Allele Frequency-Based Analyses Robustly Map Sequence Sites Under Balancing Selection in a Malaria Vaccine Candidate Antigen

Spencer D. Polley,*† Watcharee Chokejindachai*† and David J. Conway*

*Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom and †Faculty of Tropical Medicine, Mahidol University, Rajthewe, Bangkok 10400, Thailand
Manuscript received April 11, 2003
Accepted for publication June 17, 2003

ABSTRACT

The Plasmodium falciparum apical membrane antigen 1 (AMA1) is a leading candidate for a malaria vaccine. Here, within-population analyses of alleles from 50 Thai P. falciparum isolates yield significant evidence for balancing selection on polymorphisms within the disulfide-bonded domains I and III of the surface accessible ectodomain of AMA1, a result very similar to that seen previously in a Nigerian population. Studying the frequency of nucleotide polymorphisms in both populations shows that the between-population component of variance (FSt) is significantly lower in domains I and III compared to the intervening domain II and compared to 11 unlinked microsatellite loci. A nucleotide site-by-site analysis shows that sites with exceptionally high or low Fst values cluster significantly into serial runs, with four runs of low values in domain I and one in domain III. These runs may map the sequences that are consistently under the strongest balancing selection from naturally acquired immune responses.

Evidence of natural selection on gene sequences can lead to focused hypotheses on the functions of proteins and their existing alleles. Phylogenetic and codon-based methods of analyzing homologous gene sequences are powerful and are broadly used (Fitch et al. 1991; Nielsen and Yang 1998; Yáng et al. 2000; Jiggins et al. 2002; Zhang et al. 2002), but they have limitations for some applications. First, phylogenetic relationships among alleles within a species may not be accurately derived if recombination occurs frequently (Holmes et al. 1999; Schierup and Hein 2000). Second, the probability of mutational changes between particular codons (or overall summarized rates of synonymous and nonsynonymous changes among codons) may not be correctly interpreted, if there are unusual constraints on codon usage that are not completely known and accounted for (Akashi 1995).

One species that has both a high recombination rate (Conway et al. 1999; Su et al. 1999; Polley and Conway 2001) and an extreme bias in codon usage (Forsdyke 2002) is the human malaria parasite Plasmodium falciparum. There is a great need to study selection on genes of this species, as the causes of directional selection (particularly due to drug resistance) and frequency-dependent balancing selection (due to acquired immune responses) are important for understanding how
to control malaria in the future (Conway et al. 2000a; Wootton et al. 2002). Naturally acquired, nonsterile immunity to P. falciparum is seen in highly endemic populations of Africa and Asia (Marsh and Snow 1997). As antibody and T cell responses involve memory, parasites with rare antigenic types can have a frequency-dependent selective advantage. Frequency-based methods could be applied to detect this type of selection (Tajima 1989b; Fu and Li 1993; McDonald 1994). This approach offers the potential to identify targets of naturally acquired immunity (Conway et al. 2000a; Conway and Polley 2002) and aid the design of a vaccine to induce immunity against each allelic component (Conway 1997).

Apical membrane antigen 1 (AMA1) is a leading candidate in the search for a vaccine against P. falciparum. The protein is located in the microneme organelles at the apical end of the merozoite, the stage of the parasite that invades erythrocytes (Healer et al. 2002). It is encoded by a single-locus gene that is essential for parasite growth in vitro (Triglia et al. 2000). Immunization of mice with AMA1 of P. chabaudi can protect against lethal challenge in a strain-specific manner (Crewther et al. 1996), and antibodies to P. falciparum AMA1 have been shown to inhibit erythrocyte invasion in vitro with a fairly high degree of strain specificity (Hodder et al. 2001; Kennedy et al. 2002; Kocken et al. 2002). There is a strong predominance of nonsynonymous vs. synonymous polymorphisms (dN > dS) among ama1 alleles in P. falciparum (Hughes and Hughes 1995; Verra and Hughes 2000), more so than for fixed differences with the closely related species P. reichenowi (Kocken et al. 2000; Polley and Conway 2001). Constraints on synon-
younous mutation in *P. falciparum* (Forsdyke 2002) mean that an observation of $d_0 > d_1$ is not interpretable simply as positive selection on amino acids, and even the validity of the more robust comparison with *P. reichenowi* in the McDonald-Kreitman test (Kocken et al. 2000; Polley and Conway 2001) might be violated if constraints are different in that species (Akashi 1995).

A recent study of *ama1* single-nucleotide allele frequency distributions in a large sample of sequences from a Nigerian population indicates that polymorphisms are selectively maintained within domains I and III of the surface accessible ectodomain (Polley and Conway 2001). These domains were previously defined on the basis of the predicted secondary structure of AMA1 (Hodder et al. 1996), although their role in the function of AMA1 is currently unknown. It is important to know if these significant results can be generalized to other populations of *P. falciparum* and if the precision of the findings can be further increased. Asian populations of *P. falciparum* have been shown to be distinct from African populations on the basis of microsatellite allele frequencies (Anderson et al. 2000; Conway et al. 2001) and mitochondrial haplotype frequencies (Conway et al. 2000b; Joy et al. 2003). A large sample of *ama1* alleles from Thailand has been sequenced so that independent within-population and between-population analyses can be performed to evaluate signatures of balancing selection. Results show that domains I and III of *ama1* are under strong balancing selection in Asia, as well as in Africa, and identify sequential runs of polymorphic sites within these domains that may be of particular importance for the design of an AMA1-based vaccine.

**MATERIALS AND METHODS**

**Sequencing of ama1 from Thai *P. falciparum* isolates:** Genomic DNA was extracted from blood collected from malaria patients participating in clinical studies at the Hospital for Tropical Diseases (Mahidol University, Bangkok) with the approval of the institutional review board. Almost all cases were in the Karen and Maun ethnic groups who had been infected in an endemic area of the Thai-Myanmar border. DNA extractions were performed using the QiAmp DNA blood mini kits (QIAGEN, Chatsworth, CA). A 1371-nucleotide region of *ama1* (codons 144–599, which include the ectodomain), located on chromosome 11, was amplified and sequenced as three overlapping PCR fragments from 50 clinical isolates using the exact methods described previously in a study of a Nigerian population (Polley and Conway 2001). These isolates were selected from a larger panel of 100 clinical isolates because each contained a single-clone *P. falciparum* infection, as determined by microsatellite typing, thus allowing direct sequencing without cloning into bacterial plasmids. Any regions sequenced in only one direction (including primer sequences) were removed from each of the three fragments before their amalgamation into a single contiguous sequence for each isolate. Raw sequence data were checked and alignments were performed using the Seqman II and MegAlign programs (DNASTAR, Madison, WI). To ensure they were not PCR artifacts, all singletons (single-nucleotide alleles that occur in only one isolate) were verified by reamplification and resequencing from the genomic template.

**Genotyping of microsatellite loci in Thai *P. falciparum* isolates:** Alleles at 11 microsatellite loci were typed in the 50 Thai *P. falciparum* isolates using the seminested PCR method described previously for other samples (Anderson et al. 1999; Conway et al. 2001). The loci used (and their chromosomal positions) are as follows: POLYA (chr. 4), TA42 (chr. 5), TA81 (chr. 5), TA1 (chr. 6), TA87 (chr. 6), TA109 (chr. 6), ARA2 (chr. 11), PJPK2 (chr. 12), Pjg377 (chr. 12), TA102 (chr. 12), and TA60 (chr. 13).

**Analysis of sequence diversity and linkage disequilibrium:** Sequence diversity (π, average pairwise nucleotide diversity) was calculated for distinct domains of the *ama1* gene (domains I–III). Analysis of linkage disequilibrium was performed between nucleotide sites at which the frequency of the minority allele was >0.1 using $D'$ ( Lewontin 1964) and $R^2$ (Hill and Robertson 1968) indices with Fisher’s exact test of significance, calculated via the DNAsp3.53 program (Rozas and Rozas 1999). The recombination parameter C (equal to 4N, where N is the effective population size and r is the underlying recombination rate) was estimated using an approximate-likelihood coalescent method via the pairwise program (part of the LDatrace package downloadable from http://www.stats.ox.ac.uk/~mchung/LDatrace/LDatrace.html), together with the correlation coefficient of both $D'$ and $R^2$ with distance. A formal test for the presence of recombination was performed by a permutation test: 1000 randomly sampled data sets (in which the order of the polymorphic *ama1* loci was randomly shuffled) were produced and the proportion of these data sets showing a more extreme coefficient of correlation for $D'$ and $R^2$ with distance was recorded (McVean et al. 2002). A separate estimate of the population recombination parameter C was calculated according to the method of Hudson (1987) together with the minimum number of recombination events (Hudson and Kaplan 1985) using the DNAsp3.53 program.

**Within-population tests of neutrality:** Tajima’s D test (Tajima 1989b), which compares θ (nucleotide diversity predicted from the number of segregating sites) and π (observed pairwise nucleotide diversity), was used to investigate whether polymorphic single-nucleotide alleles tended to occur at higher or lower frequencies than expected under neutral drift. Fu and Li’s F test (Fu and Li 1993) was used to compare the number of singleton nucleotides in the *ama1* sequences with the number predicted under neutrality given the average number of nucleotide differences between pairs of alleles and using the *P. reichenowi ama1* sequence as an outgroup (Kocken et al. 2000). Critical values for the above tests were calculated by coalescent simulations with 10,000 replicates. As recombination tends to make these tests conservative (Tajima 1989b; Fu and Li 1993; Wall 1999), an additional series of coalescent simulations were run to account for the level of recombination (C) observed in the *ama1* sequences. All of these tests were performed with the DNAsp3.53 program.

**Comparison of between-population divergence at different loci:** Comparison of within- and between-population diversity in the *ama1* sequence data from the Thai population in this study and a Nigerian population (Polley and Conway 2001) was performed with the θ estimator of Wright’s fixation index ($F_S$) of interpopulation variance in allele frequencies (Weir and Cockerham 1984) using the FSTAT program (version 2.9.3.1; Goudet 1995). For *ama1*, mean $F_S$ values were calculated for the whole region sequenced, as well as each domain separately, together with a nucleotide site-by-site analysis. Of the 65 single-nucleotide sites that were polymorphic in the combined data set, only the 48 sites at which the minority allele(s) had a total frequency >0.1 (in the combined data set) were included for this analysis. The distributions of $F_S$
TABLE 1
Within-population analysis of amal nucleotide polymorphisms in Thailand

<table>
<thead>
<tr>
<th></th>
<th>No. of polymorphic sites</th>
<th>No. of haplotypes</th>
<th>Average pairwise diversity (π)</th>
<th>Tajima’s D value</th>
<th>Fu and Li’s F value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sequence</td>
<td>51</td>
<td>27</td>
<td>0.014</td>
<td>2.12*</td>
<td>1.95*</td>
</tr>
<tr>
<td>Domain I</td>
<td>32</td>
<td>18</td>
<td>0.025</td>
<td>1.92*</td>
<td>1.80*</td>
</tr>
<tr>
<td>Domain II</td>
<td>6</td>
<td>9</td>
<td>0.006</td>
<td>0.09</td>
<td>0.58</td>
</tr>
<tr>
<td>Domain III</td>
<td>6</td>
<td>8</td>
<td>0.014</td>
<td>2.73**</td>
<td>2.00*</td>
</tr>
</tbody>
</table>

Analyses are shown for the whole region sequenced (codons 148–582), domain I (codons 149–302), domain II (codons 320–418), and domain III (codons 443–509). Seven polymorphic sites lie outside the three domains. Three alternative single-nucleotide alleles exist at site 200b; hence the total number of mutations for the whole region and domain I is 52 and 33, respectively. For Tajima’s D test and Fu and Li’s F test, a value significantly different from zero is marked by an asterisk (*P < 0.05; **P < 0.01), as calculated by coalescent simulations with no recombination (C = 0). More highly significant results were obtained for domains I and III when recombination was accounted for (C = 66).

values for individual sites within each domain of amal and the FST values of 11 microsatellite loci were compared using the nonparametric Wilcoxon’s rank sum test using the SPSS 11.01 program (SPSS, Chicago). The Thai microsatellite data were those obtained in this study, while the Nigerian microsatellite data were those previously reported (Conway et al. 2001). The distribution of FST values among the amal polymorphic sites (considered here as a simple spatial series of 48 sequential sites) was also analyzed with Moran’s I index of global spatial autocorrelation (Moran 1950). The I index between neighboring sites was calculated using the ROOKCASE visual basic add-in for Excel (Sawada 1999) and statistically analyzed by comparison with the number of 1000 Monte Carlo randomly sampled data sets (in which the order of the 48 amal sites was randomly rearranged) that gave a higher value of I.

RESULTS

amal sequence polymorphism in Thai P. falciparum isolates: Double-stranded sequence was generated for a 1303-nucleotide portion of the amal gene (codons 148–582) from each of 50 single-clone P. falciparum infections. Sequences are available individually (EMBL accession nos. AJ494866–AJ494915) or as an alignment (EMBL-Align database: ALIGN_000459). There were 51 polymorphic sites, 48 of which have been described in previously sequenced amal alleles (Kocken et al. 2000; Escalante et al. 2001; Polley and Conway 2001). The remaining three polymorphic sites each contain singleton variants (where the rare nucleotide variant was found in only one individual), and these were all confirmed by repeat amplifications from genomic DNA and resequencing. Pairwise diversity among alleles (π) was 0.014 for the whole region sequenced and 0.025, 0.006, and 0.014 for domains I, II, and III, respectively (Table 1).

Recombination and linkage disequilibrium: The effects of recombination on amal alleles in the Thai population are shown graphically by the decline in levels of linkage disequilibrium (D’ and R^2 indices) with increasing distance between pairs of nucleotide sites (Figure 1). For D’ and R^2 the coefficient of correlation with distance was −0.275 and −0.261, respectively (none of the 1000 randomly sampled data sets produced more extreme values). Using a maximum-likelihood approach (McVean et al. 2002), we estimated the recombination parameter C to be 60 for the whole sequence. A minimum of 16 recombination events were required to ar-

![Figure 1.—Linkage disequilibrium across amal in the Thai population. The D’ and R^2 indices (A and B, respectively) for pairs of sites are plotted against the distance between them. Pairs of sites that show statistically significant linkage disequilibrium are shown by solid diamonds (crosses mark those showing nonsignificant linkage disequilibrium). Only sites containing a minority single-nucleotide allele with a frequency >0.1 were included in the analysis.](image)
Thai population are present at intermediate frequencies high sequenced shows that single-nucleotide alleles in the 0.245; Hudson of P significantly greater than zero, constant-size panmictic population (mean 395) and others with very low values (0.051 per adjacent site. sites along the gene (Figure 3). However, a significant

Within-population tests of neutrality: A highly positive Tajima’s D value of 2.12 for the entire ama1 region sequenced shows that single-nucleotide alleles in the Thai population are present at intermediate frequencies that cannot be accounted for by neutral evolution in a constant-size panmictic population (D value is significantly greater than zero, P < 0.05).

This departure from neutrality indicates that balancing selection is maintaining single-nucleotide alleles in the population. A sliding-window plot of D shows highly positive values for domains I and III and low values for domain II (Figure 2). There was significant evidence for balancing selection in domains I and III (with D values of 1.92 and 2.73, respectively), but the lower diversity in domain II produced no such trend (with a D value of 0.09; Table 1). When recombination was considered in the calculation of critical D values, even the higher level estimated above (C = 66) did not result in a significant departure from neutrality with domain II sequences, but for each of domains I and III the departure was highly significant (P < 0.01). Fu and Li’s F test shows very similar results, revealing a high proportion of polymorphisms in domains I and III that are likely to be more ancient than would be expected under neutrality (Table 1; Figure 2).

Between-population comparison of single-nucleotide allele frequencies: An analysis of ama1 single-nucleotide allele frequencies in this Thai population, together with comparable data from a previously studied Nigerian population (Polley and Conway 2001), shows that interpopulation divergence accounts for only 4.1% of total nucleotide diversity (mean $F_{ST} = 0.041$). A domain-by-domain analysis shows that $F_{ST}$ values are lowest (i.e., single-nucleotide allele frequencies are most similar) in domain I (mean $F_{ST} = 0.033$) and domain III (mean $F_{ST} = 0.032$), whereas domain II has a much higher mean $F_{ST}$ value (0.113). As a comparison, the allele frequencies of 11 putatively neutral, unlinked microsatellite loci were studied in both populations. These had $F_{ST}$ values (mean $F_{ST} = 0.106$) similar to those of domain II of ama1, but much higher than those of domains I and III (Figure 3). Wilcoxon’s rank sum test was applied to test for differences in the distribution of $F_{ST}$ values for the microsatellite loci and the individual polymorphic sites within each domain of ama1. This nonparametric analysis considers each site as an independent variable, although the $F_{ST}$ values of individual sites within ama1 may not be truly independent variables due to the physical linkage of the sites. The difference between the microsatellite $F_{ST}$ values and the ama1 nucleotide $F_{ST}$ values was significant in the case of domain I ($P = 0.006$) and of borderline significance in the case of domain III ($P = 0.056$), but no significant difference was seen with domain II ($P = 0.479$).

The analysis of individual polymorphic sites in ama1 reveals a great heterogeneity in the $F_{ST}$ values among sites along the gene (Figure 3). However, a significant spatial autocorrelation between neighboring sites is seen when analyzed with Moran’s I index (Moran’s I = 0.245; $P = 0.036$). This is due to clusters of sites with high $F_{ST}$ values (e.g., polymorphic sites in codons 330–395) and others with very low values (e.g., polymorphic sites 162c–172b, 196a–197c, 206a–225b, 283b–308a, and 493b–512b). Despite this clustering, however, sites in adjacent codons can also have very different $F_{ST}$ values (0 and 0.191 for sites 242a and 243a, respectively).

**DISCUSSION**

The within-population analyses here reveal a significantly nonneutral pattern of *P. falciparum* ama1 single-nucleotide allele frequencies in Thailand. The highly positive values of Tajima’s D and Fu and Li’s F are most likely a result of balancing selection maintaining rare single-nucleotide alleles within domains I and III (no evidence is seen for balancing selection on domain II). This closely concurs with results seen previously in a Nigerian population (Polley and Conway 2001), confirming the particular effects on domains I and III and
indicating that balancing selection is occurring in these different populations. The highly positive values of these indices are unlikely to be due to confounding population effects, as populations of *P. falciparum* are subject to expansions (Joy *et al.* 2003) and will therefore tend to have negative values under neutrality (Tajima 1989a) and the endemic populations were each chosen to avoid substructure (Tajima 1989b). Neither would codon bias be likely to lead to the positive values seen here, given that purifying selection leads to negative values of these indices (Tajima 1989b; Fu and Li 1993).

Differences between the Thai and Nigerian *P. falciparum* populations are evident, however, in the amount of recombination seen in *ama1*. The estimated recombination parameters of the Thai population are lower than those of the Nigerian population, with values of *C* calculated at 66 and 207, respectively, using one method (Hudson 1987) and 60 and >100, respectively, by another (McVean *et al.* 2002). At 35 of the polymorphic single-nucleotide sites in *ama1*, the minority alleles are present at frequencies >0.1 in both populations, and these sites can be analyzed in 595 different pairwise combinations for each data set. A significantly higher proportion of these showed statistically significant linkage disequilibrium in the Thai population (278 combinations) compared to the Nigerian population (154 combinations; *P* < 0.001). The difference in recombination rates of the two populations concurs with studies on unlinked microsatellite loci, which indicate that Asian *P. falciparum* populations generally have a lower recombination rate than African populations (Anderson *et al.* 2000). This is probably because the endemicity is lower, restricting the proportion of mixed clone infections and subsequent cross-fertilization and recombination in the mosquito stage of the life cycle (Walliker 2000).

Analysis of single-nucleotide allele frequencies between the populations shows that domains I and III of *ama1* have very low mean *F*<sub>*ST*</sub> values, indicating that virtually all polymorphism is seen within each population. The *F*<sub>*ST*</sub> values of these *ama1* domains are significantly lower than the *F*<sub>*ST*</sub> values of 11 unlinked microsatellite loci. In the absence of selection the high allelic diversity at microsatellite loci will generally result in lower *F*<sub>*ST*</sub> values than those for single nucleotide polymorphisms (SNPs; Hedrick 1999), a point that underscores the significance of this result. Thus polymorphisms in domains I and III of *ama1* are apparently under balancing selection, in concordance with the within-population analyses above.

Analysis of the *F*<sub>*ST*</sub> values on a nucleotide site-by-site basis reveals that they are nonrandomly distributed among the polymorphic sites, and runs of sites with very low values occur within both domains I and III. The resulting fine-resolution map of values among sites may allow the identification of key sites under immune selection, although due to the effects of close physical linkage it is likely that not all the sites within a run are under selection. The variability in the *F*<sub>*ST*</sub> values also emphasizes the importance of full sequence data for this kind of analysis. An analysis employing a limited subset of polymorphic sites sampled across the gene could easily miss informative signatures of selection.

A previous study of *ama1* polymorphism in a Nigerian population concluded with a prediction that there are protective human immune responses to domains I and III (Polley and Conway 2001). There is now evidence...
that antibodies to polymorphic epitopes in AMA1 can indeed inhibit invasion of \textit{P. falciparum} \textit{in vitro} (Hodder et al. 2001; Kennedy et al. 2002; Cocken et al. 2002), and examination of sequence differences between cultured parasite lines suggests that polymorphic sites within domain I in particular are targets of inhibitory antibodies (Hodder et al. 2001). It may be suggested that polymorphic changes in domain III have been selected due to their ability to compensate for some loss of function induced by changes within domain I (or vice versa). However, no obvious patterns of linkage disequilibrium between sites in domains I and III exist, and there is no significant correlation between the overall charge of the polymorphic residues within these two domains among the different alleles (Spearman’s rank correlation = 0.326; \( P = 0.097 \)). It would appear, therefore, that both domains are independently under balancing selection.

This study provides new data and analyses of a Thai population, together with a novel between-population analysis, which strongly support and extend the evidence for balancing selection on particular sites in domains I and III of the \textit{ama1} gene. Laboratory and field studies of immune responses to AMA1 provide a means of assessing the relative importance of these sequences in acquired immunity. Such studies are intended to allow the rational identification of AMA1 sequences to be represented in a multiallelic vaccine, which this study predicts would be relevant for use in different endemic populations. Thus allele frequency-based analyses offer a powerful and robust way of identifying sequences under selection in vaccine candidate antigen genes of malaria parasites and may be suitable for other endemic pathogens subject to recombination (Stothard et al. 1998; Fuduy et al. 1999; Gubbels et al. 2000; Rannala et al. 2000; Suarez et al. 2000). Such analyses can also be applied to screen for signatures of selection on a genomic scale (Akey et al. 2002).

We thank colleagues from the Faculty of Tropical Medicine in Mahidol University, Bangkok, for sample collection and Neil Alexander for helpful discussions. This work was supported by the UK Medical Research Council (grant G9803180) and the World Health Organization Special Program for Research and Training in Tropical Diseases (WHO/TDR Research Training grant to W.C).

**LITERATURE CITED**


Jiggins, F. M., G. D. Hurst and Z. Yang, 2002 Host-symbiont con-


Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated pro-

gram for molecular population genetics and molecular evolution analysis. Bioinformatics 15: 174–175.


Stothard, D. R., G. Boguslawski and R. B. Jones, 1998 Phyloge-


Zhang, J., Y.-P. Zhang and H. F. Rosenberg, 2002 Adaptive evolu-