Genome wide assessment of gene copy number and SNP variation in *Plasmodium vivax* from Ethiopia

Anthony Ford[1], Daniel Janies[1], Delenasaw Yewhalaw[2], Beka Raya[2], Richard Pearson[3], Karthigayan Gunalan[4], Louis H. Miller[4], Julian C. Rayner[3], Guiyun Yan[5], Eugenia Lo[6]

[1]Bioinformatics and Genomics, University of North Carolina at Charlotte, USA; [2]Tropical Infectious Disease Research Center, Jimma University, Ethiopia; [3]Malaria Programme, Wellcome Trust Sanger Institute, Hinxton, UK; [4]Laboratory of Malaria and Vector Research, NIAID/NIH, Bethesda, USA; [5]Public Health, University of California at Irvine, USA; [6]Biological Sciences, University of North Carolina at Charlotte, USA

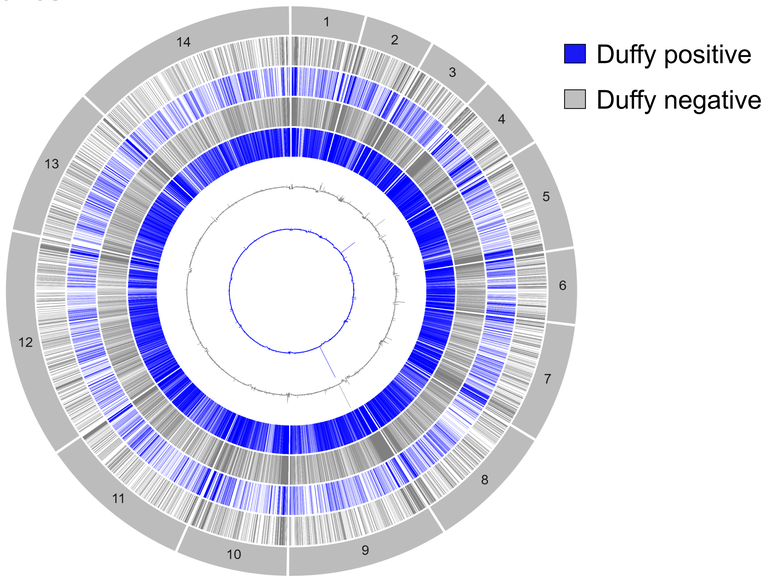
Background

- *Plasmodium vivax* is a major cause of malarial infection after *P. falciparum* with an estimated 2.5 billion people currently at risk. It poses particular challenges because it is the most widespread geographically and it has the ability to produce hypnozoites.
- *P. vivax* has been traditionally thought to be absent from most parts of sub-Saharan Africa because the populations there are protected by the lack of Duffy antigen expression that *P. vivax* needs for erythrocyte binding and invasion. However, there has been an increasing number of *P. vivax* cases reported across Africa [1] and a significant portion of those cases were confirmed in Duffy-negative individuals [e.g., 2-4]. These phenomenon leads to the hypothesis that *P. vivax* has possibly evolved with a novel, alternative invasion pathway that is Duffy independent.
- Our goal is to identify potential erythrocyte binding protein genes that play a role in the parasite-host invasion process at the genomic level.
- The purposes of this project is to develop a baseline, describing the genomic variation in *P. vivax* from Ethiopia so that we can compare these genome profiles with isolates across different regions of Africa.

Methods

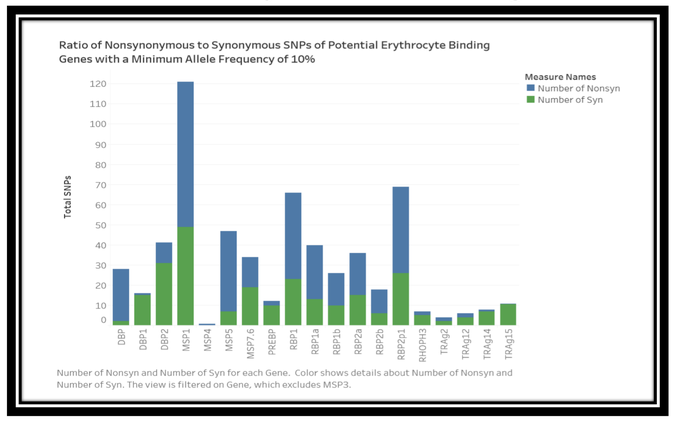
- The SNP calling was performed using the SAMtools/BCFtools pipeline, with indels removed during the calling process and using the reference genome PvivaxP01 obtained from GeneDB.org.
- The SNPS were then annotated using snpEff and subsequently filtered using SNPsift, an extension of the snpEFF program.
- The copy number variation was done using GATK4. Using window sizes of 1000, GATK4 detects potential copy number variations by standardizing counts by the panel of normal median counts followed by a log2 transformation and normalizing the counts to center around one. The second transformation denoises the copy numbers using the principal components of the panel of normal.
- The circular representation/summary of the genomes were created using the Circa software, a program that allows the user to create Circos plots of genetic data easily.
- (Ongoing) We are detecting positive selection within our panel of SNPs using the RAiSD package, which detects positive selection based on multiple signatures of selective sweeps.

Results



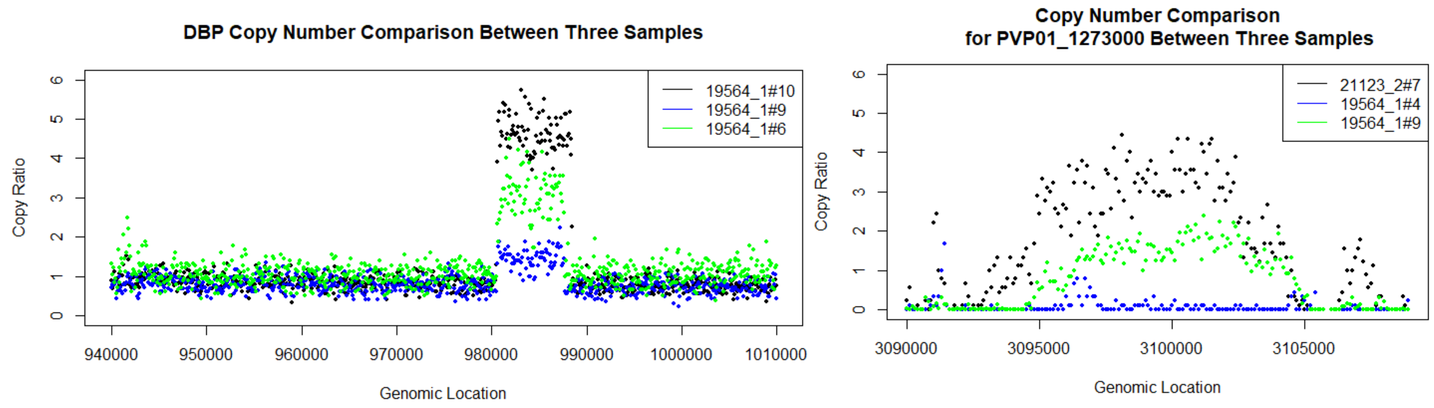
Circos plot created by the Circa software shows a "side-by-side" genome comparison between a Duffy negative and a Duffy positive *P. vivax* samples. The first ring is an ideogram representing the relative size of each chromosome; the second two rings represent the nonsynonymous SNPs and the subsequent two rings represent the synonymous SNPs. The two innermost rings represent the copy number variation detected in each sample using GATK4.

Summary of SNP Profiles generated by Samtools and Annotated using snpEff



Chromosome	Total SNPS	Nonsynonymous	Synonymous
1	17360	3668 (21.1%)	13692 (78.9%)
2	19108	4021 (21%)	15087 (79%)
3	10796	2298 (21.1%)	8498 (78.9%)
4	18553	4635 (25%)	13918 (75%)
5	17832	4375 (24.5%)	13457 (75.5%)
6	6536	1975 (30.2%)	4561 (69.8%)
7	14275	3311 (23.2%)	10964 (76.8%)
8	12111	2885 (23.8%)	9226 (76.2%)
9	33477	3572 (10.7%)	29905 (89.3%)
10	28133	6508 (23.1%)	21625 (76.9%)
11	14788	4158 (28.1%)	10630 (71.9%)
12	19610	4511 (23%)	15099 (77%)
13	11230	2721 (24.2%)	8509 (75.8%)
14	12533	3228 (25.8%)	9305 (74.2%)

Among the 20 *P. vivax* samples from Ethiopia, there were a total of 236,342 SNPs with approximately 28% as nonsynonymous and the remaining 72% as synonymous. The profiles were generated using Samtools using the reference genome *P. vivax* PVP01 available in GeneDB. The annotation was performed using snpEff with the filtering done by SNPsift, which is an extension of snpEff. Further analysis is currently ongoing with regards to identifying gene regions under positive selection based on multiple signatures in a selective sweep using the RAiSD package developed by N. Alachiotis and P. Pavlidis [5].



Comparison of the copy number variation for the Duffy binding protein gene and PVP01_1273000, an unknown gene located on chromosome 12 described as a *Plasmodium* exported protein with unknown function. The fold-coverage based on sequence read indicated that these gene regions contain variation in copy number, ranging from a single to >5 copies among samples. The high copy number of Duffy binding protein may relate to increased expression of the gene involved in erythrocyte binding invasion [6].

Summary of CNV generated by GATK4 of select genes

Gene	Copy Number	Percentage of samples
DBP (chr 6)	1-2 2-3 3-4 4+	20% 53% 26.7% 0%
MSP3	1-2 2-3 3-4 4+	53% 40% 0% 7%
VIR (chr 11)	1-2 2-3 3-4 4+	37.5% 43.75% 12.5% 6.25%
Unknown gene (chr 13)	1-2 2-3 3-4 4+	0% 25% 75% 0%

Summary of copy number variation of select genes by GATK and CNVnator. The results from both CNVnator and GATK4 were fairly similar for the above gene regions. Much of the duplications found in the *P.vivax* genomes occurred in the VIR genes that encode immunovariant proteins. *PvDBP* and *PvMSP*3 are two genes that are potentially responsible for host erythrocyte invasion.

References

1. Zimmerman PA. (2017). Plasmodium vivax infection in Duffy-Negative people. *The American Journal of Tropical Medicine and Hygiene* 97:636-638.
2. Lo EYY, Delenasaw Y, Zhong D, Zemene E, Degefa T, Tushune K, Ha M, Lee MC, James AA, Yan G. (2015). Molecular epidemiology of *Plasmodium vivax* and *Plasmodium falciparum* malaria among Duffy-positive and Duffy-negative populations in Ethiopia. *Malaria Journal* 14:84-94.
3. Abdelraheem MH, Albsheer MMA, Mohamed HS, Amin M, Mahdi Abdel Hamind M. (2016). Transmission of *Plasmodium vivax* in Duffy-negative individuals in central Sudan. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 11:258-260.
4. Russo G, Faggiono G, Paganoti GM, Djeunang Dongho GB, Pomponi A, De Santis R, Tebano G, Mibida M, Sanou Sobze M, Vullo V, Rezza G, Lista FR. (2017). Molecular evidence of *Plasmodium vivax* infection in Duffy negative symptomatic individuals from Dschang, West Cameroon. *Malaria Journal* 16:74
5. Alachiotis N, Pavlidis P. (2018) RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology* 1:79.
6. Gunalan K, Lo EYY, Hostetler JB, Yewhalaw D, Mu J, Neafsey DE, et al. (2016). Role of *Plasmodium vivax* Duffy-binding protein 1 in invasion of Duffy-null Africans. *Proceedings of the National Academy of Sciences of the United States of America* 113: 6271–6276.

Acknowledgements

We thank Mindy Shin, Jordan Connors, and Nikolaos Alachiotis for their help with the analyses. This work is supported by NIH/NIAID R15 AI138002 (Lo) and U19 AI129326 (Yan).