NOTE TO USERS

PATTERNS OF NUCLEOTIDE VARIAIBILITY WITHIN AND AROUND *G6PD*, A LOCUS UNDER POSITIVE NATURAL SELECTION IN HUMANS

By

Matthew Allan Saunders

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ECOLOGY AND EVOLUTIONARY BIOLOGY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2004

The University of Arizona ®
Graduate College

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by Matthew Allan Saunders

entitled PATTERNS OF NUCLEOTIDE VARIABILITY WITHIN AND AROUND G6PD, A LOCUS UNDER POSITIVE NATURAL SELECTION IN HUMANS.

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

MICHAEL W. NACHMAN	11/18/04 date
MICHAEL F. HAMMER	11/18/04 date
CARLOS C. CAMPBELL	18 December 2004 date
J. BRUCE WALSH	date
	11/18/04 date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

12/5/04
Dissertation Director: date
MICHAEL W. NACHMAN

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library. Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: 

ACKNOWLEDGMENTS

I am indebted to many people who have made great contributions to my efforts of completing this dissertation. First and foremost, I would like to extend the deepest gratitude to my advisor, Michael Nachman. Michael invited me into his lab at an early stage in my scientific development, before I was fully aware of the fascinating field of population genetics. Under his guidance I came to recognize that my calling lay in the study of human population genetics. Specifically, Michael first introduced me to the G6PD-malaria system that has become a major component of my life in recent years. Michael generously invested countless hours into my academic development, and under his guidance I gained invaluable insight into the evolutionary process in general. Michael's energy and excitement about science is contagious, and as I attempted to match his enthusiasm I have accomplished more than I ever imagined. In general under Michael's mentorship I learned the art of scientific design, critical thinking, and both verbal and written scientific communication. I am truly fortunate to have had Michael as a teacher and an advisor.

I feel privileged to have had the support of my committee members, who have also made invaluable contributions to my development. I especially thank Mike Hammer, who welcomed me into his lab group as one of his own students, and he generously donated both technical and intellectual resources. Without his guidance and expertise in human evolutionary biology, this project would not have been made possible. Mike's general perspectives on science and academia have always been helpful. Bruce Walsh has been an incredible resource and advisor on matters of statistics. Also, Bruce's friendship and encouragement over the years has been comforting, and made my time in department all the more pleasurable. I was delighted to have Kent Campbell on my committee. Despite Kent's peripheral interest in population genetics he never failed to amaze me with his level of investment in my work and development. Kent was an extraordinary source of knowledge about malaria and encouragement in general. Last but not least, I am grateful to Teri Markow who provided me with insight into the evolution in general, and was extremely supportive of me over the years.

Many fellow graduate students and postdocs have provided a crucial support-system both on personal and intellectual levels. I especially thank Bret Payseur, Hopi Hoekstra, Elizabeth Wood, Jeff Good and Gabriella Wlasiuk for their friendship and intellectual stimulation. Asher Cutter, Maya Metni, Dan Garrigan, Matt Kaplan, Allen Redd, and Tanya Karafet and members of the Nachman and Hammer Labs were always sources for provocative discussions and ideas. Several undergraduate students have worked with me over the years and assisted me in generating data: Jeffery Jensen, Veronique Klein, Julia Kim and Emily Landeen. I can only hope that they learned from me nearly as much as I have learned from them. Jim Krenz, Brian Coulihan and Ryan Sprissler provided incredible technical support. Finally, I would like to thank the staff of the EEB department for administrative support, especially Sue Whitworth and Kate Riley.

This list would be incomplete without mention of my two beloved dogs, Echo and Tyson, who provided companionship and brightened up my life.

DEDICATION

To my mother, Judy, and my grandfather, Henry, who inspired and encouraged my interest in science.

TABLE OF CONTENTS

ABSTRACT

The role of positive natural selection in shaping patterns of nucleotide variability in the human genome remains unclear. Only several loci in humans have been identified with alleles under selection for which we have a good understanding of the link between genotype and phenotype. These loci provide an opportunity to describe the general footprint of natural selection in the human genome. Some mutations at the gene coding for glucose-6-phosphate dehydrogenase (G6PD) cause a clinical condition of G6PD deficiency, however these alleles have also been shown to confer resistance to the parasite *Plasmodium falciparum*. Evidence for resistance to malaria comes from geographical allele distributions that coincide with malaria, epidemiology studies, and *in vitro* studies. A detailed study of nucleotide variability at *G6PD* can shed light on the signature of selection in humans and can also provide insight into the natural history of the association between malaria and humans. In Appendix A, patterns of nucleotide variability are described in a worldwide panel of humans at *G6PD* and at a neighboring locus (*L1CAM*). Patterns of nucleotide variability at *G6PD* do not significantly differ from patterns found at other loci. Nonetheless, significant long-range linkage disequilibrium (LD) associated with a selected *G6PD* deficiency allele from Africa (G6PD A-) is observed between *G6PD* and *L1CAM*. The observed LD implies that G6PD A- is young, and suggests that malaria is a recent agent of selection in humans (within the past 10,000 years). In Appendix B, patterns of nucleotide variability are examined in a larger panel of individuals of sub-Saharan African decent at *G6PD* and at

nine loci located in a region spanning ~ 1Mb on either of *G6PD*. LD and reduced nucleotide variability are observed over a region of > 1 Mb around *G6PD*. These observations confirm a young age for G6PD A- and imply that the allele has been under strong selection ($0.05 < s < 0.20$). In Appendix C, a comparable survey of nucleotide variability was conducted in relation to $G6PD_{med}$, an independently arisen selected G6PD deficiency allele, in a panel of individuals from the Middle East and the Mediterranean region. Patterns of LD provide strong evidence for independent origins of $G6PD_{med}$ in two geographic regions of Eurasia. Furthermore, these data suggest that resistance to malaria by *G6PD* mutations arose at approximately the same time in Africa and Eurasia. In Appendix D, a sample from a human population isolate, the Kurdish Jews, was studied at *G6PD*, at neighboring loci, and at 2 unlinked loci to determine if the remarkably high frequency of G6PD deficiency in this population is attributable primarily to selection or to a severe founder effect. A pattern of nucleotide variability that is consistent with selection is observed around *G6PD* among Kurdish Jews, while patterns of nucleotide variability at unlinked loci are typical of other populations that have not experienced severe founder effects. These results suggest that natural selection is largely responsible for the high frequency of G6PD deficiency among Kurdish Jews.

CHAPTER 1: INTRODUCTION

Explanation of the problem and its context

Mutations that cause phenotypic changes will potentially affect fitness. A majority of mutations are deleterious, some are neutral, and a negligible proportion of mutations are advantageous (Kimura 1983), yet this latter class often provides the most interesting displays of the evolutionary process. Even so, only a few phenotypically advantageous traits are understood at the molecular level, and few molecular changes that are putatively advantageous are clearly understood at the phenotype level. Identifying cases of recent natural selection in humans and understanding the connections between genotype and phenotype associations can teach us about the evolutionary process in general and about the natural history of our species in particular.

Evolutionary theory presents several models with predictions about the signature of selection at the molecular level (Maynard Smith and Haigh 1974; Hudson et al. 1987; Tajima 1989; McDonald and Kreitman 1991; Fu and Li 1993), and these patterns have been confirmed empirically in model organisms such as *Drosophila* (*e.g.* Kreitman and Hudson 1991; Eanes et al. 1993; Hudson et al. 1994). However, demographic effects in humans, such as small effective population sizes (that reduce genome wide levels of nucleotide variability) and population expansions (that skew the frequency spectrum of alleles) may mask signatures of recent selection. Therefore, even in light of increased availability of genomic data for humans, researchers have met with only limited success in identifying signatures of selection at the molecular level in the human genome. Recent

findings (including the present study) have revealed that the molecular signature of selection in humans is most often detectable by patterns of linkage disequilibrium, rather than by traditional tests of neutrality that are based on the frequency spectrum of alleles (Sabeti et al. 2002; Saunders et al. 2002; Toomajian et al. 2003).

This dissertation describes in detail the patterns of nucleotide variability that are associated with *G6PD*, a locus that is well-understood at the genotypic and phenotypic levels, and a locus that is known to be subject to natural selection. A clear understanding of the signature of selection at the molecular level in humans may help to identify other loci subject to selection. Patterns of nucleotide variability are described in different human populations that have been subject to different demographic effects and that bear independently arisen G6PD deficiency alleles under selection. Comparisons between these samples allow us to distinguish between the deterministic signatures of selection and stochastic patterns associated with drift and demography. The results show that the effect of selection extends over a similarly long distance (> 1.4 Mb) on the X chromosome for the independently arisen selected alleles ,implying that the different alleles are approximately the same age (< 3000 years), and have been strongly selected.

A review of the literature

Natural selection at the molecular level:

Positive natural selection will increase the frequency of an advantageous allele in a population. Importantly, the effect of selection will extend to linked neutral polymorphisms that are associated with the target of selection (*i.e.* a hitchhiking effect:

Maynard Smith and Haigh 1974). In the case of directional selection, an advantageous allele will rise in frequency quickly (*i.e.* selective sweep), causing a transient reduction in nucleotide variability surrounding the target of selection. As new mutations begin to arise in this population following a selective sweep, a sample will exhibit an excess of rare polymorphisms relative to neutral equilibrium expectations (Tajima 1989; Fu and Li 1993). Eventually, the population will return to a normal equilibrium level of neutral nucleotide variability, as mutation and genetic drift will govern the frequency of neutral polymorphisms. Although long-term balancing selection will increase levels of nucleotide variability surrounding a target of selection (Hey 1991; Navarro and Barton 2002), the effects of recent balancing selection are similar to the effects of an incomplete selective sweep by exhibiting a reduction in nucleotide variability among the selected alleles (*i.e.* an allele-specific effect). A consequence of the homogeneity in nucleotide variability among the selected alleles is that LD will be observed between the selected alleles and other alleles in a sample (Strobeck 1983). The distance over which the effects of selection are observed is directly correlated with the strength of selection and inversely correlated with the local recombination rate and the age of the selected allele (Slatkin and Rannala 2000; Garner and Slatkin 2003).

Demographic processes also have profound effects on patterns of nucleotide variability in a manner that may resemble signatures of selection at a given locus. For example, a population expansion will result in a transient excess of rare polymorphisms, similar to the effect of a selective sweep. Alternatively, a population bottleneck may increase intermediate frequency polymorphisms and create linkage disequilibrium,

similar to the effect of balancing selection. It follows that detecting a signature of selection on the background of demographic effects is challenging. In humans, generally low levels of neutral nucleotide variability due to a small effective population size, and a skew in the frequency spectrum of alleles due to recent population expansions result in low statistical power to detect the reductions in nucleotide variability and skewed frequency spectra of alleles that are caused by recent selection (Wall and Przeworski 2000). Thus traditional tests for selection based on levels of variability or the frequency spectrum of alleles have often been unable to detect selection humans (Harding et al. 1997; Hamblin and DiRienzo 2000; Sabeti et al. 2002; Saunders et al. 2002; Verrelli et al. 2002). However, patterns of long-range LD have proven to be useful to detect the effects of selection in humans (Sabeti et al. 2002; Saunders et al. 2002; Toomajian et al. 2003; Ohashi et al. 2004).

Malaria as an agent of selection:

Malaria is a major morbidity factor among humans as over 2 million people die annually from complications of the disease worldwide, and 80% of these deaths are in sub-Saharan Africa (Breman et al. 2004). The disease is caused by a protozoan parasite, *Plasmodium spp.*, which develops asexually in the erythrocytes of the human host. Four different species infect humans: *P. ovale*, *P. malariae*, *P. vivax* and *P. falciparum*. However *P. falciparum* is the most lethal of the species by virtue of its unique pathologic effects, accounting for ~ 90% of deaths by malaria (Schmidt and Roberts 1996; Miller et al. 2002). The parasite is transmitted to the human host by insect vectors, mosquitoes from

the genus *Anophelene* (Schmidt and Roberts 1996), and therefore malaria is most prevalent under conditions that promote mosquito abundance and contact with humans. Mostly children, the elderly, and individuals with impaired immune response succumb to the lethal manifestations of *Plasmodium* infection (*e.g.* cerebral malaria and renal failure) (Snow et al. 1999). Survivors of repeat infections by the parasite typically develop an acquired immunity that inhibits lethal hyper-parasitemia levels (Baird 1995). In areas where malaria is holo-endemic the disease imposes a strong selective force that acts mostly in the pre-reproductive years (Allison 1964). Due to natural selection, human populations historically exposed to malaria often exhibit high frequencies of genetic factors that confer some resistance to the parasite (*e.g.* Duffy factor, HbS and HbC alleles, G6PD deficiency: Miller 1994).

G6PD and positive natural selection:

Glucose-6-phosphate dehydrogenase (G6PD) is a housekeeping enzyme that is expressed in all tissues (Luzzatto and Battistuzzi 1985). G6PD catalyses an initial step in the hexose-monophosphate shunt of glycolysis in which glucose-6-phosphate is oxidized into 6-phosphogluconolactone, concomitantly reducing NADP to NADPH. In erythrocytes, G6PD is the sole source of NADPH which is required for oxidation in many biochemical processes including the neutralization of peroxide *via* reactions with glutathione disulfide and catalase (Kirkman et al. 1987; Gaetani et al. 1989). A reduction in the activity of G6PD may result in relatively high oxidative stress inside erythrocytes, which in turn may destabilize the integrity of the cell membrane. So, when the action of G6PD is

depressed, an individual may suffer from clinical manifestations of G6PD deficiency that include hemolytic anemia and neonatal jaundice (Beutler 1994). Many rare mutations have been described in the coding region of *G6PD* that cause G6PD deficiency (Notaro et al. 2000), however some populations exhibit particular G6PD deficiency alleles at relatively high frequencies ($0.05 < q < 0.65$: Livingstone 1985). Specifically, two G6PD deficiency alleles are commonly found at high frequencies: G6PD A- and $G6PD_{med}$. G6PD A- is found primarily in Africa, and the allele is defined by a mutation at coding site 202 (G202A) that reduces enzyme efficiency to 12% of normal (Hirono and Beutler 1988). $G6PD_{med}$ is found throughout Eurasia and the Mediterranean region, and is defined by a mutation at coding site 563 (C563T) that reduces enzyme efficiency to 5% of normal (Vulliamy et al. 1988).

Three lines of evidence indicate that G6PD deficiency alleles are subject to selection by conferring resistance to severe complications of malaria. First, the geographic distribution of populations with high frequencies of G6PD deficiency alleles strongly correlates with the distribution of historically endemic malaria (Allison 1960; Motulsky 1961). Second, *in vitro* studies demonstrated that normal parasite growth is hindered probably by virtue of the relatively high levels of peroxide found in erythrocytes due to G6PD deficiency (Roth et al. 1983). And finally, a large scale epidemiology study showed a significant under-representation of G6PD deficient individuals among severely affected malaria patients relative to the frequency of G6PD deficiency in the general population (Ruwende et al. 1995).

Recently, studies of nucleotide variability associated with G6PD deficiency alleles have shown patterns that are consistent with recent natural selection acting at the molecular level based on low levels of intra-allelic linked microsatellite variability (Tishkoff et al. 2001), an excess of nonsynonymous polymorphisms in a sample relative to neutral expectation (Verrelli et al. 2002), and unusually long range LD associated with the target of selection (Sabeti et al. 2002; Saunders et al. 2002).

In summary, *G6PD* presents some of the best-understood polymorphisms that are subject to selection in humans both at the phenotype and the genotype levels, making it an excellent model for describing patterns of nucleotide variability that are associated with the action of positive selection.

Explanation of dissertation format

The finding of different G6PD deficiency alleles that are subject to the same agent of selection, falciparum malaria, provides a unique opportunity to distinguish between patterns of nucleotide variability that are due to deterministic forces (*i.e.* selection), and patterns due to the stochastic effects of drift and demography. Appendix A describes general patterns of nucleotide variability at *G6PD* in a worldwide panel of individuals that is well-characterized at other (neutral) loci (Hammer et al. 2004). These results demonstrate that traditional single-locus measures of nucleotide variability do not show significant deviation from neutrality at *G6PD*. However, single nucleotide polymorphisms (SNPs) at a neighboring locus (*L1CAM*) exhibit a remarkable amount of LD with the target of selection of G6PD A-, suggesting that a signature of selection is

found in patterns of long-range LD. Appendix B expands upon these findings to delimit the range of LD surrounding the target of selection of G6PD A- in Africa. Ten loci were resequenced, and polymorphic sites were used to show that the ancestral long-range haplotype of G6PD A- is conserved over a region that spans > 1.6 Mb, implying that the age of the allele is < 3000 years old and bears a large selection coefficient ($0.05 < s < 0.20$). In appendix C, the extent of LD surrounding the target of selection of $G6PD_{med}$ is delimited in a Eurasian panel. An ancestral $G6PD_{med}$ long-range haplotype is significantly conserved over a distance of > 1.6 Mb, comparable to the pattern in Africa, suggesting that the alleles are of similar age. Remarkably, some $G6PD_{med}$ alleles from the Indian sub-continent have a unique long-range haplotype, implying an independent origin of the C563T mutation in India. Appendix D, describes nucleotide variability in Kurdish Jews, which exhibit the highest frequency of G6PD deficiency alleles among all human populations ($G6PD_{med}$, $q = 0.65$). By examining patterns of nucleotide variability at loci which are linked and unlinked to *G6PD*, the unusual spectrum of alleles found in this population at *G6PD* is shown to be a locus-specific pattern that is consistent with a signature of natural selection. This implies that the high frequency of G6PD deficiency in Kurdish Jews is attributable largely to natural selection, rather than to a severe founder effect.

CHAPTER 2: PRESENT STUDY

Detailed background, methods, results and conclusions of this study are presented in the papers appended to this dissertation. The following is a summary of the most important findings of these papers.

In an effort to provide a detailed description of patterns of nucleotide variability that are associated with the effect of positive Darwinian selection in humans I have surveyed nucleotide variability within and around *G6PD* in different populations that include independently arisen alleles under selection. In appendix A, I used a worldwide panel including 47 individuals from Africa, Europe, the Americas and Asia, to survey nucleotide variability at *G6PD*. I detected no significant deviation from neutral equilibrium expectations within a resequenced region of *G6PD* using traditional statistical tests of neutrality. However, a signature of selection is apparent by the absence of variability among G6PD A- alleles (n=6) and the extent of linkage disequilibrium (LD) between the target of selection and SNPs at a locus 556 kb away (*L1CAM*). Together these data were used to estimate that the G6PD A- allele arose within the last 10,000 years, suggesting that malaria has become a strong selective agent in humans relatively recently. To delimit the size of the genomic region affected by selection on *G6PD*, I used a larger pan-African panel of individuals (including 20 G6PD A- alleles, 11 G6PD A+ alleles and 20 G6PD B alleles), to resequence data from *G6PD* and ten surrounding loci in region Xq28 (Appendix B). Homogeneity among the G6PD A- individuals extends > 1Mb around *G6PD*, and LD that is associated with the target of selection extends > 1.6

Mb around *G6PD*. This extent of LD was used to refine the age estimate of G6PD A- to ~3000 years ago with a selection coefficient of $0.05 < s < 0.2$. To describe the effects of selection associated with $G6PD_{med}$, in appendix C, I surveyed Xq28 by resequencing and genotyping SNPs in a panel of individuals from the Middle East and the Mediterranean (21 $G6PD_{med}$ alleles and 23 G6PD B alleles). This study yielded two primary results. First, two distinct long-range haplotypes are associated with $G6PD_{med}$ alleles, supporting the hypothesis that the C563T mutation arose independently once within the Indian sub-continent and once in the Middle East/Mediterranean. Second, the effects of selection among $G6PD_{med}$ alleles extend over a similar genomic region as seen for the G6PD A- allele in Africa, suggesting that ages of the different alleles are roughly the same. Finally, in appendix D, I studied nucleotide variability in a sample of Kurdish Jews (n=37) to test the hypothesis that the high frequency of G6PD deficiency in this population ($q = 0.70$) is due primarily to a severe population bottleneck. In this panel I resequenced and genotyped SNPs within and around *G6PD* and at two unlinked loci, to distinguish between selection and demography as causes for observed patterns of variation. Loci unlinked to *G6PD* provide no evidence of an extreme founder effect, and a signature of selection that is reminiscent of those found in appendices A, B and C is found around *G6PD* for Kurdish Jews. Together these results imply that natural selection had a primary role in increasing the frequency of $G6PD_{med}$ to its current level in this population.

REFERENCES

Allison AC (1960) Glucose-6-phosphate dehydrogenase deficiency in red blood cells of East Africans. Nature 186:531

Allison AC (1964) Polymorphism and natural selection in human populations. Cold Spring Harbor Symposium on Quantitative Biology 29:137-149

Baird JK (1995) Host age as a determinant of naturally acquired immunity to *Plasmodium falciparum*. Parasitology Today 11:105-111

Beutler E (1994) G6PD deficiency. Blood 84:3613-3636

Breman JG, Alilio MS, Mills A (2004) Conquering the intolerable burden of malaria: What's new, what's needed: A summary. American Journal of Tropical Medicine and Hygiene 71:1-15

Charlesworth B (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. Genetical Research 63:213-227

Eanes WF, Kirchner M, Yoon J (1993) Evidence for adaptive evolution of the G6pd gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages. Proceedings of the National Academy of Sciences of the United States of America 90:7475-7479

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693-709

Gaetani GF, Galiano S, Canepa L, Ferraris AM, Kirkman HN (1989) Catalase and glutathione peroxidase are equally active in detoxification of hydrogen peroxide in human erythrocytes. Blood 73:334-339

Garner C, Slatkin M (2003) On selecting markers for association studies: Patterns of linkage disequilibrium between two and three di-allelic loci. Genetic Epidemiology 24:57-67

Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus. American Journal of Human Genetics 66:1669-1679

Hammer MF, Garrigan D, Wood E, Wilder JA, Mobasher Z, Bigham A, Krenz JG, Nachman MW (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

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. American Journal of Human Genetics 60:772-789

Hey J (1991) The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. Genetics 128:831-840

Hirono A, Beutler E (1988) Molecular cloning and nucleotide sequence of cDNA for human glucose-6-phosphate dehydrogenase variant A(-). Proceedings of the National Academy of Sciences of the United States of America 85:3951-3954

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994) Evidence for positive selection in the superoxide-dismutase (*sod*) region of *Drosophila melanogaster*. Genetics 136:1329-1340

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University press

Kirkman HN, Galiano S, Gaetani GF (1987) The function of catalase bound NADPH. Journal of Biological Chemistry 262:660-666

Kreitman M, Hudson RR (1991) Inferring the evolutionary histories of the *Adh* and *Adh-Dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics 127:565-582

Livingstone FB (1985) Frequencies of hemoglobin variants: Thalassemia, the glucose-6-phosphate dehydrogenase deficiency, G6PD variants and ovalocytosis in human populations. Oxford University Press. New York, USA.

Luzatto L, Battistuzzi G (1985) Glucose-6-phosphate dehydrogenase. Advances in Human Genetics 14:217-329

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genetical Research 23:23-35

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351:652-654

Miller LH (1994) Impact of malaria on genetic polymorphism and genetic diseases in Africans and African-Americans. Proceedings of the National Academy of Sciences of the United States of America 91:2415-2419

Miller LH, Baruch DI, Marsh K, Doumbo OK (2002) The pathogenic basis of malaria. Nature 415:673-679

Motulsky AG (1961) Glucose-6-phosphate dehydrogenase haemolytic disease of the newborn, and malaria. Lancet 1:1168

Navarro A, Barton NH (2002) The effects of multilocus balancing selection on neutral variability. Genetics 161:849-863

Notaro R, Afolayan A, Luzzatto L (2000) Human mutations in glucose 6-phosphate dehydrogenase reflect evolutionary history. FASEB 14:485-494

Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Brittenham G, Looareesuwan S, Clark AG, Tokunaga K (2004) Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. American Journal of Human Genetics 74:1198-1208

Roth EF, Raventossuarez C, Rinaldi A, Nagel RL (1983) Heterozygous and hemizygous red-cell G6PD deficiency inhibit *in vitro* growth of *falciparum*-malaria to the same extent. Clinical Research 31:A322-A322

Ruwende C, Khoo SC, Snow AW, Yates SNR, Kwiatkowski D, Gupta S, Warn P, Allsopp CEM, Gilbert SC, Peschu N, Newbold CI, Greenwood BM, Marsh K, Hill AVS (1995) Natural selection of hemizygotes and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature 376:246-249

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

Saunders MA, Hammer MF, Nachman MW (2002) Nucleotide variability at *G6PD* and the signature of malarial selection in humans. Genetics 162:1849-1861

Schmidt GD, Roberts LS (1996) Foundations of paristiology. Wm. C. Brown Publishers, Chicago

Slatkin M, Rannala B (2000) Estimating allele age. Annual Review of Genomics and Human Genetics 1:225-249

Snow RW, Craig MH, Deichmann U, Le Sueur D (1999) A preliminary continental risk map for malaria mortality among African children. Parasitology Today 15:99-104

Strobeck C (1983) Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics 103:545-555

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human *G6PD*: Recent origin of alleles that confer malarial resistance. Science 293:455-462

Toomajian C, Ajioka RS, Jorde LB, Kushner JP, Kreitman M (2003) A method for detecting recent selection in the human genome from allele age estimates. Genetics 165:287-297

Verrelli BC, McDonald JH, Argyropoulos G, Destrol-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA (2002) Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. American Journal of Human Genetics 71:1112-1128

Vulliamy TJ, Durso M, Battistuzzi G, Estrada M, Foulkes NS, Martini G, Calabro V, Poggi V, Giordano R, Town M, Luzatto L, Persico MG (1988) Diverse point mutations in the human glucose-6-phosphate dehydrogenase gene cause enzyme deficiency and mild or severe hemolytic anemia. Proceedings of the National Academy of Sciences of the United States of America 85:5171-5175

Wall JD, Przeworski M (2000) When did the human population size start increasing? Genetics 155:1865-1874

APPENDIX A: NUCLEOTIDE VARIABILITY AT *G6PD* AND THE SIGNATURE OF MALARIAL SELECTION IN HUMANS

Published: *Genetics* (2002) 162:1849-1861

GENETICS

A periodical record of investigations bearing on heredity and variation

Vol. 162 No. 4 December 2002

Genetics Society of America

PERSPECTIVES

INVESTIGATIONS

Continued on back cover

Continued from front cover

Continued on inside back cover

November 1, 2004

Dear *GENETICS* Editor,

I am a PhD graduate student and I would like to include my publication in *GENETICS* into my dissertation. I hereby request a release letter to use the copyrighted material listed below as part of my dissertation:

Saunders M. A., Hammer, M. F. and Nachman, M. W. (2002) *Nucleotide Variability at G6PD and the Signature of Malarial Selection in Humans*. Genetics **162**: 1849-1861.

This requested letter is for the use of University Microfilms Incorporated (UMI), to acknowledge permission which would extend to publication and/or microfilming by UMI. UMI may sell, on demand, single copies of the dissertation, including the copyrighted material, for scholarly purposes.

Please contact me at the following address or telephone number if you have any questions.

Matthew A. Saunders
Department of Ecology and Evolutionary Biology
Biological Sciences West bldg., Room 333
University of Arizona
Tucson, AZ 85721

Tel: 520-626-4747
Fax: 520-621-9190
E-mail: msaunder@u.arizona.edu

Your prompt reply would be greatly appreciated.
Thank you.

Sincerely,

Matthew Saunders

Genetics Society of America

The Genetics Society of America is organized to provide for association and conference among students of heredity and to encourage close relationships between workers in genetics and those in related sciences. GENETICS is its official journal. All persons interested in genetics are eligible for active membership. Application for membership may be made on the form in this journal or obtainable from the address below.

GENETICS (ISSN 0016-6731) is published monthly by the Genetics Society of America, 9650 Rockville Pike, Bethesda, Maryland 20814-3998. The 2002 subscription price for nonmembers (including postage) is $620 per year domestic, $640 outside of the USA.

Correspondence about membership, nonmember subscriptions, changes of address, missing issues, and other nonediterial matters should be directed to the appropriate address below, according to whether the inquirer is (or would be) a GSA member or nonmember subscriber. Claims for missing issues should be made within 30 days of the date of mailing, which is generally on or near the first of the month; missing issues will be supplied free only if they have been lost in the mail.

GSA MEMBERS ONLY:
GSA Administrative Office
9650 Rockville Pike
Bethesda, Maryland 20814-3998
(301) 571-1825

NONMEMBER SUBSCRIPTIONS:
FASEB Subscription Department
9650 Rockville Pike
Bethesda, Maryland 20814-3998
(301) 530-7026

Back issues of GENETICS from 1992 (Volume 130) can be obtained through the nonmember subscriber address above. GENETICS is also available in microform from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106. European inquiries may be directed to 30/32 Mortimer Street, Dept. PR, London, W1N 7RA, England.

POSTMASTER: Send address changes to GENETICS, 9650 Rockville Pike, Bethesda, MD 20814-3998. Periodicals postage paid at Bethesda, Maryland 20814-3998 and additional mailing offices. (Publication No. 215860.)

Printed by Capital City Press, Inc.

The Genetics Society of America wishes to thank Carnegie Mellon University and the Department of Biological Sciences for providing space and amenities for the GENETICS Editorial Office.

ABSTRACT

Glucose-6-phosphate dehydrogenase (G6PD) deficiency is the most common enzymopathy in humans. Deficiency alleles for this X-linked disorder are geographically correlated with historical patterns of malaria, and the most common deficiency allele in Africa (G6PD A-) has been shown to confer some resistance to malaria in both hemizygous males and heterozygous females. We studied DNA sequence variation in 5.1 kb of *G6PD* from 47 individuals representing a worldwide sample to examine the impact of selection on patterns of human nucleotide diversity and to infer the evolutionary history of the G6PD A- allele. We also sequenced 3.7 kb of a neighboring locus, *L1CAM*, from the same set of individuals to study the effect of selection on patterns of linkage disequilibrium. Despite strong clinical evidence for malarial selection maintaining G6PD deficiency alleles in human populations, the overall level of nucleotide heterozygosity at *G6PD* is typical of other genes on the X-chromosome. However, the signature of selection is evident in the absence of genetic variation among A- alleles from different parts of Africa and in the unusually high levels of linkage disequilibrium over a considerable distance of the X-chromosome. In spite of a long-term association between *Plasmodium falciparum* and the ancestors of modern humans, patterns of nucleotide variability and linkage disequilibrium suggest that the A- allele arose in Africa only within the last 10,000 years and spread due to selection.

INTRODUCTION

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

of selection are known, and in which the connections from genotype to phenotype to environment are well understood.

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

The most common G6PD deficiency allele in sub-Saharan Africa is called G6PD A-, and it typically reaches frequencies near 20% in populations living in malarial areas (Livingstone 1985). The A- allele differs from the normal allele (G6PD B) by non-synonymous changes at coding nucleotide positions 202 and 376. A minor deficiency allele, called G6PD A+, differs from the B allele only at site 376 (Figure 1). The enzymatic activities of the A+ and A- alleles are 85% and 12% of normal levels, respectively (Beutler et al. 1989; Hirono and Beutler 1988; Vulliamy et al. 1991). The mild deficiency phenotype characteristic of G6PD A+ does not cause significant clinical

manifestations and does not appear to confer resistance to malaria (Ruwende et al. 1995). However, the deficiency phenotype characteristic of G6PD A- confers ~50% reduction in risk of severe malaria in both heterozygote females and hemizygote males. Homozygous females probably have a similar level of protection from malaria, although this genotype is quite rare (Ruwende et al. 1995). In the presence of *falciparum* malaria, the G6PD A- allele is therefore beneficial, while in the absence of malaria, this allele is deleterious. Thus *G6PD* provides a rare example of a gene in humans where the selective agent and approximate form and strength of selection are known (Ruwende et al. 1995; Tishkoff et al. 2001).

As part of an ongoing project to characterize patterns of nucleotide variability at multiple loci throughout the genome for a common worldwide sample of human DNAs and to investigate the impact of selection on *G6PD*, we sequenced 5.1 kb of *G6PD* in a sample of 47 humans (Table 1). We also sequenced 3.7 kb at *L1CAM* in these same individuals. *L1CAM* is situated 556 kb from *G6PD*; thus, polymorphisms at *L1CAM* provide an opportunity to investigate the impact of selection on neighboring sites. Our nucleotide data suggest that the effects of selection on *G6PD* are more subtle than those predicted under a model of long-term diversifying selection.

MATERIALS AND METHODS

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

PCR amplification and sequencing: Maps of the human X-chromosome and the loci sampled in this study, *G6PD* and *L1CAM*, are presented in Figure 1. *L1CAM* was chosen because of its proximity to *G6PD* (556 kb); all polymorphisms detected at *L1CAM* are silent or non-coding, and there is no *a priori* reason to assume that *L1CAM* itself is a target of selection. Approximately 82 other genes are found within 1 Mb on either side of *G6PD* and none of these genes are known to be recent targets of positive

selection. PCR fragments were amplified for *G6PD* (5.2 kb), and *L1CAM* (4.2 kb) using a long-template PCR system (Roche Biochemicals). For *G6PD*, the primers Gf (5' GTT TAT GTC TTC TGG GTC AGG GAT GG 3') and Gr (5' AGT GTT GCT GGA AGT CAT CTT GGG T 3') are positioned with the 5' end of the primer at sites 206322 and 201052, respectively, in Genebank accession L44140. For *L1CAM*, the primers Lf (5' TCC TCT CCA GAG TAG CCG ATA GTG ACC 3'), and Lr (5' AAG TTT CTA CTG GCC TGA CCC TCT CG 3') are positioned with the 5' end of the primer at sites 19587 and 24251, respectively, in Genebank accession U52112 (Figure 1). Internal primers (available upon request) were used to generate overlapping sequence runs on an ABI 377 automated sequencer. Contiguous sequence that included coding and non-coding regions (5109 bp and 3691 bp for *G6PD* and *L1CAM* respectively) was assembled for each individual and aligned using the computer program *Sequencher* (GeneCodes). Sequences have been submitted to GenBank under accession numbers AY158094-AY158142 and AY167680-AY167728 for *G6PD* and *L1CAM*, respectively.

Data analysis: Nucleotide diversity, π (Nei and Li 1979), and the proportion of segregating sites, θ (Watterson 1975), were calculated using the program *PROSEQ* (Filatov et al. 2000) for the worldwide sample, African individuals, and for non-African individuals. Only the 41 individuals of the non-augmented worldwide sample were included in analyses of nucleotide diversity, and insertion-deletion polymorphisms were excluded. Under neutral equilibrium conditions both π and θ estimate the neutral parameter $3N_e \mu$ for X-linked loci, where N_e is the effective population size and μ is the neutral mutation rate. Tajima's D (Tajima 1989), Fu and Li's D (Fu and Li 1993) and

Fay and Wu's H (Fay and Wu 2000; http://crimp.lbl.gov/htest.html) were calculated to test for deviations from a neutral equilibrium frequency distribution for both loci. Ratios of polymorphism to divergence for *G6PD* and *L1CAM* were compared with the expectations under a neutral model using the HKA test (Hudson et al. 1987). Polymorphism data for these tests were derived from the 41 sequences determined in this study for *G6PD* and *L1CAM*, as well as from *DMD* (intron 44) from the same set of individuals (Nachman and Crowell 2000) and from the *PDHA1* data of Harris and Hey (1999). *DMD* and *PDHA1* were chosen for comparison because they both reside in regions of the X-chromosome with moderate to high rates of recombination and thus are expected to be relatively free of the effects of selection at linked sites. Divergence data were derived for each of these loci by comparing the homologous sequences from a chimpanzee to a single randomly chosen human allele. LD between pairs of polymorphic sites was measured using the statistics D' (Lewontin 1964) and r^2 (Hill and Robertson 1968). The age of the G6PD A- allele was estimated from the decay of linkage disequilibrium and from coalescent simulations using the computer program *GENETREE* (Harding et al. 1997; Bahlo and Griffiths 2000). The *SWST* haplotype test of Andolfatto et al. (1999) was implemented using the data from *G6PD* and *L1CAM* separately. This test compares the observed number of haplotypes with those expected under a neutral model with a specified rate of recombination.

RESULTS

Nucleotide diversity: Patterns of nucleotide variability at *G6PD* and *L1CAM* are presented in Tables 1 and 2. In the worldwide sample of 41 chromosomes (non-augmented sample) we observed 18 single nucleotide polymorphisms and three insertion/deletion (indel) polymorphisms at *G6PD*. Fifteen of these polymorphisms were in introns; of the remaining six polymorphisms, two were non-synonymous changes (coding sites 202 and 376) and four were synonymous changes. Levels of nucleotide variability were roughly four times higher in Africa than in non-African populations (Table 2), consistent with other studies that demonstrate higher diversity in Africa (*e.g.* Nachman and Crowell 2000; Harris and Hey 1999; 2001). Many of the polymorphisms found in Africa distinguish G6PD A- alleles from all other alleles in the sample. At *L1CAM* we observed 7 polymorphisms in the non-augmented sample. Levels of nucleotide variability were relatively low for *L1CAM* overall, however nucleotide variability in Africa was higher than in non-African populations.

In the worldwide sample of 41 chromosomes, there were two A- alleles in the African subset (n=10), consistent with previously documented frequencies of G6PD A- in sub-Saharan Africa of ~20%. Overall, worldwide levels of nucleotide variability at *G6PD* and *L1CAM* were close to or slightly below average values for other regions of the genome. For example, among primarily non-coding sites at 12 X-linked genes in humans, the average level of nucleotide diversity (π) is 0.06% and the average proportion of segregating sites (Waterson's θ) is 0.07% (Nachman 2001). For both *G6PD* and *L1CAM*, nucleotide diversity at intron sites is slightly below average (*G6PD* $\pi = 0.04$,

L1CAM $\pi = 0.02$), while Waterson's θ is close to average (*G6PD* $\theta = 0.08$, *L1CAM* $\theta = 0.07$). Since the A- allele represents only 5% of the worldwide sample, it is not expected to contribute substantially to levels of nucleotide variability. Within Africa, however, G6PD A- is present at high frequency (20%), yet overall levels of nucleotide variability ($\pi = 0.08\%$, Table 2) are still average. For example, the average level of nucleotide variability for 8 X-linked genes in Africa is 0.084% (Payseur and Nachman 2002).

Tests of neutrality: Tajima's D is the normalized difference between π and θ, and takes on positive values when there is an excess of intermediate-frequency polymorphisms and negative values when there is an excess of low-frequency polymorphisms (Tajima 1989). Positive Tajima's D values are generally consistent with long-term balancing selection or a population contraction, while negative values are expected following a selective sweep or a population expansion. For *G6PD*, Tajima's D is negative (but not significant) in the worldwide sample and for all subsets of the data (Table 2). Similar results are obtained with Fu and Li's D (Fu and Li 1993), which also measures the frequency distribution of polymorphisms and is sensitive to the number of singletons in the sample (Table 2). For *L1CAM* both statistics are also negative, and are significantly negative in the worldwide sample (Table 2). Fay and Wu's H statistic (Fay and Wu 2000) utilizes the frequency distribution of polymorphisms to test for an excess of high frequency derived variants compared to equilibrium neutral expectations. For both *G6PD* and *L1CAM*, Fay and Wu's H test shows no significant deviation from the neutral expectation in the worldwide sample, the African sample, or the non-African Sample.

We performed an HKA test (Hudson et al. 1987) using pairwise comparisons of polymorphism and divergence for *G6PD* and *L1CAM* and two other X-linked genes, *PDHA1* and *DMD*. In comparisons using worldwide samples or African samples alone, we failed to reject the null model (HKA $X^2 < 3.0$, $P > 0.1$ for all tests). Thus, neither the frequency spectrum nor the level of heterozygosity at *G6PD* fits the expected pattern of nucleotide variability under a simple model of long-standing diversifying or balancing selection.

To test whether the haplotype structure of the data deviates from neutral expectations we implemented the *SWST* program as described in Andolfatto et al. (1999), assuming recombination rates of 0, 1, and 2 cM/Mb. Tests were performed separately for *G6PD* and for *L1CAM*. None of these tests showed significant deviations from neutral expectations using the worldwide sample or the African sample alone.

Linkage disequilibrium: To better examine patterns of linkage disequilibrium we augmented our random sample of 10 African X-chromosomes with 4 chromosomes carrying A- alleles and 2 chromosomes carrying A+ alleles. Thus the augmented African sample in the study includes 6 chromosomes carrying *G6PD* A- alleles from South Africa, Central Africa, and West Africa (Samples YCC 9, YCC 32, G11, M115, M241, S823 in Table 1). Unusually high levels of linkage disequilibrium were observed within *G6PD*, within *L1CAM*, and between *G6PD* and *L1CAM*. D' is a measure of linkage disequilibrium that is standardized to equal 0 when there is random association among polymorphisms (*i.e.* no disequilibrium), and to equal 1 when there is complete association among polymorphisms (*i.e.* complete disequilibrium). In all comparisons

between A- alleles and other alleles, |D'|=1 for all sites in Table 1. A single most parsimonious haplotype network was inferred for all sites at *G6PD* (Figure 2), indicating that there is no evidence for recombination in this sample despite the fact that Xq28 is a genomic region with moderate to high rates of recombination (Payseur and Nachman 2000). Surprisingly, the three polymorphic sites at *L1CAM* at intermediate frequency (positions 776, 885 and 2115) can also be mapped on this same network with no homoplasy. The observed high level of linkage disequilibrium is primarily a consequence of mutations falling on the branch separating the A- deficiency allele from the normal B alleles (Figure 2). A Fisher's exact test revealed significant LD (P=0.0082) between site 202 of *G6PD* and three out of four informative polymorphisms at *L1CAM* (alignment positions 776, 885 and 2115 at *L1CAM*; Table 3).

Age of the G6PD A- allele: We estimated the age of the A- allele in two ways. First, we use a standard model for the decay of linkage disequilibrium as a function of time (t) and recombination (c), where linkage disequilibrium at time t (r^2_t) compared with time 0 (r^2_0) is given by $r^2_t/r^2_0 = (1\text{-}c)^t$ (Hedrick 1998). For this calculation, we use r^2 as a measure of linkage disequilibrium between *L1CAM* and *G6PD* because, unlike D', r^2 is sensitive to allele frequencies when only three out of four gametic types are present in a sample (Hedrick 1998). Assuming that linkage disequilibrium between site 202 of *G6PD* A- and positions 776, 885 and 2115 of *L1CAM* alleles was complete at the time of origin of the A- allele (*i.e.* $r^2_0 = 1$), we can estimate the time in generations, $t = \ln(r^2_t)/\ln(1\text{-}c)$, for the age of the allele given the observed recombinational distance between *L1CAM* and *G6PD* and the observed linkage disequilibrium in the data ($r^2 = 0.52$). Since it is

possible that r^2_0 was less than 1.0 between these sites when the G6PD A- allele arose, our estimates provide an upper bound for the age of the G6PD A- allele. This region of Xq28 is subject to moderate levels of recombination in general (1-3 cM/Mb: Payseur and Nachman 2000, Kong et al. 2002), and recombination rates near *G6PD* and *L1CAM* specifically have been estimated as low as 0.14 cM/Mb and as high as 2 cM/Mb (Small et al. 1997). Using these recombination rates we estimated the maximum age of the A- allele to be between 58 generations and 840 generations (Figure 3). With a generation time of 25 years, this implies that the *G6PD* A- allele arose 1,461-20,994 years ago.

A second estimate for the age of the A- allele was obtained from simulations using a coalescent model conditioned on the sample size and observed levels of nucleotide variability (*GENETREE*: Harding et al. 1997; Bahlo and Griffiths 2000). This model assumes neutral, equilibrium conditions, and thus may provide an overestimate of the true age of the A- allele (since the present frequency of A- has probably been determined in large measure by selection). These simulations suggest that the A- allele arose 10,575 years ago (SD ± 8,887 years).

Both of these estimates are in good agreement with an independent estimate for the age of the G6PD A- (3,840 to 11,760 years) allele that was reported by Tishkoff et al. (2001) based on intra-allelic levels of linked microsatellite variability.

DISCUSSION

Models of selection and nucleotide variability at *G6PD*: We investigated levels and patterns of nucleotide variability at *G6PD*, a locus known to be under malarial selection in some human populations, and found that nucleotide diversity was similar to average values for other X-linked genes. Moreover, several commonly employed statistical tests based on DNA sequence variation failed to reject a simple neutral model of molecular evolution. In several respects, however, the data from *G6PD* are quite striking: levels of linkage disequilibrium are high and extend over long genomic distances, much of the nucleotide variation is partitioned between functionally distinct alleles, and no nucleotide variation is observed within deficiency alleles. Below we discuss general models of selection for *G6PD* and how our observations might fit these models.

Although four different species of *Plasmodium* typically infect humans, *P. falciparum* is the most virulent species and is responsible for most malaria related deaths, especially in Africa (Schmidt And Roberts 1996). Malaria is endemic throughout most of sub-Saharan Africa where over one million people die each year due to complications from infection (Trigg and Konrachine 1998). From a population genetics perspective, such a virulent parasite serves as a strong selective agent for genetic resistance. In fact it has long been known that African populations exhibit genetic resistance factors to malaria at relatively high frequencies compared with non-African populations (*e.g.* Miller 1994). Moreover, many of the mutations that confer resistance are deleterious outside of the malaria environment. G6PD A-, for example, has an enzymatic activity that is only

about one-tenth of normal and results in significant clinical manifestations such as hemolytic anemia and neonatal jaundice (Beutler et al. 1989; Hirono and Beutler 1988; Vulliamy et al. 1991). However, this allele also confers approximately a 50% reduction in risk of severe malaria in both heterozygote females and hemizygote males (Ruwende et al. 1995).

Although *G6PD* is often assumed to be subject to balancing selection (*sensu* heterosis) (*e.g.* Tishkoff et al. 2001), the precise nature of selection on G6PD deficiency alleles is not fully understood. In the absence of malaria, deficiency alleles are at a selective disadvantage and are expected to be eliminated (Table 4, fitness array 1). In the presence of malaria, female heterozygotes and male deficiency hemizygotes appear to have a selective advantage over wild-type individuals, but the fitness of female deficiency homozygotes relative to other genotypes is not clear (Ruwende et al. 1995). If female heterozygotes have a higher fitness than either homozygote, then selection may maintain both A- and wild-type alleles in populations under malarial selection (*i.e.* heterosis; Table 4, fitness array 2). However, the conditions for maintenance of a stable X-linked polymorphism are rather restrictive; either selection must be of similar magnitude but opposite in direction for the two sexes (which seems unlikely for G6PD deficiency) or there must be heterosis in females without a large fitness difference between the two male genotypes (Hedrick 1998). Alternatively, if female deficiency homozygotes have the same fitness as male hemizygotes and female heterozygotes, then selection should drive the eventual fixation of the G6PD A- allele in populations subject to continuous malarial selection (Table 4, fitness array 3). In such a situation, the A-

allele is expected to rise to high frequencies and reach fixation in a very short period of time (*e.g.* <10,000 years; Ruwende et al. 1995). The exact time required for fixation depends on assumptions about population size, initial frequency of the A- allele, relative fitness of the different genotypes, and the average generation time. However, for a wide range of parameter values, allele frequencies are expected to quickly rise to very high levels. The observation that most African populations have A- allele frequencies below 20% (Livingstone 1985) is inconsistent with a simple model of directional selection where selection has been strong and long acting.

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

A simple model of long-term balancing selection or long-term spatially or temporally varying selection is expected to leave a distinct signature in patterns of DNA sequence variation (Figure 4). When a new advantageous mutation first appears (Figure

4a), it will rise in frequency, creating LD with other mutations on the haplotype on which it arose (Figure 4b). This transient phase will result in lowered levels of heterozygosity. Over time, linkage disequilibrium will decay through recombination around the target of selection, and heterozygosity will increase near the target of selection (Figure 4c). This simple model of long-term selection predicts elevated levels of nucleotide variability in a restricted window around the target of selection (Hudson et al. 1987) and a skew in the frequency distribution of polymorphisms with an excess of intermediate-frequency variants within this restricted window (Tajima 1989). Both of these patterns are seen in several other well studied systems. For example, at *Mhc* loci in a variety of organisms (HLA in humans), levels of heterozygosity are significantly higher than in neighboring regions (Takahata et al. 1992). At *Adh* in *Drosophila melanogaster*, heterozygosity is elevated around the fast/slow allozyme polymorphism, resulting in a significant HKA test (Hudson et al. 1987).

In contrast, patterns of nucleotide variability at *G6PD* do not support either of these predictions with respect to *G6PD* A-, and several observations suggest that patterns at *G6PD* fit the model expected in an early stage of selection (Figure 4b). First, overall levels of nucleotide diversity are close to average values for other X-linked loci. This is true for the worldwide sample, and more importantly for evaluating models of selection, it is also true for the African sample alone. An HKA test applied to our data fails to reject a neutral model. Second, there is no evidence for an excess of intermediate-frequency polymorphisms. In fact, both Tajima's D and Fu and Li's D are slightly (but not significantly) negative for the African sample (Table 2). Third, we find extensive

linkage disequilibrium within and around *G6PD*, and this disequilibrium is due almost exclusively to nucleotide differences that distinguish the A- allele from other alleles. We observed no recombination events within *G6PD*. This stands in contrast to many other human nucleotide polymorphism datasets, including intron 44 of *DMD*, surveyed in this same set of individuals (Nachman and Crowell 2000), in which numerous recombination events were observed over distances of several hundred bases. In addition to significant LD within *G6PD*, we found significant LD between *G6PD* and *L1CAM* (Table 3; D'=1 in all comparisons), loci that are separated by approximately 550 kb. This amount of LD is much higher than typical values for the human genome. For example, Reich et al. (2001) recently studied the decay of D' for 19 different genomic regions and found that in a European population the half-length of D' (the distance at which the average D' drops below 0.5) is typically 60 kb, while in an African population the half-length of D' is less than 5 kb (Reich et al. 2001). Other studies have also revealed lower levels of linkage disequilibrium in African populations compared with non-African populations (Tishkoff et al. 1996; Tishkoff et al. 2001). Interestingly, we observe much higher levels of LD than previously reported for this region of Xq28 by Taillon-Miller et al. (2000) in populations of European descent. Finally, there is no intra-allelic variation within *G6PD* A-, consistent with the notion that G6PD A- is relatively young.

Taken together, these observations argue against a model of long-term selection on the *G6PD* A- allele, but do not allow us to distinguish between recent balancing selection (*sensu* heterosis) on the one hand, and recent diversity-enhancing (*i.e.* spatially and/or temporally varying) selection on the other hand. Better fitness estimates of all

genotypes (in particular female deficiency homozygotes), as well as detailed sampling of G6PD A- frequencies across Africa, might help us to distinguish between these hypotheses.

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

Is it possible that demographic processes are primarily responsible for the high levels of LD seen in Figure 2? Linguistic and archeological evidence suggests that a Bantu expansion took place in Africa ~4000 years ago (Excoffier et al. 1987). This range expansion occurred in sub-Saharan Africa primarily from west to east and southward, a

distribution that is similar to the current distribution of African populations with elevated G6PD- allele frequencies. If admixture from this range expansion were responsible for generating the observed LD in our data, we would also expect to see G6PD B alleles with significant LD. This is not observed. Instead, most of the LD in our data is found between sites on functionally different alleles, arguing against any simple demographic explanation. Likewise no LD is observed between Bantu and non-Bantu individuals from this set of 41 individuals sampled for other loci (*e.g.* Nachman and Crowell 2001).

One intriguing observation in our dataset is the relatively high level of divergence found at *L1CAM* between individuals bearing the G6PD A- allele and all other individuals. Four of the six (66.7%) G6PD A- alleles share a common motif of three polymorphisms in complete linkage disequilibrium (C, T and T at positions 776, 885 and 2115 respectively; Table 1) while the rest of the segregating sites at *L1CAM* include only four singletons and one doubleton. This pattern along with the significant LD between *G6PD* and *L1CAM* (Table 3) suggests that the A- mutation arose on a relatively diverged haplotype, possibly as a consequence of population subdivision. Analysis of *G6PD* and *L1CAM* as well as additional neighboring loci in a larger geographic sample from Africa may shed light on this unusual pattern.

In general, the observations reported here demonstrate that even when selection is relatively strong, its signature on patterns of DNA sequence variation may be subtle, especially if selection is recent. While several of the conventional statistical tests for selection fail to reject the null hypothesis, the footprint of selection is seen in the long-range patterns of LD and in the absence of variation among A- alleles from different parts

of Africa. Similar patterns of nucleotide variability at *G6PD* have also recently been reported by Verrelli et al. (in press). The patterns of DNA sequence variation observed at *G6PD* are markedly different from those seen at another well-studied target of balancing selection, HLA, where ancient alleles result in substantially elevated levels of polymorphism (Gaudieri et al. 1999; Grimsley et al. 1998; Horton et al. 1998). The spatial and temporal scales over which selection pressures have shaped human genomic diversity are still largely unknown, but environmental changes associated with the transition from the Paleolithic to the Neolithic may have imposed substantial new selection pressures on humans, suggesting that patterns of nucleotide variability similar to those documented here for *G6PD* may be found at other loci.

Age of G6PD A- and the evolution of malarial resistance: These results have important implications for the evolution of resistance to malaria in humans. Several observations reported here suggest that the A- allele is young, including average levels of nucleotide variability at *G6PD*, negative values of Tajima's D, high levels of linkage disequilibrium between *G6PD* and *L1CAM*, and complete absence of variation among *G6PD* A- alleles from different parts of Africa (Tables 1 and 2). A recent study based on microsatellite haplotype diversity (Tishkoff et al. 2001) also suggests that the *G6PD* A- allele arose recently, within the past 4,000-12,000 years. A recent phylogeny of primate malaria parasites indicates that *P. falciparum* is closely related to *P. reichenowi*, a chimpanzee parasite. Moreover, *cytochrome b* sequence divergence between *P. falciparum* and *P. reichenowi* suggests a divergence time of 4-5 My ago (Escalante et al. 1998), in good agreement with the estimated time of the human-chimpanzee divergence.

The discrepancy between this date and the recent origin of the *G6PD* A- allele raises the possibility that *P. falciparum* has been a human parasite for most of the evolutionary history of *Homo sapiens*, but that the parasite's current level of virulence has evolved only recently (Rich et al. 1998). The estimated age of the A- allele agrees well with the spread of agriculture throughout sub-Saharan Africa (Cavalli-Sforza et al. 1996; Waters et al. 1991) and suggests that changes in human lifeway may have contributed to the transmission and/or increased virulence of *P. falciparum*, perhaps through an increase in the density and mobility of *Anopheles* mosquitoes that serve as vectors in transmission of malaria or by evolution of increased virulence factors in *P. falciparum*.

ACKNOWLEDGMENTS

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

LITERATURE CITED

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

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

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

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

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

Beutler, E., W. Kuhl, J. L. Vivescorrons And J. T. Prchal, 1989 Molecular Heterogeneity of glucose-6-phosphate dehydrogenase-a. Blood **74:** 2550-2555.

Campbell, C. C., 1997 Malaria: an emerging and re-emerging global plague. FEMS Immunol. Med. Microbiol. **18:** 325-331.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza, 1996 *The history and geography of human genes*. Princeton University Press, Princeton, N.J.

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

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

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

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

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

Fay, J. C., G. J. Wykoff and C. I. Wu, 2001 Positive and negative selection on the human genome. Genetics **158**: 1227-1234.

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

Flint, J., R. M. Harding, A. J. Boyce and J. B. Clegg, 1993 The population-genetics of the hemoglobinopathies. Baillieres Clin. Haematol. **6:** 215-262.

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

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

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

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

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

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

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

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

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

Hedrick, P. W., 1998 *Genetics of populations*. Jones and Bartlett Publishers, Sundbury , Massachusetts.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Payseur, B. A., and M. W. Nachman, (2002) Natural selection at linked sites in humans. Gene (*in press*).

Rana, B. K., D. Hewett-Emmett, L. Jin, B. H. J. Chang, N. Sambuughin et al., 1999 High polymorphism at the human melanocortin-1 receptor locus. Genetics **151:** 1547-1557.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti et al., 2001 Linkage disequilibrium in the human genome. Nature **411:** 199-204.

Rich, S. M., M. C. Light, R. R. Hudson and F. J. Ayala, 1998 Malaria's eve: Evidence of a recent population bottleneck throughout the world populations of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA **95:** 4425-4430.

Rieder, M. J., S. L. Taylor, A. G. Clark And D. A. Nickerson, 1999 Sequence variation in the human angiotensin converting enzyme. Nat. Genet. **22:** 59-62.

Roth, E., and S. Schulman, 1988 The adaptation of *Plasmodium falciparum* to oxidatve stress in G6PD deficient human erythrocytes. Brit. J. Haematol. **70:** 363-367.

Roth, E. F., C. Raventos-Suarez, A. Rinaldi and R. L. Nagel, 1983 Glucose-6-phosphate dehyrogenase deficiency inhibits *in vitro* growth of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA **80:** 298-299.

Ruwende, C., S. C. Khoo, A. W. Snow, S. N. R. Yates, D. Kwiatkowski et al., 1995 Natural-selection of hemizygotes and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature **376:** 246-249.

Schmidt, G. D., and L. S. Roberts, 1996 *Foundations of Parastiology*. Wm. C. Brown Publishers, Chicago.

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

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

Taillon-Miller, P., I. Bauer-Sardina, N. L. Saccone, J. Putzel, T. Laitinen et al., 2000 Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nat. Genet. **25**: 324-328.

Takahata N., Y. Satta and J. Klein, 1992 Polymorphism and balancing selection at major histocompatibility complex loci. Genetics **130**: 925-938.

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

Tishkoff, S. A., E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd et al., 1996 Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science **271:** 1380-1387.

Tishkoff, S. A., A. J. Pakstis, G. Ruano and K. K. Kidd, 2000 The accuracy of statistical methods for estimation of haplotype frequencies: An example from the CD4 locus. Am. J. Hum. Genet. **67:** 518-522.

Tishkoff, S. A., R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos et al., 2001 Haplotype diversity and linkage disequilibrium at human G6PD: Recent origin of alleles that confer malarial resistance. Science **293**: 455-462.

Trigg, P. I., and A. V. Konrachine, 1998 Commentary: Malaria control in the 1990s. Bull. WHO **76:** 11-16.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural et al., 2001 The sequence of the human genome. Science **291:** 1304-1351.

Vulliamy, T. J., A. Othman, M. Town, A. Nathwani, A. G. Falusi et al., 1991 Polymorphic sites in the African population detected by sequence analysis of the glucose-6-phosphate dehydrogenase gene outline the evolution of the variant A+ and variant A-. Proc. Natl. Acad. Sci. USA **88:** 8568-8571.

Verrelli, B. C., J. H. McDonald, G. Argyropolous, G. Destro-Bisol, A. Froment et al., (in press) Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. Am. J. Hum. Genet.

Wang, L. H., A. Collins, S. Lawrence, B. J. Keats and N. E. Morton, 1994 Integration of gene maps - Chromosome-X. Genomics **22:** 590-604.

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

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

Yu, N., Y. X. Fu, N. Sambuughin, M. Ramsay, T. Jenkins et al., 2001 Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. **18**: 214-222.

Zhao, Z. M., L. Jin, Y. X. Fu, M. Ramsay, T. Jenkins et al., 2000 Worldwide DNA sequence variation in a 10-kilobase non-coding region on human chromosome 22. Proc. Natl. Acad. Sci. USA **97:** 11354-11358.

TABLE 1

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

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

(*continued*)

TABLE 1

(Continued)

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

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

Table 2: Summary statistics of nucleotide variability for *G6PD* and *L1CAM*.

Geographic region	Locus	Length (bp)	Sample size	S	π (SD) (%)	θ (SD) (%)	Tajima's D	Fu and Li's D	*Divergence (SD)* Homo-Pan (%)	*Divergence (SD)* Homo-Pongo (%)
Worldwide	*G6PD* total sequence	5102	41	18	0.05 (0.04)	0.08 (0.03)	-1.429	-1.134	1.0 (0.1)	3.2 (0.3)
	G6PD introns	2918	41	10	0.04 (0.039)	0.08 (0.033)	-1.512	-1.018	1.2 (0.2)	4.0 (0.4)
	L1CAM total sequence	3691	41	7	0.01 (0.021)	0.04 (0.02)	-1.946*	-1.822*	0.8 (0.1)	2.9† (0.4)
	L1CAM introns	2087	41	6	0.02 (0.032)	0.07 (0.033)	-1.925*	-2.203*	1.1 (0.2)	3.9† (0.6)
African Sample:	*G6PD* total sequence	5103	10	14	0.08 (0.051)	0.10 (0.046)	-0.672	-0.342		
	G6PD introns	2919	10	8	0.08 (0.051)	0.10 (0.050)	-0.687	-0.087		
	L1CAM total sequence	3691	10	6	0.05 (0.03)	0.06 (0.032)	-0.886	-0.553		
	L1CAM introns	2087	10	5	0.06 (0.045)	0.08 (0.049)	-1.035	-0.884		
Non-African Sample:	*G6PD* total sequence	5108	31	7	0.02 (0.017)	0.03 (0.016)	-1.032	-1.644		
	G6PD introns	2918	31	3	0.02 (0.013)	0.03 (0.016)	-0.929	-1.532		

LICAM total sequence	3691	31	1	0.00 (0.003)	0.01 (0.007)	-1.145	-1.681
LICAM introns	2087	31	1	0.00 (0.003)	0.01 (0.007)	-1.145	-1.681

Divergence estimates were based on a comparison between a single randomly chosen human allele and the chimpanzee or orangutan alleles.

†*Homo-Pongo* divergence estimates for *LICAM* are based on 1672 bp and 1046 bp for total sequence and introns respectively.

*P<0.05

Table 3: Non-random associations between *G6PD* site 202 and polymorphisms at *L1CAM*.

G6PD polymorphism / *L1CAM* polymorphism	**G,C,C**; Pos. 776, 885, 2115	**C,T,T**; Pos. 776, 885, 2115
G; Site 202 (G6PD B or A+)	10	0
A; Site 202 (G6PD A-)	2	4

Results of Fisher's exact test (P=0.0082) for African augmented sample (n=16) is presented between coding site 202 at *G6PD* and informative polymorphisms at *L1CAM* alignment positions 776, 885 and 2115 (Table 1).

Table 4: Fitness arrays under different malarial-selection regimes.

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

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

FIGURE LEGENDS

Figure 1: Schematic ideogram of the human X-chromosome and the genomic regions sampled in this study. Genes located between *L1CAM* and *G6PD* on Xq28 are marked. Dark arrows indicate positions of the amplification primers *Gf*, *Gr*, *Lf* and *Lr*. For *G6PD*, the mutations that define G6PD A+ and G6PD A- are marked at coding position 376 and 202. *L1CAM* is located 556 kb from *G6PD*. Exons are marked with dark boxes. Polymorphic amino acid residues for alleles G6PD B, G6PD A+ and G6PD A- due to nucleotide polymorphisms and sites 202 and 376 are shown in the shaded box.

Figure 2: Haplotype network for polymorphisms of the worldwide sample at *G6PD*. Text inside circles represents sample identities (Table 1). Marks indicate polymorphisms and are labeled with the respective alignment position from the polymorphism table (Table 1). (*) indicates known haplotypes for G6PD A+ alleles that were not captured in this sample (Vulliamy et al. 1991). The dark circle represents G6PD A-, gray circles represent G6PD A+, and open circles represent G6PD B alleles.

Figure 3: Plot of linkage decay between *G6PD* (site 202) and *L1CAM* (positions 776, 885 and 2115). The expected plot of linkage decay as measured by r^2 is shown over time (t) for a range of two different recombination rates that have been suggested for the chromosomal region near *G6PD* and *L1CAM* (Small et al. 1997), 0.14 cM/Mb (gray line) and 2 cM/Mb, (bold line). The observed $r^2 = 0.52$, provides minimum (t_{min}=58 generations) and maximum (t_{max} = 840 generations) ages for the *G6PD* A- allele.

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

Figure 1:

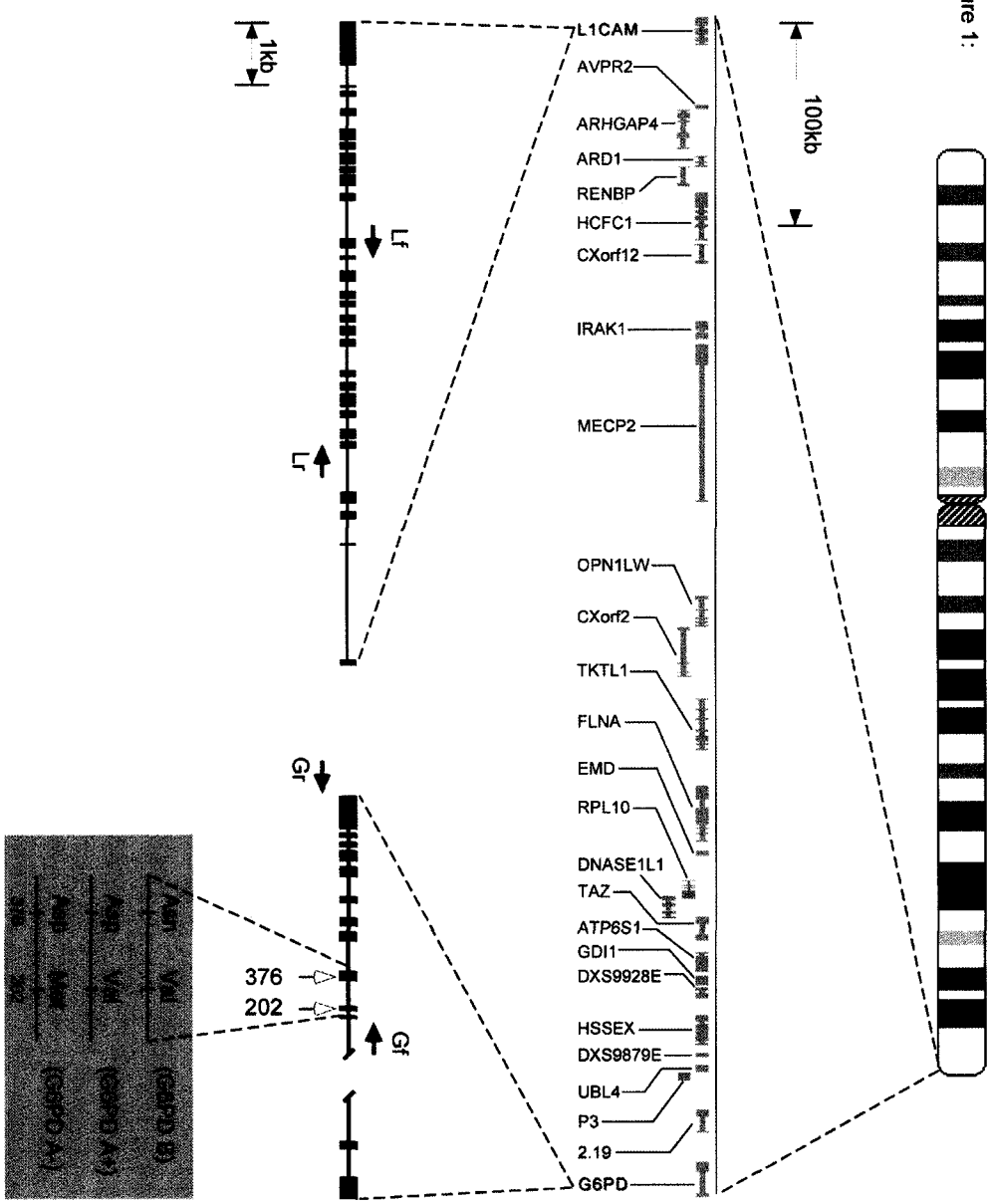

1kb
Lf
Lr
Gr
Gf
376
202
L1CAM
AVPR2
ARHGAP4
ARD1
RENBP
HCFC1
CXorf12
IRAK1
MECP2
OPN1LW
CXorf2
TKTL1
FLNA
EMD
RPL10
DNASE1L1
TAZ
ATP6S1
GDI1
DXS9928E
HSSEX
DXS9879E
UBL4
P3
2.19
G6PD
100kb

Figure 2

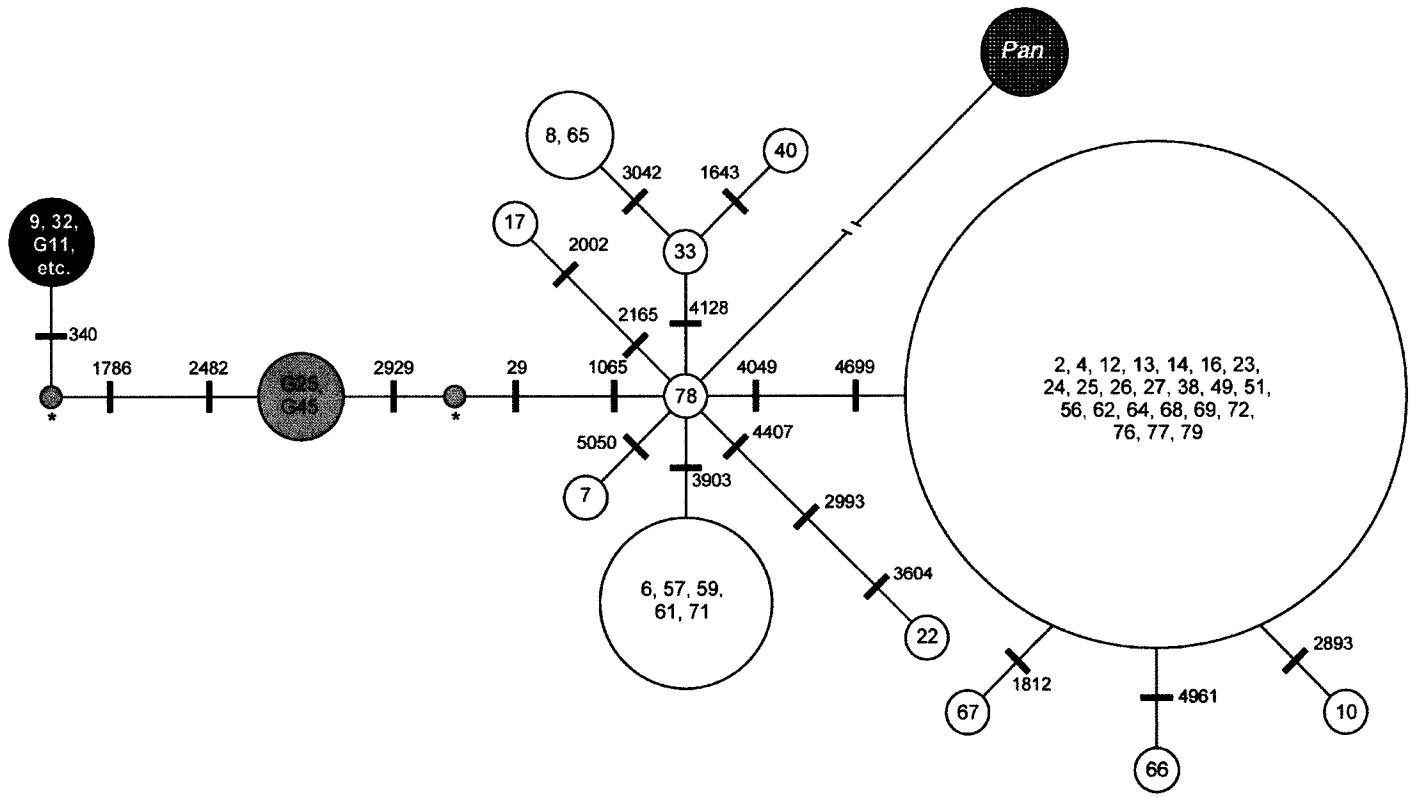

Pan
8, 65
3042
1643
40
17
2002
33
2165
4128
9, 32,
G11,
etc.
340
1786
2482
G25,
G45
2929
29
1065
78
4049
4699
2, 4, 12, 13, 14, 16, 23,
24, 25, 26, 27, 38, 49, 51,
56, 62, 64, 68, 69, 72,
76, 77, 79
5050
4407
7
3903
2993
6, 57, 59,
61, 71
3604
22
2893
1812
4961
67
10
66

Figure 3:

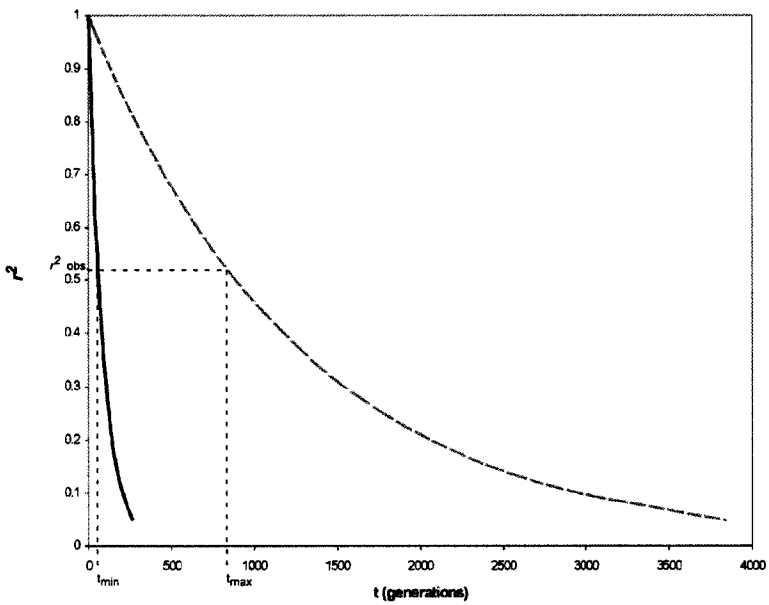
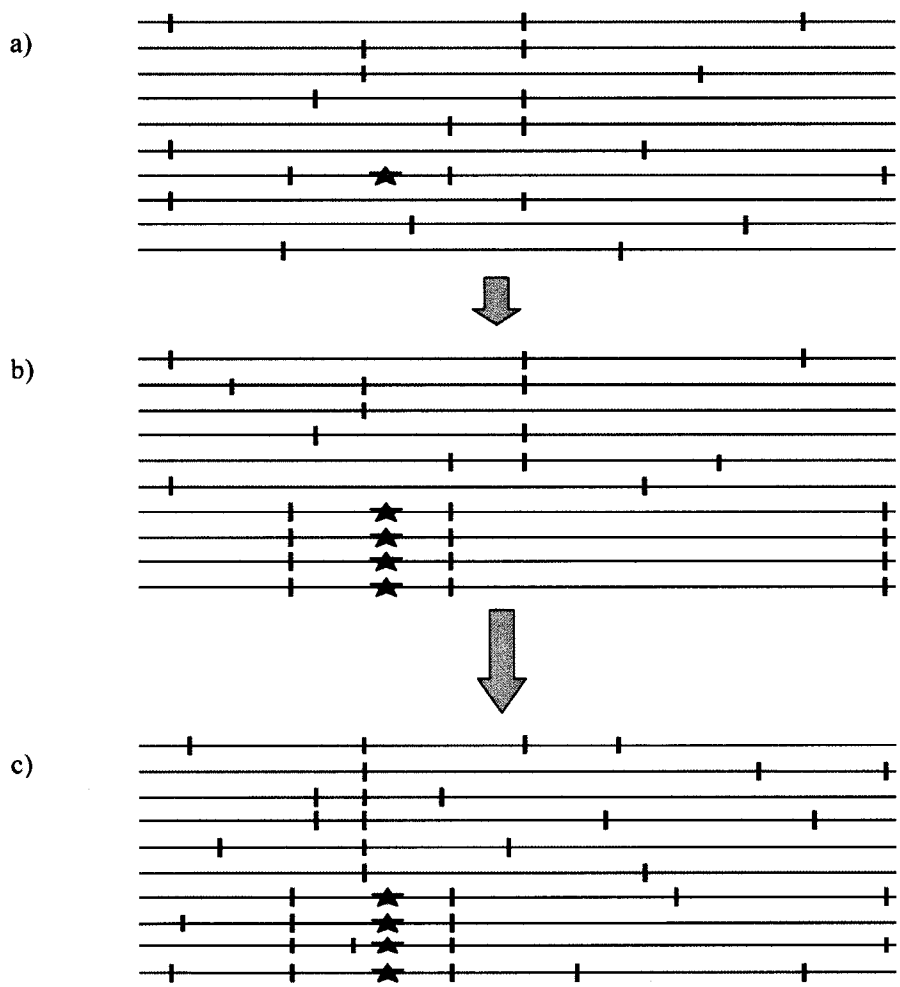

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
r^2
r^2_{obs}
0
500
1000
1500
2000
2500
3000
3500
4000
t_{min}
t_{max}
t (generations)

Figure 4:



APPENDIX B: LONG-RANGE LINKAGE DISEQUILIBRIUM AROUND *G6PD* IN AFRICA: EFFECTS OF NATURAL SELECTION BY MALARIA.

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 examine the impact of selection on patterns of linkage disequilibrium and nucleotide diversity, we resequenced 5.1 kb at *G6PD,* and ~ 2-3 kb at each of eight loci in a 2.5 Mb window 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 results extend previous findings of linkage disequilibrium at *G6PD* significantly, and show that selection can lead to non-random associations among SNPs over great physical and genetic distances, even in African populations. These patterns of nucleotide variability and LD also suggest that *G6PD A-* is younger than previous age estimates, and has increased in frequency in sub-Saharan Africa due to strong selection ($0.1 < s < 0.2$).

INTRODUCTION

Considerable work in the last five years has 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 some evidence that the spaces between these blocks correspond to recombination hotspots (*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 many human populations exhibit G6PD deficiency alleles at frequencies that range between 0.05 and 0.65 (Livingstone 1985; Oppenheim et al. 1993). Several lines of evidence indicate that some G6PD deficiency alleles confer resistance to severe malaria caused by *Plasmodium falciparum*, including (*1*) geographic correlations between the frequency of G6PD deficiency alleles and the historical prevalence of malaria (Allison 1960; Motulsky 1961), (*2*) *in vitro* experiments showing that the growth rate of *P. falciparum* is reduced in G6PD deficient erythrocytes relative to normal conditions (Roth et al. 1983; Roth and Schulman 1988), and (*3*) epidemiological evidence documenting reduced incidence of severe malaria among individuals with G6PD deficiency genotypes (Ruwende et al. 1995). 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 non synonymous changes relative to the normal allele (G6PD B) (Figure 1) which decrease enzyme activity to ~12% of normal (Hirono and Beutler 1988), and confers ~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 *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 signature of selection 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). These studies failed to detect evidence of selection using several commonly employed statistical tests (*e.g.* Hudson et al. 1987; Tajima 1989a; Fu and Li 1993), although Verrelli et al. (2002) showed that the ratio of non-synonymous to synonymous mutations was greater within than between species for *G6PD*. 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 an A→G mutation at *G6PD* coding site 376, causing a nonsynonymous change that reduces enzyme efficiency to 80% of normal (Figure 1). This mild deficiency allele does not confer resistance to malaria (Ruwende et al. 1993), and is found in sub-Saharan Africa at a frequency of ~ 0.2 (Takizawa et al. 1987). G6PD A- is defined by two mutations: an A→G mutation at *G6PD* coding site 376 co-occurring with a G→A mutation at coding site 202 (Figure 1) (Hirono and Beutler 1988). All G6PD functional alleles were determined *a priori* by restriction fragment length polymorphism analysis of a *FokI* 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 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*, *F8C*, and *G0.*9MT) were surveyed for nucleotide variability (Figure 1). Loci *G18MC*, *G1.5MC* and *G0.9MT* are 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 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 were amplified for each individual at all loci using a long-template 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 on our website (http://eebweb.arizona.edu/faculty/nachman/saunders/pubs/g6pd_2/primers.pdf). 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 *SEQUENCER* (GeneCodes). Sequences have been submitted to GenBank under accession numbers XXXXXX-XXXXXX.

Nucleotide variability data analysis: To gain insight into patterns of nucleotide variability for a random African sample across the loci surveyed, we assembled 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 θ_π (Nei and Li 1979) and θ_w (Watterson 1975) at each locus for the CRS using *dnaSP* 4.0 (Rozas and Rozas 1999). Under neutral equilibrium conditions both θ_π and θ_w estimate the neutral parameter $3N_e \mu$ for X-linked loci, where N_e is the effective population size and μ is the neutral mutation rate. Tajima's D (Tajima 1989) and Fu and Li's D (Fu and Li 1993) were calculated to test for deviations from a neutral equilibrium frequency distribution for each locus. Tajima's D is based on comparisons between θ_π and θ_w, and takes on positive values when there is an excess of intermediate-frequency polymorphisms, and negative values when there is an excess of low-frequency polymorphisms (Tajima 1989). Significant positive Tajima's D values are generally consistent with long-term balancing selection or a population contraction, while negative values are expected following a selective sweep, a population expansion, or if there are mildly deleterious alleles in a sample. Fu and Li's D and Fay and Wu's H also test the frequency distribution of alleles for deviation from neutral equilibrium expectations. Ratios of polymorphism to divergence for all loci were compared with the expectations under a neutral model using the HKA test (Hudson et al. 1987) which rests on the premise that under neutral equilibrium conditions, the ratio of polymorphism to divergence should be roughly equal between any two (or more) loci. Significant deviations of this ratio based on a χ^2 distribution may be indicative of selection. Divergence data were derived for each of these loci by comparing the homologous sequences from a chimpanzee to a single randomly chosen human allele. LD between pairs of polymorphic sites was measured using the 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). We also calculated θ_π and θ_w among individuals bearing each of the three classes of G6PD alleles (A-, A+ and B) to test for inter-allelic differences in levels of variability.

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 *L1CAM*, 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, as the results should not differ significantly from a codominance model for a young allele. 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- 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 telomoric 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 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*: Patterns of nucleotide variability at *G6PD* and 9 flanking loci are presented in Figure 2. We calculated nucleotide variability for 4 subset groups: (*i*) individuals bearing the G6PD A- allele (n=20), (*ii*) individuals bearing the G6PD A+ allele (n=11), (*iii*) individuals bearing the G6PD B allele (n=20), and (*iv*) individuals of the CRS (n=26). At *G6PD* we observed 20 segregating sites 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. Non-coding 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 $\theta_\pi = 0.069\%$ and $\theta_w = 0.082\%$ (Figure 3a; Table 2) consistent with previously reported levels of nucleotide variability at *G6PD* in sub-Saharan Africa (Sabeti *at al.* 2002; Saunders et al. 2002; Verrelli et al. 2002) and the average of 15 other X-linked loci ($\theta_\pi = 0.0755\%$, $\theta_w = 0.0815$: Hammer et al. 2004).

Nucleotide diversity around *G6PD*: Nucleotide variability for 9 loci flanking *G6PD* is presented in Figure 2. At *GAB3*, *F8C* 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 non-coding nucleotide variability for the A+ group, the B group and the CRS was $\theta_\pi = 0.027\%$, 0.024%, 0.022% respectively (Figure 3a). Although nucleotide variability in the CRS distal from *G6PD* is somewhat lower than typical

values of nucleotide variability at X-linked loci, we see even less variability in the class of A- individuals compared with the other classes of chromosomes (Figure 3a). This pattern is also seen from estimates of haplotype diversity for the different 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.

This general pattern of reduced variability among A- individuals is also seen proximal to *G6PD*, however the pattern is not as extreme as on the distal side, and the pattern decays beyond *L1CAM* (~556 kb from *G6PD*; Figure 3b). At *TAZ*, *L1CAM* and *IDH3G* the average non-coding nucleotide variability for the A- group, A+ group, B group and the CRS was θ_π = 0.032%, 0.037%, 0.040% and 0.041% respectively (Figure 3a; Table 2). At *BGN*, *G1.5MC* and *G18MC*, loci mapping 0.9 -19 Mb from *G6PD*, the average non-coding nucleotide variability for the A- group, A+ group, B group and the CRS was θ_π = 0.093%, 0.072%, 0.091% and 0.082% respectively (Figure 3a; Table 2). At these three loci the average level of nucleotide variability among the A- individuals is not reduced relative to the other allele classes. Together these 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).

We implemented several statistical tests of neutrality on the CRS. For *G6PD* and all other loci surveyed, Tajima's D is negative (but not significant except for *TAZ*) (Table 2; Figure 3c). Similarly non-significant results are obtained with Fu and Li's D, which

also measures the frequency distribution of polymorphisms and is sensitive to the number of singletons in the sample (Fu and Li 1993). A multi-locus HKA test including *G6PD* and the nine surrounding loci also failed to reject the null model.

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 *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 used, we can reject neutrality of G6PD A- and conclude that *s* is likely to be at least 0.1 even for the smallest recombination rates. 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 (HbS) 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. Across the loci surveyed in this study, levels of nucleotide variability of the CRS are consistent with estimates from other X-linked loci, and in general, there is no skew in the frequency distribution of alleles at *G6PD*, or at any of the flanking loci surveyed. These loci show a negative value for Tajima's D (although non-significant except for *TAZ*), concordant with results from many other loci surveyed to date in humans that are not under recent positive selection (*e.g.* Wall and Przeworski 2000; Stephens et al. 2001; Hammer et al. 2004). Despite the broad utility of statistical tests based on the distribution of allele frequencies, their power to detect deviations from neutrality may be limited in a species with low nucleotide diversity (Braverman et al. 1995) such as humans, or in situations where the selected alleles are still at a relatively low frequency (*e.g.* $q < 0.20$ for G6PD A-). Unfortunately, this presents a formidable challenge for identifying selection in humans, since many functional variants are present at relatively low frequencies in humans. For example, S and C *β-globin* alleles are at frequencies of ~ 0.05-0.20 (Livingstone 1985; Modiano et

al. 2001), and *HFE* alleles (Toomajian and Kreitman 2002) and $G6PD_{mediterranean}$ alleles (Livingstone 1985) are typically found at frequencies of 0.5-0.10.

Conserved extended haplotype among G6PD A- chromosomes: A remarkable feature of our data is the long-range extended haplotype common to the *G6PD* A- chromosomes. Previous studies have documented LD between *G6PD* and SNPs within ~ 600 kb around *G6PD* in Africa (Sabeti et al. 2002; Saunders et al. 2002). For example, the Long Range Haplotype test (Sabeti et al. 2002) demonstrates that G6PD A- alleles share a conserved haplotype that is significantly longer than typical alleles in an African sample. Here we have shown that the ancestral *G6PD* A- extended haplotype spans > 1.6 Mb. This extent of LD in the human genome is highly unusual, especially for Africans. 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- atypical. 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-Americans (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 similar levels of nucleotide variability as the non African-American G6PD A- samples considered alone. 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.

Other loci subject to recent natural selection exhibit relatively long range LD in association with selected alleles. 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). Furthermore, 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 (Meyer and Thomson 2001). Other examples are known where LD has been detected around putative targets of recent selection in humans. Significant LD was detected over 20 kb around *FY* in a non-African sample, consistent with recent selection by *Plasmodium vivax* (Hamblin et al. 2002). At *HB*, 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). In this context, *G6PD* provides an impressive example of the potential long range effects of selection on LD and nucleotide variability.

The decay of the G6PD A- extended haplotype (EH) is 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 psuedoautosomal 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 *G6PD*. For example, the gene *OPNMW1* (OMIM #303800) is responsible for red/green color blindness, and 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 ($G6PD_{mediterranean}$; coding site 563 C→T), all individuals with the 563 C→T mutation (n=7) were also deutan color blind based on visual acuity test. This suggests that in this population a chromosome that carried a $G6PD_{mediterranean}$ allele also harbored a mutation causing a clinical condition of deutan color blindness. Presumably, as the $G6PD_{mediterranean}$ mutation was favorably selected in this population, the deutan color blindness trait hitchhiked on the EH. A non-random association between a selected trait and another functional trait due to linkage will depend on the genetic distance between the genetic loci, the time when the chromosomal association was formed, and potential epistatic interactions between the functional traits. In our data, the polymorphism at site 41 in *IDH3G* (Figure 2) causes a non-conservative 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) based on closely linked microsatellite variability (Tishkoff et al. 2001), coalescent-based analysis of a

G6PD gene tree (Coop and Griffiths 2004), and two-locus long range LD (Saunders et al. 2002). In this study we have delimited the span of the ancestral G6PD A- haplotype that remains intact along the X-chromosome to provide additional information to resolve the evolutionary history G6PD A-. For an allele subject to recent positive selection, the strength of selection will affect the observed patterns of LD. Our 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 (2000-3000 years, assuming a 20 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. 2001). 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 HbS (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. 2001; Saunders et al. 2001). The rapid increase in frequency of G6PD A- under strong selection 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 genomic region is gene-rich, and it follows that selection at *G6PD* has the potential to increase the population level frequency of otherwise rare linked neutral or functional polymorphisms *via* hitchhiking.

ACKNOWLEDGMENTS

We thank J. Kim, D. Garrigan, A. Injap, and S. Peterson for technical assistance. Analyses of the age of the allele and strength of selection were performed in collaboration with Montgomery Slatkin and Chad Garner from the department of Integrative Biology, University of California, Berkeley. 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. This material is based upon work supported by the National Science Foundation under Grant No. 0206756.

REFERENCES

Allison AC (1956) The sickle cell and haemoglobin C genes in some African populations. Ann. Hum Genet 21:67-89

Allison AC (1964) Polymorphism and natural selection in human populations. Cold Spring Harb Symp Quant Biol 29:137-149

Allison AC (1960) Glucose-6-phosphate dehydrogenase deficiency in red blood cells of East Africans. Nature 186:531

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299-309

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J of Hum Genet 74:1111-1120

Beutler E (1994) G6PD deficiency. Blood 84:3613-3636

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency-spectrum of DNA polymorphisms. Genetics 140:783-796

Cavalli-Sforza LL, Menozzi P, Piazza A (1996) The history and geography of human genes. Princeton University Press, Princeton, New Jersey, USA

Coop G, Griffiths RC (2004) Ancestral inference on gene trees under selection. Theor Popul Biol 66:219-32

Daly MJ, Rioux JD, Schaffner SE, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229-232

Filosa S, Calabro V, Lania G, Vulliamy TJ, Brancati C, Tagarelli A, Luzzatto L, Martini G (1993) *G6PD* Haplotypes spanning Xq28 from *F8C* to red-green color-vision. Genomics 17:6-14

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693-709

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

Garner CP and Slatkin M (2002) Likelihood-based disequilibrium mapping for two-marker haplotype data. Theor Popul Biol 61:153-161

Goldstein DB (2001) Islands of linkage disequilibrium. Nature Genetics 29:109-111

Hamblin MT, Thompson EE, Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70:369-383

Hammer MF, Garrigan D, Wood E, Wilder JA, Mobasher Z, Bigham A, Krenz JG, Nachman MW (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, Beutler E (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 RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994) Evidence for positive selection in the superoxide-dismutase (*sod*) region of *Drosophila melanogaster*. Genetics 136:1329-1340

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159

Huttley GA, Smith MW, Carrington M, O'Brien SJ (1999) A scan for linkage disequilibrium across the human genome Genetics 152: 1711-1722

Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217-222

Kelly JK (1997) A test of neutrality based on interlocus associations. Genetics 146:1197-1206

Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765-777

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (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 RC (1964) Interaction of selection + linkage . I. General considerations - Heterotic models. Genetics 49:49-67

Livingstone FB (1985) Frequencies of hemoglobin variants: Thalassemia, the glucose-6-phosphate dehydrogenase deficiency, G6PD variants and ovalocytosis in human populations. Oxford university press, Oxford, UK

McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304:581-584

Meyer D, Thomson G (2001) How selection shapes variation of the human major histocompatibility complex: A review. Ann Hum Genet 65:1-26

Modiano D, Luoni G, Sirima BS, Simpore J, Verra F, Konate A, Rastrelli E, Olivieri A, Calissano C, Paganotti GM, D'Urbano L, Sanou I, Sawadogo A, Modiano G,

Coluzzi M (2001) Haemoglobin C protects against clinical *Plasmodium falciparum* malaria. Nature 414:305-308

Motulsky AG (1961) Glucose-6-phosphate dehydrogenase haemolytic disease of the newborn, and malaria. Lancet 1:1168

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Nat Acad Sci USA 76:5269-5273

Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Brittenham G, Looareesuwan S, Clark AG, Tokunaga K (2004) Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. Am J Hum Genet 74:1198-1208

Oppenheim A, Jury CL, Rund D, Vulliamy TJ, Luzzatto L (1993) $G6PD_{mediterranean}$ accounts for the high prevalence of G6PD deficiency in Kurdish Jews. Hum Genet 91:293-294

Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, et al. (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 33:382-387

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199-204

Roth E, Schulman S (1988) The adaptation of *Plasmodium falciparum* to oxidative stress in G6PD deficient human erythrocytes. Br J Haematol 70:363-367

Roth EF Jr, Raventos-Suarez C, Rinaldi A, Nagel RL (1983) Glucose-6-phosphate dehydrogenase deficiency inhibits *in vitro* growth of *Plasmodium falciparum*. Proc Natl Acad Sci USA. 80:298-299

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174-175

Ruwende C, Khoo SC, Snow AW, Yates SNR, Kwiatkowski D, Gupta S, Warn P, Allsopp CEM, Gilbert SC, Peschu N, Newbold CI, Greenwood BM, Marsh K, Hill AVS (1995) Natural selection of hemizygotes and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature 376:246-249

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

Sanchez-Mazas A, Djoulah S, Busson M, de Gouville IL, Poirier JC, Dehay C, Charron D, Excoffier L, Schneider S, Langaney A, Dausset J, Hors J (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, Ohuigin C, Takahata N, Klein J (1994) Intensity of natural-selection at the major histocompatibility complex loci. Proc Natl Acad Sci USA 91:7184-7188

Saunders MA, Hammer MF, Nachman MW (2002) Nucleotide variability at *G6PD* and the signature of malarial selection in humans. Genetics 162:1849-1861

Schliekelman P, Garner C, Slatkin M (2001) Natural Selection and resistance to HIV. Nature 411:545-546

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

Slatkin M (2001) Simulating genealogies of selected alleles in a population of variable size. Genet Res 78:49-57

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang RH, et al. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489-493

Swallow DM (2003) Genetics of lactase persistence and lactose intolerance. Ann 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, Yoneyama Y, Miwa S, Yoshida A. (1987) A single nucleotide base transition is the basis of the common human glucose-6-phosphate dehydrogenase variant A(+). Genomics. 1:228-231

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

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human *G6PD*: Recent origin of alleles that confer malarial resistance. Science 293:455-462

Toomajian C, Ajioka RS, Jorde LB, Kushner JP, Kreitman M (2003) A method for detecting recent selection in the human genome from allele age estimates. Genetics 165:287-297

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

Verrelli BC, McDonald JH, Argyropoulos G, Destrol-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA (2002) Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. Am J Hum Genet 71:1112-1128

Vogel F, Motulsky AG (1996) Human Genetics: Problems and Approaches. Springer-Verlag, New York, USA

Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 4:587-597

Wall JD, Przeworski M (2000) When did the human population size start increasing? Genetics 155:1865-1874

Walsh EC, Mather KA, Schaffner SF, Farwell L, Daly MJ, Patterson N, Cullen M, Carrington M, Bugawan TL, Erlich H, Campbell J, Barrett J, Miller K, Thomson G, Lander ES, Rioux JD (2003) An integrated haplotype map of the human major histocompatibility complex. Am J Hum Genet 73:580-590

Watterson GA (1975) Number of segregating sites in genetic models without recombination. Theor Popul Biol 7:256-276

Xu WM, Westwood B, Bartsocas CS, Malcorraazpiazu JJ, Indrak K, Beutler E (1995) Glucose-6-phosphate-dehydrogenase mutations and haplotypes in various ethnic groups. Blood 85:257-263

Table 1: Individuals sampled in study.

G6PD Allele type	Sample	Country	Ethnic/Language Group
G6PD A-:	Ivc01	Ivory Coast	Niger-Congo
	Ivc17	Ivory Coast	Niger-Congo
	* S628	Kenya	Unknown
	Alb77	South Africa	Sotho
	Mgr40	Togo	Niger-Congo
	VA010	USA	African American
	M115	USA	African American
	M241	USA	African American
	VA084	USA	African American
	VA025	USA	African American
	* VA076	USA	African American
	VA085	USA	African American
	VA088	USA	African American
	DKT338	USA	African American
	DKT381	USA	African American
	DKT382	USA	African American
	JK1031	Zaire	Mbuti Pygmy
	Sho07	Zimbabwe	Shona
	Sho18	Zimbabwe	Shona
	* Sho49	Zimbabwe	Shona
G6PD A+:	* JK785	CAR	Biaka Pygmy
	Ivc22	Ivory Coast	Niger-Congo
	* JK1071	Zaire	Mbuti Pygmy
	AU26	Kenya	Unknown
	* JW058	Mali	Malinke
	* Alb27	South Africa	Zulu
	* VA024	USA	African American
	DKT275	USA	African American
	VA012	USA	African American
	JK1058	Zaire	Mbuti Pygmy
	Sho03	Zimbabwe	Shona
G6PD B:	* JK736	CAR	Biaka Pygmy
	* JK741	CAR	Biaka Pygmy
	* Gna02	Ghana	Akan
	* Ivc04	Ivory Coast	Niger-Congo
	Ivc18	Ivory Coast	Niger-Congo
	* Ivc20	Ivory Coast	Niger-Congo

* Ivc23	Ivory Coast	Niger-Congo
* Mka21	Kenya	Maasai
* Koh188	Kenya	Meru
* Mka29	Kenya	Tsumkwe
* JR013	Namibia	Biaka
JR323	Namibia	Bagandu
* LD156	South Africa	Khoisan
* Alb74	South Africa	Unknown
* DKT331	USA	African American
* JK1029	Zaire	Mbuti
* JK1033	Zaire	Ituri
* Sho14	Zimbabwe	Shona
* Sho30	Zimbabwe	Shona
* Sho46	Zimbabwe	Shona

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

Table 2: Summary statistics of nucleotide variability for an African CRS.

Locus	X-chromosome Map Location (Kb)	Length[a] (bp)	S[b]	θ_π (SD) (%)	θ_W (SD) (%)	Tajima's D	Fu and Li's D	Divergence Homo-Pan (SD) (%)	Divergence Homo-Pongo (SD) (%)
G18MC	132610								
[c]Total sequence		1795	3	0.035 (0.007)	0.044 (0.028)	-0.50428	-1.40049	0.446 (0.157)	2.062 (0.336)
G1.5MC	149833								
[c]Total sequence		2292	9	0.087 (0.011)	0.103 (0.046)	-0.49687	-1.42383	1.140 (0.211)	2.695 (0.322)
BGN	150356								
Total sequence		2889	19	0.123 (0.014)	0.172 (0.066)	-1.01429	-0.56240	1.148 (0.202)	2.819 (0.335)
Introns only		2160	17	0.155 (0.016)	0.206 (0.080)	-0.88308	-0.46288	1.311 (0.251)	3.258 (0.428)
IDH3G	150642								
Total sequence		3036	11	0.054 (0.011)	0.095 (0.040)	-1.42553	-0.96056	0.841 (0.192)	2.985 (0.344)
Introns only		2771	10	0.057 (0.012)	0.095 (0.041)	-1.31292	-0.65798	0.807 (0.195)	3.024 (0.359)
L1CAM	150720								
Total sequence		3691	8	0.026 (0.008)	0.057 (0.026)	-1.71820	-1.72003	0.678 (0.135)	2.351 (0.372)
Introns only		2087	5	0.031 (0.011)	0.063 (0.033)	-1.42969	-1.39143	0.976 (0.207)	2.904 (0.523)
TAZ	151233								
Total sequence		3144	7	0.024 (0.007)	0.058 (0.028)	-1.81851*	-2.07827	0.539 (0.135)	2.506 (0.286)
Introns only		2906	7	0.026 (0.007)	0.063 (0.030)	-1.81851*	-2.07827	0.580 (0.145)	2.722 (0.310)
G6PD	151347								
Total sequence		5101	18	0.075 (0.008)	0.092 (0.036)	-0.68013	-0.3533	0.959 (0.136)	2.999 (0.240)
Introns only		2925	10	0.068 (0.010)	0.090 (0.039)	-0.79419	-0.1418	1.195 (0.201)	3.970 (0.363)
GAB3	151543								
[c]Introns only		2983	4	0.030 (0.006)	0.035 (0.020)	-0.39740	-0.89691	0.741 (0.157)	2.347 (0.279)
F8	151678								

Total sequence		2408	2	0.012 (0.005)	0.022 (0.016)	-0.93473	0.82564	0.249 (0.124)	0.623 (0.160)
Introns only		1944	2	0.015 (0.006)	0.027 (0.020)	-0.93473	0.82564	0.349 (0.174)	0.721 (0.192)
G0.9MT	152226								
[c]Total sequence		3082	3	0.024 (0.004)	0.026 (0.016)	-0.12971	-0.21602	0.540 (0.133)	1.898 (0.290)

N=26 individuals were used in all CRS analyses (see Materials and Methods). Divergence estimates are based on a comparison between a single randomly chosen allele (Ivc23) and a chimpanzee or orangutan allele. Map position on human X-chromosome is the center of the given window surveyed based on NCBI human map viewer version June 8, 2003.

[a]Number of bp analyzed.

[b]Number of segregating sites.

[c]Loci contained no coding sequence.

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 and *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 mutations 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 online appendix (http://eebweb.arizona.edu/faculty/nachman/saunders/publications/G6PD_AFR/sites_table_AFR.html). 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 (θ π), (b) haplotype diversity, (c) Tajima's D for CRS.

Figure 4: Segregating sites and patterns of LD at *G6PD* and flanking loci surveyed. 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 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.

Figure 1:

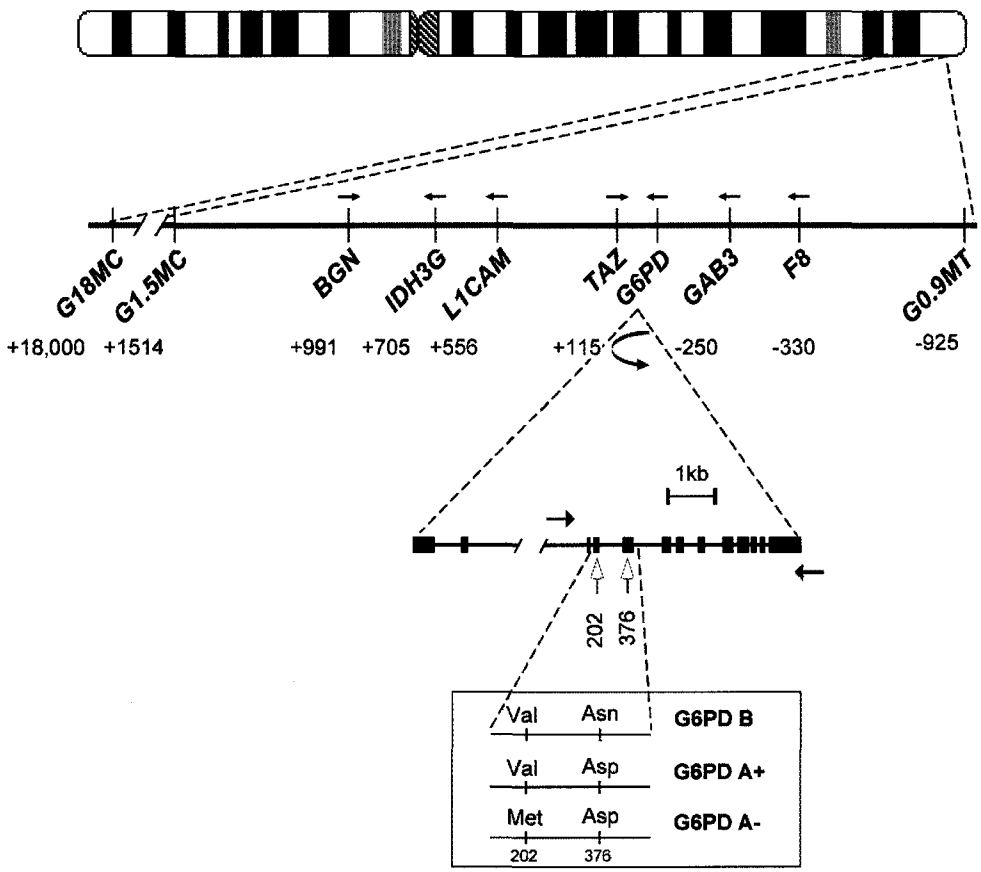

G18MC
G1.5MC
BGN
IDH3G
L1CAM
TAZ
G6PD
GAB3
F8
G0.9MT
+18,000
+1514
+991
+705
+556
+115
-250
-330
-925
1kb
202
376
Val
Asn
G6PD B
Val
Asp
G6PD A+
Met
Asp
G6PD A-
202
376

Figure 2:

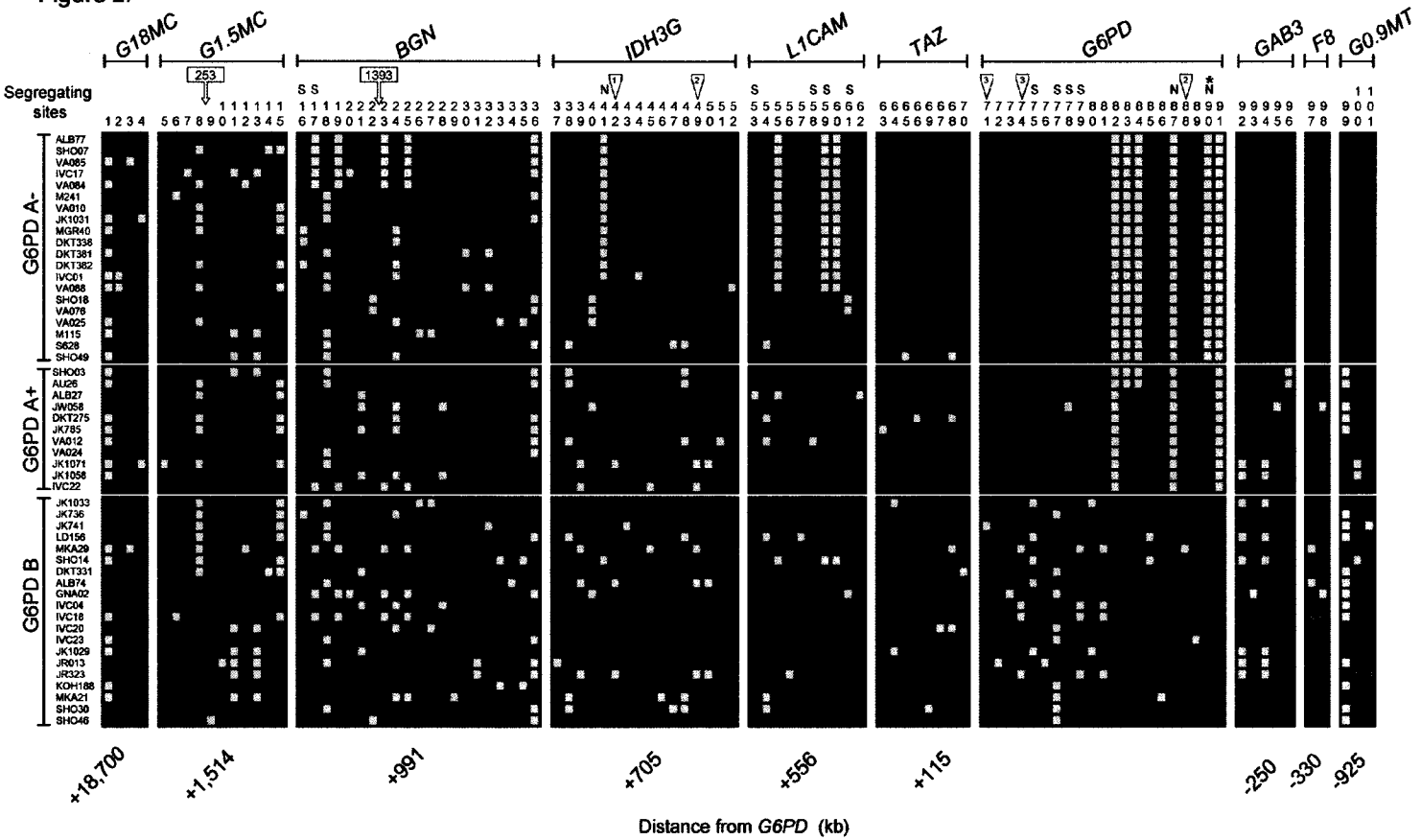

G18MC
G1.5MC
BGN
IDH3G
L1CAM
TAZ
G6PD
GAB3
F8
G0.9MT
253
1393
Segregating sites
G6PD A-
G6PD A+
G6PD B
ALB77
SHO07
VA085
IVC17
VA084
M241
VA010
JK1031
MGR40
DKT338
DKT381
DKT382
IVC01
VA088
SHO18
VA076
VA025
M115
S628
SHO49
SHO03
AU26
ALB27
JW058
DKT275
JK785
VA012
VA024
JK1071
JK1056
IVC22
JK1033
JK736
JK741
LD156
MKA29
SHO14
DKT331
ALB74
GNA02
IVC04
IVC18
IVC20
IVC23
JK1029
JR013
JR323
KOH188
MKA21
SHO30
SHO46
+18,700
+1,514
+991
+705
+556
+115
-250
-330
-925
Distance from G6PD (kb)

Figure 3:

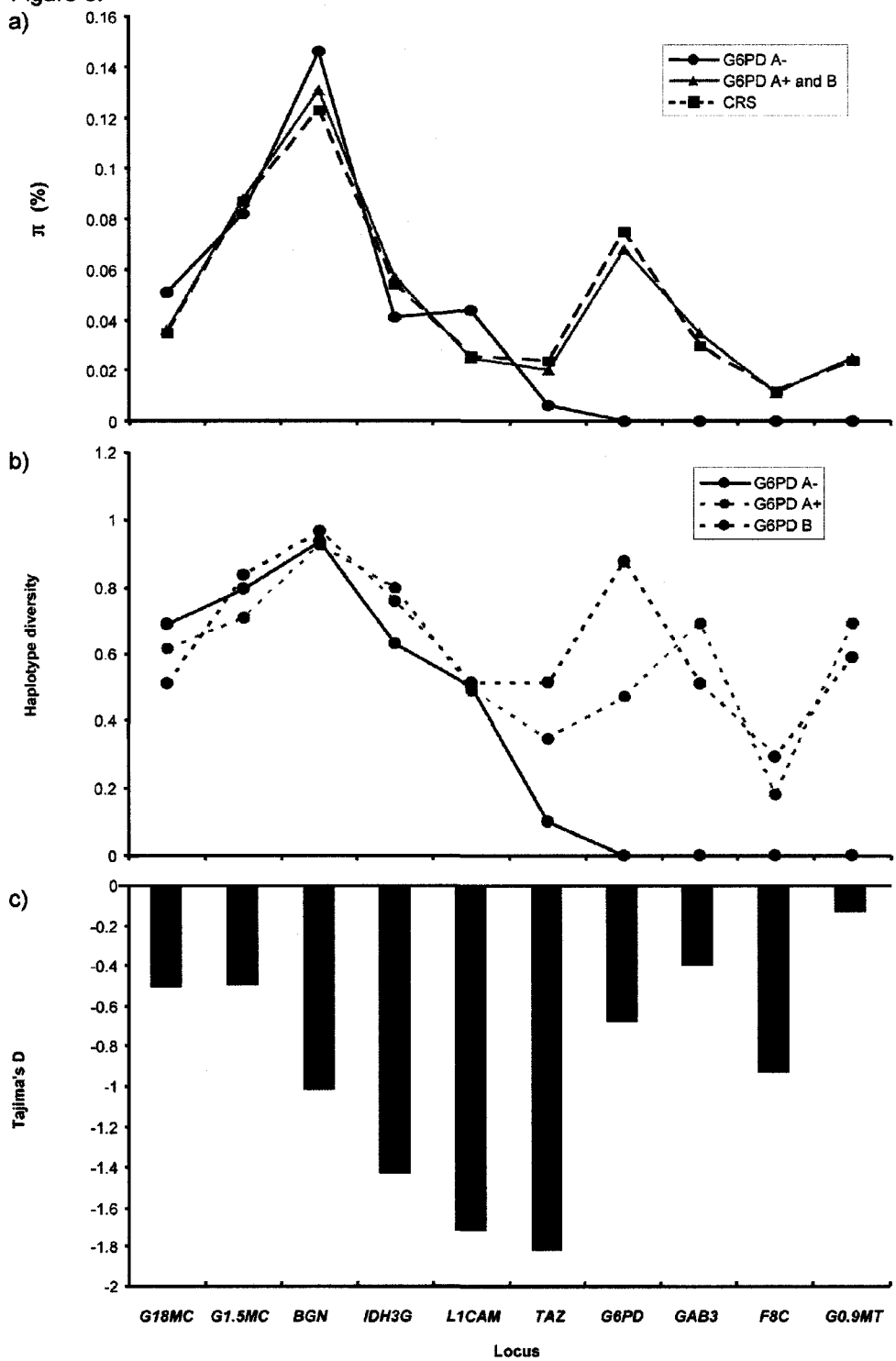

a)
π (%)
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
G6PD A-
G6PD A+ and B
CRS
b)
Haplotype diversity
1.2
1
0.8
0.6
0.4
0.2
0
G6PD A-
G6PD A+
G6PD B
c)
Tajima's D
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
-2
G18MC
G1.5MC
BGN
IDH3G
L1CAM
TAZ
G6PD
GAB3
F8C
G0.9MT
Locus

Figure 4:

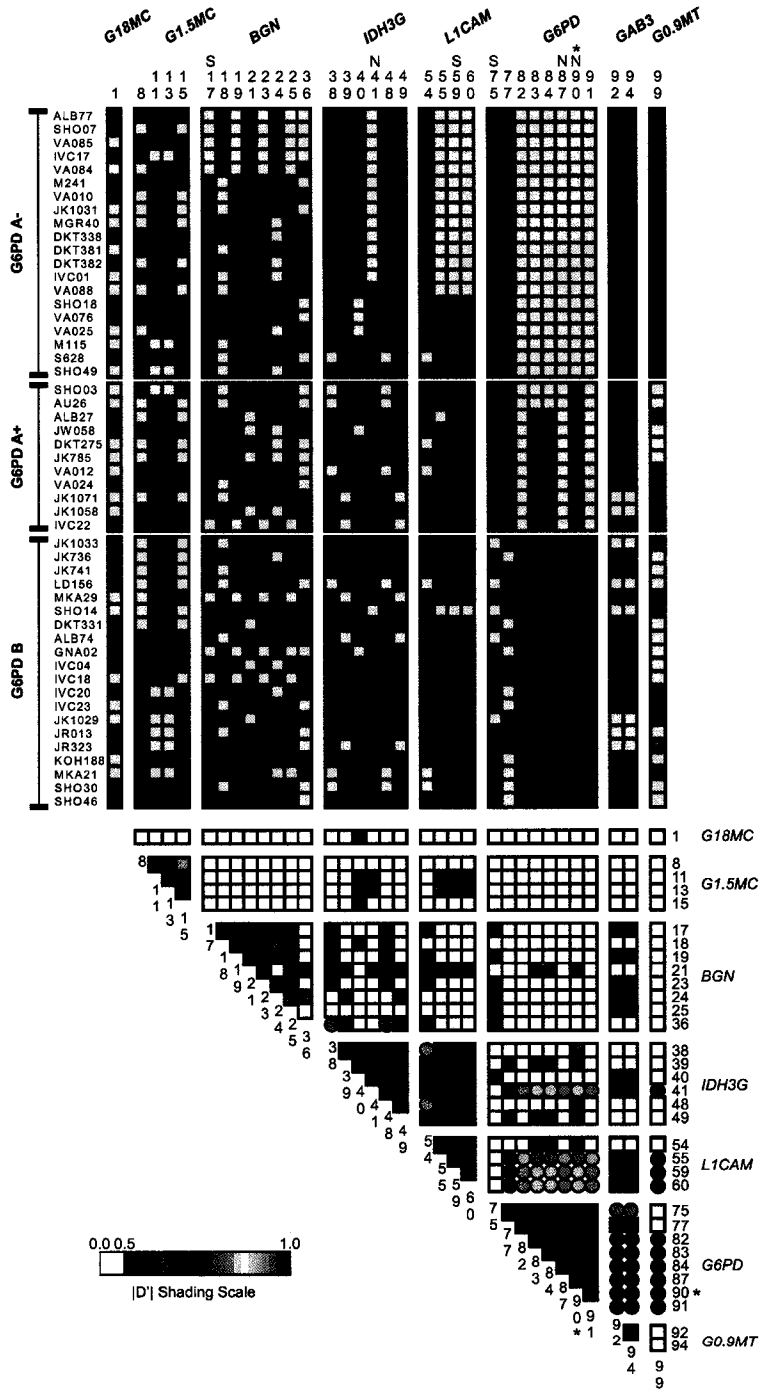
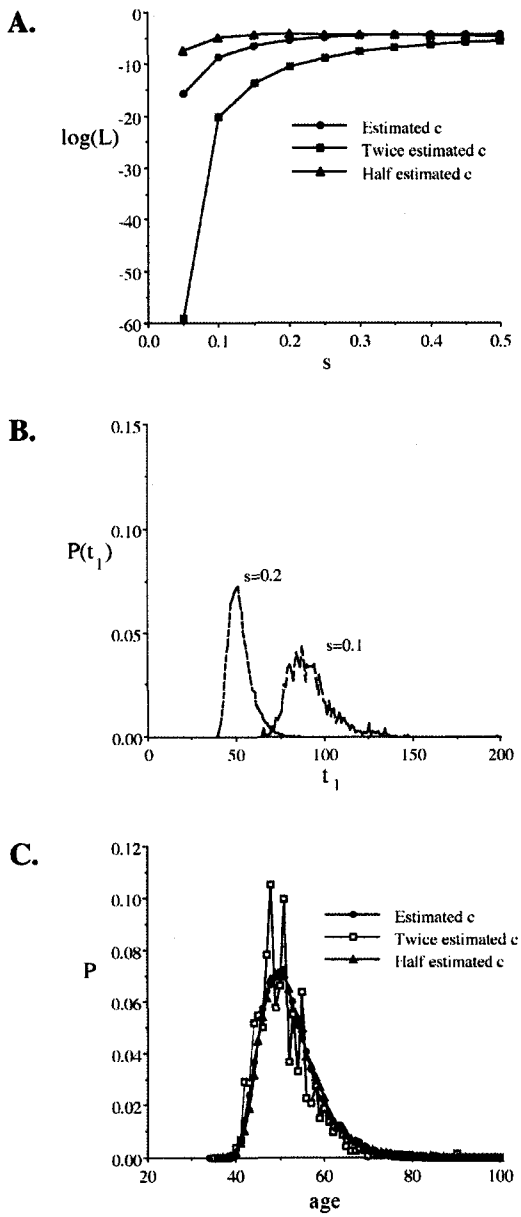

G18MC
G1.5MC
BGN
IDH3G
L1CAM
G6PD
GAB3
G0.9MT
G6PD A-
G6PD A+
G6PD B
ALB77
SHO07
VA085
IVC17
VA084
M241
VA010
JK1031
MGR40
DKT338
DKT381
DKT382
IVC01
VA088
SHO18
VA076
VA025
M115
S628
SHO49
SHO03
AU26
ALB27
JW058
DKT275
JK785
VA012
VA024
JK1071
JK1058
IVC22
JK1033
JK736
JK741
LD156
MKA29
SHO14
DKT331
ALB74
GNA02
IVC04
IVC18
IVC20
IVC23
JK1029
JR013
JR323
KOH188
MKA21
SHO30
SHO46
0.0 0.5 1.0
|D'| Shading Scale

Figure 5


A.
log(L)
0
-10
-20
-30
-40
-50
-60
0.0
0.1
0.2
0.3
0.4
0.5
s
Estimated c
Twice estimated c
Half estimated c
B.
P(t1)
0.15
0.10
0.05
0.00
0
50
100
150
200
t1
s=0.2
s=0.1
C.
P
0.12
0.10
0.08
0.06
0.04
0.02
0.00
20
40
60
80
100
age
Estimated c
Twice estimated c
Half estimated c

APPENDIX C: EXTENDED HAPLOTYPES OF G6PD$_{mediterranean}$ AND THE EVOLUTIONARY HISTORY OF RESISTANCE TO MALARIA IN EURASIA

ABSTRACT

The most common glucose-6-phosphate dehydrogenase (G6PD) deficiency allele in the Mediterranean region and in the Middle East is G6PD$_{mediterranean}$ (C563T). This allele confers resistance to malaria and is therefore maintained in many populations at relatively high frequencies by natural selection despite the severe clinical anemia that is associated with it. Previous work suggested that this allele arose twice in humans, although definitive evidence was lacking. Here we provide a test of this hypothesis. We also provide an analysis of the effects of selection on nucleotide variability, and we estimate the age of G6PD$_{mediterranean}$ (G6PD$_{med}$). We studied DNA nucleotide variability in 21 G6PD$_{med}$ chromosomes and 23 normal G6PD chromosomes, by resequencing 4 windows of ~ 3 kb each around *G6PD*, and a window of ~ 4.7 kb from *G6PD* itself. Despite relatively high background levels of linkage disequilibrium (LD) and low levels of nucleotide variability, we observed reduced heterozygosity and increased LD associated with G6PD$_{med}$ alleles compared to normal alleles spanning a distance of ~ 1 Mb around *G6PD*. Intermittent SNPs were also genotyped in the same panel to define long-range haplotypes. Two distinct ancestral G6PD$_{med}$ long-range haplotypes are significantly conserved among a majority of the selected alleles of common descent spanning > 1.6 Mb. These long-range haplotypes provide strong evidence that the

G6PD$_{med}$ C563T mutation arose twice: once within, and once outside the Indian subcontinent. These two mutations seem to have arisen at roughly the same time within the past 3500 years.

INTRODUCTION

Malaria is a major morbidity factor in tropical and temperate zones throughout the world (Breman et al. 2004). Although most deaths by malaria occur in sub-Saharan Africa, the disease is also endemic in parts of the Middle East and around the Mediterranean Sea. *Plasmodium falciparum*, the most virulent species of *Plasmodium* that infects humans, is believed to have become a significant virulence factor within the past 10,000 years (Coluzzi et al. 2002; Joy et al. 2003). In response to strong selection from malaria, some human populations exhibit resistance alleles of different genes at high frequencies (*e.g.* β-globin S and G6PD deficiency alleles: Miller 1994).

Many different G6PD deficiency alleles have been described at the molecular level and are caused by mutations that disrupt the normal function of this housekeeping enzyme (Kwok et al. 2002). The pathological manifestations of G6PD deficiency include neonatal jaundice and hemolytic anemia as well as other clinical conditions (Beutler 1994; Sirugo et al. 2004). While most G6PD deficiency alleles represent rare mutations, several alleles are known to reach population frequencies ranging from 0.05 to 0.65 (Livingstone 1985; Oppenheim et al. 1993). Interestingly, these different alleles are generally distributed across specific geographical ranges. For example, G6PD A-(G202A) is found primarily in African populations, and $G6PD_{med}$ is found primarily in populations around the Mediterranean Sea and the Middle East. Several lines of evidence suggest that G6PD deficiency confers resistance to severe *falciparum* malaria. First, there is a strong geographic correlation between populations with high G6PD deficiency frequencies and the distribution of malaria (Allison 1960; Motulsky 1961). Second, *in*

vitro evidence has demonstrated reduced parasite growth in G6PD$_{med}$ erythrocytes relative to normal erythrocytes (Roth et al. 1983). Third, a large-scale epidemiological study demonstrated that G6PD A- alleles are underrepresented among patients that suffer from severe malaria relative to the respective healthy population (Ruwende et al. 1995). Finally, recent population genetics studies have uncovered patterns of nucleotide variability spanning the *G6PD* genomic region (Xq28) that are consistent with the effects of strong recent positive selection at *G6PD* (Tishkoff et al. 2001; Sabeti et al. 2002; Saunders et al. 2002; Verrelli et al. 2002; Saunders APPENDIX B). However, the bulk of this population genetic evidence is described with respect to G6PD A- alleles.

G6PD$_{med}$ is characterized by a single mutation (C→T) at coding site 563 that causes a Ser → Phe change (Figure 1) which reduces enzyme efficiency to 5% of normal (Vulliamy et al. 1988). G6PD$_{med}$ (C563T) is commonly found at a frequency of 0.05 to 0.2 in many Mediterranean and Middle Eastern populations (Livingstone 1985). In most populations, G6PD$_{med}$ alleles bear a silent C→T change at *G6PD* coding site 1311. However, in the Indian subcontinent a large proportion of G6PD$_{med}$ alleles do not bear the C1311T mutation (Beutler and Kuhl 1990; Saha et al. 1994; Figure 1). This observation prompted the hypothesis that the mutation C563T arose twice in humans: once in the Indian subcontinent, and once in the Mediterranean/Middle east. Although the observation of two different alleles at site 1311 associated with G6PD$_{med}$ may be explained by independent origins of the C563T mutation, the results are also consistent with a possible intragenic recombination event between a G6PD$_{med}$ allele bearing 1311T and another allele bearing a 1311C polymorphism.

In this study we have determined the long-range extended haplotypes of G6PD$_{med}$ alleles (spanning ~ 2 Mb roughly centered on *G6PD*) in a panel from across the Middle East and the Indian sub-continent. Our data provide compelling evidence for an independent origin of a G6PD$_{med}$ (C563T) allele once in the Mediterranean/Middle East (G6PD$_{med\ MME}$), and once in the Indian sub-continent (G6PD$_{med\ IND}$) within the past 3500 years. These data also indicate that the age of G6PD$_{med\ MME}$ is not significantly younger than the African allele G6PD A-, as has been previously suggested.

SUBJECTS AND METHODS

Samples: We identified individuals with G6PD$_{med}$ (C563T) alleles using *MboII* restriction fragment polymorphism (RFLP) analysis. The primers "Oligo B" (5'-ACT CCC GAA GAG GGG TTC AAG G-3') and "Oligo J" (5'-GGT CGG AGG GTC CTC TCT CCT TC-3') from Kurdi-Haidar *et al.* (1990) were used to amplify a 547 bp PCR amplicon encompassing *G6PD* coding site 563. When cut with *MboII*, amplicons bearing the G6PD$_{med}$ mutation (C563T) display fragments of 277, 100, 119, 26 and 25 bp, while control amplicons display fragments of 377, 119, 26 and 25 bp. Samples were not screened *a priori* for the C1311T mutation. Based on RFLP results, we selected 21 G6PD$_{med}$ individuals from Greece (n=2), Iran (n=3), Iraqi Jews (n=4), Pakistan (n=7) and Syria (n=5) (Table 1). This panel includes individuals from across the typical geographic range of G6PD$_{med}$. We selected an additional 23 individuals bearing the normal G6PD allele (*i.e.* G6PD B) based on a similar geographic representation as the G6PD$_{med}$ individuals described above, to use as a control group (Table 1). Since *G6PD* is X-linked, we used only males to enable unambiguous phase resolution across the X-chromosome. All samples were collected with consent and were used as approved by the committee of human subjects at the University of Arizona.

Loci sampled: We resequenced a ~ 4.7 kb region from *G6PD* for all individuals. The surveyed region includes most of the coding region of *G6PD* and intervening sequence, including coding site 563 and 1311 (Figure 1). Primers and PCR conditions for resequencing *G6PD* were used as described by Saunders *et al.* (2002). To examine patterns of nucleotide variability around *G6PD*, we resequenced contiguous windows at 4

additional loci around *G6PD*: *BGN* (2889 bp), *IDH3G* (3036 bp), *GAB3* (2983 bp) and 3082 bp of an intergenic region (*G0.9MT*) as described in Saunders (APPENDIX B) (Figure 1). These loci were selected based on their physical distance from *G6PD* to serve as neutral markers, and none of these loci are known to be subject to recent positive selection themselves. All sequences were submitted to Genbank under accession numbers XXXXXXXX-XXXXXXXX. Additionally, we genotyped several SNPs at varying distances surrounding *G6PD* to define long-range haplotypes for *G6PD* alleles and to measure linkage disequilibrium (Figure 1; Table 2). These SNPs were selected from the *SNP Browser* software (Applied Biosystems*)* based on physical distance from *G6PD*, and a criterion of a minor allele frequency in a healthy Caucasian panel (*Celera*) of ≥ 0.10. SNP genotyping was performed as described by *TaqMan® Assays-on-Demand(TM)* protocols (*Applied Biosystems*).

Data Analysis: Because it has been proposed that the G6PD$_{med}$ C563T mutation arose twice (Beutler and Kuhl 1990), after identifying the *G6PD* 1311 allele state for the G6PD$_{med}$ samples by resequencing, we analyzed separately two sub-classes of G6PD$_{med}$ alleles: (*i*) G6PD$_{med\ IND}$ alleles (that bear 563T and 1311C), and (*ii*) G6PD$_{med\ MME}$ alleles (that bear 563T and 1311T) (Figure 1). We estimated nucleotide diversity using θ_π (Nei and Li 1979), θ_w (Watterson 1975), and haplotype diversity (H_d) at each resequenced locus for three subsets of the data: G6PD B alleles, G6PD$_{med\ MME}$ alleles, and G6PD$_{med\ IND}$ alleles. In a random sample of a population at mutation-drift equilibrium, the parameters θ_π and θ_w are estimators of the neutral parameter $\Theta = 3N_e\,\mu$ for X-linked loci, where N_e is the effective population size and μ is the mutation rate. This condition is

clearly not met for our sample, which was conditioned on the presence of particular alleles. Thus, θ_π and θ_w are presented as summaries of the amount of genetic variation, but not as accurate estimators of $3N_e\,\mu$. We also calculated Tajima's D (Tajima 1989) and Fu and Li's D (Fu and Li 1993) at each resequenced locus. These two statistics describe the frequency distribution of mutations in a data set; under neutral equilibrium conditions for a random sample, these test statistics should be equal to zero. Significant deviations from zero may be indicative of selection or changes in population size. However results of these tests should be interpreted with caution particularly with human data sets, because present day human populations are not expected to be at equilibrium (Wooding et al. 2004).

We calculated |D'| (Lewontin 1964) for all pairwise comparisons of segregating sites for subsets of the data: (*i*) all individuals, (*ii*) G6PD B and $G6PD_{med\ MME}$ individuals, and (*iii*) G6PD B and $G6PD_{med\ IND}$ individuals. |D'| is the standardized measure of the disequilibrium parameter D, and it ranges from 0 (complete equilibrium) to 1 (complete disequilibrium). Significance of LD measures was estimated using Fisher's exact test. All analyses of nucleotide variability and LD were performed using *dnaSP* 4.0 (Rozas and Rozas 1999).

The allele age and the selection coefficient of $G6PD_{med\ MME}$ were estimated using the method described by Saunders (APPENDIX B) for estimating the age of the African G6PD A- allele. In brief, we utilized a coalescent based program (Slatkin 2001) in combination with a maximum-likelihood "two-locus" method (Garner and Slatkin 2002) to estimate a posterior distribution of the allele age given likely selection coefficients (*s*).

We utilized the observed proportion of G6PD$_{med\ MME}$ chromosomes that bear the ancestral long-range haplotype (LRH) at *IDH3G* and at *SNP J*. At *IDH3G* the ancestral G6PD$_{med\ MME}$ LRH is conserved in 10/16 chromosomes, and at *SNP J* 15/16 chromosome still bear the G6PD$_{med\ MME}$ LRH. Accordingly, the input to the program was 10 *AMB*, 0 *AMb*, 5 *aMB*, and 1 *aMb* [in the notation of Garner and Slatkin (2002)] for G6PD$_{med\ MME}$, where *A* and *a* are the minor and major alleles respectively found at site *Id07* (Figure 2), *B* and *b* are the common and rare alleles respectively at *SNP J*, and *M* represents *G6PD* coding site 563 (*Gd03* in Figure 2). The respective population frequencies of G6PD B haplotypes for the "two-locus" data are 0.095, 0.143, 0.286, and 0.476 (n=21). Genetic distances were determined from University California-Santa Cruz human genome browser based on estimates of Kong *et al.* (2002): $c = 0.0141$ and 0.00234 Morgans for *A-M* and *M-B*, respectively. The present day (t=0) population frequency of G6PD$_{med\ MME}$ used to constrain the coalescent runs was $q = 0.05$.

RESULTS

Long-range haplotypes and linkage disequilibrium: We have generated phased SNP data to define long-range haplotypes (LRHs) among individuals bearing the alleles G6PD B and G6PD$_{med}$ (C563T) from across the Middle East and the Mediterranean region. Our panel includes 7 G6PD$_{med}$ (C563T) samples from Pakistan (Table 1), which are from a region where G6PD$_{med}$ 1311C samples have been previously described (Saha et al. 1994). Five of the Pakistani G6PD$_{med}$ samples bear the 1311C allele, while the remaining two Pakistani individuals bear the 1311T allele (site *Gd02* in Figure 2). Although our sample from Pakistan is small (n = 7), the observation of five 1311C and two 1311T alleles among G6PD$_{med}$ alleles from the Indian sub-continent is in general agreement with the ratio of 78:23 seen in a larger Indian sample of G6PD$_{med}$ alleles (n=101: Sukumar et al. 2004). None of the non-Pakistani G6PD$_{med}$ individuals bear the 1311C state, consistent with previous results showing that the 1311C polymorphism is rarely found among G6PD$_{med}$ individuals outside of the Indian subcontinent (Beutler and Kuhl 1990).

A remarkable feature of our data is the difference between the LRH of G6PD$_{med}$ alleles that bear the 1311T mutation and the LRH of G6PD$_{med}$ alleles that bear the 1311C mutation. The common (ancestral) G6PD$_{med}$ LRH of the non-Pakistani alleles is unambiguously defined in the region between *IDH3G* and *G0.9MT* (for example sample Grc865 in figure 2). The common LRH of the 5 Pakistani G6PD$_{med}$ individuals that bear 1311C (for example Pak349 in Figure 2) is distinctly different from the ancestral LRH that is common to all other G6PD$_{med}$ individuals. Importantly, these two long-range

haplotypes differ from one another on both sides of *G6PD*, and thus cannot easily be explained by recombination. The presence of the C563T mutation at *G6PD* on two distinct long-range haplotypes strongly suggests that this mutation arose independently twice (although gene conversion might also explain this pattern: see DISCUSSION). The inference of independent origins for the two G6PD$_{med}$ alleles warrants separate evolutionary analyses for G6PD$_{med\ MME}$ and G6PD$_{med\ IND}$.

Ten of the sixteen G6PD$_{med\ MME}$ alleles share an identical haplotype over a region that spans ~ 1.6 Mb from *IDH3G* to *G0.9MT* (Figures 2 and 3). This common LRH decays between *IDH3G* and *BGN*. The remainder of the G6PD$_{med\ MME}$ alleles share a smaller portion of this ancestral haplotype centered on *G6PD*; the shortest extent of conservation of the ancestral G6PD$_{med\ MME}$ LRH spans ~ 400 kb (from *SNP C* to *GAB3*) in individual Pak218 (Figure 3). In contrast, only 6 of 23 G6PD B alleles can be assigned a common LRH that spans 1.4 Mb. The remainder of the B alleles share smaller portions of this haplotype around *G6PD*. The largest region common to all B alleles spans only several kb within *G6PD* (Figure 3). A major common LRH of G6PD$_{med\ IND}$ alleles spans $\geq$ 1.9 Mb (Figure 3).

We calculated all pairwise values of |D'| for a sample composed of both G6PD B alleles (n=23) and G6PD$_{med\ MME}$ (n=16) (Figure 4). All informative intragenic pairwise values are high, with |D'| = 1 except for the *Bg08* - *Bg01* pairwise value of D' = 0.867 (Figure 4), indicating a paucity of intragenic recombination events. At the intergenic level (*i.e.* long-range) we see significant values of |D'| (by Fisher's exact test) ranging up to a distance of 925 kb (Figure 4: *Gd03* - *Mt02* pairwise |D'| = 1). Overall we see

significant LD between *G6PD* site 563 (*Gd03*) and sites up to 710 kb telomeric to *G6PD* (*SNP A*), and sites found up to 925 kb centromeric to *G6PD* (*Mt02*). When pairwise values of |D'| are calculated for B alleles alone, the longest range of significant pairwise LD is between sites *SNP D* and *SNP I* spanning ~ 335 kb (data not shown). The small number of G6PD$_{med\ IND}$ alleles prevented us from conducting robust analysis of pairwise LD. Nevertheless, significant values of |D'| appear to extend at least as far as seen with G6PD$_{med\ MME}$ alleles especially on the centromeric side (*e.g.* |D'| = 1 for *Gd03* - *Id01*; Figure 2).

Nucleotide Diversity: We calculated measures of nucleotide variability for subsets of the total panel based on the different G6PD alleles: G6PD B, G6PD$_{med\ MME}$ and G6PD$_{med\ IND}$. Our analysis of the G6PD B alleles alone may serve as a proxy for nucleotide variability in a typical Middle-Eastern population sample because G6PD$_{med}$ is typically found at a population frequency of only ~ 0.05. Analyses of a constructed random sample (Hudson et al. 1994) that includes only few G6PD$_{med}$ alleles is not significantly different in overall patterns of nucleotide variability from those estimates that are based on G6PD B alleles alone (data not shown).

Among the B alleles θ_w is 0.125, 0.059, 0.016, 0.009, and 0.026%, for *BGN*, *IDH3G*, *G6PD*, *GAB3* and *G0.9MT* respectively (Table 3). These values are lower than the respective values for the same loci in an African sample (Saunders APPENDIX B), and within the range of 15 other X-linked loci in non-African populations (Hammer et al. 2004). Nucleotide variability among G6PD$_{med\ MME}$ and G6PD$_{med\ IND}$ alleles is $\theta_w = 0$ at *G6PD*, and is lower among G6PD$_{med\ MME}$ and G6PD$_{med\ IND}$ alleles than among the B

alleles at *IDH3G*, *GAB3* and *G0.9MT*. Yet at *BGN* the value of θ_w for G6PD$_{med\ MME}$ and G6PD$_{med\ IND}$ is 0.126 and 0.133, respectively, similar to the value at *BGN* for the B alleles. Values of θ_π show a similar pattern of reduced variability for G6PD$_{med}$ alleles relative to G6PD B alleles (Table 3). None of the values of Tajima's D or Fu and Li's D show significant departures from neutral equilibrium expectations.

To consider the genotyped SNPs (Table 2) in context with the nucleotide variability of the resequenced regions, we calculated haplotype diversity (H_d) within each class of alleles at each of the genotyped SNPs and the resequenced loci (Table 3). Patterns of haplotype diversity are considerably reduced for G6PD$_{med\ MME}$ and G6PD$_{med\ IND}$ relative to G6PD B alleles (Figure 5) except at the fringes of the surveyed region, *BGN* and *G0.9MT*. We note that the small sample size of G6PD$_{med\ IND}$ alleles (n=5) creates a high standard deviation especially on these estimates of haplotype diversity.

Age of the alleles and strength of selection: Using long-range haplotype data from both sides of *G6PD* (*i.e.* "two-locus" data as described in SUBJECTS and METHODS; Garner and Slatkin 2004) we estimated the age and selection coefficient of G6PD$_{med\ MME}$ using a maximum-likelihood method. The likelihood curve as a function of the selection coefficient (s) for G6PD$_{med\ MME}$ is shown in Figure 6a. These results suggest that the likely range of the selection coefficient is > 0.1. However, values of $s \sim 0.3$ seem biologically improbable for the G6PD$_{med\ MME}$ allele in comparison with selection coefficients of other advantageous alleles in humans that are typically lower than 0.3 (*e.g.* HbS: Allison 1954; G6PD A-: APPENDIX B). Thus our analyses for the age of the allele were based on selection coefficients of $s = 0.1$ and $s = 0.2$. The age of G6PD$_{med}$

$_{MME}$ (estimated by the posterior distribution based on *s*) is 75-125 generations with an upper bound of ~ 175 generations (Figure 6b). We were unable to produce a robust analysis of the age of the allele G6PD$_{med\ IND}$ with the small sample size (n=5) in this study; However, the LRH of G6PD$_{med\ IND}$ appears to extend a similar distance as G6PD$_{med\ MME}$, implying that the age of these two alleles is approximately the same.

DISCUSSION

We generated SNP data spanning a region of ~ 2 Mb roughly centered on *G6PD* in a panel of individuals that includes G6PD$_{med}$ alleles and G6PD B alleles from diverse localities in the Mediterranean region and the Middle East. Our analysis of nucleotide variability and of long-range haplotypes sheds light on the evolutionary history of selection for resistance to malaria in this region, and also contributes to our general understanding of nucleotide variability and LD in the human genome. We find clear evidence that the C563T mutation (which defines G6PD$_{med}$) exists on two distinct LRHs. This implies that the C563T mutation arose twice in recent human history, or a recent gene conversion event occurred encompassing site *G6PD* site 563. We also document the effects of selection at *G6PD* on both levels of nucleotide variability and the extent of LD in Xq28. These patterns suggest that both G6PD$_{med}$ alleles arose about 2000 years ago. Each of these issues is discussed in turn below.

The G6PD$_{med}$ allele arose twice: A striking feature of these data is the difference between the LRH shared among most of the Pakistani G6PD$_{med}$ alleles and the LRH in the remainder of the G6PD$_{med}$ alleles. The mutation that defines G6PD$_{med}$ (*i.e.* C563T) has been associated with the derived common silent polymorphism at *G6PD* coding site 1311 (*i.e.* C1311T) in individuals from most parts of the world (Beutler and Kuhl 1990: Figure 1). However surveys of the Indian sub-continent (*i.e.* India and Pakistan) report common occurrences of G6PD$_{med}$ alleles that bear the 1311C allele (Saha et al. 1994; Kaeda et al. 1995; Sukumar et al. 2004; Figure 1). This pattern was first reported by Beutler and Kuhl (1990) and was interpreted as evidence for multiple

independent origins of G6PD$_{med}$. However, the observation of two different G6PD$_{med}$ intragenic haplotypes could also be parsimoniously explained by an intragenic recombination event (likely in the Indian sub-continent) between a "typical" G6PD$_{med}$ chromosome and another chromosome that harbored the 1311C mutation. Using our LRH data we are now able to exclude the possibility of a single intragenic recombination event between site 563 and 1311 within *G6PD* as an explanation for the two different G6PD$_{med}$ intragenic haplotypes. All G6PD$_{med}$ individuals that bear the 1311C allele in this study have a distinctly different LRH than the ancestral LRH of all other G6PD$_{med}$ chromosomes (Figure 2). Although this pattern could also be obtained by a double recombination event around site 563, this hypothesis is not probable especially given the young age of the G6PD$_{med}$ allele (see below). Two hypotheses remain likely to explain the pattern observed in LRHs: (*i*) A gene conversion event occurred that encompassed site 563 between a G6PD$_{med}$ chromosome and a G6PD B chromosome with a 1311C allele, or (*ii*) the G6PD mutation C563T arose independently once in India and once in the Middle-East/Mediterranean, each on distinct haplotypes. Distinguishing between these hypotheses may be impossible given the few segregating sites found within *G6PD*. Although these hypotheses invoke different biological mechanisms, they both suggest that different G6PD$_{med}$ alleles have independent evolutionary histories. Furthermore, because the G6PD$_{med}$ 563C alleles are found nearly exclusively in the Indian sub-continent, the given event likely occurred in this geographic region. Therefore we can effectively regard either hypothesis as an "independent origin" event, warranting that the evolutionary histories of these alleles should be analyzed separately. We refer to these

two distinct G6PD deficiency alleles as $G6PD_{med\ MME}$ (563T, 1311T) and $G6PD_{med\ IND}$ (563T, 1311C).

Evolutionary theory predicts that organisms or populations subject to similar selection pressures are candidates for convergent evolution (Hughes 1999). In fact, the occurrence of different G6PD deficiency alleles at relatively high frequencies in different populations due to malarial selection (*e.g.* G6PD A-, and $G6PD_{med}$) represents an example of convergent evolution at a broad phenotype level. Parallel adaptive evolution for identical mutations is rare in general in nature, however some cases have been noted. For example, lysozymes of ruminants and colobine monkeys display identical amino acid changes which arose independently as an adaptation for foregut fermentation (Stewart et al. 1987). Also, hypotheses have been proposed for parallel evolution in humans for the HbS allele, which confers resistance to malaria in Africa (Wainscoat 1987). However the occurrence of a recombination hotspot near the HbS mutation site (Schneider et al. 2002) complicates the distinction between hypotheses of multiple origins *vs.* multiple HbS recombinants. Mutations that cause disease in humans (but presumably are never advantageous) have also been shown to have multiple independent origins. For example, alleles of Leber's hereditary optic neuropathy (LOHN: Brown et al. 1995) and mtDNA related deafness (Hutchin and Cortopassi 1997) both display cases of recurrent independent mutations. Together, these examples show that independent parallel origins of selected/disease alleles such as that seen here for $G6PD_{med}$, may in fact occur in nature, however uncommonly.

Potential hypermutability of G6PD site 1311: We have presented strong evidence for independent origins of the C563T mutation. However, hypotheses for multiple independent origins of other G6PD deficiency alleles should be examined with caution when the premise is based the C/T polymorphism at site 1311. For example, Beutler *et al.* (1991) defined G6PD$_{viangchan}$ (from India) and G6PD$_{jammu}$ (from a Laotian population) at the molecular level both based on a G→A nonsynonymous mutation at site 871. G6PD$_{viangchan}$ bears the 1311T allele, and G6PD$_{jammu}$ bears the 1311C allele, suggesting again multiple independent origins of a G6PD deficiency allele (*i.e.* recurrent mutation or gene conversion for G871A). As mentioned above, this may also be explained parsimoniously by intragenic recombination, however we propose a third likely explanation: multiple origins of the C1311T mutation. The ancestral state for *G6PD* site 1311 is a C which is part of a CpG dinucleotide sequence. In mammalian genomes CpG sites are hotspots for mutation to TpG because of deamination of 5-methylcytosine (Waters and Swann 2000). The potential hyper-mutability of site 1311 is further suggested by recent results that have shown that ~75% of G6PD$_{kerela\text{-}kaylan}$ (G949A) alleles bear 1311C while the remainder bear 1311T, and ~50% of G6PD$_{orissa}$ (C131G) alleles bear 1311C while the remainder bear 1311T (Sukumar et al. 2004). Also, a single Italian individual with G6PD$_{med}$ has been genotyped with a 1311C allele (Beutler and Kuhl 1990). Without invoking hypermutability of site 1311, these examples of intra-allelic haplotype variability for G6PD deficiency alleles are otherwise difficult to explain given that recombination events are not common in this region of *G6PD* (Saunders et al. 2002; APPENDIX B). As site 1311 is one of the few neutral intermediate frequency

polymorphisms at *G6PD* in non-African populations, further studies of long-range haplotypes for G6PD deficiency alleles other than G6PD$_{med}$ will be required to test this hypothesis of hyper-mutability of the 1311C site.

Nucleotide variability at resequenced regions: As G6PD$_{med\ MME}$ and G6PD$_{med\ IND}$ have indistinguishable phenotypes, both alleles are probably subject to similar effects of selection. To examine how selection on *G6PD* has affected nucleotide variability in Xq28 we estimated measures of nucleotide variability at 4 resequenced windows around *G6PD* (*i.e. BGN, IDH3G, GAB3* and *G0.9MT*) for the given panel. An allele that is subject to positive selection is expected to increase in frequency rapidly and cause a reduction in levels of genetic variability at neutral linked sites (Maynard-Smith and Haigh 1974). The extent of the genomic region of reduced nucleotide variability around the selected variant will depend on the age of the allele, the strength of selection and the local recombination rate (Slatkin and Rannala 2000). However, if pre-existing levels of nucleotide variability among non-selected alleles are low, as expected for non-African populations (Wall and Przeworski 2000), a reduction in the level of nucleotide variability on G6PD$_{med}$ alleles that is due to selection may be difficult to identify.

In order to describe the background levels of nucleotide variability for non-African populations for this genomic region, we first consider only the B alleles in our panel. In general, the level of nucleotide variability for the G6PD B alleles at the loci in this study (Table 3) are within the range of values estimated for 15 other X-linked loci among non-Africans ($0.001 < \theta_{\pi} < 0.129$: Hammer et al. 2004). Specifically for *G6PD*, nucleotide variability is virtually the same in this study ($\theta_{\pi} = 0.017\%$), as in a worldwide

non-African panel of 31 individuals ($\theta_\pi = 0.016\%$; Saunders et al 2002). Furthermore, all the loci in our study show lower nucleotide variability in the panel of B alleles relative to an African sample ($\theta_\pi = 0.155\%$, 0.057%, 0.030% and 0.024% respectively for *BGN*, *IDH3G*, *GAB3* and *G0.9MT* in Africa: Saunders APPENDIX B), consistent with the typical pattern of reduced nucleotide variability in non-Africans relative to Africans. Note that as G6PD deficiency is found at low frequencies in most non-African populations, results of a constructed random sample (*i.e.* including few $G6PD_{med}$ alleles) are virtually indistinguishable from a panel of G6PD B alleles only (data not shown). All tests of neutrality based on the frequency spectrum of alleles (*e.g.* Tajima's D and Fu and Li's D) show non-significant deviations from neutrality for a panel of G6PD B alleles (and a true constructed random sample with $G6PD_{med}$ at a frequency of 0.09, n=25: data not shown). This result is similar to results obtained in Africa, where no deviation from a neutral equilibrium frequency spectrum is observed at *G6PD* (Saunders et al 2002; Sabeti et al. 2002, Verrelli et al. 2002), despite the fact that *G6PD* is known to be under selection.

When only $G6PD_{med}$ ($G6PD_{med\ MME}$ or $G6PD_{med\ IND}$) alleles are examined, we observe complete intra-allelic homogeneity ($\theta_\pi = 0$) at *G6PD*, in contrast to the normal level of intra-allelic variability among the B alleles described above ($\theta_\pi = 0.017$). Nucleotide variability at the other resequenced loci is detected among $G6PD_{med}$ alleles, however variability is still less than that found among G6PD B alleles, except for *BGN*, at a distance of 925 Kb from *G6PD*. At this distance, levels of nucleotide variability between the selected alleles and the normal alleles are indistinguishable one from the

other, indicating that the homogenizing effects of selection have been disrupted by recombination. Interestingly, the homogenizing effect of selection on nucleotide diversity is disrupted at the same distance among G6PD A- alleles in Africa relative to African G6PD B alleles (Saunders APPENDIX B). Noteworthy is the observation that at *IDH3G*, nucleotide variability is not significantly different between G6PD$_{med\ MME}$ and G6PD B, despite the fact that effects of selection seem to extend to this distance based on long-range haplotypes (see below). This apparent discordance is due to the fact that the ancestral G6PD$_{med\ MME}$ haplotype at *IDH3G* includes a relatively diverged intragenic haplotype (Figure 2) that inflates the nucleotide diversity for G6PD$_{med\ MME}$ alleles. In summary, even though general levels of nucleotide variability in non-African populations are relatively low (compared to African diversity), the effect of reduced nucleotide variability among G6PD$_{med}$ alleles due to selection is still seen in the region spanning *G6PD* to at least *G0.9MT*.

Linkage disequilibrium and long-range haplotypes: While standard statistics of nucleotide variability may capture some signature of selection, measures of LD and long-range haplotype structure may be more effective for detecting selection at the molecular level in human data sets (Sabeti et al. 2002; Saunders et al. 2002; Toomajian et al. 2003). Recently selected alleles are expected to share long-range conservation of a single (ancestral) haplotype relative to non-selected alleles, creating long range LD that is associated with the selected polymorphism. Previous studies suggest that the distance over which significant pairwise LD is observed is large in non-African samples relative to Africans (Reich et al. 2001; Wall and Pritchard 2003). The average half value of |D'| at

putatively neutral loci in a sample of Europeans was ~ 60 kb compared to ~ 5 kb in an African sample (Reich et al. 2001). In our data we observe significant LD (by Fisher's Exact test $p < 0.05$) in a sample of G6PD B and $G6PD_{med\ MME}$ that spans over 710 kb centromeric from *G6PD* (to *SNP A*), and 991 kb telomeric to *G6PD* (to *G0.9MT*). This long-range LD is associated with the target of selection at *G6PD*, coding site 563 (*i.e. Gd03* of Figures 2 and 4). Significant *intergenic* LD is not observed among the B alleles only, however *intragenic* levels of LD are high (nearly complete LD) for all relevant pairwise comparisons (data not shown). Although LD is significant between *Gd03* and sites at *IDH3G*, the value of |D'| is relatively low (|D'| = 0.364, Fisher's exact test p = 0.046). This low level of pairwise LD is an artifact of the haplotype structure of *IDH3G*, which includes two primary intragenic haplotypes that are common to both G6PD B and $G6PD_{med}$ alleles (Figure 2). Because the extended ancestral haplotype of $G6PD_{med}$ includes an already common intragenic haplotype among the B alleles, the pairwise |D'| values do not reflect the actual underlying signature of selection that may be seen by considering long-range haplotypes. To overcome this limitation of pairwise LD inferences, we examined the long-range haplotypes for the chromosomes in the sample using all the polymorphisms detected within the surveyed region. These data show conservation of a single primary LRH among the majority (10/16) of $G6PD_{med\ MME}$ chromosomes that spans from *IDH3G* to *G0.9MT* (> 1.6 Mb) (Figure 3). At *BGN* (925 kb proximal to *G6PD*) there is no significant conservation of a major common LRH, consistent with the truncation of the effects of selection seen with measures of nucleotide variability at this distance. On the telomeric side of *G6PD* we did not detect significant

decay of the ancestral G6PD$_{med\ MME}$ LRH up at distance of 991 kb (*G0.9MT*). Our survey ended at this distance because *G0.9MT* is near the pseudo-autosomal boundary, where it would be difficult to resolve phase unambiguously. In contrast to G6PD$_{med\ MME}$, the longest stretch of a conserved haplotype shared among > 0.5 of the G6PD B alleles spans < 135 kb (between *SNP E* and *G6PD*: figure 3). Saunders (APPENDIX B) found significant conservation of the ancestral LRH for G6PD A- in 13/20 chromosomes spanning the same region reported here for G6PD$_{med\ MME}$.

Age of the G6PD$_{med}$ alleles: We estimated the age of G6PD$_{med\ MME}$ along with an estimation of the selection coefficient (*s*) based on the level of decay of the ancestral haplotype seen on either side of *G6PD*. A previous study based on closely linked microsatellite variability among G6PD$_{med}$ alleles estimated that the age of the allele is ~ 3330 years, and this age was proposed to be approximately half as young as the age of G6PD A- based on the same methodology (Tishkoff et al. 2001). Our results based on long-range LD provide an age estimate for G6PD$_{med}$ of 1500-3500 years (assuming a generation time of 20 years) given a selection coefficient of $0.1 < s < 0.2$. These results are generally concordant with the previous age estimate of Tishkoff et al. (2001). By comparing the present LRH data to a parallel LRH data set from Africa which included G6PD A- alleles (Saunders APPENDIX B), we can infer that the ages of G6PD$_{med\ MME}$ and G6PD A- are roughly the same based on the fact that the ancestral LRHs of G6PD$_{med\ MME}$ and G6PD A- are conserved across the same genomic distance (*i.e. IDH3G* to *G0.9MT*) in nearly identical proportions (*i.e.* 10/16 and 12/20 of the chromosomes for this study and Africa, respectively). Given the small number of G6PD$_{med\ IND}$ alleles in the

study we lacked power to robustly estimate the age and selection coefficient of the allele. However, the similar extent of conservation of the ancestral G6PD$_{med\ IND}$ LRH relative to the G6PD$_{med\ MME}$ LRH implies that the age of G6PD$_{med\ IND}$ is similar to the age presented here for G6PD$_{med\ MME}$.

Conclusion: Different G6PD deficiency alleles have been subject to selection by malaria in different parts the world. Conservation of the long-range haplotypes seen here for G6PD$_{med}$ alleles is consistent with strong and recent selection. These LRHs suggest that resistance to malaria by G6PD deficiency alleles has arisen in Africa (*i.e.* G6PD A-) and different parts of Eurasia (*i.e.* G6PD$_{med\ MME}$ and G6PD$_{med\ IND}$) at roughly the same period of time.

ACKNOWLEDGMENTS

I thank Emily Landeen, Tanya Karafet, Zahara Mobasher, and Ryan Sprissler for technical assistance in generating the sequence data. Analyses of the age of the allele and strength of selection were preformed in collaboration with Montgomery Slatkin. Jeffery Good, Elizabeth Wood and members of the Nachman Lab and Hammer lab provided helpful discussion and comments. Qasim Mehdi provided the Pakistani samples used in this study.

REFERENCES

Allison AC (1960) Glucose-6-phosphate dehydrogenase deficiency in red blood cells of East Africans. Nature 186:531

Beutler E (1994) G6PD deficiency. Blood 84:3613-3636

Beutler E, Kuhl W (1990) The NT 1311 polymorphism of *G6PD*: G6PD mediterranean mutation may have originated independently in Europe and Asia. American Journal of Human Genetics 47:1008-1012

Beutler E, Westwood B, Kuhl W (1991) Definition of the mutations of G6PD Wayne, G6PD Viangchan, G6PD Jammu, and G6PD 'Lejeune'. Acta Haematol 86:179-182

Breman JG, Alilio MS, Mills A (2004) Conquering the intolerable burden of malaria: What's new, what's needed: A summary. American Journal of Tropical Medicine and Hygiene 71:1-15

Brown MD, Torroni A, Reckord CL, Wallace DC (1995) Phylogenetic analysis of Lebers hereditary optic neuropathy mitochondrial DNAs indicates multiple independent occurrences of the common mutations. Human Mutation 6:311-325

Coluzzi M, Sabatini A, Della-Torre A, Di Deco MA, Petrarca V (2002) A polytene chromosome analysis of the *Anopheles gambiae* species complex. Science 298:1415-1418

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693-709

Garner C, Slatkin M (2003) On selecting markers for association studies: Patterns of linkage disequilibrium between two and three diallelic loci. Genetic Epidemiology 24:57-67

Hammer MF, Garrigan D, Wood E, Wilder JA, Mobasher Z, Bigham A, Krenz JG, Nachman MW (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

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994) Evidence for positive selection in the superoxide-dismutase (*sod*) region of *Drosophila melanogaster*. Genetics 136:1329-1340

Hughes AL (1999) Adaptive evolution of genes and genomes. Oxford Press, New York, USA

Hutchin TP, Cortopassi GA (1997) Multiple origins of a mitochondrial mutation conferring deafness. Genetics 145:771-776

Joy DA, Feng XR, Mu JB, Furuya T, Chotivanich K, Krettli AU, Ho M, Wang A, White NJ, Suh E, Beerli P, Su XZ (2003) Early origin and recent expansion of *Plasmodium falciparum*. Science 300:318-321

Kaeda JS, Chhotray GP, Ranjit MR, Bautista JM, Reddy PH, Stevens D, Naidu JM, Britt RP, Vulliamy TJ, Luzzatto L, Mason PJ (1995) A new glucose-6-phosphate dehydrogenase variant, G6PD Orissa (44 Ala → Gly), Is the major polymorphic variant in tribal populations in India. American Journal of Human Genetics 57:1335-1341

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

Kurdi-Haidar B, Mason PJ, Berrebi A, Ankrabadu G, Alali A, Oppenheim A, Luzzatto L (1990) Origin and spread of the glucose-6-phosphate dehydrogenase variant (G6PD-Mediterranean) in the Middle-East. American Journal of Human Genetics 47:1013-1019

Kwok CJ, Martin ACR, Au SWN, Lam VMS (2002) G6PD*db* an integrated database of glucose-6-phosphate dehydrogenase (G6PD) mutations. Human Mutation 19:217-224

Lewontin RC (1964) Interaction of Selection + Linkage .I. General Considerations - Heterotic Models. Genetics 49:49-67

Livingstone FB (1985) Frequencies of hemoglobin variants: Thalassemia, the glucose-6-phosphate dehydrogenase deficiency, G6PD variants and ovalocytosis in human populations. Oxford University Press. New York, USA.

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genetical Research 23:23-35

Miller LH (1994) Impact of malaria on genetic polymorphism and genetic diseases in Africans and African-Americans. Proceedings of the National Academy of Sciences of the United States of America 91:2415-2419

Motulsky AG (1961) Glucose-6-phosphate dehydrogenase haemolytic disease of the newborn, and malaria. Lancet 1:1168

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences of the United States of America 76:5269-5273

Notaro R, Afolayan A, Luzzatto L (2000) Human mutations in glucose 6-phosphate dehydrogenase reflect evolutionary history. Faseb J 14:485-494

Oppenheim A, Jury CL, Rund D, Vulliamy TJ, Luzzatto L (1993) G6PD$_{mediterranean}$ accounts for the high prevalence of G6PD deficiency in Kurdish Jews. Human Genetics 91:293-294

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199-204

Roth EF, Raventossuarez C, Rinaldi A, Nagel RL (1983) Glucose-6-phosphate dehydrogenase deficiency inhibits *in vitro* growth of *Plasmodium falciparum*. Proceedings of the National Academy of Sciences of the United States of America 80:298-299

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174-175

Ruwende C, Khoo SC, Snow AW, Yates SNR, Kwiatkowski D, Gupta S, Warn P, Allsopp CEM, Gilbert SC, Peschu N, Newbold CI, Greenwood BM, Marsh K, Hill AVS (1995) Natural-Selection of Hemizygotes and Heterozygotes for G6PD Deficiency in Africa by Resistance to Severe Malaria. Nature 376:246-249

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

Saha S, Saha N, Tay JSH, Jeyaseelan K, Basair JB, Chew SE (1994) Molecular characterization of red-cell glucose-6-phosphate dehydrogenase deficiency in Singapore Chinese. American Journal of Hematology 47:273-277

Saunders MA, Hammer MF, Nachman MW (2002) Nucleotide variability at *G6PD* and the signature of malarial selection in humans. Genetics 162:1849-1861

Saunders MA, Slatkin M, Garner C, Hammer MF, Nachman MW (APPENDIX B) Long-range linkage disequilibrium around G6PD in Africa: Effects of natural selection by malaria.

Schneider JA, Peto TEA, Boone RA, Boyce AJ, Clegg JB (2002) Direct measurement of the male recombination fraction in the human beta-globin hot spot. Human Molecular Genetics 11:207-215

Sirugo G, Schaefer EA, Mendy A, West B, Bailey R, Walraven G, Sabeti P, Macciardi F, Zonta LA (2004) Is G6PD A- deficiency associated with recurrent stillbirths in The Gambia? American Journal of Medical Genetics Part A 128A:104-105

Slatkin M, Rannala B (2000) Estimating allele age. Annual Review of Genomics and Human Genetics 1:225-249

Stewart, CB, Schilling, JW, Wilson, AC (1987) Adaptive evolution in the stomach lysozymes of forgut fermenters. Nature 330: 401-404

Sukumar S, Mukhherjee MB, Colah RB, Mohanty D (2004) Molecular basis of G6PD deficiency in India. Blood Cells, Molecules, and Diseases 33:141-145

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human *G6PD*: Recent origin of alleles that confer malarial resistance. Science 293:455-462

Toomajian C, Ajioka RS, Jorde LB, Kushner JP, Kreitman M (2003) A method for detecting recent selection in the human genome from allele age estimates. Genetics 165:287-297

Verrelli BC, McDonald JH, Argyropoulos G, Destrol-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA (2002) Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. American Journal of Human Genetics 71:1112-1128

Vulliamy TJ, Durso M, Battistuzzi G, Estrada M, Foulkes NS, Martini G, Calabro V, Poggi V, Giordano R, Town M, Luzatto L, Persico MG (1988) Diverse point mutations in the human glucose-6-phosphate dehydrogenase gene cause enzyme deficiency and mild or severe hemolytic anemia. Proceedings of the National Academy of Sciences of the United States of America 85:5171-5175

Wainscoat JS (1987) The origin of mutant beta-globin genes in human populations. Acta Haematol 78:154-158

Wall JD, Przeworski M (2000) When did the human population size start increasing? Genetics 155:1865-1874

Wall JD, Frisse LA, Hudson RR, Di Rienzo A (2003) Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. American Journal of Human Genetics 73:1330-1340

Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nature Reviews Genetics 4:587-597

Waters TR, Swann PF (2000) Thymine DNA glycosylase and G to A transition mutations at CpG sites. Mutation Research-Reviews in Mutation Research 462:137-147

Watterson GA (1975) Number of segregating sites in genetic models without recombination. Theoretical Population Biology 7:256-276

Wooding S, Kim UK, Bamshad MJ, Larsen J, Jorde LB, Drayna D (2004) Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. American Journal of Human Genetics 74:637-646

Table 1: Individuals sampled in study.

G6PD allele type	Sample	Country
G6PD B:	Grc1032	Greece
	Grc870	Greece
	Irn014	Iran
	Irn031	Iran
	Irn333	Iran
	Irq004	Iraq
	Irq008	Iraq
	Irq010	Iraq
	Pak001	Pakistan
	Pak035	Pakistan
	Pak061	Pakistan
	Pak062	Pakistan
	Pak201	Pakistan
	Pak207	Pakistan
	Pak228	Pakistan
	Pak262	Pakistan
	Syr024	Syria
	Syr030	Syria
	Syr045	Syria
	Syr049	Syria
	Syr052	Syria
	Syr070	Syria
	Syr080	Syria
$G6PD_{med}$:	Grc022	Greece
	Grc865	Greece
	Irn027	Iran
	Irn049	Iran
	Irn344	Iran
	Irq005	Iraq
	Irq011	Iraq
	Irq013	Iraq
	Irq019	Iraq
	Pak033	Pakistan
	Pak050	Pakistan
	Pak099	Pakistan
	Pak206	Pakistan
	Pak212	Pakistan
	Pak218	Pakistan

Pak349	Pakistan
Syr035	Syria
Syr043	Syria
Syr053	Syria
Syr084	Syria
Syr056	Syria

Table 2: Independent SNPs assayed in study.

SNP label	Distance from *G6PD*[a]	NCBI SNP ID	Celera ID	Gene	SNP Type
A	710	rs2071123	hCV15868309	*IDH3G*	Intron
B	478	rs1059701	hCV8966366	*IRAK1*	Silent coding
C	207	rs2239466	hCV2462509	*TKTL1*	Intron
D	135	rs915943	hCV11927620	*RPL10*	Intron
E	52	rs2315325	hCV2198327	N/A	Intergenic
F	28	rs7611	hCV7493090	*FAM3A*	3' UTR
G	-100	rs4232906	hCV140027	*SPCX*	Intron
H	-118	rs4326559	hCV25649404	*CTAG2*	Silent coding
I	-201	rs2728725	hCV15924058	*GAB3*	Intron
J	-390	rs1936645	hCV11359203	*F8*	Intron

[a]Distance from *G6PD* (kilobases) as determined by distance from *G6PD* coding sites 563 found on NCBI contig NT_025965.12.

Table 3: Summary statistics of nucleotide variability resequenced regions.

Locus		n	Length[a]	S[b]	θ_π (SD) (%)	θ_W (SD) (%)	T D[c]	FL D[d]	H[e]	H_d (SD)
BGN										
G6PD B	Total sequence	23	2889	11	0.137 (0.009)	0.103 (0.045)	1.116	0.976	9	0.870 (0.046)
	Introns only	23	2160	10	0.159 (0.013)	0.125 (0.055)	0.908	0.911	8	0.842 (0.054)
$G6PD_{med\ MME}$	Total sequence	16	2889	11	0.136 (0.015)	0.115 (0.052)	0.714	0.218	6	0.783 (0.072)
	Introns only	16	2160	9	0.152 (0.017)	0.126 (0.059)	0.773	0.429	6	0.783 (0.072)
$G6PD_{med\ IND}$	Total sequence	5	2889	7	0.118 (0.044)	0.116 (0.069)	0.083	0.083	3	0.700 (0.218)
	Introns only	5	2160	6	0.139 (0.048)	0.133 (0.081)	0.286	0.286	3	0.700 (0.218)
IDH3G										
G6PD B	Total sequence	23	3036	6	0.053 (0.010)	0.054 (0.027)	-0.010	-0.219	4	0.621 (0.083)
	Introns only	23	2771	6	0.058 (0.010)	0.059 (0.030)	-0.010	-0.219	4	0.621 (0.083)
$G6PD_{med\ MME}$	Total sequence	16	3036	3	0.049 (0.007)	0.030 (0.019)	1.911	1.044	2	0.500 (0.074)
	Introns only	16	2771	3	0.054 (0.008)	0.033 (0.021)	1.911	1.044	2	0.500 (0.074)
$G6PD_{med\ IND}$	Total sequence	5	3036	0	0.00	0.00	0.00	0.00	1	0.00
Introns	only	5	2771	0	0.00	0.00	0.00	0.00	1	0.00
G6PD										
G6PD B	Total sequence	23	4679	3	0.021 (0.003)	0.017 (0.011)	0.550	-0.174	4	0.577 (0.090)
	Introns only	23	3694	2	0.017 (0.003)	0.016 (0.012)	0.177	-0.646	3	0.530 (0.071)
$G6PD_{med\ MME}$	Total sequence	16	4679	0	0.00	0.00	0.00	0.00	1	0.00
	Introns only	16	3694	0	0.00	0.00	0.00	0.00	1	0.00
$G6PD_{med\ IND}$	Total sequence	5	4679	0	0.00	0.00	0.00	0.00	1	0.00
Introns	only	5	3694	0	0.00	0.00	0.00	0.00	1	0.00
GAB3[†]										
G6PD B	Introns only	23	2983	1	0.003 (0.003)	0.009 (0.009)	-1.161	-1.591	2	0 .087 (0.078)
$G6PD_{med\ MME}$	Introns only	15	2983	0	0.00	0.00	0.00	0.00	1	0.00

G6PD$_{med\ IND}$	Introns only	5	2983	0	0.00	0.00	0.00	0.00	1	0.00
G0.9MT[†]										
G6PD B	Total Sequence	23	3069	3	0.019 (0.005)	0.026 (0.017)	-0.740	-1.334	4	0.486 (0.105)
G6PD$_{medMME}$	Total Sequence	16	3082	1	0.004 (0.003)	0.010 (0.010)	-1.162	-1.453	2	0.125 (0.106)
G6PD$_{medIND}$	Total Sequence	5	3082	1	0.019 (0.006)	0.016 (0.016)	1.225	1.225	2	0.600 (0.175)

[a]Number of bp analyzed. [b]Number of segregating sites. [c]Tajima's D. [d]Fu and Li's D. [e]Number of haplotypes. [f]Haplotype diversity.

[†]Surveyed region contained no coding sequence. Insertion deletion sites (*Id02*, *Id06* and *Mt03* in figure 2) were excluded from all analyses.

FIGURE LEGENDS

Figure 1: Map of the loci surveyed in this study. Five resequenced windows are marked with hatch marks, 10 SNPs (*A-J*) that were genotyped are marked with dark ovals. Physical distances from *G6PD* are marked below the line. Coding strand orientation is marked with dark arrows. A diagram of *G6PD* is presented below the line in reverse orientation, with dark boxes representing exons. Shaded arrows mark amplification primer positions for the region that was resequenced at *G6PD*. The genotypes and amino acid states defined by the polymorphic sites 563 and 1311 for the alleles G6PD B, $G6PD_{med\ MME}$ and $G6PD_{med\ IND}$ are depicted in white box.

Figure 2: Table of polymorphism among 44 chromosomes from the Mediterranean and Middle East for 5 resequenced windows including *G6PD*, and the genotypes for 10 SNPs intermittently spanning a region of ~ 2 Mb centered on *G6PD*. Yellow and blue boxes represent the minor and major allele, respectively, for each segregating site. Grey boxes indicate missing data. Segregating sites in each of the resequenced windows are labeled across the top in sequential order, and independent SNPs are marked (*A-J*). The three primary alleles classes (G6PD B, $G6PD_{med\ MME}$ and $G6PD_{med\ IND}$) are separated by fine white horizontal lines and bold vertical white lines represent unsurveyed inter-locus regions. The distance from *G6PD* of each resequenced window is marked below the table of polymorphism (distances for independent SNPs *A-J* are found in Table 2). N denotes nonsynonymous polymorphisms, and S denotes synonymous polymorphisms. (I) represents insertion/deletion polymorphisms at *Id02*, *Id06* and *Mt03* of sizes 2, 1, and 13

bp, respectively. Alignment site *Gd03* marked with the asterisk designates *G6PD* coding site 563, and *Gd02* designates *G6PD* coding site 1311. Explicit nucleotide states for all polymorphisms are available on the Nachman lab website (http://eebweb.arizona.edu/faculty/nachman/saunders/publications/G6PD_MME/sites_table_MME.html).

Figure 3: Schema depicting the shared long range haplotype for each of the three classes of alleles: G6PD B (white bars), G6PD$_{med\ MME}$ (black bars) and G6PD$_{med\ IND}$ (hatched bars). Site 563 of *G6PD* is marked with a vertical broken line, and distances (kb) from *G6PD* are marked across the bottom. All LRHs were extended outward on either side from *G6PD* site 563 (*i.e. Gd03* Figure 2), traced from the table of polymorphism in Figure 2. Each allele class (*i.e.* G6PD B, G6PD$_{med\ MME}$ and G6PD$_{med\ IND}$) was considered separately. For each locus outside of *G6PD*, the most common intragenic allele that maintained the shared LRH was marked. (Only within *G6PD* were haplotypes extended on an intragenic site-by-site basis). LRHs were not extended after the proportion of individuals with the shared LRH was ≤ 0.25 of the total sample size for the given group.

Figure 4: Matrix of all pairwise values of |D'| for all sites excluding G6PD$_{med\ IND}$ alleles. Segregating sites with a minor allele found in < 4 individuals were excluded from analyses. |D'| values are marked in accordance with the color legend at bottom. Asterisks inside boxes denote significant Fisher's exact test ($p < 0.05$). *G6PD* coding site 563 (*i.e.* target of selection) is marked in bold (***Gd03***).

Figure 5: Relative level of haplotype diversity (H_d) of G6PD$_{med\ MME}$ alleles / G6PD B alleles (dark line), and G6PD$_{med\ IND}$ alleles / G6PD B alleles (light broken line) for each locus surveyed.

Figure 6. Evolutionary analysis of the age and strength of selection for the G6PD$_{med\ MME}$ 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 long-range haplotype frequencies at two neighboring loci to *G6PD* (sites *Id07* and *SNP J*: Figure 2). An exponential growth rate of $r = 0.005$ and population size of 10,000 individuals were assumed. See SUBJECTS and METHODS for input data of haplotype configurations and recombination rates. For each point, 90,000 replicate sample paths were generated and 20 replicates of the Garner-Slatkin program were run for each sample path. (*a*) Log-likelihood of s, the hypothesized selective advantage of heterozygous carriers of the G6PD$_{med\ MME}$ allele. Additive selection was assumed. (*b*) The posterior distribution of allele age (t_1) for two selection coefficients consistent with the observations.

Figure 1:

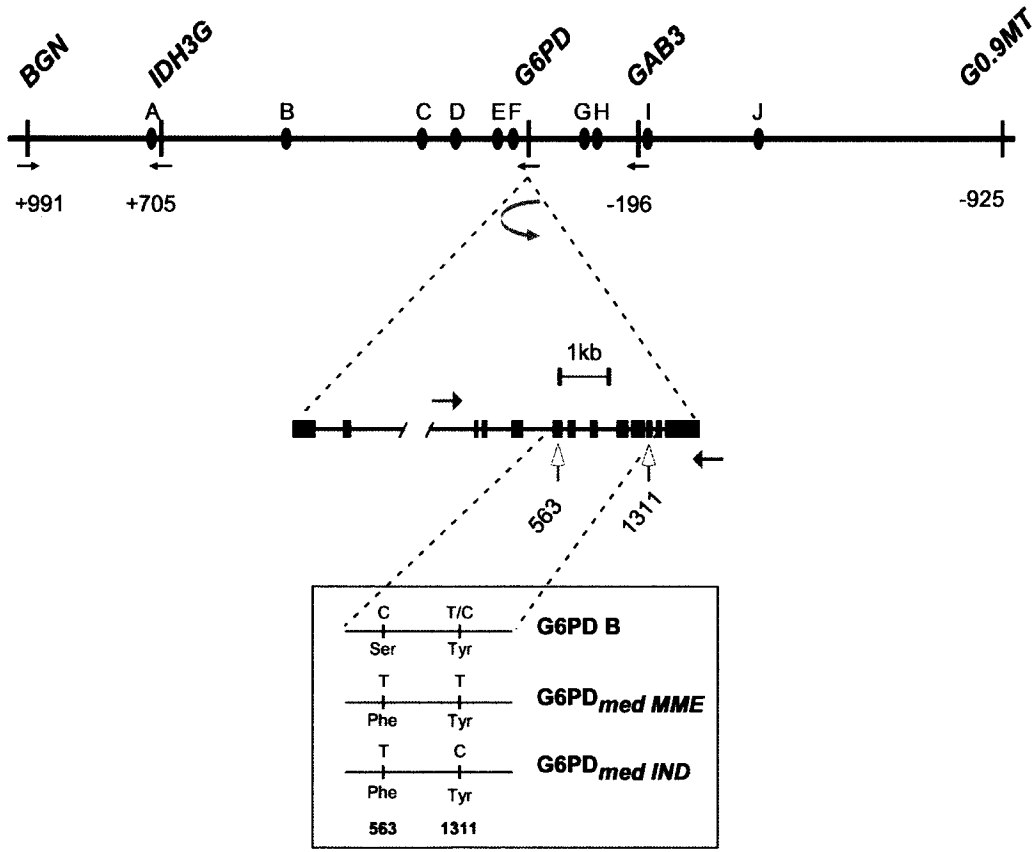

BGN
IDH3G
G6PD
GAB3
G0.9MT
A
B
C D
EF
GH
I
J
+991
+705
-196
-925
1kb
563
1311
C
T/C
Ser
Tyr
G6PD B
T
T
Phe
Tyr
G6PD med MME
T
C
Phe
Tyr
G6PD med IND
563
1311

Figure 2

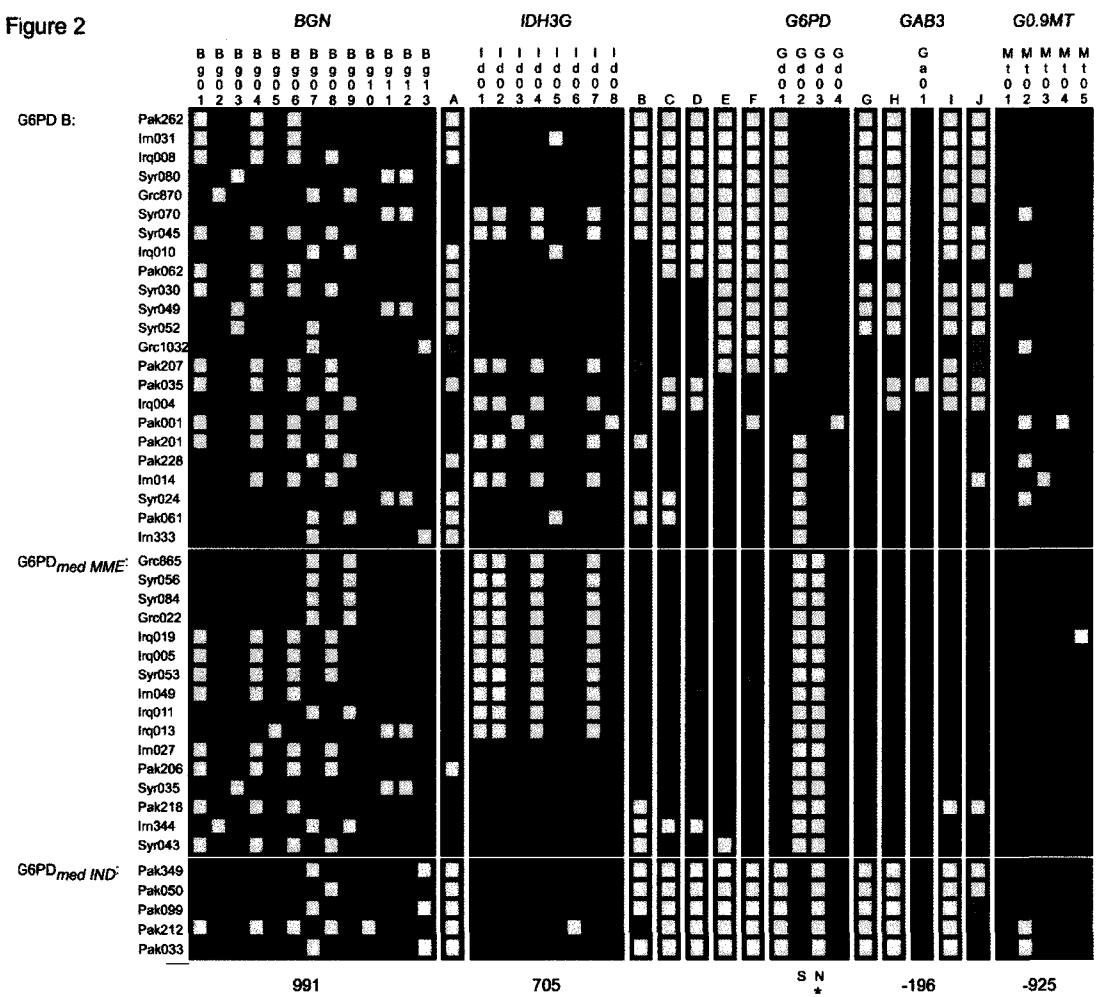

BGN
IDH3G
G6PD
GAB3
G0.9MT
Bg01 Bg02 Bg03 Bg04 Bg05 Bg06 Bg07 Bg08 Bg09 Bg10 Bg11 Bg12 Bg13
A
Id01 Id02 Id03 Id04 Id05 Id06 Id07 Id08
B C D E F
Gd01 Gd02 Gd03 Gd04
G H
Ga01
I J
Mt01 Mt02 Mt03 Mt04 Mt05
G6PD B:
Pak262
Im031
Irq008
Syr080
Grc870
Syr070
Syr045
Irq010
Pak062
Syr030
Syr049
Syr052
Grc1032
Pak207
Pak035
Irq004
Pak001
Pak201
Pak228
Im014
Syr024
Pak061
Im333
G6PDmed MME:
Grc865
Syr056
Syr084
Grc022
Irq019
Irq005
Syr053
Im049
Irq011
Irq013
Im027
Pak206
Syr035
Pak218
Im344
Syr043
G6PDmed IND:
Pak349
Pak050
Pak099
Pak212
Pak033
991
705
S N
-196
-925

Figure 3:

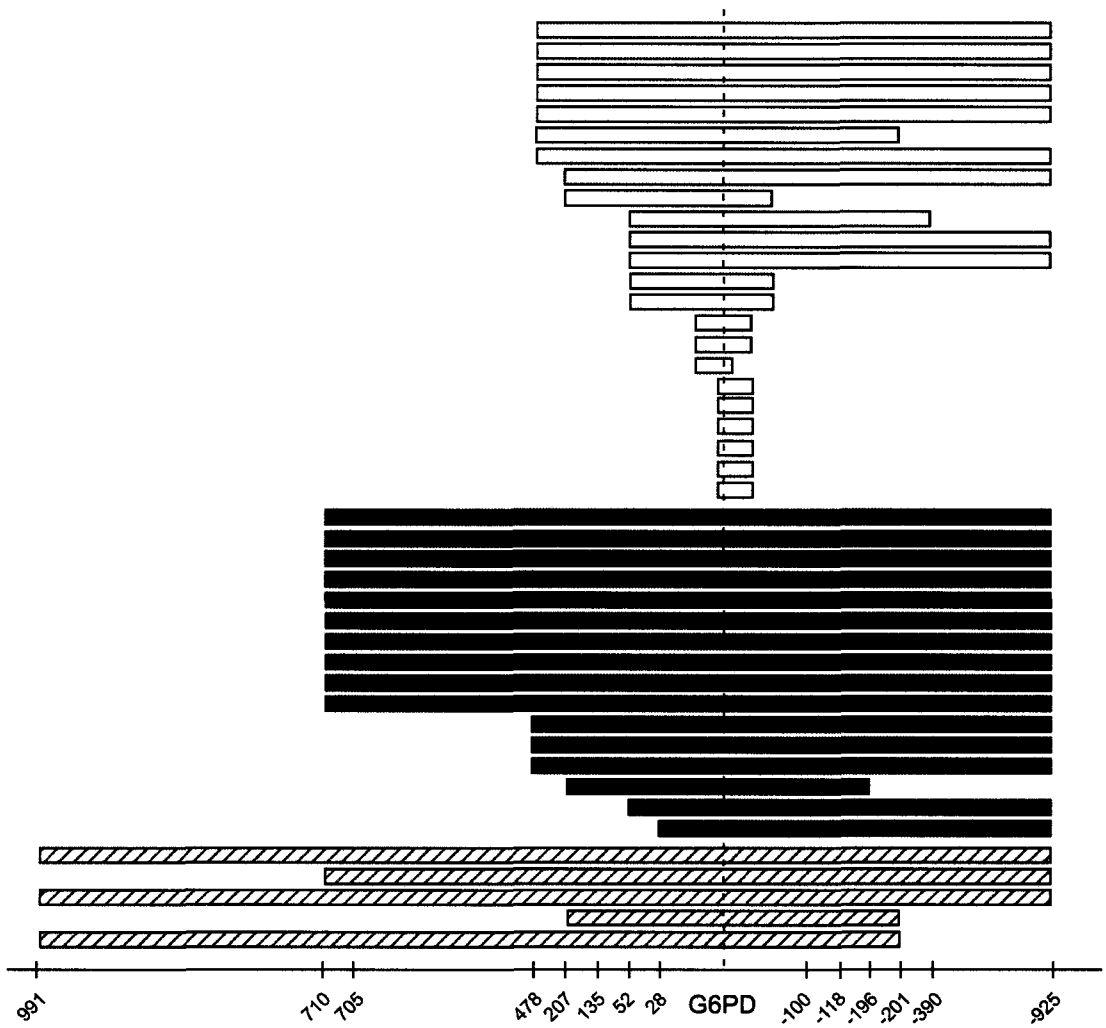


Figure 4:

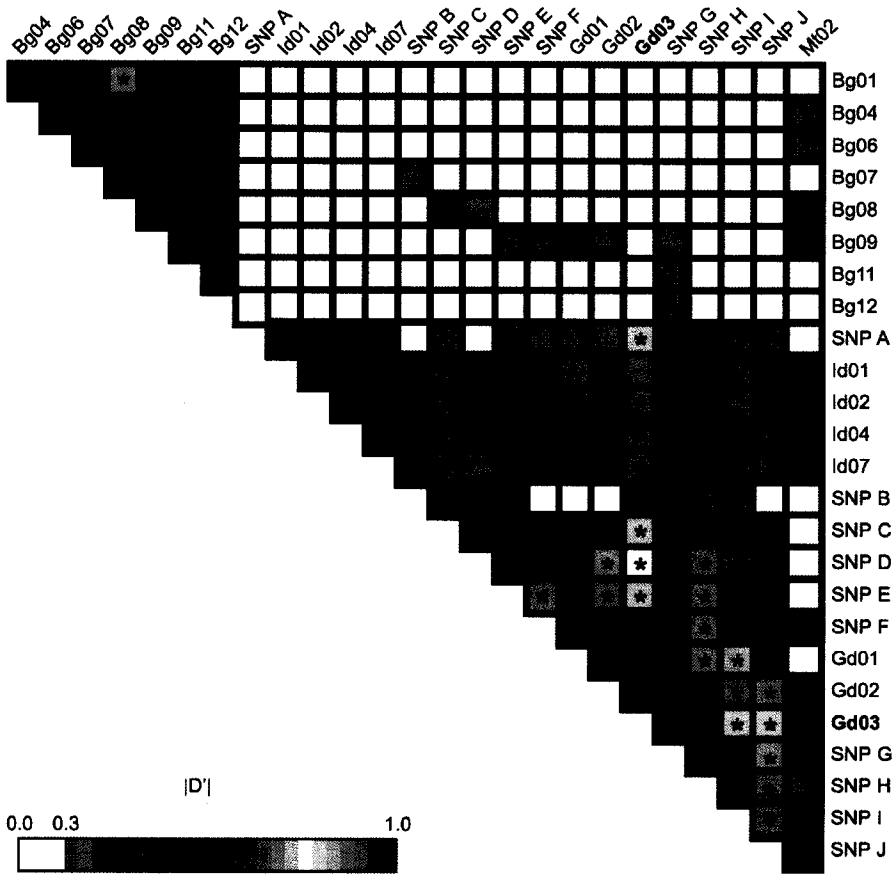

Bg04
Bg06
Bg07
Bg08
Bg09
Bg11
Bg12
SNP A
Id01
Id02
Id04
Id07
SNP B
SNP C
SNP D
SNP E
SNP F
Gd01
Gd02
Gd03
SNP G
SNP H
SNP I
SNP J
Mt02
Bg01
Bg04
Bg06
Bg07
Bg08
Bg09
Bg11
Bg12
SNP A
Id01
Id02
Id04
Id07
SNP B
SNP C
SNP D
SNP E
SNP F
Gd01
Gd02
Gd03
SNP G
SNP H
SNP I
SNP J
|D'|
0.0 0.3
1.0

Figure 5:

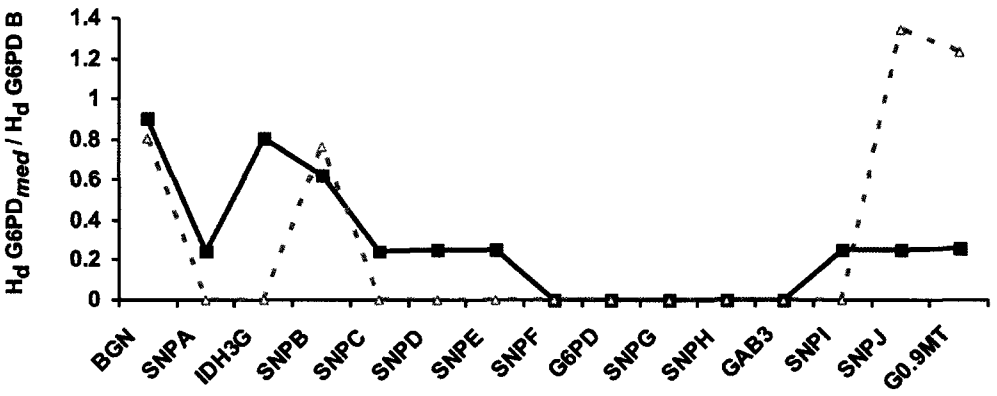
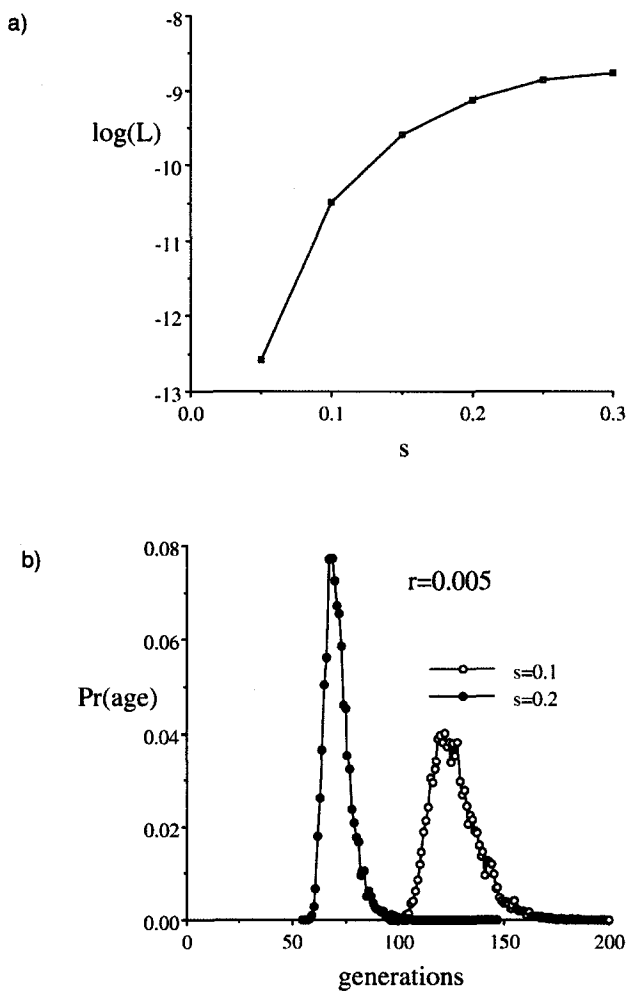

H_d G6PD$_{med}$ / H_d G6PD B
1.4
1.2
1
0.8
0.6
0.4
0.2
0
BGN
SNPA
IDH3G
SNPB
SNPC
SNPD
SNPE
SNPF
G6PD
SNPG
SNPH
GAB3
SNPI
SNPJ
G0.9MT

Figure 6:


a)
log(L)
-8
-9
-10
-11
-12
-13
0.0
0.1
0.2
0.3
s
b)
Pr(age)
0.08
0.06
0.04
0.02
0.00
r=0.005
s=0.1
s=0.2
0
50
100
150
200
generations

APPENDIX D: THE HIGH PREVALENCE OF G6PD DEFICIENCY IN KURDISH JEWS IS ATTRIBUTED LARGELY TO THE ACTION OF NATURAL SELECTION

ABSTRACT

Human population isolates often harbor relatively high frequencies of genetic disease alleles due to founder effects, inbreeding, and strong effects of genetic drift in small populations. Glucose-6-phostphate dehydrogenase (G6PD) deficiency is a common enzymopathy caused by different mutations in the coding sequence of *G6PD*. Many G6PD deficiency alleles are rare, yet some populations bear G6PD deficiency alleles at relatively high frequencies presumably due to selection by malaria. Kurdish Jews exhibit an extraordinarily high frequency of a G6PD deficiency allele ($q = 0.70$). Given this population's history of relative isolation, and the known action of selection on *G6PD* in other populations, it remains unclear if the high frequency of G6PD deficiency among Kurdish Jews is due primarily to demography or natural selection. To answer this question we surveyed nucleotide variability in a population sample of 37 Kurdish Jews at *G6PD*, at loci neighboring *G6PD*, and at two unlinked loci (*DMD* and *mtDNA* HVR1). Patterns of nucleotide variability at all loci (except *G6PD*) are similar to patterns seen in other non-African populations. Furthermore, we observe long-range linkage disequilibrium associated exclusively with the G6PD deficiency alleles. Together, our results suggest that the Kurdish Jews have not experienced a severe bottleneck or high

levels of inbreeding that would yield atypical genome-wide patterns of nucleotide variability. Patterns of nucleotide variability at *G6PD* in Kurdish Jews appear to have been strongly influenced by natural selection.

INTRODUCTION

Many genetic diseases are especially prevalent in specific human populations. A good example is the Finnish population which harbors many genetic diseases at frequencies of ~ 0.05 (Norio 1973; Kere 2001). Other examples include Samaritans, Amish and Hutterites which also have high frequencies of various diseases that are otherwise typically rare (Bonne-Tamir 1997; Arcos-Burgos 2002). The occurrence of a high frequency disease allele in a particular population may be due to a founder effect, inbreeding or genetic drift in small populations; however, a high frequency of a specific allele could also be due to population-specific natural selection. In fact, such competing hypotheses have given rise to lively debates over the past several decades in the case of Ashkenazi Jews (Zoossmann-Diskin 1995; Rish et al. 1995; Zlotogora and Bach 2003; Risch and Tang 2003). This population harbors relatively high frequencies of multiple genetic diseases (*e.g.* Tay-Sachs and lysozyme storage diseases: Motulsky 1995), and some researchers hypothesized that this may be due primarily to natural selection (Zlotogora 1988; Motulsky 1995), while recent population genetic analyses suggest that a bottleneck and/or genetic drift are sufficient to explain the high frequencies of disease alleles among Ashkenazi Jews (Risch et al. 2003; Behar et al. 2004; Slatkin 2004).

Examples of positive natural selection in humans are generally uncommon. However, selection may have strong localized effects on patterns of nucleotide variability that include a decrease in genetic variability and a high frequency of alleles associated with the selected phenotype (Maynard-Smith and Haigh 1974). Distinguishing between the effects of selection and demography in shaping nucleotide variability may be

challenging because these two forces are not mutually exclusive and because their effects may be similar. Nonetheless, several recent studies have successfully identified signatures of recent positive natural selection in humans against the background of low nucleotide variability that is typical of human populations (*e.g.* Sabeti et al. 2002; Saunders et al. 2002; Toomajian et al 2003; Bersaglieri et al. 2004).

Glucose-6-phosphate dehydrogenase (G6PD) deficiency is a well-described enzymopathy with manifestations that include neonatal jaundice and heamolytic anemia (Beutler 1994). Many rare G6PD deficiency alleles have been described worldwide (Kwok 2004). However, in some populations, G6PD deficiency alleles are found at a relatively high frequency despite disease phenotypes that are associated with these alleles (Beutler 1994). G6PD deficiency alleles confer some resistance to malaria (Roth et al. 1983; Ruwende et al. 1995), and these alleles are maintained in human populations due to natural selection (Allison 1960; Ruwende et al. 1995; Tishkoff et al. 2001; Sabeti et al. 2002; Saunders et al. 2002; Verrelli et al. 2002). In most of these populations G6PD deficiency alleles are at frequencies of 0.05 - 0.20 (Livingstone 1985). However, Kurdish Jews harbor a particularly severe G6PD deficiency allele, $G6PD_{med}$ (C563T: enzymatic activity ~ 5% of normal; Vulliamy et al. 1988), at a frequency of ~ 0.70, the highest known frequency of G6PD deficiency among human populations (Cohen 1971; Oppenheim et al. 1993), representing one of the highest frequencies of any severe genetic disease in a human population. In contrast, non-Jewish Kurdish populations and other Jewish groups do not harbor G6PD deficiency alleles (or other known malarial resistance factors) at high frequencies.

Kurdish Jews emigrated in ~ 735 BC from Samaria (present day region of Northern Israel) to the area of Kurdistan (present day region that encompasses Northern Iraq, Western Iran, and Eastern Turkey). Kurdish Jews maintained cultural isolation from other populations over many generations and returned *en masse* to Israel in ~ 1950 (Roth 1972; Brauer 1993; Levy 1999). Interestingly, Kurdish Jews also have a high frequency of β-thalassemia (Rund et al. 1991), a condition that may also confer resistance to malaria (Flint et al. 1993).

Is the remarkably high frequency of G6PD deficiency in Kurdish Jews attributable primarily to demography or to natural selection? Demographic processes are expected to affect all loci, while natural selection will affect particular alleles at a specific locus (Sabeti et al. 2001; Toomajian et al. 2003; Saunders APPENDIX B). Here we test the hypothesis that G6PD deficiency is at a high frequency among Kurdish Jews primarily due to a severe bottleneck and/or inbreeding (Oppenheim et al. 1993). We studied nucleotide variability in a population sample of 37 Kurdish Jews at *G6PD*, at multiple loci surrounding *G6PD*, and at 2 loci unlinked to *G6PD*. Overall patterns of nucleotide variability provide no evidence for a severe bottleneck in Kurdish Jews, but they do provide evidence for recent selection at *G6PD*.

SUBJECTS AND METHODS

Samples and loci surveyed: We sampled 37 unrelated male Kurdish Jews. Samples were obtained from a study of β-thalassemia, and included either patients, or relatives of non-sampled patients. This sampling is not expected to present a bias with respect to G6PD deficiency as no significant association (negative or positive) has been found between G6PD deficiency and hemoglobinopathies in individuals of other populations (*e.g.* Bouanga et al. 1998). In fact, the frequency of G6PD deficiency in this sample was found to be 0.65, in concordance with previous estimates of the frequency of $G6PD_{med}$ in this population. The genotype and phenotype of the individuals with respect to G6PD deficiency were unknown *a priori*. Only males were used in order to unambiguously resolve gametic phase across the X-chromosome. To survey nucleotide variability with respect to *G6PD* we resequenced contiguous regions from *G6PD* (5109 bp) and from 2 neighboring loci in region Xq28: *BGN* (2889 bp) and *L1CAM* (3227 bp) (Figure 1). These loci are not known to be under positive selection themselves, and have been previously utilized as neutral markers to delimit the long-range effects of selection at *G6PD* on nucleotide variability in genomic region Xq28 (Saunders et al. 2002; Saunders APPENDIX B; Saunders APPENDIX C). All primers were used as described by Saunders (APPENDIX B). We also genotyped independent SNPs (Table 1) at intermittent distances between the resequenced regions to define long-range haplotypes as described in Saunders (APPENDIX C). These SNPs were selected from the *SNP Browser* software (Applied Biosystems) based on physical distance from *G6PD*, and

were genotyped following *TaqMan® Assays-on-Demand(TM)* protocols (*Applied Biosystems*).

To look for genome-wide patterns of nucleotide variability that might be caused by demographic processes, we resequenced two unlinked loci in the same sample of Kurdish Jews: *DMD* intron 44 (*DMD44*) (2531 bp) and *mtDNA* hypervariable region 1 (HVR1) (504 bp). *DMD44* resides in genomic region Xp21 (Figure 1) and patterns of nucleotide variability at this locus suggest that it is evolving neutrally (Nachman and Crowell 2000). Furthermore, *DMD44* is in a region of high recombination and therefore is likely to be free of the effects of selection at linked sites (Nachman and Crowell 2000). Primers for *mt*DNA HVR1 and *DMD44* were used as described in Behar et al. (2004) and Nachman and Crowell (2000), respectively. All DNA sequences were determined on an ABI 3730 automated sequencer and contigs for each individual at each locus were assembled using the program *SEQUENCHER* (Gene Codes). Resequenced data have been deposited in Genbank (accession # XXXXXX-XXXXXX, XXXXXX-XXXXXX and XXXXXX-XXXXXX for Xq28 loci, *DMD44* and *mt*DNA HVR1, respectively). Sampling protocols were approved by the Human Subjects Committee at the University of Arizona.

Nucleotide variability data analysis: We performed analyses on the total sample of Kurdish Jews (n=37), and two subsets of the data: individuals that were determined by resequencing to bear G6PD$_{med}$ alleles (n=23), and individuals determined to bear G6PD B alleles (n=14). For each locus we calculated θ_π (Nei and Li 1979) and θ_w (Watterson 1975) which estimate the neutral parameter $\theta = 3N_e \mu$ for X-linked loci, and θ

$= 2N_e \mu$ for haploid loci (*e.g. mtDNA*) in a population at mutation-drift equilibrium, where N_e is the effective population size and μ is the neutral mutation rate. Tajima's D (Tajima 1989) and Fu and Li's D (Fu and Li 1993) were calculated at each locus to describe the frequency spectrum of polymorphisms. These statistics consider the difference between the parameters θ_π and θ_w, and may be used to detect an excess of intermediate frequency or low frequency polymorphisms relative to the neutral equilibrium expectation in a random population sample (*e.g.* our total panel; n=37). For subsets of the panel (*i.e.* $G6PD_{med}$ or G6PD B individuals alone) we calculated nucleotide diversity (θ_π and θ_w) and statistics of the frequency spectrum of alleles to test for potential allele-specific patterns around *G6PD*. Haplotype diversity (H_d) was calculated at each resequenced locus. Linkage disequilibrium was measured by |D'| (Lewontin 1964) between informative pairs of SNPs in region Xq28.

For *mtDNA* HVR1 we resequenced an additional n = 12 individuals to add statistical power to our inferences of demography, and we plotted the "mismatch distribution" (DiRienzo and Wilson 1991; Rogers and Harpending 1992) for this augmented sample of Kurdish Jews (n=49). This method describes the frequency spectrum of alleles by plotting the frequency of the number of differences between each pairwise comparison of alleles in the sample. This method has been commonly used in particular for *mtDNA* HVR1 data due to the statistical power gained from the large number of segregating sites and lack of recombination at this locus. A Poisson distribution is consistent with a population expansion, and is characteristic of most-non-African populations (*e.g.* DiRienzo and Wilson 1991). To further look for patterns

caused by demography and to compare our results to *mtDNA* HVR1 studies from other populations, we calculated Fu's Fs which compares the number of alleles in a sample to θ_π (Fu 1997). Although Tajima's D and Fu and Li's D may also be useful to describe the frequency spectrum of alleles and to detect non-equilibrium demographic processes (Tajima 1989; Fu and Li 1993), these methods have considerably less statistical power than Fu's F_s for detecting past demographic processes (Fu 1997), and therefore are rarely described in the literature with respect to *mtDNA* HVR1. We also calculated the raggedness (*r*) of the mismatch distribution plot (Harpending 1994). The mismatch distribution of most human populations is Poisson-shaped and smooth (*i.e.* low raggedness), while ragged distributions may indicate a constant-sized population at equilibrium or a bottleneck/founder effect (Cordaux et al. 2003; Excoffier and Schneider 1999). For all measures of nucleotide variability, individuals with any missing data at a given locus were excluded from the analyses. All analyses of nucleotide variability were calculated using *dnaSP* 4.0 (Rozas and Rozas 1999) or *ARLEQUIN* (Schneider et al. 2000).

Reference panels: To identify patterns of nucleotide variability among Kurdish Jews that are atypical relative to other human populations (that are known not to have undergone severe founder effects), we compared the observed patterns of nucleotide variability among Kurdish Jews at each locus to available nucleotide data from other populations (*i.e.* "reference panels"). In particular, we compared nucleotide variability for individuals bearing G6PD B alleles among Kurdish Jews (G6PD B_{KJ}) to individuals bearing G6PD B alleles from other non-African populations (G6PD B_{REF}). Natural

selection on *G6PD* is only expected to affect nucleotide variability associated with G6PD$_{med}$ alleles, while effects of demography will affect all alleles in the population in a similar fashion. For *BGN* and *G6PD* we used a reference panel that consisted of 23 individuals that bear G6PD B alleles from diverse localities in the Middle East and the Mediterranean region (Saunders APPENDIX C). For *DMD44* and *L1CAM* we used a reference panel composed of 31 non-African individuals from Europe, Asia, and the Americas (Nachman and Crowell 2000; Saunders et al. 2002). For *mtDNA* HVR1 we compared our results to patterns from samples of Ashkenazi Jews and Near Eastern non-Jewish populations (Behar et al. 2004). Since these reference samples are distributed over large geographic regions, they are conservative for our purpose of detecting a bottleneck in Kurdish Jews.

RESULTS

We estimated nucleotide variability in a population sample of Kurdish Jews at *G6PD*, at loci neighboring *G6PD*, and at 2 loci unlinked to *G6PD*, to assess genomewide patterns *vs.* locus-specific patterns of nucleotide variability.

Nucleotide variability at *G6PD*: We resequenced a window of 5109 bp from *G6PD* for a population sample of 37 Kurdish Jews (Figure 2). This surveyed region encompasses *G6PD* coding site 563 and has been studied in other panels from Africa (Verrelli et al. 2002; Saunders APPENDIX B), the Middle East (Saunders APPENDIX C), and a worldwide sample (Saunders et al. 2002). Twenty three of the Kurdish Jews in the sample bear the 563T mutation (site *Gd04*: Figure 2) that defines the deficiency allele $G6PD_{med}$. The remainder of the alleles had no nonsynonymous mutations, and thus were classified as G6PD B alleles. All $G6PD_{med}$ alleles in the sample bear the silent derived polymorphism (C1311T) at coding site 1311 that is typical of most Middle Eastern $G6PD_{med}$ alleles ($G6PD_{med\ MME}$: Beutler and Kuhl 1990; Saunders APPENDIX C). The frequency of $G6PD_{med}$ in this sample is 0.65 (23/37), consistent with the frequency of G6PD deficiency alleles estimated based on other samples of this population (q =0.70: Cohen 1971). We measured nucleotide variability in three subsets of the data: (*i*) the total population sample, (*ii*) individuals bearing G6PD B alleles only, and (*iii*) individuals bearing $G6PD_{med}$ alleles only. For the total population sample θ_π= 0.032% (Table 2). For G6PD B alleles from Kurdish Jews (G6PD B_{KJ}) θ_π=0.026%, while the finding for G6PD B alleles from a non-African panel (G6PD B_{REF}) is θ_π=0.021% (Table 2; Figure 3). No nucleotide variability is observed among $G6PD_{med}$ alleles from the Kurdish Jews, as in

other studies of *G6PD* (Verrelli et al. 2001; Saunders APPENDIX C), and consistent with a young age of the G6PD$_{med}$ allele (Tishkoff et al. 2001). Tajima's D (and Fu and Li's D) showed no significant deviation from neutrality for the population sample of Kurdish Jews (Table 2); however, its value (TD = 1.7074) was relatively higher (more positive) than the value for the non-African reference panel (TD = 0.550), which is representative of a typical Caucasian population (Saunders APPENDIX C) (Figure 4). The inflated value of Tajima's D (not significant), which indicates a high level of intermediate frequency polymorphisms for the total Kurdish Jew panel, is a consequence of the unusually high frequency of G6PD$_{med}$ alleles. When only G6PD B_{KJ} alleles are considered, Tajima's D is lower (TD = 1.1151) (Table 2; Figure 4). Haplotype diversity for G6PD B_{KJ} and G6PD B_{REF} alleles is H_d = 0.500 and 0.530, respectively, demonstrating no significant difference between the two groups (Table 2). Overall, at *G6PD*, we see no significant difference between the G6PD B_{KJ} alleles and G6PD B_{REF} alleles in any patterns of nucleotide variability, while the frequency of G6PD$_{med}$ alleles among Kurdish Jews is unusually high.

Nucleotide variability surrounding *G6PD*: To test for long-range patterns of nucleotide variability around *G6PD* in Kurdish Jews we resequenced regions from two loci (*BGN* and *L1CAM*) in Xq28 (Figure 1) that have been shown in other populations to bear informative sites for observing effects of selection on *G6PD* (Saunders et al. 2002; Saunders APPENDIX B). *L1CAM* exhibited very low levels of nucleotide variability (θ_π =0.005%) and included only one segregating site (Table 2). A low level of nucleotide variability at *L1CAM* was also found in a non-African panel (θ_π=0.00%; Saunders et al.

2002), and in an African panel (θ_π=0.026%; Saunders APPENDIX B), demonstrating that low variability at *L1CAM* is not unique to Kurdish Jews. Given the paucity of segregating sites at *L1CAM*, we used the single segregating site at this locus (Figure 2) only in considering the long-range haplotypes around *G6PD* (see below). In contrast to *L1CAM*, the surveyed region from *BGN* displayed an appreciable level of nucleotide variability. Nucleotide diversity was θ_π = 0.134, 0.135, 0.118 and 0.137, respectively, for the total Kurdish Jew sample, G6PD B_{KJ} alleles, G6PD$_{med}$ alleles and G6PD B_{REF} alleles (Table 2; Figure 3). None of these values are significantly different from each other, and notable is the "normal" level of variability among G6PD B_{KJ} alleles in comparison to G6PD B_{REF} alleles (Figure 3). Tajima's D (and Fu and Li's D) for the total sample is not significantly different from the reference panel (Figure 4), and haplotype diversity for G6PD B_{KJ} alleles (H_d = 0.872) also reflects similar values to the G6PD B_{REF} panel (H_d = 0.870). Although, H_d is similar between G6PD B_{KJ} alleles and G6PD$_{med}$ alleles (Table 2), the representative haplotypes are different between the two groups. A single intragenic haplotype is most common among the G6PD$_{med}$ alleles and this haplotype is otherwise rare among the G6PD B_{KJ} alleles (Figure 2). This difference is attributable to a conserved long-range haplotype centered on *G6PD* among the G6PD$_{med}$ alleles (see following section).

Long-range linkage disequilibrium in Xq28: To study linkage disequilibrium around *G6PD* we genotyped 10 independent SNPs (Table 1) at varying distances around *G6PD* (Figure 1). Along with the segregating sites of the resequenced windows, we defined long-range haplotypes (LRHs) of G6PD$_{med}$ and G6PD B alleles among Kurdish

Jews. A long-range haplotype that spans from *BGN* to *SNP J* is found among 11/23 of the G6PD$_{med}$ alleles (assuming, conservatively, that the few missing data among the SNPs bear the most common allele) (Figure 2). This LRH (*e.g.* K_10 in Figure 2) spans > 1.2 Mb, and a significant conservation of this LRH likely extends further, at least in the telomeric direction as only a single G6PD$_{med}$ individual (K_25) is seen to bear a different haplotype in this direction (Figure 2). LD between *Bg09* (representing the common LRH for G6PD$_{med}$) and *Gd04* (*i.e. G6PD* coding site 563) is significant using a Fisher's exact test (p=0.02) with |D'| = 0.763. On the telomeric side, LD between *Gd04* and *SNP I* is also similarly highly significant with |D'| = 0.858 (FET p= 0.000011). (This LD also extends to *SNP J*, however due to the missing data at this site we did not perform the analogous LD calculation). This strong LD is associated primarily with the G6PD$_{med}$ alleles. When LD is calculated among the G6PD B$_{KJ}$ alleles only, significant intergenic (long-range) LD is observed maximally between *SNP C* and *SNP I* (408 kb apart) (Figure 5).

Nucleotide variability at loci unlinked to *G6PD*: We surveyed nucleotide variability at two loci outside of Xq28 (*DMD44* and *mtDNA* HVR1) to relate to patterns at *G6PD*, and to distinguish between locus-specific and genome-wide trends in the sample of Kurdish Jews. Nucleotide variability at *DMD44* for the total sample, G6PD B$_{KJ}$ alleles and G6PD$_{med}$ alleles is θ_π = 0.141, 0.159 and 0.137%, respectively (Table 2). All these values are remarkably similar to each other and to a reference panel of 31 non-African G6PD B alleles (θ_π = 0.126; Nachman and Crowell 2000) (Figure 3). Tajima's D for the reference panel is 0.354, which is not significantly different from the values of

Tajima's D for the Kurdish samples (Table 2; Figure 4). Together, these results show no evidence of a highly skewed allele frequency at *DMD44* for Kurdish Jews.

Overall levels of nucleotide variability at mtDNA HVR1 for Kurdish Jews (θ_π = 1.1%) is similar to Ashkenazi Jews, non-Jewish Near Easterns, and Europeans (θ_π = 1.4%, 1.7% and 1.1% respectively: Behar et al 2004). *MtDNA* HVR1 is commonly used to infer signatures of past demographic events by employing the "mismatch distribution" (DiRienzo and Wilson 1991; Harpending and Rogers 1992). We plotted the mismatch distribution for a sample of Kurdish Jews that included 48 individuals (excluding incomplete sequence from individual O_02: Figure 6). Comparison of the mismatch distribution for Kurdish Jews relative to Ashkenazi Jews and non-Jewish Near Easterns for *mtDNA* HVR1 reveals general similarity between the past demographic histories of these populations. The means of the distributions (τ) are 6.2, 5.5, and 6.2, for Kurdish Jews, Ashkenazi Jews, and Near Easterns, respectively, indicating a signature of a common Pleistocene expansion (Figure 5). Both Kurdish Jews and Ashkenazi Jews have a slight increase in the low (0 difference) class of the distribution relative to the non-Jewish Near Eastern sample, possibly suggesting an old population bottleneck that is shared among Jewish populations (Behar et al. 2004). A severe founder effect or bottleneck is expected to erase the signature of a Pleistocene expansion (Excoffier and Schneider 1999), yet this signature is still obvious in the sample of Kurdish Jews. Fu's Fs for Kurdish Jews is -11.01 ($p<0.001$), and Harpending's raggedness index is $r = 0.013$: values that are similar to other populations that have not undergone severe bottlenecks (Excoffier and Schneider 1999; Behar et al. 2003). In contrast, some Ashkenazi

populations show non-significant deviations from a constant population model with relatively high values of Fs (Fs $\approx$ -5; $p > 0.02$) (Behar et al. 2004), suggesting that they are more likely than Kurdish Jews to have undergone a significant bottleneck.

DISCUSSION

Kurdish Jews exhibit the highest frequency of G6PD deficiency alleles (0.70) of any surveyed human population (Cohen 1971, Oppenheim et al. 1993). G6PD deficiency alleles are maintained in some human populations because they confer protection against severe malaria. However, in populations where G6PD deficiency alleles have been shown to be subject to selection, their frequency typically ranges between 0.05 and 0.20, making the observed frequency of G6PD deficiency among Kurdish Jews exceptional. Furthermore, the specific deficiency allele that is found among Kurdish Jews, $G6PD_{med}$, is an allele with severe clinical manifestations; the enzyme activity is < 5% of normal (class II deficiency allele [WHO classification]: Vulliamy 1988). The aim of this study was to infer the relative contribution of demography and/or selection to patterns of nucleotide variability at *G6PD* among Kurdish Jews.

Nucleotide variability at *G6PD*: An extreme founder event may result in a high frequency of deleterious alleles and a high level of homogeneity among alleles present in a sample. If a severe bottleneck and/or inbreeding have affected nucleotide variability among Kurdish Jews, we would expect to see a relatively high level of homozygosity among G6PD B_{KJ} alleles. In contrast, if recent positive natural selection is acting on an allele (*e.g.* $G6PD_{med}$), it will rise in frequency quickly (*i.e.* partial sweep), and due to a hitchhiking effect, nucleotide variability will be erased only among the recently selected alleles (Maynard Smith and Haigh 1974). Importantly, this process will not affect nucleotide variability among the non-selected alleles (*e.g.* G6PD B_{KJ}) (aside from the effect of reducing their intra-allelic effective population size). Both a founder effect and

recent selection predict low levels of variation among the G6PD$_{med}$ alleles. In fact, in other populations (without known founder effects), where a high frequency of G6PD deficiency is clearly attributed to the action of selection, almost no variability is observed among the selected alleles (Verrelli et al. 2002, Saunders et al. 2002; Saunders APPENDIX B; Saunders APPENDIX C). In Kurdish Jews, measures of nucleotide variability (*e.g.* θ_π and H_d) for G6PD B_{KJ} alleles are similar to measures of nucleotide variability from a reference panel of non-Africans (Figure 3), arguing against a severe founder effect and/or high levels of inbreeding among Kurdish Jews. As expected, due to the young age of G6PD$_{med}$ we observe no variation among this class of alleles.

Linkage disequilibrium around *G6PD*: Under a scenario of a severe founder-effect, relatively few founding chromosomes will contribute to the population. This will create a pattern of LD among the different "founding" long-range haplotypes in a population sample. Importantly, this process will affect all alleles in the population similarly. Under a scenario of recent natural selection, long-range LD is expected to be associated only with the common LRH of the selected allele (Sabeti et al. 2001; Toomajian et al. 2003; Saunders APPENDIX B). Here we defined long-range haplotypes (LRHs) roughly centered on *G6PD* in a region where LRHs of selected G6PD alleles have been previously identified in populations subject to malarial selection (Sabeti et al. 2002; Saunders APPENDIX B; Saunders APPENDIX C). A LRH common to 11/23 of G6PD$_{med}$ alleles spans a region > 1.2 Mb and creates highly significant LD in the total sample (Figure 2). A similar-sized LRH associated with G6PD deficiency alleles is seen in the Middle East (Saunders APPENDIX C) and in Africa (for G6PD A-; Saunders

APPENDIX B). In contrast, significant LD among the G6PD B_{KJ} alleles spans maximally over a region of 0.408 Mb (Figure 6), a span that is similar to the result from a "normal" Middle Eastern panel around *G6PD* (Saunders APPENDIX C). The observed long-range LD that is associated only with $G6PD_{med}$ alleles, reminiscent of other populations assumed to be subject to selection, lends support to the hypothesis that selection has shaped patterns of nucleotide variability around *G6PD* in Kurdish Jews.

Nucleotide Variability at unlinked loci: A severe bottleneck would be expected to affect not only *G6PD*, but also other loci throughout the genome. To test this idea we examined nucleotide variability at *DMD44*, which is found on the short arm of the X-chromosome (Xp21) and at *mtDNA* HVR1. Haplotype diversity and other measures of nucleotide variability in Kurdish Jews for both loci are remarkably similar to estimates from our reference panels derived from populations that have not experienced severe bottlenecks (Table 2; Figure 3; Figure 4). Since *DMD44* exhibits relatively high levels of heterozygosity in most populations surveyed to date (Nachman and Crowell 2000), it may not be the best indicator of demographic effects. However *mtDNA* HVR1 has been shown to provide good statistical power for inferring bottlenecks and population expansions using the mismatch distribution (Excoffier and Schnieder 1999; Behar et al 2004). Inspection of this distribution for Kurdish Jews (Figure 5) shows a shape that is typical of many other non-African populations (DiRienzo and Wilson 1991): the mode is centered at $\tau \approx 6$ representing a Pleistocene expansion and the raggedness of the curve is low ($r = 0.013$). Although a precise theoretical expectation for the pattern of the mismatch distribution under a recent bottleneck is unclear, it should erase the signature of

a Pleistocene expansion (Rogers and Harpending 1992; Excoffier and Schneider 1997). For example, in South-Indian tribal populations that have a recorded history of inbreeding and small population sizes, the typical smooth Poisson distribution is not seen in the mismatch distribution of *mtDNA* HVR1 (Cordaux et al. 2003). The mismatch distribution of Kurdish Jews does however exhibit a slight increase in the 0 difference class relative to other populations (*e.g.* non-Jewish Near Easterns and Europeans: Behar et al. 2004). Nonetheless, in comparison to the mismatch distribution of Ashkenazi Jews and other Jewish groups, we see a similar slight peak in the left tail of the distribution, consistent with a shared early bottleneck in the history of Jewish populations (Behar et al. 2004). This comparison is of particular importance, because it suggests that Kurdish Jews and Ashkenazi Jews have experienced an early historical bottleneck on the same order of magnitude (possibly even the same event). Although Ashkenazi Jews have often been cited as having high frequencies of genetic disorders (*e.g.* Tay Sachs, Gaucher's disease), the frequency of these disease alleles is invariably < 0.03 (Motulsky 1995; Risch et al. 2003), in contrast to the extremely high frequency of G6PD deficiency among Kurdish Jews. Together, data from loci unlinked to *G6PD* in Kurdish Jews show no evidence of atypical patterns of nucleotide variability. This suggests that the pattern seen at *G6PD* is locus-specific.

Selection at *G6PD*: Our results provide no evidence of a particularly severe founder effect and/or inbreeding in Kurdish Jews. We suggest, therefore that natural selection played a major role in increasing the frequency of $G6PD_{med}$ among Kurdish Jews. Ashkenazi Jews have relatively high frequencies ($0.008 < q < 0.030$) of different

disease alleles (Motulsky 1995), and hypotheses have been presented to explain these patterns by inferring positive natural selection (Zlotogora et al. 1988). However mounting analyses of nucleotide variability data for these populations suggest that the frequency of these diseases can be explained by a mild bottleneck (Risch et al. 2003; Behar et al. 2004; Slatkin 2004). In contrast, the extraordinary frequency of G6PD deficiency among Kurdish Jews is not likely to be explained solely by a simple demographic model (*i.e.* a bottleneck or inbreeding). We note that the effects of demography and selection are not mutually exclusive, and some combination of these forces could also contribute to the observed patterns.

With the data at hand, we may only speculate about more complex models that include the effects of both a mild bottleneck and selection. Given historical information, it is reasonable to speculate that Kurdish Jews might have experienced at least a mild bottleneck upon emigration to Kurdistan or prior to that date. Following a bottleneck, some low frequency mutations are expected to be lost from a population. Multiple genetic factors are known to affect susceptibility to malaria in humans, and are found at frequencies of < 0.25 (Migot-Nabias et al. 2000; Roberts and Williams 2003; Frodsham and Hill 2004). The interaction between these genetic factors remains unclear, yet an individual might benefit from having at least one resistance factor to malaria. If some factors are lost from a population following a bottleneck, then remaining resistance factors might be selected to increase in frequency as a compensatory effect. This "compensatory hypothesis" might explain the high frequency of G6PD deficiency among Kurdish Jews. Future tests of this hypothesis should include surveys of the allele

frequencies of the other known resistance factors to malaria in Kurdish Jews compared to other human populations.

Unrecognized population-specific interactions with the environment could also potentially explain the high frequency of G6PD deficiency among Kurdish Jews. It is known that ingestion of specific drugs and food (*e.g.* fava beans) may have adverse effects on the G6PD deficiency phenotype (Etkin 2003). Kurdish Jews may be uniquely exposed to an agent (*e.g.* food) that has an unrecognized effect of *decreasing* the adverse manifestations of G6PD deficiency. This might relieve the clinical cost of bearing $G6PD_{med}$, thus increasing the overall selection coefficient in the population and allowing the allele to rise in frequency beyond levels seen in other populations. This "gene by unique environment hypothesis" warrants detailed examination of the cultural traditions and historical living conditions of Kurdish Jews.

The exact nature of selection remains unclear in this case. Malaria caused by *Plasmodium falciparum* is the typically recognized agent of selection on G6PD deficiency alleles. However, *P. falciparum* is not endemic, in general, in Kurdistan. Furthermore, non-Jewish Kurdistani populations do not exhibit high frequencies of G6PD deficiency nor hemoglobinopathies that confer resistance to malaria (Livingstone 1985). One explanation for this discrepancy is that some unidentified agent of selection other than malaria has acted on G6PD deficiency among Kurdish Jews. In some cases, a genetic resistance factor in humans may protect against multiple pathogens. For example the ΔCCR5 polymorphism provides resistance today to HIV type I, however it is likely to have increased in frequency in the past due to selection from a different agent of

selection, possibly smallpox (Galvani and Slatkin 2003). Also, the Duffy null allele (FY*O) provides resistance to *P. vivax* and is near fixation in sub-Saharan Africa. However this polymorphism may have increased in frequency due to a different unknown pathogen (Hamblin et al. 2000). This explanation remains tentative for $G6PD_{med}$, as we are unaware of any potential alternative agents of selection for G6PD deficiency. A second, more likely, explanation is that *P. falciparum* was a strong agent of selection in Kurdish Jews in the past. Historical records suggest that Kurdish Jews inhabited Samaria (*i.e.* present day Northern Israel) until their emigration to Kurdistan in ~730 BC (Levy 1999). Parts of Northern Israel (*i.e.* the Hula Valley) have been notorious for hyper-endemic *P. falciparum* until recent eradication efforts, and this past environment may have subjected the populations inhabiting this area to a strong regime of selection (Hershkovitz et al. 1991). The possibility that malaria is the likely agent of selection in this case is further supported by the fact that β-thalassemia, another genetic resistance factor to malaria, is also at a high frequency among Kurdish Jews ($q = 0.12$), while haplotype diversity among surveyed β-thalassemia alleles in Kurdish Jews is unusually high (Rund et al. 1991). In summary, multiple genetic resistance factors to malaria may have attained high frequencies among the ancestors of Kurdish Jews in the face of extreme endemic malaria in the past.

ACKNOWLEDGMENTS

We thank Julia Kim, Zahara Mobasher, Tanya Karafet and Ryan Sprissler for technical assistance in generating the sequence data. Ariella Oppenheim donated the Kurdish Jews samples. Jeffery Good, Elizabeth Wood, Daniel Garrigan, Armando Geraldes, Jason Wilder, Kent Campbell, Bruce Walsh and members of the Nachman Lab and Hammer lab provided helpful discussion and comments. This project was funded by NSF grants to M.A.S. and M.W.N.

REFERENCES

Allison AC (1960) Glucose-6-phosphate dehydrogenase deficiency in red blood cells of East Africans. Nature 186:531

Arcos-Burgos M, Muenke M (2002) Genetics of population isolates. Clinical Genetics 61:233-247

Behar DM, Hammer MF, Garrigan D, Villems R, Bonne-Tamir B, Richards M, Gurwitz D, Rosengarten D, Kaplan M, Della Pergola S, Quintana-Murci L, Skorecki K (2004) MtDNA evidence for a genetic bottleneck in the early history of the Ashkenazi Jewish population. European Journal of Human Genetics 12:355-364

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. American Journal of Human Genetics 74:1111-1120

Beutler E (1994) G6PD deficiency. Blood 84:3613-3636

Bonne-Tamir B, Nystuen A, Seroussi E, Kalinsky H, KwitekBlack AE, Korostishevsky M, Adato A, Sheffield VC (1997) Usher syndrome in the Samaritans: Strengths and limitations of using inbred isolated populations to identify genes causing recessive disorders. American Journal of Physical Anthropology 104:193-200

Bouanga JC, Mouele R, Prehu C, Wajcman H, Feingold J, Galacteros F (1998) Glucose-6-phosphate dehydrogenase deficiency and homozygous sickle cell disease in Congo. Human Heredity 48:192-197

Brauer E (1993) The Jews of Kurdistan. Wayne State University Press, Detroit, Michigan, USA

Cohen T (1971) Genetic markers in migrants to Israel. Israel Journal of medical sciences 7:1509

Cordaux R, Saha N, Bentley GR, Aunger R, Sirajuddin SM, Stoneking M (2003) Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. European Journal of Human Genetics 11:253-264

DiRienzo A, Wilson AC (1991) Branching pattern in the evolutionary tree for human mitochondrial-DNA. Proc Natl Acad Sci U S A 88:1597-1601

Etkin NL (2003) The co-evolution of people, plants, and parasites: biological and cultural adaptations to malaria. Proceedings of the Nutrition Society 62:311-317

Excoffier L, Schneider S (1999) Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. Proc Natl Acad Sci U S A 96:10597-10602

Flint J, Harding RM, Boyce AJ, Clegg JB (1993) The Population-Genetics of the Hemoglobinopathies. Baillieres Clinical Haematology 6:215-262

Frodsham AJ, Hill AVS (2004) Genetics of infectious diseases. Human Molecular Genetics 13:R187-R194

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147:915-925

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693-709

Galvani AP, Slatkin M (2003) Evaluating plague and smallpox as historical selective pressures for the CCR5-Δ32 HIV-resistance allele. Proc Natl Acad Sci U S A 100:15276-15279

Hamblin MT (2000) Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus. American Journal of Human Genetics 66:1669-1679

Harpending H (1994) Signature of ancient population-growth in a low-resolution mitochondrial-DNA mismatch distribution. Human Biology 66:591-600

Hershkovitz I, Ring B, Speirs M, Galili E, Kislev M, Edelson G, Hershkovitz A (1991) Possible congenital hemolytic-anemia in prehistoric coastal inhabitants of Israel. American journal of physical anthropology 85:7-13

Kere J (2001) Human population genetics: Lessons from Finland. Annual Review of Genomics and Human Genetics 2:103-128

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics 22:139-144

Levy H (1999) Comprehensive history of the Jews of Iran. (The outset of the diaspora). Mazda Publishers, Costa Mesa, CA, USA

Lewontin RC (1964) Interaction of selection + linkage .I. General considerations - Heterotic models. Genetics 49:49-67

Livingstone FB (1985) Frequencies of hemoglobin variants: Thalassemia, the glucose-6-phosphate dehydrogenase deficiency, G6PD variants and ovalocytosis in human populations.

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favorable gene. Genetical research 23:23-35

Migot-Nabias F, Mombo LE, Luty AJF, Dubois B, Nabias R, Bisseye C, Millet P, Lu CY, Deloron P (2000) Human genetic factors related to susceptibility to mild malaria in Gabon. Genes and Immunity 1:435-441

Motulsky AG (1995) Jewish diseases and origins. Nature Genetics 9:99-101

Nachman MW, Crowell SL (2000) Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, *DMD*, in humans. Genetics 155:1855-1864

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A 76:5269-5273

Norio R, Nevanlin.Hr, Perheent.J (1973) Hereditary diseases in Finland - rare flora in rare soil. Annals of Clinical Research 5:109-141

Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Brittenham G, Looareesuwan S, Clark AG, Tokunaga K (2004) Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. American Journal of Human Genetics 74:1198-1208

Oppenheim A, Jury CL, Rund D, Vulliamy TJ, Luzzatto L (1993) G6PDmediterranean accounts for the high prevalence of G6PD deficiency in Kurdish Jews. Human Genetics 91:293-294

Risch N, Tang H, Katzenstein H, Ekstein J (2003) Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. American Journal of Human Genetics 72:812-822

Risch N, Deleon D, Fahn S, Bressman S, Ozelius L, Breakefield X, Kramer P, Almasy L, Singer B (1995) ITD in Ashkenazi Jews - Genetic drift or selection? - Reply. Nature Genetics 11:14-15

Risch N, Tang H (2003) Selection in the Ashkenazi Jewish population unlikely - Reply to Zlotogora and Bach. American Journal of Human Genetics 73:440-441

Risch N, Tang H, Katzenstein H, Ekstein J (2003) Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. American Journal of Human Genetics 72:812-822

Roberts DJ, Williams TN (2003) Haemoglobinopathies and resistance to malaria. Redox Report 8:304-310

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. Molecular Biology and Evolution 9:552-569

Roth C (ed) (1972) Encyclopedia Judaica. Vol 10. Keter, Jerusalem

Roth EF, Raventossuarez C, Rinaldi A, Nagel RL (1983) Heterozygous and hemizygous red-cell G6PD deficiency inhibit *in vitro* growth of *falciparum*-malaria to the same extent. Clinical Research 31:A322-A322

Rozas J, Rozas R (1999) *DnaSP* version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174-175

Rund D, Cohen T, Filon D, Dowling CE, Warren TC, Barak I, Rachmilewitz E, Kazazian HH, Oppenheim A (1991) Evolution of a genetic-disease in an ethnic Isolate - beta-thalassemia in the Jews of Kurdistan. Proc Natl Acad Sci U S A 88:310-314

Ruwende C, Khoo SC, Snow AW, Yates SNR, Kwiatkowski D, Gupta S, Warn P, Allsopp CEM, Gilbert SC, Peschu N, Newbold CI, Greenwood BM, Marsh K, Hill AVS (1995) Natural selection of hemizygotes and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature 376:246-249

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

Saunders MA, Hammer MF, Nachman MW (2002) Nucleotide variability at *G6PD* and the signature of malarial selection in humans. Genetics 162:1849-1861

Saunders MA, Slatkin M, Garner C, Hammer MF, Nachman MW (APPENDIX B) Long-range linkage disequilibrium around *G6PD* in Africa: Effects of natural selection by malaria.

Saunders MA, Slatkin M, Garner C, Hammer MF, Nachman MW (APPENDIX C) Extended haplotypes of G6PD$_{mediterranean}$ and the evolutionary history of resistance to malaria in Eurasia.

Schneider S, Roessli D, Excoffier L (2000) *Arlequin* ver. 2000: A software for population genetics data analysis., University of Geneva, Switzerland

Slatkin M (2004) A population-genetic test of founder effects and implications for Ashkenazi Jewish diseases. American Journal of Human Genetics 75:282-293

Swallow DM (2003) Genetics of lactase persistence and lactose intolerance. Annual Review of Genetics 37:197-219

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human *G6PD*: Recent origin of alleles that confer malarial resistance. Science 293:455-462

Toomajian C, Ajioka RS, Jorde LB, Kushner JP, Kreitman M (2003) A method for detecting recent selection in the human genome from allele age estimates. Genetics 165:287-297

Verrelli BC, McDonald JH, Argyropoulos G, Destrol-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA (2002) Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. American Journal of Human Genetics 71:1112-1128

Vulliamy TJ, Durso M, Battistuzzi G, Estrada M, Foulkes NS, Martini G, Calabro V, Poggi V, Giordano R, Town M, Luzzatto L, Persico MG (1988) Diverse point mutations in the human glucose-6-phosphate dehydrogenase gene cause enzyme deficiency and mild or severe hemolytic anemia. Proc Natl Acad Sci U S A 85:5171-5175

Watterson GA (1975) Number of Segregating Sites in Genetic Models Without Recombination. Theor Popul Biol 7:256-276

Zlotogora J, Zeigler M, Bach G (1988) Selection in favor of lysosomal storage disorders. American Journal of Human Genetics 42:271-273

Zlotogora J, Bach G (2003) The possibility of a selection process in the Ashkenazi Jewish population. American Journal of Human Genetics 73:438-440

Zoossmann-Diskin A (1995) ITD in Ashkenazi Jews-genetic drift or selection? Nature Genetics 11:13-14

Table 1: Independent SNPs assayed in study.

SNP label	Distance from *G6PD*[a]	NCBI SNP ID	Celera ID	Gene	SNP Type
A	710	rs2071123	hCV15868309	*IDH3G*	Intron
B	478	rs1059701	hCV8966366	*IRAK1*	Silent coding
C	207	rs2239466	hCV2462509	*TKTL1*	Intron
D	135	rs915943	hCV11927620	*RPL10*	Intron
E	52	rs2315325	hCV2198327	N/A	Intergenic
F	28	rs7611	hCV7493090	*FAM3A*	3' UTR
G	-100	rs4232906	hCV140027	*SPCX*	Intron
H	-118	rs4326559	hCV25649404	*CTAG2*	Silent coding
I	-201	rs2728725	hCV15924058	*GAB3*	Intron
J	-390	rs1936645	hCV11359203	*F8*	Intron

[a]Distance from *G6PD* (kilobases) as determined by distance from *G6PD* coding sites 563 found on NCBI contig NT_025965.12.

Table 2: Summary statistics of nucleotide variability resequenced regions for Kurdish Jews and reference panels.

Locus		n	L[a]	S[b]	θ_π(SD) (%)	θ_W (SD) (%)	T D[c]	FL D[d]	H[e]	H_d[f] (SD)
BGN										
Total sample	Total sequence	37	2889	13	0.134 (0.008)	0.108 (0.030)	0.7817	0.0943	11	0.838 (0.036)
	Introns only		2160	11	0.156 (0.009)	0.122 (0.037)	0.8482	0.3767	10	0.824 (0.037)
G6PD B_{KJ}	Total sequence	13	2889	10	0.135 (0.013)	0.112 (0.035)	0.8367	0.2007	7	0.872 (0.067)
	Introns only		2160	8	0.148 (0.016)	0.119 (0.042)	0.9422	0.3901	6	0.821 (0.082)
$G6PD_{med}$	Total sequence	24	2889	12	0.118 (0.014)	0.111 (0.032)	0.2170	-0.2825	8	0.779 (0.065)
	Introns only		2160	11	0.138 (0.017)	0.136 (0.041)	0.0458	-0.4333	8	0.779 (0.065)
$G6PD_{REF}$	Total sequence	23	2889	11	0.137 (0.009)	0.103 (0.045	1.1160	0.9760	9	0.870 (0.046)
	Introns only		2160	10	0.159 (0.013)	0.125 (0.055)	0.9080	0.9110	8	0.842 (0.054)
LICAM										
Total sample	Total sequence	35	3227	1	0.005 (0.002)	0.008 (0.008)	-0.4983	0.5770	2	0.161 (0.079)
	Introns only		1841	1	0.009 (0.004)	0.013 (0.013)	-0.4983	0.5770	2	0.161 (0.079)
G6PD B_{KJ}	Total sequence	12	3227	0	0.00	0.00	0.00	0.00	1	0.00
	Introns only		1841	0	0.00	0.00	0.00	0.00	1	0.00
$G6PD_{med}$	Total sequence	23	3227	1	0.007 (0.003)	0.008 (0.008)	-0.2132	0.6287	2	0.237 (0.105)
	Introns only		1841	1	0.013 (0.006)	0.015 (0.015)	-0.2132	0.6287	2	0.237 (0.105)
$G6PD_{REF}$	Total sequence	31	3691	1	0.00 (0.003)	0.01 (0.007)	-1.145	-1.681	2	
	Introns only		2087	1	0.00 (0.003)	0.01 (0.007)	-1.145	-1.681	2	
G6PD										
Total sample	Total sequence	37	5109	4	0.032 (0.005)	0.019 (0.009)	1.7074	1.0351	4	0.527 (0.075)
	Introns only		3694	2	0.020 (0.004)	0.013 (0.009)	1.1176	0.7802	2	0.378 (0.074)
G6PD B_{KJ}	Total sequence	13	5109	3	0.026 (0.007)	0.019 (0.011)	1.1151	1.0863	3	0.500 (0.136)
	Introns only		3694	2	0.025 (0.006)	0.017 (0.012)	1.2142	0.9528	2	0.462 (0.110)
$G6PD_{med}$	Total sequence	24	5109	0	0.00	0.00	0.00	0.00	1	0.00
	Introns only		3694	0	0.00	0.00	0.00	0.00	1	0.00

$G6PD_{REF}$	Total sequence	23	4679	3	0.021 (0.003)	0.017 (0.011)	0.550	-0.174	4	0.577 (0.090)
	Introns only		3694	2	0.017 (0.003)	0.016 (0.012)	0.177	-0.646	3	0.530 (0.071)
DMD[†]										
Total sample	Introns only	34	2531	12	0.141 (0.012)	0.116 (0.033)	0.6904	0.4985	7	0.824 (0.036)
G6PD B_{KJ}	Introns only	12	2531	12	0.159 (0.023)	0.157 (0.045)	0.0600	0.4525	6	0.864 (0.072)
$G6PD_{med}$	Introns only	22	2531	10	0.137 (0.015)	0.108 (0.034)	0.9058	1.4086*	5	0.814 (0.041)
$G6PD_{REF}$	Introns only	31	3000	19	0.125 (0.078)	0.112 (0.060)	0.3535			
mtDNA										
Augmented sample	Total Sequence	48	503	42	1.130 (0.079)	1.881 (0.589)	-1.4213	-1.3840	26	0.957 (0.014)
Total sample	Total Sequence	36	503	40	1.222 (0.083)	1.918 (0.633)	-1.2961	-1.1241	22	0.963 (0.015)
G6PD B_{KJ}	Total Sequence	13	504	27	1.277 (0.171)	1.726 (0.715)	-1.1352	-1.2781	12	0.987 (0.035)
$G6PD_{med}$	Total Sequence	23	504	33	1.217 (0.099)	1.774 (0.645)	-1.2014	-0.6629	15	0.960 (0.022)
Ashkenazi Jews	Total Sequence	528	359	N/A	1.4 (0.007)				118	0.957 (0.003)
Near Eastern	Total Sequence	327	359	N/A	1.7 (0.009)				221	0.992 (0.002)

[a]Number of bp analyzed. [b]Number of segregating sites. [c]Tajima's D. [d]Fu and Li's D. [e]Number of haplotypes. [f]Haplotype diversity.

[†]Surveyed region contained no coding sequence.

[§]Augmented sample (see SUBJECTS and METHODS)

$G6PD_{REF}$ samples represent non-African "reference panels" for respective loci. For *mtDNA* HVR1, the Ashkenazi Jews and Near Eastern (non-Jewish) samples are from Behar et al (2004) (see SUBJECTS and METHODS).

FIGURE LEGENDS

Figure 1: Loci sampled on the X chromosome. In region Xq28, three windows were resequenced (*BGN*, *L1CAM* and *G6PD*; hatch marks), and 10 independent SNPs were genotyped (Table 1; ovals). *DMD*, a locus unlinked to *G6PD*, was also resequenced. The distance of each resequenced locus from *G6PD* is marked below the line. The G6PD gene is depicted in inverted transcription orientation with black boxes representing exons. Shaded bar represents the resequenced region. *G6PD* coding site 563, that defines the alleles G6PD B and G6PD$_{med}$, is marked with a white arrow and the respective amino acids are given in parentheses.

Figure 2: Table of polymorphisms of X-linked loci surveyed. Polymorphic sites are shown for *G6PD*, *BGN*, *L1CAM*, *DMD*, and 10 SNPs (labeled *A-J*) located around *G6PD*. Shaded bar underlines loci in region Xq28. Yellow and blue boxes represent the minor and major alleles in the sample, respectively, for each segregating site. Grey boxes indicate missing data. The two primary allele groups (G6PD B, G6PD$_{med}$) are separated by a fine white horizontal line. Vertical bold white lines represent unsurveyed inter-locus regions. The distance from *G6PD* for each resequenced window is marked below the table of polymorphism. (N) represents nonsynonymous polymorphism, (S) represents synonymous polymorphism. Segregating sites in each of the resequenced windows is labeled across the top in sequential order. Alignment site *Gd04* marked with the asterisk designates *G6PD* coding site 563 (*Gd01* designates *G6PD* coding site 1311). Explicit nucleotide states for all polymorphisms are available on the Nachman lab website

(http://eebweb.arizona.edu/faculty/nachman/saunders/publications/G6PD_KJ/sites_table_KJ.html).

Figure 3: Nucleotide variability (θ_π) for Kurdish Jews (KJ) relative to reference panels of non-African G6PD B alleles. For *BGN* and *G6PD* the reference panel is from diverse localities in the Mediterranean and Middle East (n = 23; Saunders APPENDIX C). For *DMD* the reference panel is from a worldwide panel of non-African individuals (n = 31; Nachman and Crowell 2000) that bear G6PD B alleles (Saunders et al. 2002). Nucleotide variability is shown for the total sample of Kurdish Jews (hatched bars), Kurdish Jew samples bearing G6PD B alleles only (shaded bars), and Kurdish Jew samples bearing $G6PD_{med}$ only (bold bars).

Figure 4: Tajima's D for Kurdish Jews (KJ) and reference panels of non-African individuals bearing G6PD B alleles (see figure 3 for details).

Figure 5: Mismatch distribution for *mtDNA* HVR1 for a sample of Kurdish Jews (KJ; n = 48) and reference panels from Behar et al. (2004): Ashkenazi Jews (AJ; n = 528) and a Near Eastern non-Jewish sample (NE; n = 327). Sample O_02 from Kurdish Jews was excluded from the analysis due to missing data from this individual.

Figure 6: Matrix of pairwise values of |D'| for all non-singleton sites among the G6PD B alleles of Kurdish Jews. The sites are labeled as in Figure 2. Values of |D'| are marked in

accordance with the color scheme of the legend. Significant associations based on Fisher's exact test ($p < 0.05$) are marked with an asterisk within the respective box.

Figure 1:

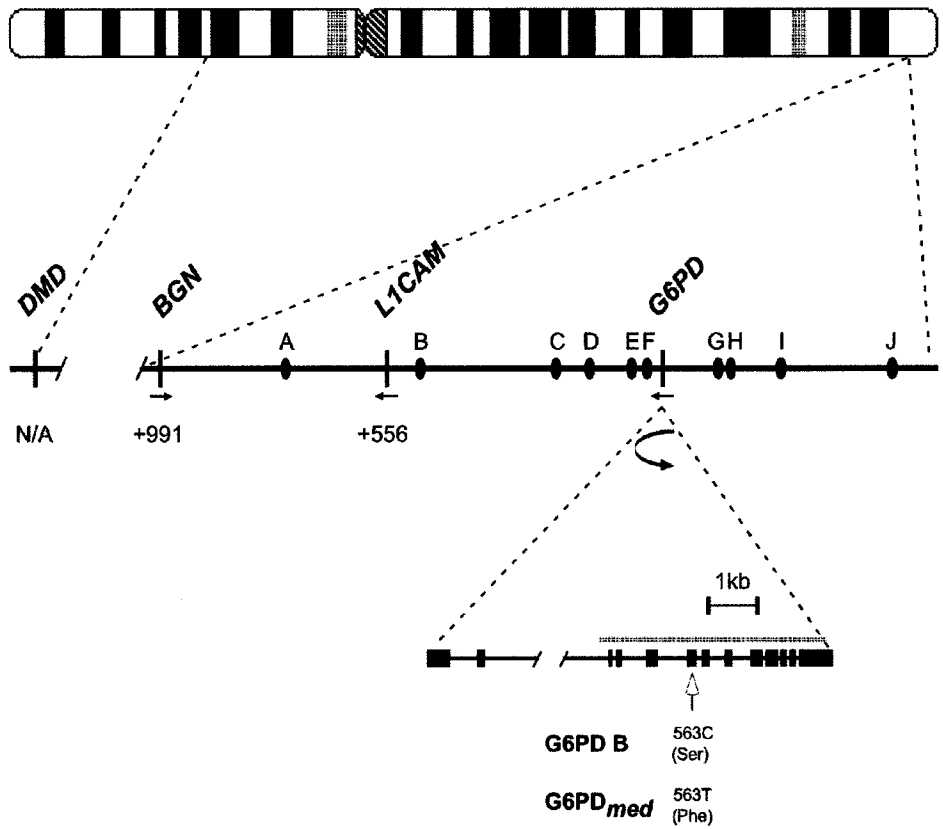

DMD
BGN
L1CAM
G6PD
A
B
C
D
EF
GH
I
J
N/A
+991
+556
1kb
G6PD B
563C
(Ser)
G6PDmed
563T
(Phe)

Figure 2

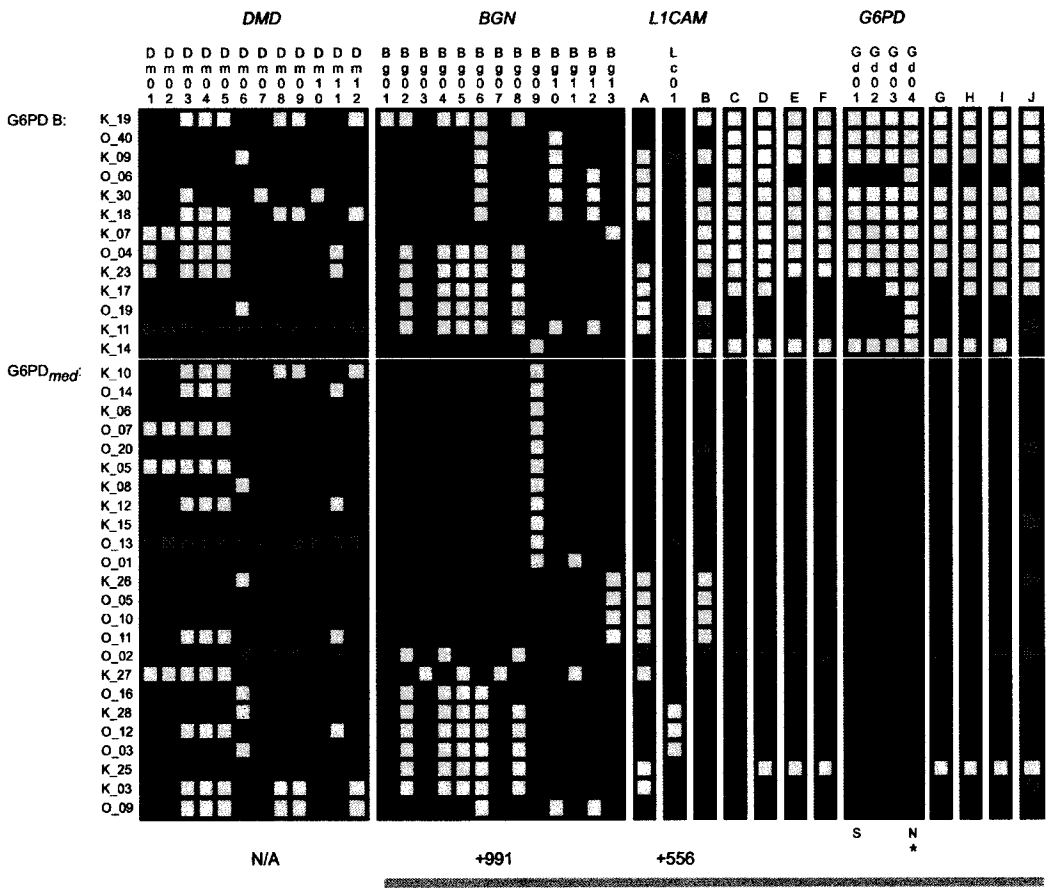

DMD
BGN
L1CAM
G6PD
Dm01 Dm02 Dm03 Dm04 Dm05 Dm06 Dm07 Dm08 Dm09 Dm10 Dm11 Dm12
Bgo01 Bgo02 Bgo03 Bgo04 Bgo05 Bgo06 Bgo07 Bgo08 Bgo09 Bgo10 Bgo11 Bgo12 Bgo13
Lco1
A B C D E F
Gd01 Gd02 Gd03 Gd04
G H I J
G6PD B:
K_19
O_40
K_09
O_06
K_30
K_18
K_07
O_04
K_23
K_17
O_19
K_11
K_14
G6PDmed:
K_10
O_14
K_06
O_07
O_20
K_05
K_08
K_12
K_15
O_13
O_01
K_26
O_05
O_10
O_11
O_02
K_27
O_16
K_28
O_12
O_03
K_25
K_03
O_09
S
N
*
N/A
+991
+556

Figure 3:

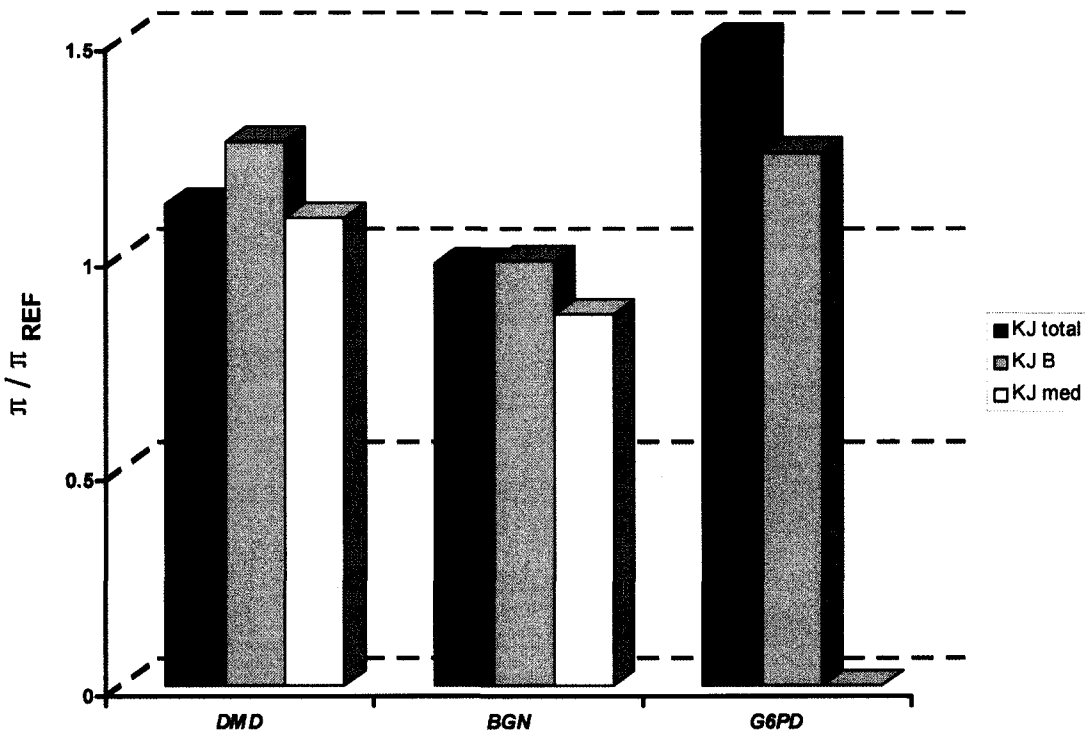


Figure 4:

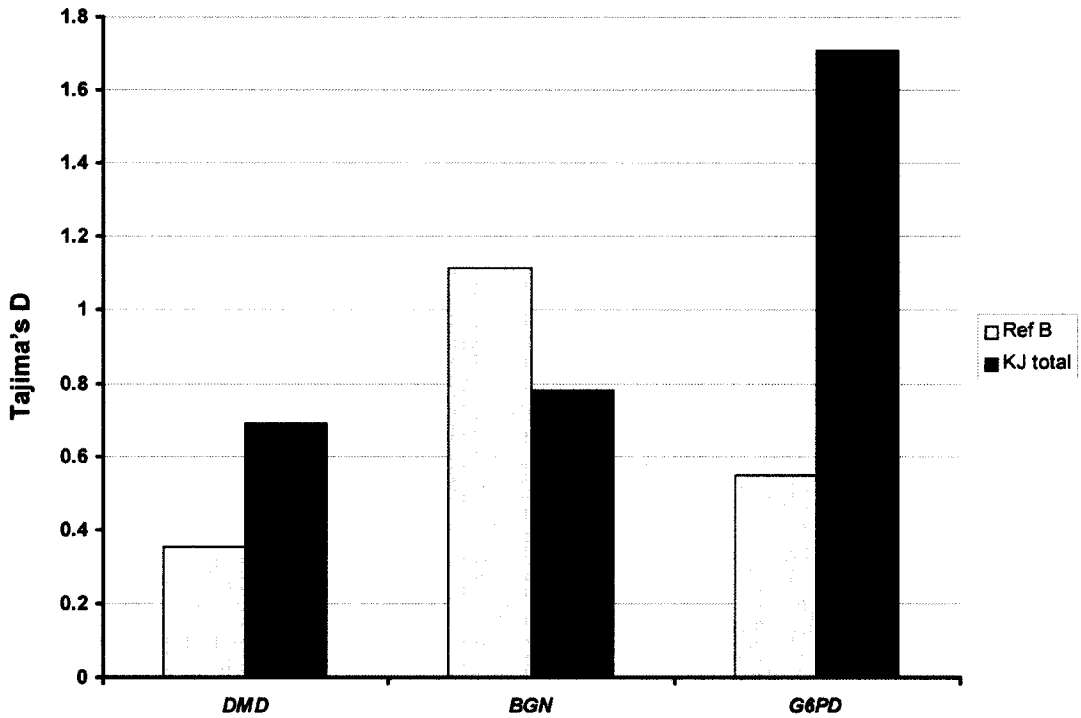


Figure 5:

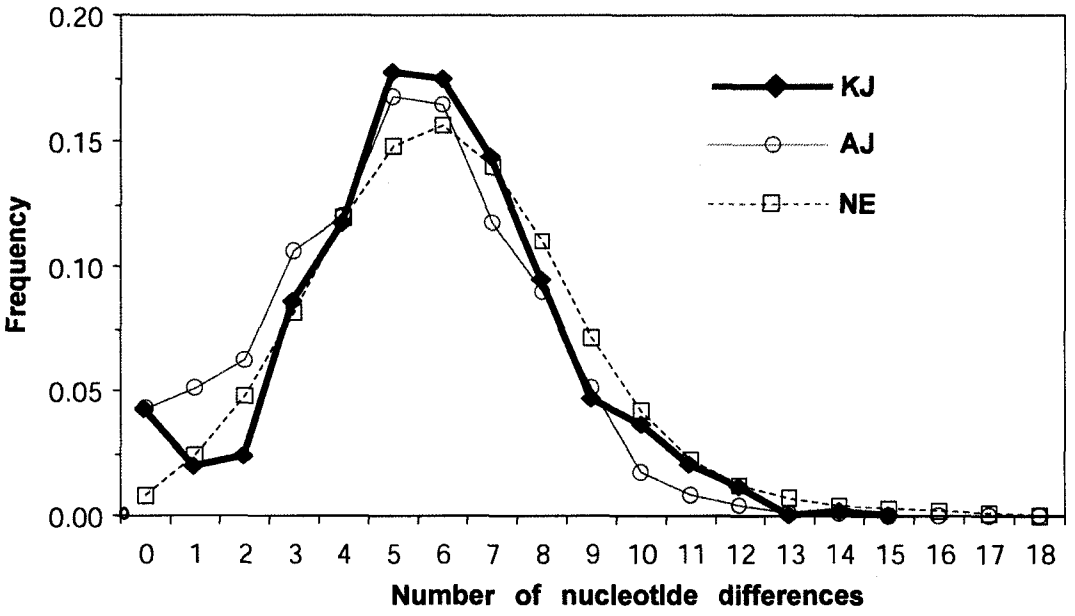
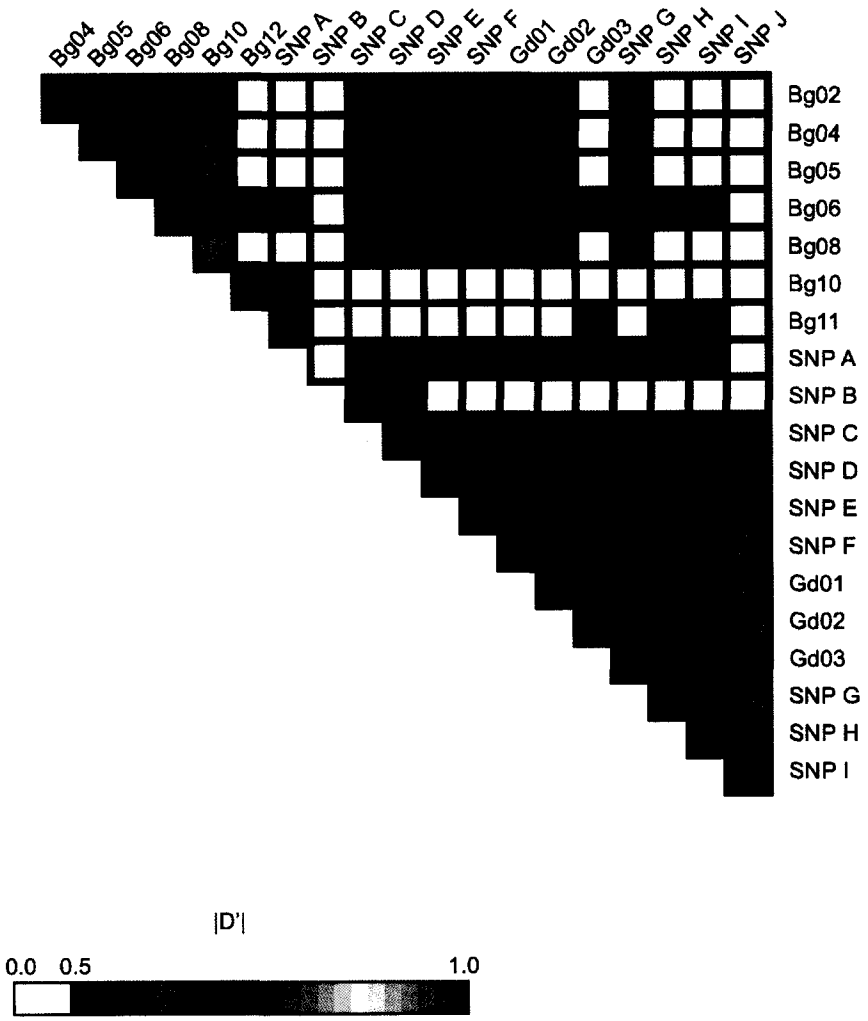

0.20
0.15
0.10
0.05
0.00
Frequency
KJ
AJ
NE
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Number of nucleotide differences

Figure 6:


Bg04
Bg05
Bg06
Bg08
Bg10
Bg12
SNP A
SNP B
SNP C
SNP D
SNP E
SNP F
Gd01
Gd02
Gd03
SNP G
SNP H
SNP I
SNP J
Bg02
Bg04
Bg05
Bg06
Bg08
Bg10
Bg11
SNP A
SNP B
SNP C
SNP D
SNP E
SNP F
Gd01
Gd02
Gd03
SNP G
SNP H
SNP I
|D'|
0.0 0.5
1.0