" exhibited ancestry from multiple genetic clusters. These results suggest a gradient of change between geographic regions, not discrete boundaries.[3]

So why has so much emphasis been placed on the results of the analysis using $K = 6$? Despite the fact that Rosenberg et al. (2002) presented no evidence that $K = 6$ represented the *most likely* number of genetic clusters in their data set, virtually all references to this study in both the scientific literature and the popular press mention the identification of either 5 or 6 genetic clusters (for examples, see Wade, 2002; Seebach, 2003; Bamshad et al., 2004; Tishkoff & Kidd, 2004; Barbujani, 2005; and Tate & Goldstein, this volume). I would suggest that these particular results have been emphasized simply because they fit the general notion in our society that continental groupings are biologically significant. This notion is a legacy of traditional racial thought and seems to persist even when not clearly supported by biological data.

Bamshad et al.'s (2003) Study of Population Structure and Group Membership

Bamshad et al. (2003) analyzed 100 *Alu* insertion polymorphisms in 565 individuals from sub-Saharan Africa, East Asia, Europe, and India, as well as 60 microsatellites in 206 of the individuals from sub-Saharan Africa, Europe, and East Asia. Like Rosenberg et al. (2002), they used the *structure* program to help determine the number of genetic clusters present in their data set. Bamshad et al. (2003) first analyzed only the samples from sub-Saharan

Africa, East Asia, and Europe and ran *structure* using values of K between 1 and 6. When all of the individuals from these three regions were included in the analysis, they found that $K = 4$ maximized the probability of observing that set of data (Pr [X|K = 4] = 1). The sampled individuals likely represented four genetic clusters, comprising (1) East Asians, (2) Europeans, (3) sub-Saharan Africans except for the Mbuti and three other individuals, and (4) the Mbuti and three other sub-Saharan Africans. This division of sub-Saharan Africans into two genetic clusters is consistent with other evidence of greater genetic diversity and greater genetic structure among Africans (Tishkoff & Williams, 2002).

Bamshad et al. (2003) also conducted this analysis excluding the Mbuti samples. In this case, *structure* found that $K = 3$ best fit the observed data (Pr [X|K = 3] = 1), indicating three genetic clusters of individuals (sub-Saharan Africans, Europeans, and East Asians). Bamshad et al. set K equal to 3 for most subsequent analyses even though those analyses used the complete data set (including the Mbuti), which most likely contained four genetic clusters. Given this, Bamshad et al. may have taken other (unnamed) factors into consideration when deciding upon the number of genetic clusters present in their data. They did note that "K provides only a rough guide for determining which models may be consistent with the data" because estimates of K depend on the number of individuals per population, the number of loci studied, and the amount of differentiation between populations (Bamshad et al., 2003, p. 579).

Bamshad et al. (2003) also investigated whether genetic data could be used to correctly infer an individual's ancestry. They used *structure* to assign individuals to genetic clusters and to estimate the proportion of an individual's ancestry from each genetic cluster. An individual "was considered assigned 'correctly' if the cluster with the greatest proportion of ancestry was the same as the continent of origin of the sample" (Bamshad et al., 2003, p. 579). Thus, Bamshad et al. assumed that continental groupings were important from the start, which perhaps explains why they chose $K = 3$ as the best estimate of human genetic structure.

Using the 100 *Alu* markers and 60 microsatellites, *structure* was able to identify the correct continent of origin for 99% of the individuals from sub-Saharan Africa, Europe, and East Asia. These results demonstrate that substantial genetic differentiation exists among the populations sampled, but they do not necessarily indicate substantial genetic differentiation among continental groupings. As Bamshad et al. (2003) note, these analyses included individuals from only a few widely separated regions of Africa, Asia, and Europe. The observed genetic differentiation may therefore reflect the large geographic distances between sampled populations rather than continental divisions per se.

Additional analyses using samples from areas closer to the continental borders support this hypothesis. When Bamshad et al. (2003) included samples from India in a data set with the European and East Asian samples, *structure* found that the optimal number of genetic clusters (K) was one. In other words, when a more representative geographic sample was analyzed, continental groupings no longer appeared to be genetically distinct. Accordingly, Bamshad et al. concluded that "the inclusion of [geographically intermediate] samples demonstrates geographic continuity in the distribution of genetic variation and thus undermines traditional concepts of race" (2003, p. 587).

The *structure* analysis in this study therefore supports our current understanding of human genetic structure. It indicates greater genetic diversity and greater genetic structure among Africans as well as little genetic differentiation among continental groupings *when a representative geographic sample is analyzed*. The results using the expanded Eurasian sample are also consistent with previous evidence that human genetic variation is clinally distributed with few sharp discontinuities.

Despite these results, Bamshad et al.'s (2003) study has been cited as showing that groups defined by continental ancestry or race are genetically differentiated (Mountain & Risch, 2004). This interpretation likely reflects the way that Bamshad et al. (2003) presented their results, rather than the results themselves. First, as noted above, they emphasized the significance of the three continental groupings even though *structure* identified four genetic clusters in the complete data set of sub-Saharan Africans, East Asians, and Europeans. Second, while Bamshad et al. (2003) recognized that the inclusion of samples from India demonstrated the continuous distribution of genetic variation in Eurasia, they excluded those samples from most analyses. As a result, many of the reported analyses implied continental discontinuities even though the more complete data set showed that such discontinuities do not exist.

Third, *structure* found that the European, Indian, and East Asian data set most likely contained a single genetic cluster, but Bamshad et al. (2003) focused primarily on an analysis of that data set using $K = 3$. In the text of their paper, Bamshad et al. wrote: "If we assumed that three clusters were present (i.e., $K = 3$), as suggested by proxy information (i.e., place of origin), three groups were distinguished. Correct assignment of samples to their place of origin was 97% for samples from East Asia, 94% for samples from Europe, and 87% for samples from southern India" (2003, p. 584). The article abstract also made no mention of the optimal clustering scheme ($K = 1$), but instead stated that "less accurate assignment (87%) to the appropriate genetic cluster was possible for a historically admixed sample from southern India" (Bamshad et al., 2003, p. 578).

Of course, as noted earlier, *structure* will identify as many groups as the program user tells it to identify, so it is not surprising—or significant—that *structure* distinguished three groups when Bamshad et al. set K equal to 3. Nor is it clear how to interpret the results of this analysis since it is statistically unlikely that three genetic clusters really exist in this data set. Bamshad et al.'s (2003) presentation of this analysis obscures these issues and makes it seem as if the three racial/ethnic groups (Europeans, East Asians, and Indians) are more genetically distinct than they really are.

Finally, Bamshad et al.'s description of the Indian population as "historically admixed" (2003, p. 578) reinforces traditional racial views of human variation and human evolutionary history. Previous studies have shown that the genetic makeup of the Indian population reflects gene flow from European and Asian sources (Bamshad et al., 2001; Majumder, 2001), but Bamshad et al.'s (2003) description suggests that such gene flow occurred only in historical times. Racial studies of the early 20th century presented a similar picture of Eurasian history. For example, Hooton (1931, 1939) suggested that populations resembling multiple races (such as Indians) formed only recently due to gene flow between the primary races, which were isolated from one another during prehistoric times. There is no evidence that a significant barrier to Eurasian gene flow existed in the more distant past, though, and other studies indicate migration and gene flow throughout Eurasia at many points in human history (Templeton, 2002; Basu et al., 2003). Thus, the way that Bamshad et al. (2003) describe their research reinforces traditional racial views of human variation even though the data do not necessarily support such views.

Ancestry and Race

Given the descriptions and interpretations of the studies by Rosenberg et al. (2002) and Bamshad et al. (2003), the relationship between ancestry and race should be examined more carefully. In recent years, ancestry has been widely promoted as an objective, scientific alternative to race. The term "ancestry" is often used without being clearly defined, but it generally refers to the geographic region or regions where one's biological ancestors lived (Collins, 2004; Jorde & Wooding, 2004; Shriver et al., 2004; Race, Ethnicity, and Genetics Working Group, 2005). Because of this focus, ancestry is seen as being more specific and objective than race (Bamshad, 2005), which is highly charged and encompasses geographic origins, political history, socioeconomic status, culture, skin color, and other perceived physical, behavioral, and genetic characteristics. Jorde and Wooding also argue that ancestry is "a more subtle and complex description of an individual's genetic makeup than is race" (2004, p. S30). An individual can have ancestry from

multiple geographic regions, and the concept of ancestry is flexible enough that those regions could be local (e.g., southwestern Nigeria) or much broader (e.g., all of Africa).

However, other aspects of ancestry are more problematic. Just as the term is rarely defined, there has been little discussion of the size of geographic regions, how they should be defined, or why specific geographic divisions are more relevant than others for studies of human genetic variation. Nor is it always clear what time frame should be considered when determining an individual's ancestry. For example, my grandparents lived in the United States, but my great-great-grandparents lived in Eastern Europe. My more distant ancestors, like those of all humans, lived in Africa. The time depth of interest depends on the question or hypothesis being addressed, but this issue is often discussed only briefly, if at all.

Furthermore, ancestry is not that different from race in practice. Like race, ancestry is sometimes defined politically or culturally (Race, Ethnicity, and Genetics Working Group, 2005). In individual ancestry studies, the ancestral regions are almost always continents (Risch et al., 2002; Mountain & Risch, 2004; The Unexamined "Caucasian," 2004). Since the contemporary Euro-American definition of race is based on continental geography, anthropologists and human geneticists use the term "ancestry" much as the general public uses the term "race." Indeed, some scientists explicitly define ancestry as an individual's racial group or the race of his or her ancestors (Risch et al., 2002; Frudakis et al., 2003).

Because an individual can have ancestry from multiple geographic regions, ancestry does differ from conceptions of race based on the one-drop rule, which allow an individual to belong to only a single race. However, contemporary understandings of race accept the existence of "mixed-race" individuals, as evidenced by the large number of Americans who checked the box associated with the Other category on the last U.S. census. Thus, while ancestry has the potential to be a more subtle, objective, and scientific alternative to race, it currently appears to be quite similar to race in practice.

Conclusion

Recent studies of individual ancestry have been cited as verifying traditional ideas about race, but these studies do not present new data suggesting that racial groups are genetically distinct. Rather, the data and *structure* analyses reported in the Rosenberg et al. (2002) and Bamshad et al. (2003) studies are consistent with our current understanding of human genetic structure. However, the results of these studies have been described and interpreted in ways that both reflect and reinforce traditional racial views of human biological diversity and the evolutionary history of our species. The disconnect

between the results and the interpretations of these studies is unfortunate since they are playing an important role in the reification of race as a biological phenomenon.

NOTES

I would like to thank Sarah Richardson, Sandra Soo-Jin Lee, Barbara Koenig, Daniel Bolnick, and Noah Rosenberg for their helpful suggestions and comments regarding an earlier draft of this chapter. I am also grateful to the participants in the authors' conference, who provided many valuable discussions.

1. Linkage disequilibrium (LD) refers to (a) the nonrandom association of alleles at different sites and (b) the length of a chromosomal segment that is inherited without recombination from a common ancestor (Kittles & Weiss, 2003; Tishkoff & Kidd, 2004).
2. Other studies have been cited as proving the same point, but I do not discuss those studies in this chapter since they are based on different methods of analysis than the Rosenberg et al. (2002) and Bamshad et al. (2003) studies. For example, Frudakis et al. (2003) used a linear classification method and Shriver et al. (2004) used a tree-based method. DNAPrint's AncestrybyDNA test, which also suggests the validity of race as a biological phenomenon (Bolnick, 2003), is based on an admixture-mapping approach (R. Malhi, personal communication).
3. Rosenberg et al. (2005) reanalyzed 1,048 individuals from the HGDP-CEPH Human Genome Diversity Panel after expanding their data set to include 993 markers. As in the 2002 article, Rosenberg et al. (2005) presented the results of *structure* analyses using $K = 2$–6 without specifying the *most likely* number of genetic clusters represented in the data set. Rosenberg et al. (2005) also examined the effects of several variables on the "clusteredness" of individuals, or the extent to which an individual was estimated as belonging to a single cluster. Although the HGDP-CEPH Human Genome Diversity Cell Line Panel does not represent a comprehensive sample of all regions occupied by humans, they found that more continuous geographic sampling would have little impact on the observed degree of clustering. Finally, Rosenberg et al. (2005) suggested that the clustering observed with $K = 5$ reflects slightly greater genetic differences across geographic barriers like oceans or the Sahara desert. Since $K = 5$ may or may not be the best estimate of the number of genetic clusters present in this data set, the biological significance of this finding is unclear.

REFERENCES

Bamshad, M. (2005). Genetic influences on health: Does race matter? *Journal of the American Medical Association, 294*, 937–946.

Bamshad, M., Kivisild, T., Watkins, W. S., Dixon, M. E., Ricker, C. E., Rao, B. B., et al. (2001). Genetic evidence on the origins of Indian caste populations. *Genome Research, 11*, 994–1004.

Bamshad, M., Wooding, S., Salisbury, B. A., & Stephens, J. C. (2004). Deconstructing the relationship between genetics and race. *Nature Reviews Genetics, 5*, 598–609.

Bamshad, M. J., Wooding, S., Watkins, W. S., Ostler, C. T., Batzer, M. A., & Jorde, L. B. (2003). Human population structure and inference of group membership. *American Journal of Human Genetics, 72*, 578–589.

Barbujani, G. (2005). Human races: Classifying people vs. understanding diversity. *Current Genomics, 6*, 215–226.

Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., et al. (2003). Ethnic India: A genomic view, with special reference to peopling and structure. *Genome Research, 13*, 2277–2290.

Bolnick, D. A. (2003). "Showing who they really are": Commercial ventures in genetic genealogy. Paper presented at the American Anthropological Association Annual Meeting. November, Chicago, IL.

Brown, R., & Armelagos, G. J. (2001). Apportionment of racial diversity: A review. *Evolutionary Anthropology, 10*, 34–40.

Burchard, E. G., Ziv, E., Coyle, N., Gomez, S. L., Tang, H., Karter, A. J., et al. (2003). The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine, 348*, 1170–1175.

Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. (1994). *The history and geography of human genes*. Princeton: Princeton University Press.

Collins, F. S. (2004). What we do and don't know about "race," "ethnicity," genetics, and health at the dawn of the genome era. *Nature Genetics, 36*, S13–S15.

Elliott, C., & Brodwin, P. (2002). Identity and genetic ancestry tracing. *British Medical Journal, 325*, 1469–1471.

Frudakis, T., Venkateswarlu, K., Thomas, M. J., Gaskin, Z., Ginjupalli, S., Guntari, S., et al. (2003). A classifier for the SNP-based inference of ancestry. *Journal of Forensic Sciences, 48*, 771–778.

Greely, H. T. (2008 [this volume]). Genetic genealogy: Genetics meets the marketplace. In B. A. Koenig, S. S.-J. Lee, & S. S. Richardson (Eds.), *Revisiting race in a genomic age* (pp. 215–234). New Brunswick, NJ: Rutgers University Press.

Helgadottir, A., Manolescu, A., Helgason, A., Thorleifsson, G., Thorsteindottir, U., Gudbjartsson, D. F., et al. (2005). A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nature Genetics, 38*, 68–74.

Hooton, E. (1931). *Up from the ape*. New York: Macmillan Company.

Hooton, E. (1939). *Twilight of man*. New York: G. P. Putnam's Sons.

Jorde, L. B., & Wooding, S. P. (2004). Genetic variation, classification, and "race." *Nature Genetics, 36*, S28–S33.

Kidd, K. K., Pakstis, A. J., Speed, W. C., & Kidd, J. R. (2004). Understanding human DNA sequence variation. *Journal of Heredity, 95*, 406–420.

Kittles, R. A., & Weiss, K. M. (2003). Race, ancestry, and genes: Implications for defining disease risk. *Annual Review of Genomics and Human Genetics, 4*, 33–67.

Klein, R. G. (1999). *The human career: Human biological and cultural origins* (2nd ed.). Chicago: University of Chicago Press.

Majumder, P. P. (2001). Ethnic populations of India as seen from an evolutionary perspective. *Journal of Biosciences, 26*, 533–545.

Malécot, G. (1969). *The mathematics of heredity*. San Francisco: W. H. Freeman.

Marks, J. (1995). *Human biodiversity: Genes, race, and history*. New York: Aldine de Gruyter.

McKeigue, P. M., Carpenter, J., Parra, E. J., & Shriver, M. D. (2000). Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: Application to African-American populations. *Annals of Human Genetics, 64*, 171–186.

Mountain, J., & Cavalli-Sforza, L. L. (1997). Multilocus genotypes, a tree of individuals, and human evolutionary history. *American Journal of Human Genetics, 61*, 705–718.

Mountain, J. L., & Risch, N. (2004). Assessing genetic contributions to phenotypic differences among "racial" and "ethnic" groups. *Nature Genetics, 36*, S48–S53.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*, 945–959.

Pritchard, J. K., & Wen, W. (2004). *Documentation for Structure software: Version 2*. Chicago, IL.

Prugnolle, F., Manica, A., & Balloux, F. (2005). Geography predicts neutral genetic diversity of human populations. *Current Biology, 15*, R159–R160.

Quintana-Murci, L., Semino, O., Bandelt, H.-J., Passarino, G., McElreavey, K., & Santachiara-Benerecetti, A. S. (1999). Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through Eastern Africa. *Nature Genetics, 23*, 437–441.

Race, Ethnicity, and Genetics Working Group. (2005). The use of racial, ethnic, and ancestral categories in human genetics research. *American Journal of Human Genetics, 77*, 519—532.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of the Sciences USA, 102*, 15942–15947.

Rannala, B., & Mountain, J. L. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of the Sciences USA, 94*, 9197–9201.

Relethford, J. H. (2004). Global patterns of isolation by distance based on genetic and morphological data. *Human Biology, 76*, 499–513.

Risch, N., Burchard, E., Ziv, E., & Tang, H. (2002). Categorization of humans in biomedical research: Genes, race, and disease. *Genome Biology, 3*, 1–12.

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., & Feldman, M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics, 1*, 660–671.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., et al. (2002). Genetic structure of human populations. *Science, 298*, 2381–2385.

Seebach, L. (2003, May 8). Biology and race: A clearer link; new genetic research establishes firmer basis for connection. *Rocky Mountain News*, p. 58A.

Shriver, M., Frudakis, T., & Budowle, B. (2005). Getting the science and the ethics right in forensic genetics. *Nature Genetics, 37*, 449–450.

Shriver, M. D., Kennedy, G. C., Parra, E. J., Lawson, H. A., Sonpar, V., Huang, J., et al. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human Genomics, 1*, 274–286.

Shriver, M. D., & Kittles, R. A. (2008 [this volume]). Genetic ancestry and the search for personalized genetic histories. In B. A. Koenig, S. S.-J. Lee, & S. S. Richardson (Eds.), *Revisiting race in a genomic age* (pp. 201–214). New Brunswick, NJ: Rutgers University Press.

Stepan, N. L. (1982). *The idea of race in science*. Hamden, CT: Archon Books.

TallBear, K. (2005). Native American DNA: Narratives of origin and race. Ph.D. Dissertation. University of California at Santa Cruz.

TallBear, K. (2008 [this volume]). Native-American-DNA.com: In search of Native American race and tribe. In B. A. Koenig, S. S.-J. Lee, & S. S. Richardson (Eds.), *Revisiting race in a genomic age* (pp. 235–252). New Brunswick, NJ: Rutgers University Press.

Tate, S. K., & Goldstein, D. B. (2008 [this volume]). Will tomorrow's medicines work for everyone? In B. A. Koenig, S. S.-J. Lee, & S. S. Richardson (Eds.), *Revisiting race in a genomic age* (pp. 102–128). New Brunswick, NJ: Rutgers University Press.

Templeton, A. (2002). Out of Africa again and again. *Nature, 416*, 45–51.

Tishkoff, S. A., & Kidd, K. K. (2004). Implications of biogeography of human populations for "race" and medicine. *Nature Genetics, 36*, S21–S27.

Tishkoff, S. A., & Verrelli, B. C. (2003). Patterns of human genetic diversity: Implications for human evolutionary history and disease. *Annual Review of Genomics and Human Genetics, 4*, 293–340.

Tishkoff, S. A., & Williams, S. M. (2002). Genetic analysis of African populations: Human evolution and complex disease. *Nature Reviews Genetics, 3*, 611–621.

Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., et al. (2000). Y chromosome sequence variation and the history of human populations. *Nature Genetics, 26*, 358–361.

The Unexamined "Caucasian." (2004). *Nature Genetics, 36*, 541.

Wade, N. (2002, December 20). Gene study identifies five main human populations, linking them to geography. *New York Times*, p. A37.

Wilson, J. F., Weale, M. E., Smith, A. C., Gratrix, F., Fletcher, B., Thomas, M. G., et al. (2001). Population genetic structure of variable drug response. *Nature Genetics, 29*, 265–269.

Wright, S. (1943). Isolation by distance. *Genetics, 28*, 114–138.