# Whole Genome Sequencing of *Plasmodium vivax* Isolates Reveals Frequent Sequence and Structural Polymorphisms in Erythrocyte Binding Genes

Anthony Ford[1,2*], Daniel Kepple[2*], Beka Raya Abagero[3], Jordan Connors[1], Richard Pearson[4], Sarah Auburn[5], Sisay Getachew[6, 7], Colby Ford[1], Karthigayan Gunalan[8], Louis H. Miller[8], Daniel A. Janies[1], Julian C. Rayner[9], Guiyun Yan[10], Delenasaw Yewhalaw[3], Eugenia Lo[2]

* Co-first authors with equal contribution

[1] Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, USA

[2] Department of Biological Sciences, University of North Carolina at Charlotte, USA

[3] Tropical Infectious Disease Research Center, Jimma University, Ethiopia

[4] Malaria Programme, Wellcome Trust Sanger Institute, Hinxton, UK

[5] Global and Tropical Health Division, Menzies School of Health Research and Charles Darwin University, Darwin, Northern Territory, Australia

[6] College of Natural Sciences, Addis Ababa University, Ethiopia

[7] Armauer Hansen Research Institute, Addis Ababa, Ethiopia

[8] Laboratory of Malaria and Vector Research, NIAID/NIH, Bethesda, USA

[9] Department of Clinical Biochemistry, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 OXY, UK

[10] Program in Public Health, University of California at Irvine, USA

**Correspondence:** Anthony Ford, Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, USA; Guiyun Yan, Program in Public Health, University of California at Irvine, USA; Eugenia Lo, Department of Biological Sciences, University of North Carolina at Charlotte

**Running Title:** Genomic characteristics of *Plasmodium vivax* in Ethiopia

## Abstract

*Plasmodium vivax* malaria is much less common in Africa than the rest of the world because the parasite relies primarily on the Duffy antigen/chemokine receptor (*DARC*) to invade human erythrocytes, and the majority of Africans are Duffy negative. Recently, there has been a dramatic increase in the reporting of *P. vivax* cases in Africa, with a high number of them being in Duffy negative individuals, potentially indicating *P. vivax* has evolved an alternative invasion mechanism that can overcome Duffy negativity. Here, we analyzed single nucleotide polymorphism (SNP) and copy number variation (CNV) in Whole Genome Sequence (WGS) data from 44 *P. vivax* samples isolated from symptomatic malaria patients in southwestern Ethiopia, where both Duffy positive and Duffy negative individuals are found. A total of 236,351 SNPs were detected, of which 21.9% was nonsynonymous and 78.1% was synonymous mutations.  The largest number of SNPs were detected on chromosomes 9 (33,478 SNPs; 14% of total) and 10 (28,133 SNPs; 11.9%). There were particularly high levels of polymorphism in erythrocyte binding gene candidates including reticulocyte binding protein 2c (*RBP*2c), merozoite surface protein 1 (*MSP*1), and merozoite surface protein 3 (*MSP*3.5, *MSP*3.85 and *MSP*3.9). Thirteen genes related to immunogenicity and erythrocyte binding function were detected with significant signals of positive selection. Variation in gene copy number was also concentrated in genes involved in host-parasite interactions, including the expansion of the Duffy binding protein gene (*PvDBP*) on chromosome 6 and several *PIR* genes. Based on the phylogeny constructed from the whole genome sequences, the expansion of these genes was an independent process among the *P. vivax* lineages in Ethiopia. We further inferred transmission patterns of *P.*

*vivax* infections among study sites and showed various levels of gene flow at a small geographical scale. The genomic features of *P. vivax* provided baseline data for future comparison with those in Duffy-negative individuals, and allowed us to develop a panel of informative Single Nucleotide Polymorphic markers diagnostic at a micro-geographical scale.

## Introduction

Vivax malaria is the most geographically widespread human malaria, causing over 130 million clinical cases per year worldwide [1]. *Plasmodium vivax* can produce dormant liver-stage hypnozoites within infected hosts, giving rise to relapse infections from months to years. This unique feature of *P. vivax* contributes to an increase in transmission potential and increases the challenge of elimination [2]. Understanding *P. vivax* genome variation will advance our knowledge of parasite biology and host-parasite interactions, as well as identify potential drug resistance mechanisms [3, 4]. Such data will also help identify molecular targets for vaccine development [5-7], and provide new means to track the transmission and spread of drug resistant parasites [8-9].

Compared to *P. falciparum*, *P. vivax* isolates from Southeast Asia (e.g., Thailand and Myanmar), Pacific Oceania (Papua New Guinea), and South America (Mexico, Peru, and Colombia) have significantly higher nucleotide diversity at the genome level [2]. This could be the result of frequent gene flow via human movement, intense transmission, and/or variation in host susceptibility [10-14]. *P. vivax* infections are also much more likely to contain multiple parasite strains in areas where transmission is intense and/or relapse is common [10, 15-18]. In Papua New Guinea, for example, *P. vivax* infections had an approximately 3.5-fold higher rate of polyclonality and nearly double the multiplicity of infection (MOI) than the *P.* falciparum infections [16]. Similar rates of polyclonality and MOI have also been reported in *P*. *vivax* in Cambodia [6]. It is possible intense transmission has sustained a large and stable parasite population in these regions [17,18]. By contrast, geographical differentiation and selection pressure over generations can lead to fixation of parasite genotypes in local populations. In the Asia-Pacific region, *P. vivax* showed a high level of genetic relatedness through inbreeding among the dominant clones, in addition to strong selection imposed in a number of antimalarial drug resistance genes [19]. In Ethiopia, the chloroquine resistance transporter gene (*Pvcrt*) of *P. vivax* on chromosome 14 had been shown with significant selection in a region upstream of the promotor, highlighting the ability of *P. vivax* to rapidly evolve in response to control measures [20]. Apart from mutations, high copy number observed in *Pvcrt* and multidrug resistant gene (*Pvmdr*1) has also been shown to be associated with increased antimalaria drug resistance [21,22].

Recent genomic studies have indicated that some highly polymorphic genes in the *P. vivax* genome are associated with red blood cell invasion and immune evasion [10, 12, 19, 23]. They include the merozoite surface protein genes *MSP*1 (PVP01_0728900) and *MSP*7 (PVX_082665), Pv-fam-b (PVX_002525), Pv-fam-e (PVX_089875), the reticulocyte binding protein gene *RBP*2c (PVP01_0534300), serine-repeat antigen 3 (*SERA*; PVX_003840), as well as virulent genes (*VIR*) such as *VIR*22 (PVX_097530) and *VIR*12 (PVX_083590) [23-29]. Polymorphisms in genes associated with immune evasion and reticulocyte invasion have important implications for the invasion efficiency and severity of *P. vivax* infections. Members of the erythrocyte binding gene family, including reticulocyte binding proteins (*RBP*s), Duffy-binding proteins (*DBP*s), and merozoite surface proteins (*MSP*3 and *MSP*7) have been previously shown to exhibit high sequence variation in *P. vivax* [20, 30]. The polymorphisms in *RBP*1 and *RBP*2 genes may relate to an increased capability of erythrocyte invasion by *P. vivax* [31-33]. It has been suggested that Pv*RBP*2b-TfR1 interaction is vital for the initial recognition and invasion of host reticulocytes [34], prior to the engagement of *PvDBP1* and Duffy antigen chemokine receptor (*DARC*) and formation of a tight junction between parasite and erythrocyte [35]. Apart from Pv*RBP*, Reticulocyte Binding Surface Antigen (Pv*RBSA*) [36], an antigenic adhesin, may also play a key role in *P. vivax* parasites binding to target cells, possessing the capability of binding to a population of reticulocytes with a different Duffy phenotype [37, 38]. Another erythrocyte binding protein gene (Pv*EBP*), a paralog of *PvDBP1*, which harbors all the hallmarks of a *Plasmodium* red blood cell invasion protein, including conserved Duffy-binding like and C-terminal cysteine-rich domains [39], has been recently shown to be variable in copy number in the Malagasy *P.* vivax [39]. Functional analyses indicated that region II of this gene bound to both Duffy-positive and Duffy-negative reticulocytes, although at a lower frequency compared to *PvDBP*, suggestive of its role in erythrocyte invasion [40]. Both Pv*EBP*1 and Pv*EBP*2 genes exhibit high genetic diversity and are common antibody binding targets associated with clinical protection [41, 42]. Other proteins such as tryptophan-rich antigen gene (*TRAg*), anchored micronemal antigen (*GAMA*), and Rhoptry neck protein (*RON*) have also been suggested to play a role in red cell invasion, especially in low-density infections [43-47]. Information of the polymorphisms in these genes will have important implications on the dynamics of host-parasite interactions.

Compared to Southeast Asia and South America where *P. vivax* is highly endemic, data on polymorphisms in erythrocyte binding gene candidates of *P. vivax* from Africa is limited. Filling the gap is critical for identifying functional genes in erythrocyte invasion, biomarkers for tracking the African *P. vivax* isolates, as well as potential gene targets for vaccine development. It was previously thought that most African populations were immune to *P.* vivax infections due to the absence of *DARC* gene expression required for erythrocyte invasion. However, several recent reports have indicated the emergence and potential spread of *P. vivax* across Africa [32, 48-50]. The objective of this study was to describe genomic variation of *P. vivax* from Ethiopia. Specifically, we examined the level of genetic polymorphisms in a panel of 64 potential erythrocyte binding protein genes that have been suggested to play a role in the parasite-host invasion process. In addition, we inferred transmission patterns of *P. vivax* infections from different study sites based on the genetic variants. A recent study by Auburn *et al*. [20] has compared the genetic variants of *P. vivax* from Ethiopia with other geographical isolates. In the present study, we focus on the genomic characteristics of *P. vivax* among different study sites in Ethiopia with the goals to establish a baseline for genome comparison with the Duffy-negative *P. vivax* in our ongoing investigation, as well as to develop a panel of informative Single Nucleotide Polymorphic (SNP) markers diagnostic at a micro-geographical scale.

## Materials and Methods

### Ethics statement

Scientific and ethical clearance was obtained from the Institutional Scientific and Ethical Review Boards of Jimma and Addis Ababa Universities in Ethiopia, and The University of North Carolina, Charlotte, USA. Written informed consent/assent for study participation was obtained from all consenting heads of households, parents/guardians (for minors under age of 18), and each individual who was willing to participate in the study.

**Study area and sample collection**

Genomic DNA was extracted from 22 clinical samples collected in Jimma, southwestern Ethiopia during peak transmission season (September – November, 2016; Figure 1). Finger-pricked blood samples were collected from malaria symptomatic (who has fever with axillary body temperature > 37.5°C and with confirmed asexual stages of malaria parasite based on microscopy) or febrile patients visiting the health centers or hospitals at each of the study sites. Thick and thin blood smears were prepared for microscopic examination, and 4-6 ml of venous blood were collected from each *P. vivax*-confirmed patient in K2 EDTA blood collection tubes. For the whole blood samples, we used the Lymphoprep/Plasmodpur-based protocol to deplete the white blood cells and enrich the red blood cell pellets [51]. DNA was then extracted from approximately 1 ml of the red blood cell pellets using Zymo Bead Genomic DNA kit (Zymo Research) following the manufacturer's procedures. The extracted DNA were first assessed by nested and quantitative PCR methods to confirm and quantify *P. vivax* of the infected samples [52]. From a larger set of samples, we then performed microsatellite analyses using seven different loci [53].  Only monoclonal samples were selected and proceeded for sequencing. Whole genome sequencing was conducted on the Illumina HiSeq 3000 Sequencing Platform at the Wellcome Sanger Institute (European Nucleotide Archive [ENA] accession number of each sample in Table 1). The generated sequence reads were mapped individually to the publicly available reference genome PvP01 from Gene

DB using Bowtie version 2 [54]. The original 22 samples were processed to remove reads other than *P. vivax*. The percentage coverage of the *P. vivax* reads in our samples were high enough to not affect the results. An additional 24 sample sequence data were obtained as FASTQ files from the ENA. These samples were collected from Arbaminch, Badowacho, Halaba, and Hawassa in southwestern Ethiopia (Figure 1), the Duffy status of each of these 24 samples is unknown. They were then aligned to the PVP01 reference genome using BWA-MEMv.2 with default settings [55, 56]. The overall quality of each resulting BAM was assessed using FASTQC. Similarly, we concluded that the percentage of the *P. vivax* reads covered in the additional 24 samples were high enough to reflect the dominant signal of the variants and negate polyclonal influences. Two of our samples displayed a significant decline in average quality in read mapping and were therefore removed from further SNP variant and copy number variation analyses.

**SNP discovery, annotation, and filtering**

Potential SNPs were identified by SAM tools v.1.6 mpileup procedure [57] in conjunction with BCF tools v.1.6 [57] across all 44 sample BAM files using the PVP01 reference genome. Compared to the Salvador-I, the PVP01 reference genome consists of 14 major chromosomal sequences, and provides a greater level of gene annotation power and improved assembly of the subtelomeres [56]. We analyzed only sequence reads that were mapped to these 14 major chromosomal sequences. The hypervariable and subtelomeric regions in our samples were retained during the variant calling procedure and each sample BAM file had duplicates marked using SAMtools 1.6 markdup procedure. For the mpileup procedure, the maximum depth threshold, which determines the number of maximum reads per file at a position, was set to 3,000 million to ensure that the maximum amount of reads for each position was not reached. Samples were pooled together using a multisampling variant calling approach. The SNPs were then annotated with SnpEff v.4.3T [58] based on the annotated gene information in GeneDB. Filtering was done using the following standard metrics, including Read Position Bias, Mapping Quality vs Strand Bias, Raw read depth, Mapping Quality Bias, Base Quality Bias, and Variant Distant Bias produced by SAM tools and BCF tools during the variant calling procedure. In Snp Sift, data was filtered by choosing SNPs that had a Phred Quality score ≥ 40, a raw read depth (DP) ≥ 30, and a base quality bias >0.1 [59]. We then calculated the allele frequency for each SNP position for all 44 samples using the frequency procedure in VCF tools v.0.1.15 [60]. The total number of SNPs across all samples, as well as the number of nonsynonymous and synonymous mutations were recorded. Mutations were compared among the 14 chromosomes in addition to a panel of 64 erythrocyte binding genes.

**Copy number variation analyses**

Copy number variation of gene regions was assessed with CNVnator [61]. CNVnator uses mean-shift theory, a partitioning procedure based on an image processing technique and additional refinements including multiple bandwidth partitioning and GC correction [61]. We first calculated the read depth for each bin and correct GC-bias. This was followed by mean-shift based segment partition and signal merging, which employed an image processing technique. We then performed CNV calling, of which segments with a mean RD signal deviating by at least a quarter from genomic average read depth signal were selected and regions with a *P*-value less than 0.05 were called. A one-sided test was then performed to call additional copy number variants. SAM tools v.1.6 was utilized in our data preprocessing step to mark potential duplicates in the BAM files and followed the CNV detection pipeline [62]. We extracted the read mappings from each of BAM files for all chromosomes. Once the root file was constructed using the extracted reads, we generated histograms of the read depths using a bin size of 100. The statistical significance for the windows that showed unusual read depth was calculated and the chromosomes were partitioned into long regions that have similar read depth.

To validate the results from CNVnator, we used the GATK4 copy number detection pipeline to further examine gene copy number [63-65]. The read coverage counts were first obtained from pre-processed genomic intervals of a 1000-bp window length based on the PvP01 reference genome. The read fragment counts were then standardized using the Denoise Read Counts that involved two transformations. The first transformation was based on median counts, including the $\log_2$ transformation, and the counts were normalized to center around one. In the second transformation, the tool denoises was used to standardized copy ratios using principal component analysis.

## Test for positive selection

Regions of positive selection were examined among the 44 Ethiopian *P. vivax* isolates using the integrated haplotype score approach, specifically the SciKit-Allel for python, a package used for analysis of large scale genetic variation data [66]. Before the samples were run through Scikit-Allel, genotypes for each of the samples were phased using BEAGLE [67]. Genes that were detected with signals of positive selection by SciKit-Allel, as well as a panel of 64 potential erythrocyte binding genes were further evaluated using the PAML package (Phylogenetic Analysis by Maximum Likelihood) [68]. Using the codeml procedure in PAML, DNA sequences were analyzed with the maximum likelihood approach in a phylogenetic framework. The synonymous and nonsynonymous mutation rates between protein-coding DNA sequences were then estimated in order to identify potential regions of positive selection. We created two models, the neutral model M1 and the selection model M2. The average $d_N/d_S$ values were estimated across all branches in both M1 and M2 models and the average $d_N/d_S$ values across all sites in the M2 model. The $d_N/d_S$ values were compared between the two models using a likelihood ratio test for significant positive selection.

### Comparison of nucleotide diversity among EBP gene regions

Based on the literature [23-33], we identified 64 gene regions that are potentially related to erythrocyte binding in *P. vivax* (Supplementary Table 1). These included the *DBP* (duffy binding protein), *EBP* (erythrocyte binding protein), *MSP* (merozoite surface protein), and *RBP* (reticulocyte binding protein) multigene families, the tryptophan rich antigen gene family (*TRAg*), GPI-anchored microanemal antigen (*GAMA)*, microneme associated antigen (*MA)*, rhoptry associated adhesin (*RA)*, high molecular weight rhoptry protein 3 (*RHOP*3), and rhoptry neck protein (*RON)* genes. Previous study has shown that the transcriptome profiles of the *TRAg* genes were differentially transcribed at the erythrocytic stages, indicating that these genes may play specific roles in blood-stage development [43].  The reticulocyte binding protein multigene family encodes genes that each have a receptor on the surface that is essential for the host-invasion stage of *P. vivax* [69]. The *MSP* multigene family, currently assumed to be a candidate for vaccine generation, also plays a role in the invasion stage of *P. vivax* and is also immunogenic [26]. The nucleotide diversity of 64 potential erythrocyte binding genes were compared among the 44 *P. vivax* sample consensus sequences using DnaSP [70]. The Pairwise-Deletion method where gaps were ignored in each pairwise comparison was used for this calculation.

**Genetic relatedness and transmission network analyses**

Phylogenetic analyses were performed to infer the genetic relatedness among the 44 Ethiopian isolates. Sequence alignment was first conducted using a multiple sequence alignment program in MAFFT v. 7 [71]. The alignment was then trimmed to remove gaps using trimal (the *gappyout* option) that trimmed the alignments based on the gap percentage count over the whole alignment. After sequence editing, we concatenated all alignment files using FASconCAT-G [72], a perl program that allows for concatenation and translation (nucleotide to amino acid states) of multiple alignment files for phylogenetic analysis. We used the maximum likelihood method implemented in the Randomized Accelerated Maximum Likelihood (RAxML) v8 to construct phylogenetic trees [73]. The GTRGAMMA model was used for the best-scoring maximum likelihood tree. The GTR model incorporates the optimization of substitution rates and the GAMMA model accounts for rate heterogeneity.  A total of 100 rapid bootstrap runs were conducted to evaluate the confidence of genetic relationships. In addition, we performed principal component analyses using the glPCA function in R, a subset of the adegenet package [74], to determine the genetic relatedness of the samples among the different study sites in Ethiopia. A transmission network was created using StrainHub, a tool for generating transmission networks using phylogenetic information along with isolate metadata [75]. The transmission network was generated using the locations of the samples as the nodes and calculating the source hub ratio between each location. The source hub ratio was calculated by the number of transitions originating from a node over the total number of transitions related to that node. A node with a ratio close to 1 indicates a source, a ratio close to 0.5 indicates a hub, and a ratio close to 0 indicates a sink for the *P. vivax* infections.

## Results

### Distribution of SNPs among the chromosomes and EBP genes

A total of 252,973 SNPs were detected among the 44 Ethiopian *P. vivax* samples (Figure 2), with 21.5% (54,336 out of 252,973) nonsynonymous and 78.5% (198,637 out of 252,973) synonymous mutations (Figure 3A). The highest number of SNPs were observed on chromosomes 7 (28,856 SNPs; 11.4%), 9 (28,308 SNPs; 11.2%), and 12 (28,190 SNPs; 11.1%); whereas the lowest number of SNPs were observed on chromosomes 3 (6,803 SNPs; 2.7%), 6 (5,044 SNPs; ~2%), and 13 (8,809 SNPs; 3.5%; Figure 3A; Supplementary Table 2).

The 64 erythrocyte binding genes accounted for 3,607 of the total SNPs, with 1685 (46.7%) identified as nonsynonymous and 1922 (53.3%) as synonymous mutations

(Figure 3B). Among these genes, the highest number of SNPs were observed in reticulocyte binding protein gene (*RBP*2c) on chromosome 5, followed by the *MSP*3 multigene family (*MSP*3.5, *MSP*3.9 and *MSP*3.8) on chromosome 10. Nucleotide diversity also showed to be highest in the *RBP* and *MSP*3 multigene families, with an average nucleotide diversity of 1.3% and 2.8%, respectively, among our samples (Figure 3B). By contrast, the lowest number of SNPs were observed in the Duffy binding protein gene (*DBP*1) on chromosome 6 with a total of 13 SNPs, of which 12 were identified as nonsynonymous and one as synonymous mutations (Figure 3B). Likewise, another erythrocyte binding protein (*EBP*2), located also on chromosome 6, was one of the least variable genes with only one nonsynonymous mutation. The *TRAg* gene family also showed a low level of nucleotide diversity when compared to the other *EBP* gene families with an average nucleotide diversity of 0.2% (Figure 3B).

**Gene regions under positive selection**

Based on the integrated haplotype scores, positive selection was detected in 13 gene regions (Figure 4). These included the sub-telomeric protein 1 (*STP*1) on chromosome 5, the membrane associated erythrocyte binding-like protein (*MAEBL*) on chromosome 9, *MSP*3.8 on chromosome 10, as well as various plasmodium interspersed repeats (*PIR)* protein genes on chromosomes 3, 5, 7, 10, 11, and 12 (Figure 4). Based on PAML, 25 out of the 64 erythrocyte binding genes showed evidence of positive selection (Table 2; Supplementary Table 3). The majority of these genes belong to the *TRAg* multigene family. The *TRAg* genes had an average $d_N/d_S$ ratio of 2.75 across all branches and an average of 5.75 across all sites for the M2 model tested for selection

(Table 2). Compared to the other *TRAg* genes, *TRAg*15 had more sites detected under positive selection, with 50 of the sites showing a posterior probability greater than 50% and 43 showing a posterior probability greater than 95% (Table 2). While the *TRAg*4 gene had the highest $d_N/d_S$ ratio across all sites among other *TRAg* genes, only six sites were shown under positive selection with a posterior probability greater than 50% and one with a posterior probability greater than 95%.

All *RBP* genes, except for *RBP*2c, showed regions with significant signals of positive selection (average $d_N/d_S$ ratio across all sites: 5.11; Table 2). Among them, *RBP2*p1 had the largest number of sites with posterior probabilities greater than 95% (Table 2). Among all the *MSP* genes, only *MSP*5, *MSP*9, and *MSP*10 indicated regions under positive selection. The *MSP*5 and *MSP*9 genes had an average $d_N/d_S$ ratio of 3.85 across all sites and 1.11 across branches (Table 2). While *MSP*10 had an average $d_N/d_S$ ratio of 1.16 across all branches and less than 1 across all sites, only seven sites were indicated with posterior probabilities greater than 50% and 95% (Table 2). Although *MSP*3.8 showed potential positive selection based on the integrated haplotype scores (Figure 4), PAML did not show significant evidence of positive selection. For the *DBP* gene family, *DBP*9 showed the highest $d_N/d_S$ ratio across all sites and branches (10.39 and 3.88, respectively; Table 2).

**Copy number variation and evolution of high-order copy variants**

According to CNVnator, 19 gene regions showed copy number variation among our samples (Figure 5; Supplementary Table 4). Among them, 11 gene regions were detected with up to 2-3 copies and 8 gene regions with 4 copies or higher. We observed copy number variation in several *PIR* genes distributed across chromosomes 1, 2, 4, 5, 7, 10 and 12 (Figure 5; Supplementary Table 4). Specifically, for the *PIR* genes located on chromosome 2 (including PVP01_0220700, PVP01_0200200, PVP01_0200300, and PVP01_0200100; Figure 5), more than 20% of the samples had 2-3 copies and approximately 2-4% of the samples had 4 copies or higher. Among the 64 erythrocyte binding genes, duplications were observed in *DBP*1 on chromosome 6 and *MSP*3 on chromosome 10. *DBP*1 ranged from one to as high as five copies, and *MSP*3 ranged from one to as high as three copies among our samples (Figure 5), consistent with previous findings [19, 20, 76]. The remaining erythrocyte binding genes were detected with a single copy across our samples.

A maximum likelihood tree constructed based on the whole genome sequences showed an admixture of *P. vivax* isolates with single and multiple *PvDBP* copy number (Figure 6A). The Ethiopian *P. vivax* isolates were divided into six subclades. Subclade I contained *P. vivax* samples mostly from Arbaminch and Badowacho with both one and two *PvDBP* copies. Subclade II contained samples from Jimma and Hawassa with two *PvDBP* copies. Subclade III contained a mixture of *P. vivax* samples from Arbaminch, Halaba, Hawassa, and Jimma with single and high-order *PvDBP* copies. This clade was sister to subclade IV that contained *P. vivax* samples mostly from Jimma (Figure 6A). In subclade IV, no distinct clusters were detected between isolates with single and multiple *PvDBP*. Subclade V contained samples from Jimma and subclade VI contained samples from Arbaminch, Badowacho, Hawassa, and Halaba. Each of the subclades had samples with both one and two *PvDBP* copies. Similar patterns were observed in the *MSP*3 and *PIR* genes where *P. vivax* isolates with single and multiple copies were clustered together in separate subclades (Figures 6B-D), suggesting that these gene regions could have expanded multiply among samples at different locations.

### Gene flow and transmission network of the Ethiopian *P. vivax*

The principal component analysis based on the SNP variants showed samples from Arbaminch, Badowacho, Hawassa, and Halaba were genetically closely related but differentiated from Jimma (Figure 7A). The transmission network indicated that Arbaminch was the major source or hub of infections where the infections in Jimma, Hawassa, Badowacho, and Halaba were originated from (Table 3; Figure 7B). On the other hand, no transmission was originated from Halaba, making this location the largest sink of transmissions. The greatest extent of gene flow was observed between Arbaminch and Badowacho (Figure 7B). Hawassa and Jimma showed a source hub ratio of 0.5, indicating that there are equally as many egress transmissions as ingress transmissions (Table 3). Although Jimma and Badowacho/Halaba are in close geographical proximity, no apparent gene flow was observed between these sites.

## Discussion

Across the genome, the total number of SNPs observed among 44 *P. vivax* isolates in Ethiopia were comparable to those previously reported in South American [77] and Southeast Asian countries [19]. For instance, 303,616 high-quality SNPs were detected in 228 *P. vivax* isolates from Southeast Asia and Oceania in a previous study, of which Sal-I was used as the reference sequence and subtelomeric regions were discarded [19]. Auburn *et al*. [20] found that the average nucleotide diversity in Ethiopia was lower than in Thailand and Indonesia, but higher than in Malaysia. Chromosomes 3, 4, and 5 have been previously shown to contain the lowest proportion of synonymous SNPs than the other parts of the genome [12]. In the present study, chromosomes 3 and 6 were found to have the lowest number of both synonymous and nonsynonymous SNPs. This follows observations made in other studies done with nucleotide diversity ranging from 0.8 SNPs per kb in North Korea to 0.59 SNPs per kb in Peru [78]. Among the 64 erythrocyte binding gene candidates, the MSP and RBP multigene families showed the highest level of genetic variation. This agrees with previous studies that reported a remarkably high diversity in *RBP*2 than in *RBP*1 and its homolog group in *P. falciparum* [31]. In the Greater Mekong Subregion, the *MSP*3 and *PIR* gene families also indicated high levels of genetic diversity with 1.96% and 1.92% SNPs per base respectively, confirming that members of multigene families are highly variable genetically [30, 79]. Such diversity suggested that the binding domains of these genes could be under differential selection pressure. This pattern has been observed in previous studies and is likely due to their critical role in reticulocyte invasion, immunogenic properties, and human migration [26, 80-82].

Both CNVnator and GATK4 showed high order copies in several *PIR* gene regions. In addition, the *PIR* and *STP*1 genes were also indicated with significant selection based on the iHS calculations. The *PIR* gene family, which includes *STP*1, are located on the subtelomere regions and is a highly variable multigene family ranging from 1,200 genes in the reference strain PvP01 to 346 genes in monkey-adapted strain Salvador-I [56, 83]. Our analyses included only SNP variants that had a quality score of 40 or higher. Also, we used the PVP01 reference genome to map and annotate the subtelomeric regions, with the goal to reflect variability and features across the entire chromosome; whereas previous studies used the Sal-I reference genome with hypervariable and subtelomeric regions removed to minimize mapping errors [19, 84]. A recent study in *P. chabaudi* suggested that polymorphisms in *PIR* genes could affect the virulence of the parasites following passage from the mosquitoes [85]. Such a variation in copy number of the *PIR* gene family has also been reported in *P. cynomolgi* and *P. vivax* [86], suggesting that gene duplication could have been occurred repeatedly in the ancestral lineages [86]. The *PIR* multigene family is one of the largest gene families identified so far in *P. vivax* with several different potential functions. Some *PIR* genes encode proteins on the surface of infected red blood cells, which could confer to immune evasion; others encode proteins involved in signaling, trafficking and adhesion functions [83]. Positive selection detected in the *PIR* genes among the Ethiopian *P. vivax* isolates may have important implications on the susceptibility of the mosquito hosts [87].

For the *P. vivax* isolates in Southeast Asia, copy number variation was observed in nine gene regions including *DBP*1, *MDR*1, and PVX_101445 (on chromosome 14) with copy number ranging from 3 to 4 [19]. *DBP*1 and *MSP*3 showed higher order copies when compared to other genomic regions. In this study, the highest and most variable copy number variations were detected in the *DBP*1, with copy numbers ranging from one to as high as five. Likewise, for the *MSP*3, copy numbers ranging from one to as high as four. Based on the phylogeny, *DBP*1 and *MSP*3 expansion had occurred multiple times as tandem copies. These findings were consistent with earlier studies [19, 76] and suggested that gene expansion may play a key role in host cell invasion

[88]. For all other putative erythrocyte binding genes, only a single copy was detected among all samples. A larger sample in future investigations would verify this observation.

In the present study, we identified a panel of 64 putative erythrocyte binding gene candidates based on the information from the literature and analyzed their polymorphisms. However, we did validate the function for each of these genes.  Among these 64 putative erythrocyte binding gene candidates, *MAEBL* was shown to be highly conserved in *Plasmodium* [89], had the highest signal for positive selection among the *P. vivax* samples in Ethiopia. In *P. berghei*, *MAEBL* is a sporozoite attachment protein that plays a role in binding and infecting the mosquito salivary gland [89]. In *P. falciparum*, *MAEBL* is located in the rhoptries and on the surface of mature merozoites, and expresses at the beginning of schizogony [89]. In *P. vivax*, *MAEBL* is a conserved antigen expressed in blood stages, as well as in the mosquito midgut and salivary gland sporozoites [89, 90]. The *MAEBL* antigen contains at least 25 predicted B-cell epitopes that are likely to elicit antibody-dependent immune responses [91]. Positive selection observed in this gene region among the Ethiopian *P. vivax* isolates could be associated with the immunity-mediated selection pressure against blood-stage antigens. Though *DBP*1 had the highest and most diverse copy number variation, no significant signal of positive selection was detected.

It is noteworthy that the calculation of integrated haplotype scores and the accuracy of phasing genotypes using BEAGLE were dependent on the levels of linkage disequilibrium of the whole genomes. The higher the levels of linkage disequilibrium, the more accurate are the phased genotypes and thus the iHS score. Pearson *et al*. [19]

found that *P. vivax* experienced drops in linkage disequilibrium after correcting for population structure and other confounders. Linkage disequilibrium of *P. vivax* genomes has been previously shown to be associated with the rate of genetic recombination and transmission intensity [92-94]. In high transmission sites of Papua New Guinea and the Solomon Islands, no identical haplotypes and no significant multilocus LD were observed, indicating limited inbreeding and random associations between alleles in the parasite populations [95, 96]. However, when transmission intensity declined, similar haplotypes and significant LD were observed possibly due to self-fertilization, inbreeding and/or recombination of similar parasite strains [92]. Multilocus LD is significantly associated with the genetic relatedness of the parasite strains [97], but inversely associated with the proportion of polyclonal infections [98]. In Southwestern Ethiopia, malaria transmission ranged from low to moderate, and LD levels varied markedly among the study sites [53, 99]. To address this limitation in BEAGLE, all genes that were detected with positive selection in BEAGLE were further analyzed with PAML for verification. Future study should include broad samples to thoroughly investigate selection pressure at the population level and the function significance of polymorphisms in the *MAEB*L and *PIR* genes.

Previous studies have shown high levels of genetic diversity among *P. vivax* isolates in endemic countries [16, 100, 101]. Such a diversity was directly related to high transmission intensity and/or frequent gene exchange between parasite populations via human movement [4, 12 , 13, 53]. For example, previous studies using microsatellites have demonstrated a consistently high level of intra-population diversity ($H_E$ = 0.83) but low between-population differentiation ($F_{ST}$ ranged from 0.001-0.1] in broader regions of

Ethiopia [53, 99]. High heterozygosity was also observed in *P. vivax* populations from Qatar, India, and Sudan (average $H_{E\,=}$ 0.78; 62), with only slight differentiation from *P. vivax* in Ethiopia ($F_{ST}$ = 0.19) [102]. Frequent inbreeding among dominant clones [92, 95] and strong selective pressures especially in relapse infections [19, 20, 102, 103] may also contribute to close genetic relatedness between and within populations. Thus, in this study, it is not surprising to detect a high level of parasite gene flow among the study sites at a small geographical scale, despite the limited number of samples. In the present study, we successfully employed a transmission network model to identify transmission paths, as well as the source and sink of infections in the region, beyond simply indicating genetic relationships.

To conclude, this study elaborated on the genomic features of *P. vivax* in Ethiopia, particularly focusing polymorphisms in erythrocyte binding genes that potentially play a key role in local parasite invasion, a critical question given the mixed Duffy positive and negative populations of Ethiopia. The findings provided baseline information on the genomic variability of *P. vivax* infections in Ethiopia and allowed us to compare the genomic variants of *P. vivax* between Duffy-positive and Duffy-negative individuals as the next step of our ongoing investigation. Further, we are in progress of developing a panel of informative SNP markers to track transmission at a micro-geographical scale.

## Data Availability

Additional information is provided as supplementary data accompanies this paper. Sequence data of this study are deposited in the European Nucleotide Archive (ENA) and the accession number of each sample is listed in Table 1.

## Acknowledgements

We are greatly indebted to the staffs and technicians from Jimma University for field sample collection, the communities and hospitals for their support and willingness to participate in this research.

## Funding

This research was funded by National Institutes of Health (NIH R15 AI138002 to EL; NIH U19 AI129326 to GY; NIH R01 AI050243 to GY; D43 TW001505 to GY) and The Wellcome Trust 206194/Z/17/Z to JR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Competing interests

The authors have declaredthat no competing interests exist.

## References

1. World Health Organization. World Malaria Report 2018. WHO, Geneva.
2. White MT, Shirreff G, Karl S, Ghani AC, Mueller I. Variation in relapse frequency and the transmission potential of *Plasmodium vivax* malaria. Proc Biol Sci. 2016;283: 20160048.
3. Hemingway J, Shretta R, Wells TNC, Bell D, Djimdé AA, Achee N, et al. Tools and strategies for malaria control and elimination: what do we need to achieve a grand convergence in malaria? PLOS Biol. 2016;14: e1002380.

4. Hupalo DN, Luo Z, Melnikov A, Sutton PL, Rogov P, Escalante A, et al. Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*. Nat Genet. 2016;48: 953–8.

5. Vallejo AF, Martinez NL, Tobon A, Alger J, Lacerda MV, Kajava AV, et al. Global genetic diversity of the *Plasmodium vivax* transmission-blocking vaccine candidate Pvs48/45. Malar J. 2016;15: 202.

6. Tham WH, Beeson JG, Rayer JC. *Plasmodium vivax* vaccine research - we've only just begun. Int J Parasitol. 2017;47: 111-118.

7. Kale S, Yadav CP, Rao PN, Shalini S, Eapen A, Srivasatava HC, Sharma SK, Pande V, Carlton JM, Singh OP, Mallick PK. Antibody responses within two leading *Plasmodium vivax* vaccine candidate antigens in three geographically diverse malaria-endemic regions of India. Malar J. 2019;18: 425.

8. Baniecki ML, Faust AL, Schaffner SF, Park DJ, Galinsky K, Daniels RF, et al. Development of a single nucleotide polymorphism barcode to genotype *Plasmodium vivax* infections. PLoS Negl Trop Dis. 2015;9: e0003539.

9. Diez Benavente E, Campos M, Phelan J, Nolder D, Dombrowski JG, Marinho CRF, et al. A molecular barcode to inform the geographical origin and transmission dynamics of *Plasmodium vivax* malaria. PLoS Genet. 2020;16: e1008576.

10. Parobek CM, Lin JT, Saunders DL, Barnett EJ, Lon C, Lanteri CA, et al. Selective sweep suggests transcriptional regulation may underlie *Plasmodium vivax* resilience to malaria control measures in Cambodia. Proc. Natl. Acad. Sci. U.S.A. 2016;113: 50.

11. Auburn S, Benavente ED, Miotto O, Pearson RD, Amato R, Grigg MJ, et al. Genomic analysis of a pre-elimination Malaysian *Plasmodium vivax* population reveals selective pressures and changing transmission dynamics. Nat Commun. 2018;9: 2585.

12. Benavente ED, Ward Z, Chan W, Mohareb FR, Sutherland CJ, Roper C, et al. Genomic variation in *Plasmodium vivax* malaria reveals regions under selective pressure. PLoS One. 2017;12: 5.

13. Lima-Junior JDC, Pratt-Riccio LR. Major histocompatibility complex and malaria: focus on *Plasmodium vivax* Infection. Front. Immunol. 2016;7: 13.

14. Kano FS, Souza AMD, Torres LDM, Costa MA, Souza-Silva FA, Sanchez BAM, et al. Susceptibility to *Plasmodium vivax* malaria associated with DARC (Duffy antigen) polymorphisms is influenced by the time of exposure to malaria. Sci. Rep. 2018;8: 13851.

15. Ventocilla JA, Nuñez J, Tapia LL, Lucas CM, Manock SR, Lescano AG, et al. Genetic variability of *Plasmodium vivax* in the north coast of Peru and the Ecuadorian amazon basin. Am. J. Trop. Med. Hyg. 2018;99: 27–32.

16. Fola AA, Harrison GLA, Hazairin MH, Barnadas C, Hetzel MW, Iga J, et al. Higher complexity of infection and genetic diversity of *Plasmodium vivax* than *Plasmodium falciparum* across all malaria transmission zones of Papua New Guinea. Am. J. Trop. Med. Hyg. 2017;16–0716.

17. Barry AE, Waltmann A, Koepfli C, Barnadas C, Mueller I. Uncovering the transmission dynamics of *Plasmodium vivax* using population genetics. Pathog. Glob. Health. 2015;109: 142–52.

18. Koepfli C, Rodrigues PT, Antao T, Orjuela-Sánchez P, Eede PVD, Gamboa D, et al. *Plasmodium vivax* diversity and population structure across four continents. PLoS Negl. Trop. Dis. 2015;9: e0003872.

19. Pearson RD, Amato R, Auburn S, Miotto O, Almagro-Garcia J, Amaratunga C, et al. Genomic analysis of local variation and recent evolution in *Plasmodium vivax*. Nat Genet. 2016;48: 959–64.

20. Auburn S, Getachew S, Pearson RD, Amato R, Miotto O, Trimarsanto H, et al. Genomic analysis of *Plasmodium vivax* in southern Ethiopia reveals selective pressures in multiple parasite mechanisms. J. Infect. Dis. 2019;220: 1738–49.

21. Costa GL, Amaral LC, Fontes CJF, Carvalho LH, Brito CFAD, Sousa TND. Assessment of copy number variation in genes related to drug resistance in *Plasmodium vivax* and *Plasmodium falciparum* isolates from the Brazilian Amazon and a systematic review of the literature. Malar. J. 2017;16: 152.

22. Lin JT, Muth S, Rogers WO, Ubalee R, Kharabora O, Juliano JJ, et al. *Plasmodium vivax* isolates from Cambodia and Thailand show high genetic complexity and distinct patterns of *P. vivax* multidrug resistance gene 1 (pvmdr1) polymorphisms. Am. J. Trop. Med. Hyg. 2013;88: 1116–23.

23. Cornejo OE, Fisher D, Escalante AA. Genome-wide patterns of genetic polymorphism and signatures of selection in *Plasmodium vivax*. Genome Biol. Evol. 2014;7: 106–19.

24. Chen E, Salinas ND, Huang Y, Ntumngia F, Plasencia MD, Gross ML, et al. Broadly neutralizing epitopes in the *Plasmodium vivax* vaccine candidate duffy binding protein. Proc. Natl. Acad. Sci. U.S.A. 2016;113: 6277–82.

25. Singh V, Gupta P, Pande V. Revisiting the multigene families: *Plasmodium* var and vir genes. J Vector Borne Dis. 2014;51: 75–81.

26. Rice BL, Acosta MM, Pacheco MA, Carlton JM, Barnwell JW, Escalante AA. The origin and diversification of the merozoite surface protein 3 (msp3) multi-gene family in *Plasmodium vivax* and related parasites. Mol. Phylogenetics Evol. 2014;78: 172–84.

27. Lu F, Li J, Wang B, Cheng Y, Kong D-H, Cui L, et al. Profiling the humoral immune responses to *Plasmodium vivax* infection and identification of candidate immunogenic rhoptry-associated membrane antigen (RAMA). J. Proteom. 2014;102: 66–82.

28. Rahul C, Krishna KS, Pawar AP, Bai M, Kumar V, Phadke S, et al. Genetic and structural characterization of PvSERA4: potential implication as therapeutic target for *Plasmodium vivax* malaria. J. Biomol. Struct. Dyn. 2013;32: 580–90.

29. Rahul C, Krishna KS, Meera M, Phadke S, Rajesh V. *Plasmodium vivax*: N-terminal diversity in the blood stage SERA genes from Indian isolates. Blood Cells Mol. Dis. 2015;55: 30–5.

30. Chen S-B, Wang Y, Kassegne K, Xu B, Shen H-M, Chen J-H. Whole-genome sequencing of a *Plasmodium vivax* clinical isolate exhibits geographical characteristics and high genetic variation in China-Myanmar border area. BMC Genom. 2017;18: 131.

31. Rayner JC, Corredor V, Tran TM, Barnwell JW, Huber CS, Galinski MR. Dramatic difference in diversity between *Plasmodium falciparum* and *Plasmodium Vivax* reticulocyte binding-like genes. Am. J. Trop. Med. Hyg. 2005;72: 666–74.

32. Gunalan K, Niangaly A, Thera MA, Doumbo OK, Miller LH. *Plasmodium vivax* infections of duffy-negative erythrocytes: historically undetected or a recent adaptation? Trends Parasitol. 2018;34: 420–9.

33. Luo Z, Sullivan SA, Carlton JM. The biology of *Plasmodium vivax* explored through genomics. Ann. N. Y. Acad. Sci. 2015;1342: 53–61.

34. Gruszczyk J, Huang RK, Chan L-J, Menant S, Hong C, Murphy JM, et al. Cryo-EM structure of an essential *Plasmodium vivax* invasion complex. Nature. 2018;559: 135–9.

35. Chan LJ, Dietrich MH, Nguitragool W, Tham WH. *Plasmodium vivax* reticulocyte binding proteins for invasion into reticulocytes. Cell. Microbiol. 2019; e13110.

36. Moreno-Pérez DA, Baquero LA, Chitiva-Ardila DM, Patarroyo MA. Characterising PvRBSA: an exclusive protein from *Plasmodium* species infecting reticulocytes. Parasites Vectors. 2017;10: 243.

37. Camargo-Ayala PA, Garzón-Ospina D, Moreno-Pérez DA, Ricaurte-Contreras LA, Noya O, Patarroyo MA. On the evolution and function of *Plasmodium vivax* reticulocyte binding surface antigen (pvrbsa). Front. Genet. 2018;9: 372.

38. Roesch C, Popovici J, Bin S, Run V, Kim S, Ramboarina S, et al. Genetic diversity in two *Plasmodium vivax* protein ligands for reticulocyte invasion. PLOS Negl. Trop. Dis. 2018;12: e0006555.

39. Hester J, Chan ER, Menard D, Mercereau-Puijalon O, Barnwell J, Zimmerman PA, et al. *De-novo* assembly of a field isolate genome reveals novel *Plasmodium vivax* erythrocyte invasion genes. PLOS Negl. Trop. Dis. 2013;7: e2569.

40. Ntumngia FB, Thomson-Luque R, Torres LDM, Gunalan K, Carvalho LH, Adams JH. A novel erythrocyte binding protein of *Plasmodium vivax* suggests an alternate invasion pathway into duffy-positive reticulocytes. mBio. 2016;7: e01261-16.

41. Carias LL, Dechavanne S, Nicolete VC, Sreng S, Suon S, Amaratunga C, et al. Identification and characterization of functional human monoclonal antibodies to *Plasmodium vivax* duffy-binding protein. J. Immunol. 2019;202: 2648–60.

42. He WQ, Shakri AR, Bhardwaj R, Franca CT, Stanisic DI, Healer J, et al. Antibody responses to *Plasmodium vivax* duffy binding and erythrocyte binding proteins predict risk of infection and are associated with protection from clinical Malaria. PLOS Negl. Trop. Dis. 2019;13: e0006987.

43. Wang B, Lu F, Cheng Y, Chen J-H, Jeon H-Y, Ha K-S, et al. Immunoprofiling of the tryptophan-rich antigen family in *Plasmodium vivax*. Infect. Immun. 2015;83: 3083–95.

44. Baquero LA, Moreno-Pérez DA, Garzón-Ospina D, Forero-Rodríguez J, Ortiz-Suárez HD, Patarroyo MA. PvGAMA reticulocyte binding activity: predicting conserved functional regions by natural selection analysis. Parasites Vectors. 2017;10: 251.

45. Arévalo-Pinzón G, Bermúdez M, Curtidor H, Patarroyo MA. The *Plasmodium vivax* rhoptry neck protein 5 is expressed in the apical pole of *Plasmodium vivax* VCG-1 strain schizonts and binds to human reticulocytes. Malar. J. 2015;14: 106.

46. Tyagi K, Hossain ME, Thakur V, Aggarwal P, Malhotra P, Mohmmed A, et al. *Plasmodium vivax* tryptophan rich antigen PvTRAg36.6 interacts with PvETRAMP and PvTRAg56.6 interacts with PvMSP7 during erythrocytic stages of the parasite. Plos One. 2016;11: e0151065.

47. Gunalan K, Sá JM, Barros RRM, Anzick SL, Caleon RL, Mershon JP, et al. Transcriptome profiling of *Plasmodium vivax* in Samira monkeys identifies potential ligands for invasion. Proc. Natl. Acad. Sci. U.S.A. 2019;116: 7053–61.

48. Zimmerman PA. *Plasmodium vivax* infection in duffy-negative people in Africa. Am. J. Trop. Med. Hyg. 2017;97: 636–8.

49. Battle KE, Lucas TCD, Nguyen M, Howes RE, Nandi AK, Twohig KA, et al. Mapping the global endemicity and clinical burden of *Plasmodium vivax*, 2000–17: a spatial and temporal modelling study. Lancet. 2019;394: 332–43.

50. Twohig KA, Pfeffer DA, Baird JK, Price RN, Zimmerman PA, Hay SI, et al. Growing evidence of *Plasmodium vivax* across malaria-endemic Africa. PLOS Negl. Trop. Dis. 2019;13: e0007140.

51. Auburn S, Campino S, Clark TG, Djimde AA, Zongo I, Pinches R, et al. An Effective Method to Purify Plasmodium falciparum DNA Directly from Clinical Blood Samples for Whole Genome High-Throughput Sequencing. Plos One. 2011;6: e22213.

52. Lo E, Yewhalaw D, Zhong D, Zemene E, Degefa T, Tushune K, et al. Molecular epidemiology of *Plasmodium vivax* and *Plasmodium falciparum* malaria among duffy-positive and duffy-negative populations in Ethiopia. Malar. J. 2015;14: 84.

53. Lo E, Hemming-Schroeder E, Yewhalaw D, Nguyen J, Kebede E, Zemene E, et al. Transmission dynamics of co-endemic *Plasmodium vivax* and *P. falciparum* in Ethiopia and prevalence of antimalarial resistant genotypes. PLOS Negl. Trop. Dis. 2017;11: e0005806.

54. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat. Methods. 2012;9: 357–9.

55. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

56. Auburn S, Böhme U, Steinbiss S, Trimarsanto H, Hostetler J, Sanders M, et al. A new *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. Wellcome Open Res. 2016;1: 4.

57. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27: 2987–93.

58. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012;6: 80–92.

59. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front. Genet. 2012;3: 35.

60. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, Depristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–8.

61. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21: 974–84.

62. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

63. 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.

64. Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43: 491–8.

65. Auwera GAVD, Carneiro MO, Hartl C, Poplin R, Angel GD, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43: 1-11.

66. Miles A, Harding N. cggh/scikit-allel: v1.1.8 [Internet]. Zenodo. 2017. Available from: https://zenodo.org/record/822784#.XKIe6yhKiUk

67. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81: 1084-97.

68. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 2007;24: 1586–91.

69. Prajapati SK, Singh OP. Insights into the invasion biology of *Plasmodium vivax*. Front. Cell. Infect. Microbiol. 2013;3: 8.

70. Rozas J. DNA sequence polymorphism analysis using DnaSP. Methods Mol. Biol. Bioinformatics DNA Seq. Anal. 2009;537: 337–50.

71. Katoh K, Standley DM. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 2013;30: 772–80.

72. Kück P, Longo GC. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front. Zool. 2014;11: 81.

73. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 2006;22: 2688–90.

74. Jombart T. Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008;24: 1403–5.

75. Schneider ADB, Ford CT, Hostager R, Williams J, Cioce M, Çatalyürek ÜV, et al. StrainHub: a phylogenetic tool to construct pathogen transmission networks. Bioinformatics. 2019; btz646.

76. Lo E, Hostetler JB, Yewhalaw D, Pearson RD, Hamid MMA, Gunalan K, et al. Frequent expansion of *Plasmodium vivax* duffy binding protein in Ethiopia and its epidemiological significance. PLOS Negl. Trop. Dis. 2019;13: e0007222.

77. Oliveira TCD, Rodrigues PT, Menezes MJ, Gonçalves-Lopes RM, Bastos MS, Lima NF, et al. Genome-wide diversity and differentiation in new world populations of the human malaria parasite *Plasmodium vivax*. PLOS Negl. Trop. Dis. 2017;11: e0005824.

78. Neafsey DE, Galinsky K, Jiang RHY, Young L, Sykes SM, Saif S, et al. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. Nat Genet. 2012;44: 1046–50.

79. Dharia NV, Bright AT, Westenberger SJ, Barnes SW, Batalov S, Kuhen K, et al. Whole-genome sequencing and microarray analysis of *ex vivo Plasmodium vivax* reveal selective pressure on putative drug resistance genes. Proc. Natl. Acad. Sci. USA 2010;107: 20045–50.

80. Chen J-H, Chen S-B, Wang Y, Ju C, Zhang T, Xu B, et al. An immunomics approach for the analysis of natural antibody responses to *Plasmodium vivax* infection. Mol Biosyst. 2015;11: 2354–63.

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

82. Mascorro CN, Zhao K, Khuntirat B, Sattabongkot J, Yan G, Escalante AA, et al. Molecular evolution and intragenic recombination of the merozoite surface protein MSP-3α from the malaria parasite *Plasmodium vivax* in Thailand. Parasitology. 2005;131: 25–35.

83.Cunning D, Lawton J, Jarra W, Preiser P, Langhorne J. The pir multigene family of *Plasmodium*: antigenic variation and beyond. Mol. Biochem. Parasitol. 2010;170: 65-73.

84. Loy DE, Plenderleithc LJ, Sundararamana SA, Liua W, Gruszczyke J, ChenG YJ, Trimbolia S, Learna GH, MacLeanc OA, Morgan ALK, Lia Y, Avittoa AN, Gilesa J, Calvignac-Spencerg S, Sachseg A, Leendertzg FH, Speedeh S, Ayoubai A, Peetersi M, Rayner JC, Tham WH, Sharp PM, Hahna BH. Evolutionary history of human *Plasmodium vivax* revealed by genome-wide analyses of related ape parasites. Proc Natl Acad Sci USA 2018;115: e8450-8459.

85. Spence PJ, Jarra W, Lévy P, Reid AJ, Chappell L, Brugat T, et al. Vector transmission regulates immune control of *Plasmodium* virulence. Nature. 2013;498: 228–31.

86. Tachibana S-I, Sullivan SA, Kawai S, Nakamura S, Kim HR, Goto N, et al. *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. Nat Genet. 2012;44: 1051–5.

87. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature. 2008;455: 757–63.

88. Urusova D, Carias L, Huang Y, Nicolete VC, Popovici J, Roesch C, et al. Structural basis for neutralization of *Plasmodium vivax* by naturally acquired human antibodies that target DBP. Nat. Microbiol. 2019;4: 1486–96.

89. Kariu T, Yuda M, Yano K, Chinzei Y. MAEBL is essential for malarial sporozoite infection of the mosquito salivary gland. J. Exp. Med. 2002;195: 1317–23.

90. Preiser P, Renia L, Singh N, Balu B, Jarra W, Voza T, et al. Antibodies against MAEBL ligand domains M1 and M2 inhibit sporozoite development *in-vitro*. Infect. Immun. 2004;72: 3604–8.

91. Doolan DL, Dobano C, Baird JK. Acquired immunity to malaria. Clin. Microbiol. Rev. 2009;22: 13–36.

92. Waltmann A, Koepfli C, Tessier N, Karl S, Fola A, Darcy AW, et al. Increasingly inbred and fragmented populations of *Plasmodium vivax* associated with the eastward decline in malaria transmission across the Southwest Pacific. PLOS Negl. Trop. Dis. 2018;12: e0006146.

93. Eede PVD, D’Alessandro U, Erhart A, Thang ND, Anné J, Overmeir CV, et al. High complexity of *Plasmodium vivax* infections in symptomatic patients from a rural community in central Vietnam detected by microsatellite genotyping. Am. J. Trop. Med. Hyg. 2010;82: 223–7.

94. Liu Y, Auburn S, Cao J, Trimarsanto H, Zhou H, Gray K-A, et al. Genetic diversity and population structure of *Plasmodium vivax* in central China. Malar. J. 2014;13: 262.

95. Jennison C, Arnott A, Tessier N, Tavul L, Koepfli C, Felger I, et al. *Plasmodium vivax* populations are more genetically diverse and less structured than sympatric *Plasmodium falciparum* populations. PLOS Negl. Trop. Dis. 2015;9: e0003634.

96. Koepfli C, Timinao L, Antao T, Barry AE, Siba P, Mueller I, et al. A large *Plasmodium vivax* reservoir and little population structure in the South Pacific. PLoS ONE. 2013;8: e66041.

97. Lin E, Kiniboro B, Gray L, Dobbie S, Robinson L, Laumaea A, et al. Differential patterns of infection and disease with *P. falciparum* and *P. vivax* in young Papua New Guinean children. PLoS ONE. 2010;5: e9047.

98. Koepfli C, Ross A, Kiniboro B, Smith TA, Zimmerman PA, Siba P, et al. Multiplicity and diversity of *Plasmodium vivax* infections in a highly endemic region in Papua New Guinea. PLOS Negl. Trop. Dis. 2011;5: e1424.

99. Getachew S, To S, Trimarsanto H, Thriemer K, Clark TG, Petros B, et al. Variation in complexity of infection and transmission stability between neighboring populations of *Plasmodium vivax* in southern Ethiopia. Plos One. 2015;10: e0140780.

100. Friedrich LR, Popovici J, Kim S, Dysoley L, Zimmerman PA, Menard D, et al. Complexity of infection and genetic diversity in Cambodian *Plasmodium vivax*. PLOS Negl. Trop. Dis. 2016;10: e0004526.

101. Lo E, Lam N, Hemming-Schroeder E, Nguyen J, Zhou G, Lee M-C, et al. Frequent spread of *Plasmodium vivax* malaria maintains high genetic diversity at the Myanmar-China border, without distance and landscape barriers. J. Infect. Dis. 2017;216: 1254–63.

102. Abdelraheem MH, Bansal D, Idris MA, Mukhtar MM, Hamid MMA, Imam ZS, et al. Genetic diversity and transmissibility of imported *Plasmodium vivax* in Qatar and three countries of origin. Sci. Rep. 2018;8: 8870.

103. Popovici J, Friedrich LR, Kim S, Bin S, Run V, Lek D, et al. Genomic analyses reveal the common occurrence and complexity of *Plasmodium vivax* relapses in Cambodia. mBio. 2018;9: e01888-17.

## Tables

**Table 1**. Information of whole genome sequences of 44 *Plasmodium vivax* isolates from Ethiopia. The European Nucleotide Archive (ENA) accession number for all files.

**Table 2.** A shortlist of 25 erythrocyte binding genes that showed signals of positive selection based on the Likelihood Ratio Test of the M1 (neutral model) and M2 models (selection model) in PAML.

**Table 3.** Transmission network metrics among study sites calculated by StrainHub.

## Figures

**Figure 1.** An overview of the *P vivax* sample collection locations including Arbaminch, Badowacho, Hawassa, Halaba, and Jimma in southwestern Ethiopia.

**Figure 2.** A summary representation of the *P. vivax* genome, with the outer ring as an ideogram representing the 14 nuclear chromosomes and sizes of each. The second track represented the average coverage for each chromosome among the 44 Ethiopian samples. The third track containing the gray vertical dashes represented the distribution of genes across the 14 chromosomes. The forth track that contained the red vertical lines represented the 64 erythrocyte binding gene candidates. The fifth inner track with the light blue background represented the $d_N/d_S$ ratio calculated by partitioning the chromosomes into genomic regions and $d_N/d_S$ directly. The three outliers (yellow dots) represented three unknown plasmodium protein genes that were

detected with significant positive selection.  The sixth track indicated the overall copy number variation calculated using CNVnator.  Red dots represented genes with copy number variation among the Ethiopian genomes.

**Figure 3.** (A) A distribution of the nonsynonymous and synonymous mutations of each chromosome. A higher proportion of synonymous mutations was observed compared to nonsynonymous mutations. Chromosomes 7, 9, and 12 have the most mutations overall, with chromosomes 6 and 3 having the fewest number of mutations. (B) Number of mutation sites and the nucleotide diversity of 64 erythrocyte binding genes.  The *PvRBP* and *PvMSP* multigene families have the highest number of polymorphic sites when compared to the others, with *PvRBP*2c the highest number of nonsynonymous and synonymous mutations, followed by *PvMSP*3 and *PvMSP*1.  Approximately 40% of the mutations were nonsynonymous.  These genes were also indicated with the highest nucleotide diversity.

**Figure 4.** Signal of positive selection across the 14 chromosomes among all *P. vivax* samples.  Genes that showed significant signal of positive selection included *STP*1, *MAEBL*, *MSP*3.8, and *PIR* gene regions.  *PvMSP*3.8 gene may play a role in the erythrocyte invasion. *MAEBL* is a membrane associated erythrocyte binding like protein that may have a function associated with erythrocyte invasion.

**Figure 5.** A total of 28 gene regions that were detected with copy number variation. Annotation of these genes can be found in Supplementary Table 4. Among them,

*PvDBP*1 (PVP01_0623800) and *PvMSP*3 (PVP01_1030900) were associated with erythrocyte invasion. Other genes that were found to have high-order copy number were *PIR* protein genes or unknown exported plasmodium proteins.

**Figure 6**. An unrooted whole genome phylogenetic tree of the 44 Ethiopian samples showing the evolution of (A) *PvDBP*; (B) *PvMSP*3; (C) *PIR* gene on chromosome 2; and (D) *PIR* gene on chromosome 11. The Ethiopian isolates were divided into three subclades. Subclade I contained samples mostly from the Arbaminch and Badowacho. Subclade II contained a mixture of isolates from Arbaminch, Halaba, Hawassa and Jimma. Subclade III contained samples from Jimma. No distinct clusters were observed between isolates with single and multiple *PvDBP*, *PvMSP*3, and *PIR* genes. These patterns suggest that these gene regions could have expanded multiply among samples at different locations.

**Figure 7.** (A) Principal component analysis plot based on the SNP information from our variant analysis. Samples obtained from Jimma were clustered together, whereas samples from Arbaminch, Badowacho, Hawassa, and Halaba were mixed together with the exception of two samples from Hawassa. This clustering pattern suggested that there was considerable genetic variation among study sites even at a small geographical scale. (B) The transmission network, created using the StrainHub program, indicated that Arbaminch was the major source of infection in Jimma, Halaba, Badowacho and Hawassa. The greatest extent of gene flow (indicated by the boldest arrow) was observed between Arbaminch and Badowacho. Even though Jimma, Badowacho and Halaba are geographically in close proximity, gene flow was not intense among these sites.

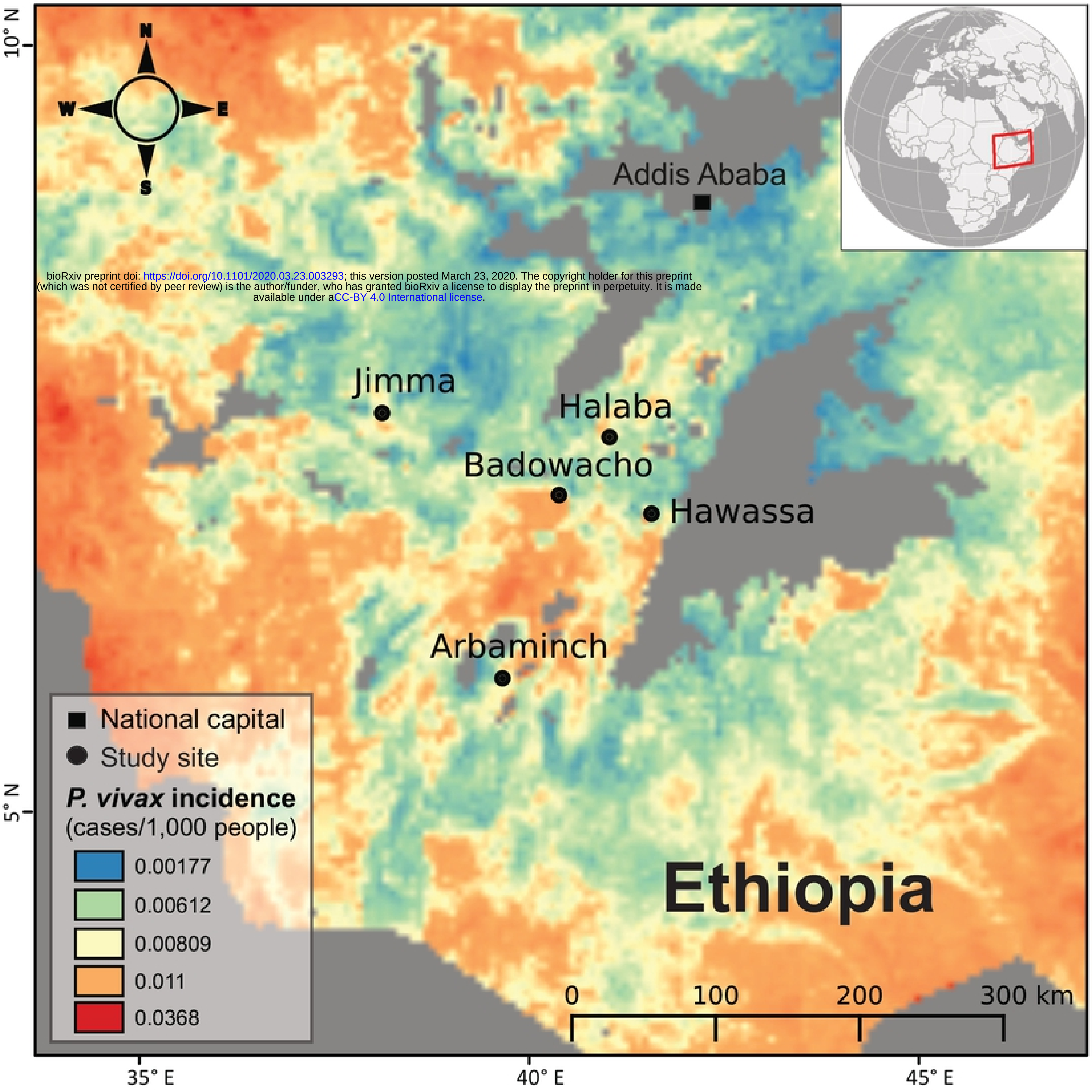
**Supplementary files**

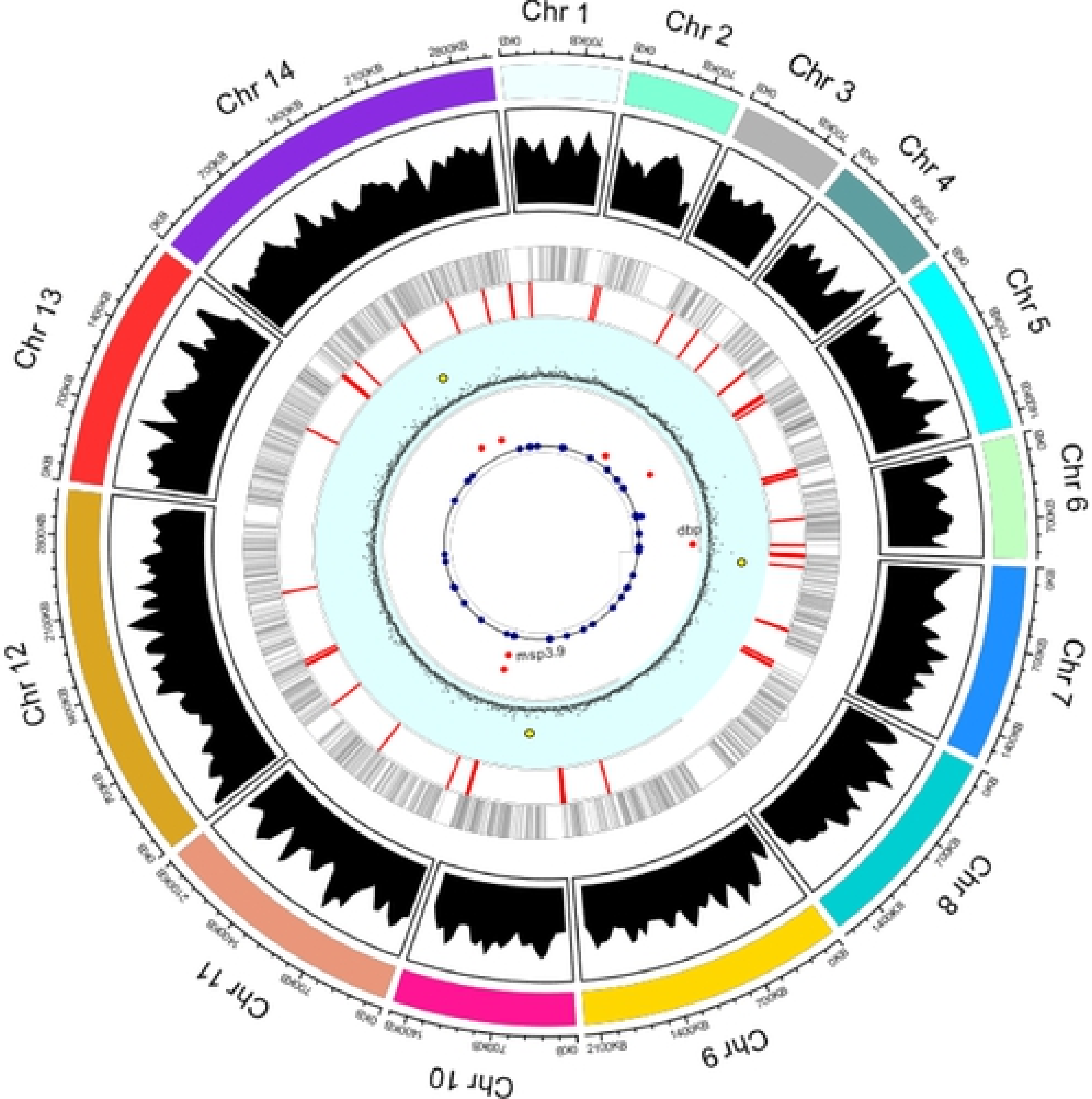
**Supplementary Table 1.** Distribution of SNP variants in the 64 *P. vivax* erythrocyte binding gene candidates among the 44 Ethiopian genomes.

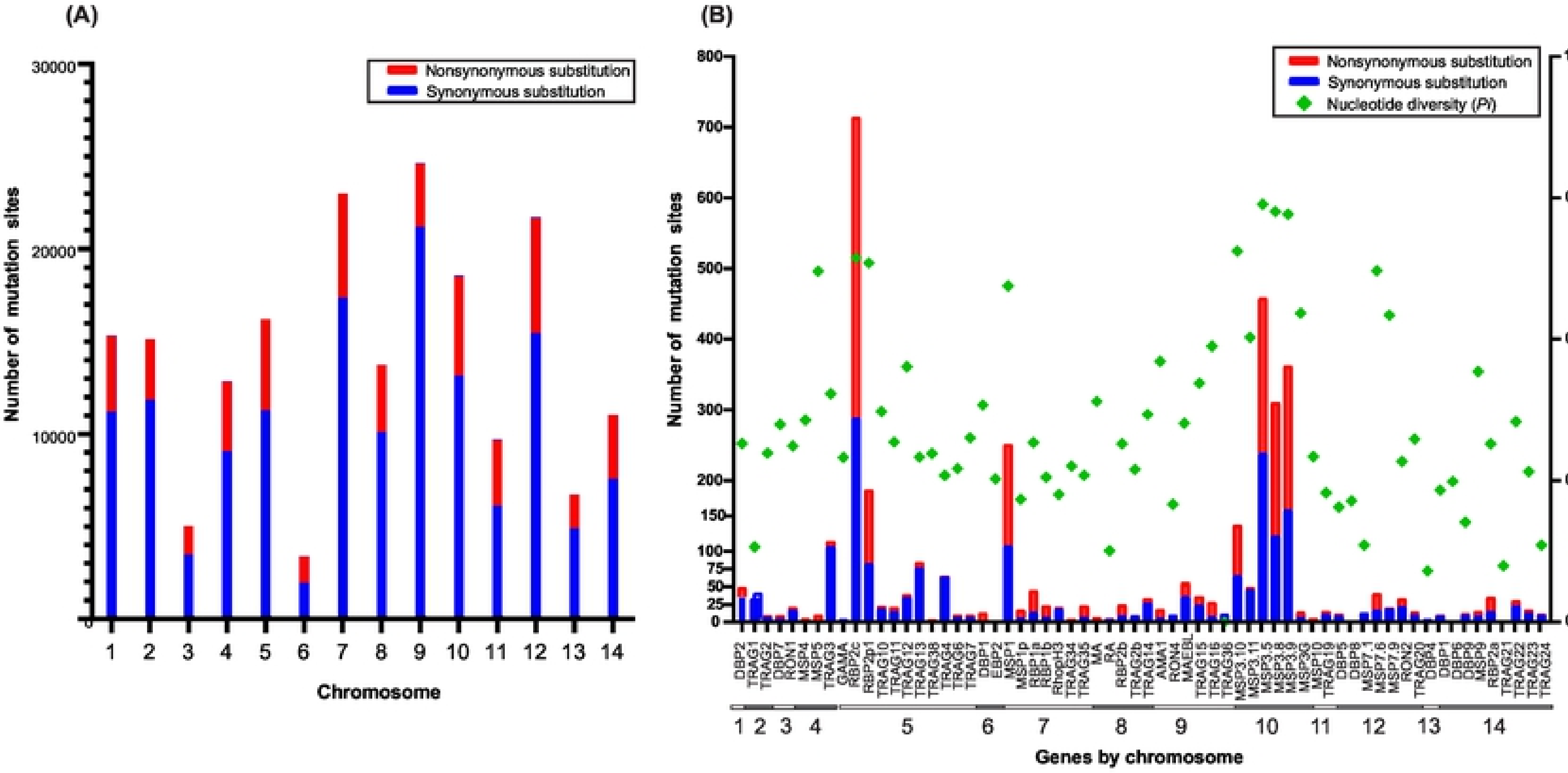
**Supplementary Table 2.** Distribution of single nucleotide polymorphism (SNP) variants across *P. vivax* chromosomes of the 44 Ethiopian genomes.

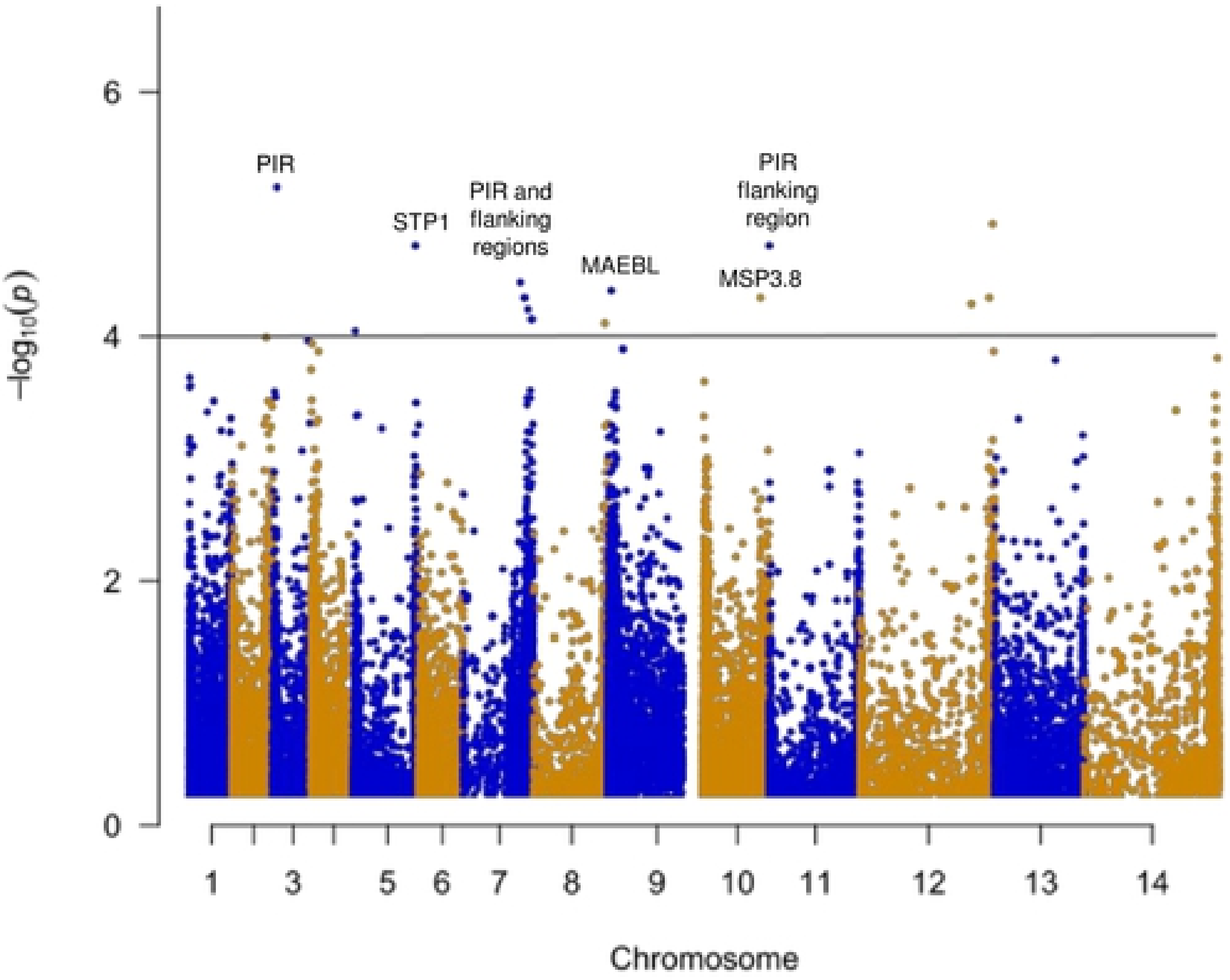
**Supplementary Table 3.** Likelihood Ratio Test results of the M1 (neutral model) and M2 models (selection model) in PAML of all the 64 erythrocyte binding gene candidates.

**Supplementary Table 4.** Gene regions that were detected with copy number variation among the 44 Ethiopian *P. vivax* isolates based on CNVnator. Among them, only two erythrocyte binding gene candidates *PvDBP*1 and *PvMSP*3 were detected with high-order copies.

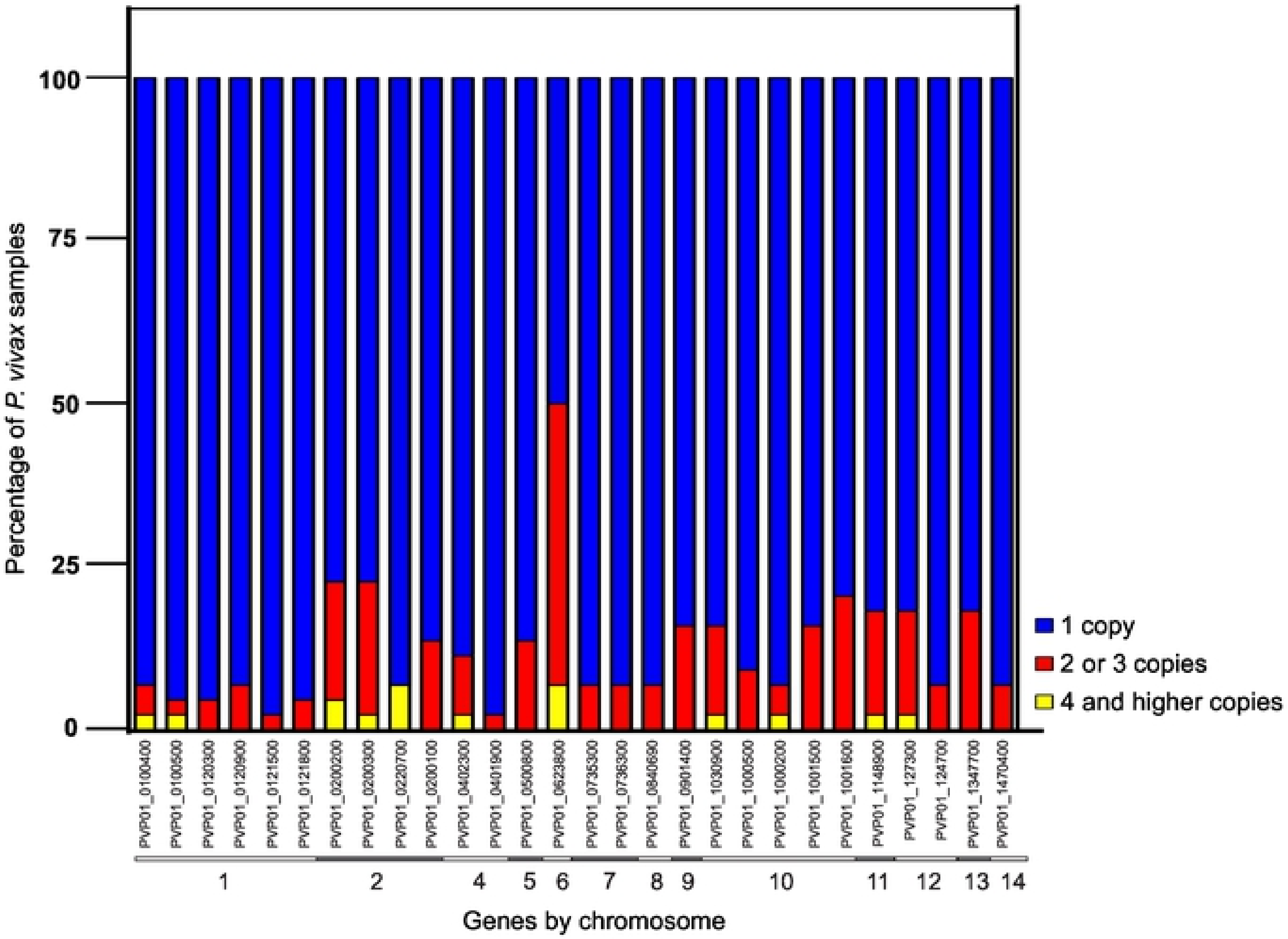

N
W
E
S
Addis Ababa
Jimma
Halaba
Badowacho
Hawassa
Arbaminch
Ethiopia
National capital
Study site
P. vivax incidence
(cases/1,000 people)
0.00177
0.00612
0.00809
0.011
0.0368
0
100
200
300 km
10° N
5° N
35° E
40° E
45° E

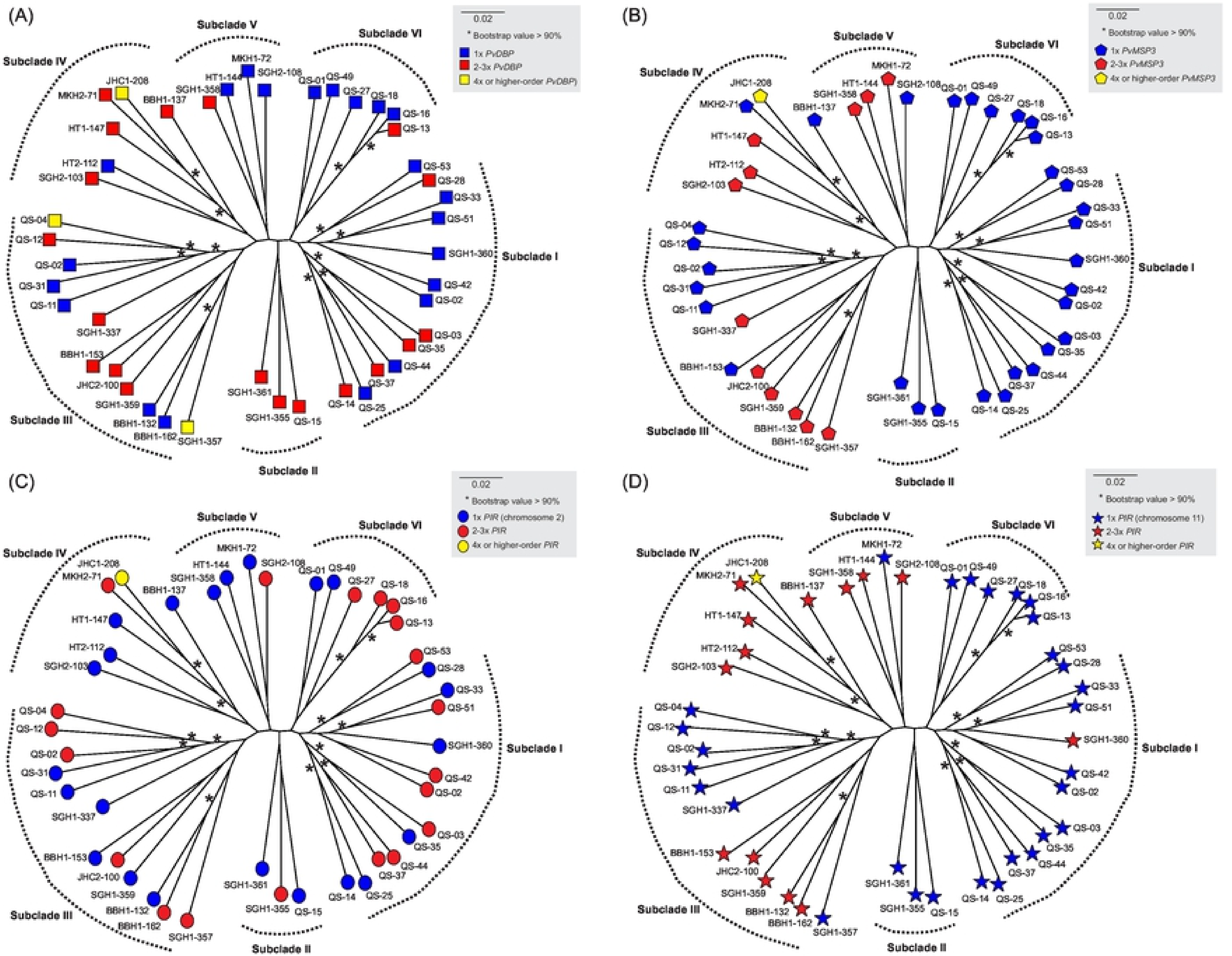

Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
dbp
msp3.9
0KB
700KB
1400KB
2100KB
2800KB

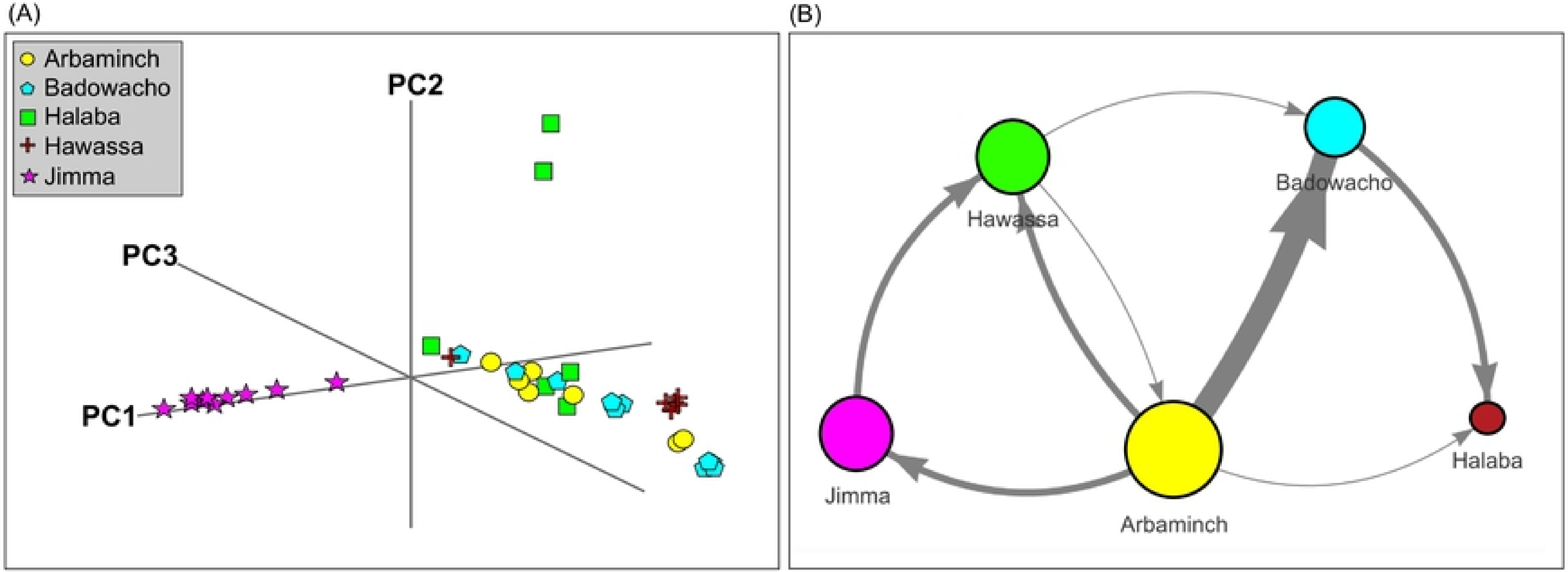

(A)
Nonsynonymous substitution
Synonymous substitution
Number of mutation sites
30000
20000
10000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Chromosome
(B)
Nonsynonymous substitution
Synonymous substitution
Nucleotide diversity (Pi)
Number of mutation sites
800
700
600
500
400
300
200
150
100
75
50
25
0
Nucleotide Diversity
1
0.1
0.01
0.001
0.0001
DBP2
TRAG1
TRAG2
DBP7
RON1
MSP4
MSP5
TRAG3
GAMA
RBP2c
RBP2p1
TRAG10
TRAG11
TRAG12
TRAG13
TRAG38
TRAG4
TRAG6
TRAG7
DBP1
EBP2
MSP1
MSP1p
RBP1a
RBP1b
RhopH3
TRAG34
TRAG35
MA
RA
RBP2b
TRAG2b
TRAG14
AMA1
RON4
MAEBL
TRAG15
TRAG16
TRAG36
MSP3.10
MSP3.11
MSP3.5
MSP3.8
MSP3.9
MSP3G
MSP10
TRAG19
DBP5
DBP8
MSP7.1
MSP7.6
MSP7.9
RON2
TRAG20
DBP4
DBP1
DBP6
DBP9
MSP9
RBP2a
TRAG21
TRAG22
TRAG23
TRAG24
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Genes by chromosome



PIR
STP1
PIR and flanking regions
MAEBL
PIR flanking region
MSP3.8
6
4
2
0
$-\log_{10}(p)$
1
3
5
6
7
8
9
10
11
12
13
14
Chromosome


Percentage of P. vivax samples
100
75
50
25
0
1 copy
2 or 3 copies
4 and higher copies
PVP01_0100400
PVP01_0100500
PVP01_0120300
PVP01_0120900
PVP01_0121500
PVP01_0121800
PVP01_0200200
PVP01_0200300
PVP01_0220700
PVP01_0200100
PVP01_0402300
PVP01_0401900
PVP01_0500800
PVP01_0623800
PVP01_0735300
PVP01_0736300
PVP01_0840690
PVP01_0901400
PVP01_1030900
PVP01_1000500
PVP01_1000200
PVP01_1001500
PVP01_1001600
PVP01_1148900
PVP01_127300
PVP01_124700
PVP01_1347700
PVP01_1470400
1
2
4
5
6
7
8
9
10
11
12
13
14
Genes by chromosome


(A)
0.02
* Bootstrap value > 90%
1x PvDBP
2-3x PvDBP
4x or higher-order PvDBP)
Subclade IV
Subclade V
Subclade VI
Subclade I
Subclade II
Subclade III
JHC1-208
MKH2-71
HT1-147
HT2-112
SGH2-103
BBH1-137
SGH1-358
HT1-144
MKH1-72
SGH2-108
QS-01
QS-49
QS-27
QS-18
QS-16
QS-13
QS-53
QS-28
QS-33
QS-51
SGH1-360
QS-42
QS-02
QS-03
QS-35
QS-44
QS-37
QS-25
QS-14
SGH1-361
SGH1-355
QS-15
QS-04
QS-12
QS-02
QS-31
QS-11
SGH1-337
BBH1-153
JHC2-100
SGH1-359
BBH1-132
BBH1-162
SGH1-357
(B)
0.02
* Bootstrap value > 90%
1x PvMSP3
2-3x PvMSP3
4x or higher-order PvMSP3
Subclade IV
Subclade V
Subclade VI
Subclade I
Subclade II
Subclade III
JHC1-208
MKH2-71
HT1-147
HT2-112
SGH2-103
BBH1-137
SGH1-358
HT1-144
MKH1-72
SGH2-108
QS-01
QS-49
QS-27
QS-18
QS-16
QS-13
QS-53
QS-28
QS-33
QS-51
SGH1-360
QS-42
QS-02
QS-03
QS-35
QS-44
QS-37
QS-25
QS-14
SGH1-361
SGH1-355
QS-15
QS-04
QS-12
QS-02
QS-31
QS-11
SGH1-337
BBH1-153
JHC2-100
SGH1-359
BBH1-132
BBH1-162
SGH1-357
(C)
0.02
* Bootstrap value > 90%
1x PIR (chromosome 2)
2-3x PIR
4x or higher-order PIR
Subclade IV
Subclade V
Subclade VI
Subclade I
Subclade II
Subclade III
JHC1-208
MKH2-71
HT1-147
HT2-112
SGH2-103
BBH1-137
SGH1-358
HT1-144
MKH1-72
SGH2-108
QS-01
QS-49
QS-27
QS-18
QS-16
QS-13
QS-53
QS-28
QS-33
QS-51
SGH1-360
QS-42
QS-02
QS-03
QS-35
QS-44
QS-37
QS-25
QS-14
SGH1-361
SGH1-355
QS-15
QS-04
QS-12
QS-02
QS-31
QS-11
SGH1-337
BBH1-153
JHC2-100
SGH1-359
BBH1-132
BBH1-162
SGH1-357
(D)
0.02
* Bootstrap value > 90%
1x PIR (chromosome 11)
2-3x PIR
4x or higher-order PIR
Subclade IV
Subclade V
Subclade VI
Subclade I
Subclade II
Subclade III
JHC1-208
MKH2-71
HT1-147
HT2-112
SGH2-103
BBH1-137
SGH1-358
HT1-144
MKH1-72
SGH2-108
QS-01
QS-49
QS-27
QS-18
QS-16
QS-13
QS-53
QS-28
QS-33
QS-51
SGH1-360
QS-42
QS-02
QS-03
QS-35
QS-44
QS-37
QS-25
QS-14
SGH1-361
SGH1-355
QS-15
QS-04
QS-12
QS-02
QS-31
QS-11
SGH1-337
BBH1-153
JHC2-100
SGH1-359
BBH1-132
BBH1-162
SGH1-357


(A)
Arbaminch
Badowacho
Halaba
Hawassa
Jimma
PC2
PC3
PC1
(B)
Badowacho
Hawassa
Jimma
Arbaminch
Halaba