Am. J. Hum. Genet. 66:1882–1899, 2000

Haplotypes and Linkage Disequilibrium at the Phenylalanine Hydroxylase Locus, *PAH,* in a Global Representation of Populations

Judith R. Kidd,[1] Andrew J. Pakstis,[1] Hongyu Zhao,[2] Ru-Band Lu,[4] Friday E. Okonofua,[5] Adekunle Odunsi,[6] Elena Grigorenko,[3], Batsheva Bonne-Tamir,[7] Jonathan Friedlaender,[8] Leslie O. Schulz,[9] Josef Parnas,[10] and Kenneth K. Kidd[1]

Departments of [1]Genetics, [2]Epidemiology and Public Health, and [3]Psychology and Child Study Center, Yale University, New Haven, CT; [4]Department of Psychiatry, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan; [5]University of Benin, Faculty of Medicine, Benin City, Nigeria; [6]Department of Gynecological Oncology, Roswell Park Cancer Institute, Buffalo; [7]Department of Genetics, Sackler School of Medicine, Tel Aviv University, Tel Aviv; [8]Department of Anthropology, Temple University, Philadelphia; [9]Department of Health Sciences, University of Wisconsin, Milwaukee; and [10]Institute of Preventative Medicine, Kommune Hospitalet, Copenhagen

Because defects in the phenylalanine hydroxylase gene (*PAH*) cause phenylketonuria (PKU), *PAH* was studied for normal polymorphisms and linkage disequilibrium soon after the gene was cloned. Studies in the 1980s concentrated on European populations in which PKU was common and showed that haplotype-frequency variation exists between some regions of the world. In European populations, linkage disequilibrium generally was found not to exist between RFLPs at opposite ends of the gene but was found to exist among the RFLPs clustered at each end. We have now undertaken the first global survey of normal variation and disequilibrium across the *PAH* gene. Four well-mapped single-nucleotide polymorphisms (SNPs) spanning ~75 kb, two near each end of the gene, were selected to allow linkage disequilibrium across most of the gene to be examined. These SNPs were studied as PCR-RFLP markers in samples of, on average, 50 individuals for each of 29 populations, including, for the first time, multiple populations from Africa and from the Americas. All four sites are polymorphic in all 29 populations. Although all but 5 of the 16 possible haplotypes reach frequencies >5% somewhere in the world, no haplotype was seen in all populations. Overall linkage disequilibrium is highly significant in all populations, but disequilibrium between the opposite ends is significant only in Native American populations and in one African population. This study demonstrates that the physical extent of linkage disequilibrium can differ substantially among populations from different regions of the world, because of both ancient genetic drift in the ancestor common to a large regional group of modern populations and recent genetic drift affecting individual populations.

Introduction

Linkage disequilibrium or, more generally, gametic-phase allelic association, is the nonrandom occurrence of alleles on chromosomes (i.e., in gametes) in a population. Linkage disequilibrium has become an important tool in the end stages of positional cloning because a recently arisen single deleterious allele will usually be nonrandomly associated with the alleles at nearby polymorphic sites that were on the chromosome on which the mutation originally occurred. Among the earliest applications of this principle for identification of distinct mutations associated with a disease were studies at the β-hemoglobin cluster to identify thalassemia mutants (Kazazian et al. 1984), as well as studies at the phenylalanine hydroxylase locus (*PAH*) to identify phenylketonuria (PKU [MIM 261600]) mutants (DiLella et al. 1986*a,* 1987).

PKU is one of the most common genetic diseases in people of northern-European descent, occurring in that group at an average rate of ~1/10,000 live births (Bickel et al. 1981). The *PAH* gene, coding for the enzyme phenylalanine hydroxylase (PAH), was implicated as the etiologic gene, by the absence of PAH enzyme activity in patients with PKU (Friedman et al. 1973). The human *PAH* cDNA was cloned (Woo et al. 1983; Kwok et al. 1985), RFLPs were identified by use of the cDNA as the probe (Woo et al. 1983; Lidsky et al. 1985*a*), the gene was mapped to human chromosome 12q22-24 (Lidsky et al. 1985*b*), and the molecular structure of the gene was described (DiLella et al. 1986*a*), relatively early in the history of recombinant-DNA studies. When the RFLPs encompassing the gene were used, it was obvious that PKU mutations occurred on several different haplotypes. By 1986 it was recognized that PKU was mutationally heterogeneous; at least some different

Received December 23, 1999; accepted for publication March 14, 2000; electronically published April 27, 2000.

Address for correspondence and reprints: Dr. Judith R. Kidd, Department of Genetics, SHM I353, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06520-8005. E-mail: kidd@biomed.med.yale.edu

0002-9297/2000/6606-0017$02.00

PKU haplotypes were found to have different mutations (DiLella et al. 1986*b,* 1987). Interestingly, in a sample from Denmark, no significant association could be demonstrated between the disease alleles and *any* single allele of the normal polymorphisms spanning the gene (Chakraborty et al. 1987). However, when haplotypes of these polymorphisms were examined, two haplotypes were significantly more common among PKU chromosomes than among normal chromosomes. The reason for the absence of allelic association with individual RFLPs was also obvious: the marker-allele frequencies among PKU chromosomes were very similar to the frequencies in normal chromosomes. In the case of normal chromosomes, the two most common haplotypes were approximately equally frequent and had alternative alleles at most sites. The PKU chromosome(s) included those two haplotypes, but two others were the most common, accounting for 20% and 38% of the PKU chromosomes in the Danish sample; these two "PKU" haplotypes also had alternative alleles at all the sites at which the two common normal haplotypes differed and had the same allele at all the sites at which the other two were the same. This allelic complementarity of the most common haplotypes in both normal and PKU chromosomes greatly reduced the power to detect allelic association of PKU with any single RFLP, although the association was obvious when haplotypes were used. The two associated haplotypes were the basis for the first identifications of specific PKU mutations (DiLella et al. 1986*b,* 1987).

As the etiologically relevant mutations for PKU and the phenylalaninemia states became known, *PAH* haplotypes of patients with PKU often signaled which mutation(s) a patient carried and/or alerted the researchers to the existence of previously unknown mutations (Daiger et al. 1989*a,* 1989*b:* Hertzberg et al. 1989; Stuhrmann et al. 1989; Apold et al. 1990; Dianzani et al. 1990; Jaruzelska et al. 1991; Konecki and Lichter-Konecki 1991; Svensson et al. 1991; Zygulska et al. 1991; Baric et al. 1992; Kozak et al. 1995). Thus, by 1989, polymorphisms, mutations, and haplotypes of the *PAH* region had finally become well characterized in patients with PKU (Woo 1988; Nowacki et al. 1997). By 1996, the PKU mutations and their haplotypes were being used to infer the natural histories both of the mutations themselves and of the populations carrying those mutations (Scriver et al. 1996). Only occasionally were non-PKU haplotypes studied in populations of non-European origin—and, even then, usually in families with PKU (Daiger et al. 1989*b;* Hertzberg et al. 1989; Hofman et al. 1991). Various summaries of *PAH* mutations have been published (e.g., see Konecki and Lichter-Konecki 1991; Eisensmith et al. 1992); an up-to-date compendium of *PAH* mutations and background haplotypes is maintained at the PAHdb Web site (Nowacki et al. 1997).

Studies of normal chromosomes in European populations have shown that disequilibrium exists among the sites at either end of the gene, at distances of 22 and 31 kb, but, in general, either does not exist or is much weaker between markers located at opposite ends of the gene, at distances $\geq$43 kb (Chakraborty et al. 1987; Daiger et al. 1989*a*). A similar pattern has been observed in a sample of 44 chromosomes from China and Japan (Daiger et al. 1989*b*) and in a sample of >600 chromosomes from several Polynesian groups (Hertzberg et al. 1989). Other analyses of various published *PAH* data have reached similar conclusions (Feingold et al. 1993; Degioanni and Darlu 1994). This pattern of the molecular extent of linkage disequilibrium agrees with the findings of Jorde et al. (1994)—that linkage disequilibrium in populations of European origin generally does not extend to >50–60 kb. Our recent studies of linkage disequilibrium in multiple populations have found that linkage disequilibrium can differ dramatically among populations from different regions of the world (Tishkoff et al. 1996*a,* 1996*b,* 1998; Kidd et al. 1998). The only substantive reports of *PAH* haplotype frequencies in specific non-European populations—Polynesians (Hertzberg et al. 1989) and eastern Asians (Daiger et al. 1989*b*)—show reduced levels of heterozygosity relative to that in Europeans. The one small study of African Americans (Hofman et al. 1991) has found that haplotype frequencies for normal and PKU chromosomes differ from each other and from frequencies in Europeans and Asians.

All of the previous studies of linkage disequilibrium at the *PAH* locus have used pairwise coefficients. The resulting matrix of disequilibrium coefficients can show a clear pattern but also often contains some pairwise values that do not fit into the general pattern of high absolute values for pairs of markers within either cluster and low absolute values for pairs of markers that bridge the two clusters. As we have begun to consider linkage disequilibrium in more-complex genetic systems, we have introduced a new coefficient to measure overall nonrandomness across the entire haplotype (Kidd et al. 1998; Zhao et al. 1999). The coefficient estimate is based on the permuted data of the observed samples. A variant of the permutation test for overall significance allows us to test significance of the disequilibrium across any segment within the haplotype (Zhao et al. 1997, 1999).

We have now examined normal, non–disease-causing polymorphisms in the region encompassing the 13 exons of *PAH,* considering a global sample of populations for the first time, to enhance our understanding of how the amount and pattern of linkage disequilibrium can differ in populations from different parts of the world.

We have chosen to study haplotypes at *PAH* specifically, for several reasons: (1) the region is well mapped, with the location and nature of several noncoding single-nucleotide polymorphisms (SNPs) clearly described; (2) some population data are already available in the literature that demonstrate allele- and haplotype-frequency variation between European and some non-European populations; (3) the well-documented reduced levels of disequilibrium between markers at either end of the gene provide an opportunity to explore the utility of the permutation-test variant that examines segment disequilibrium; and (4) both the historical importance of and interest in haplotypes at this gene already have been established.

Subjects and Methods

Population Samples

We have studied 29 populations: four from Africa (Biaka, Mbuti, Yoruba, and Ethiopian Jews), nine from Europe and southwestern Asia (Adygei, Danes, Finns, Irish, Russians, Europeans of heterogeneous ancestry [from the United States], Samaritans, Yemenite Jews, and Druze), seven from eastern Asia (two independent Han Chinese samples [one from Taiwan and one mainly from southern China and sampled in San Francisco], Hakka, Japanese, Cambodians, Ami, and Atayal), one from Siberia (Yakut), one from Australo-Melanesia (Nasioi), four from North America (Cheyenne, Arizona Pima, Mexican Pima, and Maya), and three from South America (Ticuna, Rondonian Surui, and Karitiana). Descriptions of these specific population samples, most of which have been/will be in studies of other loci, can be found in the work of Castiglione et al. (1995), Kidd et al. (1991, 1998), Tishkoff et al. (1998), and Osier et al. (1999). Additional information on these samples is available through the Internet (Kidd Lab Home Page). Sample sizes range from 23 to >100 and average ~50 individuals per population. The samples are of unrelated (at least in the first degree) members of the populations, with the exception of the three South American Indian groups. These South American samples were collected from small, endogamous populations in which everyone in the village(s) is related to everyone else. This is most evident in the Karitiana, a group who are the sole speakers of Karitiana, a Tupi language. Every Karitiana is related to every other in the village, where they live as a single extended kindred—everyone recently born in this population is descended from a single man and one or more of his four wives (often through several pathways) who lived five generations ago.

All samples were collected with both approval from the appropriate institutional review boards and informed consent from the participants. The DNA in this study was purified, by means of standard phenol-chloroform extraction and ethanol precipitation (Sambrook et al. 1989), from Epstein-Barr virus–transformed lymphoblastoid cell lines (Anderson and Gusella 1984). The Coriell Institute for Medical Research (National Institute of General Medical Sciences Human Genetic Mutant Cell Line Repository) in Camden, NJ, has available for distribution at least 5–10 cell lines and DNA of many individuals from several of these population samples. These samples were collected for purposes unrelated to PKU, and no information is available on whether any relative has PKU. We assume that all chromosomes contain a normal *PAH* allele.

Polymorphic Sites and Typing Protocols

To maximize the information on disequilibrium between the ends of the gene, we chose the two pairs of markers that were closest to each end of the gene and already converted to PCR-based typing. The four SNPs that were selected span from just downstream of exon 1 to the middle of intron 8 of the *PAH* locus, a distance of ~75 kb. Each SNP affects a specific restriction site (*Bgl*II, *Pvu*II, *Msp*I, and *Xmn*I); these were originally identified as RFLPs, by use of a *PAH* cDNA clone as the probe (Woo et al. 1983; DiLella et al. 1986*a*). These four polymorphisms are biallelic restriction-site polymorphisms (RSPs) whose primers, PCR conditions, and fragment sizes have been given by Dworniczak et al. (1991*a,* 1991*b*), Wedemeyer et al. (1991), and Goltsov et al. (1992). The *Bgl*II polymorphism is located in intron 1, just 55 bp downstream of exon 1 (GenBank AF003965; Iyengar et al. 1998); we confirmed this reported location of the *Bgl*II site (R. C. Eisensmith, personal communication) by aligning our sequence of the *Bgl*II amplicon with the exon 1 and flanking sequence found at the PAHdb Web site. The *Pvu*II polymorphism (commonly referred to as "*Pvu*IIa") is in the 5′ end of intron 2 ~1.4 kb downstream of exon 2. The *Msp*I polymorphism is ~65 kb farther downstream; by aligning our sequence of the *Msp*I amplicon (GenBank AF003967; Iyengar et al. 1998) with the exon 8 and flanking sequence found at the PAHdb Web site, we confirmed the reported position of the *Msp*I site (R. C. Eisensmith, personal communication) to 268 bp upstream of exon 8, in intron 7. The *Xmn*I site is reported to be 1.5 kb 3′ of exon 8 (R. C. Eisensmith, personal communication), which we confirmed by obtaining a 2-kb PCR product, using the 5′ primer of the *Msp*I pair with the 3′ primer of the *Xmn*I pair on genomic DNA. This places these two sites ~1.8 kb apart. Little intronic sequence is available, and no hard data have been published to give the distance between the *Pvu*IIa polymorphism and exon 2. The distance estimates that we have used derive from Goltsov et al. (1993), DiLella et al.

(1986*a*), and a personal communication from R. C. Eisensmith. Although they may be revised when the region is completely sequenced, they certainly reflect the relative sizes accurately enough for the purposes of this study (fig. 1). For each marker, the PCR product was digested with the appropriate enzyme, according to the manufacturer's protocol, and the fragments were electrophoresed on agarose gels and were stained with ethidium bromide.

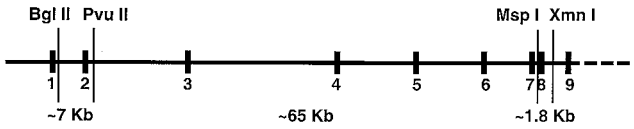

Figure 1 Map of four RSPs at the *PAH* locus. The numbered boxes correspond to the first 9 of the 13 exons. Relative spacing is the best estimate from multiple sources (see the text).

Data Management and Simple Statistics

All typing results were entered, as individual phenotypes, into PhenoDB2, our client-server database system for genetic marker data (Cheung et al. 1996). The underlying software has recently been converted from 4th-Dimension-Sybase to Access-Oracle but otherwise remains conceptually as originally described. Using the phenotype-genotype correspondences entered for each system (codominant for the SNPs in this study), PhenoDB2 calculates allele frequencies and tests for Hardy-Weinberg (H-W) ratios. Output files of specified multisite phenotypes for each individual in each population are generated for input into other programs. FENGEN (A. J. Pakstis, unpublished data; source code available from the Kidd Lab Home Page) also calculates allele frequencies and tests for H-W ratios, provides organized summary tables, and prepares input files for haplotype analyses.

Haplotype-Frequency Estimation

Since most population samples consisted of unrelated individuals, family data could not be used to set phase in multiply heterozygous individuals. Instead, maximum-likelihood estimates of haplotype frequencies and the standard errors (jackknife method) were calculated from the multisite marker-typing data, by use of either the program HAPLO (Hawley and Kidd 1995), which implements the EM algorithm (Dempster et al. 1977), or the derivative, HAPLO/P (Zhao et al. 1997, 1999). HAPLO accommodates individuals with either missing data at some sites or partial phase information, by giving them unique phenotypes corresponding to the set of underlying genotypes compatible with the information available, as explained by Hawley and Kidd (1995). In some cases, first-degree relatives could be used to fully or partially determine the haplotypes on the basis of transmission patterns, and that information was incorporated into the frequency estimates. In most cases, known relationships were distant, and such individuals were included as though they were unrelated. This does not bias the estimates but does increase the sampling error somewhat. Expected heterozygosities for individual sites and for the haplotypes have been estimated as $1 - \Sigma p_i^2$, where p_i represents the allele or haplotype frequencies for the system.

Disequilibrium

The standardized, pairwise linkage-disequilibrium value D' (Lewontin 1964) was calculated for each pair of markers, and the null hypothesis of linkage equilibrium ($D' = 0$) was tested with an asymptotically χ^2 statistic (see eq. 3.10 in Weir 1996), by means of the computer program LINKD (A. J. Pakstis, unpublished data; source code available from the Kidd Lab Home Page) and with the sample sizes and haplotype frequency estimates from HAPLO used as input. Overall disequilibrium, the deviation of observed (i.e., estimated from the data) haplotype frequencies from those expected under random association of alleles at all sites considered simultaneously, can be estimated in two ways. HAPLO calculates a likelihood-ratio statistic that can, under some circumstances, be interpreted as an asymptotic χ^2 statistic measuring overall nonrandomness. Alternatively, a permutation test can be used to measure significance of overall nonrandomness across multiple sites with multiple alleles, without the assumption of a χ^2 distribution or any specific distribution (Zhao et al. 1997, 1999; Kidd et al. 1998). We used the program HAPLO/P to generate 1,000 permuted samples for each of the 29 population samples and measured significance as the fraction of permuted samples with likelihood-ratio statistics greater than the observed value.

Quantification of Overall Disequilibrium

Since the significance of the disequilibrium does not quantify the amount of disequilibrium, we have developed a standardized coefficient, ξ, to allow comparisons among populations (Zhao et al. 1999): $\hat{\xi} = (\sqrt{2\nu}/N)[(t - \mu)/\sigma]$. This coefficient standardizes the observed likelihood-ratio statistic, using the permutation distribution, the sample size, and the complexity (or degrees of freedom [df]) of the haplotype system, where t is the observed likelihood-ratio statistic, μ and σ are, respectively, the mean and SD of the permutation distribution, N is the number of individuals in the sample, and ν is the df of the system in that population.

Segment Disequilibrium

In a multisite haplotype system, we can shift our traditional focus from disequilibrium between two sites to disequilibrium across a segment of the DNA. Specifically, in the case of *PAH,* we can consider the disequilibrium that exists across the long middle segment (fig. 1). There are four pairwise disequilibrium coefficients that provide information relevant to this segment; but how to combine them is not obvious.

The null hypothesis to be tested for the segment linkage equilibrium is that there is no linkage disequilibrium across the segment but that there can be linkage disequilibrium for markers within the two groups on either side of the segment. For the marker systems considered here, the *Bgl*II and *Pvu*II sites form one group, and the *Msp*I and *Xmn*I sites form the other group. Because the asymptotic distribution may not always provide a good approximation for assessment of statistical significance in complex data sets such as this, we use a permutation test to estimate statistical significance. For each permutation, the permuted sample is constructed by independently permuting the genotypes (phenotypes) at *Bgl*II and *Pvu*II as one group and the genotypes (phenotypes) at *Msp*I and *Xmn*I as the other group. The likelihood ratio–test statistic is calculated for each permuted sample for the null hypothesis of no linkage disequilibrium between (*Bgl*II,*Pvu*II) and (*Msp*I,*Xmn*I). For both the permuted samples and the denominator of the likelihood ratio, the phase ambiguity is preserved for double heterozygotes for either of the site pairs; thus, strictly considered, it is the paired phenotypes that are permuted. After generation of a large number of permuted samples and calculation of the likelihood ratio–test statistics, the significance level of the observed sample is estimated as the proportion of the permuted samples with likelihood-ratio statistics larger than that for the observed sample. We have used the segment disequilibrium test to determine the significance of linkage disequilibrium across the central segment of *PAH* in each of the 29 population samples.

Frequency Variation among Populations

Variation in allele and haplotype frequencies was measured as F_{ST}, estimated as $\sigma_p^2/(\overline{p}\,\overline{q})$ for each biallelic site and as the weighted average of the standardized variance for each haplotype for the combined four-site system. To determine whether the haplotype-frequency profiles for any two population samples were different from one another (i.e., whether we were sampling from the same or different groups), the genetic heterogeneity test of Workman and Niswander (1970) was applied. This genetic heterogeneity test resembles a likelihood-ratio χ^2 test and can handle the situation often found for our multisite haplotypes when we have many alleles and a number of them will have very small expected values even in large samples. The simple χ^2 test cannot be applied appropriately in such situations. The Workman and Niswander (1970) test finds the sum of the weighted and squared frequencies of each allele, in turn, across the groups being compared, and subtracts the square of the weighted average of the *i*th allele. The weighting is a function of the sample sizes, and the accumulated sum is multiplied by twice the number of subjects in the sample, so that the resulting statistic follows the χ^2 distribution. In the population comparisons performed, the df equal one less than the number of nonzero alleles.

Results

Marker typings for the four SNPs have been collected on a total of 1,485 individuals in the 29 distinct populations. Typing was >98% complete across all markers and populations, with the missing data scattered in an apparently random pattern. All individuals had multisite phenotypes, with typing data at three or four of the sites. In all, we observed 94 distinct four-site phenotypes (counting the 32 that involved missing data) across all 29 populations (data not shown).

Individual Site Results

Allele frequencies and sample sizes for all four SNPs in all 29 populations are given in ALFRED (Kidd Lab Home Page), an Internet-accessible allele frequency database (Cheung et al. 2000*a,* 2000*b*). Allele frequencies at each polymorphic site were estimated by simple gene counting, and binomial standard errors can be calculated from the information given in the database. All four sites are polymorphic in all 29 populations. In table 1, heterozygosities are given, for each RSP, as the mean and the range seen in each geographic region. Only 3 of the 120 H-W tests (4 sites in each of 30 populations) were significant at $P < .01$, one each at $P < .01$ (Druze at *Bgl*II), $P < .005$ (Ethiopian Jews at *Pvu*II), and $P < .001$ (Finns at *Xmn*I). Different populations and different sites were involved in all three; for each of those three populations, the other three sites did not show significant deviation from H-W ratios. Consequently, we do not consider any of these as being indicative of meaningful deviation from H-W ratios and random mating.

For each site, there is highly significant allele-frequency variation among the populations, but no attempt has been made to test for significance of pairwise differences in frequencies. The ranges of allele frequencies, when we focus on the site-present allele in each case, are .12–.84 for *Bgl*II, .15–.89 for *Pvu*II, .07–.81 for *Msp*I, and .06–.94 for *Xmn*I. The F_{ST} values across all 29 populations are .167 (*Bgl*II), .145 (*Pvu*II), .238 (*Msp*I), .314 (*Xmn*I), and .169 (haplotype) (table 2). The

Table 1

Expected Heterozygosity, by Site and Geographic Region

Region (No. of Populations)	Average Expected Heterozygosity (Range)[a]				
	RSP				
	*Bgl*II	*Pvu*II	*Msp*I	*Xmn*I	Haplotype[b]
Africa (4)	.46 (.41–.49)	.45 (.44–.46)	.39 (.34–.45)	.36 (.29–.41)	.81
Europe and southwestern Asia (9)	.41 (.28–.49)	.44 (.35–.50)	.49 (.46–.50)	.42 (.10–.50)	.77
Eastern Asia (7)	.32 (.21–.49)	.37 (.21–.50)	.21 (.14–.38)	.18 (.11–.34)	.55
Melanesia (1)	.23	.19	.50	.50	.67
Siberia (1)	.34	.38	.45	.50	.77
North America (4)	.41 (.27–.50)	.40 (.26–.48)	.37 (.31–.42)	.33 (.29–.38)	.58
South America (3)	.41 (.36–.49)	.46 (.39–.49)	.43 (.37–.47)	.41 (.33–.45)	.74

[a] Calculated as the unweighted average of values for each site in the populations within a region. For each population the heterozygosity values are calculated as $(1.0 - \Sigma p_i^2)$, where p_i represents the two allele frequencies obtained by simple gene counting for each sample and site.

[b] For the heterozygosities in specific population samples, see table 2.

F_{ST} values, by geographic region, are also given in table 2. In all but one case, the regional values are smaller than the global value.

Haplotype Frequencies and Patterns of Variation

The maximum-likelihood estimates of the frequencies of the 16 possible haplotypes for each population are given in table 3. From these frequencies and the total number of chromosomes (2N) (in table 3, for each population), the binomial standard errors can be estimated. For 24% of the non-zero frequency estimates, the jackknife standard-error estimates calculated by HAPLO were almost the same as the binomial standard-error estimate, mostly for the larger frequency estimates. For 72% of the frequency estimates, the jackknife estimates were larger than the binomial estimates by up to twice the amount. For only ~4% of the frequency estimates were the jackknife estimates more than twice the binomial estimates, usually for frequency estimates <2%. Those jackknife estimates of the standard errors are given in ALFRED (Kidd Lab Home Page).

Five of the 16 haplotypes never occur at a frequency >.04 and are present in only a minority of the populations. Conversely, no haplotype was definitely present in all population samples. Thus, every haplotype has a frequency range across these 29 populations, with a minimum of 0 and a maximum that ranges from .02 (for 1122 and 2211 [where "1" denotes site absence, and "2" denotes site presence]) to .78 (for 2121), depending on the haplotype.

Ancestral and Derived Alleles

On the basis of the sequence of other hominoid species, Iyengar et al. (1998) determined the ancestral states of the *Bgl*II (site present), *Pvu*II (site absent), and *Msp*I (site present) polymorphisms. As part of the present study, the same primers and PCR protocol that were used to type humans were used to amplify the region homologous to the *Xmn*I polymorphism in two chimpanzees and two gorillas. The PCR products do not cut with *Xmn*I, implying that site absence is the ancestral human state. Iyengar et al. (1998) noted that, for the *Bgl*II, *Pvu*II, and *Msp*I sites, whether the ancestral or derived allele was the more common depended on which human population was studied. The same holds for the *Xmn*I site, with the frequency of the ancestral allele (site absent, or "1") ranging from .061 in the Hakka to .944 in the Finns.

The haplotype with all four ancestral hominid alleles, 2121, was undoubtedly the original one from which the other 15 haplotypes evolved through a combination of mutations and crossovers. The ancestral haplotype has its highest frequencies in Native American populations; in six of the seven Native American populations (all but the Karitiana), it is more frequent (range .41–.78) than in any other population studied. Its next most frequent occurrence is at .26 in the Biaka. In contrast, the quadruply derived haplotype, 1212, is most frequent (range .46–.73) in the eastern-Asian populations.

Haplotype-Frequency Differences

The Workman and Niswander (1970) genetic-heterogeneity test was employed to compare *PAH* haplotype frequencies for all 29 population samples pairwise (406 tests), in order to test the null hypothesis that each paired sample was drawn from the same population. Some 368 test comparisons have $P \leq .001$, whereas only 16 of the comparisons are not significant ($P > .050$). Even such (presumably) recently separated populations as the Han from Taiwan and the Han from southern China (sampled in San Francisco) are, by the Niswander and Workman (1970) test, significantly different samples. The 39 comparisons that are either nonsignificant or only weakly to moderately significant are almost entirely be-

Table 2

F_{ST}, by Geographical Regions and Globally, for Each *PAH* RSP and for the Four-Site Haplotype

	F_{ST}					
PAH	Global	Sub-Saharan Africa	Europe and Southwestern Asia	Eastern Asia	North America	South America
*Bgl*II	.16	.05	.03	.05	.12	.18
*Pvu*II	.14	.07	.03	.06	.12	.07
*Msp*I	.23	.18	.02	.04	.01	.01
*Xmn*I	.31	.25	.10	.02	.01	.02
Haplotype	.13	.06	.04	.05	.05	.08

tween populations within the same geographic region. Of the pairwise comparisons that involved populations in different geographic regions, all but 13 were significant at $P < .001$. Eight were significant at $.001 < P < .005$: Cambodians with Yemenites; Yakut with Adygei, Russians, mixed Europeans, and Atayal; and Karitiana with Druze, mixed Europeans, and Nasioi. Only five such pairwise comparisons had $P > .005$: three, involving the Karitiana (with Adygei, Russians, and Finns), were $P < .05$; one, comparing Nasioi and Danes, was $P < .01$; and one, comparing Yakut and Yemenites, was not significant.

Pairwise Linkage Disequilibrium

All six pairwise D' values (table 4) were calculated, and the significance level was evaluated by the asymptotic χ^2 test statistic (see eq. 3.10 in the work of Weir [1996]). On a global level, only the *Bgl*II-*Pvu*II and *Msp*I-*Xmn*I values gave a consistent pattern. *Bgl*II-*Pvu*II comparisons gave mostly negative D' values, most of which were even more extreme than $-.6$ and significant at $P < .001$, with only five exceptions: two populations (Biaka and Nasioi) did not have significant disequilibrium; one population (Ethiopian Jews) had a large positive D' value of .63, significant at $P < .05$; and two populations (Mbuti and Yoruba) had less-extreme negative D' values, significant at $P < .05$ and $P < .005$, respectively. For the *Msp*I-*Xmn*I comparison, only one population (Hakka) did not have a significant D', and one population (Finns) had a value significant at only .01; all others had D' values more extreme than $-.79$, with $P < .001$. In combination, these two pairwise comparisons are sufficient to explain the significance of the overall nonrandomness indicated by the likelihood-ratio and permutation tests (see below).

The four pairwise comparisons between sites at either end of the large central region give a consistent pattern of significance for only a subset of the populations: the four North American Indian populations and the Rondonian Surui from Brazil. For these five populations, the D' values are uniformly large (i.e., $|D'| > .50$) and are significant at $P < .001$, for all four comparisons. Only one other population, the Yoruba, has a value that reaches this level of significance—and then for only one of the four comparisons. In most of the remaining comparisons, the value is not significantly different from zero.

Overall Linkage Disequilibrium

The asymptotic likelihood-ratio χ^2 for overall linkage disequilibrium is given in table 3. This χ^2 has 11 df, and the significance levels are <.0001 for all populations except Ethiopian Jews, in whom significance reaches only $P < .001$. By the permutation test with 1,000 permutations, all samples but one were significant at $P < .001$, because none of the permutated samples gave a likelihood-ratio statistic larger than the observed value; the exception was the Yoruba, in whom two permuted samples exceeded the likelihood ratio of the observed value, thereby giving a significance level of $P = .002$, with an upper confidence level of .005. Thus, we can confidently state that significant nonrandomness of alleles on chromosomes exists in all populations studied. This is not surprising, since we observed that most populations showed significant disequilibrium within each of the two pairs of sites at either end of the region.

Figure 2 graphs the estimate of the ξ coefficient for the four-site–haplotype system, in 29 populations. This standardized measure shows considerable variation among populations, even within a geographic region. The arrangement of the populations within geographic regions is arbitrary. However, the clear impression is that there are lower values within Africa and larger values outside Africa, with an increase, on average, as the distance from Africa increases. The unweighted regional average of the ξ coefficients does increase from .92 in Africa to 1.86 in Europe, 1.88 in eastern Asia, 2.64 in North America, and 1.84 in South America. Although this is not a simple linear trend, all averages outside Africa are at least twice as large as the African average, and the Native American populations, especially those in North America, show more nonrandomness than is seen elsewhere.

Table 3

PAH Four-Site Haplotype-Frequency Estimates, Sample Sizes, Expected Heterozygosities, and Global Tests of Linkage Disequilibrium, for 29 Population Samples

Population (2N)	Expected Heterozygosity	Likelihood-Ratio χ^2 [a]	Linkage Disequilibrium of Haplotype [b]															
			1111	1112	1121	1122	1211	1212	1221	1222	2111	2112	2121	2122	2211	2212	2221	2222
Biaka (140)	.85	80.9	.018	.136	.147	0	0	.019	.109	0	.042	.055	.260	0	0	.076	.131	.008
Mbuti (78)	.78	57.4	0	.382	.020	0	0	.116	.110	.038	0	.203	.075	0	.013	.043	0	0
Yoruba (112)	.83	46.6	.039	.054	.190	0	0	0	.265	.014	.029	.093	.213	.015	0	0	.087	0
Ethiopians (62)	.80	35.5	0	.122	.195	0	0	.118	.282	0	.037	0	0	0	0	0	.227	.017
Yemenites (86)	.81	77.7	.030	.025	.030	0	.044	.341	.133	.014	0	.135	.176	0	0	.030	.018	.026
Druze (154)	.74	245.0	0	.022	.044	0	.006	.331	.331	0	0	.127	.132	0	0	0	.006	0
Samaritans (80)	.85	63.1	0	0	.112	0	.143	.126	.206	0	.157	.124	.132	0	0	0	0	0
Adygei (108)	.77	162.0	0	.022	.043	0	.019	.297	.313	.010	0	.143	.153	0	0	0	0	0
Russians (96)	.75	119.8	.010	0	.062	0	.027	.233	.407	0	.035	.121	.104	0	0	0	0	0
Danes (102)	.80	125.8	0	.018	.019	0	0	.275	.268	0	.010	.146	.083	0	0	.042	.141	0
Finns (72)	.67	49.5	.032	0	0	0	.208	.028	.523	0	.065	.028	.101	0	0	0	.015	0
Irish (162)	.80	84.2	.021	.050	.177	0	.089	.124	.365	.007	.010	.073	.065	0	0	0	.020	0
Europeans (180)	.77	191.2	0	.009	.047	0	.032	.317	.291	.012	0	.090	.168	0	.003	0	.032	0
San Francisco Chinese (116)	.55	84.0	0	.070	0	0	0	.643	.070	.009	.009	.173	.007	0	0	.018	0	0
Taiwanese Chinese (100)	.51	70.3	0	.052	.040	0	.010	.672	.015	0	0	.173	.004	.010	0	.012	.011	0
Hakka (82)	.69	55.4	0	.037	0	0	.061	.465	0	.230	0	.182	0	.026	0	0	0	0
Japanese (98)	.51	80.1	0	.063	.029	0	.018	.678	.028	0	0	.157	.027	0	0	0	0	0
Ami (80)	.45	73.1	0	.043	.020	0	0	.728	.060	0	0	.117	.021	.012	0	0	0	0
Atayal (84)	.50	89.6	0	0	0	0	.040	.685	.157	0	.009	.100	.010	0	0	0	0	0
Cambodians (50)	.64	44.3	0	.041	.019	.021	0	.459	.039	0	.021	.378	0	0	0	.021	0	0
Nasioi (46)	.67	51.3	0	.072	0	0	0	.302	.474	.022	0	.037	0	0	0	.068	.026	0
Yakut (102)	.77	76.5	.032	0	.029	0	.126	.405	.180	.012	.030	.053	.111	0	.011	0	.011	0
Cheyenne (112)	.70	182.6	0	0	0	0	.064	.185	.302	.029	0	0	.411	0	0	0	.009	0
Arizona Pima (102)	.57	109.8	.036	.048	0	0	0	.129	.042	0	.029	0	.633	0	.020	0	.064	0
Mexican Pima (106)	.37	170.7	0	.009	0	0	.009	.112	.030	0	0	.058	.782	0	0	0	0	0
Maya (106)	.68	124.4	.019	0	.083	.010	.048	.214	.080	0	0	.010	.505	.010	0	.010	.012	0
Ticuna (134)	.72	88.0	0	0	.058	0	.013	.026	.126	.008	.039	.149	.484	.008	.008	.010	.070	0
Rondonian Surui (92)	.72	138.4	0	0	.155	0	.025	.291	.073	.033	0	0	.411	0	0	.013	0	0
Karitiana (108)	.78	98.9	.008	.105	.065	0	.020	.138	.395	0	.013	.086	.135	0	0	0	.034	0

[a] Comparison of data (as multisite phenotypes): estimated haplotype frequencies versus equilibrium haplotype frequencies. All values are statistically significant at $P < .0001$— except for the Ethiopian Jews, in whom the values are statistically significant at $P < .001$.

[b] The four restriction sites of each haplotypes, listed in order from left to right, are *Bgl*II, *Pvu*II, *Msp*I, and *Xmn*I. The frequencies shown are maximum-likelihood estimates calculated by the HAPLO program (Hawley and Kidd 1995).

Table 4

***D′*, χ^2, and *P* Values at the *PAH* Locus, for the Six Possible Pairings of Polymorphic Sites**

	*Bgl*II, *Pvu*II			*Bgl*II, *Msp*I			*Bgl*II, *Xmn*I			*Pvu*II, *Msp*I			*Pvu*II, *Xmn*I			*Msp*I, *Xmn*I		
Population	D'[a]	χ^2[b]	P[c]	D'[a]	χ^2[b]	P[c]	D'[a]	χ^2[b]	P[c]	D'[a]	χ^2[b]	P[c]	D'[a]	χ^2[b]	P[c]	D'[a]	χ^2[b]	P[c]
Biaka	.13	.9	NS	.13	1.5	NS	−.17	2.3	NS	.20	1.5	NS	.01	.0	NS	−.96	101.2	.001
Mbuti	−.48	4.2	.050	−.08	.1	NS	−.10	.5	NS	.42	9.6	.005	−.36	6.0	.050	−.92	57.5	.001
Yoruba	−.46	10.5	.005	−.23	2.1	NS	.31	3.0	NS	1.00	17.8	.001	−.78	8.5	.005	−.79	54.5	.001
Ethiopians	.63	5.3	.050	.52	2.6	NS	−.77	4.9	.050	.34	5.0	.050	−.18	1.3	NS	−.90	45.9	.001
Yemenites	−.68	38.2	.001	.29	6.9	.010	−.13	1.2	NS	−.21	3.6	NS	.29	6.3	.050	−.82	50.7	.001
Druze	−.96	107.1	.001	.01	.0	NS	.00	.0	NS	−.06	.2	NS	.05	.1	NS	-1.00	150.3	.001
Samaritans	-1.00	50.9	.001	−.29	3.9	.050	.14	.8	NS	−.04	.1	NS	.05	.1	NS	-1.00	21.8	.001
Adygei	-1.00	80.4	.001	−.00	.0	NS	.02	.0	NS	−.05	.1	NS	.03	.1	NS	−.96	95.8	.001
Russians	-1.00	67.9	.001	−.30	4.2	.050	.17	1.8	NS	.13	1.1	NS	−.02	.0	NS	-1.00	70.9	.001
Danes	−.77	31.5	.001	.04	.1	NS	−.07	.3	NS	.28	3.1	NS	−.22	2.0	NS	-1.00	98.0	.001
Finns	−.91	53.6	.001	−.13	.6	NS	.37	2.2	NS	.30	3.4	NS	−.35	1.8	NS	-1.00	7.6	.010
Irish	−.80	32.2	.001	−.20	2.3	NS	.24	5.6	.050	.04	.2	NS	−.15	1.8	NS	−.96	87.1	.001
Europeans	−.83	111.3	.001	.29	5.2	.050	−.28	4.5	.050	−.30	5.9	.050	.26	4.3	.050	−.95	147.1	.001
San Francisco Chinese	−.88	67.5	.001	−.61	1.0	NS	.10	.0	NS	.69	1.8	NS	−.28	.3	NS	−.88	91.0	.001
Taiwanese Chinese	−.85	49.4	.001	.13	.6	NS	.11	.0	NS	−.55	6.8	.010	.38	3.2	NS	−.86	74.7	.001
Hakka	-1.00	66.4	.001	−.51	1.9	NS	1.00	1.4	NS	.59	3.1	NS	-1.00	1.7	NS	1.00	1.8	NS
Japanese	-1.00	58.0	.001	.17	1.1	NS	−.10	.5	NS	−.54	6.9	.010	.38	4.1	.050	-1.00	79.1	.001
Ami	-1.00	52.2	.001	.17	1.6	NS	−.07	.2	NS	−.33	4.0	.050	.24	2.0	NS	-1.00	70.5	.001
Atayal	-1.00	84.0	.001	−.50	.6	NS	.26	.2	NS	.50	.6	NS	−.26	.2	NS	-1.00	61.1	.001
Cambodians	−.90	32.0	.001	-1.00	3.1	NS	.37	.4	NS	−.05	.0	NS	.05	.0	NS	−.71	25.3	.001
Nasioi	−.24	2.1	NS	−.62	2.9	NS	.60	2.5	NS	1.00	6.1	.050	-1.00	5.6	.050	-1.00	42.1	.001
Yakut	−.86	61.2	.001	.34	6.1	.050	−.48	5.7	.050	−.31	6.6	.050	.56	9.6	.005	−.93	40.4	.001
Cheyenne	-1.00	107.9	.001	1.00	26.9	.001	-1.00	22.1	.001	-1.00	25.9	.001	1.00	21.3	.001	−.82	61.8	.001
Arizona Pima	−.56	31.8	.001	.78	59.4	.001	-1.00	64.1	.001	−.44	18.8	.001	.64	25.9	.001	-1.00	61.8	.001
Mexican Pima	-1.00	99.0	.001	.77	51.9	.001	−.70	45.8	.001	−.75	46.5	.001	.68	40.6	.001	-1.00	99.8	.001
Maya	−.89	57.7	.001	.88	42.4	.001	−.78	26.7	.001	−.85	57.4	.001	.81	41.8	.001	−.89	66.0	.001
Ticuna	−.66	49.7	.001	−.31	1.3	NS	.27	.7	NS	.11	.2	NS	−.16	.3	NS	−.89	83.1	.001
Rondonian Surui	−.93	44.9	.001	.91	27.2	.001	−.91	28.3	.001	-1.00	58.6	.001	1.00	60.8	.001	−.88	69.6	.001
Karitiana	−.78	34.7	.001	.00	.0	NS	−.03	.0	NS	.27	6.8	.010	−.29	6.2	.050	-1.00	90.2	.001

[a] Value has been tested for statistical significance, under the null hypothesis that $D' = 0$.

[b] 1 df.

[c] Significance interval. "NS" indicates that $P > .050$. A probability of .050 is assigned when $.050 \leq P > .010$; a probability of .010 is assigned when $.010 \leq P > .005$; a probability of .005 is assigned when $.005 \leq P > .001$; and a probability of .001 is assigned when $P \leq .001$.

Segment Linkage Disequilibrium

The results of the segment disequilibrium test are given in table 5, for all 29 populations. The test compares the likelihood ratio for the observed data with the distribution of likelihood-ratio statistics from 1,000 permutations, summarized as the mean and variance. The probability is the fraction of the 1,000 permutations that had a likelihood-ratio statistic greater than the observed value. Not surprisingly, the test supports the consistent pairwise results for highly significant nonrandomness across the central segment in all of the North American Indian populations and in the Rondonian Surui of Brazil. Only two other populations give significant results: the Mbuti, at $P = .003$; and the Taiwanese Chinese, at $P = .039$; in the context of 29 tests, the value for the Taiwanese Chinese is probably not meaningful.

Discussion

Site and Haplotype-Frequency Variation

No previous studies of these polymorphisms have included American Indian or African populations or large numbers of eastern Asians. These SNPs were originally discovered in populations of European ancestry, as tools for genetic counseling of families with PKU; heterozygosities close to 50% were optimal for that purpose, and all four of these SNPs have heterozygosities in the range of 30%–50%, in most European populations. Although each site is a simple biallelic system, all four sites are also highly polymorphic globally. In our study, none of the alleles at any of the sites has gone to fixation in any population, and the minimum allele frequency in any population at any of the sites is .056 at the *Xmn*I site (allele "2") in Finns. The expected heterozygosity is likewise high for the four-site haplotypes, ranging from .85 in the Biaka to .37 in the Mexican Pima. Interestingly, each of the heterozygosities for the four markers individually and for the haplotype was higher, on average, in the seven Native American groups than it was in the seven Asian groups. Regional heterozygosities at a large number of other loci in some of these same population samples show lower average heterozygosities in American Indian populations than in eastern Asian popula-

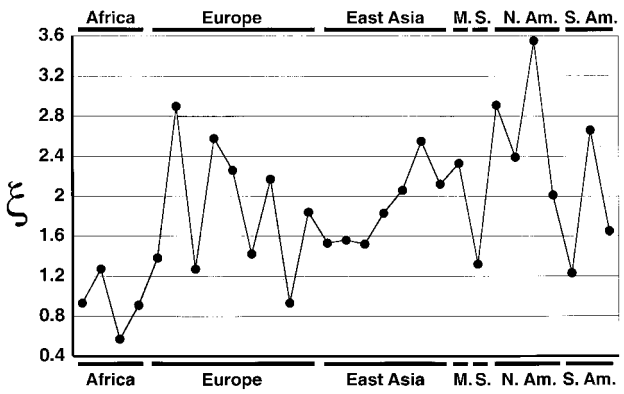

Figure 2 Overall linkage disequilibrium as the ξ coefficient for the four-site *PAH* haplotype in 29 populations. Coefficients are given for each population ordered, left to right, in the same order as the top-to-bottom sequence in tables 3 and 4. Geographic groupings are indicated across the top and bottom, as in figure 4; M. = Melanesia (Nasioi), and S. = Siberia (Yakut). All values are statistically significant at $P < .001$, except for the Yoruba, in whom the values are statistically significant at $P = .002$.

tions and similar levels of heterozygosity in the eastern Asian and in the European populations (e.g., see Kidd et al. 1991, 1993; Calafell et al. 1998; for data on other loci, see ALFRED [Kidd Lab Home Page]). The pattern here—of the lowest heterozygosities occurring in eastern Asia—is clearly unusual.

The haplotype frequencies for the population can be estimated directly by gene counting in a sample of unrelated individuals, by use of either phase-known genotypes obtained from family data (as in DiLella et al. 1986*b*) or molecular haplotyping methods (Ruano and Kidd 1991; Michalatos-Beloin et al. 1996). Alternatively, haplotype frequencies for the populations can be estimated from the multisite phenotype data in a sample of unrelated individuals by use of maximum likelihood as implemented in any of several computer programs (e.g., see Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995). We have estimated haplotype frequencies by using the maximum-likelihood method described by Hawley and Kidd (1995). The jackknife estimate of the standard error specifically accounts for the increased uncertainty in the haplotype-frequency estimates versus estimates based entirely on gene counting in a sample of the same size for which the binomial standard-error estimates would be appropriate. The fact that, in general, the jackknife estimates are not exceptionally larger than the binomial standard-error estimates is a reflection of the high percentage of unambiguous chromosomes in these samples. In more than half the populations, more than half the chromosomes in the sample were unambiguously specifiable on the basis of the multisite phenotype: either all sites were homozygous or only one site was heterozygous (table 6). This level of specification greatly constrains the estimates that the EM algorithm can produce. Additional constraint comes from the large number of individuals heterozygous at only two sites for whom only two possible genotypes (of the total of 136 possible a priori) are possible. As a fraction of all chromosomes sampled, these two categories of multisite phenotypes accounted for a minimum of 51% (the Ethiopian sample) to >90% (the Atayal and Cambodian samples) (table 6). All 16 of the possible haplotypes were observed to be present in at least one of the populations (table 3). Five haplotypes are globally quite rare. Each is an "observed" haplotype—that is, at least one individual heterozygous at only one of the four sites implies the presence of the haplotype—in at least one population. In some populations, each of these haplotypes may have been inferred to be present only by the maximum-likelihood method of estimation of haplotype frequencies.

Our analyses of genetic similarity in subsets of these 29 populations, using different data sets (Kidd and Kidd

Table 5

Linkage Disequilibrium Segment Test Results Comparing (*Bgl*II,*Pvu*II) Paired Sites versus (*Msp*I,*Xmn*I) Paired Sites, at the *PAH* Locus

Population	Observed Likelihood-Ratio χ^2	Mean[a]	Variance[a]	P[b]
Biaka	11.97	9.02	15.44	.206
Mbuti	19.40	9.12	10.43	.003
Yoruba	8.91	9.89	14.20	.560
Ethiopians	7.32	7.74	8.68	.510
Yemenites	16.27	10.04	12.89	.057
Druze	1.38	4.74	8.20	.951
Samaritans	7.75	4.95	9.40	.172
Adygei	3.31	5.77	8.22	.808
Russians	5.87	4.15	7.43	.224
Danes	6.43	5.21	6.52	.263
Finns	5.41	5.40	7.72	.417
Irish	7.19	8.30	11.88	.562
Europeans	10.85	8.44	14.31	.244
San Francisco Chinese	4.55	6.42	9.62	.691
Taiwanese Chinese	12.22	6.27	7.36	.039
Hakka	4.95	5.05	8.66	.434
Japanese	4.99	3.60	4.27	.199
Ami	3.90	3.21	4.08	.280
Atayal	.14	1.88	2.87	.887
Cambodians	7.71	6.54	9.29	.316
Nasioi	6.76	5.93	8.22	.388
Yakut	7.38	8.37	12.57	.566
Cheyenne	27.01	5.71	11.25	0
Arizona Pima	57.05	6.04	8.70	0
Mexican Pima	31.95	3.32	5.52	0
Maya	45.34	8.61	13.52	0
Ticuna	5.02	8.54	13.48	.846
R. Surui	56.33	8.34	13.35	0
Karitiana	5.08	6.23	10.02	.599

[a] For permuted distributions.

[b] Fraction of 1,000 permutations greater than that of the observed likelihood-ratio χ^2.

Table 6

Distribution of Individuals, by Number of Heterozygous Sites

Population (No.)	No. of Heterozygous Sites[a] (No. of Individuals)	
	Zero or One[b]	Two[c]
Biaka (70)	28	17
Mbuti (39)	21	8
Yoruba (56)	29	13
Ethiopians (31)	10	6
Yemenites (43)	12	19
Druze (77)	30	34
Samaritans (40)	12	13
Adygei (54)	22	21
Russians (48)	20	15
Danes (51)	21	18
Finns (36)	22	7
Irish (81)	35	23
Europeans (90)	28	27
San Francisco Chinese (58)	33	18
Taiwanese Chinese (50)	28	16
Hakka (41)	27	8
Japanese (49)	29	14
Ami (40)	25	10
Atayal (42)	23	15
Cambodians (25)	14	9
Nasioi (23)	13	7
Yakut (51)	21	18
Cheyenne (56)	29	16
Arizona Pima (51)	28	6
Mexican Pima (53)	34	9
Maya (53)	22	10
Ticuna (67)	35	23
Rondonian Surui (46)	22	3
Karitiana (54)	22	13

[a] Based on the assumption of codominant, biallelic genetic systems; only the two least ambiguous phenotype classes are shown.

[b] Gene counting. Individuals are either homozygous at all four polymorphic sites or heterozygous at only one site; both haplotypes carried by each individual are thus fully specified.

[c] Ambiguity. Individuals are heterozygous at two of the four sites, resulting in cis-trans ambiguity.

1996; Calafell et al. 1998), as well as interim analyses of multilocus data on all 29 populations (Kidd Lab Home Page), are consistent in showing four groups of populations corresponding to the geographic locations of the populations—Africa, Europe and southwestern Asia, eastern Asia, and the Americas—with genetic distances within each group that, in general, are smaller than those between groups. The Nasioi from Melanesia and the Yakut from Siberia are distinct and do not cluster either with each other or with any of the four groups. Simple inspection of site and haplotype-allele frequencies at *PAH* (table 3) suggests that this locus gives a concordant pattern that is supported by F_{ST} values being smaller for populations within each geographically defined region than they are globally, for both the haplotype data and the individual site data (with one exception) (table 2). It is also supported by a principal-components analysis (PCA) of the haplotype frequencies at *PAH* (fig. 3). All of these analyses support the validity of the regional summaries of heterozygosity (table 1) and of haplotype-frequency data (fig. 4).

For haplotype frequencies, the pattern of results for the pairwise comparison of the samples, with the Workman and Niswander (1970) genetic-heterogeneity test (results not shown), is illuminating and both supports the validity of the summaries in tables 1 and 2 and in figure 4 and demonstrates that most samples represent distinct populations, since the vast majority of the 406 pairwise comparisons differ significantly, at $P < .001$.

Nonrandomness of Alleles on Chromosomes

The presence of disequilibrium in a complex haplotype is determined by comparison of the maximum-likelihood estimates of the haplotype frequencies with the haplotype frequencies predicted by multiplication of the allele frequencies at the individual sites. Determining the overall significance levels of any linkage disequilibrium is straightforward by likelihood-ratio or "direct" χ^2 statistics if the haplotypes are simple, two-site systems; however, when there are many sites involved, experience demonstrates that the likelihood ratio–test statistic does not always closely approximate a χ^2 distribution, because (1) the expected number of some phenotypes may be small for some populations and (2) not all markers are typed on all individuals—that is, data are incomplete (Kidd et al. 1998; Zhao et al. 1999); and both of these conditions pertain in the present data set. Therefore, instead of relying on the asymptotic theory, we have also utilized the permutation test (Zhao et al. 1999), to ob-

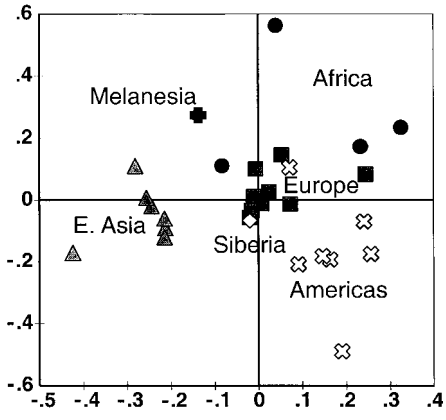

Figure 3 PCA (as described in Chang et al. 1996) of 29 populations, on the basis of *PAH* haplotype-frequency data. African populations are represented by circles, European and southwest-Asian populations by squares, eastern-Asian populations by triangles, Native American populations by crosses, the Melanesian population by a plus sign, and the Siberian population by a diamond. These first two principal components account for 68.0% of the variance.

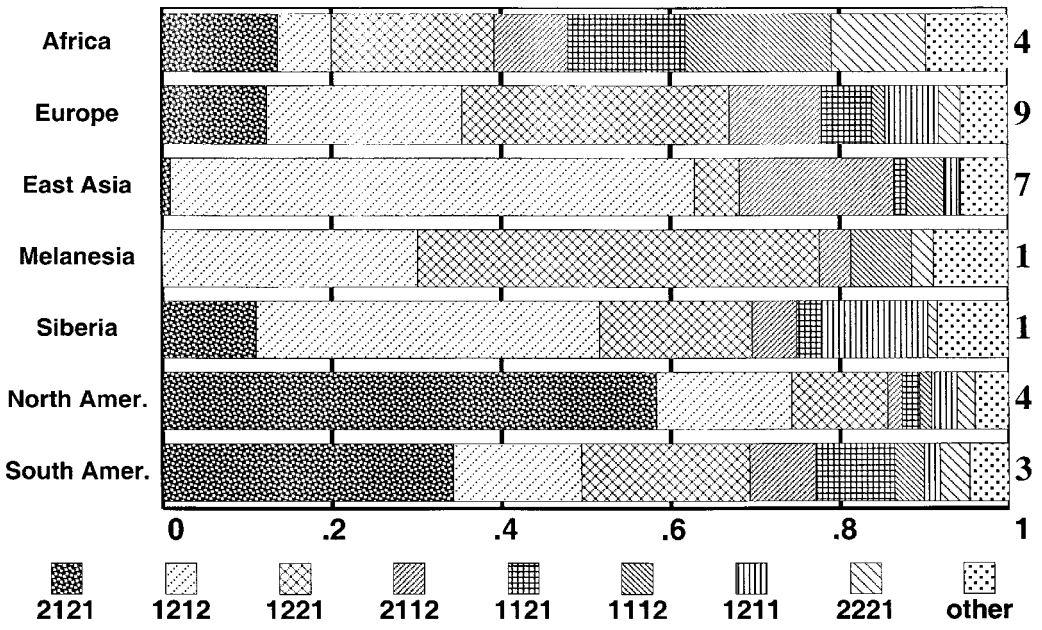

Figure 4 Average frequencies of the eight most frequent haplotypes, by geographic region. The averages are given as horizontally stacked bars, on the basis of data in table 3, with the number of populations, averaged for each region, given in the right margin. Melanesia and Siberia have only one population each—Nasioi and Yakut, respectively. The haplotypes are coded as in table 3: "1" indicates that the site is absent, and "2" indicates that the site is present, for the sites ordered as in figure 1. The frequency of the ancestral haplotype is given, starting at the left margin, and the frequency of the quadruply derived haplotype appears immediately to its right.

tain the statistical significance of observed likelihood-ratio statistics (Good 1995). The statistical-significance levels of overall linkage disequilibrium were determined for each of the 29 populations, from both asymptotic χ^2 (by HAPLO) distributions and from 1,000 permutations. Exact significance levels cannot be compared between the two approaches. Likelihood-ratio χ^2 values were significant at $P < .0001$ for all populations but the Ethiopian Jews ($P < .001$). In contrast, the permutation results could give significance values of $P < .001$ only when none of the 1,000 permutations exceeded the observed value. By this test, only one population had a significance value, for overall disequilibrium, that was >.001: the Yoruba, at $P = .002$. The Yoruba also have the smallest ξ value of any population studied (fig. 2). Interestingly, the likelihood-ratio χ^2 for overall disequilibrium was smaller for the Ethiopian Jews and the Cambodians than for the Yoruba, highlighting the nonidentity of the two measures of significance. Although the permutation test gives only an upper bound (determined by the number of permutations), it is not dependent on the assumption of an asymptotic distribution and is therefore preferable.

When overall linkage disequilibrium is significant, a subsequent question that we address is how different sites or segments of DNA contribute to that overall linkage disequilibrium. Because the number of all pairwise linkage-disequilibrium calculations is large, and because there is no meaningful integration of the statistics for multiple, nonindependent pairs, some researchers have estimated and tested higher-order disequilibrium coefficients (Piazza 1975; Long et al. 1995), but these higher-order coefficients are difficult to relate to aspects of the underlying biology, such as the distribution of disequilibrium across specific segments within a haplotype. We have, therefore, applied the Zhao et al. (1999; also, H. Zhao, A. J. Pakstis, J. R. Kidd, and K. K. Kidd, unpublished data) method of segment analysis.

Significance level is not a direct measure of amount of disequilibrium, but, in this study, it tracks the D' values reasonably well, since all sites are reasonably heterozygous in all populations and since sample sizes are all in the range of 25 to slightly >100 individuals. Thus, when all pairwise comparisons are significant for a population, there is no problem in interpretation of the results. However, in some cases, especially for the comparisons involving one site at each end of the locus, the different pairwise comparisons do not give such consistent results. Specifically, in the Yoruba one of the four comparisons was significant at $P < .001$, one was significant at $P < .005$, and the other two were not significant; in the Mbuti, one comparison was significant at $P < .005$, one was significant at $P < .05$, and two were not significant; in the Ethiopian Jews and Nasioi, two

comparisons were significant at $P < .05$, and two were not significant; in the Yemenites, Japanese, and Karitiana, one comparison was significant at $P < .01$, one was significant at $P < .05$, and two were not significant; in the Samaritans, Russians, Irish, and Ami, one comparison was significant at $P < .05$ and three were not significant; in the Europeans and Yakut, all four comparisons were significant at $P < .05$ or less; and, in the Taiwanese Chinese, one comparison was significant at $P < .01$, and three were not significant. Especially in these cases that have inconsistent significance levels, the nonindependence of the four tests and the multiple tests being done complicate interpretation. In these populations with inconsistent results of pairwise tests of disequilibrium across the middle segment, the segment test (table 5) gives a possibly significant result in only three cases: the results in the Mbuti (at $P = .003$), the Yemenites (at $P = .057$), and the Taiwanese Chinese (at $P = .039$) can be considered to be possibly significant. In the context of multiple populations being tested, the disequilibrium in the Mbuti may be the only one that is possibly significant. In the two cases in which all four pairwise comparisons gave at least borderline ($P < .05$) significance, the segment permutation test was clearly not significant, at $P = .24$ and $P = .57$ for Europeans and Yakut, respectively. In the other populations, either all four pairwise comparisons were nonsignificant or all four pairwise comparisons were significant at the $P < .001$ level. The segment test gave concordant results in these cases. We judge the segment test to be more accurate and clearer than the compilation of possibly discordant pairwise tests, because it is a single measure utilizing the information at all four sites and does not assume an asymptotic distribution. Furthermore, the segment test relates to the underlying biology in a more straightforward manner than do the higher-order coefficients. Part of the lack of concordance among the pairwise tests may be attributed to the assumption of asymptotic distribution.

Comparison of these results with earlier results shows that the distance between markers and the type of marker are probably both important and that the historical time frame for which disequilibrium at a haplotype will be informative is related to both. At least for SNPs, we see that linkage disequilibrium is highly significant at a distance of ~1.8 kb, in essentially all populations in all regions of the world. This global consistency presumably reflects a pattern, established early in human evolution, that, because of the low frequency of recombination within this short molecular distance, has not decayed. At *CD4,* the primary factor was the short tandem-repeat polymorphism (STRP), which showed little to no disequilibrium in Africa, with the biallelic marker 10 kb away, but showed essentially complete disequilibrium in non-African populations (Tishkoff et al. 1996*a*). At *PAH,* the comparably spaced markers (*Bgl*II and *Pvu*II, at ~7 kb) show linkage disequilibrium in some of the sub-Saharan populations, presumably because mutation rates are so much lower than those for an STRP. However, the disequilibrium, as determined by both ξ and D', is much stronger in the non-African populations. At *DRD2* (Kidd et al. 1998) and *DM* (Tishkoff et al. 1998), disequilibrium was essentially complete between the outermost SNPs, ~25 kb apart in both cases, in virtually all non-African populations but was much less in sub-Saharan populations. At *PAH,* there is, as yet, no pair of sites at that distance. At the larger distance of ~65 kb across the central segment of *PAH,* we see significant linkage disequilibrium in the American Indian populations and, possibly, in one African population, the Mbuti. Presumably, this longer distance, with a relatively higher recombination rate, is probing more-recent founder events. One of those is associated with the American Indian lineage prior to both the spread throughout the Americas and the diversification of those populations. The other is more difficult to identify, if we accept this as a significant result, as discussed earlier. It could be a recent founder effect specific to the Mbuti, or it could be more ancient and have involved the founder population ancestral to several modern Pygmy groups. The absence of significant linkage disequilibrium in the other Pygmy group, the Biaka, could be attributed to recent admixture with non-Pygmy groups, as has been hypothesized by Cavalli-Sforza (1986, p. 406).

Evolutionary Implications

With mutation rates for single nucleotides estimated at $\sim 10^{-8}$ (Crow 1995; Li et al. 1996), any SNP must represent effectively a single mutational event that has reached polymorphic frequencies through random genetic drift (or selection or hitchhiking). In other hominoids, the nucleotides at the position of the human SNP determine the ancestral allele of the SNP if they correspond to one of the human alleles (Iyengar et al. 1998). All of the alleles of the derived type are identical by descent (IBD) from either the original mutant or a later copy. All of the alleles of the ancestral type are IBD from some copy of the ancestral allele that may have existed either earlier than the mutation event or more recently, depending on the present frequency and the population history. Coalescent theory can predict the probability distributions of when those most recent common ancestors existed for alleles of both types, if enough is known about the history of the population(s). By extension, the ancestral haplotype will be the one composed entirely of ancestral alleles at the individual SNPs; however, the pool of ancestral haplotypes will not necessarily all be IBD, since recombination can regenerate the ancestral pattern from haplotypes with derived al-

leles at different sites. The probability of that is locus specific and population specific, depending on both recombination rates between sites and the frequencies of the appropriate heterozygotes.

Because the haplotype frequencies vary so much, even among populations in the same large geographic regions, it is difficult to make meaningful statistical statements about how the haplotypes evolved from the ancestral to the quadruply derived state. Moreover, from inspection of haplotype frequencies (table 3 and fig. 4), one sees that all populations have primarily ancestral and doubly derived configurations, for both of the closely spaced pairs of sites. Thus, for both of the two close pairs of sites at either end of the haplotype, both "intermediate" configurations—that is, 11 and 22—are rare to absent, around the world. We must conclude that the frequencies of those intermediate haplotypes in humans cannot provide information on the evolutionary history of the haplotypes. Indeed, the low frequencies and patchy occurrences of these "intermediate" configurations could as well reflect the chance survival of the uncommon crossovers in these small regions, since all populations are reasonably heterozygous for the ancestral 21 and derived 12 configurations: these transitional stages of evolution from the ancestral to derived haplotypes undoubtedly existed but may not have survived. Thus, the present "intermediates" may not be IBD with the original transitional haplotypes.

A problem in comparing our data to previously published haplotype frequencies at *PAH* is that many studies have based their frequencies only on the haplotypes that could be unambiguously assigned; haplotypes in multiply heterozygous individuals were not counted unless phase was resolved by the use of relatives. This introduces a definite bias that can be strong if two common haplotypes differ at multiple sites. As can be seen from the frequencies in table 3 and figure 4, this is commonly the case in most regions of the world. Figure 4 shows that both the ancestral pattern (2121) and the quadruply derived pattern (1212) are either the most common haplotypes or among the most common haplotypes everywhere but eastern Asia and Melanesia. The EM algorithm, in contrast, provides unbiased maximum-likelihood estimates, utilizing all available information.

Out of Africa

Our previous haplotype studies of *CD4, DM,* and *DRD2* (Tishkoff et al. 1996*a,* 1998; Kidd et al. 1998) have provided strong evidence for the out-of-Africa model of human expansion, with a very marked founder effect associated with the expansion out of Africa. Those loci show less linkage disequilibrium in sub-Saharan populations than in the non-African populations, leading to the conclusion that the founder effect established a pattern of linkage disequilibrium that is preserved in virtually all non-African populations studied. The data for *PAH* haplotypes presented here are in general agreement with this model but support it less strongly while suggesting refinements of the model. The haplotype heterozygosity is higher, on average, in Africa than elsewhere (table 1), but only barely so, and the difference from the average heterozygosity in Europe (the next highest value) is not significant. However, the *PAH* analyses do not yet incorporate an STRP into the haplotype system, in contrast to the studies of *CD4, DM,* and *DRD2*. In these latter studies, the STRPs were a major factor in the large differences seen between sub-Saharan and non-African populations. Moreover, figure 4 gives a clear impression that there are more haplotypes at moderate frequencies in Africa than there are in any non-African region.

The overall linkage-disequilibrium coefficient, ξ, is lower for African populations, on average, but some individual non-African populations have a value lower than that of at least one of the sub-Saharan populations. We also note that, in the samples of the larger African populations, there is a tendency for the strength of disequilibrium to be inversely related to the distance between sites. For the *Msp*I-*Xmn*I pair, separated by <2 kb, linkage disequilibrium is high for all the African samples. For the *Bgl*I-*Pvu*II pair, separated by ~7 kb, the linkage disequilibrium is low ($|D'| < .5$ for the sub-Saharan populations and $|D'| = .63$ for Ethiopians) and is either not significant or of only borderline significance. This contrasts sharply with the large and highly significant D' for this pair of sites in most of the European populations (table 4).

Homogeneity in Eastern Asia

In a previous study of eastern-Asian populations, Daiger et al. (1989*b*) found one haplotype at ~80% frequency. In contrast, the frequency of the comparable haplotype in our study is 46%–73% among eastern-Asian populations, with an unweighted average of 62% (fig. 4). Since the four-site haplotype in our study would comprise a superset of chromosomes, including the eight-site haplotype identified by Daiger et al. (1989*b*), as well as other haplotypes that may be present, it seems safe to conclude that the Daiger et al. (1989*b*) estimate was at the upper end of the range in eastern Asia. That range is still very homogeneous and part of a distribution of haplotypes that is much different than that seen in any other part of the world.

American Indians Compared with Eastern Asians

The North American Indians have a pattern of haplotype frequencies that is markedly different from that of eastern Asians; South American Indians have a pattern

similar to that of the North American Indians. The distinction from the eastern-Asian pattern is supported by both the PCA of haplotype frequencies (fig. 3) and the segment disequilibrium test (table 5). In our previous studies of most of these same population samples, for *CD4, DRD2,* and *DM* haplotypes, we did not see such a marked difference (Tishkoff et al. 1996*a,* 1998; Kidd et al. 1998). In those cases, the American Indian populations were similar to each other, as were the eastern-Asian populations, but the two patterns did not differ dramatically. The populations were distinguished in the second principal component—rather than in the first, which is the case in *PAH* (fig. 3). The American Indian pattern involved the same few haplotypes seen in eastern Asia, but with evidence of more drift causing one of those haplotypes to become more frequent and another to become less frequent than they are in eastern Asia. At *PAH* we see a very different pattern: a haplotype nearly absent in eastern Asia (i.e., the ancestral, 2121 haplotype) is the most common haplotype in the Americas. *PAH* is not unique in that pattern: at *DRD4,* the seven-repeat allele at the exon 3 VNTR is essentially absent in eastern-Asian populations but is the most common allele in the American Indian populations (Chang et al. 1996). *DRD4* is located on chromosome 11, at 11p15.5, a location that is not correlated—in any way that we know—with the location of *PAH* on chromosome 12, at 12q22-q24.2. In the ancestry common to American Indian populations, there is clearly an element that makes it quite distinct from the ancestry common to modern eastern-Asian populations. To the degree that tree diagrams can represent history, the trees for genetic distances calculated from multiple loci, including *PAH* treated either as a single site or as the four-site haplotype, show that the American Indian lineage diverged from the eastern-Asian lineage considerably before diversification of the modern eastern-Asian populations (Kidd and Kidd 1996; Kidd Lab Home Page).

The observation that the ancestral haplotype is most frequent in the Americas does not imply that it originated there. Drift can change haplotype frequencies in any direction, and frequency need not provide any evidence of origin. Indeed, if American Indians and eastern Asians share a remote common ancestry, as is generally accepted, and if that common ancestor was more similar to Africans and Europeans, then drift evidently has "pushed" the frequencies of ancestral and quadruply derived haplotypes in opposite directions (figs. 3 and 4), in the American Indian lineages and the eastern-Asian lineages.

Random genetic drift has affected each locus independently (if separated by just a few cM); and, for American Indian ancestry, the implications are unclear, because the data are mixed—with *CD4, DM,* and *DRD2* showing one pattern and with *PAH* and *DRD4* showing another. A distribution of similarities is expected as a function of both how long ago the lineages divided and the effective population sizes of the resulting lineages. Since only a few loci have been studied to date, and since the patterns of similarity differ substantially between the two regions, it is obvious that a single locus is not a good estimator of evolutionary histories of populations in these two geographic regions *and* that the few loci that so far have been studied do not allow a good estimate of the distribution of similarities expected for independently evolving (drifting) loci.

Future Studies

The use of haplotypes to study population and locus histories is still new. We are beginning to titrate disequilibrium with founder effects in populations and with distance between polymorphisms within haplotypes. As data are accumulated at more loci and in more populations, we shall be able to gain a much more comprehensive understanding of genome evolution and population diversification. Our next step for *PAH* will be to gather data to subdivide the distance between the *Pvu*II and *Msp*I sites and to extend the haplotype span. Subdividing the distance may provide insights into the evolution of the *PAH* locus and, in combination with increasing the span of the haplotype, provide more data on the idiosyncratic histories of the populations being studied. Global surveys of haplotypes of comparable molecular extent at other loci are also needed.

Ultimately, an understanding of the evolutionary histories of haplotypes on normal *PAH* chromosomes may be important in determining whether the distributions of haplotypes with different PKU mutations might have involved selection, as has been suggested by Kidd (1987), or can, as now seems more probable to us, be explained by the stochastic aspects of mutation and random genetic drift.

Acknowledgments

This work was supported in part by National Institutes of Health grant GM57672 (to K.K.K. and J.R.K.) and by National Science Foundation grant SBR-9632509 (to J.R.K.). Support also was provided by grants from the Alfred P. Sloan Foundation (to K.K.K. and J.R.K.), the National Science Council of Taiwan, National Science Council grant 88-2314-B-016-081 (to R.-B.L.), and 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: F. L. Black, L. L. Cavalli-Sforza, David Goldman, Kenneth Kendler, William Knowler, Frank Oronsaye, Leena Peltonen, and Kenneth Weiss. Randy C. Eisensmith has generously given us access to unpublished primers and allowed us to benefit from his specialized knowledge of the

PAH region. We also thank Neil Risch for helpful discussions on haplotype analyses. Special thanks are due to the many hundreds of individuals who volunteered to give blood samples for studies such as this. Without such participation by individuals from diverse parts of the world we would be unable to obtain a true picture of the genetic variation in our species.

Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html (for polymorphisms of *Bgl*II intron 1 [AF003965] and *Msp*I [AF003967])
Kidd Lab Home Page, http://info.med.yale.edu/genetics/kkidd (for population samples, ALFRED, FENGEN, and LINKD)
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nim.nih.gov/omim (for PKU [MIM 261600])
PAHdb, http://www.mcgill.ca/pahdb (for PAH mutations and haplotypes)

References

Anderson MA, Gusella JF (1984) Use of cyclosporin A in establishing Epstein-Barr virus-transformed human lymphoblastoid cell lines. In Vitro 20:856–858
Apold J, Eiken HG, Odland E, Fredriksen A, Bakken A, Lorens JB, Boman H (1990) A termination mutation prevalent in Norwegian haplotype 7 phenylketonuria genes. Am J Hum Genet 47:1002–1007
Baric I, Mardesic D, Gjuric G, Sarnavka V, Gobel-Schreiner B, Licther-Konecki U, Konecki DS, et al (1992) Haplotype distribution and mutations at the PAH locus in Croatia. Hum Genet 90:155–157
Bickel H, Bachmann C, Beckers R (1981) Neonatal mass screening for metabolic disorders: a collaborative study. Eur J Pediatr 137:133–139
Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK (1998) Short tandem repeat polymorphism evolution in humans. Eur J Hum Genet 6:38–49
Castiglione CM, Deinard AS, Speed WC, Sirugo G, Rosenbaum HC, Zhang Y, Grandy DK, et al (1995) Evolution of haplotypes at the DRD2 locus. Am J Hum Genet 57: 1445–1456
Cavalli-Sforza LL (1986) African Pygmies: an evaluation of the state of research. In: Cavalli-Sforza LL (ed) African Pygmies. Academic Press, Orlando
Chakraborty R, Lidsky AS, Daiger SP, Güttler F, Sullivan S, DiLella AG, Woo SLC (1987) Polymorphic DNA haplotypes at the human phenylalanine hydroxylase locus and their relationship with phenylketonuria. Hum Genet 76:40–46
Chang F-M, Kidd JR, Livak KJ, Pakstis AJ, Kidd KK (1996) The world-wide distribution of allele frequencies at the human dopamine D4 receptor locus. Hum Genet 98:91–101
Cheung KH, Miller PL, Kidd JR, Kidd KK, Osier MV, Pakstis AJ (2000*a*) ALFRED: a web-accessible allele frequency database. In: Altman RB, Dunker AK, Hunter L, Lauderdale K, Klein TE (eds) Pacific Symposium on Biocomputing 2000 Proceedings. World Scientific, Singapore, pp 639–650
Cheung KH, Nadkarni P, Silverstein S, Kidd JR, Pakstis AJ, Miller P, Kidd KK (1996) PhenoDB: an integrated client/server database for linkage and population genetics. Comput Biomed Res 29:327–337
Cheung KH, Osier MV, Kidd JR, Pakstis AJ, Miller PL, Kidd KK (2000*b*) ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. Nucleic Acids Res 28:361–363
Crow JF (1995) Spontaneous mutation as a risk factor. Exp Clin Immunogenet 12:121–128
Daiger SP, Chakraborty R, Reed L, Fekete G, Schuler D, Berenssi G, Nasz I, et al (1989*a*) Polymorphic DNA haplotypes at the phenylalanine hydroxylase (PAH) locus in European families with phenylketonuria (PKU). Am J Hum Genet 45: 310–318
Daiger SP, Reed L, Huang S-S, Zeng Y-T, Wang T, Lo WHY, Okano Y, et al (1989*b*) Polymorphic DNA haplotypes at the phenylalanine hydroxylase (PAH) locus in Asian families with phenylketonuria (PKU). Am J Hum Genet 45:319–324
Degioanni A, Darlu P (1994) Analysis of the molecular variance at the phenylalanine hydroxylase (PAH) locus. Eur J Hum Genet 2:166-176
Dempster, AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38
Dianzani I, Devoto M, Camaschella C, Saglio G, Ferrero GB, Cerone R, Romano C, et al (1990) Haplotype distribution and molecular defects at the phenylalanine hydroxylase locus in Italy. Hum Genet 86:69–72
DiLella AG, Kwok SCM, Ledley FD, Marvit J, Woo SLC (1986*a*) Molecular structure and polymorphic map of the human phenylalanine hydroxylase gene. Biochemistry 25: 743–749
DiLella AG, Marvit J, Brayton K, Woo SL (1987) An amino-acid substitution involved in phenylketonuria is in linkage disequilibrium with DNA haplotype 2. Nature 327:333–336
DiLella AG, Marvit J, Lidsky AS, Guttler F, Woo SL (1986*b*) Tight linkage between a splicing mutation and a specific DNA haplotype in phenylketonuria. Nature 322:799–803
Dworniczak B, Wedemeyer N, Eigel A, Horst J (1991*a*) PCR detection of the PvuII (Ea) RFLP at the human phenylalanine hydroxylase (PAH) locus. Nucleic Acids Res 19:1958
Dworniczak B, Wedemeyer N, Horst J (1991*b*) PCR detection of the BgIII RFLP at the human phenylalanine hydroxylase (PAH) locus. Nucleic Acids Res 19:1958
Eisensmith RC, Okano Y, Dasovich M, Wang T, Guttler F, Lou H, Guldberg P, et al (1992) Multiple origins for phenylketonuria in Europe. Am J Hum Genet 51:1355–1365
Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927
Feingold J, Guilloud-Bataille M, Feingold N, Rey F, Berthelon M, Lyonnet S (1993) Linkage disequilibrium in the human phenylalanine hydroxylase locus. Dev Brain Dysfunct 6: 26–31
Friedman PA, Fisher DB, Kang ES, Kaufman S (1973) Detection of hepatic phenylalanine 4-hydroxylase in classical phenylketonuria. Proc Natl Acad Sci USA 70:552–556
Goltsov AA, Eisensmith RC, Naughton ER, Jin L, Chakraborty R, Woo SLC (1993) A single polymorphic STR system

in the human phenylalanine hydroxylase gene permits rapid prenatal diagnosis and carrier screening for phenylketonuria. Hum Mol Genet 2:577–581

Goltsov AA, Eisensmith RC, Woo SL (1992) Detection of the XmnI RFLP at the human PAH locus by PCR. Nucleic Acids Res 20:927

Good P (1995) Permutation tests. Springer-Verlag, New York

Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409–411

Hertzberg M, Jahromi K, Ferguson V, Dahl HHM, Mercer J, Mickleson KNP, Trent RJ (1989) Phenylalanine hydroxylase gene haplotypes in Polynesians: evolutionary origins and absence of alleles associated with severe phenylketonuria. Am J Hum Genet 44:382–387

Hofman KJ, Steel G, Kazazian HH, Vallie D (1991) Phenylketonuria in US blacks: molecular analysis of the phenylalanine hydroxylase gene. Am J Hum Genet 48:791–798

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

Jaruzelska J, Henriksen K, Guttler F, Riess O, Borski K, Blin N, Slomski R (1991) The codon 408 mutation associated with haplotype 2 is predominant in Polish families with phenylketonuria. Hum Genet 86:247–250

Jorde LB, Watkins WS, Carlson M, Groden J, Albertsen H, Thuveris A, Leppert M (1994) Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. Am J Hum Genet 54:884–898

Kazazian HH Jr, Orkin SH, Markham AF, Chapman CR, Youssoufian H, Waber PG (1984) Quantification of the close association between DNA haplotypes and specific β-thalassemia mutations in Mediterraneans. Nature 310:152–154

Kidd JR, Black FL, Weiss KM, Balazs I, Kidd KK (1991) Studies of three Amerindian populations using nuclear DNA polymorphisms. Hum Biol 63:775–794

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

Kidd KK (1987) Phenylketonuria: population genetics of a disease. Nature 327:282–283

Kidd KK, Kidd JR (1996) A nuclear perspective on human evolution. In: Boyce AJ, Mascie-Taylor CGN (eds) Molecular biology and human diversity. Cambridge University Press, Cambridge, pp 242–264

Kidd KK, Morar B, Castiglione CM, Zhao H, Pakstis AJ, Speed WC, Bonne-Tamir B, et al (1998) A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. Hum Genet 103:211–227

Konecki DS, Lichter-Konecki U (1991) The phenylketonuria locus: current knowledge about alleles and mutations of the phenylalanine hydroxylase gene in various populations. Hum Genet 87:377–388

Kozák L, Kuhrová V, Blažková M, Romano V, Fajkusová L, Dvořáková D, Pijáčková A (1995) Phenylketonuria mutations and their relation to RFLP haplotypes at the PAH locus in Czech PKU families. Hum Genet 96:472–476

Kwok SC, Ledley FD, DiLella AG, Robson KJ, Woo SL (1985) Nucleotide sequence of a full-length complementary DNA clone and amino acid sequence of human phenylalanine hydroxylase. Biochemistry 24:556–561

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations: heterotic models. Genetics 49: 49–67

Li WH, Ellsworth DL, Krushkal J, Chang Bh, Hewett-Emmett D (1996) Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Mol Phylogenet Evol 5:182–187

Lidsky AS, Law ML, Morse HG, Kao FT, Rabin M, Ruddle FH, Woo SL (1985*a*) Regional mapping of the phenylalanine hydroxylase gene and the phenylketonuria locus in the human genome. Proc Natl Acad Sci USA 82:6221–6225

Lidsky AS, Ledley FD, DiLella AG, Kwok SCM, Daiger SP, Robson KJH, Woo SLC (1985*b*) Extensive restriction site polymorphism at the human phenylalanine hydroxylase locus and application in prenatal diagnosis of phenylketonuria. Am J Hum Genet 37:619–634

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799–810

Michalatos-Beloin S, Tishkoff SA, Kidd KK, Ruano G (1996) Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PR. Nucleic Acids Res 24: 4841–4843

Nowacki PM, Bick S, Prevost L, Scriver CR (1997) The PAH Mutation Analysis Consortium database update 1996. Nucleic Acids Res 25:139–142

Osier M, Pakstis AJ, Kidd JR, Lee J-F, Yin S-J, Ko H-J, Edenberg HR, et al (1999) Linkage disequilibrium at the ADH2 and ADH3 loci and risk of alcoholism. Am J Hum Genet 64:1147–1157

Piazza A (1975) Haplotypes and linkage disequilibria from three locus phenotypes. In: Kissmeyer-Nielsen F (ed) Histocompatibility testing 1975. Munksgaard, Cophenhagen, pp 923–927

Ruano G, Kidd KK (1991) Genotyping and haplotyping of polymorphisms directly from genomic DNA via coupled amplification and sequencing (CAS). Nucleic Acids Res 19: 6877–6882

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2d ed. Ford N, Nolan C, Ferguson M (eds) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Scriver CR, Byck S, Prevost L, Hoang L, PAH Mutation Analysis Consortium (1996) The phenylalanine hydroxylase locus: a marker for the history of phenylketonuria and human genetic diversity. In: Chadwick D, Cardew G (eds) Variation in the human genome. Ciba Foundation Symposium 197. John Wiley, Chichester, England, pp 73–96

Stuhrmann M, Riess O, Monch E Kurdoglu G (1989) Haplotype analysis of the phenylalanine hydroxylase gene in Turkish phenylketonuria families. Clin Genet 36:117–121

Svensson E, Von Dobeln U, Hagenfeldt L (1991) Polymorphic DNA haplotypes at the phenylalanine hydroxylase locus and

their relation to phenotype in Swedish phenylketonuria families. Hum Genet 87:11–17

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Cheung K, Kidd JR, Bonne-Tamir B, et al (1996*a*) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387

Tishkoff SA, Ruano G, Kidd JR, Kidd KK (1996*b*) Distribution and frequency of a polymorphic Alu insertion at the PLAT locus in humans. Hum Genet 97:759–774

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, et al (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

Wedemeyer N, Dworniczak B, Horst J (1991) PCR detection of the MspI (Aa) RFLP at the human phenylalanine hydroxylase (PAH) locus. Nucleic Acids Res 19:1959

Weir BS (1996) Genetic data analysis II. Sinauer Associates, Sunderland, MA

Woo SLC (1988) Collation of RFLP haplotypes at the human phenylalanine hydroxylase (PAH) locus. Am J Hum Genet 43:781–783

Woo SL, Lidsky AS, Guttler F, Chandra T, Robson KJ (1983) Cloned human phenylalanine hydroxylase gene allows prenatal diagnosis and carrier detection of classical phenylketonuria. Nature 306:151–155

Workman PL, Niswander JD (1970) Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. Am J Hum Genet 22:24–49

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

Zhao H, Pakstis AJ, Kidd KK, Kidd JR (1997) Overall and segmental significance levels of linkage disequilibrium. Am J Hum Genet Suppl 61:A17

Zygulska M, Eigel A, Aulehla-Scholz C, Pietrzyk JJ, Horst J (1991) Molecular analysis of PKU haplotypes in the population of southern Poland. Hum Genet 86:292–294