Tropical Biomedicine 40(1): 37-44 (2023)
https://doi.org/10.47665/tb.40.1.004

RESEARCH ARTICLE

Plasmodium knowlesi circumsporozoite protein: genetic characterisation and predicted antigenicity of the central repeat region

Tan, J.H.[1], Cheong, F.W.[1], Lau, Y.L.[1], Fong, M.Y.[1*]

[1]Department of Parasitology, Faculty of Medicine, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
*Corresponding author: fongmy@um.edu.my

ARTICLE HISTORY

Received: 3 November 2022
Revised: 17 December 2022
Accepted: 19 December 2022
Published: 31 March 2023

ABSTRACT

Circumsporozoite protein (CSP) central repeat region is one of the main target regions of the RTS,S/AS01 vaccine for falciparum infection as it consists of immunodominant B cell epitopes. However, there is a lack of study for *P. knowlesi* CSP central repeat region. This study aims to characterise the CSP repeat motifs of *P. knowlesi* isolates in Peninsular Malaysia. CSP repeat motifs of 64 *P. knowlesi* isolates were identified using Rapid Automatic Detection and Alignment of Repeats (RADAR). Antigenicity of the repeat motifs and linear B cell epitopes were predicted using VaxiJen 2.0, BepiPred-2.0 and BCPred, respectively. A total of 35 dominant repeat motifs were identified. The repeat motif "AGQPQAQGDGANAGQPQAQGDGAN" has the highest repeat frequency (n=15) and antigenicity index of 1.7986. All the repeat regions were predicted as B cell epitopes. *In silico* approaches revealed that all repeat motifs were antigenic and consisted of B cell epitopes which could be designed as knowlesi malaria vaccine.

Keywords: Circumsporozoite protein; central repeat; *P. knowlesi*; genetic characterisation; B-cell epitopes.

INTRODUCTION

Malaria remains a significant public health concern that burdens the lives of many. This vector-borne disease is caused by parasites of the genus *Plasmodium*. There are four human *Plasmodium* species that are transmitted around the world today, namely *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae,* and *Plasmodium ovale*. In 2004, Singh et al. discovered more than 200 microscopy-diagnosed *P. malariae* infections that were in fact caused by the simian malaria species *Plasmodium knowlesi* using molecular tools (Singh *et al*., 2004). In recent years, other zoonotic malaria species have emerged, such as *Plasmodium cynomolgi* and *Plasmodium inui*, thus emphasising the importance of active surveillance and control measures for these species (Yap *et al*., 2021). Among these zoonotic malaria species, *P. knowlesi* is the most concerning as it is the main cause of malaria in Malaysia, with 3213 and 2609 cases reported in 2019 and 2020, respectively (World Health Organization, 2020, 2021a). Complication of the disease can result in acute kidney failure, respiratory distress, and other severe clinical manifestations which require immediate management and treatment (Daneshvar *et al*., 2009).

Circumsporozoite protein (CSP) is a major surface protein expressed in the *Plasmodium* sporozoite stage (Rathore *et al*., 2002). It performs different roles in the pre-erythrocytic stages of the parasites' life cycle, such as maturation of sporozoites in the mosquito midgut as well as attachment and invasion of sporozoites into hepatocytes (Ferguson *et al*., 2014). All *Plasmodium* CSP possess similar overall organisational structure whereby a central repeat region (CRR) is flanked by N- and C-terminal regions. The N-terminal contains Region I (RI), which is a conserved five-amino acid sequence (KLKQP) found in all *Plasmodium* species except *P. gallinaceum* (McCutchan *et al*., 1996). A thrombospondin repeat-like domain, also known as Region II (RII), occurs in the C-terminal of *Plasmodium* CSP. This domain is a cell-adhesive module that is homologous to the type 1 repeat of human thrombospondin (TSP) and binds to liver heparan sulphate proteoglycans together with the RI motif (Ying *et al.*, 1997).

Unlike RI and RII, studies on the role of CRR in the parasite's life cycle are limited. Ferguson *et al*. (2014) observed that transgenic *P. berghei* expressing CSP that lacked CRR had abnormal sporozoite development and degeneration of oocysts (Ferguson *et al.*, 2014). The CRR contains species-specific tandem amino acid repeats which are B-cell immunodominant. Antibodies specific to the B-cell immunodominant epitopes of the CRR provide protective immunity against malaria by blocking sporozoite invasion of host hepatocytes (Persson *et al*., 2002). Human malaria parasite species have simple central repeat motifs, for example the major repeats for *P. falciparum* CSP are NANP and NVDP (Lê *et al*., 2018); *P. vivax* CSP are type VK210 GDRA(D/A)GQPA and type VK247 ANGA(G/D)QPG (Rosenberg *et al*., 1989; Võ *et al*., 2020); and *P. malariae* CSP are NAAG and NDAG (Saralamba *et al*., 2018).

CSP stands out among all the other potential malaria vaccine candidate due to its abundance in *Plasmodium* sporozoite stage and high immunogenicity of the CRR. Mice immunised with recombinant subunit antigen of *P. falciparum* CSP tetrapeptide repeats produced high concentration of antibodies against CSP. *In vitro* studies showed that these antibodies were able to inhibit the invasion of sporozoites into human hepatocytes, underlining the significance

of CSP (Young *et al*., 1985; Mazier *et al*., 1986). The leading malaria vaccine, RTS,S/AS01 (Mosquirix™) was designed based on NANP repeats and C-terminal (B-cell and T-cell epitopes) of *P. falciparum* CSP coupled to hepatitis B surface antigen and is administered with a novel adjuvant, AS01 (Casares *et al*., 2010). Phase 3 clinical trials reported protective effect of the vaccine against malaria in young children with tolerable side effects such as swelling at the injection site and fever (World Health Organization, 2021b).

P. knowlesi CSP (PkCSP) remains a potential pre-erythrocytic vaccine candidate for knowlesi malaria. The protein has RI and RII that are similar to PfCSP. PkCSP CRR was first thought to be simple with only two motifs, NEGQPQAQGDGA and EQPAAGAGG (McCutchan *et al*., 1996). However, recent sequencing of numerous *PkCSP* gene (Lee *et al*., 2011) revealed a highly polymorphic CRR with more than 46 repeat motifs occurring in various combinations and lengths. Thus far, several studies on non-repeat regions of PkCSP have been performed (Lee *et al.*, 2011; Fong *et al.*, 2015; Chong *et al.*, 2020) but there is no in-depth analysis of the PkCSP central region repeat motifs. Therefore, this study aims to examine the types, frequency and predicted antigenicity of the CSP central repeat motifs found in *P. knowlesi* isolates in Peninsular Malaysia.

MATERIAL AND METHODS

This study was approved by the Medical Research and Ethics Committee (MREC), Ministry of Health Malaysia (NMRR ID: NMRR-15-672-23975). Blood samples of *P. knowlesi*-infected patients (n=10) were collected from hospitals in Peninsular Malaysia states of Johor, Kelantan, Pahang, Perak, Negeri Sembilan, and Selangor between year 2019-2020. Genomic DNA was extracted from 100 µL blood samples using DNeasy® Blood and Tissue Kit (QIAGEN, Hilden, Germany) and eluted in 50 µL of Buffer AE. Extracted DNA was stored at -20! until use. Nested PCR targeting *Plasmodium 18S ribosomal RNA* gene was carried out to confirm *P. knowlesi* infection in the samples.

PCR primers flanking the *PkCSP* CRR were designed based on *P. knowlesi* strain H *circumsporozoite* (*CS*) *protein* gene (NCBI reference sequence: XM_002258966). The primers used were Pk/PvCSP_CR_F: 5'- AGAGGACTTGGTGA/GAAAG-3' and Pk/PvCSP_CR_R: 5'-CCCTGATTGTTTTGTCCC-3'. GoTaq® Long PCR master mix (Promega Corporation, USA) which consists of proofreading polymerase was used to amplify the gene to repair DNA mismatches and provide high fidelity reaction. PCR amplification conditions were as follows: initial denaturation at 95° for 4 min; 35 cycles of 95° for 1 min, 55° for 45 sec, 72° for 60 sec; final elongation at 95! for 10 min. PCR products with the expected size of ~650 bp were resolved by agarose gel electrophoresis. Upon validation, the PCR products were cloned into pGEM®-T plasmid vector (Promega Corporation, USA). One clone was selected for each sample. A total of 10 *PkCSP* CRR-pGEM®-T plasmids were sent to a commercial laboratory (Apical Scientific Sdn. Bhd., Malaysia) for Sanger sequencing. All sequences were deposited into GenBank (Accession number: ON022777 – ON022786). In addition to the 10 sequences generated in this study, 54 *PkCSP* gene sequences of Peninsular Malaysia *P. knowlesi* isolates were retrieved from GenBank (Accession number: KF861750 – KF861695). All 64 sequences were used for the analysis of *PkCSP* CRR. Their amino acid sequences were deduced from the nucleotide sequences using BioEdit.

Rapid Automatic Detection and Alignment of Repeats (RADAR) (https://www.ebi.ac.uk/Tools/pfa/radar/) was used for detection and alignment of repeats in the protein sequences. The software identifies gapped approximate repeats and complex repeat architectures involving many different types of repeats. No prior assumptions about the expected number or length of repeat units were made, and there was no limit imposed on the maximum length of the sequence. The repeat frequency was calculated using the formula:

Repeat frequency = Number of sequences having the same repeats ÷ Total number of sequences

Dominant repeat motifs identified using RADAR were evaluated using VaxiJen v2.0 (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html) for *in silico* identification of antigens. The model selected was 'parasite' and the threshold was set at 0.4. Antigenicity index of each dominant motif was generated.

Two online servers, BepiPred-2.0 (https://services.healthtech.dtu.dk/service.php?BepiPred-2.0) and BCPred (http://ailab-projects1.ist.psu.edu:8080/bcpred/predict.html), were used for the prediction of B-cell epitopes to ensure greater accuracy of the results. BepiPred-2.0 is a web server for sequence-based B-cell epitope prediction. It is trained on epitope data derived from crystal structures, therefore have significantly improved predictive power compared to BepiPred-1.0. Epitope threshold of 0.5 was used to predict the B-cell epitopes from PkCSP CRR sequences. Amino acid residues that gave epitope probability of more than 0.5 were identified as epitope residues. BCPred employs subsequence kernel-based support vector machine (SVM) classifier and was trained on homology-reduced dataset of linear B-cell epitopes (with <80% sequence identity). BCPred was run with the following characteristics: 22 amino acid epitope length, a specificity of 75%, and overlapping epitopes were selected. All the PkCSP CRR sequences were trimmed using BioEdit and the amino acid sequences used for prediction of B-cell epitopes were listed in Supplementary Files 1.1-1.7. Consensus epitopes reported from both servers were considered as potential B-cell epitopes.

RESULTS

PCR products ranged from 500 bp to 800 bp were determined from the agarose gel electrophoresis. Analysis of the *PkCSP* sequences using RADAR showed that all CRR contained different types, length, and number of motifs (Table 1). Motif with the highest copy number in each *PkCSP* CRR was classified as dominant. Dominant motifs, however, were not fully conserved as minor variations such as amino acid substitutions and indels were observed, and these were classified as variant motifs. Thirty-five dominant types of motifs were identified from the 64 *PkCSP* CRR sequences. The length of the dominant motifs ranged from 14 to 32 amino acids.

Among the sequences retrieved from GenBank, there were a few which were derived from the same blood samples, but their clones (from the ligation mixture) had different sequences. For example, one clone of UM0018 (C6) had different dominant and variant repeats from the other two clones of UM0018 (C9 and C10) which had same dominant repeat but different variant repeats. In another example, all three clones of UM0021 (C6, C10 and C19) had different dominant and variant repeats. This is likely due to multiple *P. knowlesi* infections in the blood samples. Furthermore, it was also noted that some clones from the same samples had the same dominant motif but different variant motifs, for instance, UM0047C10 had same dominant repeat but different variant repeat compared to UM0047C1. This could possibly be due to nucleotide misincorporation during PCR amplification or presence of true variants in the multiple-infected samples. Nonetheless, all the variant motifs were considered as true variants in this study to simplify the analysis since this study focused mainly on the dominant motifs.

Close examination of the sequences revealed 'AGQPQAQGDGANAGQPQAQGDGAN' as the most prevalent motif (n=15, frequency: 0.234), followed by 'PGQPEGNREAPGQPEGNRE' (n=7, frequency: 0.109), and "NAEGGANAGQPNAEGGANA" (n=5, frequency: 0.0781) (Table 2). Sixteen of the sequences contained rare motifs (n=22, frequency: 0.0156).

Table 1. Dominant and variant repeat motifs from 64 *P. knowlesi* CSP sequences. The repeat motifs were identified using RADAR. Repeat motifs that occur at higher copies in the sequence were identified as the dominant motifs. Repeat motifs that occur at lower copies in the sequence were identified as the variant motifs

Sample	CDS Accession Number	Dominant motif (copies in sequence)	Variant motif (copies in sequence)
UMMK569C20	KF861749	AGQPQAQGDGANAGQPQAQGDGAN (4)	AGQPQAQGDGANAGQPQARGDGAN (1) AGQPQAQGDRANAGQPQAQGDGAN (1)
UM0021C6	KF861746	AGQPQAQGDGANAGQPQAQGDGAN (4)	AGQPQAQGDGANTGQPQAQGDGAN (1) AGQPQAQGDRANAGQPQAQGDGAN (1)
UM0050C2	KF861744	AGQPQAQGDGANAGQPQAQGDGAN (5)	AGQPQAQGDRANAGQPQAQGDGAN (1)
UM0034C7	KF861737	AGQPQAQGDGANAGQPQAQGDGAN (5)	PDQPQAQGNGANAGQPQAQGDGAN (1) AGQPQAQGDGANAGQPQAQGDRAN (1)
UM0020C3	KF861730	AGQPQAQGDGANAGQPQAQGDGAN (4)	AGQPQAQGDRANAGQPQAQGDGAN (1) GQPQAQGDGAN (2)
UM0020C1	KF861729	AGQPQAQGDGANAGQPQAQGDGAN (4)	AGQPQAQGDRANAGQPQAQGDGAN (1) GQPQAQGDGAN (2)
UM004C4	KF861721	AGQPQAQGDGANAGQPQAQGDGAN (3)	AGQPQAQGDRANAGQPQAQGDGAN (1)
UM004C1	KF861720	AGQPQAQGDGANAGQPQAQGDGAN (3)	AGQPQAQGDRANAGQPQAQGDGAN (1)
UM001C5	KF861716	AGQPQAQGDGANAGQPQAQGDGAN (5)	AGQPQAQGDRANAGQPQAQGDGAN (1)
UMGAC7	KF861697	AGQPQAQGDGANAGQPQAQGDGAN (4)	AGQPQAQGDRANAGQPQAQGDGAN (1) GQPQAQGDGAN (2)
UMGAC4	KF861696	AGQPQAQGDGANAGQPQAQGDGAN (4)	AGQPQAQGDRANAGQPQAQGDGAN (1) GQPQAQGDGAN (2)
UMMK566C6	KF861750	AGQPQAQGDGANAGQPQAQGDGAN (4)	AGQPQAQGDRANAGQPQAQGDGAN (1) GQPQAQGDGAN (2)
UMGAC1	KF861695	AGQPQAQGDGANAGQPQAQGDGAN (4)	AGQPQAQGDRANAGQPQAQGDGAN (1) GQPQAQGDGAN (1) GQPRAQGDGAN (1)
NS3_0002	ON022778	AGQPQAQGDGANAGQPQAQGDGAN (4)	AGQPQAQGDRANAGQPQAQGDGAN (1) GQPQAQGDGAN (2)
PHG3_0002	ON022785	AGQPQAQGDGANAGQPQAQGDGAN (4)	AGQPQAQGDRANAGQPQAQGDGAN (1) GQPQAQGDGAN (2)
UM001C7	KF861717	NAGQPQAQGDGANAGQPQAQGDGA (5)	NAGQPQAQGDRANAGQPQAQGDGA (1) NEGQPQAQGDGA (1) NAGQPQAQGDGA (1)
UM0034C10	KF861738	GQPQAQGDGANAGQPQAQGDG (5)	GQPQAQGDRANAGQPQAQGDG (1)
UM0034C6	KF861736	GQPQAQGDGANAGQPQAQGDG (5)	GQPQAQGDRANAGQPQAQGDG (1)
UMHIC6	KF861715	PGQPEGNREAPGQPEGNRE (6)	PAQPEGNREAPAQPQGNGG (1)
UMHIC2	KF861714	PGQPEGNREAPGQPEGNRE (6)	PAQPEGNREAPAQPQGNGG (1)
UMJUC2	KF861711	PGQPEGNREAPGQPEGNRE (6)	PAQPEGNREAPAQPQGNGG (1)
UMJUC1	KF861710	PGQPEGNREAPGQPEGNRE (5)	PGQPEGNREAPGQPEGDRE (1) PAQPEGNREAPAQPQGNGG (1)
UMSYC3	KF861706	PGQPEGNREAPGQPEGNRE (3)	PGQPEGNREAPAQPEGNRE (2)
UMSY2C2	KF861705	PGQPEGNREAPGQPEGNRE (4)	PGQPEGNREAPAQPEGNRE (2)
UMANC1	KF861699	PGQPEGNREAPGQPEGNRE (5)	PAQPEGNREAPAQPQGNGG (1)
UMOTC3	KF861704	PGQPEGNREAPGQPEGNR (4)	PAQPEGNREAPGQPEGNR (1) QAQPEENREAPAQPQGNG (1)
UMHAC5	KF861712	PGQPEGNREAPGQPEGNR (3)	PGQPEGNREAPAQPEGNR (2)

UMRAC2	KF861698	PGQPEGNREAPGQPEGNR (3)	PGQPEGNREAPAQPEGNR (1) PGQPEGNREAPAQPQGNG (1) QAQPEGNREAPAQPQGNG (1)
UM0016C2	KF861743	RAQPEGNQDGRAQPEGN (4)	QPEGNQDGRAQPEGN (1) PAQPQGNGGAGQAQPEGN (2) PAQPQGNGGAGQAQPQKN (1)
UM0015C7	KF861742	NAEGGANAGQPNAEGGANA (4)	NAEGGANARQPNAEGGANA (2) GGGANARQPQAEGGGAN (1) GGNEGNKQAGKGQGQN (1)
UM0015C4	KF861741	NAEGGANAGQPNAEGGANA (4)	NAEGGANARQPNAEGGANA (2) GGGANARQPQAEGGGAN (1) GGNEGNKQAGKGQGQN (1)
UM0047C10	KF861740	NAEGGANAGQPNAEGGANA (4)	NAEGGANARQPNAEGGANA (2) QAEGGGANARQPQAEGGGANA (1)
UM0047C1	KF861739	NAEGGANAGQPNAEGGANA (4)	NAEGGANARQPNAEGGANA (2) GGGANARQPQAEGGGAN (1) GGNEGNKQAGKGQGQN (1)
UM009C1	KF861723	NAEGGANAGQPNAEGGANA (4)	NAEGGANARQPNAEGGANA (2) QAEGGGANARQPQAEGGGANA (1)
UM0018C10	KF861728	PNAEGGANAGQPNAEGGANAGQ (4)	PNAEGGANAGQPNAEGGANARQ (1) PNAEGGANARQPNAEGGANARQ (1) PNAEGGANARQPQAEGGGANARQ (1)
UM0018C9	KF861727	PNAEGGANAGQPNAEGGANAGQ (4)	PNAEGGANAGQPNAEGGANARQ (1) PNAEGGANARQPNAEGGANARQ (1) PNAEGGANARQPQAEGGGANARQ (1)
UM0018C6	KF861726	EGGANAGQPNAEGGANAGQPN (4)	EGGANAGQPNAEGGANARQPN (1) EGGANARQPNAEGGANARQPN (1) EGGANARQPQAEGGGANARQPQ (1)
UM0032C8	KF861735	RQPNAEGDGANARQPNAEGDG (2)	RQPNAEGDGANARQPNAEGNG (1) RQPNAEGGGANARQPNAEGDG (1) RQPNAEGDGANARQPNAEGGG (1) RQPQAEGDGANARQPQAQGDG (1)
UM0032C4	KF861733	RQPNAEGDGANARQPNAEGDG (2)	RQPNAEGDGANARQPNAEGGG (1) RQPQAEGDGANARQPQAQGDG (1)
UMNGC3	KF861708	GANAGQPNAEGGANAGQP (2) GANAGQPNAEGGANARQP (2)	NAGQPNAGGANAGQP (1) GANAGQPNAEGGANA (1)
UMNGC2	KF861707	GANAGQPNAEGGANAGQP (2) GANAGQPNAEGGANARQP (2)	NAGQPNAGGANAGQP (1) GANAGQPNAEGGANA (1)
UM009C4	KF861724	GGANAGQPNaeGGANAGQPNAE (3)	GGANARQPNAEGGANARQPNAE (1) GGANARQPNAEGGANAGQPNAE (1) GGANAGQPNAEGGANARQPQAE (1) GGANARRPQAEGGGANARQGGNE (1)
UMNGC4	KF861709	GGANAGQPNAEGGANAGQPNAE (2) GGANARQPNAEGGANAGQPNAE (2)	PNAEGGANAGQPNAE (1) GGANAGQPNAEGGANAGQPNA (1) GGANAGQPNAEGGANAGRPNAE (1) GGANAGQPNAEGGANARQPQAE (1)
UM0050C7	KF861745	AGAGGEQPAAGAGGEQPA (3)	AGARGEQPAAGAGGEQPA (2) PAPRREQPAAGAGGEQPA (1)
UMCHC1	KF861713	AGAGGEQPAAGAGGEQPA (3)	AGARGEQPAAGAGGEQPA (2) PAPRREQPAAGAGGEQPA (1)

UM0029C9	KF861732	AGARGEQPAAGAGGEQPA (3)	PAPRREQPAAGAGGEQPA (1)
KEL3_0002	ON022779	AGAGGEQPAAGAGGEQPA (5)	
UMHEC1	KF86170	AGGEQPAAGAGGEQ (3)	AGGEQPAAGARGEQ (2) PRREQPAAGAGGEQ (1)
PRK3_0004	ON022784	GEQPAAGAGGEQPAAGAG (3)	AGAGGEQPAAGAGGEQPA (1) PAPRREQPAAGAGGEQPA (1) EQPAAGAGGEQPA (2) EQPAPGAGAGDGA (1)
PRK3_0002	ON022783	GEQPAAGAGGEQPAAGAG (3)	AGAGGEQPAAGAGGEQPA (1) PAPRREQPAAGAGGEQPA (1) EQPAAGAGGEQPA (2) EQPAPGAGAGDGA (1)
UM0029C8	KF861731	GARGEQPAAGAGGEQPAA (2)	GAGGEQPAAGAGGEQPAA (1) GAGGEQPAAGGEQPAA (1) GARGEQPAAGAGGEQPAP (1) APRREQPAAGAGGEQPAP (1)
UMMAC6	KF861702	GAGGEQPAAGAGGEQP (4)	GARGEQPAAGAGGEQP (2) GAGGEQPAPAPRREQP (2)
UM0014C4	KF861725	PAPGAGGEQPAPGAGG (5)	PAPAPRREQPAPGPGG (1) PAPAPRREQPAPGAGA (1)
SEL3_0001	ON022781	PAPGAGGEQPAPGAGG (3)	PAPGAGGEQPAPEAGG (1) PAPGAGGEQPAPAPRR (2)
UM0032C7	KF861734	PEGNGGAGQAQPEGNGGAG (4)	PEGNGGAGQAQPQGNGGAG (1) PQGNGGAGQAQPQGNGGAG (1) PEGNREAPAQPQGNGGAG (1)
KEL3_0012	ON022780	PEGNGGAGQAQPEGNGGAG (4)	PEGNGGAGQAQPQGNGGAG (1) PQGNGGAGQAQPQGNGGAG (1) PEGNREA.PAQPQGNGGAG (1)
UM002C4	KF861719	PEGNGGAGPAQPEGNGGAGP (5)	PQGNGGAGPAQPQGNGGAGP (1) PQGNGGAGPAQPQGNGGAGQ (1)
UM002C3	KF861718	GGAGPAQPEGNGGAGPAQPEGN (3)	GGAGPAQPEGNGGAGPAQPGGQ (1) GGAGPAQPQGNGGAGPAQPQGN (1) MGGQVQHNRKEMGGQVQHNHKE (1) MGGQVQHNHKEMGGAGPAQPQGNGGAGPAQPQGNGGAGQAQPQKNE (1)
UM006C6	KF861722	EQPAPGGEQPAPGGEQPAPGG (3)	KQPAPGGEQPAPGGEQPAPGG (1) EQPAPGGEQPAPAPRREQPAPGG (1)
SEL3_0005	ON022782	EQPAPGGEQPAPGGEQPAPGG (2) EQPAPGGEQPAPGGEQPAPAPRR (2)	KQPAPGGEQPAPGGEQPAPGG (1)
UM0021C19	KF861748	PGGEQAGPGGEQAGPR (4)	PGGEQAGPGGEQPAPR (1)
UM0021C10	KF861747	AGGEQPAAGGEQPAAGGERP (2)	AGGEQPAAGGEQPAAGGEQP (1) AGGEQPAAGGERPAAGGEQP (1) AGGEQPAAGGEQPAP (1) PRREQPAAGGEQPAP (1)
JHR3_0001	ON022777	QPAPRPGGEQAGPGG (4)	QPAPRPGGEQPAPGPGG (1) QPAPAPRREQPAPGG (1)
PHG3_0006	ON022786	PRPGGEQAGPGGEQAGPRPGGEQAGPGGEQAG (2)	QPGGEQAGPGGEQAGPGGEQAGPGGEQAG (1) PGGEQAGPGGEQAGPGGEQ (1) PGGEQAGPGGEQPAPRPGGEQ (1) PAPAPRREQPAPGPG (1) PAPGAGAGDGARGG (1)

Table 2. Dominant motif frequency and predicted antigenicity

Sample Accession Number	Dominant repeat	Repeat frequency	Overall protective antigen prediction
KF861749 KF861746 KF861744 KF861737 KF861730 KF861729 KF861721 KF861720 KF861716 KF861697 KF861696 KF861750 KF861695 ON022778 ON022785	AGQPQAQGDGANAGQPQAQGDGAN	0.234	1.7986
KF861715 KF861714 KF861711 KF861710 KF861706 KF861705 KF861699	PGQPEGNREAPGQPEGNRE	0.109	0.8619
KF861742 KF861741 KF861740 KF861739 KF861723	NAEGGANAGQPNAEGGANA	0.0781	1.1829
KF861712 KF861704 KF861698	PGQPEGNREAPGQPEGNR	0.0469	0.8831
KF861738 KF861736	GQPQAQGDGANAGQPQAQGDG	0.0313	1.9926
KF861728 KF861727	PNAEGGANAGQPNAEGGANAGQ	0.0313	1.1234
KF861735 KF861733	RQPNAEGDGANARQPNAEGDG	0.0313	1.8109
KF861708 KF861707	GANAGQPNAEGGANAGQP GANAGQPNAEGGANARQP	0.0313	1.1682 1.1121
KF861745 KF861713	AGAGGEQPAAGAGGEQPA	0.0313	1.3878
ON022784 ON022783	GEQPAAGAGGEQPAAGAG	0.0313	1.0972
KF861743	RAQPEGNQDGRAQPEGN	0.0156	2.1919
KF861726	EGGANAGQPNAEGGANAGQPN	0.0156	1.1164
KF861724	GGANAGQPNAEGGANAGQPNAE	0.0156	1.1383
KF861709	GGANAGQPNAEGGANAGQPNAE GGANARQPNAEGGANAGQPNAE	0.0156	1.1383 1.1129
KF861732	AGARGEQPAAGAGGEQPA	0.0156	1.3233
KF861731	GARGEQPAAGAGGEQPAA	0.0156	1.3638
KF861725	PAPGAGGEQPAPGAGG	0.0156	0.7318
KF861702	GAGGEQPAAGAGGEQP	0.0156	1.4317
KF86170	AGGEQPAAGAGGEQ	0.0156	1.0735
ON022779	AGAGGEQPAAGAGGEQPA	0.0156	1.3878
ON022781	PAPGAGGEQPAPGAGG	0.0156	0.7318
KF861734	PEGNGGAGQAQPEGNGGAG	0.0156	2.0977
ON022780	PEGNGGAGQAQPEGNGGAG	0.0156	2.0977
KF861719	PEGNGGAGPAQPEGNGGAGP	0.0156	1.8987
KF861718	GGAGPAQPEGNGGAGPAQPEGN	0.0156	1.8285
KF861722	EQPAPGGEQPAPGGEQPAPGG	0.0156	1.0630
ON022782	EQPAPGGEQPAPGGEQPAPGG EQPAPGGEQPAPGGEQPAPAPRR	0.0156	1.0630 0.5916
KF861748	PGGEQAGPGGEQAGPR	0.0156	1.0957
KF861747	AGGEQPAAGGEQPAAGGERP	0.0156	0.7997
ON022777	QPAPRPGGEQAGPGG	0.0156	0.9175
ON022786	PRPGGEQAGPGGEQAGPRPGGEQAGPGGEQAG	0.0156	1.0751
KF861717	NAGQPQAQGDGANAGQPQAQGDGA	0.0156	1.7136

Prediction of the antigenicity of the dominant motifs showed that all the dominant repeats are probable antigens. Among them, 'RAQPEGNQDGRAQPEGN' had the highest antigenicity index of 2.1919 but the repeat frequency of this motif was only 0.0156 (n=1). The most prevalent motif had an antigenicity index of 1.7986 which was the eighth highest index among the dominant motifs. Majority of the dominant motifs had antigenicity indexes of more than 1 while seven dominant motifs have antigenicity indexes below 1 and above the antigenic threshold of 0.4. When the PkCSP CRR sequences were input into BepiPred-2.0 and BCPred, the servers predicted all the repeat regions as B-cell epitopes.

DISCUSSION

Being a major surface protein of sporozoites, PkCSP can be a potential pre-erythrocytic vaccine candidate for *P. knowlesi* infection. This study is the first to characterise the central repeat region of *P. knowlesi* CSP, in order to predict its immunogenicity through *in silico* approaches: identification of repeat motifs, antigenicity and B-cell epitope prediction.

As mentioned earlier, the central repeat region of *P. knowlesi* is hyperpolymorphic. This was shown by Lee *et al*. (2011) who provided a list of CSP repeat motifs in *P. knowlesi* isolated from human and macaque samples. Our study employed the bioinformatic tool RADAR to identify the repeat motifs. This program uses an automatic algorithm by segmenting query sequence into repeats and identified short composition biased as well as gapped approximate repeats. It has been used in other studies to identify internal repeats in protein sequences. The sensitivity and accuracy of RADAR repeats show good coverage, accurate alignments, and reasonable repeat borders when compared to Pfam, a database of protein domain families (Heger & Holm, 2000). With the use of RADAR, repeats of varying amino acid lengths were identified from the PkCSP sequences. In comparison to Lee *et al*. (2011), the repeats identified by RADAR were generally longer. In this study, the minimum repeat length seen was 14 amino acids, and the maximum was 32 amino acids, and the number of dominant repeat motifs ranged from 2 to 6 among the PkCSP sequences whereas in Lee *et al*. (2011), the minimum repeat length was 7 amino acids, the maximum repeat length was 12 amino acids, and the number of dominant repeat motifs ranged from 1 to 13.

Identification of B-cell epitopes is extensively employed in the development of diagnostic tests, therapeutics, and vaccines. B-cell epitopes are clusters of amino acids which can be recognized by secreted antibodies and are able to elicit humoral immune response. They can be categorized as linear and conformational epitopes. Amino acids that have surface accessibility is a common feature of B-cell epitopes. The methods to identify epitopes can generally be divided into structural (e.g., X-ray crystallography) and functional studies (e.g., screening of antigen-derived proteolytic fragments / peptides for antibody binding, evaluation of antigen-antibody reactivity of mutated antigens, and display technologies in epitope mapping) (Van Regenmortel, 2009). Various algorithms have been developed for B-cell epitopes prediction from a protein sequence or structure. The B-cell epitope prediction using both Bepipred-2.0 and BCPred showed that all identified repeat motifs of PkCSP are potential epitopes. This indicates that the repeat motifs could induce immune response, thus producing antibodies that target against the antigenic region of PkCSP protein.

Based on Table 2, the dominant motif 'AGQPQAQGDGANA GQPQAQGDGAN' is the most promising to be included in the design of *P. knowlesi* vaccine candidate. The repeat motif has a high antigenicity index and highest frequency among the 64 PkCSP sequences. No significant identity was found when compared the motif with human proteins using Protein Basic Local Alignment Search Tool (BLASTp), indicating that the motif maybe antigenic against human immune system.

T-cell epitope prediction was not included in this study despite several T-cell epitope prediction software are available online. Studies on *P. falciparum* CSP have reported the absence *P. falciparum*-specific CD8+ T-cell epitopes in the repeat region. Besides, examination of individuals vaccinated with RTS,S also exhibited inconsistent CD8+ T-cell responses (Moorthy & Ballou, 2009). In contrast, CD4+ T-cell epitopes have been reported in the repeat region of PfCSP (Esposito *et al*., 1992). Epitope mapping based on sporozoite-specific human T-cell clones revealed that the CD4+ T-cell epitope was contained in the NANPNVDPNANP sequence of the repeat region (Nardin, 1990). Nevertheless, the CD4+ T-cell response did not confer high protection level in mouse liver after parasite challenge compared to the humoral immune response (Kastenm ller *et al*., 2013). Owing to these reasons, this study focused on the *in silico* characterisation of the humoral response induced by the repeat region of PkCSP.

All the parameters studied are important in characterising and determining repeat motifs for candidate vaccine design. This study provides a fundamental analysis of *P. knowlesi* CSP as a possible pre-erythrocytic candidate for knowlesi malaria infection. Future research should focus on the characterization of PkCSP CRR of Malaysian Borneo, a region with high prevalence of knowlesi malaria. It is pertinent to compare the PkCSP CRR between the two geographical regions in Malaysia to select consensus repeat motif for vaccine design. In addition, biological characterization of the selected repeat motifs could be carried out using *in vitro* and *in vivo* models as well as field serum sample analysis.

CONCLUSIONS

The hyperpolymorphism of PkCSP CRR is due to the occurrence of repeat motifs of different amino acid composition, length, and copy numbers in a sequence. *In silico* approaches revealed that all the predicted repeats were antigenic and consisted of B-cell epitopes. This study provides fundamental knowledge for further studies such as field serum sample analysis and animal model experimentations on possible use of *P. knowlesi* repeat motif as vaccine against human knowlesi malaria.

Conflict of interest statement

The author declares that they have no conflict of interests.

Ethics statement

This study was approved by the Medical Research and Ethics Committee (MREC), Ministry of Health Malaysia (NMRR ID: NMRR-15-672-23975).

ACKNOWLEDGEMENTS

The study was supported by the Long Term Research Grant Scheme from the Ministry of Higher Education, Malaysia (Grant Code: LRGS/1/2018/UM/01/1/1). We would like to express our gratitude to Dr. Mohd Hafizi Abdul Hamid, Dr. Jenarun Jelip, and Dr. Rose Nani Mudin from Vector Borne Disease Sector, Disease Control Division, Ministry of Health, Malaysia for their contribution in coordinating nationwide sample collection. Also, we would like to thank the public health officials from the State Health Departments and District Health Offices in Johor, Kelantan, Pahang, Perak, Negeri Sembilan, and Selangor for their assistance in sample collection.

REFERENCES

Casares, S., Brumeanu, T.-D. & Richie, T.L. (2010). The RTS, S malaria vaccine. *Vaccine* **28**: 4880-4894. https://doi.org/10.1016/j.vaccine.2010.05.033

Chong, E.T.J., Neoh, J.W.F., Lau, T.Y., Lim, Y.A.-L., Chai, H.C., Chua, K.H. & Lee, P.-C. (2020). Genetic diversity of circumsporozoite protein in *Plasmodium knowlesi* isolates from Malaysian Borneo and Peninsular Malaysia. *Malaria Journal* **19**: 1-7. https://doi.org/10.1186/s12936-020-03451-x

Daneshvar, C., Davis, T.M., Cox-Singh, J., Rafa'ee, M.Z., Zakaria, S.K., Divis, P.C. & Singh, B. (2009). Clinical and laboratory features of human *Plasmodium knowlesi* infection. *Clinical Infectious Diseases* **49**: 852-860. https://doi.org/ 10.1086/605439

Esposito, F., Lombardi, S., Modiano, D., Habluetzel, A., Del Nero, L., Lamizana, L., Pietra, V., Rotigliano, G., Corradin, G. & Ravot, E. (1992). In vitro immune recognition of synthetic peptides from the *Plasmodium falciparum* CS protein by individuals naturally exposed to different sporozoite challenge. *Immunology Letters* **33**: 187-199. https://doi.org/10.1016/0165-2478(92)90046-q

Ferguson, D.J., Balaban, A.E., Patzewitz, E.-M., Wall, R.J., Hopp, C.S., Poulin, B., Mohmmed, A., Malhotra, P., Coppi, A. & Sinnis, P. (2014). The repeat region of the circumsporozoite protein is critical for sporozoite formation and maturation in *Plasmodium*. *PloS One* **9**: e113923. https://doi.org/10.1371/journal.pone.0113923

Fong, M.Y., Ahmed, M.A., Wong, S.S., Lau, Y.L. & Sitam, F. (2015). Genetic diversity and natural selection of the *Plasmodium knowlesi* circumsporozoite protein nonrepeat regions. *PloS One* **10**: e0137734. https://doi.org/10.1371/journal.pone.0137734

Heger, A. & Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins: Structure, Function, and Bioinformatics* **41**: 224-237. https://doi.org/10.1002/1097-0134(20001101)41:2<224::aid-prot70>3.0.co;2-z

Kastenm ller, K., Espinosa, D.A., Trager, L., Stoyanov, C., Salazar, A.M., Pokalwar, S., Singh, S., Dutta, S., Ockenhouse, C.F. & Zavala, F. (2013). Full-length *Plasmodium falciparum* circumsporozoite protein administered with long-chain poly (I· C) or the Toll-like receptor 4 agonist glucopyranosyl lipid adjuvant-stable emulsion elicits potent antibody and CD4+ T cell immunity and protection in mice. *Infection and Immunity* **81**: 789-800. https://doi.org/10.1128/IAI.01108-12

Lê, H.G., Kang, J.-M., Moe, M., Jun, H., Thái, T.L., Lee, J., Myint, M.K., Lin, K., Sohn, W.-M. & Shin, H.-J. (2018). Genetic polymorphism and natural selection of circumsporozoite surface protein in P*lasmodium falciparum* field isolates from Myanmar. *Malaria Journal* **17**: 1-14. https://doi.org/10.1186/s12936-018-2513-0

Lee, K.-S., Divis, P.C., Zakaria, S.K., Matusop, A., Julin, R.A., Conway, D.J., Cox-Singh, J. & Singh, B. (2011). *Plasmodium knowlesi*: reservoir hosts and tracking the emergence in humans and macaques. *PLoS Pathogens* **7**: e1002015. https://doi.org/10.1371/journal.ppat.1002015

Mazier, D., Mellouk, S., Beaudoin, R., Texier, B., Druilhe, P., Hockmeyer, W., Trosper, J., Paul, C., Charoenvit, Y. & Young, J. (1986). Effect of antibodies to recombinant and synthetic peptides on *P. falciparum* sporozoites in vitro. *Science* **231**: 156-159. https://doi.org/10.1126/science.3510455

McCutchan, T.F., Kissinger, J.C., Touray, M.G., Rogers, M.J., Li, J., Sullivan, M., Braga, E. M., Krettli, A.U. & Miller, L.H. (1996). Comparison of circumsporozoite proteins from avian and mammalian malarias: biological and phylogenetic implications. *Proceedings of the National Academy of Sciences* **93**: 11889-11894. https://doi.org/10.1073/pnas.93.21.11889

Moorthy, V.S. & Ballou, W.R. (2009). Immunological mechanisms underlying protection mediated by RTS, S: a review of the available data. *Malaria Journal* **8**: 1-7. https://doi.org/10.1186/1475-2875-8-312

Nardin, E.H. (1990). T cell responses in a sporozoite-immunized human volunteer and a chimpanzee. *Immunology Letters* **25**: 43-47. https://doi.org/10.1016/0165-2478(90)90089-9

Persson, C., Oliveira, G.A., Sultan, A.A., Bhanot, P., Nussenzweig, V. & Nardin, E. (2002). Cutting edge: a new tool to evaluate human pre-erythrocytic malaria vaccines: rodent parasites bearing a hybrid *Plasmodium falciparum* circumsporozoite protein. *The Journal of Immunology* **169**: 6681-6685. https://doi.org/10.4049/jimmunol.169.12.6681

Rathore, D., Sacci, J.B., de la Vega, P. & McCutchan, T.F. (2002). Binding and invasion of liver cells by *Plasmodium falciparum* sporozoites: essential involvement of the amino terminus of circumsporozoite protein. *Journal of Biological Chemistry* **277**: 7092-7098. https://doi.org/10.1074/jbc.M106862200

Rosenberg, R., Wirtz, R.A., Lanar, D.E., Sattabongkot, J., Hall, T., Waters, A.P. & Prasittisuk, C. (1989). Circumsporozoite protein heterogeneity in the human malaria parasite *Plasmodium vivax*. *Science* **245**: 973-976. https://doi.org/10.1126/science.2672336

Saralamba, N., Mayxay, M., Newton, P.N., Smithuis, F., Nosten, F., Archasuksan, L., Pukrittayakamee, S., White, N.J., Day, N.P. & Dondorp, A.M. (2018). Genetic polymorphisms in the circumsporozoite protein of *Plasmodium malariae* show a geographical bias. *Malaria Journal* **17**: 1-9. https://doi.org/10.1186/s12936-018-2413-3

Singh, B., Sung, L.K., Matusop, A., Radhakrishnan, A., Shamsul, S.S., Cox-Singh, J., Thomas, A., & Conway, D.J. (2004). A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *The Lancet* **363**: 1017-1024. https://doi.org/10.1016/S0140-6736(04)15836-4

Van Regenmortel, M.H. (2009). What is a B-cell epitope? In: Epitope Mapping Protocols. Methods in Molecular biology, vol 524, Schutkowski, M. & Reineke, U. (editors). Human Press, pp. 3-20. https://doi.org/10.1007/978-1-59745-450-6_1

Vγ, T.C., Lך, H.G., Kang, J.-M., Moe, M., Naw, H., Myint, M.K., Lee, J., Sohn, W.-M., Kim, T.-S. & Na, B.-K. (2020). Genetic polymorphism and natural selection of circumsporozoite protein in Myanmar *Plasmodium vivax*. *Malaria Journal* **19**: 1-17. https://doi.org/10.1186/s12936-020-03366-7

World health Organization. (2020). World malaria report 2020. https://www.who.int/publications/i/item/9789240015791. Accessed 21 Jun 2022.

World Health Organization. (2021). World malaria report 2021. https://www.who.int/publications/i/item/9789240040496. Accessed 21 Jun 2022.

World Health Organization. (2021). Full evidence report on the RTS, S/AS01 malaria vaccine. *SAGE Yellow Book*.

Yap, N.J., Hossain, H., Nada-Raja, T., Ngui, R., Muslim, A., Hoh, B.-P., Khaw, L.T., Kadir, K.A., Divis, P.C.S. & Vythilingam, I. (2021). Natural human infections with *Plasmodium cynomolgi*, *P. inui*, and 4 other simian malaria parasites, Malaysia. *Emerging Infectious Diseases* **27**: 2187. https://doi.org/10.3201/eid2708.204502

Ying, P., Shakibaei, M., Patankar, M.S., Clavijo, P., Beavis, R.C., Clark, G.F. & Frevert, U. (1997). The malaria circumsporozoite protein: interaction of the conserved regions I and II-plus with heparin-like oligosaccharides in heparan sulfate. *Experimental Parasitology* **85**: 168-182. https://doi.org/10.1006/expr.1996.4134

Young, J.F., Hockmeyer, W.T., Gross, M., Ballou, W.R., Wirtz, R.A., Trosper, J.H., Beaudoin, R.L., Hollingdale, M.R., Miller, L.H. & Diggs, C.L. (1985). Expression of *Plasmodium falciparum* circumsporozoite proteins in *Escherichia coli* for potential use in a human malaria vaccine. *Science.* **228**: 958-962. https://doi.org/10.1126/science.2988125

SUPPLEMENTARY DATA

https://msptm.org/files/Vol40No1/tb-40-1-004-Tan-J-H-supplementary-data.pdf