Gene Conversion as a Source of Nucleotide Diversity in *Plasmodium falciparum*

Kaare M. Nielsen,† Jacob Kasper,* Mehee Choi,* Trevor Bedford,* Kurt Kristiansen,† Dyann F. Wirth,‡ Sarah K. Volkman,‡ Elena R. Lozovsky,* and Daniel L. Hartl**

*Department of Organismic and Evolutionary Biology, Harvard University; †Department of Pharmacy, Faculty of Medicine, University of Tromsø, and Norwegian Institute of Gene Ecology, Tromsø, Norway; and ‡Department of Immunology and Infectious Diseases, Harvard School of Public Health

Examination of polymorphisms in the *Plasmodium falciparum* gene for falcipain 2 revealed that this gene is one of two paralogs separated by 10.8 kb in chromosome 11. We designate the annotated gene denoted chr11.gen_424 as encoding falcipain 2A and the annotated gene denoted chr11.gen_427 as encoding falcipain 2B. The paralogs are 96% identical at the nucleotide level and 93% identical at the amino acid level. The consensus sequences differ in 31/309 synonymous sites and 45/1140 nonsynonymous sites, including three amino acid replacements (V393I, A400P, and Q414E) that are near the catalytic site and that may affect substrate affinity or specificity. In six reference isolates, among 36 synonymous sites and 46 nonsynonymous sites that are polymorphic in the gene for falcipain 2A, falcipain 2B, or both, significant spatial clustering is observed. All but one of the polymorphisms appear to result from gene conversion between the paralogs. The estimated rate of gene conversion between the paralogs may be as many as 1,400 to 1,700 times greater than the rate of mutation. Owing to gene conversion, one of the falcipain 2A alleles is more similar to the falcipain 2B alleles than it is to other falcipain 2A alleles. Divergence among the synonymous sites suggests that the paralogous genes last shared a common ancestor 15.2 MYA, with a range of 8.8 to 20.6 MYA. During this period, the paralogs have acquired 0.10 synonymous substitutions per synonymous site in the coding region. The 5′ and 3′ flanking regions differ in 47.7% and 39.8% of the nucleotide sites, respectively. Hence synonymous sites and flanking regions are not conserved in sequence in spite of their high AT content and T skew.

Introduction

Microbial drug resistance, emerging diseases, and bioterrorism have all underlined the importance of understanding evolutionary processes in predicting and controlling the spread of infectious disease (Lederberg 2000). Among the most economically devastating and recalcitrant parasitic diseases in humans is falciparum malaria (Phillips 2001), caused by the protozoan *Plasmodium falciparum*. This organism has been remarkably capable of producing genetic variability affecting its responses to current therapeutic targets (World Health Organization 2000). Understanding the mechanisms generating DNA sequence variation and its differential distribution among parasite populations is important for assessing responses to drug and immunological control strategies (Hartl et al. 2002).

Analysis of genetic variability among isolates of *P. falciparum* reveals anomalously low levels of polymorphisms of synonymous sites and introns (Rich et al. 1998; Volkman et al. 2001). This finding has been interpreted to imply that the parasite underwent a severe bottleneck in population size in the human population during the last 10,000 to 40,000 years, resulting in a relatively recent common ancestor of all extant *P. falciparum*. Nevertheless, many genes in *P. falciparum* have high levels of nonsynonymous single-nucleotide polymorphisms (SNPs) (Verra and Hughes 2000; Hughes and Verra 2001), which is apparently inconsistent with the recent common ancestor hypothesis. Other analyses of SNPs have suggested an age for the most recent common ancestor of 100,000 to 200,000 years (Mu et al. 2002), 300,000 to 400,000 years (Hughes and Verra 2001), or even older (Hughes 1999).

Critical assumptions in the use of SNP polymorphisms for estimating the age of the most recent common ancestor are (1) that the SNPs undergo spontaneous mutation independently and at the same rate and (2) that the SNPs undergo mutation and random genetic drift as selectively neutral genetic markers. In this paper, we demonstrate that some polymorphisms in *P. falciparum* are in gross violation of the first assumption. In particular, we have determined the distribution of polymorphisms within two of the genes encoding the major cysteine proteases responsible for hemoglobin degradation in infected red blood cells (Rosenthal et al. 1988; Francis, Sullivan, and Goldberg 1997). Close examination of 12 kb of chromosome 11 in a region encoding falcipain 2 and falcipain 3 unexpectedly revealed a previously unrecognized paralog closely related to falcipain 2. Here, we designate the products of these paralogs as falcipain 2A and falcipain 2B. Analysis of synonymous differences between the paralogs suggests that the duplication occurred a minimum of 10 to 20 MYA, considerably prior to the divergence between *P. falciparum* and the chimpanzee parasite *P. reichenowi*.

Although both the falcipain 2A and falcipain 2B genes are highly polymorphic among natural isolates, we find that virtually all of the polymorphisms result from gene conversion between the paralogs. Among 36 synonymous and 46 nonsynonymous nucleotide sites that are polymorphic in the gene for falcipain A, falcipain B, or both, all except possibly one are consistent with an origin by gene conversion. This result is supported not only by the identity of the polymorphic nucleotides but also by the significant clustering of polymorphic nucleotides within each gene. Positioning of the polymorphisms within derived three-dimensional models of the mature protease

Key words: falcipain, gene conversion, paralogs, *Plasmodium falciparum*, polymorphisms.

E-mail: dhartl@oeb.harvard.edu.

Mol. Biol. Evol. 20(5):726–734. 2003
DOI: 10.1093/molbev/msg076

indicated that three of the five amino acid differences present in the mature protein between falcipain 2A and falcipain 2B are located near the predicted active site. All three of these are polymorphic in one or the other of the paralogs, and hence these polymorphisms may have some functional significance. These data imply that gene conversion between paralogs can be a powerful mechanism generating genetic variability among lineages of *P. falciparum*.

Materials and Methods

Strains and Strain Verification

DNA was studied from six isolates of *P. falciparum* representing accessions from throughout the world: HB3 (Honduras), 7G8 (Brazil), D6 (Sierra Leone), W2 (Laos), Muz12.4 (Papua New Guinea), and 3D7 (The Netherlands). The 3D7 strain is also the reference strain for genomic sequencing. Parasites were maintained in culture for short periods using standard culture techniques as described (Volkman et al. 2001). The isolates differ from each other by numerous microsatellite markers, and diagnostic microsatellite markers were typed as a control to detect possible cross contamination. DNA was isolated from the cultures as described (Volkman et al. 2001).

Primer Design, PCR Amplification, and Sequencing

Primers were designed based on the published 3D7 sequences as in PlasmoDB (The Plasmodium Genome Database Collaborative 2001) using Primer 3 software written by S. Rozen and H. J. Skaletsky and available at http://www-genome.wi.mit.edu/ftp/distribution/software. Primers used for amplification and sequencing correspond to the PlasmoDB coordinates in chromosome 11 indicated below. The italicized primers were used to specifically amplify each gene, and they and the others were used for DNA sequencing. The amplifications are specific because the amplification primers match the flanking sequences, which have little similarity between the paralogs. For the gene for falcipain 2A, the primers were the sequences between coordinates (590813, 590838), (578933, 578959) and (*589616*, *589642*), and complements of the sequences between coordinates (*591210*, *591185*), (580063, 580038) and (590567, 590541). For the gene for falcipain 2B, the primers were the sequences between coordinates (579655, 579681), (579339, 579365) and (*578826*, *578852*), and complements of the sequences between coordinates (*580301*, *580279*), (579754, 579729), and (579440, 579414).

PCR amplifications using the HotStarTaq DNA Master mix (Qiagen, Calif.) with 30 cycles at 2 min at 94°C, 1 min at 57°C to 62°C (usually 58°C), and 1 min 20 s at 73°C. Single PCR bands were purified for direct sequencing using shrimp alkaline phosphatase and exonuclease I (USB Biochemicals). To guard against polymerase incorporation error, products from two separate PCR amplifications were sequenced in both directions on a 3100 DNA capillary sequencer (Applied Biosystems) using Big Dye chemistry. Editing of raw chromatograms and multiple sequence alignments were carried out using the Sequencher v.3.1.1 software (GeneCodes).

Molecular Modeling

Computations

Energy minimizations and molecular dynamics simulations were performed using the all-atomic force field of the AMBER (Assisted Model Building with Energy Refinement) 6.0 programs (Pearlman et al. 1995). Models of falcipains 2A and 2B were refined with 500 steps of steepest-descent minimization, followed by conjugate gradient energy minimization until convergence with a 0.005 kcal mol^{-1} $Å^{-1}$ root mean square energy gradient difference between successive minimization steps. Molecular dynamics simulations were performed for 100 ps at 300 K. The bond lengths involving hydrogens were constrained by using the SHAKE algorithm, allowing an integration step of 0.001 ps. A distance-dependent dielectric function ($e = r_{ij}$) and a 10 Å cut-off radius for nonbonded interactions were used.

The substrate benzyloxycarbonyl-Phe-Arg-7-amino-4-methyl coumarin (Z-FR-AMC) was geometry-optimized using AM1. HF/6-31G* single-point calculation with the program Gaussian 98 (Gaussian, Carnegie, Penn.) was performed in order to obtain electrostatic potentials for the AM1 geometry-optimized structure. The atomic point charges used for the molecular mechanics and dynamics calculations were derived from electrostatic potentials using the RESP (restrained electrostatic potential) program implemented in AMBER 6.

The ICM (Internal Coordinate Mechanics) 2.8 program (Molsoft, La Jolla, Calif.) was utilized for comparative modeling, structure alignments, computer graphics visualizations, substrate docking, and calculations of molecular surfaces and electrostatics potentials.

Falcipain 2A and 2B Structures

The mature form of *P. falciparum* falcipain 2A includes only residues Q244 to E484 (Shenai et al. 2000; Sijwali, Shenai, and Rosenthal 2002). Compared with other cysteine proteases, the mature form of falcipain 2A includes 17 additional amino-terminal residues (Q244 to F260) that are important for proper folding (Sijwali, Shenai, and Rosenthal 2002). An initial model of the mature form of falcipain 2A, including residues A263 to E484, was built by comparative modeling using the homology module of ICM (Abagyan, Totrov, Kuznetsov 1994). The crystal structure of human cathepsin V associated with the irreversible vinyl sulfone inhibitor APC-3316 (pdb code: 1FH0) (Somoza et al. 2000) was selected as a template due to its high sequence similarity with falcipain 2A (36% identity over 230 residues).

Disulphide bridges were introduced between Cys282 and Cys323, Cys316 and Cys357, Cys342 and Cys362, and Cys411and Cys472. The Cys342 to Cys362 disulphide bridge was conserved among *P. falciparum* falcipains 1, 2A, 2B, and 3, whereas the three other disulphide bridges were conserved among cysteine proteases. Due to the acid environment in the *P. falciparum* food vacuole, all histidine residues in the

model (H270, H417, and H440) were modeled using the AMBER HIP residue with protonated delta and epsilon nitrogen atoms. The catalytic H417 and C285 residues were modeled as an imidazolium-thiolate ion pair. Iterative energy refinement was performed using the refineProtein macro in ICM, and the resultant model was further energy minimized using the AMBER force field.

Models of polymorphisms in the mature forms of falcipain 2A and 2B were constructed by mutating residues in the energy-minimized models of the consensus sequences of falcipains 2A and 2B. The models of the mutant proteins were energy minimized using the AMBER force field. By using ICM, substrate Z-FR-AMC was placed at its putative binding site in the falcipain 2A model utilizing knowledge from X-ray structures of inhibitor-cysteine protease complexes. The substrate-falcipain 2A complex was energy minimized and the energy-minimized structure was used as a starting structure for a 100 ps molecular dynamics simulations in which the protein backbone atoms were kept in fixed positions. The conformation observed at 100 ps was energy minimized. Near the active site, falcipain 2B differs from falcipain 2A by three amino acid replacements (V393I, A400P, and Q414E). A model of falcipain 2B was constructed by introducing these three amino acid replacements in the energy-minimized model of the falcipain-2A-Z-FR-AMC complex, and the resultant model was energy minimized.

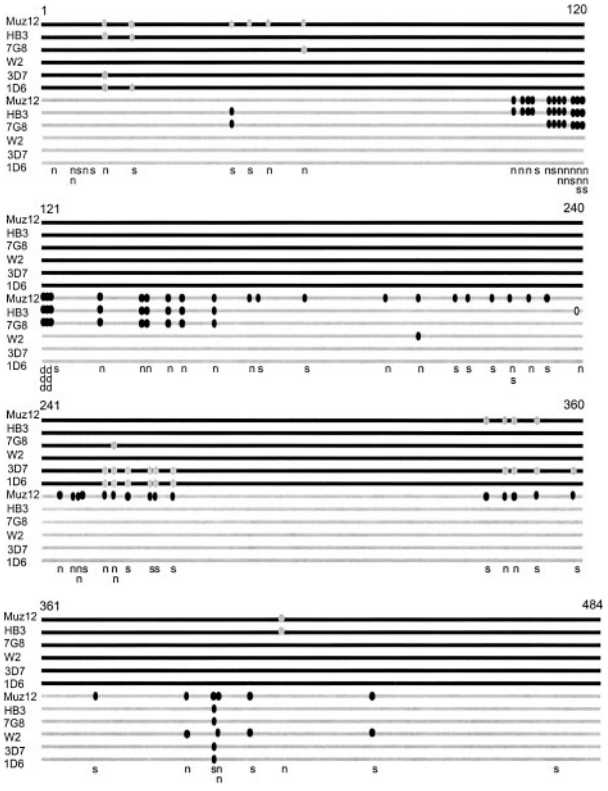

FIG. 1.—Polymorphic codons (circles) in the coding sequences of falcipain 2A and 2B in six reference strains. Falcipain 2A is in black, and falcipain 2B is in gray. Polymorphisms matching the paralogous sequence are colored gray or black, respectively. Nucleotide sites that differ between the consensus sequences are shown below the line, classified as nonsynonymous (n), synonymous (s) or deletion/insertion (d).

Results

The Falcipain 2A and Falcipain 2B Paralogs

Only one copy of falcipain 2 has previously been described (Hanspal 2000; Shenai et al. 2000) and deposited as GenBank accession numbers AF239801 and AF251193. This copy corresponds to what we have designated as the paralog falcipain 2A. The coding sequence of falcipain 2A begins at coordinate 591,186 (annotated in PlasmoDB as chr11.gen_424) and the coding sequence of falcipain 2B begins at coordinate 580,312 (annotated as chr11.gen_427) (The Plasmodium Genome Database Collaborative 2001). The paralogs are oriented in the same direction, with the gene for falcipain 3 (chr11.gen_426) between them. The gene for falcipain A is two codons longer (codons 121 and 122) than that for falcipain B, owing to an insertion/deletion difference. The aligned coding sequences are 96% identical at the nucleotide level and 93% identical at the amino acid level. The consensus sequences differ in $ds = 31/309 = 0.100$ synonymous substitutions per synonymous site, and differ in $dn = 45/1140 = 0.039$ nonsynonymous substitutions per nonsynonymous site. Because $dn/ds = 0.39$, the falcipain paralogs do not exhibit the pattern of a high dn/ds ratio that has been attributed to the high AT content of synonymous sites (Mu et al. 2002). (Here and elsewhere in this paper, we use the term consensus sequence to refer to the inferred sequences of falcipain 2A or 2B before the postulated gene-conversion events.)

Evidence for Gene Conversion

To capture geographical polymorphisms, the falcipain 2A and 2B genes were sequenced from *P. falciparum* strains originating in The Netherlands (3D7), Sierra Leone (D6), Laos (W2), Honduras (HB3), Brazil (7G8), and Papua New Guinea (Muz12). Among these strains we found 10 synonymous polymorphisms in falcipain A and 25 synonymous polymorphisms in falcipain B, and we found nine nonsynonymous polymorphisms in falcipain A and 37 nonsynonymous polymorphisms in falcipain B. The locations of the polymorphisms across the codons 1 to 484 in the gene are shown in figure 1. The falcipain A sequences are shown as black lines and those of falcipain B are shown as gray lines. Underneath the lines are designations *n* for nonsynonymous differences between the paralogous sequences, *s* for synonymous differences, and *d* for deletion in falcipain B (equivalently insertion in falcipain A). The vertical arrays of up to three symbols indicate multiple differences in a single codon.

The positions of the polymorphisms in each paralog are indicated by filled ovals. Polymorphisms in falcipain A that match the consensus sequence in falcipain B are denoted in gray, and polymorphisms in falcipain B that match the consensus sequence in falcipain A are denoted in black. Among 36 synonymous polymorphisms and 46 nonsynonymous polymorphisms in either the gene for falcipain A, the gene for falcipain B, or both, only one fails to match the consensus sequence of the paralog. This exception is in codon 239, where the consensus codon for

Table 1
Estimated Rates of Mutation, Gene Conversion, and Recombination

Polymorphic Site	Silent	Replacement	All Polymorphisms
π_w[a]	0.0247 ± 0.0053	0.0097 ± 0.0013	0.0129 ± 0.0021
π_b[a]	0.0697 ± 0.0050	0.0247 ± 0.0023	0.0339 ± 0.0028
d[a]	0.00022 ± 0.00032	0.00000 ± 0.00008	0.00005 ± 0.00012
θ[a]	0.0140 ± 0.0028	0.0050 ± 0.0006	0.0068 ± 0.0011
C[a]	0.271 ± 0.084	0.317 ± 0.074	0.294 ± 0.080
R[a]	15.1 ± 1.0	42.1 ± 25.7	35.2 ± 20.2
c/μ[a]	19.4	74.2	43.2
c/μ[b]	1,389 ± 266	1,698 ± 378	1,543 ± 344

NOTE.—Mean ± standard deviation of the mean estimated using the jackknife procedure (Efron and Stein 1981). The estimates of R exclude jackknife samples in which R is indeterminate owing to values of d excessively close to 0, and therefore the values of R should be regarded as minimum.

[a] Equilibrium calculation (polymorphisms/total): silent = 36/309, replacement = 46/1140, all polymorphisms = 82/1449.

[b] Nonequilibrium calculation (t = 45,000 generations; $\mu = 3.24 \times 10^{-9}$, r = 0.0044).

falcipain A and B are both AAA, but isolate HB3 has codon AGA (empty oval in figure 1).

The possibility that PCR artifacts account for the sequence polymorphisms in figure 1 can be ruled out for three reasons. First, the primers are specific to the divergent flanking sequences and hence specific to each of the paralogs. Second, the sequence of each paralog from each isolate is reproducible from one PCR reaction to the next. Third, note that the polymorphisms do not include the differences in the consensus sequences at each end of the paralogs. This spatial distribution rules out PCR artifacts as the cause of the differences between the sequences, since attributing them to recombinant PCR due to a change of template would require the occurrence of two template-switching events at specific sites in each independent PCR reaction.

Two lines of evidence support gene conversion as the source of most or all the polymorphisms in falcipain 2A and 2B, except that in codon 239. The first is based on sequence. With the exception of codon 239, all of the polymorphisms in each paralog are exact matches to the consensus sequence of the other paralog. The second line of evidence is that the polymorphisms are significantly clustered. For example, the 18 polymorphic codons in falcipain A include two runs, each six codons in length, that contain all the sites at which the paralogous consensus sequences differ. These runs comprise codons 15 to 59 in Muz12 and codons 255 to 266 in 3D7 and 1D6. Elementary combinatorial considerations indicate that the probability of a run of six such sites by chance alone is 0.012. Similarly, the falcipain B of Muz12 has a run of 52 codons from 105 to 434 in which all of the nucleotide sites match the consensus sequence of the gene for falcipain 2A. The probability of a run of this length by chance alone is 1.03×10^{-11}. The isolates with polymorphisms that are subsets of significant runs are most easily explained as resulting from recombination within each paralog after each gene-conversion event, rather than as resulting from multiple, independent gene-conversion events.

Estimates of the Rate of Gene Conversion

We used the theory of Innan (2002) to estimate the rates of mutation, gene conversion, and recombination affecting the paralogous falcipain 2A and 2B genes. The parameters in the theory are $\theta = 4N\mu$, where N is the effective population number and μ is the single-nucleotide mutation rate; $C = 4Nc$, where c is the rate of gene conversion between the paralogs; and $R = 4Nr$, where r is the rate of recombination between the paralogs. The data for the model fitting consist of π_w, the average number of pairwise differences within paralogs; π_b, the average number of pairwise differences between the paralogs; and d, the linkage disequilibrium between the paralogs (Innan 2002).

Innan's (2002) theory is an equilibrium theory, and the estimates assuming equilibrium are given in the top part of table 1, along with the standard deviations of the estimates calculated by means of jackknifing (Efron and Stein 1981). The rate of gene conversion, relative to that of mutation, is c/μ and estimated as C/θ. These estimates range from 19.4 to 74.2, depending on whether only synonymous or only nonsynonymous sites are analyzed. Over all sites, the ratio $c/\mu = 43.2$.

However, if *P. falciparum* has a relatively recent common ancestor, as some authors propose (Rich et al. 1998; Volkman et al. 2001), then the equilibrium calculation is not valid and c/μ may be seriously underestimated. An indication that the population is probably not in equilibrium may be seen by considering the estimated values of θ in table 1, which are approximately 10 times larger than the estimates of θ across chromosome 3 based on DNA sequencing (Mu et al. 2002).

To address the likelihood of nonequilibrium, we also estimated the value of c/μ without assuming equilibrium by making use of Innan's (2002) iterative equations. For this estimation, we hoped to bracket the true value of c/μ by using the smallest plausible estimate of the most recent common ancestor. Taking this time as 7,500 years with six generations per year yields 45,000 generations for the iterations. We also assumed $\mu = 3.24 \times 10^{-9}$, which is the average of the estimated mutation rate for fourfold degenerate synonymous sites assuming either a 5-Myr or a 7-Myr time for the divergence between *P. falciparum* and *P. reichenowi* (Rich et al. 1998). Finally, we assumed r = 0.0044 based on 1 cM per 15 to 30 kb (Su et al. 1999) across the approximately 10.8 kb between the paralogs. With these parameters, the iterative equations of Innan (2002) yield

Table 2
Estimated Time of the Falcipain 2A and Falcipain 2B Gene Duplication

Interspecific Comparison	Mutation Rate ($\times 10^{-9}$)	Number of Sites	S	MRCA ($\times 10^6$ Years)
P. berghei versus *P. falciparum*				
55 MYA divergence	$\mu_2 = 2.22$	688	16	
	$\mu_4 = 7.12$	250	13	8.8
129 MYA divergence	$\mu_2 = 0.95$	688	16	
	$\mu_4 = 3.03$	250	13	20.6
P. reichenowi versus *P. falciparum*				
5 MYA divergence	$\mu_2 = 1.86$	688	16	
	$\mu_4 = 3.78$	250	13	13.0
7 MYA divergence	$\mu_2 = 1.33$	688	16	
	$\mu_4 = 2.70$	250	13	18.3
Average of all estimates				15.2

estimates of c/μ in the range 1,389 to 1,698 (table 1). These are approximately 20 to 70 times greater than the equilibrium estimates. If the most recent common ancestor of *P. falciparum* is older than the 7,500 years assumed here as the minimum plausible value, then the true value of c/μ may be smaller than 1,000. In any case, it is clear that across the 10.8 kb separating the falcipain 2A and 2B paralogs, gene conversion has been a much more potent process for generating variation than has been mutation.

Estimated Age of the Paralogs

The synonymous differences between the consensus sequences of the paralogous genes can be used to estimate the most recent common ancestor (MRCA in table 2) of the genes. This is not necessarily the time at which the original gene duplication occurred. It is rather the time since the paralogs were last identical in sequence. Identity could be created if a large gene conversion in a lineage rendered the paralogs identical, and then this chromosome went to fixation in the population. In other words, comparison of the consensus sequences cannot distinguish between the most recent complete homogenization due to gene conversion and the time of the original gene duplication.

The estimated age of the most recent common ancestor of the falcipain 2A and 2B paralogs ranges from 8.8 to 20.6 MYA, with an average across all estimates of 15.2 MYA (table 2). The estimation procedure used in table 2 follows that of Rich et al. (1998) and makes use of estimates of the mutation rates of twofold (μ_2) and fourfold (μ_4) degenerate sites. The latter estimates are obtained by comparisons of coding sequences between *P. falciparum* and either *P. reichenowi* (a chimpanzee parasite) or *P. berghei* (a rodent parasite), taking into account the uncertainty in the time of divergence of these species (5 to 7 MYA for the former species pair and 55 to 129 MYA for the latter). In table 2, the numbers 688 and 250 refer to the number of twofold and fourfold degenerate sites in both paralogs together, and S is the number of twofold or fourfold degenerate sites at which the paralogs differ.

Gene Trees of the Falcipain 2A and Falcipain 2B Paralogs

One interesting result of gene conversion is its effect on inferred gene trees. This effect is shown in figure 2, which shows the Neighbor-Joining tree for falcipain 2A and 2B and bootstrap values among 1,000 replicates, rooted using the sequence of falcipain 3. The falcipain 2B allele in Muz12 is included among the falcipain 2A alleles with 100% bootstrap support. The misplacement occurs because of the large gene conversion event affecting falcipain 2B in isolate Muz12 (fig. 1). This result has important implications for the possible origin of highly divergent sequences in an organism with a relatively recent common ancestor (see *Discussion*).

Structural Modeling of Falcipain 2A and Falcipain 2B

Models of the mature forms of falcipain 2A and falcipain 2B were constructed by comparative modeling using a crystal structure of human cathepsin V as a template. The consensus falcipain 2A and 2B amino acid sequences differ at 10 positions within their mature form, of which five are in a region known to be important for protein folding. To avoid confusion, we adopt the convention that in the designations of amino acid residues,

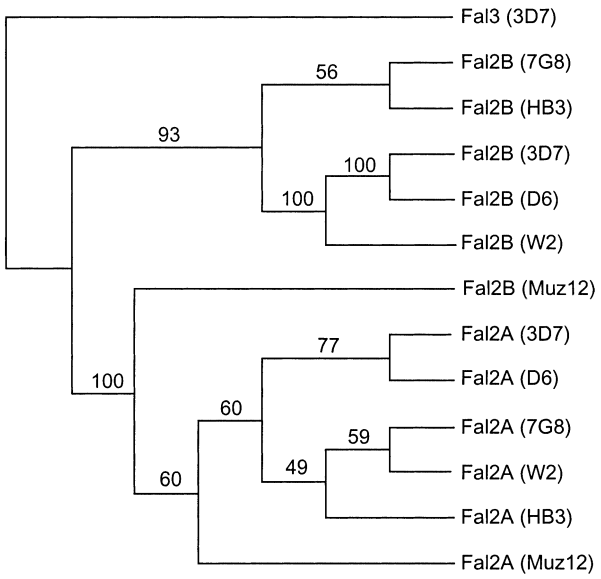

FIG. 2.—Neighbor-Joining tree and bootstrap values in 1,000 replicates for falcipain 2A and 2B alleles in six reference strain. The tree is rooted with falcipain 3.

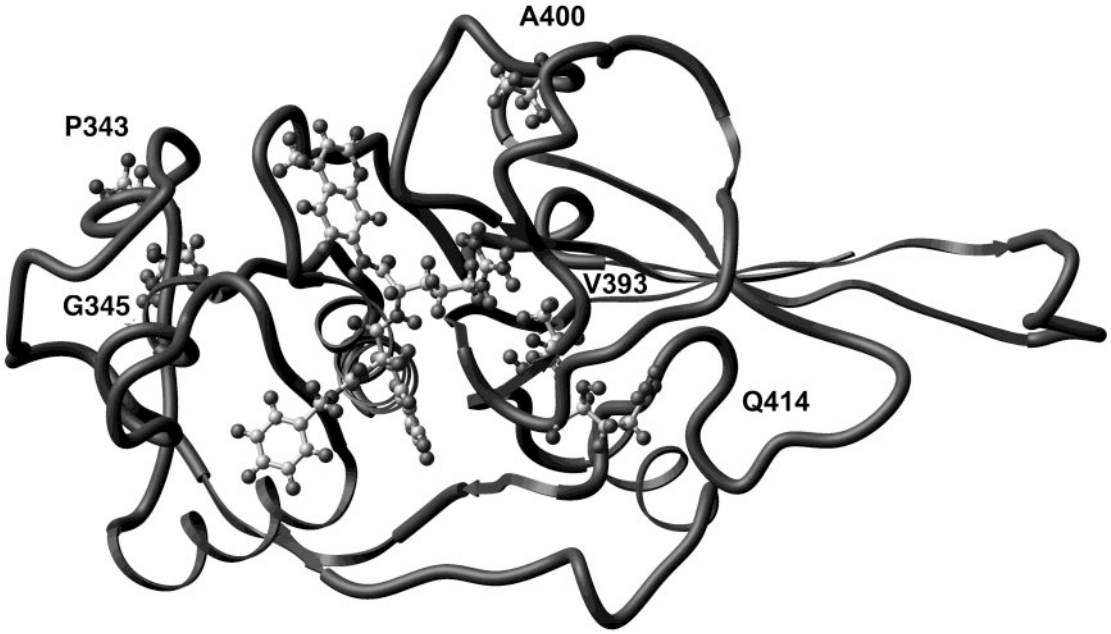

FIG. 3.—Location of replacement differences in the consensus sequences of the mature proteins of falcipain 2A and 2B. Only the falcipain 2A residue is indicated. The model also includes the substrate Z-FR-AMC.

the symbol preceding the residue number in the aligned sequences represents the amino acid present at that site in the consensus sequence of falcipain 2A, and the symbol following the residue number is the amino acid present at that site in the consensus sequence of falcipain 2B.

The mature form of falcipain 2A consists of residues Q244 to E484 (Shenai et al. 2000; Sijwali, Shenai, and Rosenthal 2002). Across this region of the protein, the consensus sequences of falcipain 2A and 2B differ in 10 residues. Five of the residues (M245I, E248D, E249A, R255L, and E257N) are in the region Q244 to F260 that is important for proper folding (Sijwali, Shenai, and Rosenthal 2002).

The localization of the remaining five differences are shown in figure 3. The two differences on the left (P343T and G345D) are localized in a loop region far away from the catalytic site. The remaining three differences between falcipain 2A and 2B (V393I, A400P, and Q414E) are localized in close proximity to residues that are predicted to interact with the substrate Z-FR-AMC (D413, D397, D398, L415, and S392). As noted in the *Discussion*, the difference Q414E is a candidate for affecting substrate affinity and specificity.

Flanking Sequences of the Falcipain Paralogs

We noted in connection with figure 1 that none of the gene conversion events includes the consensus sequence differences nearest the amino end or carboxyl end of the protein. This may be because the flanking sequences of the paralogs are quite divergent. They are divergent in spite of the fact that both paralogs are expressed in the erythrocytic stage, as indicated by reverse-transcriptase PCR (data not shown). An alignment of 107 bp of the 5′ flanking region and 103 bp of the 3′ flanking region is shown in figure 4*A*. The Xs show the position of the coding sequences. Although there is sufficient conservation to support the alignment (raised dots), there are clearly many differences. Matrices comparing the nucleotides in falcipain 2A and 2B are shown in figure 4*B* for the 5′ and 3′ flanking regions separately. The AT-richness of both flanking sequences is evident. On the other hand, the sequences are far from conserved. Among 96 sites in the 5′ region in which one or both sequences have an A or a T, 39 of the sites (41%) are discordant (occupied by A in one sequence and T in the other). Similarly, among 75 such sites in the 3′ flanking region, 27% are discordant.

Discussion

Implications for Single-Nucleotide Polymorphisms

Prior to the recognition of the falcipain 2B paralog, falcipain 2A was noted as being moderately polymorphic, and the polymorphisms in the gene were attributed to mutation-selection-drift processes occurring in the gene itself (Hughes and Verra 2001). These polymorphisms contributed to the inference that *P. falciparum* must have an ancient common ancestor and a large effective population number. The results presented here show that most or all of the polymorphisms in falcipain 2A actually result from gene conversion between it and the previously unrecognized paralog falcipain 2B. The most recent common ancestor of these paralogs has an estimated age of 8.8 to 20.6 MYA, with an average age of 15.2 MYA. Hence, the origin of the paralogs predates the human-chimpanzee divergence, and the paralogs have been accumulating genetic differences for a very long time. The differences between falcipain 2A and falcipain 2B that, through gene conversion, give rise to polymorphisms in falcipain 2A and falcipain 2B are therefore very ancient.

However, the fact that they are ancient does not imply that *P. falciparum* has an ancient common ancestor. Rather, the polymorphisms are ancient because the falcipain paralogs have an ancient common ancestor, and both paralogs passed through whatever bottleneck affected the species as a whole.

That gene conversion may play a potentially important role in generating polymorphisms among related sequences is indicated by the results in table 1. Although the falcipain paralogs are separated by a distance of 10.8 kb, the rate of gene conversion between them is estimated from non-equilibrium theory to be approximately 1,400 to 1,700 times greater than the rate of mutation (table 1). This may be somewhat of an overestimate because it is based on the assumption that *P. falciparum* shared a most recent common ancestor 7,500 years ago, but even the equilibrium theory implies that the rate of gene conversion relative to mutation is 20 or more (table 1). Gene conversion as a source of genetic variation in *P. falciparum* may therefore be more important than previously realized. Especially among the AT-rich sequences in 5′ untranslated regions, 3′ untranslated regions, and introns, there may be regions close enough together and with sufficient similarity to serve as templates for gene conversion. If these regions are short, they may be difficult to detect, but one indication of possible gene conversion would be significant clustering of polymorphisms within a sequence.

Implications for Selective Constraints Due to High-AT Content

The divergence at synonymous sites between the falcipain paralogs strongly suggests that there are no exceptionally strong constraints on the nucleotides at synonymous sites as a result of high AT content or T skew (Jongwutiwes et al. 2002; Mu et al. 2002). This is also true of the 5′ and 3′ flanking sequences (fig. 4). The flanking sequences are very AT rich, but they are not conserved. Although the falcipain paralogs have been evolving for a very long time to accumulate so many differences, the large number of differences suggests that the lack of polymorphisms observed in the introns and flanking sequences of other genes cannot easily be attributed to their skewed AT content alone (Jongwutiwes et al. 2002).

Implications for Highly Divergent Genes

The result in figure 2 is potentially important in understanding how certain alleles in *P. falciparum* can be extremely divergent, even though the population may have gone through one or more recent bottlenecks in population size. In particular, two highly divergent alleles may originally have been highly divergent paralogs (Hartl et al. 2002). Both paralogs could have passed through a bottleneck, even one as extreme as a single haploid genome. Afterward, if one lineage lost one paralog and another lost the other paralog, and these deletions replaced the duplication in the population, the coding sequences in the deleted chromosomes would once again become alleles, but extremely divergent alleles.

A

```
Fal2A TTTTTTTATA CAAATATATG TATATATTTT TTTTATTTTT CATTTAACTA ATTCATTTTT
Fal2B ·A·A·A···· T·T·····A ·G·····A·· A·AAGAA·AC ·····TTGT TAATT···A·

Fal2A TTTATTTTTT ATTTTTTATT ATTTTATTTT TATTTTTTTG TTGAAGAXXX XXXXXXXXX
Fal2B A·AT··A·A· TAA·A·ATA· ··ATATATA· ·T·····T ··TTTA·XXX XXXXXXXXX

Fal2A XXXXXXXXXX XXXXXXXATA TGTCAATAGT ATTTTAAAAA GGGACCTATC TTAAATAAAT
Fal2B XXXXXXXXXX XXXXXXX··· ·A·G·G···· ·····G ···GA····· A·····

Fal2A AAGTAATAAA ATTTTGGGAA TATTTTATAT AATTTTACGA TTTTACATAA AATATGAAGA
Fal2B ·TA··AT··· TAAGATAA·· ·T··AGGA·· CT····TAAG ··C···GATT TTACAT····
```

B

Fal2B 5′ flanking \ Fal2A 5′ flanking	A	T	G	C
A	12	27	1	0
T	12	45	2	3
G	1	1	1	0
C	0	1	0	1

Fal2B 3′ flanking \ Fal2A 3′ flanking	A	T	G	C
A	30	11	4	2
T	9	25	3	0
G	6	2	5	1
C	2	1	0	1

FIG. 4.—(*A*) Alignment of the 5′ and 3′ flanking regions of the genes for falcipain 2A and 2B. The coding sequence is indicated by the Xs. (*B*) Matrix of differences in the 5′ and 3′ regions.

A possible example of this process is found in the gene encoding merozoite surface protein 1 (*Msp1*). Based on the extreme divergence of the *MAD* and *K1* types of *Msp1* alleles, Hughes (1999) has estimated that they may have diverged from a common ancestor about 48 MYA. This would make the Msp1 polymorphism the oldest known polymorphism in any organism, and such an ancient divergence seems to be completely inconsistent with a recent common ancestor for *P. falciparum*. To account for such an apparently extreme age, we have suggested that the MAD and K1 alleles may originally have been highly divergent paralogs (Hartl et al. 2002). Both paralogs could have passed through a bottleneck, and then one lineage lost the MAD paralog and another lost the K1 paralog.

Implications for Protein Structure

The biological significance of the 26 amino acid replacements in the consensus sequences for the prodomains of falcipain 2A and 2B is unclear because of the unique N-terminal region for which structural models are not yet available. All of the falcipains have prosequences 2 to 3 times the length of other known papain family proteases. The prosequence contains a putative type II membrane-spanning domain. Similar proforms of the proteases have been identified in the aspartic proteases, which also are active in red blood cells during hemoglobin degradation (Francis, Sullivan, and Goldberg 1997; Banerjee et al. 2002). The prodomain has been reported to be a potent, reversible inhibitor of mature falcipain-2 (Sijwali, Shenai, and Rosenthal 2002), possibly involved in the controlled release of parasites from red blood cells through skeletal protein degradation (Dua et al. 2001; Sijwali et al. 2001). One might speculate that the differences in the prodomain might affect activation of the enzyme for hemoglobin degradation and control of parasite release, as well as confer resistance to drugs involved in the hemoglobin degradation pathway.

The two differences P343T and G345D affecting the loop in figure 3 may also affect protein conformation. Inspection of a structural alignment of cysteine proteases

including human cathepsins L, K, and V, *Zingiber officinale* protease II, *Trypanosoma cruzi* cruzain, *Carica papaya* papain, porcine cathepsin H, and *Actinidia chinesis* actinidin revealed that the conformation of the first part of this loop (corresponding to D344 to S351 in falcipain 2A) was conserved among cysteine proteases. It is possible that the nonconservative substitutions in the P343-D344-G345-D346 motif in falcipain 2A may affect the local loop conformation as well as the folding of the whole protein.

The differences between falcipain 2A and 2B that are near the catalytic site (V393I, A400P, and Q414E in figure 3) are likely to be important. These residues are located in close proximity to residues that are predicted to interact with the substrate Z-FR-AMC (D413, D397, D398, L415, and S392). The position of Q414 in falcipain 2A is near residues D413 and L415 that are predicted to interact directly with the substrate. It is therefore possible that the Q414E difference influences substrate affinity and specificity. In falcipain 2A, Z-LR-ACE has been reported to be slightly more effective as a substrate than Z-FR-ACE and Z-VLR-ACE (Shenai et al. 2000). Compared with falcipain 2A, all tested Z-P2-P1-ACE substrates had lower second-order rate constants (kcat/Km) in falcipain 1 and 3 (Salas et al. 1995; Francis et al. 1996; Sijwali et al. 2001).

Implications for Malaria Therapy

Owing to the essential role of proteases in *P. falciparum* pathogenicity, protease inhibitors are heavily investigated as potential chemotherapeutic agents (Rosenthal et al. 1991; Semenov, Olsen, and Rosenthal 1998; Joachimiak et al. 2001; Singh and Rosenthal 2001). Studies administering cysteine protease inhibitors that block hemoglobin hydrolysis in mice infected with *P. vinckei* or *Trypanosoma cruzi* have shown promising results (Rosenthal, Lee, and Smith 1993; Engel et al. 1998; Olsen et al. 1999). Most of the studies performed have assumed a single protein target where the falcipain 2 enzyme has been reported to constitutes 93% of soluble trophozoite cysteine protease activity, as measured with the substrate Z-Phe-Arg-AMC (Shenai et al. 2000). The discovery of a second copy of the falcipain-2 in the malaria genome is therefore of importance for further development of protease inhibitors.

Acknowledgments

K.M.N. and K.K were supported by the Norwegian Research Council, D.L.H was supported by the Burroughs-Wellcome Fund and grants from the NIH, and S.K.V. and D.F.W. were supported by grants from the NIH. We are also grateful for support from the Ellison Medical Foundation.

Literature Cited

Abagyan, R., M. Totrov, and D. Kuznetsov. 1994. ICM: a new method for protein modeling and design. Applications to docking and structure prediction from the distorted native conformation. J. Comput. Chem. **15**:488–506.

Banerjee, D., P. Mayer-Kuckuk, G. Capiaux, T. Budak-Alpdogan, R. Gorlick, and J. R. Bertino. 2002. Novel aspects of resistance to drugs targeted to dihydrofolate reductase and thymidylate synthase. Biochim. Biophys. Acta **1587**:164–173.

Dua, M., P. Raphael, P. S. Sijwali, P. J. Rosenthal, and M. Hanspal. 2001. Recombinant falcipain-2 cleaves erythrocyte membrane ankyrin and protein 4.1. Mol. Biochem. Parasitol. **116**:95–99.

Efron, B., and C. Stein. 1981. The jackknife estimate of variance. Ann. Stat. **9**:586–596.

Engel, J. C., P. S. Doyle, I. Hsieh, and J. H. A. Mckerrow. 1998. Cysteine protease inhibitors cure an experimental *Trypanosoma cruzi* infection. J. Exp. Med. **188**:725–734.

Francis, S., I. Y. Gluzman, A. Oksman, D. Banerjee, and D. E. Goldberg. 1996. Characterization of native falcipain, an enzyme involved in *Plasmodium falciparum* hemoglobin degradation. Mol. Biochem. Parasitol. **83**:189–200.

Francis, S. E., D. J. Sullivan, and D. E. Goldberg. 1997. Hemoglobin metabolism in the malaria parasite *Plasmodium falciparum*. Annu. Rev. Microbiol. **51**:97–123.

Hanspal, M. 2000. cDNA cloning of a novel cysteine protease of *Plasmodium falciparum*. Biochim. Biophys. Acta **1493**:242–245.

Hartl, D. L., S. K. Volkman, K. M. Nielsen, A. E. Barry, K. P. Day, D. F. Wirth, and E. A. Winzeler. 2002. The paradoxical population genetics of *Plasmodium falciparum*. Trends Parasit. **18**:266–272.

Hughes, A. L. 1999. Adaptive evolution of genes and genomes. Oxford University Press, New York.

Hughes, A. L., and F. Verra. 2001. Very large long-term effective population size in the virulent human malaria parasite *Plasmodium falciparum*. Proc. R. Soc. London Ser. B Biol. Sci. **268**:1855–1860.

Innan, H. 2002. A method for estimating the mutation, gene conversion and recombination parameters in small multigene families. Genetics **161**:865–872.

Joachimiak, M. P., C. Chang, P. J. Rosenthal, and F. E. Cohen. 2001. The impact of whole genome sequence data on drug discovery: a malaria case study. Mol. Med. **7**:698–710.

Jongwutiwes, S., C. Putaporntip, R. Friedman, and A. L. Hughes. 2002. The extent of nucleotide polymorphism is highly variable across a 3-kb region on *Plasmodium falciparum* chromosome 2. Mol. Biol. Evol. **19**:1585–1590.

Lederberg, J. 2000. Infectious history. Science **288**:287–293.

Mu, J., J. Duan, K. D. Makova, D. A. Joy, C. Q. Huynh, O. H. Branch, W.-H. Li, and X.-Z. Su. 2002. Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*. Nature **418**:323–326.

Olsen, J. E., G. K. Lee, A. Semenov, and P. J. Rosenthal. 1999. Antimalarial effects in mice of orally administered peptidyl cysteine protease inhibitors. Bioorg. Med. Chem. **7**:633–638.

Pearlman, D. A., D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, and P. Kollman. 1995. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comp. Phys. Comm. **91**:1–41.

Phillips, R. S. 2001. Current status of malaria and potential for control. Clin. Microbiol. Rev. **14**:208 ff inter alia.

Rich, S. M., M. C. Licht, R. R. Hudson, and F. J. Ayala. 1998. Malaria's Eve: evidence of a recent population bottleneck throughout the world population of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA **95**:4425–4430.

Rosenthal, P. J., G. K. Lee, and R. E. Smith. 1993. Inhibition of a *Plasmodium vinckei* cysteine proteinase cures murine malaria. J. Clin. Invest. **91**:1052–1056.

Rosenthal, P. J., J. H. Mckerrow, M. Aikawa, H. Nagasawa, and J. H. Leech. 1988. A malarial cysteine proteinase is necessary for hemoglobin degradation by *Plasmodium falciparum*. J. Clin. Invest. **82**:1560–1566.

Rosenthal, P. J., W. S. Wollish, J. T. Palmer, and D. Rasnick. 1991. Antimalarial effects of peptide inhibitors of a *Plasmodium falciparum* cysteine proteinase. J. Clin. Invest. **88**:1467–1472.

Salas, F., J. Fichmann, G. K. Lee, M. D. Scott, and P. J. Rosenthal. 1995. Functional expression of falcipain, a *Plasmodium falciparum* cysteine proteinase, supports its role as a malarial hemoglobinase. Infect. Immun. **63**:2120–2125.

Semenov, A., J. E. Olsen, and P. J. Rosenthal. 1998. Antimalarial synergy of cysteine and aspartic protease inhibitors. Antimicrob. Agents Chemother. **42**:2254–2258.

Shenai, B. R., P. S. Sijwali, A. Singh, and P. J. Rosenthal. 2000. Characterization of native and recombinant falcipain-2, a principal trophozoite cysteine protease and essential hemoglobinase of *Plasmodium falciparum*. J. Biol. Chem. **275**: 29000–29010.

Sijwali, P. S., B. R. Shenai, J. Gut, A. Singh, and P. J. Rosenthal. 2001. Expression and characterization of the *Plasmodium falciparum* haemoglobinase falcipain-3. Biochem. J. **360**:481–489.

Sijwali, P. S., B. R. Shenai, and P. J. Rosenthal. 2002. Folding of the *Plasmodium falciparum* cysteine protease falcipain-2 is mediated by a chaperone-like peptide and not the prodomain. J. Biol. Chem. **277**:14910–14915.

Singh, A., and P. J. Rosenthal. 2001. Comparison of efficacies of cysteine protease inhibitors against five strains of *Plasmodium falciparum*. Antimicrob. Agents Chemother. **45**:949–951.

Somoza, J. R., H. J. Zhan, K. K. Bowman, L. Yu, K. D. Mortara, J. T. Palmer, J. M. Clark, and M. E. McGrath. 2000. Crystal structure of human cathepsin V. Biochemistry **39**:12543–12551.

Su, X.-Z., M. T. Ferdig, Y. Huang, C. Q. Huynh, A. Liu, J. You, J. C. Wootten, and T. E. Wellems. 1999. A genetic map and recombination parameters of the human malaria parasite *Plasmodium falcipirum*. Science **286**:1351–1353.

The Plasmodium Genome Database Collaborative. 2001. PlasmoDB: an integrative database of the *Plasmodium falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. The Plasmodium genome resource. Nucleic Acids Res. **29**:66–69.

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.

Volkman, S. K., A. E. Barry, E. J. Lyons, K. M. Nielsen, S. M. Thomas, M. Choi, S. S. Thakore, K. P. Day, D. F. Wirth, and D. L. Hartl. 2001. Recent origin of *Plasmodium falciparum* from a single progenitor. Science **293**:482–484.

World Health Organization. 2000. WHO Expert Committee on Malaria. Twentieth report. 2000.

Rose Crozier, Associate Editor

Accepted December 20, 2002