Population structure and selection-mediated changes in *Plasmodium falciparum* by next-generation sequencing

Cheikh Cambel Dieng[1], Colby T. Ford[2], Dickson Doniou[3], Jennifer Huynh[1], Anita Lerch[4], Daniel Janies[2], Linda Amoah[3], Yaw Afrane[5], and Eugenia Lo[1]

[1]Dept. of Biological Sciences, [2]Dept. of Bioinformatics and Genomics - University of North Carolina at Charlotte, USA; [3]Noguchi Memorial Institute For Medical Research, University, Ghana
[4]Dept. of Biology - University of Notre Dame, USA; [5]Dept. of Medical Microbiology - School of Biomedical and Allied Health Sciences, University, Ghana;

Background

- An estimated 200 million malaria cases and 440,000 malaria deaths are reported each year worldwide.
- Today, *P. falciparum* remains the most dominant pathogen species in Sub-Saharan Africa that accounts for almost all mortality in malaria.
- To effectively control the disease in highly endemic regions as well as in children, the world's first malaria vaccine RTS,S will soon be implemented in three African countries including Ghana, Kenya, and Malawi.
- RTS,S is a recombinant protein vaccine that contains a portion of the central NANP repeats (B-cell epitopes) and the C-terminal region that contains the T-cell epitopes (Th2R and Th3R) of the *P. falciparum* circumsporozoite protein (PfCSP), similar to the sequence of the 3D7 laboratory strain.
- Previous studies have shown that RTS,S provides ~36% protection against clinical malaria in young children.
- The questions of whether RTS,S-induced immunity is PfCSP allele-dependent and selection favors non-3D7 strains are still uncertain.

Objectives

- To provide baseline data on within- and between-host parasite diversity in endemic regions of Ghana.
- To determine the population genetic structure of local P. falciparum and their relatedness with the 3D7 strain
- To examine the distribution of PfCSP variants and assess selection pressures driven by RTS,S vaccine.

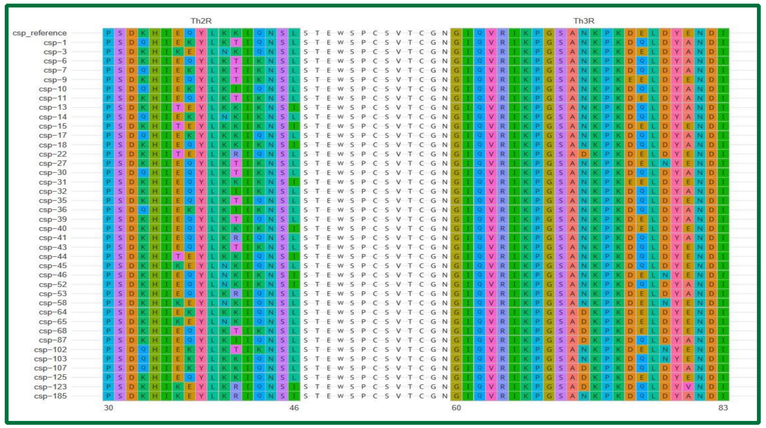


Figure 1. Sequence diversity in the Th2R and Th3R regions of PfCSP

Methods

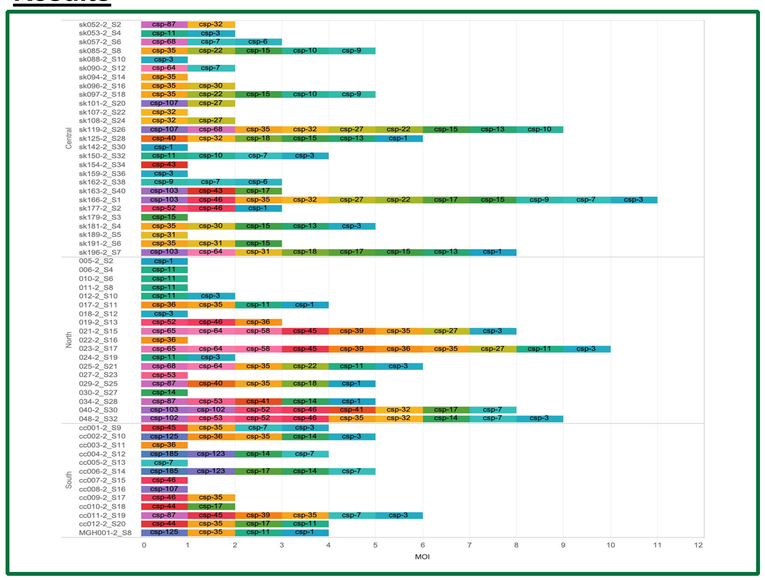
A. DNA extraction and Molecular screening of *P. falciparum*: About 200ul capillary blood samples were collected from 400 subjects in central Ghana. DNA was extracted using Zymo Quick-DNA kit following the manufacturer's protocol and samples were stored in -20°C until used. *Plasmodium* species identification and DNA quantification was performed by real-time quantitative PCR of the 18S rRNA gene for all samples.

B. PfCSP amplicon deep sequencing: The C-terminal of the PfCSP containing the Th2R and Th3R regions were amplified, and the Illumina partial adapters were added using PCR. Samples were purified using the QIAGEN kit. Multiple samples were pooled and sequenced in paired-end on the Illumina Hiseq 2500. Coverage of at least 50,000 reads were obtained for each sample amplicon.

C. Sample Haplotyping: Using the *HaplotypR* package from Lerch et al., SNPs were detected across all samples. These mutations were summarized as a final list of haplotypes based on a minimum occurrence cutoff value of 2 and mismatch rate of 50%. This list of nucleotide CSP haplotype sequences were then translated to amino acids.

D. Protein Structure Prediction: A consensus sequence of the mutations across all CSP haplotypes was created. The 3VDJ protein structure was collected from PDB and visualized as a reference (Fig. 5, A1). The mutation consensus sequence's structure was predicted using QUARK and visualized (Fig. 5, A2) and compared to the reference structure (Fig. 5, B).

Results



▲ Figure 2. Within- and between-host diversity of the PfCSP gene, showing multiplicity of infection.

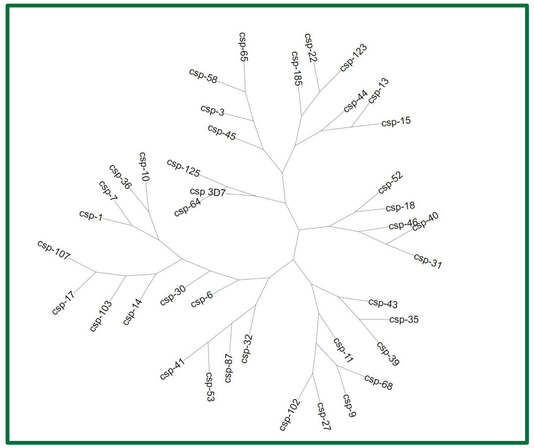
- 39 Haplotypes were detected in the 60 samples.
- The number of haplotypes found within one sample varies from 1 to 11 with sample Sk166 harboring the highest number of haplotypes. The various colors in the graph represent the different Haplotypes.
- We detect a high level of polyclonality in the majority of samples.

◄ Figure 3. Detection frequency of the different haplotypes.

(CSP-35 , CSP-3, CSP-11, CSP-7, CSP-1 are the most prevalent with 10%, 8%, 6%, and 4%, respectively.)

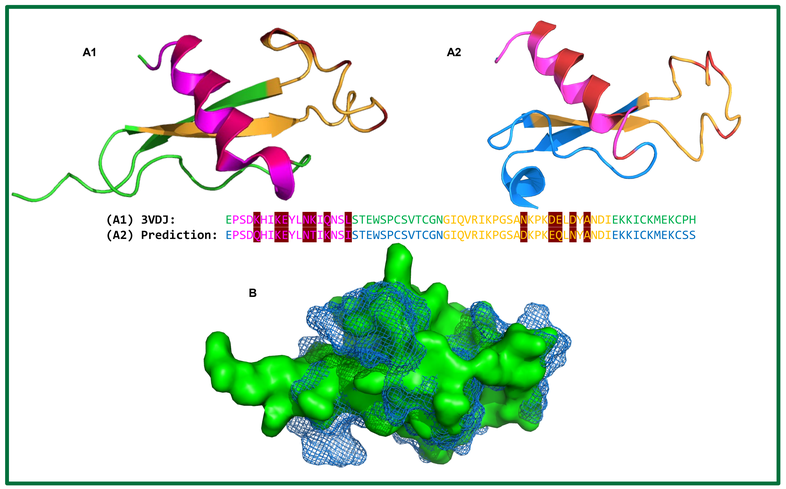
References

- Schliep K.P. (2011). phangorn: phylogenetic analysis in R. Bioinformatics, 27(4) 592-593
- Paradis E, Schliep K (2019). "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R." Bioinformatics, 35, 526-528.
- Bodenhofer U, Bonatesta E, Horejs-Kainrath C, Hochreiter S (2015). "msa: an R package for multiple sequence alignment." Bioinformatics, 31(24), 3997–3999. doi: 10.1093/bioinformatics/btv494.
- Lerch, A. et al. Development Of Amplicon Deep Sequencing Markers And Data Analysis Pipeline For Genotyping Multi-Clonal Malaria Infections. BMC Genomics (2017), 18(1), p.864. http://dx.doi.org/10.1186/s12864-017-4260-y.
- Lerch, A. et al. Longitudinal tracking and quantification of individual Plasmodium falciparum clones in complex infections. Sci. Rep. 9, 3333 (2019), http://dx.doi.org/10.1038/s41598-019-39656-7.
- Dong Xu and Yang Zhang. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins, 2012, 80, 1715-1735.



◄ Figure 4. Maximum Likelihood phylogeny of the 39 haplotypes.

- Haplotypes csp125 and csp64 are closely related to the 3d7 reference
- However they only represent 1% and 2.5% of the haplotypes present in the samples respectively and they differ by one amino acid at position 71 in the Th3R region



▼ Figure 5. Protein structural changes from mutations.

- **A1 and A2:** Th2R (magenta) and Th3R (orange) mutating residues shown in red.
- **B:** Structure changes from reference (green) to QUARK predicted mutant (blue mesh).

Summary

- We found that the MOI varies dramatically across samples and regions and the most prevalent haplotypes are not closely related to the 3D7 CSP reference.
- Mutations were found predominantly in the Th2R and Th3R regions of the PfCSP gene.
- Using the predicted structure, we observed that these mutations may change the overall conformation of the CSP protein.
- These findings may provide insights into the low vaccine efficacy (~36%) that has been described thus far in the field.