Title: Stable allele frequency distribution of the polymorphic region of SURFIN$_{4.2}$ in *Plasmodium falciparum* isolates from Thailand

Morakot KAEWTHAMASORN [1,5], Kazuhide YAHATA [1], Jean Seme Fils ALEXANDRE [1,6], Phonepadith XANGSAYARATH [1], Shusuke NAKAZAWA [1], Motomi TORII [2], Jetsumon SATTABONGKOT [3,†], Rachanee UDOMSANGPETCH [4], Osamu KANEKO [1,*]

[1] Department of Protozoology, Institute of Tropical Medicine (NEKKEN) and the Global Center of Excellence Program, Nagasaki University, Sakamoto, Nagasaki 852-8523, Japan
[2] Department of Molecular Parasitology, Ehime University Graduate School of Medicine, Shitsukawa, Toon, Ehime 791-0295, Japan
[3] Department of Entomology, Armed Forces Research Institute of Medical Sciences, Bangkok 10400, Thailand
[4] Department of Pathobiology, Faculty of Science, Mahidol University, Bangkok 10400, Thailand
[5] Parasitology Unit, Department of Pathology, Faculty of Veterinary Science, Chulalongkorn University, Bangkok 10330, Thailand
[6] Centro Nacional de Control de Enfermedades Tropicales, Santo Domingo, República Dominicana

*Corresponding author: Department of Protozoology, Institute of Tropical Medicine (NEKKEN), Nagasaki University, 1-12-4 Sakamoto, Nagasaki 852-8523, Japan. Tel (+81)-95-819-7838; Fax (+81)-95-819-7805

E-mail addresses: morakot.k@chula.ac.th, kotscmi@hotmail.com (M. Kaewthamasorn), kyahata@nagasaki-u.ac.jp (K. Yahata), semefils@yahoo.es (J.S.F. Alexandre), xangsayalathdith@yahoo.com (P. Xangsayarath), nakazawa@nagasaki-u.ac.jp (S. Nakazawa), torii@m.ehime-u.ac.jp (M. Torii), tmjetsumon@mahidol.ac.th (J. Sattabongkot), scrud@mahidol.ac.th (R. Udomsangpetch), okaneko@nagasaki-u.ac.jp (O. Kaneko)

† Present address: Mahidol Vivax Research Center, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand.

Abstract

Plasmodium falciparum SURFIN$_{4.2}$ (PFD1160w) is a polymorphic protein expressed on the surface of parasite-infected erythrocytes. Such molecules are expected to be under strong host immune pressure, thus we analyzed the nucleotide diversity of the N-terminal extracellular region of SURFIN$_{4.2}$ using *P. falciparum* isolates obtained from a malaria hypoendemic area of Thailand. The extracellular region of SURFIN$_{4.2}$ was divided into four regions based on the amino acid sequence conservation among SURFIN members and the level of polymorphism among SURFIN$_{4.2}$ sequences; N-terminal segment (Nter), a cysteine-rich domain (CRD), a variable region 1 (Var1), and a variable region 2 (Var2). Comparison between synonymous and non-synonymous substitutions, Tajima's *D* test, and Fu and Li's *D** and *F** tests detected signatures of positive selection on Var2 and to a lesser extent Var1, suggesting that these regions were likely under host immune pressure. Strong linkage disequilibrium was detected for nucleotide pairs separated by a distance of more than 1.5 kb, and 7 alleles among 19 alleles detected in 1988-1989 still circulated 14 years later, suggesting low recombination of the analyzed *surf$_{4.2}$* sequence region in Thailand. The allele frequency distribution of polymorphic areas in Var2 did not differ between two groups collected in different time points, suggesting the allele frequency distribution of this region was stable for 14 years. The observed allele frequency distribution of SURFIN$_{4.2}$ Var2 may be fixed in Thai *P. falciparum* population as similar to the observation for *P. falciparum* merozoite surface protein 1, for which a stable allele frequency distribution was reported.

Keyword(s): *Plasmodium falciparum*, SURFIN, positive diversifying selection, allele frequency distribution

1. Introduction

Malaria is a serious life-threatening disease caused by *Plasmodium* protozoan parasite. The estimated number of malaria case was 225 million and over 780,000 people died of malaria in 2009, majority of them were children under 5 years old living in Africa [1]. The erythrocytic cycle of *Plasmodium falciparum* is characterized by complex interactions between parasite-derived surface exposed antigens, host-cell receptors and immune defense proteins. Parasite adhesins expressed on the surface of malaria-infected erythrocytes provide adhesive properties and at the same time allow for host immune evasion by virtue of antigenic variation and antigenic polymorphism [2]. Recently, a new multigene family called *surf* was identified, consisting of at least 10 members in the genome of *P. falciparum* 3D7 parasite line and infected erythrocyte surface localization was shown for one of the member, SURFIN$_{4.2}$ (gene ID: PFD1160w, chr 4) [3]. All SURFIN members are predicted to be type 1 transmembrane proteins and share a similar protein makeup, containing a conserved cysteine-rich domain (CRD) in the N-terminal of their extracellular region, and tryptophan-rich domains (WRD) in the C-terminal predicted intracellular region. An orthologous gene, *Pv*STP1 has been identified in *Plasmodium vivax*. The CRD of SURFIN/*Pv*STP members shares homology with proteins encoded by the super multigene family *pir*, which are expressed on surface of infected erythrocytes and are found in numerous malaria parasite species including *Plasmodium yoelii*, *Plasmodium berghei*, *Plasmodium chabaudi*, *Plasmodium knowlesi*, and *P. vivax* [4]. The intracellular WRD of SURFIN/*Pv*STP shares homology with the intracellular region of the other parasite-encoded erythrocyte surface ligands, such as the *P. falciparum Pf*EMP-1 family and *P. knowlesi* surface variant antigen SICA*var* family. Thus, it is possible that the *pir* gene products have evolved from the extracellular region of SURFIN/*Pv*STP type proteins, and, similarly, *Pf*EMP-1 and SICA*var* proteins may be derived from the extracellular region [3]. SURFIN$_{4.2}$ is co-transported to Maurer's clefts with *Pf*EMP-1 and RIFIN, and is expected to be translocated to the erythrocyte membrane for surface exposure. SURFIN$_{4.2}$ also accumulates in the parasitophorous vacuole, and can be detected in an amorphous cap at the newly released merozoite apex, suggesting that it may have a role in the erythrocyte invasion [3]. Such molecules are expected to be under a strong host immune pressure, and indeed the extracellular region is polymorphic between 3D7 and FCR3 parasite lines [3]. Recently, a population genetics-based analysis using Kenyan *P. falciparum* isolates detected positive diversifying selection and linkage disequilibrium with a distance of ~1.5 kb [5] on *surf$_{4.2}$* exon 1 encoding the extracellular region. To further understand the host-pathogen interaction through SURFIN$_{4.2}$ and design an universal

intervention strategy targeting this molecule, it is important to understand the nucleotide diversity in different geographic areas. Thus, we obtained sequences encoding the extracellular region of $SURFIN_{4.2}$ from Thai *P. falciparum* isolates collected in a malaria hypoendemic area and evaluated allelic diversity, whether or not there was positive selection, the degree of linkage disequilibrium, and temporal changes in allele frequency distribution in a malaria hypoendemic area.

2. Materials and methods

2.1. Parasite DNA isolation

P. falciparum parasites were obtained from Thailand in three different periods. Thirty samples (1988/1989 group) were collected from November 1988 to January 1989 in Mae Sod, which lies in the north west of Thailand near the Thai-Myanmar border. They were adapted to culture, cryopreserved and kept at the Department of Protozoology, Institute of Tropical Medicine (NEKKEN), Nagasaki University as previously described [6, 7]. The parasites were thawed and maintained *in vitro* essentially as described previously [8]. Human erythrocytes and plasma used for culture were obtained from the Nagasaki Red Cross Blood Center. Parasites were harvested when parasitemia reached about 2%, and parasite genomic DNA (gDNA) was extracted using DNAzol BD (Invitrogen). Following elution, gDNA was stored at -30ºC. Twenty-eight blood samples were collected onto filter papers (Whatman 31 ETChr) in 2003 (n = 21) and 2005 (n = 7) after the approval by the Ethical Review Committee of Mahidol University. Genomic DNA was extracted using QIAamp DNA Mini Kit (Qiagen, Valencia, CA) according to the manufacturer's instruction. Filter papers were separately and carefully handled in a clean bench to avoid DNA contamination among the samples.

2.2. Polymerase chain reaction (PCR) amplification and sequencing

A DNA fragment of $surf_{4.2}$ gene encoding the extracellular region (2314 bp, nucleotide positions (nt) –22 - 2292 after 3D7 sequence) was primarily amplified with forward primer F0 (ATATTTTCCCATTTTTGATAATATG) and reverse primer R2-2 (CTTATTAAATACCAAAAACATAAAAAG). The amplification was performed in a 20 µL reaction mixture containing 200 µM dNTPs, 1x KOD -Plus- buffer, 2 mM $MgSO_4$, 500 nM of each primer and 0.4 units of KOD -Plus- DNA polymerase (Toyobo, Japan) using a GeneAmp 9700 PCR thermocycler (Applied Biosystems, Foster City, CA). Negative control was always set using distilled water as a template solution. Thermal cycling profile contained an initial denaturation at 94ºC for 2 min; 40 cycles of 92ºC for 15 sec, 54ºC for 20 sec, 68ºC for 3 min; and final-extension at 68ºC for 5 min. PCR products were then resolved by electrophoresis on a 1.0% agarose gel (Takara, Japan). After ethidium bromide staining, the PCR products were visualized by UV transillumination. Primary PCR products were then diluted with distilled water and adjusted the DNA concentration to approximately 50- 100 ng/µL to serve as a template for the nested PCR amplification. For the samples, especially those obtained from filter papers, producing undetectable or very faint target bands after primary PCR amplification, we 100-fold diluted the primary PCR product with distilled water and used for the nested PCR amplification with the same condition for the primary PCR amplification. Two sets of primers were used; F7 (CTTTTTGTTGTTGAGCTCGACAGC) and R2 (CCTGATCTGTCGAATAAATAGC) for the 5' side 1201 bp (nt 4 - 1204 after 3D7 sequence) and F3 (ATTGAAGTTGATTGTGCTGAAG) and R8 (TATCCCTTTTGAAAAAATCCCTC) for the 3' side part 1158 bp (nt 1060 - 2217 after 3D7 sequence).

PCR products were treated with ExoSAP-IT (USB Corporation) and directly sequenced from both directions with ABI PRISM® BigDye™ Terminator ver1.1 (Applied Biosystems, Foster City, CA) using an ABI 3730 DNA analyzer (Applied Biosystems) according to the manufacturer's instructions. The primers used for the sequencing are summarized in Table S1. Samples showing dual peaks, suggestive of a mixed infection, were sequenced after cloning the full length of extracellular region into pGEM-T Easy® plasmid (Promega, Madison, WI). We employed sequences supported by at least three independent plasmid clones. After sequencing reaction, sequencing mixture was subjected to ethanol precipitation to remove the fluorescent reaction mixture. All sequences were validated by at least two independently PCR-amplified DNA fragments to avoid potential error during the PCR amplification and sequencing with special care to the singletons (substitutions appearing only once among the sequences).

We analyzed four putatively neutral loci to serve as controls for allele frequency distribution. Single nucleotide polymorphisms (SNP) determined were nt 267 and 1008 of adenylosuccinate lyase (PFB0295w, chr 2), nt 165 and 319 of aspartate aminotransferase (PFB0200c, chr 2), nt 1989 of pyruvate kinase (PF10_0363, chr 10), and nt 819 and 969 of actin II (PF14_0124, chr 14). These genes were selected because they were located on different chromosomes to those harboring genes associated with drug resistance (dihydrofolate reductase-thymidylate synthase, chr 4; multidrug resistance protein, chr 5; chloroquine resistant transporter, chr 7; and dihydropteroate synthetase, chr 8), to avoid a potential hitchhiking effect by the drug pressure towards these genes. We amplified DNA region surrounding selected SNPs using primers listed in Table S2. PCR conditions and sequencing reactions were similar to those described for $surf_{4.2}$ with slight modifications. We adjusted the annealing temperature to 56°C, reduced the extension time to 45 sec ~ 1 min depending on the amplicon, and sequenced directly.

2.3. Data analysis

Nucleotide diversity (π) and its standard error (SE) were computed with the Jukes and Cantor method using MEGA 4.0 software [9]. The mean numbers of synonymous substitutions per synonymous sites (d_S) and nonsynonymous substitutions per nonsynonymous sites (d_N) and their standard errors were computed using the Nei and Gojobori method [9, 10] with the Jukes and Cantor correction, implemented in MEGA 4.0. The statistical difference between d_N and d_S was tested using a one–tailed Z–test with 500 bootstrap pseudosamples in MEGA 4.0. A value of d_N significantly higher than d_S at the 95% confidence level was taken as evidence for positive selection. Nucleotide diversity was plotted by a sliding window method (90 bases with a step size of 3 bases) using DnaSP 5.0 [11].

Departures from the predictions of the neutral model of molecular evolution were tested by a set of neutrality tests based on measures of allele frequencies or heterozygosity within species using Tajima's *D*, Fu and Li's $D^*$and F^* parameters using DnaSP 5.0, under the assumption that the size of population was stably maintained larger than effective population size. Tajima's *D* statistic relies on the difference between an average pairwise nucleotide diversity (π) and an estimated nucleotide diversity under neutrality (θ) derived from the number of segregating sites (*S*) [12]. Fu and Li's tests rely on the differences between estimates of θ based on the number

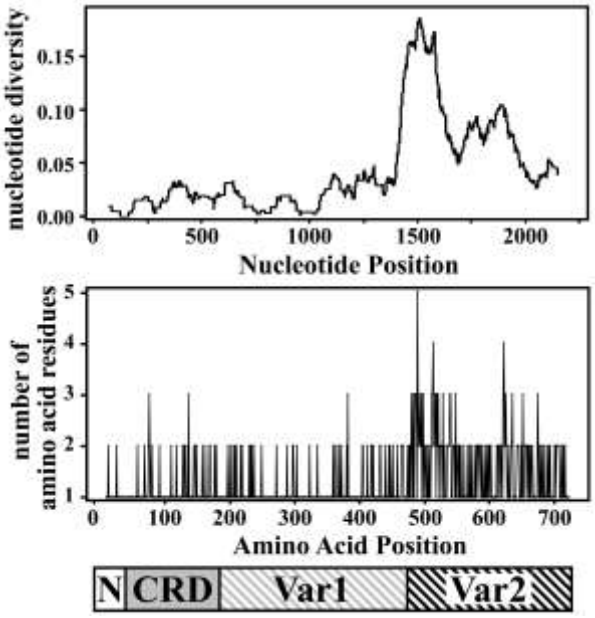

Fig. 1. Sliding window plot of nucleotide diversity and amino acid polymorphism of SURFIN$_{4.2}$ extracellular region in *Plasmodium falciparum* Thai isolates. Nucleotide diversity is plotted with a window length of 90 bp and step size of 3 bp (top) and the number of the amino acid residues at each amino acid position is plotted (middle). To visualize the location sites showing high diversity, a scheme of the extracellular region of SURFIN$_{4.2}$ is shown (bottom), which is divided into 4 parts; N-terminal (Nter), cysteine-rich domain (CRD), and variable regions (Var1 and Var2). A total of 74 sequences from Thai isolates are used. Nucleotide and amino acid positions are after the 3D7 line sequences.

of singletons and that based on S (D^* index) or π (F^* index) [13]. Minimum number of recombination events was evaluated according to Hudson and Kaplan (1985) using DnaSP 5.0 [14, 15]. Linkage analysis was performed by comparing a variance obtained from the distribution of the number of loci at which each pair of allele 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 [16]. Linkage analysis was also performed to obtain |D'| and r^2 values, indices of the linkage disequilibrium, using DnaSP 5.0 [17, 18].

3. Results

3.1. Polymorphism of the surf$_{4.2}$ gene

Single contiguous nucleotide sequences (2166 bp from nt 28 to 2193) encoding the extracellular region of SURFIN$_{4.2}$ were obtained from *P. falciparum* Thai field isolates collected at three different time points; 44 sequences (19 alleles) in 1988 - 1989 (collectively termed as 1988/1989 group), 23 sequences (14 alleles) in 2003 and 7 sequences (6 alleles) in 2005 (totally 74 sequences containing 28 alleles). The sequence names are found in the legend of figure 3**.** A total of 255 polymorphic nucleotide sites and no insertion/deletion were observed in the overall samples sequenced with an average pairwise nucleotide diversity of 0.043. In order to identify the area(s) accumulating polymorphism, we divided the extracellular region of SURFIN$_{4.2}$ 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 196–739, nt 586–2217). Then, the obtained nucleotide sequence was divided based on this definition, thus Nter was assessed based on the sequence from nt 28 to nt 150 and variable region from nt 586 to nt 2193. Although the polymorphic sites were distributed across the entire sequence, 225 of them were located in the variable region (14.0% of 1608 bp), while Nter had 2 (1.6% of 123 bp) and CRD had 28 polymorphic sites (6.4% of 435 bp). This was evident in sliding window plot of nucleotide diversity (Fig. 1).

These trends were extended to the amino acid level. Among 188 polymorphic sites, 2 were in Nter, 23 were in CRD, and 163 were in the variable region. In order to pinpoint which regions had accumulated the most polymorphisms, we plotted the location of the substitution and the number of amino acids observed at each site (Fig. 1). We noticed that more polymorphism accumulated towards the C-terminal side of the variable region. We divided SURFIN$_{4.2}$ variable region into two sub-regions for detailed analysis; the N-terminal Var1 region from aa 196 to 482 and C-terminal Var2 region from aa 483 to 739. The amount of polymorphic amino acid sites in CRD (23/145 = 15.9%) and Var1 (48/287 = 16.7%) was comparable and most of the polymorphism were dimorphic. Whereas a larger number of polymorphic sites were observed in Var2 (115/249 = 46.2%) with four different amino acids at aa 504 and 613, and five amino acids at aa 489.

3.2. Positive diversifying selection on surf$_{4.2}$

Because of the observed high polymorphism, we then evaluated signatures of a positive diversifying selection on *surf$_{4.2}$* in Thai isolates by comparing synonymous and non-synonymous substitutions using all 74 sequences (Table 1). A significant excess of non-synonymous substitutions over synonymous substitutions was detected when the entire sequence was evaluated (p = 0.0004). The same analysis performed against the Var2 region also detected a significant excess of non-synonymous substitutions over synonymous substitutions (p = 0.005). Although excess of non-synonymous substitutions over synonymous substitutions was observed for Nter, CRD, and Var1 regions, these regions were excluded from this analysis, because the numbers of synonymous

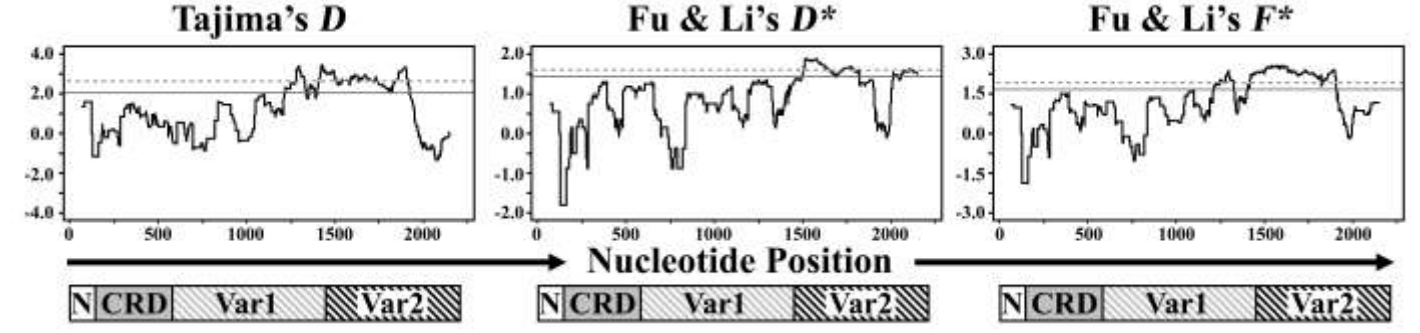

Fig. 2. Sliding window plots of Tajima's test (*D*) and Fu and Li's tests (*D and *F**) for *Plasmodium falciparum surf$_{4.2}$* sequence encoding extracellular region in Thai isolates.** Forty-four samples for 1988/1989 group are used. Sites above the solid and dashed lines are significantly departed from neutrality (two-tailed; $P < 0.05$ and $P < 0.02$, respectively), indicating diversifying selection. Nucleotide numbers are after the 3D7 line sequence. Window length is 90 bp, and step size is 3 bp.

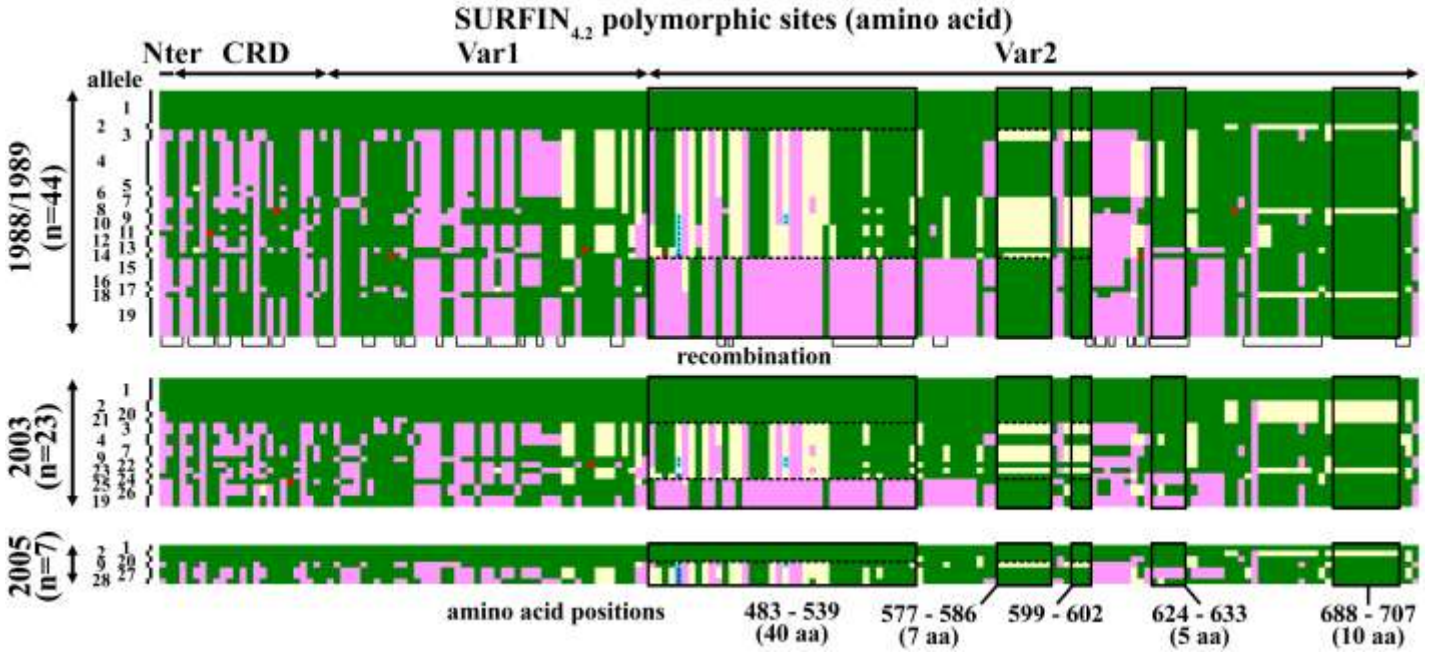

Fig. 3. Polymorphic amino acid sites of SURFIN$_{4.2}$ extracellular region among 74 Thai isolates. Only polymorphic amino acid sites were selected and aligned, then allelic type numbers were given to each sequence as follows; allele 1 (MS843, MS840, MS826, MS821, MS814A1, MS805, AA1329, AQ1132, TMPF44, AQ1125, and AQ1423), 2 (MS804, AQ1097a, AQ1099, and PA020), 3 (MS820b, MS819a, AQ1098, and AQ1142), 4 (MS844, MS946, MS818, MS816, MS815, MS813, MS808, MS824b, AQ1133, and AQ1097b), 5 (MS802b), 6 (MS811), 7 (MS809, MS833, TMPF18, and TMPF09), 8 (MS842), 9 (MS817, MS837, AQ1105b, and PA009), 10 (MS803a), 11 (MS827), 12 (MS807 and MS948), 13 (MS802a), 14 (MS829a), 15 (MS820a, MS819b, MS812, and MS825), 16 (MS828), 17 (MS947), 18 (MS806), 19 (MS831, MS830, MS829b, MS824a, MS803b, MS810, MS838, AQ1126, and AQ1129), 20 (TMPF34), 21 (AQ1139), 22 (TMPF11), 23 (AQ1127), 24 (AQ1105a, Q2D015, and TMPF338), 25 (AQ1130), 26 (AQ1101 and TMPF15), and 27 (PA021), 28 (AQ1459). MS843 sequence is used as a reference and shown with green color, because this sequence is most dominant (11/74; 14.9%). Amino acids different from MS843 are shown with pink or yellow for dimorphic sites, pink and yellow for trimorphic sites. Forth or fifth substitutions are shown with cyan or blue colors and black dots. Singleton amino acid substitutions are shown with red color. Amino acid positions 483 - 539 are clustered and can be divided into 3 patterns; allele 1, 8, and 19 pattern (boxed and separated by dashed lines). These trends can be seen at the amino acid positions 577 - 586 and 599 - 602 (boxed and separated by dashed lines). Thin lines under the scheme for 1988/1989 group connect the sites for which recombination events are detected. Note that the recombination between amino acid positions 483 - 539 are only detected within the allele 8 group and 4th and 5th amino acids are seen in this group.

We further evaluated signatures of selection by population genetics-based approaches; Tajima's *D*, Fu and Li's *D**, Fu and Li's *F** tests for 1988/1989 group (Table 2). Samples obtained from the other periods were excluded due to the low sample number. Significant positive values of Tajima's *D* were not detected for the entire extracellular region or four sub-regions. However, sliding window plot analysis depicted significant positive *D* values in Var1 and Var2 regions, suggesting the action of positive diversifying selection on these regions (Fig. 2). A significant positive value of Fu and Li's *D** (1.66) and *F** (2.02) were detected for the entire extracellular region ($p < 0.02$), respectively. When the four sub-regions were separately assessed, significant deviations greater than zero were detected for Var2 (*D** = 1.77 and *F** = 2.13; $p < 0.02$) and Var1 (*F** = 1.84; $p < 0.05$). Sliding window plot analysis of Fu and Li's *D** revealed significant positive deviation of greater than zero in the Var2 region. Sliding window plot analysis of Fu and Li's *F** revealed significant positive deviation of greater than zero in the Var1 and Var2 regions, for which the value of the Var2 region was higher than that of Var1 region. These results were consistent with the analysis based on the sub-regions. Collectively, comparison between non-synonymous and synonymous substitutions, Tajima's *D* test, and Fu and Li's *D** and *F** tests all detected the signature of positive diversifying selection on the Var2 region at the 98% confidence level and Tajima's *D* and Fu and Li's *F** tests detected that on Var1 region at the 98% confidence level.

3.3. Linkage disequilibrium and recombination of surf$_{4.2}$

To further assess the polymorphic nature of SURFIN$_{4.2}$, polymorphic amino acid sites were aligned and allele numbers were assigned (Fig. 3). We noticed that aa 483 - 539 were clustered and could be divided into three patterns; allele 1, 8, and 19 patterns. This pattern appeared to be extended further 3' side at the amino acid positions 577 - 586 and 599 - 602 where dimorphic substitutions were seen in the parasites possessing allele 3 pattern at aa 483 - 539. Thus, we evaluated linkage disequilibrium (LD) on this gene by calculating a standardized I_A (I_A^S) and found that I_A^S of 1988/1989 group sequences was 0.1459 ($p < 0.00001$), indicating significant LD of *surf$_{4.2}$* sequences. Significant |D'| and r^2 values seen in the right upper corner of the panels in figure 4 indicated LD between sites over a long distance, for example, significant values could be seen for the distance more than 1.5 kb for all plots (Fig. 4). Previous report on the Kenyan isolates detected a strong LD between nt 76 in Nter and nt sites located in nt 1473 - 1870 in Var2 (~1.5 kb apart; $p < 0.001$ after Bonferroni correction) and an epistatic relationship between these sites was proposed [5]. Our analysis also detected LD between nt 76 and nt sites located in nt 1352 - 1955 for 1988/1989 group ($p < 0.001$) after Bonferroni correction, which was consistent to the previous observation. Because more than 5 sequences belonged to the alleles 1, 4 or 19, which likely contributed to the observed LD, only one sequence from each allele were used to assess if LD was still detected or not. An obtained I_A^S value was 0.1119 ($p < 0.00001$) for 1988/1989 (n = 19), further supporting the LD on *surf$_{4.2}$* in Thai isolates.

Because LD was observed for the sites with the long distance, we evaluated the recombination events on *surf$_{4.2}$* using sequences originated from Thailand. To this end, 35 minimum recombination events were detected throughout the entire sequence obtained for 1988/1989 (detected recombination between nonsynonymous substitutions were plotted in Fig. 3), except some areas in Var2 where multiple amino acids were clustered, such as aa 483 - 539. This indicates

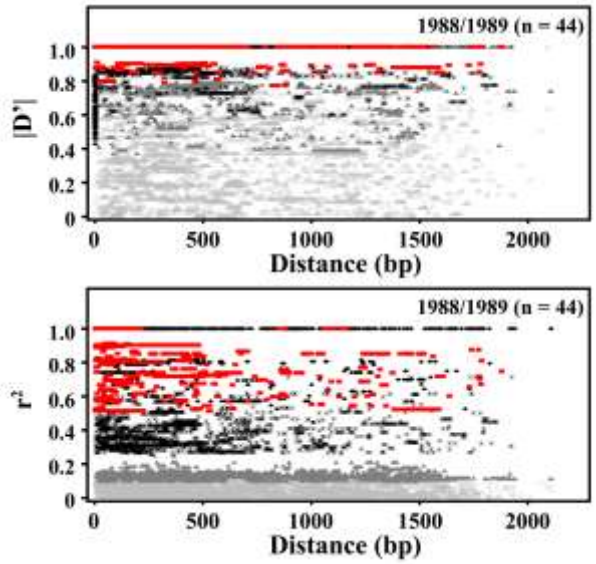

Fig. 4. Linkage disequilibrium (LD) of *surf*$_{4.2}$ region encoding extracellular region. LD is evaluated for 1988/1989 group using only informative sites after excluding sites with the frequency of the rare allele less than 10% and sites segregating for more than two nucleotides (number of polymorphic sites analyzed was 213). |D'| and r^2 were obtained by computing with DnaSP 5.0. Red filled square, $p < 0.001$ after Bonferroni correction (more conservative); asterisks, $p < 0.001$; black filled triangle, $0.001 \leq p < 0.01$; dark gray filled triangle, $0.01 \leq p < 0.05$; light gray open triangle, not significant. Correlation coefficients for |D'| and r^2 were -0.227 and -0.258, respectively.

that recombination could occur on this gene except highly polymorphic areas, which would prevent efficient recombination between too diversified sequences. The recombination and 4th and 5th amino acids at the aa 483 - 539 occurred within allele 8 pattern, suggesting that this pattern contributed to the complexity of *surf*$_{4.2}$ polymorphism more than remaining 2 patterns.

3.4. Frequency distribution of the polymorphism of surf$_{4.2}$.

When a parasite population possesses more than one allele of an antigen-encoding gene at one time point, the most common allele may be targeted by immune selection pressure, and as a result, relatively rare alleles may expand in the population. Under such a hypothesis, a fluctuation of allele-frequency distribution may be observed. Thus we evaluated if the allele-frequency distribution of *surf*$_{4.2}$ in 2003 was changed from 1988/1989. We selected four areas in Var2 region where more than four amino acids were clustered for this analysis (Fig. 3), because positive selection was detected on this region. As a control to evaluate the temporal change of the allele-frequency distribution, we obtained the information of nucleotide polymorphisms of four gene loci (nt 165 and 319 of PFB0200c, aspartate aminotransferase; nt 267 and 1008 of PFB0295w, adenylosuccinate lyase; nt 1989 of PF10_0363, pyruvate kinase; and nt 819 and 969 of PF14_0124, actin II) that were expected to be neutral to immune or drug selection pressure (Fig. S1). Because two *surf*$_{4.2}$ sequences were detected from MS802, MS803, MS819, MS820, MS824, MS829, AQ1097, and AQ1105, and one or two of the 4 loci showed mixed peaks for MS826 and MS818, sequences obtained from these parasites were excluded from the analysis. Allele frequency distribution of four putatively neutral loci were not significantly different between 1988/1989 and 2003 (Table S3), suggesting that at least there were no obvious change in the population structure that could be detected with these loci. We found also there were no statistically significant difference for the allele frequency distribution of *surf*$_{4.2}$ between two groups (Fig. 5), thus no obvious fluctuation of allele-frequency

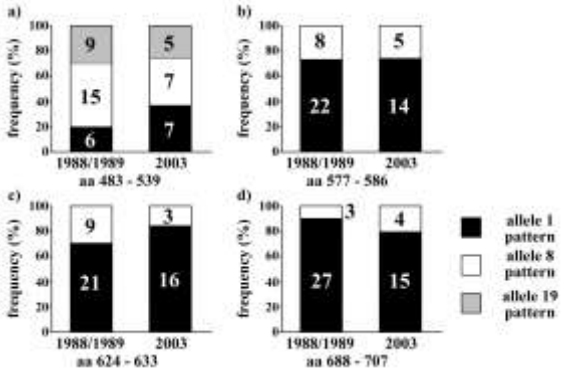

Fig. 5. Stable allele frequency distribution of the selected polymorphic areas of *Plasmodium falciparum* SURFIN$_{4.2}$ Var2 region in Thai isolates between 1988/1989 and 2003. Only sequences for which allele of the putatively neutral loci were determined are used (30 samples for 1988/1989 group and 19 samples for 2003 group). Allele frequencies are calculated for the regions where more than 4 amino acids are clustered (a, amino acid positions (aa) 483 - 539; b, aa 577 - 586; c, aa 624 - 633; d, and aa 688 - 707). Classification is detailed in figure 2. Allele frequency distributions are not significantly different between two groups (a, $p = 0.42$; b, $p = 0.98$; c, $p = 0.32$; d, $p = 0.41$). Chi-square test was used for (a) and (b), and Fisher's exact test was used for (c) and (d) (two-tail).

distribution were detected. This suggested that the frequency distribution of at least these four areas in the Var2 region were stable or did not change significantly in 14 years.

4. Discussion

In this study, we analyzed the extracellular domain-encoding nucleotide sequence of *P. falciparum surf*$_{4.2}$ for diversity, the signature of positive selection, the degree of linkage disequilibrium (LD), and temporal changes in allele frequency distribution in a *P. falciparum* population from Thailand. We found that the *surf*$_{4.2}$ gene of Thai isolates was highly polymorphic, particularly at the C-terminal side of the variable region (Var2 region), which is situated just before a predicted transmembrane region. Multiple tests detected the signature of positive selection on the Var2 region and, to a lesser extent, on the Var1 region, suggesting that they were potentially to be under host immune pressure. Similar experiments conducted for Kenyan isolates also showed similar tendency; positive selection toward the C-terminal side of extracellular region [5]. Although a significant departure from neutrality was not detected on the CRD, the nucleotide diversity of this region ($\pi = 0.0170$) is comparable to the known malaria vaccine candidate antigen *ama-1* ($\pi = 0.0166$), for which the action of the positive selection was previously detected [19]. Because sliding window analysis for Tajima's *D*, Fu and Li's *D** and *F** all showed negative values on some areas of CRD region (Fig. 2), we consider that a functional constraint likely prevented an extensive diversification of this region. Of note is that similar negative values were observed in the N-terminal side of Var1 region and this can be also seen in the analysis of Kenyan *P. falciparum* isolates [5]. The region between the CRD and transmembrane region of SURFIN members is highly diversified, and we expected that the variable (Var1 + Var2) region of SURFIN$_{4.2}$ has less functional constraint, thus the observed high diversity of C-terminal side Var1 and Var2 regions with the signature of positive selection is consistent with our expectation, whereas the negative values of Tajima's *D*, Fu and Li's *D** and *F** in N-terminal side of Var1 region is unexpected. Although Var1 region is not conserved among SURFIN members, this region might gain a functional role or structural importance during a shape-up process of current SURFIN$_{4.2}$ structure.

In contrast to the Kenyan *P. falciparum* $surf_{4.2}$ sequence, which showed 68 distinct allele in a total of 69 sequences, we found less alleles in Thai isolates (28 distinct alleles in a total of 74 sequences). Seven alleles (allele 1, 2, 3, 4, 7, 9, and 19) among 19 alleles detected in 1988/1989 were present in 2003/2005 sample collection. As recombination events were detected in both the Thai and the Kenyan sequences [5], we consider that the following factors contributed to this observation: 1) self-fertilization is dominant and frequency of the recombination of $surf_{4.2}$ gene locus is low, 2) epistatic relation between two sites with long distance (> 1.5 kb), suggested by LD analysis in both the Thai and Kenyan sequences, suppresses the emergence of recombinant types of particular combinations, 3) high nucleotide diversity, especially in Var1 and Var2 regions may prevent efficient recombination. In the African population where recombination is more frequent than Asia, LD is less than Thai population. Thus in Thailand, even though recombination may occur, the effect of the epistatic relation and/or high diversity dominates the frequency of the recombination, and as a result same $surf_{4.2}$ alleles may circulate in the same population for long time.

Amino acid substitutions with an intermediate frequency are favored and expected to be selected under a frequency-dependent selection (balancing selection) and the selection pressure against $surf_{4.2}$ is likely to be host immunity. The frequency distribution of selected areas in Var2 region did not change for 14 years, suggesting that the allele frequency distribution of this region was stable. To explain this observation, we consider following 3 scenarios: 1) there is no pressure to maintain particular allele-frequency distribution of $SURFIN_{4.2}$ Var2 region, and the current distribution was formed by chance (neutral); 2) The cycle of fluctuation is longer than 14 years so that current data set is not sufficient to detect signature of fluctuation, if any (adapted with a fluctuation); and 3) the observed allele frequency of $SURFIN_{4.2}$ Var2 region has already adapted to the local environment (e.g., human immunity and/or human genetic background) in Thai *P. falciparum* population, and stably maintained (adapted without a fluctuation). Unless the structure of Thai *P. falciparum* population is dramatically changed, for example, by a genetic drift effect due to a small population size or a migration of new alleles, it is difficult to assess the first scenario. A stable allele frequency distribution of four putative neutral loci for 14 years also failed to rule out the first scenario, however, because of the positive diversifying selections were detected on this region, we consider scenarios 2 or 3 are more likely. To assess the possibility of second scenario, further study is required with more samples from different time points. The third scenario is appealing, because such temporally stable allele frequency distribution was also seen in most of the region of *P. falciparum* merozoite surface protein 1 (MSP1), a vaccine candidate antigen under diversifying selection, for 7 years in the Gambia [20] and for 10 years in Tanzania [21]. Allele frequency distribution of block 2 (most diversified region) of MSP1 was shown to be stable among *P. falciparum* populations in the different geographic areas and antibodies against this region were strongly associated with the protection against *P. falciparum* infection [22]. Spatially and temporally stable serotype frequency distribution can be observed for *Streptococcus pneumoniae*, a causative agent of the respiratory infection, however, several reports support that human immunity is unlikely to be the selection pressure [reviewed in 23]. Thus, although we consider that the selection pressure of the $SURFIN_{4.2}$ diversity is likely human immunity by observing its extensive diversity, it is formally possible that the other factor, such as host genetic background, may be responsible. Further studies are required if Var2 region of $SURFIN_{4.2}$ could be a target of the protective immunity and observed stable allele frequency distribution are maintained by allele-specific immunity, in addition to the identification of its biological role.

Acknowledgements

We thank S. Miyashita for her expertise. We are grateful to I. Sekine, head of the Nagasaki Red Cross Blood Center for human erythrocyte and plasma. This work was supported in part by Grants-in-Aids for Scientific Research 19590428 (to OK) and the Global COE Program, Nagasaki University (to OK) from the Ministry of Education, Culture, Sports, Science and Technology, Japan. The nucleotide sequence data reported in this paper are available in the GenBank™/EMBL/DDBJ databases under the accession numbers: AB679835 – AB679908.

References

[1] World Health Organization. World Malaria Report 2010. Geneva, Switzerland: World Health Organization; 2010.

[2] Ferreira MU, da Silva Nunes M, Wunderlich G. Antigenic diversity and immune evasion by malaria parasites. Clin Diagn Lab Immunol 2004;11:987-95.

[3] Winter G, Satoru K, Haeggstrom 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;201:1853-63.

[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;269:431–6.

[5] Ochola LI, Tetteh KK, 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;27:2344-51.

[6] Nakazawa S, Culleton R, Maeno Y. In vivo and in vitro gametocyte production of *Plasmodium falciparum* isolates from Northern Thailand. Int J Parasitol 2011;41:317-23.

[7] 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.

[8] Trager W, Jensen JB. Human malaria parasites in continuous culture. Science 1976;193:673-5.

[9] Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Mol Biol Evol 2007;24:1596-9.

[10] Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 1986;3:418-26.

[11] Librado P, Rozas J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 2009;25:1451-2.

[12] Tajima F. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 1993;135:599-607.

[13] Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics 1993;133:693-709.

[14] Maynard Smith J, Smith NH, Dowson CG, Spratt BG. How clonal are bacteria? Proc Natl Acad Sci USA 1993;90:4384-8.

[15] Hudson RR, Kaplan NL. Statistical properties of the

number of recombination events in the history of a sample of DNA sequences. Genetics 1985;111:147-64.

[16] Haubold B, Hudson RR. LIAN 3.0: detecting linkage disequilibrium in multilocus data. Linkage Analysis. Bioinformatics 2000;16:847-8.

[17] Lewontin RC. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics 1964;49:49-67.

[18] Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor Appl Genet 1968;38:226–31.

[19] Polley SD, Conway DJ. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. Genetics 2001;158:1505-12.

[20] Conway DJ, Greenwood BM, McBride JS. Longitudinal study of *Plasmodium falciparum* polymorphic antigens in a malaria-endemic population. Infect Immun 1992;60:1122-7.

[21] 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;76:1037-45.

[22] 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;6:689-92.

[23] Lipsitch M, O'Hagan JJ. Patterns of antigenic diversity and the mechanisms that maintain them. J R Soc Interface 2007;22:787-802.

Table 1. Nucleotide diversity of *Plasmodium falciparum surf*$_{4.2}$ from Thai isolates (n = 74)

region (position)	number of sites (base)	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	2166	92.027	78.825	1711.288	13.203	454.712	0.0440	0.0479	0.0298	1.61	0.0004
(28–2193)		(5.484)	(5.350)	(7.900)	(2.192)	(8.182)	(0.0027)	(0.0035)	(0.0049)		
Nter	123	0.870	0.870	98.833	0	24.167	0.0071	0.0089	0	∞	
(28–150)		(0.570)	(0.589)	(1.741)	(0)	(1.692)	(0.0047)	0.0060)	(0)		
CRD	435	7.284	6.747	357.214	0.537	77.786	0.0170	0.0191	0.0070	2.73	
(151–585)		(1.599)	(1.451)	(3.469)	(0.368)	(3.443)	(0.0035)	0.0043)	(0.0050)		
Var1	861	20.563	17.924	673.883	2.639	187.117	0.0244	0.0272	0.0143	1.90	
(586–1446)		(2.598)	(2.613)	(5.348)	(1.030)	(5.457)	(0.0032)	(0.0045)	(0.0058)		
Var2	747	63.311	53.284	581.358	10.026	165.642	0.0917	0.0998	0.0641	1.56	< 0.005
(1447–2193)		(4.419)	(4.273)	(4.701)	(1.829)	(4.227)	(0.0071)	0.0088)	(0.0118)		

Extracellular, extracellular region; Nter, N-terminal segment;CRD, cysteine-rich domain; Var1, variable region 1; Var2, variable region 2. sites, sites nucleotide analyzed. k, the average number of nucleotide differences; *N* and *S*, average numbers of nonsynonymous and synonymous sites; π, pairwise nucleotide diversity; d_N, number of nonsynonymous substitutions over number of nonsynonymous 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 numbers of synonymous (*Sd*) and nonsynonymous (*Nd*) differences were calculated by Nei-Gojobori method. *p* value indicates the statistical difference between d_N and d_S, tested using a one-tail Z-test with 500 bootstrap pseudosamples implemented in MEGA4.0.

Table 2. Test of neutrality for *Plasmodium falciparum surf*$_{4.2}$ from Thai isolates collected in 1988 - 1989 (n = 44)

	nucleotide	number of			two variants		more than			Tajima's	Fu and Li's	
region	position	sites (base)	η	*S*	(singleton)	(not singleton)	two variants	π	θ	*D*	*D**	*F**
Extracellular	28–2193	(2166)	262	247	11	222	14	0.041	0.028	1.75	1.66**	2.02**
Nter	28–150	(123)	2	2	0	2	0	0.007	0.004	1.36	0.76	1.09
CRD	151–585	(435)	24	24	3	21	0	0.015	0.013	0.72	0.81	0.93
Var1	586–1446	(861)	55	55	4	51	0	0.023	0.015	1.89	1.35	1.84*
Var2	1447–2193	(747)	181	166	4	148	14	0.083	0.056	1.80	1.77**	2.13**

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.

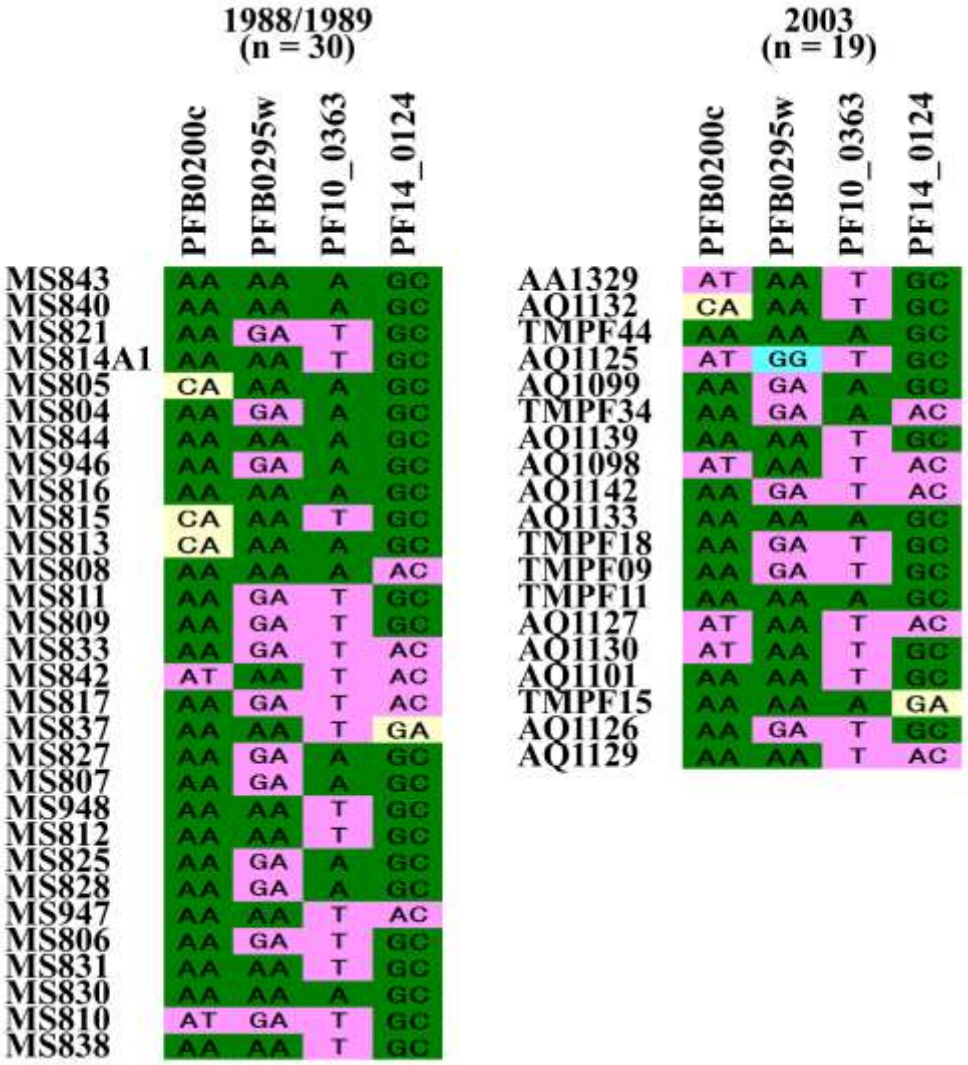

Fig. S1. Allelic type of four putative neutral markers of *Plasmodium falciparum* Thai isolates. MS843 allele is shown with green color, and the other alleles are shown with pink, yellow or cyan. Nucleotide residues at nucleotide (nt) positions 165 and 319 of FB0200c, nt 267 and 1008 of PFB0295w, nt 1989 of PF10_0363, and nt 819 and 969 of PF14_0124 are sequenced and allelic types were determined. MS802, MS803, MS818, MS819, MS820, MS824, MS826, MS829, AQ1097, and AQ1105 are excluded because of the mix-infection in one of the loci analyzed.

Table S1. Additional oligonucleotide primers used for the sequencing

Name	Sequence (5'→3')
F1	GAATTAAAAAGGTCAGGATCTG
F2	ATATGAAAAAAATGAAGGAAGATG
F2_D	ATATGAAAAAAAATAAGGAAGATG
F8	GGACTATATTCACCTGTATCAC
F9	GGAAAGTTCTGTAGGAACAGATAAG
R0	CAGTAGGTGAATTTTCCTCCT
R3	TGCACCTTCTTGAGTAGTTTC
R7	ACTTGATATATTAAGGAATAATTTATCC

Table S2. Oligonucleotide primer for PCR-amplification and sequencing for putatively neutral loci

ID	Primer	forward	reverse
PFB0295w	PCR and sequencing	TTATTATGGATGTACATGTGAACC	TTTATATATTCCTGTGAGAAGTGC
	sequencing	CATGTGAACCAACTGAAAAACATATC	GGTAGTTGGTGATGCAGGTTGTCC
	sequencing	TTGTATACATAATATAATAATACC	AAACATTTGATGAGATATATAACC
	sequencing	ATCGATTTATCTGTTGATATGTGG	TGAGAAGTGCTCCACATTTTTTGAC
PFB0200c	PCR and sequencing	CAGCTTAGAAAATATCGAAGTCG	GTAAAATAACTGAGGATCCATTTGG
	sequencing	GAGTTTAAGGAAGATACATGTGAGG	CATATTCACATGATTAATATAAGGTGG
PF10_0363	PCR and sequencing	CTAATTTGTTAGACATGGTTGAC	ATGATAAATTAAAAATGTTGGATC
	sequencing	GTAACAAGTGAATATTCATCAGG	GAGCAAGATATTCAAAATAATATATCC
PF14_0124	PCR and sequencing	GCTGAAAGGGAGATAGTTAGGG	TCAGGTGGAGCAATGACC
	sequencing	GAAAAACTATGTTACATAGCCATGG	ATTTCTTTGGTTAGCCTTTCTCC

Table S3. Temporal change of allele-frequency distribution of putatively neutral loci [a]

ID	product	chr	nucleotide position	type	1988/1989	2003	*p* value [b]
PFB0200c	aspartate aminotransferase	2	165	Syn	A/C (27/3)	A/C (18/1)	0.69*
			319	Nsyn	A/T (28/2)	A/T (14/5)	0.93*
			165 + 319		AA/AT/CA (25/2/3)	AA/AT/CA (13/5/1)	0.15
PFB0295w	adenylosuccinate lyase	2	267	Syn	A/G (17/13)	A/G (12/7)	0.65
			1008	Syn	A/G (30/0)	A/G (18/1)	0.39*
			267 + 1008		AA/GA/GG (17/13/0)	AA/GA/GG (12/6/1)	0.35
PF10_0363	pyruvate kinase	10	1989	Syn	A/T (14/16)	A/T (6/13)	0.3
PF14_0124	actin II	14	819	Syn	A/G (5/25)	A/G (5/14)	0.41
			969	Syn	A/C (1/29)	A/C (1/18)	1.00*
			819 + 969		AC/GA/GC (5/1/24)	AC/GA/GC (5/1/13)	0.66

[a] Thirty data set (MS804, MS805, MS806, MS807, MS808, MS809, MS810, MS811, MS812, MS813, MS814A1, MS815, MS816, MS817, MS821, MS825, MS827, MS828, MS830, MS831, MS833, MS837, MS838, MS840, MS842, MS843, MS844, MS946, MS947, and MS948) for the isolates collected in 1988 - 1989 (collectively termed as 1988/1989 group) and 19 data set (AA1329, AQ1098, AQ1099, AQ1101, AQ1125, AQ1126, AQ1127, AQ1129, AQ1130, AQ1132, AQ1133, AQ1139, AQ1142, TMPF09, TMPF11, TMPF15, TMPF18, TMPF34, and TMPF44) for the isolates collected in 2003 were used, after excluding isolates showing mix-infection in one of the loci analyzed including $surf_{4.2}$ gene locus.

[b] *p*-values are obtained by chi-square test or Fisher's exact test (two-tail, asterisk).