Copyright © 2002 by the Genetics Society of America

Nucleotide Variability at *G6pd* and the Signature of Malarial Selection in Humans

Matthew A. Saunders,[1] Michael F. Hammer and Michael W. Nachman

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721

Manuscript received January 14, 2002
Accepted for publication September 18, 2002

ABSTRACT

Glucose-6-phosphate dehydrogenase (G6PD) deficiency is the most common enzymopathy in humans. Deficiency alleles for this X-linked disorder are geographically correlated with historical patterns of malaria, and the most common deficiency allele in Africa (G6PD A−) has been shown to confer some resistance to malaria in both hemizygous males and heterozygous females. We studied DNA sequence variation in 5.1 kb of *G6pd* from 47 individuals representing a worldwide sample to examine the impact of selection on patterns of human nucleotide diversity and to infer the evolutionary history of the G6PD A− allele. We also sequenced 3.7 kb of a neighboring locus, *L1cam*, from the same set of individuals to study the effect of selection on patterns of linkage disequilibrium. Despite strong clinical evidence for malarial selection maintaining G6PD deficiency alleles in human populations, the overall level of nucleotide heterozygosity at *G6pd* is typical of other genes on the X chromosome. However, the signature of selection is evident in the absence of genetic variation among A− alleles from different parts of Africa and in the unusually high levels of linkage disequilibrium over a considerable distance of the X chromosome. In spite of a long-term association between *Plasmodium falciparum* and the ancestors of modern humans, patterns of nucleotide variability and linkage disequilibrium suggest that the A− allele arose in Africa only within the last 10,000 years and spread due to selection.

WITH the completion of the first drafts of the human genome (International Human Genome Sequencing Consortium 2001; Venter *et al.* 2001) considerable attention is now focused on understanding the levels and patterns of nucleotide variation among individuals. An accurate description of this variation is important for understanding processes of molecular evolution, for identifying disease genes, and for making inferences about the origin and history of *Homo sapiens*. A number of studies have described patterns of nucleotide variability in relatively large samples of individuals (Harding *et al.* 1997; Clark *et al.* 1998; Deinard and Kidd 1998; Harris and Hey 1999, 2001; Jaruzelska *et al.* 1999; Kaessmann *et al.* 1999; Rana *et al.* 1999; Rieder *et al.* 1999; Fullerton *et al.* 2000; Gilad *et al.* 2000; Hamblin and Di Rienzo 2000; Nachman and Crowell 2000; Zhao *et al.* 2000; Alonso and Armour 2001; Yu *et al.* 2001), and a large public effort recently identified and mapped >1 million single-nucleotide polymorphisms (SNPs; International SNP Map Working Group 2001). These studies have generally focused on regions of the genome in which positive natural selection is believed to be a negligible force, and as such, provide a baseline for average patterns of genomic variability. However, selection may have been an important force in shaping human genetic variation. Selection can have a powerful effect on patterns of linkage disequilibrium (LD), levels of heterozygosity, and frequencies of alleles segregating in a population, and these effects may extend to linked sites at considerable distances from the targets of selection (Hudson 1990, 1996). One way to study the impact of selection in shaping nucleotide variability is to look at regions of the genome in which the strength and form of selection are known and in which the connections from genotype to phenotype to environment are well understood.

The X-linked gene coding for glucose-6-phosphate dehydrogenase (G6PD) is subject to malarial selection in some human populations. The normal G6PD enzyme catalyzes a critical step in the pentose monophosphate shunt of glycolysis, and in cases of dysfunctional G6PD, an individual may suffer with clinical manifestations that include hemolytic anemia and neonatal jaundice (Beutler 1994). Some human populations exhibit G6PD deficiency alleles at frequencies that range up to 65% (Livingstone 1985; Oppenheim *et al.* 1993). In general, there is a geographic correlation between the frequency of G6PD deficiency alleles and the historical prevalence of malaria globally (Allison 1960; Motulsky 1961; Oppenheim *et al.* 1993). Moreover, *in vitro* studies (Roth *et al.* 1983; Roth and Schulman 1988) and epidemiological evidence (Ruwende *et al.* 1995) indicate that G6PD deficiency confers some resistance to *Plasmodium falciparum*, the primary human malaria parasite.

The most common G6PD deficiency allele in sub-

[1]*Corresponding author:* Biosciences West Bldg., Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721. E-mail: msaunder@u.arizona.edu

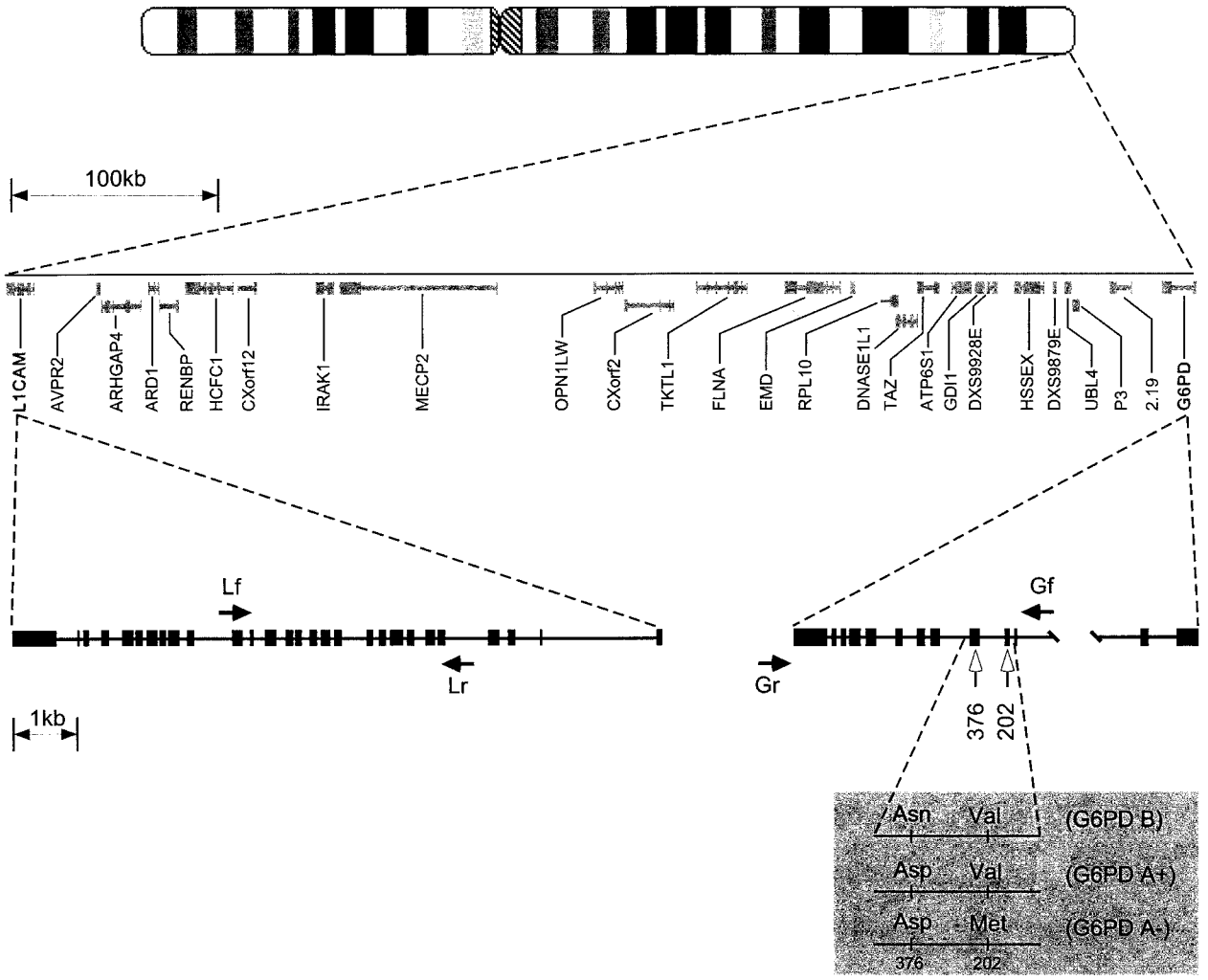

FIGURE 1.—Schematic ideogram of the human X chromosome and the genomic regions sampled in this study. Genes located between *L1cam* and *G6pd* on Xq28 are marked. Solid arrows indicate positions of the amplification primers *Gf*, *Gr*, *Lf*, and *Lr*. For *G6pd*, the mutations that define G6PD A+ and G6PD A− are marked at coding positions 376 and 202. *L1cam* is located 556 kb from *G6pd*. Exons are marked with solid boxes. Polymorphic amino acid residues for alleles G6PD B, G6PD A+, and G6PD A− due to nucleotide polymorphisms at sites 202 and 376 are shown in the shaded box.

Saharan Africa is G6PD A−, and it typically reaches frequencies near 20% in populations living in malarial areas (LIVINGSTONE 1985). The A− allele differs from the normal allele (G6PD B) by nonsynonymous changes at coding nucleotide positions 202 and 376. A minor deficiency allele, G6PD A+, differs from the B allele only at site 376 (Figure 1). The enzymatic activities of the A+ and A− alleles are 85 and 12% of normal levels, respectively (HIRONO and BEUTLER 1988; BEUTLER *et al.* 1989; VULLIAMY *et al.* 1991). The mild deficiency phenotype characteristic of G6PD A+ does not cause significant clinical manifestations and does not appear to confer resistance to malaria (RUWENDE *et al.* 1995). However, the deficiency phenotype characteristic of G6PD A− confers an ~50% reduction in risk of severe malaria in both heterozygote females and hemizygote males. Homozygous females probably have a similar level of protection from malaria, although this genotype is quite rare (RUWENDE *et al.* 1995). In the presence of *falciparum* malaria, the G6PD A− allele is therefore beneficial, while in the absence of malaria this allele is deleterious. Thus *G6pd* provides a rare example of a gene in humans where the selective agent and approximate form and strength of selection are known (RUWENDE *et al.* 1995; TISHKOFF *et al.* 2001).

As part of an ongoing project to characterize patterns of nucleotide variability at multiple loci throughout the genome for a common worldwide sample of human DNAs and to investigate the impact of selection on *G6pd*, we sequenced 5.1 kb of *G6pd* in a sample of 47 humans (Table 1). We also sequenced 3.7 kb at *L1cam* in these same individuals. *L1cam* is situated 556 kb from *G6pd*; thus, polymorphisms at *L1cam* provide an opportunity to investigate the impact of selection on neighboring

sites. Our nucleotide data suggest that the effects of selection on *G6pd* are more subtle than those predicted under a model of long-term diversifying selection.

MATERIALS AND METHODS

Samples: DNA sequences were determined in a sample of 41 human males, including 10 from Africa, 10 from the Americas, 10 from Europe, and 11 from Asia and Melanesia (Table 1). This sample was chosen as part of a long-term project in our labs to survey nucleotide variability at a number of loci throughout the genome using a common set of individuals (*e.g.*, Nachman *et al.* 1998; Nachman and Crowell 2000; M. W. Nachman and M. F. Hammer, unpublished data). However, since G6PD A− alleles are primarily found in Africa and since the effects of selection at *G6pd* are likely to be found primarily in Africa, we augmented our worldwide sample with 4 additional African individuals that were known [by restriction fragment length polymorphism (RFLP) analysis] to carry G6PD A− alleles and 2 individuals known to carry G6PD A+ alleles. This allowed us to investigate patterns of variability within G6PD A− alleles and to study LD between G6PD A− alleles and other alleles. Homologous sequences from a male chimpanzee (*Pan troglodytes*) and a male orangutan (*Pongo pygmaeus*) were also determined for divergence estimates. By studying X-linked loci in males we were able to PCR amplify single alleles and to directly recover haplotypes over long genomic distances to study patterns of linkage disequilibrium.

PCR amplification and sequencing: Maps of the human X chromosome and the loci sampled in this study, *G6pd* and *L1cam*, are presented in Figure 1. *L1cam* was chosen because of its proximity to *G6pd* (556 kb); all polymorphisms detected at *L1cam* are silent or noncoding, and there is no *a priori* reason to assume that *L1cam* itself is a target of selection. Approximately 82 other genes are found within 1 Mb on either side of *G6pd* and none of these genes are known to be recent targets of positive selection. PCR fragments were amplified for *G6pd* (5.2 kb) and *L1cam* (4.2 kb) using a long-template PCR system (Roche Biochemicals). For *G6pd*, the primers Gf (5′ GTT TAT GTC TTC TGG GTC AGG GAT GG 3′) and Gr (5′ AGT GTT GCT GGA AGT CAT CTT GGG T 3′) are positioned with the 5′ end of the primer at sites 206322 and 201052, respectively, in GenBank accession no. L44140. For *L1cam*, the primers Lf (5′ TCC TCT CCA GAG TAG CCG ATA GTG ACC 3′) and Lr (5′ AAG TTT CTA CTG GCC TGA CCC TCT CG 3′) are positioned with the 5′ end of the primer at sites 19587 and 24251, respectively, in GenBank accession no. U52112 (Figure 1). Internal primers (available upon request) were used to generate overlapping sequence runs on an ABI 377 automated sequencer. A contiguous sequence that included coding and noncoding regions (5109 and 3691 bp for *G6pd* and *L1cam*, respectively) was assembled for each individual and aligned using the computer program Sequencher (Gene Codes, Ann Arbor, MI). Sequences have been submitted to GenBank under accession nos. AY158094–AY158142 and AY167680–AY167728 for *G6pd* and *L1cam*, respectively.

Data analysis: Nucleotide diversity, π (Nei and Li 1979), and the proportion of segregating sites, Θ (Watterson 1975), were calculated using the program PROSEQ (Filatov *et al.* 2000) for the worldwide sample, for African individuals, and for non-African individuals. Only the 41 individuals of the nonaugmented worldwide sample were included in analyses of nucleotide diversity, and insertion-deletion polymorphisms were excluded. Under neutral equilibrium conditions both π and Θ estimate the neutral parameter $3N_e\mu$ for X-linked loci, where N_e is the effective population size and μ is the neutral mutation rate. Tajima's D (Tajima 1989), Fu and Li's D (Fu and Li 1993), and Fay and Wu's H (Fay and Wu 2000; http://crimp.lbl.gov/htest.html) were calculated to test for deviations from a neutral equilibrium frequency distribution for both loci. Ratios of polymorphism to divergence for *G6pd* and *L1cam* were compared with the expectations under a neutral model using the Hudson-Kreitman-Aguadé (HKA) test (Hudson *et al.* 1987). Polymorphism data for these tests were derived from the 41 sequences determined in this study for *G6pd* and *L1cam*, as well as from *Dmd* (intron 44) from the same set of individuals (Nachman and Crowell 2000) and from the *Pdha1* data of Harris and Hey (1999). *Dmd* and *Pdha1* were chosen for comparison because they both reside in regions of the X chromosome with moderate to high rates of recombination and thus are expected to be relatively free of the effects of selection at linked sites. Divergence data were derived for each of these loci by comparing the homologous sequences from a chimpanzee to a single randomly chosen human allele. LD between pairs of polymorphic sites was measured using the statistics D' (Lewontin 1964) and r^2 (Hill and Robertson 1968). The age of the G6PD A− allele was estimated from the decay of linkage disequilibrium and from coalescent simulations using the computer program *GENETREE* (Harding *et al.* 1997; Bahlo and Griffiths 2000). The *SWST* haplotype test of Andolfatto *et al.* (1999) was implemented using the data from *G6pd* and *L1cam* separately. This test compares the observed number of haplotypes with those expected under a neutral model with a specified rate of recombination.

RESULTS

Nucleotide diversity: Patterns of nucleotide variability at *G6pd* and *L1cam* are presented in Tables 1 and 2. In the worldwide sample of 41 chromosomes (nonaugmented sample) we observed 18 single-nucleotide polymorphisms and three insertion/deletion (indel) polymorphisms at *G6pd*. Fifteen of these polymorphisms were in introns; of the remaining 6 polymorphisms, 2 were nonsynonymous changes (coding sites 202 and 376) and 4 were synonymous changes. Levels of nucleotide variability were roughly four times higher in Africa than in non-African populations (Table 2), consistent with other studies that demonstrate higher diversity in Africa (*e.g.*, Harris and Hey 1999, 2001; Nachman and Crowell 2000). Many of the polymorphisms found in Africa distinguish G6PD A− alleles from all other alleles in the sample. At *L1cam* we observed 7 polymorphisms in the nonaugmented sample. Levels of nucleotide variability were relatively low for *L1cam* overall; however, nucleotide variability was higher in Africa than in non-African populations.

In the worldwide sample of 41 chromosomes, two A− alleles were in the African subset ($n = 10$), consistent with previously documented frequencies of ~20% for G6PD A− in sub-Saharan Africa. Overall, worldwide levels of nucleotide variability at *G6pd* and *L1cam* were close to or slightly below average values for other regions of the genome. For example, among primarily noncoding sites at 12 X-linked genes in humans, the average

TABLE 1

Polymorphisms for a worldwide sample of humans at *G6pd* and *L1cam*

				G6pd																			*L1cam*								
				Alignment position																												
						1	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4	4	4	5			1	1	1	2	2	2	3
					3	0	6	7	8	0	1	4	8	9	9	0	6	9	0	1	4	6	9	0	7	8	0	7	8	1	2	9	3
				2	4	6	4	8	1	0	6	8	9	2	9	4	0	0	4	2	1	9	6	5	7	8	8	3	9	1	8	4	7
				9	0	5	3	6	2	2	5	2	3	9	3	2	4	3	9	8	0	9	1	0	6	5	6	1	4	5	3	1	5
				Coding site no.																												
																	1	*1*			*1*											
					2	*3*				*6*							*1*	*3*			*4*											
					0	*7*				*4*							*1*	*1*			*3*											
					2	*6*				*2*							*6*	*1*			*1*											
Country	Ethnic/language group	G6PD allele type	Sample identity	Consensus																												
				A	G	A	C	C	T	G	C	C	C	C	C	G	G	C	T	C	G	A	C	C	G	C	C	C	G	C	A	C	C
South Africa	Pedi	B	33	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	C	T	·	G	·	·	·	·	·	·	·	·	·	·	·
South Africa	Herero	B	40	·	·	·	T	·	·	·	·	·	·	·	·	·	·	·	C	T	·	G	·	·	·	·	·	T	·	·	·	·	T
Zaire	Mbuti	B	65	·	·	·	·	·	·	·	·	·	·	·	·	A	·	·	C	T	·	G	·	·	·	·	·	·	·	·	·	·	·
Zaire	Mbuti	B	8	·	·	·	·	·	·	·	·	·	·	·	·	A	·	·	C	T	·	G	·	·	·	·	·	·	·	·	·	·	·
Namibia	Tsumkwe	B	38	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Namibia	Tsumkwe	B	22	·	·	·	·	·	·	·	·	·	·	·	T	·	A	·	C	·	—	G	·	·	·	·	·	·	T	·	·	·	·
C. African Republic	Biaka	B	6	·	·	·	·	·	·	·	·	·	·	·	·	·	·	T	C	·	·	G	·	·	·	·	·	·	·	·	·	·	·
C. African Republic	Biaka	B	7	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	C	·	·	G	·	—	·	·	·	·	·	·	·	·	·
Gambia	Unknown	*A+*	*G25*	G	·	G	·	·	·	·	·	·	·	T	·	·	·	·	C	·	·	G	·	·	·	·	·	·	·	·	·	·	T
Gambia	Serere	*A+*	*G45*	G	·	G	·	·	·	·	·	·	·	T	·	·	·	·	C	·	·	G	·	·	·	·	·	·	·	·	T	·
South Africa	Sotho	A−	32	G	A	G	·	G	·	·	·	T	·	T	·	·	·	·	C	·	·	G	·	·	C	T	·	·	·	T	·	·	·
Zaire	Mbuti	A−	9	G	A	G	·	G	·	·	·	T	·	T	·	·	·	·	C	·	·	G	·	·	C	T	·	·	·	T	·	·	·
Nigeria	Unknown	*A−*	*S823*	G	A	G	·	G	·	·	·	T	·	T	·	·	·	·	C	·	·	G	·	·	C	T	·	·	·	T	·	·	·
United States	African-American	*A−*	*M241*	G	A	G	·	G	·	·	·	T	·	T	·	·	·	·	C	·	·	G	·	·	C	T	·	·	·	T	·	·	·
United States	African-American	*A−*	*M115*	G	A	G	·	G	·	·	·	T	·	T	·	·	·	·	C	·	·	G	·	·	·	·	·	·	·	·	·	·	·
Gambia	Unknown	*A−*	*G11*	G	A	G	·	G	·	·	·	T	·	T	·	·	·	·	C	·	·	G	·	·	·	·	T	·	·	·	·	·	·
Brazil	Karitiana	B	12	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Brazil	Karitiana	B	13	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Brazil	Surui	B	14	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Brazil	Surui	B	16	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
USA	Navajo	B	23	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
USA	Tohono O'Odham	B	25	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
USA	Amerindian	B	4	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
USA	Amerindian	B	2	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·

(*continued*)

TABLE 1

(Continued)

Country	Ethnic/language group	G6PD allele type	Sample identity	Consensus																													
				A	G	A	C	C	T	G	C	C	C	C	C	G	G	C	T	C	G	A	C	C	G	C	C	C	G	C	A	C	C
USA	Porch Creek	B	27	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Mexico	Mayan	B	17	·	·	·	·	·	·	A	T	·	·	·	·	·	·	·	C	·	·	G	·	·	·	·	·	·	·	·	·	·	·
Poland	Ashkenazi	B	59	·	·	·	·	·	·	·	·	·	·	·	·	·	·	T	C	·	·	G	·	·	·	·	·	·	·	·	·	·	·
E. Europe	Ashkenazi	B	24	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
UK	British	B	26	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Germany	German	B	61	·	·	·	·	·	·	·	·	·	·	·	·	·	·	T	C	·	·	G	·	·	·	·	·	·	·	·	·	·	·
Germany	German	B	62	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Germany	German	B	64	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Turkey	Turkish	B	79	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Russia	Russian	B	72	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Russia	Russian	B	71	·	·	·	·	·	·	·	·	·	·	·	·	·	·	T	C	·	·	G	·	·	·	·	·	·	·	·	·	·	·
Russia	Adygeans	B	56	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Japan	Japanese	B	78	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	C	·	·	G	·	·	·	·	·	·	·	·	·	·	·
Cambodia	Cambodian	B	69	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Pakistan	Pakistani	B	57	·	·	·	·	·	·	·	·	·	·	·	·	·	·	T	C	·	·	G	·	·	·	·	·	·	·	·	·	·	·
Melanesia	Nasioi	B	10	·	·	·	·	·	·	·	·	·	T	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Siberia	Yakut	B	49	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Siberia	Yakut	B	51	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	C	·	·
China	S. Han	B	68	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
China	S. Han	B	66	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	T	·	·	·	·	·	·	·	·	·	·
China	S. Han	B	67	·	·	·	·	·	—	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Japan	Japanese	B	77	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Japan	Japanese	B	76	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
			Pan	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	C	·	·	G	·	·	·	·	·	·	·	·	·	·	·
			Pongo	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	C	·	·	G	·	·	·	·	·	·	·	·	·		

Samples from 41 human males representing Africa, Asia, Europe, and the Americas were obtained from the Y-Chromosome Consortium DNA collection. Additional samples, marked in italics, were selected on the basis of an *a priori* allele-type of determination of *G6pd* A− and A+ using coding sites 202 and 376. Polymorphisms at *G6pd* alignment positions 2002 (site 642), 3604 (site 1116), 3903 (site 1311), and 4128 (site 1431), and at *L1cam* alignment position 885 represent synonymous changes in coding exons. *G6pd* positions 4410, 4699, 4961, and 5050 are in the noncoding region of exon 13. All other polymorphisms are in introns (except *G6pd* coding site 202 and 376; see Figure 1). —, insertion/deletion (indel) polymorphism. The indels at coding positions 4410 and 5050 spanned three consecutive nucleotides. For outgroup taxa (Pan and Pongo) only sites that are polymorphic in the human sample are shown.

TABLE 2

Summary statistics of nucleotide variability for *G6pd* and *L1cam*

Geographic region	Locus	Length (bp)	Sample size	S	π (SD) (%)	θ (SD) (%)	Tajima's *D*	Fu and Li's *D*	Divergence (SD) Homo-Pan (%)	Divergence (SD) Homo-Pongo (%)
Worldwide	*G6pd* total sequence	5102	41	18	0.05 (0.040)	0.08 (0.030)	−1.429	−1.134	1.0 (0.1)	3.2 (0.3)
	G6pd introns	2918	41	10	0.04 (0.039)	0.08 (0.033)	−1.512	−1.018	1.2 (0.2)	4.0 (0.4)
	L1cam total sequence	3691	41	7	0.01 (0.021)	0.04 (0.020)	−1.946*	−1.822*	0.8 (0.1)	2.9[a] (0.4)
	L1cam introns	2087	41	6	0.02 (0.032)	0.07 (0.033)	−1.925*	−2.203*	1.1 (0.2)	3.9[a] (0.6)
African sample	*G6pd* total sequence	5103	10	14	0.08 (0.051)	0.10 (0.046)	−0.672	−0.342		
	G6pd introns	2919	10	8	0.08 (0.051)	0.10 (0.050)	−0.687	−0.087		
	L1cam total sequence	3691	10	6	0.05 (0.030)	0.06 (0.032)	−0.886	−0.553		
	L1cam introns	2087	10	5	0.06 (0.045)	0.08 (0.049)	−1.035	−0.884		
Non-African sample	*G6pd* total sequence	5108	31	7	0.02 (0.017)	0.03 (0.016)	−1.032	−1.644		
	G6pd introns	2918	31	3	0.02 (0.013)	0.03 (0.016)	−0.929	−1.532		
	L1cam total sequence	3691	31	1	0.00 (0.003)	0.01 (0.007)	−1.145	−1.681		
	L1cam introns	2087	31	1	0.00 (0.003)	0.01 (0.007)	−1.145	−1.681		

Divergence estimates were based on a comparison between a single randomly chosen human allele and the chimpanzee or orangutan alleles. $^{*}P < 0.05$.

[a] *Homo-Pongo* divergence estimates for *L1cam* are based on 1672 and 1046 bp for total sequence and introns, respectively.

level of nucleotide diversity (π) is 0.06% and the average proportion of segregating sites (Watterson's Θ) is 0.07% (Nachman 2001). For both *G6pd* and *L1cam*, nucleotide diversity at intron sites is slightly below average (*G6pd* $\pi = 0.04\%$, *L1cam* $\pi = 0.02\%$), while Watterson's Θ is close to average (*G6pd* $\Theta = 0.08\%$, *L1cam* $\Theta = 0.07\%$). Since the A− allele represents only 5% of the worldwide sample, it is not expected to contribute substantially to levels of nucleotide variability. Within Africa, however, G6PD A− is present at high frequency (20%), yet overall levels of nucleotide variability ($\pi = 0.08\%$, Table 2) are still average. For example, the average level of nucleotide variability for 8 X-linked genes in Africa is 0.084% (Payseur and Nachman 2002).

Tests of neutrality: Tajima's *D* is the normalized difference between π and Θ and takes on positive values when there is an excess of intermediate-frequency polymorphisms and takes on negative values when there is an excess of low-frequency polymorphisms (Tajima 1989). Positive Tajima's *D* values are generally consistent with long-term balancing selection or a population contraction, while negative values are expected following a selective sweep or a population expansion. For *G6pd*, Tajima's *D* is negative (but not significant) in the worldwide sample and for all subsets of the data (Table 2). Similar results are obtained with Fu and Li's *D* (Fu and Li 1993), which also measures the frequency distribution of polymorphisms and is sensitive to the number of singletons in the sample (Table 2). For *L1cam* both statistics are also negative and are significantly negative in the worldwide sample (Table 2). Fay and Wu's *H* statistic (Fay and Wu 2000) utilizes the frequency distribution of polymorphisms to test for an excess of high-frequency-derived variants compared to equilibrium neutral expectations. For both *G6pd* and *L1cam*, Fay and Wu's *H* test shows no significant deviation from the neutral expectation in the worldwide sample, the African sample, or the non-African sample.

We performed an HKA test (Hudson *et al.* 1987) using pairwise comparisons of polymorphism and divergence for *G6pd* and *L1cam* and two other X-linked genes, *Pdha1* and *Dmd*. In comparisons using worldwide samples or African samples alone, we failed to reject the null model (HKA $\chi^2 < 3.0$, $P > 0.1$ for all tests). Thus, neither the frequency spectrum nor the level of heterozygosity at *G6pd* fits the expected pattern of nucleotide variability under a simple model of long-standing diversifying or balancing selection.

To test whether the haplotype structure of the data deviates from neutral expectations we implemented the *SWST* program as described in Andolfatto *et al.* (1999), assuming recombination rates of 0, 1, and 2 cM/Mb. Tests were performed separately for *G6pd* and for *L1cam*. None of these tests showed significant deviations from neutral expectations using the worldwide sample or the African sample alone.

Linkage disequilibrium: To better examine patterns

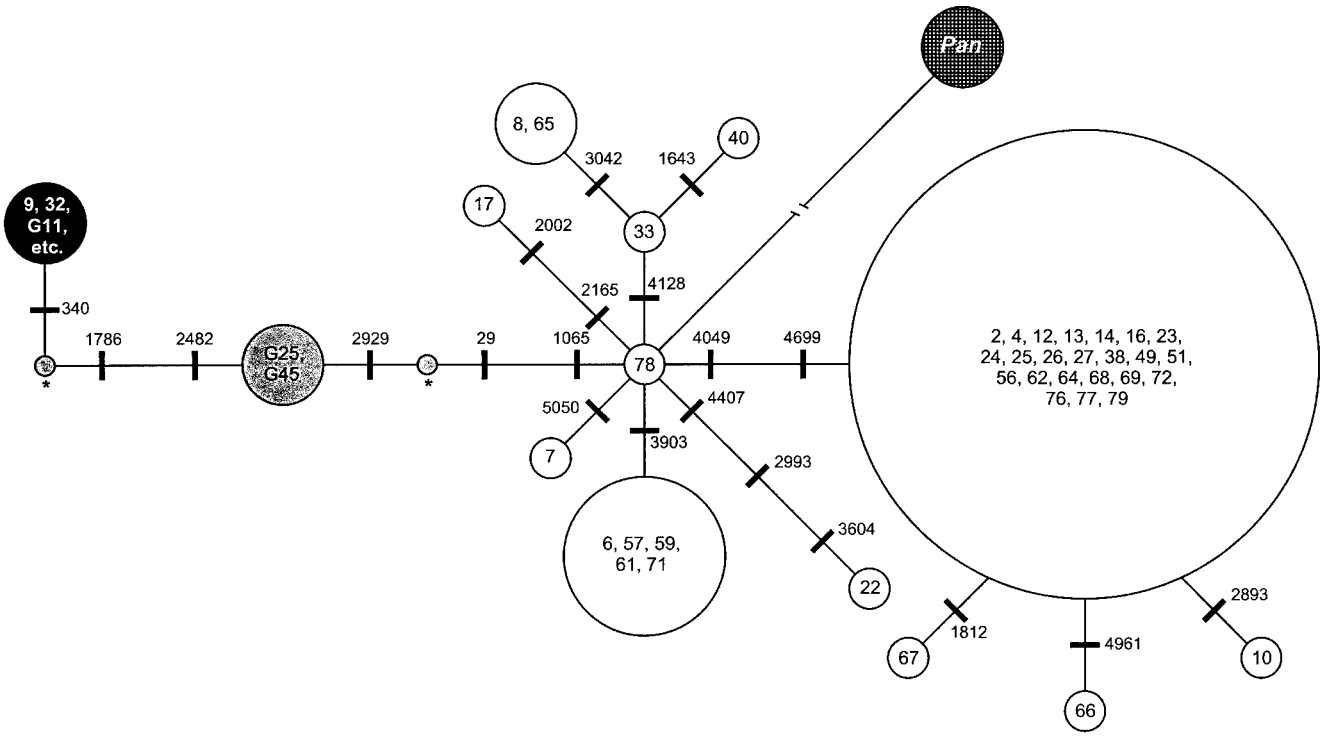

FIGURE 2.—Haplotype network for polymorphisms of the worldwide sample at *G6pd*. Text inside circles represents sample identities (Table 1). Marks indicate polymorphisms and are labeled with the respective alignment position from Table 1. * indicates known haplotypes for G6PD A+ alleles that were not captured in this sample (VULLIAMY *et al.* 1991). The solid circle represents G6PD A−, shaded circles represent G6PD A+, and open circles represent G6PD B alleles.

of linkage disequilibrium we augmented our random sample of 10 African X chromosomes with 4 chromosomes carrying A− alleles and 2 chromosomes carrying A+ alleles. Thus the augmented African sample in the study includes 6 chromosomes carrying *G6pd* A− alleles from South Africa, Central Africa, and West Africa (samples YCC 9, YCC 32, G11, M115, M241, and S823 in Table 1). Unusually high levels of linkage disequilibrium were observed within *G6pd*, within *L1cam*, and between *G6pd* and *L1cam*. D' is a measure of linkage disequilibrium that is standardized to equal 0 when there is random association among polymorphisms (*i.e.*, no disequilibrium) and to equal 1 when there is complete association among polymorphisms (*i.e.*, complete disequilibrium). In all comparisons between A− alleles and other alleles, $D' = 1$ for all sites in Table 1. A single most parsimonious haplotype network was inferred for all sites at *G6pd* (Figure 2), indicating a lack of evidence for recombination in this sample despite the fact that Xq28 is a genomic region with moderate to high rates of recombination (PAYSEUR and NACHMAN 2000). Surprisingly, the three polymorphic sites at *L1cam* at intermediate frequency (positions 776, 885, and 2115) can also be mapped on this same network with no homoplasy. The observed high level of linkage disequilibrium is primarily a consequence of mutations falling on the branch separating the A− deficiency allele from the normal B alleles (Figure 2). A Fisher's exact test revealed significant LD ($P = 0.0082$) between site 202 of *G6pd* and three out of four informative polymorphisms at *L1cam* (alignment positions 776, 885, and 2115 at *L1cam*; Table 3).

Age of the G6PD A− allele: We estimated the age of the A− allele in two ways. First, we used a standard model for the decay of linkage disequilibrium as a function of time (t) and recombination (c), where linkage disequilibrium at time t (r_t^2) compared with time 0 (r_0^2) is given by $r_t^2/r_0^2 = (1 - c)^t$ (HEDRICK 1998). For this calculation, we use r^2 as a measure of linkage disequilibrium between *L1cam* and *G6pd* because, unlike D', r^2 is sensitive to allele frequencies when only three out of four gametic types are present in a sample

TABLE 3

Nonrandom associations between *G6pd* site 202 and polymorphisms at *L1cam*

	L1cam polymorphism	
G6pd polymorphism	G, C, C: Positions 776, 885, 2115	C, T, T: Positions 776, 885, 2115
G: site 202 (G6PD B or A+)	10	0
A: site 202 (G6PD A−)	2	4

Results of Fisher's exact test ($P = 0.0082$) for African augmented sample ($n = 16$) are presented between coding site 202 at *G6pd* and informative polymorphisms at *L1cam* alignment positions 776, 885, and 2115 (Table 1).

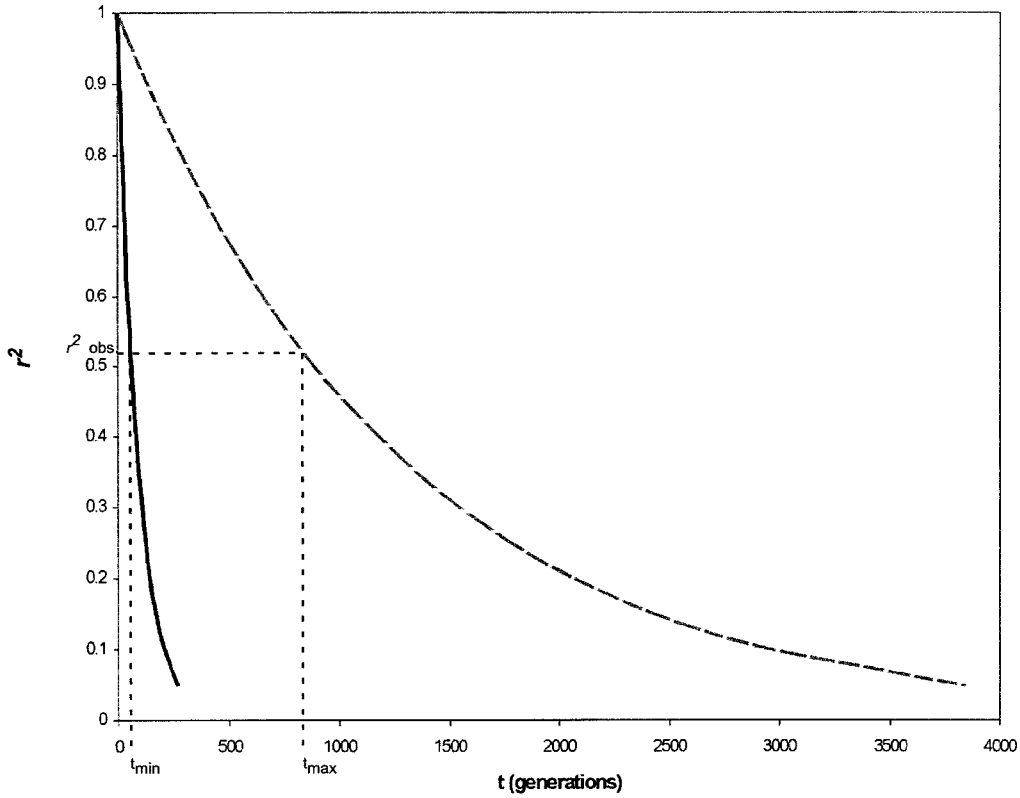

FIGURE 3.—Plot of linkage decay between *G6pd* (site 202) and *L1cam* (positions 776, 885, and 2115). The expected plot of linkage decay as measured by r^2 is shown over time (t) for a range of two different recombination rates that have been suggested for the chromosomal region near *G6pd* and *L1cam* (SMALL *et al.* 1997), 0.14 cM/Mb (long-dashed line) and 2 cM/Mb (solid line). The observed $r^2 = 0.52$ provides minimum ($t_{min} = 58$ generations) and maximum ($t_{max} = 840$ generations) ages for the *G6pd* A− allele.

(HEDRICK 1998). Assuming that linkage disequilibrium between site 202 of *G6pd* A− and positions 776, 885, and 2115 of *L1cam* alleles was complete at the time of origin of the A− allele (*i.e.*, $r_0^2 = 1$), we can estimate the time in generations, $t = \ln(r_t^2)/\ln(1 - c)$, for the age of the allele, given the observed recombinational distance between *L1cam* and *G6pd* and the observed linkage disequilibrium in the data ($r^2 = 0.52$). Since it is possible that r_0^2 was <1.0 between these sites when the G6PD A− allele arose, our estimates provide an upper bound for the age of the G6PD A− allele. This region of Xq28 is subject to moderate levels of recombination in general (1–3 cM/Mb; PAYSEUR and NACHMAN 2000; KONG *et al.* 2002), and recombination rates near *G6pd* and *L1cam* specifically have been estimated as low as 0.14 cM/Mb and as high as 2 cM/Mb (SMALL *et al.* 1997). Using these recombination rates we estimated the maximum age of the A− allele to be between 58 and 840 generations (Figure 3). With a generation time of 25 years, this implies that the *G6pd* A− allele arose 1461–20,994 years ago.

A second estimate for the age of the A− allele was obtained from simulations using a coalescent model conditioned on the sample size and observed levels of nucleotide variability (*GENETREE*: HARDING *et al.* 1997; BAHLO and GRIFFITHS 2000). This model assumes neutral equilibrium conditions and thus may provide an overestimate of the true age of the A− allele (since the present frequency of A− has probably been determined in large measure by selection). These simulations suggest that the A− allele arose 10,575 years ago (SD ± 8887 years).

Both of these estimates are in good agreement with an independent estimate for the age of the G6PD A− allele (3840–11,760 years) that was reported by TISHKOFF *et al.* (2001) on the basis of intra-allelic levels of linked microsatellite variability.

DISCUSSION

Models of selection and nucleotide variability at *G6pd*: We investigated levels and patterns of nucleotide variability at *G6pd*, a locus known to be under malarial selection in some human populations, and found that nucleotide diversity was similar to average values for other X-linked genes. Moreover, several commonly employed statistical tests based on DNA sequence variation failed to reject a simple neutral model of molecular evolution. In several respects, however, the data from *G6pd* are quite striking: levels of linkage disequilibrium are high and extend over a long genomic distance, much of the nucleotide variation is partitioned between functionally distinct alleles, and no nucleotide variation is observed within deficiency alleles. Below we discuss general models of selection for *G6pd* and how our observations might fit these models.

TABLE 4

Fitness arrays under different malarial-selection regimes

Fitness arrays	Genotype:	Females (A−)(A−) w_{11}	Females (A−)B w_{12}	Females BB w_{22}	Males (A−) w_1	Males B w_2
1. No malaria selection		$1 - s_f$	$1 - hs_f$	1	$1 - s_m$	1
2. Malaria selection: heterosis (overdominance)[a]		$1 - s_{f1}$	1	$1 - s_{f2}$	$1 - s_{m1}$	$1 - s_{m2}$
3. Malaria selection: directional (dominance)		1	1	$1 - s_f$	1	$1 - s_m$

[a] A stable polymorphism can be maintained only under the restrictive conditions $(1 - s_{f1})(1 - s_{m1}) < 1 - s_{m1}/2 - s_{m2}/2 > (1 - s_{f2})(1 - s_{m2})$; *i.e.*, if there is heterozygote advantage in females and not very strong selection in males, or if there is selection of similar magnitude in opposite directions in each sex.

Although four different species of Plasmodium typically infect humans, *P. falciparum* is the most virulent species and is responsible for most malaria-related deaths, especially in Africa (Schmidt and Roberts 1996). Malaria is endemic throughout most of sub-Saharan Africa where >1 million people die each year due to complications from infection (Trigg and Konrachine 1998). From a population genetics perspective, such a virulent parasite serves as a strong selective agent for genetic resistance. In fact, it has long been known that African populations exhibit genetic resistance factors to malaria at relatively high frequencies compared with non-African populations (*e.g.*, Miller 1994). Moreover, many of the mutations that confer resistance are deleterious outside of the malaria environment. G6PD A−, for example, has an enzymatic activity that is only about one-tenth of normal and results in significant clinical manifestations such as hemolytic anemia and neonatal jaundice (Hirono and Beutler 1988; Beutler *et al.* 1989; Vulliamy *et al.* 1991). However, this allele also confers an ~50% reduction in risk of severe malaria in both heterozygote females and hemizygote males (Ruwende *et al.* 1995).

Although G6PD is often assumed to be subject to balancing selection (*sensu* heterosis; *e.g.*, Tishkoff *et al.* 2001), the precise nature of selection on G6PD deficiency alleles is not fully understood. In the absence of malaria, deficiency alleles are at a selective disadvantage and are expected to be eliminated (Table 4, fitness array 1). In the presence of malaria, female heterozygotes and male deficiency hemizygotes appear to have a selective advantage over wild-type individuals, but the fitness of female deficiency homozygotes relative to other genotypes is not clear (Ruwende *et al.* 1995). If female heterozygotes have a higher fitness than either homozygote, then selection may maintain both A− and wild-type alleles in populations under malarial selection (*i.e.*, heterosis; Table 4, fitness array 2). However, the conditions for maintenance of a stable X-linked polymorphism are rather restrictive; either selection must be of similar magnitude but opposite in direction for the two sexes (which seems unlikely for G6PD deficiency) or there must be heterosis in females without a large fitness difference between the two male genotypes (Hedrick 1998). Alternatively, if female deficiency homozygotes have the same fitness as male hemizygotes and female heterozygotes, then selection should drive the eventual fixation of the G6PD A− allele in populations subject to continuous malarial selection (Table 4, fitness array 3). In such a situation, the A− allele is expected to rise to high frequencies and to reach fixation in a very short period of time (*e.g.*, <10,000 years; Ruwende *et al.* 1995). The exact time required for fixation depends on assumptions about population size, initial frequency of the A− allele, relative fitness of the different genotypes, and the average generation time. However, for a wide range of parameter values, allele frequencies are expected to quickly rise to very high levels. The observation that most African populations have A− allele frequencies <20% (Livingstone 1985) is inconsistent with a simple model of directional selection in which selection has been strong and long acting.

Thus the best explanation for current G6PD A− allele frequencies seems to be either heterosis (fitness array 2) or some form of spatially and/or temporally varying selection due to malaria, in which case allele frequencies may be determined primarily by changing selection pressures (*i.e.*, a combination over time or space of fitness array 1 and fitness array 2 and/or 3 in Table 4). On a large geographic scale (*e.g.*, among continents), spatially varying selection is clearly important in determining allele frequencies; the extent to which this applies to small geographic scales is less clear, although the frequency of the A− allele differs significantly among different populations in sub-Saharan Africa (Cavalli-Sforza *et al.* 1996; Tishkoff *et al.* 2001). While we cannot distinguish between heterosis and spatially/temporally varying selection, our data do allow us to address the timescale over which selection has acted.

A simple model of long-term balancing selection or

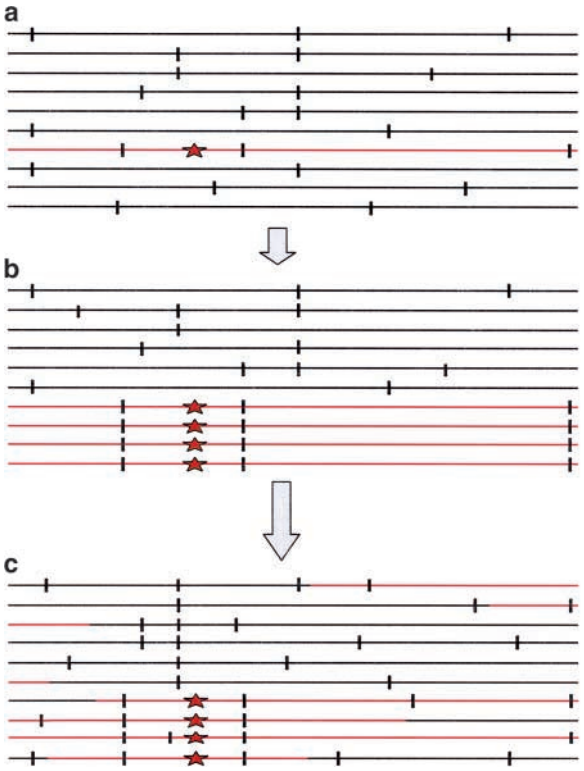

FIGURE 4.—Temporal schematic model for patterns of nucleotide variability at a locus under diversifying selection. Each group of 10 horizontal lines represents alleles sampled from a population at a given point in time. Vertical marks represent neutral polymorphisms. A red star represents the advantageous mutation under selection, and the ancestral allele bearing this mutation is marked by a red line. (a) A new advantageous mutation arises. (b) The new mutation quickly rises in frequency due to selection and is in linkage disequilibrium with neutral mutations over long distances. Heterozygosity is reduced relative to the population at the time (a). (c) Over time, for a locus under diversifying selection heterozygosity is elevated near the selected site, and linkage disequilibrium decays as a function of genetic distance.

long-term spatially or temporally varying selection is expected to leave a distinct signature in patterns of DNA sequence variation (Figure 4). When a new advantageous mutation first appears (Figure 4a), it will rise in frequency, creating LD with other mutations on the haplotype on which it arose (Figure 4b). This transient phase will result in lowered levels of heterozygosity. Over time, linkage disequilibrium will decay through recombination around the target of selection, and heterozygosity will increase near the target of selection (Figure 4c). This simple model of long-term selection predicts elevated levels of nucleotide variability in a restricted window around the target of selection (HUDSON *et al.* 1987) and a skew in the frequency distribution of polymorphisms with an excess of intermediate-frequency variants within this restricted window (TAJIMA 1989). Both of these patterns are seen in several other well-studied systems. For example, at *Mhc* loci in a variety of organisms [the human leukocyte antigen (HLA) loci in humans], levels of heterozygosity are significantly higher than those in neighboring regions (TAKAHATA *et al.* 1992). At *Adh* in *Drosophila melanogaster*, heterozygosity is elevated around the fast/slow allozyme polymorphism, resulting in a significant HKA test (HUDSON *et al.* 1987).

In contrast, patterns of nucleotide variability at *G6pd* do not support either of these predictions with respect to *G6pd* A−, and several observations suggest that patterns at *G6pd* fit the model expected in an early stage of selection (Figure 4b). First, overall levels of nucleotide diversity are close to average values for other X-linked loci. This is true for the worldwide sample and, more importantly for evaluating models of selection, it is also true for the African sample alone. An HKA test applied to our data fails to reject a neutral model. Second, there is no evidence for an excess of intermediate-frequency polymorphisms. In fact, both Tajima's *D* and Fu and Li's *D* are slightly (but not significantly) negative for the African sample (Table 2). Third, we find extensive linkage disequilibrium within and around *G6pd*, and this disequilibrium is due almost exclusively to nucleotide differences that distinguish the A− allele from other alleles. We observed no recombination events within *G6pd*. This stands in contrast to many other human nucleotide polymorphism data sets, including intron 44 of *Dmd*, surveyed in this same set of individuals (NACHMAN and CROWELL 2000), in which numerous recombination events were observed over distances of several hundred bases. In addition to significant LD within *G6pd*, we found significant LD between *G6pd* and *L1cam* (Table 3; $D' = 1$ in all comparisons), loci that are separated by ~550 kb. This amount of LD is much higher than typical values for the human genome. For example, REICH *et al.* (2001) recently studied the decay of D' for 19 different genomic regions and found that in a European population the half-length of D' (the distance at which the average D' drops below 0.5) is typically 60 kb, while in an African population the half-length of D' is <5 kb (REICH *et al.* 2001). Other studies have also revealed lower levels of linkage disequilibrium in African populations compared with non-African populations (TISHKOFF *et al.* 1996, 2001). Interestingly, we observe much higher levels of LD than those previously reported for this region of Xq28 by TAILLON-MILLER *et al.* (2000) in populations of European descent. Finally, there is no intra-allelic variation within *G6pd* A−, consistent with the notion that G6PD A− is relatively young.

Taken together, these observations argue against a model of long-term selection on the *G6pd* A− allele, but do not allow us to distinguish between recent balancing selection (*sensu* heterosis), on the one hand, and recent diversity-enhancing (*i.e.*, spatially and/or temporally varying) selection, on the other hand. Better fitness estimates of all genotypes (in particular, female deficiency homozygotes), as well as detailed sampling of

G6PD A− frequencies across Africa, might help us to distinguish between these hypotheses.

Contrary to the intra-allelic patterns of nucleotide variability for *G6pd* A−, the minor deficiency allele *G6pd* A+ shows a high level of intra-allelic diversity and greater linkage equilibrium. Although our study includes only two A+ chromosomes that represent a single haplotype, at least two additional haplotypes have been identified on the basis of RFLP analyses (Figure 2; Vulliammy *et al.* 1991). Moreover, microsatellites located up to 19 kb away from *G6pd* exhibit greater linkage equilibrium and higher diversity on A+ alleles than on A− alleles (Tishkoff *et al.* 2001). These observations, taken together with a coalescent-based estimate for the age of the mutation at coding position 376 from our study (131,250–174,375 years on the basis of *GENETREE* analysis), suggest that G6PD A+ may be relatively old. G6PD A+ has an enzymatic activity that is 80% of normal and does not appear to cause a significant clinical condition (Takizawa *et al.* 1987). Furthermore, G6PD A+ does not seem to currently confer resistance to severe *falciparum* malaria, as does G6PD A− (Ruwende *et al.* 1995). However, the age of G6PD A+ coupled with the reduced level of enzymatic activity raises the possibility that this allele has been under selection at some time in the past.

Is it possible that demographic processes are primarily responsible for the high levels of LD seen in Figure 2? Linguistic and archaeological evidence suggests that a Bantu expansion took place in Africa ~4000 years ago (Excoffier *et al.* 1987). This range expansion occurred in sub-Saharan Africa primarily from west to east and southward, a distribution that is similar to the current distribution of African populations with elevated G6PD− allele frequencies. If admixture from this range expansion were responsible for generating the observed LD in our data, we would also expect to see G6PD B alleles with significant LD. This is not observed. Instead, most of the LD in our data is found between sites on functionally different alleles, arguing against any simple demographic explanation. Likewise, no LD is observed between Bantu and non-Bantu individuals from this set of 41 individuals sampled for other loci (*e.g.*, Nachman and Crowell 2000).

One intriguing observation in our data set is the relatively high level of divergence found at *L1cam* between individuals bearing the G6PD A− allele and all other individuals. Four of the six (66.7%) G6PD A− alleles share a common motif of three polymorphisms in complete linkage disequilibrium (C, T, and T at positions 776, 885, and 2115, respectively; Table 1) while the rest of the segregating sites at *L1cam* include only four singletons and one doubleton. This pattern along with the significant LD between *G6pd* and *L1cam* (Table 3) suggests that the A− mutation arose on a relatively diverged haplotype, possibly as a consequence of population subdivision. Analysis of *G6pd* and *L1cam* as well as additional neighboring loci in a larger geographic sample from Africa may shed light on this unusual pattern.

In general, the observations reported here demonstrate that even when selection is relatively strong, its signature on patterns of DNA sequence variation may be subtle, especially if selection is recent. While several of the conventional statistical tests for selection fail to reject the null hypothesis, the footprint of selection is seen in the long-range patterns of LD and in the absence of variation among A− alleles from different parts of Africa. Similar patterns of nucleotide variability at *G6pd* have also recently been reported by Verrelli *et al.* (2002). The patterns of DNA sequence variation observed at *G6pd* are markedly different from those seen at another well-studied target of balancing selection, HLA, where ancient alleles result in substantially elevated levels of polymorphism (Grimsley *et al.* 1998; Horton *et al.* 1998; Gaudieri *et al.* 1999). The spatial and temporal scales over which selection pressures have shaped human genomic diversity are still largely unknown, but environmental changes associated with the transition from the Paleolithic to the Neolithic may have imposed substantial new selection pressures on humans, suggesting that patterns of nucleotide variability similar to those documented here for *G6pd* may be found at other loci.

Age of G6PD A− and the evolution of malarial resistance: These results have important implications for the evolution of resistance to malaria in humans. Several observations reported here, including average levels of nucleotide variability at *G6pd*, negative values of Tajima's *D*, high levels of linkage disequilibrium between *G6pd* and *L1cam*, and complete absence of variation among *G6pd* A− alleles from different parts of Africa, suggest that the A− allele is young (Tables 1 and 2). A recent study based on microsatellite haplotype diversity (Tishkoff *et al.* 2001) also suggests that the *G6pd* A− allele arose recently, within the past 4000–12,000 years. A recent phylogeny of primate malaria parasites indicates that *P. falciparum* is closely related to *P. reichenowi*, a chimpanzee parasite. Moreover, *cytochrome b* sequence divergence between *P. falciparum* and *P. reichenowi* suggests a divergence time of 4–5 million years ago (Escalante *et al.* 1998), in good agreement with the estimated time of the human-chimpanzee divergence. The discrepancy between this date and the recent origin of the *G6pd* A− allele raises the possibility that *P. falciparum* has been a human parasite for most of the evolutionary history of *H. sapiens*, but that the parasite's current level of virulence has evolved only recently (Rich *et al.* 1998). The estimated age of the A− allele agrees well with the spread of agriculture throughout sub-Saharan Africa (Waters *et al.* 1991; Cavalli-Sforza *et al.* 1996) and suggests that changes in human lifeway may have contributed to increased transmission and/or increased virulence of *P. falciparum*, perhaps through an increase in

the density and mobility of Anopheles mosquitoes that serve as vectors in transmission of malaria.

We thank J. D. Jensen and S. Peterson for technical assistance. Human DNA samples M115 and M241 were kindly donated by L. Luzzatto and K. Nafa. R. O. Ryder provided chimpanzee and orangutan samples. R. M. Harding, L. Luzzatto, E. Beutler, B. A. Payseur, E. T. Wood, C. C. Campbell, and A. J. Redd provided helpful discussion. We also thank R. G. Harrison and two anonymous reviewers who provided helpful comments about the manuscript. This work was supported by a National Science Foundation (NSF) grant to M.W.N. and M.F.H. and an NSF predoctoral fellowship to M.A.S.

LITERATURE CITED

Allison, A. C., 1960 Glucose-6-phosphate dehydrogenase deficiency in red blood cells of east Africans. Nature **186:** 531–532.

Alonso, S., and J. A. L. Armour, 2001 A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa. Proc. Natl. Acad. Sci. USA **98:** 864–869.

Andolfatto, P., J. D. Wall and M. Kreitman, 1999 Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. Genetics **153:** 1297–1311.

Bahlo, M., and R. C. Griffiths, 2000 Inference from gene trees in a subdivided population. Theor. Popul. Biol. **57:** 79–95.

Beutler, E., 1994 G6PD deficiency. Blood **84:** 3613–3636.

Beutler, E., W. Kuhl, J. L. Vivescorrons and J. T. Prchal, 1989 Molecular heterogeneity of glucose-6-phosphate dehydrogenase A−. Blood **74:** 2550–2555.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza, 1996 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan *et al.*, 1998 Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. **63:** 595–612.

Deinard, A. S., and K. K. Kidd, 1998 Evolution of a D2 dopamine receptor intron within the great apes and humans. DNA Seq. **8:** 289–301.

Escalante, A. A., D. E. Freeland, W. E. Collins and A. A. Lal, 1998 The evolution of primate malaria parasites based on the gene encoding *cytochrome b* from the linear mitochondrial genome. Proc. Natl. Acad. Sci. USA **95:** 8124–8129.

Excoffier, L., B. Pelegrini, A. Sanchez-Mazas, C. Simon and A. Langley, 1987 Genetics and history of sub-Saharan Africa. Yearb. Phys. Anthropol. **30:** 151–194.

Fay, J. C., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Filatov, D. A., F. Moneger, I. Negrutiu and D. Charlesworth, 2000 Low variability in a Y-linked plant gene and its implications for Y-chromosome evolution. Nature **404:** 388–390.

Fu, Y. X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Fullerton, S. M., A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor *et al.*, 2000 Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am. J. Hum. Genet. **67:** 881–900.

Gaudieri, S., J. K. Kulski, R. L. Dawkins and T. Gojobori, 1999 Extensive nucleotide variability within a 370 kb sequence from the central region of the major histocompatibility complex. Gene **238:** 157–161.

Gilad, Y., D. Segre, K. Skorecki, M. W. Nachman, D. Lancet *et al.*, 2000 Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. Nat. Genet. **26:** 221–224.

Grimsley, C., K. A. Mather and C. Ober, 1998 HLA-H: a pseudogene with increased variation due to balancing selection at neighboring loci. Mol. Biol. Evol. **15:** 1581–1588.

Hamblin, M. T., and A. Di Rienzo, 2000 Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. **66:** 1669–1679.

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox *et al.*, 1997 Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60:** 772–789.

Harris, E. E., and J. Hey, 1999 X chromosome evidence for ancient human histories. Proc. Natl. Acad. Sci. USA **96:** 3320–3324.

Harris, E. E., and J. Hey, 2001 Human populations show reduced DNA sequence variation at the Factor IX locus. Curr. Biol. **11:** 774–778.

Hedrick, P. W., 1998 *Genetics of Populations*. Jones & Bartlett, Sudbury, MA.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226–231.

Hirono, A., and E. Beutler, 1988 Molecular cloning and nucleotide sequence of cDNA for human glucose-6-phosphate dehydrogenase variant A(−). Proc. Natl. Acad. Sci. USA **85:** 3951–3954.

Horton, R., D. Niblett, S. Milne, S. Palmer, B. Tubby *et al.*, 1998 Large scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. J. Mol. Biol. **282:** 71–97.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

Hudson, R. R., 1996 Molecular population genetics of adaptation, pp. 291–309 in *Adaptation*, edited by M. R. Rose and G. V. Lauder. Academic Press, San Diego.

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

International Human Genome Sequencing Consortium, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860–921.

International SNP Map Working Group, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409:** 928–933.

Jaruzelska, J., E. Zietkiewicz, M. Batzer, D. E. C. Cole, J. P. Moisan *et al.*, 1999 Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron: analysis of the haplotype structure and genealogy. Genetics **152:** 1091–1101.

Kaessmann, H., F. Heissig, A. Von Haeseler and S. Paabo, 1999 DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat. Genet. **22:** 78–81.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.*, 2002 A high-resolution recombination map of the human genome. Nat. Genet. **31:** 241–247.

Lewontin, R. C., 1964 Interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

Livingstone, F. B., 1985 *Frequencies of Heamoglobin Variants: Thalassemia, The Glucose-6-Phosphate Dehydrogenase deficiency, G6PD Variants, and Ovalocytosis in Human Populations*. Oxford University Press, New York.

Miller, L. H., 1994 Impact of malaria on genetic polymorphism and genetic diseases in Africans and African-Americans. Proc. Natl. Acad. Sci. USA **91:** 2415–2419.

Motulsky, A. G., 1961 Glucose-6-phosphate-dehydrogenase deficiency, haemolytic disease of the newborn, and malaria. Lancet **1:** 1168–1169.

Nachman, M. W., 2001 Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. **17:** 481–485.

Nachman, M. W., and S. L. Crowell, 2000 Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, *Dmd*, in humans. Genetics **155:** 1855–1864.

Nachman, M. W., V. L. Bauer, S. L. Crowell and C. F. Aquadro, 1998 DNA variability and recombination rates at X-linked loci in humans. Genetics **150:** 1133–1141.

Nei, M., and W.-H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:** 5269–5273.

Oppenheim, A., C. L. Jury, D. Rund, T. J. Vulliamy and L. Luzzatto, 1993 G6PD Mediterranean accounts for the high prevalence of G6PD deficiency in Kurdish Jews. Hum. Genet. **91:** 293–294.

Payseur, B. A., and M. W. Nachman, 2000 Microsatellite variation and recombination rate in the human genome. Genetics **156:** 1285–1298.

Payseur, B. A., and M. W. Nachman, 2002 Natural selection at linked sites in humans. Gene **300:** 31–42.

Rana, B. K., D. Hewett-Emmett, L. Jin, B. H. J. Chang, N. Sambu-

UGHIN *et al.*, 1999 High polymorphism at the human melanocortin-1 receptor locus. Genetics **151:** 1547–1557.

REICH, D. E., M. CARGILL, S. BOLK, J. IRELAND, P. C. SABETI *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

RICH, S. M., M. C. LIGHT, R. R. HUDSON and F. J. AYALA, 1998 Malaria's eve: evidence of a recent population bottleneck throughout the world populations of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA **95:** 4425–4430.

RIEDER, M. J., S. L. TAYLOR, A. G. CLARK and D. A. NICKERSON, 1999 Sequence variation in the human angiotensin converting enzyme. Nat. Genet. **22:** 59–62.

ROTH, E. F., and S. SCHULMAN, 1988 The adaptation of *Plasmodium falciparum* to oxidatve stress in G6PD deficient human erythrocytes. Br. J. Haematol. **70:** 363–367.

ROTH, E. F., C. RAVENTOS-SUAREZ, A. RINALDI and R. L. NAGEL, 1983 Glucose-6-phosphate dehydrogenase deficiency inhibits *in vitro* growth of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA **80:** 298–299.

RUWENDE, C., S. C. KHOO, A. W. SNOW, S. N. R. YATES, D. KWIATKOWSKI *et al.*, 1995 Natural-selection of hemizygotes and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature **376:** 246–249.

SCHMIDT, G. D., and L. S. ROBERTS, 1996 *Foundations of Parasitology*. Wm. C. Brown Publishers, Chicago.

SMALL, K., J. IBER and S. T. WARREN, 1997 Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. Nat. Genet. **16:** 96–99.

TAILLON-MILLER, P., I. BAUER-SARDINA, N. L. SACCONE, J. PUTZEL, T. LAITINEN *et al.*, 2000 Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nat. Genet. **25:** 324–328.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

TAKAHATA, N., Y. SATTA and J. KLEIN, 1992 Polymorphism and balancing selection at major histocompatibility complex loci. Genetics **130:** 925–938.

TAKIZAWA, T., Y. YONEYAMA, S. MIWA and A. YOSHIDA, 1987 A single nucleotide base transition is the basis of the common human glucose-6-phosphate dehydrogenase variant A(+). Genomics **1:** 288.

TISHKOFF, S. A., E. DIETZSCH, W. SPEED, A. J. PAKSTIS, J. R. KIDD *et al.*, 1996 Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science **271:** 1380–1387.

TISHKOFF, S. A., R. VARKONYI, N. CAHINHINAN, S. ABBES, G. ARGYROPOULOS *et al.*, 2001 Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science **293:** 455–462.

TRIGG, P. I., and A. V. KONRACHINE, 1998 Commentary: malaria control in the 1990s. Bull. WHO **76:** 11–16.

VENTER, J. C., M. D. ADAMS, E. W. MYERS, P. W. LI, R. J. MURAL *et al.*, 2001 The sequence of the human genome. Science **291:** 1304–1351.

VERRELLI, B. C., J. H. MCDONALD, G. ARGYROPOLOUS, G. DESTRO-BISOL, A. FROMENT *et al.*, 2002 Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. Am. J. Hum. Genet. **71:** 1112–1128.

VULLIAMY, T. J., A. OTHMAN, M. TOWN, A. NATHWANI, A. G. FALUSI *et al.*, 1991 Polymorphic sites in the African population detected by sequence analysis of the glucose-6-phosphate dehydrogenase gene outline the evolution of the variant A+ and variant A−. Proc. Natl. Acad. Sci. USA **88:** 8568–8571.

WATERS, A. P., D. G. HIGGINS and T. F. MCCUTCHAN, 1991 *Plasmodium falciparum* appears to have arisen as a result of lateral transfer between avian and human hosts. Proc. Natl. Acad. Sci. USA **88:** 3140–3144.

WATTERSON, G. A., 1975 Number of segregating sites in genetic models without recombination. Theor. Popul. Biol. **7:** 256–276.

YU, N., Y. X. FU, N. SAMBUUGHIN, M. RAMSAY, T. JENKINS *et al.*, 2001 Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. **18:** 214–222.

ZHAO, Z. M., L. JIN, Y. X. FU, M. RAMSAY, T. JENKINS *et al.*, 2000 Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. USA **97:** 11354–11358.

Communicating editor: R. G. HARRISON