Plasmodium knowlesi clinical isolates from Malaysia show extensive diversity and strong differential selection pressure at the merozoite surface protein 7D (MSP7D)

Md. Atique Ahmed¹ and Fu‑Shi Quan¹,²*

**Abstract**

**Background:** The high proportion of human cases due to the simian malaria parasite Plasmodium knowlesi in Malaysia is a cause of concern, as they can be severe and even fatal. Merozoite surface protein 7 (MSP7) is a multigene family which forms a non-covalent complex with MSP-1 prior to receptor-ligand recognition in Plasmodium falciparum and thus an important antigen for vaccine development. However, no study has been done in any of the ortholog family members in P. knowlesi from clinical samples. This study investigates the level of polymorphism, haplotypes, and natural selection acting at the pkmsp7D gene in clinical samples from Malaysia.

**Methods:** Thirty-six full-length pkmsp7D gene sequences (along with the reference H-strain: PKNH_1266000) obtained from clinical isolates of Malaysia, which were orthologous to pvmsp7H (PVX_082680) were downloaded from public databases. Population genetic, evolutionary and phylogenetic analyses were performed to determine the level of genetic diversity, polymorphism, recombination and natural selection.

**Results:** Analysis of 36 full-length pkmsp7D sequences identified 147 SNPs (91 non-synonymous and 56 synonymous substitutions). Nucleotide diversity across the full-length gene was higher than its ortholog in Plasmodium vivax (msp7H). Region-wise analysis of the gene indicated that the nucleotide diversity at the central region was very high ($\pi = 0.14$) compared to the 5' and 3' regions. Most hyper-variable SNPs were detected at the central domain. Multiple test for natural selection indicated the central region was under strong positive natural selection however, the 5' and 3' regions were under negative/purifying selection. Evidence of intragenic recombination were detected at the central region of the gene. Phylogenetic analysis using full-length msp7D genes indicated there was no geographical clustering of parasite population.

**Conclusions:** High genetic diversity with hyper-variable SNPs and strong evidence of positive natural selection at the central region of MSP7D indicated exposure of the region to host immune pressure. Negative selection at the 5' and the 3' regions of MSP7D might be because of functional constraints at the unexposed regions during the merozoite invasion process of P. knowlesi. No evidence of geographical clustering among the clinical isolates from Malaysia indicated uniform selection pressure in all populations. These findings highlight the further evaluation of the regions and functional characterization of the protein as a potential blood stage vaccine candidate for P. knowlesi.

**Keywords:** Plasmodium knowlesi, Merozoite surface protein 7D, Genetic diversity, Positive selection, Haplotypes, Malaysia

*Correspondence: fquan01@gmail.com
¹ Department of Medical Zoology, School of Medicine, Kyung Hee University, Seoul 02447, Republic of Korea
Full list of author information is available at the end of the article

© The Author(s) 2019. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Background

Malaria is a major global health problem with high mortality and morbidity rates and thus a significant barrier to the socio-economic development of under-developed countries [1]. *Plasmodium knowlesi* is a simian malaria parasite which can infect humans and is an emerging infection in Southeast Asian countries [2–6]. Countries in Central Asia and almost all Southeast Asia countries has reported human infections due to *P. knowlesi*, including Malaysia [4, 7, 8], Singapore [9], Myanmar [10], Vietnam [11], Indonesia [12, 13], Philippines [14], Cambodia [15], India [16] and Thailand [17]. Malaysia being the epicenter of knowlesi malaria, infection accounts up to 90% human infections [4, 8, 18, 19]. Despite the overall reduction in malaria cases around the world, the increasing trend of *P. knowlesi* cases in Southeast Asia highlights the need for proper and effective elimination measures as well as development of effective vaccines.

*Plasmodium knowlesi* has a 24-h erythrocytic cycle thus rapid increase in parasitaemia has been associated with the development of severe malaria and also a common cause for severe and fatal malaria in Malaysian Borneo [3, 20–22]. Studies on mitochondrial and ssrRNA genes in *P. knowlesi* from patients and wild macaques identified two distinct sub-populations which clustered geographically to Peninsular Malaysia and Malaysian Borneo [23]. Additionally, genetic and genomic studies from Malaysia identified 3 distinct sub-populations; two originating from Sarawak and one from Peninsular Malaysia [24–27], highlighting the complexity of infections in humans and the challenges for control and vaccine design.

Development of an effective malaria blood-stage vaccine antigens has focused on selecting antigens that provide long term and strain transcending protection and blocks the invasion of red blood cells (RBCs). Studies have shown that parasite antigens which are recognized by the host’s immune system, accumulate polymorphism to evade host’s defense mechanism and are prime targets for vaccine development. These are achieved through the knowledge of antigen natural selection and the pattern of diversity across geographical areas. Blood-stage antigens in *Plasmodium falciparum* [e.g. merozoite surface proteins (MSPs), apical membrane antigen 1 (AMA1)] that are under strong positive balancing selection are important targets of such acquired immunity and are supported by antibody inhibition assays in culture as well as naturally acquired antibodies in endemic regions [28, 29]. Several known blood stage antigens (e.g. MSP1, MSP1P and AMA1) have been recently studied in *P. knowlesi*, but most of the antigens were found to be under the influence of strong negative/purifying selection indicating no role in immune evasion mechanism [30–33]. PfMSP1 has been a major target for anti-malaria vaccine development but recent studies have shown that MSP 6 and 7 forms a non-covalent complex with MSP1 during the invasion process [34]. Disruption of PfMSP7 gene has resulted in partial inhibition of erythrocyte invasion during merozoite stages thus indicating that it is an important molecule [35]. Furthermore, MSP7 binding to P-selectin receptor has suggested its role in modulating disease severity in mice models [36] thereby indicating that immunity induced by MSP7 could potentially disrupt parasite development.

MSP7 is a multi-gene family and the number of paralog members in each species differs [37]. For example, till date, 13 MSP7 members were identified in *Plasmodium vivax* which are arranged in head-to-tail arrangement in chromosome 12 and are named alphabetically as MSP7A to MSP7M [37]. Three of its members i.e. PvMSP7C, PvMSP7H and PvMSP7I have been found to be highly polymorphic at the central region and under strong positive natural selection in Colombian population indicating it is an important molecule for consideration as a vaccine candidate [38]. *Plasmodium falciparum* has 8 MSP7 genes and low genetic polymorphism has been observed within the field isolates [39]. Certain MSP7 members forms part of the protein complex interacting with host cells [40] and they are localized at the merozoite surface [41, 42]. Despite the fact that *P. knowlesi* is phylogenetically closely related to *P. vivax* and there are five paralog members in *pkmsp7* gene family [37], no genetic study has been done to characterize the *P. knowlesi* MSP7 ortholog members from clinical samples. Thus in this study *pkmsp-7D* gene which is an ortholog to *pvmsp-7H* was chosen for genetic analysis.

In this study, 36 *pkmsp-7D* full-length sequences (32 clinical isolates and 4 laboratory lines of Malaysia) were obtained from published databases based on the ortholog gene sequence in *P. vivax msp-7H* (PVX_082680) [37, 43]. The level of sequence diversity, natural selection using full-length genes as well at each of the three regions of the gene (5′, 3′ and the central region) were determined (along with the H-strain) of Malaysia. The information obtained from this study will be helpful for designing functional studies and future rational design of a vaccine against *P. knowlesi*.

Methods

**pkmsp-7D** sequence data

Based on the identification of orthologs in *Plasmodium* species [37], the *pkmsp-7D* sequences were downloaded for 36 isolates (32 clinical and 4 long term isolated lines) originating from Kapit, Betong and Sarakei in Malaysian Borneo and Peninsular Malaysia along with the H-strain (PKNH_1266000, old ID; PK13_3510c).
characterized based on the published ortholog of PvMSP7D prediction software [44, 45]. The PkMSP7D regions were predicted using Signal IP 3.0 and the transmembrane regions using the Phobious prediction software. The full-length gene had the highest D values (Fig. 3b). Interestingly, the amino acid positions 22 and 23 of the PkMSP7D protein were excluded from analysis by DnaSP software. A statistical significance test was conducted using Fisher’s exact test and $\chi^2$ test departs from randomness and the value of $r^2$ ranges from 0 to 1 in the software.

**Analysis of recombination and linkage disequilibrium**

Analysis of the minimum number of recombination events (Rm) within the pkmsp7D genes was performed using DnaSP software v5.10. Linkage disequilibrium (LD) is the non-random association of sequences at two or more loci. Linkage disequilibrium index, $r^2$ (square of the correlation coefficient of allelic states at each pair of loci) against nucleotide distance were plotted for pkmsp7D sequences across the gene using DnaSP 5.10 software. A statistical significance test was conducted using Fisher’s exact test and $\chi^2$ test departs from randomness and the value of $r^2$ ranges from 0 to 1 in the software.

**Phylogenetic analysis**

Phylogenetic analysis was conducted using deduced amino acid sequences from 32 PkMSP7 full-length sequences from Malaysian Borneo, 5 laboratory lines from Peninsular Malaysia; reference H-strain (PKNH_1266000), the Malayan Strain (PKNH_S09532900), MR4H strain, Philippine strain and the Hackeri strain along with other P. knowlesi MSP7 paralog members PkMSP7A (PKNH_1266300), PkMSP7E (PKNH_1265900) and PkMSP7C (PKNH_1266100) were included. Other ortholog members of P. vivax MSP7H (PVX_082680), MSP7E (PVX_082665) and Plasmodium coatneyi (PCOAH_00042440) were also included for comparative analysis.

**Results**

**Polymorphism within full-length pkmsp7D genes and natural selection**

The Signal IP server detected a signal peptide in between amino acid positions 22 and 23 of the PkMSP7D protein (Additional file 2) however, no transmembrane region was detected (data not shown). Alignment and comparison of the amino acid sequences of the full-length P. knowlesi H reference strain MSP7D sequence with P. vivax MSP7H Sal-1 (PVX_082680) reference strain showed 60.4% identity and the next nearest PvMSP7 ortholog gene was E which had 47.3% identity. The amino acid identity within the PkMSP7 paralog members (A, C, D and E) were in the range of 17–26%. The schematic representation of PkMSP7D protein with demarcated inter-species level, the robust McDonald and Kreitman (MK) test were also performed with the closest msp7 orthologs in both P. vivax (PVX_082680) and Plasmodium cynomolgi (PCYB_122830) as outgroups individually using DnaSP v5.10 software.

**Polymorphism within full-length pkmsp7D genes and natural selection**

The Signal IP server detected a signal peptide in between amino acid positions 22 and 23 of the PkMSP7D protein (Additional file 2) however, no transmembrane region was detected (data not shown). Alignment and comparison of the amino acid sequences of the full-length P. knowlesi H reference strain MSP7D sequence with P. vivax MSP7H Sal-1 (PVX_082680) reference strain showed 60.4% identity and the next nearest PvMSP7 ortholog gene was E which had 47.3% identity. The amino acid identity within the PkMSP7 paralog members (A, C, D and E) were in the range of 17–26%. The schematic representation of PkMSP7D protein with demarcated inter-species level, the robust McDonald and Kreitman (MK) test were also performed with the closest msp7 orthologs in both P. vivax (PVX_082680) and Plasmodium cynomolgi (PCYB_122830) as outgroups individually using DnaSP v5.10 software.

**Analysis of the minimum number of recombination events (Rm) within the pkmsp7D genes was performed using DnaSP software v5.10. Linkage disequilibrium (LD) is the non-random association of sequences at two or more loci. Linkage disequilibrium index, $r^2$ (square of the correlation coefficient of allelic states at each pair of loci) against nucleotide distance were plotted for pkmsp7D sequences across the gene using DnaSP 5.10 software. A statistical significance test was conducted using Fisher’s exact test and $\chi^2$ test departs from randomness and the value of $r^2$ ranges from 0 to 1 in the software.

**Phylogenetic analysis**

Phylogenetic analysis was conducted using deduced amino acid sequences from 32 PkMSP7 full-length sequences from Malaysian Borneo, 5 laboratory lines from Peninsular Malaysia; reference H-strain (PKNH_1266000), the Malayan Strain (PKNH_S09532900), MR4H strain, Philippine strain and the Hackeri strain along with other P. knowlesi MSP7 paralog members PkMSP7A (PKNH_1266300), PkMSP7E (PKNH_1265900) and PkMSP7C (PKNH_1266100) were included. Other ortholog members of P. vivax MSP7H (PVX_082680), MSP7E (PVX_082665) and Plasmodium coatneyi (PCOAH_00042440) were also included for comparative analysis. The Maximum likelihood (ML) method based on Poisson correction model was used as described in MEGA 5.0 with 1000 bootstrap replicates to test the robustness of the trees.

**Common linkage disequilibrium index, $r^2$ (square of the correlation coefficient of allelic states at each pair of loci) against nucleotide distance were plotted for pkmsp7D sequences across the gene using DnaSP 5.10 software. A statistical significance test was conducted using Fisher’s exact test and $\chi^2$ test departs from randomness and the value of $r^2$ ranges from 0 to 1 in the software.

**Phylogenetic analysis**

Phylogenetic analysis was conducted using deduced amino acid sequences from 32 PkMSP7 full-length sequences from Malaysian Borneo, 5 laboratory lines from Peninsular Malaysia; reference H-strain (PKNH_1266000), the Malayan Strain (PKNH_S09532900), MR4H strain, Philippine strain and the Hackeri strain along with other P. knowlesi MSP7 paralog members PkMSP7A (PKNH_1266300), PkMSP7E (PKNH_1265900) and PkMSP7C (PKNH_1266100) were included. Other ortholog members of P. vivax MSP7H (PVX_082680), MSP7E (PVX_082665) and Plasmodium coatneyi (PCOAH_00042440) were also included for comparative analysis. The Maximum likelihood (ML) method based on Poisson correction model was used as described in MEGA 5.0 with 1000 bootstrap replicates to test the robustness of the trees.
the 3’ region also had some SNPs, which had positive D values (Fig. 3b). The natural selection analysis of the full-length \textit{pkmsp7D} genes indicated $d_N - d_S = -2.1$, however, Tajimas’ D and Fu and Li’s D* and F* values were positive (Table 1). The robust inter-species McDonald-Kreitman test indicated a positive selection when \textit{P. vivax} (NI = 1.25, $P > 0.1$) and \textit{P. cynomolgi} (NI = 1.38, $P > 0.1$) were used as out-groups but were not significant (Table 3). Nucleotide polymorphisms across the full-length gene with hyper-variable SNPs within the sequences are shown in Additional file 3.

**Region-wise analysis of diversity and natural selection**

Based on the sequence alignment of \textit{pkmsp7D} and the diversity pattern across the full-length gene (Fig. 3a), three regions were defined; 5’ region (from amino acid 1 to 144), central region (from amino acid 145 to 257) and the 3’ region (from amino acid 258 to 397). Nucleotide diversity at the 5’ and the 3’ regions were of similar levels ($\pi = 0.010$–0.017) (Table 1). The number of haplotypes in both the regions were 30 with similar levels of haplotype diversities ($Hd = 0.992$–$0.990$) (Table 1). Within the 5’ region, there were 28 SNPs (18 synonymous and 10 non-synonymous substitutions) and the 3’ region had 30 SNPs (22 synonymous and 8 non-synonymous substitutions) (Table 2).

The central region constituted the highest number of SNPs (151) leading to 32 haplotypes ($Hd = 0.995$) and the nucleotide diversity was very high ($\pi = 0.149$) compared to the other regions (Table 1). Of the 151 SNPs, 89 SNPs (16 synonymous and 73 non-synonymous substitutions) could be analysed using DnaSP software. Size variations within the \textit{msp7D} genes were observed due insertion/deletion at two specific locations within the central domain; 6 nucleotides deletion (from 435 to 440 nt) and 6 to 12 nucleotides insertion/deletion (from 590 to 601 nt) in some isolates (Additional file 3). Twenty-three tri to tetra morphic nucleotides (hyper-variable SNPs) were observed within the central region (Additional file 3).

To determine whether natural selection contributes to the polymorphism in the 3’, 5’ and the central regions of \textit{pkmsp7D} gene, the difference of $d_N - d_S$ and codon-based Z test was evaluated independently for each region. The negative values at both the 5’ and 3’ region obtained indicated negative/purifying selection (Table 1). The 3’ region was under strong negative selection pressure ($dN - dS = 2.68$, $P < 0.01$). Tajimas’ D test, Fu and Li’s D* and F* test were all negative implying negative/purifying selection due to functional constraints at both these regions (Table 1). Negative selection was also evident as there were an excess of synonymous substitutions in both these regions (Table 2). Additionally, the robust MK test, which compares natural selection in the inter-species level, also showed significant values for both \textit{P. vivax} and \textit{P. cynomolgi} as outgroup sequences for the 3’ region (Table 3). Though MK did not yield statistically significant results for the 5’ region, the results were indicative of a negative selection.

The test results for natural selection for the highly diverse central region of \textit{pkmsp7D} indicated that the region was under the influence of strong positive/balancing selection. The difference of $d_N - d_S = 1.84$ ($P < 0.03$) and Tajimas’ D, Fu and Li’s D* and F* values were also positive (Table 1). This was also evident due the presence of excess of non-synonymous substitutions within the region (nsSNPs = 73, sSNPs = 16) (Table 2). MK test,
Table 1 Estimates of nucleotide diversity, natural selection, haplotype diversity and neutrality indices of *pkmsp7*

| Region       | No. samples | SNPs | No. haplotype | Diversity ± SD         | \(d_N - d_S\) | Codon-based z test | Taj D | Fu and Li's \(D^*\) | Fu and Li's \(F^*\) |
|--------------|-------------|------|---------------|------------------------|----------------|---------------------|-------|---------------------|---------------------|
| Full-length  | 36          | 203  | 34            | 0.998 ± 0.007          | 0.052 ± 0.002 | -2.0                | \(P > 0.1\) | 0.295               | 0.190               | 0.270               |
| 5' region    | 28          | 30   | 0.992 ± 0.009 | 0.010 ± 0.001         | -3.0          | \(P > 0.1\)        | -1.25 | -2.24              | -2.26              |
| Central region | 151        | 32   | 0.995 ± 0.008 | 0.1495 ± 0.006        | 1.84          | \(P < 0.05\)       | 0.673 | 0.896              | 0.973              |
| 3' region    | 32          | 30   | 0.990 ± 0.010 | 0.017 ± 0.0007        | -2.68         | \(P < 0.01\)       | -0.275 | -1.027            | -0.917             |

*SNPs* single nucleotide polymorphisms, *SD* standard deviation
which also showed significant values when both P. vivax (NI = 2.49, P < 0.05) and P. cynomolgi (NI = 2.28, P < 0.05) were used as outgroup sequences for the central region (Table 3).

Recombination and Linkage disequilibrium (LD) analysis

Analysis of recombination within the aligned pkmsp7D sequences estimated a minimum number of 29 recombination events (Rm) indicating intragenic recombination could be a factor adding to the extensive diversity at the central domain of the gene. LD analysis across the full-length PkMSP7D genes showed the relationship between nucleotide distance and $R^2$ index and the decline in regression trace line indicated that intragenic recombination might exist within the msp7D genes in P. knowlesi (Fig. 4).

Phylogenetic analysis

Phylogenetic analysis of the 37 full-length PkMSP7D amino acid sequences indicated that there was no geographical clustering of knowlesi sequences (Fig. 4). The PkMSP7D sequences were found to be more closely related to P. vivax MSP7H compared to its paralogs in P. knowlesi and ortholog in P. vivax and P. coatneyi. The H-strain and the Malayan strain clustered together along with sequences from Malaysian Borneo. The Philippine, MR4 and the Hackeri Strains formed a separate group along with Malaysian Borneo isolates. The distinct sub-populations observed in other invasion genes of P. knowlesi like MSP1P, DBP1R, NBPXA and TRAP was not observed in MSP7D [23, 25–27, 31, 49] (Fig. 5).

Discussion

The merozoite surface protein (MSP7) gene family is a group of surface proteins which are involved in the initial interaction with MSP1 through a complex formation at the merozoite surface during the invasion process and orthologs and paralog members have been identified across Plasmodium species [37, 43]. In P. falciparum, MSP7 is expressed during schizont stages and undergoes two step proteolytic processing similar to MSP1 and disruption of the MSP7 gene leads to invasion impairment.
indicating the molecule is essential for the parasite [35]. Considering previous studies on MSP7 members in *P. vivax* (i.e. PvMSP7C, MSP7E, PvMSP7H and PvMSP7I) the central region were found to be under the influence of strong positive natural selection and extensive diversity indicating host immune pressure [38, 50]. Extensive functional and field-based studies have been conducted on *P. falciparum* and *P. vivax* MSP7 gene family members, but no studies have been done on *P. knowlesi*. Thus, this study was conducted to genetically characterize the MSP7D genes from clinical isolates of Malaysia, which is the closest ortholog of *P. vivax* MSP7H.

Sequence alignment of 36 full-length amino acid sequences of *pkmsp7D* genes from Malaysia showed that it shares approximately 60.4% sequence identity with its ortholog *Pvmsp7H* and the overall diversity across the full-length gene was higher than *P. vivax* [38]. Interestingly, the sequence identity within the PkMSP7 paralog member were low. Analysis of polymorphism across the full-length *pkmsp7D* gene indicated that only the central region acquired high number of polymorphism (SNPs = 151) compared to the 5′ and 3′ regions. This finding was similar to the diversity in the *msp7* members (C, E, H and I) in *P. vivax* [38, 50]. The genetic diversity at the central region of the gene was 14-fold higher (π = 0.1495) than the 5′ and 3′ region. This was due to mainly two reasons; firstly, due to size variations in some of the clinical isolates in two specific locations within the central region and, second, due the high number of non-synonymous hyper-variable SNP sites (tri-morphic and tetra-morphic). The number of singletons (low-frequency polymorphisms) in each of the regions were high leading a to high number of haplotypes and haplotype diversity (Table 1). Similar hyper-variable sites were also observed in *P. vivax* *msp7* sequences from Thailand [50]. Tests for natural selection (using Taj D, Li and Fu’s D*, F*, dN-dS and MK test) at each region indicated that only the central region was under strong positive selection indicating immune evasion mechanism of parasite at this region, however, the 5′ and the 3′ prime regions are under functional constraints (purifying selection). These

---

**Table 3 McDonald–Kreitman tests on msp7D of Plasmodium knowlesi and its regions with *P. vivax* and *P. cynomolgi* orthologs as outgroup species**

| MSP7          | Polymorphic changes within *P. knowlesi* | Fixed differences between species | Neutrality index |
|---------------|------------------------------------------|-----------------------------------|------------------|
|               | Syn | NonSyn | Pk vs Pv | Pk vs Pcy | Pk vs Pcy | Pk vs Pv | Pk vs Pcy |
|               | Syn | NonSyn |         |          |          |         |           |
| Full-length   | 56  | 91     | 145     | 188      | 112      | 132      | 1.25      | 1.38      |
| 5′ region     | 18  | 10     | 47      | 51       | 52       | 51       | 0.51      | 0.56      |
| Central region| 16  | 73     | 31      | 57       | 15       | 30       | 2.49*     | 2.28*     |
| 3′ region     | 22  | 8      | 67      | 78       | 54       | 61       | 0.31*     | 0.32*     |

Syn synonymous sites, NonSyn non synonymous sites, Pk Plasmodium knowlesi, Pv Plasmodium vivax, Pcy Plasmodium cynomolgi

* Fisher’s exact test P-value < 0.05; ¥ Fisher’s exact test P-value < 0.01
regions could potentially be the putative B cell epitope binding regions. This was probably due to the structural conformity attained by the protein during the invasion process and exposure of the central region to host immune system. Indeed, it remains to be proven through functional studies whether PkMSP7D forms a complex with PkMSP1 as observed in PfMSP1-MSP7 complex [34, 42]. A recent genetic and structural study on P. vivax msp7E sequences from Thailand predicted 4–6 α-helical domains at the 5′ and 3′ regions suggesting structural and functional constraints in these domains [50]. Sliding window analysis of Tajima’s D across the full-length Pkmsp7D identified high D values within the central region and a short region within the 3′ region. These regions with elevated D values were probable epitope binding regions.

Previous studies identified two sympatric sub-populations from Sarawak and geographical separation in samples from Peninsular Malaysia [24], however, phylogenetic analysis of the current study did not show any specific separation of the P. knowlesi MSP7D isolates. This is an interesting finding because the pkmsp7D sequences used in this study is from the same genomic data [24]. Bifurcation of trees, indicating dimorphism and strong negative/purifying selection on invasion genes like DBPαII (PkDBPαII) [31], PkNBPXa [26], PkAMA1 [30] and Pk41 [51] have been reported from clinical samples. This probably indicates that Pkmsp7D is not influenced by geographical origin of the parasite and both the parasite sub-populations were under strong host immune pressure at the central region similar to the non-repeat region in PkcsP from Malaysia [52]. High recombination values and LD analysis indicated that intragenic recombination could play a crucial role in increasing diversity in the central region. These findings will be helpful for future designing functional binding assays to validate the molecule as a vaccine candidate.

(See figure on next page.)

**Fig. 4** Linkage disequilibrium across the full-length pkmsp7D genes from Malaysia. Relationship between nucleotide distance and $R^2$ index and the decline in regression trace line indicated intragenic recombination within P. knowlesi isolates.

**Fig. 5** Phylogenetic relationship of deduced amino acid sequences of PkMSP7D from Sarawak, Malaysia and its paralog members PkMSP7A (PKNH_1266300), PkMSP7E (PKNH_1265900) and PkMSP7C (PKNH_1266100) and ortholog members of Plasmodium vivax MSP7H (PVX_082680) and Plasmodium coatneyi (PCOAH_00042440). The tree was constructed using Maximum Likelihood (ML) method based on Poisson correction model was used as described in MEGA 5.0 with 1000 bootstrap replicates to test the robustness of the trees. MB and PM indicates samples from Malaysian Borneo and Peninsular Malaysia respectively.
Conclusion
This study is the first to investigate genetic diversity and natural selection of pkmsp7D gene from clinical samples from Malaysia. High level of genetic diversity and strong positive selection pressure was observed in the central region of the pkmsp7D gene indicating exposure to host immune pressure. Absence of both dimorphism and geographical clustering within the parasite populations indicated uniform exposure of the central region to host immunity. Future studies should focus on the functional characterization of pkmsp7D as a vaccine candidate.

Additional files

Additional file 1: Table S1. Study samples and origin.
Additional file 2. Signal peptide identified by Signal IP server.
Additional file 3. Neucleotide polymorphism across the full-length pkmsp7D genes. Hyper-variable SNPs are marked in bold and white and SNPs with deletions marked in red.

Abbreviations
MSP1: merozoite surface protein 1 paralog; kDa: kilodalton.

Authors' contributions
MAA and FQS designed the study. MAA retrieved Pkmsp7D gene sequences from genomic databases, performed the genetic analysis. MAA and FQS wrote the manuscript. All authors read and approved the final manuscript.

Author details
1 Department of Medical Zoology, School of Medicine, Kyung Hee University, Seoul, Republic of Korea. 2 Medical Research Center for Bioreaction to Reactive Oxygen Species and Biomedical Science Institute, School of Medicine, Graduate School, Kyung Hee University, Seoul, Republic of Korea.

Acknowledgements
The authors would like to thank Dr. Syeda Wasfiea Wazid for helping in data retrieval and organizing the data for analysis.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
The datasets analysed during the current study were derived from the following public region resources: https://doi.org/10.1371/journal.pone.0121303 and https://doi.org/10.1073/pnas.1509534112.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
This work was supported by the grants from the National Research Foundation of Korea (NRF) (2018R1A2B600353S and 2018R1A6A1A3025124).

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 27 September 2018 Accepted: 16 April 2019 Published online: 29 April 2019

References
1. WHO. World malaria report. Geneva: World Health Organization; 2016.
2. White NJ. Plasmodium knowlesi: the fifth human malaria parasite. Clin Infect Dis. 2008;46:172–3.
3. Cox-Singh J, Davis TM, Lee KS, Shamsul SS, Matusop A, Ratnam S, et al. Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening. Clin Infect Dis. 2008;46:165–71.
4. Singh B, Kim Sung L, Matusop A, Rakhadakrishnan A, Shamsul SS, et al. A large focus of naturally acquired Plasmodium knowlesi infections in human beings. Lancet. 2004;363:1017–24.
5. Garnham PCC. Malaria parasites and other haemosporidia. Oxford: Blackwell Scientific Publications; 1966.
6. Ahmed MA, Cox-Singh J. Plasmodium knowlesi - an emerging pathogen. ISBT Sci Ser. 2015;10(Suppl 1):134–40.
7. Vythilingam I, Noorazian YM, Huat TC, Jiram AI, Yusri YM, Azahari AH, et al. Plasmodium knowlesi in humans, macaques and mosquitoes in peninsular Malaysia. Parasit Vectors. 2008;1:26.
8. Barber BE, William T, Jikal M, Jilip J, Dhararaj P, Menon J, et al. Plasmodium knowlesi malaria in children. Emerg Infect Dis. 2011;17:814–20.
9. Ng OT, Ooi EE, Lee CC, Lee PJ, Ng LC, Pei SW, et al. Naturally acquired human Plasmodium knowlesi infection, Singapore. Emerg Infect Dis. 2008;14:814–6.
10. Jiang N, Chang Q, Sun X, Lu H, Yin J, Zhang Z, et al. Co-infections with Plasmodium knowlesi and other malaria parasites, Myanmar. Emerg Infect Dis. 2010;16:1476–8.
11. Van den Eede P, Van HN, Van Overmeir C, Vythilingam I, Duc TN, Le Hung X, et al. Human Plasmodium knowlesi infections in young children in central Vietnam. Malar J. 2009;8:249.
12. Figtree M, Lee R, Bain L, Kennedy T, Mackertich S, Urban M, et al. Plasmodium knowlesi in human, Indonesian Borneo. Emerg Infect Dis. 2010;16:672–4.
13. Herdiana H, Irnawati I, Coutrier FN, Munthe A, Mardiasi M, Yuniarti T, et al. Two clusters of Plasmodium knowlesi cases in a malaria elimination area, Sabang Municipality, Aceh, Indonesia. Malar J. 2018;17:186.
14. Luchavez J, Espino F, Curameng P, Espina R, Bell D, Chiодini P, et al. Human Infections with Plasmodium knowlesi, the Philippines. Emerg Infect Dis. 2008;4:811–3.
15. Khim N, Siv S, Kim S, Mueller T, Fleischmann E, Singh B, et al. Plasmodium knowlesi infection in humans, Cambodia, 2007–2010. Emerg Infect Dis. 2011;17:1900–2.
16. Tyagi RK, Das MK, Singh SS, Sharma YD. Discordance in drug resistance-associated mutation patterns in marker genes of Plasmodium falciparum and Plasmodium knowlesi during coinfections. J Antimicrob Chemother. 2013;68:1081–8.
17. Sermsri Watayongwong N, Singh B, Nishibuchi M, Sawangjaroen N, Vuddhakul V. Human Plasmodium knowlesi infection in Ranong province, southwestern border of Thailand. Malar J. 2012;11:36.
18. Yusuf R, Lau YL, Mahmud R, Fong MY, Jelip J, Ngian HU, et al. High proportion of knowlesi malaria in recent malaria cases in Malaysia. Malar J. 2014;13:168.
19. Rajahram GS, Barber BE, William T, Grigg MJ, Menon J, Yeo TW, et al. Falling Plasmodium knowlesi malaria death rate among adults despite rising incidence, Sabah, Malaysia, 2010–2014. Emerg Infect Dis. 2016;22:41–8.
20. Daneshvar C, Davis TM, Cox-Singh J, Rafiee MZ, Zakaria SK, Divis PC, et al. Clinical and laboratory features of human Plasmodium knowlesi infection. Clin Infect Dis. 2009;49:852–60.
21. William T, Menon J, Rajahram G, Chan L, Ma G, Donaldson S, et al. Severe Plasmodium knowlesi malaria in a tertiary care hospital, Sabah, Malaysia. Emerg Infect Dis. 2011;17:1248–55.
22. Willmann M, Ahmed A, Siner A, Wong IT, Woon LC, Singh B, et al. Laboratory markers of disease severity in Plasmodium knowlesi infection: a case control study. Malar J. 2012;11:363.
23. Yusuf R, Ahmed MA, Jelip J, Ngian HU, Mustakim S, Hussin HA, et al. Phylogographic evidence for 2 genetically distinct zoonotic Plasmodium knowlesi parasites, Malaysia. Emerg Infect Dis. 2016;22:1371–80.
Ahmed and Quan Malar J (2019) 18:150

24. Assefa S, Lim C, Preston MD, Duffy CW, Nair MB, Adroub SA, et al. Population genetic structure and adaptation in the zoonotic malaria parasite Plasmodium knowlesi. Proc Natl Acad Sci USA. 2015;112:13027–32.

25. Pinheiro MM, Ahmed MA, Millar SB, Sanderson T, Otto TD, Lu WC, et al. Plasmodium knowlesi genome sequences from clinical isolates reveal extensive genomic dimorphism. PLoS ONE. 2015;10:e0121303.

26. Ahmed AM, Fong MY, Lau YL, Yusuf R. Clustering and genetic differentiations of the normocyte binding protein (nbpxa) of Plasmodium knowlesi clinical isolates from Peninsular Malaysia and Malaysia Borneo. Malar J. 2016;15:241.

27. Ahmed MA, Pinheiro MM, Divis PC, Siner A, Zainudin R, Wong JT, et al. Disease progression in Plasmodium knowlesi is linked to variation in gene family members. PLoS Negl Trop Dis. 2014;8:e3086.

28. Kennedy MC, Wang J, Zhang Y, Miles AP, Chitsaz F, Saul A, et al. In vitro studies with recombinant Plasmodium falciparum apical membrane antigen 1 (AMA1): production and activity of an AMA1 vaccine and generation of a multiallelic response. Infect Immun. 2002;70:948–60.

29. Polley SD, Tetteh KK, Lloyd JM, Akpogheneta OJ, Greenwood BM, Bojang KA, et al. Plasmodium falciparum merozoite surface protein 3 is a target of allele-specific immunity and alleles are maintained by natural selection. J Infect Dis. 2007;195:279–87.

30. Faber BW, Abdul Kadir K, Rodriguez-Garcia R, Remarque EJ, Saul FA, Vulle-iz-Le Normand B, et al. Low levels of polymorphisms and no evidence for diversifying selection on the Plasmodium knowlesi apical membrane antigen 1 gene. PLoS ONE. 2015;10:e0124400.

31. Fong MY, Lau YL, Chang PY, Anthony CN. Genetic diversity, haplotypes and allele groups of Duffy binding protein (PkdBPalphal) of Plasmodium knowlesi clinical isolates from Peninsular Malaysia. Parasit Vectors. 2014;7:161.

32. Yap NJ, Goh XT, Koehler AV, William T, Yeo TW, Vythilingam I, et al. Genetic diversity in the C-terminus of merozoite surface protein 1 among Plasmodium knowlesi isolates from Selangor and Sabah Borneo, Malaysia. Infect Genet Evol. 2017;54:39–46.

33. Ahmed MA, Fauzi M, Han ET. Genetic diversity and natural selection of Plasmodium knowlesi merozoite surface protein 1 paralog gene in Malaysia. Malar J. 2018;17:115.

34. Kauth CW, Woehlbier U, Kern M, Mekonnen Z, Lutz R, Mucke N, et al. Deletion of the merozoite surface protein 7 gene impairs invasion of the malaria parasite Plasmodium falciparum. J Biol Chem. 2006;281:31517–27.

35. Kadekoppala M, O’Donnell RA, Grainger M, Trucco C, Howell S, Fernandez-Reyes D, et al. The 22 kDa component of the protein complex on the surface of Plasmodium falciparum merozoites is derived from a larger precursor, merozoite surface protein 7. Mol Biochem Parasitol. 2001;117:83–9.

36. Garzon-Ospina D, Cadavid LF, Patarroyo MA. Differential expansion of Plasmodium knowlesi. Proc Natl Acad Sci USA. 2015;112:13027–32.

37. Roy SW, Weedall GD, da Silva RL, Polley SD, Ferreira MU. Sequence diversity and evolutionary dynamics of the dimorphic antigen merozoite surface protein-6 and other Msp genes of Plasmodium falciparum. Gene. 2009;443:12–21.

38. Pachebat JA, Jing IT, Grainger M, Trucco C, Howell S, Fernandez-Reyes D, et al. The 22 kDa component of the protein complex on the surface of Plasmodium falciparum merozoites is derived from a larger precursor, merozoite surface protein 7. Mol Biochem Parasitol. 2001;117:83–9.

39. Mello K, Daly TM, Long CA, Burns JM, Bergman LW. Members of the merozoite surface protein 7 family with similar expression patterns differ in ability to protect against Plasmodium yoelii malaria. Infect Immun. 2004;72:1010–8.

40. Mello K, Daly TM, Morrisey J, Vaidya AB, Long CA, Bergman LW. A multigene family that interacts with the amino terminus of Plasmopdium MSP-1 identified using the yeast two-hybrid system. Eukaryot Cell. 2002;1:915–25.

41. Castillo AI, Andreina Pacheco M, Escalante AA. Evolution of the merozoite surface protein 7 (msp7) family in Plasmodium vivax and P. falciparum: a comparative approach. Infect Genet Evol. 2017;50:7–19.

42. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignAp 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785–6.

43. Kall L, Krogh A, Sonnhammer EL. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res. 2007;35:W429–32.

44. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25:1451–2.

45. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123:585–95.

46. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGAS: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28:2731–9.

47. Ahmed MA, Lau YL, Quan FS. Diversity and natural selection on the thrombospondin-related adhesive protein (TRAP) gene of Plasmodium knowlesi in Malaysia. Malar J. 2018;17:274.

48. Cheng CW, Putapatrump C, Jongvubwise S. Polymorphism in merozoite surface protein-7E of Plasmodium vivax in Thailand: natural selection related to protein secondary structure. PLoS ONE. 2018;13:e0196765.

49. Ahmed MA, Chu KB, Quan FS. The Plasmodium knowlesi PK41 surface protein diversity, natural selection, sub-population and geographical clustering: a 6-cysteine protein family member. PeerJ. 2016;6:e6141.

50. Fong MY, Ahmed MA, Wong SS, Lau YL, Sitam F. Genetic diversity and natural selection of the Plasmodium knowlesi circumsporozoite protein nonrepet region PLoS ONE. 2015;10:e0137734.