Positive Selection and Interallelic Recombination at the Merozoite Surface Antigen–1 (MSA-1) Locus of *Plasmodium falciparum*[1]

Austin L. Hughes

Department of Biology, Institute of Molecular Evolutionary Genetics,
The Pennsylvania State University

DNA sequences of alleles at the merozoite surface antigen–1 (MSA-1) gene locus of the malaria parasite *Plasmodium falciparum* show evidence of repeated past recombination events between alleles. These include both (1) nonreciprocal recombination events that have homogenized certain gene regions among alleles and (2) reciprocal recombination events that have combined allelic segments with divergent evolutionary histories, thereby enhancing allelic diversity. In three different gene regions, the rate of nonsynonymous nucleotide substitution significantly exceeds that of synonymous nucleotide substitution, implying that positive Darwinian selection has acted to diversify alleles at the amino acid level. The MSA-1 polymorphism seems to be quite ancient; the two major allelic types have been maintained for ~35 Myr.

Introduction

The immune system of vertebrates, which is adapted to recognize foreign proteins, is expected to exert natural selection on parasitic organisms favoring avoidance of immune recognition (Sher 1988; Capron and Dessaint 1989). In the case of the human malaria parasite *Plasmodium falciparum,* there is evidence of such selection at the molecular level. Epitopes of the circumsporozoite (CS) protein which are known to be bound by host major-histocompatibility-complex (MHC) molecules and recognized by T-cell receptors (TCR) show an accelerated rate of nonsynonymous nucleotide substitution (Hughes 1991*a*). The CS protein is a major surface protein of the sporozoite, the stage of the malaria parasite that infects the vertebrate host. After infection, the sporozoites migrate to liver cells, where they mature and divide mitotically, eventually releasing merozoites into the bloodstream. The merozoites in turn invade erythrocytes, and, after several cell divisions, release further merozoites (Mahmoud 1989). Reasoning that merozoite surface proteins may be subject to selective pressures similar to those affecting the CS protein, I here analyze published DNA sequences of alleles at the merozoite surface antigen–1 (MSA-1) locus in *P. falciparum.*

I test for evidence of positive selection by comparing rates of synonymous and nonsynonymous nucleotide substitution (Hughes and Nei 1988). Since Tanabe et al. (1987) proposed that recombination has occurred among alleles at this locus, I also test for evidence of interallelic recombination and examine the interaction between

1. Key words: interallelic recombination, merozoite, MSA-1, *Plasmodium falciparum,* positive selection.

Address for correspondence and reprints: Austin L. Hughes, Department of Biology, Mueller Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802.

Mol. Biol. Evol. 9(3):381–393. 1992.

0737-4038/92/0903-0001$02.00

recombination and selection in generating and maintaining allelic polymorphism at this locus.

DNA Sequences Analyzed

MSA-1 (also known as major merozoite protein, p190, and p195) of *Plasmodium falciparum* is encoded by a single polymorphic locus (Tanabe et al. 1987). The gene consists of a single open reading frame, the length of which can vary considerably among alleles. Four complete and two partial DNA sequences of MSA-1 genes were obtained from the GenBank DNA sequence data base. The following allelic designations were used (GenBank accession number and reference are in parentheses): (1) M (complete) (X02919; Holder et al. 1985), (2) PA (complete) (X15063; Myler 1989), (3) K1 (complete) (X03371; Stunnenberg and Bujard 1985), (4) MAD (complete) (X05624; Tanabe et al. 1987), (5) CAM (partial) (X03831; Weber et al. 1986) and (6) RO (partial) (Y00087; Certa et al. 1987).

Tanabe et al. (1987) divided *P. falciparum* MSA-1 alleles into two groups on the basis of sequence similarity. (The division between these groups is most pronounced in an extensive region of the gene designated "region 6" below.) On the basis of their classification, M, PA, and K1 belong to group I; and MAD, CAM, and RO belong to group II. In the most polymorphic region of MSA-1 alleles, which includes a repeated tripeptide motif, RO along with certain other alleles partially amplified by the polymerase chain reaction (Kimura et al. 1990) differs markedly from both group I and group II. However, outside this region, RO tends to resemble group II alleles and is here classified with them to simplify presentation of data.

The deduced amino acid sequences of these alleles were aligned by Gotoh's (1987) method. When the alignment postulated a gap, the corresponding codon was excluded from all DNA sequence comparisons so that a comparable data set was used in all pairwise comparisons. On the basis of preliminary examination of the DNA sequences, each sequence was divided into 11 regions (fig. 1). These regions were identified by the following methods: (1) The signal peptide and the region of tripeptide repeats were identified on the basis of previous literature (Tanabe et al. 1987; Kimura et al. 1990). (2) Stephens's (1985) method for identifying clustered partitions of polymorphic sites was used to locate potential regions of interallelic recombination. (3) Preliminary comparisons of rates of nucleotide substitution in different regions were used to identify regions with unique characteristics at the level of DNA sequence. The regions were classified as 5′ regions (regions 1–7), for which sequence was available for all six alleles, and 3′ regions (regions 8–11), for which sequence was available for all group I alleles but only for MAD among group II alleles (fig. 1).

Results

Recombination in 5′ Regions

To understand the evolutionary relationships among MSA-1 alleles, I computed the number of synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions per site in each 5′ region for all pairwise comparisons among alleles, by using Nei and Gojobori's (1986) method. Mean d_S and mean d_N were computed for comparisons within and between the two allelic groups, and their standard errors (SEs) were estimated by Nei and Jin's (1989) method (table 1). Differences in the pattern of d_S and d_N between regions provide evidence of recombination. Because purifying selection can cause similarity at nonsynonymous sites, and because shared G+C content bias can cause similarity at synonymous sites (Wolfe et al. 1989), it is important to examine

Table 1
Mean $d_S \pm$ SE and Mean $d_N \pm$ SE, in Comparisons of 5′ Regions of MSA-1 Gene Alleles

	GROUP I (3)		GROUP II (3)		GROUP I VS. GROUP II (9)	
REGION	d_S	d_N	d_S	d_N	d_S	d_N
1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0†††	0.0 ± 0.0†††
2	1.5 ± 1.9	0.8 ± 0.4	6.1 ± 2.1††	3.7 ± 0.9††‡‡	7.5 ± 2.1†††	2.8 ± 0.6*†††
3	12.8 ± 5.1†††	50.9 ± 8.9***†††	30.4 ± 8.6†††	83.5 ± 10.7***†††‡	10.8 ± 3.9†††	56.5 ± 7.2***
4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.5 ± 0.5	0.0 ± 0.0†††	6.5 ± 2.1**†††
5	0.0 ± 0.0	0.8 ± 0.6	12.8 ± 5.1†	21.6 ± 3.3†††‡‡‡	8.5 ± 3.0†††	16.5 ± 2.7*†††
6	0.0 ± 0.0	0.1 ± 0.1	0.3 ± 0.2	0.8 ± 0.2‡‡	63.6 ± 6.0	46.0 ± 2.2**
7	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	14.2 ± 7.3†††	10.5 ± 2.9†††

NOTE.—Numbers in parentheses are number of comparisons. Tests of significance of difference between d_S and d_N: * = $P < 0.05$; ** = $P < 0.01$; and *** = $P < 0.001$. Tests of significance of difference between value of either d_S or d_N and corresponding value for region 6: † = $P < 0.05$; †† = $P < 0.01$; and ††† = $P < 0.001$. Tests of significance of difference between value of either d_S or d_N in comparisons among group II alleles and corresponding value among group I alleles: ‡ = $P < 0.05$; ‡‡ = $P < 0.01$; and ‡‡‡ = $P < 0.001$.

```
                    1 2                                                2 3
MKIIFFLCSFLFFIINTQC|VTHESYQELVKKLEALEDAVLTGYSLFQKEKMVL|NEGTSGTAVTTSTPGSSGSVT
                                                  3 2
SGGSVASVASVASGGSGGSVASGGSGNSRRTNPSDNS|SDSNTKTYADLKHRVQNYLFTIKELKYPELFDLTNHML
                                                                          2
TLSKNVDGFKYLIDGYEEINELLYKLNFYYDLLRAKLNDACANSYCQIPFNLKIRANELDVLKKIVFGYRKPLDN|
4
IKDNVGKMEDYIKKNKTTIANINELIEGSKKTIDQNKNADNEEGKKKLYQAQYNLFIYNKQLQEAHNLISVLEKRI
      4 5                                                5 6
DTLKKNEN|IKKLLEDIDKIKTDAENPTTGSKPNPLPENKKKEVEGHE|EKIKEIAKTIKFNIDSLFTDPLELEYY

LREKNKKVDVTPKSQDPTKSVQIPKVPYPNGIVYPLPLTDIHNSLAADNDKNSYGDLMNPDTKEKINEKIITDNKE

RKIFINNIKKQIDLEEKNINHTKEQNKKLLEDYEKSKKDYEELLEKFYEMKFNNNFDKDVVDKIFSARYTYNVEKQ

RYNNKFSSSNNSVYNVQKLKKALSYLEDYSLRKGISEKDFNHYYTLKTGLEADIKKLTEEIKSSENKILEKNFKGL

THSANASLEVSDIVKLQVQKVLLIKKIEDLRKIELFLKNAQLKDSIHVPNIYKPQNKPEPYYLIVLKKEVDKLKEF

IPKVKDMLKKEQAVLSSITQPLVAASETTEDGGHSTHTLSQSGETEVTEETEVTEETVGHTTTVTITLPPKEESAP

KEVKVVENSIEHKSNDNSQALTKTVYLKKLDEFLTKSYICHKYILVSNSSMDQKLLEVYNLTPEEEKELKSCDPLD

LLFNIQNNIPAMYSLYDSMNNDLQHLFFELYQKEMIYYLHKLKEENHIKKLLEEQKQITGTSSTSSPGNTTVNTAQ

SATHSNSQNQQSNASSTNTQNGVAVSSGPAVVEESHDPLTVLSISNDLKGIVSLLNLGNKTKVPNPLTISTTEMEK
                                                  6 7
FYENILKNNDTYFNDDIKQFVKSNSKVITGLTETQKNA|LNDEIKKLKDTLQLSFDLYNKYKLKLDRLFNKKKELG
                        7 8
QDKMQIKKLTLLKEQLESKLN|SLNNPHNVLQNFSVFFNKKKEAEIAETENTLENTKILLKHYKGLVKYYNGESSP
                                                                      8 9
LKTLSEVSIQTEDNYANLEKFRALSKIDGKLNDNLHLGKKKLSFLSSGLHHLITELKEVIKNKNYTGNSPS|ENNK

KVNEALKSYENFFPEAKVTTVVTPPQPDVTPSPLSVRVSGSSGSTKEETQIPTSGSLLTELQQVVQLQNYDEEDDS

LVVLPIFGESEDNDEYLDQVVTGEAISVTMDNILSGFENEYDVIYLKPLAGVYRSLKKQIEKNIITFNLNLNDILN

SRLKKRKYFLDVLESDLMQFKHISSNEYIIEDSFKLLNSEQKNILLKSYKYIKESVENDIKFAQEGISYYEKVLAK

YKDDLESIKKVIKEEKEKFPSSPPTTPPSPAKTDEQKKESKFLPFLTNIETLYNNLVNKIDDYLINLKAKINDCNV
                                                                  9 10
EKDEAHVKITKLSDLKAIDDKIDLFKNTNDFEAIKKLINDDTKKDMLGKLLSTGLVQ|IFPNTIISKLIEGKFQDM
                                                  10 11
LNISQHQCVKKQCPENSGCFRHLDEREECKCLLNYKQEGDKCEENPNP|TCNENNGGCDADATCTEEDSGSSRKKI

TCECTKPDSYPLFDGIFCSSSNFLGISFLLILMLILYSFI
```

FIG. 1.—Amino acid sequence of MSA-1 allele MAD, indicating regions analyzed separately. The 5′ regions are as follows (numbers of aligned codons are given in parentheses): 1 = the portion of the gene encoding the signal peptide (18 codons); 2 = the 5′ region of the gene surrounding the region of tripeptide repeats (147 codons); 3 = the tripeptide repeat region (49 codons); 4 = a short region in which group II alleles are identical to group II alleles at synonymous nucleotide sites but not at nonsynonymous nucleotide sites (see below) (54 codons); 5 = a region in which the group II allele CAM is identical to certain group I alleles (64 codons); 6 = an extensive region in which group I and group II alleles are highly divergent from each other and in which there is no evidence of recombination between the two groups of alleles (628 codons); and 7 = a region in which group I alleles are rather similar to group II alleles (58 codons). The 3′ regions (in which MAD is the only group II allele for which data are available) are as follows: 8 = a region in which, as in region 7, MAD is very similar to group I alleles (125 codons); 9 = a region in which MAD is very divergent from group I alleles (332 codons); 10 = a short region in which the group I allele PA is divergent from other group I alleles and resembles MAD (59 codons); and 11 = a region in which MAD is highly similar to group I alleles (74 codons).

both synonymous and nonsynonymous sites when testing for recombination (Hughes 1991*b*).

The results (table 1) reveal striking differences among the 5′ regions. For example, group I alleles are far more divergent from group II alleles at both synonymous and nonsynonymous sites in certain regions (especially region 6) than they are in others. Thus, relatively recent events of nonreciprocal recombination have apparently homogenized the two groups in certain regions. Regions 1 and 4 seem to have been homogenized by quite recent events, since the two groups are very similar in these regions. Regions 2, 5, and 7 seem to have been homogenized by somewhat more distant events. Still, groups I and II are significantly closer at both synonymous and nonsynonymous sites in these latter three regions than they are in region 6 (table 1).

In region 3, which contains the tripeptide repeats, the aligned codons are highly divergent both within and between the two groups of alleles (table 1). The explanation for this pattern appears to be a reciprocal recombination between groups I and II in this region. In a phylogenetic tree of region 3 sequences, the group II allele MAD clustered with the group I alleles M and PA, whereas the group I allele K1 clustered with the group II allele CAM (fig. 2A). At synonymous sites in aligned codons in this region, K1 is identical to CAM, and at nonsynonymous sites these two genes are only about as divergent from each other as are MAD and PA (table 2). On the other hand, in a phylogenetic tree based on region 6 (fig. 2B), group I alleles and group II alleles form sharply separated clusters.

In the phylogenetic tree based on *d* in region 3, RO does not cluster with either group I or group II alleles (fig. 2A). At nonsynonymous sites in particular, region 3

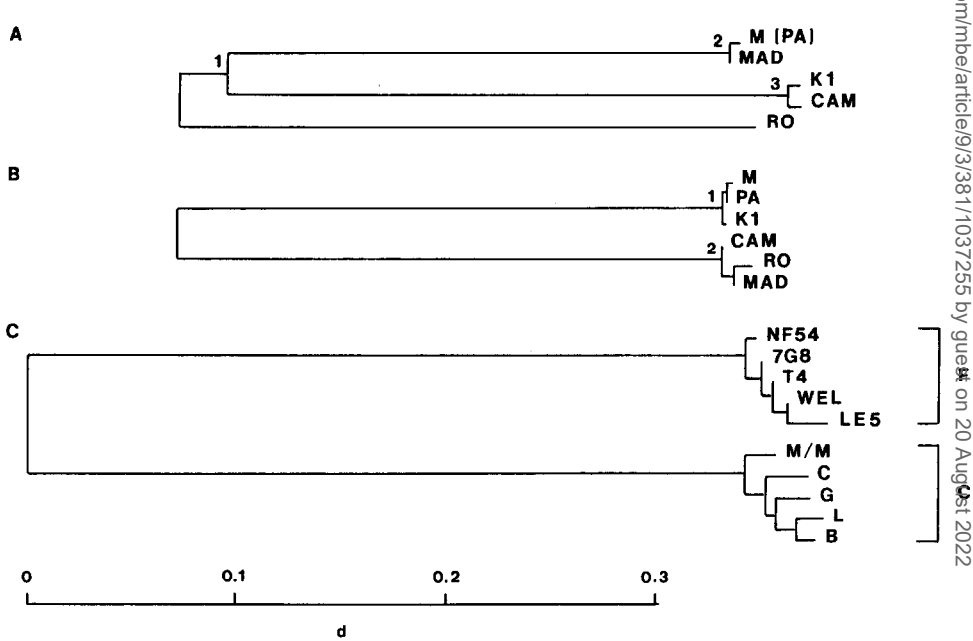

FIG. 2.—Phylogenetic trees constructed by neighbor-joining method (Saitou and Nei 1987), based on *d* (d) (Jukes and Cantor 1969). A, Tree for region 3 of MSA-1 alleles. The lengths of branches 1-2 (0.252 ± 0.063) and 1-3 (0.266 ± 0.065) are significantly different from zero, at the 0.1% level [method of Li (1989)]. B, Tree for region 6 of MSA-1 alleles. The length of branch 1-2 (0.522 ± 0.023) is significantly different from zero, at the 1% level. C, Tree for nonrepeat region (143 aligned codons) of CS protein gene alleles for two *Plasmodium* species: *P. falciparum* (F) and *P. cynomolgi* (C).

Table 2
Mean $d_S \pm$ SE (below Diagonal) and Mean $d_N \pm$ SE (above Diagonal), in Pairwise Comparisons of Region 3 of MSA-1 Gene Alleles

	M	PA	K1	MAD	CAM	RO
M		0.0 ± 0.0	76.4 ± 13.4***	3.9 ± 2.0†	76.4 ± 13.4***†	89.3 ± 15.9***
PA	0.0 ± 0.0		76.4 ± 13.4***	3.9 ± 2.0†	76.4 ± 13.4***†	89.3 ± 15.9***
K1	19.2 ± 7.6	19.2 ± 7.6		76.4 ± 13.4***	1.9 ± 1.3	90.7 ± 16.0
MAD	4.8 ± 3.4	4.8 ± 3.4	13.0 ± 6.1		76.4 ± 13.4***†	83.3 ± 14.7***
CAM	19.2 ± 7.6	19.2 ± 7.6	0.0 ± 0.0	13.1 ± 6.1		90.7 ± 16.0
RO	14.5 ± 6.5	14.5 ± 6.5	6.9 ± 4.4	8.8 ± 4.9	6.9 ± 4.4	

NOTE.—Tests of significance of difference between d_S and d_N: * = $P < 0.05$; ** = $P < 0.01$; and *** = $P < 0.001$. Tests of significance of difference between value of d_N and value of d_N, for comparison with K1 in same column: † = $P < 0.001$.

of RO is divergent from that of other alleles (table 2). On the other hand, at synonymous sites in the same region, RO resembles group I alleles more closely than it does group II alleles. In region 6, RO is clearly a member of group II (fig. 2B). Thus, region 3 of RO may have been donated by a group I allele and then may have diverged from other group I alleles at nonsynonymous sites, as a result of positive selection. Alternatively, region 3 of RO may have been donated by a member of yet another allelic lineage, a complete member of which has so far not been identified (for discussion of the possibility that such a lineage exists, see Kimura et al. 1990).

In region 5, the group II allele CAM is identical to the group I alleles M and PA but quite different from the other group II alleles (mean $d_S \pm$ SE = 12.8 ± 4.4 substitutions/100 sites; mean $d_N = 24.6 \pm 4.0$ substitutions/100 sites). Thus, this region of CAM seems to have recently been donated by either M or PA or an allele closely related to them.

In addition to recombination events which have homogenized certain regions between two groups of alleles, there is evidence that in some regions additional events have homogenized the group II alleles with one another. In regions 2 and 5, group II alleles are more divergent from each other at both synonymous and nonsynonymous sites than are group I alleles from each other (table 1). In regions 1, 4, and 7, on the other hand, group II alleles are just as similar to each other as group I alleles are to each other (table 1). This suggests that events of nonreciprocal recombination have homogenized group II alleles in the latter three regions.

Recombination in 3′ Regions

Table 3 shows d_S and d_N among group I alleles and between group I alleles and MAD in the 3′ regions. In regions 8 and 11, group I alleles and MAD seem to have been homogenized recently. In region 10, PA and MAD are comparatively similar (mean $d_S \pm$ SE = 9.7 ± 5.7 substitutions/100 sites; mean $d_N \pm$ SE = 9.6 ± 2.7 substitutions/100 sites), as are M and K1 (mean $d_S \pm$ SE = 2.9 ± 2.9 substitutions/100 sites; mean $d_N \pm$ SE = 1.4 ± 1.0 substitutions/100 sites). However, in this same region, PA and MAD are very different from M and K1 (mean $d_S \pm$ SE = 55.1 ± 16.6 substitutions/100 sites; mean $d_N \pm$ SE = 39.3 ± 5.6 substitutions/100 sites). This region of PA then appears to have been donated by a sequence similar to MAD, perhaps by another group II allele.

Table 3
Mean $d_S \pm$ SE and Mean $d_N \pm$ SE, in Comparisons among 3′ Regions of MSA-1 Gene Group I Alleles and MAD (Group II)

	GROUP I (3)		GROUP I VS. MAD (3)	
REGION	d_S	d_N	d_S	d_N
8	2.7 ± 1.6	0.2 ± 0.2	5.0 ± 2.0†††	5.8 ± 1.4†††
9	0.6 ± 0.5	0.3 ± 0.2	53.2 ± 6.8	37.9 ± 2.7*
10	33.5 ± 10.5	21.1 ± 3.7†††	44.2 ± 11.9	28.3 ± 4.2
11	2.5 ± 1.9	2.9 ± 1.0†	2.3 ± 1.6†††	4.1 ± 1.2†††

NOTE.—Numbers in parentheses are number of comparisons. Tests of significance of difference between d_S and d_N: * = $P < 0.05$. Tests of significance of difference between value of either d_S or d_N and corresponding value for region 9: † = $P < 0.05$; and ††† = $P < 0.001$.

Natural Selection on 5′ Regions

Under purifying selection, which operates on most protein-coding genes, d_S is expected to exceed d_N, whereas, under complete neutrality, as seen in pseudogenes, d_S and d_N should be approximately equal (Nei 1987, pp. 79–86). On the other hand, positive selection favoring diversity at the amino acid level will cause d_N to exceed d_S (Hughes and Nei 1988). Two regions of MSA-1 genes show evidence of positive selection. In region 3, d_N is significantly higher than d_S, both in comparisons between M and PA, on the other hand, and in comparisons between K1, CAM, and RO, on the other (table 2). In region 3, alleles may also differ in the number of repeat segments; CAM has a much greater number of repeats than do other alleles (fig. 3A).

In region 4, d_N is significantly higher than d_S in the comparison between group I and group II alleles (table 1). A similar trend is seen in region 5. In the latter region, it is of interest that, when CAM (which apparently has recently received region 5 from either M or PA) is compared with other group II alleles, d_N significantly exceeds d_S ($P < 0.05$; values of d_S and d_N are given above).

Since a G+C-content bias at third-codon positions can lower d_S (Wolfe et al. 1989), I examined third-position G+C content in the different regions of MSA-1 genes (table 4). As is typical with genes of *Plasmodium* (Weber 1988), MSA-1 genes are AT rich in all regions. This AT richness is particularly remarkable in region 3 of K1 and in the group II alleles (table 4). In M and PA, on the other hand, third-position G+C content is not significantly lower in region 3 than in other 5′ regions (table 5). The AT richness in region 3 has probably lowered d_S in certain comparisons in this region. Given that d_N between groups I and II is approximately the same in region 3 as in region 6 (table 1), it might be proposed that regions 3 and 6 of the two groups of alleles diverged at about the same time. On this hypothesis, the explanation for the much lower d_S between groups I and II in region 3, in comparison with that in region 6 (table 1), would be that the AT richness of region 3 has lowered d_S there. However, it seems unlikely that a G+C-content difference of 6%–10% can account for a sixfold difference in d_S. Thus, the hypothesis that d_N has been increased in region 3 as a result of positive selection seems more plausible. In the case of regions 4 and 5, there is no unusual G+C content bias (table 4); thus, the difference between d_N and d_S in this region is most easily explained as a result of positive selection favoring diversification between the two allelic groups.

In contrast to regions 3–5, regions 2 and 6 show some evidence of purifying

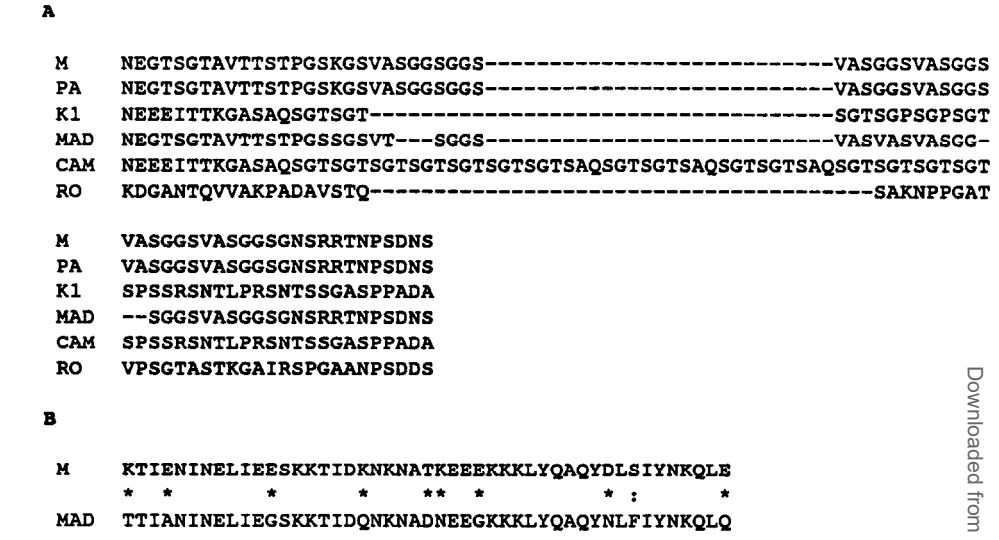

FIG. 3.—MSA-1 regions showing evidence of positive selection. A, Region 3 (tripeptide repeat region). B, Comparison of most variable portion of region 4 from group I allele (M) and group II allele (MAD): an asterisk (*) denotes amino acid replacement involving residue charge change, and a colon (:) denotes amino acid replacement not involving charge change.

selection in the comparison between the two allelic groups (table 1). Thus, in these regions there appears to be some constraint on the amino acid sequence of the MSA-1 protein. Region 7, on the other hand, appears to be evolving with little or no constraint, while in region 1 the two allelic groups have been homogenized too recently for any nucleotide differences to have occurred.

I applied the method of Hughes et al. (1990) to test whether particular kinds of amino acid replacements are favored in regions 3–5. In brief, this method divides nonsynonymous nucleotide sites into conservative and radical sites (or fractions of sites) on the basis of whether a nucleotide substitution at the site is conservative or radical (nonconservative) with respect to an amino acid property of interest; and the proportion of conservative nonsynonymous differences per site (p_{NC}) and the proportion radical nonsynonymous substitutions per site (p_{NR}) are estimated. By this method, it has been found that there is a statistically significant bias toward charge change both in binding regions of class I MHC molecules and in epitopes of *P. falciparum* CS protein that are bound by MHC molecules; in other words $p_{NR} > p_{NC}$ in these regions. In the case of MSA-1, a similar bias toward charge change was found in region 4 but not in regions 3 or 5 (table 5 and fig. 3B). When other amino acid properties (polarity, hydrophobicity, and volume) were tested, no significant bias was seen in any region.

Age of MSA-1 Polymorphism

The absence of a fossil record makes it difficult to estimate the age of *P. falciparum* MSA-1 alleles. MSA-1 sequences are available for two species of *Plasmodium* parasitic on rodents, *P. yoelii* (Lewis 1989) and *P. chabaudi* (Deleersnijder et al. 1990). Similarity between these genes and those of *P. falciparum* is low, but Lewis (1989) has identified two conserved portions of region 6. Table 6 shows d_S and d_N in this region, both between the *P. yoelii* and *P. chabaudi* sequences and between each of them and

Table 4
Percent G+C at Third Codon Position in MSA-1 Gene Regions

	G+C AT THIRD CODON OF ALLELE (%)					
REGION	M	PA	K1	MAD	CAM	RO
1	22.2	22.2	22.2	22.2	22.2	22.2
2	13.6	13.6	15.0	15.6	16.3	14.3
3	8.2	8.2	4.1**	2.0***	4.1**	4.1**
4	12.8	12.8	12.8	12.8	12.8	12.8
5	22.7	22.7	22.7	21.2	22.7	24.2
6	13.7	13.7	13.7	12.4	12.3	12.6
7	6.9	6.9	6.9	10.3	10.3	10.3
8	11.2	10.4	11.2	8.8	...	...
9	14.8	14.5	14.8	17.5	...	...
10	20.3	15.3	18.6	11.9	...	...
11	25.6*	24.3*	24.3*	24.3*	...	...

NOTE.—Test of significance of difference between percent G+C and that in region 6 (binomial test): * = $P < 0.05$; ** = $P < 0.01$; *** = $P < .001$.

the two groups of *P. falciparum* alleles. On the assumption that *P. falciparum* diverged from the two rodent malarias when rodents and primates diverged ~80 Mya, comparison of d_N values provides an estimate of ~35 Mya for the divergence, in region 6, between the two groups of *P. falciparum* MSA-1 alleles. This would correspond to a rate of 2.6×10^{-9} nonsynonymous substitutions per site per year. In this case, d_S probably cannot provide an accurate estimate of divergence time, since d_S values are quite high; indeed, d_S between the two groups of *P. falciparum* alleles is nearly as high as that between *P. falciparum* and the two rodent malarias (table 6). In region 6, the two allelic groups are about as divergent at both synonymous and nonsynonymous sites as are nonrepeat regions of CS protein genes from *P. falciparum* and *P. cynomolgi* (fig. 2). An estimate of 35 Mya for the divergence time of the two groups would be consistent with this result if *P. falciparum* and *P. cynomolgi* can be assumed to have diverged when hominids diverged from Old-World monkeys.

In region 3, MSA-1 alleles are about as divergent at nonsynonymous sites as they are in region 6; however, since selection has apparently favored diversification, nonsynonymous sites have presumably evolved much more rapidly than in region 6. If in region 6 the two groups of alleles diverged 35 Mya, then the ratio of d_S values between regions 3 and 6 suggests that in region 3 the most divergent pairs of alleles may have diverged only ~10 Mya. By similar reasoning, the two groups of alleles may have diverged 5–10 Mya. in regions 2, 5, and 7. In region 4, on the other hand, the groups of alleles have diverged only very recently.

Discussion

Known *Plasmodium falciparum* MSA-1 alleles show the effects of repeated recombination events of two kinds: (1) nonreciprocal recombination events that have homogenized alleles in the regions involved and (2) reciprocal recombinations that have recombined highly divergent domains, thereby enhancing allelic diversity. It is possible that similarity among alleles is favored in certain regions, as a result of functional constraints. Other regions may be relatively free to vary, and in some cases

Table 5
Mean $p_{NC} \pm$ SE and Mean $p_{NR} \pm$ SE, with Respect to Amino Acid Residue Charge in 5′ Regions of MSA-1 Gene Alleles

	GROUP I (3)		GROUP II (3)		GROUP I VS. GROUP II (9)	
REGION	p_{NC}	p_{NR}	p_{NC}	p_{NR}	p_{NC}	p_{NR}
3	32.2 ± 4.0	31.5 ± 5.6	54.9 ± 5.5	41.8 ± 7.3	37.2 ± 3.8	29.5 ± 5.0
4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.7 ± 1.7	9.8 ± 3.4*
5	0.8 ± 0.8	0.9 ± 0.9	18.7 ± 3.6	18.4 ± 3.7	14.3 ± 2.9	13.7 ± 2.8

NOTE.—Numbers in parentheses are number of comparisons. Test of significance of difference between p_{NC} and p_{NR}: * = $p < 0.05$.

divergence between alleles may actually be selectively favored as a means of evading the host immune system.

Although it is generally accepted that *P. falciparum* undergoes an obligate sexual stage in the mosquito host, Tibayrenc et al. (1991) argue that some population genetic evidence supports the hypothesis of a clonal population structure in this species. This would require a yet undiscovered asexual mode of reproduction. The evidence of repeated recombination at the MSA-1 locus certainly implies sexual reproduction. However, the data are not incompatible with the hypothesis that sexual reproduction occurs relatively infrequently. If recombinant MSA-1 alleles have been selectively favored, then they would increase in frequency even if rare. An analogous process may occur in bacteria. In bacteria, loci involved in interaction with the host immune system show much greater evidence of recombination (Kroll and Moxon 1990; Smith et al. 1990) than do housekeeping enzyme loci (Nelson et al. 1991); the discrepancy can be explained by selection favoring recombinant types at the former loci.

The three regions of MSA-1 genes showing evidence of positive selection (regions 3–5) may be subject to balancing selection arising from contrasting aspects of the host immune system. Region 3 has a repeat structure, as do many immunogenic regions of *Plasmodium* proteins (Kemp et al. 1987). Experimentally a simple amino acid polymer elicits a T cell–independent antibody response. Enea and Arnot (1988) argue that the role of repeat regions in parasite antigens is to elicit just such a response, which is relatively inefficient in eliminating the infection and does not confer lasting immunity. The repeat region of the CS protein gene of *P. cynomolgi* also shows some evidence of positive selection favoring diversification among alleles (Hughes 1991*a*). In the case of MSA-1 the evidence for positive selection is even stronger, suggesting that some form of balancing selection favors allelic diversity in this repeat region.

Region 4, in contrast to region 3, has apparently been recently homogenized between the two allelic groups. Nonetheless, positive selection is acting in this region to promote diversity between the two groups. Region 4 has characteristics suggesting that it may contain peptides bound by MHC molecules and presented to T cells. The fact that selection on this region is acting to diversity charge profile (pattern of residue charges) is reminiscent of the pattern seen in T cell epitopes of CS proteins (Hughes 1991*a*) and presumably represents an adaptation to evade recognition by MHC molecules, which themselves are characterized by different patterns of residue charge in the antigen-binding region (Hughes et al. 1990). In the case of region 5, on the other hand, since obvious distinguishing marks of an epitope for either antibody or T cells

Table 6
Mean $d_S \pm$ SE (Above Diagonal) and Mean $d_N \pm$ SE (below Diagonal) in Conserved Portion of MSA-1 Gene, in Comparisons Among Genes from Two Rodent Malarias and Two Major Groups of *Plasmodium falciparum* Alleles

			P. falciparum	
	P. yoelii	*P. chabaudi*	Group I	Group II
P. yoelii		20.2 ± 4.6	76.7 ± 13.6	63.7 ± 11.2
P. chabaudi	5.3 ± 1.2		85.8 ± 15.4	75.1 ± 13.1
P. falciparum:				
Group I	44.1 ± 4.1	45.7 ± 4.2		68.1 ± 12.2
Group II	36.8 ± 3.6	39.2 ± 3.7	18.1 ± 2.2	

NOTE.—Number of codons compared is 179.

are not apparent, it is so far not possible to speculate regarding the source of the apparent positive selection.

An allelic polymorphism under balancing selection can be maintained for much longer than can a neutral polymorphism (Takahata and Nei 1990). In the case of the mammalian MHC, certain polymorphisms seem to have been maintained for >5 Myr and possibly for as long as 20–40 Myr (Mayer et al. 1988; Gyllensten and Ehrlich 1989; Gyllensten et al. 1990). If the MSA-1 allelic polymorphism has been maintained 35 Myr, then it is one of the oldest polymorphisms so far known in any eukaryote. Because balancing selection is required to maintain a polymorphism for such a long time, positive selection on MSA-1 must considerably predate its encounter with the human immune system as it currently exists. Thus, this selection must have arisen as a result of features general to the vertebrate (or at least the mammalian) immune system and not specific to humans.

Acknowledgment

This research was supported by a grant from the National Institutes of Health.

LITERATURE CITED

CAPRON, A., and J. P. DESSAINT. 1989. Molecular basis of host-parasite relationship: towards the definition of protective antigens. Immunol. Rev. **112**:27–48.

CERTA, U., D. ROTMANN, H. MATILE, and R. REBER-LISKE. 1987. A naturally occurring gene encoding the major surface antigen precursor P190 of *Plasmodium falciparum* lacks tripeptide repeats. EMBO J. **6**:4137–4142.

DELEERSNIJDER, W., D. HENDRIX, N. BENDAHMAN, J. HANEGREEFS, L. BRUS, C. HAMERS-CASTERMAN, and R. HAMERS. 1990. Molecular cloning and sequence analysis of the gene encoding the major merozoite surface antigen of *Plasmodium chabaudi chabaudi* IP-PC1. Mol. Biochem. Parasitol. **43**:231–244.

ENEA, V., and D. ARNOT. 1988. The circumsporozoite gene in Plasmodia. Pp. 5–11 *in* M. J. TURNER and D. ARNOT, eds. Molecular genetics of parasitic protozoa. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

GOTOH, O. 1987. Pattern matching of biological sequences with limited storage. Cabios **3**:17–20.

GYLLENSTEN, U. B., and H. A. ERLICH. 1989. Ancient roots for polymorphism at the HLA-DQα locus in primates. Proc. Natl. Acad. Sci. USA **86**:9986–9990.

GYLLENSTEN, U. B., D. LASHKARI, and H. A. ERLICH. 1990. Allelic diversification at the class II DQB locus of the mammalian major histocompatibility complex. Proc. Natl. Acad. Sci. USA **87**:1835–1839.

HOLDER, A. A., M. J. LOCKYER, K. G. ODINK, J. S. SAUDHU, V. RIVEROS-MORENO, S. C. NICHOLLS, Y. HILLMAN, L. S. DAVEY, M. L. V. TIZARD, R. T. SCHWARZ, and R. R. FREEMAN. 1985. Primary structure of the precursor to the three major surface antigens of *Plasmodium falciparum* merozoites. Nature **317**:270–273.

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

———. 1991*b*. Testing for interlocus genetic exchange in the MHC: a reply to Andersson and co-workers. Immunogenetics **33**:243–246.

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., T. OTA, and M. NEI. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol. Biol. Evol. **7**:515–524.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KEMP, D. J., R. L. COPPEL, and R. F. ANDERS. 1987. Repetitive proteins and genes of malaria. Annu. Rev. Microbiol. **41**:181–208.

KIMURA, E., D. MATTEI, S. MANA DE SANTI, and A. SCHERF. 1990. Genetic diversity in the major merozoite surface antigen of *Plasmodium falciparum:* high prevalence of a third polymorphic form detected in strains derived from malaria patients. Gene **91**:57–62.

KROLL, J. S., and E. R. MOXON. 1990. Capsulation in distantly related strains of *Haemophilus influenzae* type b: genetic drift and gene transfer at the capsulation loci. J. Bacteriol. **172**: 1374–1379.

LEWIS, A. P. 1989. Cloning and analysis of the gene encoding the 230-kilodalton merozoite surface antigen of *Plasmodium yoelii.* Mol. Biochem. Parasitol. **36**:271–287.

LI, W.-H. 1989. A statistical test of phylogenies estimated from sequence data. Mol. Biol. Evol. **6**:424–435.

MAHMOUND, A. A. F. 1989. Parasitic protozoa and helminths: biological and immunological challenges. Science **246**:1015–1022.

MAYER, W. E., M. JONKER, D. KLEIN, P. IVANYI, G. VAN SEVENTER, and J. KLEIN. 1988. Nucleotide sequences of chimpanzee MHC class I alleles: evidence for *trans*-species mode of evolution. EMBO J. **7**:2765–2774.

MYLER, P. J. 1989. Nucleotide and deduced amino acid sequence of the gp195 (MSA-1) gene from *Plasmodium falciparum* palo alto PLF-3/B11. Nucleic Acids Res. **17**:5401.

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.

NEI, M., and L. JIN. 1989. Variances of the average numbers of nucleotide substitutions within and between populations. Mol. Biol. Evol. **6**:290–300.

NELSON, K., T. S. WHITTAM, and R. K. SELANDER. 1991. Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gap*A) in natural populations of *Salmonella* and *Escherichia coli.* Proc. Natl. Acad. Sci. USA **88**:6667–6671.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SHER, A. 1988. Vaccination against parasites: special problems imposed by the adaptation of parasitic organisms to the host immune response. Pp. 169–182 *in* P. T. ENGLUND and A. SHER, eds. The biology of parasitism. Alan R. Liss, New York.

SMITH, N. H., P. BELTRAN, and R. K. SELANDER. 1990. Recombination of *Salmonella* phase 1 flagellin genes generates new serovars. J. Bacteriol. **172**:2209–2216.

STEPHENS, J. C. 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2**:539–556.

STUNNENBERG, H., and H. BUJARD. 1985. Polymorphism of the precursor for the major surface antigens of *Plasmodium falciparum* merozoites: studies at the genetic level. EMBO J. **4**: 3823–3829.

TAKAHATA, N., and M. NEI. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics **124**:967–978.

TANABE, K., M. MACKAY, M. GOMAN, and J. G. SCAIFE. 1987. Allelic dimorphism in a surface antigen of the malaria parasite *Plasmodium falciparum.* J. Mol. Biol. **195**:273–287.

TIBAYRENC, M., F. KJELLBERG, J. ARNAUD, B. OURY, S. F. BRENIÈRE, M.-L. DARDÉ, and F. J. AYALA. 1991. Are eukaryotic microorganisms clonal or sexual? a population genetics vantage. Proc. Natl. Acad. Sci. USA **88**:5129–5133.

WEBER, J. L. 1988. Molecular biology of malaria parasites. Exp. Parasitol. **66**:143–170.

WEBER, J. L., W. M. LEININGER, and J. A. LYON. 1986. Variation in the gene encoding a major merozoite surface antigen of the human malaria parasite *Plasmodium falciparum.* Nucleic Acids Res. **14**:3311–3323.

WOLFE, K. H., P. M. SHARP, and W.-H. LI. 1989. Mutation rates vary among regions of the mammalian genome. Nature **337**:283–285.

JAN KLEIN, reviewing editor

Received August 30, 1991; revision received September 24, 1991

Accepted October 24, 1991