Strong Diversifying Selection on Domains of the *Plasmodium falciparum* Apical Membrane Antigen 1 Gene

Spencer D. Polley and David J. Conway

Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom

Manuscript received January 31, 2001
Accepted for publication May 14, 2001

ABSTRACT

The surface-accessible ectodomain region of the *Plasmodium falciparum* apical membrane antigen 1 (AMA1) is a malaria vaccine candidate. The amino acid sequence may be under selection from naturally acquired immune responses, and previous analyses with a small number of allele sequences indicate a non-neutral pattern of nucleotide variation. To investigate whether there is selection to maintain polymorphism within a population, and to identify the parts of the ectodomain under strongest selection, a sample of 51 alleles from a single endemic population was studied. Analyses using Fu and Li's *D* and *F* tests, Tajima's *D* test, and the McDonald-Kreitman test (with the chimpanzee parasite *P. reichenowi* as outgroup) show significant departure from neutrality and indicate the selective maintenance of alleles within the population. There is also evidence of a very high recombination rate throughout the sequence, as estimated by the recombination parameter, *C*, and by the rapid decline in linkage disequilibrium with increasing nucleotide distance. Of the three domains (I–III) encoding structures determined by disulfide bonds, the evidence of selection is strongest for Domains I and III. We predict that these domains in particular are targets of naturally acquired protective immune responses in humans.

A major factor in the evolution of parasites is likely to be the evasion of host immune responses. Where such responses have a memory component, as in vertebrates, rare pathogen alleles encoding antigenic types infrequently recognized by the host will have a selective, frequency-dependent advantage (Brunham *et al.* 1993). Patterns of nucleotide polymorphism within antigen genes under such selection should statistically depart from those predicted by neutral evolution models (Tajima 1989a; Fu and Li 1993). The ability to identify regions of antigen genes that are under such selection may help in finding targets of naturally acquired immunity in complex pathogens, such as the malaria parasite *Plasmodium falciparum* (Conway 1997). A study on a major blood stage antigen gene (*msp1*) in *P. falciparum* identified a particular region likely to be under balancing selection, and a clear predictive association between the presence of antibodies to this region of the protein and protection from clinical malaria was subsequently shown (Conway *et al.* 2000a).

There is evidence that positive selection may be operating on several other antigen genes of *P. falciparum* (Hughes and Hughes 1995; Conway 1997; Escalante *et al.* 1998). These include the apical membrane antigen 1 (*AMA1*) gene, which encodes a membrane-bound protein on the erythrocyte-invading merozoite stage. Experimental studies in primates (infected with *P. fragile* or *P. knowlesi*) and rodents (infected with *P. chabaudi*) have shown that immunization with AMA1 protects against challenge infection (Deans *et al.* 1988; Collins *et al.* 1994; and Crewther *et al.* 1996). This protection was parasite strain-dependent in the mouse *P. chabaudi* system, suggesting that protective immune responses may be directed to polymorphic epitopes of AMA1 (Crewther *et al.* 1996). Preliminary studies have shown the existence of antibodies, and responsive T cells, to *P. falciparum* AMA1 in humans living in endemic African populations (Thomas *et al.* 1994; Lal *et al.* 1996; Riley *et al.* 2000), but protective epitopes have not yet been demonstrated.

Previous sequence analysis of the *AMA1* gene in *P. falciparum* has revealed an excessive level of nonsynonymous *vs.* synonymous polymorphism, and this suggests that the locus is under diversifying selection, even though part of the effect is due to codon bias in this A + T rich parasite (Hughes and Hughes 1995; Escalante *et al.* 1998; Verra and Hughes 2000). A recent comparison with the sequence from *P. reichenowi* (a closely related parasite of chimpanzees), applying the McDonald-Kreitman test, demonstrates a non-neutral excess of nonsynonymous polymorphic sites in *P. falciparum* (Kocken *et al.* 2000). This effect is apparently on the region of the gene encoding the mature protein ectodomain (rather than the N-terminal signal and prosequence, which is cleaved). Analyses that require a large sample of sequences from a single population (Tajima 1989a,b; Fu and Li 1993) have not been per-

Corresponding author: David J. Conway, Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel St., London WC1E 7HT, United Kingdom.
E-mail: david.conway@lshtm.ac.uk

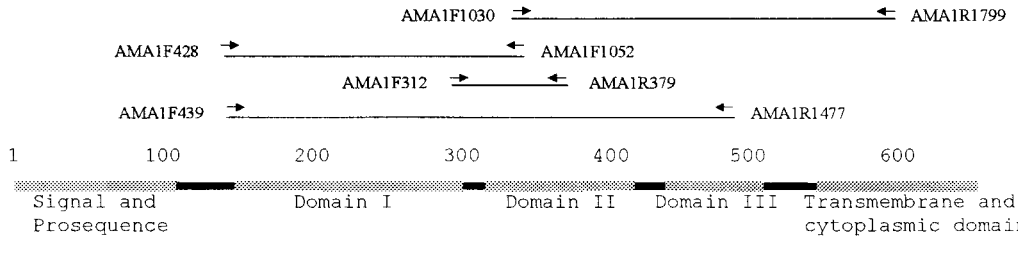

FIGURE 1.—Sequencing strategy for the *AMA1* gene and the positions of primers used to amplify products. PCR products are shown above the coding region of *AMA1* as solid lines, while primers are represented by an arrow. The codon numbers for the open reading frame (shown by the bottom scheme) are taken from the HB3 sequence (EMBL reference U33277), with the domains of *AMA1* as described by HODDER *et al.* (1996) shown as shaded boxes.

formed effectively due to the small number and diverse origin of sequences available.

Here, 51 *AMA1* ectodomain gene sequences are sampled from one endemic African population. Comprehensive analyses with several independent tests show evidence of positive selection and recombination across the entire sequence. The signature of selection is seen most strongly in two regions previously identified as encoding domains defined by several disulfide bridges (Domains I and III). These new data provide very strong evidence for diversifying selection on AMA1, and the specific distribution of this effect in regions of the antigen is of relevance to vaccine research.

MATERIALS AND METHODS

Sample selection and DNA extractions: Blood samples from subjects in Ibadan, South-West Nigeria, who were naturally infected with *P. falciparum*, had been previously prepared for parasite DNA analysis (CONWAY *et al.* 1999, 2000a,b). The *P. falciparum AMA1* gene was amplified from 49 of these samples that had each appeared to contain only a single clone infection as determined by typing of the polymorphic *msp1* gene (CONWAY *et al.* 1999, 2000a).

***AMA1* amplification and sequencing:** Malaria parasites in the human host are haploid, and thus each contains only one allele of the single locus *AMA1* gene. Codons 143–598, encompassing the region of the *AMA1* gene encoding the mature protein ectodomain, were sequenced in each sample. Three overlapping templates encoding codons 143–350, codons 312–379, and codons 344–598 were separately amplified by PCR using primers *AMA1*F143 (5′-GACTTCCATCAGGGA AATGTCC-3′) and *AMA1*R350 (5′-TTAGGTTGATCCGAAGC ACTCAA-3′); *AMA1*F312 (5′-CGGATTATGGGTCGATGGAAA TTG-3′) and *AMA1*R379 (5′-CTGCTTTAAAAGCACCAGTGGG AAG-3′ and *AMA1*F344 (5′-TTGAGTGCTTCGGATCAACC TAA-3′) and *AMA1*R598 (5′-GCCTCAGGATCTAACATTTCA TC-3′), respectively, using 40 rounds of amplification (codon numbers taken from HODDER *et al.* 1996). Figure 1 shows a representation of the regions of the *AMA1* gene sequenced and the primers used. The resulting 624-bp, 202-bp, and 769-bp amplification products were agarose gel purified using the Gel Extraction spin kit (QIAGEN, Crawley, UK), as per manufacturer's protocol for preparation of templates for direct sequencing, and sequenced using big dye (Applied Biosystems, Warrington, UK) termination chemistry in both forward and reverse directions. The reactions were run on an ABI prism 377 automated sequencer and forward and reverse reactions from each allele were aligned and analyzed to ensure sequence fidelity using the Sequence Navigator software (Applied Biosystems). The three fragments from each allele were then aligned against a reference sequence (EMBL reference U33277) to create a single contiguous sequence. For two isolates, which were each found to contain two different parasite alleles, the PCR products were cloned into pGEM-T vector using the TA cloning kit (Promega, Madison, WI) and five random clones were sequenced from each. From these two isolates, a fourth region (1038 bp, codons 147–492) was also amplified from the original genomic DNA using primers *AMA1*F146 (5′GGA AATGTCCAGTATTTGGTAAAGG3′) and *AMA1*R492 (5′CAC ATGGGCATTTTAAACTGTC3′) and cloned and sequenced in forward and reverse orientations, to assemble the allelic sequences into distinct haplotypes. Any singleton polymorphisms that occurred in the data set were confirmed by independent amplification of the template from the original genomic DNA and resequencing to ensure that they were not errors.

Sequence alignment and analysis: A single contiguous sequence of 1311 nucleotides (excluding outer primers) was derived for each of the 51 *AMA1* alleles in the sample (one from each of 47 isolates and two from each of the remaining 2 isolates). These were aligned using the CLUSTAL program in MEGALIGN (DNAStar) and exported as a NEXUS alignment for statistical analyses using the DnaSP3.52 software (ROZAS and ROZAS 1999). Analyses of recombination and tests of neutrality were as follows.

Analysis of recombination and linkage disequilibrium: Analysis of recombination in *AMA1* was performed on the alignment of 51 sequences to calculate the minimum number of recombination events that have occurred throughout the sequence (HUDSON and KAPLAN 1985) and to estimate the recombination parameter, C, equal to $4Nr$ where N is the effective population size and r is the probability of recombination between adjacent nucleotides per generation (HUDSON 1987). The D' (LEWONTIN 1964) and R^2 (HILL and ROBERTSON 1968) indices of linkage disequilibrium were considered quantitatively between sites at which rare alleles had a frequency of >0.1, and the relationship between linkage disequilibrium and distance between nucleotide sites was plotted.

Tajima's *D* test: This tests for a departure from the neutral evolution model by comparing the estimations of nucleotide diversity (θ), calculated alternatively from π (average pairwise nucleotide diversity) and the total number of mutations. Under a panmictic constant-size neutral model, the expectations of π and θ are the same, but under balancing selection rare alleles have an advantage and are thus maintained at intermediate frequencies, yielding a positive value of D (TAJIMA 1989a). The analysis was performed on each of the three domains of the ectodomain of *AMA1*, in addition to the whole sequence, which was also analyzed using a sliding window approach.

Fu and Li's *D* and *F* tests: In these tests, departures from neutrality are identified as a deviation between estimates of θ, derived from the number of mutations in external branches of the phylogeny, and from the total number of mutations

(giving the index *D*) or from the the average pairwise diversity (giving the index *F*). A deficit of mutations in external branches results in positive values of *D* and *F*, indicative of the presence of unusually ancient alleles that are probably maintained by balancing selection (Fu and Li 1993). The *AMA1* sequence of the closely related species *P. reichenowi* (accession no. AJ252087; Kocken *et al.* 2000) was used as an outgroup for these analyses. The analysis was performed on each of the three domains (I–III) and on the whole sequence, which was also analyzed using a sliding window approach.

Coalescent simulations: Tajima's and Fu and Li's tests have been predicted to give conservative estimates of the departure from neutrality when recombination has occurred between alleles (Tajima 1989a; Wall 1999). Therefore, critical values of Tajima's and Fu and Li's indices under neutrality were also calculated, assuming differing levels of recombination using coalescent simulations (Hudson 1990), and compared to the actual values observed with the data set. The 95% confidence intervals of the neutral distributions were calculated using 10,000 coalescent simulations in DnaSP3.52, and statistical significance was inferred where the observed value lay outside these ($P < 0.05$). In these cases, the 99% confidence intervals were also calculated to determine whether the data were also significant at a higher level ($P < 0.01$).

The McDonald-Kreitman test: This tests for different ratios of synonymous to nonsynonymous changes within and between species (McDonald and Kreitman 1991). The *P. reichenowi AMA1* sequence was used for the interspecific comparison with *P. falciparum* (Kocken *et al.* 2000). The analysis was performed on each of the three domains (I–III) of the ectodomain of *AMA1*, in addition to the whole ectodomain sequence. Codons with a complex evolutionary history that were ignored by the computer analysis were analyzed by eye according to the guidelines given in DNAsp3.5. Fisher's exact test was then applied to the data to test for significant nonrandomness, and skew from randomness was also calculated as the "neutrality index" (Rand and Kann 1996).

RESULTS

Sequence diversity: The sequences of 51 alleles of *AMA1* sampled from the Nigerian population were determined and aligned for analysis. Sequences are available as an alignment (EMBL accession no. ALIGN_000019; individual sequence accession nos. AJ408300–50). Within the 1311-bp region sequenced there were 62 polymorphic sites (the polymorphic nucleotides of all 51 alleles are shown in Figure 2). Fifty-seven of these sites were dimorphic (two alleles), and the remaining 5 were trimorphic (three alleles). Eighteen of the nucleotide polymorphisms are novel to this data set, all of which cause amino acid replacements. Average pairwise nucleotide diversity per site (π) was 0.01642. Figure 3A shows a sliding window plot of π across the entire region. A representation of the positions of the three previously determined domains of *AMA1* is shown above the plot, each bounded by terminal cysteine codons (as in Table 1). Although the polymorphic sites are distributed throughout the gene, 38 of them are in Domain I, while Domains II and III have only 9 polymorphic sites each. Similarly, the pairwise diversity is highest in Domain I ($\pi = 0.027$).

Recombination and linkage disequilibrium: In total there were 45 different allele sequences among the 51 sampled (Figure 2). The Hudson and Kaplan (1985) algorithm estimated a minimum number of 24 recombination events having occurred among the sequences. The recombination parameter, *C*, equal to $4Nr$ (Hudson 1987), was estimated at 0.158 between adjacent sites and 207 for the whole sequence. Linkage disequilibrium was calculated for the data set using dimorphic sites at which the frequency of the minor allele was at least 0.1. Figure 4 shows the relationship between nucleotide distance between pairs of sites and R^2 and D' indices of linkage disequilibrium. Both R^2 and D' decline rapidly with distance, and most of the statistically significant linkage disequilibrium is between sites <500 bp apart. The high value of the recombination parameter, and the very rapid decay of linkage disequilibrium, are independent and concordant indicators of a very high meiotic recombination rate in the gene.

Tajima's *D* test: Table 1 shows positive values of *D* for the region as a whole and each domain separately. The *D* value for Domain III shows a significant positive departure from zero. Figure 3B shows a sliding window plot for Tajima's *D* value, illustrating that Domain III emerges as the most highly positive. This results from a greater number of nucleotide alleles at intermediate frequencies compared with expectations under neutrality.

Fu and Li's *D* and *F* tests: Table 1 gives the *D* and *F* values for each domain in isolation and the region as a whole. Values for the whole of the ectodomain, and for Domain I on its own, are significantly greater than zero. These result from a significantly fewer number of mutations occurring in external branches of a phylogeny compared to that predicted by either the average pairwise diversity or the total number of mutations. This is evidence that there are long internal branches in the phylogeny of this gene, which is consistent with balancing selection-maintaining alleles in both Domain I and the sequenced region as a whole, as seen in a sliding window plot of *D* and *F* (Figure 3C).

Coalescent simulations: Table 2 shows the upper 95% confidence intervals for the expected values of Tajima's *D* and Fu and Li's *F* under neutrality and the observed values for the sequence as a whole and for each domain (I–III). These are given for different levels of recombination (*C*), ranging from 0 to 100 (the maximum range available in the DnaSP 3.52 program, which is less than the value of 207 estimated above, and so still tends toward conservatism). The simulations show that when *C* is set at the relatively modest value of 20, the observed data depart from neutrality more significantly than was apparent in Table 1. For example, Tajima's *D* value shows a significant departure from zero for the sequence as a whole ($P < 0.01$), and Domain III is also highly significant ($P < 0.01$). When *C* is set at 50, the observed Tajima's *D* value for Domain I on its own is also significant ($P < 0.05$). For Fu and Li's *F* test, the sequence as a

```
                                                                  1111111111111111111
          4555555555555555566666666677777788888889999111222333344445555567
          8011256678899999001117788222339044578927897781120145578901033 4
          6059391606901689230692448479059658366823857841314522488378452 1

NIG-002   TCATGGATGGCGTGGATTAGGGATAGAAAACAACCAGGGCCTCAGCAATCATAATGAATGCC
NIG-005   .....AT.A..A....C...T...CT..G........AC....GAG...GG..C.....AAG
NIG-006   ..GA.AT.AAG..A....G....AGTG.GTG......A..TA..AG.T..GAT.G.CGA...
NIG-015   .AG.....A..AA.......T....TG.G....T..CA..TAT.AG....GATC.....AAG
NIG-016   A....A.....AA.C...GAT.T.....G.G.T....AC.T..GA.....G..........
NIG-026   .....AT.A..A....C...T...CT..G........AC....GAG...GG..C.....AAG
NIG-029   A....ATC..GA..CT.GG.........G.G.........TA...G....GAT.G.CGA.AG
NIG-030   A....ATC.AG......A..T..AGT..G.G......A..TA.GAG....GAT.G.CG..AG
NIG-034A  A.........GA..CT.G..T....T.TGTG......A..TA.GAG....GAT......A.G
NIG-034B  .....AT.A..A....C...T...CT..G........AC....GAG....G.T......A.G
NIG-035   .AG.....A.GA........T...CTG.G.......CA..TA...G....GATC.....AAG
NIG-036   .....A.....AA.C...GAT.T..TG.G.G.T....AC.T...A.G..A.........AAG
NIG-039   .....A.....AA.......T...GT..G....TG.CAC.T...A.....GAT.G.CGA...
NIG-043   .....A.....AA.......T...GT..G....TG.CAC.TA..A.....GAT.G.CGA...
NIG-044   ..G.T.....GAA.C...GAT.T.....G.G.T..G.AC......G....GAT......A..
NIG-046   A....ATC.AG......A..T....T.TGTG......A..TA.GA.G..A...C.....AAG
NIG-050   ....................T.............................GATC.C...AAG
NIG-052   ..........................................GATC.....AAG
NIG-054   ..GA.AT.A..A........T...GT..G.......CAC.TA.GAG....GAT.G.CG..AG
NIG-056   A....ATC.AG......A..T..AGT..G.G......A..TA..A.G..AG...G.CG....
NIG-057   ..GA.AT.A..A........T...GT..G....T..C.C.TA..AGG..AG...G.CG....
NIG-058   A..........AA.C...G..................A..TA.....................
NIG-059   .AG.....A.GA........T....TG.G....TG.....TA.GAG...A.........AAG
NIG-060   .....A.....AA.C...GAT.T.....G..T.....AC.TA..A.G..A.........AAG
NIG-062   .....A.....AA.......T...GT..G....TG.CAC.TA..A.....GAT.G.CGA...
NIG-064   .AG.....A.GA........T....TG.G....TG.....TA.GAG...A.........A..
NIG-065   .....A.....AA.C...GAT.T.....G..TT....A.GTA..A.GT.A....G.CG..AG
NIG-066   .....AT.A..A....C...T...CT..G........AC....GAG...GG..C.....AAG
NIG-067   .AG.....A.GA........T....TG.G....TG.....TA.GAG...A.........AAG
NIG-077   .....A.....AA.C...GAT.T..TG.G.G..TG.....TA..AGG.............AAG
NIG-078   .....ATC...AA.C...GAT.T.....G.G..TG..AC.TA..AGG..AG...G.CG..AG
NIG-083   A....ATC.AG......A..T...G...G.G......A..TA.GAG....GAT.G.CG..A.
NIG-080   ..G.T.....GAA.C...GAT.T.....G.G.T..G.AC.....A.G..AG...G.CG..AG
NIG-087   ..........................................A.....GAT.G.CGA..G
NIG-088   A....A....GAA.C...GAT.T.....G.G.T....AC.T.....G..AG...G.CG....
NIG-089   .AG.....AAG.........T..GG...G.G......A..TA..AG.T.A....G.CG....
NIG-091   .AG..AT.AAG......A.....................AC.TA.GAG...GG..C.....AAG
NIG-092   ..G.T.....GAA.C...GAT.T.....G...........TA...G....GATC.....A..
NIG-094   ..G.T.....GAA.C...GAT.T.....G.G.T..G.AC.....A.G..AG...G.CG..AG
NIG-096   A....A.....AA.C...GAT.T.....G.G.T....AC.T..GA.....G..........
NIG-100   .AG.....A..AA.C...G.T....TG.G....T..CAC.T...A.G..AG...G.CG..AG
NIG-101   .AG.....AAG.........T..GG...G.G......A.GTA..A.GT.A....G.CG..AG
NIG-105   .AG.....A.GA........T....TG.G....T..CAC.TA..AGG...GAT.G.CGA...
NIG-107   .AG.....A..AA.C...G.T....TG.G....T..CA..TA..A.G..AGATCG.CG.AAG
NIG-108   ..........AA.C...GAT.T.....G..TT....AA....GAGG.A.GAT......A..
NIG-110   A....ATC.AG......A..T....T.TG.G......A..TA...G..A.GATCG.CG....
NIG-113   A.........GA..CG..G.T....T.TG.G......A..TA.GAG..A.GAT......AAG
NIG-114A  .AG.....A.GA........T....TG.G....TG.....TA.GAG...A.........A.G
NIG-114B  .AG.......GA........T....TG.G....TG.....TA.GAG...A.........A.G
NIG-117   ..GA.AT.AAG..A....G....AGT..G.G......A..TA.GAG....GAT.G.CG..AG
NIG-119   A....A....GA..CT.GG...T.....G.G..TG..AC.TA..AGG.A.GATC.....AAG
```

FIGURE 2.—Polymorphic nucleotides in 1311 bp sequenced from 51 Nigerian alleles of *P. falciparum AMA1* (full alignment available as EMBL accession no. ALIGN_000019 at the following URL address: ftp://ftp.ebi.ac.uk/pub/databases/embl/align/). The three synonymous polymorphisms are shown in italics, including site 603, which is part of a codon with a complex evolutionary history. Nucleotide positions in the gene are written vertically above the polymorphic sites.

whole and Domain I on its own show a highly significant depature from 0 when C is set at 20 ($P < 0.01$), and the test for Domain III approaches significance. In contrast, Domain II fails to show a significant departure from neutrality with any test, assuming any level of recombination.

McDonald-Kreitman test: Table 3 gives the results of the McDonald-Kreitman test for the whole sequence and for the domains taken individually, comparing polymorphism within *P. falciparum* and fixed differences between *P. falciparum* and *P. reichenowi*. The sequence as a whole shows a significant departure from neutrality with a high excess of intraspecific nonsynonymous polymorphisms. All three of the domains show the same trend, indicating that diversifying selection is occurring across the whole of the sequence, and this is statistically significant for Domain I. The small number of fixed differences between species in Domains II and III means that the test has less power for these domains.

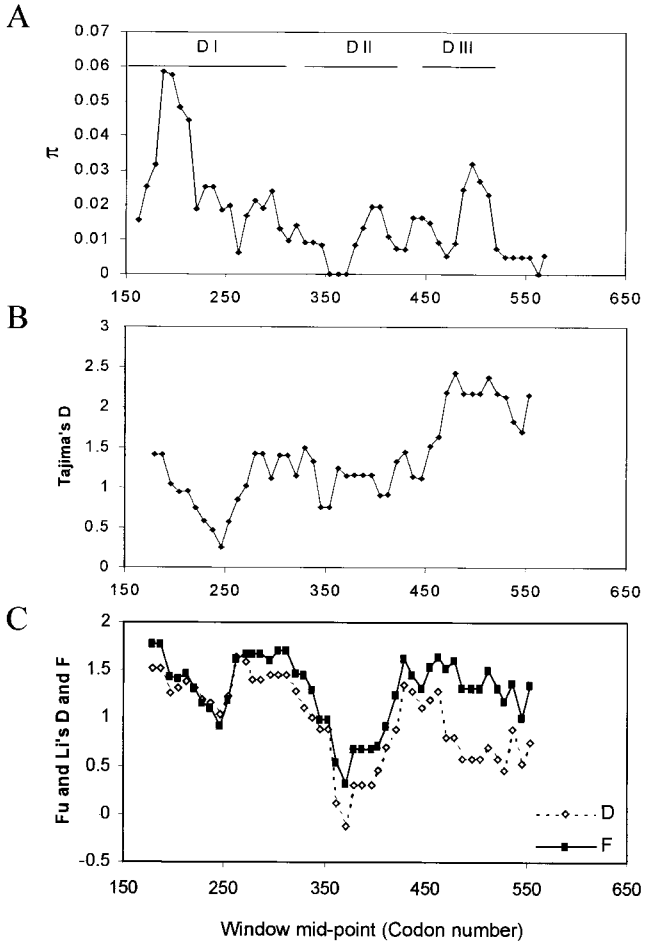

FIGURE 3.—Sliding window plot of (A) average pairwise diversity (π), (B) Tajima's *D* value, and (C) Fu and Li's *D* and *F* values across *AMA1*. A window size of 100 nucleotides was used for the analysis of π, while a window size of 200 nucleotides was used for the analysis of Tajima's *D* and Fu and Li's *D* and *F*. The three domains of *AMA1* are represented within the first plot as lines labeled DI, DII, and DIII, representing Domains I, II, and III, respectively.

DISCUSSION

The signature of natural selection on particular domains of the *P. falciparum AMA1* gene is very strong. Previous analyses of a small number of alleles indicated that positive selection had probably operated (HUGHES and HUGHES 1995; ESCALANTE *et al.* 1998; VERRA and HUGHES 2000), in particular on the surface-associated ectodomain (KOCKEN *et al.* 2000). By studying a large number of alleles sampled from a single population, the present study aimed to test independently for selection and to locate the regions of the sequence under the strongest selection. Several different tests were used, which complement and re-enforce each other, giving strongly consistent results.

The three domains (I–III) defined as disulfide-bonded structures (HODDER *et al.* 1996) show different patterns of nucleotide diversity, with Domain I having a higher number of polymorphic sites and nucleotide diversity than Domains II and III. There is significant evidence for balancing selection on Domain I, as shown by Tajima's *D* and Fu and Li's *D* and *F* tests. There is also significant evidence for balancing selection on Domain III, as shown by Tajima's *D* test and Fu and Li's *F* tests approaching significance (note that the smaller number of polymorphic sites in this domain gives less power than for the analysis of Domain I). In contrast with Domains I and III, there was no significant evidence of selection on Domain II.

Recombination is very frequent in the *AMA1* gene, as shown by the high estimate of the recombination parameter, *C*, and the very rapid decline in linkage disequilibrium with increasing distance between nucleotide sites (there is no significant linkage disequilibrium between pairs of sites more than ~600 nucleotides apart). This is similar to the rapid decline in linkage disequilibrium with distance already reported in the *P. falciparum* gene *msp1*, in this and most other African populations studied (CONWAY *et al.* 1999), and reflects

TABLE 1

Summary of statistics for *AMA1* and tests for departure from neutrality

	S	*S(si)*	*H*	π	*D* (Tj)	*D* (F&L)	*F*
Total region	62 (67)	4	45	0.016	1.559	1.699*	2.000*
Domain I	38 (42)	2	35	0.027	1.110	1.694*	1.780*
Domain II	9 (9)	1	20	0.010	1.187	0.701	1.028
Domain III	9 (9)	1	13	0.018	2.235*	0.701	1.448

The total region sequenced comprised codons 147–584; Domain I codons 149–302; Domain II codons 320–418; and Domain III codons 443–509 (HODDER *et al.* 1996). Six polymorphic sites (seven mutations) map to locations outside Domains I–III; hence the number in these domains do not equal the total. For both Tajima's *D* test and Fu and Li's *D* and *F* tests the total number of mutations, rather than the number of segregating sites, was used to estimate θ because the latter does take into account several instances of multiple mutations at the same sites. * indicates a significantly positive value of a given index ($P < 0.05$). *S*, number of segregating sites (parentheses give the number of mutations); *S(si)*, number of singleton mutations; *H*, number of haplotypes; π, average pairwise nucleotide diversity; *D* (Tj), Tajima's *D* value; *D* (F&L), Fu and Li's *D* value; *F*, Fu and Li's *F* value.

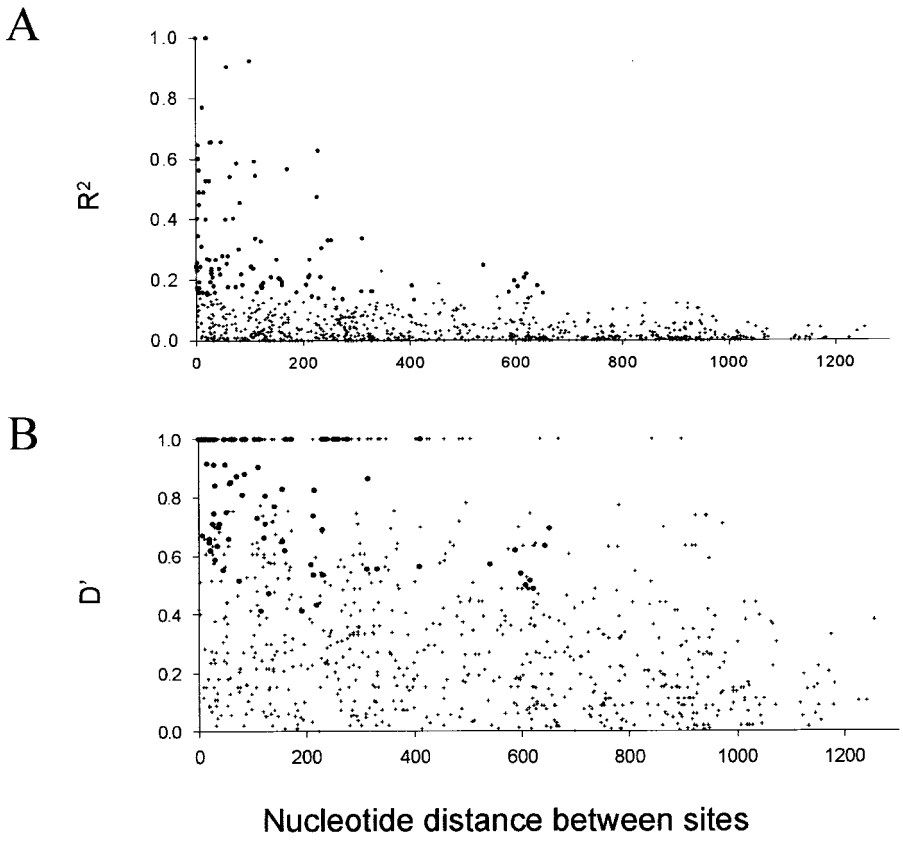

FIGURE 4.—Linkage disequlibrium across the *AMA1* gene as calculated using (A) R^2 and (B) D'. Those pairs of sites that show significant linkage disequilibrium as calculated by Fisher's exact test are shown as solid circles, while all other pairs of sites are represented by small crosses.

the high meiotic recombination rate throughout the *P. falciparum* genome (SU *et al.* 1999). If the average recombination rate in the genome is considered, determined experimentally by SU *et al.* (1999) as 1 cM per 17 kb of sequence, this equates to a probability of recombination between adjacent nucleotides, $r = 6 \times 10^{-7}$ (*i.e.*, the reciprocal of 1 M = 1.7 Mb). With this value of r, we can estimate the genetically effective population size N, using the recombination parameter equation $C = 4Nr$ (HUDSON 1987), as we have an estimate here of $C = 0.158$ (between adjacent nucleotides). Thus, $N = C/4r = 6.6 \times 10^4$, which is very similar to estimates for African populations of $\sim 10^4$ derived by ANDERSON *et al.* (2000) using an entirely different approach (based on

TABLE 2

Results of Tajima's *D* test and Fu and Li's *F* test with recombination

	Total sequence	Domain I	Domain II	Domain III
Tajima's *D*				
Observed	1.559	1.110	1.187	2.235
Expected upper 95% CI ($C = 0$)	1.784	1.815	1.924	1.920*
Expected upper 95% CI ($C = 20$)	1.051**	1.171	1.583	1.592**
Expected upper 95% CI ($C = 50$)	0.863**	0.986*	1.463	1.469**
Expected upper 95% CI ($C = 100$)	0.699**	0.880*	1.416	1.439**
Fu and Li's *F*				
Observed	2.000	1.780	1.028	1.448
Expected upper 95% CI ($C = 0$)	1.755*	1.779	1.714	1.733
Expected upper 95% CI ($C = 20$)	1.291**	1.344**	1.623	1.623
Expected upper 95% CI ($C = 50$)	1.152**	1.249**	1.591	1.585
Expected upper 95% CI ($C = 100$)	1.089**	1.234**	1.567	1.567

Expected values are the upper 95% confidence limits expected under a panmictic constant-size neutral model. C refers to the recombination parameter used in calculating the expected value by coalescent simulation. * and ** signify when observed values of D or F lie outside the neutral distribution (* $P < 0.05$, 95% CI as shown; ** $P < 0.01$, 99% CI not shown). CI, confidence interval.

TABLE 3

Summary of results of McDonald and Kreitman analysis of *AMA1*

Region	Polymorphic changes within *P. falciparum*		Fixed differences between species		Neutrality index	Fisher's exact test
	Synonymous	Nonsynonymous	Synonymous	Nonsynonymous		
Total region	3	64	12	22	11.8	0.00008***
Domain I	3	39	4	8	6.5	0.03616*
Domain II	0	9	1	4	∞	0.357143
Domain III	0	9	1	3	∞	0.307692

Neutrality index is that of Rand and Kann (1996), similar to an odds ratio on a 2×2 table. Several complex codons were analyzed by eye using the methodology given in DnaSP 3.5 (Rozas and Rozas 1999). * $P < 0.05$; *** $P < 0.001$.

estimates of heterozygosity and mutation rates of microsatellite loci).

Recombination is predicted to make Tajima's and Fu and Li's tests conservative in identifying statistically significant positive deviations from neutrality. When recombination is accounted for by use of coalescent simulations, more significant departures from neutrality can be observed. It is only by the use of this approach that the ectodomain sequence as a whole, and Domain I on its own, shows a clear positive departure from neutrality with Tajima's *D* test. The *D* value for Domain II, by contrast, stays well within the 95% confidence intervals, even when the highest level of recombination is used to generate the neutral expectations. With Fu and Li's *F* test, Domains II and III do not show significant departure from neutrality, although the results do tend toward significance for Domain III ($P < 0.1$) when the 95% confidence intervals under neutrality are calculated assuming a high level of recombination.

Some effects on Tajima's (and similarly on Fu and Li's) indices can be due to population substructure, or changes in population size, in ways that can mimic the patterns of selection (Tajima 1989b; Fu 1996; Wall 1999). Here, we studied alleles from a single population sample to avoid substructure generated by pooling alleles from different populations. Theoretically, if a population size declines drastically, the number of segregating sites is reduced more markedly than the average number of nucleotide differences (Tajima 1989c), and this can result in transiently positive values for *D*. Analysis of microsatellite allele frequency distributions does not suggest that any such decline has occurred recently in African populations of *P. falciparum* (Anderson *et al.* 2000). The stable, high endemicity of malaria in the area from which samples were studied here is typical in much of Africa, and thus the positive Tajima's *D* values obtained here can be attributed to balancing selection. In an expanding population, conversely, there would be a tendency for Tajima's *D* values to be negative (Tajima 1989c). There is evidence to indicate that *P. falciparum* emerged and expanded from a small founding population in Africa some thousands, or tens of thousands, of years ago (Conway *et al.* 2000b), and thus any skewing of neutral Tajima's *D* values would be in the opposite direction to that observed in the data.

The McDonald-Kreitman test using the allele sequences obtained here gives a consistent, but even more highly statistically significant, result than that obtained previously with a small number of other alleles. The present analysis has more power, as there are more polymorphic nucleotides in the new sample of 51 alleles. This test could theoretically be affected if there were differences in the relative codon usage of the two species analyzed (Nei and Kumar 2000). However, comparison of codon prevalence between the *P. falciparum* and *P. reichenowi AMA1* sequences shows a virtually identical pattern, and there is no evidence from other genes to suggest a difference in the codon bias in these closely related species. The large number of polymorphic sites in this study allows an analysis of individual domains, with Domain I clearly showing a significant excess of intraspecific nonsynonymous polymorphisms. This positive trend is also seen, although not significantly, in the other two domains, which have fewer polymorphic nucleotides.

Taken together, the sequence analyses presented here on a new population data set clearly show that the *P. falciparum AMA1* gene is under selection, which is maintaining nucleotide polymorphisms, particularly those that are nonsynonymous. The strongest selection appears to act on sequences encoding Domains I and III, and we hypothesize that this is produced by the host immune system mounting an effective response to epitopes within these domains of the protein. We predict that immunological studies will demonstrate that human antibodies to polymorphic sequences in one or both of these domains inhibit parasites and protect against malaria.

We thank Ayoade Oduola and Olumide Ogundahunsi for original sample collection and Jennie Lloyd and Zsusanna Mikes for technical assistance. We are also grateful to Ziheng Yang and Michael Gaunt for comments on the analyses and to Martin Holland and Kevin Tetteh for comments on the manuscript. This work was supported by the UK Medical Research Council (grant G9803180).

LITERATURE CITED

Anderson, T. J. C., B. Haubold, J. T. Williams, J. G. Estrada-Rfanco, L. Richardson *et al.*, 2000 Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. Mol. Biol. Evol. **17:** 1467–1482.

Brunham, R. C., F. A. Plummer and R. S. Stephens, 1993 Bacterial antigenic variation, host immune response, and pathogen-host coevolution. Infect. Immun. **61:** 2273–2276.

Collins, W. E., D. Pye, P. E. Crewther, K. L. Vandenberg, G. G. Galland *et al.*, 1994 Protective immunity induced in squirrel monkeys with recombinant apical membrane antigen-1 of *Plasmodium fragile*. Am. J. Trop. Med. Hyg. **51:** 711–719.

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

Conway, D. J., C. Roper, A. M. Oduola, D. E. Arnot, P. G. Kremsner *et al.*, 1999 High recombination rate in natural populations of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA **96:** 4506–4511.

Conway, D. J., D. R. Cavanagh, K. Tanabe, C. Roper, Z. S. Mikes *et al.*, 2000a A principal target of human immunity to malaria identified by molecular population genetic and immunological analyses. Nat. Med. **6:** 689–692.

Conway, D. J., C. Fanello, J. M. Lloyd, B. M. A.-S. Al-Joubori, A. H. Baloch *et al.*, 2000b Origin of *Plasmodium falciparum* malaria is traced by mitochondrial DNA. Mol. Biochem. Parasitol. **111:** 163–171.

Crewther, P. E., M. L. Matthew, R. H. Flegg and R. F. Anders, 1996 Protective immune responses to apical membrane antigen 1 of *Plasmodium chabaudi* involve recognition of strain-specific epitopes. Infect. Immun. **64:** 3310–3317.

Deans, J. A., A. M. Knight, W. C. Jean, A. P. Waters, S. Cohen *et al.*, 1988 Vaccination trials in rhesus-monkeys with a minor, invariant, *Plasmodium knowlesi* 66 kd merozoite antigen. Parasite Immunol. **10:** 535–552.

Escalante, A. A., A. A. Lal and F. J. Ayala, 1998 Genetic polymorphism and natural selection in the malaria parasite *Plasmodium falciparum*. Genetics **149:** 189–202.

Fu, Y.-X., 1996 Statistical tests of neutrality for DNA samples from a population. Genetics **143:** 557–570.

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

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

Hodder, A. N., P. E. Crewther, M. L. Matthew, G. E. Reid, R. L. Moritz *et al.*, 1996 The disulfide bond structure of *Plasmodium* apical membrane antigen-1. J. Biol. Chem. **271:** 29446–29452.

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.

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

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

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

Kocken, C. H., D. L. Narum, A. Massougbodji, B. Ayivi, M. A. Dubbeld *et al.*, 2000 Molecular characterisation of *Plasmodium reichenowi* apical membrane antigen-1 (AMA-1), comparison with *P. falciparum* AMA-1, and antibody-mediated inhibition of red cell invasion. Mol. Biochem. Parasitol. **109:** 147–156.

Lal, A. A., M. A. Hughes, D. A. Oliveira, C. Nelson, P. B. Bloland *et al.*, 1996 Identification of T-cell determinants in natural immune responses to the Plasmodium falciparum apical membrane antigen-1 (AMA-1) in an adult population exposed to malaria. Infect. Immun. **64:** 1054–1059.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations: heterotic models. Genetics **49:** 49–67.

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

Nei, M., and S. Kumar, 2000 *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Rand, D. M., and L. M. Kann, 1996 Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. Mol. Biol. Evol. **13:** 735–748.

Riley, E. M., G. E. Wagner, M. F. Ofori, J. G. Wheeler, B. D. Akanmori *et al.*, 2000 Lack of association between maternal antibody and protection of African infants from malaria infection. Infect. Immun. **68:** 5856–5863.

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

Su, X-Z., M. T. Ferdig, Y. Huang, C. Q. Huynh, A. Liu *et al.*, 1999 A genetic map and recombination parameters of the human malaria parasite *P. falciparum*. Science **286:** 1351–1353.

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

Tajima, F., 1989b DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. Genetics **123:** 229–240.

Tajima, F., 1989c The effect of population size on DNA polymorphism. Genetics **123:** 597–601.

Thomas, A. W., J. F. Trape, C. Rogier, A. Goncalves, V. E. Rosario *et al.*, 1994 High prevalence of natural antibodies against *Plasmodium falciparum* 83-kilodalton apical membrane antigen (PF83/AMA-1) as detected by capture-enzyme-linked immunosorbent assay using full-length baculovirus recombinant PF83/AMA-1. Am. J. Trop. Med. Hyg. **51:** 730–740.

Verra, F., and A. L. Hughes, 2000 Evidence for ancient balanced polymorphism at the Apical Membrane Antigen-1 (AMA-1) locus of *Plasmodium falciparum*. Mol. Biochem. Parasitol. **105:** 149–153.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. Camb. **74:** 65–79.

Communicating editor: D. Charlesworth