BIOINFORMATICS DISCOVERY NOTE

Vol. 30 no. 5 2014, pages 597–600
doi:10.1093/bioinformatics/btt587

Genome analysis

Advance Access publication October 15, 2013

Thymine distribution in genes provides novel insight into the functional significance of the proteome of the malaria parasite *Plasmodium falciparum* 3D7

Balamurugan Palanisamy[1], Rajasekaran Ekambaram[1] and Klaus Heese[2,*]

[1]School of Biotechnology and Health Sciences, Karunya University, Karunya Nagar, Coimbatore 641114, Tamil Nadu, India and [2]Graduate School of Biomedical Science and Engineering, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul 133-791, Republic of Korea

Associate Editor: John Hancock

ABSTRACT

Summary: *Plasmodium falciparum* (*Pf*)-mediated malaria is one of the most devastating diseases in the world, and the search for suitable antimalarial drugs remains an extraordinary challenge for scientists working in this area. Novel unconventional approaches could reveal new potential targets that may be useful for the treatment of malaria. We used a bioinformatics approach to analyze the entire genome of the *Pf*3D7 strain. Because the carbon (C-) content is a pivotal parameter that determines the hydrophobicity of a protein, which in turn controls protein folding and function, we analyzed the entire *Pf*3D7 proteome based on the gene's thymine (T)-controlled amino acid expression. Our data disclose a total of 14 proteins encoded by chromosome-4 and chromosome-9 that have an outstanding T-encoded and C-controlled hydrophobic character. The identification of these proteins could open new pivotal drug-targeting avenues.

Contact: klaus@hanyang.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 5, 2013; revised on October 3, 2013; accepted on October 4, 2013

1 INTRODUCTION

The most prevalent and lethal malaria parasite that infects humans, *Plasmodium falciparum* (*Pf*), may have recently undergone host switching to become a *Plasmodia* adaptation that is specific for humans (Baron *et al*., 2011). *P.falciparum* is the most deadly of the five *Plasmodium* species that cause human malaria, which is the world's second largest cause of death after tuberculosis (Cohen *et al*., 2012; Langhorne *et al*., 2008). Thus we have chosen to analyze the completely sequenced and annotated *Pf*3D7 strain. The *Pf*3D7 nuclear genome is composed of 22.8 megabases, which are distributed among 14 chromosomes (Chrs) (Gardner *et al*., 2002).

Despite the completion of the *Pf*3D7 genome and comparative proteomics studies (Florens *et al*., 2002; Lasonder *et al*, 2002), we are far from a strategy for curing this disease because there are still too many uncharacterized proteins to obtain a reasonable understanding of the disease-causing activity of *P.falciparum*.

*To whom correspondence should be addressed.

The base thymine (T) defines the stability of an organism at the nucleotide level and, in turn, also controls the proteins' carbon (C-) content and thus hydrophobicity for stability and activity. Hydrophobic residues are important for dictating topological features of proteins because hydrophobic interactions are the dominant force in proper protein folding and in organizing the self-assembly of water-soluble globular proteins (Aftabuddin and Kundu, 2007; Dyson *et al*., 2006).

In the current study, we analyzed the entire *Pf*3D7 genome with regard to the T-content to identify potential novel proteins that could have pivotal significance in *Pf*3D7-mediated malaria.

2 METHODS

2.1 Dataset

Coding regions (CDRs) of the complete set of 5334 messenger RNA sequences available for the *Pf*3D7 species studied here were downloaded from a public database (http://www.ncbi.nlm.nih.gov) (Supplementary Fig. S1).

2.2 Thymine distribution calculation

Using the programing language Perl, T-distribution analyses were conducted for all of the CDRs of the genes of each chromosome of *Pf*3D7 {fraction of T = [number of T residues in the gene (CDR)]/[total number of nucleotides in the gene (CDR)]} (Supplementary Fig. S2).

2.3 Calculating the absolute and relative amino acid numbers and protein lengths encoded by chromosome-1-14 of *Pf*3D7

We used previously described vectors (Hua and Sun, 2001) to calculate the proteins' length and amino acid (AA) composition (Supplementary Figs S3–S6).

2.4 Hydrophobic and hydrophilic AA distribution in the *Pf*3d7 proteome

We calculated the protein hydrophobicity under the consideration that XTX-encoded (X = A, T, G or C; T = thymine) AAs, such as phenylalanine, isoleucine, leucine, methionine and valine (FILMV; group A), and non–XTX-encoded AAs, such as glycine, alanine, proline, cysteine and tryptophan (GAPCW; group B), were considered separately. The composition FILMV = [(sum of the residues F+I+L+M+V in a protein)/(total number of residues in the protein)] was calculated for all of the

proteins encoded by all of the 14 chromosome of *Pf*3D7, followed by calculation of the average composition for each chromosome. Similarly, the composition of GAPCW = [(sum of residues G+C+A+P+W in a protein)/(total number of residues in the protein)] (Supplementary Figs S7 and S8).

2.5 Carbon distribution in all of the proteins of *Pf*3d7

We calculated the C-content in all of the AAs and proteins of *Pf*3D7 by translating the protein sequences into their corresponding atomic sequences and applying the carbon distribution (CARd) program (Rajasekaran, 2012) (Supplementary Fig. S9).

3 RESULTS

3.1 Thymine distribution and protein length encoded by all 14 chromosomes of *Pf*3D7

Chr-4 has a relatively low T-fraction, whereas the T-content in Chr-9 is rather high. Considering the length of the proteins (Chr-4 encodes for relatively long proteins), we show that the amounts of the AAs asparagine (D) and glutamate (glutamic acid, E) are relatively high, but the amount of phenylalanine (F) is relatively low in the proteins encoded by Chr-4 (Supplementary Figs S2–S6).

3.2 Hydrophobic residue compositions of all of the proteins of *Pf*3D7

On the one hand, Chr-4 has relatively few XTX-encoded hydrophobic AAs, whereas Chr-9 has a relatively high number of XTX-encoded hydrophobic AAs. On the other hand, Chr-4 and Chr-9 encode for a moderate number of non–XTX-encoded hydrophobic AAs, whereas Chr-12 encodes for a low number of non–XTX-encoded hydrophobic AAs (Supplementary Figs S7 and S8).

3.3 Carbon fraction analysis of all of the proteins of *Pf*3D7

Furthermore, we found that the proteins encoded by Chr-4 have a lower protein C-content than the proteins of all of the other chromosomes, whereas Chr-1 and Chr-9 encode for proteins with a relatively high C-content (Supplementary Fig. S9).

3.4 Identification of 10 proteins encoded by Chr-4 of *Pf*3D7

Recognizing that Chr-4 has a low T-content overall, encodes for a relatively high average protein length with a low content of XTX-encoded hydrophobic AAs and a low C-content (Supplementary Figs S2, S6, S7 and S9), we examined 10 proteins encoded by this Chr-4 more closely. We found that the relatively high XTX-encoded hydrophobicity of proteins 8 (329 AAs) and 10 (778 AAs) could be explained by their transmembrane (TM) domains, and that of protein-9 could be explained by its length (1111 AAs). The particularly high XTX-encoded hydrophobicity of protein-7 [413 AAs, probably *S*-adenosyl-methionine-dependent methyltransferase (SAM- or AdoMet-MTase, class I)], however, cannot be explained by its length nor a TM domain (Fig. 1; Supplementary Fig. S10 and Supplementary Tables S1 and S3).

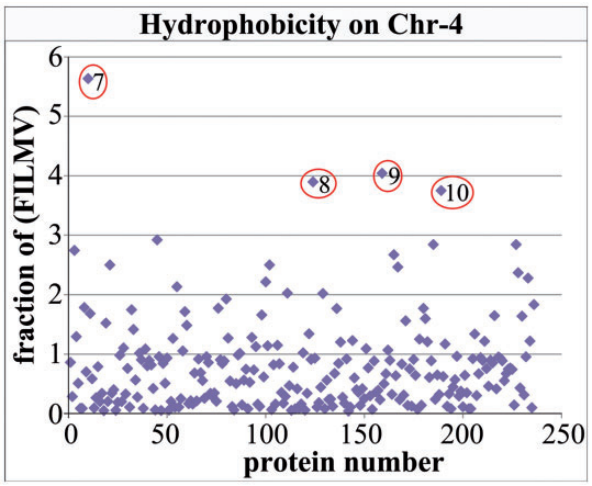

Fig. 1. Identified proteins 7–10 (red circle). The XTX-encoded hydrophobicity of all of the 236 proteins encoded by Chr-4 of *Pf*3D7. Composition of FILMV = {[(sum of residues F+I+L+M+V in a protein)/(total number of residues in the protein)]/(total number of residues in the protein)} × 1000. Protein-7: 413 AAs; protein-8: 329 AAs (TM); protein-9: 1111 AAs; protein-10: 778 AAs (TM)

3.5 Proteins 11–14 encoded by Chr-9 of *Pf*3D7

We also examined Chr-9 more closely because it had the highest content of XTX-encoded hydrophobic AAs; its carbon content was also high (Supplementary Figs S7 and S9). Among the 375 XTX-encoded hydrophobic proteins, we determined that proteins 11–14 had a particularly high FILMV fraction (Fig. 2, Supplementary Fig. S11 and Supplementary Tables S2, S4–S6).

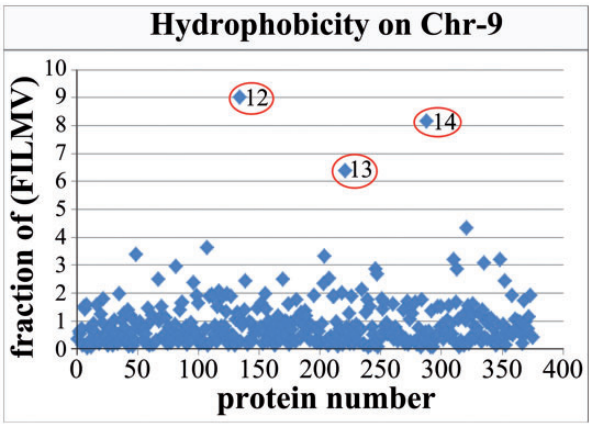

Fig. 2. Identified proteins 12–14 (red circle). The XTX-encoded hydrophobicity of all of the 375 proteins encoded by Chr-9 of *Pf*3D7. Composition of FILMV = {[(Sum of residues F+I+L+M+V in a protein)/(total number of residues in the protein)]/(total number of residues in the protein)} × 1000. Protein-12 = 235 AAs; protein-13 = 982 AAs (TM); protein-14 = 521 AAs, (TM)

It is noteworthy that proteins 1–14 are all relatively rich in Asn (N, polar AA, uncharged side chain) and/or Lys (K, basic AA charged side chain); both of these AAs are non-hydrophobic AAs that have an additional NH_2 group in their side chains. It has already been shown that *Pf* has a higher Asn frequency than other *Plasmodium* species (Yadav and Swati, 2012).

Our approach also identified the well-known protein rifin. Independent of the length of the proteins and across chromosomes 4 and 9, we found three ATPases (proteins 9, 10 and 12) and several membrane-associated proteins, all of which are further discussed below.

4 DISCUSSION

4.1 XTX- and non-XTX–encoded AAs and C-content in the *Pf*3D7 proteome

We found that chromosomes 4 and 7 of *Pf*3D7 had the lowest level of XTX-encoded AAs and the lowest C-content. In contrast, Chr-9 has the highest level of XTX-encoded AAs and the highest C-content (Supplementary Figs S7 and S9) and hence could encode for more globular proteins. Most of the proteins on these chromosomes have not been characterized yet. However, their characterization is of utmost importance because it is tempting to speculate that the relative T-encoded C-content of Chr-4 could account for the higher virulence of *Pf* (Fard *et al*., 2009).

The 14 proteins disclosed here have a divergent hydrophobic feature, thus attracting special attention for future functional investigations. Noticeably, protein-3 (its Cwc25 domain indicates its involvement in catalyzing spliceosome reactions) has an outstanding C-content, whereas the overall low T-content of gene-14 (serine/threonine protein kinase, FIKK family) correlates with a low protein C-content, therefore providing these two proteins exceptional qualities (Fig. 3).

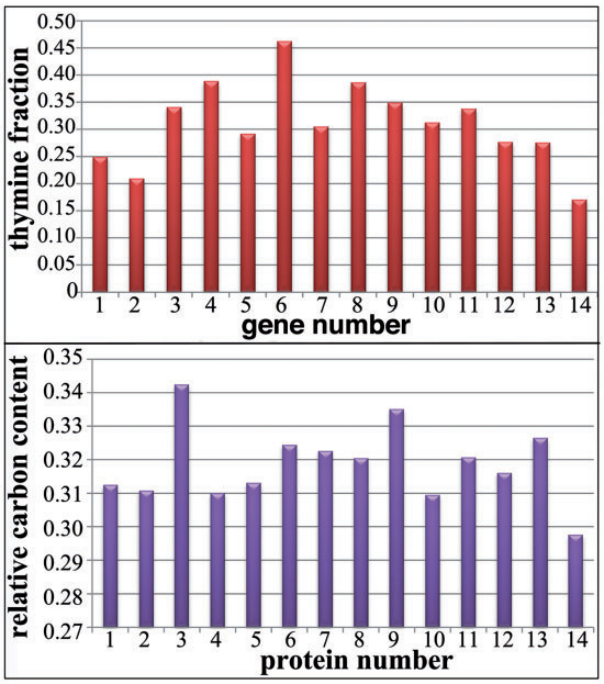

Fig. 3. Top: relative T-distribution in all of the 14 genes (Supplementary Tables S1–S6) (open reading frames only). In each gene, the numbers of A, C, G and T residues were calculated. The total count of A, C, G and T was then calculated. The number of T was then divided by the total number of [T/(A+C+G+T)]. Bottom: the relative C-content of all of the 14 proteins. For each protein, the total number of C-atoms was divided by the total number of atoms: [C/(C+S+N+O+H)]

4.2 Protein-7: AdoMet-MTase

AdoMet-MTases use SAM as a substrate for methyl transfer to create the product *S*-adenosyl-*L*-homocysteine (AdoHcy), which is used by AdoHcy hydrolases. AdoMet-MTases and AdoHcy hydrolases are currently under intense investigation as putative drug targets for the treatment of malaria (Bobenchik *et al*., 2011; Mpangase *et al*., 2013; Singh *et al*., 2013; Tanaka *et al*., 2013). The AdoMet-MTases are putatively redox-regulated proteins in malarial parasites and provide a functional link between AdoMet-MTases (protein-7, Chr-4) and Tlp2 (protein-11, Chr-9) (Sturm *et al*., 2009).

4.3 Protein-8: Rifin

Rifin and stevor are proteins with variable surface antigens that are uniquely found in the malaria parasites *P.falciparum* and *Plasmodium reichenowi*. These two pathogens have evolved strategies to survive within the hosts they infect by varying the antigens they expose to the host immune system, which usually results in the proliferation of multicopy protein families that are commonly called variable surface antigens (Joannin *et al*., 2011; Templeton, 2009).

4.4 Protein-9: AAA^+ ATPase

This AAA^+ ATPase contains a von Willebrand factor type A domain [MDN1, midasin homolog; the members of this subgroup have a conserved MIDAS (metal ion-dependent adhesion site) motif; however, the function of this domain is not known]. AAA^+ ATPases are involved, *inter alia*, in promoting the degradation of ubiquinated proteins by the proteasome (Chou *et al*., 2010). In *P.falciparum*, these proteins have also been described as being involved in DNA replication and transcription, chromatin remodeling, genome stability (Ansari and Tuteja, 2012) and intra-erythrocytic schizogony (Ahmad *et al*., 2012) and also as being part of a novel export machine in malarial parasites (de Koning-Ward *et al*., 2009),

4.5 Protein-10: ATPase

Protein-10 contains a TM- and a LITAF-ABC2 domain. The *P.falciparum* ATP-binding cassette (ABC)-type transporter superfamily is composed of 16 members, and current knowledge of their physiological function and contribution to antimalarial drug resistance is limited (Alcantara *et al*., 2013; Koenderink *et al*., 2010; Sauvage *et al*., 2009). In humans, an association of ABCA1 promoter polymorphisms with susceptibility to severe malaria has been reported (Sahu *et al*., 2013). The *P.falciparum* multidrug resistance gene 1 (pfmdr1) encodes a food vacuolar membrane transporter and has been linked with parasite drug resistance. *P.falciparum* develops resistance by gradually acquiring genetic polymorphisms that decrease drug susceptibility (Malmberg *et al*., 2013).

4.6 Protein-11: Tlp2

P.falciparum possesses a functional thioredoxin and glutathione system comprising the dithiol-containing redox proteins thioredoxin (Trx), glutaredoxin (Grx) and plasmoredoxin (Plrx) (Rahlfs *et al*., 2003). To maintain the intracellular redox balance,

P.falciparum uses this complex thioredoxin and glutathione system (Jortzik and Becker, 2012). Among their target proteins, several enzymes involved in SAM metabolism have been described, which suggests that redox control is required to balance the metabolic fluxes of SAM between methyl group transfer reactions and polyamine synthesis (Sturm *et al*., 2009).

4.7 Protein-12: V-ATPase

Ion transport processes are expected to play a role in both pathogenicity and adaptation. The V-type H^+-ATPase is critical during the intra-erythrocytic stage of the human malaria parasite *P.falciparum*; it is responsible for maintaining a near-neutral cytosolic pH (pH 7.3), an acidic digestive vacuole (pH 4.5–5.5) and an inside-negative plasma membrane potential. Disruption of parasite pH regulation through the inhibition of its V-type H^+-ATPase results in parasite death and has thus been considered as an antimalarial approach (Moriyama *et al*., 2003; van Schalkwyk *et al*., 2010).

4.8 Protein-13

Protein-13 is a TM protein with a putative rpoC2 domain that is usually found in DNA-directed RNA polymerase subunits. Localization of RNA replication to intracellular membranes, such as the mitochondria, is a universal feature of positive-strand RNA viruses and has been described previously (Miller and Ahlquist, 2002).

In conclusion, we disclose 14 proteins of the *Pf*3D7 proteome with a pivotal T-encoded and C-controlled hydrophobic feature that determines the proteins' folding, structure and function. Further molecular and cellular experiments could confirm these molecules as functional therapeutic targets for the treatment of *P.falciparum*-induced malaria.

Funding: This study was supported by Hanyang University.

Conflict of Interest: none declared.

REFERENCES

Aftabuddin,M. and Kundu,S. (2007) Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophys. J.*, **93**, 225–231.

Ahmad,M. *et al.* (2012) Novel RuvB nuclear ATPase is specific to intraerythrocytic mitosis during schizogony of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.*, **185**, 58–65.

Alcantara,L.M. *et al.* (2013) Chemosensitization potential of P-glycoprotein inhibitors in malaria parasites. *Exp. Parasitol.*, **134**, 235–243.

Ansari,A. and Tuteja,R. (2012) Genome wide comparative comprehensive analysis of *Plasmodium falciparum* MCM family with human host. *Commun. Integr. Biol.*, **5**, 607–615.

Baron,J.M. *et al.* (2011) A revised timeline for the origin of *Plasmodium falciparum* as a human pathogen. *J. Mol. Evol.*, **73**, 297–304.

Bobenchik,A.M. *et al.* (2011) Phosphoethanolamine methyltransferases in phosphocholine biosynthesis: functions and potential for antiparasite therapy. *FEMS Microbiol. Rev.*, **35**, 609–619.

Chou,T.F. *et al.* (2010) Selective, reversible inhibitors of the AAA ATPase p97. In: *Probe Reports from the NIH Molecular Libraries Program*. National Center for Biotechnology Information, Bethesda, MD.

Cohen,J.M. *et al.* (2012) Malaria resurgence: a systematic review and assessment of its causes. *Malar. J.*, **11**, 122.

de Koning-Ward,T.F. *et al.* (2009) A newly discovered protein export machine in malaria parasites. *Nature*, **459**, 945–949.

Dyson,H.J. *et al.* (2006) The role of hydrophobic interactions in initiation and propagation of protein folding. *Proc. Natl Acad. Sci. USA*, **103**, 13057–13061.

Fard,A.T. *et al.* (2009) In silico comparative genome analysis of malaria parasite *Plasmodium falciparum* and *Plasmodium vivax* chromosome 4. *Parasitol. Res.*, **104**, 1361–1364.

Florens,L. *et al.* (2002) A proteomic view of the *Plasmodium falciparum* life cycle. *Nature*, **419**, 520–526.

Gardner,M.J. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498–511.

Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Joannin,N. *et al.* (2011) RSpred, a set of Hidden Markov Models to detect and classify the RIFIN and STEVOR proteins of *Plasmodium falciparum*. *BMC Genomics*, **12**, 119.

Jortzik,E. and Becker,K. (2012) Thioredoxin and glutathione systems in *Plasmodium falciparum*. *Int. J. Med. Microbiol.*, **302**, 187–194.

Koenderink,J.B. *et al.* (2010) The ABCs of multidrug resistance in malaria. *Trends Parasitol.*, **26**, 440–446.

Langhorne,J. *et al.* (2008) Immunity to malaria: more questions than answers. *Nat. Immunol.*, **9**, 725–732.

Lasonder,E. *et al.* (2002) Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature*, **419**, 537–542.

Malmberg,M. *et al.* (2013) *Plasmodium falciparum* drug resistance phenotype as assessed by patient antimalarial drug levels and its association with pfmdr1 polymorphisms. *J. Infect. Dis.*, **207**, 842–847.

Miller,D.J. and Ahlquist,P. (2002) Flock house virus RNA polymerase is a transmembrane protein with amino-terminal sequences sufficient for mitochondrial localization and membrane insertion. *J. Virol.*, **76**, 9856–9867.

Moriyama,Y. *et al.* (2003) Vacuolar proton pumps in malaria parasite cells. *J. Bioenerg. Biomembr.*, **35**, 367–375.

Mpangase,P.T. *et al.* (2013) Discovery-2: an interactive resource for the rational selection and comparison of putative drug target proteins in malaria. *Malar. J.*, **12**, 116.

Rahlfs,S. *et al.* (2003) *Plasmodium falciparum* thioredoxins and glutaredoxins as central players in redox metabolism. *Redox Rep.*, **8**, 246–250.

Rajasekaran,E. (2012) CARd: carbon distribution analysis program for protein sequences. *Bioinformation*, **8**, 508–512.

Sahu,U. *et al.* (2013) Promoter polymorphisms in the ATP binding cassette transporter gene influence production of cell-derived microparticles and are highly associated with susceptibility to severe malaria in humans. *Infect. Immun.*, **81**, 1287–1294.

Sauvage,V. *et al.* (2009) The role of ATP-binding cassette (ABC) proteins in protozoan parasites. *Mol. Biochem. Parasitol.*, **167**, 81–94.

Singh,D.B. *et al.* (2013) Docking and in silico ADMET studies of noraristeromycin, curcumin and its derivatives with *Plasmodium falciparum* SAH hydrolase: a molecular drug target against malaria. *Interdiscip. Sci.*, **5**, 1–12.

Sturm,N. *et al.* (2009) Identification of proteins targeted by the thioredoxin superfamily in *Plasmodium falciparum*. *PLoS Pathog.*, **5**, e1000383.

Tanaka,N. *et al.* (2013) [Structural biology for developing antimalarial compounds]. *Yakugaku Zasshi*, **133**, 527–537.

Templeton,T.J. (2009) The varieties of gene amplification, diversification and hypervariability in the human malaria parasite, *Plasmodium falciparum*. *Mol. Biochem. Parasitol.*, **166**, 109–116.

van Schalkwyk,D.A. *et al.* (2010) Inhibition of *Plasmodium falciparum* pH regulation by small molecule indole derivatives results in rapid parasite death. *Biochem. Pharmacol.*, **79**, 1291–1299.

Yadav,M.K. and Swati,D. (2012) Comparative genome analysis of six malarial parasites using codon usage bias based tools. *Bioinformation*, **8**, 1230–1239.