Whole-genome sequencing and analysis of Plasmodium falciparum isolates from China-Myanmar border area

Hai-Mo Shen, Shen-Bo Chen, Yan-Bing Cui, Bin Xu, Kokouvi Kasenge, Eniola Michael Abe, Yue Wang, and Jun-Hu Chen

Abstract

Background: China has made progress in malaria control and aims to eliminate malaria nationwide, but implementing effective interventions along the border regions remain a huge task. The Plasmodium falciparum cases imported from Southeast Asia has frequently reported especially in the China-Myanmar border (CMB) area. Though, information is scant on P. falciparum genetic variability in this area.

Methods: This study reported P. falciparum isolates genome sequence of six clinical isolates in the CMB area. Furthermore, we estimated the nucleotide diversity, Watterson’s estimator and Tajima’s D value for the whole genome mutation rate in slide window.

Results: Our data were aligned onto 96.05–98.61% of the reference 3D7 genome in high fold coverages. Principal component analysis result showed that P. falciparum clustered generally according to their geographic origin. A total of 91 genes were identified as positive selection with Ka/Ks ratio significantly higher than 1, and most of them were multigene families encoding variant surface antigens (VSAs) such as var, rif and stevor. The enrichment of the positive selection on VSA genes implied that the environment complexity subjected CMB’s P. falciparum to more pressure for survival.

Conclusions: Our research suggests that greater genetic diversity in CMB area and the positive selection signals in VSA genes, which allow P. falciparum to fit the host immune system well and aggravate the difficulty of treatment. Meanwhile, results obtained from this study will provide the fundamental basis for P. falciparum population genomic research in CMB area.

Keywords: Plasmodium falciparum, Genome, Variant surface antigen, China-Myanmar border

Multilingual abstracts

Please see Additional file 1 for translations of the abstract into the five official working languages of the United Nations.

Background

Malaria was one of the most prevalent parasitic diseases in the Greater Mekong sub-region (GMS) historically.

© The Author(s). 2018 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
to identify risk factors for clinical malaria [5]. It showed that Myanmar does not only share border but also export malaria to China, which seriously impedes the progress of malaria elimination [6]. In addition to the remarkable increase in the number of cases, imported *Plasmodium falciparum* malaria also fuel concerns of re-introduction, and this had been shown to be partly attributed to increases in the intrinsic potential for malaria transmission [7]. Also, antimalarial drug resistance (parasites resistance to chloroquine and pyrimethamine) was first experienced in Southeast Asia before it spread out to Africa. A recent study used in vitro drug assay data from CMB area as phenotypes in genome-wide association study and found several loci associated with in vitro drug resistance to multiple antimalarials [8]. Therefore, there is urgent need to curtail the continued resistance by malaria parasite in the region.

Whole genome studies provide us with the opportunity to develop new control methods, including new drugs and vaccines, improved diagnostics and effective vector control techniques [9]. Several hundreds of *P. falciparum* isolates whole genome had been sequenced in the past, most of them were meant to identify the artemisinin resistance loci [10, 11]. There is paucity of information on whole-genome sequencing of malaria isolates from CMB area. Hence, we performed the whole genome sequencing of *P. falciparum* isolates from CMB area. Genomic DNA was extracted using the QIAGEN DNeasy Blood & Tissue Prep Kit (Illumina, San Diego, USA), and the direct sequencing libraries with Illumina TruSeq DNA LT Sample Prep Kit (Illumina, San Diego, USA), and the direct sequencing approach was used in our previous study [13]. We filtered all reads by removing the adapter sequences and low quality sequences with Trimmomatic-3.0 [14]. In addition, the genome and annotation data of the 3D7 reference from PlasmoDB database (http://plasmodb.org) was downloaded [15].

### Identification of SNPs from *P. falciparum* isolates and data analysis

Sequencing reads were mapped from all the six samples of *P. falciparum* 3D7 genome using Burrows-Wheeler Aligner and Sequence Alignment/Map (SAMtools-1.3) [16], which is the most complete whole genome standard reference. The genotyping was performed using in-house pipeline based on GATK and SnpEff workflows [17]. We performed the principal component analysis (PCA) of all samples and compared the single nucleotide polymorphism (SNP) with those of 34 *P. falciparum* isolates collected worldwide [18, 19]. Then, the nucleotide diversity ($\pi$), Watterson’s estimator ($\theta_w$) for the whole genome mutation rate in 4 kb sliding window and 2 kb step across each chromosome were estimated in ARLEQUIN-Ver3.5 [20]. Also, the Tajima’s D value for each sliding window and the corresponding gene were calculated. Here, the Tajima’s D test help us to distinguish between genes evolving neutrally and under pressure, including selection, demographic expansion or genetic hitchhiking. Genome alignment data were further used to calculate the gene non-synonymous (Ka) and synonymous (Ks) substitution rates with NG (Nei & Gojobori) and YN (Yang and Nielsen) model [21, 22] in KaKs_Calculator1.2 [23], which helps to estimate the balance between neutral mutations, purifying selection and beneficial mutations acting on homologous.

### Results

**Whole genome sequencing of parasites and mapping**

We sequenced the library on Illumina HiSeq X10 (Illumina, San Diego, USA) and generated 78 to 248 M paired-end reads of 150 bp. Illumina sequencing reads have been submitted to the NCBI Short Read Archive (Bio-Project no. PRJNA393218). A variable proportion of reads (4.08–27.08%) from all the isolates were mapped to the reference, and aligned onto 96% of the reference genome in high fold coverage (20.57–121.78×). A total of 369 700 SNPs were captured and only 18 953 common loci were available for analysis after quality filtering (Table 1). PCA of all strains was performed to assess the regional differences. As part of the Asia isolates, the CMB isolates illustrated a higher discrepancies with the 3D7
genome (Fig. 1a). The major axis of differentiation (F1) showed that *P. falciparum* clustered generally according to their geographic origin and the Asian samples exhibited greater genetic diversity than African. The second and third principal components (F2 and F3) defined a distinct South-American cluster and distinguished the African samples better (Fig. 1b). It is important to note that CMB samples were widely separated in our PCA result, suggesting higher diversity from border area.

**Scan of important gene under selection**

On genome scale, we estimated $\bar{\pi}$ to be 0.0205 and $\bar{\theta}_{\omega}$ to be 0.0287, and genetic diversity was lower in exonic regions but higher in intronic and intergenic regions. As expected, the Watterson’s estimator ($\bar{\theta}_{\omega}$) is higher than global samples (where $\bar{\theta}_{\omega}$ has been estimated to be 1.03 $\times$ 10$^{-3}$ using isolates from Africa, America, Asia and Oceania) [24]. Mean pairwise divergence is higher in gene families associated with red blood cell invasion and immune evasion. The Tajima’s D values obtained were mostly negative, with a mean value of −1.76, and 103 genes (2.5%) have positive values (Fig. 2).

Of the 5566 genes analyzed, 3485 have valid Ka/Ks value, while only 91 genes were identified as positive selection with Ka/Ks ratio significantly higher than 1 (Fig. 2). However, a high proportion of genes encoding variant surface antigens like the *var* genes, the repetitive interspersed family (*rif*) and the subtelomeric variant open reading frame (*stevor*) was found among these 91 positive selection genes. The variant antigens *P. falciparum* erythrocyte membrane protein 1 (PIEMP1) and repetitive interspersed families of polypeptides (RIFINs) are adhesins implicated in severe *P. falciparum* malaria, and STEVORS are erythrocyte binding protein that mediates merozoite invasion and resetting [25–28].

**Table 1** Sequencing and mapping summary of 6 isolates of *Plasmodium falciparum* from China-Myanmar border area

| Samples | Pf7       | Pf54      | Pf87       | Pf237     | Pf289     | Pf297     |
|---------|-----------|-----------|-----------|-----------|-----------|-----------|
| Number of clean reads | 101,315,479 | 140,050,984 | 248,893,994 | 78,515,613 | 80,809,032 | 94,593,018 |
| Mapped on *P. falciparum* | 5,430,602 | 6,342,044 | 10,150,675 | 21,265,252 | 10,626,161 | 5,929,181 |
| Mapped (%) | 5.36 | 4.53 | 4.08 | 27.08 | 13.15 | 6.27 |
| Mean mapping quality | 47.37 | 46.59 | 46.07 | 53.2 | 51.98 | 48.88 |
| Coverage fold | 20.57 | 22.76 | 35.50 | 121.78 | 54.68 | 23.98 |
| Genome covered > 1 (%) | 96.43 | 96.05 | 96.96 | 98.61 | 98.04 | 96.10 |
| Variation | | | | | | |
| Raw SNP | 53,314 | 50,032 | 58,753 | 83,456 | 70,415 | 53,730 |
| Filtered SNP | 11,838 | 10,928 | 606 | 15,410 | 14,361 | 1230 |
| Filtered Indel | 15,657 | 11,474 | 3094 | 32,402 | 27,814 | 8953 |

SNP single nucleotide polymorphism, Indel insertion-deletion

**Discussion**

Recent studies in Myanmar revealed that malaria incidence and mortalities showed significant decreasing trend even in artemisinin-resistant areas [29]. The effective collaboration between China and Myanmar promptly carried out the interventions through simplified processes, and dramatically decreased malaria burden in CMB area [30]. Meanwhile, the relative role of the hidden reservoir of resistant parasites needs to be assessed, particularly in regions that are low-transmission settings and pre-elimination phases [31]. The control of malaria requires intensive efforts, which should be guided by a thorough understanding of adaptive processes occurring in pathogen populations in different endemic areas. Malaria transmission intensity and parasite genetic diversity are known to vary greatly in different parts of Southeast Asia due to variation in rainfall abundance and seasonality [32]. Our result showed that the genetic diversity estimated from CMB *P. falciparum* isolates is higher than the global samples. The genetic structure of CMB parasites was similar with other Asian countries despite the vector species abundance generally enhance the environmental compatibility of parasites [32–35]. Meanwhile, we performed the Tajima’s test to identify genes not fit for the neutral model at equilibrium. These predominantly negative values were remarkably similar to previous analyses, which indicated population expansion of *P. falciparum* in Africa [36]. The greater genetic diversity and selection signals in genes associated with red blood cell invasion and drug resistance are consistent with previous research [37] and also suggested the malaria control programs of Myanmar imposed huge pressure on *P. falciparum* in CMB area and play an important role in the process of diversification.

On the other, it is advantageous to apply positive selection test where environment factors apply consistent...
Fig. 1 (See legend on next page)
Fig. 1 Principal component analysis based on 205,189 common SNP loci in CMB isolates and reference strains. The samples are dyed by their geographic origin: red for South-American, blue for African, grey for Asian, and pink for CMB area. **a**. The major fact (F1) of differentiation showed that *P. falciparum* clustered generally according to their geographic origin. **b**. The second and third facts (F2 and F3) define a distinct South-American cluster and distinguished the African samples better according to their locations.

Fig. 2 Genomic map of *P. falciparum* CMB isolates. Segments from outside to inside: The outer circle of the diagram depicts the whole genome of 3D7 reference strain and arranged in chromosome order. The second track shows the histogram of $\pi$ in 4 kb windows. The third track shows the Tajima's D value. The inner circle shows the Ka/Ks ratio of each gene. This figure was performed using Circos [41].
pressure over generations in favor of specific beneficial trait. Early analyses identified loci showing evidence of recent positive directional selection and balancing selection confirmed that antimalarial drugs and host immunity have been major selective agents [38]. In our study, genes encoding variant surface antigens exhibit greater diversity and positive selection than other genes. These genes play an important role in \textit{P. falciparum} malaria pathogenesis and in immune evasion by the malaria parasite. Studies in Africa have shown that severe malaria is associated with the ability of erythrocytes infected with the parasite to bind uninfected erythrocytes and form rosettes [39]. It is well-known that different members of these protein families bind to different adhesion receptors. The proteins are antigenically unique and switching of the individual PfEMP1 proteins during an infection is important for the maintenance of chronic infections. The enrichment of the positive Ka/Ks values on these genes is consistent with the previous reports that the high degree of diversity could help \textit{P. falciparum} to fit the host immune system better. For example, STEVOR plays a role in creating antigenic diversity of schizont stage parasites, thereby adding additional complexity to the immunogenic properties of the infected red blood cell [40]. In our result, a total of 17 STEVOR genes have got valid value higher than 1 in Ka/Ks test, and 6 of them were under significant positive selection. These STEVOR genes offered enough modifications which enable the parasite to establish long-lasting chronic infection by evading antibody mediated immune recognition and splenic clearance. Similar selection pattern arose on \textit{rfl} genes family which contribute to the rosetting of \textit{P. falciparum} mediated by blood antigen and help to express clonally variant antigens at the surface of the infected erythrocyte. Among all the 158 \textit{rfl} genes downloaded from plasmoDB, 66 genes have got valid value higher than 1, with 12 of them under significant positive selection. The signatures of positive selection suggested that the local environment complexity subjects CMB's \textit{P. falciparum} to more pressure for survival [26, 27].

Conclusions

Our study assessed the genome sequences of six clinical isolates of \textit{P. falciparum} in the CMB area. We found greater genetic diversity in CMB area and the positive selection signals in variant surface antigens genes. As gene flow is relatively unrestricted in East-south Asia, highly recombining populations of the \textit{P. falciparum} are closely related, but markedly varying ecology and transmission intensity should cause distinct local selective pressures. Findings from this study provide more insight on the current epidemiological scenario of malaria in China. Lastly, our results will help deepen our understanding of \textit{P. falciparum} evolution, and also provide the fundamental basis for further studies.
3. Feng J, Zhang L, Huang F, Yin JH, Tu H, Xia ZG, et al. Ready for malaria elimination: zero indigenous case reported in the People’s republic of China. Malar J. 2018;17:315.

4. Yin JH, Zhou SS, Xia ZG, Wang RB, Qian YJ, Yang WZ, et al. Historical patterns of malaria transmission in China. Adv Parasitol. 2014;86:1–19.

5. Zhou X, Huang JI, Njage MT, Li SG, Chen JH, Zhou XN. A molecular survey of febrile cases in malaria-endemic areas along China-Myanmar border in Yunnan province, People’s republic of China. Parasite. 2014;21:27.

6. Zhou G, Sun L, Xia R, Duan Y, Xu J, Yang H, et al. Clinical malaria along the China-Myanmar border, Yunnan Province, China, January 2011-august 2012. Emerg Infect Dis. 2014;20:575–8.

7. Wang D, Li S, Cheng Z, Xiao N, Cotter C, Hwang J, et al. Transmission risk from imported Plasmodium vivax malaria in the China-Myanmar border region. Emerg Infect Dis. 2015;21:1861.

8. Shen HM, Chen SB, Wang Y, Xu B, Abe EM, Chen JH. Genome-wide scans for the human malaria parasite Plasmodium falciparum. Nature. 2002;419:498.

9. Gardner MJ, Hall N, Fung E, Bertram M, Hyman RW, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature. 2002;419:498.

10. Miotto O, Almagro-Garcia J, Manske M, MacInnis B, Campino S, Rockett KA, et al. Genome-wide analysis of selection on the malaria parasite Plasmodium falciparum in Cambodia. Nat Genet. 2013;45:648–55.

11. Takala-Harrison S, Clark TC, Jacob CG, Cummings MP, Miotto O, Dondorp AM, et al. Multiple populations of artemisinin-resistant Plasmodium falciparum following artemisinin treatment in Southeast Asia. Proc Natl Acad Sci U S A. 2013;110:240–5.

12. Shi B, Zheng J, Qiu H, Yang GJ, Xia S, Zhou XN. Risk assessment of malaria transmission at the border area of China and Myanmar. Infect Dis Poverty. 2017;6:108.

13. Shen H, Chen SB, Wang Y, Xu B, Abe EM, Chen JH. Genome-wide scans for the identification of Plasmodium vivax genes under positive selection. Malar J. 2017;16:238.

14. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114.

15. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986;3:418–20.

16. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and windows. Mol Ecol Resour. 2010;10:564–7.

17. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

18. Aurrecoechea C, Breitkrell J, Brunk BP, Dommern J, Fischer S, Gajria B, et al. PlasmaDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 2008;37:539–43.

19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.

20. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 2002;19:32–43.

21. Zhang Z, Li J, Zhao XQ, Wang J, Gong W, Yu J, KaKs Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics. 2006;4:259–63.

22. Nygaard S, Braustein A, Malisen G, Van Dongen S, Gardner PP, Koehg A, et al. Long-and short-term selective forces on malaria parasite genomes. PLoS Genet. 2010;6:e1000199.

23. Wahlsten M, Goel S, Akhouri RR. Variant surface antigens of Plasmodium falciparum and their roles in severe malaria. Nat Rev Microbiol. 2017;15(8):479–91.

24. Bultman E, Brick K, Mukherjee S, Zhang Y, Silverstine F, Alano P, et al. Revisiting the Plasmodium falciparum RIFIN family: from comparative genomics to 3D-model prediction. BMC Genomics. 2009;10:445.

25. Koel S, Palmkvist M, Moll K, Joannin N, Lara P, Akhouri RR, et al. RFPINs are adhesins implicated in severe Plasmodium falciparum malaria. Nat Med. 2015;21:314–7.

26. Niang M, Bei AK, Madhani KG, Pelly S, Dankwa S, Kanjee U, et al. STEVOR is a Plasmodium falciparum erythrocyte binding protein that mediates merozoite invasion and rosetting. Cell Host Microbe. 2014;16:81–93.

27. Neve TQ, Oo T, Wai KT, Zhou S, Van Giessen J, Chinnakkal P, et al. Malaria profiles and challenges in artemisinin resistance containment in Myanmar. Infect Dis Poverty. 2017;6:76.

28. Xu JW, Li Y, Yang HL, Zhang J, Zhang ZX, Yang YM, et al. Malaria control along China-Myanmar border during 2007–2013: an integrated impact evaluation. Infect Dis Poverty. 2016;5:75.

29. Abdul-Ghani R, Mahdy MA, Beier JC, Basco LK. Hidden reservoir of resistant parasites: the missing link in the elimination of falciparum malaria. Infect Dis Poverty. 2017;6:12.

30. Yu G, Yan G, Zhang N, Zhong D, Wang Y, He Z, et al. The Anopheles gambiae community and the role of Anopheles minimus on malaria transmission on the China-Myanmar border. Parasit Vectors. 2013;6:264.

31. Wang Y, Ma A, Chen SB, Yang YC, Chen JH, Yin MB. Genetic diversity and natural selection of three blood-stage 6-Cys proteins in Plasmodium vivax populations from the China-Myanmar endemic border. Infect Genet Evol. 2014;28:167–74.

32. Zhu X, Zhao Z, Fung Y, Li P, Li F, Liu J, et al. Genetic diversity of the Plasmodium falciparum apical membrane antigen I gene in parasite populations from the China-Myanmar border area. Infect Genet Evol. 2016;39:155–62.

33. Wang Y, Zhong D, Cui L, Lee MC, Yang Z, Yan G, et al. Population dynamics and community structure of Anopheles mosquitoes along the China-Myanmar border. Parasit Vectors. 2015;8:445.

34. Chang HH, Park DJ, Galinsky KJ, Schaffner SF, Ndaye D, Ndir O, et al. Genetic sequencing of Plasmodium falciparum malaria parasites from Senegal reveals the demographic history of the population. Mol Ecol Biol. 2012;22:3427–39.

35. Feng J, Zhou D, Lin Y, Xiao H, Yan H, Zhou H, et al. Amplification of pfdmrl, pfcr, pvmrd1 and K13-propeller polymorphism associated with Plasmodium falciparum and Plasmodium vivax at the China-Myanmar border. Antimicrob Agents Chemother. 2015;59(5):2554–9.

36. Mobegi VA, Duffy CW, Amambua-Ngwa A, Loua KM, Laman E, Nwakanma DC, et al. Genome-wide analysis of selection on the malaria parasite Plasmodium falciparum in west African populations of differing infection endemicity. Mol Biol Evol. 2014;31:1490–9.

37. Carlson J, Helmby H, Wahlgren M, Hill A, Brewster D, Greenwood B. Human cerebral malaria: association with erythrocyte rosetting and lack of anti-rosetting antibodies. Lancet. 1990;336:1457–60.

38. Niang M, Yam XY, Peiser PR. The Plasmodium falciparum STEVOR multigene family mediates antigenic variation of the infected erythrocyte. PLoS Pathog. 2009;5:e1000307.

39. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.