Title
Genetic polymorphism and natural selection in the malaria parasite Plasmodium falciparum.

Permalink
https://escholarship.org/uc/item/1tn1p501

Journal
Genetics, 149(1)

ISSN
0016-6731

Authors
Escalante, Ananias A
Lal, Altaf A
Ayala, Francisco J

Publication Date
1998-05-01

DOI
10.1093/genetics/149.1.189

Genetic Polymorphism and Natural Selection in the Malaria Parasite *Plasmodium falciparum*

Ananias A. Escalante,* Altaf A. Lal* and Francisco J. Ayala[†]

**Division of Parasitic Diseases, National Center for Infectious Diseases, Centers for Disease Control and Prevention, US Public Health Service, Chamblee, Georgia 30341 and* [†]*Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525*

Manuscript received November 13, 1997
Accepted for publication January 22, 1998

ABSTRACT

We have studied the genetic polymorphism at 10 *Plasmodium falciparum* loci that are considered potential targets for specific antimalarial vaccines. The polymorphism is unevenly distributed among the loci; loci encoding proteins expressed on the surface of the sporozoite or the merozoite (AMA-1, CSP, LSA-1, MSP-1, MSP-2, and MSP-3) are more polymorphic than those expressed during the sexual stages or inside the parasite (EBA-175, Pfs25, PF48/45, and RAP-1). Comparison of synonymous and nonsynonymous substitutions indicates that natural selection may account for the polymorphism observed at seven of the 10 loci studied. This inference depends on the assumption that synonymous substitutions are neutral, which we test by analyzing codon bias and G+C content in a set of 92 gene loci. We find evidence for an overall trend towards increasing A+T richness, but no evidence for mutation bias. Although the neutrality of synonymous substitutions is not definitely established, this trend towards an A+T rich genome cannot explain the accumulation of substitutions at least in the case of four genes (AMA-1, CSP, LSA-1, and PF48/45) because the G↔C transversions are more frequent than expected. Moreover, the Tajima test manifests positive natural selection for the MSP-1 and, less strongly, MSP-3 polymorphisms; the McDonald-Kreitman test manifests natural selection at LSA-1 and PF48/45. We conclude that there is definite evidence for positive natural selection in the genes encoding AMA-1, CSP, LSA-1, MSP-1, and Pfs48/45. For four other loci, EBA-175, MSP-2, MSP-3, and RAP-1, the evidence is limited. No evidence for natural selection is found for Pfs25.

ELUCIDATING the processes that maintain genetic polymorphism is an issue of considerable interest and contention in population genetics (*e.g.*, Kimura 1983; Gillespie 1991; Ohta 1992, 1996). In the case of *Plasmodium falciparum*, the agent of malignant malaria, it is also a matter of great clinical and epidemiological importance. Every year, between 300 and 500 million people in the world are infected with malaria; at least one million children under the age of five die each year in sub-Saharan Africa, and more than two billion people are at risk throughout the world (World Health Organization 1995). The design of antimalarial vaccines and the use of antimalarial drugs are hampered by extensive polymorphism in Plasmodium's proteins, particularly those expressed on the parasite's surface, which are obvious targets for the development of highly specific vaccines (McCutchan *et al.* 1988; Anders and Saul 1994; Kaslow 1994; Conway 1997). Highly polymorphic regions have been observed in the genes encoding surface antigenic proteins such as the circumsporozoite protein (CSP), merozoite surface proteins 1 and 2 (MSP-1 and MSP-2), and the S-antigen (Anders and Saul 1994).

Given that these genes encode antigenic proteins that are recognized by the host's immune system, the observed high levels of heterozygosity and rates of evolution have been attributed to natural selection, an outcome of the accumulation and frequent switch of suitable mutations, by means of which the parasite escapes the host's immune defenses (Hughes 1991, 1992; Anders and Saul 1994; Hughes and Hughes 1995). This interpretation is buttressed by the widespread observation that nonsynonymous nucleotide substitutions are more common than synonymous substitutions (Lockyer *et al.* 1989; Thomas *et al.* 1990; Shi *et al.* 1992a). Synonymous substitutions are likely to be neutral, or nearly so, whereas nonsynonymous substitutions may be functionally constrained and, thus, subject to natural selection (Kimura 1977; Ohta 1996).

The matter is, however, far from settled. It is possible in certain cases to account for an excess of nonsynonymous over synonymous substitutions while assuming that the substitutions are neutral (Sawyer and Hartl 1992; Maynard-Smith 1994). A ratio favoring nonsynonymous substitutions could occur, for example, if the compared sequences are so different from one another that a steady state between forward and backward mutations has been reached, as shown in Neisseria (Maynard-Smith 1994). The genome of *Plasmodium falciparum* is A+T rich and exhibits a strong codon bias

Corresponding author: Francisco J. Ayala, Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, CA 92697-2525. E-mail: fjayala@uci.edu

(Hyde and Sims 1987; Musto *et al.* 1995), circumstances that constrain synonymous substitutions, which will thus accumulate slowly, affecting the observed ratio of synonymous to nonsynonymous substitutions (Sharp and Li 1987; Gillespie 1994). Biases in nucleotide frequencies and in the transition/transversion ratio may affect the estimates of synonymous and nonsynonymous substitutions, as noticed for mitochondrial genes (Ina 1995).

In the present study, we analyze 10 genes that are expressed at different stages of *P. falciparum*'s complex life cycle. These genes encode antigens that are considered candidates for antimalarial vaccines. Our study includes genes that have not been investigated previously, such as those coding for EBA-175, MSP-3, Pfs25, and RAP-1, and includes new sequences for four other genes. We first estimate their polymorphism and investigate whether the synonymous–nonsynonymous substitution rates are consistent with neutrality. We consider, in particular, the effects of A+T content, biased codon use, and the transition/transversion ratio. We then apply the Tajima (1989) and McDonald-Kreitman (MK) tests (McDonald and Kreitman 1991) as additional, reliable methods for ascertaining natural selection. To apply the MK test, we compare *P. falciparum* with *P. reichenowi*, its most closely related species (Coatney *et al.* 1971; Collins and Aikawa 1993; Escalante and Ayala 1994; Escalante *et al.* 1995), which is parasitic to chimpanzees.

MATERIALS AND METHODS

Life cycle of *P. falciparum*: *P. falciparum* belongs to the phylum Apicomplexa, which consists of parasitic taxa characterized by the presence, in at least one stage of their life cycle, of a structure called the "apical complex" that is involved in the penetration of the host cell (Cheng 1986; Collins and Aikawa 1993).

The invasive stage to the vertebrate host consists of haploid sporozoites, which are injected by the mosquito vector during its blood meal. These sporozoites are carried by the blood to the liver, where they multiply within the hepatocyte and develop into liver merozoites, which start the erythrocyte life stage. Some merozoites differentiate into gametocytes, which are the forms taken up with the mosquito's blood meal. Fusion of gametes occurs in the mosquito, where the zygote is formed, develops into the ookinete, and further differentiates into the oocyst. This is the only part of the life cycle where the parasite is diploid. Meiosis takes place in the oocyst, resulting in the formation of haploid sporozoites (Collins and Aikawa 1993).

Genes and DNA sequences: We analyze 10 loci in *P. falciparum* that encode proteins expressed at different stages of the parasite's life cycle.

Apical membrane antigen-1 (AMA-1): The AMA-1 (also known as PF83) protein has 622 residues and molecular weight of 83 kD. AMA-1 appears first in the apical complex and migrates to the merozoite's surface. The nine sequences used in our study are from Peterson *et al.* (1989), Thomas *et al.* (1990), and Oliveira *et al.* (1996).

Circumsporozoite protein (CSP): The CSP has ~420 residues and molecular weight of 58 kD; it has a variable central region consisting of multiple repeats of four-residue-long motifs (McCutchan *et al.* 1988). It is the predominant protein on the surface of the sporozoite, the invasive stage transmitted by the mosquito vector. We use 22 sequences from Dame *et al.* (1984), del Portillo *et al.* (1987), Lockyer and Schwarz (1987), Campbell (1989), Caspers *et al.* (1989), and Jongwutiwes *et al.* (1994).

Erythrocyte-binding antigen of 175 kD (EBA-175): This is a merozoite protein involved in the initial erythrocyte binding by the merozoite (Sim 1995). We use 14 sequences from Ware *et al.* (1993) and Liang and Sim (1997).

Liver stage antigen 1 (LSA-1): This 200-kD protein is detected only during the liver stage and is accumulated in the parasitophorous vacuole (Zhu and Hollingdale 1991). It has a large central repeat region plus short, nonrepetitive N and C terminals. We use 14 sequences of the N-terminal portion from Yang *et al.* (1995).

Merozoite surface protein-1 (MSP-1): This protein has variable size determined by a central repeat region; the molecular weight is ~200 kD. The MSP-1 is proteolytically cleaved, and the C-terminal region remains on the merozoite after erythrocyte invasion. We analyze only a 42-kD fragment encoding the C-terminal region. We use 40 sequences from Chang *et al.* (1988), Peterson *et al.* (1988), Weber *et al.* (1988), Jongwutiwes *et al.* (1993), Pan *et al.* (1995), Tolle *et al.* (1995), plus unpublished sequences reported by Y. P. Shi, M. P. Alpers, M. M. Pova, B. L. Nahlen, A. G. Oloo and A. A. Lal under GenBank accession numbers U20726–U20733 and U20653–U20656. The alignment was made following the description in Miller *et al.* (1993).

Merozoite surface protein-2 (MSP-2): This, like MSP-1, is a membrane protein located on the merozoite, and has a molecular weight of 45–54 kD. It consists of N- and C-terminal regions and a central variable segment made up of repetitive and nonrepetitive motifs (Thomas *et al.* 1990; Marshall *et al.* 1992). We used 30 complete sequences from Smythe *et al.* (1990), Thomas *et al.* (1990), Fenton *et al.* (1991), Marshall *et al.* (1991), and Bhattacharya *et al.* (1995).

Merozoite surface protein-3 (MSP-3): This protein, also known as the secreted polymorphic antigen associated with the merozoite (SPAM), has a molecular weight of ~43 kD (McColl *et al.* 1994). It consists of N- and C-terminal regions and a central variable segment of repetitive motifs (McColl *et al.* 1994). The MSP-3 is secreted by the parasite into the parasitophorous vacuole space of the erythrocyte cytoplasm (McColl *et al.* 1994). We have included 19 sequences: two from McColl *et al.* (1994) and 17 partial sequences from Huber *et al.* (1997).

Ookinete protein (Pfs25): This is a 25-kD surface protein expressed in maturing gametocytes and in the zygote. Our 13 sequences are from Kaslow *et al.* (1989) and Shi *et al.* (1992b).

Pfs48/45: These two proteins of 48 and 45 kD are detected from day 2 of gametocytogenesis through gametogenesis and fertilization (Kaslow 1994). They are encoded by a single gene (with likely post-transcriptional modifications; see Kaslow 1994), and they are candidates for a transmission-blocking vaccine (Kaslow 1994). We use eight sequences from Kocken *et al.* (1993, 1995).

Rhoptry antigen protein-1 (RAP-1): This protein of 83 kD is present in the rhoptries, which are organelles located in the apical complex. We use two complete sequences from Ridley *et al.* (1990) and one from Y. P. Shi with GenBank accession number U20985.

For interspecific comparisons, we use five sequences (only one available at each locus) from *P. reichenowi*: CSP (Lal and Goldman 1991), the N-terminal fragment of LSA-1 (Y. Chang, personal communication), Pfs25 (Lal *et al.* 1990), Pfs48/45 (R. L. B. Milek, C. H. M. Kocken, H. Meijers, J. G. G. Schoenmakers and R. N. H. Konings, GenBank accession number L33882), and RAP-1 (Y. P. Shi, GenBank accession number U20986).

Statistical analysis: We use four measures of genetic polymorphism (see Nei 1987 and Tamura 1992). The parameter π estimates the average number of substitutions per site between any two sequences, assuming that the sample is random. This average is also estimated by d, which is based on Tamura's (1992) three-parameter model and corrects for bias in G+C content and transition/transversion ratio. The parameter θ is related to heterozygosity per site, or the effective number of alleles ($n_e = 1 + \theta$); under neutrality equilibrium assumptions, $\theta = 4N\mu$, where N is the effective population size and μ is the rate of neutral mutations. S is simply the number of sites segregating in the sample and is dependent on sample size and length of the sequence. We provide this parameter as a measure of the polymorphism observed, but we do not use it for comparisons between loci.

We test for intragenic recombination (which has been suggested to occur in CSP, MSP-1, and MSP-2; see Conway *et al.* 1991, Marshall *et al.* 1991, Hughes 1992, and McCutchan *et al.* 1992; Rich *et al.* 1997) with Sawyer's permutation test sum of squares of the condensed fragment lengths (SSCF; Sawyer 1989; Hartl and Sawyer 1991). We calculate the SSCF score using all polymorphic sites rather than only the synonymous ones, as would be appropriate, owing to the scarcity of synonymous substitutions. This test could, therefore, be affected by convergence among the nonsynonymous sites as a result of natural selection. Where possible, however, we also calculate the SSCF score separately for synonymous and nonsynonymous substitutions. The significance of the SSCF score is obtained by means of 10,000 random computer permutations (Hartl and Sawyer 1991).

We first test for evidence of positive natural selection by comparing the number of synonymous and nonsynonymous substitutions. Without positive selection favoring amino acid polymorphism, the incidence of synonymous substitutions should be higher owing to purifying selection against nonsynonymous substitutions; a higher incidence of nonsynonymous than of synonymous substitutions is taken as evidence that positive natural selection is promoting polymorphism (Kimura 1977; Kreitman and Akashi 1995; Ohta 1996). The numbers of synonymous and nonsynonymous substitutions per site are estimated using two methods: Nei and Gojobori's (1986) with the Jukes and Cantor (1969) correction, as implemented in the MEGA program (Kumar *et al.* 1994), and the method of Li (1993) based on Kimura's (1980) two-parameter model. Intra- and interspecific transitions and transversions are estimated using the pairwise differences without any correction, as implemented in the MEGA program (Kumar *et al.* 1994); 95% confidence intervals for these statistics are estimated with 5000 bootstrap replications (Efron and Tibshirani 1993).

The previous test assumes that synonymous substitutions are neutral, as is generally taken to be the case. We test this assumption by ascertaining the consequences of codon bias. The effective number of codons, *Nc*, is defined as the number of codons that would yield the observed level of codon usage if all codons were equally frequent (Wright 1990). *Nc* is a measure of the codon bias, and its value can range between 20 and 61. A value of $Nc = 20$ indicates that only one codon per amino acid is used; a value of 61 indicates that all synonymous codons are equally used. *Nc* and G+C content are calculated for a data set of 92 loci using the program CODONS (Lloyd and Sharp 1992); deviation from 50% G+C content in silent sites affects *Nc* values so that a correlation is expected if there is codon bias (Wright 1990). The 92 sequences are obtained from the GenBank database (http://www.ncbi.nlm.nih.gov/Web/Genbank/). Only one allele is included for each locus, which was chosen at random when more than one sequence is available. Only genes and gene fragments with >100 codons are included in this set. The significance of the correlation of *Nc* with (1) total G+C content and (2) G+C content in the third position is ascertained by means of Pearson's correlation test, with $\alpha = 0.05$ and Bonferroni's correction for multiple tests, as appropriate (Dunn 1961; Hancock and Klockars 1996).

We used two additional tests for detecting positive natural selection in maintaining genetic polymorphism. The Tajima (1989) test is based on the statistic D:

$$D = (k - S/a_1)/\sqrt{e_1 S + e_2 S(S - 1)},$$

where S is the total number of segregating sites, $a_1 = \Sigma\,(1/i)$ from $i = 1$ to $n - 1$ (n is the number of nucleotide sequences), k is the average number of nucleotide differences between pairs of sequences, e_1 and e_2 are constants fixed so that the mean and variance of D are $\sim$ 0 and 1, respectively. Tajima's test is based on the neutral model prediction that estimates of S/a_1 and k are unbiased estimates of θ. We use the critical values for $\alpha = 0.05$ reported by Simonsen *et al.* (1995). They are generated by simulating the D values under neutrality, based in the coalescent model described by Hudson (1993).

The intra- and interspecific numbers of synonymous and nonsynonymous sites (McDonald and Kreitman 1991) are compared using Fisher's exact test for a 2×2 contingency table (Conover 1980).

RESULTS

Table 1 gives estimates of genetic variation at each of 10 gene loci of *P. falciparum.* Genetic diversity is greater in the five genes expressed on the surface of either the merozoite ($\pi = 0.016$, 0.088, 0.044, and 0.097, respectively, for AMA-1, MSP-1, MSP-2, and MSP-3) or the sporozoite ($\pi = 0.006$ for CSP) than in the four other genes ($\pi = 0.004$, 0.004, 0.002, and 0.002, respectively, for EBA-175, Pfs25, Pfs48/45, and RAP-1). The N-terminal region of LSA-1 is fairly polymorphic ($\pi = 0.009$), even though the protein has been said to be conserved (Fidock *et al.* 1994). LSA-1 has not been detected on the surface of the hepatocyte, although no experiments specifically designed for the purpose have been performed. The presence of cytotoxic T lymphocytes or cytotoxic T lymphocyte epitopes suggests immune protection (Hill *et al.* 1992), which has stimulated the working hypothesis that LSA-1 is transported outside the parasitophorous vacuole to the hepatocyte surface, where it would interact with major histocompatibility complex class I molecules (Fidock *et al.* 1994). The intermediate level of genetic diversity we find in LSA-1 (N-terminal region) is consistent with this hypothesis.

MSP-3 is very polymorphic ($\pi = 0.097$), comparable with MSP-1 (the 42-kD region herein analyzed, see materials and methods) and other surface proteins. Some authors suppose that MSP-3 is located on the merozoite surface (Oeuvray *et al.* 1994); however, MSP-3 is part of a group of proteins that are not an integral part of the merozoite membrane but are secreted into the parasite parasitophorous vacuole space or the erythrocyte cytoplasm (McColl *et al.* 1994). These proteins are often referred to as "surface pro-

TABLE 1

Polymorphism in 10 *P. falciparum* genes

Gene	*n*	Sites	π	*d*	θ	S	*P*
AMA-1	9	1866	0.01635	0.01655	0.0150	76	0.0044
CSP	22	780	0.00645	0.00647	0.0084	24	0.0013
EBA-175	14	1842	0.00366	0.00369	0.0034	20	NS
LSA-1	14	480	0.00872	0.00880	0.0105	16	0.0441
MSP-1	40	1056	0.08792	0.10581	0.0550	247	0.0000
MSP1-MAD[a]	30	1116	0.00413	—	0.0043	19	—
MSP1-Well[a]	10	1065	0.00100	—	0.0010	3	—
MSP-2	30	417	0.04409	0.04660	0.0268	44	0.0000
MSP-3	19	575	0.09694	0.11038	0.0633	104	NS
Pfs25	13	654	0.00357	0.03590	0.0059	12	NS
Pfs48/45	8	1300	0.00181	0.00174	0.0018	6	NS
RAP-1	3	2349	0.00199	0.00600	0.0020	7	NS

Genes are identified by an abbreviation of the encoded protein; *n*, number of sequences sampled; sites, number of sites excluding alignment gaps; π, average number of nucleotide substitutions; *d*, average number of nucleotide substitutions under Tamura's three-parameter model; θ, $n_e - 1$ (n_e is the effective number of alleles); S, the number of segregating sites; *P* comes from the randomization test of SSCF; NS, not significant.

[a] MAD and Well refer to MAD-20 and Wellcome allele types (Tanabe *et al.* 1987).

teins” but their relationship with the merozoite membrane is unknown.

EBA-175 and RAP-1 are not present on the surface of the merozoite, but they are involved in the erythrocyte invasion that occurs as an essential stage of the parasite's life cycle. Pfs25 is a surface protein expressed in the mature gametocyte close to exflagellation and in the zygote, *i.e.*, inside the mosquito host, and thus is not exposed to the immune system of the human host, where no antibodies against it have been found (Kaslow 1994). Similarly, Pfs48/45 is a surface protein expressed during gametogenesis. Naturally occurring antibodies against this protein have been found in 9–60% individuals in exposed populations (Kaslow 1994). Epidemiologic studies suggest that individuals, after their first malarial infection, produce antibodies that gradually disappear in endemic areas with long-term exposure (Kaslow 1994).

Among the surface-expressed, more polymorphic loci, MSP-1 is distinctive in that it consists of two allelic families (MAD-20 and Wellcome, identified as MSP-1-MAD and MSP-1-Well in Table 1), with low divergence among the members of a family but great differentiation between the families ($\pi = 0.089$ for all 40 alleles, but only 0.004 for the 30 MAD alleles, and 0.001 for the 10 Well alleles).

If we use all polymorphic sites, synonymous as well as nonsynonymous, intragenic recombination is detected by the SSCF test in almost all fairly polymorphic genes: the three merozoite surface antigens (AMA-1, MSP-1, and MSP-2), the sporozoite surface antigen (CSP), and LSA-1; the only exception is MSP-3. The SSCF test, however, is not significant when synonymous substitutions alone are considered, which is only possible in the four genes that exhibit nonsynonymous polymorphism within the regions included in our alignment (six sites in AMA-1, three sites in CSP, four sites in MSP-2, and five sites in Pfs25). When only amino acid replacement sites are taken into account, there is evidence of intragenic recombination at AMA-1 (SSCF = 5495; $P = 0.001$), CSP (SSCF = 27,669; $P = 0.045$), and MSP-2 (SSCF = 18,503; $P = 0.029$). We consider in this paper only the C-terminal region of MSP-1, but we have made an additional analysis of six complete sequences (using the alignment of Miller *et al.* 1993). Evidence of intragenic recombination exists for synonymous (SSCF = 134,495; $P = 0.000$) and replacement sites (SSCF = 759,275; $P = 0.000$).

Table 2 gives the incidence of synonymous and nonsynonymous substitutions, estimated by two methods. As previously reported (Hughes 1991, 1992; Hughes and Hughes 1995), the incidence of nonsynonymous substitutions is greater (significantly so in most cases) for several loci. The exceptional cases are MSP-1, MSP-3, and Pfs25. The number of synonymous substitutions is zero in three cases (LSA-1, Pfs48/45, and RAP-1). The pattern of synonymous and nonsynonymous substitutions in MSP-1 (the 42-kD region herein studied) is consistent with previous observations. Hughes (1992) found in regions 8–11 (using the classification of Tanabe *et al.* 1987) a greater number of synonymous than nonsynonymous substitutions. The excess of nonsynonymous substitutions (comparing K_s and K_n) indicates positive natural selection in seven of the 10 loci: AMA-1, CSP, EBA-175, LSA-1, MSP-2, Pfs48/45, and RAP-1. The method of Li (1993) leads to similar results for the same loci except for MSP-2; at this locus, the number of nonsynonymous substitutions is higher than the number of synonymous substitutions, but not significantly.

TABLE 2

Synonymous and nonsynonymous substitutions in 10 *P. falciparum* genes

Gene	LK_s	LK_n	K_s	K_n
AMA-1	0.0052 [0.0042–0.0062]	0.0198 [0.0184–0.0211]	0.0054 [0.0045–0.0064]	0.0190 [0.0181–0.0207]
CSP	0.0065 [0.0051–0.0079]	0.0155 [0.1044–0.0126]	0.0034 [0.0027–0.0042]	0.0110 [0.0099–0.0121]
EBA-175	0.0002 [0.0001–0.0004]	0.0047 [0.0042–0.0051]	0.0004 [0.0002–0.0006]	0.0044 [0.0040–0.0049]
LSA-1	0.00	0.0105 [0.0094–0.0115]	0.00	0.0108 [0.0097–0.0120]
MSP-1	0.2302 [0.2095–0.2509]	0.0925 [0.0848–0.1004]	0.1509 [0.1380–0.1644]	0.0933 [0.0859–0.1011]
MSP-2	0.0451 [0.0412–0.0492]	0.0492 [0.0453–0.0530]	0.0336 [0.0308–0.0363]	0.0504 [0.0463–0.0544]
MSP-3	0.1087 [0.0949–0.1227]	0.1108 [0.0973–0.1238]	0.1064 [0.0931–0.1195]	0.1104 [0.0976–0.1236]
Pfs25	0.0040 [0.0034–0.0046]	0.0033 [0.0029–0.0037]	0.0068 [0.0057–0.0080]	0.0028 [0.0024–0.0032]
PF48/45	0.00	0.0027 [0.0018–0.0027]	0.00	0.0022 [0.0018–0.0026]
RAP-1	0.00	0.0026	0.00	0.0025

LK_s and LK_n are synonymous and nonsynonymous substitutions, respectively, using the method of Li (1993). K_s and K_n use the method of Nei and Gojobori (1986) with the Jukes and Cantor (1969) correction. Bootstrap 95% confidence intervals are indicated within brackets (except for RAP-1 because only three sequences are available).

Table 3 gives the incidence of transitions and transversions for the 10 *P. falciparum* genes. All the genes, except Pfs25, exhibit a higher number (often much higher) of transversions. Usually, transitional substitutions are about twice as common as transversions (Gojobori *et al.* 1982; Li *et al.* 1984; Collins and Jukes 1994) and almost 10 times more common in animal mitochondria (Brown *et al.* 1982; Kondo *et al.* 1993). In *P. falciparum*, transversions are more common than transitions in six genes (AMA-1, LSA-1, MSP-1, MSP-2, MSP-3, and Pfs8/45) and about equal in three genes (CSP, EBA-175, and RAP-1). The observed pattern is comparable with those found in the mitochondrial genome of Drosophila (Wolstenholme and Clary 1985; Gleason *et al.* 1997; Inohira *et al.* 1997) and *Apis mellifera* (Crozier and Crozier 1993), which are also A+T-rich genomes.

Table 4 gives the total G+C content and the *Nc* values for 92 loci. The average G+C for the *P. falciparum* genes is, for all three codon positions, 30.22% [with a 95% confidence interval (C.I.) around the mean of 29.2–31.2%, using a *t* distribution range 22–51%]. However, for the third position, the G+C content is only 15.16% (C.I. 14.2–16.1%, range 7–31%). The average *Nc* is 36.82 (C.I. 36.0–37.6, range 31.33–51.44), which is comparable to other A+T-rich genomes, such as the proteobacteria *Rickettsia prowazekii*, with an average *Nc* of 40.84 and a range from 33.4 to 51.1 (Andersson and Sharp 1996). Figure 1 shows that *Nc* is highly correlated with G+C content at the third codon position (Pearson's coefficient: $r = 0.737$, $P = 0.0001$) but not with G+C content at the first ($r = -0.134$, $P = 0.20$) or at the second ($r = 0.07$, $P = 0.51$) codon position. There is no significant correlation, either, between *Nc* and total G+C ($r = 0.178$, $P = 0.09$). The significant correlation between *Nc* and G+C content at the third codon position persists even if we use a level of significance of $\alpha/4 = 0.0125$ [on the grounds that four separate tests are performed; Bonferroni correction (see Dunn 1961; Hancock and Klockars 1996)]. Codon use is, thus, constrained by the reduction in G+C content in the third positions, where most of the synonymous substitutions occur.

TABLE 3

Transitions and transversions in 10 *P. falciparum* genes

Gene	Transitions	Transversions
AMA-1	13.39 [12.42–14.36]	17.11 [15.75–18.42]
CSP	2.41 [2.17–2.67]	2.62 [2.40–2.85]
EBA-175	3.29 [2.96–3.60]	3.46 [3.01–3.90]
LSA-1	0.92 [0.78–1.05]	3.26 [2.87–3.68]
MSP-1	33.09 [30.35–35.99]	57.94 [53.00–62.86]
MSP-2	5.09 [4.73–5.44]	13.16 [12.02–14.28]
MSP-3	17.30 [15.20–19.37]	28.26 [24.92–31.72]
Pfs25	2.18 [2.00–2.37]	0.15 [0.08–0.24]
Pfs48/45	0.68 [0.46–0.89]	1.68 [1.36–2.00]
RAP-1	2.00	2.66

Bootstrap 95% confidence intervals are in brackets (except for RAP-1 because only three sequences are available).

Table 5 reports the results of two additional tests that seek to ascertain whether positive natural selection contributes to the genetic polymorphism in *P. falciparum*. Tajima's test shows positive values and a significant departure from neutral expectations for MSP-1, where the synonymous and nonsynonymous substitution ratio failed to detect selection. MSP-3 is very close to significance; we repeated the test for combinations of all sequences minus one, and 90% of the tests were statistically significant. Positive and significant values of *D* indicate strong overdominance selection (Gillespie 1994).

The MK test (McDonald and Kreitman 1991) ascertains whether the intraspecific number of sites with nonsynonymous substitutions is significantly greater than the neutral expectation, which is determined by the interspecific nonsynonymous/synonymous ratio. The interspecific comparisons are made with *P. reichenowi*, which

TABLE 4

The 92 *P. falciparum* genes included in the analysis of G+C content and codon use

Accession number	Gene name	Sites	*Nc*	GC3	GC2	GC1	GC
A00661[a]	Ring-infected erythrocyte surface antigen 1 (RESA-1)	1073	36.15	0.14	0.25	0.49	0.30
A00663[a]	Interspersed repeat antigen (FIRA)	976	34.23	0.12	0.42	0.53	0.36
A16141[a]	Serine-rich protein (SERP)	733	34.13	0.13	0.28	0.38	0.27
D86573	Flavoprotein subunit of succinate dehydrogenase	620	33.20	0.09	0.39	0.45	0.32
D86574	Iron-sulfur subunit of succinate dehydrogenase	321	33.47	0.12	0.31	0.33	0.27
J03998[a]	Glutamic acid-rich protein	678	35.17	0.14	0.17	0.54	0.29
J04000[a]	Serine-repeat Antigen (SERA)	989	33.73	0.12	0.34	0.38	0.29
L01655[a]	Triosephosphate isomerase	222	34.32	0.12	0.31	0.44	0.29
L02513	Pfmdr1	1208	44.47	0.30	0.32	0.15	0.26
L02822[a]	Heat shock protein	655	32.72	0.13	0.31	0.48	0.31
L08135	Transmission blocking target antigen (Ps230)	3135	34.31	0.11	0.25	0.39	0.25
L11172[a]	RNA polymerase I	2743	35.17	0.11	0.24	0.33	0.23
L18785[a]	DNA polymerase alpha	1855	36.43	0.14	0.21	0.30	0.22
L22057[a]	Ribonucleotide reductase R1 subunit	804	34.46	0.12	0.35	0.40	0.30
L28825	MSA-3	389	37.91	0.10	0.25	0.43	0.27
L32150	Carbamoyl phosphate synthetase II	2391	34.22	0.11	0.26	0.34	0.24
L34028[a]	Heat shock protein 86	750	39.25	0.21	0.31	0.36	0.30
L46348[a]	d-aminolevulinic acid synthetase	630	38.23	0.15	0.24	0.30	0.24
M18824	S-antigen	640	32.22	0.13	0.45	0.57	0.38
M19146	Actin I	376	31.97	0.12	0.40	0.47	0.34
M19881[a]	Knop protein (KP)	634	35.20	0.18	0.40	0.52	0.37
M22159	Dihydrofolate reductase-thymidylate synthetase	608	36.20	0.13	0.25	0.35	0.25
M22718[a]	Actin II	376	39.54	0.18	0.38	0.44	0.35
M25769	parasitophorous vacuole antigen	427	33.68	0.12	0.29	0.37	0.27
M27133	AMA-1	622	33.67	0.12	0.33	0.43	0.30
M28398[a]	β-Tubulin	445	37.48	0.18	0.40	0.49	0.37
M28881[a]	Aldolase	369	32.51	0.17	0.41	0.47	0.36
M32153	β-Galactosidase fusion protein	201	32.70	0.07	0.12	0.86	0.35
M34390[a]	α-Tubulin II	450	43.24	0.22	0.40	0.50	0.38
M59249	3-Phosphoglycerate kinase	416	32.09	0.12	0.31	0.46	0.30
M59706	Glutamate-rich protein (Glurp)	1271	34.08	0.08	0.21	0.58	0.29
M64705	16-kD sexual stage (Pfs16)	159	51.44	0.24	0.37	0.45	0.36
M69147	Sexual stage DNA sequence	1048	37.22	0.15	0.18	0.47	0.27
M69164	Erythrocyte membrane-associated giant protein Ag332	458	31.33	0.09	0.22	0.66	0.33
M69183[a]	MESA antigen	1510	33.12	0.13	0.20	0.48	0.28
M73770[a]	RNA polymerase III	2298	36.08	0.14	0.23	0.32	0.24
M73810	MSA-2	274	38.83	0.13	0.51	0.43	0.36
M77834	Membrane-associated calcium-binding protein	343	36.52	0.11	0.22	0.45	0.27
M80655	Glucose-6-phosphate deyhdrogenase	736	35.71	0.10	0.25	0.31	0.23
M81341	Cysteine proteinase	569	33.92	0.13	0.22	0.34	0.24
M83163	CSP	437	40.55	0.20	0.38	0.45	0.35
M91672[a]	RESA-2	461	40.88	0.18	0.26	0.36	0.28
M92054	Hexokinase	493	33.33	0.12	0.39	0.41	0.32
M93720	l-lactate dehydrogenase	316	34.88	0.17	0.33	0.47	0.33
M94013	Sporozoite surface protein 2	574	35.91	0.12	0.35	0.52	0.34
M99442[a]	Calmodulin	149	42.43	0.18	0.30	0.49	0.34
U00152	Enolase	446	33.33	0.16	0.36	0.45	0.33
U14189	MCP1	361	43.60	0.21	0.25	0.43	0.30
U16955	ATPase 2	1501	34.97	0.11	0.28	0.35	0.26
U16995	ATPase	1103	33.72	0.11	0.34	0.41	0.29
U18984	Putative phosphatidylethanolamine-binding protein	190	37.56	0.12	0.32	0.41	0.29
U20985	RAP-1	782	34.38	0.12	0.31	0.37	0.27
U25814	Chromodomain protein	266	39.32	0.16	0.29	0.35	0.27
U27338[a]	Erythrocyte membrane protein 1	2924	46.22	0.22	0.36	0.44	0.35
U34363	CTRP	2098	37.32	0.16	0.36	0.39	0.31
U37225	Elongation factor 3-related protein (pfgcn20)	816	38.71	0.14	0.22	0.38	0l.25
U38963	60-kD heat-shock protein PfHsp60	577	33.22	0.12	0.35	0.45	0.32

(*continued*)

TABLE 4

(*continued*)

Accession number	Gene name	Sites	*Nc*	GC3	GC2	GC1	GC
U39298	P-type ATPase	1228	33.90	0.11	0.35	0.40	0.30
U54642	Phosphoribosylpyrophosphate synthetase	322	34.68	0.13	0.37	0.47	0.33
U56663	Acidic ribosomal phosphoprotein	319	42.49	0.22	0.33	0.44	0.34
U69552	Histidine-rich protein III (HRP III)	264	37.19	0.21	0.38	0.74	0.44
U70366	Rab6	240	38.85	0.14	0.28	0.34	0.26
U73195	MO15-related protein kinase (Pfmrk)	324	33.30	0.09	0.26	0.35	0.24
X03371	MSP1	1630	35.59	0.13	0.25	0.37	0.26
X05074[a]	Exp-1	162	36.23	0.20	0.34	0.51	0.35
X07802	Pfs25	217	44.61	0.15	0.32	0.35	0.28
X15979[a]	α-Tubulin I	453	35.98	0.15	0.40	0.49	0.36
X17483	Asparagine-rich antigen	391	38.12	0.21	0.17	0.28	0.23
X52524	EBA-175	1435	35.98	0.14	0.30	0.37	0.28
X53030	Sexual stage-specific protein	145	46.62	0.24	0.38	0.47	0.37
X56203	LSA-1	1909	33.52	0.23	0.21	0.60	0.35
X58777	RAP-2 protein	398	36.92	0.14	0.28	0.33	0.26
X61921[a]	Pfc2 gene for p34cdc2 protein kinase	288	41.79	0.19	0.32	0.46	0.33
X67288[a]	Cpk	524	41.81	0.21	0.26	0.37	0.29
X69769[a]	GBPH2	309	32.61	0.18	0.37	0.53	0.37
X69922[a]	HRPII	327	41.25	0.25	0.51	0.59	0.51
X73954[a]	Ras-related nuclear protein	214	38.54	0.18	0.32	0.47	0.33
X75420[a]	Cpn60	700	36.30	0.14	0.26	0.39	0.27
X75787	Aspartic hemoglobinase	452	34.54	0.14	0.31	0.38	0.28
X79836[a]	Glycophorin-binding protein gene family	339	32.96	0.18	0.38	0.50	0.36
X80759	Cdc2-related protein kinase	719	40.99	0.18	0.24	0.35	0.26
X81648	Pfs48/45	488	35.05	0.11	0.30	0.35	0.25
X83707[a]	Protein kinase 1	909	36.80	0.13	0.23	0.31	0.23
X84904	Pfg27/25	217	50.46	0.31	0.26	0.45	0.34
X85956	Cyclophilin	195	38.86	0.17	0.32	0.44	0.31
X92977	Rab6 GTPase protein	207	37.90	0.14	0.32	0.39	0.29
X93462[a]	Glutathione reductase	500	37.49	0.13	0.29	0.41	0.28
Y00060[a]	Knob-associated, histidine-rich protein	657	35.25	0.19	0.40	0.51	0.37
Z11832[a]	Gene for putative serine kinase	332	33.09	0.11	0.29	0.33	0.26
Z26314	STARP antigen	604	33.72	0.11	0.36	0.21	0.23
Z31584[a]	Hydroxymethyl-dihydropterin pyrophosphokinase	706	37.00	0.13	0.19	0.36	0.24
Z68200[a]	Glutathione peroxidase	205	36.02	0.18	0.23	0.30	0.25

Nc, the effective number of codons (Wright 1990); GC, G+C content for codon positions 1, 2, 3, and total; GC3 excludes the codons for methionine and stop.

[a] Genes with introns.

is parasitic to chimpanzees and is the closest known species to *P. falciparum* (Coatney *et al.* 1971; Collins and Aikawa 1993; Escalante and Ayala 1994; Escalante *et al.* 1995, 1997). The nucleotide sequence is known for only five of the nine genes surveyed in the present study. Evidence of natural selection emerges at two of the five loci, LSA-1 and Pfs48/45, also shown to be under selection by the test on the basis of excess nonsynonymous substitutions.

DISCUSSION

Although *P. falciparum* has been one of the most extensively investigated parasites, there are severe limitations when seeking to assess its genetic diversity. The gene loci studied remain few, and in several of them, the number of sequences is small. Only sequences from cultured parasites are available for some loci, which introduces bias into the samples. In the case of field isolates, sampling efforts have focused on areas with low genetic diversity (perhaps resulting from transmission differences; Jongwutiwes *et al.* 1993). For example, the CSP data set is mostly limited to Asian samples, and half of the MSP-2 sequences come from India. Moreover, malaria research has focused on those protein parts that are immunologically relevant, so the polymorphism of the complete gene cannot be assessed properly.

This study shows that loci encoding proteins expressed on the surface of the sporozoite and the merozoite are more polymorphic than those expressed during the sexual stages or inside the parasite. These results agree with the general observation that stage-specific

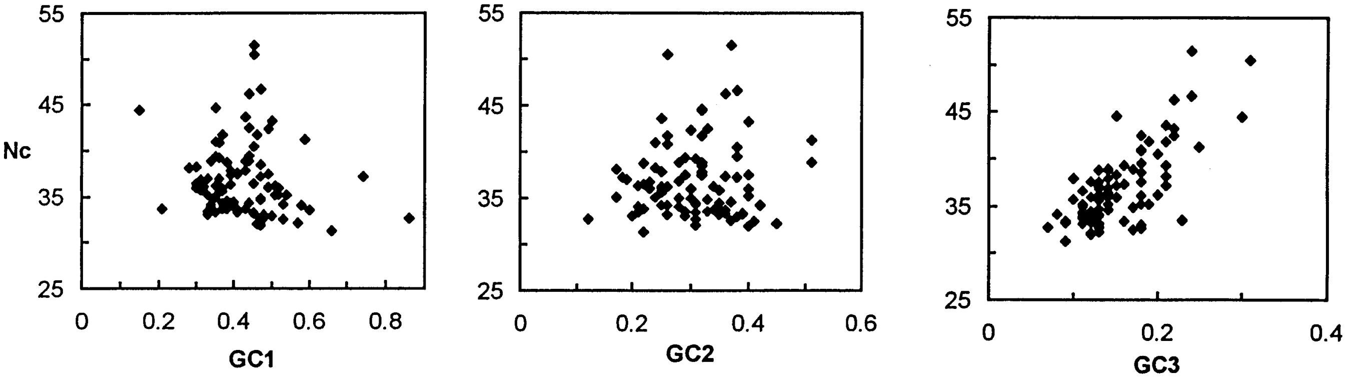

Figure 1.—Correlation in *P. falciparum* between the effective number of codons (*Nc*) and G+C content in the first (GC1), second (GC2), or third (GC3) codon position, on the basis of 92 gene loci. Only the correlation with the third codon position is statistically significant ($r = 0.737$, $P < 0.001$ with $\alpha = 0.05/4 = 0.0125$).

surface proteins exhibit high polymorphism when compared with internal antigens (McCutchan *et al.* 1988; Riley *et al.* 1994). The general inference is that proteins that are involved in the parasite's recognition by the host's immune system are under strong selection pressure for accumulating polymorphism as a means for evading the host's defenses.

The 10 loci of *P. falciparum* we have surveyed are fairly polymorphic, with a weighted average of $\pi = 0.0197$ and a range from 0.002 (RAP-1) to 0.097 (MSP-3). This is higher than the observed diversity in many eukaryotes, such as humans, $\pi = 0.0011$ (Li and Sadler 1991), or in *Drosophila melanogaster*; *e.g.*, $\pi = 0.009$ for Cu, Zn superoxide dismutase, which has an intron of 706 bp (Hudson *et al.* 1994), and $\pi = 0.010$ for the exon of the gene encoding glucose dehydrogenase (Hamblin and Aquadro 1997), even though the genes of *P. falciparum* considered in this study do not include noncoding segments. Intragenic recombination needs to be taken into account as a possible generator of genetic diversity in *P. falciparum.* Concerted evolution has been postulated as the process accounting for the intragenic conservation observed in genes such as CSP, which has a large middle region of tandem repeats, and the highly differentiated alleles of MSP-1 and MSP-2 (Tanabe *et al.* 1987; McCutchan *et al.* 1988; Marshall *et al.* 1991; Frontali 1994; Rich *et al.* 1997). Intragenic similarities between alleles with respect to nonsynonymous substitutions, however, may arise by convergence, *i.e.*, selection for homoplastic substitutions (McCutchan *et al.* 1992; Rich *et al.* 1997). One possible test for discerning between intragenic recombination and homoplasy is to test for recombination only between silent sites. When using only silent sites, we detected recombination only at MSP-1, but the scarcity of silent substitutions makes it difficult to exclude intragenic recombination at the other loci (in addition to the constraints imposed by low G+C content; McCutchan *et al.* 1992). When using

TABLE 5

Polymorphism of *P. falciparum* genes

			Intraspecific		Interspecific		
Gene	*k*	*D*	Syn	Nsyn	Syn	Nsyn	*P*
AMA-1	30.50	0.467	—	—	—	—	—
CSP	5.04	−0.889	4	20	4	34	NS***
EBA-175	6.75	0.305	—	—	—	—	—
LSA-1	4.19	−0.689	0	16	15	25	0.01
MSP-1	91.27	2.157*	—	—	—	—	—
MSP-2	18.25	1.663	—	—	—	—	—
MSP-3	45.56	1.808**	—	—	—	—	—
Pfs25	2.33	−1.619	5	7	10	15	NS***
PF48/45	2.36	0.087	0	6	13	12	0.03
RAP-1	4.67	0.000	0	7	16	48	NS***

k, the number of pairwise nucleotide differences between *P. falciparum* sequences; *D* is for Tajima's test. The number of *P. falciparum* sequences and of segregating sites (S) used for Tajima's test are the same as in Table 1. For the MK test (McDonald and Kreitman 1991): number of synonymous (Syn) and nonsynonymous (Nsyn) substitution sites within *P. falciparum* (intraspecific) and between *P. falciparum* and *P. reichenowi* (interspecific). *P* is for Fisher's exact test. * Significant, $\alpha = 0.05$; ** $P < 0.10$; *** not significant.

both silent and nonsilent substitutions, the SSCF test was significantly positive for four additional surface-expressed loci but not for other genes (Table 1, last column); however, this result may be clouded by homoplasy, as noted.

The polymorphism is unevenly distributed among the loci studied. Loci-encoding proteins expressed on the surface of the sporozoite or the merozoite (AMA-1, CSP, LSA-1, MSP-1, MSP-2, and MSP-3) are more polymorphic than those expressed during the sexual stages or inside the parasite (weighted average of $\pi = 0.040$, range 0.008–0.096 *vs.* weighted average of $\pi = 0.003$, range 0.002–0.004). The level of polymorphism observed in MSP-1 and MSP-3, for example, is comparable to that found in the locus DRB1 of the major histocompatibility complex in humans ($\pi = 0.071$ based on 58 sequences obtained from Marsh and Bodmer 1993). The DRB1 polymorphism is considered to be ancient polymorphism under balancing selection (Ayala *et al.* 1994). In the case of genes such as CSP, MSP-1, and MSP-2, π estimates should be taken as minimum estimates because the polymorphism present at repeat motives or highly divergent regions was excluded from our analysis because of the difficulty in obtaining a reliable alignment. The greater polymorphism obtained in the surface antigens is commonly attributed to natural selection, which is assumed to favor polymorphism in those genes directly exposed to the vertebrate host's immune system as a strategic mechanism for evading the host's defense.

Evidence that surface-expressed proteins exhibit high polymorphism as a consequence of positive natural selection can be seen in that amino acid replacement sites are more polymorphic than synonymous sites. For the most part, synonymous substitutions are generally thought to be selectively neutral rather than restrained by purifying natural selection. The higher incidence of replacement substitutions would then be driven by positive natural selection (Anders and Saul 1994; Kaslow 1994). In our *P. falciparum* loci, the incidence of nonsynonymous substitutions is higher at all loci (and most often significantly so), except for MSP-1 and Pfs25.

The nonsynonymous/synonymous substitution ratio has been used in Drosophila and other organisms for testing whether the nonsynonymous substitutions are under positive selection, which is thought to be the case when the ratio is high (Kreitman and Akashi 1995; Endo *et al.* 1996; Ohta 1996). In particular, the high nonsynonymous/synonymous ratios observed in human HLA and other major histocompatibility complex loci are commonly attributed to natural selection favoring diversity in the antibodies and other specific components of the immune defense (Hughes and Nei 1988, 1989). Recent studies have concluded that the high level of replacement substitutions observed in exposed surface proteins of *P. falciparum* is evidence of positive natural selection (Hughes 1991, 1992; Hughes and Hughes 1995). Endo *et al.* (1996) considered 3595 homologous proteins and found evidence for positive natural selection on 17 gene groups; nine of them are the surface antigens of parasites or viruses, and among them is the MSP-2 of *P. falciparum.* The observation that selection is contributing to the maintenance of genetic polymorphism has also been made for the *env* gene in HIV-1 (Seibert *et al.* 1995). The *env* gene encodes the envelope glycoproteins gp120 and gp41 with regions V1–V5, which are involved in immune evasion and have a higher number of nonsynonymous than synonymous substitutions (a ratio of 2.0 in V3 and 6.4 in V2; other differences are not statistically significant). The remainder parts of the glycoproteins have significantly more synonymous than nonsynonymous substitutions, providing additional evidence for positive natural selection operating in the maintenance of the nonsynonymous polymorphisms observed in the antigen regions (Seibert *et al.* 1995). Similar studies with similar conclusions have been carried out in other genes, such as the hemagglutinin gene of the human influenza A viruses (Ina and Gojobori 1994), and in proteins mediating sperm-egg recognition in marine invertebrates (Vacquier *et al.* 1997).

The conclusion that a high nonsynonymous/synonymous substitution ratio is an indication of natural selection favoring nonsynonymous polymorphisms depends, however, on certain assumptions. Foremost is the assumption that synonymous substitutions are neutral so that they are not subjected to purifying selection or affected by various constraints such as codon bias or G+C content (Ohta 1996). We have noted that for 92 *P. falciparum* loci the overall G+C content is 30.2% and only 15.2% for third codon positions, whereas the average G+C content in first and second codon positions is 42.59 and 30.64%, respectively. This variation in G+C content is in agreement with previous studies that included fewer genes and smaller fragments (Musto *et al.* 1995). Frontali (1994) has observed that about half of *P. falciparum*'s genome consists of noncoding regions with very low G+C, even lower than for coding regions, but no quantitative analysis has been provided. We have studied the 37 introns reported in the list of the 92 loci in Table 4 and found an average G+C content of only 13.03% (95% C.I. 11.8–14.3, using a *t* distribution). The overall G+C content of the *P. falciparum* genome is 18% (McCutchan *et al.* 1984). Selection or mutation bias may thus be favoring substitutions towards A+T, whether or not they are synonymous. Be that as it may, the high A+T content indicates that constraints exist so that synonymous substitutions may not be completely neutral. The existence of these constraints can be observed in Figure 1, showing a strong correlation between *Nc* and G+C content in the third position (see also Hyde and Sims 1987; Musto *et al.* 1995). Musto *et al.* (1995) have concluded that there is a composition constraint operating in all the translated sequences and codon positions, but Hughes (1991, 1992; Hughes and Hughes 1995) has argued that G+C content does not account for the high nonsynonymous/synonymous substitution ratio observed in *P. falciparum*, although his argument has not been accepted definitively. Endo *et al.* (1996) have thus suggested that the mutation pressure towards increasing A+T content may account for the low level of synonymous substitutions in *P. falciparum*. A similar argument has been made in other cases where a strong base composition bias is present; for example, the excess of nonsynonymous with respect to synonymous substitutions in hepatitis delta virus has been explained by a strong preference of G+C at the third codon positions (Krushkal and Li 1995).

Codon bias is associated with genome G+C content, but it can be accounted for by either selection or mutation pressure (Gillespie 1991; Ohta 1996). The "mutation pressure" could be caused by mutation bias (Sueoka 1962, 1988), which may be detected if a correlation is found between the G+C content of introns and exons in the same genes (Shields *et al.* 1988; Moriyama and Hartl 1993; Powell and Moriyama 1997). We found no correlation between intron G+C composition and (1) G+C content of exons ($r = -0.176$, $P = 0.283$), (2) G+C content at the third codon positions ($r = -0.170$, $P = 0.300$), or (3) the effective number of codons ($r = -0.198$, $P = 0.226$). Because there is no evidence supporting a mutation bias, it may be possible that the observed codon usage is caused by natural selection (Gillespie 1991). Selection favoring specific codons may impose strong constraints on synonymous substitutions.

This trend for keeping a strong codon bias and an A+T-rich genome, whether it is selection or mutation pressure, will affect the number of synonymous substitutions. The substitutions observed in a set of alleles will then be in the direction of increasing the A+T content. The difficulty is that there is no direct way to observe whether a mutation was from A to G or G to A because we don't have enough information for establishing the ancestral state at the polymorphic sites under question. However, we can quantify the transversions A↔T and G↔C because these preserve the G+C content regardless of the direction of the change. We also can estimate the G+C content maintained in a given pair of sequences using those sites that do not change. This allows us to build a very conservative test. If the observed substitutions are caused by a genomic trend towards increasing A+T richness or keeping it high, we expect that (1) the ratio between the average number of G+C *vs.* A+T substitutions on all sequence pairs should not differ from the ratio of G+C *vs.* A+T sites estimated using invariant sites (assuming equilibrium in G+C composition at that specific gene) or (2) that the ratio of G+C *vs.* A+T substitutions will be lower because the G+C substitutions will be affected by purifying selection while A+T substitutions will be neutral or favored by selection. The results are summarized in Table 6. The G+C substitutions are more abundant than expected in the four genes AMA-1, CSP, LSA-1, and Pfs48/45. Although this is not a test for neutrality, it shows that the observed substitutions cannot be explained by a trend for increasing A+T content. This test is highly conservative because it requires a pattern leading to an increase in G+C content that can be observed in the accumulation of G+C transversions. This result thus favors the conclusion that the excess of nonsynonymous substitutions is caused by positive natural selection.

An additional source of concern is that the estimation of synonymous substitutions can be affected by nucleotide composition. Ina (1995) noted that all methods have some degree of bias when they are used on sequences with uneven nucleotide frequencies. On the whole, the method of Li (1993) provides better estimates than that of Nei and Gojobori (1986); both methods are biased, but in different directions. We noted in results that in the case of MSP-2, the excess of nonsynonymous substitutions is significant according to the method of Nei and Gojobori, but not according to Li's. This might occur because Nei and Gojobori's method underestimates the number of synonymous substitutions and overestimates the number of nonsynonymous substitutions when there is a strong bias in nucleotide composition, while Li's (1993) method overestimates the number of synonymous substitutions and underestimates the number of nonsynonymous substitutions (Ina 1995). To the best of our knowledge, no systematic studies have been performed about the effect of overall G+C content on the statistics used for estimating synonymous and nonsynonymous substitutions.

The conclusion at this point is that although the nonsynonymous/synonymous substitution ratio suggests that natural selection may account for the high levels of amino acid polymorphism observed at seven of the 10 loci studied, the evidence is clouded by the constraints imposed by the particular characteristics of codon bias and A+T content in the *P. falciparum* genome. These characteristics evidence an overall trend for keeping or increasing A+T richness in this genome. This may decrease the number of synonymous substitutions and, thus, affect the nonsynonymous *vs.* synonymous substitution ratio. However, this trend cannot explain the accumulation of substitutions in the case of AMA-1, CSP, LSA-1, and Pfs48/45 because transversions towards G+C are more frequent than expected. This G+C accumulation, in addition to the observed ratio of nonsynonymous *vs.* synonymous substitutions, supports that selection is operating at least on these four genes.

In the case of MSP-1 and MSP-3, we could not detect evidence for positive natural selection using the nonsynonymous/synonymous substitution method in the regions under study, although Tajima's test shows that MSP-1 (Table 5) and perhaps MSP-3 are subject to selec-

TABLE 6

The ratio GC/AT in 10 *P. falciparum* genes

Gene	Invariant sites	Substitution sites
AMA-1	0.4313 [0.4303–0.4323]	0.7333 [0.6362–0.8452]
CSP	0.4391 [0.4388–0.4895]	2.1212 [1.7395–2.6606]
EBA-175	0.5259 [0.5257–0.5261]	0.0429 [0.0137–0.0809]
LSA-1	0.3292 [0.3282–0.3304]	0.8061 [0.5877–1.0602]
MSP-1	0.2111 [0.2087–0.2135]	0.1826 [0.1799–0.1853]
MSP-2	0.4262 [0.4250–0.4277]	0.1561 [0.1492–0.1640]
MSP-3	0.4301 [0.4282–0.4319]	0.3368 [0.2995–0.3816]
Pfs25	0.3932 [0.3927–0.3936]	0.00
Pfs48/45	0.3414 [0.3414–0.3420]	0.5161 [0.3529–0.7273]

Bootstrap 95% confidence intervals are indicated within brackets.

tion. It is possible that selectively favored nonsynonymous substitutions become saturated over time, producing a synonymous/nonsynonymous substitution ratio that is consistent with neutrality in distantly related sequences but could be detected in closely related sequences (Hughes and Nei 1988; Hughes 1992). Lack of significance at other loci using Tajima's test may not be interpreted as negative evidence because this test is highly conservative and has limited power whenever there has been a recent bottleneck or selective sweep (Gillespie 1994; Simonsen *et al.* 1995), as most likely seems to have occurred in *P. falciparum* (Rich *et al.* 1997). Moreover, several sequences are partial, which further limits the power of the test.

The MK test uses the number of fixed substitutions between two closely related species (nonsynonymous/synonymous ratio) as the expected value under neutrality (McDonald and Kreitman 1991). Of the nine loci surveyed in *P. falciparum*, five have been sequenced in *P. reichenowi*, which are thus the only ones to which the MK test can be applied. The test is significant at two loci, LSA-1 and Pfs48/45 (Table 5); the number of replacement substitutions is greater than expected under neutrality at both of these loci. The MK test corrects for various constraints, such as those arising from G+C content and codon bias, if we assume that the pattern of divergent substitutions between closely related species is also affected by the same kind of constraints as are intraspecific substitutions.

The closest known relative of *P. falciparum* is *P. reichenowi*, a chimpanzee parasite. The time of divergence between *P. falciparum* and *P. reichenowi* has been estimated to be ~5–8 mya, about the same time when the chimpanzee and human lineages diverged (Coatney *et al.* 1971; Collins and Aikawa 1993; Escalante and Ayala 1994; Escalante *et al.* 1995). The genetic distance between *P. reichenowi* and *P. falciparum* alleles for the CSP is about five times as large as the distance between intraspecific *P. falciparum* alleles (Escalante *et al.* 1995). The information available about *P. reichenowi* is limited to 10 gene loci (including the 18S rRNA); however, the assumption that the G+C content or codon bias does not affect the MK results appears to be reasonable. For example, the overall G+C content of *P. reichenowi* is 0.29 (95% C.I. 0.255–0.325, *t* distribution), the G+C content at the third codon position is 0.158 (95% C.I. 0.105–0.21), and the effective number of codons is 39.6 (95% C.I. 34.3–44.9). These values are not statistically different from those observed in the same genes of *P. falciparum*. Table 7 gives the number of transitions and transversions, as well as the number of synonymous and nonsynonymous substitutions between *P. reichenowi* and the *P. falciparum*. The transversion bias appears to be less pronounced than the one found intraspecifically, specially for LSA-1, one of the genes for which we found evidence of selection using the MK test. This may result from the presence of more synonymous than nonsynonymous substitutions between *P. reichenowi* and *P. falciparum* for all loci, except for the one encoding the CSP.

One potential problem with the MK test in the present case is that we have only one *P. reichenowi* sequence at each of the five loci tested (only one isolate of *P. reichenowi* is known to be available; Coatney *et al.* 1971, p. 309). If some sites are polymorphic in *P. reichenowi*, we are likely to overestimate the number of fixed differences between the species. This potential problem is not likely to be important because the MK test compares proportions and there is no reason to expect that the proportion of nonsynonymous/synonymous substitutions will be affected by *P. reichenowi* polymorphism. The MK test is detecting a disproportionate accumulation of nonsynonymous substitutions in *P. falciparum* when compared to *P. reichenowi*, which is apparent when K_s and K_n are compared within *P. falciparum* (Table 2) and between *P. falciparum* and *P. reichenowi* (Table 7). A similar approach was used by Mindell (1996) for studying the effect of natural selection in the maintenance of the genetic polymorphism in HIV-1 viruses.

In conclusion, there is evidence of natural selection contributing to amino acid polymorphism at nine loci: MSP-1 and MSP-3 (Tajima's test); LSA-1 and Pfs48/45 (MK test); AMA-1, CSP, EBA175, LSA-1, MSP-2, Pfs48/

TABLE 7

Transitions, transversions, synonymous, and nonsynonymous substitutions between *P. falciparum* alleles and *P. reichenowi*

Gene	Transitions	Transversions	K_s	K_n
CSP	23.41 [23.05–23.73]	23.09 [22.55–23.59]	0.0450 [0.0425–0.0473]	0.0667 [0.0659–0.0674]
LSA-1	24.29 [24.00–24.57]	21.64 [21.29–22.00]	0.1543 [0.1519–0.1573]	0.0908 [0.0891–0.0925]
Pfs25	16.23 [15.92–16.54]	10.08 [10.00–10.23]	0.0750 [0.0725–0.0769]	0.0332 [0.0326–0.0338]
Pfs48/45	14.00 [13.50–14.40]	14.63 [13.88–15.38]	0.0475 [0.0475–0.0475]	0.0149 [0.0141–0.0155]
RAP1	29.33	37.33	0.0382	0.0265

K_s and K_n are as described in Table 2. Bootstrap 95% confidence intervals are indicated within brackets.

45, and RAP-1 (synonymous/nonsynonymous rates). The evidence derived from the intraspecific nonsynonymous/synonymous ratio may be questionable if a trend towards increasing A+T richness could account for this pattern, but this is not the case for AMA-1 and CSP. The evidence for MSP-2 may be questioned because the excess of nonsynonymous substitutions was not significant according to the method of Li (1993). The evidence for MSP-3 may also be questioned because of the ambiguous significance obtained with Tajima's test. Positive natural selection seems definitely established in at least five of the 10 loci investigated.

AMA-1, CSP, LSA-1, and MSP-1 are host-exposed surface proteins in which, as noted above, natural selection is generally assumed to favor polymorphism as an evasion strategy from the host's immune system. Pfs48/45 is only moderately immunogenic, with antibody levels that vary geographically and with grade of exposure (Kaslow 1994). It may be that even moderate or low immune host activity generates sufficient selective pressure to be detectable in the parasite's polymorphism.

A. A. Escalante is supported by a fellowship from the American Society for Microbiology. Research in F. J. Ayala's laboratory is supported by National Institutes of Health grant GM42397. This work was supported in part by U.S. Agency for International Development grant HRN-60010-A-00-4010-00 to A. A. Lal.

LITERATURE CITED

Anders, R. F., and A. J. Saul, 1994 Candidate antigens for an asexual blood stage vaccine against falciparum malaria, pp. 169–208 in *Molecular Immunological Considerations in Malaria Vaccine Development*, edited by M. F. Good and A. J. Saul. CRC Press, Boca Raton, FL.

Andersson, S. G. E., and P. M. Sharp, 1996 Codon usage and base composition in *Rickettsia prowazekii*. J. Mol. Evol. **42:** 525–536.

Ayala, F. J., A. Escalante, C. O'Huigin and J. Klein, 1994 Molecular genetics of speciation and human origins. Proc. Natl. Acad. Sci. USA **91:** 6787–6794.

Bhattacharya, P., P. Malhotra, P. Sharma, D. M. N. Okenu and V. S. Chauhan, 1995 Merozoite surface antigen 2 (MSA-2) gene of *Plasmodium falciparum* strains from India. Mol. Biochem. Parasitol. **74:** 125–128.

Brown, W. M., E. M. Prager, A. Wang and A. C. Wilson, 1982 Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18:** 225–239.

Campbell, J. R., 1989 DNA sequence of the gene encoding a *Plasmodium falciparum* malaria candidate vaccine antigen. Nucleic Acids Res. **17:** 5854.

Caspers, P., R. Gentz, H. Matile, J. R. Pink and F. Sinigaglia, 1989 The circumsporozoite protein gene from NF54, a *Plasmodium falciparum* isolate used in malaria vaccine trials. Mol. Biochem. Parasitol. **35:** 185–190.

Chang, S. P., K. J. Kramer, K. M. Yamaga, A. Kato, S. E. Case *et al.*, 1988 *Plasmodium falciparum*: gene structure and hydropathy profile of the major merozoite surface antigen (gp195) of the Uganda-Palo Alto isolate. Exp. Parasitol. **67:** 1–11.

Cheng, T. C., 1986 *General Parasitology*, Ed. 2. Academic Press, Orlando, FL.

Coatney, G. R., W. E. Collins, M. Warren and P. G. Contacos, 1971 *The Primate Malarias*, U.S. Government Printing Office, Washington, DC.

Collins, W. E., and M. Aikawa, 1993 Plasmodia of nonhuman primates, pp. 105–133 in *Parasitic Protozoa*, Vol. 5, edited by J. P. Kreier. Academic Press, New York.

Collins, D. W., and T. H. Jukes, 1994 Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics **20:** 386–396.

Conover, W. J., 1980 *Practical Nonparametric Statistics*. John Wiley & Sons, New York.

Conway, D. J., 1997 Natural selection on polymorphic malaria antigens and the search for a vaccine. Parasitol. Today **13:** 26–29.

Conway, D. J., V. Rosario, A. M. J. Oduola, L. A. Salako, B. M. Greenwood *et al.*, 1991 *Plasmodium falciparum*: intragenic recombination and nonrandom associations between polymorphic domains of the precursor to the major merozoite surface antigens. Exp. Parasitol. **73:** 469–480.

Crozier, R. H., and Y. C. Crozier, 1993 The mitochondrial genome of the honeybee *Apis mellifera* complete sequence and genome organization. Genetics **133:** 97–117.

Dame, J. B., J. L. Williams, T. F. McCutchan, J. L. Weber, R. A. Wirtz *et al.*, 1984 Structure of the gene encoding the immunodominant surface antigen on the sporozoite of the human malaria parasite *Plasmodium falciparum*. Science **225:** 593–599.

del Portillo, H. A., R. S. Nussenzweig and V. Enea, 1987 Circumsporozoite gene of a *Plasmodium falciparum* strain from Thailand. Mol. Biochem. Parasitol. **24:** 289–294.

Dunn, O. J., 1961 Multiple comparisons among means. J. Am. Stat. Assoc. **56:** 52–64.

Efron, B., and R. J. Tibshirani, 1993 *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Endo, T., K. Ikeo and T. Gojobori, 1996 Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13:** 685–690.

Escalante, A. A., and F. J. Ayala, 1994 Phylogeny of the malarial genus *Plasmodium* derived from rRNA gene sequences. Proc. Natl. Acad. Sci. USA **91:** 11373–11377.

Escalante, A. A., E. Barrio and F. J. Ayala, 1995 Evolutionary origin of human and primate malarias: evidence from the circumsporozoite protein gene. Mol. Biol. Evol. **12:** 616–626.

Escalante, A. A., I. F. Goldman, P. De Rijk, R. De Wachter, W. E. Collins *et al.*, 1997 Phylogenetic study of the genus *Plasmodium* based on the secondary structure-based alignment of the small subunit ribosomal RNA. Mol. Biochem. Parasitol. **90:** 317–321.

Fenton, B., J. T. Clark, C. M. Khan, J. V. Robinson, D. Walliker *et al.*, 1991 Structural and antigenic polymorphism of the 35- to 48-kilodalton merozoite surface antigen (MSA-2) of the malaria parasite *Plasmodium falciparum*. Mol. Cell. Biol. **11:** 963–974.

Fidock, D. A., H. Gras-Masse, J. P. Lepers, K. Brahimi, L. Benmohamed *et al.*, 1994 *Plasmodium falciparum* liver stage antigen-1 is well conserved and contains B and T cell determinants. J. Immunol. **153:** 190–204.

Frontali, C., 1994 Genome plasticity in *Plasmodium*. Genetica **94:** 91–100.

Gillespie, J. H., 1991 *The Causes of Molecular Evolution*, Oxford University Press, New York.

Gillespie, J. H, 1994 Alternatives to Neutral Theory, pp. 1–17 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

Gleason, J. M., A. Caccone, E. N. Moriyama, K. P. White and J. R. Powell, 1997 Mitochondrial DNA phylogenies for the *Drosophila obscura* group. Evolution **51:** 433–440.

Gojobori, T., W. H. Li and D. Graur, 1982 Patterns of nucleotide substitutions in pseudogenes and functional genes. J. Mol. Evol. **18:** 360–369.

Hamblin, M. T., and C. F. Aquadro, 1997 Contrasting patterns of nucleotide sequence variation at the glucose dehydrogenase (Gld) locus in different populations of *Drosophila melanogaster*. Genetics **145:** 1053–1062.

Hancock, G. R., and A. J. Klockars, 1996 The quest for α: developments in multiple comparison procedures in the quarter century since Games (1971). Rev. Edu. Res. **66:** 269–306.

Hartl, D. L., and S. A. Sawyer, 1991 Inference of selection and recombination from nucleotide sequence data. J. Evol. Biol. **4:** 519–532.

Hill, A. V. S., J. Elvin, A. C. Willis, M. Aidoo, C. E. M. Allsopp *et al.*, 1992 Molecular analysis of the association of HLA-B53 and resistance to severe malaria. Nature **360:** 434–439.

Huber, W., I. Felger, H. Matile, H. J. Lipps, S. Steiger *et al.*, 1997 Limited sequence polymorphism in the *Plasmodium falciparum* merozoite surface antigen 3. Mol. Biochem. Parasitol. **87:** 231–234.

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology*, edited by N. Takahata and A. G. Clark. Sinauer Associates, Inc., Sunderland, MA.

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

Hughes, A. L., 1991 Circumsporozoite protein genes of malaria parasites (*Plasmodium* spp.): evidence for positive selection on immunogenic regions. Genetics **127:** 345–353.

Hughes, A. L., 1992 Positive selection and intrallelic recombination at the merozoite surface antigen-1 (MSA-1) locus of *Plasmodium falciparum*. Mol. Biol. Evol. **9:** 381–393.

Hughes, M. K., and A. L. Hughes, 1995 Natural selection on *Plasmodium* surface proteins. Mol. Biochem. Parasitol. **71:** 99–113.

Hughes, A. L., and M. Nei, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167–170.

Hughes, A. L., and M. Nei, 1989 Nucleotide substitution at major histocompatibility complex class II loci: evidence of overdominant selection. Nature **335:** 167–170.

Hyde, J. E., and F. G. Sims, 1987 Anomalous dinucleotide frequencies in both coding and non-coding regions from the genome of the human malaria parasite *Plasmodium falciparum*. Gene **61:** 177–187.

Ina, Y., 1995 New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. **40:** 190–226.

Ina, Y., and T. Gojobori, 1994 Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses. Proc. Natl. Acad. Sci. USA **91:** 8388–8392.

Inohira, K., T. Hara and E. T. Matsura, 1997 Nucleotide sequence divergence in the A+T-rich region of mitochondrial DNA in *Drosophila simulans* and *Drosophila mauritiana*. Mol. Biol. Evol. **14:** 814–822.

Jongwutiwes, S., K. Tanabe and H. Kanbara, 1993 Sequence conservation in the C-terminal part of the precursor to the major merozoite surface proteins (MSP-1) of *Plasmodium falciparum* from field isolates. Mol. Biochem. Parasitol. **59:** 95–100.

Jongwutiwes, S., K. Tanabe, M. K. Hughes, H. Kanbara and A. L. Hughes, 1994 Allelic variation in the circumsporozoite protein of *Plasmodium falciparum* from Thai field isolates. Am. J. Trop. Med. Hyg. **51:** 659–668.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kaslow, D. C., 1994 Progress toward a transmission-blocking vaccine, pp. 209–244 in *Molecular Immunological Considerations in Malaria Vaccine Development*, edited by M. F. Good and A. J. Saul. CRC Press, Boca Raton, FL.

Kaslow, D. C., I. A. Quakyi and D. B. Keister, 1989 Minimal variation in a vaccine candidate from the sexual stage of *Plasmodium falciparum*. Mol. Biochem. Parasitol. **32:** 101–104.

Kimura, M., 1977 Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267:** 275–276.

Kimura, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111–120.

Kimura, M., 1983 The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Kocken, C. H. M., J. Jansen, A. M. Kann, P. J. A. Beckers, T. Ponnudurai *et al.*, 1993 Cloning and expression of the gene coding for the transmission blocking target antigen Pfs48/45 of *Plasmodium falciparum*. Mol. Biochem. Parasitol. **61:** 59–68.

Kocken, C. H. M., R. L. B. Milek, T. H. W. Lensen, D. C. Kaslow, J. G. G. Schoenmakers *et al.*, 1995 Minimal variation in the transmission-blocking vaccine candidate Pfs48/45 of the human malaria parasite *Plasmodium falciparum*. Mol. Biochem. Parasitol. **69:** 115–118.

Kondo, R., H. Satoshi, Y. Satta and N. Takahata, 1993 Evolution of hominoid mitochondrial DNA with special reference to the silent substitution rate over the genome. J. Mol. Evol. **36:** 517–531.

Kreitman, M., and H. Akashi, 1995 Molecular evidence for natural selection. Annu. Rev. Ecol. Syst. **26:** 403–422.

Krushkal, J., and W. H. Li, 1995 Substitution rates in hepatitis delta virus. J. Mol. Evol. **41:** 721–726.

Kumar, S., K. Tamura and M. Nei, 1994 MEGA-Molecular Evolutionary Genetics Analysis Software for Microcomputers. Comput. Appl. Biosci. **10:** 189–191.

Lal, A. A., and I. F. Goldman, 1991 Circumsporozoite protein gene from *Plasmodium reichenowi*, a chimpanzee malaria parasite evolutionary related to the human malaria parasite *Plasmodium falciparum*. J. Biol. Chem. **266:** 6686–6688.

Lal, A. A., I. F. Goldman and G. H. Campbell, 1990 Primary structure of the 25-kilodalton ookinete antigen from *Plasmodium reichenowi*. Mol. Biochem. Parasitol. **43:** 143–146.

Li, W.-H., 1993 Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36:** 96–99.

Li, W.-H., and L. A. Sadler, 1991 Low nucleotide diversity in man. Genetics **129:** 513–523.

Li, W.-H., C.-I. Wu and C.-C. Luo, 1984 Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. **21:** 58–71.

Liang, H., and B. K. L. Sim, 1997 Conservation of structure and function of the erythrocyte-binding domain of *Plasmodium falciparum* EBA-175. Mol. Biochem. Parasitol. **84:** 241–245.

Lloyd, A. T., and P. M. Sharp, 1992 CODONS: a microcomputer program for codon usage analysis. J. Hered. **83:** 239–240.

Lockyer, M. J., and R. T. Schwarz, 1987 Strain variation in the circumsporozoite protein gene of *Plasmodium falciparum*. Mol. Biochem. Parasitol. **22:** 101–108.

Lockyer, M. L., K. Marsh and C. I. Newbold, 1989 Wild isolates of *Plasmodium falciparum* show extensive polymorphism in T cell epitopes of the circumsporozoite protein. Mol. Biochem. Parasitol. **37:** 275–280.

Marsh, S. G. E., and J. G. Bodmer, 1993 HLA Class-II nucleotide sequences. Immunogenetics **37:** 79–94.

Marshall, V. N., R. L. Coppel, R. K. Martin, A. M. J. Oduola, R. F. Anders *et al.*, 1991 A *Plasmodium falciparum* MSA-2 gene apparently generated by intragenic recombination between the two allelic families. Mol. Biochem. Parasitol. **45:** 349–352.

Marshall, V. M., R. L. Coppel, R. F. Anders and D. J. Kemp, 1992 Two alleles within subfamilies of the merozoite surface protein antigen 2 (MSA-2) of *Plasmodium falciparum*. Mol. Biochem. Parasitol. **50:** 181–184.

Maynard-Smith, J., 1994 Estimating selection by comparing synonymous and substitutional changes. J. Mol. Evol. **39:** 123–128.

McColl, D. J., A. Silva, M. Foley, J. F. Kun, J. M. Favaloro *et al.*, 1994 Molecular variation in a novel polymorphic antigen associated with *Plasmodium falciparum* merozoites. Mol. Biochem. Parasitol. **68:** 53–67.

McCutchan, T. F., J. B. Dame, L. H. Miller and J. Barnwell, 1984 Evolutionary relatedness of *Plasmodium* species as determined by the structure of DNA. Science **225:** 808–811.

McCutchan, T. F, V. F. de la Cruz, M. F. Good and T. E. Wellems, 1988 Antigenic diversity in *Plasmodium falciparum*. Prog. Allergy **41:** 173–192.

McCutchan, T. F., A. A. Lal, V. de Rosario and A. P. Waters, 1992 Two types of sequence polymorphism in the circumsporozoite gene of *Plasmodium falciparum*. Mol. Biochem. Parasitol. **50:** 37–46.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. Nature **351:** 652–654.

Miller, L. H., T. Roberts, M. Shahabuddin and T. F. McCutchan, 1993 Analysis of sequence diversity in the *Plasmodium falciparum* merozoite surface protein-1 (MSP-1). Mol. Biochem. Parasitol. **59:** 1–14.

Mindell, D. P., 1996 Positive selection and rates of evolution in immunodeficiency viruses from humans and chimpanzees. Proc. Natl. Acad. Sci. USA **93:** 3284–3288.

Moriyama, E. N., and D. L. Hartl, 1993 Codon usage bias and base composition of nuclear gene in *Drosophila*. Genetics **134:** 847–858.

Musto, H., H. Rodriguez-Maseda and G. Bernardi, 1995 Compositional properties of nuclear genes from *Plasmodium falciparum*. Gene **152:** 127–132.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

Oeuvray, C., H. Bouharoun-Tayoun, H. Gras-Masse, E. Bottius, T. Kaidoh *et al.*, 1994 Merozoite surface protein-3: a malaria protein inducing antibodies that promote *Plasmodium falciparum* killing by cooperation with blood monocytes. Blood **84:** 1594–1602.

Ohta, T., 1992 The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. **23:** 263–286.

Ohta, T., 1996 The neutralist-selectionist debate. Bioessays **18:** 673–677.

Oliveira, D. A., V. Udhayakumar, P. Bloland, Y. P. Shi, B. L. Nahlen *et al.*, 1996 Genetic conservation of the *Plasmodium falciparum* apical membrane antigen-1 (AMA-1). Mol. Biochem. Parasitol. **76:** 333–336.

Pan, W., R. Tolle and H. Bujard, 1995 A direct and rapid sequencing strategy for the *Plasmodium falciparum* antigen gene gp190/MSA1. Mol. Biochem. Parasitol. **73:** 241–244.

Peterson, M. G., R. L. Coppel, P. McIntyre, C. J. Langford, G. Woodrow *et al.*, 1988 Variation in the precursor to the major merozoite surface antigens of *Plasmodium falciparum*. Mol. Biochem. Parasitol. **27:** 291–302.

Peterson, M. G., V. M. Marshall, J. A. Smythe, P. E. Crewther, A. Lew *et al.*, 1989 Integral membrane protein located in the apical complex of *Plasmodium falciparum*. Mol. Cell. Biol. **9:** 3151–3154.

Powell, J. R., and E. N. Moriyama, 1997 Evolution of codon usage bias in *Drosophila*. Proc. Natl. Acad. Sci. USA **94:** 7784–7790.

Rich, S. M., R. R. Hudson and F. J. Ayala, 1997 *Plasmodium falciparum* antigenic diversity: evidence of clonal population structure. Proc. Natl. Acad. Sci. USA **94:** 13040–13045.

Ridley, R. G., B. Takacs, H. W. Lahm, C. J. Delves, M. Goman *et al.*, 1990 Characterisation and sequence of a protective antigen from *Plasmodium falciparum*. Mol. Biochem. Parasitol. **41:** 125–134.

Riley, E. M., L. Hviid and T. G. Theander, 1994 Malaria, pp. 119–143 in *Parasitic Infections and the Immune System*, edited by F. Kierszenbaum. Academic Press, Inc., San Diego.

Sawyer, S. A., 1989 Statistical test for detecting gene conversion. Mol. Biol. Evol. **6:** 526–538.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132:** 1161–1176.

Seibert, S. A., C. A. Howell, M. K. Hughes and A. L. Hughes, 1995 Natural selection on the gag, pol, and env genes of the human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. **12:** 803–813.

Sharp, P. M., and W.-H. Li, 1987 The rate of synonymous substitutions in eubacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:** 222–230.

Shi, Y. P., M. P. Alpers, M. M. Povoa and A. A. Lal, 1992a Diversity in the immunodominant determinants of the circumsporozoite protein of *Plasmodium falciparum* parasites from malaria endemic regions of Papua New Guinea and Brazil. Am. J. Trop. Med. Hyg. **47:** 844–855.

Shi, Y. P., M. P. Alpers, M. M. Povoa and A. A. Lal, 1992b Single amino acid variation in the ookinete vaccine antigen from field isolates of *Plasmodium falciparum*. Mol. Biochem. Parasitol. **50:** 179–180.

Shields, D. C., P. M. Sharp, D. G. Higgins and W. Wright, 1988 'Silent' sited in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

Sim, B. K. L., 1995 EBA-175—an erythrocyte-binding ligand of *Plasmodium falciparum*. Parasitol. Today **6:** 213–217.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

Smythe, J. A., M. G. Peterson, R. L. Coppel, A. Saul, D. J. Kemp *et al.*, 1990 Structural diversity in the 45-kilodalton merozoite surface antigen of *Plasmodium falciparum*. Mol. Biochem. Parasitol. **39:** 227–234.

Sueoka, N., 1962 On the genetic basis of variation and heterogeneity on DNA base composition. Proc. Natl. Acad. Sci. USA **48:** 582–592.

Sueoka, N., 1988 Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA **85:** 2653–2657.

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

Tamura, K., 1992 Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. Mol. Biol. Evol. **9:** 678–687.

Tanabe, K., M. Mackay, M. Goman and J. G. Scaife, 1987 Allelic dimorphism in a surface antigen gene of the malaria parasite *Plasmodium falciparum*. J. Mol. Biol. **195:** 273–287.

Thomas, A. W., D. A. Carr, J. M. Carter and J. A. Lyon, 1990 Sequence comparison of allelic forms of the *Plasmodium falciparum* merozoite surface antigen MSA-2. Mol. Biochem. Parasitol. **43:** 211–220.

Tolle, R., H. Bujard and J. A. Cooper, 1995 *Plasmodium falciparum*: variations within the C-terminal region of merozoite surface antigen-1. Exp. Parasitol. **81:** 47–54.

Vacquier, V. D., W. J. Swanson and Y. H. Lee, 1997 Positive Darwinian selection on two homologous fertilization proteins: what is the selection pressure driving their divergence? J. Mol. Evol. **44:** S15–S22.

Ware, L. A., K. C. Kain, B. K. L. Sim, J. D. Haynes, J. K. Baird *et al.*, 1993 Two alleles of the 175-kilodalton *Plasmodium falciparum* erythrocyte binding antigen. Mol. Biochem. Parasitol. **60:** 105–109.

Weber, J. L., B. K. Sim, J. A. Lyon and R. Wolff, 1988 Merozoite surface protein sequence from the Camp strain of the human malaria parasite *Plasmodium falciparum*. Nucleic Acids Res. **16:** 1206.

Wolstenholme, D. R., and D. O. Clary, 1985 Sequence evolution of *Drosophila* mitochondrial DNA. Genetics **109:** 725–744.

World Health Organization, 1995 Tropical Disease Research. TDR Twelfth Program Report, pp. 57–76. World Health Organization, Geneva, Switzerland.

Wright, F., 1990 The 'effective number of codons' used in a gene. Gene **87:** 23–29.

Yang, C., Y. P. Shi, V. Udhayakumar, M. P. Alpers, M. M. Povoa *et al.*, 1995 Sequence variation in the non-repetitive regions of the liver stage-specific antigen-1 (LSA-1) of *Plasmodium falciparum* from field isolates. Mol. Biochem. Parasitol. **71:** 291–294.

Zhu, J., and M. Hollingdale, 1991 Structure of *Plasmodium falciparum* liver stage antigen-1. Mol. Biochem. Parasitol. **48:** 223–226.

Communicating editor: W.-H. Li