3

The Biology of Ancestry

DNA, Genomic Variation, and Race

Marcus W. Feldman

Emerging knowledge about genomic variation within and across different human populations suggests the ways genes affect human phenotypes (e.g., skin color, hair texture, and disease), how natural selection works, and what has been the history of human migrations out of Africa and onto other continents. After noting the role biology has played in the historical classification of human beings, this essay discusses the implications of recent technological advances on researchers' ability to detect human genomic variation at minute levels. Following a review of the methods used to determine different populations' continental ancestries, the essay turns to the relationship between natural selection, phenotypes, and genotypes, especially as they are associated with continental origin. The essay concludes by noting that only a small fraction of human DNA is responsible for the visible physical differences that are used to assign people to different "races," and moreover, that phenotypes are never attributable to genes alone. While further research into humans' genetic ancestries should increase knowledge of the role of genes in the manifestation of diseases, the social category of "race" is less useful to scientists than what we might more accurately call "ancestry groups."

WHO WERE THE ETRUSCANS? FOR A PERIOD FROM ABOUT THE NINTH century BCE to the first century BCE, a region of Italy that includes Tuscany and parts of Umbria was the center of an advanced and distinctive

culture that we now recognize as Etruscan. Etruscan culture then disappeared. Although no Etruscan community remains today, by examining the genetic constitution of people living in a part of Tuscany thought to have been a center of Etruscan culture more than two thousand years ago, the origin of the Etruscans can be inferred. One part of this genetic constitution—which is contained in mitochondria (maternally inherited bodies found inside all human cells) taken from today's residents of the town of Murlo in Tuscany—shows great similarity to the corresponding genetic material from people in the Near East (Achilli et al. 2007). Modern population and molecular genetics have, in this way, helped to solve a long-standing conundrum; it is most likely that the Etruscans were part of or the result of a post-neolithic migration from the Near East, possibly Anatolia (Renfrew 1987).

This story of the Etruscans is introduced as an example of how our understanding of relationships among humans—both historical and contemporary—can be informed by modern genetic and statistical technologies. Population geneticists are able to analyze genetic material from humans sampled around the world, from different populations on different continents, and to ask how and why these individuals, and the populations from which they derive, are related. Understanding the patterns of human diversity around the world and how these patterns came about is essential if we are to understand patterns of human migration and cultural contact.

The same molecular and statistical techniques can be helpful to scientific efforts to understand the relationship between genetics and observable human phenotypes. Phenotypes are the measurable or observable features of an organism. For a fruit fly, we might be interested in the number of bristles on its abdomen; for a mouse, how it reacts to threatening situations, and for humans, phenotypes might include diseases (versus non-disease states) or height or hair color or intelligence. So, for example, when people with a disease are compared to healthy people who have been selected to match them on criteria such as age, socioeconomic status (SES), and ethnicity (these are called matched controls), then scientists can more easily isolate which parts of humans' genetic material contribute to producing that disease. If families are available with some members having the disease and others healthy, even greater reliability can be achieved for an inference that some gene is involved as a cause or origin of that disease. An important part of current genetic research—the search for genetic factors related to diseases, especially complex diseases such as hypertension or coronary heart disease—continues apace.

As helpful as new genetic research is in finding biological contributors to diseases or how populations came to have their current genetic relationships, this research can be confusing and even pernicious if it is misinterpreted by those without a sophisticated understanding of the possible relationships between genes, phenotypes, and environments. This is especially dangerous in the context of contentious issues such as race and ethnicity and phenotypic differences between presumed races or ethnic groups. Accordingly, this essay reviews the data pertaining to human genetic diversity and asks how these data relate to the definition and meaning of racial classification. At stake is the question of whether there is any connection between recent discoveries about human genomic variation and the social science concept of "race." The essay concludes with the finding that only a small fraction of all genomic variation is responsible for the differences such as skin color, facial features, or hair form that are commonly used to assign people to different races. More important, the small fraction of genomic variation that does exist between people of different continental ancestry has nothing to do with supposed racial differences related to intelligence, moral character, or tendency to criminality.

VARIATION IN HUMAN DNA

The problem of classification has always played a central role in the biological sciences. Indeed, Darwin's great contribution to science was to show how the well-studied level of biological classification, namely, the subdivision of life forms into species, could arise by natural selection within accepted classes. Although the variation within species had been observed and recorded, it was not until Darwin's groundbreaking work that its relevance to between-species variation was recognized.

In his travels, Darwin observed and was intrigued by the great deal of variation among humans and by attempts to classify them into groups commonly identified at the time as races. In his review of attempts at classification in his 1871 book, *The Descent of Man*, Darwin expressed doubt about the usefulness of phenotypic variation for the purpose of racial classification, stating that "it may be doubted whether any character can be named which is distinctive of race" (225). He added, moreover, that the "variability of all the characteristic differences between the races . . . indicates that these differences cannot be of much importance" (249). Nevertheless, scholars that Darwin saw as "capable judges" had attempted to divide humans into subgroups; Darwin cites claims

that there were from two to sixty-three "races" (226). The most famous of these "judges" was Johann Friedrich Blumenbach, whose partition of human phenotypes into five major racial groups became widely accepted.

Since *The Descent of Man* was published, a vast amount of biological knowledge has accumulated about the relationships between species and the genetic material—DNA, or deoxyribonucleic acid—that is passed from a parent to its offspring. In the past fifteen years, billions of dollars have been spent to produce a picture of the human DNA, which is referred to as the human genome, and now billions more are being devoted to an assessment of how this genome varies between people from the same and different parts of the world.

DNA is the genetic material in the cells of all kinds of organisms from bacteria to humans that contains the information needed to produce things like the red cell antigens that make human blood groups such as A, B, and O, or the pigment molecules that determine the color patterns on butterfly wings. We now know that the part of the DNA that is responsible for these molecules, called the coding regions, comprises only about 5 percent of the three billion units of DNA that make up the human genome. These units are called nucleotides, and the coding strings of nucleotides are the genes. The function of the rest of the human genome is not yet fully understood, although some of it acts to control how much product a gene makes or when and in what tissue the gene product is produced.

Before the 1990s, biologists predominantly studied human diseases whose patterns of occurrence among relatives exhibited the same rules as the blood groups referred to above, namely, Mendel's rules. But since the human genome was sequenced, techniques have become available to detect many kinds of DNA sequence variation, including variation at single nucleotides (single nucleotide polymorphisms, or SNPs), variation in the number of times a short arrangement of nucleotides is repeated (short tandem repeats, STRs), or variation in the presence, absence, or duplication of a segment of DNA (copy number variants, CNVs). These classes of variation, most often in nongenic regions that make up 95 percent of the human genome, have emerged as the foci of study. To date, large amounts of money have been spent on documenting these new kinds of genomic variation in four samples of the human population: European Americans (from Utah), Africans (members of the Yoruba people from Nigeria), Japanese, and Chinese. Millions of SNPs have been discovered in these samples, and there have been hundreds of studies

devoted to interpreting this variation (e.g., Barreiro et al. 2008). These data and much of their analysis are referred to as the HapMap project. We now know that the world's human population exhibits thousands of CNVs, thousands of STR polymorphisms, and millions of SNPs—levels of DNA variation not suspected a mere fifteen years ago.

These new types of genomic variation have begun to dominate the study of human genetics for three reasons. The first is the search for genetic contributions to complex phenotypes such as height, weight, schizophrenia, coronary artery disease, or intelligence. No single gene is responsible for these phenotypes, and any genetic contributions to their variation are likely to be small, to involve many parts of the genome, and to interact with environments in complex ways. However, if variation in a part of the genome is found to contribute to variation in the risk of having some disease, it might provide clues as to the biochemical nature of the causes of the disease and perhaps suggest biochemical interventions, namely, drugs. On the other hand, knowledge that you have a variant that adds slightly to your risk of eventually having a specific disease might make you particularly conscientious about elements of your behavior or your diet that also are known to affect your risk of getting that disease.

The second reason the study of human genetics has begun to focus on the new types of genomic variation comes from the hope that the patterns of variation across the genome will enable scientists to understand more about natural selection. Investigators are interested, for example, in knowing which parts of the genome exhibit signals of the action of natural selection due to the different environments that different human populations have occupied. These patterns of variation are much more cryptic than the well-known hemoglobin A–hemoglobin S polymorphism that gives some people a genetic advantage against malaria in the parts of the world where this parasitic disease is common.

Finally, the third reason for the new focus in the study of human genetics is that if we know the pattern of genomic variation around the world—say, which populations seem to be genetically close to one another, and which are far apart—we can say how these relationships came about. For example, we can develop quantitative models of human migration and see which models give the closest fit to the observed pattern of genomic relationships. In this third endeavor, geneticists incorporate knowledge from paleobiology and linguistics as aids in confirming inferences about how today's patterns of variation arose. It is this pattern of worldwide genomic variation that will be our main focus in this piece,

although we shall allude to the issues of associations between DNA and phenotypes and signals of natural selection at the end of this essay.

PHENOTYPES AND GENOTYPES, CONTINENTS AND POPULATIONS

It has been known for centuries that certain physical features of humans are concentrated within families: hair, eye, and skin color; height; inability to digest milk; curliness of hair; and so on. These phenotypes also show obvious variation among people from different continents. Indeed, skin color, facial shape, and hair form are examples of phenotypes whose variation among populations from different geographical regions is noticeable. No one has any difficulty seeing visible physical differences between a random person chosen from West Africa, from China, from Norway, or from the tropical rainforest of the Orinoco basin. Some classic blood group polymorphisms are also geographically differentiated, although most are not (Cavalli-Sforza et al. 1994). All of the above-mentioned phenotypic variation is largely due to variation in genes, although the specific parts of the DNA responsible for this variation are just beginning to be discovered and the genes responsible are only now beginning to be located in the genome.

Over the course of history, the typological and geographic notions of race have become merged because of the observation that continental regions are characterized by some phenotypic differences of color, facial features, or hair form. Of course, variation in these traits also occurs among individuals whose ancestors come from the same continent—and who thus might be considered to be of the same "race"—which means that categorization according to these phenotypes can be problematic (Brown and Armelagos 2001). Nevertheless, many people are interested in knowing what fraction of all of the variation in the DNA is of the kind that produces the typological distinctions most commonly used to assign racial categories.

The technology for genotyping DNA at the individual human level—and thus our ability to answer some of our questions about human genetic differentiation and continental origin—has advanced rapidly over the past ten years, and the cost has also begun to decline. This progress has been driven by the search for statistical association between one or a few SNPs and disease phenotypes; such studies are now called "whole genome association studies" (WGAS), although they do not involve the whole (three billion nucleotides) genome but up to a million

SNPs. An important by-product of the improved technology and lower price of genotyping is an explosion of knowledge about variation in the DNAs of individuals sampled from different populations around the world.

The above-mentioned HapMap project has represented one approach to producing a picture of the world's genomic variation, but since it involves samples from only four of the world's populations, HapMap cannot claim to be very comprehensive. There is, however, another approach whose aim has been to produce a much broader picture of the world's genomic variation. This is the human genome diversity project (HGDP). HGDP was begun in the early 1990s, at the impetus of the population geneticist L. L. Cavalli-Sforza, with the ambitious objective of sampling about twenty-five people from each spoken language on Earth—perhaps five thousand to six thousand populations. The idea was to ask each population to donate blood samples and convert these samples into cell lines, each of which could, in principle, provide a near inexhaustible source of DNA for current and future research. A combination of political and financial limitations led to the restriction of HGDP to a collection of 1,064 cell lines representing fifty-two sampled populations. These cell lines were provided by researchers who had them in their laboratories as a result of previous collections of blood from donors who had participated in earlier projects. An organization in Paris, France, the Centre Etude Polymorphism Humaine (CEPH), became the repository for these 1,064 cell lines, and the provider of DNA from them to qualified investigators (Cann et al. 2002; Cavalli-Sforza 2005). The collection is now called the CEPH-HGDP panel, and although it is seriously deficient in representation of indigenous (and other) populations from many parts of the world, it represents the broadest and best-documented sampling of human populations available today. An important component of the process that culminated in the CEPH-HGDP panel was an investigation of the ethical protocols that should characterize any such collection, taking account of different tribal authority systems and avoiding exploitation of indigenous blood donors. This document, by Hank Greely, a Stanford law professor, was published in the *Houston Law Review* and can be accessed online at www.stanford.edu/group/morrinst/hgdp/protocol.html.

It should be noted that there is an important difference between "populations" as used in this chapter and what are generally referred to as "races." "Population" is a term used by biologists to refer to a circumscribed group of people identified by geographical and/or linguistic

characters that differentiate them from another population. Depending on how fine we draw the distinctions, one could say that there are as many as 6,000 populations as identified by distinct languages. "Races," on the other hand, is an imprecise term generally used to refer to groups of people who differ in some way (physically or culturally) from one's own group. Only a small fraction of all of the variation in the DNA is of the kind that produces the typological distinctions most commonly used to assign racial categories. What this means is that while "race" can be a useful term in the study of societal dynamics, it is not, as the studies below suggest, helpful in a discussion of human genomic variation. This is addressed in more detail below.

BETWEEN- AND WITHIN-POPULATION VARIATION

Ever since the human blood groups were discovered, geneticists have accumulated data on how the frequencies of the genes determining these blood groups varied within and among different human populations. This interest arose partly from issues at the individual level, such as paternity testing and avoiding incompatible transfusions, but also from issues at the population level, such as the necessity of maintaining correct inventories at blood banks. The application of biochemical techniques to detect variation in enzymes, first used in 1966, led to research on patterns of variation in the genes coding for such proteins in many species, including humans.

The first attempt to statistically partition variation in human blood group genes and enzymes into between- and within-group components suggested that about 85 percent of the variation occurred *within* populations and at most 11 percent was *between* what we usually refer to as "races" (Lewontin 1972). It took twenty-five years until DNA-level variation, in STRs, was able to augment the earlier blood group and enzyme data to a total of 109 variable sites in the DNA. The statistical result hardly changed; about 85 percent of the variation in these genes resided within the fourteen populations studied (Barbujani et al. 1997). The availability of the CEPH-HGDP panel with its fifty-two populations has massively expanded these initial geographically-limited sets of data.

The first statistical study of the CEPH-HGDP panel of fifty-two populations used 377 STRs and estimated that about 94 percent of the variation in frequencies of the STR alleles was within populations and

only 3 percent to 4 percent between populations whose origins lie on different continents (Rosenberg et al. 2002). A recent assessment of variation at 650,000 single nucleotides within and between the CEPH-HGDP populations gave an estimate that 89 percent of the variation was within populations and about 9 percent between populations from different continents (Li et al. 2008). It follows that the genomic variation among people from the same population is about ten times that between people from different continents. Thus, the modern data confirm Lewontin's original conclusion that the overwhelming majority of genomic variation among humans occurs within populations.

People who have identified themselves as belonging to the same population are often indigenous to the same continent and might, therefore, be considered to be of the same "race." It is in the many cases of people whose ancestry derives from more than one continent that the arbitrariness of race and its social construction emerge. Thus, in the United States, a person with an African father and a European mother would probably be assigned by most people to a black race. This is a matter of historical precedent and social identification by the majority white population. It would be correct to say that such a person has ancestry on two continents (or more in many cases) and remove any racial designation which, as just mentioned, is equally European and African. In the same way, a racial designation of African American can be regarded as arbitrary because such a person can have ancestry of 1 percent or 90 percent in either Europe or Africa. It is in such cases that the social designation of race today usually trumps the much more complicated but more exact notion of genetic ancestry.

Some of the DNA differences that produce the 9 percent of between-continent variance can be assumed to be associated with genes whose functions produce the differences in skin color, hair texture, facial shape, and height of people from different geographical regions and whose detection does not require sophisticated biochemical tools. We can conclude from this that only a small fraction of all genomic variation is responsible for the differences that are commonly used to assign people to different races.

GEOGRAPHY AND LEVEL OF HETEROZYGOSITY

Although the fraction of genomic variation that is attributable to between-continent differences is very small, we can still ask whether this fraction contains enough information to tell us something about human evolution

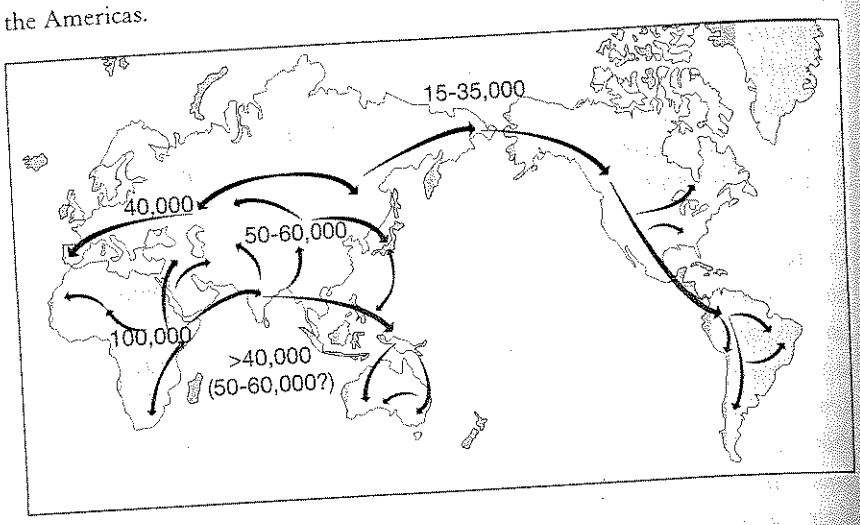
or even about the historical relationships between different populations of humans. In fact, it is possible to make a number of such inferences. The principle is one that is well known and widely exploited in the social and behavioral sciences: namely, that if each of the variables to be studied contributes a tiny fraction to the overall variation in some measure (or measures), then with a sufficient number of such variables a pattern in the variation of the measure may be detectable. In the genetic case, if most of the 650,000 positions on the DNA show small differences between continents, correlation between and accumulation of these differences combine to allow important inferences to be made.

A natural first step in the analysis of such large sets of STR (Ramachandran et al. 2005) or SNP data is to compare the amount of genomic variation among all fifty-two populations in the HGDP-CEPH panel. One measure of this variation is the level of the heterozygosity. At a particular position in the DNA, if the contribution that an individual received from his or her mother differs from that given by the father, then that individual is a heterozygote and the average over all tested positions in the DNA of the fraction of heterozygotes in all individuals in the population is that population's heterozygosity. For both STRs and SNPs, African populations have the highest heterozygosity while indigenous American populations have the lowest. Moreover, the plot of the heterozygosity of each population against geographical distance from Addis Ababa, in Ethiopia, shows a steady decline all the way to the Americas (Ramachandran et al. 2005). In other words, populations are less genetically variable the farther they are from Africa.

Population genetic theory provides the explanation for this steady loss of within-population variation with distance from Africa: Africa has the most genomic variation, and the first small group of humans that ventured out of Africa and colonized either Asia or Europe represented a small sample of the original variation. Then the next group of migrants going south or east was a sample of that first sample and had even less within-population variation. This continued loss of heterozygosity as a function of the continued sampling of earlier samples is called the serial founder effect. The observed pattern of decline in heterozygosity provides strong genetic evidence that all modern humans originated in Africa. Figure 3.1 illustrates the consensus view among population geneticists of the migration of modern humans over the past 100,000 years. Undoubtedly the dates in Figure 3.1 will be made more precise as more research is carried out on the CEPH-HGPD panel and more population samples from around the world become available.

FIGURE 3.1 | **THE MIGRATION OF MODERN HUMANS.**

This map shows the likely routes and approximate dates of the major migrations of modern humans that began from East Africa, with the most recent arrivals in the Americas.



SOURCE | Cavalli-Sforza and Feldman 2003.

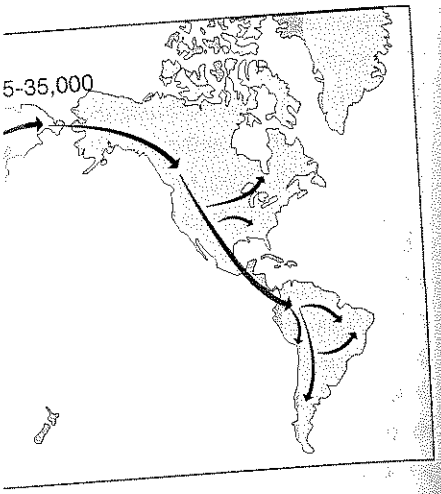
CONTINENTAL ANCESTRY

The explosion of data concerning human genomic variation has produced new statistical and computational methods that help us understand these data. Two methods have been particularly useful in addressing questions concerning the continental ancestry of different ancestral groups and of the individuals who form those groups. These are STRUCTURE, a statistical algorithm devised by Pritchard et al. (2000), and FRAPPE, which was developed by Tang and colleagues (Tang et al. 2005a, 2006). This is not the appropriate place to present the statistical details that underpin these methods. Suffice it to say that both methods produce for each individual in the study an estimate of the probability that a randomly chosen STR or SNP from that individual originates from one of a set of ancestral groups. The number of such ancestral groups, K, is chosen to produce the (statistically) best estimate of these probabilities, which are averaged over all STRs or SNPs to assign a membership coefficient, namely, a fraction of each individual's ancestry, to one of the ancestral groups. In principle, each person in the study could have part of his or her ancestry in many groups of populations, and the membership coefficient represents the fraction of his or her genome that has ancestry in each of these groups.

ty

IUMANS.

oximate dates of the major migrations
Africa, with the most recent arrivals in



TRY

human genomic variation has produced
methods that help us understand these
ticularly useful in addressing questions
try of different ancestral groups and of
groups. These are STRUCTURE, a sta-
chard et al. (2000), and FRAPPÉ, which
eagues (Tang et al. 2005a, 2006). This is
sent the statistical details that underpin
hat both methods produce for each indi-
of the probability that a randomly chosen
ial originates from one of a set of ancestral
cestral groups, K, is chosen to produce the
ese probabilities, which are averaged over
membership coefficient, namely, a fraction
one of the ancestral groups. In principle,
l have part of his or her ancestry in many
ne membership coefficient represents the
hat has ancestry in each of these groups.

It is important to emphasize that when researchers do this analysis, the ancestral groups are not specified in advance and the population memberships of all individuals are removed prior to the analysis. Only the genetic information enters this analysis. Graphical representation of the pattern of membership coefficients in each group makes use of a program, DISTRUCT (Rosenberg 2004), that uses colors to represent the fractions of membership in each group. The population from which each individual actually comes is noted after the statistical analysis has been carried out. This method of analysis allows us to graphically represent, using bars of various shades of gray, genomic variation as it relates to the ancestral origin of populations from differing geographic regions of the world.

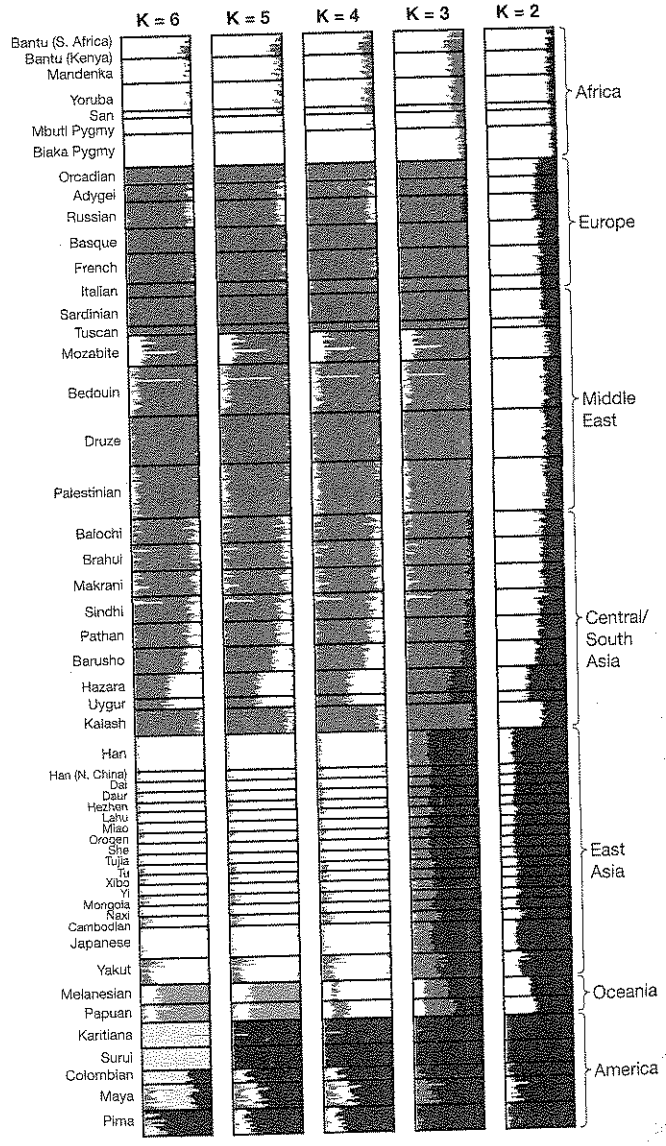
The results of two recent studies on the genomic variation among populations who have been sampled by the CEPH-HGDP panel are illustrated in Figure 3.2. The membership coefficients shown in Figure 3.2a are inferred using STRUCTURE on 783 STR polymorphisms (Rosenberg et al. 2005), while Figure 3.2b is based on the 642,690 autosomal SNPs from the study by Li et al. (2008), analyzed using the program FRAPPÉ. The similarities between the two sets of results are important. With $K = 5$ groups in Figure 3.2a, the populations from Africa, Eurasia, East Asia, Oceania, and the Americas form statistically separable continental clusters. With $K = 6$ groups, the Surui, a small indigenous Brazilian population, forms its own cluster, as do the Kalash. With the vastly increased number of polymorphisms in Figure 3.2b, the huge Eurasian "supercontinent" west of the Himalayas separates into three statistically recognizable subgroups: Middle Eastern populations, European populations, and Central-South Asian populations (Note: all sampled populations in this last group are in Pakistan; there are no populations from India in the HGDP panel). Because of the long history of migration and colonization across Eurasia, resolution of the CEPH-HGDP population samples from this part of the world required much more data than was provided by just the STRs.

Thus, although only 9 percent of the genomic variation in the SNP study is attributable to differences between large geographic regions (i.e., continents), the huge amount of data allows multivariate statistical analysis to produce from this 9 percent the regional clusters represented by the colored bars in Figure 3.2. The clusters contain within them individuals who are genetically more similar than they are to members of other clusters; the large amount of data makes this clustering statistically detectable.

Geographic clustering and spatial gradients are features of these large data sets of human genomic variation. Between pairs of populations

FIGURE 3.2a | **REPRESENTATION OF ANCESTRY IN THE CEPH-HGDP SAMPLE OF 1,056 INDIVIDUALS FROM 52 POPULATIONS AROUND THE WORLD.**

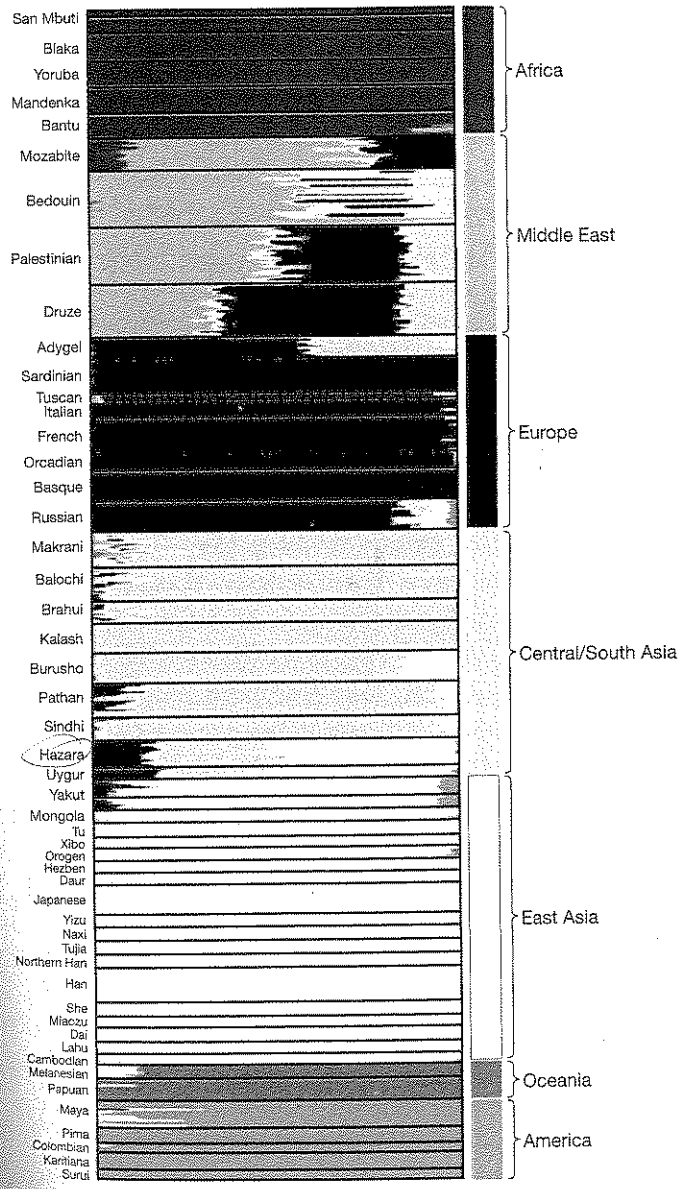
Each individual is represented by a vertical line, and the location of each person's ancestry is shown with different levels of shading. The K values 2 through 6 refer to the number of regions assumed to encompass the possible range of ancestries. The figure is based on genetic variation at 783 STR (short tandem repeats) in the DNAs of the sampled people. Many people have more than one level of shading, indicating ancestry in more than one place.



SOURCE | Rosenberg et al. 2005.

FIGURE 3.2b | **REPRESENTATION OF ANCESTRY IN 938 PEOPLE FROM THE CEPH-HGDP PANEL GENOTYPED AT 642,690 POSITIONS ON 22 CHROMOSOMES.**

With this large number of polymorphisms, seven regions are identified as sources of ancestry for these 938 people. Note that many have ancestry in three regions, which indicates the geographic origins of the ancestors of the people sampled.

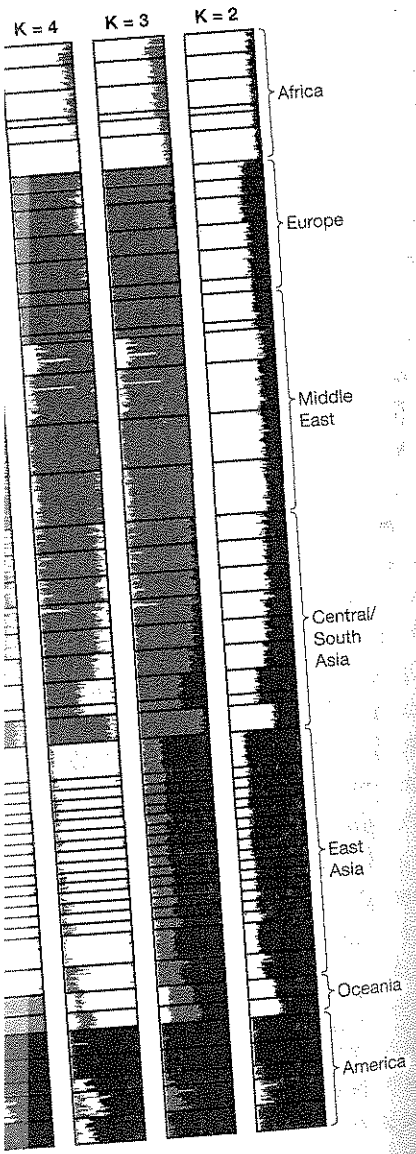


SOURCE | Li et al. 2008.

.city

RY IN THE CEPH-HGDP SAMPLE OF 1,056 .ATIONS AROUND THE WORLD.

cal line, and the location of each person's f shading. The K values 2 through 6 refer compass the possible range of ancestries. at 783 STR (short tandem repeats) in the ople have more than one level of shading, ace.



from the same cluster, genetic distance increases as geographic distance increases. But for pairs of populations from different clusters, genetic distance is generally larger than that between pairs of populations from the same clusters that are separated by the same geographic distance. This suggests that major geographic barriers, such as oceans, deserts, or mountain ranges, result in genetic discontinuities because they prevent migration from being geographically uniform. It is conceivable that some details of Figures 3.2a and 3.2b will have to be modified as more population samples expand the CEPH-HGDP panel. But it is unlikely that our general picture of the migration out of Africa to the rest of the world will change.

ADMIXTURE, MIGRATION, AND ASSIGNMENT

In Figures 3.2a and 3.2b, we see that members of many populations have ancestry in two or more continental regions. This is particularly true of the Middle Eastern populations whose ancestry includes components from Europe and Pakistan; the Mozabite population who have ancestry in Europe, Africa, and the Middle East; and populations from Pakistan and North West China. The Hazara are an ethnic group from the Hazarajat region of Afghanistan, many of whom now live in Pakistan, which is where this sample was taken. Three regions—Europe, East Asia, and Pakistan—are represented in the genomic ancestry of the Hazara. The same is true of the Uygur from Xinjiang province in China. They descend from groups who speak Turkic, Indo-European, and Mongol languages and show the same pattern of ancestry in three regions as the Hazara, although the historical events that have produced this pattern are different for these two ethnic groups.

The multiple ancestries of the Middle East on populations that we see in Figure 3.2b undoubtedly reflect that region's importance as a migration corridor between Europe and Asia. Also, the Siberian group, the Yakuts, shares common ancestry with the indigenous American populations even though the latter were sampled south of the United States, and this shared ancestry most likely reflects the migration across the Bering bridge by the Siberian founders of the indigenous American peoples.

Most people today would not refer to a "Middle Eastern race," even though the pattern of Middle Eastern ancestries in Figure 3.2b differs from all others. This phenomenon of multiple ancestry is referred to as "admixture," and its extent in any population reflects that population's history of subjugation, colonization, invasion, and even trade. The fact

that a population is admixed does not preclude its assignment to a cluster, even a cluster of which it is the sole member, provided enough data are available.

Indeed, even in the population of the United States, with the use of sufficient DNA polymorphisms, clusters of African Americans, Hispanic Americans, Asian Americans, and European Americans are clearly identifiable (Tang et al. 2005a, 2006). The first three of these obviously have ancestry in two or more of the colored blocks representing continental origin in Figure 3.2, but within the U.S. population they form their own statistically identifiable clusters. This suggests that the term "race," with its negative baggage, could be replaced by a less pejorative and more accurate term "ancestry group." This term would apply equally well to any of the clusters in Figure 3.2, including those which, like the Middle Eastern bar, comprise individuals with ancestry from three continental regions. And without any pejorative connotation, it could be applied to a sample of individuals from the United States, who if studied in a worldwide context, would have ancestries in multiple clusters of Figure 3.2, or if studied only in the context of the United States, would form separate statistically detectable ancestry groups.

The statistical analyses that produced Figures 3.2a and 3.2b show that given an individual's genotype and enough worldwide genomic data, he or she can be assigned to specific genetic clusters with a high degree of certainty. So we can pinpoint a person's geographical origin (or origins, in the case of admixture) by knowing his or her genotype. However, it is very important to note that the reverse is not the case. Merely knowing a person's geographical origin does not give us enough information to predict his or her genotype. This is so because, as in the example discussed above with people from Africa, the majority of genomic variation occurs within, rather than between, populations.

GEOGRAPHIC PATTERNS OF NATURAL SELECTION

Natural selection occurs when some individuals in a population have a higher likelihood of surviving to adulthood, or reproducing, or finding mates with whom to reproduce because they carry some phenotype that others in their population do not. We then say that this phenotype is under natural selection, and because of its greater chance of being represented in future generations, the phenotype will increase in frequency over time. In the same way, a phenotype that has a disadvantage will decrease in frequency and would be expected to eventually be rare in

the population. Thus the frequency of a phenotype on a population is expected to change under evolution by natural selection.

Human phenotypes emerge from a complex interaction between genotypes and environments, the latter including gestational, early developmental, educational and other environments that might be expected to affect a phenotype. To the extent that human phenotypic differences have a genotypic contribution, natural selection on phenotypes will be reflected in changes in genotype frequencies. The observation that individuals vary in both phenotypes and genotypes raises the question of whether it is possible to infer whether natural selection has produced the current pattern of variation, and if so, what was the environmental agent (or agents) that caused the natural selection. The availability of large genomic data sets and rapid computational methods for their analysis has opened new avenues in the search for evidence of selection.

The classical case of natural selection on a single human gene concerns variation in the hemoglobin gene. The most common hemoglobin allele has undergone a number of small changes that result in an altered hemoglobin molecule. The best known of these mutations of the normal hemoglobin A is called hemoglobin S. In the malarial environments of West Africa, southern Europe, the Middle East, and India, individuals of genotype AS have a selective advantage over those with the normal genotype AA (as well as those who are severely anemic and have genotype SS) due to the superior ability of blood cells in AS individuals to resist invasion by the malarial parasite. The result is that in areas where malaria is endemic, A and S types are both prevalent. Without malaria, S would be very rare. This is a case of natural selection that favors the AS genotype, thereby maintaining both A and S forms in malarial environments. More on the biology and population genetics of this system can be found in any human genetics text.

Now that we have the millions of SNPs from the four HapMap populations and the 650,000 SNPs from the fifty-two CEPH-HGDP populations, can we detect signals that part of the genome has responded to selection in one or more populations or continental regions? And if a statistical signal is detectable in the genome, can we identify the phenotypes that have been subject to this natural selection and the environmental agent(s) that are likely to be responsible for the selection?

A number of methods have been used to detect a signal of selection that favors a particular part of the genome, that is, positive selection, or that keeps parts of the genome related to deleterious phenotypes at low frequencies. One method, suggested originally by Cavalli-Sforza (1966),

y of a phenotype on a population is
ı by natural selection.
ı a complex interaction between gen-
er including gestational, early devel-
nvironments that might be expected
ıt that human phenotypic differences
tural selection on phenotypes will be
frequencies. The observation that in-
s and genotypes raises the question of
ther natural selection has produced the
f so, what was the environmental agent
l selection. The availability of large ge-
ıtational methods for their analysis has
h for evidence of selection.
lection on a single human gene concerns
ne. The most common hemoglobin al-
f small changes that result in an altered
known of these mutations of the normal
lobin S. In the malarial environments of
, the Middle East, and India, individuals
e advantage over those with the normal
who are severely anemic and have geno-
bility of blood cells in AS individuals to
parasite. The result is that in areas where
ypes are both prevalent. Without malaria,
case of natural selection that favors the AS
g both A and S forms in malarial environ-
ıd population genetics of this system can be
text.
ions of SNPs from the four HapMap popu-
's from the fifty-two CEPH-HGDP popu-
s that part of the genome has responded to
ulations or continental regions? And if a sta-
the genome, can we identify the phenotypes
ıis natural selection and the environmental
responsible for the selection?
ave been used to detect a signal of selection
t of the genome, that is, positive selection, or
ome related to deleterious phenotypes at low
suggested originally by Cavalli-Sforza (1966),

examines the degree of differentiation among the different populations for each genetic variant—for example, all the microsatellites or SNPs in a study. The idea is that all parts of the genome should be equally affected by any demographic forces, such as migration or changes in population size. Hence, if one DNA segment gives a statistical signal of differentiation that is either too high or too low relative to the distribution of that signal across all DNA segments, this extreme value may be due to selection.

Barreiro et al. (2008) examined 2.8 million SNPs assayed in the four HapMap project samples using a standard measure of population differentiation for individual SNPs. They found many signals of selection against SNPs located in genes known to be related to diseases. They also found evidence of strong positive selection for SNPs associated with genes for skin color, olfaction, immune response to pathogens, as well as other phenotypes. An earlier study by Voight et al. (2006) of HapMap data had also detected positive selection on these classes of genes as well as others related to spermatogenesis and fertilization. Over the past few years, genomic analyses have repeatedly confirmed that genes involved in skin pigmentation show strong signs of positive selection in Europeans for the lighter colored pigmentation alleles. The suggestion is that in the European environment with reduced solar radiation, lighter skin probably had a selective advantage.

A fascinating example of natural selection in both Europeans and Africans has emerged from analyses of SNPs associated with the enzyme lactase, which controls the ability of adults to digest milk. Lactase persistence, which enables milk to be digested into adulthood, had previously been thought to be a European trait. Indeed, there is a SNP associated with lactase persistence that shows strong signs of selection in Europeans (Enattah et al. 2002, 2007). However, a recent study by Tishkoff et al. (2007) identified a different SNP that is associated with tolerance of milk in some Kenyan and Tanzanian populations. In Europeans and these African populations, the selected allele apparently spread rapidly along with the culturally transmitted practice of pastoralism and drinking milk. These examples of recent analyses of selection at the genomic level illustrate that at least a small part of the variation in the human genome can be attributed to environmental effects and/or cultural practices.

SELF-IDENTIFICATION AND GENETIC ANCESTRY

Admixture (mixed ancestry) in a population is known to produce spurious associations between a phenotype and genes. This can be important

in epidemiological studies where cases of a disease are matched with healthy controls, and large numbers of DNA variants, usually SNPs, are tested to see if they differ in frequencies between the two phenotypic samples. Especially in admixed populations, such as ethnic minorities in the United States, it is important to assess the extent of genetic substructure that may affect estimates of association of genes with chronic complex diseases. Such substructure may occur if the population from which the cases and/or controls are taken is stratified in any way—for example, by socioeconomic status, routine use of different diets or exercise regimes, or ancestry from different parts of the world. For this reason it is important to ask whether self-reported population ancestry is concordant with that determined by genetic studies.

From their analysis of the CEPH-HGDP panel, Rosenberg et al. (2002) pointed out that "self-reported ancestry likely provides a suitable proxy for genetic ancestry." These authors worried, however, that in recently admixed populations variation in genetic ancestry might correlate with risk of disease as a consequence of genetic or cultural factors. Thus, in such recently admixed groups as the ethnic minorities in the United States, it is important to ascertain how well self-identified race or ethnicity matches the genetic structure of these populations when the latter is estimated from a larger number of markers.

An important attempt to answer this question was made by Tang et al. (2005b), who genotyped 326 STRs on 3,636 people of varying race or ethnicity sampled from fifteen locations in the United States and Taiwan. Subjects identified themselves as being white non-Hispanic, black non-Hispanic, Hispanic, Chinese, and Japanese. As in the study by Rosenberg et al. (2002), the statistical determination of genetic clusters was done by researchers who had knowledge of how the people being studied self-identified. Using STRUCTURE (Pritchard et al. 2000), four genetic clusters were identified, and these corresponded almost perfectly to the self-identified categories with one cluster containing the Chinese and Japanese subjects; only five of the people sampled were differentially classified. The results, taken from Tang et al. (2005b), are presented in Table 3.1.

Two of the conclusions reached by Tang et al. should be emphasized. The first is that the genetic clustering in Table 3.1 reflects ancient geographic ancestry rather than the geography of current residence. Second, in the absence of self-identified ethnic labels, it is dangerous to attribute phenotypic differences—for example, in disease incidence

TABLE 3.1 | GENETIC CLUSTER ANALYSIS VERSUS SELF-IDENTIFIED ETHNICITY OF 3,636 PEOPLE

Self-Identified Ethnicity	Number of Subjects in Genetic Cluster			
	A	B	C	D
White non-Hispanic	1,348	0	0	1
Black non-Hispanic	3	0	1,305	0
Hispanic	1	0	0	411
Chinese	0	407	0	0
Japanese	0	160	0	0

SOURCE | After Tang et al. 2005b.

or life expectancy—between genetically defined clusters as being due to genes because these differences may actually be due to environmental, cultural, or behavioral factors.

GENES, DISEASE, AND ANCESTRY

Since the sequencing of the human genome and the subsequent development of technologies that allow detection of hundreds of thousands or millions of SNPs, hundreds of studies have attempted to find associations between SNPs and disease phenotypes. From the position of a SNP found to be associated with a disease, the genes in the neighborhood of that SNP can be located from one of the public databases. To the extent that the nature and amount of the products of these genes are known, the biochemical causes of the disease might then be revealed. Such studies have focused on complex diseases, that is, those for which no single gene is known to cause the disease. Important examples include hypertension, coronary artery disease, rheumatoid arthritis, type 2 diabetes, and behavioral disorders such as bipolar disease, autism, and schizophrenia.

The largest study of this kind was carried out in the United Kingdom and involved 2,000 cases of each of seven major diseases and a shared set of 3,000 controls. All subjects were of European ancestry (Wellcome Trust Case Control Consortium 2007). One example of the many interesting associations found was the SNP on chromosome 9 that showed the highest signal of association with coronary artery disease.

We can ask whether a SNP found to be associated with coronary artery disease in Europeans (i.e., British) is also a risk factor for people

of other ancestries. In other words, can we extrapolate from a genome-wide association study of people with ancestry on one continent to those with ancestry on another continent? An elegant study by McPherson et al. (2007) shows that the answer is no. These investigators in a very carefully replicated study found an association between two SNPs and chronic artery disease in people of European ancestry. These were located very close to the part of the genome identified in the British study. When these same SNPs were assessed in the HapMap sample of Yoruba people (the Yoruba are a large ethnic group from Nigeria), the alleles strongly associated with chronic artery disease in Europeans were either absent or very rare. Further, the two SNPs which were significant risk factors in European Americans were not associated with the disease in African Americans from two different samples. The conclusion is inescapable: alleles associated with risk factors for this disease have widely different frequencies in different ancestry groups.

The amount of RNA produced by a gene can be measured and treated as a continuously varying phenotype for each gene and is called the "expression level" for that gene. Variation in the expression level of a gene can be due to difference in other parts of the genome called "regulators." Spielman et al. (2007) examined expression levels in 4,197 genes in people of European and East Asian (Chinese and Japanese) ancestry and found significant differences between the expression levels in the two groups for 1,097 genes. This raises the possibility that other phenotypic differences between populations of different ancestries could also be due to variation in the regulatory parts of the genome. It is unlikely that such phenotypic differences will be large. Indeed it is worth quoting a major conclusion of the British study referred to above:

> The novel variants we have uncovered are characterized by modest effect size . . . and even these estimates are likely to be inflated. . . . The observed distribution of effect sizes is consistent with models based on theoretical considerations, and empirical data from animal models that suggest that, for any given trait there will be few (if any) large effects, a handful of modest effects, and a substantial number of genes generating small or very small increases in disease risk. (674–675)

It seems reasonable, then, to assume that the differences between people who belong to different ancestry groups in the added risk of disease from such genes is also likely to be modest, small, or very small.

CONCLUSION

After reviewing the available data pertaining to human genetic diversity and asking how these data relate to the definition and meaning of racial classification, we have shown that people can be assigned to ancestry clusters that usually agree with their self-identified ethnicity. We have also shown that only a small fraction of all genomic variation is responsible for the visible morphological differences such as skin color, facial features, or hair form that are commonly used to assign people to different races. Moreover, that small fraction of genomic variation has nothing to do with supposed racial differences related to intelligence, moral character, or tendency to criminality. For that reason, we suggest that a more accurate name for such genetically defined clusters—rather than race or ethnicity—might be "ancestry groups."

While most genomic variation occurs within ancestry groups, an examination of the pattern of differences within and between ancestry groups can reveal those parts of the genome that have been subject to selection in different parts of the world. Admixture and mixed ancestry can generate spurious correlations between genes and disease, so proper development of control groups for epidemiological studies should take continental ancestry into account. However, as pointed out by Epstein (2006), "genetic risk assessment will attain sufficient predictive power to be of use if—and only if— analyses of many loci are combined with evaluations of non-genetic lifestyle and environmental factors" (436). There is potential for lines of ancestry to provide medically useful information. But because phenotypes are never attributable to genes alone, for diagnosis and treatment, individual properties—both genetic and environmental—are likely to provide the most useful information.

Works Cited

Achilli, A., A. Olivieri, M. Pala, E. Metspalu, Simona Fornarino, V. Battaglia, M. Accetturo, I. Kutuev, E. Khusnutdinova, E. Pennarun, N. Cerutti, C. Di Gaetano, F. Crobu, D. Palli, G. Matullo, A.S. Santachiara-Benerecetti, L.L. Cavalli-Sforza, O. Semino, R. Villems, H.-J. Bandelt, A. Piazza, and A. Torroni. 2007. Mitochondrial DNA variation of modern Tuscans supports the Near Eastern origin of Etruscans. *Am. J. Hum. Genet.* **80**: 759–768.

Barbujani, G., A. Magagni, E. Minch, and L. L. Cavalli-Sforza. 1997. An apportionment of human DNA diversity. *Proc. Natl. Acad. Sci. USA* **94**: 4516–4519.

4

Which Differences Make a Difference?

Race, DNA, and Health

Barbara A. Koenig

This essay considers the implications of new findings in molecular genetics for our understanding of the meaning of race, particularly within the domain of biomedicine. New methods of studying the full complexity of genetic ancestry in human populations—scientific practices that pinpoint millions of points of variation in the DNA—are employed in the study of disease etiology, in treatment, and in explaining persistent health inequalities between minority and majority groups in the United States. Although it is well-known that illnesses cluster, the exact reasons are often unclear: Is it genetics, shared social environment, or some combination? Less well-examined are the consequences of searching for the roots of health disparities utilizing new genomic tools. This essay argues that genomics may inadvertently reinstate a biological understanding of human racial categorization, replacing the hard-won scientific compromise that differences among human groups are primarily cultural in nature. Ironically, well-intentioned scientific practices for "doing race" through genetic studies may obscure the health implications of how race is "lived."

> *New York, perhaps more than any other big city, harbors all the ingredients for a continued epidemic. It has large numbers of the poor and obese who are at higher risk. It has a growing population of Latinos, who get the disease in disproportionate numbers, and of Asians, who can develop it at much lower weights than* people of other races.
>
> NEW YORK TIMES, JANUARY 9, 2006 (EMPHASIS ADDED)

Differences
Difference?
IA, and Health

Koenig

of new findings in molecular genetics for our
e, particularly within the domain of biomedi-
full complexity of genetic ancestry in human
t pinpoint millions of points of variation in the
disease etiology, in treatment, and in explaining
en minority and majority groups in the United
hat illnesses cluster, the exact reasons are often
environment, or some combination? Less well-
earching for the roots of health disparities utiliz-
rgues that genomics may inadvertently reinstate
an racial categorization, replacing the hard-won
ces among human groups are primarily cultural
med scientific practices for "doing race" through
alth implications of how race is "lived."

haps more than any other big city, harbors all the
a continued epidemic. It has large numbers of the
who are at higher risk. It has a growing popula-
, who get the disease in disproportionate numbers,
who can develop it at much lower weights than
er races.

MES, JANUARY 9, 2006 (EMPHASIS ADDED)

THE STATEMENT ABOVE APPEARED UNDER THE HEADLINE "DIABETES and Its Awful Toll," in a series of articles alerting readers to an emerging public health threat. This quote includes several clues about which differences "matter" in health care and biomedicine. There are hints that disease is somehow associated with certain populations, and that rates of ill health may vary by race. Asians and Latinos are described as separate "races," yet without much explanation. Poverty is implicated as another factor, yet its relationship to race is not mentioned. We might ask: Why are such categories of difference used? What logic, what practices of classification undergird this method of engaging in health research? Will new genetic research tools affect health research, and as a consequence, change the ways race is talked about and understood? And what are the policy consequences?

WHAT'S AT STAKE: GENETICS AND HEALTH DISPARITIES

The United States' persistent health disparities are widely recognized. In 2003 the Institute of Medicine (IOM 2003) published a major report examining the state of minority health in the United States. The report documented differences in rates of common diseases like cancer or heart disease, as well as significant variations in mortality rates and overall life expectancy. Black–white differences are especially troubling; U.S. whites consistently outlive blacks. And these differences persist even among those with equal access to health insurance, such as those older than 65 who are covered by Medicare. The experts convened by the IOM cited a large body of empirical research that underscores the existence of these and other health disparities among U.S. groups.

For example, minorities are *less likely* than whites to be given appropriate heart medications or to undergo cardiac bypass surgery and are less likely than whites to receive kidney dialysis or transplants. By contrast, they are *more likely* to undergo certain less desirable procedures, such as lower limb amputations for diabetes. Eliminating the root cause of such persistent inequalities has become a shared national goal—and it is a laudable objective, though hard to realize.

Although lack of access to health services, and unequal treatment when care is sought, account for a significant portion of racial and ethnic health disparities, the completion of the Human Genome Project has

focused attention on *biological* (hence genetic) variation across human populations. The tools of the new genetics are increasingly aimed at targets like soaring rates of diabetes and obesity. Even complex behaviors like smoking are being investigated using genetic hypotheses (NCI 2009). Even without meaning to, this genetic research deflects attention away from well-documented social determinants of ill health.

For decades, research investigating the causes of addiction has used difference across human populations as a starting point for analysis. Addiction—a trait that inevitably combines biology and behavior—provides a useful example when considering the implications of genetics research targeted by race. The use of some illicit substances is more common in U.S. "racialized" population groups, and addiction is associated with differential health outcomes among these groups. For example, certain Native American communities have high rates of alcoholism and alcohol-related disease, and U.S. minority populations are addicted to nicotine at higher rates than middle-class whites. It is appealing to cite genetic explanations for these differences (Caron et al. 2005), since locating the cause in the body, at the level of the gene, emphasizes personal responsibility for health. In turn, genetic causes tend to privilege pharmaceutical solutions, rather than social strategies for harm reduction, such as smoking bans or tax increases (Dingel and Koenig 2008). Viewing addiction through a genomic lens can direct attention away from interventions designed to change the underlying social conditions that support higher rates of addiction in racialized minorities.

NIH CENTER FOR GENOMICS AND HEALTH DISPARITIES

In 2008, the U.S. National Institutes of Health (NIH) announced the formation of a special center to study the intersection of genomics and health disparities. Note that *genetic* research examines specific genes and their health consequences; in contrast, *genomic* approaches target an organism's full complement of genes acting together, which is called the genome. The center initially was to be named the NIH Intramural Center for Genomics and Health Disparities. Said its director, Charles Rotimi, "The priority of our center will be to understand how we can use the tools of genomics to address some of the issues we see with health disparities" (NIH 2008). In the United States, health disparities are measured, tracked, and compared using categories of race. Perhaps because of the political sensitivity of linking genes and racial disparities in health,

within months both the name and the mission of the new agency had changed. Now called the Center for Research on Genomics and Global Health (CRGGH), its research goal is:

> to facilitate a global understanding of the relationship between human genetic variation and population differences in disease distribution, with the ultimate goal of informing health disparities. Investigators in the CRGGH will develop genetic epidemiology models that will explore the patterns and determinants of common complex diseases in populations in the United States and other human populations around the world. (CRGGH 2009)

This framing avoids granting scientific privilege to U.S. racial categories or to a priori assumptions that health disparities are predominantly caused by biological differences, rather than by lack of access to health services or structural social inequalities (Sankar et al. 2004). Within NIH, the intersection of race and genetics is clearly understood to be politically sensitive territory.

LINKING GENES, DISEASE, AND POPULATION

NIH's decision to create a new center, and the initial choice to link health disparities firmly to genetics, is hardly novel. The idea that certain diseases occur in specific human groups appears self-evident and noncontroversial. Indeed, it has long been established that some diseases occur at greater frequency in populations that share what I will call in this essay "biogeographical ancestry": descent from a particular region of the world. Note that biogeographical ancestry does not map perfectly onto common understandings of race. The two concepts are not synonymous.

We often speak casually about the link between illness and race. The classic cases in which disease is linked to race stem from the study of Mendelian (single-gene) disorders by population geneticists. Rare diseases caused by a mutation in a single gene often cluster in certain human populations. Two kinds of *chance* events lead to clustering: (1) population "bottlenecks" happen when only a small number of ancestors form the basis of a larger population, for example, after an environmental disaster significantly reduces the population size; and (2) "random genetic drift" occurs over time when groups are isolated from each other because of geographic barriers like oceans or mountain ranges. Clusters may also

occur when a genetic variant confers a selective advantage. It is often difficult to ascertain whether random events or selective advantages have caused the differences in gene frequencies between different populations. For example, sickle cell anemia, an autosomal recessive genetic condition, is widely understood as a "black" disease in the United States, even though in reality it is distributed throughout the world (Wailoo 2001). Variations in the sickle cell gene are found throughout Africa, the Mediterranean, the Middle East, and South Asia—areas where falciparum malaria is common. Those with sickle cell often die young. But because of a selective advantage among those who carry one (but not two) copies of a single-gene mutation that affects the blood's oxygen carrying ability, the variant persists. Having one copy of the gene variant protects against malaria; having two copies leads to lethal sickle cell disease. Sickle cell anemia was the first disease to be fully understood at the molecular level: It is caused by a mutation in a single gene that codes for hemoglobin.

Likewise, Tay-Sachs disease is widely understood as a condition affecting those of Ashkenazi Jewish background (even though it also occurs in other groups, such as French Canadians), and cystic fibrosis is considered a disease of northern Europeans (even though it is actually found throughout the world). Thus the idea of what might be called racialized diseases is commonplace, even though it is scientifically inaccurate. The problem, however, was that initial public policies for sickle cell disease screening so intricately tied the disease to African origins that in some U.S. states only newborns identified (socially) as black were screened for the condition. This practice led to undetected cases and consequent harm for those with the diease who were not diagnosed early. Testing for cystic fibrosis provides another example. The condition is caused by any one of more than 1000 distinct mutations in a gene called CFTR. Mutation *patterns* vary by ancestry, meaning that commercial genetic tests are of greater or lesser efficacy depending on the geographical origin of the individual screened. The *frequency* of both the disease and mutation pattern varies by biogeographical ancestry, but few genetic conditions are specific to any human group. The racialization of disease is a *social* practice, a consequence of "doing race." Racialized diseases are created by linking genetic conditions to historically and socially constructed categories of group difference. The linkage of sickle cell disease to blackness is historically and socially contingent; it is not biological fact. Although genes control the expression of skin pigmentation, blackness cannot be found in biology.

Similarly, both Tay-Sachs disease and certain rare mutations in a gene that predispose to breast cancer risk were linked early on to

city

a selective advantage. It is often dif-
events or selective advantages have
ncies between different populations.
autosomal recessive genetic condi-
k" disease in the United States, even
roughout the world (Wailoo 2001).
found throughout Africa, the Medi-
outh Asia—areas where falciparum
kle cell often die young. But because
se who carry one (but not two) copies
cts the blood's oxygen carrying ability,
py of the gene variant protects against
to lethal sickle cell disease. Sickle cell
ully understood at the molecular level:
le gene that codes for hemoglobin.
videly understood as a condition affect-
kground (even though it also occurs in
adians), and cystic fibrosis is considered
ven though it is actually found through-
what might be called racialized diseases
s scientifically inaccurate. The problem,
olicies for sickle cell disease screening so
ican origins that in some U.S. states only
black were screened for the condition.
ases and consequent harm for those with
sed early. Testing for cystic fibrosis pro-
dition is caused by any one of more than
e called CFTR. Mutation *patterns* vary by
rcial genetic tests are of greater or lesser
raphical origin of the individual screened.
se and mutation pattern varies by biogeo-
netic conditions are specific to any human
sease is a *social* practice, a consequence of
es are created by linking genetic conditions
nstructed categories of group difference.
ase to blackness is historically and socially
fact. Although genes control the expression
ss cannot be found in biology.
; disease and certain rare mutations in
reast cancer risk were linked early on to

"Jewishness"—at least in part because many American Jews endorse scientific values and participate in research at high rates. The availability of blood specimens collected from Jewish volunteers enabled the early identification of several mutations found in higher frequency (but not exclusively) in those of Ashkenazi background, leading genetic testing companies to offer targeted testing for "Jewish mutations" (Parthasarathy 2007, 83). In the last decade, fundamental changes in the *practice* of genetic research—stimulated by the mapping of the full human genome—have further complicated our understanding of the links among genes, ancestry, and human disease.

A POST-GENOMIC MOMENT

One specific, unified message accompanied the official announcement of the completion of the Human Genome Project in 2000: human beings are essentially the same.[1] Once the full genome reference sequence was published, this central theme was widely proclaimed by scientists and politicians alike (Collins and Mansoura 2001; California Newsreel 2003). Human genetic sequences are 99.9 percent identical; of the 0.1 percent of the human genome that varies from person to person, only 3 to 10 percent of that variation is associated with geographic ancestry or "race" as classically understood (Koenig, Lee, and Richardson 2008). This message was nothing new. For decades, the finding that there is greater genetic variability "within groups" than "between groups" had generally been accepted as evidence that the human species is not divided into discrete races. Rather, genetic variation is continuous, or "clinal," the technical term used by population genetics researchers (Lewontin 1991; Serre and Pääbo 2004). Geneticists publicly interpreted these results as disproving a biological concept of race and voiced their hopes that such scientific findings might help counter racism. This unified message has been conveyed by professional associations (AAA 1998), major travelling exhibits mounted by science museums (AAA 2007), and popular television documentaries such as *The Human Family Tree*, aired by the National Geographic Society in 2009. Voice-overs in that program announce authoritatively, "We are all related. . . . We are basically identical." The notion that modern humans—despite widespread geographic

[1] On June 26, 2000, President Clinton noted, "I believe one of the great truths to emerge from this triumphant expedition inside the human genome is that in genetic terms, all human beings, regardless of race, are more than 99.9 percent the same" (NHGRI 2000).

migrations following their departure "out of Africa"—are 99.9 percent similar, is invoked throughout the program as the camera pans to an extraordinarily diverse-looking group of people attending a street fair in Queens, New York. With the completion of the Human Genome Project, many imagined a future in which "race would become obsolete as attention shifted from the body politics of skin color, hair texture, and eye shape to the molecular-level biopolitics of the gene" (Koenig, Lee, and Richardson 2008, 1).

In actuality, since the mapping of the human genome, research has focused on human genetic *variation* and *differences* among groups (Jakobsson et al. 2008). This focus has rekindled debates about the connection between genetic traits (at the level of DNA) and human "racial" difference (Wilson 2001). "Scholars are divided on the question of whether racial categorization is an appropriate means of organizing potentially useful genetic data or a pernicious reification of historically destructive typologies" (Lee et al. 2008, 404).

In the nineteenth century, ideas of human difference were deeply linked with the biological theories of the day (see this volume's introduction); most aspects of human behavior, including intelligence, were ranked according to a social Darwinist evolutionary schema. That approach changed profoundly at the turn of the twentieth century with Franz Boas, the creator of American anthropology. The work of Boas and his students uncoupled biological race from theories of human culture, according diverse populations equal status and setting the stage for an anti-racist cultural relativism that became dominant following the atrocities of World War II (Pierpont 2004).

Thus, in the second half of the twentieth century, prior to the Human Genome Project's completion, most historians, social scientists, and race theorists afforded biology little status in theories of human difference (Fredrickson 2002). Contrary to these expectations and hopes, postgenomic science has *focused* on difference, reviving the idea of racial categories as proxies for true biological differences (Lee, Mountain, and Koenig 2001). Population genetics research and work by anthropological geneticists has focused on that small 0.1 percent margin of difference among human groups, using the tools of genetic difference to investigate human origins and ancient movements of populations (Li et al. 2008). Although highly contested and disputed, these minute differences are used in forensic investigations to solve crimes or even to adjudicate claims about national origins (Travis 2009). In addition, human geneticists argue that the genome holds the key to medically significant biological differences

among human racial and ethnic populations. Increasingly, genetic variation among human populations—races, ethnicities, nationalities—is an object of keen biomedical interest (Bamshad et al. 2004). And ironically, this work—which reinstates the link between biology and lay concepts of race—is justified by the claim that it will lead to the elimination of racial and ethnic health disparities (Krieger 2005; Montoya 2007; Sankar 2006).

CAN GENETICS "PROVE" THE EXISTENCE OF RACE AND ITS LINK TO DISEASE?

"Race Is Seen as a Real Guide to Track Disease," announced a 2002 *New York Times* article, reporting on a new paper by Neil Risch et al. in *Genome Biology* (Wade 2002a). In that paper, Risch and colleagues argued that genetic differences among populations cluster into five major groups corresponding to "classical definition of races based on continental ancestry," and they boldly made the case for the "validity of racial/ethnic self-categorization" in genetic epidemiology research (Risch, Burchard, and Tang 2002). In challenging the seemingly unified chorus among scholars and scientists that race is a set of historically created and reinforced social ideas and practices, not rooted in human genes, the paper was a pivotal event in redirecting the emerging discourse on race and genetics. It was followed eight months later by a *New England Journal of Medicine* paper (Burchard et al. 2003) that not only advocated the use of race in human genetics research but framed its necessity in terms of the public policy goal of mitigating health disparities among racially identified populations. At the same time, other voices disputed the claim that genetic science had demonstrated the validity of traditional racial categories in studies of common illness, such as heart disease or diabetes (Cooper 2003).

The vigorous reassertion of the coupling of race and genes, once seen as antiquated, accompanies the shift of human genetic variation research into the genomic age. The new research revives old debates over the existence of a biological basis for race. But the new genetic race concept differs in important ways from its predecessors, as does the context of the debate.

THE DEVELOPMENT OF TOOLS TO INTERROGATE THE GENOME

Technologically, the genomic age institutes a shift from relatively limited gene-hunting research possible in Mendelian genetics to an era of

whole-genome analysis. Confronted with a vast pool of largely undifferentiated genomic data, researchers must find ways to make sense of it. Much of the new genomics is not hypothesis-driven. Instead, researchers query the human genome, seeking distinctions and patterns as leads for further research. The data derived from the human genome, like any large, multidimensional database, can be probed and organized in various ways. Because it is an important and highly salient social category, race (as understood and lived in the social world) has rapidly become a prominent "search tool" in scientific efforts to make sense of human genetic variation.

Intensive work of this sort has resulted in large-scale studies seeking patterns in human genetic variation across the world. Most prominently, a paper published in 2002 by Noah Rosenberg became the most cited article that year in the prestigious journal *Science* (Rosenberg et al. 2002). The study used a set of 377 genetic markers (microsatellites) that were assessed in 1,056 people. The 1,056 individual DNA samples were collected from 52 ethnically and linguistically diverse populations drawn from Africa, Asia, Europe, and the Americas. The paper graphically displayed the differences among human populations distributed across the world (also see Chapter 3). The team concluded that without using prior information about the origins of individuals, using a computer program called Structure, they were able to identify six genetic clusters. Five of these correspond to major geographic regions. Although the authors did not use the word "race" in the paper, their findings were widely reported in the popular media as demonstrating genetic *proof* of racial differences. In the *New York Times*, the science writer Nicholas Wade observed, "The issue of race and ethnicity has forced itself to biomedical researchers' attention because human populations have different patterns of disease" (Wade 2002b). Wade acknowledged that the authors themselves disavowed a racial interpretation of their findings, but he nonetheless invoked the term "race," linking it tightly to the needs of health research.

Slates of "neutral" genetic markers like those Rosenberg used are employed widely by researchers seeking associations between genes and disease occurrence. (These are variants that occur by chance, as discussed above.) If a disease is known to occur at different frequencies in groups of patients with varied biogeographical ancestry, then it is critical to correct for what is called "population stratification." Specifically, when searching for new associations between genetic mutations and disease, it is necessary to make sure that your genetic sample does not include DNA provided by different population groups that have different underlying frequency

rates of the disease under study. If the sample is stratified in this way, the results will be incorrect. This is a legitimate and scientifically important use of ancestry information and of panels of genetic "ancestry-informative" genetic markers. Unfortunately, the population clusters created by using these markers are often assumed to map directly onto lay categories of race, even though this is not technically accurate (Bamshad et al. 2004).

The technology continues to change and develop. Once a single reference human genome was available, researchers' next step was to create tools to account for the full spectrum of human diversity. One such tool was the human "haplotype map," which sought to determine the patterns in the distribution of genetic polymorphisms across the globe (International HapMap Consortium 2005). Although the haplotype map project focused on seeking human genetic diversity, its sampling methods (perhaps unsurprising, but very important) reflected conventional understandings of race. Sampling was not driven by a "bottom-up" examination of genetic variation around the world but by a "top-down" approach that employed specific, socially and historically contingent, racial categories. DNA samples were not collected at random from throughout the globe (a strategy called "grid sampling" that involves drawing a grid on a world map and sampling from all the boxes) but in three identified human "populations." The populations selected, Northern Europeans, Nigerians from a single African tribe, and Asians (Han Chinese and residents of Tokyo), correspond to conventional racial thinking. Almost as if working from the same script, the most recent effort to categorize human genetic variation, the 1000 Genomes project, uses similar categories of European, Asian, and African (1000 Genomes 2008). This massive research effort seeks to collect and make available the *full* sequence data of 1,000 humans, essentially repeating the Human Genome Project 1,000 times.

These databases have lay conceptions of race literally *built into them*. Once instituted, these data become a reference point for further research, as the information is analyzed and transformed for use by specialists in other fields, ranging from the "recreational" genetic testing pursued by amateur genealogists via the Internet to mainstream biomedical research. When these race-inscribed categories are used uncritically and outside the context in which they were derived, they may become naturalized, reified, institutionalized ways of conceptualizing the human genome. These practices carry serious implications for all subsequent human genome research and for our shared social understanding of the category of race itself.

GENETIC ANCESTRY TESTING

Market forces further link genetic technologies with categories of human difference, specifically race. Claims about racial background can allegedly be documented through the services provided by online genetic genealogy companies, which provide public access to inexpensive genetic testing via the Internet (Koenig, Lee, and Richardson 2008). Companies offer to provide evidence of Native American or African ancestry, links to certain tribes, or claims to Jewish priestly status. These services use ancestry-informative markers, the technology developed to study the scientific concept of population stratification (discussed previously) to provide customers with recreational estimates of the percentages of their ancestry that are European, African, or Asian.

Recreational genetics introduces new and challenging frontiers in racial identity formation, fostering the notion of a tight linkage between genes and race. These testing services, like any commercial venture, sell both a product and a desire for the product. Marketing literature is laced with the discourse of racial purity and racial mixture, as well as constructs such as blood, kinship, ancestry, and homeland. The implications are as yet unclear; genetic testing could either complicate notions of racial purity or build them up. As ancestry testing becomes cheaper and more widespread, new configurations of racial and national identity may emerge. At the policy level, genetic race verification services have potentially serious implications for community concepts of kinship and nationhood. In the case of entitlements that are tied to race, such as affirmative action or tribal membership, genetic ancestry testing may inflame long-standing debates about eligibility and the social recognition of race as a class. The technology of biological race verification will change the terms of debate and analysis.

BIOMEDICAL APPLICATIONS OF HUMAN GENETIC VARIATION: RESEARCH, RACE, AND PHARMACOGENOMICS

As noted, genetic research on race increasingly takes place in a medical context. Throughout the twentieth century, human population genetics was most closely associated with anthropological efforts to reconstruct the history of human migration. This research succeeded in offering impressive corroboration for the "out of Africa" hypothesis of human colonization of the globe and demonstrated the associations among

ity

ING

: technologies with categories of
Claims about racial background
gh the services provided by online
ı provide public access to inexpen-
Koenig, Lee, and Richardson 2008).
ıce of Native American or African
aims to Jewish priestly status. These
markers, the technology developed
ɔulation stratification (discussed pre-
recreational estimates of the percent-
pean, African, or Asian.
new and challenging frontiers in ra-
he notion of a tight linkage between
vices, like any commercial venture,
or the product. Marketing literature
il purity and racial mixture, as well as
ancestry, and homeland. The implica-
esting could either complicate notions
. As ancestry testing becomes cheaper
igurations of racial and national iden-
level, genetic race verification services
ons for community concepts of kinship
ıtitlements that are tied to race, such as
nbership, genetic ancestry testing may
ıbout eligibility and the social recogni-
ology of biological race verification will
analysis.

ATIONS OF HUMAN
J: RESEARCH, RACE,
NOMICS

race increasingly takes place in a medical
tieth century, human population genetics
ith anthropological efforts to reconstruct
ion. This research succeeded in offering
the "out of Africa" hypothesis of human
ıd demonstrated the associations among

time, geographical distance, and genetic variation. In a departure from this anthropological context, current investment in genetic research on human population variation is increasingly driven by the goal of "personalized medicine," as well as that of alleviating "health disparities." Research on genetic variation among populations stratified by race is widely pursued as a step toward therapies tailored to individual genetic signatures, or personalized medicine (Wilson et al. 2001; Lee 2007).

Pharmacogenomics, or the search for genomic markers that may help physicians determine safe and effective drug dosage, is the first likely application of personalized medicine. Individually targeted therapies have long been promised as the hallmark of a future "molecular medicine." The same drug, at the same dose, works well in some people, has no effect in others, and may lead to serious adverse outcomes in some who take it. Speed of metabolism, like skin color, is under partial genetic control. Variation in individual response to the ingestion of substances—whether naturally occurring substances or manufactured drugs—varies across the globe. The classic example is the occurrence of a "flush response" after drinking alcohol, which occurs more commonly, but not exclusively, among those with Asian biogeographical ancestry. As the pharmaceutical industry seeks marketable technologies to remain profitable during an unexpected drought in medical breakthroughs, pharmacogenomics has become a particularly attractive investment. Converging with this trend, increased government interest in alleviating health disparities, which often fall along racial lines, has also directed resources toward research on genetic differences among races. The discourse of "health disparities" once focused primarily on differences in health outcomes and access to quality health care. Now a concern with health disparities fuels investment in research on, for instance, possible genetic causes of asthma in Hispanics. The pharmaceutical industry has also used the promise of remedying health disparities to lend a politically correct image to its efforts to market drugs or genetic tests to racial subgroups. The twin emphases on redressing health disparities and individualizing health care effectively shield race and genetics research from appearing as fringe or retrogressive as it once might have.

In their search for genetic signatures associated with differences in drug response, pharmacogenomics researchers focused first on differences between racial and ethnic populations (Genes, drugs, and race 2001). However, claims that such differences exist are often anecdotal, rarely validated, and lie beyond the scope of U.S. Food and Drug Administration (FDA) scrutiny (Tate and Goldstein 2008). Nonetheless, clinical

practice reflects these trends of pharmacogenomics and has also begun to incorporate assumptions about racial ancestry and genetics. When attempting to individualize drug therapy, a clinician ideally should directly assess the genes controlling expression of a drug metabolizing enzyme rather than relying on either their impression of the patient's biogeographical ancestry or the patient's self-reported race. Since such tests are not yet available, doctors are likely to be moved by the new paradigm of research on racial variation in drug response to rely increasingly on hunches about an individual's likely ancestry, assessments that are particularly complex in plural societies like the United States. This reveals yet another way in which the scientific practices of "doing race" can have consequences in the social world.

In Chapter 3, Marcus Feldman argues forcefully that classic racial categories are not biologically useful for medical purposes. In a detailed review of recent human population genetics findings, he demonstrates why these findings yield insufficient information about the genetic signature of an individual. Feldman notes that the hope that race translates into useful clinical information remains premature. However, as both the hope of racially targeted drug therapy and the case of BiDil, discussed below, illustrate, prior assumptions about race and health frequently trump evidence.

RACE-TARGETED RESEARCH AND THERAPEUTICS: THE CASE OF BIDIL

Genetic research utilizes racially coded genomic data with the goals of understanding the distribution of disease in the human population and developing population-specific drugs. In 2005, BiDil became the first drug approved by the U.S. FDA for a racialized population (African Americans). The story of BiDil is complex and multifaceted (Sankar and Kahn 2005). On the surface, the approval of a drug for a specific population seems justified by claims of significant racial differences in the rate of heart failure between blacks and whites and by the results of a clinical trial of a combination drug product conducted exclusively in African Americans (Taylor et al. 2004; Temple and Stockbridge 2007).

As noted above, the Institute of Medicine has documented poor cardiovascular disease outcomes in blacks; that fact is not in dispute. However, through detailed historical and ethnographic analysis, Kahn (2004) has demonstrated that market forces—not improvements in care—actually motivated the creation of the first racially targeted drug

twentieth century, research quickly documented that the frequency of certain blood types varied across the world. In fact, in an era before the development of modern molecular genetics, blood group analyses were among the first techniques used to study human genetic variation and the history of migration across the globe. To some extent, individuals' biogeographical ancestry relates to the likelihood that they will have a particular blood type. For example, those with ancestry from Asia are more likely than other biogeographical groups to carry blood type B. Note that, with some exceptions, all blood groups are found throughout the world; what varies is the *relative frequency* of blood types O, A, B, or AB. (The distribution of other blood types, such as the Rh or Rhesus factor, also varies by biogeographic ancestry.) Karl Landsteiner's discovery of human blood groups was recognized with the Nobel Prize in 1930. The implications of this important discovery for ideas about race was drawn by the chairman of the Nobel committee at the award ceremony, who noted,

> The varying frequency of the individual blood groups in different races *points to essential constitutional differences*. Here Landsteiner's discovery opened up new fields for research on the racial purity of a people. Blood group determinations have shown that if an alien race is present within a population this race retains its specific blood group characteristics, even if it has lived away from its main and original homeland for centuries. (Kenny 2006, 473, emphasis added)

New scientific findings are immediately linked into existing social concerns, in this case notions of racial purity.

Biological difference is, of course, meaningful and of great consequence. Making correct distinctions among blood types—variations caused by antigens on the surface of red blood cells—is critical in the safe transfusion of blood. Blood typing according to ABO and Rh classifications (as well as other, less well-known systems) is routinely carried out prior to transfusion. Nonetheless, long after the practice of transfusion medicine became safe and commonplace because all blood was carefully typed for specific compatibility prior to use, American hospitals maintained a strict segregation of blood from black and white donors. "Black" blood was reserved for black patients, and "white" blood was reserved for whites, reflecting social practices of segregation common in the Jim Crow South. The American Red Cross initially refused to send "black blood" overseas for U.S. troops during World War II, though the group eventually compromised by sending it and strictly labeling the race of

the donor (Kenny 2006). Symbolic apartheid of human blood persisted in American hospitals into the 1950s and 1960s (Lederer 2008).

In 1959 a major public dispute arose in response to the proposal that racially segregating blood could help reduce adverse reactions to the blood transfusions required for major cardiac surgery (Kenny 2006). A newly identified antigen named "Kidd" was found to vary in frequency between blacks and whites, sometimes leading to sensitization, a condition making it extremely difficult to find compatible blood. The *New York Times* played a major role in reporting about the proposal, put forth by several leading blood bankers. They suggested that segregating donor pools was a cost-effective way to make transfusions safer. The proponents argued that markers of human population difference were simply neutral, noting, "this may sound wrong sociologically . . . but it is scientifically correct" (Kenny 2006, 462). Only with the origins of the civil rights movement did a revised reading of the science of human blood groups arise. In 1959, proposals to segregate human blood were reported as a symptom of widespread racial discrimination, not as a necessary procedure to protect racial purity, supported by scientific fact.

THE ROLE OF THE MEDIA

Scientific research concerned with race or racial categories is often misinterpreted or simplified when journalists translate complex findings for the public (Condit et al. 2004). The BiDil case, for example, was widely—and simplistically—reported. Yet the media plays an inevitably critical role in shaping public understanding of research findings about genetic population differences. Geneticists may use great care to avoid inaccuracies in their writing about human population genetic difference. But this care can be undone at the moment of translation to the public. Nicholas Wade's coverage of the Rosenberg findings, discussed above, provides a useful illustration. The scientists on the team had carefully avoided the use of lay conceptions of race in their reports, but journalists turned, nonetheless, to the cultural logic of race in their explanations.

In a valuable essay, the journalist and science writer Sally Lehrman (2008) characterizes the news environment that makes it difficult for reporters to decouple race, genetics, and biology. Conventions in story framing and source selection often result in reductionist thinking about race and genetics. Most important, stories must attract attention; they must convey "news." Thus universities' press officers and scientific journals often oversimplify, noting that "the gene for x, y, or z" has

been "found," or a "new drug found to treat heart disease in blacks" has been discovered. The fact that reporters are assigned to specialized "beats," such as sports or crime, increases the possibility of mistranslation. The scientific proposal to segregate blood for transfusion, noted above, coincided with the beginning of the civil rights movement. This social context provided journalists with a new language for discussing the implications of a seemingly "neutral" set of scientific facts describing allele frequency differences among human groups. A new "beat" for reporters had emerged. Few journalists become experts in the intricacies (and limitations) of genetic science as it applies to complex topics like the behavioral genetics of addiction or human population genetic studies. But journalists serve as the connectors and mediators who convey and interpret findings linking race with genetics for the public. Thus their use of racial categories and language must be accounted for when considering the policy implications of advances in genomic science. Reporters reflect lay conceptions and interpretations of race, a discourse that is also influenced heavily by U.S. regulatory requirements.

RACE AND THE CONDUCT OF SCIENCE: MANDATED ETHNIC AND RACIAL CATEGORIES IN BIOMEDICAL RESEARCH

Federal guidelines, in particular those promulgated by the U.S. Office of Management and Budget (OMB), provide specific, targeted incentives to see and use race and ethnicity in a manner that promotes the reification of race as a genetic category (Lee, Mountain, and Koenig 2001; Kahn 2006). Beginning in 1994, researchers with funding from the NIH were required to categorize all human participants in their research by sex and also according to racial and ethnic background (Epstein 2007; Hunt and Megyesi 2008a, 2008b). The motivation for this bureaucratic change was a claim based on justice: Many large clinical studies had only included white men, and thus the findings could not be applied to women. Rates and age of onset of heart disease, for instance, differ significantly between men and women; conducting research exclusively in white male subjects, because they were easier to recruit into clinical studies, clearly limited the findings. Similar logics were applied in arguments requiring the inclusion of minorities in federally funded research (Epstein 2007).

The U.S. National Institutes of Health, the largest funder of genomic research related to human health, enforces these rules (NIH 2001), as does the U.S. Food and Drug Administration (FDA). Although these

regulations were initially envisioned as applying to the conduct of large-scale clinical trials, they are now enforced in all human research. Every scientist and clinician requesting funding from NIH (or reporting study results to the FDA) must report data using standard categories of race and ethnicity that are regulated by the OMB. The OMB categories, developed (and refined each decade) to guide the collection of census data, are required for all government data reporting.

However, there is a problem with applying these categories unreflectively in biomedical research (Kahn 2006). Developed for nonhealth purposes, such as ensuring diverse voting districts or aiding the enforcement of antidiscrimination legislation in schools and workplaces, census categories are fundamentally political tools (see Chapter 1). They are an inadequate reflection of *actual genetic variation* in human populations. Thus the race and ethnicity reporting categories that scientists are required to use may, ironically, lead to bad science. Researchers following regulatory recipes may set up comparisons according to social and political categories of race or ethnicity, while missing true sources of biological difference. People who identify as Hispanic, for example, have biogeographical ancestry from around the globe and yet are combined in a single group (Montoya 2007). Similarly, those classified as "Asian" in the U.S. Census have diverse geographical origins, from South Asia to Japan. There are harms in the uncritical use of race as a proxy for biological relatedness. This practice not only makes the science less accurate, but research methods and reporting of results strongly convey the notion that race is "real," a biological fact, and that the human population can be easily divided into distinct races (Lee 2007).

Although regulations set standards for reporting of human data using racial and ethnic categories, empirical studies have documented that in practice these categories are inconsistently, if not haphazardly, applied (Shanawani et al. 2006; Ma et al. 2007; Hunt and Megyesi 2008a). Self-report by research subjects is complicated by identity politics; many participants do not understand how questions soliciting "ethnic" affiliation differ from questions about classic racial distinctions (black, white, Asian, Native American). True measures of biogeographical ancestry have not been developed and validated. As noted above, Lee (2005) and others have documented how lay conceptions of race and ethnicity are built into the DNA storage repositories that support genetic research on topics such as pharmacogenomics or studies of disease etiology. The outcomes of studies using samples collected using OMB categories recreate the flaws of classification built into the databases themselves.

RACIAL REALISM

Lee (2008) points out an additional ethical dimension of what she terms "racial realism." Racial realists leverage recent population genetics research to assert that racial difference is a "hard" biological reality, easily discernible in the human body (Satel 2002). And they have enthusiastically incorporated into their platform the endorsement of racially targeted therapeutics in the name of social justice. At the same time, racial realist discourse allies itself with health policies that deemphasize social solutions to problems, such as disparities in health outcomes among racially identified groups. This approach is, in turn, tethered to a moral discourse of "personal responsibility." Genetic research on race is mobilized as a response to liberal models of race that focus on environmental and historical conditions as the root causes for health disparities. From this perspective, for example, the comparatively higher rates of addiction in Native American populations are hypothesized to stem from genetic variations, not from the complex history of dislocation and alienation that characterize reservation life set apart from American society (Dingel and Koenig 2008). The ways that scientists frame and explain genetic differences filter into political discourse and influences the *range* of perceived solutions to questions of inequality and justice. In health disparities research, the result is that social views of race lose ground to biological interpretations of difference.

CONCLUSION: THE APPROPRIATE USE OF "RACE" IN GENETICS RESEARCH

Discussing race across disciplinary boundaries is challenging. There are long-standing distinctions in the ways that scholars in the humanities, social sciences, and genetics approach the concept of race (Koenig, Lee, and Richardson 2008). In the humanities and social sciences, the view that race reflects social hierarchies rather than biological or genetic difference is the starting point of analysis. Visible differences in skin color or hair texture, while linked to concepts of race and self-evidently "genetic," do not fully explain the meaning of race in social life—that is, the way race functions as an empirically measurable determinant of social status, and thus of health outcomes. For historians, the very concept of race is imbued with the notion of hierarchical ordering; in the West, modern understandings, often justified by the science of the

thical dimension of what she terms
/erage recent population genetics
ence is a "hard" biological reality,
(Satel 2002). And they have enthu-
atform the endorsement of racially
of social justice. At the same time,
with health policies that deempha-
ch as disparities in health outcomes
his approach is, in turn, tethered to
onsibility." Genetic research on race
l models of race that focus on envi-
s as the root causes for health dispari-
mple, the comparatively higher rates
opulations are hypothesized to stem
he complex history of dislocation and
vation life set apart from American
. The ways that scientists frame and
nto political discourse and influences
o questions of inequality and justice.
result is that social views of race lose
ns of difference.

ROPRIATE USE OF "RACE"
H

y boundaries is challenging. There are
ways that scholars in the humanities,
oach the concept of race (Koenig, Lee,
manities and social sciences, the view
hies rather than biological or genetic
of analysis. Visible differences in skin
d to concepts of race and self-evidently
he meaning of race in social life—that
n empirically measurable determinant
alth outcomes. For historians, the very
the notion of hierarchical ordering, in
ngs, often justified by the science of the

time, arose with the need to justify colonial expansion and domination (Fredrikson 2002).

Among researchers studying human genetic variation at the molecular level, however, the story is different. Genetic scientists disagree on the utility of racial classification for studying common disease, but they generally accept new data showing the range of human genetic diversity and its global patterning (Rosenberg et al. 2002; Feldman, this volume). Many consider the *possibility* of biological determinants of a range of differences among human populations defined by racial labels to be an open scientific question. They enthusiastically embrace new genomic technologies to elucidate the etiology of common illness, making use of the categories of difference required within U.S. regulatory regimes—categories that correspond to differences in lived experience in societies historically divided and structured by race. Asthma, diabetes, hypertension, and heart failure all are examined through a genetic lens. Unlike historians or sociologists, geneticists may assume that categories of difference "found in nature" can be deployed in *neutral* ways, acting as simple classifiers of true genetic variability.

There is an irony associated with the day-to-day performance of race in health disparities research that uses genetic tools. On one hand, the attention to group difference, the increasing focus on genetic variability, provides powerful scientific explanations of variation in disease susceptibility and pathophysiology. On the other hand, the very conduct and practice of the new genomic medicine has the unintended consequence of molecularizing the concept of race and of racial difference (Fullwiley 2007). As in nineteenth-century racial science, when difference is located in biology, the social world is erased, even if unintentionally.

What is the answer? The category of race, fundamentally a social category (although long intertwined with ideas of biological difference), *must* be used in biomedical research in order to assure that health disparities based on injustices may be tracked and addressed. At the same time, researchers must track and account for actual genetic variation associated with biogeographical ancestry. Can we accomplish both objectives at once?

I believe that we can. A recent statement published by a Stanford University faculty working group proposes ten principles to guide the use of racial categories in human genetics research (Lee et al. 2008). The authors note that using racial categories appropriately is an ethical obligation of all researchers, particularly those whose work targets health

disparities. Changed practices of "doing race" in science—in particular, careful attention to language when discussing actual genetic differences across human groups—may alleviate the harm of a newly geneticized or molecularized understanding of race. The signers warn, "The 'gene' remains a powerful icon in the public imagination and is often misunderstood as deterministic and immutable. . . . [H]istory reminds us that science may easily be used to justify racial stereotypes and racist policies" (Lee et al. 2008, 3). It is critical for researchers to recognize that racial and ethnic categories are created and maintained within particular social and political contexts, and that those contexts change over time. Since racial classifications in medicine have sometimes reflected racist ideologies or prejudice—as when blood banks separated black and white blood for transfusion—it is critical to use care in assigning labels to participants in research studies or when rendering clinical care. One recommendation, suggested by the editors of several major scientific journals, is that authors specify exactly how and why human subjects or specimens like DNA are assigned a specific racial or ethnic identifier (Sankar and Cho 2002). Specific classifications, and comparisons across groups, must be justified, not simply assumed (Schwartz 2001; Braun et al. 2007). This obligation applies in spite of U.S. regulatory requirements for inclusion of all populations when conducting clinical research.

The concept of polysemy—the capacity for a word (or sign) to have multiple meanings simultaneously—may be helpful. Scientists and clinicians attentive to the complex history of the race concept will be able to apply a polysemic racial vocabulary in their everyday work, a set of definitions that avoids confusion between social race and true biological difference. When they do race, when they classify human research participants into categories or seek the root causes of health disparities, scientists must maintain a conceptual distinction between two senses of "race": race linked to genetic variation deriving from biogeographical ancestry *and* race as a reflection of the lived experience of social difference. Both senses are needed to make progress in alleviating health disparities. We must not ask "whether" race should be used in genetic research; rather, we must ask, "under what conditions and for what purposes?"[2]

2 I wish to thank Sandra Soo-Jin Lee, Joanna Mountain, Sarah Richardson, and all the participants in the volume *Revisiting Race in a Genomic Age* for many stimulating discussions about the intersections of race, health research, and genetics. Troy Duster's thinking deeply informed the policy recommendation in the conclusion of this essay. I am extremely grateful to Hazel Markus and Paula Moya for inviting me to lecture in their Stanford course and for recognizing the critical importance of biomedical practices to our shared social understanding of "doing race" in the United States.