Genomic insights into *Plasmodium vivax* population structure and diversity in central Africa

Valerie Gartner [1,2], **Benjamin D. Redelings** [1,3,4], **Claudia Gaither**[5], **Jonathan B. Parr** [5], **Albert Kalonji**[6], **Fernandine Phanzu**[6], **Nicholas F. Brazeau** [5], **Jonathan J. Juliano** [5], **Gregory A. Wray** [1] ✉

[1]Biology Department, Duke University, Durham, NC 27708, USA; [2]University Program in Genetics and Genomics, Duke University, Durham, NC 27708, USA; [3]Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS 66045, USA; [4]Ronin Institute, Durham, NC 27705, USA; [5]University of North Carolina, Chapel Hill, NC, 27599, USA; [6]SANRU asbl, 149 A/B, Boulevard du 30 Juin, Kinshasa/Gombe, Democratic Republic of Congo

✉ **For correspondence:** gwray@duke.edu

Competing interests: The author declare no competing interests.

Abstract

Plasmodium vivax malaria has not traditionally been a major concern in central Africa given the high prevalence of the human Duffy-negative phenotype that is believed to prevent infection. Increasing reports of asymptomatic and symptomatic infections in Duffy-negative individuals throughout Africa raise the possibility that *P. vivax* is evolving to evade host resistance, but there are few parasite samples with genomic data available from this part of the world. In this study, we perform whole genome sequencing of a new *P. vivax* isolate from the Democratic Republic of the Congo (DRC) and assess how this central African isolate fits into the global context of this species. We use population genomics methods to show that *P. vivax* from DRC is similar to other African parasite populations and is not closely related to the non-human primate parasite *P. vivax*-like.

Significance

The second most common malaria species to infect humans, *Plasmodium vivax*, is not considered a major threat to human health in central Africa because people in this region frequently have a genetic variant that prevents the *P. vivax* species from being able to cause illness. Recent research shows that *P. vivax* can be found in individuals who should be immune, but there is insufficient data to understand why. Our study investigates the genome of one *P. vivax* sample collected from central Africa to show that the DRC population is closely related to other *P.vivax* populations in Africa.

Keywords

Malaria, Plasmodium vivax, Genome, Duffy negative, Sub-Saharan Africa, Central Africa

Introduction

The widespread fixation of the Duffy-negative phenotype in the human population in Sub-Saharan Africa, which provides immunity from *Plasmodium vivax*, is one of the most remarkable cases of natural selection documented in human populations (***Hamblin and Di Rienzo, 2000***; ***Hamblin et al., 2002***; ***Kwiatkowski, 2005***). The Duffy-negative phenotype occurs in humans with two copies of a silencing mutation in the promoter region of the Duffy Antigen Receptor for Chemokines (DARC) gene, resulting in the absence of receptor expression exclusively in erythrocytes necessary for the progression of the *P. vivax* life cycle (***Parasol et al., 1998***; ***Miller et al., 1976***). Despite this, there are an increasing number of reports of asymptomatic and symptomatic *P. vivax* infections in people with the Duffy-negative mutation suggesting that *P. vivax* persists in central Africa at low levels in people with the Duffy-negative resistance allele (***Brazeau et al., 2021***; ***Russo et al., 2017***; ***Motshoge et al., 2016***; ***Ryan et al., 2006***; ***Mendes et al., 2011***).

An alternate explanation for the persistence of *P. vivax* in central Africa comes from the recent discovery of a closely related parasite species that infects non-human primates, *Plasmodium vivax*-like in Western Africa (***Liu et al., 2014***; ***Loy et al., 2018***; ***Gilabert et al., 2018***). Though there is only one confirmed report of *P. vivax*-like infecting humans, a study using *P. vivax*-like recombinant binding proteins did not reveal species-specific barriers to erythrocyte invasion of human, gorilla,

or chimpanzee red blood cells, suggesting *P. vivax*-like likely is able to infect humans (***Prugnolle et al., 2013***; ***Loy et al., 2018***).

A third possible explanation for the presence of *P. vivax* in central Africa despite human resistance alleles might be that *P. vivax* is adapting to overcome the Duffy negative resistance allele, which would be a serious concern for malaria elimination efforts in central Africa. Genomics can potentially aid in understanding the source of these infections in central Africa, however none of the seventy-seven publicly available African *P. vivax* genomes are from regions with high levels of Duffy negativity except for three samples from Uganda. Importantly, these Ugandan samples were collected from people of unknown Duffy status after returning to the United Kingdom (***Gunalan et al., 2018***; ***Benavente et al., 2021***).

In this study, we perform whole genome sequencing of a new *P. vivax* isolate from the Eastern region of Democratic Republic of the Congo (DRC) and assess how *P. vivax* from central Africa fits into the global context of this pathogen. Though the original study design excludes the possibility of genotyping the human host of this *P. vivax* sample, the patient had no known travel history and resides in a region where the Duffy-negative phenotype frequency is at or above 80% (***Howes et al., 2011***), thus this patient has a high chance of having the Duffy-negative phenotype. We confirm the presence of a *P. vivax* population in central Africa that is not closely related to the ape-infection *P. vivax*-like species. Further, we investigate whether this sample has duplications of the Duffy binding ligand genes PvDBP and PvDPB2 (also referred to as EBP and EBP2) which might potentially enable *P. vivax* to evade host immunity. Though Copy Number Variation of these genes is not conclusively linked to *P. vivax* infection of Duffy-negative individuals (***Lo et al., 2019***, ***2021***), both genes contain the Duffy Binding Protein II domain, one of the foremost vaccine target candidates (***Roesch et al., 2018***). We show a duplication of the gene PvDBP in the DRC *P. vivax* sample and a single copy of PvDBP2.

Results

P. vivax from DRC falls within African diversity in context of global population structure

To determine where this new central African sample fits within global *P. vivax* populations, we first performed principal components analysis (PCA) from 696 *P. vivax* samples using only biallelic SNPs

(excluding hyper-variable regions defined by (**Pearson et al., 2016**; **Auburn et al., 2016**)). The global PCA analysis in Figure 1A shows a population structure as defined by geography is reproduced by the first two principal components, as has been reported previously (**Daron et al., 2021**; **Benavente et al., 2021**). Three main sub-populations are formed: 1. samples from the Americas, 2. African and South Asian samples, and 3. East Asian and Southeast Asian samples. Within the cluster of African and South Asian samples, the new DRC sample is most similar to those from Uganda and Madagascar in their position alongside South Asian *P. vivax*.

To further understand how this new DRC *P. vivax* sample relates to other global populations, we constructed a maximum likelihood tree from 349,353 SNPs across the genome for no more than 10 samples per country. Figure 1B shows the DRC sample clusters within African *P. vivax* variation, and again clusters most closely with Uganda and Madagascar samples. *P. vivax*-like samples were used to root this tree, and notably, as in **Daron et al.** (**2021**), the root for our tree is located centrally in the *P. vivax* tree and not inside African variation, which might be expected if this sample represented an ancestral source population of *P. vivax* in humans.

Though clustering of genetic variation is expected due to geographic separation, PCA and phylogenetic trees do not quantify how much genetic ancestry is shared across geographic populations. To determine the fraction of shared ancestry across subgroups, we assessed the same SNP data set used in Figure 1A-B through Admixture analysis. We modeled *P. vivax* ancestry proportions for population sizes of 2 through 17 and found that a population size of K=14 was supported based on mean Cross Validation Error value. Figure 1C shows the global ancestry proportions of *P. vivax* when modeled at K=14. Ancestry proportions for all calculated K values is shown in Supplementary Figure 3. Additionally, F4 statistics (**Patterson et al., 2012**; **Maier et al., 2022**) were calculated assess the correlation in allele frequencies of the DRC sample with other *P. vivax* populations around the world. We use the form (*P. vivax*-like, DRC; Papua New Guinea, Y), where *P. vivax*-like is the outgroup species, shown in figure 1D. Higher F4 estimates indicate the DRC sample has more gene flow with the test population than it does with samples from Papua New Guinea (PNG). All test populations except for North Korea and the Philippines resulted in a significant absolute Z score ($|Z| > 3$). F4 estimates and related data are available in Supplemental table 4.

P. vivax from DRC has similar levels of population diversity as other African populations

In order to explore *P. vivax* population diversity despite having only a single sample from the DRC, we looked at the number of private alleles in each country. Private alleles are variants present in one population and in none of the others, making them unique to a population. Table S1 shows the full set of summary statistics calculated for all countries. When normalizing the private allele count by dividing by the number of samples, as shown in Figure S1B, the DRC sample had a similar amount of variation as other African populations despite having a low absolute private allele count (Figure S1A). The genome-wide within-population diversity value, π, calculated for *P. vivax* from different sub-regions shown in Figure S2 indicates that when combined, all African samples have similar genome-wide diversity as populations in Asia. *P. vivax* nucleotide diversity of central African samples (DRC and Uganda) is similar to but slightly lower than that of East African (Ethiopia, Eritrea, and Sudan).

P. vivax in central Africa is distinct from *P. vivax*-like

To assess the relatedness of *P. vivax* in humans in central Africa to the *P. vivax*-like malaria species found in non-human primates, we constructed a maximum likelihood tree including publicly available *P. vivax*-like genome sequences mapped to the PvP01 reference genome. Figure 2 shows that the *P. vivax* sample from the DRC clusters with other African *P. vivax* samples, while all *P. vivax*-like samples are separate from *P. vivax* populations. Additionally, the longer branch lengths for the *P. vivax*-like samples in figure 2 illustrate the higher level of diversity in this species than is found in any population of the human-infecting *P. vivax*.

Copy number variation is present in binding proteins

Copy Number Variation (CNV) in certain binding proteins is potentially important for pathogenesis of *P. vivax* in Duffy-negative individuals (***Gunalan et al., 2018***). We used BAM files aligned to PvP01 with optical duplicates removed to compare read depth within the gene region to coverage in the region 10 Kb upstream and downstream of the coding region for several genes related to erythrocyte binding and invasion: PvDBP, PvDBP2, PvRBP1a, PvRBP1b, PvRBP2a, PvRBP2b, and PvRBP2c based on results from ***Gunalan et al.*** (***2018***). In the DRC *P. vivax* sample, only one gene, PvDBP, had evidence of a potential gene duplication (Supplemental table 2). Lumpy was used to determine the

number of paired-end and split reads that support a duplication in PvDBP, which showed evidence of a duplication of 8,216 base pairs in length at chr6: 980,472-988,688 with 292 paired-end reads and 419 split reads supporting the structural variant. The ratio of the coverage for the duplicated PvDBP region compared to the surrounding intergenic region was 2.47. Based on the IGV pileup view in Figure 3, there appears to be two distinct copies of this gene being mapped to the single PvDBP reference annotation. We note that the region of higher read depth extends into the intergenic regions on either side of the gene annotation for PvDBP, consistent with the duplication type first reported in Malagasy samples (***Menard et al., 2013***; ***Hostetler et al., 2016***).

Discussion

Despite the publication of recent studies on *P. vivax* diversity that include African samples (***Chan et al., 2012***; ***Menard et al., 2013***; ***Auburn et al., 2019***; ***Twohig et al., 2019***; ***Daron et al., 2021***), there is still very little known about this pathogen in central Africa. Our analyses of one new *P. vivax* genome collected from the Idjwi island of Lake Kivu in DRC show that this sample falls within the scope of African parasite diversity and is distinct from *P. vivax*-like samples. This suggests that an endemic *P. vivax* population is present in central Africa, as previously proposed by ***Brazeau et al. (2021)***.

Our results shown in Figure 1 suggest the DRC *P. vivax* sample is most like those from Uganda and Madagascar. While the similarity between *P. vivax* from eastern DRC and Uganda is not surprising, it is interesting that *P. vivax* populations in the DRC, Uganda, and Madagascar all share ancestry with South Asian samples, as shown in Figure 1C and D, but no measurable ancestry from Southeast Asian *P. vivax* populations despite a well-documented history of Austronesian human migration into this region (***Anderson et al., 2018***; ***Brucato et al., 2018***, ***2019***).

The phylogenetic tree in Figure 1B shows that this new *P. vivax* sample clusters with other African samples and not with any *P. vivax*-like samples that have been sequenced previously, suggesting that there is a *P. vivax* population in the DRC separate from potential zoonosis from an animal reservoir. However, all publicly available *P. vivax*-like sequences to date have been collected from animals in countries on the West coast of Africa. Further sampling of both humans and non-human primates throughout the broad geography of central Africa is needed to determine whether there truly is no transfer of parasites across species, as *in vitro* studies indicate that there is little host specificity of *P. vivax*-like (***Loy et al., 2018***).

Our analyses replicate previous studies showing that global *P. vivax* populations are distinct from each other based on geographic distance, and most sharing of haplotype backgrounds occurs within geographic regions and only rarely across geographic borders. This geographic separation of ancestral groups, along with the summary statistics calculated in Supplementary table 1 and illustrated in Supplementary Figures S1 and S2, possibly indicate that DRC has a comparable *P. vivax* population size relative to Ethiopia and Uganda, though this interpretation is greatly limited by the reductive nature of genome-wide summary statistics.

Though the evidence linking copy number of Duffy binding ligand genes with *P. vivax* infection of Duffy-negative individuals is not conclusive (***Lo et al., 2019***, ***2021***), it remains a subject of concern, especially since the Duffy Binding Protein-II domain is one of the foremost vaccine target candidates (***Roesch et al., 2018***). We show that this *P. vivax* sample from the DRC has a duplication in PvDBP relative to the PvP01 reference genome that corresponds with the longer Malagasy-type duplication, as opposed the shorter PvDBP duplication first detected in Cambodian samples (***Hostetler et al., 2016***). Duplications in PvDBP may play a role in Duffy-independent mechanisms of infection in Duffy-negative individuals and should be considered in future studies (***Lo et al., 2021***).

Our findings are largely limited by the uncertainty of whether this single sample is representative of the larger population of *P. vivax* in central Africa. This *P. vivax* sample was collected from an individual with no known travel history in a region with an estimated 98% homozygosity for Duffy-negativity (***Howes et al., 2011***), however as the patient's Duffy genotype was not collected in the study, caution should be exercised until future studies can provide further context.

It has become clear from epidemiological studies that *P. vivax* is much more common in central Africa than previously thought (***Ryan et al., 2006***; ***Twohig et al., 2019***; ***Mendes et al., 2011***; ***Brazeau et al., 2021***). Generating genomes from these infections however is difficult, as most of them are extremely low parasite densities (***Brazeau et al., 2021***). Thus, they are not amenable to whole genome sequencing. To increase our understanding of this parasite in Africa, the research community needs to continue to try to identify samples amenable to analysis and deposit them for community use. Intentional sampling across Africa would further contextualize this sample within African *P. vivax* diversity and shed light on the mechanisms of infection in Duffy negative individuals.

Methods

Genome Sequencing of *P. vivax* sample from the Democratic Republic of Congo

One *P. vivax* sample was collected from Idjwi, DRC (***Parr et al., 2021***). The patient was an 11-year-old with reported fever, diarrhea, and headache who tested positive for *P. vivax* via 18s qPCR assay as previously described (***Brazeau et al., 2021***) at 957 parasites per µL. The patient tested negative for *P. falciparum* via rapid HRP2 test and real time PCR. Due to the original study design, the patient's Duffy genotype was not assessed. Travel history was not taken.

DNA from three 6mm punches from a dried blood spot were extracted using Chelex-Tween as previously described (***Topazian et al., 2020***). *P. vivax* infection was confirmed using a Taqman real time PCR assay (***Brazeau et al., 2021***). *Plasmodium* DNA was enriched from human DNA using a custom Twist hybrid capture array and in-house pipeline (Twist Biosciences, San Francisco, CA, USA). The array was designed by single tiling of the PvP01 genome with baits complementary to human removed. Capture and library preparation were completed per manufacturer's instructions. Sequencing was completed on a NovaSeq 6000 at the University of North Carolina High Throughput Sequencing Facility.

Genomic Data Processing

1,408 FASTQ files for *P. vivax* with metadata about the geographic location from which they were sampled were downloaded from the Sequence Read Archive (***Kodama et al., 2012***). BAM files were created using bwa mem (***Li and Durbin, 2009***) to align short reads to the PvP01 reference genome (***Auburn et al., 2016***). Picard MarkDuplicates version 2.18.15 (***Broad Institute, 2019***) was used to remove optical duplicates, and variants underwent hard filtering using the Genome Analysis Toolkit (GATK) HaplotypeCaller version 3.8.1, followed by joint calling (***Auwera and O'Connor, 2020***). To avoid confounding analyses with *P. vivax* samples made up of more than one haplotype background (i.e. a multiplicity of infection greater than one), samples were filtered out based on haplotype number estimates generated by Octopus (***Cooke et al., 2018***). Sample accessions and location metadata of this final sample set of 696 *P. vivax* samples, one being the new DRC genome sequenced in this paper, are available via the project GitHub repo: https://github.com/vlrieg/DRC_vivax/blob/main/sample_info/metadata_table.csv.

Population Genetics Analysis

Analyses were performed only on assembled chromosomes as defined by the PvP01 reference genome (***Auburn et al., 2016***). Hyper-variable regions determined in ***Pearson et al.*** (***2016***) were converted to PvP01 coordinates using the alignment-smc tool in bali-phy with the translate-mask option (***Suchard and Redelings, 2006***), then removed from the data set using bedtools version 2.25.0 (***Quinlan and Hall, 2010***) and VCFtools version 0.1.15 (***Danecek et al., 2011***). The data were reduced to biallelic Single Nucleotide Polymorphisms (SNPs) only, and LD pruning was performed using PLINK to obtain unlinked singletons and variants from the data set as previously described (***Daron et al., 2021***; ***Benavente et al., 2021***), resulting in 94,083 SNPs. Principal Component Analysis was performed using Plink version 1.9 (***Purcell et al., 2007***) and plotted in R using ggplot2 (***Wickham, 2016***). Admixture analysis was performed using Admixture software version 1.3.0 (***Alexander et al., 2009***) via admixturePipeline version 2.0.2 (***Mussmann et al., 2020***). The resulting Q matrices were visualized using Pong version 1.5 (***Behr et al., 2016***). Cross validation error supports K=14 populations. F4 statistics were calculated on 467,205 biallelic SNPs that had no more than 5% missing sites from 696 *P. vivax* and 56 *P. vivax*-like samples using Admixtools2 version 2.0.0 (***Maier et al., 2022***). Summary statistics were computed on 696 *P. vivax* samples. π was computed on biallelic SNPs using Pixy version 1.2.4.beta1 (***Korunes and Samuk, 2021***).

Phylogenetics Analysis

Phylogenies of global and African subsamples of *P. vivax* were made using IQtree version 1.6.12 for Linux 64-bit (***Minh et al., 2020***; ***Hoang et al., 2018***). Trees were constructed using biallelic SNPs from PvP01-defined nuclear chromosomes except for hypervariable regions. Phylogenies were inferred using the GTR+ASC model to account for ascertainment bias. Trees were rooted using two *P. vivax*-like samples from two recent studies of this closely related species (***Gilabert et al., 2018***; ***Loy et al., 2018***). Trees were visualized using FigTree version 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) and modified with Adobe Illustrator.

Duffy Binding Gene Copy Number Variation Analysis

The read depths of important genes related to *P. vivax* pathogenesis were investigated by extracting the genomic regions from the BAM file (after removing optical duplicates) using Samtools version 1.3.1 (***Danecek et al., 2021***) and visualized in IGV version 2.4.14 (***Robinson et al., 2011***). Genomic

coverage as defined by read depth was calculated for PvDBP using bedtools version 2.25.0 (***Quinlan and Hall, 2010***). Breakpoint evidence to support a duplication of PvDBP was estimated using Lumpy version 0.2.13 (***Layer et al., 2014***).

Acknowledgment

This work was supported by: the National Institutes of Health [R01TW010870 and K24AI134990 to J.J.J.], the North Carolina Biotechnology Center support for high-performance computing facility [2016-IDG-1013, 2020-IIG-2109], and the Global Fund to Fight AIDS, Tuberculosis, and Malaria. Thanks to Krista Pipho for feedback on early drafts of this manuscript.

Author contributions

Conceptualization: V.G., B.D.R., N.F.B., J.J.J., G.W. **Sample Collection and Sequencing:** J.B.P., C.G., A.K., F.P., J.J.J. **Conducted Analyses:** V.G. **Advised on Analyses:** B.D.R., N.F.B., J.J.J., G.W. **Aided in Manuscript Preparation:** V.G., B.D.R., C.G., J.B.P, A.K., F.P., N.F.B., J.J.J., G.W.

Data Availability

P. vivax WGS data from DRC available under BioProject accession: PRJNA909777. Accession numbers for previously published data used in this study are available on the project GitHub repository.

References

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009; 19(9):1655–1664. doi: 10.1101/gr.094052.109.

Anderson A, Clark G, Haberle S, Higham T, Kemp MN, Prendergast A, Radimilahy C, Rakotozafy LM, Ramilisonina, Schwenninger JL, Virah-Sawmy M, Camens A. New evidence of megafaunal bone damage indicates late colonization of Madagascar. PLoS ONE. 2018; 13(10):1–14. doi: 10.1371/journal.pone.0204368.

Auburn S, Böhme U, Steinbiss S, Trimarsanto H, Hostetler J, Sanders M, Gao Q, Nosten F, Newbold CI, Berriman M, Price RN, Otto TD. A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. Wellcome Open Res. 2016; 1:4. https://wellcomeopenresearch.org/articles/1-4/v1, doi: 10.12688/wellcomeopenres.9876.1.

Auburn S, Getachew S, Pearson RD, Amato R, Miotto O, Trimarsanto H, Zhu SJ, Rumaseb A, Marfurt J, Noviyanti R, Grigg MJ, Barber B, William T, Morgado Concalves S, Drury E, Sriprawat K, Anstey NM, Nosten F, Petros B, Aseffa A, et al. Genomic analysis of Plasmodium vivax in southern Ethiopia reveals selective pressures in multiple parasite mechanisms. The Journal of Infectious Diseases. 2019; https://academic.oup.com/jid/advance-article/doi/10.1093/infdis/jiz016/5298331, doi: 10.1093/infdis/jiz016.

Auwera GAVd, O'Connor BD. Genomics in the Cloud. 1 ed. O'Reilly Media, Inc; 2020.

Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics. 2016; 32(18):2817–2823. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw327, doi: 10.1093/bioinformatics/btw327.

Benavente ED, Manko E, Phelan J, Campos M, Nolder D, Fernandez D, Velez-Tobon G, Castaño AT, Dombrowski JG, Marinho CRF, Aguiar ACC, Pereira DB, Sriprawat K, Nosten F, Moon R, Sutherland CJ, Campino S, Clark TG. Distinctive genetic structure and selection patterns in Plasmodium vivax from South Asia and East Africa. Nature Communications. 2021; 12(1):1–11. https://doi.org/10.1038/s41467-021-23422-3, doi: 10.1038/s41467-021-23422-3, publisher: Springer US.

Brazeau NF, Mitchell CL, Morgan AP, Deutsch-Feldman M, Watson OJ, Thwai KL, Gelabert P, van Dorp L, Keeler CY, Waltmann A, Emch M, Gartner V, Redelings B, Wray GA, Mwandagalirwa MK, Tshefu AK, Likwela JL, Edwards JK, Verity R, Parr JB, et al. The epidemiology of Plasmodium vivax among adults in the Democratic Republic of the Congo. Nat Commun. 2021; 12(1):4169. https://www.nature.com/articles/s41467-021-24216-3, doi: 10.1038/s41467-021-24216-3, number: 1 Publisher: Nature Publishing Group.

Broad Institute, Picard; 2019. https://broadinstitute.github.io/picard/.

Brucato N, Fernandes V, Kusuma P, Cerny V, Mulligan CJ, Soares P, Rito T, Besse C, Boland A, Deleuze JF, Cox MP, Sudoyo H, Stoneking M, Pereira L, Ricaut FX, Saitou N. Evidence of Austronesian Genetic Lineages in East Africa and South Arabia: Complex Dispersal from Madagascar and Southeast Asia. Genome Biology and Evolution. 2019; 11(3):748–758. doi: 10.1093/gbe/evz028.

Brucato N, Fernandes V, Mazières S, Kusuma P, Cox MP, Ng'ang'a JW, Omar M, Simeone-Senelle MC, Frassati C, Alshamali F, Fin B, Boland A, Deleuze JF, Stoneking M, Adelaar A, Crowther A, Boivin N, Pereira L, Bailly P, Chiaroni J, et al. The Comoros Show the Earliest Austronesian Gene Flow into the Swahili Corridor. American Journal of Human Genetics. 2018; 102(1):58–68. doi: 10.1016/j.ajhg.2017.11.011.

Chan ER, Menard D, David PH, Ratsimbasoa A, Kim S, Chim P, Do C, Witkowski B, Mercereau-Puijalon O, Zimmerman PA, Serre D. Whole Genome Sequencing of Field Isolates Provides Robust Characterization of Genetic Diversity in Plasmodium vivax. PLoS Neglected Tropical Diseases. 2012; 6(9). doi: 10.1371/journal.pntd.0001811, ISBN: 1935-2735 (Electronic)\n1935-2727 (Linking).

Cooke DP, Wedge DC, Lunter G. A unified haplotype-based method for accurate and comprehensive variant calling. bioRxiv. 2018; p. 456103. https://www.biorxiv.org/content/10.1101/456103v1.full, doi: 10.1101/456103.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156–2158. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr330, doi: 10.1093/bioinformatics/btr330.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. GigaScience. 2021; 10(2):giab008. https://doi.org/10.1093/gigascience/giab008, doi: 10.1093/gigascience/giab008.

Daron J, Boissière A, Boundenga L, Ngoubangoye B, Houze S, Arnathau C, Sidobre C, Trape JF, Durant P, Renaud F, Fontaine MC, Prugnolle F, Rougeron V. Population genomic evidence of a Southeast Asian origin of Plasmodium vivax. Science Advances. 2021; 7(18). doi: 10.1126/sciadv.abc3713.

Gilabert A, Otto TD, Rutledge GG, Franzon B, Okouga P, Ngoubangoye B, Makanga B, Boundenga L, Paupy C, Renaud F, Prugnolle F, Rougeron V. Plasmodium vivax-like genome sequences shed new insights into Plasmodium vivax biology and evolution. PLoS Biology. 2018; p. 1–24. doi: 10.1371/journal.pbio.2006035.

Gunalan K, Niangaly A, Thera MA, Doumbo OK, Miller LH. Plasmodium vivax Infections of Duffy-Negative Erythrocytes: Historically Undetected or a Recent Adaptation? Trends in Parasitology. 2018; 34(5):420–429. https://linkinghub.elsevier.com/retrieve/pii/S147149221830028X, doi: 10.1016/j.pt.2018.02.006.

Hamblin MT, Di Rienzo A. Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus. The American Journal of Human Genetics. 2000; 66(5):1669–1679. http://linkinghub.elsevier.com/retrieve/pii/S0002929707629966, doi: 10.1086/302879, ISBN: 0002-9297.

Hamblin MT, Thompson EE, Di Rienzo A. Complex Signatures of Natural Selection at the Duffy Blood Group Locus. The American Journal of Human Genetics. 2002; 70(2):369–383. http://linkinghub.elsevier.com/retrieve/pii/S0002929707639524, doi: 10.1086/338628, ISBN: 0002-9297.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution. 2018; 35(2):518–522. https://doi.org/10.1093/molbev/msx281, doi: 10.1093/molbev/msx281.

Hostetler JB, Lo E, Kanjee U, Amaratunga C, Suon S, Sreng S, Mao S, Yewhalaw D, Mascarenhas A, Kwiatkowski DP, Ferreira MU, Rathod PK, Yan G, Fairhurst RM, Duraisingh MT, Rayner JC. Independent Origin and Global Distribution of Distinct Plasmodium vivax Duffy Binding Protein Gene Duplications. PLoS neglected tropical diseases. 2016; 10(10):e0005091. doi: 10.1371/journal.pntd.0005091.

Howes RE, Patil AP, Piel FB, Nyangiri OA, Kabaria CW, Gething PW, Zimmerman PA, Barnadas C, Beall CM, Gebremedhin A, Ménard D, Williams TN, Weatherall DJ, Hay SI. The global distribution of the Duffy blood group. Nature Communications. 2011; 2(1). doi: 10.1038/ncomms1265, ISBN: 2041-1723 (Electronic)\r2041-1723 (Linking).

Kodama Y, Shumway M, Leinonen R, on behalf of the International Nucleotide Sequence Database Collaboration. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Research. 2012; 40:D54–D56. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr854, doi: 10.1093/nar/gkr854.

Korunes KL, Samuk K. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. Molecular Ecology Resources. 2021; p. 1359–1368. doi: 10.1111/1755-0998.13326.

Kwiatkowski DP. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. The American Journal of Human Genetics. 2005; 77(2):171–192. http://linkinghub.elsevier.com/retrieve/pii/S0002929707629097, doi: 10.1086/432519, ISBN: 0002-9297 (Print)\r0002-9297 (Linking).

Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014; 15(6):R84. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-6-r84, doi: 10.1186/gb-2014-15-6-r84.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–1760. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp324, doi: 10.1093/bioinformatics/btp324.

Liu W, Li Y, Shaw KS, Learn GH, Plenderleith LJ, Malenke JA, Sundararaman SA, Ramirez MA, Crystal PA, Smith AG, Bibollet-Ruche F, Ayouba A, Locatelli S, Esteban A, Mouacha F, Guichet E, Butel C, Ahuka-Mundeke S, Inogwabini BI, Ndjango JBN, et al. African origin of the malaria parasite Plasmodium vivax. Nature Communications. 2014; 5. doi: 10.1038/ncomms4346, iSBN: 2041-1723 (Electronic)\r2041-1723 (Linking).

Lo E, Hostetler JB, Yewhalaw D, Pearson RD, Hamid MMA, Gunalan K, Kepple D, Ford A, Janies DA, Rayner JC, Miller LH, Yan G. Frequent expansion of Plasmodium vivax Duffy Binding Protein in Ethiopia and its epidemiological significance. PLOS Neglected Tropical Diseases. 2019; 13(9):e0007222. doi: 10.1371/journal.pntd.0007222.

Lo E, Russo G, Pestana K, Kepple D, Abagero BR, Dongho GBD, Gunalan K, Miller LH, Hamid MMA, Yewhalaw D, Paganotti GM. Contrasting epidemiology and genetic variation of Plasmodium vivax infecting Duffy-negative individuals across Africa. International Journal of Infectious Diseases. 2021; 108:63–71. https://doi.org/10.1016/j.ijid.2021.05.009, doi: 10.1016/j.ijid.2021.05.009, publisher: International Society for Infectious Diseases.

Loy DE, Plenderleith LJ, Sundararaman SA, Liu W, Gruszczyk J, Chen YJ, Trimboli S, Learn GH, MacLean OA, Morgan ALK, Li Y, Avitto AN, Giles J, Calvignac-Spencer S, Sachse A, Leendertz FH, Speede S, Ayouba A, Peeters M, Rayner JC, et al. Evolutionary history of human Plasmodium vivax revealed by genome-wide analyses of related ape parasites. Proceedings of the National Academy of Sciences. 2018; p. 201810053. http://www.pnas.org/lookup/doi/10.1073/pnas.1810053115, doi: 10.1073/pnas.1810053115.

Maier R, Flegontov P, Flegontova O, Changmai P, Reich D, On the limits of fitting complex models of population history to genetic data. bioRxiv; 2022. https://www.biorxiv.org/content/10.1101/2022.05.08.491072v2, doi: 10.1101/2022.05.08.491072.

Menard D, Chan ER, Benedet C, Ratsimbasoa A, Kim S, Chim P, Do C, Witkowski B, Durand R, Thellier M, Severini C, Legrand E, Musset L, Nour BYM, Mercereau-Puijalon O, Serre D, Zimmerman PA. Whole Genome Sequencing of Field Isolates Reveals a Common Duplication of the Duffy Binding Protein Gene in Malagasy Plasmodium vivax Strains. PLoS Neglected Tropical Diseases. 2013; 7(11). doi: 10.1371/journal.pntd.0002489, ISBN: 1935-2735 (Electronic)\n1935-2727 (Linking).

Mendes C, Dias F, Figueiredo J, Mora VG, Cano J, de Sousa B, do Rosário VE, Benito A, Berzosa P, Arez AP. Duffy negative antigen is no longer a barrier to Plasmodium vivax - molecular evidences from the African West Coast (Angola and Equatorial Guinea). PLoS Neglected Tropical Diseases. 2011; 5(6):2–7. doi: 10.1371/journal.pntd.0001192, ISBN: 1935-2735 (Electronic)\n1935-2727 (Linking).

Miller LH, Mason SJ, Clyde DF, McGinniss MH. The Resistance Factor to Plasmodium vivax in Blacks — The Duffy-Blood-Group Genotype, FyFy. New England Journal of Medicine. 1976; https://www.nejm.org/doi/10.1056/NEJM197608052950602.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution. 2020; 37(5):1530–1534. https://doi.org/10.1093/molbev/msaa015, doi: 10.1093/molbev/msaa015.

Motshoge T, Ababio GK, Aleksenko L, Read J, Peloewetse E, Loeto M, Mosweunyane T, Moakofhi K, Ntebele DS, Chihanga S, Motlaleng M, Chinorumba A, Vurayai M, Pernica JM, Paganotti GM, Quaye IK. Molecular evidence of high rates of asymptomatic P. vivax infection and very low P. falciparum malaria in Botswana. BMC Infectious Diseases. 2016; 16(1):1–8. http://dx.doi.org/10.1186/s12879-016-1857-8, doi: 10.1186/s12879-016-1857-8, publisher: BMC Infectious Diseases.

Mussmann SM, Douglas MR, Chafin TK, Douglas ME. AdmixPipe: population analyses in Admixture for non-model organisms. BMC Bioinformatics. 2020; 21(1):337. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03701-4, doi: 10.1186/s12859-020-03701-4.

Parasol N, Reid M, Rios M, Castilho L, Harari I, Kosower NS. A novel mutation in the coding sequence of the FY*B allele of the Duffy chemokine receptor gene is associated with an altered erythrocyte phenotype. Blood. 1998; 92(7):2237–2243. doi: 10.1016/S0887-7963(99)80027-3, iSBN: 0006-4971 (Print) 0006-4971 (Linking).

Parr JB, Kieto E, Phanzu F, Mansiangi P, Mwandagalirwa K, Mvuama N, Landela A, Atibu J, Efundu SU, Olenga JW, Thwai KL, Morgan CE, Denton M, Poffley A, Juliano JJ, Mungala P, Likwela JL, Sompwe EM, Rogier E, Tshefu AK, et al. Analysis of false-negative rapid diagnostic tests for symptomatic malaria in the Democratic Republic of the Congo. Scientific Reports. 2021; 11(1):6495. https://www.nature.com/articles/s41598-021-85913-z, doi: 10.1038/s41598-021-85913-z, number: 1 Publisher: Nature Publishing Group.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. Ancient Admixture in Human History. Genetics. 2012; 192(3):1065–1093. https://doi.org/10.1534/genetics.112.145037, doi: 10.1534/genetics.112.145037.

Pearson RD, Amato R, Auburn S, Miotto O, Almagro-Garcia J, Amaratunga C, Suon S, Mao S, Noviyanti R, Trimarsanto H, Marfurt J, Anstey NM, William T, Boni MF, Dolecek C, Tran HT, White NJ, Michon P, Siba P, Tavul L, et al. Genomic analysis of local variation and recent evolution in Plasmodium vivax. Nature Genetics. 2016; 48(8):959–964. doi: 10.1038/ng.3599, ISBN: 1546-1718; 1061-4036.

Prugnolle F, Rougeron V, Becquart P, Berry A, Makanga B, Rahola N, Arnathau C, Ngoubangoye B, Menard S, Willaume E, Ayala FJ, Fontenille D, Ollomo B, Durand P, Paupy C, Renaud F. Diversity, host switching and evolution of Plasmodium vivax infecting African great apes. Proceedings of the National Academy of Sciences. 2013; 110(20):8123–8128. http://www.pnas.org/cgi/doi/10.1073/pnas.1306004110, doi: 10.1073/pnas.1306004110, iSBN: 1091-6490 (Electronic)\n0027-8424 (Linking).

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007; 81(3):559–575. https://linkinghub.elsevier.com/retrieve/pii/S0002929707613524, doi: 10.1086/519795.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–842. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033, doi: 10.1093/bioinformatics/btq033.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011; 29(1):24–26. https://www.nature.com/articles/nbt.1754, doi: 10.1038/nbt.1754, number: 1 Publisher: Nature Publishing Group.

Roesch C, Popovici J, Bin S, Run V, Kim S, Ramboarina S, Rakotomalala E, Rakotoarison RL, Rasoloharimanana T, Andriamanantena Z, Kumar A, Guillotte-Blisnick M, Huon C, Serre D, Chitnis CE, Vigan-Womas I, Menard D. Genetic diversity in two Plasmodium vivax protein ligands for reticulocyte invasion. PLOS Neglected Tropical Diseases. 2018 Oct; 12(10):e0006555. https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0006555, doi: 10.1371/journal.pntd.0006555, publisher: Public Library of Science.

Russo G, Faggioni G, Paganotti GM, Bruna G, Dongho D, Pomponi A, Santis RD, Tebano G, Mbida M, Sobze MS, Vullo V, Rezza G, Lista FR. Molecular evidence of Plasmodium vivax infection in Duffy negative symptomatic individuals from Dschang , West Cameroon. Malaria Journal. 2017; p. 1–9. doi: 10.1186/s12936-017-1722-2, publisher: BioMed Central.

Ryan JR, Stoute JA, Amon J, Dunton RF, Mtalib R, Koros J, Owour B, Luckhart S, Wirtz RA, Barnwell JW, Rosenberg R. Evidence for transmission of Plasmodium vivax among a Duffy antigen negative population in Western Kenya. American Journal of Tropical Medicine and Hygiene. 2006; 75(4):575–581. doi: 75/4/575 [pii], ISBN: 0002-9637 (Print)\r0002-9637 (Linking).

Suchard MA, Redelings BD. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioin-

formatics. 2006; 22(16):2047–2048. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl175, doi: 10.1093/bioinformatics/btl175.

Topazian HM, Gumbo A, Puerto-Meredith S, Njiko R, Mwanza A, Kayange M, Mwalilino D, Mvula B, Tegha G, Mvalo T, Edwards JK, Emch M, Pettifor A, Smith JS, Hoffman I, Meshnick SR, Juliano JJ. Asymptomatic Plasmodium falciparum malaria prevalence among adolescents and adults in Malawi, 2015–2016. Sci Rep. 2020; 10(1):18740. https://www.nature.com/articles/s41598-020-75261-9, doi: 10.1038/s41598-020-75261-9, number: 1 Publisher: Nature Publishing Group.

Twohig KA, Pfeffer DA, Baird JK, Price RN, Zimmerman PA, Hay SI, Gething PW, Battle KE, Howes RE. Growing evidence of Plasmodium vivax across malaria-endemic Africa. PLoS Neglected Tropical Diseases. 2019; 13(1). doi: 10.1371/journal.pntd.0007140.

Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. 2016 ed. Use R!, Springer International Publishing : Imprint: Springer; 2016. doi: 10.1007/978-3-319-24277-4.

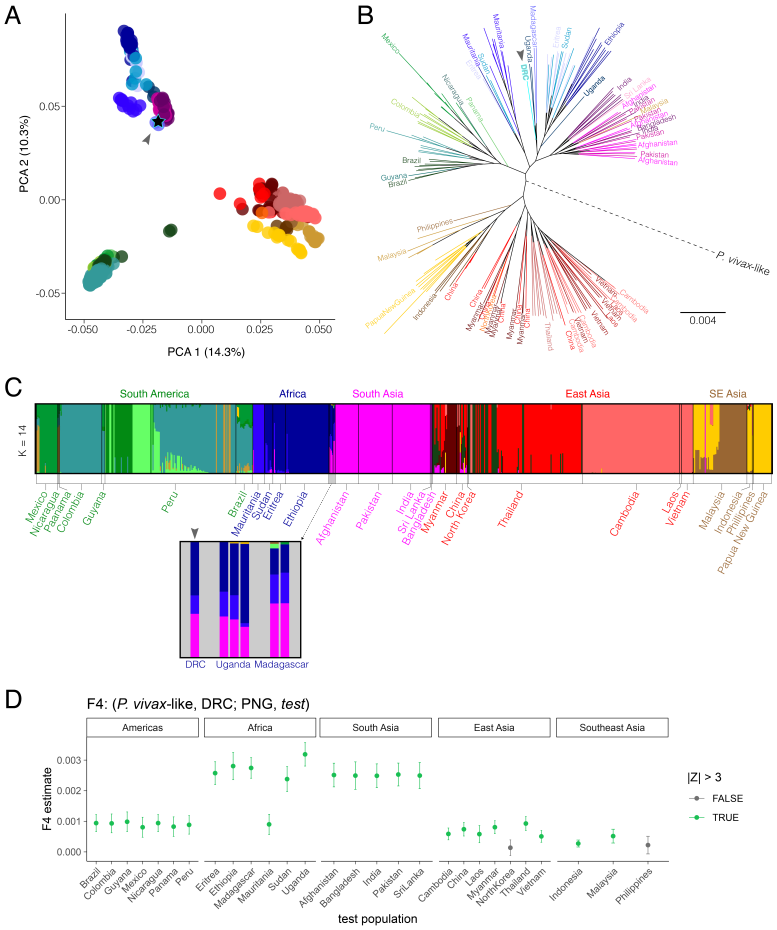

Figure 1. DRC P. vivax falls within African variation in global species context. DRC sample indicated by arrowhead in each panel. **A** Principal components analysis of global genetic diversity reveals *P. vivax* from the DRC grouped with African and South Asian populations. The first principal component explains 14.3% of variation in the data and the second principal component represents 10.3% of variation. **B** Maximum Likelihood tree shows *P. vivax* from the DRC clusters with Uganda and Madagascar. Tree constructed using whole-genome SNP data and rooted using two *P. vivax*-like sequences (red dotted line). Continent colors: Americas (green), Africa (blue), South Asia (pink), East Asia (red), Southeast Asia (brown). Tree constructed using IQtree with the GTR+ASC model to account for ascertainment bias in SNP data. **C** Admixture analysis of global population structure indicates *P. vivax* ancestry is geographically structured. Each vertical bar represents the proportion of genetic ancestry belonging to one individual *P. vivax* sample for each simulated population size (K). Population size K=14 is most-supported by Cross Validation Error. **D** F4 statistics calculated using Admixtools2 using the relationship (*P. vivax*-like, DRC; Papua New Guinea (PNG), *test*). Higher F4 estimates indicate DRC has a closer relationship with the test population than it does to the PNG population. Error bars indicate ±3 SE.

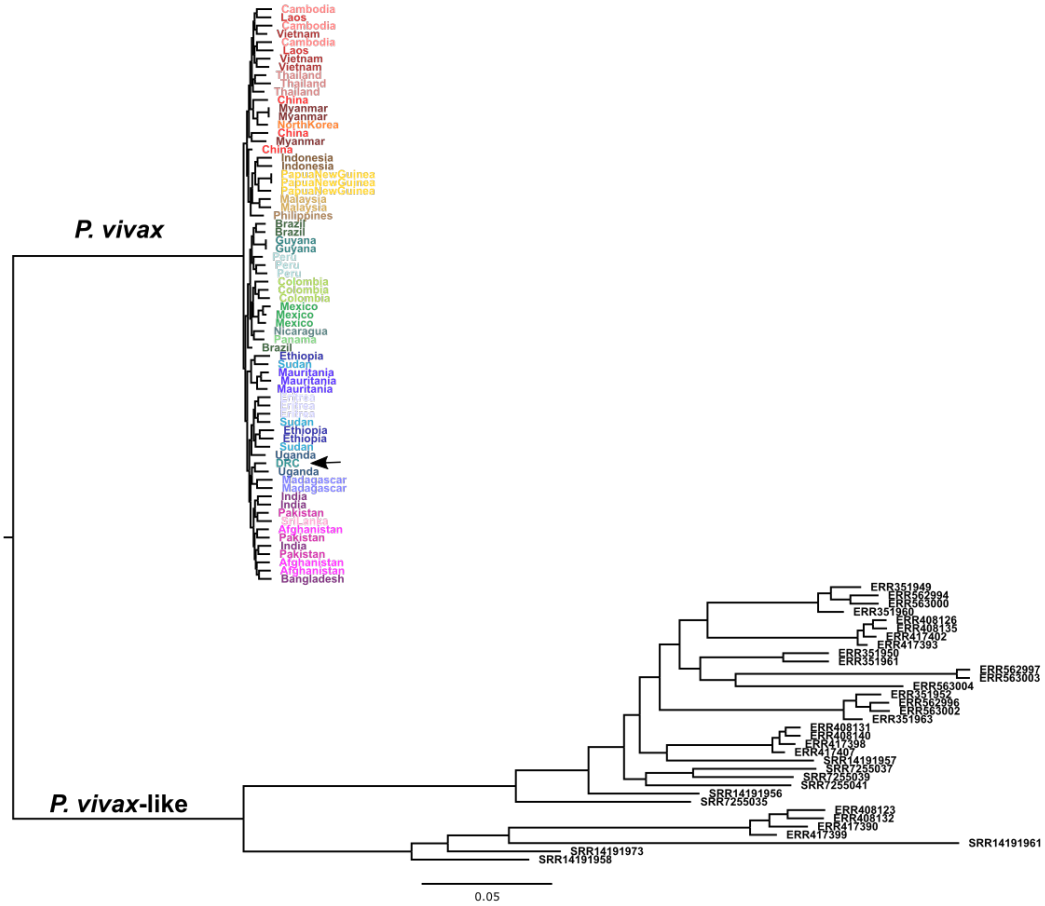

Figure 2. Maximum Likelihood tree shows DRC *P. vivax* sample branches with *P. vivax* populations and not *P. vivax*-like. Maximum Likelihood tree of nuclear genome SNP data shows that *P. vivax* from the DRC does not branch with *P. vivax*-like samples (black tip labels). The DRC sample, indicated here with an arrow, clusters with other African samples.

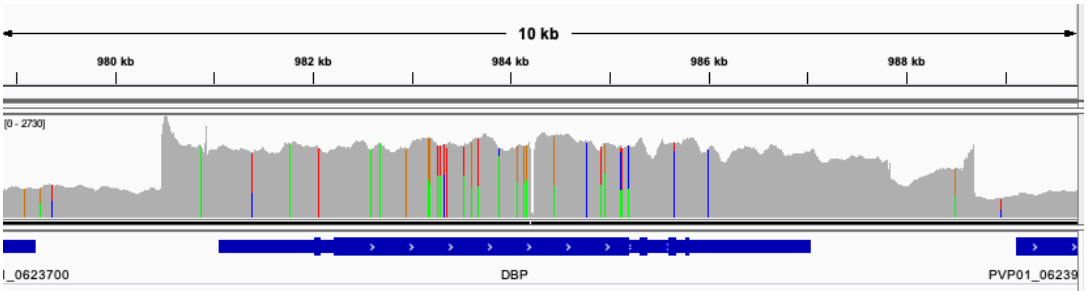

Figure 3. Read Pileup image of duplication of PvDBP in *P. vivax* from DRC. IGV view of the read depth for PvDBP in the DRC *P. vivax* sample where genome coordinate is on the X-axis and read depth is on the Y-axis. This shows an increase in read depth around PvDBP indicating a duplication, and the different variants present along this region suggest two distinct haplotypes.