This paper was presented at the session "Exploring the Nature of Human Biological Diversity: Myth v. Reality" at the American Anthropological Association (AAA) 2003 Annual Meeting on November 21, 2003 in Chicago, Illinois. The session was sponsored by the AAA *Understanding Race and Human Variation* project, organized by AAA's Society for Medical Anthropology and Committee on Minority Affairs, and funded by the Ford Foundation. The *Understanding Race and Human Variation* project is funded by the National Science Foundation and the Ford Foundation. This paper represents the views of the author and not the AAA *Understanding Race and Human Variation* project.

HUMAN GENETIC VARIATION: THE MECHANISMS AND RESULTS OF MICROEVOLUTION

Jeffrey C. Long
Professor of Human Genetics
Department of Human Genetics
University of Michigan Medical School
Ann Arbor, MI 48109

INTRODUCTION

Only a few decades ago, our knowledge about genetic diversity was far more limited than today. Most of what we knew was obtained from a few genetic marker systems such as blood groups and serum proteins. These systems were attributable to variations in genes, but our knowledge of genes was rudimentary. Each *gene* was known to reside at specific location on a chromosome called its *locus*. It was also known that a gene contained a set of instructions for a particular product. Nevertheless, little was known about the chemistry of the gene or the nature of variability in genes. Nothing was known about how well variation in the few genes whose products could be observed represented variation in the vast number of genes that were yet to be discovered. Little was known about how well variation in the products of genes represented the variation in the genes themselves. Recent advances in DNA analysis have changed the playing field, and the nature of variation is now observable in the details of the DNA molecule (9, 63). It has been necessary to refine and extend basic genetic concepts, methods, and terminology in order to fully utilize these new data.

A QUICK LOOK AT THE HUMAN GENOME AND GENOMIC DIVERSITY

Research today considers genes in the broad context of genomes. A *genome* is defined to be one complete copy of all the genes and accompanying DNA for a species. Each person holds two copies of the genome. One copy obtained from their mother's egg and other from their father's sperm. Each copy of the human genome consists of over three billion pairs of the repetitive building blocks of DNA. These building blocks are called *nucleotides*, of which, there are four different kinds, denoted by the letters A, C, G, and T. The letter designations for nucleotides are used as a short hand for the chemical bases that give the four kinds of nucleotides their distinctive properties. The DNA molecule (Fig. 1A) consists of two complementary strands in which each

nucleotide, or base, from one strand is paired with a complementary base on the other strand. Because the base pairing is complementary, each strand contains all of the information for the nucleotide sequence of the other strand. The information in a gene is encoded in the sequence of nucleotides at its locus. The complexity of our genome is great, and it defies simple descriptions. For example, our genes vary greatly in size; the largest gene, dystrophin, is about 2.4 million nucleotides long, while the smallest gene, which encodes an rRNA is only 73 nucleotides long. Alternative forms of a gene, called *alleles*, arise from minor changes in the nucleotide sequence at the gene locus.

Surprisingly, the genome contains far more DNA than is required to encode all of the information in our genes. In fact, only about two percent of the genome encodes genes, and about half of the genome consists of repeated nucleotide sequences with no known function. The functions of repetitive DNA and the reason why our genome contains so much more DNA than is necessary to encode all the genetic information are active topics of investigation. Interestingly, there is little difference in genome size or the number and kinds of genes across different mammalian species. The existence of so much DNA outside of gene loci has necessitated that we broaden the concepts of the locus and allele. Now, a locus refers to any specific location on a chromosome, whether or not the DNA sequence encodes a gene. An allele at a locus is any alternative sequence of nucleotides at that locus.

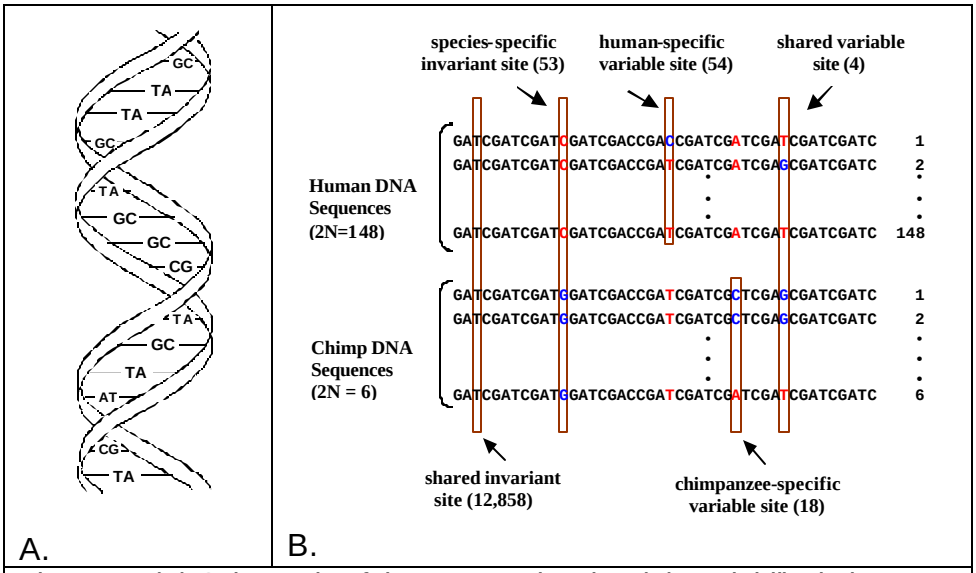

Figure 1. (A) Schematic of the DNA molecule. (B) Variability in human and chimpanzee DNA sequences. The sequence of only one strand is shown to avoid redundancy. Redrawn after reference (17)

No two copies of the genome are identical. The differences are usually minor, at some places one base is substituted for another, or a small number of nucleotides is inserted or deleted from the DNA. These differences are created at random by mutation. We can get some idea about the level and pattern of genomic variation by making comparisons within and between species. This is illustrated in Fig. 1B, which is redrawn from a study of genes with functions related to blood pressure and hypertension (17). The nucleotides for a sequence of 12,987 sites was obtained from a

total of N=74 people and N=3 chimpanzees. At 12,858 sites (99%), the same nucleotides were present in all examined copies of the human and chimpanzee genomes. At 53 sites (0.4%), there was a fixed difference between species, i.e., all copies of the genome from one species possessed one nucleotide while all copies from the other species possessed another nucleotide. At 54 sites (0.4%) humans were variable while chimpanzees were constant. At 18 sites (0.01%) chimpanzees were variable while humans were constant. Finally, at 4 sites (0.03%) humans and chimpanzees were both variable and appeared to harbor the same variants. The challenge is to determine what these patterns of variation mean in terms of the biological processes that create, maintain, and distribute variation. How many variable sites should we expect in a species? How many species-specific sites should be expected? How will the size of our sample influence our observations? Will a new mutation ultimately be lost from or fixed in a species? How many generations will elapse before that new mutation is lost or fixed?

FOUR MECHANISMS OF EVOLUTIONARY CHANGE

Four processes shape the level and pattern of variation. They are random sampling, mutation, exchange of members, and natural selection. The biological concept of a population is essential to understanding how these processes work. In principle, a population is defined as a group of individuals from the same species who, excepting barriers imposed by sex and developmental stage, are likely to mate and reproduce (13). In practice, it is hard to precisely identify the boundaries and membership of a population. Populations are often bounded by geography, but ethnicity and language also play roles in delimiting human populations (5, 13). Suffice it to say that human populations are stratified and related to each other in complex ways, and research often requires operational definitions (29). Despite this difficulty, the population concept is necessary for deriving the relationships between processes and outcomes.

Random Sampling of Gametes

The first of these processes is random sampling of gametes in the formation of a new generation. As a simplification, we envision a process where a finite number, N, of individuals produce an infinite number of gametes, from which 2N gametes are randomly

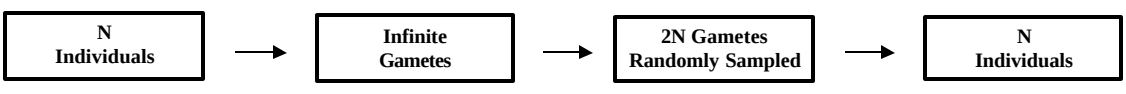

Figure 2. The Wright-Fisher sampling model.

chosen to produce N new individuals in the next generation (Fig. 2). The process is repeated over endless generations. Random sampling produces a phenomenon called *genetic drift*. If no other forces are operating, the finite sampling process will eventually lead to the loss of all variation. The chance that a particular variant is lost or retained is totally random. In general, the rate of genetic drift is inversely proportional to population size. Variation is lost very slowly in large populations but it can be lost rapidly in small

populations. In natural settings, genetic diversity is lost faster than would be predicted from population sizes alone. One reason for this is that the pool of reproductive individuals is always smaller than the total population. There are other reasons that include individual differences in expected fertility and changes in population size. Basic principles show that if the population size fluctuates, genetic diversity is lost at a rate related to the smallest size. There are two noteworthy circumstances that greatly accelerate genetic drift. The first, called a *bottleneck*, occurs when a population size is reduced for a protracted period of time and then rebounds. The second, called a *founder effect*, occurs when all individuals in a population trace back to a small number of founding individuals.

Mutation

Mutation is the second force that affects the extent and pattern of variation in populations. A mutation occurs when the nucleotide sequence of a DNA molecule is altered. New mutations occur randomly in both time and space. Many mutations are functionally silent because they occur outside of genes or they miss the important coding regions of genes. These silent mutations are often used for studying genetic ancestry and the demography of populations. However, when functional mutations do occur, they usually impair function (11). In the long run, the production of new mutations offsets the loss of genetic variation by finite sampling. Population geneticists devote a good deal of effort to determining the balance between these two opposing forces. The most common approach used for finding this balance is called *coalescent* modeling (22, 39, 51). While coalescent models can be taken to levels of great complexity, the basic ideas are rather accessible and they provide a good deal of insight into the structure of genetic variation in our species. Each coalescent model looks at the diversity between DNA sequences in terms of three basic components: (i) the expected time back to a common ancestral sequence, (ii) the mutation rate, and (iii) the outcome of a mutation.

The coalescent approach is illustrated here by considering the expected number of nucleotide substitutions (*S*) between two copies of a locus sampled at random from a population. We begin by asking, what is the average number of generations back to the most recent common ancestor (MRCA), and what is the average number of mutations per generation. With these quantities, and the assumption that each mutation leads to a nucleotide change, the expected number of sites that differ between the two sequences is $S = 2\bar{t}m$, where $\bar{t}$ denotes the average time back to the common ancestor and μ denotes the average rate at which mutations hit the locus (22, 55). The coefficient 2 comes from the fact there are separate lineages for the time leading back to the common ancestor. Basic probability gives the intuitive result that, on average, the waiting time to the common ancestor is the number of genomes in the population $\bar{t} = 2N$, which is twice the number of people. Thus, we can rewrite our equation as $S = 2\bar{t}m = 4Nm$. Figure 3 illustrates an outcome of this process for two copies of a locus. A and B share a common ancestor at a time in the past, t_{MRCA}. After that time, two mutations occurred on the lineage leading to B and one mutation occurred on the

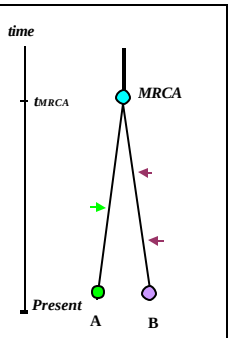

Figure 3. The coalescent model.

lineage leading to A. Assuming that each mutation led to a unique substitution of one nucleotide for another, then A and B will differ at three nucleotide positions. This simple example hides the fact that there is a great deal of variability in the process. If two other copies of the same locus were chosen from the same population, their time of coalescence would likely be different, and even if the time were the same, the actual number of mutations would likely be different.

The expected number of nucleotide differences between two random copies of a locus (*S*) is an ideal measure of variation at the DNA level because it shows the degree of difference between alleles. However, coalescent methods can also be used to derive expectations for more conventional measures of genetic variation. For example, it is easy to show that in a random mating population the expected probability of a homozygous genotype is $f = 1/(4Nm+1)$ (22). It follows that the expected probability of a heterozygous genotype is $h = 4Nm/(4Nm+1)$. Interestingly, all three measures of variation, *S*, *f* and *h,* are related to the underlying process through the same parameter, $4Nm$. Notice that the effect of mutation is inextricably tied to the population size. There is no distinction between a large population with a low mutation rate and a small population with a high mutation rate. It should be noted that these simple expectations are strongly dependent on the assumption that the parameter $4Nm$ has not changed over a very long time. Fortunately, coalescent theory has been developed for more complex population histories (22, 39).

Just as any two copies of locus in a population eventually coalesce to a common ancestor all copies of a locus in a sample or population eventually coalesce to a common ancestor. The mathematics of coalescence in larger samples is well understood and there are large sample formulas for basic quantities such as the expected number of variable sites in a sample and the variance of the number of variable sites (22). Without delving too deeply into the theory for larger samples, an intuitive feel for some important results can be obtained from the diagram in Figure 4. First, there is a correlation between the age of a neutral allele and its frequency in a sample. More common alleles tend to be older. Most rare alleles are young, but very old alleles are also rare. Second, because of their age, copies of an older allele will tend to carry more background mutations than copies of younger alleles. Interestingly, these expectations will be distorted if a population's structure is more complex than the simple model outlined in Figure 2. For example, an excess of rare alleles is an indication that population size has recently expanded (41, 46, 47, 54).

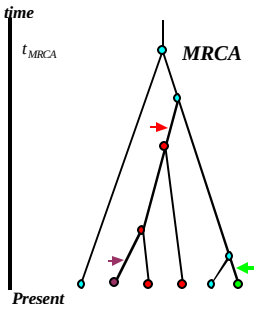

Figure 4. Coalescent for larger samples.

Subdivision, Migration and Genetic Exchange

Migration is the third force that affects the extent and pattern of variation. In the population genetic sense, migration requires that the total population is divided into local groups, and that individuals who migrate reproduce in their new group (14). This is clearly not migration in the sense of the large-scale cyclical movements of entire populations typical of some species such as geese and butterflies. It is also distinct from the migrations deep in history by which humans moved into unoccupied regions on the globe.

It is useful to present some basic ideas for subdivided populations before turning to migration in detail. As shown in Figure 5, the existence of local groups does not prevent all copies of a locus in the species from eventually coalescing to a common ancestor. Interestingly, the copies of a locus coexisting in one local group do not always coalesce together before they coalesce with copies from another group. Notice in Figure 5 that B and C coalesce with D before they coalesce with A. The general property that more frequent alleles tend to be older has an important implication for subdivided populations. That is, an allele that is common in one local population will be common throughout the species. However, the presence of local groups does increase the average time to coalescence(51-53). This increase ultimately leads to more genetic differences between copies of a locus sampled from different local groups than for copies of a locus sampled from the same local group.

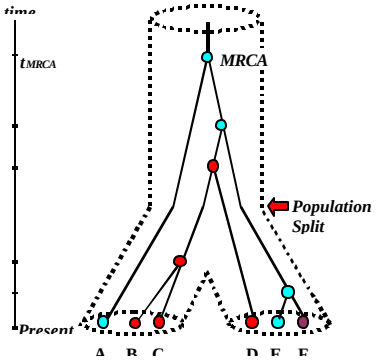

Figure 5. Coalescent in a subdivided population. Drawn after reference (37).

Turning back to migration, genetic exchanges neither create nor remove variation. They only reshuffle it. The more exchange between local groups the less genetic differentiation will occur. The process of genetic exchange between neighboring local populations is often called *gene flow*. A large body of theory exists for the effects of gene flow (6, 30, 52). Without belaboring details of this theory, the important parameter that emerges is *Nm*, the product between the local population size and the probability that an individual has migrated. The product *Nm* can be interpreted as the actual number of individuals who migrated into a local group. This has an interesting effect, a collection of large local groups with low migration will show levels of differentiation that are about the same as a collection of small local groups with high migration.

Natural Selection

The fourth force that affects the extent and pattern of variation in populations is *natural selection*. The theory of natural selection is complex and many new developments are being made. Natural selection does not create variation. However, natural selection can remove a harmful variant, spread a favorable one, or even maintain variation at a locus.

The incidence of many rare genetic disorders most likely reflects a balance between new mutations and selection against old ones. This mode of selection is often referred to as negative selection. A good example is human oculocutaneous albinism type 1 (OCA1). OCA1 results from mutations of the tyrosinase gene and presents with the absence of melanin pigment and results in acute sensitivity to sunlight. A recent study (26) found over 60 different alleles that cause OCA1 in a sample of only 102 patients. Most albino patients carried copies of two mutant alleles that differed at the DNA level. This finding suggests that mutation provides a constant supply of new deficiency alleles, but selection removes them from the gene pool before long. In general, negative selection has important implications for general health and reproductive planning, but it does not account for the adaptive characteristics of our species, and it does not seem to impose a low limit to the store of variability in our species.

By contrast, a favorable allele may occasionally rise rapidly and sweep through the population replacing all other alleles at that locus. This mode of selection is often referred to as positive selection. One consequence of a selective sweep is that the level of background genetic variation in the vicinity of the favored allele is reduced. Indeed, this signature can be detected by examining genome-wide patterns of variability (35, 40). A good example of a selective sweep is provided by the null allele at the Duffy blood group (18). It is now established that the vivax malaria parasite uses the functional product of the Duffy blood group to gain entry into the red blood cell. People who lack this cell surface molecule are completely resistant to vivax malaria.

The long-term consequence of both negative selection and selective sweeps is the reduction of genetic variation, yet natural selection can also maintain variability in a population when heterozygotes are favored. The best-documented examples of natural selection follow this mode of selection. For example, the sickle cell allele *(Hb^S)* at the beta hemoglobin locus is maintained along with the most common allele *(Hb^A)* in populations where malaria is prevalent. People who carry one copy of each allele *(Hb^A/Hb^S)* are most fit because they are resistant to malaria, while those who carry two copies of the normal allele *(Hb^A/Hb^A)* are vulnerable to malaria, and those who carry two copies of the sickle cell allele *(Hb^S/Hb^S)* suffer from sickle cell disease. The balance in frequencies for the Hb^A or Hb^S alleles is stable in the presence of malaria. Any tendency for one of the alleles to become too prevalent is reversed by the production of too many unfit homozygotes of that particular type.

HUMAN GENETIC DIVERSITY

Nucleotide Substitutions

A robust picture of human variability at the genomic level is now emerging. This picture is illustrated well by a study (68) where DNA sequences comprising a total of 25,000 nucleotides from 50 different genetic loci were obtained for each member in a sample of 30 individuals. Ten of these individuals were African, ten Asian, and ten European. The individuals on each continent were selected from local groups residing in widely spaced regions. Two key observations were made. First, the level of genetic

diversity in the human species is more consistent with a small population than it is with a large population. Second, nested subsets describe the structure of genetic diversity across geographic regions.

The average heterozygosity per nucleotide (nucleotide diversity) for the combined sample was only 0.08%. This value is very low considering the billions of people on earth today. By applying the formula $h = 4Nm/(4Nm+1)$ with reasonable estimates for N and μ (Say, N = 6,000,000,000 and μ = 0.000,000,01), we would expect a heterozygosity value over 99%. In fact, the observed nucleotide diversity in our species (0.08%) is on par with a breeding population of about 18,000 people. A plausible explanation for this discrepancy is that the human population size increased vastly during the Pleistocene. It is well known that the level of diversity lags behind population growth because mutation rates are low relative to our potential for intrinsic increase.

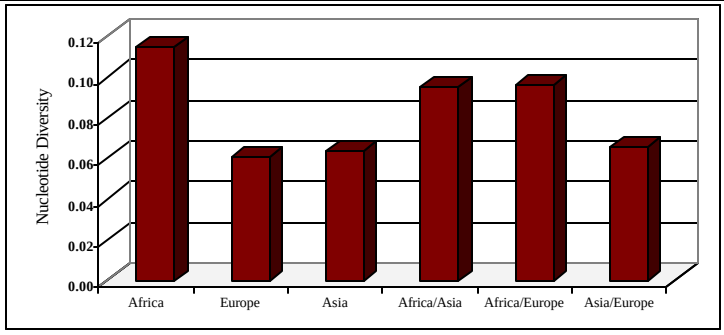

Figure 6

These data also shed a good deal of light on the pattern of nucleotide diversity within and among geographic regions, (see Figure 6). Nucleotide diversity within Europeans is nearly that within Asians while it is much higher in Africans. But, the story is more interesting than this. Notice that pairs of nucleotides, one drawn from Asia and the other drawn from Europe, are about equally likely to differ as a pair both drawn from Europe, or a pair both drawn from Asia. As might be expected, for nucleotide pairs, one drawn from Africa and the other from Asia, are more likely to differ than a pair both drawn from Asia. The observation also holds for African/European pairs. However, for pairs of nucleotides, both drawn from Africa, there is a greater chance of difference than occurs between African/Asian and African/European pairs. This perplexing result reflects the fact that most nucleotide diversity in non-Africans is a subset of nucleotide diversity in Africans. An interesting implication of this finding is that most common variation in the genome could be identified by studying a sample composed only of Africans, but a good deal of the common variation would be missed by studying a sample composed only of Europeans or Asians.

Short Tandem Repeats

A large independent set of human DNA data confirms the basic features of the worldwide pattern just described (4, 48). This collection contains samples from over 1,000 individuals who belong to 52 widely dispersed local populations. Each individual was assayed at 377 short tandem repeat (STR) loci. These loci do not encode a protein or RNA product; rather, they possess runs of simple nucleotide motifs, e.g. CA·CA·CA·CA·CA. Alleles at STR loci vary with respect to the number of motif units that they possess. STRs are abundant throughout the genome and many STR loci are highly variable. The principles of coalescence can be applied to interpreting variability at STRs (53, 69, 70). For example, whenever the number of repeats differs between two copies of an STR locus, at least one mutation has occurred after they shared a common ancestor.

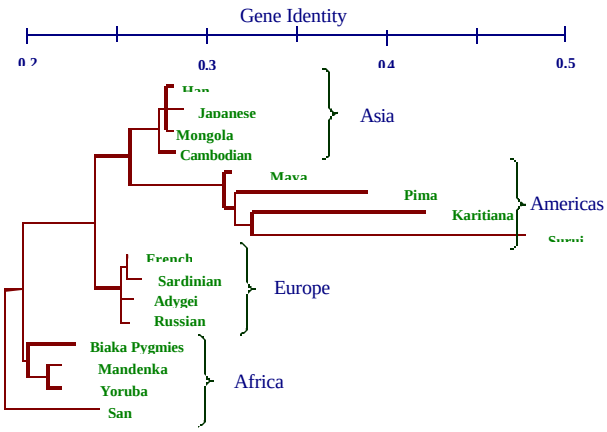

Figure 7

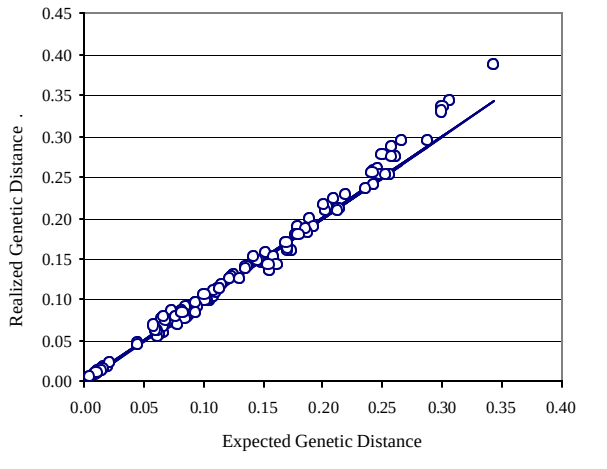

Figure 7 presents and original analysis of a subset of 16 populations from the complete STR dataset. Four populations were selected from each of four continental regions in order to have a balanced statistical analysis. A hierarchical model consisting of local populations within regions within the total population was fitted to these data using the method of (62) . The scale is *gene identity* (37), which is the probability that two copies of locus chosen at random (from the same or different populations) will have the same DNA sequence. Gene identity is equal to the homozygosity that would result from random mating. To some extent gene identity is affected by the ways in which populations are defined and samples are collected (23). Some larger groups may actually consist of aggregates of smaller groups, each with more gene identity. Likewise, some small groups may actually be isolates that are not representative of large geographic regions. Nevertheless, this effect should not bias comparisons across regions (8). The model fitted here nests all non-Africans together, and Native American populations are nested with Asian populations.

The population structure depicted in Figure 7 is uneven. For example, gene identity varies considerably among the samples. It ranges from only 0.22 in Yoruba and

Mandenka to 0.50 in Sururi. There are large differences in gene identity among regions. For example, the gene identity between the San and the Biaka Pygmy is 0.19, while the gene identity between the French and Sardinian is 0.24 and between the Maya and Sururi is 0.30. It is of great interest that the gene identity between the San and any non-African population is about 0.19, which is the same as between the San and the Biaka, Mandeka, or Yoruba. The scatter plot in Fig. 7B provides a test of the fit of the tree model to the actual data by comparing Nei's minimum genetic distance (37) between all pairs of populations to an estimate of genetic distance between pairs of populations obtained by all of the horizontal branch lengths separating the populations (62). The tree model fits the data very well, as evidenced by $R^2 = 0.990$.

GEOGRAPHIC PATTERN AND GENERATING MECHANISMS

The pattern of genetic diversity displayed by the STR loci is consistent with a model that postulates a succession of ancient founder effects that occurred with range expansion and the human occupation of new continents. In this view, the origin of the species was in Africa about 200,000 years ago and the species expanded out of Africa beginning only 100,000 years ago (47). The pathways and sequence of events in Figure 8 (drawn after (38)) are reasonable approximations to history but the dates are speculative. While the present data agree with this scenario, it must be recognized that there are hot debates on the timing of human origins and the global expansion of the species (65).

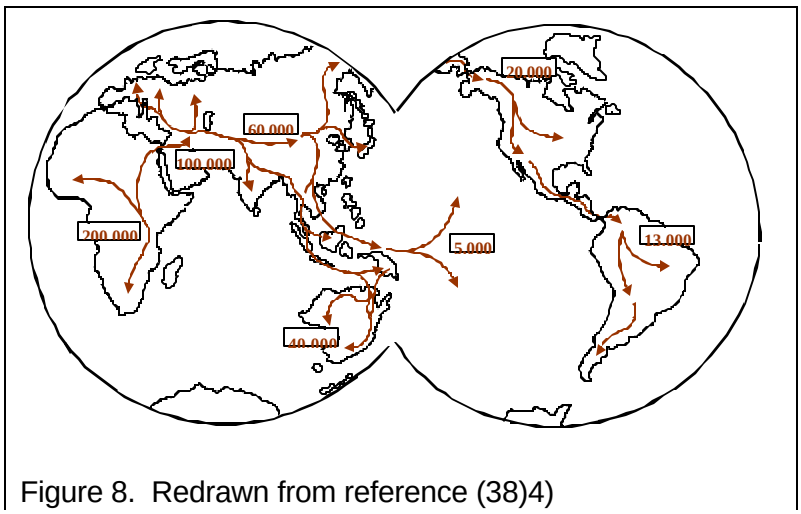

Figure 8. Redrawn from reference (38)4)

Many researchers have noted that the genetic diversity between human populations increases as a function of geographic distance. Figure 9 provides a view of the STR data from this perspective. Here, Nei's minimum genetic distance (37) computed between each pair of local populations is plotted against the over-land geographic distance separating them. There is a clear trend for the genetic distance between local populations to increase as the geographic distance increases. While the tree model is a better predictor of genetic distance, it is important to consider the relationship between geography and genetics in greater detail. It raises a plausible alternative to the range expansion model. That is, the pattern of genetic divergence reflects a balance between the processes of finite sampling and local genetic exchanges (gene flow). In population genetics, there is a formal model called isolation by distance that applies to this situation (30, 66, 67).

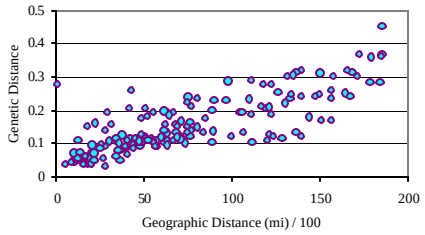

Figure 9. Plot of genetic distance versus geographic distance .

The isolation by distance model assumes that individuals are evenly spread over the range of the species, and that individuals disperse from the location of their birth to a new location to reproduce. In addition, dispersal distances are assumed to be random and to follow a probability distribution that is the same for all members of the species. When these conditions are met, local population size and dispersal distance predict a quantity called genetic kinship, which is a standardized measure of the increase in gene identity caused by geographically restricted mating (30, 33). The isolation by distance model makes two strong predictions. First, genetic kinship should be the same for all local populations. Second, genetic kinship between local populations should decay geometrically with geographic distance (Figure 10, descending line). It is easy to convert genetic kinship to genetic distance (Figure 10, ascending line), and some geneticists have preferred to do so (7, 59). However, information about genetic kinship within groups is lost in the translation to genetic distance. For the purpose of a visual test of the fit of isolation by distance to the STR data (Figure 11) the genetic kinship metric is used. The complete set of genetic kinship estimates within and between local populations is presented in Figure 11A. Superficially, it supports the isolation by distance model. Genetic kinship is highest in local groups and it declines as distance increases. However, a closer look at the data shows some real problems. By contrast to the first prediction of isolation by distance, the genetic kinship in local populations is highly variable. Moreover, the genetic kinship estimated between many pairs of local populations is negative, but genetic kinship is defined only in the interval between zero and one. Figures 11B through 11D show other departures from isolation by distance in the STR data. Figure 11B plots those points from Figure 11A that compare the four African samples with the twelve non-African samples. The range of genetic kinship estimates is very limited. In other words, all pairs of local populations, African with non-African, show about the same degree of genetic kinship, regardless of the geographic distance that separates them. There is no trend for genetic kinship to decrease with increasing geographic distance in accord with isolation by distance. Figure 11C plots those points from Figure 11A that compare the four European samples with the twelve non-European samples. In terms of genetic kinship, there are two strata. The bottom stratum represents comparisons between European/African pairs of local populations, while the top stratum contains comparisons between European/Asian and European/Native American pairs of local populations. There is no trend for genetic kinship to decrease with increasing geographic distance in accord with isolation by distance.

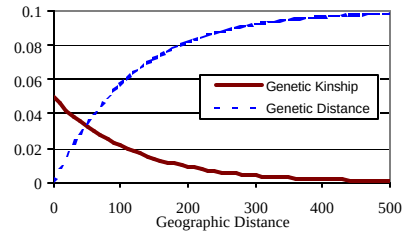

Figure 10. Theoretical predictions of isolation by distance.

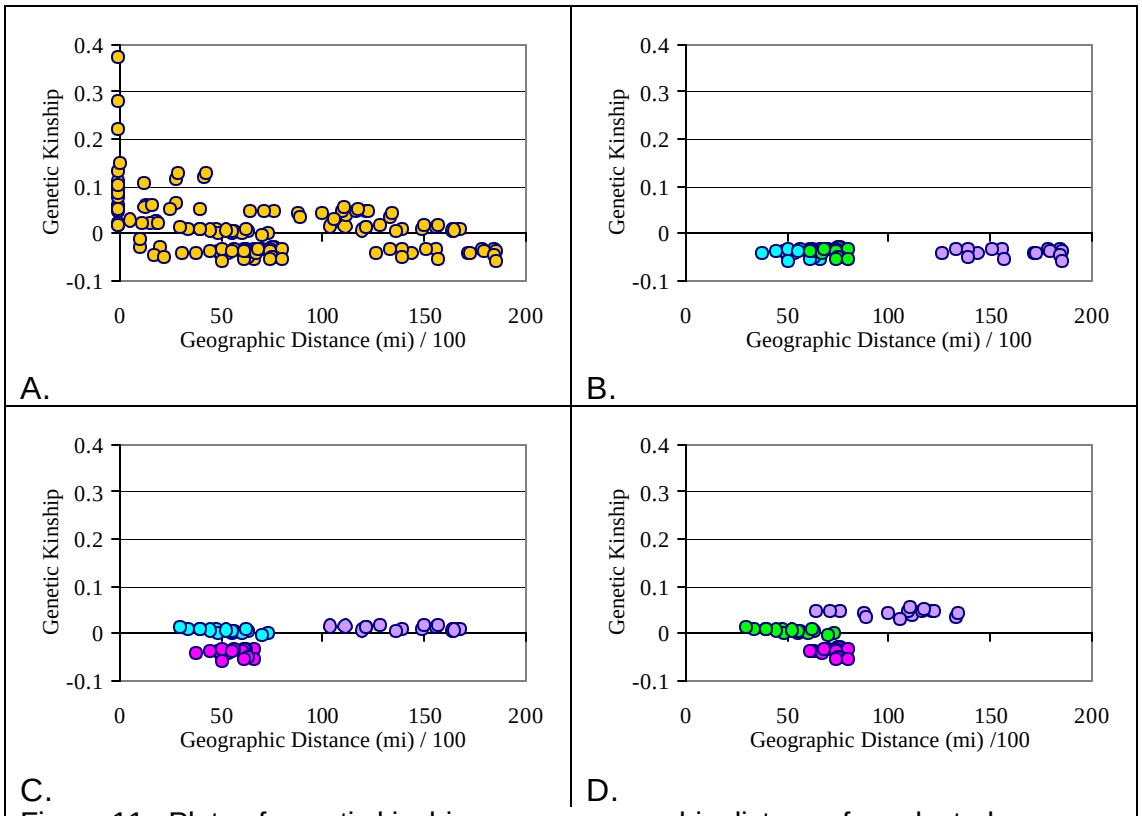

Figure 11. Plots of genetic kinship versus geographic distance for selected populations.

Figure 11D plots those points from Figure 11A that compare the four Asian samples with the twelve non-Asian samples. There are now three strata of points, the lowest stratum compares Asian/African pairs, the middle stratum compares Asian/European pairs, and the highest stratum compares Asian/Native American pairs. The pattern is remarkable. Indeed, it recreates the nested subsets pattern of variation discerned for the DNA sequences in Figure 6.

GENETIC VARIATION AND RACE

It is natural to ask, what relevance do these findings have for the use of race concepts in understanding human variation? However, this question is not as simple as it appears. The term race is used in many different ways, and as often as not, it is used without a precise definition. Twentieth century biologists debated many views on race. There was a trend to replace concepts based on unchanging essential types with concepts related to evolving populations. Despite this trend, a single widely accepted definition for race failed to emerge. Examples of typological and population definitions of race are discussed below in terms of their logical consistency and their fit to genetic data.

Race Concepts

Essentialist: Hooton (21) defined a race as a great division of the species identified by differences in anatomical features whose expression was primarily influenced by heredity. He felt that these features were discerned best in groups and were obscured by individual variations. Hooton defined primary and secondary races. The primary races were unmixed, and represented in modern times only by isolated remnant groups. The secondary races arose as mixtures among the primary races. Most extant populations belonged to secondary races. Hooton saw that individuals within groups vary. However, to him, the variation within groups was noise that obscured the underlying ideal types. By contrast, the most salient concept in Darwinian evolution is that the variation within groups is the ultimate source of variation between groups (28). Without meaningful variation within groups, Hooton's primary races could not have evolved in the first place. The conceptual divide between Hooton and Darwin is illustrated by imagining the history of populations. With Hooton, tracing populations back in time eventually leads to a collection of distinct non-overlapping groups. With Darwin, all populations eventually merge back into a single gene pool unified by common descent.

Population: In population biology, a race is defined to be a cluster of local populations that differs genetically from other clusters of local populations (13). This concept differs from the essentialist concept because a race is defined to be the entire group. Each individual is recognized to be different; no single individual or ideal type can represent all of the individuals. Notice that the population concept of race embraces the central role of variation. The population concept focuses on two key aspects of local populations. The first is genetic differentiation, which precedes speciation in the adaptive process. The second is reproductive isolation, which must be present to maintain genetic differentiation and is the final step in the origin of a new species. In biology, races are sometimes given the formal label of subspecies, and their formation is thought to be a step in the process of forming new species.

Even with this focus on process, the race concept has been difficult to apply. Taken to the extreme, every population qualifies as a race and the two terms are reduced to being synonyms. In order to draw a greater distinction, Mayr (32) defined a race as a phenotypically identifiable subpopulation, with a restricted geographic range, that differs from other subpopulations. He recommended that a population could be declared a race when its members can be correctly identified 75%, or more, of the time on the basis of diagnostic traits. However, the 75% decision rule is operational and does not represent a natural threshold in the evolutionary process. More importantly, the percentage of correct identification depends on two factors, the degree of divergence between local populations and the number of traits measured. With many traits, members of trivially different groups can be correctly classified most of the time. Therefore, it is unclear what achieving 75% correct placement represents: meaningful difference among local populations, or a thorough data analysis. With enough measurements, the taxonomic approach allows every local population to be declared a race.

Templeton (59) suggested that a race, if it exists, is a distinct evolutionary lineage that has been isolated for a long time. This view forcefully maintains the basic requirements related to speciation: divergence and reproductive isolation. However, it has practical problems too. For example, it is easy to define a lineage for a single gene, but it is difficult to define a lineage for an individual or a population. Each individual carries two copies of each genetic locus; therefore, with respect to recent ancestry, an individual represents two different lineages. If we look at copies of different genetic loci in the same person, they will have different lineages and the most recent common ancestral sequence for each locus is likely to have been carried by different people, at different times (15). If a population lineage is defined by an increase in the average number of shared gene lineages, or a decrease in the average coalescent time, then a population lineage implies nothing more than a genetically distinct local population.

Race and Human DNA Sequence Variation

Notice that the essentialist and population concepts of race describe the same basic pattern of variation, but the two views differ in how they attach meaning to it. The essentialist view of race is illustrated in Figure 12A. According to it, variation among individuals (represented by dotted lines) is insignificant, while the ideal types (represented by solid lines) are paramount. The population race concept, as illustrated in Figure 12B, recognizes that variation (represented only by solid lines) is the central necessity of the evolutionary process. In both Figures 12A and 12B, the three races are drawn to show about the same amount of variation, the same overlap, and same distance between centroids. This was done because, by using the same name, we imply that each race attains about the same level of distinctiveness, or, at least, that all populations have crossed the same threshold. This implication stands regardless of how much variation exists or the meaning that is attached to it. Surprisingly, the variation in human DNA sequences does not show this pattern at all. The distinction between the nested subset pattern of actual DNA variation and the hypothetical pattern inherent in all race concepts (Figures 12A and 12B) is highlighted in Figure 12C in which the data from Figure 6 are redrawn.

One important consequence of the actual pattern of variation is that we can expect, at most genetic loci, an uneven distribution of alleles. African populations will harbor some relatively common alleles that will be absent in non-African populations; however, all of the alleles that are common in non-African populations will also be common in African populations. Indeed, the tree diagram in Figure 7 reveals the same pattern of variation that is portrayed in the Venn diagram in Figure 12C. For example, the probability that two copies of an STR locus will carry the same number of repeats when one copy is drawn from the San and the other copy from another African population is same as when one copy is drawn from the San and the other copy is drawn from any non-African population.

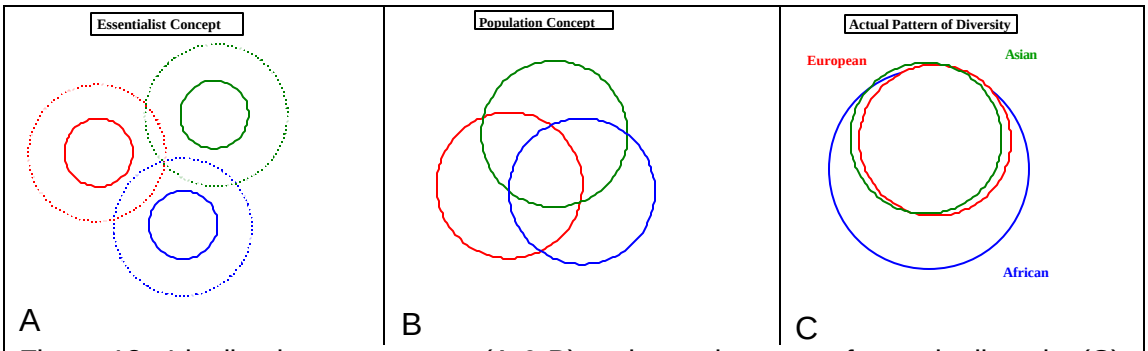

Figure 12. Idealized race concepts (A & B) and actual pattern of genetic diversity (C).

The actual pattern of DNA variation creates some unsettling problems for the lineage definition of races. For example, it implies that non-Africans constitute a race with respect to Africans, but Africans are not a race with respect to non-Africans. Moreover, Asians and Native Americans would be a single race with respect to Europeans, but Native Americans would be a distinct race with respect to Asians. These findings show, indeed, that even the lineage concept of race is a poor descriptor of human genetic variation.

Why do Human Race Concepts Persist?

There are several reasons why simple race concepts continue to be used, despite their poor fit to the actual pattern of genetic variation. First, the large body of DNA data that reveals the nested structure of genetic variation among human groups became available only recently. These data are side benefits from advances in molecular biology and gene mapping. Before this, there were not enough data to reveal the asymmetrical distribution of genetic variation. Second, many current studies are focused on the distinctiveness between groups, or the percent of individuals that can be correctly assigned to their population of origin (1, 42, 48). Classification is the principal concern of these studies. The fact that accurate classification is possible with a great deal of variation within groups is not a mystery. It is expected when a large number of traits can be measured and computers perform heavy computation without effort. Unfortunately, the results from classifying individuals are equally compatible with essentialist and population concepts of race. They hide the nested pattern of variation in DNA that typifies human populations. As noted by Lewontin many years ago (28), a world-view is at issue, not the facts. Third, many people in both the lay and scientific communities use what I call the implicit definition of race (12). In this view, races represent a pattern of variation that is difficult to pinpoint but clear to most people. It is impossible to assess or refute this definition. It takes an article of faith as a given, that races display a pattern of variation that is already clear to most people. Because the implicit concept is so fuzzy, it is easy for its users to unwittingly fall back on typological race thinking. Fourth, population geneticists who have tried to debunk race have often attacked straw man concepts (2, 28). For the purposes of debunking, they define races to be groups of near uniform people that are marked by a few visible traits such as skin color. The ample store of genetic variation in even small isolated human populations shows the fallacy of this definition. However, there are no clear examples of race advocates explicitly using this extreme definition. Recall, even outright essentialists

such as Hooton realized that there is a great deal of variation within groups. Moreover, pernicious social and political concepts such as the 'One Drop Rule' (16) permit a great deal of genetic variation within perceived races, and recent advocates of a genetic basis for racial differences in intelligence wax *ad nauseam* about the variation in intelligence within groups (20). Fifth, recognizing genetic differences among groups can be useful in medicine and public health. Organ transplantation requires careful genetic matching. Failure to recognize genetic differences among populations can create spurious results in genetic epidemiology and mapping disease genes. While some investigators argue that race is a reliable and useful proxy for the pattern of genetic differences (3, 45, 57) others are quite skeptical (10). There is a lively discussion ongoing about the utility of race as a variable in health and medicine (25, 34, 56, 58, 61, 64). Despite the imperfection of race concepts for describing patterns of genetic variation, a universally accepted alternative does not exist.

SUMMARY AND CONCLUSIONS

This paper has presented a broad overview of the patterns of human genetic variation and the processes that generate them. One primary objective was to introduce the concepts, terms, and methods that have recently emerged to meet the needs of studying variation at the level of DNA. A second major objective was to look at the evolutionary record embedded within the human genome. The final objective was to evaluate whether race concepts are consistent with new findings.

The coalescence model is a powerful new concept for exploring the outcomes from evolutionary processes. It enables us to derive expectations that fully utilize the information in DNA sequences, e.g. the expected number of variable sites at a locus. It also enables derivations of traditional allele frequency based measures such as homozygosity and heterozygosity. One interesting result from coalescent models is that common alleles tend to be older. Because of this, at most genetic loci, an allele that is common in people from one geographic region is common throughout the species (43, 44, 50), either because of descent from a common ancestral population, or because it has spread by local gene flow and migration. Another result from coalescent models is that in subdivided populations, the lines of descent for a sample of DNA sequences may not replicate the lines of descent for the population (37). In fact, it is unlikely that trees for DNA sequences and trees for populations will match unless the populations have been diverged for a very long time.

One clear signal from our genome is that the level of diversity is at odds with the large number of people living today. In fact, it is more in line with a population of fewer than 20,000 breeding individuals. This is not a new conclusion. The nucleotide sequences and STRs discussed here (68) agree with findings from many studies and for many other types of genetic data (1, 19, 24, 31, 36, 41, 44, 46, 47, 60, 69, 71). Another robust finding is that the genetic diversity in people outside of Sub-Saharan Africa is for the most part a subset of that found in Sub-Saharan Africans. The STR data confirm that the predominant genetic diversity pattern among populations is one of nested subsets. This pattern is consistent with a model that postulates a succession of ancient founder effects and bottlenecks that occurred as the human species expanded its range and occupied new continents (47). It is perhaps more surprising that the

geographic pattern of genetic diversity departs from the predictions of the isolation by distance model. Others have suggested (27, 59) that a false tree-like appearance can be produced by sampling widely dispersed regions from a geographic continuum, and that once populations are sampled to fill the gaps in geography, the tree-like properties of genetic diversity will disappear. Interestingly, the graphs of genetic kinship versus geographic distance (Figure 11) do not support this suggestion. Isolation by distance cannot explain why the genetic kinship between an African population and a European population is the same as between an African population and a Native South American population, when the distance to South America from Africa is enormously further than the distance to Europe. It therefore appears that the vast store of genetic variation shared by all humans owes to the fact that all humans trace back to a common source population that existed recently on an evolutionary time scale. Moreover, the existing patterns of local gene flow have not persisted long enough to reshape the genetic structure of our species.

It should be noted that the picture for a single gene can be quite different from that given by an average taken over many genes. One reason is that the order and timing of evolutionary change is highly variable because coalescence and mutation both occur at random (19); another reason is that the patterns of genetic variation can look quite different if natural selection has biased the persistence or removal of alleles at a locus. Both the STRs and nucleotide sequences presented here were chosen for analysis because they do not encode functional products that would be subject to natural selection. By contrast, natural selection has left a signature on the variation at several gene loci that encode functional products (18, 26, 49, 50). Two expectations emerge. Population history and relationships are read best from DNA sequences without function. By contrast, DNA sequences that encode expressed genes will show patterns of variation that are more directly related to natural selection and human adaptation.

The nested pattern seen in both nucleotide sequences and STRs reveals a new problem with the use of race to describe patterns of human variation. It contradicts the intuitive expectation that a race classification is symmetrical, i.e. that if A is a race with respect to B, then B is a race with respect to A. For example, we see that the non-African samples analyzed here form a lineage with respect to the African samples, but the African samples do not form a unique lineage. Although race concepts describe the patterns of genetic variation poorly, breaking the tradition of using them will be difficult. Race concepts are used in vague and imprecise ways by both those who embrace them and those who reject them. Neither attacks against nor defenses for race can be successful as long as this practice persists. Despite the inadequacy of race concepts for describing patterns of genetic variation, genetic differences among human populations do exist. Rare alleles are generally found in one local population or one geographic region. Common alleles are generally shared across all populations, but their frequencies differ. Populations differ with respect to the overall level of genetic variation. The organization of this variation is ultimately explained by the evolutionary history of our species and best understood in that context.

Acknowledgements

The author thanks Drs. Suzanne Cole, Malia Fullerton, Chuck Hilton, Keith Hunley, Connie Mulligan, and Yasuko Takezawa for their comments on earlier drafts of this paper. The contents of the paper reflect only the author's views and the author accepts sole responsibility for any errors of omission or commission.

Research support from the National Science Foundation BSC-0321610 is gratefully acknowledged.

REFERENCES

1. Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB. 2003. Human population genetic structure and inference of group membership. *American Journal of Human Genetics* 72: 578-89.
2. Brown RA, Armelagos GJ. 2001. Apportionment of racial diversity: A review. *Evolutionary Anthropology* 10: 34-40.
3. Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, et al. 2003. The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine* 348: 1170-5.
4. Cann HM, de Toma C, Cazes L, Legrand M-F, Morel V, et al. 2002. A Human Genome Diversity Cell Line Panel. *Science* 296: 261b-2.
5. Cavalli-Sforza LL. 1997. Genes, peoples, and languages. *Proceedings of the National Academy of Sciences USA* 94: 7719-24.
6. Cavalli-Sforza LL, Bodmer WF. 1970. *The Genetics of Human Populations.* San Francisco: W. F. Freeman.
7. Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. *The History and Geography of Human Genes*. Princeton: Princeton University Press. 413 pp.
8. Cavalli-Sforza LL, Piazza A. 1975. Analysis of evolution: evolutionary rates, independence and treeness. *Theoretical Population Biology* 8: 127-65.
9. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, et al. 2003. Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios. *Science* 302: 1960-3.
10. Cooper RS, Kaufman JS, Ward R. 2003. Race and genomics. *New England Journal of Medicine* 348: 1166-70
11. Crow JF. 1997. The high spontaneous mutation rate: is it a health risk? *Proceedings of the National Academy of Sciences USA* 94: 8380-6.
12. Crow JF. 2002. Unequal by nature: A geneticist's perspective on human differences. *Daedalus* 131: 81.
13. Dobzhansky T. 1970. *Genetics of the Evolutionary Process*. New York: Columbia University Press.
14. Endler JA. 1977. *Geographic Variation, Speciation, and Clines*. Princeton, NJ: Princeton University Press. 246 pp.
15. Felsenstein J. 2004. *Inferring phylogenies*. Sunderland, Mass: Sinauer Associates. 664 pp.
16. Fredrickson GM. 2002. *Racism - A Short History*. Princeton: Princeton University Press.

17. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, et al. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics* 22: 239-47.
18. Hamblin MT, Thompson EE, Di Rienzo A. 2002. Complex signatures of natural selection at the Duffy blood group locus. *American Journal of Human Genetics* 70: 369-83.
19. Harpending H, Rogers A. 2000. Genetic perspectives on human origins and differentiation. *Annual Review of Genomics and Human Genetics* 1: 361-85.
20. Herrnstein RJ, Murray C. 1994. *The Bell Curve: Intelligence and Class Structure in American Life*: The Free Press.
21. Hooton EA. 1926. Methods of racial analysis. *Science* 63: 75-81.
22. Hudson RR. 1990. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* 7: 1-44.
23. Jorde LB. 1980. The genetic structure of subdivided human populations: A review. In *Current Developments in Anthropological Genetics. Vol. 1. Theory and Methods*, ed. JH Mielke, MH Crawford, pp. 135-208. New York and London: Plenum Press.
24. Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, et al. 1995. Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. *American Journal of Human Genetics* 57: 523-38.
25. Keita SO, Kittles RA, Royal CD, Bonney GE, Furbert-Harris P, et al. 2004. Conceptualizing human variation. *Nature Genetics* 36: S17-20.
26. King RA, Pietsch J, Fryer JP, Savage S, Brott MJ, et al. 2003. Tyrosinase gene mutations in oculocutaneous albinism 1 (OCA1): definition of the phenotype. *Human Genetics* 113: 502-13.
27. Kittles RA, Weiss KM. 2003. Race, ancestry, and genes: implications for defining disease risk. *Annual Review of Genomics and Human Genetics* 4: 33-67.
28. Lewontin RC. 1974. *The Genetic Basis of Evolutionary Change*. New York: Columbia University Press.
29. Long JC, Kittles RA. 2003. Human genetic diversity and the nonexistence of biological races. *Human Biology* 75: 449-71.
30. Malecot G. 1969. *The Mathematics of Heredity*. San Francisco: W. F. Freeman
31. Marth G, Schuler G, Yeh R, Davenport R, Agarwala R, et al. 2003. Sequence variations in the public human genome data reflect a bottlenecked population history. *Proceedings of the National Academy of Sciences USA* 100: 376-81.
32. Mayr E. 1969. *Principles of Systematic Zoology*. New York: McGraw-Hill.
33. Morton NE. 1969. The genetic structure of populations. *Annual Review of Genetics* 3: 53-73.
34. Mountain JL, Risch N. 2004. Assessing genetic contributions to phenotypic differences among 'racial' and 'ethnic' groups. *Nature Genetics* 36: S48-53.
35. Nachman MW, Bauer VL, Crowell SL, Aquadro CF. 1998. DNA variability and recombination rates at X-linked loci in humans. *Genetics* 150: 1133-41.
36. Nachman MW, Brown WM, Stoneking M, Aquadro CF. 1996. Nonneutral mitochondrial DNA variation in humans and chimpanzees. *Genetics* 142: 953-63.
37. Nei M. 1987. *Molecular Evolutionary Genetics*. New York: Columbia University Press.
38. Nei M, Roychoudhury AK. 1993. Evolutionary relationships of human populations on a global scale. *Molecular Biology and Evolution* 10: 927-43.

39. Nordborg M. 2001. Coalescent theory. In *Handbook of Statistical Genetics*, ed. MB D. Balding, and C. Cannings, pp. 179-212. Chichester,UK: Wiley
40. Payseur BA, Cutter AD, Nachman MW. 2002. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Molecular Biology and Evolution* 19: 1143-53.
41. Pluzhnikov A, Di Rienzo A, Hudson RR. 2002. Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* 161: 1209-18.
42. Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945-59.
43. Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. *Trends in Genetics* 17: 502-10.
44. Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, et al. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genetics* 32: 135-42.
45. Risch N, Burchard E, Ziv E, Tang H. 2002. Categorization of humans in biomedical research: genes, race and disease. *Genome Biology* 3: comment2007
46. Rogers AR, Harpending H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution* 9: 552-69.
47. Rogers AR, Jorde LB. 1995. Genetic evidence on modern human origins. *Human Biology* 67: 1-36.
48. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. 2002. Genetic structure of human populations. *Science* 298: 2381-5.
49. Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V, Wallace DC. 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* 303: 223-6.
50. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832-7.
51. Slatkin M. 1991. Inbreeding coefficients and coalescence times. *Genetical Research* 58: 167-75.
52. Slatkin M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* 47: 264-79
53. Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457-62.
54. Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129: 555-62.
55. Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437-60.
56. Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol* 28: 289-301
57. Tang H, Quertermous T, Rodriguez B, Kardia SL, Zhu X, et al. 2005. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *American Journal of Human Genetics* 76: 268-75.
58. Tate SK, Goldstein DB. 2004. Will tomorrow's medicines work for everyone? *Nature Genetics* 36: S34-42.

59. Templeton AR. 1999. Human Races: A Genetic and Evolutionary Perspective. *American Anthropologist* 100: 632-50.
60. Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, et al. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 271: 1380-7.
61. Tishkoff SA, Kidd KK. 2004. Implications of biogeography of human populations for 'race' and medicine. *Nature Genetics* 36: S21-7.
62. Urbanek M, Goldman D, Long JC. 1996. The apportionment of dinucleotide repeat diversity in Native Americans and Europeans: a new approach to measuring gene identity reveals asymmetric patterns of divergence. *Mol Biol Evol* 13: 943-53.
63. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291: 1304-51.
64. Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, et al. 2001. Population genetic structure of variable drug response. *Nat Genet* 29: 265-9.
65. Wolpoff MH, Hawks J, Frayer DW, Hunley K. 2001. Modern Human Ancestry at the Peripheries: A Test of the Replacement Theory. *Science* 291: 293-7.
66. Wright S. 1943. Isolation by distance. *Genetics* 28: 114-38.
67. Wright S. 1969. *Evolution and the Genetics of Populations. Vol. 2. The Theory of Gene Frequencies*. Chicago: University of Chicago Press.
68. Yu N, Chen FC, Ota S, Jorde LB, Pamilo P, et al. 2002. Larger genetic differences within Africans than between Africans and Eurasians. *Genetics* 161: 269-74.
69. Zhivotovsky LA. 2001. Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. *Molecular Biology and Evolution* 18: 700-9.
70. Zhivotovsky LA, Bennett L, Bowcock AM, Feldman MW. 2000. Human population expansion and microsatellite variation. *Molecular Biology and Evolution* 17: 757-67.
71. Zhivotovsky LA, Rosenberg NA, Feldman MW. 2003. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *American Journal of Human Genetics* 72: 1171-86.

