Genetic Polymorphism of *Plasmodium vivax* msp1p, a Paralog of Merozoite Surface Protein 1 from Worldwide Isolates

Yue Wang,‡ Osamu Kaneko,† Jetsumon Sattabongkot, Jun-Hu Chen, Feng Lu, Jong-Yil Chai, Satoru Takeo, Takaumi Tsuboi, Francisco J. Ayala, Yong Chen, Chae Seung Lim, and Eun-Taek Han*

Department of Parasitology, Kangwon National University College of Medicine, Chuncheon, Gangwon-do, Republic of Korea; Institute of Parasitic Diseases, Zhejiang Academy of Medical Sciences, Hangzhou, People’s Republic of China; Department of Protozoology, Institute of Tropical Medicine and the Global Center of Excellence Program, Nagasaki University, Nagasaki, Japan; Department of Entomology, Armed Forces Research Institute of Medical Sciences, Bangkok, Thailand; Jiangsu Institute of Parasitic Diseases, Wuxi, People’s Republic of China; Department of Parasitology and Tropical Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea; Cell-Free Science and Technology Research Center, Ehime University, Ehime, Japan; Department of Ecology and Evolutionary Biology, University of California, Irvine, California; Zhejiang Medical College and Zhejiang Academy of Medical Sciences, Hangzhou, People’s Republic of China; Department of Laboratory Medicine, College of Medicine, Korea University, Seoul, Republic of Korea

Abstract. *Plasmodium vivax* msp1p, a paralog of the candidate vaccine antigen *P. vivax* merozoite surface protein 1, possesses a signal peptide at its N-terminus and two epidermal growth factor–like domains at its C-terminus with a glycosylphosphatidylinositol attachment motif. Twenty-one distinguishable allelic types (A1–A21) of the E/Q-rich region were identified from worldwide isolates. Among them, four types were detected in isolates from South Korea. The length polymorphism of the E/Q-rich region might be useful as a genetic marker for population structure studies in malaria-endemic areas.

INTRODUCTION

Among the species of malarial parasite that infect humans, *Plasmodium vivax* is the most globally prevalent and threatens almost 40% of the world’s population, resulting in approximately 250 million clinical infections each year. Although *P. vivax* malaria had been considered relatively benign, compared with that of *P. falciparum*, this view is now being challenged. Additionally, resistance to chloroquine is appearing in *P. vivax*. Merozoite surface proteins have been characterized and are highly immunogenic in natural infection. Among the species of malarial parasites that infect humans, *P. vivax* is the most globally prevalent and threatens almost 40% of the world’s population, resulting in approximately 250 million clinical infections each year. Although *P. vivax* malaria had been considered relatively benign, compared with that of *P. falciparum*, this view is now being challenged. Additionally, resistance to chloroquine is appearing in countries where malaria is endemic. Thus, there are good reasons to pursue an effective *P. vivax* vaccine. In this regard and in contrast to *P. falciparum*, research into *P. vivax* is limited, due in part to difficulties in culturing blood-stage parasites in vitro. Nevertheless, several *P. vivax* vaccine candidates from different parasitic stages have been characterized. Among them, various merozoite surface proteins (MSPs), apical membrane protein-1, duffy binding protein, Pvs25, Pvs28, circumsporozoite protein, and thrombospondin-related anonymous proteins have been studied. Merozoite surface proteins have been characterized and are highly immunogenic in natural infection. Among them, several major vaccine candidate antigens (including MSP1, MSP4, MSP5, MSP8, and MSP10) are either known or presumed glycosylphosphatidylinositol (GPI)–anchored membrane proteins. *Plasmodium vivax* MSP1 is the largest and most abundant protein on the *P. vivax* merozoite surface. The gene that encodes this protein (*Pvmsp1p*) has been studied. 5

The gene that encodes this protein (Pvmsp1p) is highly polymorphic and consists of a mosaic of conserved and variable blocks with numerous recombination sites distributed throughout the gene. However, the fragment that encodes the 19-kDa C-terminal epidermal growth factor (EGF)–like domain is relatively conserved. This gene has been used as a polymorphic marker for investigations of the genetic structure of *P. vivax* populations and in molecular epidemiology.

With the completion of *P. vivax* genome sequencing, GPI-anchored proteins of *P. vivax* have been predicted by comparison with validated *P. falciparum* GPI-anchored proteins. *Plasmodium vivax* msp1p (Pvmsp1p), a novel paralog of the Pvmsp1 gene, was found immediately upstream of Pvmsp1. This gene is predicted to encode a 1,854-amino-acid protein (predicted molecular mass of 215 kDa) with an N-terminal signal sequence, C-terminal EGF-like domains, and a GPI-attachment motif (Figure 1). The functions of this molecule remain unknown. Thus, we have analyzed available genomic data from PlasmoDB (http://www.plasmodb.org/) to search for distinctive pattern of diversity in msp1p and msp1 genes among *Plasmodium* species. We have also assessed the nature and extent of polymorphisms in PvMSP1P from worldwide isolates and laboratory lines of *P. vivax*.

MATERIALS AND METHODS

Gene sequences. The following sequences of malarial parasites were used for the analyses: human *P. falciparum* PfMSP1 (CA27070), PfMSP8 (PFE0120c), and PfMSP10 (PFF0995c), and *P. vivax* PvMSP1 (PVX_099980), PvMSP8 (PVX_097625), PvMSP10 (PVX_114145), and PvMSP1P (PVX_099975); rodent malaria *P. berghei* PbMSP1 (AAC28851), PbMSP8 (PBANKA_110220), and PbMSP10 (PBANKA_111960), *P. yoelii* PyMSP1 (PY05748), and *P. chabaudi* PchMSP1 (PCAS_083080); primate *P. knowlesi* PrMSP1 (PKH_072850) and PrMSP1P (PKH_072840), *P. reichenowi* PrMSP1 (CAH10285), and *P. cynomolgi* PcyMSP1 (BAI82251); and avian malaria *P. gallinaceum* PgMSP1 (CAH10838). The *P. gallinaceum* sequence database (http://www.sanger.ac.uk/) was used to search for homologs of PvMSP1 and PvMSP10.

Blood samples and DNA preparation. Blood samples were collected, after informed consent had been obtained, from 81
symptomatic patients diagnosed by microscopic examination with *P. vivax* infection at Korea University Ansan Hospital, local health centers, and clinics in Gyeonggi and Gangwon provinces, Republic of Korea. Genomic DNA was purified from 200 μl of whole blood by using a QIAamp DNA Blood Mini Kit (QIAGEN, Valencia, CA), according to the manufacturer's protocol. Genomic DNA of *P. vivax* isolates (n = 33) obtained from Thailand (n = 26), Indonesia (n = 3), India (n = 1), Papua New Guinea (n = 1), Western Samoa (n = 1), and Pakistan (n = 1), and nine *P. vivax* laboratory lines (Africa Mauritania, New Guinea, Honduras III, Brazil I, Salvador I, Vietnam IV, Indonesia I, India VII, and Columbia Rio Meta) were used for polymerase chain reaction (PCR).

**Amplification and sequencing of target genes.** Genomic DNA from 20 isolates from the Republic of Korea and 9 isolates from other locations was used for amplification of the *Pvmsp1p* full-length gene. Primers SeqF1 (5′-TGC ATA TTC ATA ATG CCG AGT TCG TGA T-3′) and SeqR8 (5′-GCT GTC AAA TCG TGG CAG-3′) were designed based on the *Pvmsp1p* sequence of the *P. vivax* Sal I strain. These were used to amplify DNA fragments from 90 basepairs upstream to 80 basepairs downstream of the *Pvmsp1p* coding sequence by using LA Taq DNA polymerase (TaKaRa, Tokyo, Japan). The PCR amplification was performed on a MyCycler Thermal Cycler (Bio-Rad, Hercules, CA) by using the following temperature profile: 94°C for 2 minutes; 35 cycles at 94°C for 30 seconds, 64°C for 30 seconds, 72°C for 6 minutes; and a final extension at 72°C for 10 minutes.

Both strands of the PCR products were directly sequenced by using a series of sequencing forward primers (SeqF1; SeqF2: 5′-ATC AAC CGG AAG AAC TCC CT-3′; SeqF3: 5′-CAA AGG GAG AAG AAA AAA ATG TAC C-3′; SeqF4: 5′-GTT GAG CTA ATC GAA CCG G-3′; SeqF5: 5′-TGG GGC GCA CAT AAC CT-3′; SeqF6: 5′-CCC GTC TAC TCC AAG GAT GTG ATG ATA AG-3′; SeqF7: 5′-TGA AGT GCA ACA CGT GGA AT-3′; SeqF8: 5′-GTG GAC TACTAGGG CTAG CATG A-3′ and SeqF9: 5′-ATT CTC TAT TCA GAC AAC AAG GAG GTG GTG-3′) and reverse primers (SeqR1: 5′-AGG GAA GGA TTA GAG ACGG; SeqR2: 5′-GCAGCTTTATGGTGG TAG TC-3′; SeqR3: 5′-AGC GTT AAA TGG TGG CAGG-3′; SeqR4: 5′-TTC GTG ATG ATC GCG TTG GGT AGC AG-3′; SeqR5: 5′-TCC CCG ATG AAA TAT GC-3′; SeqR6: 5′-ACT GCA GAT GGA TGG TCA TCT-3′; SeqR7: 5′-AAC TGC ATC GCG TTC GTA T-3′; and SeqR8) using an ABI Prism 377 DNA sequencer (Genotek, Seoul, South Korea).

Analysis of the full-length gene sequences showed one highly polymorphic region. This region was amplified by PCR from genomic DNA of 61 samples from the Republic of Korea and 24 samples from other locations, as well as nine laboratory strains, and sequenced. To examine variable tandem repeat regions, as found in the *Pvmsp1p* gene sequence, primers (TR-F: 5′-CCT ACA CGG GAT GGG AGA T-3′ and TR-R: 5′-CGG AGA CGT ACG TCG TGA T-3′) were used to amplify a 200-basepair fragment encompassing this region from 33 worldwide isolates and 9 laboratory lines.

**Data analysis.** Sequence data were submitted to GenBank under accession numbers GU556592–GU556620. Amino acid sequence alignments were constructed using the MUSCLE program, with manual corrections. The number of nonsynonymous substitutions per nonsynonymous site (dN) and the number of synonymous substitutions per synonymous site (dS) were computed by using the Nei-Gojobori method, accompanied by bootstrap analysis with 1,000 replicates for the neighbor-joining method and 100 for the maximum parsimony method implemented in PHYLIP version 3.68 after excluding insertions/deletions (indels) and unreliable amino acid sites.

**RESULTS**

An MSP1P homolog can be found in the *P. vivax* and *P. knowlesi* genome databases, but not in the *P. falciparum*, *P. yoelii*, or *P. berghei* genome databases. To investigate the evolutionary relationship of *MSP1P* with other MSPs, we searched for *Pvmsp1p* homologs in the available *Plasmodium* genome database and analyzed their relationship with MSP1 by using the distantly related MSP8 and MSP10 sequences as outgroups. A TBLASTN search of the *P. gallinaceum* sequence database (http://www.sanger.ac.uk) was conducted by using the PvMSP1P amino acid sequence as a query. This search identified a contig (28a.d00006175.Contig1) that contained a putative partial sequence of the *Pgmsp1p* gene (encoding the C-terminal end included EGF-like domains) (Figure 2). Based on the BLOSUM matrix, the amino acid sequence identity/similarity of the EGF-like domains to those of *PvMSP1P* and *PkmSP1P* were 56.71% and 58.75%, respectively (Figure 3A). The identity/similarity of the N-terminal region of this gene

**Figure 2.** Alignment of the partial amino acid sequences of *Plasmodium vivax* merozoite surface protein 1 paralog (PvMSP1P), *P. knowlesi* merozoite surface protein 1 paralog (PkMSP1P), and *P. gallinaceum* merozoite surface protein 1 paralog (PgMSP1P). Dashes indicate deletions. Cys residues with light areas indicate Cys residues conserved among all sequences and those with dark areas and the arrowhead indicate the additional two Cys residues conserved among MSP8, MSP10, MSP1P, and Pf/Pp/PmMSP1P (Figure 3). Asterisks, dots, and colons under the alignments indicate identical, conserved, and semi-conserved substitutions, respectively, based on BLOSUM. The glycosylphosphatidylinositol (GPI) modification site is indicated with arrowhead.
Second, beside the EGF-like domains, N-terminal side of the PgMSP1P and was separated from the MSP1 clade (Figure 3B). That PvMSP1P formed one clade with PkMSP1P and sons. First, the dendrogram using EGF-like domains indicated that PvMSP1P might be evolutionarily closer to PfMSP1 than PvMSP1. However, this appears not to be so for three reasons. First, the dendrogram using EGF-like domains indicated that PvMSP1P formed one clade with PkMSP1P and PgMSP1P and was separated from the MSP1 clade (Figure 3B). Second, beside the EGF-like domains, N-terminal side of the PgMSP1P showed greater similarity to Pv/PK/MSP1P (>60%) than PgMSP1 (<30%). Third, EGF-like domains of the distantly related MSP8 and MSP10 proteins contain two extra Cys residues, similar to MSP1P and Pf/Pr/PgMSP1. This finding indicates that the common ancestral protein of these MSPs possessed 12 Cys residues (Figure 3A). Collectively, these data suggest that the two Cys sites of the first MSP1 EGF-like domain in P. vivax, P. knowlesi, P. yoelii, and P. berghei were substituted with other amino acids during their evolution.

To assess the global genetic diversity of the Pmvmsp1p gene, we determined the Pmvmsp1p full-length sequence of 20 isolates of P. vivax from the Republic of Korea and 9 strains from other locations worldwide. One highly polymorphic glutamate (Glu, E)/glutamine (Gln, Q)-rich region and a polymorphic hepta-peptide motif (SAYSYVS) with number variation (single to triple) were detected. There was no amino acid polymorphism in the EGF-like domains (Figure 1). No diversifying selection was detected, and there was no significant excess of $d_{N}$ over $d_{S}$.

To determine the repeat variation of the hepta-peptide motif, additional genomic DNA from P. vivax worldwide, isolates and laboratory strains was amplified by PCR. All field isolates and laboratory strains had double repeats of this hepta-peptide motif, except for the isolates from Western Samoa (triple repeats) and Pakistan (single repeat). Outside of one E/Q-rich region and the hepta-peptide motif, the sequences were highly conserved with relatively few amino acid substitutions (Table 1).
Of these substitutions, only the S755I or P1686T (or both) mutations were found in 15 of the isolates from the Republic of Korea and Papua New Guinea, whereas more mutants were found in the worldwide isolates (Table 1).

A short and highly diverse region, composed of Glu and Gln as 3–5 Glu residues, followed by one or several basic E/Q (n = 1–6) units, was found in the Pvmsp1p gene. Twenty-one distinguishable allelic types (A1–A21) were identified in 127 isolates (clones), based on a comparison with corresponding regions in the P. vivax Sal I strain (Table 2). Type A1 had an identical sequence to that of the Sal I strain, which was found in only two laboratory strains, from Central and South America. Type A2 predominated (33.1%, 42 of 127) in all P. vivax samples, and in the Korean (35.8%, 29 of 81), Thai (26.7%, 8 of 30), Pacific (67%, 2 of 3), and African isolates (100%, 1 of 1), which share 96.7% amino acid identity with type A1. The 81 Korean isolates appeared to have limited diversity because only four genotypes (allelic types A2, A7, A20, A21) were found, whereas isolates from other locations worldwide, and laboratory strains, showed 20 allelic types (the exception being A21). Interestingly, type A21 was detected only in Korean isolates (18.5%, 15 of 81).

Polymerase chain reaction amplification resulted in two or three target bands in each of three Thai isolates, which suggested multi-clone infection. To confirm this finding, PCR products of the polymorphic region amplified from these samples was cloned and sequenced. Three types (A2, A16, A20) were detected from the Thai T21 isolate, two (A7, A15) from isolate T25, and two (A19, A20) from isolate T29.

**DISCUSSION**

We have assessed the evolutionary relationship of the msp1p gene with other msp genes and propose that a duplication event (msp1 and msp1p) occurred before the diversification of the clades in P. vivax and P. gallinaceum. This account requires two independent deletions of msp1p, one in the rodent lineage (after its divergence from the primate lineage to P. knowlesi and P. vivax) and another deletion in the lineage to P. falciparum (after its divergence from the primate lineage to P. knowlesi and P. vivax). We also propose that the common ancestor of P. vivax, P. knowlesi, P. yoelii, and P. berghei possessed MSP1 that had 12 Cys residues in the first EGF-like domain, and that two Cys sites were substituted to other amino acids during their evolution. We further investigated the genetic diversity of the Pvmsp1p gene in isolates from locations worldwide, including the Republic of Korea. We found an E/Q-rich polymorphic region, a hepta repeat region, and several polymorphic sites. However, no diversifying selection was apparent by comparing dN and dS. Although the molecular data (e.g., size, molecular mass, number, location of Cys residue) were similar to those of PvmSP1, PvmSP1P is not polymorphic and appears to not be under noticeable host immune pressure. However, the repeat-length polymorphism of the E/G-rich region may prove useful as a genetic marker for epidemiologic studies.

High conservation of the double EGF-like domains was also detected in other merozoite surface proteins, such as MSP1 and MSP4. These are involved in putative ligand-receptor interactions during erythrocyte invasion by merozoites. Thus, the lack of variation in the C-terminus sequence of PvmSP1P is not polymorphic and appears to not be under noticeable host immune pressure. However, the repeat-length polymorphism of the E/G-rich region may prove useful as a genetic marker for epidemiologic studies.

The overall nucleotide diversity of Pvmsp1p is much lower than that of other P. vivax antigens, such as MSP1, MSP3β, and apical membrane antigen 1. In the PvmSP1P sequences, the E/Q-rich region was shown to be highly polymorphic (21 allelic types in 127 clones/isolates). In the cases of PvmSP1 and PfSP230 (AF269242), the E/Q-rich region of PvmSP1P was highly polymorphic and represented the principal source of genetic diversity. In a low-complexity region analysis of P. vivax, Gln appeared with a somewhat higher frequency in the repetitive than in the non-repetitive motifs. The E/Q-rich regions and repeat motif of PvmSP1P and PfSP230 were located in a low-complexity region. These low-complexity regions harbor tandem repeats identified in P. vivax and correspond to species-specific and rapidly diverging regions.

This variation in E/Q-rich regions and the number of repeats could be generated by slipped-strand mispairing mechanisms. These result in duplication, deletion, or mutation of certain repeat units. The tandem repeat regions of PvmSP1P may result from rapid diversification, which enables the parasite to evade the immune response of the host by antigenic polymorphism.

Finally, the highly polymorphic E/Q-rich region sequence of PvmSP1P might be useful as a genetic marker for studies on the population structure and dynamics of P. vivax in malaria-endemic areas.
Table 2
Allelic types of the E/Q-rich region of Plasmodium merozoite surface protein 1 paralog and geographic prevalence of each *P. vivax* isolate from various locations*

<table>
<thead>
<tr>
<th>Allelic types</th>
<th>Amino acid sequence variation of E/Q-rich region sequence</th>
<th>GenBank accession no.</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>South Korea</td>
<td>Thailand</td>
</tr>
<tr>
<td>Sal I</td>
<td>EE—EQOQ—EQQEKQKK</td>
<td>1157</td>
<td>1174</td>
</tr>
<tr>
<td>A1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>E—E———EQQQ——EQQQKQKK</td>
<td>29 (35.8)</td>
<td>8 (26.6)</td>
</tr>
<tr>
<td>A3</td>
<td>EE—E——EQQQ——EQQQKQKK</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>A4</td>
<td>EE—E——EQQQ——EQQQKQKK</td>
<td>1 (3.3)</td>
<td>1 (3.3)</td>
</tr>
<tr>
<td>A5</td>
<td>EE—E——EQQQ——EQQQKQKK</td>
<td>1 (3.3)</td>
<td>1 (3.3)</td>
</tr>
<tr>
<td>A6</td>
<td>EE—E——EQQQ——EQQQKQKK</td>
<td>2 (6.7)</td>
<td>2 (6.7)</td>
</tr>
<tr>
<td>A7</td>
<td>EE—E——EQQQ——EQQQKQKK</td>
<td>1 (3.3)</td>
<td>1 (3.3)</td>
</tr>
<tr>
<td>A8</td>
<td>EE—E——EQQQ——EQQQKQKK</td>
<td>19 (22.2)</td>
<td>5 (16.7)</td>
</tr>
<tr>
<td>A9</td>
<td>EE—E——EQQQ——EQQQKQKK</td>
<td>1 (12.5)</td>
<td>1 (12.5)</td>
</tr>
<tr>
<td>A10</td>
<td>EE——E——EQQQ——EQQQKQKK</td>
<td>1 (12.5)</td>
<td>1 (12.5)</td>
</tr>
<tr>
<td>A11</td>
<td>EE——E——EQQQ——EQQQKQKK</td>
<td>1 (12.5)</td>
<td>1 (12.5)</td>
</tr>
<tr>
<td>A12</td>
<td>EE——E——EQQQ——EQQQKQKK</td>
<td>1 (12.5)</td>
<td>1 (12.5)</td>
</tr>
<tr>
<td>A13</td>
<td>EE——E——EQQQ——EQQQKQKK</td>
<td>1 (12.5)</td>
<td>1 (12.5)</td>
</tr>
<tr>
<td>A14</td>
<td>EE——E——EQQQ——EQQQKQKK</td>
<td>1 (12.5)</td>
<td>1 (12.5)</td>
</tr>
<tr>
<td>A15</td>
<td>EE——E——EQQQ——EQQQKQKK</td>
<td>1 (12.5)</td>
<td>1 (12.5)</td>
</tr>
<tr>
<td>A16</td>
<td>EE——E——EQQQ——EQQQKQKK</td>
<td>1 (12.5)</td>
<td>1 (12.5)</td>
</tr>
<tr>
<td>A17</td>
<td>EE——E——EQQQ——EQQQKQKK</td>
<td>2 (6.7)</td>
<td>2 (6.7)</td>
</tr>
<tr>
<td>A18</td>
<td>EE——E——EQQQ——EQQQKQKK</td>
<td>2 (6.7)</td>
<td>2 (6.7)</td>
</tr>
<tr>
<td>A19</td>
<td>EE——E——EQQQ——EQQQKQKK</td>
<td>2 (6.7)</td>
<td>2 (6.7)</td>
</tr>
<tr>
<td>A20</td>
<td>EE——E——EQQQ——EQQQKQKK</td>
<td>1 (3.3)</td>
<td>1 (3.3)</td>
</tr>
<tr>
<td>A21</td>
<td>EE——E——EQQQ——EQQQKQKK</td>
<td>1 (3.3)</td>
<td>1 (3.3)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>81 (100)</td>
<td>30† (100)</td>
</tr>
</tbody>
</table>

*   Dots and dashes represent identical residues and deletions, respectively.
†   Asia except South Korea and Thailand.
‡ Including four clones found in mixed infection from Thailand isolates.

*Allelic types of the E/Q-rich region of *Plasmodium* merozoite surface protein 1 paralog and geographic prevalence of each *P. vivax* isolate from various locations.*
REFERENCES


