Analysis of Genetic Variability of *Plasmodium vivax* Isolates from Different Brazilian Amazon Areas Using Tandem Repeats

Antônio M. Rezende, Eduardo Tarazona-Santos, Álvaro D’ Almeida Couto, Cór Jésus F. Fontes, José Maria De Souza, Luzia H. Carvalho, and Cristiana F. A. Brito*

**Laboratory of Malaria, Centro de Pesquisas René Rachou/Fiocruz, Belo Horizonte, Brazil**

**Abstract.** Few genetic markers have been described to analyze populations of *Plasmodium vivax*. The genetic variability of *P. vivax* has been analyzed mainly among isolates taken from areas ranging from hyper- to holoendemic areas. These studies of genetic variability have neglected many areas with different epidemiologic profiles. The purpose of this study was to analyze the genetic variability of *P. vivax* isolates from four different Brazilian Amazon areas. We chose to study the five most polymorphic tandem repeats (TRs) identified so far. All TRs studied were polymorphic in at least one studied population, with a modal allele at nearly all loci. Expected heterozygosity ranged from 0.462 to 0.666 and did not correlate with the repeat array length. The genetic distances among the populations varied from 0.027 to 0.241, and did not correlate with their geographic separation. Tandem repeats identified in *P. vivax* isolates failed to allow geographic clustering.

**INTRODUCTION**

Malaria is a public health problem in over 100 countries; it causes at least 300 million cases of acute illness, and it kills more than one million people every year. *Plasmodium vivax* is the most widespread malaria species, and up to 40% of the world population is at risk of infection from this species. Currently, malaria caused by *P. vivax* is endemic in many countries of Asia, the South Pacific, North Africa, the Middle East, and South and Central America. In Brazil, 540,000 new cases of malaria were reported in the year 2006. Nearly all of them (99.5%) occurred in the Amazon states, where *P. vivax* was responsible for more than 70% of cases. Despite a lower mortality than *Plasmodium falciparum* infection, *P. vivax* is responsible for intense morbidity and significant economic losses. Parasite control strategies depend on an understanding of the genetic variability and population structure of the parasite and the dynamics of disease transmission. Moreover, studies of genetic variability, particularly of field isolates collected directly from patients, can provide markers to predict drug resistance, virulence, and periodic relapse. Several studies have assessed the genetic variability, population structure, and evolution of *P. falciparum*. In contrast, the lower virulence of *P. vivax* and the difficulty of propagating it continuously in culture mean that less is known about its genetic variability. Furthermore, few studies have analyzed the genetic diversity among *P. vivax* populations, which could lead to inferences about its population dynamics and patterns of gene flow.

Molecular markers for population studies of *P. vivax* have mostly been based on orthologs of *P. falciparum* antigen genes, such as merozoite surface antigen, erythrocyte-binding proteins, and circumsporozoite protein, which are under selection by host immunity. However, some population parameters are better inferred from neutral or nearly neutral markers, including microsatellite loci. In contrast with the ~1,000 polymorphic microsatellites described for *P. falciparum*, only 40 or so have been described for *P. vivax*. Microsatellite variability in *P. vivax* is strongly dependent on the repeat array length: short arrays exhibit little variation and long arrays are highly polymorphic. When repeat array length is taken into account, *P. vivax* microsatellites show levels of variation comparable to those in *P. falciparum*. However, the geographic clustering of *P. vivax* isolates was not detected using microsatellite loci. This may be a result of high within-population heterogeneity or to the complex phylogeny of *P. vivax*. On the other hand, Feng and colleagues described 33 polymorphic tandem repeats (TRs) or minisatellites containing repeating units of five nucleotides or more. Leclerc and colleagues showed a limited variability among isolates from different areas using microsatellites and TR loci. Tandem repeat loci tend to be less variable than microsatellite loci and, for this reason, they can allow for geographic clustering of isolates. In addition, polymerase chain reaction (PCR)-amplified fragments of TRs may be separated using simple agarose gel electrophoresis and easily analyzed because of their significant variation in the number of repeats.

Thus, the aim of this study was to analyze the variability of *P. vivax* isolates from different endemic areas in the Amazon using the five most polymorphic TRs used by Leclerc and colleagues. Our specific objectives were to assess the genetic diversity, structure, and differentiation of *P. vivax* populations in relation to geographic distances between these populations and the level of malaria transmission in different areas of Brazilian Amazon. We report the results of the genetic analysis and discuss the evolutionary forces that may explain the observed patterns of diversity.

**MATERIALS AND METHODS**

*Plasmodium vivax* isolates. Blood samples were collected from 44 *P. vivax*-infected patients (11 individuals from each area) in different areas of the Brazilian Amazon (Figure 1): Macapá in Amapá State (AP; 00°02′20″S, 51°03′59″W), Cuiabá in Mato Grosso State (MT; 15°35′45″S, 56°05′49″W), Augusto Correa in Pará State (PA; 01°01′18″S, 46°38′06″W), and Manaus in Amazonas State (AM; 03°06′07″S, 60°01′30″W). The samples were collected from the different regions as follows: AM, May 2003; AP, November 2004; PA, October 2005; and MT, from July 2003 to July 2004. The available epidemiologic data indicated that the studied regions were areas of hypo- to mesoendemic malaria with intermittent transmission. The rate of transmission for each area was measured by the annual parasite index (API), which reflects the number of positive blood
smears per 1,000 inhabitants. By following the guidelines of the Health Surveillance Secretariat of the Ministry of Health, the API index was used to stratify the risk of malaria infection in the different areas at the time of blood collection: high risk (API > 50), Augusto Correa (PA); medium risk (API 10–50), Manaus (AM); and low risk (API < 10), Macapá (AP). Patients who came to the health center in Cuiabá (MT) reported many locations of infection with different API, so the area was classified as variable risk. Patient infection was confirmed by microscopic analysis on Giemsa-stained blood smears. All patients were between 16 and 58 years of age (mean, 32 years).

**Genomic DNA isolation and PCR amplification.** The DNA was extracted from whole blood samples using the Puregene DNA isolation Kit (Gentra Systems, Minneapolis, MN), according to the manufacturers protocol. The five loci (MN2, MN7, MN21, MN23, and MN25) were amplified using the primers and conditions described by Leclerc and colleagues, but forward primers were labeled with fluorescent dye (FAM). The repeat unit ranged from 4 (MN23) to 20 nucleotides (MN25) (Table 1). These markers were assigned to chromosome 8 of *P. vivax* Sal-1 (unfinished sequence, whole genome shotgun sequencing project. GenBank accession no. AAKM00000000). The markers were separated from each other at a distance of 1,727 bp (MN21/MN23) to 32,934 bp (MN2/MN7) (Table 1). The PCR products were separated using agarose gel electrophoresis and the exact size of PCR products was identified using the automated laser fluorescent (ALF) automatic DNA sequencer (Amersham Pharmacia Biotech, Uppsala, Sweden), comparing to the ALFexpress sizer 50–500 bp (5 fmol/μL; Amersham), and analyzed using Allele Link software (version 1.0, Amersham).

**Variability analysis.** We measured allele frequencies using the single or predominant allele at each locus from each isolate; non-predominant alleles were recorded and used to estimate multiplicity of infection, because all markers are single-copy loci and blood stage malaria parasites are haploid. The highest peak in the electropherogram traces was defined as the predominant allele and electropherogram traces with at least one-third the height of the predominant allele were defined as additional alleles in a multiple-clone infection. We measured intra-population diversity using the statistic \( H_e \), which is the average probability that two alleles randomly obtained for each locus are different. An analysis of molecular variance (AMOVA) was performed to measure whether the resulting variability could be explained by variations among and within populations. We performed a Kruskal–Wallis test to verify whether the differences between the \( H_e \) of each locus for individual populations were significant.

We assessed the linkage disequilibrium (LD) for each locus pair and each population using the Fisher exact test of LD available in Arlequin 3.1 software with 100,000 steps in the Markov Chain and 1,000 dememorization steps. We also used LIAN 3.5 software to test for evidence of multilocus LD in each population. This test compares the variance \( V_{\text{obs}} \) of the number \( D \) of alleles shared between all pairs of haplotypes observed in the population \( (D) \) with the variance expected when alleles are assigned randomly \( V_e \) as follows: \( F_A = (V_{\text{obs}}/V_e - 1) \), where \( r \) is the number of loci analyzed. The \( V_{\text{obs}} \) is derived from 100,000 simulated data sets in which alleles were randomly reshuffled among haplotypes. Significant LD is detected if \( V_{\text{obs}} \) is 95% of the values derived from the reshuffled data sets. Two analyses were done: one considering all haplotypes and another with unique haplotypes to remove the effect of the epidemic expansion of a particular haplotype on LD in randomly mating populations.

Between-population genetic diversity was quantified using pairwise \( F_{\text{ST}} \), as implemented in Arlequin 3.1. We also estimated the correlation among the matrices of pairwise genetic and geographic distances using the Mantel test.

We attempted to assess the population structure of samples using Structure 2.1 software. This uses Bayesian approaches.

---

**Table 1** Description of tandem repeats (TRs) used to analyze variability in *P. vivax* isolates

<table>
<thead>
<tr>
<th>Name</th>
<th>Repeat unit</th>
<th>First position</th>
<th>Gene ID†</th>
<th>Description of coding gene or flanking gene (Genbank accession no.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN2</td>
<td>(CCGTTGATGAGAATCAT)</td>
<td>1625109</td>
<td>Coding</td>
<td>A protein containing a kinase domain (5472301)</td>
</tr>
<tr>
<td>MN7</td>
<td>(CATTITG)</td>
<td>1591966</td>
<td>Non-coding</td>
<td>1683 bp at 5′ side: hypothetical protein (5472308); 1224 bp at 3′ side: hypothetical protein (5472307)</td>
</tr>
<tr>
<td>MN21</td>
<td>(CCACT)</td>
<td>1567132</td>
<td>Non-coding</td>
<td>358 bp at 5′ side: hypothetical protein (5472314); 739 bp at 3′ side: dihydrodipolamide acyltransferase, putative (5472313)</td>
</tr>
<tr>
<td>MN23</td>
<td>(CACCC)</td>
<td>1556219</td>
<td>Coding</td>
<td>1021 bp at 5′ side: hypothetical protein (5472317); 2408 bp at 3′ side: membrane skeletal protein, putative (5472316)</td>
</tr>
<tr>
<td>MN25</td>
<td>(AAGTCAACATGGCAAAAAA)</td>
<td>1556283</td>
<td>Non-coding</td>
<td>Hypothetical protein (5472314)</td>
</tr>
</tbody>
</table>

*First position at chromosome 8 (GenBank accession no. NC 009913).† Top hit obtained in a BlastN search of the *P. vivax* Sal-1 database at NCBI genomes (http://blast.ncbi.nlm.nih.gov/).
Five TR loci were selected to analyze 44 isolates of Plasmodium vivax taken from four different endemic areas in the Brazilian Amazon; their characteristics are shown in Table 1.

The K values ranged from 2 to 8. For each K value, we ran the software three times, each one with 100,000 burn-ins and 1,000,000 iterations. The admixture model and linkage model that assumed dependent alleles allowed the individual parasites to represent the diversity within each population.

RESULTS

Five TR loci were selected to analyze 44 isolates of Plasmodium vivax from four different endemic areas in the Brazilian Amazon; their characteristics are shown in Table 1. The TR markers used in this study map to chromosome 8, and they include non-coding loci (MN7, MN21, and MN25), a hypothetical protein locus (MN23), and a protein containing a kinase domain locus (MN2). The number of alleles per locus ranged from 2 to 9, with a modal allele in most cases (frequency ≥ 0.5) (Table 2). For all the TR loci studied, isolates were polymorphic in at least one population, and only MN2 was monomorphic in isolates from the PA population. We used a Kruskal–Wallis test to verify whether the difference between the heterozygosity expected \( (H_e) \) for each locus was significant (Table 3). We noted no difference between the loci, indicating that the length of the repeating unit of each locus did not correlate with its variability. Multiplicity of infection was estimated at 27% and 66% in the MT and AM populations, respectively. However, we did not find any individual infected with more than one haplotype in the PA and AP populations.

The comparative analysis of the four Amazon populations studied showed that all of them were highly variable, with \( H_e \) ranging from 0.46182 to 0.66600. In addition, when AMOVA was performed, we found that approximately 84% of the diversity detected was a result of variation within each population. Only 16% of the diversity could be explained by variation among populations.

The \( F_{ST} \) values between populations (Table 3) were between 0.02 and 0.24, and they did not correlate with geographic distance between populations (Mantel Test, \( Z = 0.031, P = 0.295 \)). The AM and AP populations could not be differentiated, and all of the populations diverged significantly (\( P ≤ 0.05 \)) from the PA population. This \( F_{ST} \) distribution and the derived estimations for the number of effective migrants (\( M = Nm \)) suggest that gene flow is higher among the AM and AP populations.

The LD was estimated using the Fisher test. The number of steps in the Markov Chain used in the analysis allowed that the results converged. We found LD to vary as a function of population and loci pairs, but these results should be interpreted with caution given our small sample sizes. Analysis of LD for the loci pairs in all isolates showed that six of 10 possible pair combinations showed high LD and did not correlate with the physical distance between the marker loci (Figure 2). Presence of marker loci in LD varied among populations. In the AP population, the loci pairs MN7/MN21, MN2/MN23, and MN21/MN25 showed high LD; in the MT population, the loci pairs MN7/MN23 and MN2/MN27 showed high LD; in the AM population, only the loci pair MN7/MN25 showed LD. Linkage disequilibrium was not found in the PA population. When LD analysis was performed with AP and AM as one population, because they both had low \( F_{ST} \) values, we noted only the locus pair MN7/MN21 in LD. The AP and AM populations showed significantly higher multilocus LD. However, when the analysis was carried out only on the unique haplotype infections, only the AP population showed significant LD (Table 4).

### Table 2

| Alleles detected at five tandem repeats (TRs) among Plasmodium vivax isolates from different endemic areas in Brazil |
|-----------------|-----|-----|-----|-----|-----|
| Locus*          | Allele† | Number of alleles (frequency) |
| MN2             | 133  | 2 (22.2) | 1 (9.1) | 2 (25) |
|                 | 152  | 6 (77.8) | 10 (90.9) | 3 (37.5) |
|                 | 171  | 1 (12.5) |
|                 | 189  | 2 (25) |
| MN7             | 209  | 6 (60) | 9 (81.8) | 7 (70) |
|                 | 220  | 6 (54.7) | 2 (18.2) | 3 (30) |
| MN21            | 254  | 2 (18.2) | 6 (54.5) | 5 (55.7) |
|                 | 258  | 3 (25.0) | 2 (18.2) | 1 (17.1) |
|                 | 266  | 1 (8.4) | 2 (14.5) |
|                 | 270  | 1 (9.1) |
|                 | 278  | 3 (21.4) |
|                 | 290  | 1 (8.4) | 3 (27.2) | 3 (21.4) |
| MN23            | 219  | 1 (8.4) | 5 (45.5) | 8 (53.3) |
|                 | 223  | 4 (33.3) | 3 (27.2) |
|                 | 237  | 3 (27.2) | 1 (6.7) |
|                 | 240  | 1 (8.4) | 6 (40.0) |
|                 | 247  | 5 (41.6) |
|                 | 251  | 1 (8.4) |
|                 | 278  | 5 (45.5) |
|                 | 282  | 3 (27.2) |
|                 | 302  | 1 (9.1) |
| MN25            | 169  | 8 (72.7) | 1 (9.1) | 1 (11.1) |
|                 | 191  | 1 (9.1) |
|                 | 212  | 1 (11.1) |
|                 | 220  | 1 (9.1) |
|                 | 234  | 3 (27.3) | 2 (18.2) | 4 (44.4) |
|                 | 241  | 1 (9.1) | 2 (18.2) |
|                 | 256  | 1 (9.1) | 6 (54.5) | 4 (36.4) | 1 (11.1) |
|                 | 261  | 2 (22.2) |
|                 | 276  | 3 (27.2) |

*Locus named according to Leclerc and colleagues.†Mean size of amplicons determined using Allele Link software (version 1.0, Amersham).

### Table 3

| Expected heterozygosity (\( H_e \)) for the tandem repeat (TR) loci and \( F_{ST} \) among Plasmodium vivax isolates from different endemic areas in Brazil |
|-----------------|-----|-----|-----|-----|-----|-----|-----|
| Locus           | Allele | PA | MT | AP | AM | Mean in population | \( F_{ST} \) (X100) |
| MN2             | 0     | 0.51 | 0.56 | 0.74 | 0.49 | 0.46 (0.248) |
| MN7             | 0.47 | 0.54 | 0.56 | 0.71 | 0.69 | 0.59 (0.092) | 20.355‡ |
| MN21            | 0.18 | 0.33 | 0.65 | 0.71 | 0.82 | 0.54 (0.242) | 19.697‡ | 9.373 |
| MN23            | 0.82 | 0.47 | 0.84 | 0.51 | 0.69 | 0.67 (0.153) | 24.160‡ | 16.175 |
| Mean at locus   | 0.37 (0.310) | 0.46 (0.083) | 0.65 (0.111) | 0.67 (0.092) | 0.67 (0.117) |

*\( H_e \) expected heterozygosity and \( F_{ST} \) were calculated using Arlequin 3.1 software.†Geographic regions: PA = Para State; MT = Mato Grosso State; AP = Amapá State; AM = Amazonas State.

‡Statistically significant (\( P < 0.005 \)).
It was not possible to cluster the samples using Structure 2.1 (data not shown). This was probably because of the limited number of markers used and the LD between the markers. Consequently, little independent information was available for the software to use for clustering.

**DISCUSSION**

All molecular markers that were previously reported to be polymorphic in populations from different countries (Azerbaijan, Ethiopia, Turkey, Venezuela, and Thailand) have also turned out to be polymorphic in Brazilian Amazonic populations. As reported by Leclerc and colleagues, we observed a predominant allele at almost all loci, yet at only one locus was the same allele prevalent in all geographic areas studied. We detected in our isolates 68% of the alleles previously described.

Our data showed a high variability within populations, with high values of \( H_{\text{e}} \). Similar values of \( H_{\text{e}} \) were also reported using microsatellite analysis of \( P. falciparum \) isolates from different regions of Brazil. However, a study by Ferreira and colleagues analyzing a multilocus microsatellite in \( P. vivax \) isolates from one Brazilian population in Acre State reported slightly higher \( H_{\text{e}} \) values than ours. This population showed a high percentage of multiple-clone infections, whereas only one population in our study showed a high multiplicity of infection: AM population, 66%. All of our populations showed a similar level of within-population variability. The population studies here were found to be structured with minimal gene flow, based on \( F_{\text{ST}} \) values. Our results corroborate the findings of multilocus microsatellite analysis in \( P. falciparum \) populations in Brazil.

Moreover, our study and that of Machado and colleagues did not show a correlation between genetic and geographic distances for populations of \( P. vivax \) or \( P. falciparum \), respectively.

Despite the fact that all the molecular markers map to the same chromosome, we did not observe a correlation between physical distance and LD. For instance, the loci MN21 and MN23 lay less than 2,000 bp away from each other, but they did not show significant LD. When we performed multilocus LD analysis, we observed a correlation between the \( F_{\text{ST}} \) value and number of loci pairs in LD in each population. If we excluded the MT population, the correlation was \( r = 0.9425 \). The MT population was excluded because their isolates were obtained from individuals infected in many localities. Moreover, the blood samples in this group were collected over the longest period (one year). These two factors may increase the variability in these populations. Interestingly, we did not find a negative correlation between multiplicity of infection and LD in populations, contrary to what has been observed with \( P. falciparum \). This result is similar to that of Ferreira and others, and it may indicate that these markers are under some selection pressure, which would alter the pattern of LD. Nevertheless, according to the Ewens–Watterson neutrality test, the loci used in this work do not suffer from selection pressure. Indeed, none of the TR markers used here map to known surface antigens likely to be under immune-mediated diversifying natural selection. A higher variability was observed at loci MN7, MN21, and MN25, which lie in non-coding regions (Table 1). However, these molecular markers may not be considered to be strictly neutral, because of their physical proximity and LD with genes encoding proteins under natural selective pressure.

In \( P. falciparum \), the genetic structure of the population is strongly linked to patterns of transmission. A widely accepted model is that high levels of single infection, inbreeding, and rare recombination characterize low transmission regions with low genetic variability. In contrast, frequent mosquito inoculation, multiple infections, frequent outbreeding, and extensive recombination characterize high transmission regions with high genetic variability. A similar result was obtained in \( P. falciparum \) from the Amazon region, which showed a negative correlation between LD and multiplicity of infection. This correlation between genetic variability and patterns of transmission was less clear in \( P. vivax \) infections. For South America (Colombia), Laos, and Thailand, the isolates fit the model well, as demonstrated by Anderson and others. However, Indian isolates were found to have a low LD and low frequency of multiple-clone infections. Conversely, Ferreira and colleagues showed a high LD and high frequency of multiple-clone infection in their study of Brazilian isolates from Acre State. Here, we observed high LD in the absence of multiple infections only in one of our populations.

According to Inwong and colleagues and Russell and colleagues, the repeat array length of microsatellites is directly associated with genetic variability. The fact that we did not observe this correlation for TRs suggests that the relationship is too weak to be detected by our small number of TRs, or it may not exist for these particular markers. These molecular markers were shown to be highly polymorphic, though slightly less so than the microsatellites that have already been described. In addition, this approach allows easy analysis based only on agarose gel electrophoresis, which suggest that high-throughput processing is possible using an automatic DNA sequencer.

**Table 4**

<table>
<thead>
<tr>
<th>Infection type</th>
<th>Amapa</th>
<th>Amazonas</th>
<th>Mato Grosso</th>
<th>Para</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>All infections</td>
<td>0.2360*</td>
<td>0.2038†</td>
<td>0.0626</td>
<td>0.0358</td>
<td>0.0278</td>
</tr>
<tr>
<td>Infection with unique haplotype</td>
<td>0.1870* (N = 9)</td>
<td>0.1122 (N = 6)</td>
<td>-0.0007 (N = 9)</td>
<td>-0.0159 (N = 9)</td>
<td>0.0088 (N = 33)</td>
</tr>
</tbody>
</table>

* \( P < 0.01 \)
† \( P < 0.05 \)
However, there is no evidence for a putative mutation mechanism to explain the variability at these markers, and this makes it difficult to understand how TR diversity is maintained. Therefore, further research on this issue is necessary.

REFERENCES


