Gene: X 4 (2019) 100024

Contents lists available at ScienceDirect

Gene: X

journal homepage: www.journals.elsevier.com/gene-x

Research paper

Putative circumsporozoite protein (CSP) of *Plasmodium vivax* is considerably distinct from the well-known CSP and plays a role in the protein ubiquitination pathway

Manoswini Dash[a], Veena Pande[b], Abhinav Sinha[a,*]

[a] Division of Epidemiology and Clinical Research, ICMR-National Institute of Malaria Research, New Delhi, India
[b] Department of Biotechnology, Bhimtal Campus, Kumaun University, Nainital, Uttarakhand, India

ARTICLE INFO

Keywords:
Malaria
Plasmodium vivax
Circumsporozoite protein
Tandem repeat
Computational analysis
Hypothetical protein

ABSTRACT

Amidst technical challenges which limit successful culture and genetic manipulation of *P. vivax* parasites, we used a computational approach to identify a critical target with evolutionary significance. The putative circumsporozoite protein on chromosome 13 of *P. vivax* (*Pv*puCSP)is distinct from the well-known vaccine candidate *Pf*CSP. The aim of this study was to understand the role of *Pv*puCSP and its relatedness to the well-known CSP. The study revealed *Pv*puCSP as a membrane bound E3 ubiquitin ligase involved in ubiquitination. It has a species-specific tetra-peptide unit which is differentially repeated in various *P. vivax* strains. The *Pv*puCSP is different from CSP in terms of stage-specific expression and function. Since E3 ubiquitin ligases are known antimalarial drug targets targeting the proteasome pathway, *Pv*puCSP, with evolutionary connotation and a key role in orchestrating protein degradation in *P. vivax*, can be explored as a potential drug target.

1. Introduction

In the era, when malaria has been successfully eliminated from many countries and some are on the verge of elimination, it is still a health concern for countries of the tropical and sub-tropical regions. Global strategies and intervention policies have brought down the cases of malaria and associated mortality by 41% and 62%, respectively, in the past fifteen years leading up to 2015 (World Health Organization (WHO), 2016). Nevertheless, high number of cases (219 million) and deaths (435,000) are still estimated in the annual World Malaria Report for 2018 (World Health Organization (WHO), 2018). Malaria biology is considered to be very intricate because of the parasite's complex life-cycle where it dwells between female *Anopheles* mosquito vectors and a gamut of hosts including rodents, aves, and primates. Two apicomplexan parasites namely *Plasmodium falciparum* and *Plasmodium vivax* are the major causative agents of human malaria, accountable for the global malaria burden. *P. falciparum* is the more lethal species while *P. vivax* believed to be non-lethal, was neglected for long (Baird, 2007; Bassat and Alonso, 2011; Tham et al., 2017). However, the assumption of the benign nature of *P. vivax* has progressively changed, with increasing cases of severity and death by *P. vivax* malaria reported globally (Aashish and Manigandan, 2015; Genton et al., 2008; Geleta and Ketema, 2016; Douglas et al., 2012; Tjitra et al., 2008). Besides the 'not-so-fatal' assumption, lack of accurate diagnosis, dormant liver stage formation, early transmission and technical difficulties for the continuous in vitro culture have also generated hindrances in *P. vivax* research (Price et al., 2007; Baird et al., 2012). Primaquine is the only licensed drug against dormant liver stage parasites, but it is contra-indicated in G6PD[1] deficient malaria patients due to chances of hemolysis, thereby limiting its extensive use (Cappellini and Fiorelli, 2008). Tafenoquine is an alternative single dose medicine that was approved for the radical cure of *P. vivax* malaria by US Food and Drug Administration in 2018 and might offer a better and more compliant radical treatment. Moreover, the rate at which *P. falciparum* is developing resistance against the current first-line treatment (Artemisinin Combination Therapy), it will not be incorrect to assume that the similar situation can also arise in *P. vivax* though resistance in *P. vivax* against ACT[2] has not been reported yet (World Health Organization (WHO), 2018). It is, therefore, crucial to devise explicit prevention, treatment and control measures for *P. vivax*, along with *P. falciparum*, to effectively accomplish the goal of malaria elimination (Lover et al., 2018).

* Corresponding author.
E-mail address: abhinav@mrcindia.org (A. Sinha).
[1] Glucose 6-phosphate dehydrogenase.
[2] Artemisinin Combination Therapy.

https://doi.org/10.1016/j.gene.2019.100024
Received 1 July 2019; Received in revised form 21 October 2019; Accepted 7 November 2019
Available online 12 November 2019

There is no efficient immunization available against malaria till date, especially in *P. vivax* with very limited numbers of candidates studied in-depth (Tham et al., 2017). In *P. falciparum*, partial success has been achieved with the development of vaccine RTS, S targeting circumsporozoite protein (CSP), a sporozoite membrane protein and was launched in 2019 (Gordon et al., 1995). However, the *P. vivax* homologue, *Pv*CSP could not induce sterile protection despite significantly delaying the infection (Bennett et al., 2016; De Camargo et al., 2018). It is therefore important to identify novel and unique candidates to be used as a critical drug or vaccine target to combat the parasite.

This study focused on the putative circumsporozoite protein (puCSP; PVX_086150), which although annotated as circumsporozoite protein, lacks the functional thrombospondin domain. The gene is present on chromosome 13 and is hypothesized to be under selection pressure based on the hitchhiking model of molecular evolution (Gupta et al., 2010; Gupta et al., 2012). Since PVX_086150 holds evolutionary significance and shares annotation with well-known CSP (PVX_119355), it is assumed that the protein might have significant role in *P. vivax* biology. However, there is no experimental evidence available to date to confirm its function. Since half of the protein coding genes in the *P. vivax* genome are unannotated, lack of knowledge about these proteins could be one of the reasons why *P. vivax* biology is still not well understood. In the absence of direct experimental evidence, computational methods are the best measures to annotate the uncharacterized genes by utilizing the knowledge from orthologs and the resources that are generated from other studies and available in public repositories. We present a comprehensive and systematic computational study to annotate the *Pv*puCSP to understand its role in *P. vivax* and detangle the *Pv*puCSP from the well-known CSP of *P. vivax* using an *in-silico* framework.

2. Results

2.1. PvpuCSP have transmembrane domain and RDNA as repeat unit towards C-terminal

The *Pv*puCSP is a 2.2 kb single exon gene, encodes a 739 amino acid long peptide and belongs to PA-TM-RING family. The PA-TM-RING family includes transmembrane E3 ubiquitin ligases and is characterized by an N-terminal transient signal peptide, a Protease Associated (PA) domain, a transmembrane domain and a C-terminal $C_3H_2C_3$ type RING H2 finger domain (cd16454). *Pv*puCSP carries five pass transmembrane domain (236–258 aa, 277–299 aa, 316–333 aa, 338–355 aa, and 362–379 aa) and a C-terminal $C_3H_2C_3$ type RING zinc finger domain (689–733 aa), whereas N-terminal signal peptide and PA domain was not found (Fig. 1A). The 44 residues long RING zinc finger domain is the functional domain in *Pv*puCSP, with characteristic $C_3H_2C_3$ arrangement (Cys-X_2-Cys-X_{14}-Cys-X_3-His-X_2-His-X_2-Cys-X_{10}-Cys-X_2-Cys). *Pv*puCSP also possesses a transcription termination factor-Rho (420–616 aa) (cl28310) in the central region.

A scan of *Pv*puCSP for repeats identified a 17 tandem repeat of R[D/C]NA (Arg-Asp/Cys-Asn-Ala) sequence towards the C-terminal (546–613 aa). Multiple sequence alignment of *Pv*puCSP and its *Plasmodium* orthologs revealed that the repeat is very much unique to *P. vivax*, and is not present in other *Plasmodium* species except *P. knowlesi*, in which the repeating unit is present only once (Fig. 1B-i). Furthermore, comparison of the repeat region among *P. vivax* strains displays variation in the number of repeats (Fig. 1B-ii).

The topology, orientation and sub-cellular localization of a protein governs its cellular function. *Pv*puCSP is located in association with plasma membrane (GO: 0005886) and contains transmembrane domains. Orientation of N- and C-terminals of *Pv*puCSP with respect to the plasma membrane was predicted using topology prediction servers (Table 1). A total of 12 servers were used to predict the *Pv*puCSP topology, of which 50% (6 out of 12) of servers (OCTOPUS, PHILLIUS, SPOCTOPUS, S-TMHMM, TOPCONS, and TOPPRED) predicted four transmembrane domains and both N- and C-terminals as cytoplasmic while rest of the servers predicted 5 transmembrane domains in *Pv*puCSP except MEMSAT which predicted *Pv*puCSP as single-pass transmembrane protein. Four prediction servers (MEMSAT, PHOBIUS, POLYPHOBIUS, and SCAMPI) predicted N-terminal exposed to outside while the other two (HMMTOP and TMHMM) predicted C-terminal outside.

2.2. Orthologs of PvpuCSP present across eukaryotes

DELTA-BLAST search against the Apicomplexan Refseq_protein database using *Pv*puCSP as a query retrieved 280 BLAST hits. Most of the retrieved hits were from *Plasmodium* species and are annotated as putative circumsporozoite protein or RING zinc finger protein or E3 ubiquitin ligase or conserved hypothetical protein with unknown function. The likelihood tree generated using the closest orthologs revealed *P. inui*, a simian parasite, as the closest to *Pv*puCSP, followed by *P. knowlesi* among other *Plasmodium* species (Fig. 2).

The phylogenetic tree was constructed using orthologous proteins across species, sharing similar domain architecture (Supplementary Table S1) to *Pv*puCSP to find out the relatedness and to ascertain the function. Orthologs of *Pv*puCSP were found in other protozoans, algae, fungi, plants and higher organisms (Fig. 3). Other than *Plasmodium*, *Cryptosporidium* and *Toxoplasma* were found to be close homologs among Apicomplexans. In higher eukaryotes, the orthologs are characterized as an E3 ubiquitin ligases, while in lower eukaryotes, including the apicomplexans (around 70% of the total orthologs retrieved), majority are unannotated (Supplementary Table S1). The ortholog of *Pv*puCSP is annotated as circumsporozoite protein only in *P. coatneyi*, *P. cynomolgi*, and *P. knowlesi* while in other apicomplexans, it is named as RING zinc finger protein based on its functional domain. The constructed phylogeny revealed conserved functional domains in diverse groups. However, the annotated proteins form a clade with substantial phylogenetic distance from *Pv*puCSP and the proteins sharing a similar clade with *Pv*puCSP are mostly hypothetical or putative.

*Pv*puCSP is found to be a single copy gene from the self BLASTn search with an expectation value of $\geq 1E-10$. However, *Pv*puCSP also has paralogs in *P. vivax*, present on different chromosomes (PVX_111310, PVX_094410, PVX_113525, PVX_119830, and PVX_079770). From eggNOG search, a total of 6623 proteins in 237 species were retrieved as orthologs of *Pv*puCSP across species, which were predicted to be involved in protein metabolic processes like protein ubiquitination and also possess metal and/or ion binding activity.

2.3. Moderate expression of PvpuCSP across all parasitic stages

Examination of published transcriptome data from two distinct *P. vivax* isolates during the intra-erythrocytic developmental cycle (IDC) revealed moderate expression with minimum transcriptional changes of *Pv*puCSP, with an average Log_2 Fragments Per Kilobase of transcript per Million mapped reads (FPKM) value of 5.5, across the blood stages (Zhu et al., 2016). However, *Pv*CSP (PVX_119355) shows significant transcriptional changes though less expressed as compared to *Pv*puCSP across different time points in the erythrocytic stage (Fig. 4A-i).

The transcriptome data generated by Westenberger and group revealed that *Pv*puCSP (PVX_086150) was expressed with a fold change (FC) value 2.8 and pANOVA < 0.05 (Westenberger et al., 2010). Expression of *Pv*puCSP was marked in all stages (sporozoite, blood stage, gametocytes and oocyst) of parasite along with the sporozoite stage with average expression value of 116. However, the well-known CSP (PVX_119355) was expressed with the FC value of 144.14, with highest expression in the sporozoite stage (13,188.85). Comparison of the stage-specific expression of two CSPs showed that *Pv*puCSP had moderate expression across all parasitic stages while the well-known CSP

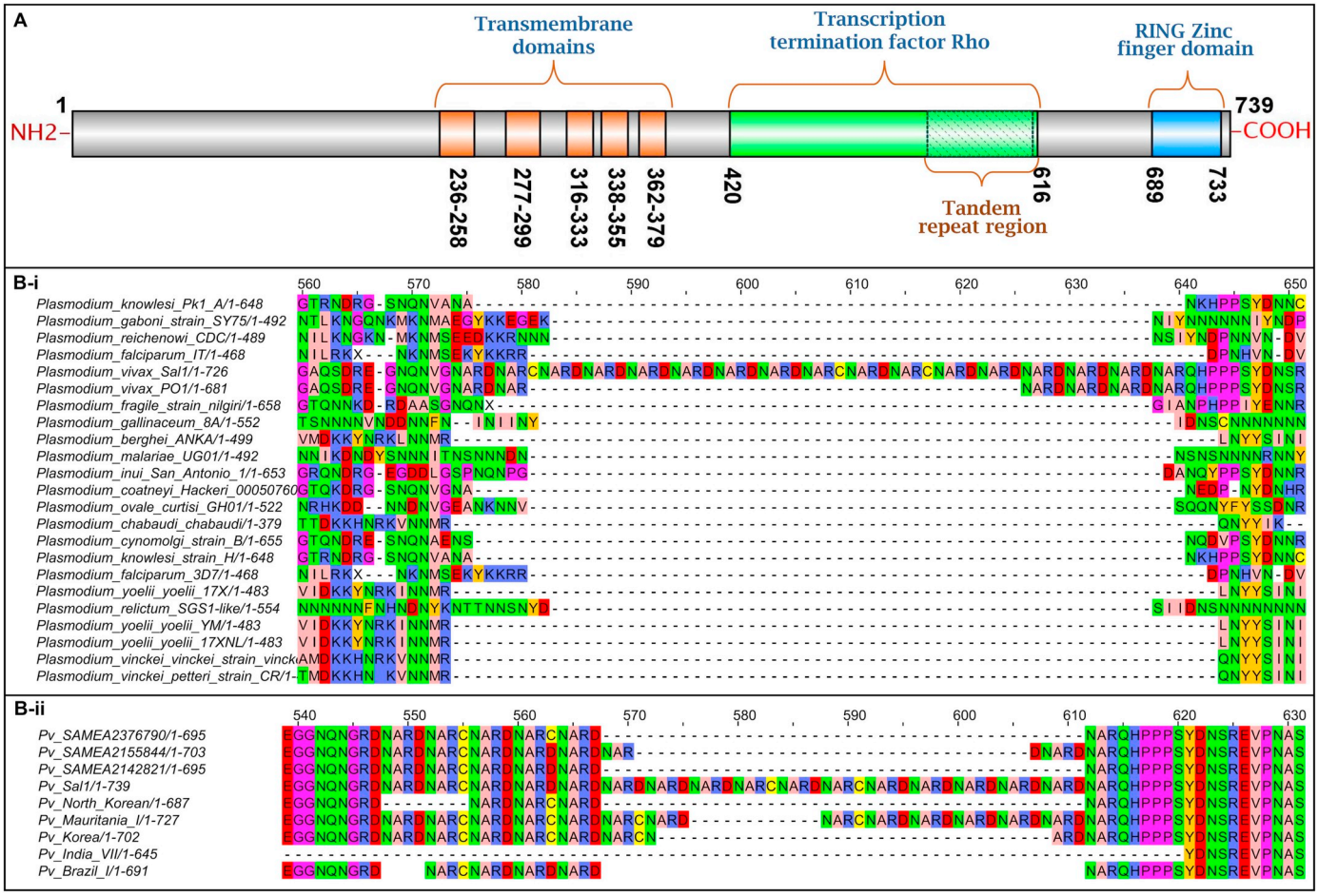

Fig. 1. Schematic diagram of *Pv*puCSP and multiple sequence alignment of *Pv*puCSP orthologs. (A) Transmembrane domains (orange), transcription termination factor Rho (green) and RING zinc finger domain (blue) are marked according to their corresponding amino acid position in *Pv*puCSP. The repeat region (marked with dotted diagonal lines) is predicted to lie within the transcription termination factor Rho. The above figure was prepared using DOG (Domain Graph) version 1.0 (Ren et al., 2009; Liu et al., 2015). (B-i) Multiple sequence alignment of *Pv*puCSP and its *Plasmodium* orthologs showing the R[D/C]NA repeat region. (B-ii) Multiple sequence alignment of *Pv*puCSP from different strains of *P. vivax* available, showing variation in the number of repeats. Only repeat region of the multiple sequence alignment has been shown in the above figure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
List of tools available in public resources (and their references) used to predict in and out the orientation of N- and C-terminals, number of transmembrane domains and their corresponding positions in *Pv*puCSP are mentioned.

Transmembrane prediction servers and references	Orientation of N-terminal	Orientation of C-terminal	Number of predicted TM domains	The position of TM domains with respect to *Pv*puCSP aa position				
				TM1	TM2	TM3	TM4	TM5
HMMTOP	Inside	Outside	5	239–256	277–299	316–333	338–355	362–379
MEMSAT	Outside	Inside	1	237–258	–	–	–	–
OCTOPUS	Inside	Inside	4	237–257	274–294	311–331	345–375	–
PHILLIUS	Inside	Inside	4	237–258	273–295	314–333	351–378	–
PHOBIUS	Outside	Inside	5	240–261	273–294	314–333	340–357	363–379
POLYPHOBIUS	Outside	Inside	5	239–261	273–295	314–333	338–357	362–379
SCAMPI	Outside	Inside	5	238–258	274–294	314–334	337–357	360–380
SPOCTOPUS	Inside	Inside	4	237–257	274–294	311–331	345–375	–
S-TMHMM	Inside	Inside	4	237–257	273–293	313–334	351–378	–
TMHMM	Inside	Outside	5	236–258	273–295	316–333	338–357	362–379
TOPCONS	Inside	Inside	4	237–257	274–294	313–333	351–371	–
TOPPRED	Inside	Inside	4	237–257	273–293	314–334	351–371	–

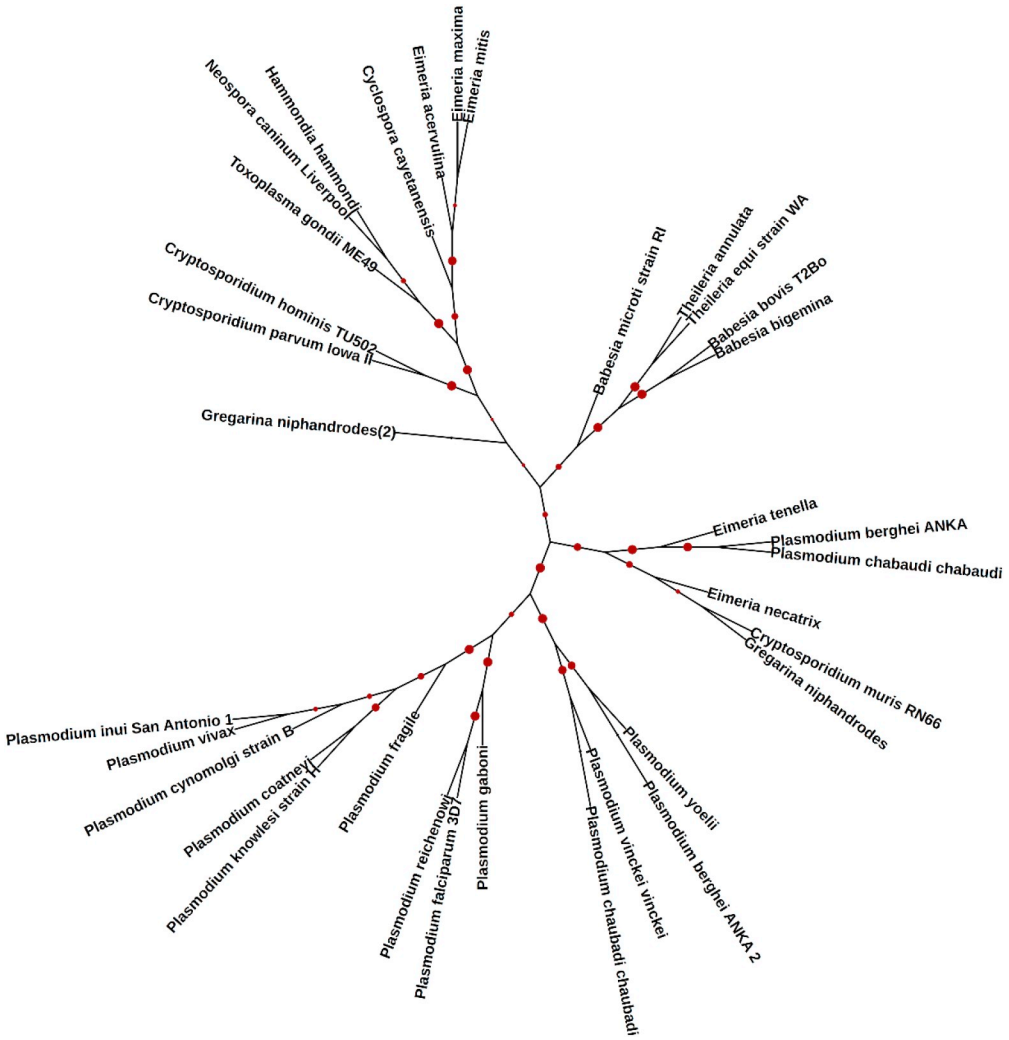

Fig. 2. Reconstruction of phylogeny using *Pv*puCSP orthologs sharing similar domain. The orthologous proteins obtained from DELTA-BLAST using *Pv*puCSP (with > 35% identity) as query inferred using maximum likelihood method. Evolutionary distances were computed using the JTT method, with bootstrap support at each join based on 1000 replicates. The size of the circles on the branch represents the bootstrap value.

had significant expression in sporozoite stage. Change in expression was > 100 times more in the case of well-known CSP from the sporozoite to erythrocytic stage while in *Pv*puCSP, change in expression was only about two-fold. Analysis of two other known stage-specific proteins (AMA1 and *Pv*s25) also suggested that the *Pv*puCSP is expressed almost at a similar rate in every stage and not confined to any particular parasite stage (Fig. 4A-ii; Supplementary Table S2). Analysis of hypnozoite transcriptome showed that *Pv*puCSP is also getting expressed during the dormant stage with average Transcripts Per Kilobase Million (TPM) 24, following a similar pattern (Gural et al., 2018). The above array of expression of *Pv*puCSP in almost all parasitic stages at a moderate rate further corroborates that it might be associated with some function crucial to the parasite across all stages. Analysis of transcriptome data from other *Plasmodium* species didn't provide any significant evidence about *Pv*puCSP.

2.4. PvpuCSP is a ubiquitin ligase, co-expressed with proteins involved in the protein ubiquitination pathway

A PPI network was developed using the DEGs[3] to find out the proteins that are co-expressed with *Pv*puCSP (Dharia et al., 2010). A total of 4326 genes were selected with FC ≥ 2 and p-value ≤ 0.05 from the transcriptome (Westenberger et al., 2010) (Supplementary Table S3). The PPI network was generated using the above DEGs based on the interaction data available in UniProt database. It resulted in 4091 nodes and 78,299 edges, where each node represented DEG and edges as their connected proteins (Supplementary Table S3). The data showed that *Pv*puCSP has 37 degrees i.e. it interacts with 37 different proteins (Table 2). Nearly half of the proteins that showed interaction with *Pv*puCSP, were ubiquitin-conjugated enzymes or ubiquitin-like proteins, which suggests that *Pv*puCSP might play a role in the ubiquitination process. Other proteins showing interaction with *Pv*puCSP were mostly hypothetical, while few of them were serine proteases and DNA-dependent RNA polymerases. Interestingly, more than half of the proteins interacting with *Pv*puCSP were hub genes with the highest degree (Table 2). Five out of the top 10 hub genes are showing interaction with the *Pv*puCSP, with a closeness centrality value ≥0.5. The network is strongly connected as the closeness centrality measure ranges from 0.5 to 1 in all the connected nodes. Interaction of *Pv*puCSP with the hub genes showcases its importance (Fig. 4B). The Gene Ontology analysis of the above set of proteins revealed them as ubiquitin conjugative enzyme, ATP binding, and zinc ion binding and mostly shown to be involved in ubiquitin-mediated proteolysis (Supplementary Table S3).

[3] Differentially expressed genes.

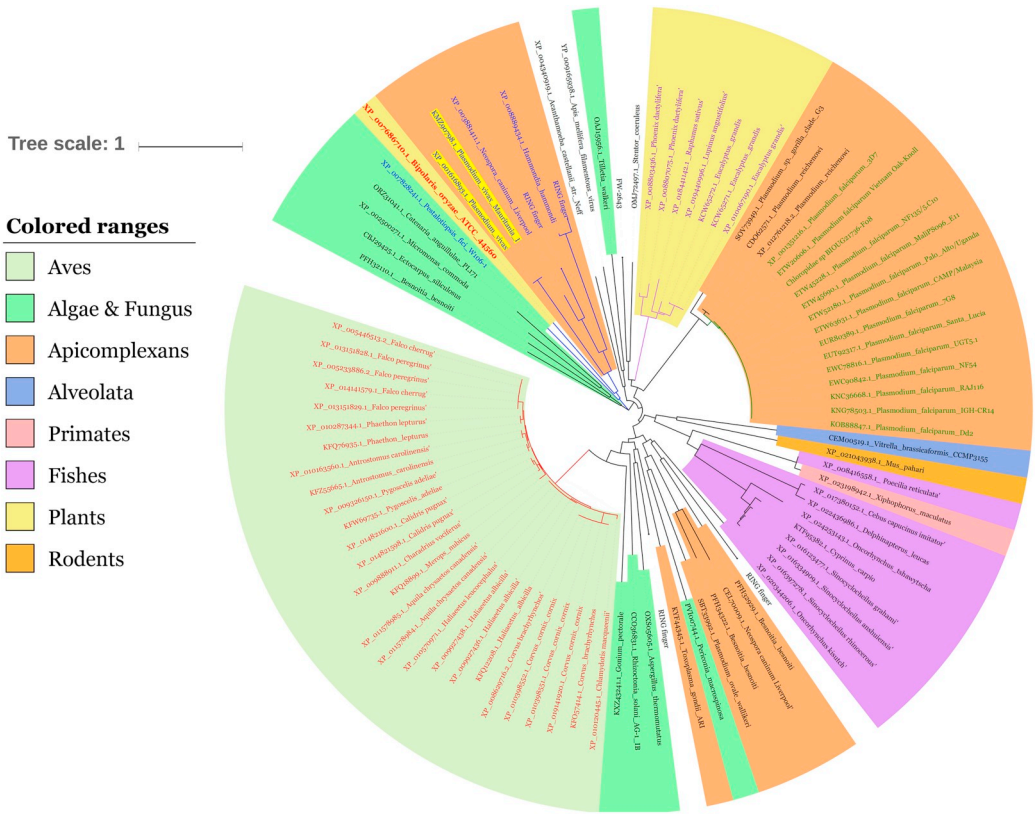

Fig. 3. Phylogenetic tree of *Pv*puCSP orthologs across species with similar domain architecture (Rho and RING-H2_PA-TM-RING domain-containing protein). The phylogeny was re-constructed using the maximum like-lihood method. The tree was designed and color coded using a web-based tool, Interactive Tree Of Life (iTOL https://itol.embl.de/) (Letunic and Bork, 2016).

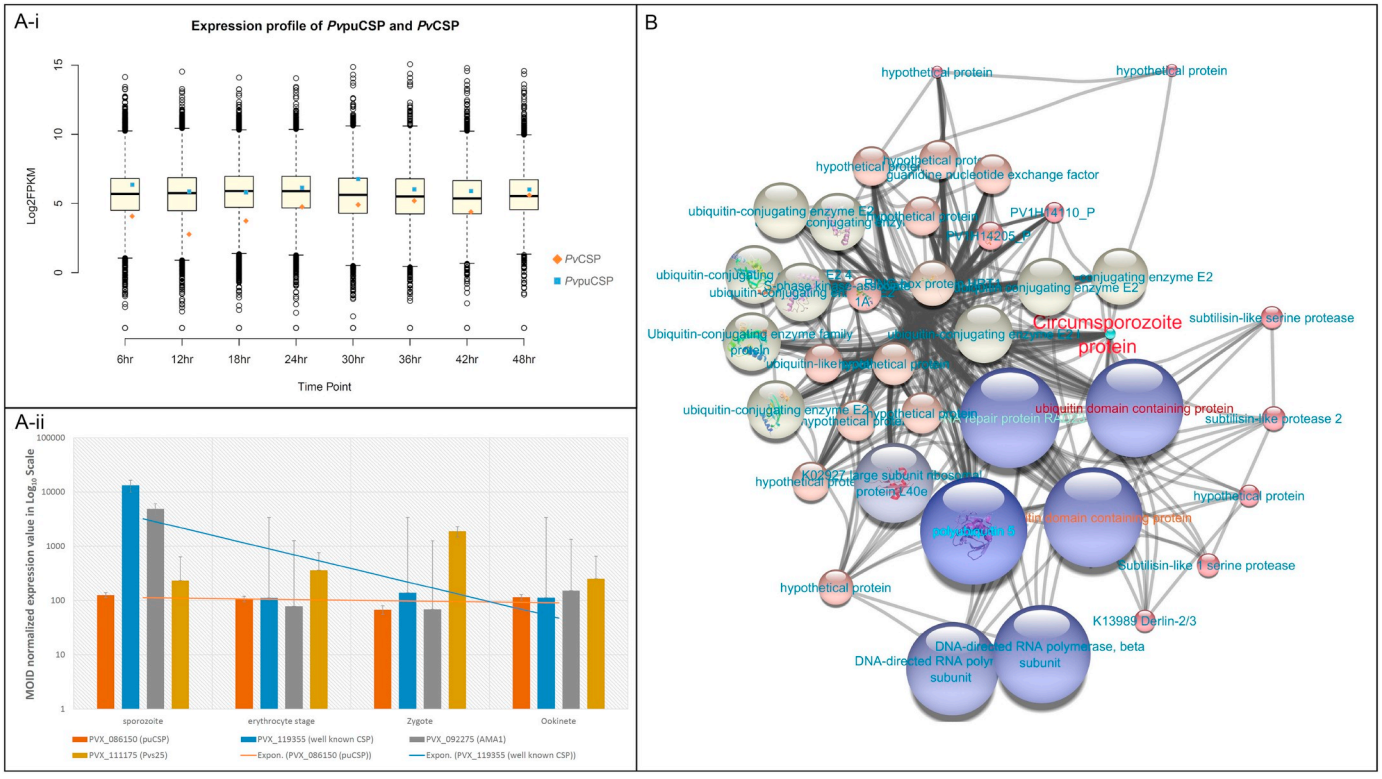

Fig. 4. Expression profiling and PPI network of *Pv*puCSP (A-i) The Log_2 FPKM value of *Pv*puCSP (PVX_086150) and well-known CSP (PVX_119355) was plotted at 6 time points during the intra-erythrocytic life cycle. The dots in the box plot show the exact Log_2FPKM value of each protein at a different time point of erythrocytic life cycle. (A-ii) MOID normalized expression value of known stage-specific proteins (well-known CSP for sporozoite stage; AMA1 for erythrocytic stage and Pvs25 for gametocyte stage) were plotted for all parasitic stages (in human and mosquitoes). The trend lines show the exponential change in expression value of *Pv*puCSP and well-known CSP. (B) PPI network of *Pv*puCSP and interacting DEGS, where size and color represent the degree of the node (bigger size and dark color to the nodes with a higher degree). The expression data generated by Westenberger et al. was used for the above network.

Table 2
List of DEGs co-expressed with *Pv*puCSP (PVX_086150) obtained from the STRING enrichment analysis in Cytoscape, using expression data generated by Westenberger group.

PlasmoDB ID	UniProt ID	Description of the protein	Degree	GO term	GO description
*PVX*_084620	A5K0L7	Polyubiquitin 5	549	GO:PM16005087	Polyubiquitin 5, putative
*PVX*_094805	A5K7E2	DNA repair protein RAD23	505	GO:0020002	Host cell plasma membrane
*PVX*_122475	A5JZB9	Ubiquitin domain containing protein	499	GO:0008104	Protein localization
*PVX*_091315	A5K4E9	Ubiquitin domain containing protein	491	GO:0006511	Ubiquitin-dependent catabolic process
*PVX*_092345	A5K506	DNA-directed RNA polymerase, beta subunit	464	GO:GNF0218	Merozoite development:
*PVX*_084365	A5K0G6	DNA-directed RNA polymerase III subunit	433	GO:0043064	DNA-directed RNA polymerase III subunit, putative
*PVX*_115255	A5K2R4	K02927 large subunit ribosomal protein L40e	370	GO:0003735	Structural constituent of ribosome
*PVX*_113480	A5K1R0	Ubiquitin-conjugating enzyme E2 I	232	GO:0004148	Dihydrolipoyl dehydrogenase activity
*PVX*_123140	A5JZQ3	Ubiquitin-conjugating enzyme E2	232	GO:0009308	Amine metabolic process
*PVX*_084235	A5K0E0	Ubiquitin-conjugating enzyme E2 4	232	GO:0015935	Ubiquitin-conjugating enzyme E2 4, putative
*PVX*_079885	A5K986	Ubiquitin-conjugating enzyme E2	232	NA	NA
*PVX*_114795	A5K2H0	Ubiquitin-conjugating enzyme family protein	232	NA	NA
*PVX*_099465	A5K6S4	Ubiquitin conjugating enzyme E2	228	GO:0003735	Structural constituent of ribosome
*PVX*_099185	A5K6L9	Ubiquitin conjugating enzyme	225	GO:0009308	Amine metabolic process
*PVX*_083175	A5K893	Ubiquitin-conjugating enzyme E2	216	NA	NA
*PVX*_085805	A5K1A0	Ubiquitin-conjugating enzyme E2	213	GO:0015935	Ubiquitin-conjugating enzyme E2, putative
*PVX*_095280	A5K7N6	RING-box protein HRT1	170	GO:PM16005087	NA
*PVX*_087775	A5KA52	Hypothetical protein	158	NA	NA
*PVX*_089055	A5K5J0	Hypothetical protein	151	GO:0006261	DNA-dependent DNA replication
*PVX*_114510	A5K2B3	Hypothetical protein	147	GO:0004221	Ubiquitin thiolesterase activity
*PVX*_091605	A5K4K7	Hypothetical protein	146	GO:0045047	Protein targeting to ER
*PVX*_097850	A5KAY6	Ubiquitin-like protein	142	GO:CCYCL09	NA
*PVX*_097985	A5KB13	Guanidine nucleotide exchange factor	141	GO:0004540	Ribonuclease activity
*PVX*_088110	A5KAC0	Hypothetical protein	140	NA	NA
*PVX*_092370	A5K511	Hypothetical protein	139	NA	NA
*PVX*_114810	A5K2H3	Hypothetical protein	139	NA	NA
*PVX*_115310	A5K2S5	S-phase kinase-associated protein 1A	124	GO:0044085	Cellular component biogenesis
*PVX*_101110	A5K8U5	Hypothetical protein	114	NA	NA
*PVX*_119410	A5KB74	*PV*1H14205_P	93	GO:PM15774020	NA
*PVX*_119315	A5KB55	*PV*1H14110_P	62	NA	NA
*PVX*_092460	A5K529	Subtilisin-like protease 2	61	GO:0031124	mRNA 3′-end processing
*PVX*_097935	A5KB03	Subtilisin-like 1 serine protease	57	GO:PM15032632	NA
*PVX*_097920	A5KB00	Subtilisin-like serine protease	53	GO:GNF0218	Merozoite development
*PVX*_117040	A5K372	K13989 Derlin-2/3	51	GO:CCYCL09	NA
*PVX*_111100	A5KDX2	Hypothetical protein	50	GO:CCYCL01	NA
*PVX*_086150	A5K1G9	Circumsporozoite protein	37	GO:0009308	Amine metabolic process
*PVX*_094820	A5K7E5	Hypothetical protein	16	NA	NA
*PVX*_081725	A5KA05	Hypothetical protein	8	NA	NA

GO, Gene Ontology; NA, not available.

2.5. Tertiary PvpuCSP model possesses long IDRs and characteristic cross-brace zinc finger motif

The PDB search against *Pv*puCSP to obtain template(s) for homology modeling, retrieved hits with maximum 48% identity, 5% query coverage and the hits were only confined to the C-terminal of the protein. Therefore, the templates were not found suitable for homology modeling and thus fold recognition method was implemented to generate a tertiary model. The validation study confirmed the model generated by I-TASSER to have better quality and was considered for further analysis (Table 3). The final model after loop refinement and energy minimization had the DOPE score changed from −3899.25415 to −61,670.992188.

The tertiary model of *Pv*puCSP shows that majority of the regions were alpha helices, followed by the coil and a beta sheet. The predicted transmembrane domains fold into alpha helices, which further supports that *Pv*puCSP is an integral membrane protein (Fig. 5B). The zinc RING finger domain (689–733 amino acids) presented with its characteristic alpha helix, small beta sheet, and variable loop length, binding with zinc metal ion with $C_3H_2C_3$ arrangement (Fig. 5C). Some portion of

Table 3
Model validation report of *Pv*puCSP models generated by homology modeling and fold recognition methods.

Servers	Parameters checked	I-TASSER	Orion	SPARK-X	Modeller
Verify 3D	Averaged 3D-1D score > 0.2	65.22%	48.57%	59.13%	42.22%
PROCHECK	Residues in most favoured region	62.2%	78.9%	76.9%	61.9%
	Additionally allowed region	30.3%	16.7%	17%	26.8%
	Generously allowed region	5.2%	2.3%	3.5%	6.6%
	Disallowed region	2.3%	2.2%	2.6%	4.7%
ERRAT	Overall quality factor	84.583	21.216	11.44	22.8532
ProSA	Z score	−0.0369	1.5	−0.07	−0.02
Modeller	DOPE score	−58,795.65234	−43,805.5391	−46,093.4375	−36,789.57422
	GA-341 score	1	0.041819	0.961289	0.07887

*Pv*puCSP seemed to be inadequately modeled, which can be evidenced from the *Pv*puCSP model where a region was showing thread-like structure. These regions might be IDRs as *Pv*puCSP is predicted to have 57% of IDRs by DISOPRED3. Confirmation of the presence of transmembrane domains and RING domain towards the C terminal from the tertiary structure adds to the fact that *Pv*puCSP might be a membrane-bound E3 ubiquitin ligase.

A detailed analysis of *Pv*puCSP with respect to existence of IDRs was done through IDR profiling. The per-residue disorder propensity analysis of *Pv*puCSP revealed enrichment of disorder-promoting residues (49.12%) followed by order-promoting and neutral residues, 36.53% and 14.61%, respectively (Supplementary Table S4). Composition profiling of *Pv*puCSP with a set of naturally abundant proteins (SwissProt 51) revealed that *Pv*puCSP is significantly enriched with order promoting residues Cys (C) and Asn (N), but depleted with Phe (F), Val (V) and Leu (L). The disorder promoting residues like Arg (R), Glu (E), Ser (S), Gly (G) and Ala (A) are enriched in PvpuCSP while Pro (P) and Gln (Q) are depleted (Fig. 6A). Depletion of amino acids with certain physico-chemical properties, like hydrophobicity, bulkiness, and propensity to form linker and beta structure, was observed while amino acids that are more frequent in coils were enriched.

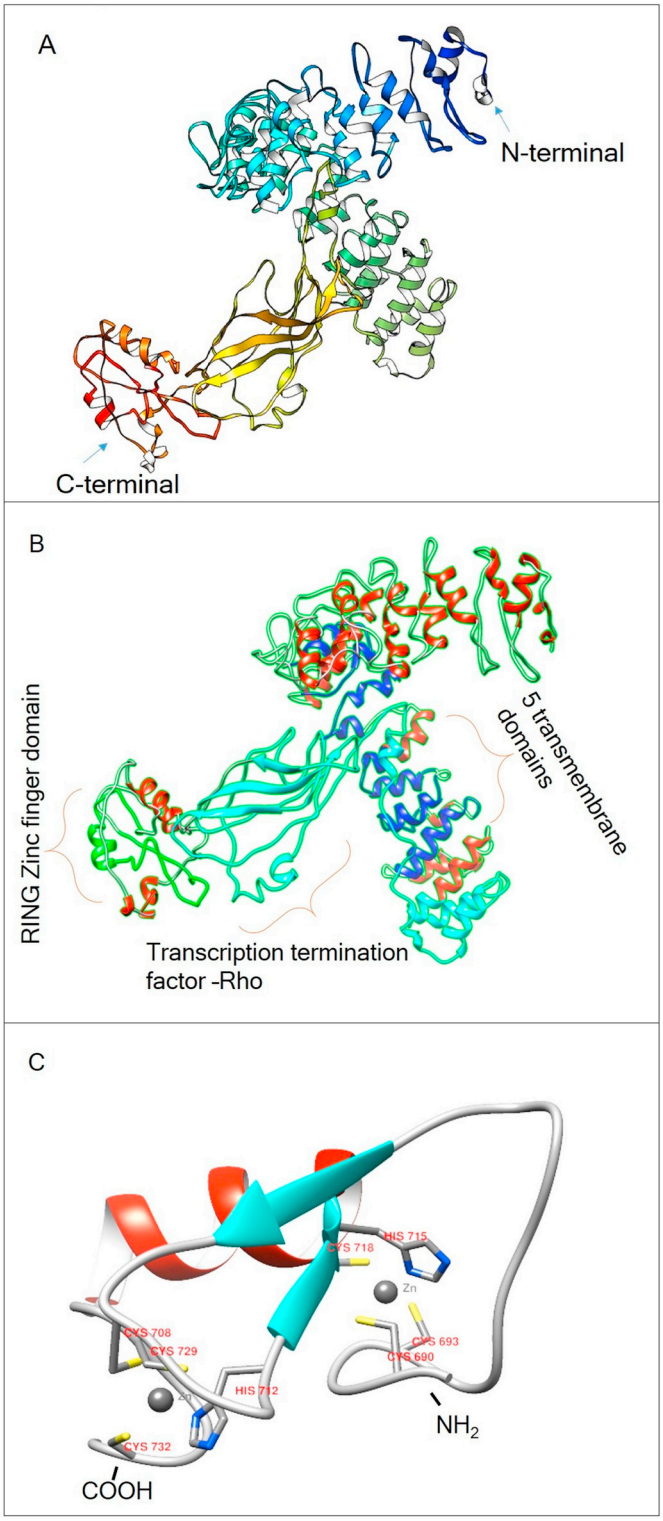

Fig. 5. Cartoon representation of the model of *Pv*puCSP. (A) The backbone of the model of *Pv*puCSP is rainbow colored from blue to red, from the N- to C-terminal. (B) Domain architecture of *Pv*puCSP showing five transmembrane domains (236–258, 277–299, 316–333, 338–355 and 362–379 aa), transcription termination factor–Rho (420–616 aa) and RING (Really Interesting New Gene) zinc finger domain (689–733 aa) marked in blue, cyan blue and green respectively. (C) Homology model RING zinc finger domain of *Pv*puCSP showing characteristics alpha helix, small beta strands, and variable length loops. Zinc atoms are shown as silver sphere and residues interacting with zinc atoms are shown with their corresponding amino acid position. It has $C_3H_2C_3$ motif to bind with two zinc ions.
Note: The predicted tertiary model carries two long IDRs, tertiary conformation of which couldn't be accurately predicted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Prediction of disordered regions using PONDR algorithms revealed a substantial portion of *Pv*puCSP as intrinsically disordered. PONDR VL-XT predicts 382 residues (51.69%) as disordered and 12 disordered regions with two longest regions of 90 (133–222) and 62 (550–612) residues at N and C-terminal, respectively. Consensus from all the IDR predictors used revealed that there are two major IDRs, one at the N-terminal (1–220) and the other towards C-terminal (420–680) (Fig. 6B). Scanning of *Pv*puCSP for MoRFs within IDRs using ANCHOR tool exposed 8 MoRFs (1–51, 56–80, 108–205, 446–453, 459–471, 503–522, 599–633 and 640–655) (Fig. 6B).

The charge-hydropathy plot suggests *Pv*puCSP as an ordered protein as a whole (Fig. 6C). However, charge-hydropathy plot of the predicted disordered region reveals its intrinsic disordered nature. The CDF plot revealed *Pv*puCSP as a disordered protein (Fig. 6D).

3. Discussion

Research on *P. vivax*, one of the two major parasites that contribute to the global malaria burden has not received the same attention as *P. falciparum*. This might be due to the fatal nature of *P. falciparum*, which kills the host if not treated. On the contrary, as *P. vivax* is believed to be more ancient than *P. falciparum* it has probably learned to survive and grow inside the host, without being noticed by the immune system (Das, 2015). *P. vivax* maintains itself at low parasitemia, sometimes by forming an inactive form (hypnozoites) without killing the host. This nature of *P. vivax*, being hidden from the host immune system provides a survival advantage and might serve as a reservoir for malaria, thus negatively influencing the malaria elimination goal. Recent reports of severity by *P. vivax* from different parts of the globe have begun to change our outlook towards the infectious disease and research interest has also increasingly been focused on the hitherto neglected parasite. For the improvement of available treatment and prevention tools, a number of targets (defined and novel), need to be explored to identify suitable and critical leads. Even though many proteins have been identified and proved to have biological significance in the *P. vivax* life cycle from past genomic, transcriptomic and proteomic studies, most of them are hypothetical or putative. Therefore, annotating more and

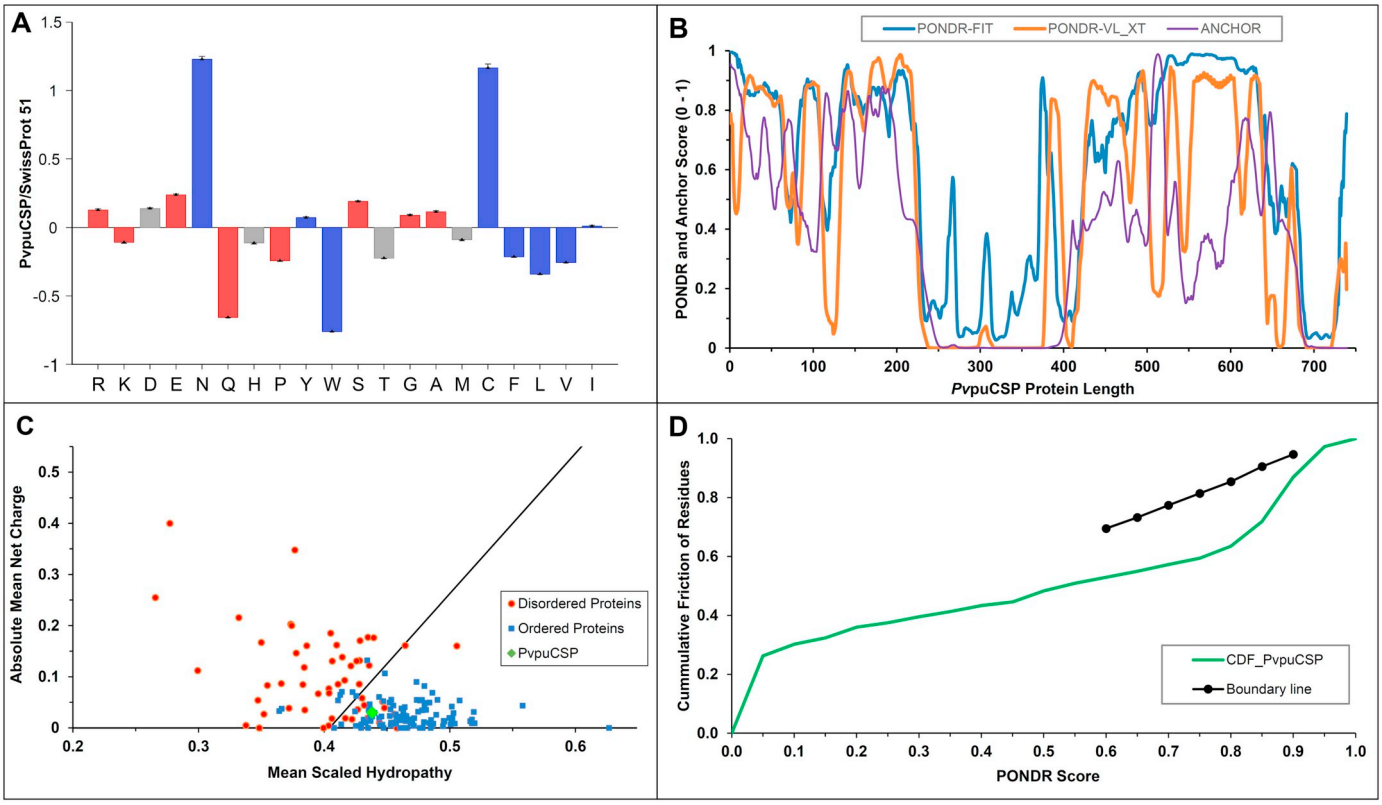

Fig. 6. (A) Composition profiling analysis of *Pv*puCSP by comparing against SwissProt 51 as reference data set. Amino acids are arranged according to their hydrophobicity (Kyte-Doolittle) and color coded as per disorder propensity. Blue, red and grey color bar represents order-promoting, disorder-promoting and neutral amino acid residues, respectively. Positive and negative values represent the enriched and depleted amino acids in the *Pv*puCSP, respectively. (B) Prediction of natural disordered region in *Pv*puCSP using PONDR-FIT and PONDR VL-XT. PONDR score below 0.5 is considered suggestive of ordered protein. MoRFs prediction using ANCHOR is shown as purple line. Residues with ANCHOR score above 0.5 are predicted as possible binding site for substrates. (C) Charge-hydropathy plot of *Pv*puCSP with hydropathy value ranging from 0 to 1. It compares the mean net charge and hydropathy. Red and blue color dots represent the disordered and ordered residues, respectively and the *Pv*puCSP is represented in green. (D) CDF plot of *Pv*puCSP using cumulative histogram of the PONDR VL-XT score as an input. Black line with dots represents the boundary line and green line shows the cumulative distribution fraction of amino acids in PvpuCSP. Location of *Pv*puCSP curve below the boundary line suggest it as a disordered protein as a whole. It is to be specially noted that the tertiary model shown in the figure carries two long IDRs, tertiary conformation of which couldn't be accurately predicted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

more predicted genes to their biological processes and molecular functions is one of the essential pre-requisites for identification of critical targets before proceeding to expensive and tedious wet-lab experiments. Computational approaches to annotate the unknown genes using extensive real data available in public resources is a cost-effective and time-saving process.

Putative CSP was one of the genes present in a 200 kb genomic region, which was presumed to be under selection pressure (Gupta et al., 2012). On preliminary observation, it was believed that the identified putative CSP must be coding for the abundant sporozoite surface protein. However, on a closer look, it was found to be different from the well-known circumsporozoite protein in terms of its amino acid sequence and functional domain although the annotation 'putative CSP' was retained in the databases. Therefore, the present study was conceived to annotate the putative CSP through an integrated and systematic computational approach. It was also important to distinguish from or correlate the putative CSP and the well-known CSP in order to avoid the muddle in terms of sharing a similar name.

The only similarity observed between the well-known CSP and *Pv*puCSP was the presence of a central repeat region, though the repeat composition (amino acid) was different. Moreover, the tandem repeats of tetra-peptide R-[DC]-N-A were found to be exclusive to *P. vivax* and variation observed within the species in terms of a number of repeats make it a suitable marker to inspect variability in a population.

Although domain architecture of *Pv*puCSP is defined, its role in *P. vivax* is uncertain. The protein family with this architecture is not explored till date despite the fact that it is present in other clinically important *Plasmodium* species such as *P. falciparum*. The RING zinc finger ($C_3H_2C_3$) domain present at the C-terminal of the protein, which is the functional domain in E3 ubiquitin ligase, provides insights that *Pv*puCSP might be involved in the ubiquitination process in *P. vivax* (Fig. 1). Ubiquitination is the process of targeted protein degradation through a chain reaction involving ubiquitin (Ub), enzymes to activate Ub (E1), Ub conjugating enzymes (E2) and ubiquitin ligases (E3) (Lorick et al., 1999). E3 ubiquitin ligase is a class of diverse proteins which plays an important role in protein ubiquitination by selecting appropriate protein for degradation. More specifically, it catalyzes the transfer of E2-ubiquitin complex or in some cases only modified Ub(s) to the lysine side chain of a specific protein to be degraded (Hershko, 1996; Ponts et al., 2008).

Accurate prediction of the topology of a transmembrane protein and its orientation are few of the big challenges since the experimental determination of membrane protein structure is hard to achieve with the existing crystallography method (Almeida et al., 2017). Membrane proteins are essential for signaling and small molecule transport and are also effective drug targets (Terstappen and Reggiani, 2001; Ubarretxena-Belandia and Stokes, 2010). There are several prediction servers available to assign the membrane topology. Most of them use physicochemical properties of the protein like hydropathicity of amino acids, (Kyte and Doolittle, 1982) positive-inside rule (von Heijne,

1992), free energy contribution of amino acids to span the membrane (Hessa et al., 2007) and charge bias analysis to predict the topology (Krogh et al., 2001). The ambiguity in predicting *Pv*puCSP topology especially the 'in and out' orientation by different prediction servers may be addressed by lack of experimental evidence about the fold and structure of this protein family. While the majority of the tools predicted both ends of the *Pv*puCSP to be cytoplasmic, prediction servers like PHOBIUS and SCAMPI revealed N-terminal to be outside (Table 1). Since the functional domain is present towards the C-terminal of the *Pv*puCSP, it is of keen interest to find out the right orientation as erroneous prediction may impede the functionalities of the protein as a whole. From the above exercise, though a rudimentary idea about the transmembrane domains and their spanning region is obtained, the information is certainly not enough to suggest *Pv*puCSP as a transmembrane protein unless the exact location and orientation of *Pv*puCSP is confirmed by crystallography tools and cellular localization studies. Besides, whether it is a part of the cell membrane or nuclear membrane or any other organelle membrane also needs to be established.

Function prediction of a hypothetical or a novel protein based on the phylogenetic profiling of its orthologous sequences whose function is already known, is a highly reliable approach (Tatusov et al., 1997; Sivashankari and Shanmughavel, 2006). To annotate *Pv*puCSP, phylogenetic analysis of *Pv*puCSP orthologs from *Plasmodium* and its other apicomplexan counterparts might be helpful to assign certain biological function to it. The presence of the gene in all *Plasmodium* species, with conserved synteny strongly suggests that the gene might have some essential function for the parasite life cycle (Fig. 2). Presence of orthologous sequences in higher, as well as lower organisms, substantiates its functional significance (Fig. 3). Even though the protein is characterized as an ubiquitin ligase in higher organisms, it is still to be explored in lower organisms including apicomplexans.

Analyses of transcriptomes from a wide range of species from varying origin have drawn an impression that a group of genes that are involved in common cellular processes get co-expressed or follow a similar expression pattern. The above idea has been implemented as a basic strategy to annotate the function of an unknown protein by tracing the pathway in which the protein is involved, thereby its probable function.

In malaria parasite, a number of transcriptomic studies have been conducted to answer many biological questions (Lee et al., 2018). The stage-specific transcriptome profiling studies of malaria parasite have established that malaria parasites have a well-regulated time and life cycle dependent gene and protein expression system because of their complex life cycle in two very different hosts (Bozdech et al., 2003). Since E3 ubiquitin ligase has a critical role in determining the fate of most of the proteins and is a major checkpoint in protein ubiquitination, strict control on its expression and regulation is important in the life of parasite for survival and virulence. Moderate expression of *Pv*puCSP across all parasitic stages suggests that *Pv*puCSP have a regulatory function. This further adds one more layer of evidence to the speculation that *Pv*puCSP is different from *Pv*CSP in terms of biological function and is not a "true" circumsporozoite protein. Inhibiting the activity of *Pv*puCSP will provide a better picture about the importance of it in the life cycle of *P. vivax*. Since E3 ubiquitin ligases are the suitable drug target in apicomplexans and under research focus (Ng et al., 2017; Jain et al., 2017; Gupta et al., 2018), the conservedness of *Pv*puCSP across *Plasmodium* species will make it a more suitable target.

Functions of proteins are intrinsically linked with their unique three-dimensional conformation adopted in the native cellular environment (Anfinsen, 1973). There are a limited number of experimental methods (X-ray crystallography, Nuclear Magnetic Resonance-spectroscopy, and electron microscopy) available to determine the three-dimensional atomic coordinates of a biomolecule, with their own proficiencies and challenges (Lacapère et al., 2007; Slabinski et al., 2007). The complications increase many fold when it comes to transmembrane proteins, as the process of extraction and purification of the transmembrane protein from the lipid bilayer is extremely challenging (Lacapère et al., 2007). That is the reason why very less number of protein folds and tertiary structure are resolved till date although a vast number of protein sequences are available. Molecular modeling is another alternative to the protein structure determination when structure could not be determined by the above-mentioned tools. However, the lack of a suitable template with threshold homology to the protein of interest certainly restricts the prediction accuracy.

In the absence of appropriate template(s), there are many algorithms that have been proposed for accurately threading the target protein to the known folds and to obtain the best fit. The automated fold recognition servers used in this study to model *Pv*puCSP have the basic fold recognition algorithm with added features to enhance its prediction accuracy. Since there is variation in approaches and algorithms used to predict the tertiary structure, all the models were validated to obtain the best fit model. A loop is a small and flexible form of a protein's secondary structure which, helps to interconnect two secondary structures (alpha helices and beta sheets) and biologically has many vital functions because of its flexible nature, however it does not follow a regular and observable pattern. That is why; it is difficult to accurately model the loop region of a protein. Knowledge-based loop refinement increases the stability of the protein by changing the empirical distribution of amino acids.

The classical concept of structure-function paradigm, i.e. the activity of a protein is determined by its unique three-dimensional conformations, has been changing with the discovery of IDRs and their comparable abundance with structured proteins in nature. Nevertheless, the conformational plasticity is found to be fundamental for many crucial biological activities such as DNA binding, recognition, regulation and signaling (Iakoucheva et al., 2002; Uversky and Obradovic, 2008). Therefore, any change during the process of the protein formation and folding would eventually lead to many diseased states like neurodegeneration, cardiovascular disease, amyloidogenesis and many more (Cheng et al., 2006; *Table of Contents* 1, 2009; Kulkarni and Uversky, 2019; Du and Uversky, 2017). In the context of folds that proteins can acquire, they can be roughly classified into three categories: structured proteins, intrinsically disordered proteins and proteins with both structured and disordered regions. A large portion of the eukaryote proteome belongs to the third category i.e. hybrid proteins with both ordered and disordered regions (Dunker et al., 2013).

The tertiary model of *Pv*puCSP shows that a significant portion of the protein forms coils as secondary structure. The above observation might be due to lack of a suitable template, however, presence of long IDRs couldn't be denied as a probable reason. In *Pv*puCSP, around 50% of the amino acids are disorder promoting, which adds first line evidence to its disordered nature. However, *Pv*puCSP seemingly belongs to the category of hybrid proteins with a mix of well-defined domains of compact structure and disordered regions with structural flexibility. Furthermore, IDRs are mostly low complexity regions rich with different kind of repeats and lack of structural stability in the tandem repeats has been established from earlier studies. Therefore, co-localization of tetra-peptide tandem repeats and IDRs at the C-terminal of *Pv*puCSP support of each other. It has also been reported that post-translational modifications are more frequent in IDRs (Xue et al., 2009). Therefore, the presence of glycosylation and phosphorylation sites profoundly in predicted IDRs of *Pv*puCSP further confirms its intrinsic disordered nature (Supplementary Table S4).

Since a major portion of *Pv*puCSP is predicted to have disordered regions, it might be correlated with the presence of zinc finger domains that might act as a DNA binding domains. It has been reported that proteins which act as chaperones for other proteins also carry the unfolded segments in order to bind with misfolded protein and RNA molecule (Tompa and Csermely, 2004). Besides, IDRs are more common in proteins that are de novo translated locally from their transcripts rather than in the proteins that are pre-translated before getting transferred (Lacapère et al., 2007). The enrichment of IDRs in

distally translated protein provides conformational flexibility to the proteins that lead to a larger surface area to interact with a diverse group of molecules (Van et al., 2014). Compositionally biased amino acid and multiple linear motifs are the common features of IDRs. Since *Pv*puCSP also possesses multiple stretches of a repeat of amino acid [R-D-N-A], it is suggested that it might undergo localized translation and the presence of zinc finger domain further supports the fact that it may bind with different biomolecules more specifically with the proteins to initiate ubiquitination. To summarize, both direct and indirect evidence supports the finding that *Pv*puCSP is a protein that carries N- and C-terminal IDRs and well defined transmembrane and RING zinc finger domains.

The PPI network built on DEGs in *P. vivax* revealed the interaction of putative CSP with other ubiquitin ligases and with DNA-directed RNA polymerase and DNA repair proteins which also suggests *Pv*puCSP as an integral part of ubiquitination along with DNA repair and signaling.

In conclusion, in the light of existing difficulties in maintaining long term *P. vivax* culture, the present study analyzed the *Pv*puCSP (PVX_086150) at every possible level starting from its primary sequence to the tertiary structure, using existing resources like orthologous sequences across species, high throughput transcriptome data, atomic coordinates of the template, etc. with necessary computational tools, to gather and correlate the information to assign putative function(s) to it. Comparison of both genes (*Pv*CSP and *Pv*puCSP) revealed low sequence level similarity other than a similar repeat pattern. Moreover, presence of RING domain, which is entirely different from that present in *Pv*CSP, leads to the conclusion that *Pv*puCSP is an integral membrane protein, with both terminals exposed on the cytosolic side of the parasite. However, the expression of *Pv*puCSP is not restricted to the sporozoite stage. The presence of repeat unit only in *P. vivax* and variation within *P. vivax*, make it a suitable marker to study the diversity of *P. vivax* in different population. However, the basis on which the protein was named as circumsporozoite protein in *P. vivax* and in few other *Plasmodium* species like *P. coatneyi*, *P. cynomolgi*, and *P. knowlesi* is not clear. This study proposes that *Pv*puCSP is a completely different protein from the well-known CSP irrespective of sharing a similar name. Since many of the ubiquitination related proteins are verified drug targets, and *Pv*puCSP is implicated in the ubiquitination pathway, it should be further explored as a novel antimalarial target. Taken together all the observations from this study, it can be concluded that *Pv*puCSP lacks direct evidence of its function at present, however, biochemical and functional studies to characterize *Pv*puCSP might project it as a critical target and open new paths towards vivax malaria treatment and control strategy.

4. Materials and methods

4.1. *PvpuCSP characterization and domain analysis*

The protein sequence of *Pv*puCSP was retrieved from PlasmoDB (PVX_086150) and domain architecture was analyzed using multiple platforms viz. Conserved Domain Database CDD,[4] SPARCLE,[5] (Marchler-Bauer et al., 2013) PROSITE (Sigrist et al., 2013), and SMART[6] (Letunic and Bork, 2018; Schultz et al., 1998). Proteins with similar domain architecture were searched in other organisms including apicomplexans through the CDART[7] to understand its function based on homology (Supplementary Table S1) (Geer et al., 2002). Since presence of repeats in central region is a characteristic feature of well-known circumsporozoite protein (PVX_119355), *Pv*puCSP was scanned for tandem repeats using the Tandem Repeat Finder (Benson, 1999).

In order to confirm whether *Pv*puCSP is a single or multi-copy gene, BLASTn was performed against the *P. vivax* genome, with the presumption that the most similar hits with expect value [E] more than 1E−10 would be considered as single copy gene (Wu et al., 2006). Topology and orientation of *Pv*puCSP were predicted using transmembrane prediction servers (Table 1). Presence of signal peptide in *Pv*puCSP was also investigated using SignalP4.1 (Petersen et al., 2011) and signalHsmm (Burdukiewicz et al., 2018), which is specifically designed to identify signal peptides from malaria parasite and related species.

A DELTA-BLAST[8] search was performed using *Pv*puCSP as the query against Refseq_protein database to obtain homologous sequence carrying a similar domain as *Pv*puCSP in order to perform homology-based annotation (Boratyn et al., 2012). Retrieved hits with > 35% identity with *Pv*puCSP were aligned using iterative refined method present in MAFFT (Katoh et al., 2005). The duplicate sequences were removed manually and poorly aligned regions were trimmed from the alignment using G-BLOCKS (Talavera and Castresana, 2007). The maximum likelihood phylogeny was reconstructed using the JTT matrix-based model applying 1000 bootstraps in MEGA7 and the tree with the highest log likelihood value was chosen for analysis. More information about orthologs and paralogs of *Pv*puCSP was also retrieved from EggNOG (Huerta-Cepas et al., 2016).

4.2. *Analysis of stage-specific P. vivax transcriptome data*

The transcriptome data provides a better idea about the role of a protein based on its level of expression and the other protein along with which it gets expressed. The stage specific transcriptome data generated from previous studies were utilized to gather information about *Pv*puCSP expression and to distinguish it from well-known CSP by comparing their expression at different stages of parasite (Zhu et al., 2016; Westenberger et al., 2010; Gural et al., 2018; Bozdech et al., 2003; Bozdech et al., 2008; Hoo et al., 2016). The stage-specific transcriptome data of the intra-erythrocytic developmental cycle (IDC) of two distinct *P. vivax* isolates were screened (Zhu et al., 2016; Bozdech et al., 2003; Bozdech et al., 2008; Hoo et al., 2016). The dataset included the transcriptome data during the intra-erythrocytic cycle at 9-time points (TP) covering all erythrocytic stages (early ring, late ring, trophozoite, early schizont, late schizont) at 6-hour gaps. The transcriptome data of all stages of *P. vivax* in human and mosquito (sporozoite, erythrocytic stage at 8 different time points, gametes, zygotes, and ookinetes) were also scanned in depth to check the stage of expression and interaction of *Pv*puCSP with other proteins (Westenberger et al., 2010). The known stage-specific proteins (*Pv*AMA1 for blood stage and *Pvs*25 for gametocyte stage) were chosen as a control to check the change in expression. The recently published hypnozoite transcriptome data were also checked (Gural et al., 2018). Since limited expression data is available for *P. vivax*, the transcriptome of other *Plasmodium* species (*P. berghei*, *P. yoelii* and *P. falciparum*) were also explored using *Pv*puCSP orthologs (Bozdech et al., 2003; Hoo et al., 2016).

The transcriptome data covering all stages of the parasite in human and mosquito were further analyzed employing STRING[9] databaseV10.5 (Szklarczyk et al., 2017) and Cytoscape (Shannon et al., 2003), to identify the other genes co-expressed with *Pv*puCSP and infer its biological function through protein-protein interaction network analysis. From the transcriptome data, the DEGs[10] with a cutoff of Fold Change value (FC) ≥2 and p value ≤ 0.05 were filtered out. A PPI[11]

[4] Conserved Domain Database.
[5] Subfamily Protein Architecture Labelling Engine.
[6] Simple Modular Architecture Research Tool.
[7] Conserved Domain Architecture Retrieval Tool.
[8] Domain Enhanced Lookup Time Accelerated BLAST.
[9] Search tool for the retrieval of interacting genes.
[10] Differentially Expressed Genes.
[11] Protein-protein interaction.

network of the DEGs was built using Cytoscape, and functional enrichment was performed by mapping the DEGs to STRING app in-built in Cytoscape (Szklarczyk et al., 2017; Shannon et al., 2003) to obtain the degree and other centrality measures for each node.

In network analysis, each gene is referred as a node and its interaction with other proteins represented as edges. The nodes with the highest degrees (number of interacting proteins) were considered to be biologically significant and called as a hub gene (Barabási and Oltvai, 2004). The proteins with which *Pv*puCSP showed interaction were clustered using MCODE plug-in (Sun et al., 2017). To get a more comprehensive idea about the molecular function of *Pv*puCSP, Gene Ontology enrichment analysis was performed using the proteins showing interaction with *Pv*puCSP. All the genes in the network were uploaded to DAVID[12] v6.8 online tool to perform gene enrichment analysis (Huang et al., 2009a; Huang et al., 2009b).

4.3. Molecular modeling and binding site prediction

*Pv*puCSP was searched against PDB[13] to screen out the suitable templates for homology modeling using DELTA-BLAST (Boratyn et al., 2012). Since the PDB hits obtained against the *Pv*puCSP had significantly low sequence similarity and insufficient query coverage, *Pv*puCSP was modeled via fold recognition method using automated modeling servers like Orion (Ghouzam et al., 2016), SPARK-X (Yang et al., 2011) and I-TASSER server (Yang et al., 2014). Besides, the template information obtained from the fold recognition servers were selected as a template (PDB ID: 5ALU, 5WTJ, 2M6M, and 5XJY) based on its score and similarity with *Pv*puCSP and the structure was predicted using the multi-template method in the modeler (Šali and Blundell, 1993). The functional domain [RING domain] of *Pv*puCSP was modeled independently via homology modeling using the best match templates (PDB ID-2L0B and 2ECT) along with zinc ion using Modeller9.19 (Webb and Sali, 2017). The quality of the model was assessed based on DOPE[14] score (Marchler-Bauer et al., 2013; Shen and Sali, 2006) and GA341 score (Melo et al., 2009). The model with the lowest DOPE score and GA341 score close to 1 considered to be a best-fit model for further analysis.

The best (with highest C score and Z score) out of all the models generated by I-TASSER, Orion, SPARK-X and Modeler were analyzed using the validation servers viz. using PROCHECK (Laskowski et al., 1993), ERRAT (Colovos and Yeates, 1993; MacArthur et al., 1994), Verify-3D (Eisenberg, 1997) and PROVE (Pontius et al., 1996) available in SAVES meta-server (Table 3), and the one with a good score and qualified the validation process was chosen for loop refinement (Modloop server) and energy minimization was done using YASARA energy minimization server (Krieger et al., 2009). *Pv*puCSP was also scanned for IDRs[15] using DISOPRED3 (Jones and Ward, 2003; Jones and Cozzetto, 2015).

4.4. IDR profiling of PvpuCSP

Composition profiling analysis of *Pv*puCSP with respect to the propensity and number of order-promoting (N, C, I, L, F, W, Y and V) and disorder-promoting (A, S, R, G, E, P, E and K) residues was carried out using Composition Profiler tool, which detects enriched and depleted amino acids in the query protein based on their many physico-chemical and structural properties (Vacic et al., 2007). A group of proteins with specific attributes that provides background amino acid distribution was compared against the query protein. Here, SwissProt 51 was used as a reference set because the database closely resembles the distribution of proteins in nature. The IDR profiling was done with 10,000 bootstrap iteration and Bonferroni correction to enhance the accuracy and reduce the error probability. The disordered regions were also predicted using PONDR-FIT, a variant of PONDR algorithm, and a meta-predictor that incorporates the output of other individual disorder predictors using artificial neural network (ANN) method (Xue et al., 2010). PONDR VL-XT, a variant of PONDR algorithm with a higher sensitivity to locate disordered stretches within a protein, was also utilized. ANCHOR algorithm was used to find out potential Molecular Recognition Features or MoRFs, which are small active sites present within the IDRs and play an important role to bind and interact with other substrates, in *Pv*puCSP (Dosztányi et al., 2009). Stretches with < 10 residues long were filtered out and residues with ANCHOR score > 0.5 were considered to be MoRFs.

Charge-hydropathy (CH) plot and cumulative distribution fraction plot (CDF) are the two other binary classification measures followed to confirm the structural conformation of *Pv*puCSP as a whole. Binary classification of a protein is to predict the nature of a protein as a whole by analyzing basic properties like charge, hydropathy, bulkiness, surface propensity of the amino acids. Relative high net charge with low mean hydropathy is the characteristic feature of a disordered protein or regions. In CH plot, the mean net charge and hydropathy value of a group of ordered and disordered proteins are used to generate a linear boundary line. Presence of query protein towards left and right of this boundary line predicts the query protein as ordered or disordered protein, respectively. In CDF plot, based on the distribution of a disorder score generated by VL-XT for each amino acid, the ordered state of protein is predicted. The proteins located towards the lower-right half of the CDF plot are assumed to be disordered whereas those located towards the upper-left half are compact proteins (Xue et al., 2009).

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2019.100024.

Funding

This work was supported by the Indian Council of Medical Research (ICMR-https://www.icmr.nic.in/), New Delhi, India in the form of Senior Research Fellowships to MD. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Author contributions

MD and AS designed the study. MD collected, analyzed and interpreted the data. MD and AS jointly wrote the manuscript. VP edited and reviewed the manuscript. AS conceptualize the idea and supervise the work. All authors read and approved the final manuscript.

Data availability statement

All data generated or analyzed during this study are included in this published article and its supplementary information files. The transcriptome data used in the study was taken from http://carrier.gnf.org/publications/Pv/ (Westenberger et al., 2010).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Dr. Neena Valecha, Regional Advisor-Malaria WHO-SEARO, for providing necessary research facilities and moral support during the study. We thank Dr. Sanju Kumari (ICMR PostDoctoral

[12] Database for Annotation, Visualization and Integrated Discovery.
[13] Protein Data Bank.
[14] Discrete Optimized Protein Energy.
[15] Intrinsically disordered regions.

Fellow) and Ms. Aparna Tiwari (CSIR Senior Research Fellow) for critically reviewing the manuscript. Thanks are also owing to all the lab members for their helpful comments and suggestions while preparing the manuscript. We extend our special thanks to Dr. Anubhav Srivastava (Research Fellow, D4 Research Team, Monash University, Australia) for his extensive review for language editing of the manuscript.

References

Aashish, A., Manigandan, G., 2015. Complicated vivax malaria, an often underestimated condition - case report. J. Fam. Community Med. 22 (3), 180. https://doi.org/10.4103/2230-8229.163040.

Almeida, J.G., Preto, A.J., Koukos, P.I., Bonvin, A.M.J.J., Moreira, I.S., 2017. Membrane proteins structures: a review on computational modeling tools. Biochim. Biophys. Acta Biomembr. 1859 (10), 2021–2039. https://doi.org/10.1016/j.bbamem.2017.07.008.

Anfinsen, 1973. Principles that govern the folding of protein chains. Science (80-) 181 (4096), 223–230.

Baird, J.K., 2007. Neglect of *Plasmodium vivax* malaria. Trends Parasitol. 23 (11), 533–539. https://doi.org/10.1016/j.pt.2007.08.011.

Baird, K.J., Maguire, J.D., Price, R.N., 2012. Diagnosis and Treatment of *Plasmodium vivax* Malaria. vol 80 Elsevierhttps://doi.org/10.1016/B978-0-12-397900-1.00004-9.

Barabási, A.L., Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. Nat Rev Genet 5 (2), 101–113. https://doi.org/10.1038/nrg1272.

Bassat, Q., Alonso, P.L., 2011. Defying malaria: fathoming severe *Plasmodium vivax* disease. Nat. Med. 17 (1), 48–49. https://doi.org/10.1038/nm0111-48.

Bennett, J.W., Yadava, A., Tosh, D., et al., 2016. Phase 1/2a trial of *Plasmodium vivax* malaria vaccine candidate VMP001/AS01B in malaria-naive adults: safety, immunogenicity, and efficacy. PLoS Negl. Trop. Dis. 10 (2), 1–16. https://doi.org/10.1371/journal.pntd.0004423.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27 (2), 573–580. https://doi.org/10.1093/nar/27.2.573.

Boratyn, G.M., Schäffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J., Madden, T.L., 2012. Domain enhanced lookup time accelerated BLAST. Biol. Direct 7, 1–14. https://doi.org/10.1186/1745-6150-7-12.

Bozdech, Z., Llinás, M., Pulliam, B.L., Wong, E.D., Zhu, J., DeRisi, J.L., 2003. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol. 1 (1), 85–100. https://doi.org/10.1371/journal.pbio.0000005.

Bozdech, Z., Mok, S., Hu, G., et al., 2008. The transcriptome of *Plasmodium vivax* reveals divergence and diversity of transcriptional regulation in malaria parasites. Proc. Natl. Acad. Sci. 105 (42), 16290–16295. https://doi.org/10.1073/pnas.0807404105.

Burdukiewicz, M., Sobczyk, P., Chilimoniuk, J., Gagat, P., Mackiewicz, P., 2018. Prediction of signal peptides in proteins from malaria parasites. Int. J. Mol. Sci. 19 (12), 3709. https://doi.org/10.3390/ijms19123709.

Cappellini, M.D., Fiorelli, G., 2008. Glucose-6-phosphate dehydrogenase defi ciency. Lancet 371 (6), 64–74. https://doi.org/10.2514/2.3643.

Cheng, Y., Legall, T., Oldfield, C.J., Dunker, A.K., Uversky, V.N., 2006. Abundance of Intrinsic Disorder in Protein Associated With Cardiovascular Disease. pp. 10448–10460.

Colovos, C., Yeates, T.O., 1993. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. 2 (9), 1511–1519. https://doi.org/10.1002/pro.5560020916.

Das, A., 2015. The distinctive features of Indian malaria parasites. Trends Parasitol. 31 (3), 83–86. https://doi.org/10.1016/j.pt.2015.01.006.

De Camargo, T.M., De Freitas, E.O., Gimenez, A.M., et al., 2018. Prime-boost vaccination with recombinant protein and adenovirus-vector expressing *Plasmodium vivax* circumsporozoite protein (CSP) partially protects mice against Pb/Pv sporozoite challenge. Sci. Rep. 8 (1), 1–14. https://doi.org/10.1038/s41598-017-19063-6.

Dharia, N.V., Bright, A.T., Westenberger, S.J., et al., 2010. Whole-genome sequencing and microarray analysis of ex vivo *Plasmodium vivax* reveal selective pressure on putative drug resistance genes. Proc. Natl. Acad. Sci. 107 (46), 20045–20050. https://doi.org/10.1073/pnas.1003776107.

Dosztányi, Z., Mészáros, B., Simon, I., 2009. ANCHOR: web server for predicting protein binding regions in disordered proteins. 25 (20), 2745–2746. https://doi.org/10.1093/bioinformatics/btp518.

Douglas, N.M., Anstey, N.M., Buffet, P.A., et al., 2012. The anaemia of *Plasmodium vivax* malaria. Malar. J. 11 (1). https://doi.org/10.1186/1475-2875-11-135.

Du, Z., Uversky, V.N., 2017. A comprehensive survey of the roles of highly disordered proteins in type 2 diabetes. https://doi.org/10.3390/ijms18102010.

Dunker AK, Babu MM, Barbar E, et al. Why these proteins are intrinsically disordered what's in a name? 2013:1–5.

Eisenberg, D., 1997. Verify 3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. 277, 396–404.

Geer, L.Y., Domrachev, M., Lipman, D.J., Bryant, S.H., 2002. CDART: Protein Homology by Domain Architecture. pp. 1619–1623. https://doi.org/10.1101/gr.278202.CDART.

Geleta, G., Ketema, T., 2016. Severe malaria associated with *Plasmodium falciparum* and *P. vivax* among children in Pawe Hospital, Northwest Ethiopia. Malar Res Treat 2016, 1–7. https://doi.org/10.1155/2016/1240962.

Genton, B., D'Acremont, V., Rare, L., et al., 2008. *Plasmodium vivax* and mixed infections are associated with severe malaria in children: a prospective cohort study from Papua New Guinea. PLoS Med. 5 (6), 0881–0889. https://doi.org/10.1371/journal.pmed.0050127.

Ghouzam, Y., Postic, G., Guerin, P.-E., de Brevern, A.G., Gelly, J.-C., 2016. ORION: a web server for protein fold recognition and structure prediction using evolutionary hybrid profiles. Sci. Rep. 6 (1), 28268. https://doi.org/10.1038/srep28268.

Gordon, D.M., McGovern, T.W., Krzych, U., et al., 1995. Safety, immunogenicity, and efficacy of a recombinantly produced *Plasmodium falciparum* circumsporozoite protein-hepatitis B surface antigen subunit vaccine. J. Infect. Dis. 171 (6), 1576–1585. https://doi.org/10.1093/infdis/171.6.1576.

Gupta, B., Dash, A.P., Shrivastava, N., Das, A., 2010. Single nucleotide polymorphisms, putatively neutral DNA markers and population genetic parameters in Indian *Plasmodium vivax* isolates. Parasitology 137, 1721–1730. https://doi.org/10.1017/S0031182010000533.

Gupta, B., Srivastava, N., Das, A., 2012. Inferring the evolutionary history of Indian *Plasmodium vivax* from population genetic analyses of multilocus nuclear DNA fragments. Mol. Ecol. 21 (7), 1597–1616. https://doi.org/10.1111/j.1365-294X.2012.05480.x.

Gupta, I., Aggarwal, S., Singh, K., Yadav, A., Khan, S., May 2018. Ubiquitin proteasome pathway proteins as potential drug targets in parasite Trypanosoma cruzi. Sci. Rep. 1–12. https://doi.org/10.1038/s41598-018-26532-z.

Gural, N., Mancio-Silva, L., Miller, A.B., et al., 2018. In vitro culture, drug sensitivity, and transcriptome of *Plasmodium vivax* hypnozoites. Cell Host Microbe 23 (3), 395–406.e4. https://doi.org/10.1016/j.chom.2018.01.002.

von Heijne, G., 1992. Membrane protein structure prediction. J. Mol. Biol. 225, 487–494. https://doi.org/10.1016/0022-2836/92/100487-08.

Hershko, A., 1996. Lessons from the discovery of the ubiquitin system. Trends Biochem. Sci. 21 (11), 445–449. https://doi.org/10.1016/S0968-0004(96)10054-2.

Hessa, T., Meindl-Beinker, N.M., Bernsel, A., et al., 2007. Molecular code for transmembrane-helix recognition by the Sec61 translocon. Nature 450 (7172), 1026–1030. https://doi.org/10.1038/nature06387.

Hoo, R., Zhu, L., Amaladoss, A., et al., 2016. Integrated analysis of the Plasmodium species transcriptome. EBioMedicine 7, 255–266. https://doi.org/10.1016/j.ebiom.2016.04.011.

Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009a. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4 (1), 44–57. https://doi.org/10.1038/nprot.2008.211.

Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009b. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37 (1), 1–13. https://doi.org/10.1093/nar/gkn923.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., et al., 2016. EGGNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44 (D1), D286–D293. https://doi.org/10.1093/nar/gkv1248.

Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Dunker, A.K., 2002. Intrinsic disorder in cell-signaling and cancer-associated proteins. 2836 (2), 573–584. https://doi.org/10.1016/S0022-2836(02)00969-5.

Jain J, Jain SK, Walker LA, Tekwani BL. Inhibitors of ubiquitin E3 ligase as potential new antimalarial drug leads. 2017:1–10. doi:https://doi.org/10.1186/s40360-017-0147-4.

Jones, D.T., Cozzetto, D., 2015. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31 (6), 857–863. https://doi.org/10.1093/bioinformatics/btu744.

Jones, D.T., Ward, J.J., 2003. Prediction of disordered regions in proteins from position specific score matrices. Proteins Struct. Funct. Genet. 53 (Suppl. 6), 573–578. https://doi.org/10.1002/prot.10528.

Katoh, K., Kuma, K.I., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33 (2), 511–518. https://doi.org/10.1093/nar/gki198.

Krieger, E., Joo, K., Lee, J., et al., 2009. NIH Public Access. Proteins 77 (Suppl. 9), 114–122. https://doi.org/10.1002/prot.22570.Improving.

Krogh, A., Larsson, B., Von Heijne, G., Sonnhammer, E.L.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305 (3), 567–580. https://doi.org/10.1006/jmbi.2000.4315.

Kulkarni, P., Uversky, V.N., 2019. Intrinsically Disordered Proteins in Chronic Diseases. pp. 1–6.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydrophativ character of a protein. J. Mol. Biol. 105–132.

Lacapère, J.J., Pebay-Peyroula, E., Neumann, J.M., Etchebest, C., 2007. Determining membrane protein structures: still a challenge!. Trends Biochem. Sci. 32 (6), 259–270. https://doi.org/10.1016/j.tibs.2007.04.001.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 26 (2), 283–291. https://doi.org/10.1107/S0021889892009944.

Lee, H.J., Georgiadou, A., Otto, T.D., et al., 2018. Transcriptomic studies of malaria: a paradigm for investigation of systemic host-pathogen interactions. Microbiol. Mol. Biol. Rev. 82 (2), 1–37. https://doi.org/10.1128/MMBR.00071-17.

Letunic, I., Bork, P., 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44 (W1), W242–W245. https://doi.org/10.1093/nar/gkw290.

Letunic, I., Bork, P., 2018. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. 46 (D1), D493–D496. https://doi.org/10.1093/nar/gkx922.

Liu, W., Xie, Y., Ma, J., et al., 2015. IBS: an illustrator for the presentation and visualization of biological sequences. Bioinformatics 31 (20), 3359–3361. https://doi.org/10.1093/bioinformatics/btv362.

Lorick, K.L., Jensen, J.P., Fang, S., Ong a, M., Hatakeyama, S., Weissman a, M., 1999. RING fingers mediate ubiquitin-conjugating enzyme (E2)-dependent ubiquitination. Proc. Natl. Acad. Sci. U. S. A. 96 (20), 11364–11369. https://doi.org/10.1073/pnas.

96.20.11364.
Lover, A.A., Baird, J.K., Gosling, R., Price, R., May 2018. Malaria elimination: time to target all species. Am J Trop Med Hyg. https://doi.org/10.4269/ajtmh.17-0869.
MacArthur, M.W., Laskowski, R.A., Thornton, J.M., 1994. Knowledge-based validation of protein structure coordinates derived by X-ray crystallography and NMR spectroscopy. Curr. Opin. Struct. Biol. 4 (5), 731–737. https://doi.org/10.1016/S0959-440X(94)90172-4.
Marchler-Bauer, A., Zheng, C., Chitsaz, F., et al., 2013. CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. 41 (D1), 348–352. https://doi.org/10.1093/nar/gks1243.
Melo, F., Sánchez, R., Sali, A., 2009. Statistical potentials for fold assessment. Protein Sci. 11 (2), 430–448. https://doi.org/10.1002/pro.110430.
Ng, C.L., Fidock, D.A., Bogyo, M., 2017. Protein degradation systems as antimalarial therapeutic targets. Trends Parasitol. xx, 1–13. https://doi.org/10.1016/j.pt.2017.05.009.
Petersen, T.N., Brunak, S., Von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8 (10), 785–786. https://doi.org/10.1038/nmeth.1701.
Pontius, J., Richelle, J., Wodak, S.J., 1996. Deviations from standard atomic volumes as a quality measure for protein crystal structures. J. Mol. Biol. 264 (1), 121–136. https://doi.org/10.1006/jmbi.1996.0628.
Ponts, N., Yang, J., Chung, D.W.D., et al., 2008. Deciphering the ubiquitin-mediated pathway in apicomplexan parasites: a potential strategy to interfere with parasite virulence. PLoS One 3 (6). https://doi.org/10.1371/journal.pone.0002386.
Price, R.N., Tjitra, E., Guerra, C.A., Yeung, S., White, N.J., Anstey, N.M., 2007. Vivax malaria: neglected and not benign. Am J Trop Med Hyg 77 (Suppl. 6), 79–87 (doi:77/6_Suppl/79 [pii]).
Ren, J., Wen, L., Gao, X., Jin, C., Xue, Y., Yao, X., 2009. DOG 1.0: illustrator of protein domain structures. Cell Res. 19 (2), 271–273. https://doi.org/10.1038/cr.2009.6.
Šali, A., Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234 (3), 779–815. https://doi.org/10.1006/jmbi.1993.1626.
Schultz, J., Milpetz, F., Bork, P., Ponting, C.P., 1998. SMART, a simple modular architecture research tool: identification of signaling domains. Proc. Natl. Acad. Sci. 95 (11), 5857–5864. https://doi.org/10.1073/pnas.95.11.5857.
Shannon, P., Markiel, A., Ozier 2, O., et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. (13), 2498–2504. https://doi.org/10.1101/gr.1239303.metabolite.
Shen, M., Sali, A., 2006. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2507–2524. https://doi.org/10.1110/ps.062416606.Instead.
Sigrist, C.J.A., De Castro, E., Cerutti, L., et al., 2013. New and continuing developments at PROSITE. Nucleic Acids Res. 41 (D1), 344–347. https://doi.org/10.1093/nar/gks1067.
Sivashankari, S., Shanmughavel, P., 2006. Functional annotation of hypothetical proteins - a review. Bioinformation 1 (8), 335–338. https://doi.org/10.6026/97320630001335.
Slabinski, L., Jaroszewski, L., Rodrigues, A.P.C., et al., 2007. The challenge of protein structure determination-lessons from structural genomics. Protein Sci. 16 (11), 2472–2482. https://doi.org/10.1110/ps.073037907.
Sun, C., Yuan, Q., Wu, D., Meng, X., Wang, B., 2017. Identification of core genes and outcome in gastric cancer using bioinformatics analysis. Oncotarget 8 (41), 70271–70280. https://doi.org/10.18632/oncotarget.20082.
Szklarczyk, D., Morris, J.H., Cook, H., et al., 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 45 (D1), D362–D368. https://doi.org/10.1093/nar/gkw937.
Table of Contents 1. pp. 5188–5238.
Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56 (4), 564–577. https://doi.org/10.1080/10635150701472164.
Tatusov, R.L., Koonin, E.V., Lipman, D.J., 1997. A genomic perspective on protein families. Science (80-) 278 (5338), 631–637. https://doi.org/10.1126/science.278.5338.631.
Terstappen, G.C., Reggiani, A., 2001. In silico research in drug discovery. Trends Pharmacol. Sci. 22 (1), 23–26. https://doi.org/10.1016/S0165-6147(00)01584-4.
Tham, W.H., Beeson, J.G., Rayner, J.C., 2017. Plasmodium vivax vaccine research – we've only just begun. Int. J. Parasitol. 47 (2–3), 111–118. https://doi.org/10.1016/j.ijpara.2016.09.006.
Tjitra, E., Anstey, N.M., Sugiarto, P., et al., 2008. Multidrug-resistant *Plasmodium vivax* associated with severe and fatal malaria: a prospective study in Papua, Indonesia. PLoS Med. 5 (6), 0890–0899. https://doi.org/10.1371/journal.pmed.0050128.
Tompa, P., Csermely, P., 2004. The role of structural disorder in the function of RNA and protein chaperones. FASEB J. 18 (11), 1169–1175. https://doi.org/10.1096/fj.04-1584rev.
Ubarretxena-Belandia, I., Stokes, D.L., 2010. Present and future of membrane protein structure determination by electron crystallography. Adv Protein Chem Struct Biol. 81, 33–60. https://doi.org/10.1111/j.1743-6109.2008.01122.x.Endothelial.
Uversky, V.N., Obradovic, Z., 2008. NIH Public Access 6 (5), 1882–1898. https://doi.org/10.1021/pr060392u.Functional.
Vacic, V., Uversky, V.N., Dunker, A.K., Lonardi, S., 2007. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. 7, 1–7. https://doi.org/10.1186/1471-2105-8-211.
Van, R.K., Uyar, B., Weatheritt, R.J., et al., 2014. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem. Rev. 114 (13), 6733–6778. https://doi.org/10.1021/cr400585q.
Webb, B., Sali, A., 2017. Comparative protein structure modeling using MODELLER. Curr Protoc Bioinforma 54 (ii), 1–55. https://doi.org/10.1002/cpbi.3.Comparative.
Westenberger, S.J., McClean, C.M., Chattopadhyay, R., et al., 2010. A systems-based analysis of *Plasmodium vivax* lifecycle transcription from human to mosquito. PLoS Negl. Trop. Dis. 4 (4). https://doi.org/10.1371/journal.pntd.0000653.
World Health Organization (WHO), 2018. World Malaria Report 2018.
World Health Organization (WHO), 2016. World Malaria Report 2016.
Wu, F., Mueller, L.A., Crouzillat, D., Pétiard, V., Tanksley, S.D., 2006. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics 174 (3), 1407–1420. https://doi.org/10.1534/genetics.106.062455.
Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K., Uversky, V.N., 2010. PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1804 (4), 996–1010. https://doi.org/10.1016/j.bbapap.2010.01.011.
Xue, B., Oldfield, C.J., Dunker, A.K., Uversky, V.N., 2009. CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. FEBS Lett. 583 (9), 1469–1474. https://doi.org/10.1016/j.febslet.2009.03.070.
Yang, Y., Faraggi, E., Zhao, H., Zhou, Y., 2011. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 27 (15), 2076–2082. https://doi.org/10.1093/bioinformatics/btr350.
Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., Zhang, Y., 2014. The I-TASSER suite: protein structure and function prediction. Nat. Methods 12 (1), 7–8. https://doi.org/10.1038/nmeth.3213.
Zhu, L., Mok, S., Imwong, M., et al., 2016. New insights into the *Plasmodium vivax* transcriptome using RNA-Seq. Sci. Rep. 6 (January), 1–13. https://doi.org/10.1038/srep20498.