Supplementary file 1. Nucleotide sequence alignment of sequenced *PkCSP* gene (Seq *PkCSP*) with full reference *PkCSP* gene (Ref *PkCSP*) (NCBI reference sequence: NC_011909, Location: 1757461-1758552). BLASTn analysis showed percentage identity of 99%, confirming the identity of the amplified gene sequence.

```
Ref PkCSP    ATGAAGAACT TCATTCTCTT GGCCGTCTCC TCCATCCTGC TGGTGGACTT GCTCCCCACA  60
Seq PkCSP    ---------- ---------- ---------- ---------- ---------- -------...  3

Ref PkCSP    CACTTCGAAC ATAATGTAGA TCTCTCCAGG GCCATAAATG TAAATGGAGT AAGCTTCAAT  120
Seq PkCSP    .......... .......... .......... .......... .......... ..........  63

Ref PkCSP    AATGTAGACA CCAGTTCACT TGGCGCAGCA CAGGTGAGAC AAAGTGCTAG CCGAGGCAGA  180
Seq PkCSP    .......... .......... .......... .......... .......... ..........  123

Ref PkCSP    GGACTTGGTG AGAAGCCAAA AGAAGGAGCT GATAAAGAAA AGAAAAAAGA AAAAGGAAAA  240
Seq PkCSP    .......... .......... .......... .......... .......... ..........  183

Ref PkCSP    GAAAAAGAAG AAGAACCAAA GAAGCCAAAT GAAAATAAGC TGAAACAACC GAATGAAGGA  300
Seq PkCSP    .......... .......... .......... .......... .......... ..........  243

Ref PkCSP    CAACCACAAG CACAGGGTGA TGGAGCAAAT GCAGGACAAC CACAAGCACA AGGAGATGGA  360
Seq PkCSP    .......... .......... .......... .......... .......... ..........  303

Ref PkCSP    GCAAATGCAG GACAACCACA AGCACAGGGT GATGGAGCAA ATGCAGGACA ACCACAAGCA  420
Seq PkCSP    .......... .......... .......... .......... .......... ..........  363

Ref PkCSP    CAGGGTGATG GAGCAAATGC AGGACAACCA CAAGCACAAG GAGATGGAGC AAATGCAGGA  480
Seq PkCSP    .......... .......... .......... .......... .......... ..........  423

Ref PkCSP    CAACCACAAG CACAGGGTGA TGGAGCAAAT GCAGGGCAAC CACAAGCACA GGGTGATGGA  540
Seq PkCSP    .......... .......... .......... .......... .......... ..........  483

Ref PkCSP    GCAAATGCAG GACAACCACA AGCACAAGGA GATGGAGCAA ATGCAGGACA ACCACAAGCA  600
Seq PkCSP    .......... ...G...... .......... .......... .......... ..........  543

Ref PkCSP    CAAGGAGATG GAGCAAATGC AGGACAACCA CAAGCACAGG GTGATGGAGC AAATGCAGGA  660
Seq PkCSP    .......... .......... .......... .......... .......... ..........  603

Ref PkCSP    CAACCACAAG CACAGGGTGA TAGGGCGAAT GCAGGACAAC CACAAGCACA AGGAGATGGG  720
Seq PkCSP    .......... .......... .......... .......... .......... ..........  663

Ref PkCSP    GCAAATGTAC CACGACAAGG AAGAAACGGG GGAGGTGCAC CAGCAGGAGG AAATGAGGGG  780
Seq PkCSP    .......... .......... .......... .......... .......... ..........  723

Ref PkCSP    AATAAACAAG CAGGAAAAGG ACAGGGACAA AACAATCAGG GTGCGAATGC CCCAAATGAA  840
Seq PkCSP    .......... .......... .......... .......... .......... ..........  783

Ref PkCSP    AAAGTTGTGA ATGATTACCT ACACAAAATT AGATCTAGCG TTACCACCGA GTGGACTCCA  900
Seq PkCSP    .......... .......... .......... .......... .......... ..........  843

Ref PkCSP    TGCAGTGTAA CCTGTGGAAA TGGTGTAAGA ATTAGAAGAA AAGCTCATGC AGGTAATAAA  960
Seq PkCSP    .......... .......... .......... .......... .......... ..........  903

Ref PkCSP    AAGGCAGAGG ACCTTACTAT GGATGACCTT GAGGTGGAAG CTTGTGTAAT GGATAAGTGC  1020
Seq PkCSP    .......... .......... .......... .......... .......... ..........  963

Ref PkCSP    GCTGGCATAT TTAACGTTGT GAGTAATTCA TTAGGCTTAG TCATATTGTT AGTCCTAGCA  1080
Seq PkCSP    .......... .......... .......... .......... .......... ..........  1023

Ref PkCSP    TTATTCAATT AA 1092
Seq PkCSP    .......... .. 1035
```

Supplementary file 2. *Plasmodium knowlesi* circumsporozoite protein sequence deduced from nucleotides of the amplified *PkCSP* gene. Signal peptide region, RI and RII motif, central repeat region, Pk Th2R, and PkTh3R of the PkCSP were identified and labelled accordingly.

Signal peptide region

```
1   ------------------THFEHNVDLSRAINVNGVSFNNVDTSSLGAAQVRQSASRGRGLGEKPKEGA 51

                              RI motif
52  DKEKKKEKGKEKEEEPKKPNENKLKQPNEGQPQAQGDGANAGQPQAQGDGANAGQPQAQGDGANAGQPQA 121

                                          Central repeat region
122 QGDGANAGQPQAQGDGANAGQPQAQGDGANAGQPQAQGDGANAGRPQAQGDGANAGQPQAQGDGANAGQP 191

192 QAQGDGANAGQPQAQGDRANAGQPQAQGDGANVPRQGRNGGGAPAGGNEGNKQAGKGQGQNNQGANAPNE 261

      Pk Th2R                 R2 motif          Pk Th3R
262 KVVNDYLHKIRSSVTTEWTPCSVTCGNGVRIRRKAHAGNKKAEDLTMDDLEVEACVMDKCAGIFNVVSNS 331

332 LGLVILLVLALFN 344
```

Supplementary file 3. MALDI-TOF result from the University of Malaya Proteomics Research Centre. The recombinant PkCSP was labelled as S6 in sample position B7. Results of protein summary indicated protein was significantly matched as a *Plasmodium knowlesi* protein via the SwissProt database.

UM/MBL/TM-1/F1A

MEDICAL BIOTECHNOLOGY LABORATORY
Block B, Level 4
Faculty of Medicine
University of Malaya
50603 Kuala Lumpur
Tel: 603-79677535
Email: proteomics@um.edu.my

UNIVERSITI MALAYA

STANDARDS MALAYSIA ACCREDITED LABORATORY
MS ISO/IEC 17025
TESTING
SAMM NO. 837

IN CONFIDENCE

Tan Jia Hui/Prof. Lau Yee Ling
Department of Parasitology
Faculty of Medicine
University of Malaya
50603 Kuala Lumpur

25/01/2021

TEST REPORT
Report: M2021/001

SAMPLES

Date of receipt: 05/1/2021
Number of samples: 6 samples
Source of sample: *Plasmodium knowlesi* (Gel plugs from SDS PAGE experiments)
Service required: Identification of proteins
Date of performance of the test: 22/1/2021

TEST METHOD: IN-GEL TRYPSIN DIGESTION OF PROTEINS FOR MASS SPECTROMETRY ANALYSIS (REFER NO.: SAMM 837)

The protein samples were trypsin digested and peptides extracted according to the in-house method developed based on Shevchenko *et al.* (Analytical Chemistry, 1996; Nature Protocols, 2007). Peptides were analysed by MALDI ToF/ToF 5800 mass spectrometer (Sciex, Framingham, USA). Spectra were analysed to identify protein of interest using Mascot search engine (Perkins *et al.*, Electrophoresis, 1999).

Database: *Plasmodium knowlesi* (Last update: January 2021, 16981 sequences)
Taxonomy: All Entries

The performance of the instrument was verified using trypsin-digested beta-galactosidase from *Escherichia coli* (Sciex, Framingham, USA). In-house prepared Coomassie-stained bovine serum albumin (5 µg/ml) gel plugs served as internal control.

RESULTS

Results including Mascot search results (MHTML file format) and MS/MS spectral data (t2d file format) are provided in a CD. The former can be viewed using any internet browsers whilst the latter, using Sciex Data Explorer software (available at MBL). The labels of the Mascot search results containing folders correspond to the position of the samples on the Opti-ToF MALDI plate as listed below.

No.	Sample Position	Sample Name	Results (Protein summary)
1	B3	S1	+
2	B4	S2	+
3	B6	S3	+
4	B7	S4	+
5	B9	S5	+
6	B10	S6	+

+ - Significant result(s).
NS - No significant result(s).

Note: Protein/peptide scores were derived from ions scores as a non-probabilistic basis for ranking protein hits ($p < 0.05$) indicates identity and extensive homology.

All results are stored on a secure and password protected server; data will be available until 2026.

NOTES ON INTERPRETING THE RESULTS

Database: SwissProt

UniProtKB/Swiss-Prot is a high quality, manually annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions. For further information please refer:

http://www.matrixscience.com/help/seq_db_setup.html

Viewing the results

To view the results, click on the respective sample-spot position corresponding folders found in the CD. Each folder contains a complete Mascot search results consisting of a protein and peptide summary report and its corresponding protein view. The **protein summary report** is intended for peptide mass fingerprint results whilst, **peptide summary report** provides the clearest and most complete picture of MS/MS search results, especially if the sample was a

protein mixture. If there are no peptide matches, only molecular weight matches, then only a Protein Summary report will be generated.

Ion scores is -10*Log(P), where P is the probability that the observed match between the experimental data and the database is a random event. Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits. This means protein scores are the sum of a series of peptide scores and this determines the ranking of protein hits.

Each peptide is fragmented within the mass spectrometer to produce ions that give amino acid sequence information. In each case, the peptide ion data is matched to possible amino acid sequences in the database. This data frequently lends itself to more than one sequence interpretation.

a) Data in **red** indicates that the protein hit ranked number one in the list of possible sequences (move the cursor over the number in the QUERY column to see the list).
b) **Black** indicates the protein hit ranked lower down the list of possible sequences. Clicking on the number in the QUERY column shows the MS/MS peptide spectra that matched the sequence.
c) When data appears in **bold (red)** this is the first time a peptide found in your dataset has been matched to a protein.
d) Peptides not in bold **(red)** are seen further down the list of hits, and show the peptide has already been matched to a protein at a higher level of significance.

e) The peptide matches not assigned to protein hits at the end of the report are sequences of low significance also contained within the sample.

   For further information please refer:
   http://www.matrixscience.com/help/msms_summaries_help.html

   In all cases the best results are achieved when two or more peptides map to the same protein. One matched peptide at high confidence is indicative. Search results are not absolute and matches near the significance threshold should be closely examined. You should be concerned if your molecular mass data does not support the Mascot hit, and a hit lower down the list but of the correct size may indicate a better match.

   The results shown are generated by automatic database searching. Where no significant hit is obtained this may indicate that there is insufficient protein concentration or the protein is not in the database. Analysis against an alternative database or further de novo peptide sequencing may be beneficial.

f) The search parameters used are as follow:

   - Fixed modification: **Carbamidomethyl (C)**
   - Variable modification: **Oxidation (M)**
   - Peptide tolerance: **100 ppm**
   - MS/MS tolerance: **0.2 Da**

- Peptide charge: **+1**
- Mass: **Monoisotopic**
- Enzyme: **Trypsin**
- Missed cleavage: **1**

g) Above parameters can be used to conduct new searches against other available databases.

In the search result file, locate the **Format As** button and change the option to **Protein Summary (deprecated)**. Click on **Format As** to update the page. The **Re-Search All** button will be seen at the updated page. Click on the **Re-Search All** button to display all available databases and enable new searches.

To re-search, select the respective database and taxonomy to the corresponding taxonomic group of the target organism. The above search parameters are recommended and will usually appear as the default parameters when the re-search option is selected. Note that, search results might differ if the search parameters are altered.

Prepared by,	Approved by,
Nur Athirah Hussin	Dr. Jaime Jacqueline J.
Technical Staff	Approved Signatory
25/01/2021	25/01/2021

[THE REMAINDER OF THE PAGE IS INTENTIONALLY LEFT BLANK]