Tropical Medicine and Health Vol. 40 No. 3, 2012, pp. 79-87
doi:10.2149/tmh.2012-12
Copyright© 2012 by The Japanese Society of Tropical Medicine

Original article

Positive Diversifying Selection on the *Plasmodium falciparum surf$_{4.1}$* Gene in Thailand

Phonepadith Xangsayarath[1,2], Morakot Kaewthamasorn[1,3], Kazuhide Yahata[1], Shusuke Nakazawa[1], Jetsumon Sattabongkot[4], Rachanee Udomsangpetch[5] and Osamu Kaneko[1]*

Received 8 May, 2012 Accepted 13 May, 2012 Published online 23 August, 2012

Abstract: *Plasmodium falciparum* SURFIN$_{4.1}$ is a type I transmembrane protein thought to locate on the merozoite surface and to be responsible for a reversible adherence to the erythrocyte before invasion. In this study, we evaluated *surf$_{4.1}$* gene segment encoding extracellular region for polymorphism, the signature of positive selection, the degree of linkage disequilibrium, and temporal change in allele frequency distribution in *P. falciparum* isolates from Thailand in 1988–89, 2003, and 2005. We found that SURFIN$_{4.1}$ is highly polymorphic, particularly at the C-terminal side of the variable region located just before a predicted transmembrane region. A signature of positive diversifying selection on the variable region was detected by multiple tests and, to a lesser extent, on conserved N-terminally located cysteine-rich domain by Tajima's *D* test. Linkage disequilibrium between sites over a long distance (> 1.5 kb) was detected, and multiple SURFIN$_{4.1}$ haplotype sequences detected in 1988/89 still circulated in 2003. Few of the single amino acid polymorphism allele frequency distributions were significantly different between the 1988/89 and 2003 groups, suggesting that the frequency distribution of SURFIN$_{4.1}$ extracellular region remained stable over 14 years.
Key words: *Plasmodium falciparum*, positive diversifying selection, allele frequency distribution

INTRODUCTION

Malaria is a serious and often fatal disease caused by parasites of the genus *Plasmodia* killing nearly 800,000 people each year [1]. The malaria parasite is an obligate intracellular protozoa, and invasion into the host erythrocyte by merozoite-stage parasite is essential for its survival in the vertebrate hosts. Thus, the parasite's ligands that recognize and adhere to the host erythrocytes are important for the parasite and potential targets for intervention [2]. Recently, a large type I transmembrane protein SURFIN$_{4.1}$ (gene ID: PFD0100c) in *Plasmodium falciparum* was reported to present on the merozoite surface and was proposed to be responsible for the reversible association with target erythrocytes [3]. SURFIN$_{4.1}$ is a member of SURFINs that are encoded by a family of 10 *surf* genes located within or close to the subtelomeres of 5 chromosomes in the 3D7 parasite line. SURFIN also has a homolog in *P. vivax*, called PvSTP1, and the extracellular region of SURFIN/PvSTP members share homology with proteins encoded in the other super multigene family *pir*, which is expressed on the infected erythrocyte surface and found in a variety of malaria species including *P. yoelii*, *P. berghei*, *P. chabaudi*, *P. knowlesi* and *P. vivax* (homologous region was named cysteine-rich domain (CRD)) (Fig. 1) [4]. The intracellular region of SURFIN/PvSTP possess homology with the intra-

[1] Department of Protozoology, Institute of Tropical Medicine (NEKKEN) and the Global COE Program, Nagasaki University, Sakamoto, Nagasaki 852-8523, Japan
[2] National Institute of Public Health, Vientiane, Lao PDR
[3] Parasitology Unit, Department of Pathology, Faculty of Veterinary Science, Chulalongkorn University, Bangkok 10330, Thailand
[4] Mahidol Vivax Research Center, Faculty of Tropical Medicine, Mahidol University, Bangkok 10400, Thailand
[5] Department of Pathobiology, Faculty of Science, Mahidol University, Bangkok 10400, Thailand
*Corresponding author:
Department of Protozoology, Institute of Tropical Medicine, Nagasaki University, 1-12-4 Sakamoto, Nagasaki 852-8523, Japan
Tel: (+81) 95 819 7838
Fax: (+81) 95 819 7805
E-mail: okaneko@nagasaki-u.ac.jp
Sequence data from this article have been deposited with the GenBank™/EMBL/DDBJ databases under accession numbers: AB480049–AB480068 and AB712293–AB712335.
Abbreviations: DNA, deoxyribo nucleic acid; indels, insertions/deletions; nt, nucleotides; PCR, polymerase chain reaction; *surf* gene, surface-associated interspersed gene

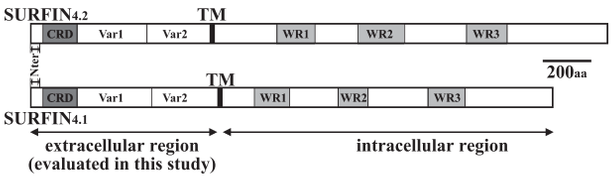

Fig. 1. Schematic of the $SURFIN_{4.1}$ and $SURFIN_{4.2}$ domain structure. Extracellular region is divided into 4 parts: N-terminal (Nter), CRD, and variable regions 1 and 2 (Var1 and Var2). WR domain in the intracellular region is indicated. Positions are after 3D7 line sequences based on the previous report [3, 5].

cellular region of the other parasite-encoded erythrocyte surface ligands (homologous region was named tryptophan-rich (WR) domain), such as *P. falciparum* PfEMP-1 family and *P. knowlesi* surface variant antigen SICAvar family [5]. Thus, the SURFIN protein structure appears to consist of components from two major virulence factors in different malaria species, *P. vivax* VIR and *P. falciparum* PfEMP-1.

Malaria proteins that exhibit high polymorphism are considered to be targets of the host immune system and therefore potential candidates for vaccine development. Most of the leading vaccine candidate proteins are indeed highly polymorphic, for example, apical membrane antigen 1 (AMA1) [6], merozoite surface protein 1 (MSP1) [7] and EBA175 [8]. Detection of a positive diversifying selection on the gene is a useful approach in the search for new vaccine candidates. Ochola et al. (2010) performed a genome wide survey to identify novel polymorphic genes and found that another SURFIN member, $SURFIN_{4.2}$, was highly polymorphic and that the signature of positive selection was acting on this gene locus using *P. falciparum* samples from Kenya [9]. Positive diversifying selection on this gene was further validated in the Thai *P. falciparum* population [10]. However, the analysis of $surf_{4.1}$ diversity using Kenyan isolates found $surf_{4.1}$, encoding $SURFIN_{4.1}$, under purifying selection rather than positive selection. $SURFIN_{4.1}$ and $SURFIN_{4.2}$ are the closest members among SURFIN family members, and it is unclear if the purifying selection detected on $surf_{4.1}$ in the Kenyan population is universal. Thus, in this study, we evaluate $surf_{4.1}$ gene for its polymorphism, positive selection, linkage disequilibrium, and temporal changes in the frequency distribution to obtain further insights into this molecule.

MATERIALS AND METHODS

Parasite and DNA preparation

Thirty-seven *P. falciparum* isolates (the name starts with "MS" followed by the number) were collected in Mae Sod, Thailand, close to the Thai-Myanmar border, between Nov/21/1988 and Jan/16/1989 (1988/89 group) [11], then adapted to the in vitro culture more or less as described previously [12]. Human erythrocytes and plasma used for the culture were obtained from the Nagasaki Red Cross Blood Center. The samples MS814, MS822 and MS835 were subjected to clone by limiting dilution to obtain MS814H1, MS814A1, MS822G8 and MS835A1. Twenty-one and 4 blood samples were collected onto the filter paper in 2003 and 2005, respectively, from *P. falciparum* patients in Thailand. The known origin in Thailand is: AQ1097, AQ1099, AQ1105, PA021, Q2D015, TMPF09, TMPF11 and TMPF15 from Tak; AA1329, AQ1098, AQ1126, AQ1129, AQ1132, AQ1133, AQ1139, AQ1423, AQ1459, TMPF34 and TMPF44 from Kanchanaburi; AQ1101 and AQ1125 from Chiangmai; AQ1130 from Chaiyaphumi; and AQ1142 from Saraburi. One sample (AQ1095) was also obtained from a patient who visited Myanmar in 2003. The sampling was authorized by the Ethical Review Committee of Mahidol University.

Genomic DNA (gDNA) was extracted from the culture-adapted parasites using DNAzol BD (Invitrogen) according to the manufacturer's instructions. Filter papers containing blood were cut into 3-mm disc using a sterile hole puncher in the clean bench to avoid contamination, and gDNA were extracted using a QIAamp DNA Mini Kit (Qiagen, Valencia, CA). The extracted gDNA were stored at –30°C until use.

Polymerase chain reaction (PCR) amplification and sequencing

For 20 samples (MS804, MS806, MS807, MS808, MS809, MS810, MS811, MS813, MS814H1, MS815, MS816, MS821, MS827, MS828, MS838, MS840, MS842, MS843, MS844 and MS947), DNA fragment for $surf_{4.1}$ encoding the extracellular region were amplified under the following conditions: an initial denaturation step of 94°C for 2 min followed by 40–45 amplification cycles of 94°C for 15 sec, 50°C for 20 sec and 68°C for 3 min, then final extension step of 68°C for 5 min. PCR amplification was performed in a 20-μL reaction mixture containing 0.5 μM each of forward (F4, CTTTATAGTAATAAAAAATAAAC AATG) and reverse (R4, GAACTCCAAAAACTGCTAAA GC) primers, 200 μM dNTP, 0.4 units of KOD-Plus DNA polymerase (Toyobo, Japan), 2.0 μL of 10 × KOD-Plus buffer II, 1 mM $MgSO_4$, and 0.4 μl of the template DNA solution. A pair of semi-nested PCR amplifications was performed for some samples for 30–40 amplification cycles in a 20-μl reaction mixture in the same way as the initial PCR, except using 0.04–0.4 μl of initial PCR solution as a template solution and a primer pair for the 5' part (F4 and R9,

GGATGAGCGTATTTATTTATTTGTTC) or 3' part (F5, AAATAAAACAATATTTGGAAAAGTC and R4).

For 43 samples (MS805, MS812, MS817, MS818, MS819, MS825, MS830, MS833, MS836, MS837, MS841, MS946, MS948, MS814A1, MS822G8, MS835A1, MS829. AA1329, AQ1095, AQ1097, AQ1098, AQ1099, AQ1101, AQ1105, AQ1125, AQ1126, AQ1127, AQ1129, AQ1130, AQ1132, AQ1133, AQ1139, AQ1142, TMPF09, TMPF11, TMPF15, TMPF18, TMPF34, TMPF44, AQ1423, AQ1459, PA021 and Q2D015), PCR was performed under the following conditions: an initial denaturation step of 94°C for 2 min followed by 40 amplification cycles of 98°C for 10 sec, 50°C for 30 sec, and 68°C for 90 sec, then final extension step of 68°C for 5 min. Reaction was performed in a 12.5-μL reaction mixture containing 200 μM dNTP, 0.4 units of KOD-Plus NEO DNA polymerase (Toyobo, Japan), 2.0 μL of 10× KOD-Plus NEO buffer, 1.5 mM $MgSO_4$, 0.4 μl of the template DNA solution and 0.3 μM each of forward (F4.1, AAAGTTTTATTAAAC CAGAAATGTAAAC) and reverse (R4.1, ATTACTTTGTT AAATAATATAAAAACG) primers, which were designed at the outside of the priming sites of primers F4 and R4, respectively. Following initial amplification, nested PCR amplification was performed for 30–40 cycles in a 12.5-μl reaction mixture in the same way as the initial PCR, except using 0.04–0.4 μl of the initial PCR product solution as a template solution and F4 and R4 as a primer set. PCR products were subjected to a 1.5% agarose gel electrophoresis and visualized with ethidium bromide under UV transillumination. The negative control reaction was always set using distilled water as a template solution. DNA standard marker was used to evaluate the size of the PCR products.

When the PCR band on the agarose gel was confirmed to be single with little or no background, amplified DNA fragments were treated with ExoSAP-IT (GE Healthcare, Buckinghamshire, UK) and sequenced directly with ABI PRISM® BigDye™ Terminator ver1.1 according to the manufacturer's instructions using a panel of oligonucleotide primers (Table 1). Sequences were then analyzed with an ABI3730 DNA analyzer (Applied Biosystems, Foster City, CA). The samples showing multiple peaks in the chromatogram, which indicates a mix infection, were sequenced after cloning of the target region into pGEM®-T Easy (Promega, Madison WI). We employed the sequence that supported at least three independent plasmid clones.

Table 1. Oligonucleotide primers used for sequencing of *surf*$_{4.1}$

Name	Sequence
F6_7G8	GATGAAAACGATGCATTTGTTC
F6_842	GTGGATAACTATGCATTTGTTC
F6	GATGGACTTAATGGATTTGATC
F7	TATTGTTCTGGAGATGAATGTG
F8	GAGAGAATGTAGTTTCTACTGCTGG
R5	TCCCTTCATATCTTTTGGATTTA
R6	ACGCAATATCATTATGATGAGG
R6_D10	AAGACAATTTCATTATGATCAGG
R6_7G8	CAACGCAATTTCATTATAATCTGG
R6_842	ACCGCAATTTCATTGTTTTCTGA
R8	CATTTGTTATATAAGAATGAGTACCTTCA

Statistical analyses

Sequences were aligned using a CLUSTAL W program [13]. The mean numbers of synonymous substitutions per synonymous site (d_S) and nonsynonymous substitutions per nonsynonymous site (d_N) and their standard errors were computed using the Nei and Gojobori method [14] with the Jukes and Cantor correction, implemented in MEGA4.0 [15]. The statistical difference between d_S and d_N was tested using a one-tailed Z-test with 500 bootstrap pseudosamples using MEGA. A value of d_N significantly higher than d_S at the 95% confidence level was taken as evidence for positive selection. Nucleotide diversity (π) was computed using DnaSP5.0 [16]. Sliding window plots of the nucleotide diversity (90 bases with a step size of 3 bases) was generated using DnaSP5.0. Images of the sliding window plot results were modified using Adobe Photoshop. Nucleotide (nt) and amino acid (aa) positions are after 3D7 line sequence.

Tajima's *D* test was used to evaluate a departure from the neutral evolution model by comparing θ (nucleotide diversity estimated based on the number of segregating site, *S*) and π (observed pairwise nucleotide diversity) to investigate whether polymorphic single nucleotide alleles tend to occur at a higher or lower frequency than expected under neutral drift [17]. Fu and Li's *D** and *F** tests evaluate departures from neutrality by comparing the number of mutations in the external (considered to be "new" mutations) and internal (considered to be "old" mutations) branches of the genealogy. The number of external mutations would be deviated from neutral expectation by the selective pressure, whereas the number of internal mutations is less affected. Under positive diversifying selection, the number of internal "old" mutations is expected to be higher than the number of external "new" mutations. Fu and Li's *D** compares the estimated θ based on the number of singletons (mutations appearing only once among the sequences, which is new and locates in the external branches) and that based on *S*. Fu and Li's *F** compares the estimated θ based on the number of singletons and that based on k (average number of pairwise nucleotide difference) [18].

Linkage analysis was performed by comparing a variance obtained from the distribution of the number of loci at

which each pair of alleles was different, and an expected variance obtained by a Monte Carlo simulation (100,000 iterations) and a standardized I_A (I_A^S), a function of the recombination rate with zero value indicating a linkage equilibrium, were calculated using LIAN3.5 [19]. Linkage analysis was also performed to obtain |D'| [20] and r^2 [21] values, indices of the linkage disequilibrium, using only informative sites after excluding sites with the frequency of the rare allele less than 10% and sites segregating for more than two nucleotides. Fisher's exact test of significance was calculated with DnaSP5.0. Minimum number of recombination events was estimated according to Hudson and Kaplan (1985) using DnaSP5.0 [22]. The differences in proportion of the major amino acid polymorphism allele at each amino acid site were assessed using Fisher's exact test ($p < 0.05$ was taken as significant).

RESULTS

Polymorphism of the *surf*$_{4.1}$

Nucleotide sequences (nt 4 to 2310) encoding the extracellular region of SURFIN$_{4.1}$ were obtained from *P. falciparum* Thai field isolates collected at three different time points: 37 sequences (24 haplotypes) in 1988–1989 (collectively termed as 1988/89 group), 21 sequences (17 haplotypes) in 2003 and 4 sequences (4 haplotypes) in 2005 (a total of 62 sequences containing 37 haplotypes). In addition, one sequence was obtained from a sample originating from Myanmar, which was different from all 37 haplotypes found in Thai parasite lines. Because 11 haplotypes contained more than one sequence, we checked other gene loci (*surf*$_{4.2}$, *clag2*, *clag8*, *clag9*, and 18 putatively neutral loci), which were determined for most of the isolates evaluated in this study [10, 23, 24] to see if any of them showed the same genetic background. We found that all sequences possessed distinct allele status for at least one SNP, except MS841, for which other gene loci were not determined, confirming that they had different genetic backgrounds. A total of 405 polymorphic nucleotide sites and 9 insertion/deletions (indels; AAT at nt 1348–1350, TAC between nt 1746–1747 and GGA between nt 2274–2274) were observed among 63 sequences with an average pairwise nucleotide diversity of 0.069. These values were higher than those for *surf*$_{4.2}$ obtained from the 74 sequences originating from a similar set of Thai samples (among 67 isolates used for *surf*$_{4.2}$, 9 isolates were excluded and 5 isolates were newly added for this *surf*$_{4.1}$ analysis), which showed 255 polymorphic sites out of 2166 bp with an average pairwise nucleotide diversity of 0.043 [10]. To evaluate the area(s) accumulating polymorphisms as we did for SURFIN$_{4.2}$, we again divided the extracellular region of SURFIN$_{4.1}$ into three regions based on amino acid sequence conservation among SURFIN members: N-terminal segment (Nter; amino acid positions (aa) 1–50, nt 1–150), CRD (aa 51–195, nt 151–585), and a variable region (aa 196–770, nt 586–2310). Then, the obtained nucleotide sequence was divided based on this definition, and Nter was assessed based on the sequence from nt 4 to nt 150. The polymorphic sites were predominantly distributed in the variable region; 394 of them were located in the variable region (22.9% of 1725 bp), while Nter had 2 (1.4% of 147 bp) and CRD had 9 polymorphic sites (2.1% of 434 bp). These trends can be seen from the sliding window plot of the nucleotide diversity (Fig. 2 top panel).

These high levels of nucleotide diversity are reflected in the amino acid level. Among 260 polymorphic substitution sites, 2 were in Nter region, 7 in CRD, and 251 in the variable region. All 3 indels were located in the variable region. To visualize the regions accumulating the polymorphism, we plotted the location of the substitution and the

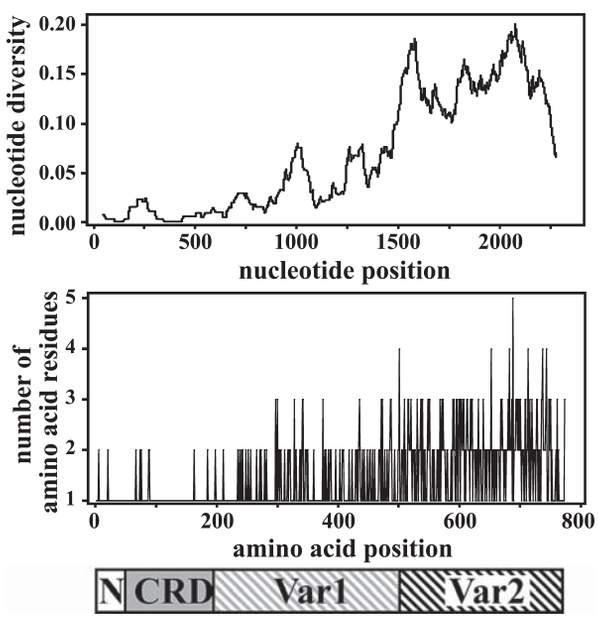

Fig. 2. Sliding window plot of nucleotide diversity and amino acid polymorphism of SURFIN$_{4.1}$ extracellular region in *Plasmodium falciparum* Asian isolates. Window length of 90 bp and step size of 3 bp is used for the sliding window plot (top). The number of the amino acid type at each amino acid position (middle) and a scheme of SURFIN$_{4.1}$ extracellular region (bottom) is shown in scale to visualize the location of the polymorphic sites. SURFIN$_{4.1}$ extracellular region was divided into 4 parts: N-terminal (Nter), CRD, and variable regions 1 and 2 (Var1 and Var2). A total of 62 sequences from Thai isolates and 1 sequence from Myanmar isolate were used. Nucleotide and amino acid positions are 3D7 line sequences.

number of amino acids observed at each site (Fig. 2 bottom panel). More polymorphism accumulated toward the C-terminal side of the variable region as observed for SURFIN$_{4.2}$. Thus, for detailed analysis, we further divided the variable region into two sub-regions, Var1 (aa 196–502, nt 586–1506) and Var2 (aa 503–770, nt 1507–2310), based on the homology between SURFIN$_{4.1}$ and SURFIN$_{4.2}$, for which the variable region was also divided into sub-regions [10]. The amount of polymorphic amino acid sites in Var2 (172/268 sites = 64.2%) was much larger than that in Var1 (79/307 sites = 25.7%), and trimorphic sites were abundant in the Var2 region (Fig. 2 and Fig. 3).

Positive diversifying selection on the *surf$_{4.1}$* gene

In view of the fact that purifying selection was detected on *surf$_{4.1}$* using the Kenyan *P. falciparum* population in a previous study [9], we evaluated signatures of a selection on *surf$_{4.1}$* implementing samples collected mainly in Thailand by comparing synonymous and non-synonymous substitutions using all 63 sequences (Table 2). A significant excess of non-synonymous over synonymous substitutions was detected when the entire sequence was evaluated (p = 0.002). The same analysis performed against the Var2 region also detected a significant excess of non-synonymous over synonymous substitutions (p = 0.0003). This result suggests that positive selection acts on the Var2 region of *surf$_{4.1}$* gene. We used Fisher's exact test for Nter, CRD, and Var1 regions to assess the difference between non-synonymous and synonymous substitutions, because the number of synonymous differences (*Sd*) was too low (0.00, 0.73 and 5.31, respectively) for the optimal analysis with Z-test. We found no significant difference.

To further evaluate the signatures of selection, we employed a population-based approach: Tajima's *D*, Fu and Li's *D**, Fu and Li's *F** tests for the 1988/89 group (Table 3). Samples obtained from the other periods were excluded due to the low sample number. Significant positive values of all Tajima's *D*, Fu and Li's *D**, Fu and Li's *F** were detected using the entire obtained sequence (D = 2.13, $p < 0.05$; D^* = 1.76, $p < 0.02$; and F^* = 2.25, $p < 0.02$, respectively). When four sub-regions were separately evaluated, significant positive values of all tests were detected only on Var2 regions (D = 2.31, $p < 0.05$; D^* = 1.95, $p < 0.02$; and F^* = 2.47, $p < 0.02$, respectively). Sliding window plot analysis of all tests revealed a significant positive deviation of greater than zero in the Var1 ($p < 0.02$) and Var2 ($p < 0.02$), the value of the Var2 region being higher than that of the Var1 region (Fig. 3). Tajima's test also revealed significant positive deviation in the CRD region ($p < 0.05$). Fu and Li's *D** and *F** showed negative values at nt 843–849 (mid point) in the Var1 region, but this was not significant by two-tailed test ($0.05 \leq p < 0.1$; asterisk in Fig. 4). Collectively, the comparison between non-synonymous and synonymous substitutions, Tajima's *D*, Fu and Li's *D**, and Fu and Li's *F** tests all revealed the signature of positive diversifying selection in the Var2 region at a 98% confidence level. Tajima's *D*, Fu and Li's *D**, and Fu and Li's *F** tests showed this in the Var1 region at a 98% confidence level, and Tajima's *D* test showed it in the CRD region at a 95% confidence level by the sliding window plot method, despite the purifying selection on this gene by analyzing the Kenyan *P. falciparum* population [9].

Linkage disequilibrium (LD) of *surf$_{4.1}$*

To gain further insights into SURFIN$_{4.1}$ polymorphism, polymorphic amino acid sites were aligned and haplotype numbers were assigned (Fig. 4; haplotype 1, MS804, MS807, MS836 and AQ1139; 2, MS812; 3, MS828; 4, MS814H1, MS815, MS841 and AQ1133; 5, MS818 and MS833; 6, AQ1127 and MS811; 7, MS840; 8, MS808; 9, MS843; 10, MS830; 11, MS829; 12, MS813; 13, MS844; 14, MS814A1 and MS835A1; 15, MS947 and TMPF09; 16, AQ1098, MS827 and Q2D015; 17, MS838; 18, AQ1105, AQ1129, MS806, MS821, MS842; 19, AQ1126, MS810,

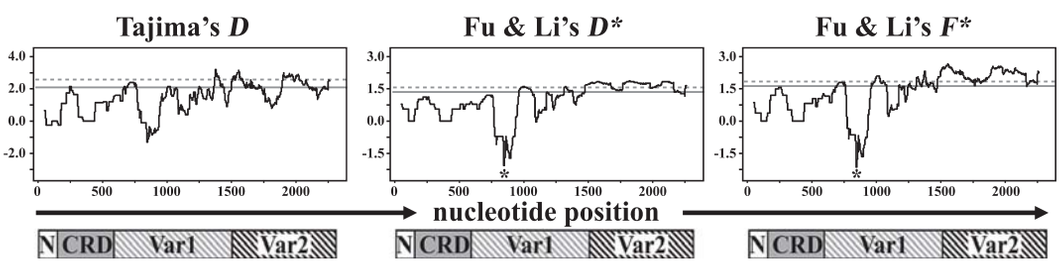

Fig. 3. Sliding window plots of Tajima's *D*, Fu and Li's *D** and *F** for *Plasmodium falciparum surf$_{4.1}$* sequence encoding extracellular region in Thai isolates. Thirty-seven sequences from the 1988/89 group are used. Sites above the solid line (percentage point $\alpha = 0.975$) and broken line ($\alpha = 0.995$ for Tajima's *D* and $\alpha = 0.99$ for Fu and Li's *D** and *F**) depart significantly from neutrality (two-tailed), suggesting diversifying selection. Nucleotide numbers are after 3D7 line sequence. Window length is 90 bp and step size is 3 bp. Asterisks indicate positions showing a negative peak value with $0.05 \leq p < 0.1$.

Table 2. Nucleotide diversity of *Plasmodium falciparum surf4.1* from Asian isolates (n = 63)

region (position)	Number of sites (base)	indels	k (SE)	*Nd* (SE)	*N* (SE)	*Sd* (SE)	*S* (SE)	π (SE)	d_N (SE)	d_S (SE)	d_N/d_S	*p* ($d_N > d_S$)
Extracellular	2304	9	149.87	125.77	1811.79	24.10	492.21	0.069	0.074	0.051	1.45	0.002
(4–2310)			(6.86)	(7.47)	(8.59)	(2.91)	(8.59)	(0.003)	(0.005)	(0.006)		
Nter	147	0	0.69	0.69	117.86	0.00	29.14	0.005	0.006	0.000	∞	ns
(4–150)			(0.45)	(0.49)	(2.02)	(0.00)	(2.06)	(0.003)	(0.004)	(0.000)		
CRD	435	0	3.13	2.40	354.58	0.73	80.42	0.007	0.007	0.009	0.78	ns
(151–585)			(1.13)	(0.91)	(3.53)	(0.41)	(3.48)	(0.003)	(0.003)	(0.006)		0.00006
Variable region	1722	9	146.06	122.68	1339.35	23.37	382.65	0.092	0.100	0.065	1.54	
(586–2310)			(6.80)	(6.44)	(6.80)	(2.87)	(7.05)	(0.005)	(0.006)	(0.008)		
Var1	918	3	35.56	30.25	722.26	5.31	195.74	0.040	0.044	0.028	1.57	ns
(586–1506)			(3.40)	(3.40)	(5.18)	(1.34)	(5.40)	(0.004)	(0.005)	(0.007)		
Var2	804	6	110.50	92.44	617.09	18.06	186.92	0.157	0.174	0.106	1.64	0.0003
(1507–2310)			(5.49)	(5.10)	(4.18)	(2.38)	(4.33)	(0.009)	(0.011)	(0.016)		

Extracellular, extracellular region; Nter, N-terminal segment; CRD, cysteine-rich domain; Var1, variable region 1; Var2, variable region 2; sites, sites nucleotide analyzed; indels, insertion/deletion polymorphism; k, the average number of nucleotide differences; *N* and *S*, average numbers of non-synonymous and synonymous sites; π, pairwise nucleotide diversity; d_N, number of non-synonymous substitutions over number of non-synonymous sites; d_S, number of synonymous substitutions over number of synonymous sites; SE, standard error computed using the Nei-Gojobori method with Jukes-Cantor correction. SE was estimated using the bootstrap method with 500 replication. The number of synonymous (*Sd*) and non-synonymous (*Nd*) differences was calculated using the Nei-Gojobori method. *p* value indicates the statistical difference between d_N and d_S, tested using one-tail Z-test with 500 bootstrap pseudosamples implemented in MEGA. ns indicates not significant by two-tailed Fisher's exact tests. Number is after 3D7 line sequence.

Table 3. Test of neutrality for *Plasmodium falciparum surf4.1* from Thai 1988/89 isolates (n = 37)

region	nucleotide position	number of sites (base)	η	*S*	two variants		more than two variants	π	θ	Tajima's *D*	Fu and Li's	
					(singleton)	(not singleton)					*D**	*F**
Extracellular	4-2310	2304	425	396	15	352	29	0.06	0.04	2.13*	1.76**	2.25**
Nter	4-150	147	2	2	0	2	0	0.004	0.003	0.68	0.78	0.87
CRD	151-585	435	9	8	0	7	1	0.007	0.004	1.58	1.31	1.64
Variable region	586-2310	1722	414	386	15	343	28	0.08	0.05	2.14*	1.75**	2.25**
Var1	586-1506	918	105	102	11	88	3	0.038	0.027	1.61	1.12	1.53
Var2	1507-2310	804	309	284	4	255	25	0.137	0.08	2.31*	1.95**	2.47**

Extracellular , extracellular region; Nter, N-terminal segment; CRD, cysteine-rich domain; Var1, variable region 1; Var2, variable region 2; sites, nucleotide sites analyzed; η, the total number of mutations; *S*, number of segregating sites; π, observed nucleotide diversity; θ, the expected nucleotide diversity under neutrality derived from *S*. * indicates $p < 0.05$ and ** indicates $p < 0.02$. Sequence number is after 3D7 line sequence.

MS948 and TMPF34; 20, MS817; 21, MS825; 22, AA1329, MS805, MS809, MS816, MS822G8 and MS946; 23, MS837; 24, MS819; 25, AQ1132; 26, AQ1142; 27, AQ1097 and TMPF44; 28, AQ1130; 29, TMPF11; 30, AQ1125; 31, TMPF18; 32, AQ1101; 33, TMPF15; 34, AQ1099; 35, AQ1459; 36, AQ1423; and 37, PA021). Although at least 48 recombination events were detected throughout the entire sequence (Fig. 4), we found that many sites were clustered and so evaluated linkage disequilibrium on *surf4.1* by calculating a standardized I_A (I_A^S). We found that the I_A^S of 1988/89 group sequences was 0.2255 ($p < 1 \times 10^{-05}$), indicating significant LD of *surf4.1* sequences. To further rule out potential deviation by the inclusion of multiple identical sequences for the LD analysis, we used only 24 haplotype sequences and obtained I_A^S with 0.2010 ($p < 1 \times 10^{-05}$), further supporting significant LD of *surf4.1* sequences in Thai isolates. Significant LD between sites over a long distance (> 1.5 kb) was evident by |D'| and r^2 values seen in the right upper corner of the panels in Fig. 5.

Frequency distribution of the amino acid polymorphism of SURFIN4.1

When a parasite population possesses more than one haplotype of an antigen-encoding gene at one time point,

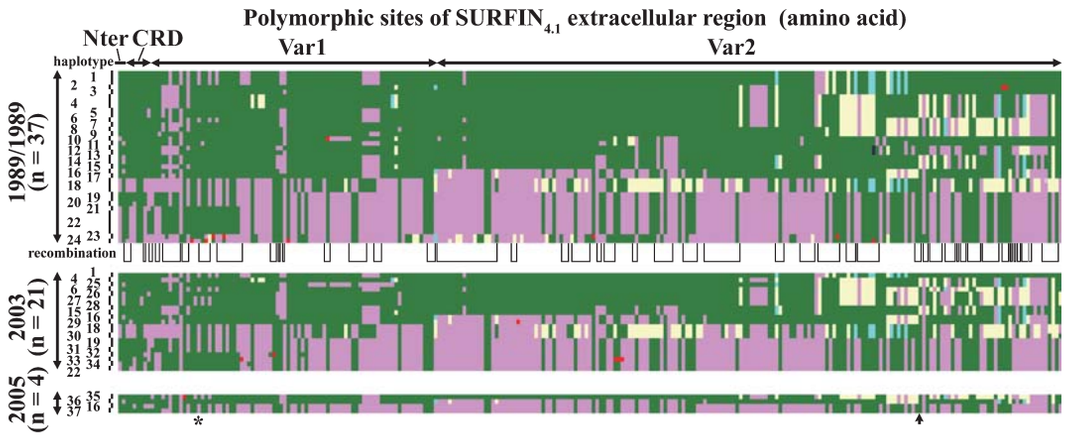

Fig. 4. Polymorphic amino acid sites of SURFIN$_{4.1}$ extracellular region among 62 Thai isolates. Only polymorphic sites were selected and aligned, then haplotype numbers were given to each sequence as described in the text. Amino acid residues are shown with green, pink or yellow for dimorphic sites, and green, pink and yellow for trimorphic sites. Fourth or fifth substitutions are shown with cyan or black colors. Singleton amino acid substitutions are shown with red color. The arrow and asterisk on the bottom of the scheme indicate amino acid position for which significant change was detected in the frequency distribution and the position where Fu and Li's *D** and *F** showed negative peak value, respectively. Thin lines under the scheme for the 1988/89 group connect the sites for which recombination was detected.

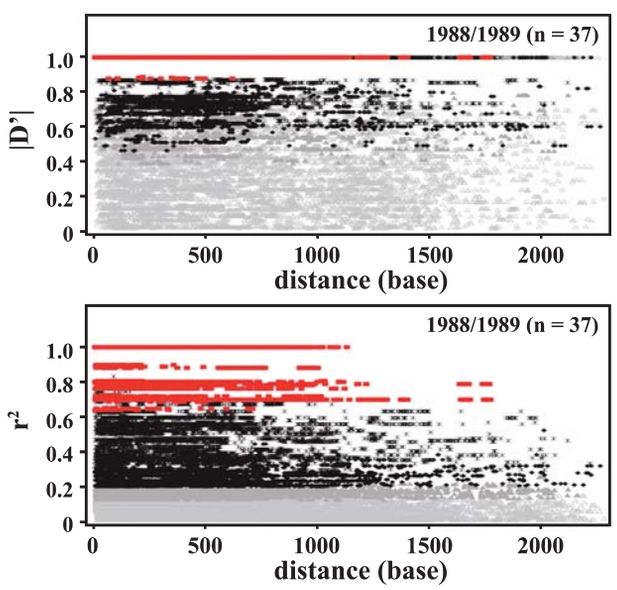

Fig. 5. Linkage disequilibrium of *Plasmodium falciparum surf$_{4.1}$* sequence encoding extracellular region in Thai isolates. The number of polymorphic sites analyzed was 283. Red filled square, $p < 0.001$ after Bonferroni correction (more conservative); asterisks, $p < 0.001$; black filled triangle, $0.001 \le p < 0.01$; dark gray filled triangle, $0.01 \le p < 0.05$; light gray open triangle, not significant. Correlation coefficients for |D'| and r^2 were –0.189 and –0.273, respectively.

the most common haplotype may be targeted by immune selection pressure, and, as a result, the frequency of the most common haplotype may decrease and relatively rare haplotypes, which encode a different set of amino acids from those encoded by the original major haplotype at each polymorphic site, may expand in the population. We evaluated this for SURFIN$_{4.2}$ previously and found that the frequency distribution of the selected polymorphic region was stable over a period of 14 years [10]. We expanded this study to SURFIN$_{4.1}$ by comparing the change in the single amino acid polymorphism allele frequency distribution for all polymorphic sites (257 substitution and 3 indel polymorphisms) between the 1988/89 and 2003 groups. Stable frequency distribution of the putatively neutral 18 SNPs between these groups was validated previously, and no detectable change was found in the background of the parasite population structure [24]. We found that the allele frequency distribution of almost all single amino acid polymorphisms on SURFIN$_{4.1}$ extracellular region did not differ significantly between the two groups. The only exception was the Gly/Arg substitution at aa 703, where the proportion of Gly in the 1988/89 group (68%, 25/37) was significantly higher than that in the 2003 group (95%, 20/21) (p = 0.02).

DISCUSSION

In this study, we evaluated the *P. falciparum surf$_{4.1}$* gene sequence encoding extracellular region for polymorphism, the signature of positive selection, the degree of linkage disequilibrium, and temporal changes in allele frequency distribution in *P. falciparum* isolates from Thailand. We found that SURFIN$_{4.1}$ is highly polymorphic, particu-

larly at the C-terminal side of the Var2 region. This pattern is slightly different from that of SURFIN$_{4.2}$, where the N-terminal side of Var2 is more polymorphic than the C-terminal. Then we noticed that SURFIN$_{4.1}$ appears to have 4 types of sequence in the C-terminal side of the Var2 region, whereas only 3 types of sequence were seen for the C-terminal side of SURFIN$_{4.2}$ (Fig. 3 in reference 10), a fact that probably contributes to the difference. We also noticed that SURFIN$_{4.1}$ has significantly more polymorphic sites than SURFIN$_{4.2}$ by chi-square test ($p < 1 \times 10^{-05}$; 396 sites/2304 bp = 17.2% and 247/2166 = 11.4%, respectively [10]). At this point, the nature of the difference between SURFIN$_{4.1}$ and SURFIN$_{4.2}$ is unclear. However, a difference in their location was reported previously: SURFIN$_{4.2}$ was located on both the infected-erythrocyte membrane and the merozoite surface, while SURFIN$_{4.1}$ was located on the merozoite surface only, a fact that may explain the observed difference.

For the Thai isolates, we detected a signature of positive diversifying selection on Var1 and Var2 by multiple tests and, to a lesser extent, on CRD region by Tajima's *D* test. This observation is different from the previous report demonstrating purifying selection on the C-terminal side of the Var1 region and the N-terminal side of the Var2 region using Kenyan *P. falciparum* population [9]. We compared the *surf$_{4.1}$* sequence set from Kenya (accession numbers HM000393–HM000443) and our sequence set from Thailand and noticed that there were significantly more singleton segregating sites/polymorphic sites in the Kenyan sequence set [38.2% (126/330), n = 51] than the Thai sequence set [3.8% (15/396) for 37 sequences from the 1988/89 group and 4.9% (20/405) for 62 Thai sequences] ($p < 1 \times 10^{-05}$ by two-tailed Chi-square test). Ochola et al. speculated that the copy number polymorphism of *surf$_{4.1}$* gene [6 copies in FCR3 line (originating from Southeast Asia) and 1 copy in 3D7 (presumably Africa) and 7G8 (South America) lines] might shed light on the complex history of the evolution of the *surf*$_{4.1}$ gene locus generating the current *surf$_{4.1}$* sequence set in Kenya. Thus, is it possible that more Kenyan *P. falciparum* isolates possess a multiple copy number of *surf$_{4.1}$* than Thai isolates? Multiple copies of *surf$_{4.1}$* gene in the Asian FCR3 line indicate that at least some Asian *P. falciparum* parasites possess more than one copy. It is noteworthy that a window in the N-terminal side of Var1 showed negative values for Fu and Li's *D** and *F** for Thai sequences that could be significant if the one-tailed test is employed ($p < 0.05$). To gain a clear picture of the diversity of *surf$_{4.1}$*, copy number polymorphism needs to be investigated in combination with *surf$_{4.1}$* sequence information from the parasites circulating in the different geographical areas.

Although over 40 recombination events were detected throughout the entire sequence, LD between sites over a long distance (> 1.5 kb) was detected. Multiple SURFIN$_{4.1}$ haplotype sequences detected in the 1988/89 group still existed in the 2003 group (haplotypes 1, 4, 6, 15, 16, 18, 19 and 22), and the 2005 group (haplotype 16), which was similar to the SURFIN$_{4.2}$ reported previously [10]. Different combinations of the allele at 18 putatively neutral sites and 3 *clag* gene loci indicated that the shared *surf$_{4.1}$* gene sequence between the 1988/89 and 2003 groups is not due to a cross-contamination during the experiment. Thus, as we proposed for SURFIN$_{4.2}$, we propose that the epistatic relation, if any, and/or high diversity dominates the frequency of recombination for SURFIN$_{4.1}$.

The single amino acid polymorphism allele frequency distribution did not show a statistically significant difference between the 1988/89 and 2003 groups except for one site, suggesting that the frequency distribution of the SURFIN$_{4.1}$ extracellular region was stable over a period of 14 years. Recently, we evaluated the allele frequency distribution for 4 other antigens (*surf$_{4.2}$*, *clag2*, *clag8*, and *clag9*) using a similar set of samples collected in Thailand in 1988/89 and 2003, and the allele frequency distribution for these genes was also found to be stable over the 14-year period with only one or two exceptional sites [10, 24]. The stable frequency distribution of *P. falciparum* merozoite surface protein 1 (MSP1) has been reported for 10 years in Tanzania [25] and for 7 years in The Gambia [26], indicating that stable frequency distribution of polymorphism in antigen may be a universal phenomenon for malaria parasite. However, it is still possible that the observed Gly/Arg change at aa 703 is a response to a selection pressure. It is also possible that substitutions occurred during the *in vitro* culture of the 1988/89 group parasites, which were isolated more than 20 years ago. However, this probably did not significantly affect our conclusion, because the parasites were stored in liquid nitrogen most of the time and the total culture period was less than half a year. Nevertheless, since the 1988/89 group parasites were exposed to selection pressure to adapt to the *in vitro* culture environment, we can't exclude the possibility that the observed Gly/Arg change in frequency distribution at aa 703 was selected during the adaptation to the culture system. Another candidate for the selection pressure is the host immune system. Further studies are required to determine whether the observed change is reproducible.

ACKNOWLEDGEMENTS

We are grateful to I. Sekine, head of the Nagasaki Red Cross Blood Center for human erythrocytes and plasma. This work was supported in part by Grants-in-Aids for Scientific Research 19590428, 22390079 and 24406012 (to

OK) and the Global COE Program, Nagasaki University (to OK) from the Ministry of Education, Culture, Sports, Science and Technology, Japan. M.K. and P.X. are recipients of a Nagasaki University Ph.D. research scholarship and Yeh Kuo Shii scholarship, Nagasaki University, Japan.

REFERENCES

1. WHO. World Malaria Report 2010. Geneva: World Health Organization. 2010.
2. Kaneko O. Erythrocyte invasion: vocabulary and grammar of the *Plasmodium* rhoptry. Parasitol Int 2007; 56: 255–262.
3. Mphande FA, Ribacke U, Kaneko O, Kironde F, Winter G, Wahlgren M. SURFIN$_{4.1}$, a schizont-merozoite associated protein in the SURFIN family of *Plasmodium falciparum*. Malar J 2008; 7: 116.
4. Janssen CS, Barrett MP, Turner CM, Phillips RS. A large gene family for putative variant antigens shared by human and rodent malaria parasites. Proc Biol Sci 2002; 269: 431–436.
5. Winter G, Kawai S, Haeggström M, Kaneko O, von Euler A, Kawazu S, Palm D, Fernandez V, Wahlgren M. SURFIN is a polymorphic antigen expressed on *Plasmodium falciparum* merozoites and infected erythrocytes. J Exp Med 2005; 201: 1853–1863.
6. Polley SD, Conway DJ. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. Genetics 2001; 158: 1505–1512.
7. Conway DJ, Cavanagh DR, Tanabe K, Roper C, Mikes ZS, Sakihama N, Bojang KA, Oduola AM, Kremsner PG, Arnot DE, Greenwood BM, McBride JS. A principal target of human immunity to malaria identified by molecular population genetic and immunological analyses. Nat Med 2000; 6: 689–692.
8. Baum J, Thomas AW, Conway DJ. Evidence for diversifying selection on erythrocyte-binding antigens of *Plasmodium falciparum* and *P. vivax*. Genetics 2003; 163: 1327–1336.
9. Ochola LI, Tetteh KKA, Stewart LB, Riitho V, Marsh K, Conway DJ. Allele Frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in *Plasmodium falciparum*. Mol Biol Evol 2010; 27: 2344–2351.
10. Kaewthamasorn M, Yahata K, Alexandre JSF, Xangsayarath P, Nakazawa S, Torii M, Sattabongkot J, Udomsangpetch R, Kaneko O. Stable allele frequency distribution of the polymorphic region of SURFIN$_{4.2}$ in *Plasmodium falciparum* isolates from Thailand. Parasitol Int 2012; 61: 317–323.
11. Nakazawa S, Culleton R, Maeno Y. In vivo and in vitro gametocyte production of *Plasmodium falciparum* isolates from Northern Thailand. Int J Parasitol 2011; 41: 317–323.
12. Trager W, Jensen JB. Human malaria parasites in continuous culture. Science 1976; 193: 673–675.
13. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22: 4673–4680.
14. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 1986; 3: 418–426.
15. Kumar S, Tamura K, Nei M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 2004; 5: 150–163.
16. Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 2003; 19: 2496–2497.
17. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989; 123: 585–595.
18. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics 1993; 133: 693–709.
19. Haubold B, Hudson RR. LIAN 3.0: detecting linkage disequilibrium in multilocus data. Linkage Analysis. Bioinformatics 2000; 16: 847–848.
20. Lewontin RC. The interaction of selection and linkage. I. general considerations; heterotic models. Genetics 1964; 49: 49–67.
21. Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 1968; 38: 226–231.
22. Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 1985; 111: 147–164.
23. Alexandre JSF, Kaewthamasorn M, Yahata K, Nakazawa S, Kaneko O. Positive selection on the *Plasmodium falciparum clag2* gene encoding a component of the erythrocyte-binding rhoptry protein complex. Trop Med Health 2011; 39: 77–82.
24. Alexandre JSF, Xangsayarath P, Kaewthamasorn M, et al. Stable allele frequency distribution of the *Plasmodium falciparum clag* genes encoding components of the high molecular weight rhoptry protein complex. Trop Med Health (in the same issue)
25. Tanabe K, Sakihama N, Rooth I, Björkman A, Färnert A. High frequency of recombination-driven allelic diversity and temporal variation of *Plasmodium falciparum msp1* in Tanzania. Am J Trop Med Hyg 2007; 76: 1037–1045.
26. Conway DJ, Greenwood BM, McBride JS. Longitudinal study of *Plasmodium falciparum* polymorphic antigens in a malaria-endemic population. Infect Immun 1992; 60: 1122–1127.