The span of linkage disequilibrium caused by selection on $G6PD$ in humans.

Matthew A. Saunders$^{1*}$, Montgomery Slatkin$^2$, Chad Garner$^3$, Michael F. Hammer$^1$, and Michael W. Nachman$^1$

$^1$Department of Ecology and Evolutionary Biology
University of Arizona
Tucson, AZ 85721

$^2$Department of Integrative Biology
University of California, Berkeley
Berkeley, CA 94720

$^3$Department of Medicine
University of California, Irvine
Irvine, CA 92697

*Current address: University of Chicago, 1101 E. 57th St., Chicago, IL 60637.

saunders@uchicago.edu
Running head: Nucleotide variability around G6PD

Keywords: G6PD deficiency, linkage disequilibrium, natural selection, Plasmodium falciparum.

Corresponding Author:
Matthew A. Saunders
1101 E. 57th Street
Chicago, IL 60637
Phone: (773) 834-3964
Fax: (773) 702-9740
E-mail: saunders@uchicago.edu
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 non-random associations among SNPs over great physical and genetic distances, even in African populations.
INTRODUCTION

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 forces, 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.
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 \textit{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 many human populations exhibit G6PD deficiency alleles at frequencies that range between 0.05 and 0.20 due to selection (Allison 1960; Motulsky 1961; Ruwenande et al. 1995). G6PD A- is a common deficiency allele in sub-Saharan Africa that reaches frequencies of \textasciitilde 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 \textasciitilde12\% of normal (Hirono and Beutler 1988), and confers \textasciitilde50\% 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 \textit{P. 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 \textit{G6PD} and at loci relatively close to \textit{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 8 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.

SUBJECTS 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, co-occurring 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 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
Saunders 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 are 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. *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 non-coding regions was assembled for each
individual for each locus, using the computer program *SEQUENCHER* (GeneCodes). Sequences have been submitted to GenBank under accession numbers XXXXXX-XXXXXX.

**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* (Verrelli *et al.* 2002; Saunders *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 ($H_d$), $\theta_\pi$ (Nei and Li 1979) and $\theta_w$ (Watterson 1975) at each locus using *dnaSP* 4.0 (Rozas and Rozas 1999). Under neutral equilibrium conditions both $\theta_\pi$ and $\theta_w$ for a random sample estimate the neutral parameter $3N_e \mu$ for X-linked loci, where $N_e$ is the effective population size and $\mu$ is the neutral mutation rate, assuming a sex ratio of one. However, the structure of the sample in the present study is non-random, and therefore $\theta_\pi$ and $\theta_w$ are used simply as measures of nucleotide variability. We also analyzed nucleotide variability in a constructed random sample (CRS: Hudson 1994). This subset of chromosomes ($n=26$) contains G6PD alleles at frequencies that are representative for a typical sub-Saharan African population subject to malarial selection, based on extensive allele frequency surveys (G6PD A-: 0.11; G6PD A+: 0.20; G6PD B: 0.69; Livingstone 1985). We calculated $\theta_\pi$, $\theta_w$, Tajima’s D (Tajima 1989) and Fu and Li’s D (Fu and Li 1993) at each locus for the CRS to compare to non-biased 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 intra-allelic 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 based on long-range haplotypes using the intra-locus combination of sites 55, 59 and
60 at LICAM, SNP 90 at G6PD (i.e. coding site 202), and the intra-locus 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 intra-allelic genealogy of G6PD A- since it arose by mutation. The
intra-allelic 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 intra-allelic genealogy. For each value of $s$,
the selective advantage of A- was considered, 90,000 sample paths and intra-allelic 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 multi-site data set by this method, we used the fact that all 20 A- chromosomes carried the same haplotype for 925 kb telomeric to \textit{G6PD} (to locus \textit{G0.9MT}), and that 14 of 20 chromosomes carried the same haplotype for 556 kb centromeric to \textit{G6PD} (see SNPs 55, 59 and 60 at \textit{L1CAM}; Figure 2). The recombination parameters in the two directions were assumed to be $c = 0.01675$ and $0.00555$ Morgans for \textit{L1CAM} and \textit{G0.9MT} respectively (KONG \textit{et al.} 2002). Therefore, the two-locus data set was 14 \textit{AMB}, 6 \textit{aMB}, 0 \textit{AMb}, 0 \textit{aMb} (in the notation of GARNER and SLATKIN [2002]), where \textit{A} represents the haplotype of rare alleles at \textit{L1CAM} SNPs 55, 59 and 60 (which are very rare on non-A- chromosomes), \textit{M} represents the site under selection, and \textit{B} represents the multilocus haplotype at \textit{G0.9MT} which is rare on non-A-chromosomes.

\section*{RESULTS}

\textbf{Nucleotide diversity at \textit{G6PD}:} Polymorphic sites at \textit{G6PD} are presented in Figure 2. We calculated nucleotide variability for three subset groups: (\textit{i}) individuals bearing the \textit{G6PD} A- allele (n=20), (\textit{ii}) individuals bearing the \textit{G6PD} A+ allele (n=11), (\textit{iii}) individuals bearing the \textit{G6PD} B allele (n=20). At \textit{G6PD} we observed 18 segregating sites (excluding 3 indels) in the entire sample, consistent with previous findings (SAUNDERS \textit{et al.} 2002; VERRELLI \textit{et al.} 2002). Among the \textit{G6PD} A- individuals (n=20) there was no nucleotide variability in 5109 bp of contiguous DNA sequence. Non-coding nucleotide variability among \textit{G6PD} A+ and \textit{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) consistent with previously reported
levels of nucleotide variability at \textit{G6PD} in sub-Saharan Africa (\textit{Sabeti et al.} 2002; \textit{Saunders et al.} 2002; \textit{Verrelli et al.} 2002) and the average of 15 other X-linked loci ($\theta_x = 0.0755\%, \theta_w =0.0815$: \textit{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), also consistent with the previous studies at \textit{G6PD}.

**Nucleotide diversity around \textit{G6PD}**: The segregating sites for 9 loci flanking \textit{G6PD} are presented in Figure 2. At \textit{GAB3}, \textit{F8} and \textit{G0.9MT}, loci distal to \textit{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 non-coding nucleotide variability for the A- group, the B group and the CRS was $\theta_x = 0.027\%$, $0.024\%$, $0.022\%$ 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 \textit{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 \textit{G6PD}, however the pattern is not as extreme as on the distal side, and the pattern decays beyond \textit{L1CAM} (~556 kb from \textit{G6PD}; Figure 3a and 3b). At \textit{TAZ}, \textit{L1CAM} and \textit{IDH3G} the average non-coding nucleotide variability for the A- group, A+ group, B group and the CRS was $\theta_x = 0.032\%$, $0.037\%$, $0.040\%$ and $0.041\%$ respectively (Figure 3a; Table S2). At \textit{BGN}, \textit{G1.5MC} and \textit{G18MC}, loci mapping 0.9 -19 Mb from \textit{G6PD}, the average non-coding nucleotide variability for the A- group, A+ group, B group and the CRS was $\theta_x = 0.093\%$, $0.072\%$, $0.091\%$ and $0.082\%$ respectively (Figure 3a; Table S2). At these three loci the average level of nucleotide variability among the A- individuals is not reduced relative to 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 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 ≥ 5 individuals (Figure 4). Intragenic pairwise comparisons show strong LD within each of the loci surveyed, consistent with a null expectation 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 6 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 3 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 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 SUBJECTS 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 LICAM 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 used $s$ 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 larger than 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 roughly 100
generations with an upper bound of less than 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.
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 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 approximately 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 kb and ~5 kb for a Caucasian and 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, four of the African-American samples (VA088, VA076, VA025 and M115) have a disrupted G6PD A- ancestral extended haplotype at \textit{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 similar levels of nucleotide variability as 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 \textit{Sabeti et al.} (2002) that included 252 sub-Saharan African (\textit{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 \textit{Sanchez-Mazas et al.} (2000) detected LD in this region spanning \(~1.3\text{ Mb.} \) 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 (\textit{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 non-random combinations of alleles that are advantageous for immune response (\textit{Meyer} and \textit{Thomson} 2001). Significant LD was detected over 20 kb around \textit{FY} (Duffy) in a non-African sample, consistent with recent selection by \textit{Plasmodium vivax} (\textit{Hamblin et al.} 2002). At \textit{HB} (\textit{\(\beta\)-globin}), long range LD was detected among SNPs spanning nearly 100 kb in association with \textit{HB-E} alleles in a Thai population, consistent with selection by malaria (\textit{Ohashi et al.} 2004). And at \textit{LCT}, significant LD has been reported spanning \(>800\text{ kb} \) in association with lactase persistence alleles in a European-American population (\textit{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 \textit{IDH3G}) and 991 kb (at \textit{BGN}) proximal to
\textit{G6PD}, whereas it remains fully conserved among all 20 G6PD A- chromosomes at 925 kb (at
\textit{G0.9MT}) distal to \textit{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 \textit{G6PD} to \textit{BGN} is \(~2.0\)
cM/Mb, while the estimated recombination rate between \textit{G6PD} and \textit{G0.9MT} is \(~0.6\) cM/Mb
(UCSC human map viewer based on Kong \textit{et al.} 2002). The relatively low local recombination
rate distal to \textit{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 \textit{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 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 \textit{G6PD}. For example, the gene \textit{OPN1MW} (OMIM #303800) that is responsible for
green color blindness is located within this region (\(~350\) kb proximal to \textit{G6PD}). A study by
Filosa \textit{et al.} (1993) demonstrated that in a region of Calabria that bears the Mediterranean type
G6PD deficiency allele (\textit{G6PD} \(_{med}\) coding site \(563\ C\rightarrow T\)), all individuals with the \(563\ C\rightarrow T\)
mutation \((n=7)\) were also deutan (green) color-blind based on 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 non-conservative amino acid change (Arg$\rightarrow$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) based on 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 roughly 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 uncertainty in microsatellite mutation rate and/or recombination rates in Xq28. Coalescent-based analyses utilizing ~ 5 kb from $G6PD$ provides a relatively old age estimate of > 9,500 years (Coop and Griffiths 2004; Verrelli et al. 2002). 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 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 towards recognizing loci under strong selection.

**Conclusion:** Selection at G6PD is strong and recent, consistent with the idea of 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.
ACKNOWLEDGMENTS

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. Two anonymous reviewers provided helpful comments. This material is based upon work supported by the National Science Foundation under Grant No. 0206756.
LITERATURE CITED


Saunders et al.


Ohashi, J., I. Naka, J. Patarapotikul, H. Hananantachai, G. Brittenham et al., 2004

Phillips, M. S., R. Lawrence, R. Sachidanandam, A. P. Morris, D. J. Balding et al., 2003


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.


Table 1: Individuals sampled in study.

<table>
<thead>
<tr>
<th>G6PD Allele type</th>
<th>Sample</th>
<th>Country</th>
<th>Ethnic/Language Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>G6PD A-</td>
<td>Ivc01</td>
<td>Ivory Coast</td>
<td>Niger-Congo</td>
</tr>
<tr>
<td></td>
<td>Ivc17</td>
<td>Ivory Coast</td>
<td>Niger-Congo</td>
</tr>
<tr>
<td></td>
<td>* S628</td>
<td>Kenya</td>
<td>Unknown</td>
</tr>
<tr>
<td></td>
<td>Alb77</td>
<td>South Africa</td>
<td>Sotho</td>
</tr>
<tr>
<td></td>
<td>Mgr40</td>
<td>Togo</td>
<td>Niger-Congo</td>
</tr>
<tr>
<td></td>
<td>VA010</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>M115</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>M241</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>VA084</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>VA025</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>* VA076</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>VA085</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>VA088</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>DKT338</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>DKT381</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>DKT382</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>JK1031</td>
<td>Zaire</td>
<td>Mbuti Pygmy</td>
</tr>
<tr>
<td></td>
<td>Sho07</td>
<td>Zimbabwe</td>
<td>Shona</td>
</tr>
<tr>
<td></td>
<td>Sho18</td>
<td>Zimbabwe</td>
<td>Shona</td>
</tr>
<tr>
<td>* Sho49</td>
<td>Zimbabwe</td>
<td>Shona</td>
<td></td>
</tr>
<tr>
<td>G6PD A+</td>
<td>* JK785</td>
<td>CAR</td>
<td>Biaka Pygmy</td>
</tr>
<tr>
<td></td>
<td>Ivc22</td>
<td>Ivory Coast</td>
<td>Niger-Congo</td>
</tr>
<tr>
<td></td>
<td>* JK1071</td>
<td>Zaire</td>
<td>Mbuti Pygmy</td>
</tr>
<tr>
<td></td>
<td>AU26</td>
<td>Kenya</td>
<td>Unknown</td>
</tr>
<tr>
<td></td>
<td>* JW058</td>
<td>Mali</td>
<td>Malinke</td>
</tr>
<tr>
<td></td>
<td>* Alb27</td>
<td>South Africa</td>
<td>Zulu</td>
</tr>
<tr>
<td></td>
<td>* VA024</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>DKT275</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>VA012</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>JK1058</td>
<td>Zaire</td>
<td>Mbuti Pygmy</td>
</tr>
<tr>
<td></td>
<td>Sho03</td>
<td>Zimbabwe</td>
<td>Shona</td>
</tr>
<tr>
<td>G6PD B</td>
<td>* JK736</td>
<td>CAR</td>
<td>Biaka Pygmy</td>
</tr>
<tr>
<td></td>
<td>* JK741</td>
<td>CAR</td>
<td>Biaka Pygmy</td>
</tr>
<tr>
<td></td>
<td>* Gna02</td>
<td>Ghana</td>
<td>Akan</td>
</tr>
<tr>
<td></td>
<td>* Ivc04</td>
<td>Ivory Coast</td>
<td>Niger-Congo</td>
</tr>
<tr>
<td></td>
<td>Ivc18</td>
<td>Ivory Coast</td>
<td>Niger-Congo</td>
</tr>
<tr>
<td></td>
<td>* Ivc20</td>
<td>Ivory Coast</td>
<td>Niger-Congo</td>
</tr>
<tr>
<td></td>
<td>* Ivc23</td>
<td>Ivory Coast</td>
<td>Niger-Congo</td>
</tr>
<tr>
<td></td>
<td>* Mka21</td>
<td>Kenya</td>
<td>Maasai</td>
</tr>
<tr>
<td></td>
<td>* Koh188</td>
<td>Kenya</td>
<td>Meru</td>
</tr>
<tr>
<td></td>
<td>Individuals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-------------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td></td>
<td>* Mka29</td>
<td>Kenya</td>
<td>Tsumkwe</td>
</tr>
<tr>
<td></td>
<td>* JR013</td>
<td>Namibia</td>
<td>Biaka</td>
</tr>
<tr>
<td></td>
<td>JR323</td>
<td>Namibia</td>
<td>Bagandu</td>
</tr>
<tr>
<td></td>
<td>* LD156</td>
<td>South Africa</td>
<td>Khoisan</td>
</tr>
<tr>
<td></td>
<td>* Alb74</td>
<td>South Africa</td>
<td>Unknown</td>
</tr>
<tr>
<td></td>
<td>* DKT331</td>
<td>USA</td>
<td>African American</td>
</tr>
<tr>
<td></td>
<td>* JK1029</td>
<td>Zaire</td>
<td>Mbuti</td>
</tr>
<tr>
<td></td>
<td>* JK1033</td>
<td>Zaire</td>
<td>Ituri</td>
</tr>
<tr>
<td></td>
<td>* Sho14</td>
<td>Zimbabwe</td>
<td>Shona</td>
</tr>
<tr>
<td></td>
<td>* Sho30</td>
<td>Zimbabwe</td>
<td>Shona</td>
</tr>
<tr>
<td></td>
<td>* Sho46</td>
<td>Zimbabwe</td>
<td>Shona</td>
</tr>
</tbody>
</table>

Individuals used for constructed random sample (CRS) analyses are marked with (*).
FIGURE LEGENDS

**Figure 1:** Schematic 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 orientation 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 3 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.

**Figure 2:** Table of polymorphism for $G6PD$ and surrounding loci. Fifty one unrelated human males of sub-Saharan African decent were surveyed for nucleotide variability at $G6PD$ and 9 surrounding loci. Individual samples were selected based on *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 bi-allelic marker (*i.e.* SNP or indel). Segregating sites are listed in numerical order. For exact alignment positions (in bp) of segregating sites and the identity of the polymorphic nucleotides (*i.e.* A, G, C or T) see our website (http://eebweb.arizona.edu/faculty/nachman/saunders/pubs/G6PD_AFR/sites_table_AFR.pdf). 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 grey 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 a white triangle including the size of the indel in bp. 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 bp.

**Figure 3:** Nucleotide variability for *G6PD* and 9 surrounding loci for subset groups of the dataset (G6PD A- alleles, G6PD A+ alleles, G6PD B alleles and CRS): (a) nucleotide diversity (θπ), (b) haplotype diversity.

**Figure 4:** Patterns of LD between segregating sites at *G6PD* and flanking loci. The table of polymorphism only includes segregating sites at which the less common allele (minor allele) is found in ≥ 5 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 grey 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 a circle.
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 number 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 Morgans 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 (t1) for two selection coefficients consistent with the observations. (c) The posterior distribution of allele age (t1) for s = 0.2 for the three sets of recombination rates used.
Figure 1:
Figure 2:
Figure 3:

(a) Haplotype diversity

(b) Locus distribution

Locus:
- G18MC
- G1.5MC
- BGN
- IDH3G
- L1CAM
- TAZ
- G6PD
- GAB3
- F8
- G0.9MT

Haplotype diversity:
- G6PD A-
- G6PD A+
- G6PD B
- CRS
Figure 5

A.

\[ \log(L) \]

- Estimated c
- Twice estimated c
- Half estimated c

B.

\[ P(t_1) \]

s=0.2
s=0.1

C.

\[ P \]

age