Journal of Heredity 2004:95(5):406–420
doi:10.1093/jhered/esh060

Understanding Human DNA Sequence Variation

K. K. KIDD, A. J. PAKSTIS, W. C. SPEED, AND J. R. KIDD

From the Department of Genetics, Yale University School of Medicine, 333 Cedar St., New Haven, CT 06520-8005.

Address correspondence to Kenneth K. Kidd at the address above, or e-mail: Kenneth.Kidd@yale.edu.

Abstract

Over the past century researchers have identified normal genetic variation and studied that variation in diverse human populations to determine the amounts and distributions of that variation. That information is being used to develop an understanding of the demographic histories of the different populations and the species as a whole, among other studies. With the advent of DNA-based markers in the last quarter century, these studies have accelerated. One of the challenges for the next century is to understand that variation. One component of that understanding will be population genetics. We present here examples of many of the ways these new data can be analyzed from a population perspective using results from our laboratory on multiple individual DNA-based polymorphisms, many clustered in haplotypes, studied in multiple populations representing all major geographic regions of the world. These data support an "out of Africa" hypothesis for human dispersal around the world and begin to refine the understanding of population structures and genetic relationships. We are also developing baseline information against which we can compare findings at different loci to aid in the identification of loci subject, now and in the past, to selection (directional or balancing). We do not yet have a comprehensive understanding of the extensive variation in the human genome, but some of that understanding is coming from population genetics.

Introduction: The First Century

Humans have always been fascinated by the variation among individuals. That certain traits were hereditary has been understood in many different cultural traditions for a long time, but understanding genetics and how those traits were in fact inherited only began a century ago with the rediscovery of Mendel's laws and the demonstration that one inherited disorder, alkaptonuria, followed Mendel's laws (Garrod 1902). Although the ABO blood group was discovered in 1900 by Landsteiner (1900), the actual mode of inheritance was determined by Bernstein in 1924–1925 (as cited in Crow 1993) using population genetics approaches. Over the ensuing decades, a few more common Mendelian traits were identified in humans and other inherited disorders were demonstrated to show single-gene Mendelian inheritance. However, population geneticists had relatively few genetic markers to use to study normal populations. This time after 1925 was not a period of stagnation, however; Fisher, Wright, and Haldane were laying the foundations of theoretical population genetics. Theory was racing ahead of our ability to test/apply that theory with real data. Breakup of this logjam occurred in 1966 with the demonstration by Harry Harris that many red cell enzymes and serum proteins showed electrophoretic variation and could be used for population genetics studies (Harris 1966). Suddenly available data for studies of human populations resulting from Harris' discovery, and that of Hubby and Lewontin (1966), demonstrating the same high level of polymorphism for *Drosophila*, were beginning to catch up with theory. Human populations from around the world have subsequently been studied for many of those normal polymorphisms and a large compendium of the data collected over the ensuing quarter century has been published (Cavalli-Sforza et al. 1994).

Those markers are now generally considered collectively as the "classical" polymorphisms because, starting in 1978, a new surge in identification of polymorphisms, this time directly in the DNA, began with the discovery of a DNA sequence polymorphism near the beta globin gene (Kan and Dozy 1978). By 1980 it was obvious that enough DNA-based polymorphisms (then called restriction fragment length polymorphisms, RFLPs, because of the methodology for studying them) would be available to generate a linkage map of *Homo sapiens* (Botstein et al. 1980). It was felt that with a genetic linkage map in place, any genetic disease or trait could be mapped and the genes causing them could be identified. During the 1980s the Human Gene Mapping Workshops catalogued the numbers of RFLPs that had been identified, and demonstrated an exponential growth in those numbers (Williamson et al. 1991).

In 1985 Jeffreys et al. described and focused attention on a different type of polymorphism, the hypervariable "minisatellite" regions. Some multiallelic loci described previously (e.g., D14S1) (Wyman and White 1980) were of this type, involving segments of DNA that are dozens of nucleotides long and are tandemly repeated a large but variable number of times. These markers are also known as variable number of tandem repeats (VNTRs) (Nakamura et al. 1987). In early applications of DNA technology to forensics, these VNTRs were used extensively. Some population genetics studies were done with VNTRs, mostly to define the distributions, essentially continuous, of allelic sizes for forensic statistics. Relationships among populations were difficult to infer using VNTRs because allelic similarity or correspondence between populations was difficult to determine unless individual alleles were analyzed for the pattern of sequence variation among the individual repeated units, as was done by Armour et al. (1996).

In 1989 another quantum leap forward was made with the simultaneous discovery by two research groups of a new class of polymorphism, the short tandem repeat (Litt and Luty 1989; Weber and May 1989). These markers are multiallelic, varying in the number of repeated units on a DNA strand, but in contrast to the VNTRs, the units are small. Short tandem repeat polymorphisms (STRPs, also called microsatellites) have higher heterozygosities and are statistically much more informative for most studies than biallelic markers. They are also much more common in the human genome than VNTRs. It was these STRPs that were the basis for the first comprehensive human linkage map (Dib et al. 1996) and the best linkage map to date (Kong et al. 2002). STRPs are also now the standard in many forensic applications (Budowle et al. 1998, 2001).

After the early 1990s, printed compendia of all known polymorphisms were no longer maintained and instead the National Center for Biotechnology Information (NCBI) instituted a database, dbSNP (Sherry et al. 2001), with information on DNA sequence variants. Although dbSNP includes only information submitted to it, and many of the variants are not true polymorphisms, but rare variants or even sequencing errors, it now contains 5.8 million distinct single nucleotide polymorphisms (SNPs), of which 2 million have been validated (Build 116, August 2003). Other databases with subsets of the information in dbSNP and other types of information that researchers find useful also catalogue DNA sequence variation, including HGVbase (Fredman et al. 2002) and JSNP (Hirakawa et al. 2002). One database, ALFRED, the Allele Frequency Database (Osier et al. 2002; Rajeevan et al. 2003), is specifically devoted to cataloging allele frequencies for DNA-based polymorphisms in defined populations. None of those sources of information catalogues all of the polymorphisms of any one type that are currently known and described somewhere in the scientific literature, but any one of them will identify a very large number of polymorphisms. Thus today there are millions of SNPs known, of which at least thousands are known to have at least moderate levels of heterozygosity in at least one population. In addition to SNPs, there are thousands of small insertion/deletion polymorphisms (indels) and thousands of short tandem repeat polymorphisms (STRPs). A recent estimate is that there are at least 4800 VNTRs or minisatellites, longer-sequence tandem repeats that are polymorphic, a few of which are so highly polymorphic they would appear to qualify as hypermutable (Denoeud et al. 2003). One century after the first human genetic polymorphism was discovered we have a plethora of genetic variation detectable at the DNA level in humans.

The Next Century

One of the challenges for the next century formulated by the National Human Genome Research Institute is understanding this variation: "Grand Challenge I-3: Develop a detailed understanding of the heritable variation in the human genome" (Collins et al. 2003). Understanding that variation has many components. How much of that variation has any functional significance? Of the variation that does have some functional relevance, what does it do? How do those functional variants operate to cause differences in the development or metabolism of individuals? What fraction of the variation is common? How is the variation distributed within the species and what factors—selection, mutation, migration, and/or random genetic drift—caused the distribution of variation that we see? Thus one must be inclusive of all *Homo sapiens* and carefully examine the genetic variation that occurs within the various ethnic groups and the variation that is different among the ethnic groups. Such studies become very suspect because of the numerous traditions of ethnic and racial discrimination among human populations, but it is important to recognize the reality and understand its significance. It is important also to recognize that such statements as "We hold these truths to be self-evident, that all men are created equal. . ." (U.S. Declaration of Independence, 1776) are statements of morality and individual rights, not statements of science. While one can accept the equality of all individual humans as a moral imperative, one must also recognize from the biological perspective that all human beings independently conceived are genetically unique. It was eloquently stated by James F. Crow (2001): "In a sexual population, each genotype is unique, never to recur. The life expectancy of a genotype is a single generation. In contrast, the population of genes endures."

While recognizing the genetic uniqueness of every human, we must also recognize the great genetic similarity of all humans. Current estimates of how much variation occurs species-wide are poor, but are on the order of 1 nucleotide in 500 to 1000 differ between two copies chosen at random from the population. A different pair might differ at other nucleotides so that the varying fraction of the genome may be closer to 1 in 200 to 1 in 500 nucleotides. Thus *Homo sapiens* has 99.5% to 99.8% of its genome essentially identical in everyone. Of course, that 0.2% to 0.5% of 3 billion nucleotides is still a very large number—6 to 15 million. Thus roughly 10 million variants that can potentially occur in all different combinations is vastly more

than enough to ensure individual uniqueness at the DNA level while still representing a very small fraction of the total genome. Of course, not all combinations occur. If one considers a few of the variants in a small segment of DNA, it will often be the case that only a few of the possible combinations are found. This nonrandomness is itself an interesting phenomenon of importance, as discussed later. However, in the genome at large, all possible combinations of these small regions can potentially occur. The limiting factor is the number of humans—many more combinations are possible than there are people. Much of this variation occurs in DNA of no known function and may not affect the phenotype of individuals. However, dbSNP documents many SNPs in coding regions of genes, altering amino acid sequences of proteins. Additional variation exists in regulatory regions and may often have an effect on the expression of the gene. This expressed variation, known to alter gene products and function, is part of normal genetic variation and is itself likely sufficient to ensure individuality.

In the remainder of this article, we present data on common sequence variation, that is, DNA polymorphisms, studied on large numbers of individuals from around the world to illustrate the types of data that are becoming available and are beginning to help us to understand some of the factors that are involved in determining the ways that common genetic variation is distributed in human populations around the world. Simply documenting the patterns of variation is only the initial step toward the deep understanding that is one of the "grand challenges" for this next century. Interpretation of these patterns will be the way we meet that challenge.

Exploring Human Variation

The following examples provide an overview of several aspects of global human genetic variation. These examples use data we have been collecting at Yale over the past several years. This growing dataset is among the more comprehensive datasets in terms of multiple independent loci studied on multiple populations representing most major regions of the world. The populations being studied are listed in some of the following figures and tables, and descriptions of the populations and the individual samples studied can be found in ALFRED. Analyses have been done on 38 populations, averaging slightly more than 50 individuals per population. Different numbers of polymorphisms were involved in the various analyses presented, but all involved more than 100 individual polymorphisms. Many researchers have contributed to this dataset (see Acknowledgements) and continue to be coauthors on various more detailed analyses of these data, in part or in whole.

Classical markers show considerable gene frequency variation among populations, and DNA-based markers are no different in that respect. While early studies of classical markers tended to attribute selective factors to the frequency differences, it is increasingly clear that much of the variation in allele frequencies is simply the result of random genetic

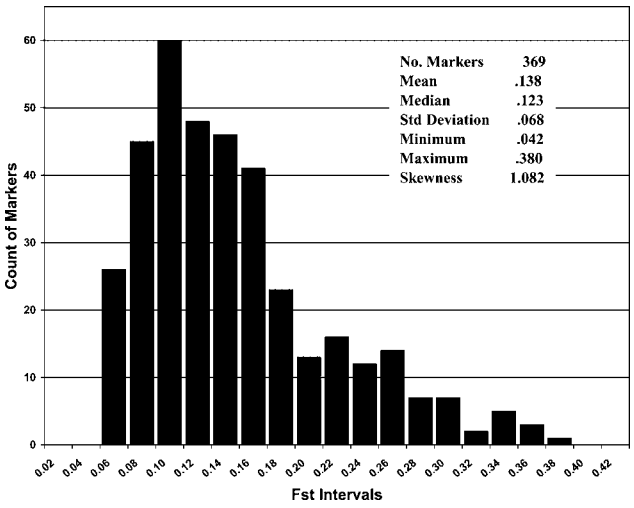

Figure 1. F_{ST} distribution for 369 biallelic markers. Values are calculated for 38 populations, all of those in Figure 2. This represents an update of the 94 sites described in Pakstis et al. (2002).

drift, with different populations showing different frequencies because of limited gene flow between populations. Indeed, for most of the DNA polymorphisms that have been identified, it is exceedingly difficult to imagine how selection could directly operate to alter allele frequencies since most occur in noncoding regions of the DNA and in regions of no known or plausible regulatory function. However, the phenomenon of hitchhiking—changes in allele frequencies at loci tightly linked to a locus under directional selection—could account for frequency changes at some of these presumably neutral markers (Maynard Smith and Haigh 1974). Some examples of hitchhiking exist (e.g., Sabeti et al. 2002).

One standard measure of gene frequency variation among populations is the statistic F_{ST} (Wright 1969). This statistic can be related theoretically to random genetic drift as a function of time and effective population size, but can more simply be considered as a measure of the relative amounts of variation among populations shown by different genetic polymorphisms. Figure 1 shows the distribution of the F_{ST} statistic for 369 individual polymorphisms across 38 populations representing all major continental regions. All sites have been studied on the same individuals in all of these populations; none of these sites has any known or likely functional relevance. The distribution is essentially unimodal, with a mean of 0.138, a range of 0.042 to 0.380, and is obviously skewed. It is expected that loci independently subject to random genetic drift through identical historical demography, as is the case here, would show a distribution of realized variation among populations. However, it is not immediately obvious why this distribution is skewed. Other studies of multiple loci in smaller numbers of ethnic groups also show a similar skewed distribution (Bamshad and Wooding 2003). This distribution may have been affected by

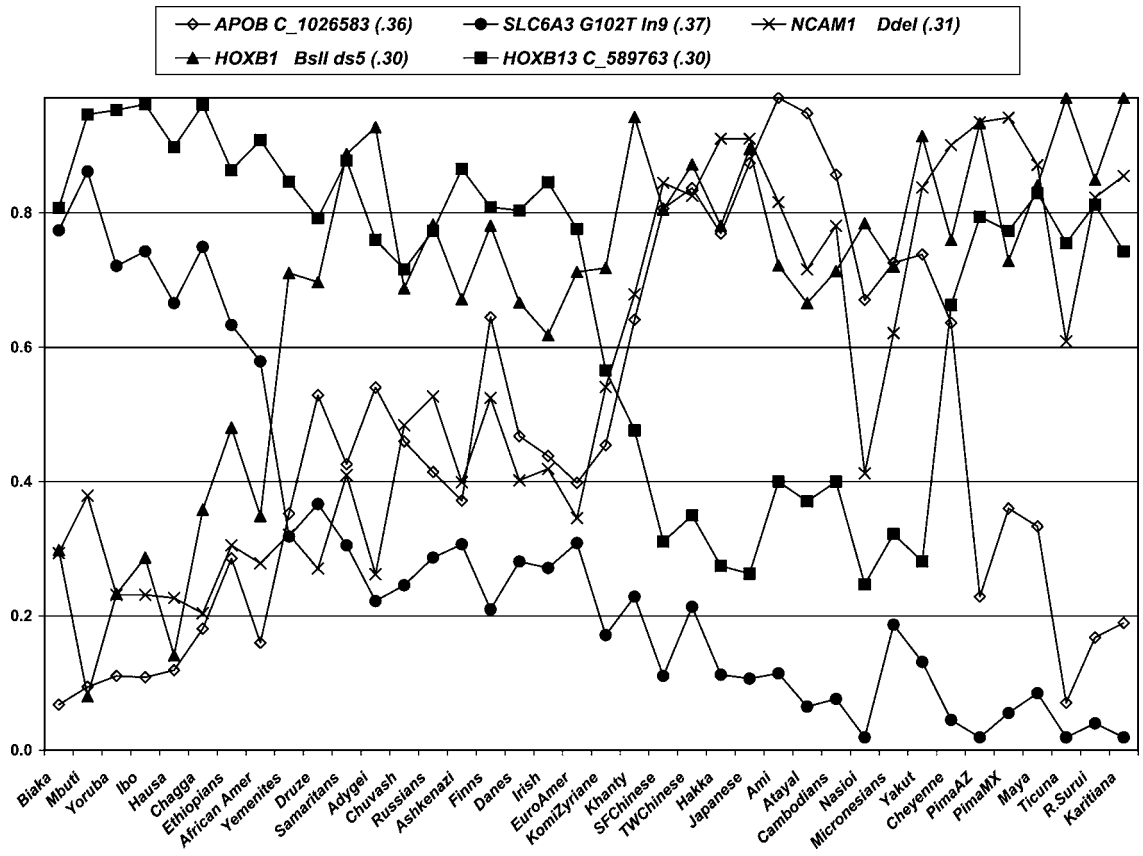

Figure 2. Allele frequencies of five SNPs from the upper tail of the F_{ST} distribution in Figure 1. Frequencies are given for the 38 populations indicated.

the complex ascertainment of the markers included. That is an area to be investigated. For example, almost all markers selected for study were initially shown to be polymorphic with moderate to high heterozygosities in a non-African population and/or to be polymorphic in both African and non-African populations.

While none of the sites in this figure seems a likely candidate for being subject to selection, a priori, one would suspect that loci subject to strong positive selection for one allele in one geographic region would fall at the far upper end of this distribution, while loci subject to balancing selection everywhere would fall at the lower end of this distribution. Thus recently investigators have used F_{ST} on various datasets to identify sets of loci that might be targets of such selective forces (reviewed in Bamshad and Wooding 2003). Unfortunately we do not understand fully historical human demography and hence how high the upper tail of this skewed distribution of neutral markers can extend. Thus, while loci that seem fairly definitely to have undergone selection, such as *ADH1B* (Osier et al. 2002) and *ALDH2* (Oota et al. 2004), show very high F_{ST} values, one must in general consider exceedingly high F_{ST} values as only weak evidence that selection has operated.

The presumably neutral markers, giving rise to the distribution in Figure 1, show a variety of different patterns of variation among populations. Figure 2 shows the allele frequency variation for several of the markers from the upper tail of the distribution. Some of these markers show one allele to have high frequency in Africa, but low frequency elsewhere. Others have an allele with low frequency in Africa and the Native American populations, intermediate frequency in Europe, and high frequency in East Asia. In contrast, the markers at the low end of the F_{ST} distribution tend to have low heterozygosities everywhere, but a few have quite high heterozygosities in populations all around the world (Figure 3). High heterozygosity but low F_{ST} is unusual, but cannot necessarily be considered the result of balancing selection operating globally. Again, understanding how frequently such patterns can occur by chance alone is important in identifying markers subject to global balancing selection. Markers such as the hemoglobin S allele are subject to balancing selection only in regions of endemic malaria, and the allele is only present in areas of endemic malaria except for populations, such as African Americans, that have more recently moved to another location. Such loci may have quite intermediate values of F_{ST} because the allele frequencies in malarial areas are not sufficiently high for the F_{ST} values to be very high.

Understanding historical demography and the consequences it has on genetic variation is another important area.

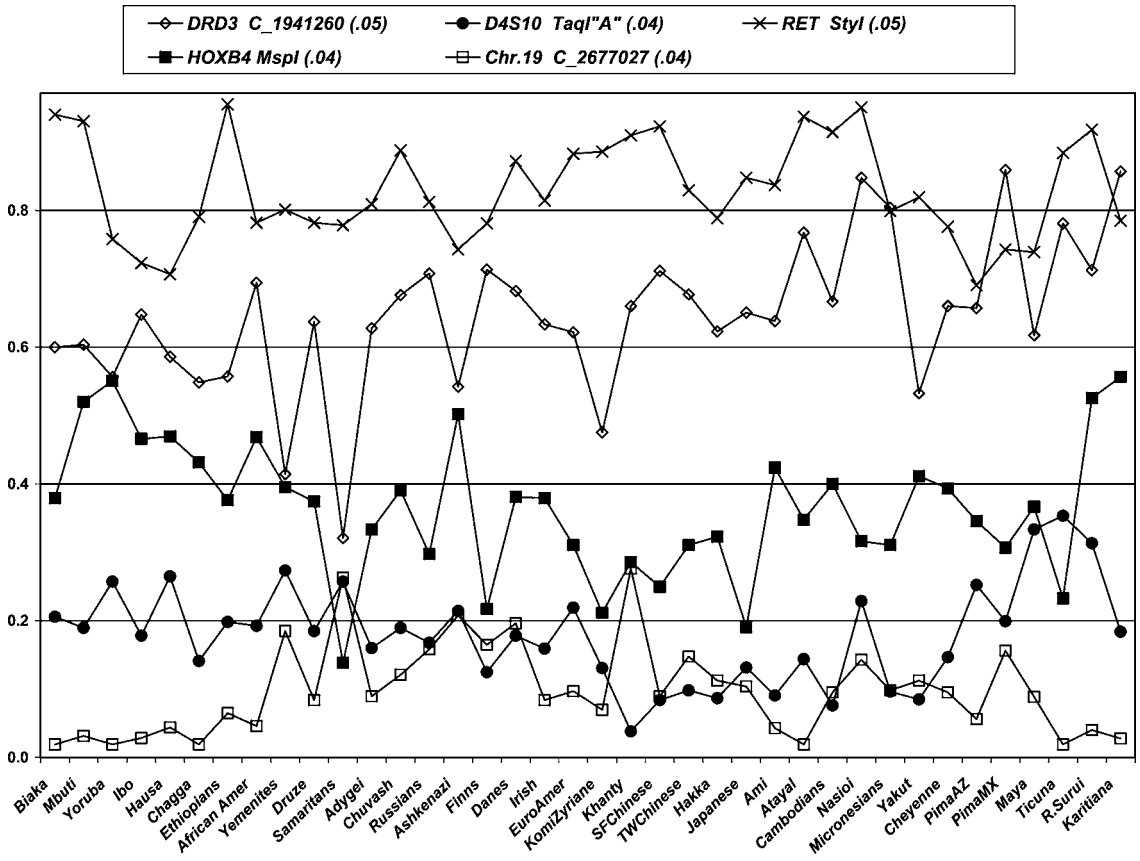

Figure 3. Allele frequencies of five SNPs from the lower tail of the F_{ST} distribution in Figure 1. Frequencies are given for the 38 populations indicated.

Figure 4 shows the average heterozygosity across 50 independent loci for 38 populations. It is interesting to compare not only the geographic regions, but also the populations within a geographic region. First, in comparing geographic regions, we notice that Europeans have the highest average heterozygosity. This is almost certainly the result of ascertainment bias in many of the polymorphisms being studied. Many of these polymorphisms were identified in Europeans, and those polymorphisms with high heterozygosity in Europeans were preferentially identified in the small screening sets used for polymorphism discovery. Within Africa, the two Pygmy groups have the lowest heterozygosity, and the Mbuti markedly so. This particular Pygmy group is relatively smaller and more isolated than the other African populations studied and hence the lower heterozygosity is not unexpected. In the Middle East, the Samaritans show the lowest heterozygosity. Again, this is not surprising, because in addition to being essentially reproductively isolated for at least the past 2000 years, the population had dwindled to only about 100 individuals at the beginning of the 20th century. Although the population is larger now, the consequences of that isolation and recent bottleneck are evident. In East Asia, the population with the lowest heterozygosity is the Atayal, one of the aboriginal populations on Taiwan. Again, this has never been a particularly large population and is but one of several aboriginal populations with considerable endogamy on the island of Taiwan. The sample of Melanesians, the Nasioi from the island of Bougainville, is similarly a small population, one of many distinct populations speaking mutually unintelligible languages and hence quite endogamous on the island of Bougainville in the northern Solomon Islands. Finally, we notice that all three South American groups show relatively low average heterozygosity. Not only does there seem to be a reduction in genetic variation associated with the migration into South America, but each of these is a relatively small group. The Karitiana, in particular, are essentially only one large family of 150 individuals, all of whom are related in many ways in the last few generations (Kidd et al. 1993).

Ancestral Alleles

With the advent of polymerase chain reaction (PCR) and rapid and relatively inexpensive sequencing of DNA, it is easy to obtain sequences in other higher primates for regions homologous to any specified human sequence. We showed in the very early days of PCR that cross-species PCR worked most of the time in apes using primers designed from human

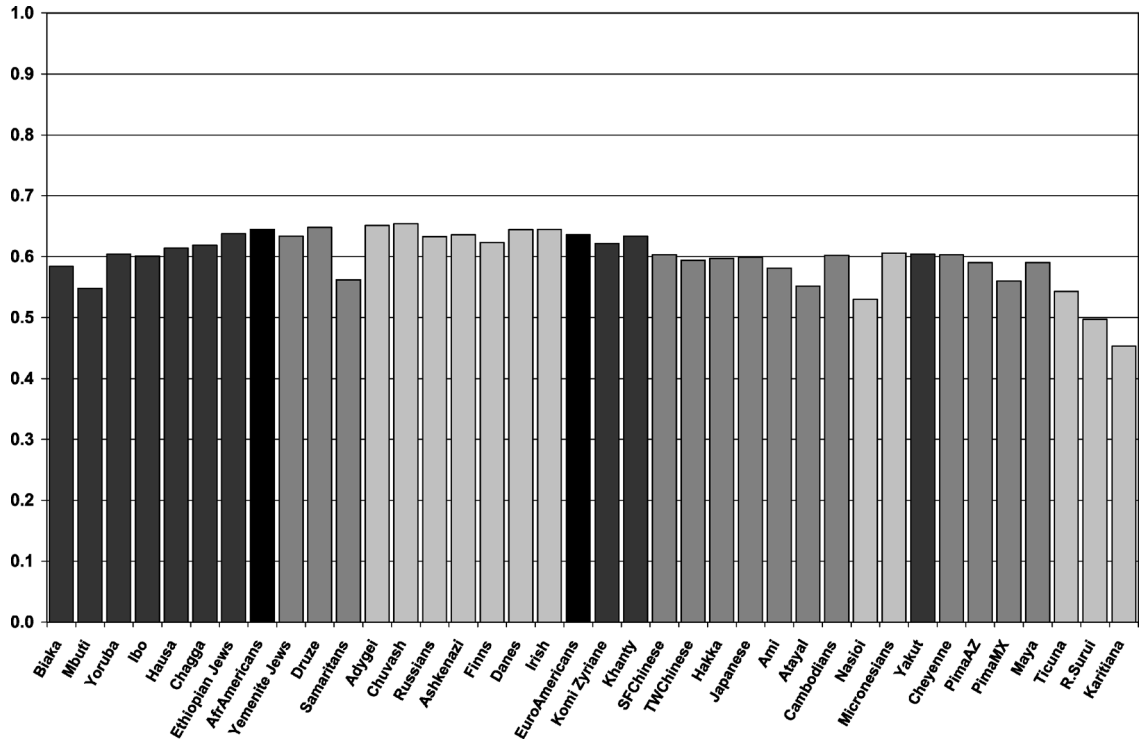

Figure 4. Average heterozygosity of populations based on 50 loci. These 50 loci represent a snapshot of the data continuing to accumulate on these populations and involve a total of 154 different site polymorphisms (145 SNPs and indels plus 9 STRPs). These loci are distributed across 14 autosomes and the sites are organized for this analysis as 29 multisite haplotype loci (from two to nine sites) and 21 single-site loci. For the analysis of the multisite haplotypes and the single-site multiallelic polymorphisms, alleles that were rare or low in frequency across all population samples were pooled together. Thus, of the 515 alleles (haplotypes) analyzed, 465 are considered statistically independent. Because most of these loci are multiallelic haplotyped loci, the average heterozygosity is greater than 0.5 in almost all populations.

sequence (Ruano et al. 1989). That ability now allows us to determine the ancestral alleles for most human SNPs (Hacia et al. 1999; Iyengar et al. 1998). This has become routine, but is not always perfect. Iyengar et al. (1998) gave an example in which the ancestral state was indeterminate, even with sequence from several different ape species. Hacia et al. (1999) showed that occasionally one other ape shared the same polymorphism as humans. If that were the only ape sequenced, inference would not be possible. We have unpublished results of a very few cases where the two closest ape genera, *Pan* and *Gorilla*, have different nucleotides as fixed differences (in samples of several individuals) and those correspond to the two alleles in humans. While those situations are usually resolved by sequencing an orangutan, they do show that one individual of one nonhuman species does not provide infallible evidence of the human ancestral state. The situation of recurrent mutation is especially likely for a C/T polymorphism at the hypermutable CpG dinucleotides. We have one example of both *Pan paniscus* and *Gorilla gorilla* polymorphic for the same G/A polymorphism seen in humans, but based on other evidence, it is clearly due to recurrent mutation at a CpG (unpublished results). Still, one can be essentially certain of the ancestral allele if both chimpanzee and gorilla agree with one of the human alleles.

Figure 5 shows the frequency distributions for the ancestral alleles at 178 SNPs for groups of populations from different geographic regions. If we dichotomize the distribution and consider whether the ancestral allele is the more frequent or less frequent allele, the difference between African populations and virtually all non-African populations is highly significant ($P \ll 10^{-3}$). Clearly the neutral theory cannot explain those different distributions with a single set of parameters (Watterson and Guess 1977). Some of the difference can be attributed to ascertainment bias for SNPs heterozygous in non-African populations, but what fraction of the difference is unclear because the ascertainment has been unsystematic and often undocumented for many of these SNPs. Nor does ascertainment bias explain the ancestral allele being significantly more common in Africa. Hacia et al. (1999) found the ancestral allele to be more common in a "diverse human population set" 75.7% of the time for a set of 214 SNPs. That is a higher value than we see in any population for our independent set of SNPs, and the distributions we see are different from the one they observed. Thus, while the two studies agree that the ancestral

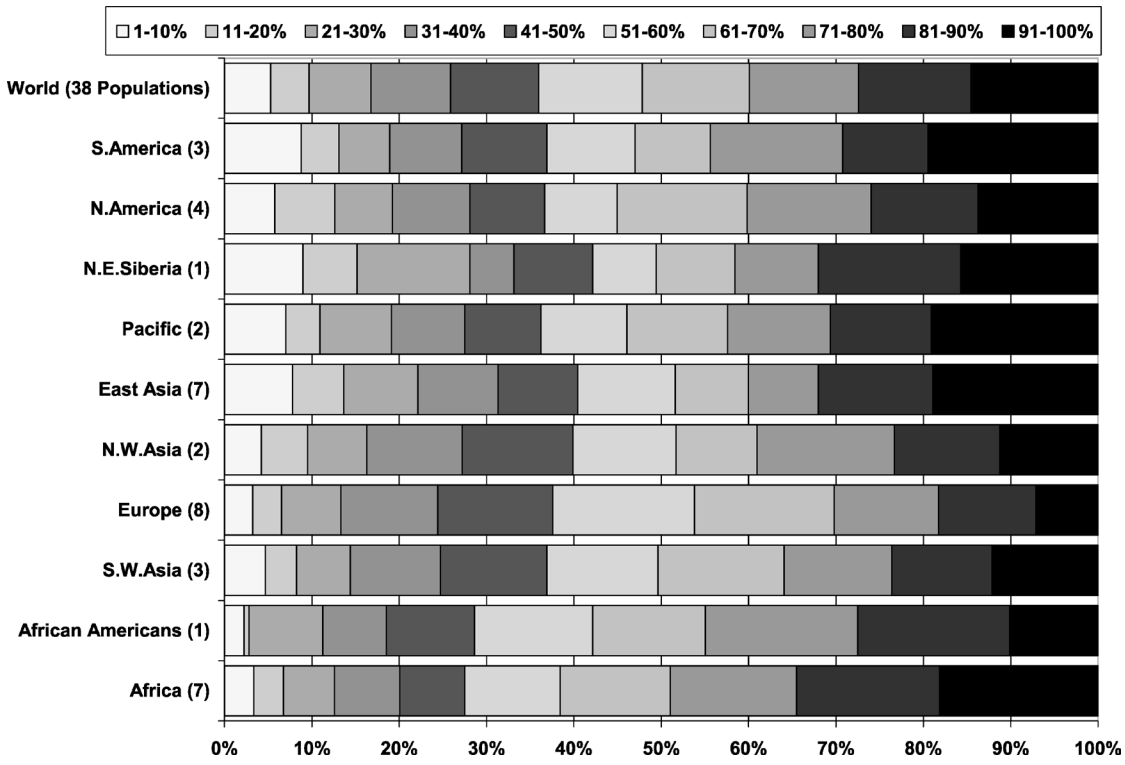

Figure 5. Distributions of ancestral alleles by geographic region. The length of each interval is the fraction in 38 populations of the 178 ancestral alleles that occurred in that frequency range.

allele is usually the more common one, the differences illustrate the difficulties in relating these empiric data to theoretical predictions.

Patterns of Haplotype Frequencies

Allele frequency variation, as illustrated in Figures 2 and 3, is essentially the consequence of random genetic drift. Recurrent mutation is too rare to have had any systematic effect, and for the bulk of such markers, differential selection is a highly unlikely phenomenon (though balancing selection is possibly more likely for the low F_{ST} polymorphisms). When one begins to consider haplotypes, two additional considerations become highly relevant: the order in which sites have mutated to become polymorphic and recombination among the polymorphic sites. If recombination is effectively zero, the haplotypes in the population can be determined by the order in which mutations accumulated and the background on which each mutation occurred plus random genetic drift to make some of those haplotypes frequent in the population. There are loci, such as the myotonic dystrophy locus, where for three polymorphic sites, the two common chromosomes are the haplotype with the ancestral allele at each of the sites and the haplotype with the derived alleles at all three sites (Tishkoff et al. 1998). Extremely few chromosomes are found for any of the intermediate possibilities. In such situations, it is impossible to determine the order in which mutations accumulated and whether or not recombination was required to produce the triply derived haplotype. In other cases, many of the intermediate forms exist and it is possible to determine how the common forms arose because recombination is very low across the segment being analyzed. One such example is the *BRCA1* gene, which in European populations has only two common haplotypes that differ at multiple sites (e.g., Bonnen et al. 2002). However, in African populations there are several other haplotypes that allow the sequence of accumulation of mutations to be determined (Kidd et al. 2003). One pattern, however, seems to be consistent, and that is that sub-Saharan African populations have more haplotypes at common frequencies than do non-African populations. As with any generalization, there are exceptions, but the rule is well supported (e.g., DeMille et al. 2002; Gabriel et al. 2002). Other than that generalization, different haplotypes show a variety of different global patterns. Three haplotypes will serve to demonstrate these differing patterns. Schematics of the three loci and the polymorphisms studied are illustrated in Figure 6; haplotype frequencies for those loci are graphed in Figure 7.

Figure 7A gives the regional average frequencies of different haplotypes defined by five biallelic, polymorphic sites across 76 kb of the *HOXB* cluster. Fourteen of the 32 possible haplotypes occur at a frequency of at least 5% in at least one of the 38 populations studied. While almost all regions of the world have many of these 14 haplotypes, only

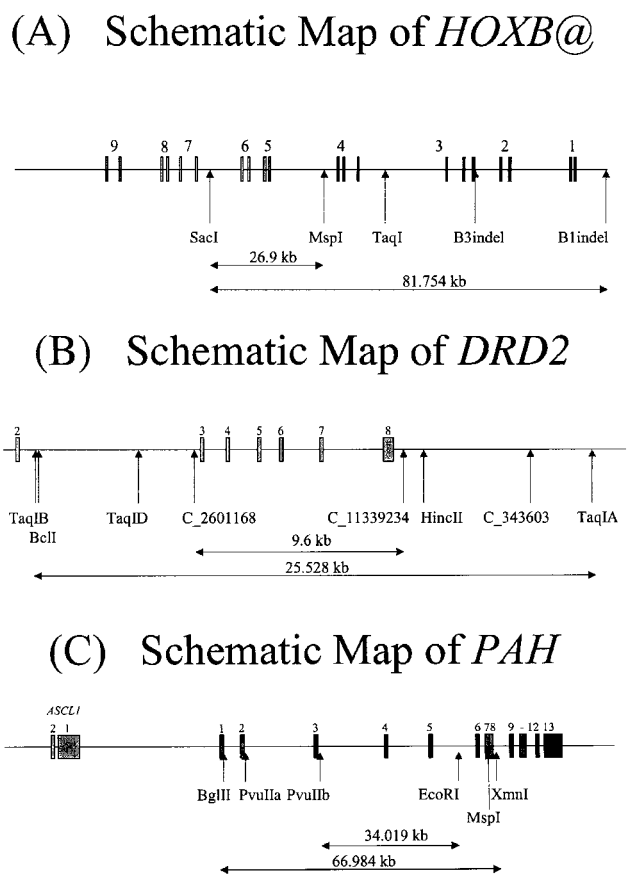

Figure 6. Schematic maps of polymorphic sites at three haplotyped loci: **(A)** *HOXB*, **(B)** *DRD2*, and **(C)** *PAH*.

in Africa and in African Americans do we see six of those haplotypes at quite common frequencies and no single haplotype accounting for a disproportionate fraction of the chromosomes. This haplotype system is unusual in that outside of Africa, all of the regions have the same three to five haplotypes as the most common and a single haplotype at nearly 50% everywhere. Thus, except for Africa, this haplotype system would appear to have a quite low F_{ST}, analogous to some of the SNPs in Figure 3.

A quite different pattern is presented in Figure 7B by the haplotypes at *DRD2*, comprised of eight biallelic sites across 25 kb. Here, 12 of the 256 possible haplotypes account for about 90% of the chromosomes in most parts of the world. The predominant haplotype in Africa is not the predominant haplotype in Europe and southwest Asia and the predominant European haplotype is very uncommon in east Asia and in the Americas. The most common haplotype in the Americas and second most common in east Asia is in fact a fairly uncommon haplotype in sub-Saharan Africa. Global patterns such as this are difficult to explain except through random genetic drift. Previous attempts to determine the evolutionary pathway for haplotypes comprised of the restriction sites indicated in Figure 7B could not distinguish between alternative scenarios of mutation, recombination, and recurrent mutation (Castiglione et al. 1995; Kidd et al. 1998; Lobos and Todd, 1998).

The six sites at *PAH* define a third global pattern (Figure 7C). In common with most loci, and confirming an earlier study of *PAH* with only four SNPs (Kidd et al. 2000), there are more haplotypes at moderate frequencies in African populations and in African Americans than in non-African populations. However, outside of Africa, the major regions of Europe, east Asia, and the Americas exhibit three different patterns. The same two haplotypes are the most common ones in Europe and east Asia, but in the Americas a third haplotype becomes the most common haplotype. Additional haplotype data involving a seventh SNP and an STRP at *PAH* are presented in Kidd and Kidd (2003).

In all of these examples, what is not shown is the level of variation within the larger geographic regions of Africa, Europe, east Asia, and the Americas. Just as there are more common haplotypes in Africa as a whole, each individual African population has multiple common haplotypes. However, there is considerable variation in the frequencies of those haplotypes among the African populations. That variation can be seen from the data on these haplotypes in ALFRED. In contrast to the African populations, there is relatively less variation among European populations, among the east Asian populations, and among the Native American populations. This can also be seen in the graphical representation of the frequencies of these haplotype systems in ALFRED. As typified by these three haplotype systems, a variety of different patterns of haplotype distribution occur among loci from different parts of the genome.

Nonrandomness Along the DNA

Linkage disequilibrium (LD) is the term usually used to describe nonrandom combinations of alleles at multiple sites on chromosomes in the population. LD is not an all or nothing phenomenon, but ranges from nearly random combinations of alleles to complete correlation such that a given allele at one site always occurs with a specific allele at another site and vice versa. A variety of statistics can be used to quantify the level of nonrandomness in any segment of DNA (Devlin and Risch 1995; Zhao et al. 1999). As one can imagine from differences in haplotype frequencies, these statistical measures, which are based on those haplotype frequencies, show differences in the magnitude of LD among populations from different parts of the world. Thus polymorphic sites that show a nonrandom association of alleles in populations outside of Africa often show random or nearly random association of alleles in populations within Africa. Any general statement is difficult to formulate because recombination rates differ by orders of magnitude across segments of DNA such that loci differ considerably in strength and pattern of LD, as can be inferred from Figure 7.

Figure 8 shows the distribution of LD across 130 segments of the genome. For each major geographic region, the values plotted are the average LD values for all populations in the region for all DNA segments falling into

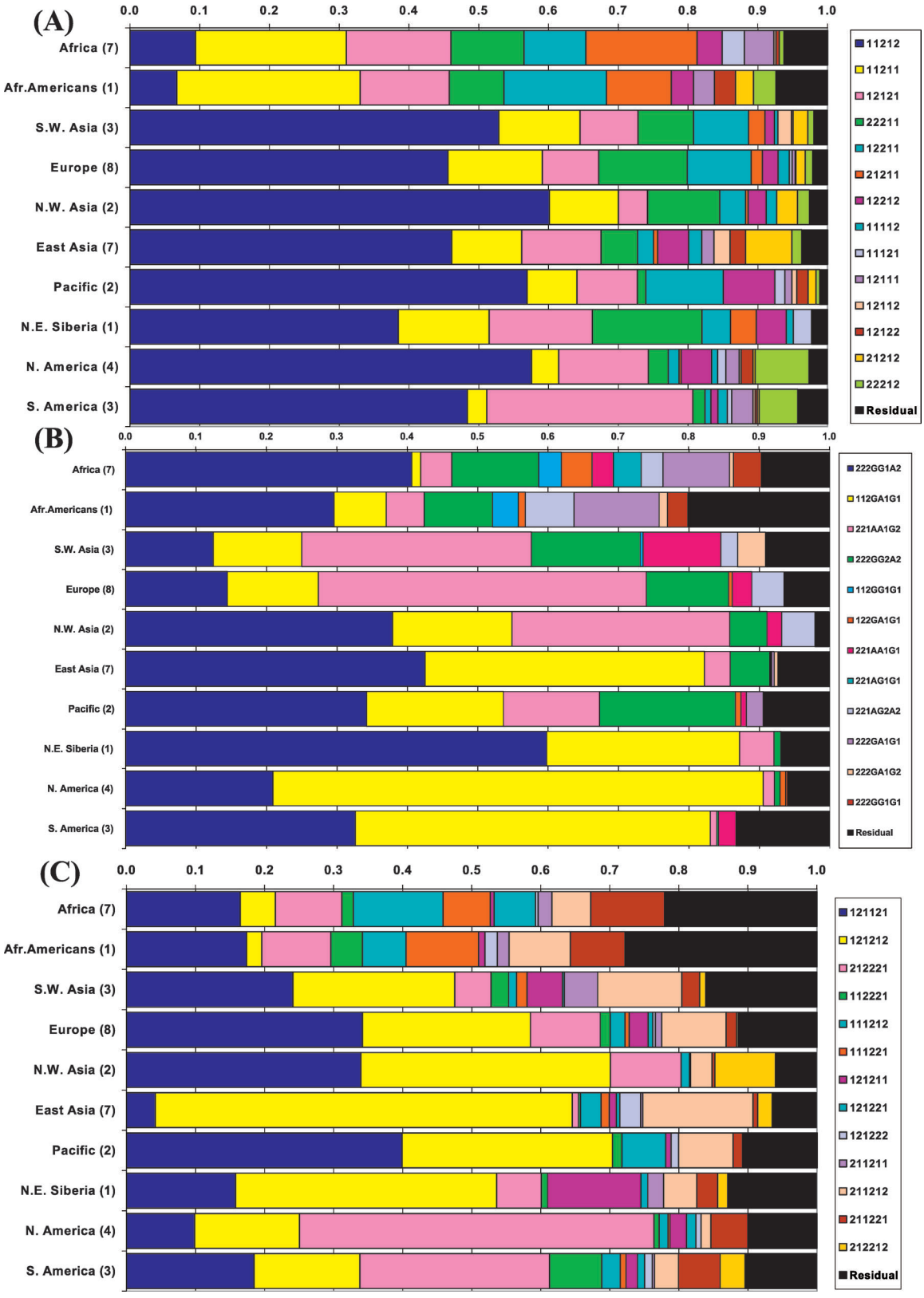

Figure 7. Regional average haplotype frequencies for three haplotyped loci. The haplotypes are defined by the allelic combinations for the sites illustrated in Figure 6 in the order illustrated. **(A)** *HOXB* cluster haplotype frequencies—five sites across 76 kb; **(B)** *DRD2* haplotype frequencies—eight sites across 25 kb; **(C)** *PAH* haplotype frequencies—six sites across 75 kb. Details of the sites and haplotypes can be found in ALFRED under each specific locus.

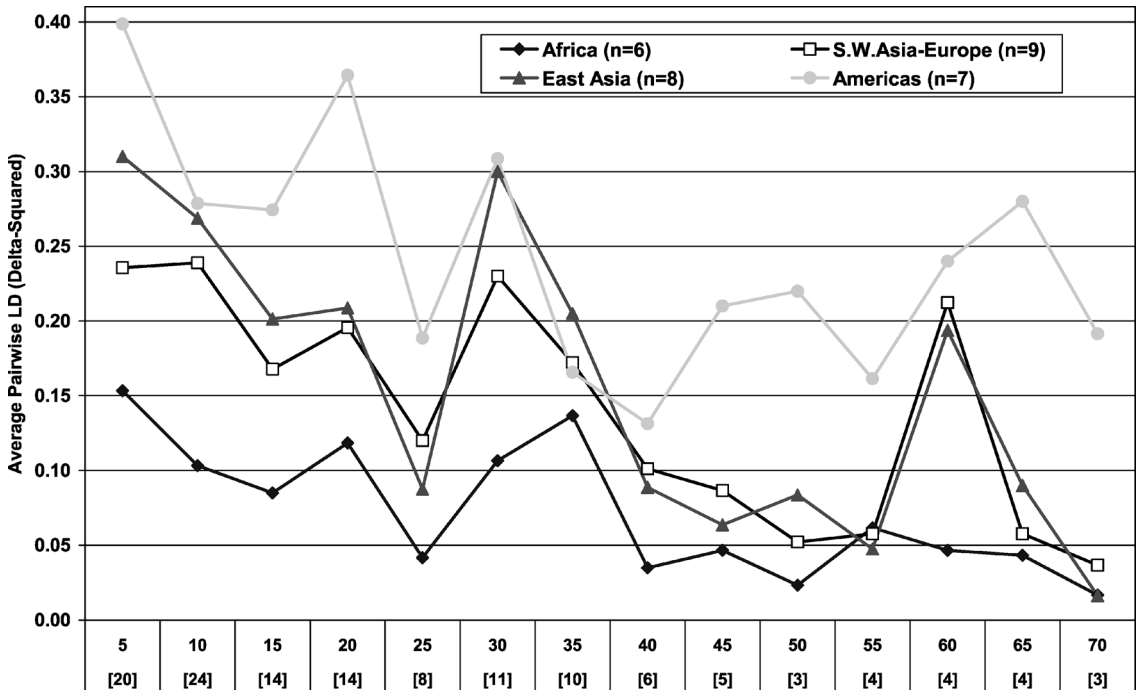

Figure 8. Decline of LD with distance. The measure of LD used is Δ^2 (Devlin and Risch 1995). The 130 segments derive from 21 regions of the genome and include 62 segments that are nonoverlapping, as well as 68 segments that are composed of two or more adjacent smaller segments. The DNA segments are grouped by length in increments of 5 kb; the number of independent, nonoverlapping segments in each group is indicated in brackets. All populations are evaluated for all segments and the averages by geographic region and segment length grouping are graphed. The size (in kb) of each interval is shown across the *x* axis.

each 5 kb interval. There are obvious irregularities attributable to different regions having inherently different values of LD. On the other hand, for each 5 kb interval region, exactly the same segments of DNA are involved for all populations, allowing meaningful comparisons among populations. Thus one can conclude that different populations have different average amounts and extents of LD and a geographic pattern emerges. In summary, one can say that African populations rarely show significant LD for segments larger than 20 kb, European and east Asian populations generally show significant LD up to 35 kb, and Native American populations show significant LD up to 70 kb. Analyses of both a subset and a superset of the six sites at *PAH* in Figure 7C have shown quite definitively this pattern of LD extending different distances in different regions of the world (Kidd and Kidd 2003; Kidd et al. 2000). In those analyses, it is clear that during the course of the expansion of modern humans, different crossover events have occurred at a sufficient frequency that, interacting with random genetic drift, diverse combinations of alleles have arisen in Africa, resulting in little LD. In contrast, only a subset of those haplotypes is common in the rest of the world, presumably because of a founder effect in the expansion out of Africa. However, the frequency of recombination has not had time outside of Africa to regenerate the diversity seen within Africa. This picture is seen at multiple loci in addition to those shown in Figures 6 and 7: *CD4* (Tishkoff et al. 1996), *DM* (Tishkoff et al. 1998), *COMT* (DeMille et al. 2002), and *RET* (Chattopadhyay et al. 2003).

Genetic Similarities Among Populations

The individual SNP frequency variation in Figures 2 and 3 and the haplotype frequency variation by geographic region in Figure 7 show there is a tendency for populations that are geographically closer to have allele and haplotype frequencies that are more similar. On a global basis there are many methodologies available for studying genetic similarity among populations. Figure 9 shows principal component plots of the 38 populations in Figure 4. Figure 9A shows the genetic relationships based on pairwise genetic distances calculated using the same 50 loci as in Figure 4. This two-dimensional representation of those pairwise genetic distances accounts for 62.6% of the total variation among the populations. The third principal component accounts for an additional 11.5% and primarily separates the east Asian from the Native American populations in the third dimension. The most informative aspect of this view is the comparison with the geographic distances represented by the same statistical analysis in Figure 9B. In this panel, the pairwise distances analyzed are geographic "migrational" distances requiring that large bodies of water be avoided, but otherwise using the shortest great circle distances between populations. This

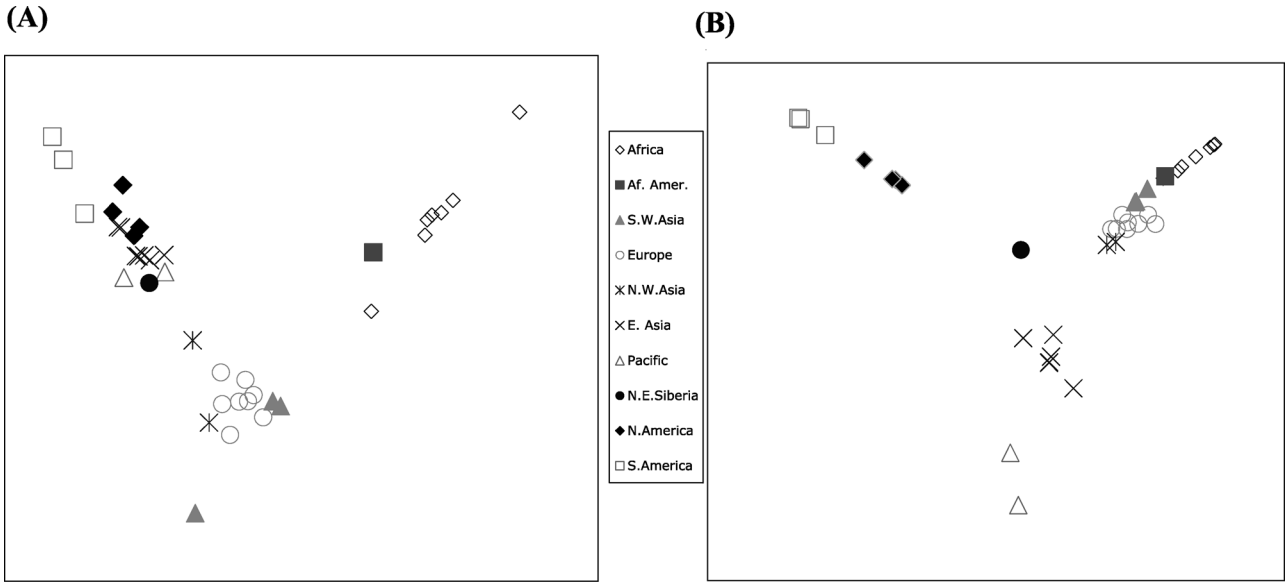

Figure 9. Principal component analyses (PCAs) for 38 populations. **(A)** PCA of pairwise genetic distances based on the 50 locus dataset described in Figure 4. **(B)** PCA of pairwise migrational distances (see text). Note: African Americans are arbitrarily placed in the middle of the Mediterranean and European Americans are arbitrarily placed in the middle of western Europe for the purposes of "migrational" distances.

two-dimensional representation accounts for 91.5% of the variation in these "migrational" distances among the populations.

Clearly, geographically close populations tend to be similar genetically, but the genetic clusters show a different relationship to each other than the geographic clusters. Most noticeable are the differences between the Africans and Europeans. African populations are more different from each other than their geography would predict, and the European populations are much more different from the African populations than the geography would suggest. Although the relationships of the Europeans, east Asians, and Native Americans look different, they are quite similar if one takes into account the third dimension and realizes that the different analyses give different perspectives. The major differences are the closer relative positions of the regional groupings (except for the African populations) in the genetic space than in the geographic space and the greater dispersal within the regions in the genetic space than in the geographic space.

Another representation of population relationships is a tree diagram based on genetic distances. An underlying assumption in this analysis is that the allele frequency differences between populations are due to random genetic drift and a history of little gene flow once populations have separated (clearly an invalid assumption in many cases). For each pair of populations, one can calculate genetic distances that should be related to time in generations, since the populations separated divided by twice the effective population size (Kidd and Cavalli-Sforza 1974). The effective population size considers all of the ancestral populations from the common ancestor of the two populations to the present sizes of the populations. Using these genetic distances, one can then attempt to identify a tree structure in which the branch lengths represent additive components in units of $t/2N_e$. The tree that is best is the one that most closely matches each pairwise distance with the sum of the branch lengths connecting the two populations. One such additive tree that comes close to the minimum amount of evolution required, along with no negative segments, is illustrated in Figure 10. Many of the segments of this tree have bootstrap support that is very high, often greater than 70%, and in one case 100%, of the 1000 bootstrap analyses that were done.

Several of the relationships among populations shown by this tree (Figure 10) are concordant with the groupings of populations shown in Figure 9A. Also, some elements of this tree are supported by historical knowledge. Thus it is not surprising that the three Native American populations from the Amazon cluster and appear more "derived" than do the North American populations. Similarly the tree clusters together the two Pima groups that are known to speak similar languages and are thought to have been separated for only 600 to 1000 years. The large bootstrap values basically define four strong clusters: the African populations (including African Americans); the European populations (including European Americans) and Middle Eastern populations, along with one population from northwestern Siberia (Komi); the east Asian populations; and the Native American populations. Three populations are definitely outside of those four clusters: the Khanty from northwestern Siberia, the Micronesians, and the Nasioi Melanesians. There

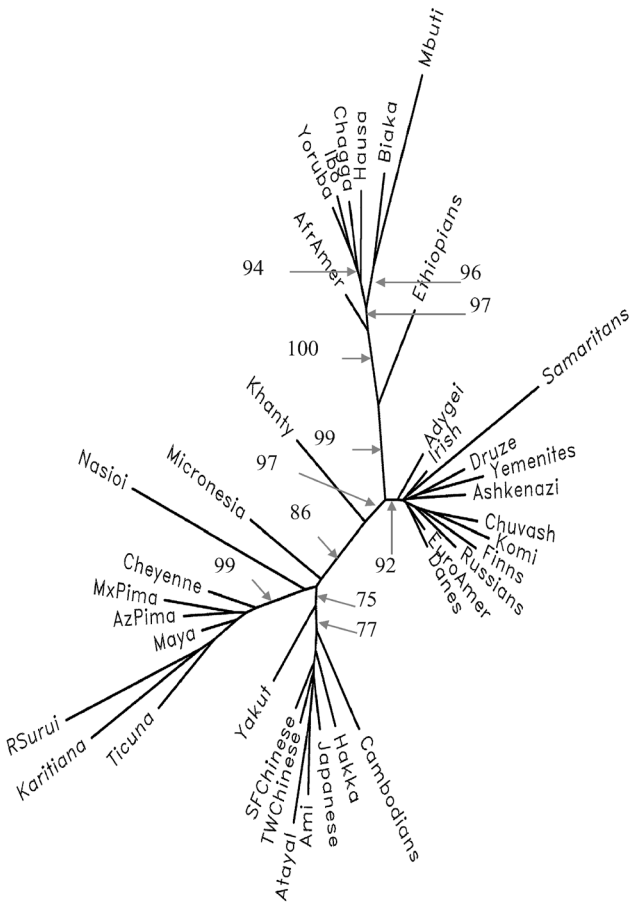

Figure 10. Least squares tree of 38 populations based on the 50 locus pairwise τ genetic distances used in Figure 9A. Note: This is not necessarily the best possible representation of the data because it is impossible to examine all 2.395×10^{55} possible tree structures. While there may be a tree structure that gives a better fit to the data, it is likely to differ only around very short branches. Hence this tree is probably indistinguishable from the absolute best tree within the graphic resolution of those very tiny segments. The bootstrap values for several segments are given as the percentages of the 1000 bootstrap analyses that divided the 38 populations into two groups separated by the indicated segment.

is considerable statistical uncertainty about how the Micronesians and Melanesians connect in the part of the tree where Native Americans and east Asians begin to diverge, but there is very strong support that the Micronesian and Melanesian samples are largely derived from ancestral east Asian peoples and not derived from an independent migration out of Africa. The Khanty, in contrast, has high bootstrap support for a clearly intermediate position between European populations on one hand and east Asian populations on the other.

While the underlying assumptions of this tree analysis are not strictly true for some populations, most notably African Americans, they may be reasonable approximations for some aspects of the tree. Thus the structure indicates that Native Americans and east Asians are derived from an ancestral population that diverged considerably from the European and African populations. But the implication is that two separate ancestral populations diverged and only subsequently did each diverge into multiple populations within each region. Thus the Native Americans are not descended from one of the existing east Asian lineages, but from a much more remote common ancestor. The small branch common to all European and Middle Eastern populations indicates they share some common ancestry distinct from that of all other populations. Among the longest and best supported branches are those associated with the separation of the sub-Saharan African populations from all others. These collectively represent the "out of Africa" bottleneck. What cannot be determined from this representation is the degree to which recent migration among geographically close populations, such as those in Europe and those in east Asia, is responsible for the short branches. However, some of the recent bottlenecks are quite obvious in exceptionally long branches, most noticeably the long branch to the Samaritans.

Finally, the tree makes a strong statement about effective population sizes for populations from different parts of the world. The genetic distances summed along the branches are in units of $t/2N_e$, yet the time from the origin of modern humans to the present is identical for all extant populations at the termini of the branches. Placing that origin at the point where the majority of the African populations join indicates roughly a fourfold difference in N_e between African and east Asian populations. Thus there is no single effective population size for modern humans.

Conclusion

The "out of Africa" model, first supported molecularly by studies of mitochondrial DNA (Cann et al. 1987), is supported by most of these data except those heterozygosity data in Figure 4. The distributions of ancestral alleles and the global F_{ST} data are neutral with respect to this model, but the pairwise genetic distances underlying the principal component and tree analyses (Figures 9 and 10) clearly support the model. The greater haplotype diversity in African populations and the greater LD in non-African populations are related and supportive of the model. The model can be described as follows: Genetic variation had already accumulated in anatomically modern humans in Africa between 150,000 years before present (BP) and 100,000 years BP. That variation was not evenly distributed across the continent, as expected in an isolation-by-distance model, but considerable randomness among closely linked sites had accumulated. That variation and low levels of LD still exist in most modern African populations. About 100,000 years BP, some peoples from northeast Africa migrated into southwest Asia. Since the people who migrated originated from the populations of northeast Africa, they sampled from that already partially diverged gene pool and the sampling error of the migration accentuated the loss of variation. Only a fraction of the genetic variation in Africa as a whole was

represented in that initial "non-African" population. It was that population in southwest Asia that then increased in numbers and spread geographically to occupy all of Eurasia and Australo-Melanesia by about 40,000 years BP. There has not been enough time for much new genetic variation to arise outside of Africa; and, in fact, some of the variation in southwest Asia that spread into Eurasia was lost by the populations through accumulating genetic drift as they spread eastward and eventually reached far East Asia.

At some time more recent than 40,000 years BP, some of the populations from Siberia migrated to the Americas and expanded to occupy first North and then South America. Additional variation was lost during that colonization, but the effect was less than that associated with the migration out of Africa. At all of those stages where variation was lost, nonrandomness (LD) increased among the remaining variants in small segments of DNA. Thus much of the LD seen in non-African populations is the result of the founder effect associated with the expansion out of Africa. An abstract, artistic rendition of this model can be found at http://info.med.yale.edu/genetics/kkidd/point.htm.

Data and analyses such as those presented here represent some of the components of the work/research involved in understanding genetic variation in humans. These data and analyses are beginning to define empirically the distribution of common variation around the world and to establish an evolutionary framework based on the random consequences of the historical demographics of modern human populations. That framework begins to set a baseline against which individual loci can be compared to identify loci acted upon by region-specific directional selection or by balancing selection. The challenge of understanding variation will not be met solely by population genetics studies such as those presented here, but they are an integral part of the quest.

Electronic Databases Cited

ALFRED: http://alfred.med.yale.edu/alfred/index.asp
dbSNP: http://www.ncbi.nlm.nih.gov/SNP/
HGVbase: http://hgvbase.cgb.ki.se/
JSNP: http://snp.ims.u-tokyo.ac.jp/
Kidd Lab Web Site: http://info.med.yale.edu/genetics/kkidd/

Acknowledgments

This work was supported in part by NIH grants AA09379, GM57672, and MH62495 (to K.K.K.), and by NSF grant SBR-9632509 (to J.R.K.). Support for sample collection was also provided by a grant from the Alfred P. Sloan Foundation (to K.K.K. and J.R.K.). Support was also provided by a contract from the National Institute of Diabetes and Digestive and Kidney Diseases (to K.K.K.). We want to acknowledge and thank the following individuals for their help over the years in assembling the samples from the diverse populations: Francis L. Black, Batsheva Bonne-Tamir, William F. Byerly, L. L. Cavalli-Sforza, Jonathon Friedlaender, David Goldman, Elena Grigorenko, Sylvester L. B. Kajuna, Nganyirwa J. Karoma, Kenneth Kendler, William Knowler, Selemani Kungulilo, Ru-Band Lu, Adekunle Odunsi, Friday Okonofua, Frank Oronsaye, Josef Parnas, Leena Peltonen, Leslie O. Schulz, Kenneth Weiss, and Olga V. Zhukova. We also thank the several technicians, graduate students, and postdoctorals who have helped collect the data on these populations over the past 18 years. Special thanks are due to the many hundreds of individuals who volunteered to give blood samples for studies such as this. Without such participation of individuals from diverse parts of the world we would be unable to obtain a true picture of the genetic variation in our species. This paper was originally presented at the American Genetics Association 2003 Annual Meeting and Centennial Celebration at the University of Connecticut, Storrs, July 18–30, 2003.

References

Armour JAL, Anttinen T, May CA, Vega EE, Sajantila A, Kidd JR, Kidd KK, Bertranpetit J, Paabo S, and Jeffreys AJ, 1996. Minisatellite diversity supports a recent African origin of modern humans. Nat Genet 13: 154–160.

Bamshad M and Wooding SP, 2003. Signatures of natural selection in the human genome. Nat Rev Genet 4:99–111.

Bonnen PE, Wang PJ, Kimmel M, Chakraborty R, and Nelson DL, 2002. Haplotype and linkage disequilibrium architecture for human cancer-associated genes. Genome Res 12:1846–1853.

Botstein D, White RL, Skolnick M, and Davis RW, 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331.

Budowle B, Masibay A, Anderson SJ, Barna C, Biega L, Brenneke S, Brown BL, Cramer J, DeGroot GA, Douglas D, Duceman B, Eastman A, Giles R, Hamill J, Haase DJ, Janssen DW, Kupferschmid TD, Lawton T, Lemire C, Llewellyn B, Moretti T, Neves J, Palaski C, Schueler S, Sgueglia J, Sprecher C, Tomsey C, and Yet D, 2001. STR primer concordance study. Forensic Sci Int 124:47–54.

Budowle B, Moretti TR, Niezgoda SJ, and Brown BL, 1998. CODIS and PCR-based short tandem repeat loci: law enforcement tools. In: Proceedings of the Second European Symposium on Human Identification. Madison, WI: Promega, Corp.; 73–88.

Cann R, Stoneking M, and Wilson AC, 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.

Castiglione CM, Deinard AS, Speed WC, Sirugo G, Rosenbaum HC, Zhang Y, Grandy DK, Grigorenko EL, Bonne-Tamir B, Pakstis AJ, Kidd JR, and Kidd KK, 1995. Evolution of haplotypes at the DRD2 locus. Am J Hum Genet 57:1445–1456.

Cavalli-Sforza LL, Menozzi P, and Piazza A, 1994. The history and geography of human genes. Princeton, NJ: Princeton University Press.

Chattopadhyay P, Pakstis AJ, Mukherjee N, Iyengar S, Odunsi A, Okonofua F, Bonne-Tamir B, Speed WC, Kidd JR, and Kidd KK, 2003. A global survey of haplotype frequencies and linkage disequilibrium at the RET locus. Eur J Hum Genet 11:760–769.

Collins FS, Green ED, Guttmacher AE, and Guyer MS, 2003. A vision for the future of genomics research. Nature 422:835–847.

Crow JF, 1993. Felix Bernstein and the first human marker locus. Genetics 133:4–7.

Crow JF, 2001. The beanbag lives on. Nature 409:771.

DeMille MMC, Kidd JR, Ruggeri V, Palmatier MA, Goldman D, Odunsi A, Okonofua F, Grigorenko E, Schulz LO, Bonne-Tamir B, Lu R-B, Parnas J, Pakstis AJ, and Kidd KK, 2002. Population variation in linkage disequilibrium across the *COMT* gene considering promoter region and coding region variation. Hum Genet 111:521–537.

Denoeud F, Vergnaud G, and Benson G, 2003. Predicting human minisatellite polymorphism. Genome Res 13:856–867.

Devlin B and Risch N, 1995. A comparison of linkage disequilibrium measures for fine scale mapping. Genomics 29:311–322.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Missasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, and Weissenbach J, 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152–154; supplement A1–A138.

Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, and Brookes AJ, 2002. HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. Nucleic Acids Res 30: 387–391.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel Bhiggins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, and Altshuler D, 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Garrod E, 1902. The incidence of alkaptonuria: a study in chemical individuality. Lancet 2:1616–1620.

Hacia JG, Fan JB, Ryder O, Jin L, Edgemon K, Ghandour G, Mayer RA, Sun B, Hsie L, Robbins CM, Brody LC, Wang D, Lander ES, Lipshutz R, Fodor SP, and Collins FS, 1999. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nat Genet 22:164–167.

Harris H, 1966. Enzyme polymorphisms in man. Proc R Soc Lond B Biol Sci 22:298–310.

Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, and Nakamura Y, 2002. JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 30:158–162.

Hubby JL and Lewontin RC, 1966. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. Genetics 54:577–594.

Iyengar S, Seaman M, Deinard AS, Rosenbaum HC, Sirugo G, Castiglione CM, Kidd JR, and Kidd KK, 1998. Analyses of cross-species polymerase chain reaction products to infer the ancestral state of human polymorphisms. DNA Sequence 8:317–327.

Jeffreys AJ, Wilson V, and Thein SL, 1985. Hypervariable "minisatellite" regions in human DNA. Nature 314:67–83.

Kan YW and Dozy AM, 1978. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. Proc Natl Acad Sci USA 75:5631–5635.

Kidd JR and Kidd KK, 2003. The population genetics of PAH. External update attached to chapter 77 (the Hyperphenylalaninemias). In: Metabolic and molecular bases of inherited disease, online version (Scriver C, ed). New York: McGraw-Hill.

Kidd JR, Pakstis AJ, and Kidd KK, 1993. Global levels of DNA variation. Proceedings of the 4th International Symposium on Human Identification. Madison, WI: Promega, Corp.; 21–30.

Kidd JR, Pakstis AJ, Zhao H, Lu R-B, Okonofua FE, Odunsi A, Grigorenko E, Bonne-Tamir B, Friedlaender J, Schulz LO, Parnas J, and Kidd KK, 2000. Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus (*PAH*) in a global representation of populations. Am J Hum Genet 66:1882–1899.

Kidd JR, Speed WC, Pakstis AJ, and Kidd KK, 2003. A 100 kb block encompassing BRCA1. Am J Hum Genet 73(suppl):173.

Kidd KK and Cavalli-Sforza LL, 1974. The role of genetic drift in the differentiation of Icelandic and Norwegian cattle. Evolution 28:381–395.

Kidd KK, Morar B, Castiglione CM, Zhao H, Pakstis AJ, Speed WC, Bonne-Tamir B, Lu R-B, Goldman D, Lee C, Nam YS, Grandy DK, Jenkins T, and Kidd JR, 1998. A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. Hum Genet 103:211–227.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher Jr, and Stefansson K, 2002. A high-resolution recombination map of the human genome. Nat Genet 31:241–247.

Landsteiner K, 1900. Zur Kenntnis der antifermentativen, lytischen and agglutinierenden Wirkungen des Blutserums und der Lymphe. Zentralblatt für Bakteriologie, Parasitekunde und Infektionskrankheiten, Abteilung 1, 27:357–362.

Litt M and Luty JA, 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397–401.

Lobos EA and Todd RD, 1998. Association analysis in an evolutionary context: cladistic analysis of the DRD2 locus to test for association with alcoholism. Am J Med Genet (Neuropsychiatr Genet) 81:411–419.

Maynard Smith J and Haigh J, 1974. The hitch-hiking effect of a favorable gene. Genet Res 23:23–35.

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, and White R, 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622.

Oota H, Pakstis AJ, Bonne-Tamir B, Goldman D, Grigorenko E, Kajuna SLB, Karoma NJ, Kungulilo S, Lu R-B, Odunsi A, Okonofua F, Zhukova OV, Kidd JR, and Kidd KK, 2004. The evolution and population genetics of the ALDH2 locus: random genetic drift, selection, and low levels of recombination. Ann Hum Genet 68:93–109.

Osier MV, Cheung K-H, Kidd JR, Pakstis AJ, Miller PL, and Kidd KK, 2002. ALFRED: an allele frequency database for anthropology. Am J Phys Anthropol 119:77–83.

Pakstis AJ, Kidd JR, and Kidd KK, 2002. A reference distribution of F_{st} values for biallelic DNA markers. Am J Hum Genet 71(suppl):371.

Rajeevan H, Osier MV, Cheung H-K, Deng H, Druskin L, Heinzen R, Kidd JR, Stein S, Pakstis AJ, Tosches NP, Yeh CC, Miller PL, and Kidd KK, 2003. ALFRED: the ALlele FREquency Database Update. Nucleic Acids Res 31:270–271.

Ruano G, Rogers J, and Kidd KK, 1989. Comparative mapping utilizing cross-species PCR (CS-PCR): detection of a beta-globin region rearrangement in the gibbon. Cytogenet Cell Genet 51:1071.

Sabeti P, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJK, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, and Lander ES, 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, and Sirotin K, 2001. DbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311.

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Cheung K, Kidd JR, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Watson E, Krings M, Pääbo S, Risch N, Jenkins T, and Kidd KK, 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387.

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, and Kidd KK, 1998. A global haplotype analysis of the DM locus: implications for the evolution of modern humans and the origin of myotonic dystrophy mutations. Am J Hum Genet 62:1389–1402.

Watterson GA and Guess HA, 1977. Is the most frequent allele the oldest? Theor Popul Biol 11:141–160.

Weber JL and May PE, 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396.

Williamson R, Bowcock A, Kidd KK, Pearson P, Schmidtke J, Ceverha P, Chipperfield M, Cooper DN, Coutelle C, Hewitt J, Klinger K, Langley K, Beckmann J, Tolley M, and Maidak B, 1991. Report of the DNA committee and catalogues of cloned and mapped genes, markers formatted for PCR and DNA polymorphisms. Cytogenet Cell Genet 58:1190–1832.

Wright S, 1969. Evolution and the genetics of populations. Volume 2: The theory of gene frequencies. Chicago: University of Chicago Press; 511.

Wyman A and White RW, 1980. A highly polymorphic locus in human DNA. Proc Natl Acad Sci USA 77:6754–6758.

Zhao H, Pakstis AJ, Kidd JR, and Kidd KK, 1999. Assessing linkage disequilibrium in a complex genetic system. I. Overall deviation from random association. Ann Hum Genet 63:167–179.

Corresponding Editor: Kent E. Holsinger