Haplotype Diversity and Linkage Disequilibrium at Human *G6PD*: Recent Origin of Alleles That Confer Malarial Resistance

Sarah A. Tishkoff,[1,2]* Robert Varkonyi,[2] Nelie Cahinhinan,[2] Salem Abbes,[3] George Argyropoulos,[4] Giovanni Destro-Bisol,[5] Anthi Drousiotou,[6] Bruce Dangerfield,[7] Gerard Lefranc,[8] Jacques Loiselet,[9] Anna Piro,[10] Mark Stoneking,[11] Antonio Tagarelli,[10] Giuseppe Tagarelli,[10] Elias H. Touma,[9] Scott M. Williams,[12]† Andrew G. Clark[2]

The frequencies of low-activity alleles of glucose-6-phosphate dehydrogenase in humans are highly correlated with the prevalence of malaria. These "deficiency" alleles are thought to provide reduced risk from infection by the *Plasmodium* parasite and are maintained at high frequency despite the hemopathologies that they cause. Haplotype analysis of "A−" and "Med" mutations at this locus indicates that they have evolved independently and have increased in frequency at a rate that is too rapid to be explained by random genetic drift. Statistical modeling indicates that the A− allele arose within the past 3840 to 11,760 years and the Med allele arose within the past 1600 to 6640 years. These results support the hypothesis that malaria has had a major impact on humans only since the introduction of agriculture within the past 10,000 years and provide a striking example of the signature of selection on the human genome.

Malaria, resulting from infection by the *Plasmodium falciparum*, *P. vivax*, *P. malariae*, or *P. ovale* parasites, is the leading cause of death in the global human population. Each year 500 million people suffer from malaria, resulting in about 2 million deaths. During the course of human evolution in regions where malaria is prevalent, naturally occurring genetic defense mechanisms have evolved for resisting infection by *Plasmodium*. Most of the human genes that are thought to provide reduced risk from malarial infection are expressed in red blood cells or play a role in the immune system. These loci include human leukocyte antigen (HLA), α- and β-globin, Duffy factor (*FY*), tumor necrosis factor (*TNF*), and glucose-6-phosphate dehydrogenase (*G6PD*).

G6PD catalyzes the first step of the hexose monophosphate pathway and plays a critical role in the metabolism of glucose and the maintenance of balance of reduced/oxidized states of glutathione (important for coping with oxidative stress). G6PD enzyme deficiency, caused by mutations in the *G6PD* gene, is the most common enzymopathy of humans, affecting an estimated 400 million people and resulting in a number of hemopathologies, often triggered by certain foods (e.g., fava beans), drugs, or infection (*1*–*3*). The *G6PD* locus is located on the telomeric region of the long arm of the X chromosome (Xq28) and is flanked 300 kb on either side by the factor VIII and red/green color pigment genes. Nearly 400 G6PD variants have been identified on the basis of electrophoretic and biochemical properties (*1*). The normal activity G6PD B variant is present worldwide, but other variants, particularly those resulting in enzyme deficiency, are restricted to specific geographic regions (e.g., G6PD A and A− in sub-Saharan Africa and G6PD Med in Southern Europe, the Middle East, and India), although they may occur at a low frequency in regions where there has been recent gene flow (*1*, *2*). At the molecular level, more than 130 different mutations have been identified in the *G6PD* gene that result in enzyme deficiency, nearly all of which are single-base substitutions that cause an amino acid substitution (*1*, *2*, *4*).

The distribution of G6PD deficiency is highly correlated with the distribution of current or past malaria endemicity. This observation led to the widely accepted hypothesis that G6PD deficiency confers reduced risk from infection by the *Plasmodium* parasite (*2*). This hypothesis is supported by the observation that patients with G6PD deficiency have lower *P. falciparum* parasite loads than controls and by in vitro studies showing that parasite growth is inhibited in the first few cycles of infection in G6PD-deficient cells [summarized in (*2*, *3*)]. Additionally, a large case-control study of more than 2000 African children demonstrated that the most common form of G6PD deficiency in Africa (G6PD A−) is associated with a 46 to 58% reduction in risk of severe malaria for both female heterozygotes and male hemizygotes (*5*). Ruwende *et al.* (*5*) suggest that the selective advantage conferred by resistance to malarial infection is counterbalanced by a selective disadvantage associated with the hemopathologies associated with enzyme deficiency. Thus, the genetic variability maintained at the *G6PD* locus appears to be an example of a balanced polymorphism that, with the classic examples of sickle cell anemia and thalassemia, represents one of the best examples of natural selection acting on the human genome.

The *G6PD* gene spans about 18 kb and has 13 exons (Fig. 1). The three most common G6PD electrophoretic variants in Africa are G6PD B, which has normal enzyme activity (60 to 80% frequency range), G6PD A, which has 85% normal enzyme activity (15 to 40% frequency range), and G6PD A−, which has 12% normal enzyme activity (0 to 25% frequency range) (*3*, *5*, *6*). Only the A− variant is thought to provide protection against malarial infection in Africa (*5*). The G6PD Med variant has 3% normal enzyme activity and usually ranges in frequency from 2 to 20%, but is as high as 70% among Kurdish Jews (*2*).

To reconstruct the evolutionary history of G6PD deficiency mutations, we have identified three highly polymorphic microsatellite repeats within 19 kb of the *G6PD* locus. Using these microsatellites and restriction fragment length polymorphisms (RFLPs) within the *G6PD* gene, we have examined haplotype variability in geographically diverse human populations, originating from Africa, the Middle East, the Mediterranean, Europe, and Papua New Guinea (*7*), to estimate the age of *G6PD* alleles that confer resistance to malarial infection.

RFLP haplotype analysis. The *G6PD* B, A, A−, and Med alleles can all be detected

[1]Department of Biology, Biology/Psychology Building, University of Maryland, College Park, MD 20742, USA. [2]Institute of Evolutionary Genetics, Department of Biology, Pennsylvania State University, University Park, PA 16802, USA. [3]Faculty of Medicine and Pasteur Institute, Tunis, Tunisia. [4]Pennington Biomedical Research Center, Louisiana State University, Baton Rouge, LA 70808, USA. [5]Department of Human and Animal Biology, University "La Sapienza," Rome, Italy. [6]Department of Biochemical Genetics, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. [7]Department of Human Genetics, South African Institute of Medical Research, University of the Witwatersrand, Johannesburg, South Africa. [8]University of Sciences and CNRS, Montpellier, France. [9]University St. Joseph, Beirut, Lebanon. [10]Istituto di Medicina Sperimentale e Biotecnologie–CNR, Mangone (Cosenza), Italy. [11]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [12]Department of Microbiology, Meharry Medical College, Nashville, TN 37208, USA.

*To whom correspondence should be addressed. E-mail: st130@umail.umd.edu

†Temporary address: Montreal Genome Center, Montreal General Hospital Research Institute, Montreal, Quebec H3G 1A4, Canada.

by polymerase chain reaction (PCR) and RFLP analysis (Fig. 1) (*8*). Frequencies of the A− allele in our sample of sub-Saharan African populations range from 3% to 19% (Table 1) and exhibit significant heterogeneity across populations ($X^2_{7df} = 17.57$, $P < 0.03$). In the North African, Middle Eastern, and Mediterranean populations, RFLP analysis to distinguish *G6PD* A, A−, or Med alleles was performed only in individuals with deficient enzyme activity levels (Table 1) (*9*). Therefore, we were not able to determine unbiased allele frequencies. Other studies have estimated the frequency of G6PD Med to range from ~2 to 10% in these populations (*10*, *11*). We refer to chromosomes outside of Africa without deficiency mutations as Norm to distinguish them from B chromosomes in Africa, because they have distinct patterns of haplotype variation and were ascertained differently. Only four polymorphisms in noncoding regions and one polymorphism at a synonymous site of the *G6PD* gene have been identified in human populations (*1*, *12*). Only one of these RFLPs (Bcl I) is polymorphic outside of Africa (*13*) and has been analyzed with the Mbo II RFLP that defines the Med allele. Our results confirm previous reports that Norm chromosomes are nearly always associated with the Bcl I (−) allele, whereas the Med allele is most frequently associated with the Bcl I (+) alleles in North African, Middle Eastern, and Mediterranean populations, but with the Bcl I (−) allele in Eastern Indian populations (*11*, *13*).

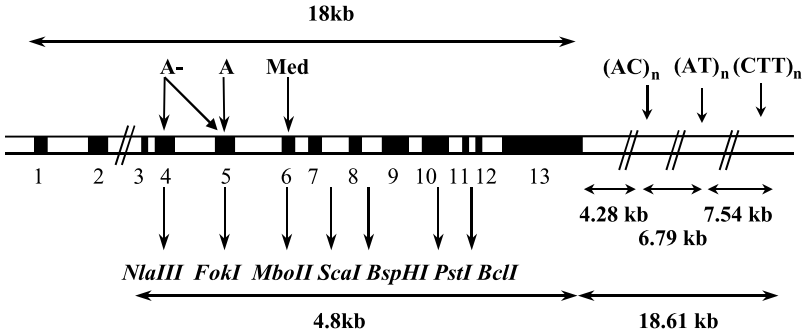

Fig. 1. Diagram of *G6PD* gene structure showing the location of the RFLPs and microsatellites used in the haplotype analysis. Exons are shown as solid boxes. The G6PD A allele results from an A to G transition at nucleotide 376 in exon 5, causing an amino acid change from Asn to Asp (*58*). The most common G6PD A− variant in Africa has the mutation at nucleotide 376 and a second G to A transition at nucleotide 202, causing a Val to Met amino acid change (*59*). The G6PD Med variant results from a mutation at nucleotide position 563, causing an amino acid change from Ser to Phe (*13*). The mutation resulting in G6PD A creates a Fok I site, the mutations resulting in G6PD A− create an Nla III site in addition to the Fok I site (*8*, *12*), and the mutation resulting in G6PD Med creates an Mbo II site (*13*). G6PD B lacks these three restriction sites. The Sca I, Bsp HI, Pst I, and Bcl I RFLP sites detect noncoding, or silent, substitution mutations (*8*, *12*). The Sca I and Bsp I restriction sites were created by mismatch-containing primers (*12*). The $(AC)_n$ repeat, located 4.28 kb downstream of *G6PD* at GenBank sequence position 22,359 (*14*), is a highly compound repeat consisting of the sequence $(TA)_5(AA)_1(TA)_6(CA)_6(CT)_1(CA)_1(TA)_1(CA)_{10}$ (corresponding to a 178-bp repeat) (*60*). The $(AT)_n$ repeat, located 11.07 kb downstream from *G6PD* at GenBank sequence position 29,191, consists of a perfect $(AT)_{14}$ repeat (corresponding to a 135-bp allele). The $(CTT)_n$ repeat, located 18.61 kb downstream from *G6PD* at GenBank sequence position 36,756, is a compound repeat consisting of the sequence $(CTT)_{11}(ATT)_7$ (corresponding to a 198-bp allele).

Microsatellite analysis. Few RFLP haplotypes have previously been identified at the *G6PD* locus because the RFLPs have low heterozygosity, and there is strong linkage disequilibrium (LD) between markers located within a 3-kb region of the gene (*12*). Additionally, because only one RFLP is polymorphic outside of Africa, RFLP haplotype analyses were not informative for reconstructing the evolutionary history of the *G6PD* gene in non-African populations. Therefore, we screened 52,173 base pairs (bp) containing the *G6PD* gene and flanking sequence (*14*) for potentially variable microsatellite repeats. Three microsatellite repeats, referred to as AC, AT, and CTT, were identified within a

Table 1. Population samples, number of chromosomes typed, and G6PD allele counts. G6PD A and A− variants in the sample of sub-Saharan African populations were detected by screening populations for the Fok I and Nla III restrictions sites (*8*). The G6PD deficiency phenotype in the North African, Lebanese, Cypriot, and Italian populations was identified on the basis of enzyme activity levels (*9*), and therefore unbiased allele frequencies could not be obtained. *N* represents the number of chromosomes included in the study. For Sub-Saharan African populations, chromosomes were typed and the frequency of G6PD B, A, and A− alleles was counted. Standard errors are shown in parentheses. For North African and non-African populations, chromosomes were typed and G6PD alleles were counted. # def M and # def F denote the number of males and females who had the G6PD deficiency phenotype; # Other indicates chromosomes for which we have not yet identified the mutation underlying the deficiency phenotype. PNG, Papua New Guineans.

Sub-Saharan African	*N*	#Males	#Females	B	A	A−
South African Bantu-speakers	124	42	41	0.815 (0.035)	0.097 (0.027)	0.089 (0.026)
Sierra Leone (mixed)	30	30	0	0.733 (0.072)	0.233 (0.069)	0.033 (0.029)
Mende	99	99	0	0.697 (0.046)	0.273 (0.045)	0.030 (0.017)
Temne	38	38	0	0.553 (0.081)	0.342 (0.077)	0.105 (0.050)
Ghana (mixed)	19	5	7	0.842 (0.084)	0.053 (0.051)	0.105 (0.070)
Fante	35	3	16	0.743 (0.074)	0.200 (0.068)	0.057 (0.039)
Ga	69	13	28	0.536 (0.060)	0.275 (0.054)	0.189 (0.047)
Cameroon (mixed)	16	16	0	0.563 (0.124)	0.313 (0.116)	0.125 (0.083)
Bakaka	23	19	2	0.783 (0.086)	0.174 (0.079)	0.043 (0.042)
Total	453	265	94			

North African/non-African	*N*	#Males	#Females	#Norm	#def M	#def F	#Med	#A	#A−	#Other
Tunisians	56	40	8	49	13	0	0	2	5	6
Lebanese	79	45	17	47	32	2	28	2	2	2
Cypriots	47	19	14	34	11	1	13	0	0	0
Italians	72	66	3	36	46	3	33	2	1	13
Europeans	15	5	5	15	0	0	0	0	0	0
PNG	25	23	1	25	0	0	0	0	0	0
Total	294	198	48	206	108	6	74	6	8	2

19-kb region downstream of the *G6PD* gene (Fig. 1). In total, we observed 10 $(AC)_n$ alleles (ranging from 164 to 188 bp), 26 $(AT)_n$ alleles (ranging from 125 to 179 bp), and 8 $(CTT)_n$ alleles (ranging from 195 to 216 bp) (*15*, *16*). Allele frequencies and heterozygosity values for these microsatellites are presented in Appendix A (*17*). Allele number and heterozygosity levels are higher for the perfect AT repeat than for the interrupted repeats. Bootstrap samples of equal size from African and non-African populations revealed more alleles in African populations for the $(AC)_n$ and $(AT)_n$ repeats ($P < 0.001$) and higher heterozygosity levels in sub-Saharan African populations for all three microsatellite repeats ($P < 0.001$).

Microsatellite haplotypes and linkage disequilibrium. Haplotypes consisting of the $(AC)_n$, $(AT)_n$, and $(CTT)_n$ microsatellites and the RFLPs distinguishing B, A, A−, Med, and Norm alleles were typed in 591 chromosomes from individuals originating from ethnically diverse sub-Saharan African, Tunisian, Lebanese, Cypriot, Italian, European, and Papua New Guinean populations (*7*). Chromosomes with Norm and Med alleles were further characterized by presence (+) or absence (−) of the Bcl I site. Linkage phase could be determined unambiguously in males, which constitute ~80% of the sample (Table 1). In multiply heterozygous females, linkage phase could not be determined unambiguously, so statistical inference had to be applied (*18*). A total of 149 distinct AC/AT/CTT haplotypes were identified and are presented in Appendix B (*17*).

Generally, the greatest haplotype diversity is found on B and A chromosomes from Africa ($H = 0.96 \pm 0.02$ and 0.91 ± 0.04, respectively), moderate levels of diversity on Norm/Bcl I(−) and Norm/Bcl I(+) chromosomes outside of Africa ($H = 0.87 \pm 0.03$ and 0.86 ± 0.10, respectively), and the most restricted variability on A− ($H = 0.72 \pm 0.08$), Med/Bcl I(+) ($H = 0.18 \pm 0.04$), and Med/Bcl I(−) ($H = 0.38 \pm 0.15$) chromosomes (Fig. 2) (*19*). Distinct patterns of microsatellite haplotype variability and of LD were associated with the various *G6PD* alleles (Figs. 2 and 3). G6PD A− alleles are always associated with a 166-bp AC allele, and G6PD A alleles are always associated with either a 164- or 166-bp AC allele. There are broad ranges of AT and CTT alleles on A chromosomes, whereas A− alleles are associated with only large-sized AT alleles (ranging from 165 to 179 bp in size) and nearly always with a 195-bp CTT allele. By contrast, B chromosomes from Africa have primarily large AC alleles, 176 to 186 bp in size (with 182- to 184-bp alleles most common), as well as a broad range of AT and CTT alleles.

Norm/Bcl I(−) chromosomes from Tunisia and outside of Africa appear to have a subset of the haplotype variability present on B chromosomes in sub-Saharan Africa, with only three common microsatellite haplotypes (containing a 178-bp AC allele, 137- to 141-bp AT alleles, and a 198-bp CTT allele) (Fig. 2). Only 17 Norm/Bcl I(+) chromosomes were in our sample and the majority have a 182-bp AC allele, large-sized AT alleles (147 to 159 bp), and a 210-bp CTT allele. Most G6PD Med alleles were associated with the Bcl I(+) allele. Of the Med/Bcl I(+) chromosomes, 57 out of 63 carried the 182/151/210 haplotype. Of the six chromosomes that

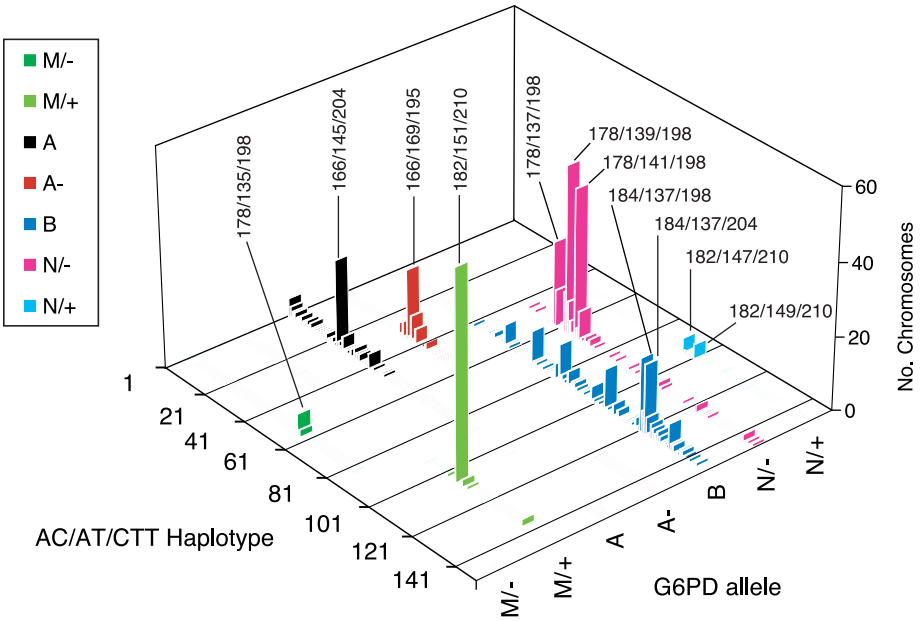

Fig. 2. Relative frequencies of AC/AT/CTT microsatellite haplotypes on B chromosomes ($n = 183$), A chromosomes ($n = 90$), A− chromosomes ($n = 42$), Norm/Bcl I(−) chromosomes ($n = 188$), Norm/Bcl I(+) chromosomes ($n = 17$), Med/Bcl I(+) chromosomes ($n = 63$), and Med/Bcl I(−) chromosomes ($n = 8$). The 149 microsatellite haplotypes identified are ordered by size of the AC, then AT, then CTT repeats, and full haplotype identities and frequencies are given in Appendix B (*17*).

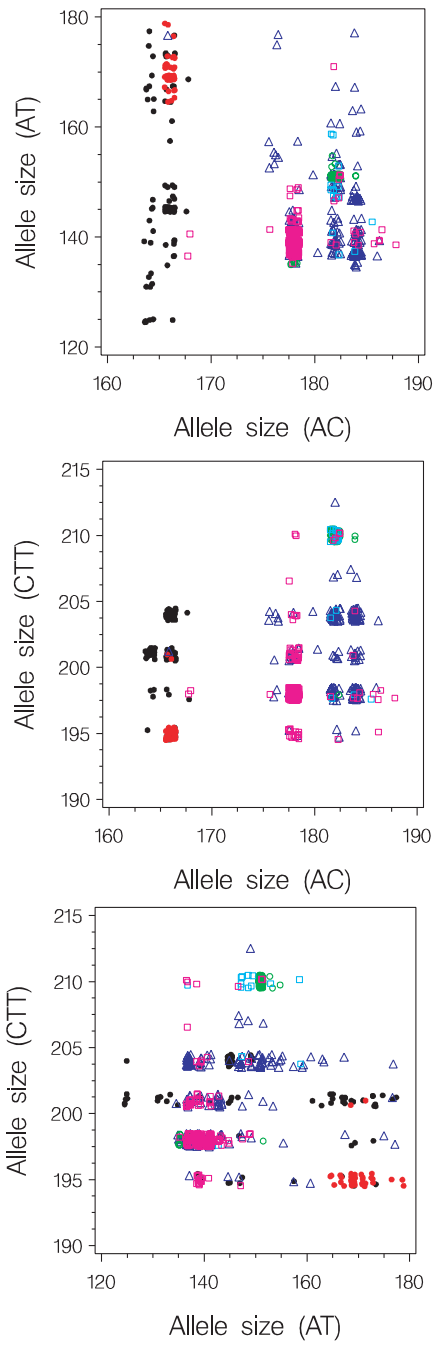

Fig. 3. Plot of the distribution of microsatellite alleles on chromosomes with different *G6PD* alleles. Linkage disequilibrium between the *G6PD* alleles and microsatellites is indicated by the clustering of points. A: filled black circles; A−: filled red circles; B: open, dark blue triangles; Norm/Bcl I(−): open, magenta squares; Norm/Bcl I(+): open, light blue squares; Med/Bcl I(+): open, light green circles; Med/Bcl I(−): open, dark green circles.

do not have this haplotype, five differ only by a single AC or AT repeat, and one (182/151/198) appears to be a recombinant at the CTT site. The 182/151/210 haplotype is very rare in the global sample and has been observed on only two Norm/Bcl I(−) chromosomes from Cyprus and Italy, consistent with a rapid expansion in frequency of G6PD Med alleles. The clustering of microsatellite alleles on A− and Med chromosomes indicates pairwise LD with the G6PD deficiency alleles that is highly significant for all three repeats ($P < 0.01$, Fisher's exact test) (Fig. 3) (*20*). The pattern of haplotype variability and LD was nearly identical in all populations sampled across geographically diverse regions, indicating a single common origin of the A− allele in Africa and the Med allele in the Mediterranean and Middle East.

Eight Med/Bcl I(−) chromosomes were identified in the southern Italian population originating from the Calabria region (21). All carry microsatellite haplotypes identical to the most common haplotypes on Norm/Bcl I(−) chromosomes. Without more detailed haplotype analysis of markers upstream of *G6PD*, it is not possible to distinguish whether the Med/Bcl I(−) haplotypes arose from recombination between Med/Bcl I(+) and Norm/Bcl I(−) haplotypes or whether the Med allele arose independently on a Norm/Bcl I(−) background. All individuals with the Med/Bcl I(−) haplotype are also red-green color blind, indicating that LD extends at least 300 kbp downstream in these individuals (*21*).

RFLP/microsatellite haplotype analysis. To reconstruct the evolutionary history of the A, A−, and B chromosomes in the sub-Saharan African populations, we selected a subset of these chromosomes for a more extensive RFLP/microsatellite haplotype analysis. In addition to the Fok I and Nla III sites that distinguish A, A−, and B alleles, four noncoding, or silent, RFLPs located within a 3-kb region of the *G6PD* gene were typed (Fig. 1). Two distinct clades are formed by the A/A- chromosomes and the B chromosomes (Fig. 4). All A/A− chromosomes carry a 164- to 166-bp AC allele, and all but one B chromosome possesses a 176- to 186-bp AC allele. B chromosomes have the greatest RFLP haplotype diversity and also have high levels of microsatellite haplotype diversity on the two most common G6PD B haplotypes. This observation supports the hypothesis that the B allele is ancestral, as indicated by sequence data from humans and chimpanzees (*6*). On the A chromosomes, one major RFLP haplotype, "− + − − + −", has high levels of microsatellite haplotype diversity and is most likely the ancestral A haplotype. It can be derived from the most common B haplotype, "− − − + + −", by two mutational steps. All but 2 of the 42 A− chromosomes have a single RFLP haplotype, "+ + + − + −", and have reduced microsatellite haplotype diversity, consistent with the hypothesis of a recent origin of the A− allele from an A chromosome (*12*).

Coalescent simulations to assess fit to strict neutrality. Three features that stand out in both the African and Mediterranean data are reduced microsatellite variability in the A− (and Med) clades, striking patterns of LD between the major *G6PD* alleles and the microsatellites, and AT and CTT microsatellite allele sizes of the A− (and Med) clades that are distinct from others in the rest of the genealogy. Coalescence analysis was applied to reveal whether patterns of variation in 315 African haplotypes and 294 non-African haplotypes are consistent with a neutral model (*22*). The coalescent simulations generated null distributions under neutrality for haplotype diversity within the A− or Med allelic lineages. The simulations generally produced levels of microsatellite variation in the A− clade that were greater than those observed in the population samples (Table 2). In the population samples, the A− clade also had significantly fewer microsatellite alleles, lower variance in allele size, and higher maximum LD (δ_{max}) than those generated by the neutral coalescent model. An analysis of variance (ANOVA) *F* statistic (*22*) also indicated that observed AT microsatellite alleles exhibited significantly greater differences in allele sizes between A− and B clades than would be expected under neutrality. For the Mediterranean data, all three microsatellite loci differ significantly from the neutral coalescent tree for the number of observed alleles, the microsatellite size variance, δ_{max}, and the *F* statistic. Overall, the A− and Med clades are far more constrained in their variability and exhibit greater LD than the neutral coalescent would predict. Hence, forces other than drift have resulted in a rapid expansion of these alleles, giving them a relatively high frequency and broad geographic distribution without sufficient time to generate within-clade heterogeneity (*23*).

Age estimates of the A− and Med alleles. Although methods exist to incorporate natural selection into the framework of the coalescent (*24*, *25*), these methods are not easily adapted to infer the age of an allele when positive selection is present. Therefore, we simulated a Poisson branching process in a growing population to estimate the age of the deficiency alleles (A− and Med), drawing parameters from prescribed prior distributions and subjecting each simulation run to rejection criteria (*26*, *27*). In these simulations, the A− deficiency mutation was assumed to occur uniquely within sub-Saharan Africa, and the Med deficiency was assumed to occur uniquely in the Mediterranean or north African populations. Our analysis required the following parameters: the mutation

Table 2. Coalescence simulations. Results of coalescence simulations designed to test the correspondence of the four given sample statistics for the A− clade in the African sample and the Med-containing samples to those generated by a neutral coalescent (*22*). $n_{alleles}$ is the number of distinct microsatellite alleles in the given class; STR variance is the variance in microsatellite allele size; δ_{max} is the LD value for the microsatellite allele in strongest LD with the A− or Med allele; and F_{ANOVA} is the *F* statistic from the ANOVA contrasting the microsatellite size of A− versus non-A− alleles (and Med versus non-Med alleles). Table entries are the respective statistics obtained from the actual data, and the asterisks indicate the significance of the deviation of these observed figures from the null distributions generated by the coalescent.

Allele	STR	$n_{alleles}$	STR variance	δ_{max}	F_{ANOVA}
A−	AC	1***	0***	0.994***	54.13
	AT	7	9.92	0.976**	157.0**
	CTT	2***	2.04***	0.936***	126.99
Med	AC	2**	0.122***	0.897**	152.78**
	AT	4*	2.66**	0.986***	128.97**
	CTT	2***	2.25***	0.925**	121.50**

*$P < 0.05$. **$P < 0.01$. ***$P < 0.001$.

Table 3. Mean and 95% credibility intervals of allele age and model parameters. Results of simulations of a Poisson branching process incorporating selection for the A− (and Med) deficiency chromosomes (*28*). Table entries are the means of the posterior distributions obtained from the simulations, and figures in parentheses are the 95% credibility intervals. Although we expect the mutation rate and recombination rate to be the same in Africa as in the Mediterranean, these two parameters were estimated independently in the two samples.

Parameter	A−	Med
Mutation rate (μ)	3.05×10^{-4} [$(1.4–4.8) \times 10^{-4}$]	4.84×10^{-4} [$(1.6–8.9) \times 10^{-4}$]
Selection coefficient (*s*)	0.044 (0.019–0.048)	0.034 (0.014–0.049)
Recombination rate (*r*)	1.99×10^{-4} [$(0.2–4.3) \times 10^{-4}$]	2.25×10^{-4} [$(0.2–4.6) \times 10^{-4}$]
Allele age (years)	6,357 (3,840–11,760)	3,330 (1,600–6,640)

rate (μ) at the microsatellite, the selective advantage (s) of genotypes bearing A− (or Med) alleles, and the recombination rate (r) between the *G6PD* locus and the microsatellites. A fitness of $1 + s$ was assumed for the A−/Y males relative to non-A−/Y males and for the A−/non-A− heterozygous females relative to the non-A−/non-A− females. We used both a dominance model, where A−/A− homozygous females had a fitness $1 + s$, and an overdominance model, where A−/A− females had a fitness 1 (*28*).

The credibility intervals for the mean mutation rate of microsatellite loci, the selection coefficient for deficiency alleles, and the rate of recombination to the closest microsatellite locus were very similar for the A− and Med chromosomes (Table 3). The local rate of recombination in this region of the X chromosome is ~1.64 cM/Mbp (*29*), which corresponds to a recombination rate of 2.95×10^{-4} for an 18-kb span, a figure that is within the credibility interval. Although mutation and recombination rates appear consistent across the A− and Med alleles, the A− lineage appears to be older than the Med lineage (Table 3). The mean age of the A− allele in these runs was 6357 years, with a 95% credibility interval extending from 3840 to 11,760 years. For the Med alleles, the mean age was 3330 years with a 95% credibility interval of 1600 to 6640 years (*30*).

Evolutionary history of the *G6PD* locus. Our RFLP and microsatellite haplotype analyses support the hypothesis that B alleles are ancestral and that A alleles are more recently derived (*1*, *6*, *12*) (Fig. 4). The data are also consistent with the expectation that the highly compound AC repeat should be more stable than the AT and CTT repeats and should, therefore, remain in LD with the *G6PD* alleles. However, there has been sufficient time to accumulate considerable variation at the AT and CTT microsatellites on both B and A chromosomes as a result of microsatellite mutation and/or recombination. The maintenance of LD between *G6PD* deficiency alleles and the microsatellite alleles is consistent with previous reports indicating strong LD in this part of the X chromosome (*21*), as well as a recent origin of the G6PD deficiency alleles. Sequence analysis of a 5-kb region of the *G6PD* gene in 50 African individuals and in chimpanzees indicates that although the B allele currently dominates in frequency worldwide and is the inferred ancestral state, A chromosomes also show high levels of nucleotide variation (*31*), suggesting that the A allele was historically greater in frequency. This observation is also consistent with the high level of microsatellite variation linked to the A haplotype class in this current study. The maintenance of two distinct clades (B and A/A−) at the *G6PD* locus, as indicated by sequence and microsatellite haplotype analysis, could indicate historical balancing or directional selection acting on B and/or A alleles (or at closely flanking loci).

RFLP and microsatellite haplotype analyses suggest that the G6PD A− allele arose on an A chromosome containing a 164/169/195 haplotype (the most common A− haplotype) and then spread rapidly across a broad geographic range in Africa, with time for only a limited amount of variation at the AT repeat to accumulate. The similar pattern of haplotype variability and LD across geographically diverse African populations at the *G6PD* locus contrasts markedly with the divergent pattern of haplotype variation and LD observed across African populations at the *CD4*, *DM*, *PLAT*, and *PAH* loci (*31*–*35*) and likely reflects the effects of selection at the *G6PD* locus. The few A and A− chromosomes observed outside Africa have patterns of haplotype variation that are identical to that observed in Africa and likely originate from recent gene flow from Africa.

The Norm chromosomes outside Africa may descend from a subset of the B chromosomes that were carried by a small founding population(s) during the migration of modern humans out of Africa within the past 100,000 years (*32*). Genetic drift at the time of this founding event may have resulted in the distinct pattern of haplotype variability observed on normal chromosomes outside Africa (Fig. 2). The Med mutation most likely arose on a normal chromosome with a 182/151/210 haplotype background [possibly on a Norm/Bcl I(+) chromosome] and spread rapidly throughout the Middle East and Mediterranean region. The presence of Med and A− alleles on distinct microsatellite haplotypes supports the conclusion that they arose independently. The high frequency and broad geographic range of these deficiency mutations, in the face of low haplotype variability and high LD, is inconsistent with a model of neutrality. Rather, our results support the hypothesis that the A− and Med mutations have attained high frequency as a result of selection at this locus, most likely in response to malaria infection caused by the *Plasmodium* parasite. Thus, the pattern of haplotype variability and LD at *G6PD* represents an excellent example of the signature of selection on the human genome.

Origins of malarial resistance in humans. The high variability and mutation rate of the three microsatellite markers at *G6PD* make it possible to obtain a reasonably well-bounded estimate of the origin of alleles that confer protection against malaria. We estimate that the A− mutation arose within the past 3840 to 11,760 years. This estimate

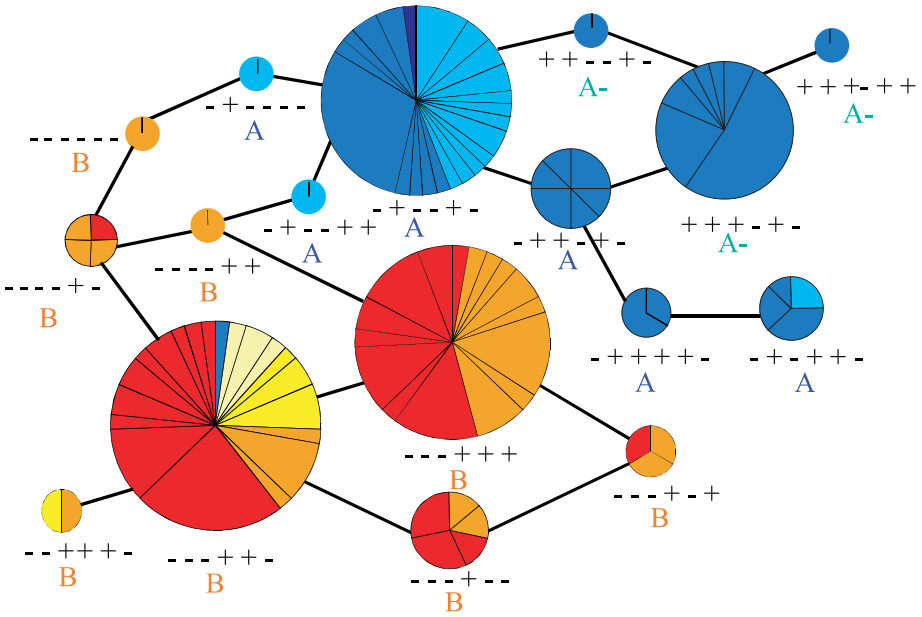

Fig. 4. RFLP/microsatellite haplotype network. Each circle represents an RFLP haplotype indicated by the presence (+) or absence (−) of restriction sites for the Nla III, Fok I, Sca I, Bsp HI, Pst I, and Bcl I RFLPs (Fig. 1), and the size of the circle is approximately proportional to the number of chromosomes observed. Each line between pie charts indicates a single mutation and/or recombination event, and the length of the line is not correlated with the number of mutational events separating haplotypes. Microsatellite haplotype diversity on each RFLP haplotype background is indicated by slices of the pie chart. Microsatellite haplotypes are color coded as follows on the basis of the size of the AC repeat: 164-bp alleles (light blue), 166-bp alleles (dark blue), 168-bp alleles (darkest blue), 176-bp alleles (light yellow), 178-bp alleles (dark yellow), 182-bp alleles (orange), and 184-bp alleles (red). Of the 250 chromosomes typed for all nine polymorphisms, only the 186 chromosomes that had no missing data and could be unambiguously phased are included in the network analysis. In total, 97 B, 61 A, and 28 A− chromosomes are included in the cladogram.

is consistent with archaeological and historical documents indicating that malaria has had a significant impact on humans only within the past 10,000 years, coincident with the origination and spread of agriculture in the Middle East and Africa (*36*–*39*). According to Livingstone (*36*, *37*), the introduction of slash and burn agriculture in West Africa about 2000 to 4000 years ago resulted in the clearing of tropical forests and an increase in sunlit pools of water. The increased number of *Anopheles gambiae* breeding places resulted in an increase in the population density of *A. gambiae*, the major vector for *P. falciparum* parasite, the *Plasmodium* species associated with more severe, stable, hyperendemic malaria (*38*). Additionally, agriculture enabled increased human population density, facilitating the spread of malaria.

However, a number of factors may have caused malaria to become hyperendemic slightly earlier in Africa, as our date estimate suggests. Africa underwent an increase in both temperature and humidity between 12,000 and 7000 years ago, along with a concurrent increase in the number of sunlit lakes and ponds (*40*, *41*); these conditions support the spread and rapid adaptive speciation of the *A. gambiae* vector (*42*). Two other pieces of evidence also indicate an earlier increase in human population density in the Sahara and northeast Africa, allowing for the importance of malaria as a selective agent. First, plant and animal domestication originated about 8000 to 10,000 years ago in this region (*37*, *40*), leading to conditions that could facilitate the spread of infectious disease. Second, archaeological evidence indicates denser and more permanent populations around lakeshores owing to the spread of fishing industries, as well as to incipient cattle domestication in these regions (*40*, *43*). These population settlements on or near lakeshores and water pools could have served as adequate preconditions for the spread of mosquito-borne pathogens (*40*).

Our date estimates are consistent with studies of genetic diversity in the *P. falciparum* genome suggesting a recent population bottleneck followed by rapid population expansion within the past 5000 to 50,000 years (*44*) [but for an alternative perspective see (*45*)]. Our date estimates are also consistent with studies of genetic diversity in the *A. gambiae* genome, suggesting rapid adaptive speciation and the emergence of more anthropophilic taxa within the past 10,000 years (*42*). Although mild forms of malaria may have existed in humans throughout much of their evolutionary history, our data suggest that more severe malaria did not become hyperendemic until the past 10,000 years, likely in response to climatic and/or cultural changes that facilitated population expansion and diversification of the *Anopheles* vector, the *P. falciparum* parasite, and the human host.

The more recent spread of the Med allele within the past 1600 to 6640 years is consistent with historical Greek and Egyptian documents indicating that, despite the earlier presence of more mild forms of malaria resulting from infection by *Plasmodium malariae* and *P. vivax*, the more severe *P. falciparum* malaria may not have been prevalent in the Mediterranean until after 500 B.C. (*39*). Thus, the selective pressure of severe malarial infection may have increased more recently in the Mediterranean region. It is possible that the recent and rapid spread of the Med allele across a broad geographic region may correspond with the spread of agriculture during a Neolithic expansion and migration across Europe from the Middle East 10,000 to 5000 years ago (*46*). However, our date estimate suggests that this mutation could have been spread by more recent migration events, perhaps as a result of the extensive trade routes and colonizations of the Greeks into these regions in the first several millennia B.C. (*11*, *47*). It is even conceivable that the Med mutation was spread throughout this region by the army of Alexander the Great, which invaded and conquered territories ranging from the Mediterranean to India, the Middle East, and even North Africa during the fourth century B.C. (*47*). Thus, the study of polymorphism at the *G6PD* locus demonstrates how the environment, culture, genes, and history interact to shape variation in the modern human genome.

Note added in proof: Since the submission of this manuscript, Saunders *et al.* have reported sequencing 41 G6PD alleles and similarly found low A− haplotype diversity and a recent time of origin of the A− allele (*61*).

References and Notes

1. T. Vulliamy, P. Mason, P. L. Luzzatto, *Trends Genet.* **8**, 138 (1992).
2. E. Beutler, *Blood* **84**, 3613 (1994).
3. C. Ruwende, A. Hill, *J. Mol. Med.* **76**, 581 (1998).
4. L. Luzatto, A. Mehta, T. Vulliamy, in *The Metabolic and Molecular Bases of Inherited Disease*, C. R. Scriver, A. L. Beaudet, W. S. Sly, D. Valle, Eds. (McGraw-Hill, New York, ed. 8, 2001), pp. 4517–4553.
5. C. Ruwende *et al.*, *Nature* **376**, 246 (1995).
6. E. Beutler, W. Kuhl, J. L. Vives-Corrons, J. T. Prchal, *Blood* **74**, 2550 (1989).
7. DNA was isolated by standard phenol-chloroform extraction and/or a Gentra PureGene kit. All samples are from unrelated individuals and were obtained with informed consent and with the approval of the human subjects review panels of participating institutions. The Papua New Guinean samples come from both highland and coastal locations (*48*). The Bantu-speaking populations are from ethnically mixed South African Bantu speakers. The Mende and Temne ethnic groups originate from Sierra Leone (*49*), the Ga originate from Ghana (*50*), and the Bakaka from Cameroon (*51*). We also included samples from assorted ethnic groups originating from Sierra Leone (Creole, Fula, Limba, Loko, Mandingo), Ghana (Akan, Ashante, Ewe), and Cameroon (Mandara, Podoko, Uldeme, Bassa) for which the samples sizes were very small and, therefore, were grouped together and analyzed as "Mixed." The Cypriot samples are from individuals whose parents are both Greek Orthodox and originate from both coastal and mountainous districts of the island. The Italian samples all originate from the Ionian coast of the Calabria region of southern Italy (*11*, *21*). The Tunisian samples are from unrelated patients from different parts of Tunisia, and the Lebanese samples originate from unrelated samples collected in Beirut.
8. The G6PD B, A, and A− alleles in Africa were distinguished by PCR amplification of exons 3 and 4 followed by digestion with Nla III restriction enzyme, and by amplification of exon 5 followed by digestion with Fok I. Individuals who lack both restriction sites were classified as B, those who lacked the Nla III site but contained the Fok I site were classified as A, and those who had both Nla III and Fok I restriction sites were classified as A−. Primers used to amplify exons 3 and 4 were G6PD3/4F (5′-AACCACACACCTGTTCCCTC-3′) and G6PD 3/4R (5′-GCTGGTAGAGAGGGCAGAAC-3′). Amplification with these primers produces a 320-bp product, and digestion with Nla III produces 207- and 113-bp fragments in individuals with the A− mutation and a 320-bp fragment in individuals who lack the mutation. Primers used to amplify exon 5 were G6PD5F (5′-CAAAGAGAGGGGCTGACATC-3′) and G6PD5R (5′-GCTCATAGAGTGGTGGGAGC-3′). Amplification with these primers produced a 342-bp product. Digestion with Fok I produces 173- and 169-bp fragments if the individual has the A mutation. The Med mutation can be detected by amplification of exon 6 followed by digestion with Mbo II. Primers used to amplify exon 6 were G6PD6F (5′-TGCAGCTGTGATCCTCACTC-3′) and G6PD6R (5′-AGGTGGAGGAACTGACCTTG-3′). Amplification produces a 388-bp product. Digestion with Mbo II produces 253-, 99-, and 36-bp fragments in individuals with the Med mutation and produces 352- and 36-bp fragments in individuals who lack the mutation. Amplifications were performed with 50 to 100 ng of genomic DNA in a 25-μl (total volume) reaction mixture. The reaction mixture contained 10 pmol of each forward and reverse primer, 200 μM of each deoxynucleotide triphosphate (dNTP), 50 mM KCl, 10 mM tris-HCl, 1.5 mM $MgCl_2$, and 0.625 U of *Taq* polymerase. Exons 3/4 and exon 5 samples were denatured for 1 min at 94°C, followed by 30 cycles of 94°C for 1 min, 58°C for 1 min, and 72°C for 1 min, followed by a 10-min extension at 72°C. Exon 6 samples were denatured for 1 min at 94°C, followed by 30 cycles of 94°C for 1 min, 59°C for 1 min, and 72°C for 1 min, followed by a 10-min extension at 72°C. PCR products were digested for 2 hours overnight with the appropriate restriction enzymes, buffers, and incubation temperatures. Amplification products were run on a 1% agarose gel, stained with ethidium bromide, and visualized with an ultraviolet transilluminator and photoimager. The Pst I and Bsp HI RFLPs were detected with the primers and protocols as described in (*13*) and (*12*), respectively. The Sca I and Bsp I restriction sites were created by mismatch-containing primers to detect the underlying mutations as described (*12*).
9. The Tunisian and non-African samples were selected on the basis of G6PD protein electrophoretic properties and/or enzyme activity levels. For the Tunisian populations, electrophoretic mobility and enzyme activity of G6PD was performed with the standard World Health Organization protocol as described (*52*). Enzyme activity levels for the Lebanese and Cypriot samples were measured by a semiquantitative fluorescence test (*53*). Enzyme activity level in the Italian population was measured by differential pH-metry as described (*54*). For the Italian sample, only those individuals with enzyme activity levels less than 5% of normal were included in the study.
10. C. C. Plato, D. L. Rucknagel, H. Gershowitz, *Am. J. Hum. Genet.* **16**, 267 (1964).
11. A. Tagarelli *et al.*, *Gene Geogr.* **5**, 141 (1991).
12. T. J. Vulliamy *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8568 (1991).
13. E. Beutler, W. Kuhl, *Am. J. Hum. Genet.* **47**, 1008 (1990).
14. GenBank accession number X55448.
15. The G6PD A, A−, and Med mutations were not detected in a sample of 11 great apes. In 16 chromosomes screened from chimps ($n = 13$) and gorillas ($n = 3$), we observed microsatellite allele sizes that

overlap the range observed in humans, with the exception of six 196-bp and three 152-bp AC alleles observed in chimps and gorillas, respectively, and 180- to 192-bp CTT alleles observed in both chimps and gorillas (*55*).

16. The $(AC)_n$ repeat was amplified with primers ACF (5′-TCACTTGGGCCATGATCAC-3′) and ACR (5′-TTAATTTGTATCATGGGGTCCTAG-3′) that produce fragments from 164- to 188-bp long. The $(AT)_n$ repeat was amplified with primers ATF (5′-CATGGTTTCTGTGGAGTCTAGC-3′) and ATR (5′-GGTGGGAGGATTGCTTGAAG-3′) that produce fragments from 125- to 179-bp long. The $(CTT)_n$ repeat was amplified with primers CTTF (5′-GTTCAAGCGATTCTAGTGCCC3′) and CTTR (5′-CGGGTAGATTGCTTGAGCC-3′) that amplify fragments from 195- to 216-bp long. Amplification was performed with 50 ng of genomic DNA in a 25-μl (total volume) reaction mixture. The reaction mixture contained 5 pmol each of fluorescently labeled forward and reverse primer, 200 μM of each dNTP, 50 mM KCl, 10 mM tris-HCl, 1.5 mM $MgCl_2$, and 0.625 U of *Taq* polymerase. Samples were denatured for 1 min at 94°C, followed by 25 cycles of 94°C for 1 min, 56°C for 1 min, and 72°C for 1 min, followed by a 10-min extension at 72°C. Amplification products were run on a 6% polyacrylamide gel on an ABI 373 DNA sequencer, and fragment sizes were determined with Genescan software. Amplification of the CTT repeat, which is located within an *Alu*-rich region, produces some nonspecific fragments (predominantly a 154-bp band) that do not vary among individuals. Only bands between 195 and 216 bp in size are polymorphic.
17. Supplementary Web material is available on *Science* Online at www.sciencemag.org/cgi/content/full/1061573/DC1.
18. Although data from males provide unambiguous linkage phase of the restriction sites and microsatellite alleles, linkage phase in females can be ambiguous. The usual expectation-maximization (EM) algorithms that have been used to infer haplotype phase (*56*) are inadequate for such data, because these algorithms exhaustively enumerate potential haplotypes. With three microsatellites having 10, 26, and 8 alleles segregating in the population, there are 2080 ($10 \times 26 \times 8$) potential haplotypes and the likelihood surface is nearly flat. Instead, haplotypes were inferred in a two-tiered process, first by applying the algorithm of Clark (*57*), followed by EM to obtain maximum likelihood estimates of frequencies and likelihoods of alternative phases for ambiguous individuals. The algorithm of Clark (*57*) identifies haplotypes that are unambiguous either by virtue of homozygosity (or hemizygosity as in the case of X-linked genes in males), or if the two haplotypes in a female differ by only a single site. Because the data consisted of ~80% males (Table 1), we have an extensive starting set of known haplotypes. Remaining individuals with ambiguous linkage phase are then tested to see if any of the confirmed haplotypes might be borne by those individuals. If so, then a possible phase consists of that haplotype and the complement that gives the observed genotype. When the likely phasings are compiled for all individuals, the EM is run to maximize the likelihood for a set of phasings.
19. The allele frequencies of B, A, and A− variants were estimated by gene counting. Heterozygosities for individual sites and for the haplotypes were estimated as $n(1 - \Sigma p_i^2)/(n - 1)$, where p_i represents the frequency of the ith allele or haplotype for any given system, and n is the sample size.
20. Owing to the large allele number of the microsatellites near *G6PD*, classical tests of LD among microsatellites are inappropriate. However, LD between the major *G6PD* alleles (B, A, A−, Med) and each microsatellite locus could be tested with Fisher's exact test. Contingency tables were constructed from the haplotype counts in which columns were the G6PD functional allele and rows were the microsatellite alleles. Fisher's exact tests assessed whether the observed sample deviates from that expected under independent sampling of RFLPs and microsatellites.
21. S. Filosa *et al.*, *Genomics* **17**, 6 (1993).
22. Coalescent simulations were run for each combination of A− versus Med allelic lineages and microsatellite locus (total of six runs). The sample from sub-Saharan Africa was drawn without ascertainment bias, so the observation of 42 A− alleles in a sample of 315 chromosomes represents the population frequency used in the simulations. The other populations were ascertained nonrandomly, so we used an estimate of 0.05 as the frequency of the Med allele in those populations (*10*, *11*), and we drew a sample from the observed Med haplotypes to reconstruct a population sample. In each simulation run, 10,000 random coalescence trees were generated that satisfy the rejection criteria [that the derived clade have a frequency equal to the observed A− (or Med) clade and that the variance in the microsatellite allele size on the remaining clades be within ±20% of the observed variance]. The mutation rate of the microsatellite was sampled from a prior distribution determined from extrinsic data (the variance in the microsatellite allele size). From each coalescence tree that was accepted, we calculated four sample statistics: (i) the number of distinct haplotypes; (ii) the variance in the microsatellite allele size; (iii) the linkage disequilibrium parameter δ for the microsatellite allele with the highest LD {where $\delta = 1 -$ frequency[modal microsatellite allele not on A− (or Med)clade]/frequency[modal microsatellite allele on the A− (or Med) clade]}; and (iv) the *F* statistic from an ANOVA testing equal microsatellite size between the A− (or Med) clade and the rest of the genealogy. This process generated the null distributions of the four sample statistics, conditioned on having a clade with the observed deficiency allele frequency. If a sample statistic obtained from the observed data falls far into the tails of the respective null distribution, we conclude that a neutral model does not fit the data satisfactorily, and that another factor(s) contributes to the variation.
23. Because all the tests ended up being one-tailed (A− and Med chromosomes are depauperate in microsatellite variation), performing the tests assuming zero recombination is conservative. Furthermore, incorporation of population growth would result in a more starlike phylogeny and would also inflate the variance of the A− and Med subclades relative to the rest of the tree. Thus, simulations without population growth provide a conservative test. Finally, human populations have shown some level of subdivision, whereas the A− and Med alleles appear to have spread across fairly broad geographic regions. Incorporation of geographic structure would increase the number of haplotypes and inflate the variance of allele size and would make the observed patterns of high frequency, high LD, low microsatellite variability, and broad geographic distribution even less likely under neutrality.
24. C. Neuhauser, S. M. Krone, *Genetics* **145**, 519 (1997).
25. J. H. Gillespie, *Genetics* **155**, 909 (2000).
26. M. Slatkin, B. Rannala, *Annu. Rev. Genomics Hum. Genet.* **1**, 225 (2000).
27. The rejection method was applied to accept only those simulation runs that bore sufficient resemblance to the observed data. For each of the AC, AT, and CTT microsatellites, we estimated the variance in the A− (or Med) clades, the variance in the non-A− (or non-Med) clades, and the LD between the A− (or Med) allele and the most common microsatellite allele. Each simulation produced values of these statistics. If the difference between any of the simulated statistics and the observed values exceeded 20% of the observed value, that simulation was rejected. This criterion reflects a compromise between overly small values rejecting too large a proportion of the simulations, and too large a value, which leads to spuriously inflated variance in the posterior distributions. Altogether, 37,052 runs that satisfied the rejection criteria were collected for the A− simulations and 38,362 for the Med simulations. For the simulation runs not rejected, many attributes were collected, including the time required for the A− or Med clade to expand to its current frequency.
28. For each iteration of the simulation, values of μ, *s*, and *r* were drawn from a uniform prior distribution with the following ranges: μ (10^{-3} to 10^{-4}), *s* (0 to 0.1), and *r* (0 to 10^{-4}). Given an initial population size of 10,000 and a growth rate of $\gamma = 0.001$ per generation, the allele frequencies were iterated until the population attained a frequency of the A− clade equal to that observed in the African sample ($q_{crit} = 0.11$) or of the Med clade equal to that estimated for the Mediterranean populations ($q_{crit} = 0.05$). The dominance and overdominance selection models produced estimates of mutation rate, recombination rate, and allele age that were all within 10% of each other (because A−/A− females are relatively rare in the early stages of allele expansion). The estimates of mean age of deficiency clades varied only by about 5% across the 10-fold range of growth rates, suggesting that the estimates were robust across a range of population growth rates, as long as the population is growing fast enough to make the branching process supercritical (i.e., growing sufficiently rapidly that the deficiency lineage is not lost). Similarly, mean allele age estimates were relatively insensitive to the initial population size (varying only 2% across a range of sizes from 10,000 to 100,000). Although we expect the mutation rate and recombination rate to be the same in Africa as in the Mediterranean, these two parameters were estimated independently in the two samples. The consistency of estimates of mutation rate and recombination in these two samples serves to test one aspect of validity of the model, namely, whether it converges on biologically consistent estimates.
29. A. Collins, J. Frezal, J. Teague, N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 14771 (1996).
30. To examine the effect of our assumption that the Med allele frequency was 0.05, we repeated the simulations described above assuming a G6PD Med allele frequency (*q*) equal to 0.02 or 0.10, representing the range of observed allele frequencies in the populations studied. With $q = 0.10$ we estimated a mean allele age of 4244 years with a 95% credibility interval of 2280 to 8400 years, and with $q = 0.02$ we estimated a mean allele age of 2183 years with a 95% credibility interval of 840 to 4720 years.
31. B. C. Verrelli, S. A. Tishkoff, unpublished data.
32. S. A. Tishkoff *et al.*, *Science* **271**, 1380 (1996).
33. S. A. Tishkoff *et al.*, *Am. J. Hum. Genet.* **62**, 1389 (1998).
34. S. A. Tishkoff *et al.*, *Am. J. Hum. Genet.* **67**, 901 (2000).
35. J. R. Kidd *et al.*, *Am. J. Hum. Genet.* **66**, 1882 (2000).
36. F. B. Livingstone, *Am. Anthropol.* **60**, 533 (1958).
37. _____, *Annu. Rev. Genet.* **5**, 33 (1971).
38. S. L. Wiesenfeld, *Science* **157**, 1134 (1967).
39. I. W. Sherman, in *Malaria: Parasite Biology, Pathogenesis, and Protection*, I. W. Sherman, Ed. (American Society for Microbiology, Washington, DC, 1998), pp. 3–10.
40. A. Brooks, personal communication.
41. A. S. Brooks, P. T. Robershaw, in *The World at 18,000 BP*, vol. 2, *Low Latitudes*, O. Soffer, C. Gamble, Eds. (Unwin Hyman, London, 1990), pp. 121–169.
42. M. Coluzzi, *Parassitologia* **41**, 277 (1999).
43. J. E. Yellen. *Afr. Archaeol. Rev.* **15**, 173 (1998).
44. S. M. Rich, F. J. Ayala, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6994 (2000).
45. F. Verra, A. L. Hughes, *Mol. Biochem. Parasitol.* **105**, 149 (2000).
46. L. L. Cavalli-Sforza, A. Piazza, P. Menozzi, *History and Geography of Human Genes* (Princeton Univ. Press, Princeton, NJ, 1994).
47. F. Durando, *Ancient Greece; The Dawn of the Western World* (Syewart, Tabori, and Chang, New York, 1997).
48. M. Stoneking, L. B. Jorde, K. Bhatia, A. C. Wilson, *Genetics* **124**, 717 (1990).
49. G. Argyropoulos *et al.*, *J. Clin. Invest.* **102**, 1345 (1998).
50. S. M. Williams *et al.*, *Hypertension* **36**, 2 (2000).
51. G. Spedini *et al.*, *Am. J. Phys. Anthropol.* **110**, 143 (1999).
52. E. Beutler, *Blood* **49**, 467 (1977).
53. E. Touma, R. Kruithof, G. Reclos, in *7th International Congress of Inborn Errors and Metabolism*, Vienna, Austria, 21 to 25 May 1997, p. 26.
54. A. Tagarelli, A. Piro, L. Bastone, G. Tagarelli, *FEBS Lett.* **466**, 139 (2000).
55. S. A. Tishkoff *et al.*, data not shown.

56. M. E. Hawley, K. K. Kidd, *J. Hered.* **86**, 409 (1995).
57. A. G. Clark, *Mol. Biol. Evol.* **7**, 111 (1990).
58. T. Takizawa, Y. Yoneyama, S. Miwa, A. Yoshida, *Genomics* **1**, 228 (1987).
59. A. Hirono, E. Beutler, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 3951 (1988).
60. To examine the possibility of sequence heterogeneity, we sequenced $(AC)_n$ and $(AT)_n$ alleles in a set of geographically diverse individuals. The $(AT)_n$ repeat consists of perfect, uninterrupted repeats in one European, three Africans, one Papua New Guinean, and one chimpanzee sequence (with alleles ranging from 135 to 139 bp). At the AC repeat, the 178-bp allele was sequenced in two Europeans and three Papua New Guineans and was found to be identical to the 178-bp allele published in GenBank. However, two 166-bp alleles sequenced in two African individuals (a South African Bantu-speaker and a Mende individual from Sierra Leone) had the sequence $(TA)_5(AA)_1(TA)_9(CA)_{10}$, which is 12 bp smaller than the 178-bp allele. This shorter allele was most likely the product of a single event, because geographically diverse Africans have identical sequences, suggesting that these A− alleles are identical by descent. Further sequencing analysis will be required to obtain a detailed understanding of the evolutionary history of the three microsatellite alleles. Sequencing conditions were as follows: for PCR products from individuals homozygous for microsatellite alleles, alleles from $(AC)_n$ were amplified with primers ACSEQF (5′-GAGACTGAGGTGGGAGGTTC-3′) and ACSEQR (5′-AAGGAAAAGTTCCTGGGTGG-3′), which produce a 235-bp product. Alleles from $(AT)_n$ were amplified with newly designed primers ATSEQF (5′-TGCATTTTATCACCCCTTC-3′) and ATSEQR (5′-CAGCTAAGGTGGGCATAGTG-3′), which produce a 248-bp product. Amplification was performed with 50 to 100 ng of genomic DNA in a 25-μl (total volume) reaction mixture. The reaction mixture contained 10 pmol of each forward and reverse primer, 200 μM of each dNTP, 50 mM KCl, 10 mM tris-HCl, 1.5 mM $MgCl_2$, and 0.625 U of *Taq* polymerase. Samples were denatured for 1 min at 94°C, followed by 25 cycles of 94°C for 1 min, 60°C for 1 min, and 72°C for 1 min, followed by a 10-min extension at 72°C. The amplified products were purified with a Qiagen PCR purification kit and were cycle sequenced with a Beckman CEQ DTCS sequencing kit. Products were run and analyzed on a Beckman CEQ2000 automated DNA sequencer.
61. M. Saunders, M. Hammer, M. Nachman, in preparation.
62. Funded by a Burroughs Wellcome Fund Career Award, NSF Sloan fellowship, and NSF grants BCS-9905396 (S.A.T.) and DEB 9806655 (A.G.C.). S.M.W. was supported by grants K14-HL03321 (National Heart, Lung, and Blood Institute), G12-RR03032 (National Center for Research Resources), and T37-TW00043 (Fogarty Center, NIH). E.H.T. and J.L. were supported by a grant from CNRS-Lebanon. We thank C. Gallo for technical assistance, A. Deinard for providing chimpanzee and gorilla samples, S. Gevao for assistance with collecting samples from Sierra Leone, and M. Angastiniotis (Cyprus Thalassaemia Centre) for assistance with collecting samples in Cyprus. We also thank A. Brooks, J. Friedlaender, H. Harpending, P. Hedrick, N. Risch, and B. Verrelli for critical review of the manuscript and for helpful discussion. S.A.T. thanks B. Dangerfield, A. Krause, and T. Jenkins for stimulating her interest in the *G6PD* locus and thanks the South African Institute of Medical Research and Trefor Jenkins' laboratory for hosting her as a visiting research scientist in 1997 during which time this work was initiated.

12 April 2001; accepted 7 June 2001
Published online 21 June 2001;
10.1126/science.1061573
Include this information when citing this paper.

Optical Control of Electrons During Electron Transfer

Ignacio B. Martini, Erik R. Barthel, Benjamin J. Schwartz*

The dynamics of electron transfer reactions in solution can be controlled with the use of a sequence of femtosecond laser pulses. In the charge transfer to solvent (CTTS) reaction of sodide (Na^-) in tetrahydrofuran, an initial light pulse launched the CTTS reaction, ejecting an electron into either an immediate or a solvent-separated Na^0:solvated electron contact pair. A second pulse was used to excite the electrons in the contact pairs, and a third pulse monitored the amount of Na^- produced through the back electron transfer. Excitation of the electrons in immediate contact pairs shut off the back electron transfer, whereas excitation of the electrons in solvent-separated pairs both enhanced and hindered the back electron transfer.

Examples of electron transfer (ET) reactions abound in biology, chemistry, and physics (*1*, *2*), but the large number of degrees of freedom in most ET systems makes it difficult to obtain a fundamental understanding of the charge transfer process. To build a complete microscopic understanding of ET reactions, it makes sense to study model systems that consist of atomic reactants. In this report, we show that femtosecond laser pulses can be used to control the motion of the electron as ET takes place in a model charge transfer system that has only electronic degrees of freedom, in this case, the transfer of a single electron from an atomic anion in solution to a nearby solvent cavity.

This reaction is an example of the well-known phenomenon of CTTS (*3*). Vertical excitation of a CTTS transition produces a localized excited state that is bound only by the polarization of the surrounding solvent. The resulting motions of the solvent molecules then cause the electron to be ejected from the excited parent anion, which produces a solvated neutral atom and a solvated electron (*4*). The photoinitiation of such reactions makes them amenable for study with ultrafast lasers because the entire photoexcited ensemble undergoes ET from the same temporal starting point (*5*–*8*).

In previous work, we have extensively characterized the CTTS transition of sodide (Na^-) in tetrahydrofuran (THF) (*5*–*7*). Alkali metal anions (as opposed to the more familiar cations) are formed in solution by the disproportionation of solid alkali metals (M) into the solvated ions M^+ and M^-. This reaction is usually catalyzed by cation complexing agents such as crown ethers (*9*). Our particular choice of Na^- is based on its spectroscopic convenience. The sodide CTTS absorption band is easily accessible in the visible (*10*), and the absorption spectra of the solvated electron (*11*) and sodium atom (*12*, *13*) products in THF are well known and spectrally well isolated (Fig. 1A). Because there are only electronic degrees of freedom, this system allows for detailed investigations of how the solvent affects the electronic energy of each species (*5*–*7*, *14*, *15*). Moreover, the lack of spectral congestion from vibrations and rotations allows for facile optical control over the dynamics of the back ET reaction.

Our previous femtosecond pump-probe experiments on the Na^-/THF system showed that excitation of the CTTS band produces a solvated electron (e_s^-) and a solvated sodium atom (Na^0) in 700 fs (*5*). The back ET reaction (recombination) to reform the parent sodium anion is not diffusion controlled (*6*) but can be explained by assuming that most of the ejected electrons remain in the immediate vicinity of their Na^0 partners in contact pairs. In some contact pairs, the electron resides in the same solvent cavity as the Na^0, which allows for direct nonadiabatic recombination within 1.5 ps; we refer to these species as "immediate" contact pairs. In other contact pairs, the electron and Na^0 products localize in adjacent solvent cavities and do not recombine for hundreds of picoseconds, depending on the solvent; we refer to these as "solvent-separated" contact pairs. The energy of the excitation pulse controls the branching ratio for formation of immediate and solvent-separated contact pairs (*6*). Our basic understanding of the kinetics of the CTTS process of Na^- is summarized in Fig. 1B (*7*).

Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, CA 90095–1569, USA.

*To whom correspondence should be addressed. E-mail: schwartz@chem.ucla.edu