DOI: 10.1534/genetics.105.048140

The Extent of Linkage Disequilibrium Caused by Selection on *G6PD* in Humans

Matthew A. Saunders,*[,1] Montgomery Slatkin,[†] Chad Garner,[‡] Michael F. Hammer* and Michael W. Nachman*

**Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, [†]Department of Integrative Biology, University of California, Berkeley, California 94720 and [‡]Department of Medicine, University of California, Irvine, California 92697*

Manuscript received July 12, 2005
Accepted for publication July 13, 2005

ABSTRACT

The gene coding for glucose-6-phosphate dehydrogenase (G6PD) is subject to positive selection by malaria in some human populations. The G6PD A− allele, which is common in sub-Saharan Africa, is associated with deficient enzyme activity and protection from severe malaria. To delimit the impact of selection on patterns of linkage disequilibrium (LD) and nucleotide diversity, we resequenced 5.1 kb at *G6PD* and ~2–3 kb at each of eight loci in a 2.5-Mb region roughly centered on *G6PD* in a diverse sub-Saharan African panel of 51 unrelated men (including 20 G6PD A−, 11 G6PD A+, and 20 G6PD B chromosomes). The signature of selection is evident in the absence of genetic variation at *G6PD* and at three neighboring loci within 0.9 Mb from *G6PD* among all individuals bearing G6PD A− alleles. A genomic region of ~1.6 Mb around *G6PD* was characterized by long-range LD associated with the A− alleles. These patterns of nucleotide variability and LD suggest that *G6PD A−* is younger than previous age estimates and has increased in frequency in sub-Saharan Africa due to strong selection ($0.1 < s < 0.2$). These results also show that selection can lead to nonrandom associations among SNPs over great physical and genetic distances, even in African populations.

RECENT studies have focused on describing and understanding the general structure of linkage disequilibrium (LD) in the human genome, primarily to provide a sound basis for mapping disease loci in association studies (Kruglyak 1999; Goldstein 2001). Patterns of LD are expected to be complicated because LD is affected by many factors, including genetic drift, population structure, migration, admixture, selection, mutation, gene conversion, and recombination (Ardlie *et al.* 2002). Moreover, some of these factors, such as recombination, are not constant across the genome (McVean *et al.* 2004), and thus LD is expected to vary in different genomic regions. Despite this expected complexity, several general results have emerged from empirical studies of LD in humans. First, the human genome is divided into haplotype blocks, with regions of high LD over fairly long stretches, separated by regions with little LD (Daly *et al.* 2001; Gabriel *et al.* 2002; Phillips *et al.* 2003; Wall and Pritchard 2003). There is clear evidence that the spaces between these blocks correspond to recombination hotspots in some cases (*e.g.*, Jeffreys *et al.* 2001), although simulations suggest that a block-like pattern may be expected even in the absence of recombination hotspots (Phillips *et al.* 2003). Recombination hotspots may occur in the human genome roughly every 200 kb (McVean *et al.* 2004). Second, there appears to be less LD in African populations than in non-African populations (Tishkoff *et al.* 1996; Reich *et al.* 2001). This observation is consistent with the presumed larger long-term effective population size for African populations. Third, selection on individual genes can elevate levels of LD in a given genomic region (*e.g.*, Huttley *et al.* 1999; Sabeti *et al.* 2002; Saunders *et al.* 2002; Toomajian and Kreitman 2002; Swallow 2003). In fact, this expectation has motivated several statistical tests of a neutral model of molecular evolution (*e.g.*, Hudson *et al.* 1994; Kelly 1997; Slatkin and Bertorelle 2001; Sabeti *et al.* 2002; Toomajian *et al.* 2003). It is difficult to predict exactly how far the effects of selection will extend because the observed patterns will depend on multiple factors, including the type of selection (*e.g.*, balancing, purifying, directional), the time over which selection has acted, the strength of selection, the local recombination rate, and various demographic factors. To study this problem empirically, we have chosen to focus on the genomic region surrounding *G6PD*, a gene known to be subject to selection in humans.

Glucose-6-phosphate dehydrogenase (G6PD) is a housekeeping enzyme that catalyzes a critical step in the pentose monophosphate shunt of glycolysis. G6PD deficiency mutations cause hemolytic anemia and neonatal jaundice (Beutler 1994); however, some deficiency mutations also confer resistance to severe malaria

[1]*Corresponding author:* Department of Ecology and Evolution, University of Chicago, 1101 E. 57th St., Chicago, IL 60637.
E-mail: saunders@uchicago.edu

(Allison 1960; Motulsky 1961; Ruwende *et al.* 1995). Many human populations exhibit G6PD deficiency alleles at frequencies that range between 0.05 and 0.20 as a consequence of selection. G6PD A− is a common deficiency allele in sub-Saharan Africa that reaches frequencies of ~0.2 in populations living in malarial areas (Livingstone 1985; Cavalli-Sforza *et al.* 1996). This allele is characterized by two nonsynonymous changes relative to the normal allele (G6PD B) (Figure 1), which decrease enzyme activity to ~12% of normal (Hirono and Beutler 1988) and confer ~50% reduction in risk of severe malaria in both females and males (Ruwende *et al.* 1995). It follows that the G6PD A− allele is beneficial in the presence of malaria caused by *Plasmodium falciparum*, while in the absence of malaria this allele is deleterious. The wealth of knowledge and the clear understanding of genotype-phenotype connections for G6PD make it a useful model for studying the effects of selection on patterns of LD in humans.

Recently, several studies have investigated patterns of nucleotide variability at *G6PD* and at loci relatively close to *G6PD* (Tishkoff *et al.* 2001; Sabeti *et al.* 2002; Saunders *et al.* 2002; Verrelli *et al.* 2002). All of these studies documented LD associated with the G6PD A− allele. In particular, Sabeti *et al.* (2002) and Saunders *et al.* (2002) showed that LD extended over ~550 kb in an African sample. Neither of these studies surveyed loci beyond this distance, and therefore they were unable to delimit the full extent of LD caused by selection on *G6PD*. Here, we extend these results to delimit the genomic region over which selection at *G6PD* has created LD. We resequenced ~3-kb windows from each of eight loci in a 2.5-Mb region, centered roughly on *G6PD* in a panel of 51 individuals from sub-Saharan Africa. In this panel, we also resequenced 5.1 kb at *G6PD* and ~2 kb at an unlinked "control" locus, situated 19 Mb proximal to *G6PD*. Our data show that selection at *G6PD* has affected a region that spans >1.6 Mb of the human X chromosome, demonstrating that selection can have considerable effects on nucleotide variability over remarkably long genomic distances in humans.

MATERIALS AND METHODS

Samples: DNA sequences were determined in a sample of 51 human males of African descent (Table 1) that includes 20 G6PD A− alleles, 11 G6PD A+ alleles, and 20 G6PD B alleles. The G6PD A− allele is defined by two mutations: an A → G nonsynonymous mutation at *G6PD* coding site 376, cooccurring with a G → A nonsynonymous mutation at coding site 202 (Figure 1) (Hirono and Beutler 1988). G6PD A+ is defined by the A → G nonsynonymous mutation at *G6PD* coding site 376 that reduces enzyme efficiency to 80% of normal (Figure 1). This mild deficiency allele does not confer resistance to malaria (Ruwende *et al.* 1995) and is found in sub-Saharan Africa at a frequency of ~0.2 (Takizawa *et al.* 1987). All G6PD functional alleles were determined *a priori* by restriction fragment length polymorphism analysis of a *Fok*I restriction site at coding position 376 and a *Nla*III restriction

TABLE 1

Individuals sampled in this study

G6PD allele type	Sample	Country	Ethnic/language group
G6PD A−	Ivc01	Ivory Coast	Niger-Congo
	Ivc17	Ivory Coast	Niger-Congo
	S628[a]	Kenya	Unknown
	Alb77	South Africa	Sotho
	Mgr40	Togo	Niger-Congo
	VA010	United States	African-American
	M115	United States	African-American
	M241	United States	African-American
	VA084	United States	African-American
	VA025	United States	African-American
	VA076[a]	United States	African-American
	VA085	United States	African-American
	VA088	United States	African-American
	DKT338	United States	African-American
	DKT381	United States	African-American
	DKT382	United States	African-American
	JK1031	Zaire	Mbuti Pygmy
	Sho07	Zimbabwe	Shona
	Sho18	Zimbabwe	Shona
	Sho49[a]	Zimbabwe	Shona
G6PD A+	JK785[a]	Central African Republic	Biaka Pygmy
	Ivc22	Ivory Coast	Niger-Congo
	JK1071[a]	Zaire	Mbuti Pygmy
	AU26	Kenya	Unknown
	JW058[a]	Mali	Malinke
	Alb27[a]	South Africa	Zulu
	VA024[a]	United States	African-American
	DKT275	United States	African-American
	VA012	United States	African-American
	JK1058	Zaire	Mbuti Pygmy
	Sho03	Zimbabwe	Shona
G6PD B	JK736[a]	Central African Republic	Biaka Pygmy
	JK741[a]	Central African Republic	Biaka Pygmy
	Gna02[a]	Ghana	Akan
	Ivc04[a]	Ivory Coast	Niger-Congo
	Ivc18	Ivory Coast	Niger-Congo
	Ivc20[a]	Ivory Coast	Niger-Congo
	Ivc23[a]	Ivory Coast	Niger-Congo
	Mka21[a]	Kenya	Maasai
	Koh188[a]	Kenya	Meru
	Mka29[a]	Kenya	Tsumkwe
	JR013[a]	Namibia	Biaka
	JR323	Namibia	Bagandu
	LD156[a]	South Africa	Khoisan
	Alb74[a]	South Africa	Unknown
	DKT331[a]	United States	African-American
	JK1029[a]	Zaire	Mbuti
	JK1033[a]	Zaire	Ituri
	Sho14[a]	Zimbabwe	Shona
	Sho30[a]	Zimbabwe	Shona
	Sho46[a]	Zimbabwe	Shona

[a] Individuals used for constructed random sample (CRS) analyses.

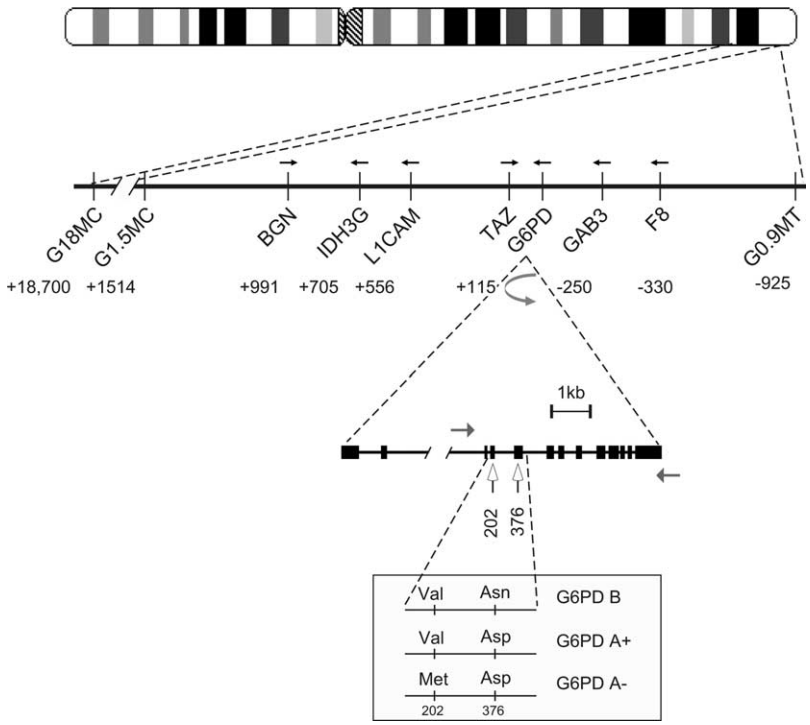


FIGURE 1.—Ideogram of the human X chromosome and the genomic regions surveyed in this study. Approximate distances between each of the surveyed windows including *G6PD* are marked on the scale. Transcription orientations of the genic regions are marked with solid arrows. The exon/intron structure of *G6PD* is designated (displayed in inverted orientation relative to chromosomal orientation) along with the defining substitutions of the three allelic classes: A−, A+, and B (in box). Positions of amplification primers used to survey the 5.1-kb window of *G6PD* are marked with shaded arrows.

site at coding position 202 (XU *et al.* 1995). As G6PD A− is believed to be of a single origin (SAUNDERS *et al.* 2002; VERRELLI *et al.* 2002) we selected individuals to represent diverse localities in sub-Saharan Africa. By studying X-linked loci in males we were able to PCR amplify single alleles and directly recover haplotypes over long genomic distances to study patterns of LD. Homologous sequences from a chimpanzee (*Pan troglodytes*) and an orangutan (*Pongo pygmaeus*) were also determined at each locus for divergence estimates. All sampling protocols were approved by the Human Subjects Committee at the University of Arizona.

Loci surveyed: *G6PD* and nine flanking loci (*G18MC*, *G1.5MC*, *IDH3G*, *BGN*, *L1CAM*, *TAZ*, *GAB3*, *F8*, and *G0.*9MT) were surveyed for nucleotide variability (Figure 1). Loci *G18MC*, *G1.5MC*, and *G0.9MT* (our nomenclature) are anonymous intergenic regions while the remainder of the regions surveyed are primarily introns. All loci were chosen because of their physical distance from *G6PD*, and there is no evidence that any of these loci are themselves targets of selection. Approximately 30 other genes are found within 1 Mb on either side of *G6PD* and none of these genes is known to be recent targets of positive directional selection in sub-Saharan Africa. All loci chosen are single copy in the genome and are situated outside the pseudoautosomal region. *G0.9MT* is situated near the boundary of the pseudoautosomal region, and we did not survey loci distal to this locus.

PCR amplification and sequencing: Single PCR fragments (2–5 kb) were amplified for each individual at each locus using a long-range PCR system (Invitrogen HiFi Taq). Amplification primers for *G6PD* and *L1CAM* are found in SAUNDERS *et al.* (2002) and primers for all other loci can be found at www.genetics.org/supplemental (Table S1). Internal primers were used to generate overlapping sequence runs on an ABI 3730 automated sequencer. Contiguous sequence that included coding and noncoding regions was assembled for each individual for each locus, using the computer program *SEQUENCHER* (Gene Codes, Ann Arbor, MI). Sequences have been submitted to GenBank under accession nos. DQ173562–DQ173642.

Nucleotide variability data analysis: Nucleotide variability and the frequency spectrum of alleles in an unbiased African sample have been reported in detail elsewhere for *G6PD* (SAUNDERS *et al.* 2002; VERRELLI *et al.* 2002). To gain insight into the long-range effects of selection on nucleotide variability we analyzed three subsets of our data: G6PD A−, G6PD A+, and G6PD B alleles. We calculated haplotype diversity (H_d), θ_π (NEI and LI 1979), and θ_w (WATTERSON 1975) at each locus using *dnaSP* 4.0 (ROZAS *et al.* 2003). Under neutral equilibrium conditions both θ_π and θ_w for a random sample 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, assuming a sex ratio of one. However, the structure of the sample in the present study is nonrandom, and therefore θ_π and θ_w are used simply as measures of nucleotide variability. We also analyzed nucleotide variability in a constructed random sample (CRS) (HUDSON *et al.* 1994). This subset of chromosomes ($n = 26$) contains G6PD alleles at frequencies that are representative of a typical sub-Saharan African population subject to malarial selection, on the basis of extensive allele frequency surveys (G6PD A−, 0.11; G6PD A+, 0.20; G6PD B, 0.69) (LIVINGSTONE 1985). We calculated θ_π, θ_w, Tajima's D (TAJIMA 1989), and Fu and Li's D (FU and LI 1993) at each locus for the CRS to compare to nonbiased African samples that are available for different X-linked loci. Detailed statistics of nucleotide variability for the CRS are available at www.genetics.org/supplemental (Table S2). Divergence data were derived for each of these loci by calculating the average of all pairwise comparisons between the homologous sequence from an outgroup and the samples in the CRS. LD between pairs of polymorphic sites was measured using the statistic $|D'|$ (LEWONTIN 1964). This measure of linkage disequilibrium 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).

The age of the G6PD A− allele: The age of the G6PD A− allele and the intensity of past selection it experienced were

estimated by combining the method of Slatkin (2001) for generating intraallelic genealogies of selected alleles with the method of Garner and Slatkin (2002) for estimating the probability of haplotypes at two linked loci. All analyses were performed on the basis of long-range haplotypes, using the intralocus combination of sites 55, 59, and 60 at *L1CAM*; SNP 90 at *G6PD* (*i.e.*, coding site 202); and the intralocus combination of sites 99, 100, and 101 at *G0.9MT* (Figure 2). The computer program described by Slatkin (2001) was used to generate sample paths of allele frequency from the time of the mutation (t_1, the allele age) until the present ($t = 0$), with the constraint that the frequency at $t = 0$ is the observed frequency, 0.1. An additive dominance model was used. We assumed a constant population size of $N_e = 10{,}000$ and $N_e = 20{,}000$ individuals. For each sample path, a neutral coalescent model was used to generate an intraallelic genealogy of G6PD A− since it arose by mutation. The intraallelic coalescence times from this genealogy were then passed as parameters to a program that estimates the probability of obtaining the observed configuration of the data (the numbers of the four haplotypes found on the 20 A−-bearing chromosomes), given the recombination rates and haplotype frequencies on non-A−-bearing chromosomes (assumed constant). That probability is the likelihood of the data, given the intraallelic genealogy. For each value of *s*, the selective advantage of A− was considered, 90,000 sample paths and intraallelic genealogies were generated, and 10 replicates of the Garner-Slatkin program were used to estimate the likelihood for each sample path. Likelihoods were averaged across sample paths, using the weighting method described by Slatkin (2001). For each parameter value, this method provided an estimate of the likelihood of the data under the model and an estimate of the posterior distribution of allele age.

To allow analysis of the multisite data set by this method, we used the fact that all 20 A− chromosomes carried the same haplotype for 925 kb telomeric to *G6PD* (to locus *G0.9MT*) and that 14 of 20 chromosomes carried the same haplotype for 556 kb centromeric to *G6PD* (see SNPs 55, 59, and 60 at *L1CAM*; Figure 2). The recombination parameters in the two directions were assumed to be $c = 0.01675$ and 0.00555 M for *L1CAM* and *G0.9MT*, respectively (Kong *et al.* 2002). Therefore, the two-locus data set was 14 *AMB*, 6 *aMB*, 0 *AMb*, 0 *aMb* (in the notation of Garner and Slatkin 2002), where *A* represents the haplotype of rare alleles at *L1CAM* SNPs 55, 59, and 60 (which are very rare on non-A− chromosomes), *M* represents the site under selection, and *B* represents the multilocus haplotype at *G0.9MT*, which is rare on non-A− chromosomes.

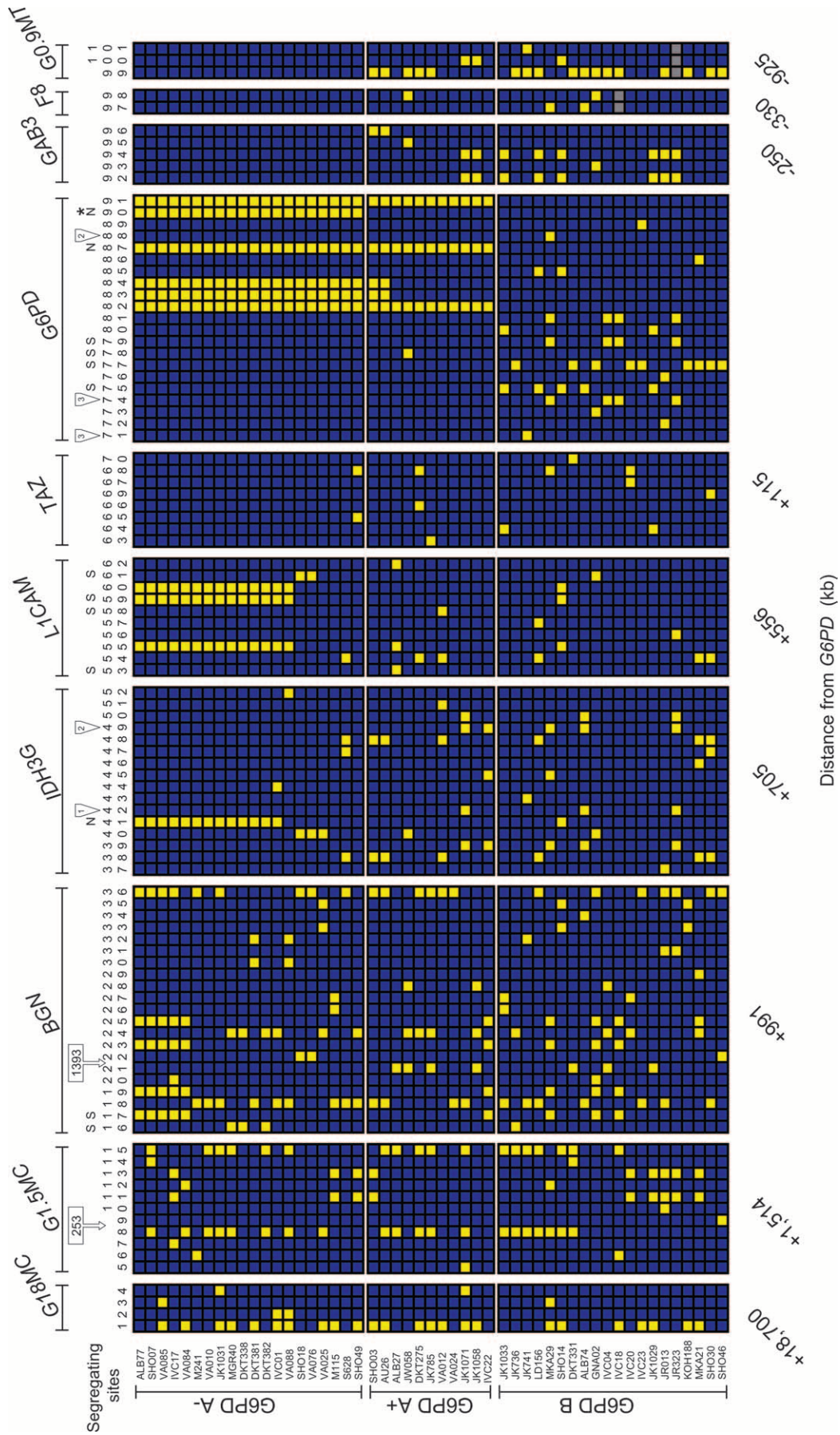
RESULTS

Nucleotide diversity at *G6PD*: Polymorphic sites at *G6PD* are presented in Figure 2. We calculated nucleotide variability for three subset groups: (i) individuals bearing the G6PD A− allele ($n = 20$), (ii) individuals bearing the G6PD A+ allele ($n = 11$), and (iii) individuals bearing the G6PD B allele ($n = 20$). At *G6PD* we observed 18 segregating sites (excluding three INDELs) in the entire sample, consistent with previous findings (Saunders *et al.* 2002; Verrelli *et al.* 2002). Among the G6PD A− individuals ($n = 20$) there was no nucleotide variability in 5109 bp of contiguous DNA sequence. Noncoding nucleotide variability among G6PD A+ and G6PD B alleles was $\theta_\pi = 0.024$ and 0.04%, respectively (Figure 3a). Nucleotide variability for the CRS was $\theta_\pi = 0.068\%$ and $\theta_w = 0.090\%$ (Figure 3a; Table S2 at http://www.genetics.org/supplemental/), consistent with previously reported levels of nucleotide variability at *G6PD* in sub-Saharan Africa (Sabeti *et al.* 2002; Saunders *et al.* 2002; Verrelli *et al.* 2002) and with the average of 15 other X-linked loci ($\theta_\pi = 0.0755\%$, $\theta_w = 0.0815\%$) (Hammer *et al.* 2004). Tajima's *D* and Fu and Li's *D* for the CRS were −0.794 and −0.142, respectively (Table S2 at http://www.genetics.org/supplemental/), also consistent with the previous studies at *G6PD*.

Nucleotide diversity around *G6PD*: The segregating sites for nine loci flanking *G6PD* are presented in Figure 2. At *GAB3*, *F8*, and *G0.9MT*, loci distal to *G6PD*, we found no nucleotide variability in the A− group (Figure 3a). This portion of the data includes a cumulative survey of 13,582 bp, thus exhibiting a remarkable degree of nucleotide homogeneity among 20 unrelated individuals of African decent. At these same loci, the average noncoding nucleotide variability for the A+ group, the B group, and the CRS was $\theta_\pi =$ 0.027, 0.024, 0.022% respectively (Figure 3a). Estimates of haplotype diversity for the different allele classes also demonstrate a significant contrast between the A− alleles and the other allele classes. At loci distal to *G6PD*, the average haplotype diversity was $H_d = 0.0$, 0.435, and 0.879 for A−, A+, and B, respectively (Figure 3b).

This general pattern of reduced variability among A− individuals is also seen proximal to *G6PD*; however, the pattern is not as extreme as that on the distal side, and the pattern decays beyond *L1CAM* (~556 kb from *G6PD*; Figure 3, a and b). At *TAZ*, *L1CAM*, and *IDH3G* the average noncoding nucleotide variability for the

Figure 2.—Table of polymorphism for *G6PD* and surrounding loci. Fifty-one unrelated human males of sub-Saharan African descent were surveyed for nucleotide variability at *G6PD* and nine surrounding loci. Individual samples were selected on the basis of *a priori* allele type determination based on coding sites *202* and *376* of *G6PD* to define three allele classes: A−, A+, and B. Each segregating site (in columns) represents a biallelic marker (*i.e.*, SNP or INDEL). Segregating sites are listed in numerical order. For exact alignment positions (in base pairs) of segregating sites and the identity of the polymorphic nucleotides (*i.e.*, A, G, C, or T), sequences are available from GenBank or upon request from the author. At each segregating site one allelic state is marked with a blue box and the alternate allelic state is marked with a yellow box. Missing data for individual Ivc18 at locus *F8* and for individual JR323 at locus *G0.9MT* are indicated with gray boxes. Boundaries between loci are marked by vertical white bars. S and N denote synonymous and nonsynonymous changes, respectively. *G6PD* coding site 202 (segregating site 90) is marked with an asterisk. Polymorphisms at sites 72, 73, 74, and 75 in *G6PD* are in the 3′-untranslated region of exon 13. All other polymorphisms are in introns or intergenic regions. INDELs are marked with an open triangle including the size of the INDEL in base pairs. At *G1.5MC* and *BGN* unsurveyed regions in the otherwise contiguous windows are marked by an arrow with numbers in box indicating the number of contiguous unsurveyed base pairs.


G18MC
G1.5MC
BGN
IDH3G
L1CAM
TAZ
G6PD
GAB3
F8
G0.9MT
253
1393
Segregating sites
G6PD A-
G6PD A+
G6PD B
-18,700
+1,514
+991
+705
+556
+115
-250
-330
-925
Distance from G6PD (kb)

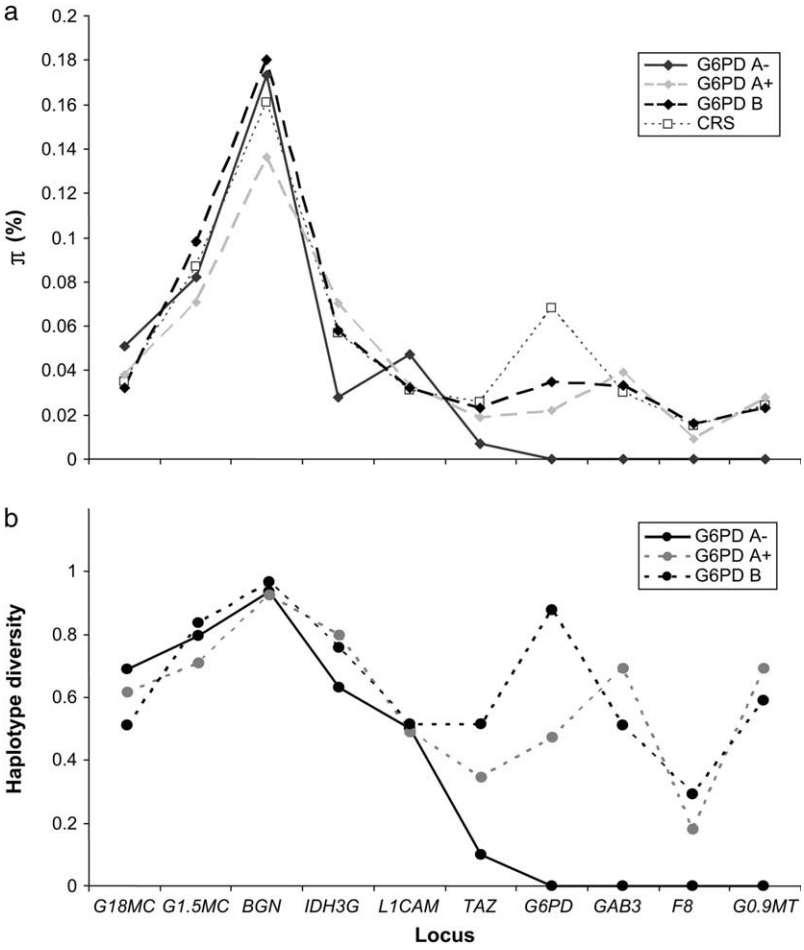


FIGURE 3.—Nucleotide variability for *G6PD* and nine surrounding loci for subset groups of the data set (G6PD A− alleles, G6PD A+ alleles, G6PD B alleles, and CRS): (a) nucleotide diversity ($\theta\pi$); (b) haplotype diversity.

A− group, A+ group, B group, and CRS was $\theta_\pi = 0.032$, 0.037, 0.040, and 0.041%, respectively (Figure 3a; Table S2 at http://www.genetics.org/supplemental/). At *BGN*, *G1.5MC*, and *G18MC*, loci mapping 0.9–19 Mb from *G6PD*, the average noncoding nucleotide variability for the A− group, A+ group, B group, and CRS was $\theta_\pi =$ 0.093, 0.072, 0.091, and 0.082%, respectively (Figure 3a; Table S2 at http://www.genetics.org/supplemental/). At these three loci the average level of nucleotide variability among the A− individuals is not reduced relative to that of the other subsets of the data. Together our results demonstrate that the A− chromosomes exhibit reduced variability relative to other allelic classes at loci around *G6PD* over a region that spans ~1.5 Mb (roughly from *L1CAM* to *G0.9MT*). This effect may extend further distally, but we were unable to survey loci beyond *G0.9MT*, which lies near the border of the pseudoautosomal region (see MATERIALS AND METHODS).

Linkage disequilibrium: To examine patterns of linkage disequilibrium we calculated $|D'|$ (LEWONTIN 1964) for all pairwise comparisons of segregating sites for which the minor allele was found in five or more individuals (Figure 4). Intragenic pairwise comparisons show strong LD within each of the loci surveyed, consistent with expectations over short distances. However, a striking feature of the data is seen in intergenic comparisons. Within *G6PD*, all A− individuals share a common haplotype that differs from the consensus *G6PD B* allele at six sites (segregating sites 82, 83, 84, 87, 90, and 91). These sites exhibit strong LD (significant by Fisher's exact test) with sites at *L1CAM* (sites 55, 59, and 60), *IDH3G* (site 41), *GAB3* (sites 92 and 94), and *G0.9MT* (site 99). Furthermore, site 99 at *G0.9MT* exhibits complete LD ($D' = 1$) with the aforementioned sites at *L1CAM* and *IDH3G*. Together, this pattern defines a conserved G6PD A− haplotype that spans >1.6 Mb encompassing *G6PD*. Although site 21 at *BGN* also exhibits strong LD ($D' = 1$) with three sites associated with the common *G6PD A*− haplotype (sites 83, 84, and 90), this pattern does not represent conservation of the extended A− haplotype, as sites such as 17, 19, 23, 25, and 36 at *BGN* do not exhibit strong LD with *G6PD*. When G6PD A− individuals are excluded from the analysis, few intergenic associations in significant LD are found. At *IDH3G* a haplotype that consists of the minor alleles at sites 38 and 48 is found in significant LD with site 36 of *BGN* and site 54 of *L1CAM*. Significant intergenic LD is also found between site 75 of *G6PD* and sites 92 and 94 of *GAB3*. This LD is not associated with any known functional alleles. As the minor allele polymorphisms in each of these cases are not associated with the extended G6PD A− haplotype, this LD remains

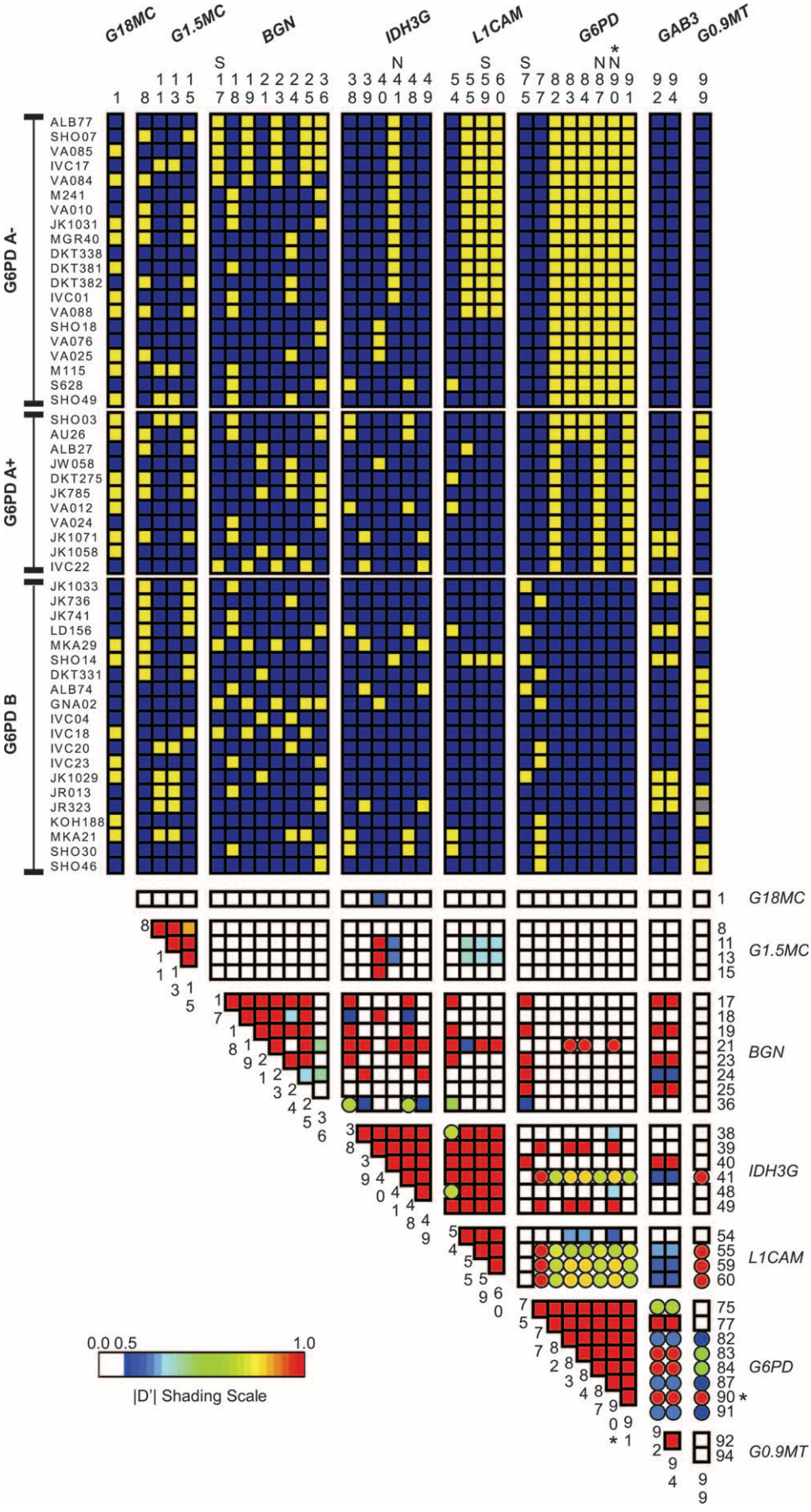


FIGURE 4.—Patterns of LD between segregating sites at *G6PD* and flanking loci. The table of polymorphism includes only segregating sites at which the less common allele (minor allele) is found in five or more individuals. At each segregating site, one allele is marked with a blue box and the alternate allele is marked with a yellow box. Missing data are marked with a gray box. Segregating sites are numbered and labeled according to Figure 2. Boundaries between loci surveyed are marked by vertical white bars. Below the table of polymorphism is a matrix of estimates of $|D'|$ for all pairwise comparisons of sites. Values of $|D'|$ are shown between 0.5 and 1.0 in accordance with the shading scale. Intergenic pairwise associations in significant LD ($P < 0.05$) by Fisher's exact test are marked in the matrix by circles.

intact when G6PD A− individuals are excluded from analyses. In summary, these data demonstrate that a majority of the intergenic LD in the surveyed region is due to the extended G6PD A− haplotype.

Age of the G6PD A− allele and strength of selection: Figure 5A shows the likelihood of the two-locus data set described in MATERIALS AND METHODS as a function of *s*, the assumed selective advantage of A−-bearing chromosomes. Three curves are shown, one based on the recombination rates ($c = 0.01675$ and 0.00555 for *L1CAM* and *G0.9MT*, respectively), one based on assuming half those values ($c = 0.008375$ and 0.002775), and one based on assuming twice those rates ($c = 0.0335$ and 0.0111). Although the choice of recombination rate affects the estimated likelihoods, the qualitative results are the same. For all three sets of values, *s* is likely to be at least 0.1. This method does not allow us to place an upper bound on *s*, but on other grounds we can exclude values >0.2, which is roughly the selective advantage of individuals heterozygous for the S allele at the β-globin locus (HB-S) in malarial regions, which is thought to have a higher selection coefficient than G6PD deficiency with respect to malarial protection (ALLISON 1964).

The posterior distribution of the age of A− depends on *s*. Figure 5B shows the distributions for two values of *s*, assuming the estimated recombination rates. For $s = 0.1$, which is the smallest value consistent with the observations, the estimated age is ~100 generations with an upper bound of <150 generations. The posterior distribution depends only slightly on the recombination rates, as shown in Figure 5C. Most of the information about the age of a strongly advantageous allele is contained in the frequency, not in the extent of linkage disequilibrium with nearby marker alleles.

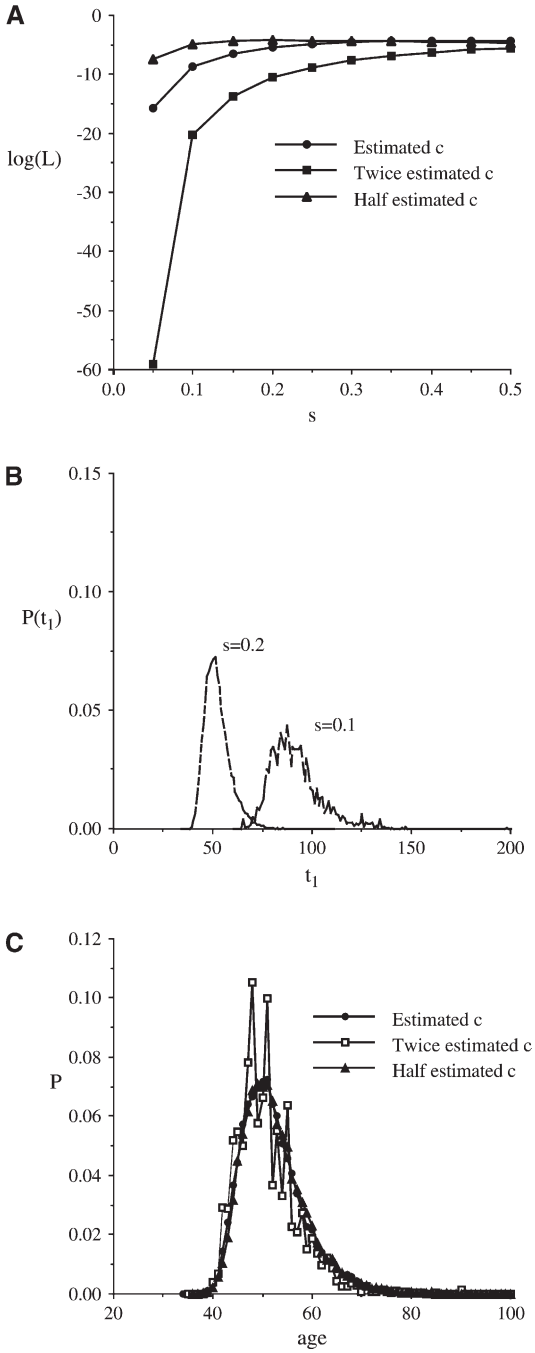


FIGURE 5.—Results of the evolutionary analysis of the G6PD A− allele. Results were obtained by combining the importance sampling method of SLATKIN (2001) for averaging over replicate sample paths with the method of GARNER and SLATKIN (2002) for computing the probability of a configuration of haplotype frequencies at two linked loci. A population frequency for G6PD A− of 0.1 and a constant population size of 10,000 individuals were assumed. The numbers of G6PD A− chromosomes with the two-locus haplotypes at *L1CAM* and *G0.9MT* were 14, 6, 0, and 0. For each point, 90,000 replicate sample paths were generated and 20 replicates of the Garner-Slatkin program were run for each sample path. The estimated recombination rates from *G6PD* were $c = 0.008375$ and 0.002775 M for *L1CAM* and *G0.9MT*, respectively. Other results shown were obtained by doubling and halving those values. (A) Log-likelihood of *s*, the hypothesized selective advantage of heterozygous carriers of the G6PD A− allele. Additive selection was assumed. (B) The posterior distribution of allele age (t_1) for two selection coefficients consistent with the observations. (C) The posterior distribution of allele age (t_1) for $s = 0.2$ for the three sets of recombination rates used.

DISCUSSION

Nucleotide variability around *G6PD*: We investigated patterns of nucleotide variability at *G6PD*, a locus known to be under natural selection by malaria in Africa, and at nine flanking loci at varying distances from *G6PD*. Previous studies have shown that in sub-Saharan Africa levels of nucleotide variability at *G6PD* are typical of those at other X-linked loci, and tests of neutrality based on the frequency spectrum of alleles do not deviate from neutral equilibrium expectations (SABETI *et al.* 2002; SAUNDERS *et al.* 2002; VERRELLI *et al.* 2002). Our data corroborate these results. Nonetheless, we find that the levels of nucleotide variability among *G6PD* A−

alleles are reduced over ~1.5 Mb of the X chromosome, roughly from *L1CAM* to *G0.9MT*, demonstrating the strong impact of selection at *G6PD* on neighboring genes.

Conserved extended haplotype among G6PD A− chromosomes: The effect of selection on *G6PD* is also seen in the extent of LD. Previous studies documented LD between *G6PD* and SNPs within ~600 kb around *G6PD* in Africa (Sabeti *et al.* 2002; Saunders *et al.* 2002). Here we have shown that the ancestral *G6PD* A− extended haplotype spans >1.6 Mb. A genome-wide survey of the half distance of $|D'|$ (the distance at which D' decays to half its maximal value, $|D'| = 0.5$) was ~100 and ~5 kb for a Caucasian and an African population, respectively (Reich *et al.* 2001). Short-range LD in African populations relative to non-African populations is common for most human data sets (*e.g.*, Tishkoff *et al.* 1996; Wall and Pritchard 2003), making the finding of such extensive LD associated with *G6PD* A− in Africa highly unusual.

The pattern of extensive LD seen in these data is consistent with recent strong selection at *G6PD* accompanied by hitchhiking of SNPs that preexisted on the ancestral G6PD A− chromosome. However, patterns of LD may also be created by population admixture and/or underlying population subdivision in a sample. This is a potential concern in this study because 11 of the 20 G6PD A− individuals are African-American. However, 4 of the African-American samples (VA088, VA076, VA025, and M115) have a disrupted G6PD A− ancestral extended haplotype at *IDH3G* (contributing to more than half of the G6PD A− extended haplotype recombinants at this locus). Furthermore, the African-American G6PD A− samples considered alone have levels of nucleotide variability similar to those of the non African-American G6PD A− samples considered alone (data not shown). Along with the detection of a portion of this extended G6PD A− haplotype by Sabeti *et al.* (2002) that included 252 sub-Saharan African (*i.e.*, non-African-American) samples, these results suggest that an overrepresentation of African-American G6PD A− samples is not a major factor contributing to the long-range LD seen here, and that selection is the most likely explanation for the atypical pattern of LD.

The effect of selection on LD has been studied at several other genes. The HLA region exhibits long-range LD in general, and in a non-African panel Sanchez-Mazas *et al.* (2000) detected LD in this region spanning ~1.3 Mb. However, given the local recombination rate in this region, the genetic distance over which LD is found is not significantly higher than the genome average (Walsh *et al.* 2003). The long-range LD found in the HLA region might not be due to physical linkage, but instead may be a result of epistatic interactions that create nonrandom combinations of alleles that are advantageous for immune response (Meyer and Thomson 2001). Significant LD was detected over 20 kb around *FY* (Duffy) in a non-African sample, consistent with recent selection by *P. vivax* (Hamblin *et al.* 2002). At *HB* (β-*globin*), long-range LD was detected among SNPs spanning nearly 100 kb in association with *HB-E* alleles in a Thai population, consistent with selection by malaria (Ohashi *et al.* 2004). And at *LCT*, significant LD has been reported, spanning >800 kb in association with lactase persistence alleles in a European-American population (Bersaglieri *et al.* 2004).

The decay of the G6PD A− extended haplotype (EH) is slightly asymmetrical around the target of selection. The EH decays between 705 kb (at *IDH3G*) and 991 kb (at *BGN*) proximal to *G6PD*, whereas it remains fully conserved among all 20 G6PD A− chromosomes at 925 kb (at *G0.9MT*) distal to *G6PD*. Genetic hitchhiking around a target of selection is not necessarily expected to exhibit a symmetrical pattern, even in the face of homogeneous recombination rates across the affected region (Kim and Stephan 2002). Nonetheless, we note that the extended G6PD A− haplotype spans a region exhibiting heterogeneity in recombination rate. For example, the sex-averaged local recombination rate for the region spanning from *G6PD* to *BGN* is ~2.0 cM/Mb, while the estimated recombination rate between *G6PD* and *G0.9MT* is ~0.6 cM/Mb (UCSC human map viewer based on Kong *et al.* 2002). The relatively low local recombination rate distal to *G6PD* is consistent with the absence of recombinant G6PD A− extended haplotypes in this region. It seems unlikely, however, that the conserved G6PD A− EH extends much further in the distal direction, since *G0.9MT* is adjacent to the q-arm pseudoautosomal region of the X chromosome, where recombination rates are substantially higher.

The observation that the extended G6PD A− haplotype spans >1.6 Mb has interesting implications given that >60 additional genes (including 38 Online Mendelian Inheritance in Man, OMIM, loci) have been identified in this region. In the event that a functional trait (not related to G6PD deficiency) is associated with an ancestral G6PD deficiency EH, this trait could increase in frequency along with the target of selection at *G6PD*. For example, the gene *OPN1MW* (OMIM no. 303800) that is responsible for green color blindness is located within this region (~350 kb proximal to *G6PD*). A study by Filosa *et al.* (1993) demonstrated that in a region of Calabria that bears the Mediterranean-type G6PD deficiency allele ($G6PD_{med}$ coding site 563 C → T), all individuals with the 563 C → T mutation ($n = 7$) were also deutan (green) color-blind on the basis of a visual acuity test. This suggests that in this population a chromosome that carried a $G6PD_{med}$ allele also harbored a mutation causing a clinical condition of deutan color blindness. Presumably, as the $G6PD_{med}$ mutation was favorably selected in this population, the deutan color blindness trait hitchhiked on the EH. In our data, the polymorphism at site 41 in *IDH3G* (Figure 2) causes a nonconservative amino acid change (Arg → Cys) that is found on 14 of the 20 G6PD A− chromosomes and is

rarely found on any other G6PD allelic background. The phenotypic consequences of this polymorphism, if any, are unknown. However, given that this polymorphism is in significant LD with *G6PD* site 202, our data suggest that it may be at its current frequency in Africa due to a hitchhiking event with the G6PD A− allele.

Age of the G6PD A− allele and magnitude of selection: Previous estimates for the age of G6PD A− suggest that the allele is young (<20,000 years) on the basis of closely linked microsatellite variability (Tishkoff *et al.* 2001), coalescent-based analysis of a *G6PD* gene tree (Coop and Griffiths 2004), and intergenic LD (Saunders *et al.* 2002). Our current analysis suggests that the likely age of the G6PD A− allele is ~100 generations with an upper bound of 150 generations given a selection coefficient against the normal (G6PD B) homozygotes of $s \approx 0.1$. This age estimate (2500–3750 years, assuming a 25-year generation time) is somewhat younger than a previous age estimate based on intra-allelic microsatellite variability (3840–11,760 years with $s = 0.044$) (Tishkoff *et al.* 2001). The discrepancy between these two age estimates may be due to the different selection coefficients that were estimated or to uncertainty in microsatellite mutation rate and/or recombination rates in Xq28. Coalescent-based analyses utilizing ~5 kb from *G6PD* provide a relatively old age estimate of >9500 years (Verrelli *et al.* 2002; Coop and Griffiths 2004). Although this analytical method is generally powerful, in this case the structure of the data (*i.e.*, homogeneity among G6PD A− alleles) precludes a precise estimate of the age of the A− allele.

Our likelihood analysis suggests that the selection coefficient for the G6PD A− allele is large; however, it is similar in magnitude to other selection coefficients estimated in humans. For example, selection coefficients of $s = 0.26$, 0.30, and ~0.15 have been proposed for HB-S (Allison 1956), CCR5Δ32 (Schliekelman *et al.* 2001), and some HLA alleles (Satta *et al.* 1994), respectively, for protection from infectious diseases in humans. It is possible that these large values may reflect an ascertainment bias toward recognizing loci under strong selection.

Conclusion: Selection at *G6PD* is strong and recent, consistent with an adaptive response to a recent increase in virulence of *P. falciparum* in sub-Saharan Africa (Ruwende *et al.* 1995; Tishkoff *et al.* 2001; Sabeti *et al.* 2002; Saunders *et al.* 2002). The rapid increase in frequency of G6PD A− has resulted in retention of the ancestral haplotype among the majority of G6PD A− chromosomes spanning >1.6 Mb (~1% of the human X chromosome). This provides a dramatic example of the extent to which recent positive selection can generate long-range LD in human populations. Moreover, if selection is specific to particular geographic regions, as is the case here, it may lead to large differences in patterns of LD for the same genomic region in different populations. This highlights the need for population-specific haplotype maps in association studies.

We thank J. Kim, D. Garrigan, and A. Indap for technical assistance. Some human DNA samples were kindly donated by L. Luzzatto, K. Nafa, Rex Riis, and Jeffrey Ban. R. O. Ryder provided chimpanzee and orangutan samples. B. A. Payseur, E. T. Wood, H. E. Hoekstra, and A. J. Redd provided helpful discussion. D. Begun and two anonymous reviewers provided helpful comments. This material is based on work supported by the National Science Foundation under grant no. 0206756.

LITERATURE CITED

Allison, A. C., 1956 The sickle cell and haemoglobin C genes in some African populations. Ann. Hum. Genet. **21:** 67–89.

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

Allison, A. C., 1964 Polymorphism and natural selection in human populations. Cold Spring Harbor Symp. Quant. Biol. **29:** 137–149.

Ardlie, K. G., L. Kruglyak and M. Seielstad, 2002 Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. **3:** 299–309.

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner *et al.*, 2004 Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. **74:** 1111–1120.

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

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

Coop, G., and R. C. Griffiths, 2004 Ancestral inference on gene trees under selection. Theor. Popul. Biol. **66:** 219–232.

Daly, M. J., J. D. Rioux, S. E. Schaffner, T. J. Hudson and E. S. Lander, 2001 High-resolution haplotype structure in the human genome. Nat. Genet. **29:** 229–232.

Filosa, S., V. Calabro, G. Lania, T. J. Vulliamy, C. Brancati *et al.*, 1993 *G6PD* haplotypes spanning Xq28 from *F8C* to red-green color-vision. Genomics **17:** 6–14.

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

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225–2229.

Garner, C. P., and M. Slatkin, 2002 Likelihood-based disequilibrium mapping for two-marker haplotype data. Theor. Popul. Biol. **61:** 153–161.

Goldstein, D. B., 2001 Islands of linkage disequilibrium. Nat. Genet. **29:** 109–111.

Hamblin, M. T., E. E. Thompson and A. DiRienzo, 2002 Complex signatures of natural selection at the Duffy blood group locus. Am. J. Hum. Genet. **70:** 369–383.

Hammer, M. F., D. Garrigan, E. T. Wood, J. A. Wilder, Z. Mobasher *et al.*, 2004 Heterogeneous patterns of variation among multiple human X-linked loci: the possible role of diversity-reducing selection in non-Africans. Genetics **167:** 1841–1853.

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.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski and F. J. Ayala, 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster.* Genetics **136:** 1329–1340.

Huttley, G. A., M. W. Smith, M. Carrington and S. J. O'Brien, 1999 A scan for linkage disequilibrium across the human genome. Genetics **152:** 1711–1722.

Jeffreys, A. J., L. Kauppi and R. Neumann, 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. **29:** 217–222.

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197–1206.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

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.

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139–144.

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

Livingstone, F. B., 1985 *Frequencies of Hemoglobin Variants: Thalassemia, the Glucose-6-Phosphate Dehydrogenase Deficiency, G6PD Variants and Ovalocytosis in Human Populations.* Oxford University Press, Oxford.

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581–584.

Meyer, D., and G. Thomson, 2001 How selection shapes variation of the human major histocompatibility complex: a review. Ann. Hum. Genet. **65:** 1–26.

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

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.

Ohashi, J., I. Naka, J. Patarapotikul, H. Hananantachai, G. Brittenham *et al.*, 2004 Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. Am. J. Hum. Genet. **74:** 1198–1208.

Phillips, M. S., R. Lawrence, R. Sachidanandam, A. P. Morris, D. J. Balding *et al.*, 2003 Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat. Genet. **33:** 382–387.

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.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer and R. Rozas, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

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.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832–837.

Sanchez-Mazas, A., S. Djoulah, M. Busson, I. L. De Gouville, J. C. Poirier *et al.*, 2000 A linkage disequilibrium map of the MHC region based on the analysis of 14 loci haplotypes in 50 French families. Eur. J. Hum. Genet. **8:** 33–41.

Satta, Y., C. Ohuigin, N. Takahata and J. Klein, 1994 Intensity of natural-selection at the major histocompatibility complex loci. Proc. Natl. Acad. Sci. USA **91:** 7184–7188.

Saunders, M. A., M. F. Hammer and M. W. Nachman, 2002 Nucleotide variability at *G6PD* and the signature of malarial selection in humans. Genetics **162:** 1849–1861.

Schliekelman, P., C. Garner and M. Slatkin, 2001 Natural selection and resistance to HIV. Nature **411:** 545–546.

Slatkin, M., 2001 Simulating genealogies of selected alleles in a population of variable size. Genet. Res. **78:** 49–57.

Slatkin, M., and G. Bertorelle, 2001 The use of intra-allelic variability for testing neutrality and estimating population growth rate. Genetics **158:** 865–874.

Swallow, D. M., 2003 Genetics of lactase persistence and lactose intolerance. Annu. Rev. Genet. **37:** 197–219.

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

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:** 228–231.

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.

Toomajian, C., and M. Kreitman, 2002 Sequence variation and haplotype structure at the human HFE locus. Genetics **161:** 1609–1623.

Toomajian, C., R. S. Ajioka, L. B. Jorde, J. P. Kushner and M. Kreitman, 2003 A method for detecting recent selection in the human genome from allele age estimates. Genetics **165:** 287–297.

Verrelli, B. C., J. H. McDonald, G. Argyropoulos, G. Destrol-Bisol, A. Froment *et al.*, 2002 Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. Am. J. Hum. Genet. **71:** 1112–1128.

Wall, J. D., and J. K. Pritchard, 2003 Haplotype blocks and linkage disequilibrium in the human genome. Nat. Rev. Genet. **4:** 587–597.

Walsh, E. C., K. A. Mather, S. F. Schaffner, L. Farwell, M. J. Daly *et al.*, 2003 An integrated haplotype map of the human major histocompatibility complex. Am. J. Hum. Genet. **73:** 580–590.

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

Xu, W. M., B. Westwood, C. S. Bartsocas, J. J. Malcorraazpiazu, K. Indrak *et al.*, 1995 Glucose-6-phosphate-dehydrogenase mutations and haplotypes in various ethnic groups. Blood **85:** 257–263.

Communicating editor: D. Begun