



Genomic epidemiology of *Plasmodium vivax* malaria in Latin America and the Caribbean

Citation

Manrique-Valverde, Paulo Cesar. 2026. Genomic epidemiology of *Plasmodium vivax* malaria in Latin America and the Caribbean. Doctoral Dissertation, Harvard University Graduate School of Arts and Sciences.

Link

<https://dash.harvard.edu/handle/1/42740305>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.

Please share how this access benefits you. [Submit a story](#)

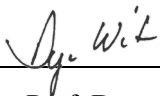
HARVARD
Kenneth C. Griffin



GRADUATE SCHOOL
OF ARTS AND SCIENCES

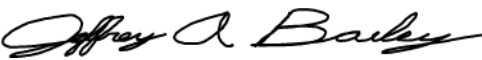
THESIS ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Committee on Higher Degrees in Biological Sciences in Public Health
have examined a dissertation entitled
Genomic epidemiology of *Plasmodium vivax* malaria
in Latin America and the Caribbean
presented by Paulo Cesar Manrique-Valverde
candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature 
Typed name: Prof. Dyann Wirth

Signature 
Typed name: Prof. Manoj Duraisingh

Signature 
Typed name: Prof. Phyllis Kanki

Signature 
Typed name: Prof. Jeffrey Bailey

Date: May 4, 2026

Genomic epidemiology of *Plasmodium vivax* malaria in Latin America and the Caribbean

A Dissertation presented

by

Paulo Cesar Manrique-Valverde

to

The Committee on Higher Degrees in Biological Sciences in Public Health

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological Sciences in Public Health

Harvard University

Cambridge, Massachusetts

May 2026

© 2026 Paulo Cesar Manrique-Valverde

All rights reserved.

Genomic epidemiology of *Plasmodium vivax* malaria in Latin America and the Caribbean**Abstract**

Plasmodium vivax is the main cause of malaria outside of Africa. Each year, between 6 and 7 million cases are reported, and in the Americas, this parasite represents over 70% of all malaria cases. Despite efforts to reduce malaria, *P. vivax* persists. Progress in reducing cases across the region has been uneven. Between 2010 and 2018, Panama, Honduras, and Guatemala succeeded in reducing, or at least containing, the number of cases per year within their territories. However, starting in 2019, this progress stalled. By 2023, Panama was reporting more than 9,000 cases, and beginning in 2021, Honduras and Guatemala began to follow the same trend. This increase coincides with the intensification of the Venezuelan exodus and other demographic displacements in the region. In this way, genetic relatedness from parasite populations in LAC could allow us to test if the increment on cases is due to the introduction of new parasite populations or the expansion of local populations that are resilient to control and elimination strategies.

In Chapter One, we review *P. vivax* transmission in Latin America and the Caribbean over the past 15 years. We also discuss the close relationship between transmission intensity and the genetic relatedness among parasite populations, and how these relationships can help classify infections as either imported or locally acquired. In Chapter Two, I introduce two new Illumina-based multiplexed PCR amplicon sequencing panels for *P. vivax*: PvGTSeq and PvCRiSP. The PvGTSeq panel incorporates 249 amplicons, covering candidate genes associated with drug resistance as well as amplicons capturing polymorphic regions particularly useful for finescale spatial resolution of parasite populations in Latin America. PvCRiSP consists of only four

highly-polymorphic and ultra-sensitive amplicons to estimate complexity of infection (COI) and identify instances of clonal transmission, without a costly and effort-intensive pre-amplification step. Both panels reliably generate genotypic data for patient samples with parasitemia levels as low as five parasites/ μ L and do not use proprietary reagents. These panels are suited for use in diverse geographic regions, with particular utility for studying *P. vivax* in the Americas—a region that is facing significant challenges in malaria control and elimination. At the time of submitting this thesis, the results presented in Chapter Two had been accepted for publication in the scientific journal PLOS Neglected Tropical Diseases, and the work will be published under the Creative Commons Attribution (CC BY) 4.0 license.

In Chapter Three, by using PvGTSeq and whole genome sequencing data I evaluate whether the increase in *P. vivax* cases in Panama, Honduras and Guatemala from 2020 onwards is due to the resilience of local parasite populations or the introduction of foreign parasite populations. Our results suggest that despite the presence of imported and introduced cases in Guatemala and Panama, the *P. vivax* parasite population remains largely unchanged, with the increase in cases primarily attributed to clonal lineages that have proven resilient to existing control measures. Finally, in Chapter Four I discuss the challenges posed by the epidemiological characteristics of Latin America to generating large-scale genomic data cost-effectively and interpreting it correctly. These findings are crucial for guiding local authorities in managing migrant populations, who often become vulnerable to local parasite strains.

Table of Contents

<i>Title page</i>	<i>i</i>
<i>Copyright</i>	<i>ii</i>
<i>Abstract</i>	<i>iii</i>
<i>Table of Contents</i>	<i>v</i>
<i>List of Figures and Tables</i>	<i>vii</i>
<i>Acknowledgments</i>	<i>x</i>
<i>Chapter One: Introduction</i>	<i>1</i>
1.1 <i>Plasmodium vivax in Latin America</i>	2
1.2 <i>Biological cycle of malaria and of P. vivax</i>	6
1.3 <i>Transmission characteristics of P. vivax</i>	8
1.4 <i>Transmission and parasite population diversity</i>	10
1.5 <i>References</i>	16
<i>Chapter Two: PvGTSeq and PvCRiSP: two amplicon-based targeted sequencing panels for Plasmodium vivax</i>	20
2.1 <i>Abstract</i>	21
2.2 <i>Introduction</i>	22
2.3 <i>Methods</i>	24
2.3.1 <i>Clinical samples and genomic databases</i>	24
2.3.2 <i>Parasite DNA quantification</i>	25
2.3.3 <i>Selection of amplicons and primer design</i>	26
2.3.4 <i>Whole genome sequencing and variant call</i>	28
2.3.5 <i>Panel protocols and optimization</i>	29
2.3.6 <i>Amplicon data analysis</i>	31
2.3.7 <i>Genotyping performance: amplification rate and limit of detection (LOD)</i>	32
2.3.8 <i>Consistency across assays</i>	33
2.3.9 <i>Detection of major and minor clones in polyclonal infections</i>	34
2.3.10 <i>Diversity and utility of the amplicon panels</i>	35
2.3.11 <i>Population Structure and Geographic attribution</i>	37

2.4	<i>Results.....</i>	38
2.4.1	<i>Selection of amplicons and protocol optimization</i>	38
2.4.2	<i>Genotyping performance.....</i>	41
2.4.3	<i>Consistency across sample replicates</i>	43
2.4.4	<i>Detection of major and minor clones in polyclonal infections.....</i>	45
2.4.5	<i>Diversity and utility of the amplicon panels</i>	47
2.4.6	<i>Population structure and geographic attribution.....</i>	51
2.5	<i>Discussion.....</i>	54
2.6	<i>References</i>	59
<i>Chapter Three: Resolving the recent demographic history of Plasmodium vivax populations in Latin America and the Caribbean</i>		67
3.1	<i>Abstract.....</i>	68
3.2	<i>Introduction.....</i>	69
3.3	<i>Methods</i>	71
3.3.1	<i>Sample collection.....</i>	71
3.3.2	<i>Sample processing and sequencing</i>	74
3.3.3	<i>Sequencing data analysis</i>	76
3.3.4	<i>Polyclonal infections and haplotypic diversity</i>	79
3.3.5	<i>Genetic relationships among infections across LAC, clonal propagation and population subdivision</i>	79
3.3.6	<i>Contribution of imported cases to transmission in Mesoamerica</i>	82
3.4	<i>Results.....</i>	83
3.4.1	<i>Haplotypic diversity of the study areas</i>	83
3.4.2	<i>Population structure patterns in LAC</i>	85
3.4.3	<i>Spatial and temporal distribution of clonal and ecological groups.....</i>	90
3.4.4	<i>Contribution of imported cases to transmission in Mesoamerica</i>	95
3.5	<i>Discussion.....</i>	96
3.6	<i>References</i>	102
<i>Chapter Four: Discussion</i>		105
4.1	<i>References</i>	112

List of Figures and Tables

Chapter One

Figure 1.1: <i>P. vivax</i> cases 2022 with respect to 2015.	2
Figure 1.2: Trend in <i>P. vivax</i> cases and people crossing the Darien Gap between 2010 to 2023.	3
Figure 1.3: Route of migrants through the Darien Gap, and number of imported cases in Panama, Honduras, and Guatemala from 2010 to 2023.	4

Chapter Two

Table 2.1: Nucleotides and amino acids in antimalarial resistance candidate genes covered by PvGTSeq.	40
Figure 2.1: Limit of detection of PvGTSeq and PvCRiSP in Colombia, Guyana and Venezuela.	43
Figure 2.2: Detectability of alleles from major and minor clones in mock samples using PvGTSeq and PvCRiSP.	46
Figure 2.3: Number of microhaplotypes (MHaps) and nucleotide diversity (π) per amplicon use groups and across seven global regions	49
Figure 2.4: Panel discrimination power.	51
Figure 2.5: Population subdivision with PvGTSeq.	53

Chapter Three

Table 3.1: Percentage of polyclonal infections in LAC	83
Table 3.2: Haplotypic and genetic diversity of LAC countries using PvGTSeq	85
Figure 3.2: Identification of clonal and low-diversity groups using PvGTSeq genetic relatedness (IBD) in a network analysis.	87
Figure 3.2: Identification of ecological groups using PvGTSeq.	89
Figure 3.4: Geographic distribution of clonal and ecological groups in Panama.	91
Figure 3.5: Temporal distribution of clonal and ecological groups in Mesoamerica.	93
Table 3.3: Consistency of malaria case classification between genetic information and travel history	95

APENDIX

Supplementary figure 2.1: Flowchart of the selection, optimization and validation process for PvGTSeq and PvCRiSP.	114
Supplementary figure 2.2: Examples of Pseudo-cigar representation of polymorphisms observed from multiple alignment of amplicon microhaplotypes with their reference sequence.	115

Supplementary figure 2.3: Principal coordinate analysis of world regions comparing whole genome sequencing (A), PvGTSeq (B), PvAmpliSeq (C) and rhAmpSeq (D).	116
Supplementary figure 2.4: Principal coordinate analysis of Latin America and Caribbean <i>P. vivax</i> populations	117
Supplementary figure 2.5: Deviation of IBD estimates from the amplicon panels relative to WGS.	118
Supplementary figure 2.6: Bar plot showing the prevalence of haplotypes of genes that carry mutations associated with antimalarial resistance.	119
Supplementary figure 2.7: Performance of PvGTSeq and PvCRiSP amplicons.	120
Supplementary figure 2.8: Contribution of each amplicon and the type of polymorphism to the discrepancies within and between PvGTSeq sequencing runs, and between PvGTSeq with respect to WGS.	121
Supplementary figure 2.9: Detectability of major and minor clones in mock polyclonal infections.	122
Supplementary figure 2.10: Read depth of major and minor clones within mock samples.	123
Supplementary figure 2.11: Diversity of PvGTSeq amplicons across world regions.	124
Supplementary figure 2.12: Diversity of PvGTSeq amplicons across countries.	125
Supplementary figure 2.13: Panel discrimination power.	126
Supplementary figure 2.14: Scatter plots depicting Tajima's D values in each amplicon included in the PvGTSeq panel.	127
Supplementary figure 2.15: Population differentiation using PvGTSeq in other world regions.	128
Supplementary figure 2.16: Network analysis in Mesoamerica and Colombia.	129
Supplementary figure 2.17: Harmonization of PvGTSeq and WGS data.	130
Supplementary figure 3.1: Pairwise IBD distribution using PvGTSeq and WGS within LAC countries.	131
Supplementary figure 3.2: Scanning of critical $IBD_{PvGTSeq}$ values that indicate cluster merges.	132
Supplementary figure 3.3: Identification of the appropriate $IBD_{PvGTSeq}$ value to define clonal or low-diversity groups using PvGTSeq.	133
Supplementary figure 3.4: Fusion of parental clonal groups and their recombinant progeny in Panama and Guatemala using PvGTSeq.	134
Supplementary figure 3.5: Origin of CL-04. Identification of long IBS and IBD segments between CL-04 and its two parental genetic groups, CL-01 and CL-02.	135
Supplementary figure 3.6: Scanning of critical IBD_{WGS} values that indicate cluster merges.	136
Supplementary figure 3.7: Identification of the appropriate IBD_{WGS} value to define clonal or low-diversity groups using WGS.	137
Supplementary figure 3.8: Identification of clonal and low-diversity groups using WGS genetic relatedness (IBD_{WGS}) in a network analysis.	138
Supplementary figure 3.9: Fusion of parental clonal groups and their recombinant progeny in Panama and Guatemala using PvGTSeq.	139

Supplementary figure 3.10: Identification of the appropriate IBD value to define ecological populations using PvGTSeq.	140
Supplementary figure 3.11: Identification of the appropriate IBD value to define ecological populations using WGS.	141
Supplementary figure 3.12: Identification of ecological groups using WGS genetic relatedness (IBD_{WGS}) in a network analysis.	142
Supplementary figure 3.13: Principal coordinate analysis of LAC <i>P. vivax</i> populations using PvGTSeq markers.	143
Supplementary figure 3.14: Principal coordinate analysis of LAC <i>P. vivax</i> populations using WGS.	144
Supplementary figure 3.15: Possible number of ancestral populations in ADMIXTURE.	145
Supplementary figure 3.16: Pattern of population structure according to ADMIXTURE.	146
Supplementary figure 3.17: Heatmap of PvGTSeq pairwise genetic relatedness between clonal and ecological groups in LAC.	147
Supplementary figure 3.18: Heatmap of WGS pairwise genetic relatedness between clonal and ecological groups in LAC.	148
Supplementary figure 3.19: Origin of PAN-02.1.	149
Supplementary figure 3.20: Origin of CL-03.	150
Supplementary figure 3.21: Origin of CL-05.	151
Supplementary figure 3.22: Origin of PAN-02.2.	152
Supplementary figure 3.23: Origin of CL-19-Like.	153
Supplementary figure 3.24: Origin of CL-03-Like.	154
Supplementary figure 3.25: Origin of COL-05-Like in Panama.	155
Supplementary figure 3.26: Fusion of parental clonal groups and their recombinant progeny in Honduras and USA using PvGTSeq.	156
Supplementary figure 3.27: Fusion of parental clonal groups and their recombinant progeny in Honduras and USA using WGS.	157
Supplementary figure 3.28: Origin of the outbreak in USA.	158
Supplementary figure 3.29: Origin of CL-08 in Honduras.	159

Acknowledgments

I want to begin by thanking **Daniel Neafsey**, my thesis advisor and mentor. **Dan**—Thank you for your patience throughout the five years we have worked together. I know that, on several occasions, I was not an easy student. I came to the program with prior experience in the field, and it can be difficult to relearn things or set aside old habits. Thank you for helping me understand, sometimes by force, that being meticulous does not necessarily mean communicating clearly, and that for our message to have impact, we must strike a balance between rigorous analysis, clarity and simplicity, and a sense of opportunity, or timing. I will always keep in mind the phrase you repeated to me weekly: “Paulo, please keep it simple.” Over this time, I also learned to admire the dedication and discipline you bring to every aspect of your life, from the care you give your plants, to your daily routine of biking and staying fit, to how systematic you are in carrying out your work during the day, and, most importantly, the attention you pay to the needs of each member of your group. I will also be eternally grateful that you enabled me to be with my family, especially during the difficult time we went through with the health of my father. Thank you for everything.

Neafsey Labs—It has been fantastic to spend these five years alongside such wonderful people. Some have since left the group to continue growing professionally, and others joined over time, but I will always be grateful to everyone for making the group feel like a second home. In particular, I want to thank **Philipp Schwabl** for supporting me and for indulging each of my attempts to rebel against Dan’s instructions. Most of the ideas I developed and brought to life came from the long conversations we had about how we should represent and analyze the data, my most creative scientific moments happened during those conversations. I also want to thank **Annie Laws**. You were the first friend I made in the group, always ready to start a completely

random conversation to lighten the mood in the office. You also encouraged some of my vices, like cookies, but more importantly, you helped me connect with the needs of all the collaborators we worked with over the course of these five years. Thank you for being the first to give me a purpose within the group beyond my thesis. I also cannot forget **Katrina Kelley**, another person who helped rekindle my caffeine addiction with our almost daily walks to “Flour”, but who, above all, contributed to processing the thousands of samples analyzed in this work. Without your help, I would not have been able to finish the work on time.

I also want to thank the four members of my DAC, **Dyann Wirth**, **Yonatan Grad**, **Marcia Castro**, and **Marcelo Ferreira**, who, at every meeting, focused on contributing constructively to this work while also supporting my growth as a scientist. They accurately identified some cognitive biases that I tend to fall into and guided me on how to address them. I also want to include in these acknowledgments the entire BPH community and the IID Department of Immunology and Infectious Diseases, and mentors such as Manoj Duraisingh, Jeff Dvorin, and Selina Bob, who, in every departmental meeting, share their knowledge to improve the quality of the presentations. To **Flaminia Catteruccia**, **Tatevik Holmgren**, and **Eric Creighton**, thank you for always being attentive to any need or concern I had during the PhD program. I also want to include my friends at BPH, **Eric Neubauer**, **Sarah Perkins**, **Chloe Hasund**, **Malhar Khushu**, **Branden Wang**, and all the members of my cohort and the program, for their friendship and support throughout this time. It has been rewarding to share this journey with a community that always makes you feel welcome and heard.

To my roommates and friends, **Teguo**, **Ryan**, and **Jaz**: Thank you for making me feel at home and for sharing so many great conversations and experiences. I feel very fortunate that our paths crossed and that I gained your friendship.

To my previous mentors: **Oscar Nolasco**, for first guiding me into the field of Molecular Biology and Bioinformatics. **Dionicia Gamboa**, for opening the doors to malaria research, for your support, and for helping me grow as a professional over more than nine years in your group. **Ananias Escalante**, for introducing me to the field of population genetics. **Alejandro Llanos** and **Michelle Hsiang**, for motivating me to continue in public health and to pursue work that can have an impact on decision-making.

Finally, my family. My brother, **Edgar Manrique**, with whom I shared the adventure of starting graduate school. It helped so much to have someone close by on this journey, and someone to share the challenges of being separated from our family, adapting to a new language, and navigating new cultures. Thank you for always being there, even when we were on different continents. To my beloved wife, **Yesi Correa**. Without a doubt, you have been the person who supported me most throughout this journey. Even though we lived apart for these five years, we found ways to stay close. Thank you for always finding creative ways to spend time together, and for keeping me company through long nights on video calls as I finished the analyses and wrote the thesis. You listened, again and again, to topics that must have seemed boring, but because I was the one talking about them, you did your best to stay awake. Thank you, too, for taking care of our chubby little ones, **Chimu** and **Guchi**, and for sharing all their antics with me so I could distract myself and ease the strain of being far from home. I would not have reached this point without your support and the confidence you gave me over these years. Finally, I want to dedicate this work to my dear parents, **Maria Valverde** and **Miguel Manrique**. You sacrificed so much to give me the best education you could, and you instilled in me the values that make me who I am today. You taught me resilience, humility in the face of my limitations,

and the courage to start over and reinvent myself to overcome obstacles and achieve my goals.

Thank you for all your love and support. This thesis is for you.

Chapter One: Introduction

1.1 *Plasmodium vivax* in Latin America

Malaria is one of the tropical diseases that causes the highest morbidity and is a serious public health problem worldwide. This disease is caused by parasites of the genus *Plasmodium* and transmitted by female mosquitoes of the genus *Anopheles*. The five species that cause malaria in humans are *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale*, and *P. knowlesi*, and their geographic distribution varies by species. *P. falciparum* is the predominant species on the African continent, whereas outside Africa *P. vivax* is the species that causes the greatest morbidity. *P. vivax*, has certain biological characteristics that favor its transmission,¹ including its low developmental temperature range in the mosquito,² low parasitemias and rapid development of immunity that favor asymptomatic infections, early and continuous production of gametocytes,^{3,4} and perhaps most notably, the presence of hypnozoites in the liver.^{5,6} These characteristics allow this parasite to persist despite control efforts, posing a challenge for current strategies.

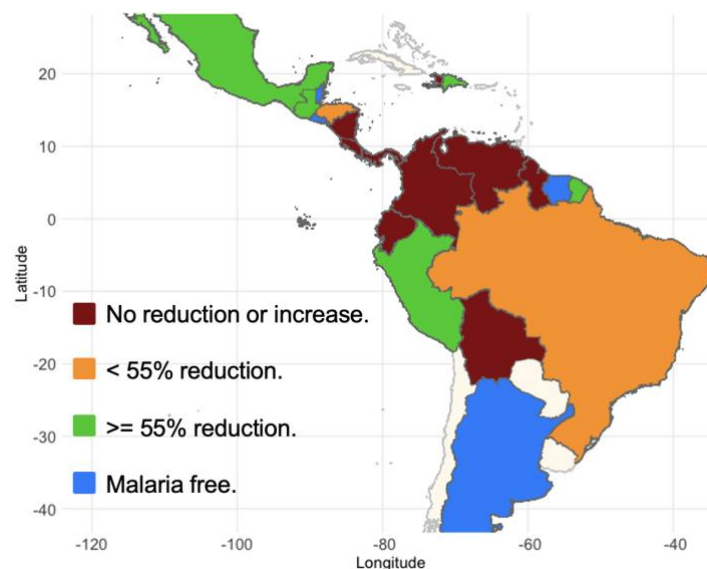


Figure 1.1: *P. vivax* cases 2022 with respect to 2015. Adaptation from WHO Malaria report 2023.⁷

Each year, between 6 and 7 million malaria cases are reported, and in the Americas, *P. vivax* represents over 70% of all malaria cases.⁷ Since 2015, many countries in LAC have shown an increase in cases in their territories, reaching their historic peak of 1 million malaria cases in 2017 and 2018.⁷ Venezuela leads the list with approximately 50% of the cases in the region.⁸ Other countries such as Colombia, Guyana, Bolivia, Ecuador, Nicaragua, and Costa Rica, and more recently Panama, Honduras and Guatemala have shown a stall in the decrease or increase in the number of cases of *P. vivax*.⁷ Between 2010 and 2018, Panama, Honduras, and Guatemala succeeded in reducing, or at least containing, the number of cases per year within their territories. Panama sustained fewer than 1,000 cases per year, while Honduras and Guatemala showed a significant decline. However, starting in 2019, this progress stalled. By 2023, Panama was reporting more than 9,000 cases, and beginning in 2021, Honduras and Guatemala began to follow the same trend.⁷

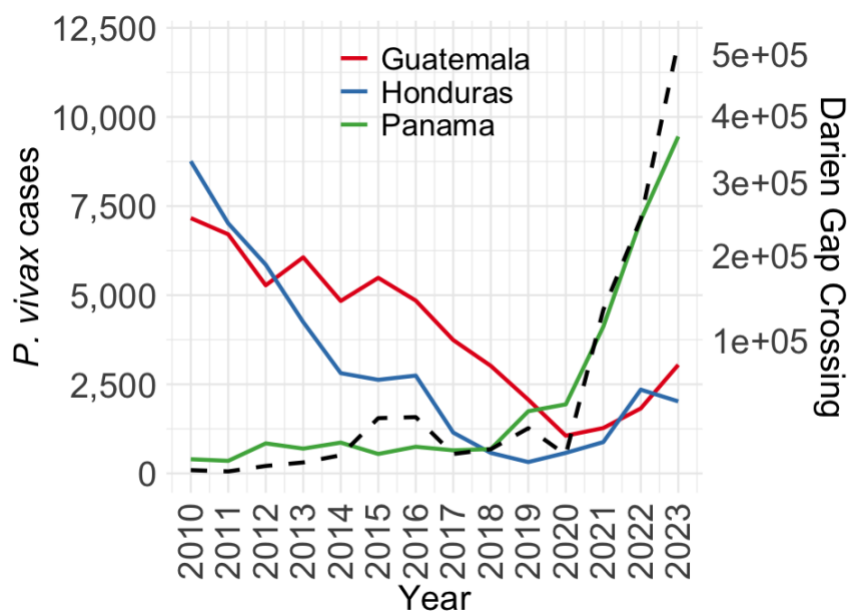


Figure 1.2: Trend in *P. vivax* cases and people crossing the Darien Gap between 2010 to 2023. Data taken from WHO Malaria report 2023⁷ and the Department for migration from Panama.^{9,10}

It is striking to observe that this increase coincides with the intensification of the Venezuelan exodus and other demographic displacements in the region through the Darien Gap.¹¹ By 2021, there were 4.99 million Venezuelan refugees and migrants in LAC,¹¹ however, other countries are also experiencing similar problems. This rises two hypotheses, the first is that the increase in cases in this region is explained mainly to the importation and introduction of parasites from countries with political and migratory problems such as Venezuela. The second is that this increase is due mainly to the expansion of indigenous populations resilient to control and elimination strategies.

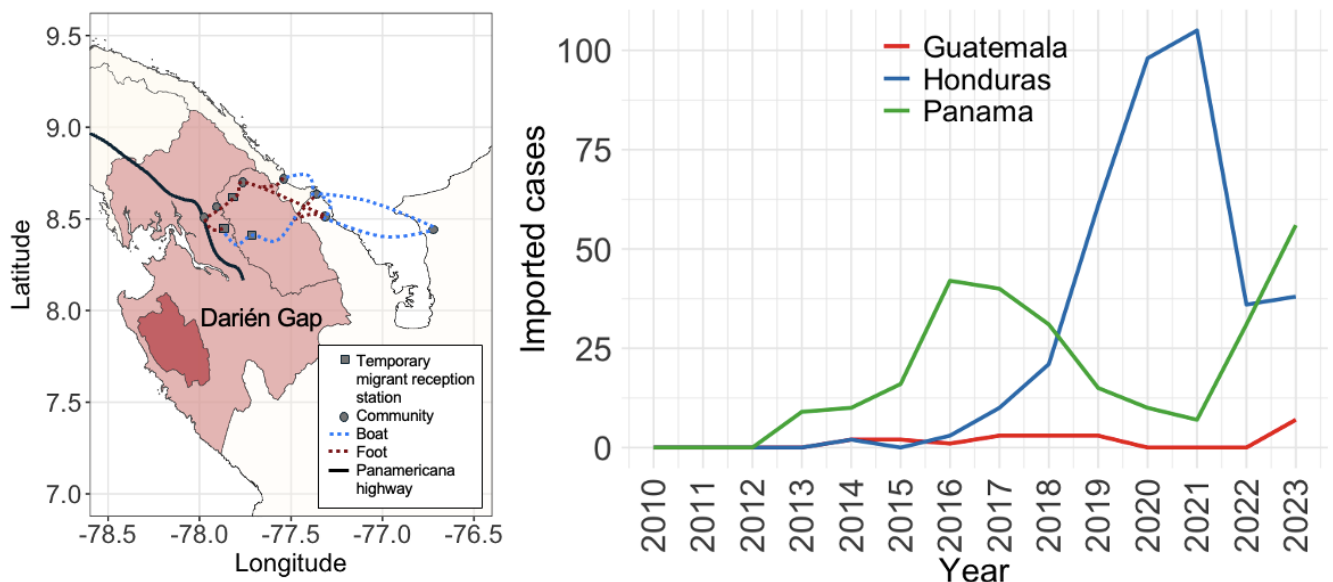


Figure 1.3: Route of migrants through the Darien Gap, and number of imported cases in Panama, Honduras, and Guatemala from 2010 to 2023. Data taken from WHO Malaria report 2023.⁷

Operationally the control of malaria is carried out at the subnational level. Transmission intensity is not distributed homogeneously across geography, instead transmission occurs in small geographic units known as “hot spots”,¹² within which the risk of exposure is influenced by factors such as the proportion of children in the household, socioeconomic status, host habits or risk activities, lack of use of vector control measures, vector distribution and behavior,

housing type, and micro-environmental factors. Thus, for the control and elimination of this disease to be effective, monitoring tools must not only collect information on the presence of transmission foci, but also identify transmission patterns, consider their temporal and spatial heterogeneity, and elucidate the risk factors that favor the spread of parasite populations.¹³

At the individual level, the plan for malaria elimination in the region of the Americas recommended by the Pan American Health Organization (PAHO)¹⁴ is based on four components: diagnosis, treatment, investigation, and response (DTI-R). Each malaria case must be diagnosed by microscopy or rapid diagnostic tests within no more than 48 hours, and treatment must be provided on the same day the diagnosis is confirmed. In the case of *P. vivax*, the recommended treatment is 25 mg/kg chloroquine for 3 days, and 0.5 mg/kg primaquine for 7 days. Then each confirmed case must be investigated and classified no more than 3 days after confirmation to determine the appropriate response. This response may include reactive case detection or vector-control activities (long-lasting insecticidal nets and indoor residual spraying).

At the level of operational units (foci, microareas or hotspots), malaria elimination is based on the cyclical implementation of the following activities: 1) Delimit and characterize the operational units; 2) Organize diagnosis, treatment, investigation, and response actions; 3) Implement DTI-R in the foci; and 4) Management and supervision of operations.

In addition, the World Health Organization stratifies malaria risk in operational units based on the receptivity and vulnerability to malaria transmission.¹⁵ Where, receptivity is understood as an ecosystem's capacity to allow malaria transmission, while vulnerability refers to the risk of importation of the parasite (Imported cases or immigration from endemic areas). The risk strata are: Stratum 1, non-receptive; Stratum 2, receptive with no autochthonous cases

and not vulnerable; Stratum 3, receptive with no autochthonous cases but vulnerable; and Stratum 4, receptive with autochthonous cases and vulnerable. It is on the basis of this stratification that the DTI-R response action is decided.

1.2 Biological cycle of malaria and of *P. vivax*

All *Plasmodium* species that cause malaria in humans are transmitted through the bite of an infected gravid female mosquito of the genus *Anopheles*. During feeding, the mosquito inoculates less than 100 haploid sporozoites from its salivary glands. These travel through the person's bloodstream to liver cells in an average of about one hour. Once in the liver, asexual exoerythrocytic multiplication begins, producing more than 10,000 infective forms called merozoites; this phase lasts about 12 days. In the case of *P. vivax*, dormant stages called hypnozoites can persist in the liver for months or even years.¹⁶

Once the exoerythrocytic cycle is complete, the merozoites are released into the bloodstream and invade red blood cells (*P. vivax* invades only reticulocytes, immature nucleated red blood cells), where they differentiate into trophozoites and undergo mitotic division to form blood schizonts. The new schizonts then lyse the host cell, releasing new merozoites into the bloodstream and establishing the erythrocytic cycle. These merozoites can invade new uninfected red blood cells, and with each cycle the parasitemia increases to about ten times its concentration.¹⁷ At this stage, the infection remains asymptomatic until the infected individual reaches the pyrogenic threshold (the minimum parasitemia required to produce fever). At that point, toxins released by schizont rupture (for example, the parasite's glycosylphosphatidylinositol, or GPI, DNA motifs, and hemozoin, a byproduct of heme-group

degradation from hemoglobin) are recognized by the host's phagocytic cells, generating a pro-inflammatory immune response and triggering symptoms.^{18,19}

Less than 1% of merozoites produced will begin gametocytogenesis rather than continue the erythrocytic cycle.¹ This requires that all merozoites within a schizont commit to differentiating into male or female gametocytes. The genetic information required for gametocyte development is located on chromosome 9. Gametocyte development is subdivided into five morphologically distinct stages. Key morphological characteristics include a subpellicular microtubule cytoskeleton and a double membrane, and the most apparent differences between micro- and macrogametocytes are observed at stage 5. In *P. falciparum*, only stage 5 can be seen in peripheral blood, whereas in *P. vivax* all stages can be observed due to the absence of cytoadherence proteins. From day 6 after the start of gametocyte formation, they become arrested in G0 and only RNA synthesis occurs. Once stage 5 is reached, it takes 2 to 3 days for gametocytes to become infective, and replication is reactivated only when the gametocyte enters the mosquito midgut, culminating in the formation of macro- and microgametes.

Gametocytogenesis can be induced by multiple factors, including host immunity, the density of asexual parasites, hemoglobin concentration in blood, and the presence of reticulocytes (*P. falciparum*). In *P. vivax*, gametocytogenesis can begin from the first generation of merozoites (in the liver), and gametocytes are detected by microscopy within 3 days after the first asexual parasite is seen. However, *P. vivax* gametocytes have a shorter lifespan than those of *P. falciparum*. Studies in patients with neurosyphilis with *P. falciparum* showed that gametocyte production can be continuous throughout the first month of infection, and transmission to mosquitoes can occur even at submicroscopic concentrations.²⁰

Once the mosquito ingests gametocytes, the likelihood of infection depends on their maturity, the concentration of both sexes, and the immune response of the human host and the vector. Fifteen to 20 minutes after ingestion of gametocytes by the female mosquito, male gametocytes resume replication and divide to form eight flagellated microgametes. This process, called exflagellation, lasts about 10 to 12 minutes and marks the start of the sporogonic cycle.²¹ Microgametes move in search of the macrogamete to fertilize it and form the zygote. This new diploid cell differentiates into a motile ookinete over 15 to 30 hours, after which the ookinete moves from the mosquito midgut wall to the basal membrane, where it secretes a thin cyst wall and forms the oocyst.²² Four to fifteen days after the mosquito's blood meal, sporozoite formation begins. Each oocyst can release up to 5,000 sporozoites into the mosquito hemocoel, although only 20% will reach the salivary glands. For both *P. falciparum* and *P. vivax*, the sporogonic cycle lasts 8 to 10 days at 28°C and 16 days at 20°C. However, *P. vivax* can develop even at 14.5°C, whereas *P. falciparum* develops only above 16°C, which has favored the wide geographic spread of *P. vivax* even into temperate areas.

1.3 Transmission characteristics of *P. vivax*

The most notable characteristic is the presence of hypnozoites, which remain latent in the liver of the human host, and the factors that trigger their reactivation are still unknown. Episodes caused by hypnozoite reactivation are called relapses. In high-transmission settings, and in individuals with a long and constant history of exposure, hypnozoites have been observed to reactivate over short time periods (~3 months after the primary infection), and relapses are mostly heterologous,^{23,24} meaning that the genotype responsible for the relapse episode is genetically different from the genotype of the first infection. In contrast, in low-transmission settings, reactivation time tends to be longer (more than 6 months, or even years), and relapses

are usually homologous: the genotypes of the primary infection and the relapse are the same.^{25,26} This ability to remain hidden for long periods allows the parasite to evade implemented control strategies and also confers greater capacity for geographic dispersal.

Thus, the presence of hypnozoites strongly affects gene flow among parasite populations because it favors the geographic movement of parasites.¹¹ It also increases genetic diversity by increasing the likelihood of polyclonal infections in the human host and subsequent recombination in the mosquito.^{27,28} Finally, patterns of parasite population structure can be obscured because relapses provide an additional source of parasites that may or may not be similar to the parasite populations that are being actively transmitted.

Another important biological characteristic of *P. vivax* is the low parasite density associated with infection, which is partly explained by its restricted invasion of reticulocytes.³ This low parasite burden is also supported by the rapid acquisition of non-sterilizing immunity that has been observed in infections caused by *P. vivax* compared with infections caused by other *Plasmodium* species.¹⁷ Together, these factors cause human populations exposed to this parasite to rapidly acquire immunity, and subsequently control parasite density, generating a higher proportion of submicroscopic asymptomatic infections compared to *P. falciparum*. This characteristic allows parasite populations to evade case-finding and case-management strategies (primarily focused on symptomatic carriers), as well as detection tools (based on microscopy as the reference test), thereby favoring transmission and dispersal.

An additional characteristic that favors transmission of this parasite is the early and continuous production of gametocytes, even before the infected individual develops symptoms and goes to a health center to receive medication.¹ This factor is important because it allows even

symptomatic infections to evade established disease-control strategies. In addition, asymptomatic infections, both those that are asymptomatic because they have not yet reached the infected individual's pyrogenic threshold and those that are asymptomatic as a result of non-sterilizing immunity, can generate gametocyte densities sufficient to transmit these sexual stages to the mosquito and thus continue the transmission cycle. Moreover, this capacity to transmit disease before receiving treatment means that the selection pressure exerted by drug administration on resistant and susceptible clones is less pronounced, and fixation of resistant alleles is slow and less readily detected.²⁹ This effect is observed not only for parasite-favorable genetic traits but also for other neutral traits, and therefore *P. vivax* populations tend to fix alleles less readily and maintain population diversity more effectively than *P. falciparum* populations.

Finally, the ability of this parasite to maintain its life cycle at temperate temperatures (14.5°C) is noteworthy. This has enabled a broader global geographic distribution and the invasion of regions not inhabited by *P. falciparum*.³⁰ Together, these characteristics facilitate *P. vivax* transmission and create serious difficulties for control and elimination. However, for transmission to occur effectively, interaction between the vector and the human host is required in a favorable ecological and epidemiological context.

1.4 Transmission and parasite population diversity

In the epidemiological context of malaria, the four main actors in transmission, the vector, humans, parasite biology, and the surrounding or environmental context (including climatological factors and intervention strategies), have a direct effect on the evolutionary processes that affect the parasite population. Therefore, when we make inferences about these evolutionary processes, we also indirectly make inferences about vector, human, parasite, and

environmental factors that favor transmission. For this reason, a variety of genetic indicators have been used in malaria to measure the effect of transmission on parasite populations.

One such indicator is the prevalence of polyclonal infections, or the coexistence of two or more distinct clones within an infected individual.³¹ Because the parasite genome in humans is haploid, when genotyping an infected sample we expect to observe only one allele at each locus analyzed, that is, a single haplotype. The presence of two or more alleles at one or more loci results from the presence of more than one parasite clone (two or more distinct haplotypes) in the sample. This multiplicity of infection may be due to transmission of two or more distinct genotypes through a single bite (co-infection), or acquisition of parasites through multiple independent infections or bites (superinfection). Both processes are related to transmission intensity.

When transmission is low, there are few infectious bites from mosquito to human and vice versa (entomological inoculation rate and infection rate, respectively). Therefore, there is a low probability that a person acquires a superinfection or co-infection, and most infections are monoclonal (the presence of a single clone in the infection).³² When conditions that favor an increase in mosquito populations occur and mosquito-human contact is enhanced, mosquito infection rates and entomological inoculation rates also increase. This leads to an increase in the prevalence of infected people, which in turn increases the likelihood of superinfections and co-infections, producing an increase in the prevalence of polyclonal infections.^{27,33}

Another genetic indicator widely used in population genetic studies of *Plasmodium* parasites is linkage disequilibrium, or the non-random association of alleles at different loci. Allele inheritance mechanisms follow Mendel's law of independent assortment. Thus, if we

analyze two biallelic loci with alleles A, a, B, and b, we expect to observe four possible haplotypes in the population: AB, Ab, aB, and ab. Because allele segregation is independent, the frequency of these four haplotypes is determined by the frequencies of the alleles that make them up. For example, the frequency of haplotype AB (f_{AB}) equals the product of the frequencies of alleles A (f_A) and B (f_B) in the population.³⁴

However, many factors can affect the independence of allele inheritance in a population, causing alleles at different loci to segregate together and not at random. Formally, linkage disequilibrium is defined as the difference (D) between the observed haplotype frequency (f_{AB}) and the expected haplotype frequency ($f_A f_B$). If D equals zero, the alleles at the two loci are in linkage equilibrium. In contrast, if D differs from zero, the alleles at the two loci are in linkage disequilibrium (LD). Multiple coefficients are used to measure LD . The standardized index of association (IAS) is the most commonly used in *Plasmodium* studies because it provides a summary measure of D across alleles at multiple loci.³⁵

Genetic relatedness based on identity by descent has become increasingly relevant in malaria genetic studies in recent years.³⁴ Although the concept is not new, analytical tools that account for the lack of independence in allele inheritance were not available until 2018.^{36–39} Genetic relatedness is defined as the probability that, at any locus in the genome, the alleles sampled from two different individuals are identical by descent (IBD). Two alleles are identical-by-descent (also IBD) if descended from a recent common ancestor in some ancestral reference population. IBD can also be interpreted in terms of shared genome segments unbroken by recombination since a recent common ancestor, where the segment length distribution relates to ancestor generation under a coalescent model.⁴⁰ Thus this index ranges from zero between two unrelated individuals to one between clones.

As noted above, low-transmission settings favor monoclonal infections. Therefore, when an infected person is bitten, a single haplotype is transmitted and the parasite undergoes self-fertilization in the mosquito (sexual reproduction between two genetically identical or related haplotypes). This propagation and reproduction mechanism via selfing is termed clonal propagation, and it causes all alleles within a haplotype to segregate together rather than independently, thereby increasing linkage disequilibrium. For this reason, the presence of LD , or an IAS value different from zero, reflects the presence of self-fertilization and clonal propagation in the population.

The effective size of the parasite population (the number of individuals in the population that determines the magnitude of genetic drift) and, along with it, parasite diversity are also influenced by transmission. Increased transmission reflects the reproductive success of the different circulating parasite lineages and an increase in their effective population size. Under these circumstances, the introduction of new variants into the population is favored (an increase in the population mutation rate, or θ). This increases the number of polymorphic sites, the allelic richness of each polymorphic site (the number of alleles per locus), and the genetic diversity of the population in terms of heterozygosity (H_{exp}). In contrast, when transmission decreases and effective population size is reduced, genetic drift causes the random fixation of alleles, and population diversity decreases. However, effective population size and diversity are also strongly influenced by the introduction of new variants through migration and by the genetic structuring of parasite populations. Therefore, the relationship between these metrics and transmission is not necessarily linear.

Finally, there is genetic structuring, or population subdivision. Unlike the traditional concept of a population in epidemiology or ecology, where a population is the set of individuals

coexisting in the same place and time, in population genetics a population is defined as the set of individuals that share a common genealogy, are connected by gene flow, and for which mating among all individuals belonging to the population is random.³⁹ Population subdivision arises from reproductive isolation (non-random mating) among individuals in the global population. This non-random mating results from natural and historical processes that include geographic isolation and parasite reproductive behavior and preferences, as well as interactions with human host risk behavior, vector behavior, and mobility patterns of both hosts.

This subdivision pattern becomes more pronounced as transmission declines, to the point where what is promoted is clonal propagation through self-fertilization. In practical terms, observing infections caused by parasites belonging to the same clonal group or subpopulation might imply that the individuals carrying these parasites share similar behaviors or risk factors (occupation, geographic proximity, movement or recreational habits, ancestral origin, among others). In this way, determining the pattern of population subdivision makes it possible to identify and measure the differential contribution of each risk factor involved in the origin of cases.

A wide variety of methods are used to detect structuring. Perhaps the most traditional is the so-called Wright's F-statistics (*FST*), also called fixation indices or differentiation indices.⁴¹ When subdivision occurs, the reduction in population size within subpopulations promotes random allele fixation within each subpopulation due to genetic drift, which leads to increased genetic differentiation among subpopulations. This differentiation can occur at hierarchical levels (infections within communities, communities within districts, and districts within provinces). In this way, three coefficients are described: *FIT*, the correlation of alleles within individuals in the total population; *FIS*, the correlation of alleles within subpopulations; and *FST*,

the correlation of alleles among subpopulations. Many variants of these statistics have been proposed to correct biases due to sampling, marker mutation rates, or substitution models. One such variant is Analysis of Molecular Variance (AMOVA), which provides a framework for evaluating different hypotheses about the factors that promote population subdivision.⁴² However, the main limitations of these approaches are that they depend on prior knowledge of the factors or sources that caused the subdivision (this information is rarely known), and they do not allow an objective assignment of genotypes to the proposed subpopulations.

To address this limitation, multiple model-based clustering methods have been proposed. Among them, the implementation of Bayesian models and Markov chain Monte Carlo (MCMC) sampling in the program STRUCTURE stands out.⁴³ If N individuals are sampled from a total population, *a priori* there is no knowledge of the number of subpopulations (k) that may be present, nor which individuals belong to the same population. Using MCMC, STRUCTURE simulates different scenarios or ways of grouping individuals while assuming different values of k . Based on the posterior probability of ancestry parameters (ancestry coefficient, or Q) and the admixture proportion (α), the program identifies the model whose posterior probability is maximized given the likelihood of the data. In this way, STRUCTURE estimates allele frequencies within each subpopulation, the ancestral origin of each individual (Q), and the number of subpopulations present in the total population. For the purpose of scaling this approach to a larger volume of data, ADMIXTURE was developed, a deterministic tool based on maximum likelihood.

Another simple way to analyze patterns of population structure is to visualize genetic differences using principal component analysis (PCA) or principal coordinates analysis (PCoA), an approach that has become popular in genomic studies.⁴⁴ To do this, it is essential to define

how to represent pairwise genetic differences between samples (i.e., a genetic relatedness matrix). The simplest option is to calculate Euclidean distances for all pairwise comparisons, but other genetic distance methods that account for allele frequencies or are based on IBD often yield better results.

PCA-based or STRUCTURE-like methods are the most commonly used in malaria to identify population structure. They are effective at detecting ancestral populations when diversity and population differentiation are high. However, in the presence of clonal groups, these methods can treat clonal groups as independent populations; in PCA, this can also distort the genetic relationships between the clonal group and other population groups. A practical solution is to use a single representative member of each clonal group; however, the epidemiological interpretation becomes more diffuse because these analyses alone do not explicitly show the interactions (recombination) between clonal groups with different ancestry that coexist in the same time period and territory (for example, when outcrossing occurs between local populations and recently introduced parasites). One way to describe these recent recombination events is to search for long IBD segments, although holistic methods that integrate this approach within the theoretical framework of coalescent theory are still under development.⁴⁵

1.5 References

1. Bousema, T. & Drakeley, C. Epidemiology and Infectivity of *Plasmodium falciparum* and *Plasmodium vivax* Gametocytes in Relation to Malaria Control and Elimination. *Clin. Microbiol. Rev.* **24**, 377–410 (2011).
2. Guerra, C. A., Snow, R. W. & Hay, S. I. Defining the Global Spatial Limits of Malaria Transmission in 2005. *Adv. Parasitol.* **62**, 157–179 (2006).
3. McKenzie, F. E., Jeffery, G. M. & Collins, W. E. PLASMODIUM VIVAX BLOOD-STAGE DYNAMICS. *J. Parasitol.* **88**, 521–535 (2002).

4. McKenzie, F. E., Jeffery, G. M. & Collins, W. E. GAMETOCYTEMIA AND FEVER IN HUMAN MALARIA INFECTIONS. *J. Parasitol.* **93**, 627–633 (2007).
5. Imwong, M. *et al.* Relapses of Plasmodium vivax Infection Usually Result from Activation of Heterologous Hypnozoites. *J. Infect. Dis.* **195**, 927–933 (2007).
6. Chen, N., Auliff, A., Rieckmann, K. & Cheng, Q. Relapses of Plasmodium vivax Infection Result from Clonal Hypnozoites Activated at Predetermined Intervals. *J. Infect. Dis.* **195**, 934–941 (2007).
7. World malaria report 2024. <https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2024>.
8. Grillet, M. E. *et al.* Malaria in Southern Venezuela: The hottest hotspot in Latin America. *PLoS Negl. Trop. Dis.* **15**, e0008211 (2021).
9. Estadísticas: Tránsito irregular por el Darién 2022 – Migración Panamá. <https://www.migracion.gob.pa/estadisticas/> (2023).
10. Estadísticas: Tránsito irregular por el Darién 2025 – Migración Panamá. <https://www.migracion.gob.pa/estadisticas/> (2023).
11. Where Are Venezuelan Migrants and Refugees Going? An Analysis of Legal and Social Contexts in Receiving Countries. *The Center for Migration Studies of New York (CMS)* <https://cmsny.org/publications/venezuelan-migrants-legal-contexts-wolfe-010421/>.
12. Bousema, T. *et al.* Hitting Hotspots: Spatial Targeting of Malaria for Control and Elimination. *PLOS Med.* **9**, e1001165 (2012).
13. Arnott, A., Barry, A. E. & Reeder, J. C. Understanding the population genetics of Plasmodium vivax is essential for malaria control and elimination. *Malar. J.* **11**, 14 (2012).
14. Pan American Health Organization (PAHO). Microplanning Manual for Malaria Elimination in the Region of the Americas. <https://iris.paho.org/items/23bb3bf5-db26-4f7c-895a-fb2bb8c85e0f>.
15. WHO malaria terminology, 2021 update. <https://www.who.int/publications/i/item/9789240038400>.
16. White, N. J. Why Do Some Primate Malarias Relapse? *Trends Parasitol.* **32**, 918–920 (2016).
17. Oakley, M. S., Gerald, N., McCutchan, T. F., Aravind, L. & Kumar, S. Clinical and molecular aspects of malaria fever. *Trends Parasitol.* **27**, 442–449 (2011).
18. López, C., Yepes-Pérez, Y., Hincapié-Escobar, N., Díaz-Arévalo, D. & Patarroyo, M. A. What Is Known about the Immune Response Induced by Plasmodium vivax Malaria Vaccine Candidates? *Front. Immunol.* **8**, 126 (2017).

19. Schofield, L. & Grau, G. E. Immunological processes in malaria pathogenesis. *Nat. Rev. Immunol.* **5**, 722–735 (2005).
20. Eichner, M. *et al.* Genesis, sequestration and survival of *Plasmodium falciparum* gametocytes: parameter estimates from fitting a model to malariatherapy data. *Trans. R. Soc. Trop. Med. Hyg.* **95**, 497–501 (2001).
21. Enger, A., Strand, Ø. A., Ranheim, T. & Hellum, K. B. Exflagellation of Microgametocytes in *Plasmodium vivax* Malaria: A Diagnostic Conundrum. *Med. Princ. Pract.* **13**, 298–300 (2004).
22. Bousema, T. *et al.* Mosquito Feeding Assays to Determine the Infectiousness of Naturally Infected *Plasmodium falciparum* Gametocyte Carriers. *PLOS ONE* **7**, e42821 (2012).
23. de Araujo, F. C. F., de Rezende, A. M., Fontes, C. J. F., Carvalho, L. H. & Alves de Brito, C. F. Multiple-clone activation of hypnozoites is the leading cause of relapse in *Plasmodium vivax* infection. *PloS One* **7**, e49871 (2012).
24. Betuela, I. *et al.* Relapses Contribute Significantly to the Risk of *Plasmodium vivax* Infection and Disease in Papua New Guinean Children 1–5 Years of Age. *J. Infect. Dis.* **206**, 1771–1780 (2012).
25. Eede, P. V. den *et al.* *Plasmodium vivax* Sub-Patent Infections after Radical Treatment Are Common in Peruvian Patients: Results of a 1-Year Prospective Cohort Study. *PLOS ONE* **6**, e16257 (2011).
26. Kim, J.-R. *et al.* Genotyping of *Plasmodium vivax* Reveals Both Short and Long Latency Relapse Patterns in Kolkata. *PLOS ONE* **7**, e39645 (2012).
27. Koepfli, C. *et al.* Multiplicity and Diversity of *Plasmodium vivax* Infections in a Highly Endemic Region in Papua New Guinea. *PLoS Negl. Trop. Dis.* **5**, e1424 (2011).
28. Koepfli, C. *et al.* A High Force of *Plasmodium vivax* Blood-Stage Infection Drives the Rapid Acquisition of Immunity in Papua New Guinean Children. *PLoS Negl. Trop. Dis.* **7**, e2403 (2013).
29. Schneider, K. A. & Escalante, A. A. Fitness components and natural selection: why are there different patterns on the emergence of drug resistance in *Plasmodium falciparum* and *Plasmodium vivax*? *Malar. J.* **12**, 15 (2013).
30. Escalante, A. A. *et al.* A monkey's tale: The origin of *Plasmodium vivax* as a human malaria parasite. *Proc. Natl. Acad. Sci.* **102**, 1980–1985 (2005).
31. Escalante, A. A. *et al.* Malaria Molecular Epidemiology: Lessons from the International Centers of Excellence for Malaria Research Network. *Am. J. Trop. Med. Hyg.* **93**, 79–86 (2015).

32. Volkman, S. K., Neafsey, D. E., Schaffner, S. F., Park, D. J. & Wirth, D. F. Harnessing genomics and genome biology to understand malaria biology. *Nat. Rev. Genet.* **13**, 315–328 (2012).
33. Nkhoma, S. C. *et al.* Population genetic correlates of declining transmission in a human pathogen. *Mol. Ecol.* **22**, 273–285 (2013).
34. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).
35. Brown, A. H. D., Feldman, M. W. & Nevo, E. MULTILOCUS STRUCTURE OF NATURAL POPULATIONS OF HORDEUM SPONTANEUM. *Genetics* **96**, 523–536 (1980).
36. Neafsey, D. E., Taylor, A. R. & MacInnis, B. L. Advances and opportunities in malaria population genomics. *Nat. Rev. Genet.* **22**, 502–517 (2021).
37. Zhu, S. J., Almagro-Garcia, J. & McVean, G. Deconvolution of multiple infections in *Plasmodium falciparum* from high throughput sequencing data. *Bioinformatics* **34**, 9–15 (2018).
38. Schaffner, S. F., Taylor, A. R., Wong, W., Wirth, D. F. & Neafsey, D. E. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *Malar. J.* **17**, 196 (2018).
39. Henden, L., Lee, S., Mueller, I., Barry, A. & Bahlo, M. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLoS Genet.* **14**, e1007279 (2018).
40. Taylor, A. R., Jacob, P. E., Neafsey, D. E. & Buckee, C. O. Estimating Relatedness Between Malaria Parasites. *Genetics* **212**, 1337–1351 (2019).
41. Meirmans, P. G. & Hedrick, P. W. Assessing population structure: FST and related measures. *Mol. Ecol. Resour.* **11**, 5–18 (2011).
42. Meirmans, P. G. Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evol. Int. J. Org. Evol.* **60**, 2399–2402 (2006).
43. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
44. Elhaik, E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Sci. Rep.* **12**, 14683 (2022).
45. Camponovo, F., Buckee, C. O. & Taylor, A. R. Measurably recombining malaria parasites. *Trends Parasitol.* **39**, 17–25 (2023).

**Chapter Two: PvGTSeq and PvCRiSP: two amplicon-based targeted sequencing panels for
*Plasmodium vivax***

2.1 Abstract

Plasmodium vivax is the main cause of malaria outside of sub-Saharan Africa, and in many settings it presents significant challenges to malaria elimination efforts. Despite some control successes in the Americas, regional annual case counts of malaria have increased by over 25% between 2014 and 2023, largely driven by *P. vivax*. Genomic surveillance can play a key role in understanding the extent to which disease persistence represents indigenous transmission as opposed to introduction of new strains through migration, and whether specific variants evade control measures. Efforts to make *P. vivax* genomic surveillance more cost-effective have led to the development of targeted sequencing-based methods, which strike a varying balance between assay sensitivity and breadth/informativeness. We introduce two new highly sensitive multiplexed amplicon sequencing panels for *P. vivax*: PvGTSeq and PvCRiSP. PvGTSeq requires selective whole-genome amplification (sWGA) and contains 249 amplicons—36 for antimalarial resistance and 213 for population structure—optimized for Latin America but applicable to all continents. PvCRiSP features four highly polymorphic amplicons that operate without sWGA and is designed to estimate complexity of infection (COI), identify instances of clonal transmission, and characterize recurrent episodes. Both panels use a single multiplex PCR with non-proprietary reagents, achieve $\geq 75\%$ amplicon recovery at parasitemias as low as five parasites/ μL , and PvCRiSP remains effective with low quality DNA. PvGTSeq showed high sequencing accuracy (error rate $3.85\text{e-}4\%$ - $2.87\text{e-}3\%$), and both panels efficiently detected alleles from minority clones in simulated polyclonal infections. We validated both panels with samples from Colombia, Guyana, Honduras, Panama, and Venezuela, and performed in-silico assessments using data from 16 countries worldwide, confirming that these two panels have high power to discriminate samples and assign global geographic origin to imported cases. These

panels will therefore be useful tools for *P. vivax* molecular surveillance in diverse geographic settings. The results presented in this Chapter had been accepted for publication in the scientific journal PLOS Neglected Tropical Diseases, and the work will be published under the Creative Commons Attribution (CC BY) 4.0 license.¹

2.2 Introduction

Plasmodium vivax is the main cause of malaria outside of Africa. Each year, 6-7 million cases are reported, and in the Americas this *Plasmodium* species causes 76% of all malaria cases.² Despite efforts to reduce malaria, *P. vivax* possesses biological characteristics that make it resilient to current control measures. Compared to *Plasmodium falciparum*, *P. vivax* has an earlier developmental commitment to transmissible sexual stages³ and can persist as a dormant hypnozoite in the liver that can reactivate weeks or months after the primary infection.⁴ In addition to these biological advantages, several factors complicate the control and elimination of this parasite in the Americas: the high frequency of submicroscopic and asymptomatic infections,⁵ its prevalence in remote rural and border regions with limited health surveillance,⁶ and inadequate patient compliance with radical cure treatment (7 to 14 days primaquine) to clear hypnozoites from the liver.⁷

Despite successful malaria control and elimination efforts in some countries across the Americas, there has been an increase in *P. vivax* cases since 2014—even in areas that previously reported fewer than 1,000 cases annually or where *P. falciparum* was the dominant parasite.² Although Venezuela has experienced the most dramatic increase in cases,⁸ other countries such as Colombia, Costa Rica, Guyana, Ecuador, Guatemala, Honduras, Nicaragua, and Panama have also shown increases.² The drivers of *P. vivax* persistence and resurgence are complex, and

control programs therefore require cost-effective tools to monitor these population changes and enable evidence-based decision-making.

Genomic surveillance can play an important role in this goal, and the development of targeted sequencing-based methods for *P. falciparum* has made this tool more cost-effective and easy to transfer to National Malaria Control Programs.^{9–13} Recently, the use of genomic surveillance has expanded to *P. vivax* with a variety of use cases.^{14–21} The most frequent use case has been characterizing population connectivity and distinguishing imported infections from those that are locally acquired. Other applications include characterizing candidate antimalarial resistance-related genes and potential blood-stage vaccine targets, as well as identifying amplicons to distinguish between infection relapses, recrudescences, and reinfections.

Here, we introduce two new amplicon panels for targeted Illumina sequencing of *P. vivax*. The first, PvGTSeq, is a multipurpose panel composed of 249 amplicons, including 213 highly heterozygous amplicons for assessing parasite diversity, relatedness, and population structure at three geographic scales: global, intra-continental, and sub-national in Latin America and the Caribbean. This panel has enhanced sensitivity for identifying imported infections in the Americas, while maintaining utility for malaria molecular surveillance worldwide. Additionally, this panel includes 36 amplicons targeting ten candidate genes associated with antimalarial resistance. The second panel, PvCRiSP, consists of four highly-polymorphic amplicons. This high-sensitivity panel is ideal when dozens or hundreds of amplicons do not improve genotype discrimination or enhance detection of minor clones in polyclonal infections. PvCRiSP is best suited for efficiently estimating the number of genetically distinct parasites within a single infection (complexity of infection, or COI); identifying instances of clonal transmission, for

example in localized outbreaks; or characterizing recurrent infection episodes in therapeutic efficacy studies. Both panels work on all Illumina sequencing platforms and require a single multiplex PCR reaction prior to the indexing reaction. Both also involve non-proprietary and easily-sourced reagents, facilitating global access and sustainable programmatic use. In this manuscript, we describe the performance of the panels to assess their utility for various use cases.

2.3 Methods

2.3.1 Clinical samples and genomic databases

Multiple genomic data sources and clinical samples were used for the selection, standardization, and validation of PvGTSeq and PvCRiSP amplicons. In the selection process for PvGTSeq amplicons, genomic data from four different sources were used: 1) 1895 genomes from the MalariaGen Pv4 database,²² 2) 89 Brazilian genomes available from the European Nucleotide Archive under project codes PRJEB56411, PRJEB44419, PRJEB36199,²³ 3) 705 genomes obtained from Guyana not present in MalariaGen Pv4,²⁴ and 4) 532 new genomes obtained from clinical samples from Colombia (n = 200), Peru (n = 90) and Venezuela (n = 342). For the validation of PvGTSeq and PvCRiSP, a total of 821 additional clinical blood spot or DNA samples from Colombia (n = 259), Guyana (n = 460), Honduras (n = 39), Panama (n = 53), and Venezuela (n = 10) were genotyped using one of these methods (**Supplementary figure 2.1**). All sequenced clinical samples were collected with ethical approval from local institutions. Written informed consent was obtained from participants before clinical samples were collected and in the case of children, formal consent was obtained from parents or guardians. In Colombia, approval was granted by the Comité de Ética para Investigación en Humanos Centro

Internacional de Vacunas (CECIV) under the protocol titled “Vigilancia Molecular de la Malaria en Colombia”. In Guyana, approval was obtained from the Ministry of Health Institutional Review Board for protocol 645/2019. In Honduras approval was granted by the UNAH Comité de Ética de Investigación de la Maestría de Enfermedades Infecciosas y Zoonóticas for protocols 02-2023 and 03-2020. In Panama, approval was obtained from Comité de Bioética de la Investigación del Instituto Conmemorativo Gorgas de Estudios de la Salud for protocols 468/CNBI/ICGES/06, 413/CNBI/ICGES/12, and 314/CBI/ICGES/24. In Peru, approval was obtained from UNTRM Comité de Ética de Investigación (CIEI) for protocols CIEI-N° 011, and CIEI-N° 015. In Venezuela, approval was granted by the Comité Independiente de Bioética para la Investigación del Centro Nacional de Bioética Venezuela (CIBI-CENABI) for protocol CIBI-CENABI-03/2022. Sample analysis was approved by the Harvard University Longwood Institutional Review Board (Protocols: IRB19-1779, IRB20-1636, IRB23-0621, IRB23-1567).

2.3.2 Parasite DNA quantification

We measured parasitemia for a subset of clinical samples used in assay optimization, limit of detection measurement (LOD), and minor clone sensitivity analysis. Parasitemia was measured by real-time quantitative polymerase chain reaction (qPCR) using the 18S ribosomal gene²⁵ and a standard curve from the 3D7 strain of *P. falciparum*.¹ Due to the lack of a quantified *P. vivax* culture or large volumes of a high-parasitemia sample that could be used as a control for all assays in this study, to quantify *P. vivax* parasitemia on clinical samples we used genomic DNA from the monoclonal *P. falciparum* reference strain, 3D7, as a quantitation standard. The parasite strain was grown using leukocyte depleted red blood cells and we prepared a positive control template representing infected human blood with 10,000 parasites/ μ L by mixing human genomic DNA (13.76 ng/ μ L) with 3D7 genomic DNA (0.92 ng/ μ L) at 2.66:1

ratio (v/v), yielding 10 ng/μL human gDNA and 0.25 ng/μL 3D7 gDNA. The 0.25 ng represents 10,000 *P. falciparum* genomes (based on 23-Mbp genome size and 660 g/mol bp average mass), assuming one haploid genome per infected cell. We created controls representing 10,000, 1000, 100, 10 and 1 parasites/μL by serial 1:10 dilution with 10 ng/μL human gDNA. Because primer pair amplification efficiency may differ between *P. falciparum* and *P. vivax*, we measured the amplification efficiency in the logarithmic PCR phase for each sample as a correction factor relative to the standard curve efficiency.

2.3.3 Selection of amplicons and primer design

To describe patterns of population structure and connectivity at three geographic scales (global, Latin America, and subnational within Latin America), we selected discriminatory amplicons using Discriminant Analysis of Principal Components (DAPC).²⁶ Whole genome sequencing information (WGS) from the MalariaGEN Pv4 dataset²² and lab-generated WGS were filtered to remove low-quality regions and genotyping error-prone loci. Genomes were divided into 150 bp segments, with segment selection based on: 1) >0.1% contribution to top discriminant components, 2) highest contribution within 200,000 bp windows, and 3) location in functionally annotated genes. Segments in uncharacterized genes were selected only when no other segments met the first criterion within a window. We performed this selection process independently for each of the three geographical scales, generating three subsets of segments. Segments shared across all three subsets were prioritized for primer design.

To confirm that our amplicon panel can detect population structure, we compared the patterns revealed by principal coordinate analysis (PCoA) with those generated using whole-genome data and two previously reported amplicon panels, PvAmpliSeq and rhAmpSeq.^{15,18,21} We performed this analysis at two geographic scales: across world regions and among countries

in Latin America and the Caribbean. In addition to WGS from MalariaGEN Pv4 ²² and our lab generated genomes, we include 89 Brazilian genomes available from the European Nucleotide Archive.²³ To avoid bias from missing data, we included only samples with WGS information that had $\geq 75\%$ coverage both genome-wide and across the markers for each of the three amplicon panels analyzed. PCoA was performed using the inverse of identity by descent (IBD), with IBD calculated using the R package *paneljudge*.^{12,28,29} For WGS data, IBD was calculated using only biallelic SNPs whose minor alleles are present in at least five samples in the dataset, while for the amplicon panels we used all variant sites (with minor alleles present in at least five samples), excluding those located in homopolymers and dinucleotide short tandem repeats (STRs). We assessed the accuracy of IBD estimates for each amplicon panel by measuring the root mean squared error (RMSE) relative to WGS, and we assessed precision by quantifying the coefficient of variation (CV). We calculated RMSE and CV in 0.05 IBD windows spanning 0 to 1. We tested differences between amplicon panels in each window using a paired Student's t-test. For surveillance of antimalarial drug resistance, we identified 150 bp segments (with 50 bp flanking regions) covering polymorphic sites in 11 candidate resistance genes: *CRT* (PVP01_0109300),^{30–36} *MRPI* (PVP01_0203000),^{37–40} *DMT2* (PVP01_0312700),³⁹ *DHFR* (PVP01_0526600),^{34,37,39,41–46} *MDR1* (PVP01_1010900),^{31,37,46–51} *PI3K* (PVP01_1018600),^{50,52} *ABCE1* (PVP01_1103800),³⁶ *KELCH13* (PVP01_1211100),^{34,39,53} *MDR2* (PVP01_1259100),^{54,55} *DHPS* (PVP01_1429500),^{32,34,45,46,55–59} and *MRP2* (PVP01_1447300).^{39,54} Finally, we contracted the services of GTseek LLC ⁶⁰ to design multiplexable primer sets following Genotyping in Thousand Sequencing (GT-Seq) protocol specifications.⁶¹ We named this amplicon panel PvGTSeq in honor of its design and its capacity to genotype up to 1,152 samples and 249 amplicons per sample in a single sequencing run using a NovaSeq instrument. We then

selected four amplicons from PvGTSeq—CG2_related, **RIPR**, VPS11, and **PIGM**—to form the PvCRiSP panel. These four amplicons were chosen based on their heterozygosity, ability to detect polyclonal infections, and superior amplification efficiency across the various Latin American countries analyzed with PvGTSeq.

2.3.4 Whole genome sequencing and variant call

For the 1271 clinical samples from Colombia (n = 459), Guyana (n = 460), and Venezuela (n = 352) we extracted total genomic DNA from dried blood spot samples (20–35 mm² spotted area punched per sample) using KingFisher Ready DNA Ultra 2.0 Prefilled Plates on the Kingfisher Flex instrument (ThermoFisher Scientific). For the 182 DNA samples from DNA Honduras (n = 39), Panama (n = 53), and Peru (n = 90), our collaborators sent DNA they had stored in their sample banks from previous studies. We subsequently applied selective whole genome amplification (sWGA) to DNA extracts.⁶² Each 50 µL sample reaction consisting of 5 µL 10x phi29 polymerase buffer (NEB B0269S), 0.125 µL (2.5 µg) recombinant albumin (NEB B9200S), 0.5 µL primer mix (10 oligos combined at 250 µM, see below), 5 µL 10 mM dNTPs (Thermo Scientific), 16.375 µL nuclease-free water, 3 µL (30 units) phi29 DNA polymerase (M0269L), and 20 µL DNA. Reactions were prepared on ice, with components added in the order shown. We used ‘Pvset1’ for *P. vivax* as reported by Cowell et. al. 2017. Amplifications were generated using step down incubation (35 °C for 5 min, 34 °C for 10 min, 33 °C for 15 min, 32 °C for 20 min, 31 °C for 30 min, 30 °C for 16 h, and 65 °C for 15 min), followed by cooling to 4 °C. Note that we applied this sWGA protocol only to obtain whole genome sequencing information. We subsequently modified this protocol to enhance its sensitivity for low parasitemia samples, and that final version which is based on the Genomiphi v2 kit (Cytiva) was used to generate amplicon sequencing data through our PvGTSeq panel. We then applied

AMPure XP magnetic beads (Beckman Coulter A63881) at room temperature to exchange post-reaction sample buffer to 10 mM Tris-HCl + 0.1 mM EDTA. Final library construction using the NEBNext Ultra II FS DNA Library Prep Kit (NEB E6177) and 2×151 bp sequencing on the Illumina NovaSeq 6000 platform was completed at the Broad Institute. We aligned reads to the *P. vivax* P01 reference genome version 1⁶³ assemblies using BWA-MEM v0.7.17-r1188⁶⁴ and called SNPs and INDELs using GATK v3.5-0-g36282e4⁶⁵ ‘HaplotypeCaller’ and ‘GenotypeGVCFs’ according to best practices defined by the Pf3k consortium (http://www.malariagen.net/data_package/pf3k-5/). The mapping and joint variant call process also included samples from the MalariaGEN Pv4 dataset. All downstream genetic analyses focused on variant sites, single nucleotide polymorphisms (SNPs) and insertions and deletions (INDELs), in core regions of the genome. Samples missing genotype calls for >25% variant sites at $\geq 2\%$ minor allele frequency (after application of the core region filter) were also excluded from further analysis. Following sample exclusions, we removed variant sites located in coding regions prone to alignment errors (coding regions with observed heterozygosity above the 95th percentile and variant site density exceeding the 92.5th percentile of all coding regions in all populations). We also removed variant sites located in homopolymers and di-nucleotide short tandem repeats, as these are prone to PCR errors.

2.3.5 Panel protocols and optimization

To facilitate standardization of library generation procedures, we started with the conditions described for a previously developed *P. falciparum* protocol.¹² From there, we made modifications to annealing temperature and primer concentrations, the final volume of the multiplex PCR reaction, and the primer sets and enzymes for sWGA. These modifications were applied to improve sequencing yield in terms of the number of on-target reads per sample per

amplicon. For this standardization, we selected 60 clinical samples from Colombia (n = 30), Peru (n = 15), and Venezuela (n = 15), each with parasitemia levels above 100 parasites/ μ L.

Individual protocol modifications were evaluated in a paired manner, using the previous protocol conditions as a reference. Comparisons were made in terms of the number of reads per sample per amplicon, the total number of reads per sample, and the percentage of amplicons amplified per sample. Additionally, at each step, the rate of non-target product generation for each primer was evaluated, and the primer pairs that generated a higher amount of non-target reads compared to specific reads were removed from the protocol for subsequent optimization rounds (**Supplementary figure 2.1**).

The final protocol for sWGA consisted on two consecutive rounds of amplification of the DNA extracts. For both rounds of sWGA, the 10 μ L sample mix consisted of 0.52 μ L of primer set (Pvset1920 for the first round and Pvset1 for the second round), 6.25 μ L of 4 mM dNTPs (Thermo Scientific), 0.23 μ L of nuclease-free water, and 3 μ L of DNA. After heating the sample mix for 3 minutes at 95 °C, we added 10 μ L of reaction mix containing 9 μ L of reaction buffer and 1 μ L of Phi29 enzyme (Genomiphi v2 kit, Cytiva). Amplification was performed using step-down incubation (35 °C for 5 min, 34 °C for 10 min, 33 °C for 15 min, 32 °C for 20 min, 31 °C for 30 min, 30 °C for 24 h, and 65 °C for 15 min), followed by cooling to 4 °C. The amplified product from each round was diluted 1 over 50 using nuclease-free water.

The final protocol for the library generation and sequencing for PvGTSeq and PvCRiSP consisted of a nested PCR, a multiplex PCR to specifically amplify the 249 targets in the panel, followed by a conventional PCR to add the i5 and i7 Illumina adapters (Nextera indexing). Each multiplex PCR reaction consisted of 10 μ L of Qiagen Plus Master Mix, 3 μ L of primer mix (0.067–0.3 μ M), 1 μ L of nuclease-free water, and 6 μ L of DNA template. Amplification was

performed with 15 min of initial denaturation, followed by 5 cycles of step-up incubation (30 s denaturation at 95 °C, 30 s annealing starting at 57 °C and increasing by 5% in each cycle, and 2 min extension at 72 °C), then 20 cycles of amplification (30 s denaturation at 95 °C, 30 s annealing at 65 °C, and 30 s extension at 72 °C), followed by cooling to 4 °C. The Nextera indexing PCR mix consisted of 5 µL of KAPA HiFi HotStart Ready Mix (2x), 2.2 µL of unique dual index, and 7.2 µL of a 1/13 dilution of the multiplex PCR1 product. Samples from two 96-well plates were pooled, and two rounds of cleanup and size selection were performed using AMPure XP beads at 1X. Library quantification was performed by qPCR using the KAPA SYBR FAST kit (Roche), and amplicon sizes were verified on a Bioanalyzer. Finally, sequencing was performed using the Illumina iSeq 100, MiSeq v2, or NovaSeq 6000 kits for a total of 300 cycles.

2.3.6 Amplicon data analysis

We used the pipeline previously described for *P. falciparum* with some modifications to generalize its use for *P. vivax*.¹² Briefly, we processed paired-end Illumina sequencing data in the form of FASTQ files using a custom analysis pipeline for which documentation can be found at <https://github.com/broadinstitute/malaria-amplicon-pipeline>. This pipeline utilizes the Divisive Amplicon Denoising Algorithm (DADA2)⁶⁶ to obtain microhaplotypes (i.e., alleles or ‘amplicon sequence variants’). We first aligned microhaplotypes obtained from dada2 against a custom-built database of PvP01 reference sequences for each amplicon locus. We then summarized observed sequence polymorphism into a concise format by converting individual microhaplotypes into ‘pseudo-CIGAR’ strings using a custom python script that can be found at <https://github.com/Paulonvnnv/MHap-Analysis>. The rules used for this conversion are also detailed in **Supplementary figure 2.2**. Next, microhaplotypes were discarded if supported by

fewer than 10 read-pairs or by less than 1% of total read-pairs within a heterozygous locus. We also filtered out microhaplotypes exhibiting less than 80% identity to the reference sequence. We then masked variants in both the FASTA sequence of the microhaplotype and its pseudo-cigar string representation according to three criteria: 1) SNPs or INDELs located in homopolymer regions (≥ 5 consecutive identical nucleotides), 2) INDELs positioned at either the beginning or end of the microhaplotype sequence, and 3) variants predominantly ($> 66\%$) present as minor alleles and mainly found in heterozygous sites ($> 66\%$) across the population. All downstream analyses were performed based on the pseudo-cigar strings using custom R scripts that can be found at https://github.com/Paulonvnnv/PvGTSeq_PvCRiSP_paper. All functionalities created to handle pseudo-cigar strings are documented at <https://github.com/Paulonvnnv/MHap-Analysis>. This pipeline is also implemented in a Terra workspace—a user-friendly cloud-based platform for processing amplicon sequencing data.

2.3.7 Genotyping performance: amplification rate and limit of detection (LOD)

Following pseudo-cigar generation and filtration steps, we evaluated the amplification rate and LOD using a set of 160 and 240 samples for PvGTSeq and PvCRiSP, respectively, and from three different endemic countries: Colombia, Guyana, and Venezuela. For PvGTSeq, parasite concentrations ranged from 0.26 to 1309.38 parasites/ μL . For PvCRiSP, which contains only four amplicons and was expected to have greater sensitivity, we used lower concentrations. We performed a 1/10 dilution of the samples previously used for PvGTSeq, resulting in final concentrations ranging from 0.03 to 130.94 parasites/ μL . This approach ensured sufficient samples with low parasite concentrations that would not amplify any of the four PvCRiSP markers, allowing us to obtain a more precise estimate of its LOD. The amplification rate was analyzed at both the amplicon and sample level, and was defined as the percentage of samples

with ≥ 10 reads per amplicon and as the percentage of amplified amplicons per sample, respectively. The LOD was calculated for each individual amplicon as well as the complete panel of amplicons that pass the optimization step (249 for PvGTSeq and 4 for PvCRiSP). For each amplicon, the LOD was defined as the minimum parasite concentration needed to generate a microhaplotype with ≥ 10 reads with a 75% probability in the case of PvGTSeq, or 95% in the case of PvCRiSP. LOD at the level of the amplicon was calculated through logistic regression with a quasi-binomial distribution, where parasite concentration was the explanatory variable and success (1) or failure (0) of amplification in the samples was the response variable. The quasi-binomial distribution helped adjust for outliers caused by poor DNA quality or other factors that increase data dispersion not included in the model. Amplicons were classified as highly sensitive if they amplified samples with concentrations < 10 parasites/ μL , and only this set of amplicons was used for subsequent analysis.

The LOD at panel level was defined as the minimum parasite concentration needed to detect ≥ 10 reads per microhaplotype at 75% of all amplicons in the case of PvGTSeq and 100% of all amplicons in the case of PvCRiSP. In both cases, LOD was calculated through linear regression between the logarithm of parasite DNA concentration (explanatory variable) and the logarithm of the odds of amplified amplicons per sample (dependent variable).

2.3.8 Consistency across assays

The consistency of genotyping was determined through reproducibility of the microhaplotype calls between technical replicates of monoclonal infections within the same sequencing run (30 duplicate samples), between sequencing runs (68 duplicate samples), and between genotypes obtained from amplicon sequencing and genotypes extracted from WGS

VCF files (39 duplicate samples). Due to differences in filtering parameters for sequencing and alignment errors between GATK and our dada2-based pipeline, no quality filters were applied to the VCF files obtained from joint calling. Instead, VCF files were reconverted into FASTA sequences that were incorporated into our analysis pipeline as if they were reads obtained from amplicon sequencing. Thus, for each technical replicate we calculated the error rate of the sequencing, defined as the percentage of nucleotide differences between replicates with respect to the total number of base pairs across all amplicons.

2.3.9 Detection of major and minor clones in polyclonal infections

The ability of PvGTSeq and PvCRiSP to detect microhaplotypes from different clones that co-occur within the same infection at different concentrations was measured using mock polyclonal samples generated by combining DNA extracts from different geographical areas. Four samples were selected, two from Colombia and two from Guyana, with similar parasite concentrations. The selected clinical samples for generating mock polyclonal infections were the following: SP0112286092, G4GWM400, SP0112286092, and G4GWM400; henceforth we will refer to them as samples S1, S2, S3, and S4 respectively. Samples S1 and S3 were from Colombia and their concentrations were 198.25 and 203.05 parasites/ μ L, respectively. Samples S2 and S4 were from Guyana and their concentrations were 114.78 and 119.06 parasites/ μ L, respectively. When analyzing these samples individually through the amplicons from PvGTSeq, samples S1, S2, and S3 were monoclonal, while sample S4 was polyclonal, with a fraction of heterozygous amplicons equal to 0.22 and a maximum number of microhaplotypes per amplicon of 3. The proportion of loci that differed between the haplotypes of samples S1 and S2 was 0.52, while for samples S3 and S4 it was 0.48. The mock sample mixtures generated with samples S1 and S2 were named Mock Sample 1 (MS1), and those generated from samples S3 and S4 were

named MS2. Each combination was performed at seven different proportions between the samples (1:10, 1:5, 1:2, 1:1, 2:1, 5:1, and 10:1) and in duplicate, generating 28 mock samples in total. In all mock samples, the parasite concentration of the minority clone was always above the detection limit of the technique, so that the absence of detection should only be explained by competition between alleles during the amplification or sequencing steps.

To evaluate whether the amplicon panels could detect both major and minor clones within polyclonal infections, we calculated the percentage of clone-specific alleles detected (private alleles). This percentage was determined by dividing the number of detected private alleles by the total expected private alleles and multiplying by 100.

2.3.10 Diversity and utility of the amplicon panels

We subsequently evaluated the ability of both amplicon panels to discriminate whether two infections are identical by state or not in different geographic regions. This analysis only included populations, world regions, or countries with 20 or more monoclonal samples with coverage for at least 75% of the amplicons in the PvGTSeq panel. The genetic sequences of monoclonal samples were obtained from three different sources: 1) WGS from the MalariaGEN Pv4 database, 2) WGS from our laboratory, and 3) amplicon sequencing data from samples sequenced using the PvGTSeq amplicon panel. A monoclonal sample was defined as one that had an F_{ws} score greater than or equal to 0.975⁶⁷ and a fraction of heterozygous loci less than or equal to 0.05, with loci being defined as variant sites in the case of a VCF file for WGS samples, or as microhaplotypes in the case of samples sequenced by PvGTSeq. To maintain consistency in variant calling between WGS and PvGTSeq sequences, VCF files were converted into FASTA

sequences and incorporated into our analysis pipeline as if they were reads obtained from amplicon sequencing.

To determine the diversity of the amplicons in each geographic region, each amplicon was described in terms of the 1) number of polymorphic sites (S), 2) the number of microhaplotypes, and 3) the per-site nucleotide diversity. The nucleotide diversity, or π , was defined as the average number of nucleotide differences per site (each single nucleotide position in the amplicon) between two randomly selected sequences from the population. All these metrics were described independently for the set of amplicons used for geographic attribution at different administrative levels, for amplicons of PvCRiSP and for amplicons covering candidate drug resistance loci. We compared π between amplicon subsets using the Wilcoxon test for unpaired samples, and compared π between geographic regions using the Wilcoxon test for paired samples.

To directly measure the utility of PvGTSeq and PvCRiSP, we measured the capacity of these panels to distinguish between 1) parasites with distinct genomic sequences and 2) parasites that do not belong to the same clonal group. To assess discrimination of genomically distinct parasites, we used WGS data from genetically distinct monoclonal samples (we kept samples with pairwise identity by state or IBS < 0.999 or $> \sim 10$ SNPs difference) and samples sequenced only by PvGTSeq. For samples sequenced exclusively by PvGTSeq, we assumed all haplotypes were genomically distinct since they originated from different infected patients. For assessing whether parasites do not belong to the same clonal group, we relied solely on WGS information, defining a clonal group as haplotypes with genetic similarity (IBS) ≥ 0.99 or < 100 SNPs difference in pairwise comparisons. For both scenarios and both amplicon panels, the discriminatory power D (due to its similarity with Simpson's index) was quantified as the proportion of in silico pairwise sample comparisons from the same population yielding IBS

values less than one. IBS was calculated as the proportion of microhaplotypes shared between sample pairs across all amplicons in each panel, or as the proportion of biallelic SNPs shared when comparing WGS information. D was reported for each country, and we selected country as the unit of analysis because it is the smallest geographic unit with sufficient sample size across our dataset, and it is also the context in which most future studies will use these two panels.

2.3.11 Population Structure and Geographic attribution

We evaluated the ability of PvGTSeq to detect population structure at three geographic levels (global, intra-continental, and sub-national). PvGTSeq data was generated by applying our amplicon panel to clinical samples or by extracting PvGTSeq coordinates from whole genome sequencing data—either from the MalariaGEN Pv4 database or from our laboratory. At the global and intra-continental levels, population structure was assessed through PCoA, while at the sub-national level, a network analysis was used instead. For both approaches, pairwise genetic similarity between infections was determined using IBS. To avoid introduction of bias due to directional selection,⁶⁸ amplicons from candidate drug resistance loci, that in most cases also showed low diversity, were excluded from this analysis. We performed Tajima's D test on each amplicon in each country to confirm that none were under directional selection pressure. To prevent clonal expansions from skewing the results, we selected only one haplotype from each clonal group ($IBS > 0.99$) in each population. We conducted the analysis only in countries with more than 20 non-clonal haplotypes. We excluded amplicons under strong directional selection pressure (Tajima's $D < -2$) from the population structure analysis. However, we reported but did not exclude amplicons under strong balancing selection pressure (Tajima's $D > 2$), because this study does not aim to infer historical genealogical relationships among populations—the focus is to observe whether population differentiation exists.

2.4 Results

2.4.1 Selection of amplicons and protocol optimization

We identified 60,056 variant sites meeting our inclusion criteria: located in coding regions not prone to alignment errors, not in homopolymers or di-nucleotide short tandem repeats. A total of 1,292 genomes had coverage $\geq 75\%$ supported by ≥ 5 reads for these variant sites (**Supplementary figure 2.1**). Using DAPC, we identified 263 segments (150 bp each) contributing to population differentiation across three geographic scales (**Supplementary figure 2.1**). We also identified 57 non-overlapping segments containing polymorphisms in 11 genes previously associated with antimalarial resistance. From this pool of 320 segments, we successfully designed primers for 271, which underwent successive rounds of protocol optimization. Of these, 249 amplicons amplified efficiently and specifically in the 60 clinical samples used for the standardization step. The final protocol for library generation and sequencing is detailed in the S1 Text and S2 Text, and the complete list of 249 amplicons, including primer information and reference genome coordinates, is provided in the **S1 Table**. The finalized panel consists of 213 amplicons specialized for geographic differentiation (112 for global-scale, 122 for between-country, and 110 for subnational differentiation in Latin America). Note that many of the amplicons nevertheless function at multiple geographic scales.

While we did not evaluate the minimum number of amplicons needed for genetic differentiation between populations, we did assess whether our panel contains enough amplicons to distinguish samples that are genetically different at genome level. In addition, PCoA showed that the 213 amplicons in PvGTSeq reproduces the differentiation patterns observed with WGS and outperforms the other two amplicon panels at both geographic scales analyzed: world

regions and countries in Latin America and the Caribbean (**Supplementary figure 2.3 and 2.4**). PvGTSeq successfully distinguished the eight world regions. Within LAC, PvGTSeq distinguished samples from Mesoamerican countries, the Colombian Pacific coast, and the Amazon. However, differentiation between the Amazon and the Guiana Shield was not clear, likely due to the smaller numbers of samples from Venezuela (9) and Guyana (15) compared with other populations (> 20 samples per country). All three amplicon panels showed high accuracy and precision when WGS-based pairwise IBD was close to one (RMSE and CV < 0.1) (**Supplementary figure 2.5**). In contrast, when IBD was below 0.5, accuracy and precision declined, with a clear tendency to overestimate IBD relative to WGS (**Supplementary figure 2.1A**). Among the three methods, PvGTSeq was the most accurate: RMSE was below 0.1 for IBD values above 0.45 and approximately 0.16 when IBD was in the range 0.15–0.45. In this range (0.15 - 0.45), PvGTSeq RMSE was significantly lower than PvAmpliSeq and rhAmpSeq (paired Student's t-test, p -value < 0.001) (**Supplementary figure 2.5B**). In contrast, rhAmpSeq was the most precise. rhAmpSeq and PvGTSeq maintained a CV below 0.1 for IBD values above 0.35; however, at lower IBD values, PvGTSeq CV was higher than rhAmpSeq CV (**Supplementary figure 2.5C**). Because of differences in IBD estimates between the amplicon panel and WGS, we opted to use IBS for subsequent analyses.

The finalized panel additionally targets 36 amplicons within 10 genes associated with antimalarial resistance. The segment identified for the *CRT* gene failed at the primer design stage. For the DHFR gene, we included only the segment spanning amino acids 6–24; other polymorphic regions in this gene are not covered by the final panel. As a result, the previously reported mutations 50I, 57L, 58R, 61M, 117N, and 173L cannot be monitored with this panel.^{42,43,45,69–72} For DHPS, the panel includes the 383G mutation and mutations located

between amino acids 28–43 and 622–650; however, the 553G mutation fall outside the regions covered by our amplicons.^{43,57,72} For *MDR1*, the PvGTSeq panel includes the 976F and 1076L polymorphisms, which have been proposed as markers of chloroquine resistance.^{47,49,72–77} The number of amplicons for each candidate antimalarial resistance gene, and the coding regions covered by those amplicons are shown in **Table 2.1**. The haplotype frequencies of these 10 genes are shown in **Supplementary figure 2.6**, and we will interpret the temporal and geographic distribution of these haplotypes in a subsequent publication. Finally, we selected four highly polymorphic amplicons from this panel that were heterozygous in most polyclonal samples and showed high amplification efficiency. These four amplicons—CG2_related, **RIPR**, **VPS11**, and **PIGM**—form the PvCRiSP panel. All amplicons are named according to their corresponding genes.

Table 2.1: Nucleotides and amino acids in antimalarial resistance candidate genes covered by PvGTSeq.				
Gene ID	Gene name	Number of amplicons	Covered nucleotides in the coding sequence	Covered amino acids
PVP01_0203000	<i>MRP1</i>	7	755-862, 1930-1983, 2704-2862, 3032-3202, 3673-3853, 4573-4635, 4751-4843	252-288, 644-661, 902-954, 1011-1068, 1225-1285, 1525-1545, 1584-1615
PVP01_0312700	<i>DMT2</i>	1	655-663	219-221
PVP01_0526600	<i>DHFR</i>	1	17-72	6-24

Table 2.1 (Continued)

PVP01_1010900	<i>MDR1</i>	5	642-719, 1493-1607, 2698-2750, 2917-3091, 3203-3249	214-240, 498- 536, 900-917, 973-1031, 1068- 1083
PVP01_1018600	<i>PI3K</i>	10	231-354, 551-602, 879-1043, 1270-1325, 2396-2525, 2649-2796, 3350-3407, 3535-3578, 4349-4449, 4695-4745	77-118, 184-201, 293-348, 424- 442, 799-842, 883-932, 1117- 1136, 1179-1193, 1450-1483, 1565- 1582
PVP01_1103800	<i>ABC-E1</i>	2	50-106, 1420-1475	17-36, 474-492
PVP01_1211100	<i>K13</i>	1	1838-1900	613-634
PVP01_1259100	<i>MDR2</i>	4	907-994, 1394-1565, 2525-2578, 3226-3288	303-332, 465- 522, 842-860, 1076-1096
PVP01_1429500	<i>DHPS</i>	3	84-127, 1115-1172, 1865-1948	28-43, 372-391, 622-650
PVP01_1447300	<i>MRP2</i>	2	176-280, 433-547	59-94, 145-183

2.4.2 Genotyping performance

The amplification rate of the amplicons and the limit of detection varied among the three populations analyzed with PvGTSeq due to sample quality, but PvCRiSP yielded consistently

high performance across populations. Regarding PvGTSeq and using the MiSeq Illumina instrument, in Colombia, 190 out of the 249 amplicons amplified in more than 75% of the samples (**Supplementary figure 2.7A-B**), with a median read depth of 95.5 (interquartile range or IQR = 409) across the 249 amplicons. The median LOD of these amplicons analyzed individually was 4.64 (IQR = 14.6) (**Supplementary figure 2.7C-D**), and the minimum parasite concentration (LOD) to amplify 75% of the total 249 amplicons was 4.05 parasites/ μ L (95% CI from 2.5 to 5.9) (**Figure 2.1**). When considering only the amplicons that worked in more than 75% of the samples in Colombia, the LOD to amplify 95% of these 190 amplicons was 3.24 parasites/ μ L (95% CI from 1.18 to 6.15). A similar result was obtained with samples from Guyana, where 201 amplicons amplified in more than 75% of the processed samples; however, parasite concentrations were above 5.73 parasites/ μ L, and LOD could not be determined either at the individual amplicon level or for the complete panel. Furthermore, there were 201 amplicons that had amplification rates above 75% in Colombia or Guyana. In contrast, Venezuelan samples showed poor amplification success, where only 21 amplicons were amplified in more than 75% of the samples.

We next investigated reproducibility, ability to detect minor alleles in polyclonal infections and population structure signals using the group of 201 amplicons that showed over 75% amplification success in Colombia or Guyana to avoid bias from missing data.

The PvCRiSP mini multiplex exhibited similar amplification success and LOD among the three populations using the iSeq Illumina instrument. The four amplicons amplified on average more than 86% of the samples and had a median read depth of 131 (IQR = 446). The amplicon *CG2_{related}* showed the lowest amplification rate (77%) and the lowest read depth (median = 45.5, IQR = 127.25) across the three populations. The minimum concentration needed to amplify

the four amplicons was 6.33 parasites/ μ L (95% CI from 4.92 to 8.23) and no significant differences were observed between the three countries when comparing their confidence intervals.

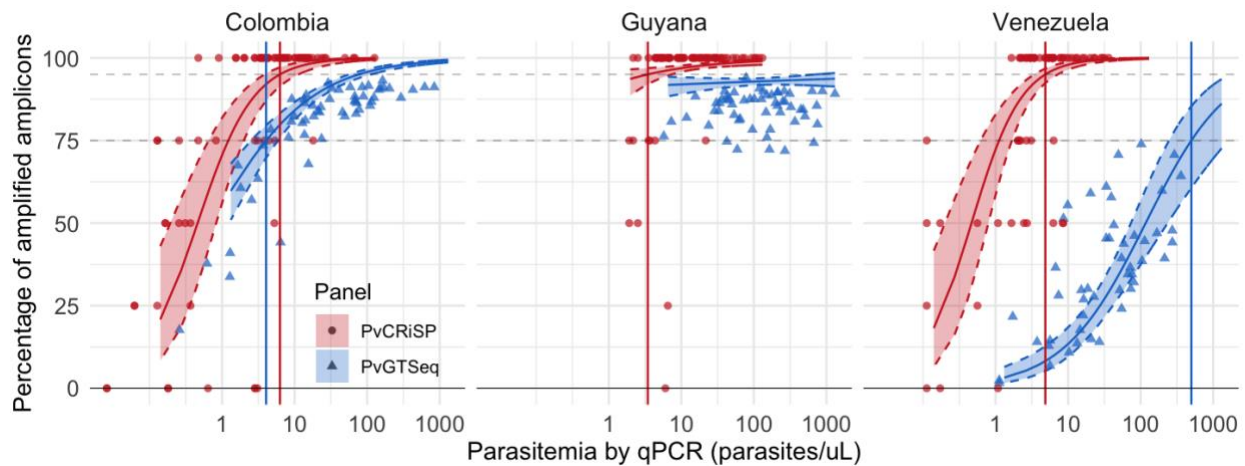


Figure 2.1: Limit of detection of PvGTSeq and PvCRiSP in Colombia, Guyana and Venezuela. The percentage of amplified amplicons per sample as a function of parasite concentration is illustrated for PvGTSeq (Blue) and PvCRiSP (Red). Each dot represents an individual sample and regression model (solid curve line) was constructed using the logarithm of parasite concentration (independent variable) and the logarithm of the odds of the amplified amplicons per sample (dependent variable) in each country. Vertical lines represent the minimum concentration a sample must have to amplify 75% of the 249 amplicons in PvGTSeq (blue lines) or the four amplicons of PvCRiSP (red lines). The shaded area between the dashed curve lines represents the 95% confidence interval.

2.4.3 Consistency across sample replicates

The results showed high reproducibility between technical replicates of monoclonal infections. Within the same sequencing run, of the 30 technical replicates analyzed, only in three cases were there differences between any of the amplicons between replicates. In these three cases, the difference was only one nucleotide between replicates. In two cases, the discrepancies were found in a marker against the gene *RAMA* (PVP01_0107500). In both of these two cases one of the replicates had a haplotype similar to the reference strain PVP01 while the other replicate contained only one non-synonymous polymorphism with respect to the reference. These

two polymorphisms were present in 58 and 43 samples respectively in our group of samples used for the estimation of LOD, an indication of cross-contamination. The third discrepancy was a singleton polymorphism in our full data set of sequenced samples that could be caused by a PCR or sequencing error. The error rate including only the genotyping error was 3.85e-4% (phred Q34; permutation test, 95% CI 7.93e-5% to 1.12e-3%).

When comparing replicates between different sequencing runs, we observed that 16 out of the 68 replicates differ in at least one amplicon. Most discrepant cases consisted of a single nucleotide polymorphism (11 cases). Other discrepant cases included the occurrence of two single nucleotide variants (SNVs) in a single amplicon (1 case), one INDEL plus one SNV in a single amplicon (2 cases), one INDEL and one SNV in two different amplicons (1 case), and one INDEL plus three SNVs in a single amplicon (1 case). Moreover, in five cases the discrepancy was in an amplicon that was heterozygous in one of the two replicates. Based on these observed discrepancies between replicates that were not likely caused by contamination the error rate was 1.42e-3% (phred Q28.5; 95% CI from 8.53e-4% to 2.21e-3%).

PvGTSeq panel replicates showed greater differences when compared with WGS replicates (**Supplementary figure 2.8**). Thirty-two out of 39 replicates showed differences in one or more amplicons and 30 out of them were due to an INDEL. Most of these INDELs occur in amplicons for seven genes: PVP01_0109700, PVP01_0304700 (LISP2), PVP01_0530700, PVP01_1018600 (PI3K), PVP01_1240400 (ApiAP2), PVP01_1421800, PVP01_1465800. Our amplicons are approximately 150 base pairs, and our sequencing uses 300 total cycles. This allows complete coverage and overlap of the amplicon by both reads generated. Therefore, these discrepancies were explained by inconsistent representation of INDELs in the variant calling

process of WGS data. Excluding amplicons against these seven genes, the error rate was 2.87e-3% (phred Q25.4; 95% CI from 1.75e-3% to 4.44e-3%).

2.4.4 Detection of major and minor clones in polyclonal infections

Next, we tested the utility of the 201 highly sensitive amplicons from PvGTSeq and the four highly sensitive amplicons of PvCRiSP for detecting minor alleles in mock polyclonal infections. The selected clinical samples were named S1, S2, S3, and S4. We mixed S1 and S2 to generate mock samples at seven different proportions between the samples (1:10, 1:5, 1:2, 1:1, 2:1, 5:1, and 10:1) in duplicates, and we labeled those mixtures as Mock Sample 1 (MS1). The mixtures generated from samples S3 and S4 were labeled as MS2. Additionally, the combinations were performed at a total parasitemia of 100.33 parasites/ μ L for MS1 and 58.46 parasites/ μ L for MS2 such that the concentration of the minority clone was always above 5.85 parasites/ μ L, therefore the absence of amplification of the clone is expected to reflect competition between alleles in PCR and/or sequencing, not the clone occurring below panel detection limits.

The PvGTSeq and PvCRiSP panels identified all mock samples as polyclonal, including those samples in which the major clone was in a 10:1 ratio compared to the minor clone (**Figure 2.2**). In PvGTSeq, the percentage of private alleles detected was >76% whenever the major clone was in a proportion of 1:1, 2:1, 5:1, or 10:1 relative to the other clone present in the sample. However, for the minor clone, the detection rate of its private alleles decreased linearly with respect to its proportion in the sample (**Figure 2.2A**). Thus, only when both clones were in similar proportions in the sample was the observed fraction of heterozygous loci close to the expectation (75%) (**Supplementary figure 2.9**), and as the ratio between the major clone to the

minor clone increased, the ratio between the observed and expected fraction of heterozygous loci decreased, reaching only 30% when the major:minor clone was in a 10:1 ratio. Additionally, the read depth of the major clone was consistently higher than the read depth of the minor clone across all analyzed amplicons and all technical and biological replicates (**Supplementary figure 2.10**). Thus, a positive correlation was observed between the median ratio of read depth of private alleles between major and minor clones with respect to the ratio of DNA input concentrations per clone (**Supplementary figure 2.9C**). In the case of PvCRiSP, the detection rate of private alleles was always above 67%, even when the major:minor clone ratio was 10:1 (**Figure 2.2B**). Therefore, the ratio between observed and expected fraction of heterozygous loci was always one regardless of the proportion of clones within the mock sample (**Supplementary figure 2.11B**). This result is similar to that obtained with the 4CAST panel designed for *P. falciparum*, where the sensitivity to detect alleles from the minor clone was 94% even with ratios of 10:1.¹¹

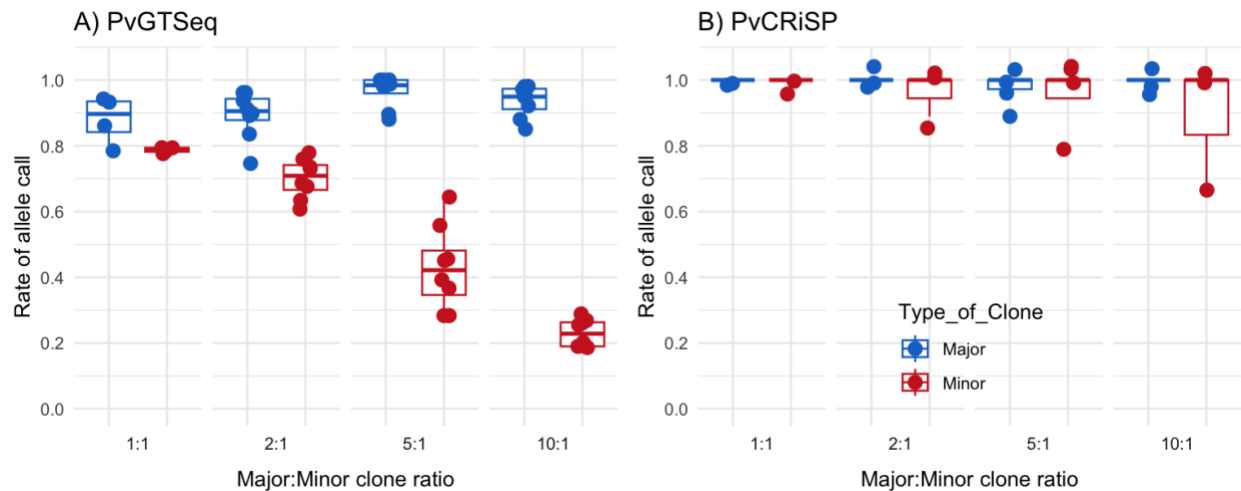


Figure 2.2: Detectability of alleles from major (blue) and minor (red) clones in mock samples using PvGTSeq (A) and PvCRiSP (B) panels. The X-axis represents the ratio of the concentration of the major clone (numerator) with respect to the minor clone (denominator) in the mock sample. The Y-axis represents the proportion of private alleles of the clone of interest detected (numerator) with respect to the total number of private alleles expected (denominator). Dots represent each clone of interest in the 28 mock samples.

2.4.5 Diversity and utility of the amplicon panels

Genetic diversity varied across amplicon subsets and geographic areas at different spatial scales. Globally, from the 201 highly sensitive amplicons in PvGTSeq, the median number of polymorphic sites (S, which includes SNVs and INDELs) and microhaplotypes per amplicon was 7 (IQR = 6) and 9 (IQR = 7), respectively. The nucleotide diversity (π) was 4.38e-03 (IQR = 5.13e-03) (**Figure 2.3**). Among the four PvGTSeq amplicon subsets (differentiation at three geographical scales and genes associated with antimalarial resistance), those selected for their discrimination power between world regions and within Latin America and Caribbean countries showed the highest diversity (Wilcoxon Rank-Sum p-value < 0.005 with respect to the other subsets) with no significant differences (p-value > 0.005) observed between these two subsets (**Supplementary figure 2.11**). The sub-set of amplicons for differentiation between world regions showed a median 9 microhaplotypes (IQR = 7), and a median nucleotide diversity of 4.95e-03 (IQR = 5.04e-03) while in the sub-set of amplicons for differentiation within Latin America and Caribbean countries, the median number of microhaplotypes and nucleotide diversity were 11 (IQR = 12.5) and 5.61e-03 (IQR = 5.58e-03), respectively. Conversely, amplicons targeting antimalarial resistance genes showed significantly lower diversity compared to the other subsets (p-value < 0.05, S = 3, number of microhaplotypes = 4, π = 2.8e-03) (**Figure 2.3**), except for two amplicons, pvpi3k_7 and pvpi3k_13, targeting the *PI3K* gene (PVP01_1018600) which exhibited more than 30 microhaplotypes due to short tandem repeats in their sequences. While only the amplicon pvpi3k_1 was monomorphic across all global monoclonal samples, it showed polymorphism in three polyclonal samples from Cambodia, Thailand and Myanmar. Regarding PvCRiSP, it showed more than six polymorphic sites

(median = 8.5, IQR = 2.75) and eight microhaplotypes per amplicon (median = 13, IQR = 17.25), with global nucleotide diversity above 3.15×10^{-3} (median = 1.06×10^{-2} , IQR = 6.55×10^{-3}).

Across the seven global regions analyzed (East Asia was excluded due to its low number of samples), the median nucleotide diversity of the 169 amplicons for geographic differentiation of the PvGTSeq panel exceeded 3.01×10^{-3} in Africa (AF), Eastern Southeast Asia (ESEA), Latin America and Caribbean (LAC), Western Southeast Asia (WSEA) and Western Asia (WAS). Significant inter-regional differences (Wilcoxon Signed-Rank p-value < 0.005) were observed except between AF and WSEA, ESEA and WAS, and WSEA and WAS (**Figure 2.3**). Maritime Southeast Asia (MSEA) and Oceania (OCE) showed lower nucleotide diversity (1.75×10^{-3} and 1.7×10^{-3} , respectively, p-value < 0.005), with up to 73 (36%) monomorphic amplicons in these regions. This pattern persisted across various PvGTSeq panel subsets (use groups) as well as in the PvCRiSP panel, and the amplicon against the gene *PIGM* in PvCRiSP was monomorphic in OCE. This amplicon was also monomorphic in ESEA. At the country level, Brazil showed the highest nucleotide diversity ($\pi = 5.01 \times 10^{-3}$, p-value < 0.005) while Malaysia showed the lowest ($\pi = 5.72 \times 10^{-4}$, p-value < 0.005; **Supplementary figure 2.12**). Indonesia and Papua New Guinea were excluded from the country level analysis due to insufficient monoclonal samples, but they were kept at the world region level analysis.

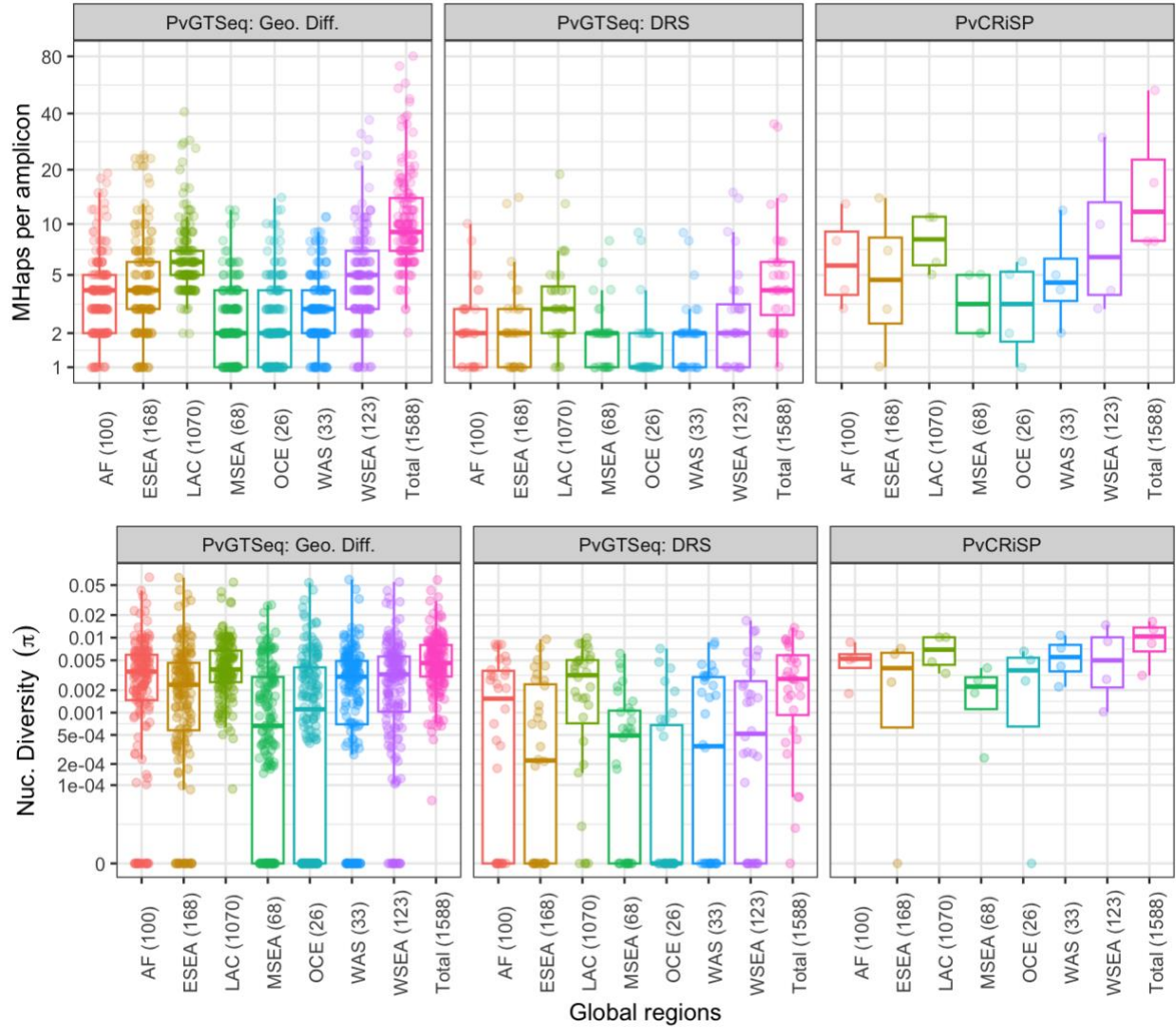


Figure 2.3: Number of microhaplotypes (MHaps) and nucleotide diversity (π) per amplicon use groups and across seven global regions: Africa (AF), Eastern Southeast Asia (ESEA), Latin America and Caribbean (LAC), Maritime Southeast Asia (MSEA), Oceania (OCE), Western Asia (WAS), and Western Southeast Asia (WSEA). Numbers within parenthesis indicate the number of monoclonal clinical samples. Each dot represents an amplicon. Use case groups were defined as 169 amplicons for geographic differentiation, 32 amplicons for drug resistance surveillance (DRS), and the four amplicons comprising PvCRiSP.

To directly measure the utility of PvGTSeq and PvCRiSP in each country, we defined the discriminatory power of the panels as the ability of: 1) differentiating parasites given that they have different genomic sequences (**Supplementary figure 2.13**), and 2) differentiating parasites given that they do not belong to the same clonal group (**Figure 2.4**). Regarding the ability of the

panels to differentiate parasites with different genomic sequences, for PvGTSeq, it was found that in 15 of the 16 analyzed countries, D was ≥ 0.95 , but lower in Panama (0.767, 95% CI from 0.748 to 0.785) (**Supplementary figure 2.3**), a country where transmission is dominated by two highly clonal groups (**Figure 2.5D**). For PvCRiSP, D was ≥ 0.95 in seven countries, between 0.9 and 0.95 in four countries, and below 0.9 in five countries (Ethiopia, Mexico, Malaysia, Panama and Papua New Guinea). In both panels, D was influenced by the distribution of IBS of the pairwise comparisons in the population (Pearson correlation p -value < 0.001 , $\text{adj-}R^2 = 0.71$).

Because D is associated with the IBS distribution in the geographic unit of interest (diversity and degree of clonality of the population), we redefined D as the probability that two random samples have different genotypes given that these two infections do not belong to the same clonal group. Under this definition, PvGTSeq demonstrated high discriminatory power in all analyzed countries. D was 1.0 for all countries except Honduras, where it was 0.986. For PvCRiSP, D was above 0.8 for all countries, but remained limited in Malaysia (0.786) and Papua New Guinea (0.523). In these two countries, the amplicon targeting *PIGM* was monomorphic and the amplicon targeting *VPS11* has only 2 alleles, reducing the ability of PvCRiSP to discriminate between genomically similar samples or between samples that come from endemic areas with low population diversity.

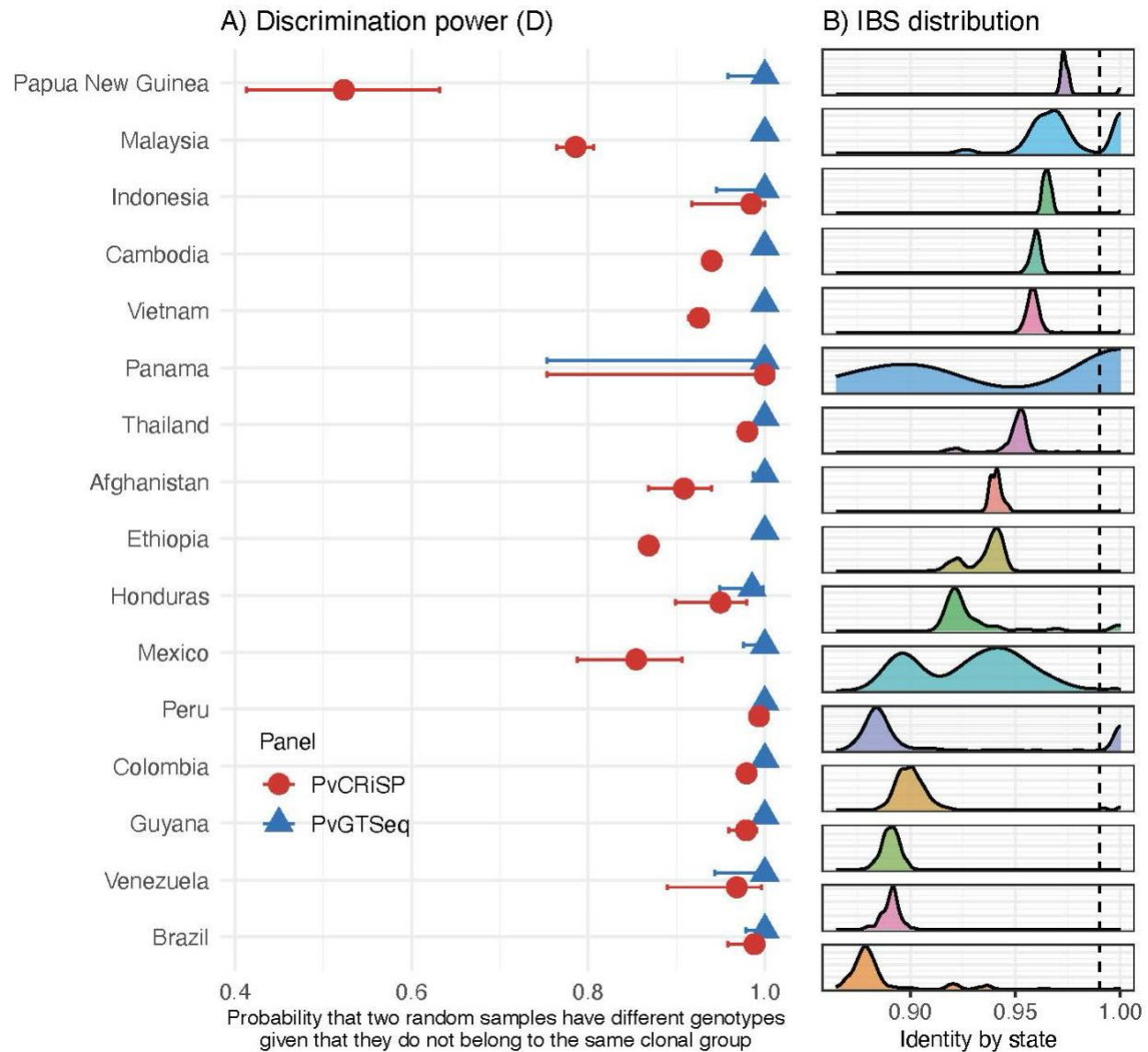


Figure 2.4: Panel discrimination power. The figure A on the left shows the probability that two random samples have a different genotype (differ in at least one amplicon) given that they do not belong to the same clonal group (x-axis) across global *P. vivax* populations (y-axis) using PvGTSeq (blue) and PvCRiSP (red). The figure B) on the right shows the distribution of identity by state within each population measured from whole genome sequencing data and the vertical dashed line indicates the threshold used to define clonal groups with whole genome information ($IBS \geq 0.99$).

2.4.6 Population structure and geographic attribution

None of the amplicons used for the population structure analysis showed Tajima's D values below -2, and only the amplicon targeting gene PVP01_0834800 showed a value above 2 in

Guyana (**Supplementary figure 2.14**). Across all countries, amplicons targeting *MRP1* (PVP01_0203000), *MDR1* (PVP01_1010900), *PI3K* (PVP01_1018600), *ABC-E1* (PVP01_1103800), *K13* (PVP01_1211100), and *DHPS* (PVP01_1429500) had Tajima's D values below -1. Genetic differentiation using a PCoA and a network method revealed that genetic differences align with administrative boundaries at three geographical levels: global, intra-continental, and subnational. The global PCoA divided samples into three major groups: the first composed of LAC, the second including AF and WAS, and the third comprising ESEA, MSEA, OCE and WSEA. Each of these groups showed internal differentiation patterns (**Figure 2.5A** and **Supplementary figure 2.15**). In Latin America and the Caribbean, PCoA analysis revealed that the first principal coordinate, explaining 43% of genetic differentiation, separates parasites from the Guiana shield region (Eastern Colombia, Guyana, and Venezuela) from those in the Colombian Pacific and Central America (**Figure 2.5C**). The second coordinate distinguished parasites from Colombia from those in Honduras and Panama. Network analysis of samples from Colombia, Honduras, and Panama further reveals distinct clonal groups specific to each sub-national region (**Figure 2.5D** and **Supplementary figure 2.16**). Finally, the sequencing method used (amplicon or WGS) did not cause large batch effects (**Supplementary figure 2.17**).

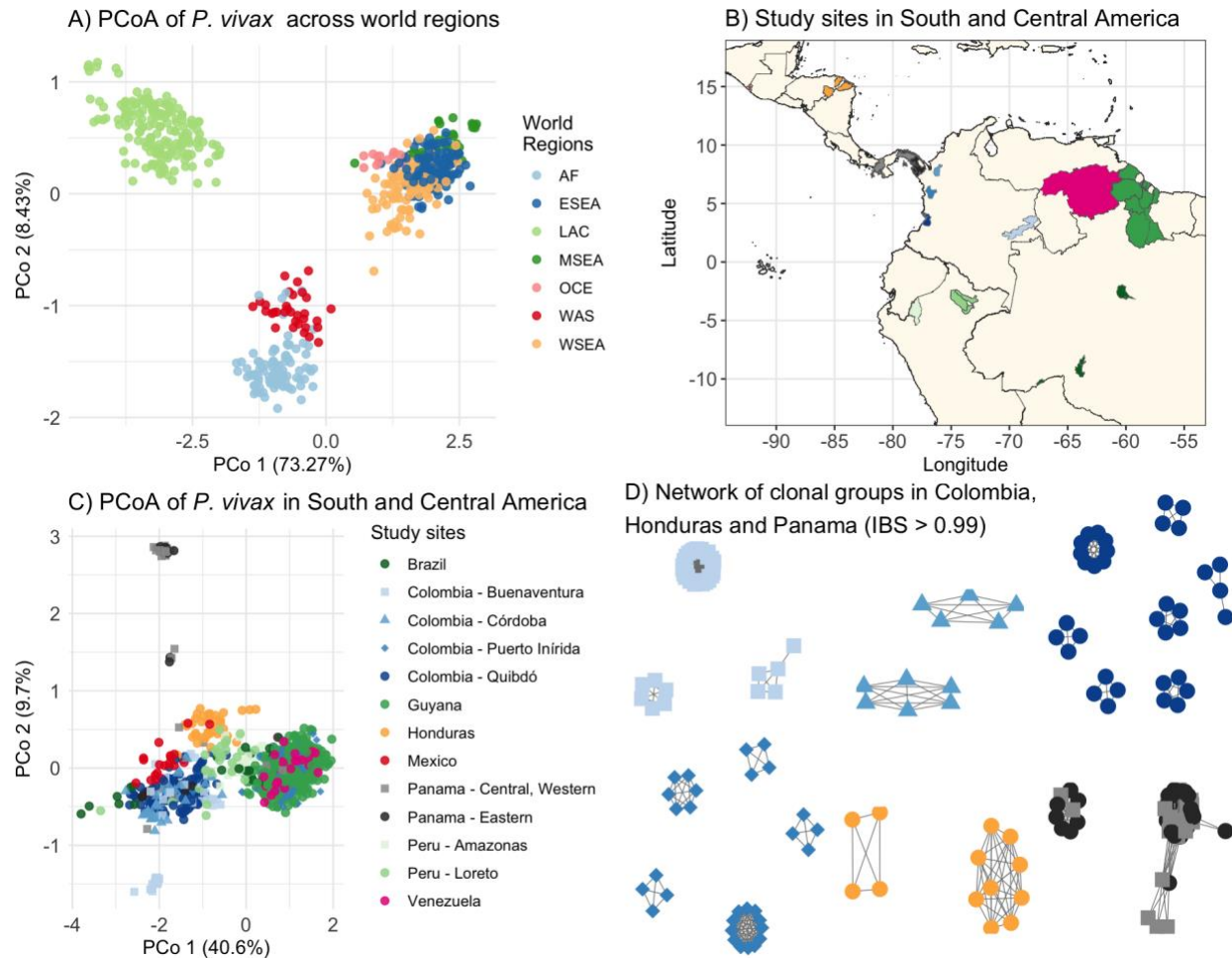


Figure 2.5: Population subdivision with PvGTSeq. A) Principal coordinate analysis (PCoA) of *P. vivax* samples from seven global regions: Africa (AF), Eastern Southeast Asia (ESEA), Latin America and Caribbean (LAC), Maritime South-East Asia (MSEA), Oceania (OCE), Western Asia (WAS), and Western South-East Asia (WSEA). LAC samples were sequenced either by WGS or the PvGTSeq amplicon panel or were obtained from the MalariaGEN Pv4 database. Genetic information for other regions was extracted from the MalariaGEN Pv4 database. Each point represents a monoclonal sample, with colors indicating geographical origin. The x- and y-axes represent principal coordinates 1 and 2, respectively. B) *P. vivax* sampling areas across South and Central America for genetic characterization using PvGTSeq. Colored regions show the administrative areas where samples were collected—first subnational level for Guyana, Mexico and Venezuela, and second subnational level for Brazil, Colombia, Honduras, and Peru. The map was produced using ggplot2 in R,^{78,79} and shape files were obtained from the Database of Global Administrative Areas (<https://gadm.org/>), under CC BY 4.0 using R package GADMTTools.^{79,80} C) PCoA of *P. vivax* populations in South and Central America. Points represent monoclonal samples, with colors and shapes indicating country of origin, except for Colombia, where colors represent municipality (second subnational level). The x- and y-axes represent principal coordinates 1 and 2, respectively. D) Network graph showing genetic relationships (IBS) among samples within clonal groups (IBS > 0.99) in Colombia, Honduras, and Panama. Each node represents a monoclonal sample, with edges indicating genetic relationships exceeding 0.99 IBS. Figures B, C and D share the same color scheme.

2.5 Discussion

We describe two new amplicon-based targeted sequencing panels for *P. vivax*, PvGTSeq and PvCRiSP. PvGTSeq is a multipurpose panel that requires sWGA pre-amplification and contains 201/249 highly sensitive amplicons for monitoring variants of interest in nine genes associated with antimalarial resistance (32/36 amplicons) and identifying population structure (169/213 amplicons), suitable for use in any global region but particularly optimized for Latin America. All amplicons are multiplexed in a single PCR reaction using non-proprietary reagents. With sWGA, PvGTSeq offers high sensitivity and amplifies $\geq 75\%$ of amplicons in samples with parasitemias as low as four parasites/ μL . The PvCRiSP panel consists of only four highly polymorphic amplicons and works with parasitemias ≥ 5 parasites/ μL without sWGA, enabling characterization of samples with compromised DNA integrity. PvCRiSP is tailored towards estimating COI and distinguishing between infections caused by distinct clonal groups, information useful for characterizing outbreaks. We validated both panels directly with samples from five Latin American countries (Colombia, Guyana, Honduras, Panama, and Venezuela) and *in silico* across 16 countries worldwide. This information serves as a reference for potential users to evaluate the utility of these panels in their endemic areas of interest.

The cost per sample of PvGTSeq library preparation and sequencing is \$23.64 with 186 samples per MiSeq v2 kit 300 cycles (median of >200 reads/amplicon/sample), plus \$15.99 per sample for sWGA using the Genomiphi v2 kit (Cytiva). In contrast, PvCRiSP costs \$19.04 per sample due to its simplified design. The number of samples that can be multiplexed in these two panels depends on the Illumina instrument, the i5 and i7 indexes, and the desired average read depth per amplicon (e.g. >200 reads). For the PvGTSeq, in the iSeq we can multiplex up to 96 samples including controls with a read depth above 200 reads per amplicon per sample. For the

MiSeq, we can multiplex up to 192 samples and controls. When using a HiSeq or NovaSeq instrument, the number of samples depends on the i5 and i7 indexes. We use up to 1,152 unique combinations of these indexes—the maximum number of samples (including controls) we can multiplex in a single sequencing run.

These panels contribute to the expanding repertoire of targeted next-generation sequencing approaches for *P. vivax* characterization, all of which have strengths and weaknesses for various use cases. The most common applications of targeted sequencing include analyzing population structure, determining geographic origin, tracking resistance variants, measuring COI, determining pairwise relatedness, and distinguishing between new infections, relapses, and recrudescences.^{14,15,17–20,81,82} Spatial resolution depends on the genomic dataset used for target selection and the criteria applied to maximize resolution per area of interest. While MalariaGEN Pv4 data has enabled new panel development, some regions remain underrepresented—particularly Latin America—causing variable panel performance across regions. PvAmpliSeq, the first validated amplicon sequencing panel, was designed to identify patterns of population structure between global regions and within Vietnam¹⁵ but showed poor discrimination in South America, leading to a second version with 41 additional SNPs for Peruvian samples. A recent MIP-based panel targeting ~1200 SNPs tested with Peruvian samples offers deep profiling of samples but struggles with parasitemias below 100 parasites/μL.¹⁹ Others have grouped markers from multiple panels to characterize autochthonous *P. vivax* cases that occurred in the United States in 2023.⁸³ While previous panels were designed to characterize SNPs, two additional panels were developed to amplify microhaplotypes (multiple polymorphic sites within a single amplicon), though these had constraints with low parasitemia samples.^{21,82} In creating our new PvGTSeq and PvCRiSP panels, we incorporated extensive WGS data from MalariaGEN Pv4

(genomes from Brazil, Colombia, Mexico, and Peru) and we additionally leveraged in-house WGS representing Colombia, Guyana, Peru, and Venezuela. Then we validated both panels with performed samples from Colombia, Guyana, Honduras, Panama and Venezuela. Additionally, we coupled PvGTSeq library preparation with sWGA. As a result, these two panels are the most sensitive and extensively evaluated in the Americas.

Distinguishing between new infections, relapses, and recrudescences requires high sensitivity to detect alleles from minority clones in polyclonal infections. In this context, panels with large numbers of multiplexed amplicons present a significant disadvantage.^{20,81} Our results show that while PvGTSeq can correctly identify whether an infection is monoclonal or polyclonal, alleles from the minority clone may remain undetected up to 70% of the time when the proportion between clones is 10:1. Although reducing the number of samples per sequencing run could increase read depth and minimize allele loss from minority clones, this approach would raise costs, making this technology prohibitively expensive. This challenge is effectively addressed by using panels with fewer amplicons, such as PvCRiSP, which offers high sensitivity, efficiently detects all alleles present in polyclonal samples without requiring sWGA, and offers low procedural complexity.

As amplicon sequencing becomes more widespread for generating genetic data for malaria parasites and other pathogens, there is a need for versatile analysis tools that work with any amplicon panel. Here we used a dada2-based analysis pipeline for denoising PCR errors and achieving sensitive minor allele detection in polyclonal infections. We validated our error-filtering approach using technical replicates at three levels: within runs, between runs, and against WGS. This multi-level validation achieved lower error rates ($3.85 \times 10^{-4}\%$ within sequencing runs and $2.86 \times 10^{-3}\%$ when comparing amplicon sequencing to WGS) than comparable

studies (from 0.008% to 0.02%)¹¹ and demonstrated that INDELs do not reduce reproducibility in amplicon sequencing. This indicates that a genotyping error will occur approximately once every 142,000 base pairs or one error every five genotyped samples with the full set of 249 amplicons—a rate comparable to non-proofreading polymerases. Because 36 of the amplicons target genes associated with antimalarial resistance, approximately one in every 50 analyzed samples will have a random error in one of these amplicons. These errors can be filtered out by establishing allele frequency thresholds at the population level (e.g. discarding singletons) or by using technical replicates in therapeutic efficacy studies with few samples. This panel-agnostic pipeline is available on GitHub and embedded in a Terra workspace—a cloud-based platform designed for users with limited bioinformatics expertise. A comprehensive description of this analysis pipeline and all its functionalities will be presented in a subsequent publication.

Despite their significant advantages, both PvGTSeq and PvCRiSP have limitations. These include ascertainment bias inherent in the design of any targeted sequencing panel, limited overlap with existing panels, PvGTSeq allelic dropout in polyclonal samples with skewed strain ratios, and lack of representation of some antimalarial resistance markers. However, despite these limitations, PCoA and network analysis confirm that PvGTSeq can identify population structure on a global scale as well as detect clonal expansions at subnational levels in Colombia and Panama, where persistent clonal groups have been previously documented using WGS.^{84,85}

While in this work we report the diversity, neutrality and the discriminatory power of our amplicons in up to 16 countries, it is important to note that these analyses were performed with genetic information from MalariaGEN Pv4 or clinical samples provided by our collaborators. Because sampling intensity, strategy, and sequencing criteria may vary among these datasets, we limit our interpretations to the technical applicability of these tools—specifically, whether the

amplicon panel is diverse enough to characterize a specific population—and not to comparisons of parasite diversity among populations. We recommend that the use of these or other amplicon panels must be evaluated on a case-by-case basis in future studies. We do not recommend using this panel, or any targeted genotyping panel, for selection screening or demographic inference due to ascertainment bias. Most amplicons were designed to maximize variant sites using data that may not reflect diversity where the panel is implemented. Whole genome sequencing remains the best method for population genetic inference. This highlights the importance of generating and maintaining representative genomic databases over time. Such databases will help to evaluate existing panels based on the context they are intended to be used and guide panel redesign when necessary. To achieve this, it is essential to define representative sampling for different endemic areas and establish criteria for how frequently this information should be generated.

PvGTSeq does not include amplicons targeting the *CRT* gene. This gene is considered a candidate resistance marker due to the role of its ortholog in *Plasmodium falciparum*.⁷² In *P. vivax*, this gene has few polymorphisms at very low frequency, and none have been associated with chloroquine resistance.^{33,47} However, overexpression of this gene may mediate chloroquine resistance through an increase in the number of TGAAGH motif repeats in the 5' upstream region or through a deletion within intron nine.⁸⁶ To date, no existing panels include amplicons targeting these noncoding motifs, and the original PvGTSeq design excluded these regions as well. PvGTSeq also does not cover all reported polymorphisms in *DHFR* and *DHPS* genes, which are associated with resistance to pyrimethamine and sulfadoxine, respectively. Although these two drugs are not part of first-line treatment for *P. vivax* due to the widespread of mutations conferring resistance in *P. vivax* populations, monitoring them remains important

because, in some Sub-Saharan African Counties where *P. falciparum* and *P. vivax* are co-endemic, sulfadoxine-pyrimethamine is still used for *P. falciparum* intermittent preventive treatment in pregnant women and seasonal chemoprevention for children.⁸⁷ Primers were designed for all polymorphic regions in these genes; however, some exhibited low sensitivity and specificity in multiplex PCR. Future redesigns may address this by including them in a separate PCR1 reaction (not tested in this study) and subsequently combining the PCR1 products for PCR2. This approach highlights the flexibility of PvGTSeq, which utilizes open-source reagents and primers that users can obtain independently.

The results presented here demonstrate that PvGTSeq and PvCRiSP are two highly sensitive panels that allow genetic characterization of infections above five parasites/ μ L. Through analysis of samples from five endemic countries in Latin America and the Caribbean, we demonstrate that PvGTSeq enables monitoring of most candidate antimalarial resistance variants and identification of population structure at different geographic scales. Meanwhile, PvCRiSP provides reliable estimation of COI and identifies cases of clonal transmission even in samples with compromised DNA integrity. This information will serve as a reference for describing demographic changes in the *P. vivax* population, particularly in Latin America and the Caribbean—a region that in recent years has faced significant challenges for malaria control and elimination.

2.6 References

1. Manrique-Valverde, P. C. et al. PvGTSeq and PvCRiSP: two amplicon-based targeted sequencing panels for *Plasmodium vivax*. 2025.11.04.25339464 Preprint at <https://doi.org/10.1101/2025.11.04.25339464> (2025).
2. World malaria report 2024. <https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2024>.

3. Bousema, T. & Drakeley, C. Epidemiology and Infectivity of *Plasmodium falciparum* and *Plasmodium vivax* Gametocytes in Relation to Malaria Control and Elimination. *Clin. Microbiol. Rev.* **24**, 377–410 (2011).
4. White, N. J. Why Do Some Primate Malarias Relapse? *Trends Parasitol.* **32**, 918–920 (2016).
5. Ferreira, M. U. *et al.* Relative contribution of low-density and asymptomatic infections to *Plasmodium vivax* transmission in the Amazon: pooled analysis of individual participant data from population-based cross-sectional surveys. *Lancet Reg. Health – Am.* **9**, (2022).
6. Battle, K. E. *et al.* Mapping the global endemicity and clinical burden of *Plasmodium vivax*, 2000–17: a spatial and temporal modelling study. *The Lancet* **394**, 332–343 (2019).
7. Thriemer, K. *et al.* Quantifying primaquine effectiveness and improving adherence: a round table discussion of the APMEN Vivax Working Group. *Malar. J.* **17**, 241 (2018).
8. Grillet, M. E. *et al.* Malaria in Southern Venezuela: The hottest hotspot in Latin America. *PLoS Negl. Trop. Dis.* **15**, e0008211 (2021).
9. Aydemir, O. *et al.* Drug-Resistance and Population Structure of *Plasmodium falciparum* Across the Democratic Republic of Congo Using High-Throughput Molecular Inversion Probes. *J. Infect. Dis.* **218**, 946–955 (2018).
10. Early, A. M. *et al.* Detection of low-density *Plasmodium falciparum* infections using amplicon deep sequencing. *Malar. J.* **18**, 219 (2019).
11. Kattenberg, J. H. *et al.* Malaria Molecular Surveillance in the Peruvian Amazon with a Novel Highly Multiplexed *Plasmodium falciparum* AmpliSeq Assay. *Microbiol. Spectr.* **11**, e00960-22 (2023).
12. LaVerriere, E. *et al.* Design and implementation of multiplexed amplicon sequencing panels to serve genomic epidemiology of infectious disease: A malaria case study. *Mol. Ecol. Resour.* **22**, 2285–2303 (2022).
13. Aranda-Díaz, A. *et al.* Sensitive and modular amplicon sequencing of *Plasmodium falciparum* diversity and resistance for research and public health. *Sci. Rep.* **15**, 10737 (2025).
14. Fola, A. A. *et al.* SNP barcodes provide higher resolution than microsatellite markers to measure *Plasmodium vivax* population genetics. *Malar. J.* **19**, 375 (2020).
15. Kattenberg, J. H. *et al.* Novel highly-multiplexed AmpliSeq targeted assay for *Plasmodium vivax* genetic surveillance use cases at multiple geographical scales. *Front. Cell. Infect. Microbiol.* **12**, (2022).
16. Trimarsanto, H. *et al.* A molecular barcode and web-based data analysis tool to identify imported *Plasmodium vivax* malaria. *Commun. Biol.* **5**, 1411 (2022).

17. Siegel, S. V. *et al.* Lineage-informative microhaplotypes for recurrence classification and spatio-temporal surveillance of *Plasmodium vivax* malaria parasites. *Nat. Commun.* **15**, 6757 (2024).
18. Kattenberg, J. H. *et al.* *Plasmodium vivax* genomic surveillance in the Peruvian Amazon with Pv AmpliSeq assay. *PLoS Negl. Trop. Dis.* **18**, e0011879 (2024).
19. Popkin-Hall, Z. R. *et al.* High-throughput genotyping of *Plasmodium vivax* in the Peruvian Amazon via molecular inversion probes. *Nat. Commun.* **15**, 10219 (2024).
20. Rosado, J. *et al.* Understanding *Plasmodium vivax* recurrent infections using an amplicon deep sequencing assay, PvAmpSeq, identity-by-descent and model-based classification. 2025.05.26.25327775 Preprint at <https://doi.org/10.1101/2025.05.26.25327775> (2025).
21. Kleinecke, M. *et al.* Microhaplotype deep sequencing assays to capture *Plasmodium vivax* infection lineages. *Nat. Commun.* **16**, 7192 (2025).
22. Adam, I. *et al.* An open dataset of *Plasmodium vivax* genome variation in 1,895 worldwide samples. *Wellcome Open Res.* **7**, 136 (2022).
23. Ibrahim, A. *et al.* Population-based genomic study of *Plasmodium vivax* malaria in seven Brazilian states and across South America. *Lancet Reg. Health Am.* **18**, 100420 (2023).
24. Schwabl, P. *et al.* Contrasting genomic epidemiology between sympatric *Plasmodium falciparum* and *Plasmodium vivax* populations. *Nat. Commun.* **15**, 8450 (2024).
25. Mangold, K. A. *et al.* Real-Time PCR for Detection and Identification of *Plasmodium* spp. *J. Clin. Microbiol.* **43**, 2435–2440 (2005).
26. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94 (2010).
27. MalariaGEN, I, A. & Alemu, M. S. An open dataset of *Plasmodium vivax* genome variation in 1,895 worldwide samples. <https://wellcomeopenresearch.org/articles/7-136/v1>.
28. Schaffner, S. F., Taylor, A. R., Wong, W., Wirth, D. F. & Neafsey, D. E. hmlIBD: software to infer pairwise identity by descent between haploid genotypes. *Malar. J.* **17**, 196 (2018).
29. Taylor, A. R., Jacob, P. E., Neafsey, D. E. & Buckee, C. O. Estimating Relatedness Between Malaria Parasites. *Genetics* **212**, 1337–1351 (2019).
30. Barnadas, C. *et al.* *Plasmodium vivax* Resistance to Chloroquine in Madagascar: Clinical Efficacy and Polymorphisms in *pvm-dr1* and *pvcrt-o* Genes. *Antimicrob. Agents Chemother.* **52**, 4233–4240 (2008).
31. Melo, G. C. *et al.* Expression Levels of *pvcrt-o* and *pvm-dr-1* Are Associated with Chloroquine Resistance and Severe *Plasmodium vivax* Malaria in Patients of the Brazilian

- Amazon. *PLOS ONE* **9**, e105922 (2014).
32. Joy, S. *et al.* Drug resistance genes: pvcrt-o and pvmdr-1 polymorphism in patients from malaria endemic South Western Coastal Region of India. *Malar. J.* **17**, 40 (2018).
 33. Silva, S. R. *et al.* Chloroquine resistance is associated to multi-copy pvcrt-o gene in *Plasmodium vivax* malaria in the Brazilian Amazon. *Malar. J.* **17**, 267 (2018).
 34. Tantiamornkul, K., Pumpaibool, T., Piriyaongsa, J., Culleton, R. & Lek-Uthai, U. The prevalence of molecular markers of drug resistance in *Plasmodium vivax* from the border regions of Thailand in 2008 and 2014. *Int. J. Parasitol. Drugs Drug Resist.* **8**, 229–237 (2018).
 35. Sá, J. M. *et al.* *Plasmodium vivax* chloroquine resistance links to pvcrt transcription in a genetic cross. *Nat. Commun.* **10**, 4300 (2019).
 36. Rovira-Vallbona, E. *et al.* High Proportion of Genome-Wide Homology and Increased Pretreatment pvcrt Levels in *Plasmodium vivax* Late Recurrences: a Chloroquine Therapeutic Efficacy Study. *Antimicrob. Agents Chemother.* **65**, 10.1128/aac.00095-21 (2021).
 37. Bright, A. T. *et al.* Genetic Analysis of Primaquine Tolerance in a Patient with Relapsing Vivax Malaria. *Emerg. Infect. Dis.* **19**, 802–805 (2013).
 38. Gonçalves, L. A., Cravo, P. & Ferreira, M. U. Emerging *Plasmodium vivax* resistance to chloroquine in South America: an overview. *Mem. Inst. Oswaldo Cruz* **109**, 534–539 (2014).
 39. Flannery, E. L. *et al.* Next-Generation Sequencing of *Plasmodium vivax* Patient Samples Shows Evidence of Direct Evolution in Drug-Resistance Genes. *ACS Infect. Dis.* **1**, 367–379 (2015).
 40. Pearson, R. D. *et al.* Genomic analysis of local variation and recent evolution in *Plasmodium vivax*. *Nat. Genet.* **48**, 959–964 (2016).
 41. Tjitra, E., Baker, J., Suprianto, S., Cheng, Q. & Anstey, N. M. Therapeutic Efficacies of Artesunate-Sulfadoxine-Pyrimethamine and Chloroquine-Sulfadoxine-Pyrimethamine in Vivax Malaria Pilot Studies: Relationship to *Plasmodium vivax* dhfr Mutations. *Antimicrob. Agents Chemother.* **46**, 3947–3953 (2002).
 42. Hastings, M. D. *et al.* Novel *Plasmodium vivax* dhfr Alleles from the Indonesian Archipelago and Papua New Guinea: Association with Pyrimethamine Resistance Determined by a *Saccharomyces cerevisiae* Expression System. *Antimicrob. Agents Chemother.* **49**, 733–740 (2005).
 43. Auliff, A. *et al.* AMINO ACID MUTATIONS IN PLASMODIUM VIVAX DHFR AND DHPS FROM SEVERAL GEOGRAPHICAL REGIONS AND SUSCEPTIBILITY TO ANTIFOLATE DRUGS. *Am. J. Trop. Med. Hyg.* **75**, 617–621 (2006).

44. Rungsihirunrat, K., Sibley, C. H., Mungthin, M. & Na-Bangchang, K. Geographical Distribution of Amino Acid Mutations in Plasmodium vivax DHFR and DHPS from Malaria Endemic Areas of Thailand. *Am. J. Trop. Med. Hyg.* **78**, 462–467 (2008).
45. Cubides, J. R. *et al.* Simultaneous detection of Plasmodium vivax dhfr, dhps, mdr1 and crt-o resistance-associated mutations in the Colombian Amazonian region. *Malar. J.* **17**, 130 (2018).
46. Auburn, S. *et al.* Genomic Analysis of Plasmodium vivax in Southern Ethiopia Reveals Selective Pressures in Multiple Parasite Mechanisms. *J. Infect. Dis.* **220**, 1738–1749 (2019).
47. Suwanarusk, R. *et al.* Chloroquine Resistant Plasmodium vivax: In Vitro Characterisation and Association with Molecular Polymorphisms. *PLOS ONE* **2**, e1089 (2007).
48. Imwong, M. *et al.* Gene Amplification of the Multidrug Resistance 1 Gene of Plasmodium vivax Isolates from Thailand, Laos, and Myanmar. *Antimicrob. Agents Chemother.* **52**, 2657–2659 (2008).
49. Marfurt, J. *et al.* Molecular Markers of In Vivo Plasmodium vivax Resistance to Amodiaquine Plus Sulfadoxine-Pyrimethamine: Mutations in pvdhfr and pvmdr1. *J. Infect. Dis.* **198**, 409–417 (2008).
50. Hupalo, D. N. *et al.* Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax. *Nat. Genet.* **48**, 953–958 (2016).
51. Ngassa Mbenda, H. G. *et al.* Evolution of the Plasmodium vivax multidrug resistance 1 gene in the Greater Mekong Subregion during malaria elimination. *Parasit. Vectors* **13**, 67 (2020).
52. Mbenda, H. G. N. *et al.* Genetic diversity of the Plasmodium vivax phosphatidylinositol 3-kinase gene in two regions of the China-Myanmar border. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **61**, 45–52 (2018).
53. Mint Deida, J. *et al.* Assessment of drug resistance associated genetic diversity in Mauritanian isolates of Plasmodium vivax reveals limited polymorphism. *Malar. J.* **17**, 416 (2018).
54. Cowell, A. N., Valdivia, H. O., Bishop, D. K. & Winzeler, E. A. Exploration of Plasmodium vivax transmission dynamics and recurrent infections in the Peruvian Amazon using whole genome sequencing. *Genome Med.* **10**, 52 (2018).
55. Noisang, C. *et al.* Molecular detection of drug resistant malaria in Southern Thailand. *Malar. J.* **18**, 275 (2019).
56. Korsinczky, M. *et al.* Sulfadoxine Resistance in Plasmodium vivax Is Associated with a Specific Amino Acid in Dihydropteroate Synthase at the Putative Sulfadoxine-Binding Site. *Antimicrob. Agents Chemother.* **48**, 2214–2222 (2004).
57. Imwong, M. *et al.* Limited Polymorphism in the Dihydropteroate Synthetase Gene (dhps) of

- Plasmodium vivax Isolates from Thailand. *Antimicrob. Agents Chemother.* **49**, 4393–4395 (2005).
58. Mint Lekweiry, K. *et al.* Molecular surveillance of drug-resistant Plasmodium vivax using pvdhfr, pvdhps and pvmdr1 markers in Nouakchott, Mauritania. *J. Antimicrob. Chemother.* **67**, 367–374 (2012).
 59. Ganguly, S., Saha, P., Chatterjee, M. & Maji, A. K. Prevalence of Polymorphisms in Antifolate Drug Resistance Molecular Marker Genes pvdhfr and pvdhps in Clinical Isolates of Plasmodium vivax from Kolkata, India. *Antimicrob. Agents Chemother.* **58**, 196–200 (2014).
 60. Home - GTseek - GT-seq. *GTseek LLC* <https://gtseek.com/>.
 61. Campbell, N. R., Harmon, S. A. & Narum, S. R. Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Mol. Ecol. Resour.* **15**, 855–867 (2015).
 62. Cowell, A. N. *et al.* Selective Whole-Genome Amplification Is a Robust Method That Enables Scalable Whole-Genome Sequencing of Plasmodium vivax from Unprocessed Clinical Samples. *mBio* **8**, e02257-16 (2017).
 63. Amos, B. *et al.* VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res.* **50**, D898–D911 (2022).
 64. Fast and accurate short read alignment with Burrows–Wheeler transform | Bioinformatics | Oxford Academic. <https://academic.oup.com/bioinformatics/article/25/14/1754/225615?login=false>.
 65. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
 66. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
 67. Manske, M. *et al.* Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature* **487**, 375–379 (2012).
 68. Guo, B. *et al.* Strong positive selection biases identity-by-descent-based inferences of recent demography and population structure in Plasmodium falciparum. *Nat. Commun.* **15**, 2499 (2024).
 69. Asih, P. B. S. *et al.* Distribution of Plasmodium vivax pvdhfr and pvdhps alleles and their association with sulfadoxine-pyrimethamine treatment outcomes in Indonesia. *Malar. J.* **14**, 365 (2015).
 70. Hastings, M. D. & Sibley, C. H. Pyrimethamine and WR99210 exert opposing selection on dihydrofolate reductase from Plasmodium vivax. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 13137–

13141 (2002).

71. Imwong, M. *et al.* Association of genetic mutations in *Plasmodium vivax* dhfr with resistance to sulfadoxine-pyrimethamine: geographical and clinical correlates. *Antimicrob. Agents Chemother.* **45**, 3122–3127 (2001).
72. Buyon, L. E., Elsworth, B. & Duraisingh, M. T. The molecular basis of antimalarial drug resistance in *Plasmodium vivax*. *Int. J. Parasitol. Drugs Drug Resist.* **16**, 23–37 (2021).
73. Orjuela-Sánchez, P. *et al.* Analysis of single-nucleotide polymorphisms in the crt-o and mdr1 genes of *Plasmodium vivax* among chloroquine-resistant isolates from the Brazilian Amazon region. *Antimicrob. Agents Chemother.* **53**, 3561–3564 (2009).
74. Suwanarusk, R. *et al.* Amplification of pvmdr1 associated with multidrug-resistant *Plasmodium vivax*. *J. Infect. Dis.* **198**, 1558–1564 (2008).
75. Marques, M. M. *et al.* *Plasmodium vivax* chloroquine resistance and anemia in the western Brazilian Amazon. *Antimicrob. Agents Chemother.* **58**, 342–347 (2014).
76. Schousboe, M. L. *et al.* Multiple Origins of Mutations in the mdr1 Gene--A Putative Marker of Chloroquine Resistance in *P. vivax*. *PLoS Negl. Trop. Dis.* **9**, e0004196 (2015).
77. Spotin, A. *et al.* Global assessment of genetic paradigms of Pvmdr1 mutations in chloroquine-resistant *Plasmodium vivax* isolates. *Trans. R. Soc. Trop. Med. Hyg.* **114**, 339–345 (2020).
78. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. - Google Search. <https://link.springer.com/book/10.1007/978-3-319-24277-4>.
79. R Core Team (2024). *_R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. - Google Search. <https://www.scirp.org/reference/referencespapers?referenceid=3887298>.
80. IamKDO. IamKDO/GADMTTools. (2025).
81. Kleinecke, M. *et al.* Microhaplotype deep sequencing assays to capture *Plasmodium vivax* infection lineages. 2024.10.14.24315131 Preprint at <https://doi.org/10.1101/2024.10.14.24315131> (2024).
82. Hubbard, A. *et al.* PvGAP: Development of a globally-applicable, highly-multiplexed microhaplotype amplicon panel for *Plasmodium vivax*. 2025.04.30.25326751 Preprint at <https://doi.org/10.1101/2025.04.30.25326751> (2025).
83. Barratt, J. L. N. *et al.* Genetic characterization of *Plasmodium vivax* linked to autochthonous malaria transmission in the US (2023) using Illumina AmpliSeq technology: a genetic epidemiology study. *Lancet Reg. Health – Am.* **48**, (2025).
84. Sutanto, E. *et al.* Genomics of *Plasmodium vivax* in Colombia reveals evidence of local

- bottle-necking and inter-country connectivity in the Americas. *Sci. Rep.* **13**, 19779 (2023).
85. Buyon, L. E. *et al.* Population genomics of *Plasmodium vivax* in Panama to assess the risk of case importation on malaria elimination. *PLoS Negl. Trop. Dis.* **14**, e0008962 (2020).
86. Sá, J. M. *et al.* *Plasmodium vivax*: allele variants of the *mdr1* gene do not associate with chloroquine resistance among isolates from Brazil, Papua, and monkey-adapted strains. *Exp. Parasitol.* **109**, 256–259 (2005).
87. CDC. Core Drug-based Malaria Prevention Strategies. *Malaria*
<https://www.cdc.gov/malaria/php/public-health-strategy/drug-strategies.html> (2024).

Chapter Three: Resolving the recent demographic history of *Plasmodium vivax* populations in Latin America and the Caribbean

3.1 Abstract

Each year, between 6 and 7 million malaria cases are reported, and in the Americas, *P. vivax* accounts for more than 70% of cases.¹ Contrary to the global trend, since 2015 many countries in LAC have reported increasing case counts, reaching a historic peak of 1 million malaria cases in 2017 and 2018.¹ Venezuela accounts for approximately 50% of cases in the region.² Other countries, including Colombia, Guyana, Bolivia, Ecuador, Nicaragua, and Costa Rica, have experienced stalled progress, with decreases slowing or reversing for *P. vivax*.¹

Between 2012 and 2019, Honduras and Guatemala made notable progress in malaria control, reducing *P. vivax* cases to fewer than 1,000 in 2019.^{3,4} Similarly, Panama reported fewer than 1,000 cases per year between 2011 and 2018.^{5,6} However, during the past decade this region has been affected by human migratory waves and the Darien Gap is a critical point in this route. Since 2019 Panama began reporting a resurgence in malaria cases, followed by Honduras and Guatemala in 2020 and 2021, respectively.^{3,4} To allocate resources efficiently it is important to determine whether these resurgences are driven by the importation and introduction of new parasite groups through human migration from the southern part of the continent, or by indigenous parasite populations that have remained resilient to interventions in the region.

Genomic tools can support malaria case classification and better distinguish locally acquired cases from imported cases. Using amplicon sequencing and whole genome sequencing data from the *P. vivax* parasite across 12 Latin American and Caribbean countries (LAC), we found that indigenous parasites represent 70 to 100% of *P. vivax* cases after 2020. We also identified clonal lineages in Panama that were recently introduced from the Pacific coast of Colombia and the Guiana Shield, as well as evidence of northward movement of parasite

genotypes from Panama and Honduras into Guatemala and the United States, respectively. Taken together, these findings emphasize that migrants are vulnerable to infections during their transit and suggest that malaria control in these countries depends largely on case management and local vector control.

3.2 Introduction

Each year, between 6 and 7 million malaria cases are reported, and in the Americas, *P. vivax* accounts for more than 70% of cases.¹ Contrary to the global trend, since 2015 many countries in LAC have reported increasing case counts, reaching a historic peak of 1 million malaria cases in 2017 and 2018.¹ Venezuela accounts for approximately 50% of cases in the region.² Other countries, including Colombia, Guyana, Bolivia, Ecuador, Nicaragua, and Costa Rica, have experienced stalled progress, with decreases slowing or reversing for *P. vivax*.¹ Between 2012 and 2019, Honduras and Guatemala made notable progress in malaria control, reducing *P. vivax* cases to fewer than 1,000 in 2019.^{3,4} Similarly, Panama reported fewer than 1,000 cases per year between 2011 and 2018.^{5,6} In 2015, several Central American countries, including Panama, Honduras, and Guatemala, joined the challenge of eliminating malaria within their territories by 2030.⁷

However, this region is affected by migratory waves moving north toward the United States. In the Darién Gap, between 2015 and 2016, around 60,000 regular migrants were reported, mainly Haitian and Cuban people returning from South America. A second major wave began in 2022 and lasted until late 2024, consisting largely of Venezuelan nationals traveling to the United States in search of employment.^{8–10} The Darién Gap, on the border between Colombia and Panama, is the initial critical point along this route. Reports from the United Nations

International Organization for Migration (IOM) indicate that, within Panama, the first arrival point is the community of Canaán Membrillo, from which migrants seek to reach the Pan-American Highway to continue north by bus.¹⁰ Other important stops include Danlí in Honduras, near the border with Nicaragua,¹¹ and southeastern Guatemala, near the border with Honduras.¹²

Panama began reporting a resurgence in cases in 2019, reaching 7,102 *P. vivax* cases in 2022 and 9,451 cases in 2023.⁵ Honduras and Guatemala also reported increases beginning in 2020 and 2021, respectively, reaching nearly 2,000 cases each in 2022.^{3,4} To define the most appropriate intervention strategies to contain the disease and reduce risk, the WHO recommends stratifying risk areas based on receptivity (the capacity to sustain transmission) and vulnerability (the risk of receiving imported cases from other endemic areas).¹³ It is therefore important to distinguish between locally acquired cases, whether due to indigenous transmission or recently introduced parasites, and imported cases. This classification can be feasible when importation flows into non-receptive, non-endemic areas, or into receptive endemic areas without active indigenous transmission. However, when importation occurs in receptive areas with active transmission foci, classifying cases using epidemiological information alone becomes more difficult.

Genetic information from clinical cases can support this classification, and studies in Honduras show that indigenous cases have genotypes that differ from imported cases.¹⁴ Distinguishing these cases is already important for local decision-making, but it does not necessarily identify the source of importation or enable action at the regional level. To trace the origin of infections with higher resolution, it is important to catalog parasite diversity in the region.

In the present study, we tested whether the rise in cases in Panama, Honduras, and Guatemala is due to the importation and introduction of parasites or the expansion of indigenous populations resilient to control and elimination strategies. To do so, using our PvGTSeq amplicon panel in combination with whole genome sequencing information we generated an extensive catalog of genomic data of *P. vivax* populations across 12 Latin American and Caribbean countries (LAC). This information allowed us to infer the geographic origins of infections detected in Panama, Honduras, and Guatemala, focusing mainly on 2020–2024 while also incorporating earlier years.

Our results indicate that transmission in Mesoamerica is maintained primarily through the clonal expansion of indigenous parasites as they represent more than 70% of the total cases in the period after 2020. We also identified clonal lineages in Panama that were recently introduced from the Pacific coast of Colombia and the Guiana Shield, as well as evidence of northward movement of parasite genotypes from Panama and Honduras into Guatemala and the United States, respectively. Taken together, these findings emphasize the vulnerability of migrants to acquire infections during transit and suggest that malaria control in these countries depends largely on case management and local vector control.

3.3 Methods

3.3.1 Sample collection

For this study, 2,819 samples were sequenced using whole-genome sequencing (WGS; $n = 1,087$) and/or amplicon sequencing ($n = 2,059$) with the PvGTSeq panel. Samples were collected by our collaborators in Colombia ($n = 576$), Guatemala ($n = 154$), Guyana ($n = 1,271$), Honduras

(n = 96), Panama (n = 254), Peru (n = 88), Trinidad and Tobago (n = 45), and Venezuela (n = 358).

Samples from Colombia were obtained by the Caucesco Scientific Research Center through passive surveillance of clinical cases at diagnostic centers in the municipalities of Quibdó (n = 307), Buenaventura (n = 120), Guapi (n = 8), Tumaco (n = 2), and Puerto Inírida (n = 139) between 2021 and 2022.

In Guatemala, sample collection was carried out in coordination with Clinton Health Access Initiative (CHAI) and the Guatemalan Ministry of Health through passive case and reactive collection in the departments of Alta Verapaz (n = 74), Chiquimula (n = 10), Escuintla (n = 27), Izabal (n = 23), Petén Sur Occidente (n = 11), and Petén Norte (n = 8) during 2024. The community of Esquipulas in Chiquimula was selected as a sentinel site due to its role as a migration control point for people coming from Honduras and El Salvador.

In Guyana, samples were collected through passive case collection at diagnostic posts in cooperation with the country's Ministry of Health Vector Control Services and came from the regions of Upper Takutu–Upper Essequibo (n = 56), Potaro–Siparuni (n = 49), Barima–Waini (n = 33), Cuyuni–Mazaruni (n = 169), and Demerara–Mahaica (n = 964). They were collected in 2020 (n = 251), 2021 (n = 695), and 2022 (n = 325).

In Honduras, samples were selected from the malaria DNA bank in the department of Parasitology at the Universidad Autonoma de Honduras and were originally collected through passive and active case detection by the Ministry of Health (SESAL) mainly in the department of Gracias a Dios (n = 81) in 2020 (n = 2), 2021 (n = 20), 2022 (n = 43), and 2023 (n = 3). Samples

from the departments of Colón (n = 5), Islas de la Bahía (n = 1), and Olancho (n = 6) were also included, as well as historical samples from 2011 (n = 12).

In Panama, samples were provided by the Gorgas Institute and spanned a broad time window from 2004 to 2024. Original sample collection was done through passive and active case detection by the Vector Control Department at the Panama Ministry of Health (DCV-MoH) as part of routine malaria surveillance. These samples were collected in the jurisdictions of East Panama (n = 98), Guna Yala (n = 53), Darién (n = 36), Ngäbe-Buglé (n = 24), Madugandí (n = 9), Veraguas (n = 9), and Colón (n = 1). In addition, 24 samples were received without information on the sampling location or date.

In Peru, collection was led by the research group at the Universidad Nacional Toribio Rodríguez de Mendoza in coordination with the regional health directorate of the Condorcanqui district in the department of Amazonas. Collection took place in 2020 (n = 22), 2021 (n = 13), and 2022 (n = 53).

In Trinidad and Tobago, samples were shared by the University of West Indies School of Veterinary Medicine and collected through passive, active, and reactive case detection carried out by the Insect Vector Control Division (IVCD) of the Ministry of Health. Samples were collected from patients presenting with febrile illness, recent travel history, or residence in malaria-endemic regions. When a blood smear was confirmed positive, the IVCD conducted follow-up investigations, including testing of household members and nearby individuals, to assess potential local transmission. This collection was conducted mainly in 2019 (n = 24), but samples were also collected in 2018 (n = 2) and from 2020 to 2023 (n = 19).

In Venezuela, samples were collected by the Biomedical Research and Therapeutic Vaccines Institute in Ciudad Bolívar, Venezuela, through passive surveillance of patients with a positive malaria diagnosis at University Hospital Ruiz y Paez in Ciudad Bolívar in the state of Bolívar in 2019 (n = 11), 2021 (n = 51), and 2022 (n = 296).

Additionally, the full genomes of 1,368 samples from Latin America and the Caribbean (n = 623) and from other geographic world regions (n = 745) were downloaded from the MalariaGEN Pv4 database (n = 891) and from the SRA (n = 493), requiring at least 75% of the genome with > 5 reads of coverage and/or > 75% coverage of the PvGTSeq amplicons. These Latin American samples included the countries of Brazil (n = 101), Colombia (n = 67), El Salvador (n = 1), French Guiana (n = 13), Guyana (n = 325), Honduras (n = 7), Mexico (n = 20), Nicaragua (n = 1), Panama (n = 18), Peru (n = 55), United States of America (n = 3), and Venezuela (n = 12).

3.3.2 Sample processing and sequencing

For clinical samples from Colombia (n = 459), Guatemala (n = 154), Guyana (n = 460), and Venezuela (n = 352), total genomic DNA was extracted from dried blood spots (20–35 mm² punched area per sample) using KingFisher Ready DNA Ultra 2.0 Prefilled Plates on the KingFisher Flex instrument (ThermoFisher Scientific). For samples from Honduras (n = 39), Panama (n = 53), Peru (n = 90), and Trinidad and Tobago (n = 44) collaborators provided stored DNA from previous studies.

We then applied two consecutive rounds of selective whole-genome amplification (sWGA) to the DNA extracts. For both rounds of sWGA, the 10 µL sample mix consisted of 0.52 µL of primer set (Pvset1920 for the first round and Pvset1 for the second round), 6.25 µL of 4 mM

dNTPs (Thermo Scientific), 0.23 μ L of nuclease-free water, and 3 μ L of DNA. After heating the sample mix for 3 minutes at 95 °C, we added 10 μ L of reaction mix containing 9 μ L of reaction buffer and 1 μ L of Phi29 enzyme (Genomiphi v2 kit, Cytiva). Amplification was performed using step-down incubation (35 °C for 5 min, 34 °C for 10 min, 33 °C for 15 min, 32 °C for 20 min, 31 °C for 30 min, 30 °C for 24 h, and 65 °C for 15 min), followed by cooling to 4 °C. The amplified product from each round was diluted 1 over 50 using nuclease-free water.

For Guyana, samples were sequenced only using PvGTSeq on an Illumina NovaSeq instrument. For Honduras and Panama, all samples were sequenced using PvGTSeq on an iSeq 100 instrument, and 84 and 179 samples, respectively, were sequenced by WGS. For Colombia, 362 samples were sequenced using PvGTSeq on an Illumina MiSeq instrument, 195 by WGS, and 19 by both methods. For Guatemala, 35 samples were sequenced using PvGTSeq on an Illumina iSeq instrument, and 119 by WGS. For Venezuela, 45 samples were sequenced using PvGTSeq on an Illumina NovaSeq instrument and 358 by WGS. Finally, for Peru and Trinidad and Tobago, all samples were sequenced using WGS. WGS was performed on the Illumina NovaSeq 6000 platform.

Library generation using the PvGTSeq panel consisted of a nested PCR, a multiplex PCR to specifically amplify the 249 targets in the panel, followed by a conventional PCR to add the i5 and i7 Illumina adapters (Nextera indexing). Each multiplex PCR reaction consisted of 10 μ L of Qiagen Plus Master Mix, 3 μ L of primer mix (0.067–0.3 μ M), 1 μ L of nuclease-free water, and 6 μ L of DNA template. Amplification was performed with 15 min of initial denaturation, followed by 5 cycles of step-up incubation (30 s denaturation at 95 °C, 30 s annealing starting at 57 °C and increasing by 5% in each cycle, and 2 min extension at 72 °C), then 20 cycles of amplification (30 s denaturation at 95 °C, 30 s annealing at 65 °C, and 30 s extension at 72 °C),

followed by cooling to 4 °C. The Nextera indexing PCR mix consisted of 5 µL of KAPA HiFi HotStart Ready Mix (2x), 2.2 µL of unique dual index, and 7.2 µL of a 1/13 dilution of the multiplex PCR1 product. Samples from two 96-well plates were pooled, and two rounds of cleanup and size selection were performed using AMPure XP beads at 1X. Library quantification was performed by qPCR using the KAPA SYBR FAST kit (Roche), and amplicon sizes were verified on a Bioanalyzer. Finally, sequencing was performed using the Illumina iSeq 100, MiSeq v2, or NovaSeq 6000 kits for a total of 300 cycles. WGS libraries were constructed using the NEBNext Ultra II FS DNA Library Prep Kit (NEB E6177) and sequenced (2 × 151 bp) on the Illumina NovaSeq 6000 platform at the Broad Institute.

3.3.3 Sequencing data analysis

For amplicon raw sequencing data of PvGTSeq, we processed paired-end Illumina sequencing data in the form of FASTQ files using a custom analysis pipeline for which documentation can be found at <https://github.com/broadinstitute/malaria-amplicon-pipeline>. This pipeline utilizes the Divisive Amplicon Denoising Algorithm (DADA2)¹⁵ to obtain microhaplotypes (i.e., alleles or ‘amplicon sequence variants’). We aligned microhaplotypes obtained from dada2 against a custom-built database of PvP01 reference sequences for each amplicon locus. We then summarized observed sequence polymorphism into a concise format by converting individual microhaplotypes into ‘pseudo-CIGAR’ strings using a custom python script that can be found at <https://github.com/Paulonvnnv/MHap-Analysis>. The rules used for this conversion are also detailed in **Supplementary figure 2.2**. Next, microhaplotypes were discarded if supported by fewer than 5, 10, or 20 read-pairs on the iSeq, MiSeq or NovaSeq generated data respectively, or by less than 1% of total read-pairs within a heterozygous locus.

For WGS information, we aligned reads to the *P. vivax* P01 reference genome version 2 assemblies¹⁶ using BWA-MEM v0.7.17-r1188 and called SNPs and INDELs using GATK v3.5-0-g36282e4¹⁷ ‘HaplotypeCaller’ and ‘GenotypeGVCFs’ according to best practices defined by the Pf3k consortium (http://www.malariagen.net/data_package/pf3k-5/). The mapping and joint variant call process also included samples from the MalariaGEN Pv4 dataset and information from SRA. All downstream genetic analyses focused on variant sites, single nucleotide polymorphisms (SNPs) and insertions and deletions (INDELs), in core regions of the genome. Samples missing genotype calls for >25% variant sites at $\geq 2\%$ minor allele frequency (after application of the core region filter) were also excluded from further analysis. Following sample exclusions, we removed variant sites located in coding regions prone to alignment errors (coding regions with observed heterozygosity above the 95th percentile and variant site density exceeding the 92.5th percentile of all coding regions in all populations). We also removed variant sites located in homopolymers and di-nucleotide short tandem repeats, as these are prone to PCR errors.

Our downstream analyses of population structure and identification of clonal groups were based on the combined use of PvGTSeq and WGS data. To do this, we extracted variant sites within the coordinates of the PvGTSeq amplicons from the gVCF files generated for each sample. We then reconstructed the polymorphisms observed in each gVCF into the original FASTA sequence(s) for each amplicon and represented those polymorphisms in pseudo-CIGAR format. This gVCF-to-FASTA and pseudo-CIGAR conversion was performed only if all variant positions for a given sample at a given amplicon had at least 5 supporting reads. If any position had fewer than 5 reads, or if information was missing, all data for that amplicon in that sample were considered missing. The datasets were harmonized by identifying identical FASTA

sequences between PvGTSeq and WGS and comparing their corresponding pseudo-CIGAR strings. When identical FASTA sequences between PvGTSeq and WGS had different pseudo-CIGAR strings, we manually curated the results to identify possible causes: 1) unequal representation of repetitive regions between PvGTSeq and WGS, or 2) off-target alignment of the amplicon during paired-read alignment in the lab's malaria-amplicon-pipeline. In the first case, we imputed the WGS data, assuming that the correct INDEL representation is that of PvGTSeq because this representation is based on a multiple alignment performed with all ASVs specific to the amplicon of interest, and because there is greater confidence in correct INDEL placement for PvGTSeq since primers generally flank repetitive regions and each amplicon is generated through the near-complete overlap of the two paired-reads (each amplicon is on average 150 bp and was sequenced with 300 cycles). In the second case, we first verified by BLAST that the aligned sequence variant (ASV) generated by PvGTSeq had been incorrectly assigned to the amplicon, and if confirmed we imputed the PvGTSeq pseudo-CIGAR information with that of WGS. This occurred for only 23 ASVs.

Once both datasets were harmonized, we filtered out microhaplotypes exhibiting less than 80% identity to the reference sequence. We then masked variants in both the FASTA sequence of the microhaplotype and its pseudo-cigar string representation according to three criteria: 1) SNPs or INDELs located in homopolymer regions (≥ 5 consecutive identical nucleotides), 2) INDELs positioned at either the beginning or end of the microhaplotype sequence, and 3) variants predominantly ($> 66\%$) present as minor alleles and mainly found in heterozygous sites ($> 66\%$) across the population. Finally, samples and amplicons with more than 25% missing data were removed.

3.3.4 Polyclonal infections and haplotypic diversity

Using the combined dataset, we distinguished between monoclonal and polyclonal infections based on the fraction of heterozygous loci (Het_f) within each sample and the within-sample genetic differentiation (F_{ws}). We classified an infection as monoclonal when $Het_f \leq 0.03$ (3%) and $F_{ws} \geq 0.975$. Using monoclonal infections, we quantified each country's haplotypic diversity based on the total number of haplotypes detected, the effective number of haplotypes (the number of haplotypes that explain the observed haplotypic diversity), and Shannon's haplotypic diversity index. These indices were reported only for countries in Latin America and the Caribbean.

3.3.5 Genetic relationships among infections across LAC, clonal propagation and population subdivision

We used identity by descent (IBD) to quantify genetic relationships between each pair of monoclonal samples in the combined LAC dataset (PvGTSeq plus PvGTSeq regions from WGS). We used PvGTSeq microhaplotype data as an initial screen because it provides the largest number of samples per geographic unit. We also included all polyclonal samples from Guatemala, Honduras, and Panama to identify the origin of infections in these countries. IBD was estimated using the R package `paneljudge`.¹⁸ For polyclonal infections, we first phased haplotype clones based on read depth and estimated IBD for each clone against all samples in the dataset. We then took the lowest IBD value across all possible comparisons with each other sample.

For each country, we summarized the distribution of IBD across all pairwise comparisons. We also identified critical IBD values to describe: 1) clonal propagation, 2) recombination

between unrelated or partially related clonal groups and/or genotypes, and 3) parasite population structure driven by geographic subdivision (ecological groups). We define clonal propagation as a group of highly related genotypes that expands across space and time through inbreeding among its members, in the absence of introgression (recombination with unrelated genotypes). In clonal groups, we expect long IBD segments among members across the entire genome. In contrast, recombination between unrelated or partially related groups or genotypes is expected to produce outcrossing products with long, complementary IBD segments shared with each parental group. Finally, under population subdivision, members within each subpopulation are connected by historical gene flow, but we do not expect to observe many long IBD segments.

In population genetics, IBD values of 0.5 (50%) and 0.25 (25%) are traditionally used to describe expected genetic relationships between parent and offspring and between siblings, respectively, and to identify inbreeding. However, IBD estimates from amplicon data tend to be inflated, particularly at intermediate to low values (0.5–0). This inflation reflects the limited number of genomic regions analyzed, ascertainment bias introduced by the dataset used to select amplicons, and bias introduced by population inbreeding when estimating allele frequencies.

We identified these critical IBD values by evaluating node aggregation (or branch formation among nodes or monoclonal samples) using a network analysis. We scanned IBD thresholds starting at $IBD = 1$ and decreasing by 0.01 until all samples were connected in a single cluster. A cluster was defined as a group of four or more samples in the network in which each sample is connected by at least one edge to another member, and edges were drawn between samples when their pairwise IBD was greater than or equal to the specified threshold. For each IBD threshold, we described: 1) the number of clusters (groups of four or more samples in which each sample is connected by at least one edge to another member), 2) the number of samples not

in clusters (singletons, doubletons, and tripletons), 3) the number of samples in clusters, 4) the number of new samples incorporated into a cluster relative to the previous IBD value (the first-order negative derivative of the number of samples in clusters), and 5) the number of merges between clusters (M). The number of merges M was defined as the change in the number of clusters relative to the previous IBD value or ΔC minus the number of new clusters formed. And ΔC is the first-order derivative of the number of clusters.

Thus, $M > 0$ indicates that two or more clusters merge due to 1) diversification events when the peak occurs at IBD close to 1, or 2) recombination among unrelated clusters and/or genotypes when the peak occurs at lower IBD values. We used network topology to distinguish between diversification (core-periphery clusters) and recombination (overlapping or fuzzy clusters with hubs representing the recombinant lineage).

We then analyzed the critical IBD values corresponding to peaks where $M > 0$ and defined clonal groups based on the minimum IBD value at which recombination was not observed in the whole population (the IBD critical value plus 0.01). To identify parasite population structure due to geographic subdivision or ecological groups, we used the minimum critical IBD value that kept clusters from the administrative areas of interest (Country) disconnected. We also confirmed these patterns, and the genetic relationships among clonal groups, using WGS data. To corroborate population structure, we performed an unsupervised ADMIXTURE analysis and a principal coordinates analysis (PCoA) using the R package princomp. To avoid bias from the presence of clonal groups in these analyses, we randomly selected three clonal samples from each clonal group. Finally, we describe the frequency and the spatial and temporal distribution of clonal groups across LAC, as well as the genetic relationships among clonal and ecological groups within and between countries in LAC. Genetic relationships among clonal or ecological

groups were evaluated based on the fraction of pairwise IBD comparisons between groups that were equal to or greater than the lowest critical IBD value observed.

3.3.6 Contribution of imported cases to transmission in Mesoamerica

The WHO classifies malaria cases into three categories based on infection origin:¹³ 1) imported—when travel history suggests the infection was acquired outside the region of interest; 2) introduced—when the infection is locally acquired but can be linked to an external source, despite no travel outside the area; and 3) indigenous—when the infection is locally acquired with no connection to an imported source. This classification works well when people move between endemic and non-endemic areas. However, when movement occurs between endemic regions, classification based solely on travel history becomes difficult.

Using our genetic dataset, we classified infections in Panama and Guatemala into three categories: 1) Autochthonous: samples that belong to clonal or ecological groups, or to an ancestral population, that are confined to a subnational or national area of interest; 2) Introduced: samples that belong to a clonal group that shows no interaction with the ecological groups present in the area of interest and do not belong to the ancestral population; 3) Imported: unique genotypes that do not belong to clonal or ecological groups, or ancestral populations, from the area of interest. We describe the frequency of these categories in each country and compare this genetic classification with travel history records from Panama and Guatemala.

3.4 Results

3.4.1 Haplotypic diversity of the study areas

Of the total 4,187 sequences, 2,954 had >75% coverage of the PvGTSeq markers and 1,860 had >75% coverage of the whole genome (**Figure 3.1**). Of these 2,954 unique samples by PvGTSeq 2,088 were from LAC, and 1,537 out of them were categorized as monoclonal and 551 as polyclonal. The number and proportion of monoclonal samples varied across the different LAC countries analyzed are presented in **Table 3.1**. The percentage of polyclonal samples was around 1% in Guatemala (0.8%) and Honduras (1.1%), around 10% in Panama (9%), Mexico (10%), and Peru (11%), and above 20% in Brazil (23.9%), Colombia (26.89%), Venezuela (33.3%), and Guyana (42.4%). Moreover, in Panama we observed an increase in the percentage of polyclonal infections from 2020 onward (5/32, 15.6%, CI95% 6.9–31.8%) compared with previous years (4/59, 6.8%, CI95% 2.7–16.2%); however, the difference was not significant.

Table 3.1: Percentage of polyclonal infections in LAC

Region	Country	N	Polyclonals	% Polyclonals	Lower limit	Upper limit
Amazon	BRA	117	22	18.8	12.8	26.8
	PER	108	10	9.3	5.1	16.2
Guiana shield	COL - Guainia	135	46	34.1	26.6	42.4
	VEN	64	18	28.1	18.6	40.1
	GUY	949	358	37.7	34.7	40.9
	TTO	42	13	31	19.1	46
Pacific coast	COL	339	70	20.6	16.7	25.3
Mesoamerica	PAN	117	11	9.4	5.3	16.1
	HND	61	0	0	0	5.9
	GTM	118	1	0.8	0	4.6
	MEX	20	1	5	0.3	23.6

Only the 1,537 monoclonal samples were used to analyze patterns of population structure and clonal propagation across LAC countries. Additionally, 12 polyclonal samples from Guatemala, and Panama, and 593 monoclonal samples from other world regions were included to identify possible imported cases from Africa and Asia into Mesoamerica.

Across the 1,537 LAC monoclonal samples, we identified 1,003 unique haplotypes. PvGTSeq haplotypic Shannon (H') and Simpson (D) diversity were 6.747 and 0.993, respectively. The expected heterozygosity of the PvGTSeq marker was 0.494, and the median nucleotide diversity across all coding regions in the genome was 1.12×10^{-4} (IQR: 1.54×10^{-4}) (**Table 3.2**).

The geographic area with the highest haplotypic diversity was the Guiana Shield region (Guyana, Trinidad, Venezuela, and the Guainía region in Colombia). Across this region, we found 681 unique haplotypes, and H' was 6.66. In contrast, the Mesoamerica region (Panama, Honduras, Guatemala, and Mexico) had the lowest haplotypic diversity, with 62 unique haplotypes and a $H' = 2.96$. In Panama, haplotypic diversity increased after 2020. The number of unique haplotypes changed from 5 to 9, and the number of effective haplotypes (the number of haplotypes that explain the observed diversity) increased from 1.98 to 5.48. The Amazon and the Colombian Pacific coast showed similar levels of haplotypic diversity (H' from 4.16 to 4.486). However, at the subnational level, in Colombia diversity was closer to that in Mesoamerica, with the exception of Quibdó in 2022, where H' was 3.89 and 59.55 effective haplotypes were observed.

The genetic diversity of PvGTSeq markers was similar across most countries ($H_{(exp)}$ from 0.41 to 0.495), with the exception of Guatemala ($H_{(exp)} = 0.265$), Mexico ($H_{(exp)} = 0.353$),

Panama ($H_{\text{(exp)}} = 0.354$), and Peru ($H_{\text{(exp)}} = 0.374$). Genome-wide nucleotide diversity ranged from $7.13\text{e-}05$ to $1.49\text{e-}04$.

Table 3.2: Haplotipic and genetic diversity of LAC countries using PvGTSeq

Region	Country	N	S	D	H'	A	H(exp)	pi(median)
Amazon	Brazil	95	76	0.992	4.199	4.404	0.444	1.25E-04
	Peru	98	44	0.822	2.775	3.574	0.374	8.48E-05
	Total	193	119	0.95	4.16	4.87	0.45	1.16E-04
Guiana shield	COL - Guainia	89	58	0.972	3.737	3.235	0.424	1.59E-04
	Venezuela	46	40	0.993	3.636	3.093	0.422	1.00E-04
	Guyana	591	544	0.999	6.419	4.339	0.430	9.72E-05
	Trinidad	29	26	0.985	3.176	2.940	0.424	1.55E-04
	Total	768	681	1.00	6.66	4.66	0.43	1.02E-04
Pacific coast	Colombia	269	141	0.979	4.486	4.322	0.495	1.03E-04
Mesoamerica	Panama	106	14	0.682	1.625	3.672	0.354	8.32E-05
	Honduras	61	26	0.950	2.993	2.978	0.416	1.16E-04
	Guatemala	117	8	0.548	1.076	2.891	0.265	7.13E-05
	Mexico	19	16	0.982	2.726	2.415	0.353	1.14E-04
	Total	307	62	0.89	2.96	4.70	0.44	1.06E-04
Total		1537	1003	0.993	6.747	7.536	0.494	1.12E-04

N: Number of monoclonal samples, S: Richness or number of unique genotypes, D: Simpson's diversity index, H': Shannon's diversity index, A: Average number of alleles among loci, H(exp): Expected Heterozygosity, pi(median): median nucleotide diversity among all coding regions in the parasite genome.

3.4.2 Population structure patterns in LAC

Pairwise IBD values using PvGTSeq (IBD_{PvGTSeq}) and WGS (IBD_{WGS}) showed a multimodal distribution (**Supplementary figure 3.1**). In a network analysis scanning different IBD thresholds, the number of clusters remained nearly constant (~40 clusters) when IBD_{PvGTSeq} was > 0.85 (**Supplementary figure 3.2**). Between 0.73 and 0.85, the number of clusters

increased slightly (~50 clusters). Below 0.73, the number of clusters declined steadily, reaching one when $IBD_{PvGTSeq}$ was 0.58.

As we lowered the $IBD_{PvGTSeq}$ threshold to evaluate cluster merging, we identified up to 10 critical points at which one or more merges occurred. The first two merges occurred at $IBD_{PvGTSeq} = 0.91$ and 0.89 , and both reflected unions among highly related clusters that had diverged through diversification processes (**Supplementary figure 3.3**). The third merge occurred at $IBD_{PvGTSeq} = 0.85$. Below this threshold, many merges reflected recombination between unrelated genetic groups, such as the merge of clonal groups CL-01, CL-04, and CL-02, where CL-04 is the product of recombination between CL-01 and CL-02 (**Figure 3.2**, **Supplementary figure 3.3**, and **Supplementary figure 3.4**), and these last two clonal groups belongs to two different ancestral populations (Mesoamerica and the pacific coast of Colombia respectively). We corroborated this result by scanning long IBD segments among these three lineages (the two parental lines and the recombinant lineage) (**Supplementary figure 3.5**) and analyzing the aggregation pattern with WGS information (**Supplementary figures from 3.6 to 3.9**). The critical points used to define clonal groups was lower with WGS than with $PvGTSeq$ -generated data (**Supplementary figure 3.6** and **3.8**). Although $PvGTSeq$ overestimates genetic relationships, the order or hierarchy of interactions among clusters (for example, merges by recombination) was similar between $PvGTSeq$ and WGS (**Supplementary figure 3.3** and **3.7**).

$$IBD_{PvGTSeq} \geq 0.86$$

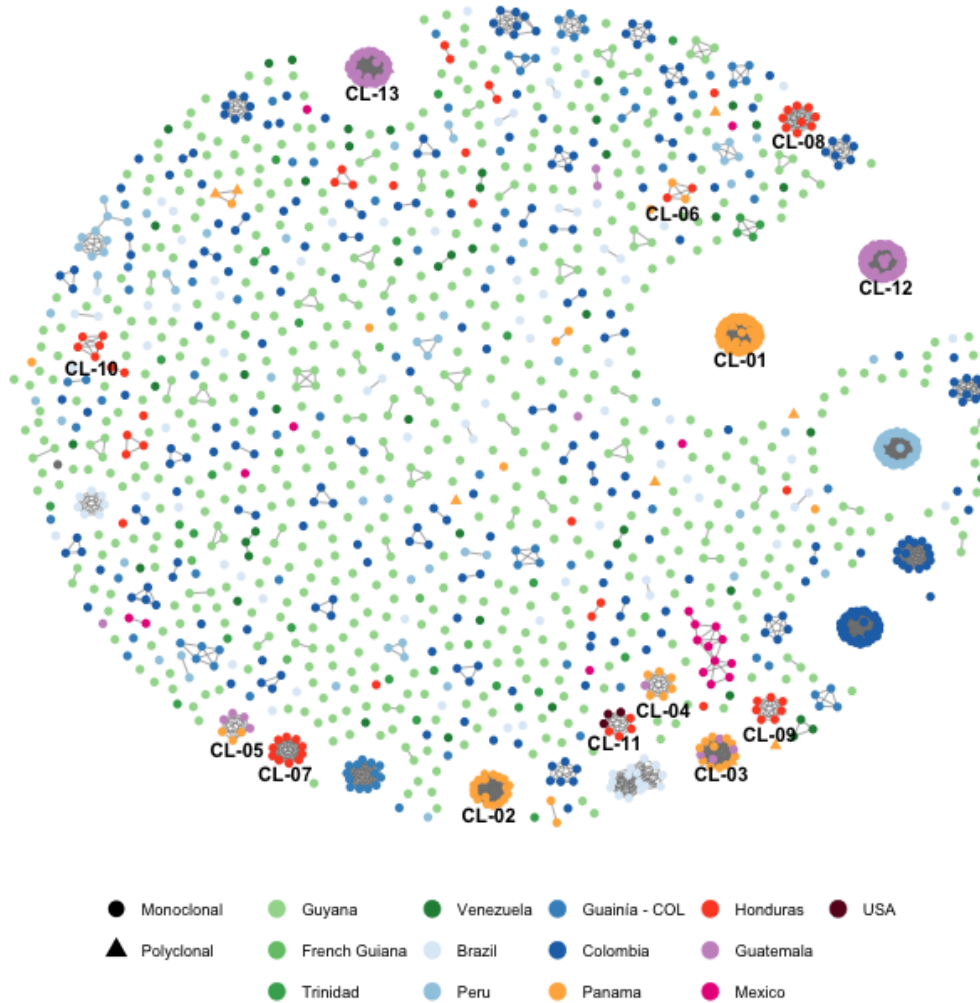


Figure 3.2: Identification of clonal and low-diversity groups using PvGTSeq genetic relatedness (IBD) in a network analysis. Each dot represents a genotyped sample. Colors are assigned based on the geographic origin of the sample (country). Shapes are based on the complexity of the sample infection. Polyclonal infections from Panama, Honduras, and Guatemala were only included in this analysis. Polyclonal infections from other countries were excluded. Branches represent pairwise IBD greater than 0.86. Clonal groups were defined as groups of four or more samples connected by branches. Labels above clusters are the labels used in the main text.

For this reason, we set $IBD_{PvGTSeq} \geq 0.86$ as the cutoff to define clonal groups or genetic groups with low diversity. In our set of 1,549 samples, we identified 41 clonal groups, and 32.1% (498/1,549) of all analyzed samples belonged to a clonal group (95% CI 29.9–34.5%). Most clonal groups were specific to a given country, with the exception of CL-03, CL-04, and

CL-05, which were shared between Panama and Guatemala; CL-06 was present in Panama and Honduras; and CL-11 was present in Honduras and in the three samples from the 2023 *P. vivax* outbreak in Florida, USA.

When we lowered the $IBD_{PvGTSeq}$ threshold from 0.86 to 0.74, we observed that all merging occurred between clusters within the same geopolitical subdivision (ecological groups within countries) (**Supplementary figures 3.10 and 3.11**). We therefore attributed merges at $IBD_{PvGTSeq} \geq 0.74$ to parasite population structure driven by geographic subdivision (ecological groups) (**Figure 3.2 and Supplementary figure 3.12**). This definition differs from a purely genetic definition of a population because panmixia within groups is not necessarily present, and it more closely reflects an ecological-epidemiological definition. Using this approach, we identified 52 ecological groups explained by geographic subdivision, and 47.1% of analyzed samples belonged to one of these clusters (CI95% from 44.7 to 49.6%). To assign the remaining 52.9% of the population, we matched each unclassified infection to the closest classified genotype. In Panama, Honduras, and Guatemala, we confirmed these assignments by searching for long IBD segments between infections.

We confirmed this aggregation pattern using WGS, and we found that branching and merging among nodes on the Colombian Pacific coast occur earlier than among nodes in the Guiana Shield region using WGS. With *PvGTSeq*, the opposite pattern is observed, indicating ascertainment bias toward the Pacific coast (**Supplementary figures 3.10 and 3.11**).

$$IBD_{PvGTSeq} \geq 0.74$$



Figure 3.2: Identification of ecological groups using PvGTSeq. Each dot represents a genotyped sample. Colors are assigned based on the geographic origin of the sample (country). Shapes are based on the complexity of the sample infection. Polyclonal infections from Panama, Honduras, and Guatemala were only included in this analysis. Polyclonal infections from other countries were excluded. Branches represent pairwise IBD greater than 0.86. Clonal groups were defined as groups of four or more samples connected by branches. Labels above clusters are the labels used in the main text.

PCoA and ADMIXTURE analyses using PvGTSeq confirmed population structure driven by geographic subdivision, with the exception of clonal groups CL-02, CL-03, and CL-04 in Panama, which these analyses suggest originated on the Colombian Pacific coast (Supplementary figures 3.13 - 3.16). The same pattern was observed in the network analysis when using critical values of $IBD_{PvGTSeq} \leq 0.63$ or $IBD_{WGS} \leq 0.36$. We also observed that

parasites from the Guiana Shield (Guyana, French Guiana, Trinidad and Tobago, Venezuela, and the Guainía region in Colombia) behave as a single panmictic population, and that clonal group CL-05 in Panama and Guatemala has an ancestral origin in that region (**Supplementary figure 3.16**). Pairwise genetic relatedness between all clonal and ecological groups found in the data set are shown in **Supplementary figures 3.17** and **3.18**. This pattern is also supported by the WGS network analysis, which links CL-05 to parasites in Trinidad and Tobago, where *P. vivax* cases occur mainly in migrants from Venezuela. Our WGS-based network analysis also shows slight population differentiation between Venezuela and Guyana, a pattern not captured by PvGTSeq or by PCoA or ADMIXTURE analyses (**Supplementary figure 3.11**).

3.4.3 Spatial and temporal distribution of clonal and ecological groups

Transmission in Mesoamerica is dominated by clonal propagation. This region showed the highest percentage of infections belonging to a clonal or low-diversity group (13 clonal groups, 83.1%, 95% CI, 78.6-86.8%), and 90.9% belonged to ecological groups (6 ecological groups, CI95% 87.2-93.6%). Moreover, 88.9% (CI95%, 84.8% - 92%) of infections are caused by clonal groups that belong to the autochthonous ancestral Mesoamerican population, clonal groups that have been introduced before 2020, or clonal groups that are the product of admixture between autochthonous and recently introduced groups.

In Panama, we found evidence of introductions of genotypes from the Colombian Pacific coast region and from the Guiana Shield. In this country six clonal groups were observed (CL-01 to CL-06), and all but CL-03 were assigned to the same ecological group PAN-02 (**Figure 3.2**). CL-01 was the most prevalent (50.4%, CI95% 41.5-59.3%), the most widely spread (**Figure 3.3**), the oldest clonal group in Panama (from 2004 to 2024) (**Figure 3.4**), and it belongs to the

ancestral Mesoamerican lineage (**Supplementary figure 3.16**). CL-02 had a prevalence of 15.4% (95% CI, 10-23%) and was distributed mainly in the southeastern region of the country. This group belongs to the ancestral population of the Colombian Pacific coast, but the presence of recombinants between CL-02 and CL-01 since 2013 (PAN-02.1 and CL-04 in **Supplementary figures 3.5** and **3.19**), suggest the presence of CL-02 in Panama before that year.

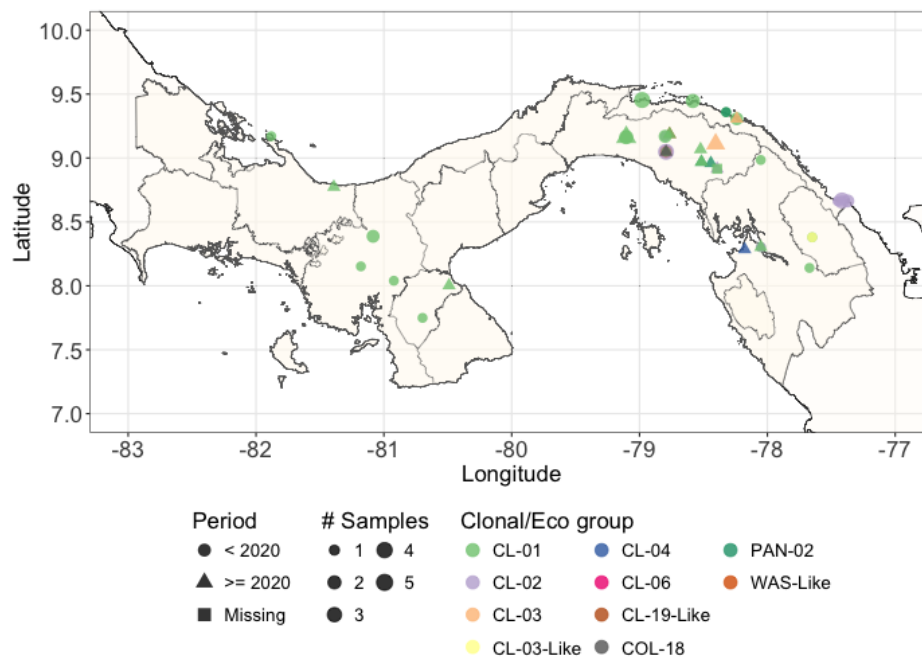


Figure 3.4: Geographic distribution of clonal and ecological groups in Panama.

A possible recently introduced clone in Panama is CL-03. It was observed in twelve infections (10.3%, 95% CI, 6-17.1%), and only four of them have information on site and date of collection, and the first reported infection was in the community of Playon Chico in Guna-Yala in May 2023, and three infections reported in Rio Sabalo in the Madugandi region in May 2024. Moreover, we found one sample with 100% IBD_{WGS} against CL-03 in Buenaventura, Colombia,

in 2020, and other samples and clonal groups in Cordoba and Buenventura in Colombia showed long IBD_{WGS} segments with respect to CL-03 (**Supplementary figure 3.20**), which indicates a recent introduction of this clone in Panama.

Another introduced clone is CL-05, which consists of three infections. One sample did not report site or date of collection, and the other two were collected in Darien (name of village not provided) in 2024. It originated in the Guiana Shield region, and the genotype closest to this group was from Trinidad and Tobago in 2019 - country where all cases are imported from Venezuela - and has an IBD_{WGS} of 0.27 (**Supplementary figure 3.21**). We also observed a genotype resulting from recombination between CL-05 and CL-04 (**Supplementary figure 3.22**). This genotype also lacks site or date of collection information, but the presence of this recombinant suggests that CL-05 is being transmitted locally in Panama.

The last clonal group observed in Panama was CL-06, with two samples, one from 2016 and the second with no reported date or site of collection. This group was also reported in two other samples in Honduras in 2021 and belongs to the HND-02 ecological cluster, indicating that CL-06 has had more observed interactions with the parasites from Honduras than with those from Panama, thus making it hard to establish the directionality of the split. Finally, WAS-Like was composed of two infections in migrant people from India. The abbreviation WAS indicates that the genetic analysis suggests an origin in West Asia, including India. The possible origin of other haplotypes observed in Panama as singletons are in **Supplementary figures 3.23-3.25**.

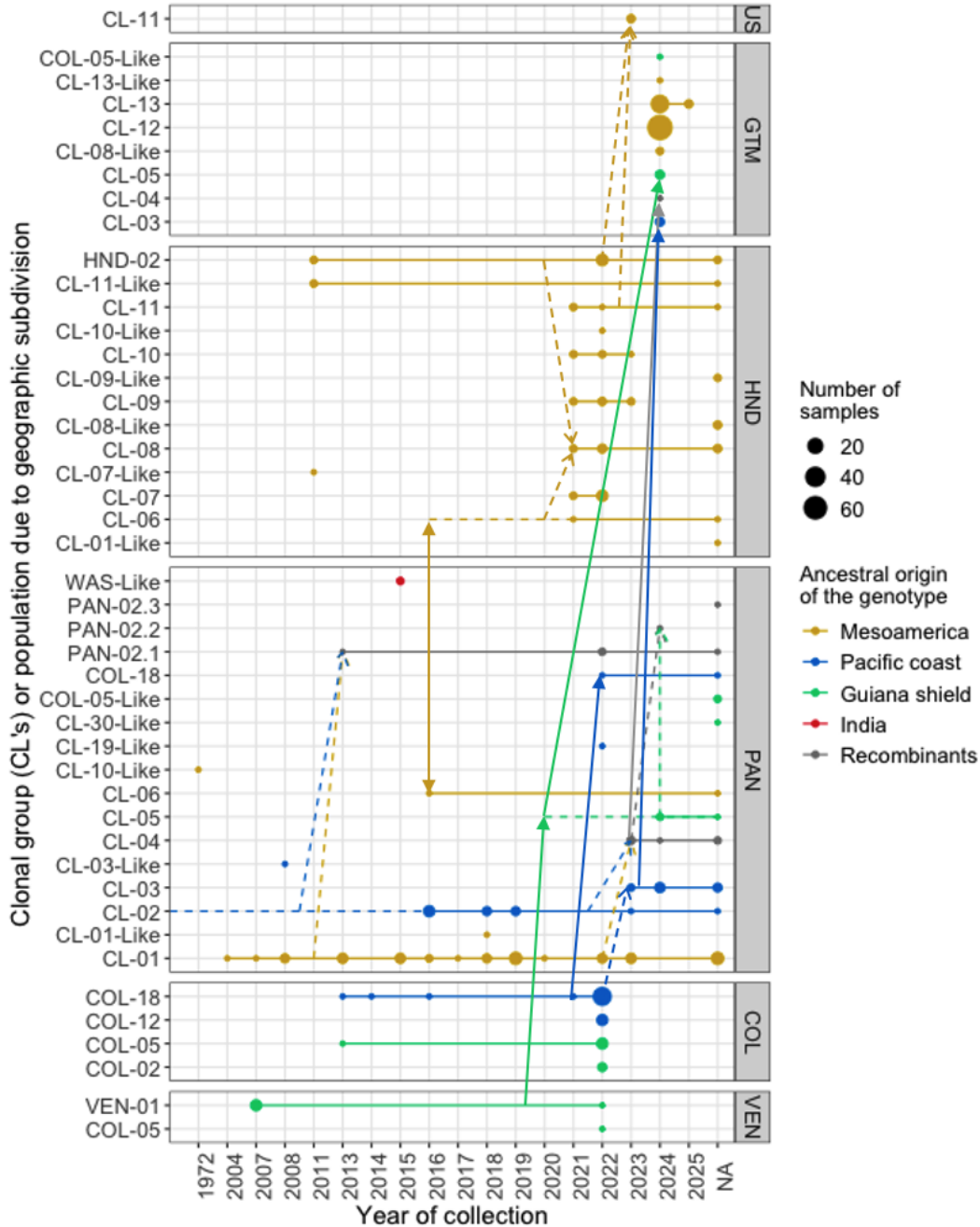


Figure 3.5: Temporal distribution of clonal and ecological groups in Mesoamerica. Y-axis represents clonal clusters (CLs), groups of closely related parasites sustained by inbreeding, or parasite ecological groups, defined as genotypes present in the same area of interest that show recent introgression events among members, with or without similar ancestral origins. Dots and their sizes represent the number of samples with a given parasite clone at a specific time (x-axis) and in a specific country (panels). Colors represent each clone's ancestral origin (Supplementary figure 3.16). Horizontal solid lines represent the time period for which there is evidence of a clone's presence in a country. Horizontal dashed lines represent the period for which there is suggestive evidence of a clone's presence in a country. Directional diagonal or vertical dashed arrows within panels represent recombination events between clonal groups. For example, although CL-01 and CL-02 have different ancestral origins, they belong to the same

Figure 3.4 (Continued):

parasite ecological group (PAN-02), because there are two clonal groups that are the product of outcrossing between CL-01 and CL-02 (CL-04 and PAN-02.1) and these recombinant have been observed since 2013, these evidences suggest the presence of CL-02 since 2013. Directional diagonal or vertical solid arrows between panels represent population split events. Arrow direction indicates the source and sink populations. Bidirectional vertical solid arrows represent population split events for which there is no consistent evidence of split directionality. For instance, CL-06 is observed twice in Panama in 2016 and twice in Honduras in 2021. However, CL-06 shows no outcrossing with any other clones in Panama, unlike in Honduras. NA's on the x-axis represent samples without date-of-collection information. The suffix -Like indicates that the observed genotype likely comes from the same ancestral population to which the clonal group belongs, but that there is an absence of long IBD blocks.

In contrast to Panama, the Gracias a Dios region in Honduras showed no evidence of parasites introduced from the south. In addition to the clonal group CL-06, six other clonal groups were observed, all belonging to two ecological populations, HND-01 and HND-02 and to the ancestral population of Mesoamerica. Several ungrouped genotypes or clonal groups appeared to be products of recombination among members of these ecological populations (**Supplementary figure 3.26 and 3.27**). One example is CL-08 which is the product of the outcrossing between CL-06 and a genotype that belongs to the ecological group HND-02 (**Supplementary figure 3.28**). Another example is the clonal group identified during the Florida outbreak in 2013, which resulted from recombination between CL-11 and a genotype belonging to the HND-02 ecological population (**Supplementary figure 3.29**). This suggest that Honduras may have served as a source of the *P. vivax* outbreak in Florida, United States.

In Guatemala, the most prevalent group was CL-12 (61%, 95% CI, 52% to 69.3%), with higher prevalence in the northern region of the country, where it accounted for 78.4% (95% CI, 68.7% to 85.7%) of cases. The second most frequent clonal group was CL-13 (28%, 95% CI, 20.7% to 36.7%), which had higher prevalence in Escuintla (75%, 95% CI, 53.1% to 88.8%) in the south. However, these two groups were present at all collection sites in Guatemala except Esquipulas in Chiquimula. In Esquipulas, clonal groups CL-03, CL-04, and CL-05 were

observed, and all of these cases were in migrant people who had recently entered the country from the south. CL-03 was observed in four individuals who arrived in Esquipulas between October 3th and November 24th in 2024, the migrant with CL-04 arrived in July 2024, and the four migrants with CL-05 arrived between July 17th and December 5th in 2024.

3.4.4 Contribution of imported cases to transmission in Mesoamerica

Based on our genetic classification, 85.6% (95% CI, 80.5–99.5%) of samples from Panama and Guatemala were classified as autochthonous, 6.8% as recently introduced (95% CI, 4.2–10.8%), and 7.7% as imported (95% CI, 4.9–11.8%). When we compared these results with travel history, we observed perfect concordance in Guatemala but not in Panama (**Table 3.3**). In Panama, we observed six cases in Venezuelan migrants before 2020 that were classified as imported; however, the genetic data indicate these cases were caused by clonal group CL-02, the autochthonous clonal group in the southeastern of Panama that has likely been present before 2013. Additionally, we observed six other cases for which the genetic data suggest importation (five originating from the Pacific coast and one from the Guiana Shield); however, these cases occurred in Panamanian individuals who did not report travel in the previous 15 days, so they may instead represent introduced cases whose frequency remains low in the population.

Country	Period	Genetic assignment	Travel History	# Sample	Percentage	Lower limit	Upper limit
Guatemala	>= 2020	Indigenous	Locally acquired	108	91.5%	85.1%	95.3%
		Imported	Imported	10	8.5%	4.7%	14.9%
Panama	>= 2020	Indigenous	Locally acquired	25	71.4%	54.9%	83.7%
		Introduced	Locally acquired	4	11.4%	4.5%	26.0%
		Imported	Locally acquired	6	17.1%	8.1%	32.7%

Table 3.3 (Continued)

Panama	< 2020	Indigenous	Locally acquired	50	84.7%	73.5%	91.8%
		Introduced	Locally acquired	1	1.7%	0.1%	9.0%
		Indigenous	Imported	6	10.2%	4.7%	20.5%
		Imported	Imported	2	3.4%	0.9%	11.5%

Bold indicates discrepancies between the genetic data and travel history

3.5 Discussion

Using genomic information from the *P. vivax* parasite, we identified the geographic origin of infections occurring in Panama, Honduras, and Guatemala, mainly between 2020 and 2024 but also in earlier years. To do so, we generated the most extensive catalogue of *P. vivax* genomic information from populations in Latin America and the Caribbean. Our results suggest that *P. vivax* transmission in Mesoamerica is clonal and sustained primarily by local transmission. We also observed clonal parasite groups in Panama that were recently introduced from the Pacific coast of Colombia (CL-03) and from the Guiana Shield (CL-05), as well as the importation of cases from Panama and Honduras to the north of the region, reaching Guatemala and the United States of America. These results highlight the vulnerability of migrants to infection during transit through the region and show that malaria control in these countries depends mainly on case management and internal vector control.

Over the last 25 years, Mesoamerica has experienced abrupt shifts in malaria transmission intensity,¹³ with consequences for the diversity of *P. vivax*. Between 2001 and 2005, Panama saw a rapid rise in cases, increasing from fewer than 1,000 per year to more than 5,000 in four years.⁶ In response, the government implemented the Vector Control Task Force, an intensive campaign based on indoor residual spraying (IRS) and active case detection. As a

result, cases declined to fewer than 500 in 2010 and 2011. The region is also characterized by spatially autocorrelated outbreaks concentrated in comarcas, administrative subregions formed by Indigenous communities, with high inter-community mobility.¹⁹ These factors help explain the persistence, limited genetic diversification, and broad geographic dispersal of clonal group CL-01, which appears in our data as early as 2004 and remains the dominant clone through 2024. Because this intervention strategy was not sustained, malaria control stagnated in subsequent years, and Panama recorded more than 500 cases per year between 2012 and 2018.

In addition, Mesoamerica is a major corridor for people from around the world who migrate to North American countries—the United States and Canada—in search of better opportunities. In this context, the Darién Gap plays an important role in both human migration and the importation and introduction of new parasite haplotypes into the region. Between 2000 and 2019, 361 imported cases based on travel history were recorded, with 81.6% originating in the Americas, particularly Colombia (48.6%).⁶ Furthermore, the number of migrants using the Darién Gap as an entry route has increased since 2011, rising from fewer than 500 that year to almost 30,000 in 2015 and 2016. This first migratory wave consisted mainly of Haitian and Cuban people coming from Brazil and Chile. In subsequent years, and during COVID, the flow of people decreased; however, in 2022 a second wave of nearly 250,000 people was observed, this time Venezuelans coming not only from Venezuela but also from other South American countries.^{8,10}

These characteristics make eastern Panama highly vulnerable to imported infections. Previous genetic studies have shown that parasite populations in the comarcas east of the Panama Canal (Guna Yala, Madugandí, and Wargandí) are more diverse than those in the west (Ngäbe-Buglé comarca and Veraguas province).²⁰ In a previous study by our research group, they

identified not only clonal group CL-01 but also CL-02 in samples collected in Puerto Obaldía (on the border with Colombia in Guna Yala) between 2017 and 2019.²¹ In addition, they identified four other samples that did not cluster with either of these clonal groups and were dispersed across eastern Panama. In our current study, we also detected CL-02 and confirmed that it represents an introduction from the Pacific coast of Colombia. However, our data further show that this clonal group is more widely distributed across eastern Panama and has been present since at least 2013.

Additionally, we identified a larger number of clonal groups, some produced by outcrossing between the indigenous groups CL-01 and CL-02 (CL-04 and PAN-02.1), other clonal groups recently introduced into the region from the Pacific coast and the Guiana shield (CL-03 and CL-05, respectively), and others produced by outcrossing between indigenous and introduced groups (outcrossing between CL-04 and CL-05). We also observed five haplotypes in single infections with no travel history whose ancestral origin is on the Colombian Pacific coast (CL-03-Like, CL-19-Like, and COL-18) or in the Guiana Shield region (COL-05 and CL-30-Like). This confirms the vulnerability of this region and suggests that the area is highly receptive to transmission of both local and non-local parasites. It also suggests that transmission intensity is sufficiently high to increase diversity through outcrossing among circulating parasites, whether indigenous or introduced. In our data, the percentage of polyclonal infections increased starting in 2020, rising from 6.8% to 15.6%; however, our small sample size did not allow us to confirm whether these differences are statistically significant.

Another factor that may contribute to the outcrossing observed in Panama and Honduras is *P. vivax* hypnozoites—the dormant liver stages that can remain latent and reactivate weeks, months, or even years after the initial infection. A new infection can trigger the activation of

heterologous hypnozoites in the liver (hypnozoites whose haplotype differs from that of the new infection), which may in turn promote outcrossing when gametocytes from both lineages are taken up by a mosquito.

In Panama, alongside the increase in cases since 2019, an increase in the number of recurrent cases has also been reported—recurrent episodes in patients from urban areas or communities that are not endemic for the parasite and who report no recent travel history to endemic communities.⁵ These relapses account for around 20% of all cases; however, this rise does not necessarily imply an increase in the relapse rate (i.e., the frequency with which an individual experiences relapse). Estimating relapse rates would require selecting a cohort of people from non-endemic areas who acquired recent infections while traveling to endemic regions and then following them to measure recurrence after they receive radical cure treatment.

From our genotyped cases, it is impossible to distinguish new infections from relapses. During sample collection, no mechanism was established to identify patients with recurrent episodes: only sample codes are available, and each sample is assumed to correspond to a new case. Even if that information were available, the clonal propagation of CL-01 and CL-02 would make it very difficult to determine whether a recurrent episode in a patient caused by the same parasite haplotype as the primary episode reflects relapse or a true new infection.

Our results also show that Panama and other Mesoamerican countries are not only vulnerable to imported infections but can also serve as a source of parasites for migrant populations who remain in these countries temporarily. Our data show that Venezuelan migrants who had recently entered Panama were infected with clonal group CL-02, which has been indigenous since at least 2013, yet these cases were incorrectly classified as imported. In

addition, in Guatemala we detected migrants infected with parasites indigenous to Panama (CL-04) or with parasite groups recently introduced into Panama (CL-03 and CL-05). Although it is possible that CL-03 and CL-05 were introduced directly from the Colombian Pacific coast and the Guiana Shield, this seems unlikely given the high diversity and the apparent absence of clonal propagation observed, at least in Venezuela. The four CL-05 cases reported in Esquipulas (Chiquimula) were spaced five months apart; although we do not have information about their point of departure from Venezuela, the most parsimonious interpretation is that these individuals acquired CL-05 infection in Panama, where this clonal group has been introduced and has recombined with indigenous parasites.

Honduras has also been shown to be a source of infections toward countries to the north, specifically the United States. In 2023, locally acquired cases were reported in the states of Florida, Texas, and Arkansas.^{22,23} Studies conducted at the time revealed a Central American origin for these introductions. In this study, we used genomic information from three of the infections reported in Florida²³ and traced their origin not only at the regional level, but also at the country level, successfully identifying the Honduran clonal group (CL-11) that served as the source for these three U.S. cases. The Gracias a Dios department is not considered part of the migration corridor used by people traveling north across the continent, so the geographic spread of the clonal group CL-11 is likely broader.

Being able to monitor the contribution of locally acquired cases relative to imported cases in the context of malaria elimination is an important pillar for stratifying risk areas and defining appropriate control and elimination strategies. Along these lines, our results show that the genetic characterization of retrospective malaria-positive samples is a necessary activity that complements case classification. However, it is important to note that this sampling strategy

introduces limitations in interpreting the results because the criteria for storing samples differ across study areas. As a result, sampling is often opportunistic and the data may not reflect the true diversity of endemic areas. In addition, many samples lack records of the collection date or site, which biases estimates of the prevalence of certain genetic groups and, therefore, precise estimates of indigenous, introduced, and imported cases. For example, in Panama, a group of 24 samples lacked information on collection date and site, and other samples contained only one of these two pieces of information. Most of these samples belonged to the clonal group CL-01, and they were not included in the prevalence estimates. Therefore, these results need to be corroborated with more representative sampling of the region. In addition, in this work we limited ourselves to case classification, and in the near future it is required to analyze whether newly introduced genetic groups carry mutations that may be relevant for decision-making in antimalarial treatment management.

Overall, our results show that the increase in malaria cases in Mesoamerica has been driven mainly by increased local transmission and the dispersal of parasite lineages that are indigenous to the region. In Panama, we also observed local transmission of recently introduced parasites (11.4%); however, indigenous parasites still account for 71% of total cases. The remaining 17.6% were classified as imported because they were singletons in Panama with genetic signatures from foreign countries. However, none of those cases reported travel history, so they might represent recent introductions. These findings highlight the difficult conditions migrant populations face during their journey through Mesoamerica. To provide humanitarian assistance and improve the registration of migratory transit through the Darién, the Panamanian government opened four temporary migrant reception stations (Lajas Blancas, San Vicente, Bajo Chiquito, Canaán Membrillo) starting in September 2020 and the last one opened in 2022.^{10,24}

However, reports from the United Nations International Organization for Migration (IOM) indicate that up to 97% of migrants spend the night outdoors for two to three weeks and have limited or no access to health services before reaching these stations.¹⁰ As a result, communities such as Puerto Limón, Puerto Obaldía, La Miel, Armila, and Aswemullu—located near this migration route and endemic for malaria—are a source of malaria infections for people who pass through this corridor.¹⁹

Since the change in the U.S. administration in February 2025, the number of migrants crossing the Darién has dropped dramatically to fewer than 100 people per month.⁹ For this reason, the Panamanian government announced in March 2025 the progressive closure of these stations.²⁵ However, the comarcas in eastern Panama remain an active malaria transmission focus, with more than 9,000 malaria cases in 2025. It is therefore necessary to intensify vector-control activities and case management in this region, and to provide adequate health services at border checkpoints to reduce the harms migrants face along this corridor and to prevent the spread of new parasite haplotypes whose phenotypic characteristics remain uncertain.

3.6 References

1. World malaria report 2024. <https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2024>.
2. Grillet, M. E. *et al.* Malaria in Southern Venezuela: The hottest hotspot in Latin America. *PLoS Negl. Trop. Dis.* **15**, e0008211 (2021).
3. Malaria 2024 Honduras country profile. <https://www.who.int/publications/m/item/malaria-2024-HND-country-profile>.
4. Malaria 2024 Guatemala country profile. <https://www.who.int/publications/m/item/malaria-2024-gtm-country-profile>.
5. Malaria 2024 Panama country profile. <https://www.who.int/publications/m/item/malaria-2024-pan-country-profile>.

6. Hurtado, L. *et al.* Long-term transmission patterns and public health policies leading to malaria elimination in Panamá. *Malar. J.* **19**, 265 (2020).
7. Global technical strategy for malaria 2016-2030, 2021 update.
<https://www.who.int/publications/i/item/9789240031357>.
8. Estadísticas: Tránsito irregular por el Darién 2022 – Migración Panamá.
<https://www.migracion.gob.pa/estadisticas/> (2023).
9. Estadísticas: Tránsito irregular por el Darién 2025 – Migración Panamá.
<https://www.migracion.gob.pa/estadisticas/> (2023).
10. International Organization for Migration (IOM), U. N. (UN). Displacement Tracking Matrix. <https://dtm.iom.int/report-product-series/other-30>.
11. Refugio de Esperanza: Historias de migrantes en el Centro. *IOM (NTCA)*
<https://nortedecentroamerica.iom.int/es/stories/refugio-de-esperanza-historias-de-migrantes-en-el-centro-sagrada-familia-de-honduras> (2025).
12. Centros de Atención para Personas Migrantes y Refugiadas (CAPMiR). *ACNUR Guatemala* <https://help.unhcr.org/guatemala/acnur-en-guatemala-2/centros-de-atencion-capmir/>.
13. WHO malaria terminology, 2021 update.
<https://www.who.int/publications/i/item/9789240038400>.
14. Godoy, A. *et al.* Molecular Characterization of Plasmodium Species to Strengthen Malaria Surveillance in Migrant Populations in Honduras. *Trop. Med. Infect. Dis.* **10**, 292 (2025).
15. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
16. Amos, B. *et al.* VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res.* **50**, D898–D911 (2022).
17. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
18. Taylor, A. R., Jacob, P. E., Neafsey, D. E. & Buckee, C. O. Estimating Relatedness Between Malaria Parasites. *Genetics* **212**, 1337–1351 (2019).
19. Cumbreira, A., Calzada, J. E., Chaves, L. F. & Hurtado, L. A. Spatiotemporal Analysis of Malaria Transmission in the Autonomous Indigenous Regions of Panama, Central America, 2015–2022. *Trop. Med. Infect. Dis.* **9**, 90 (2024).
20. Santamaría, A. M. *et al.* Plasmodium vivax Genetic Diversity in Panama: Challenges for Malaria Elimination in Mesoamerica. *Pathogens* **10**, 989 (2021).

21. Buyon, L. E. *et al.* Population genomics of *Plasmodium vivax* in Panama to assess the risk of case importation on malaria elimination. *PLoS Negl. Trop. Dis.* **14**, e0008962 (2020).
22. Barratt, J. L. N. *et al.* Genetic characterization of *Plasmodium vivax* linked to autochthonous malaria transmission in the US (2023) using Illumina AmpliSeq technology: a genetic epidemiology study. *Lancet Reg. Health – Am.* **48**, (2025).
23. Muneer, A. *et al.* Autochthonous *Plasmodium vivax* Infections, Florida, USA, 2023. *Emerg. Infect. Dis.* **30**, 1214–1217 (2024).
24. estaciones temporales de recepción migratoria en panama - Google Search.
<https://cnnespanol.cnn.com/2022/11/26/panama-inaugura-nueva-estacion-para-recepcion-temporal-de-migrantes-orix>.
25. Yahila. CIERRE TOTAL DE LA ESTACIÓN TEMPORAL DE RECEPCIÓN MIGRATORIA DE LAJAS BLANCAS – Migración Panamá.
<https://www.migracion.gob.pa/cierre-total-de-la-estacion-temporal-de-recepcion-migratoria-de-lajas-blancas/> (2025).

Chapter Four: Discussion

This thesis presents a large-scale genomic analysis of *P. vivax* in Latin America and the Caribbean. In Chapter Two, I introduce PvGTSeq and PvCRiSP, two affordable Illumina-based molecular-surveillance approaches that use non-proprietary reagents. The 249 amplicons of PvGTSeq provide high fine-scale spatial resolution for assigning the origin of infections in LAC, as well as for surveillance of ten drug-resistance gene candidates. PvCRiSP is a simplified method based on only four amplicons. It does not require laborious and expensive pre-amplification steps, and the amplicons are highly diverse, enabling estimation of complexity of infection (COI) and identification of clonal transmission. In Chapter Three, I evaluate whether the resurgence of *P. vivax* in Panama, Honduras, and Guatemala from 2020 onward is due to transmission of local parasite populations or the introduction of foreign parasite populations from the south. Our results suggest that, despite the presence of imported and introduced cases in Guatemala and Panama, the autochthonous *P. vivax* parasite population remains largely unchanged. These findings are crucial for guiding local authorities in managing migrant populations, who often become vulnerable to local parasite strains. They also suggest that control and elimination depend mainly on sustained case management and vector control.

These findings highlight the importance of systematically and continuously cataloguing the region's genetic diversity to answer questions that inform decision-making for disease control and elimination. In this work, we focused on supporting the classification of case origin and distinguishing between locally acquired cases (autochthonous or introduced) and imported cases, but the potential applications extend further: identifying and evaluating the spread of genetic variants with phenotypic characteristics of interest (e.g., antimalarial drug resistance), delineating transmission foci, and monitoring the effectiveness of interventions. These applications were discussed at the first meeting of the regional network for malaria molecular

surveillance (MMS Americas) in August 2025.¹ MMS Americas aims to promote the sharing of experiences, protocols, and molecular data to support decision-making for malaria control and elimination at the regional level. However, the epidemiological characteristics of the region add complexity to this goal in at least four areas: 1) selecting the most appropriate sampling strategies, 2) choosing data-generation tools, 3) choosing data-analysis tools, and 4) harmonizing data from different sources.

In hyperendemic countries or subnational regions with clear evidence of panmixia, simple random sampling is typically sufficient to obtain genetically relevant, non-redundant information. Depending on the objective, investigators may choose either whole-genome sequencing or amplicon sequencing. In these settings, however, the prevalence of polyclonal infections must be above 30%, so initial screening with a short amplicon panel would help optimize costs.

In contrast, in hypoendemic regions it is first necessary to identify the main transmission foci, or hotspots. As we observed in Mesoamerica and on the Colombian Pacific coast, parasite diversity within hotspots is expected to be low and may even reflect clonal propagation;^{2,3} however, different clonal groups are likely circulating between transmission foci (or even between hotspot within transmission foci). In such cases, sampling should first be stratified using hotspots as the sampling unit. Then, because clonal expansion reduces the value of sequencing many similar infections, whole-genome sequencing of all samples from a hotspot would not provide much additional information unless the goal is to quantify *de novo* mutations within the clonal group to estimate how long it has been propagating (a procedure that remains highly uncertain even with genomic data).⁴ Similarly, large amplicon panels would not be cost-effective. In these scenarios, it would be advisable to first screen with small amplicon panels to

guide sample selection. Then, one sample per unique haplotype detected by these panels would be sequenced, either with a longer amplicon panel or with whole-genome sequencing.

Establishing a single analysis framework is complicated when clonal propagation and limited admixture are present. It is also important to recognize that the genetic definition of a population is not the same as the epidemiological definition. In the genetic definition, random mating is an intrinsic characteristic of a population; under the epidemiological/ecological definition, this is not necessarily true. Traditionally, sources of infection are assigned through population-structure analyses using tools such as PCoA or STRUCTURE - Like methods. These tools are appropriate for identifying ancestral populations and assume that mating within each subpopulation is random (panmixia). As a result, repeated genotypes due to clonal expansion can bias these results by effectively treating each clonal group as a subpopulation, making the relationships among subpopulations (clonal groups) difficult to interpret. Conversely, if only a representative member of each clonal group is used, the ancestral pattern of population subdivision can be identified and the genetic relationships among haplotypes become clearer, but the epidemiological interpretation becomes more diffuse because these analyses alone do not explicitly show the interactions between clonal groups with different ancestry that coexist in the same time period and territory.

In the absence of a holistic method that accounts for clonal propagation as an intrinsic part of this parasite's biological cycle, we staged the analysis in steps and used network analysis, corroborated by a search for long IBD segments. This strategy allowed us not only to identify clonal groups, but also to detect outcrossing among their members and, in doing so, to derive an epidemiologically meaningful population definition ("ecological groups") from the genetic data. For example, clonal group CL-02 belongs to the ancestral population from Colombia's Pacific

coast, and epidemiologic information indicates that it has been present since 2016. However, the network analysis revealed haplotypes generated by outcrossing between CL-02 and CL-01 in Panama since 2013, and no evidence of outcrossing between CL-02 and haplotypes present on Colombia's Pacific coast. This suggests not only that CL-02 is part of the Panamanian ecological population, but also that it is no longer part of the Colombian population. In addition, these results allowed us to trace the entry of CL-02 into Panama back to 2013, even though the first genotyped cases are from 2016.

Initiatives to develop holistic methods that account for the parasite's biology are already underway.⁵ These efforts are converging on the use of ancestral recombination graphs under the framework of coalescent theory—an approach analogous to the empirical strategy used in our work, but with a more robust mathematical formulation. Future advances in this area will make it possible to jointly model epidemiological and genomic processes and enable a more direct epidemiological interpretation of genetic findings.

One of the most basic—but also most complex—tasks in building a genetic data bank is to systematize the collection of the clinical/epidemiological information associated with each biological sample.^{1,6} This process spans the full chain of events, from sample collection to the integration of clinical/epidemiological and laboratory data: sample collection, administration of clinical/epidemiological surveys, shipment of samples to local laboratories in the countries of origin, shipment to reference centers for genetic analysis, generation of genetic data, and, finally, integration of all information for the generation of reports to policy makers. It is important to recognize that retrospective sample sources often use different strategies to collect samples and associated information; these differences should be clearly documented so that the scope of future interpretations can be appropriately delimited.

A starting point for this harmonization is to establish clear rules for coding sample identifiers.^{7,8} This system must make it possible to determine whether a patient has repeated samples and, if so, the reason for the repetition: monitoring disease progression, recurrent malaria episodes, biological or technical replicates, or repeating assays after an initial laboratory failure. Data-entry errors should also be minimized by implementing mechanisms that promptly identify when such errors occur. Finally, it is important to define the minimum epidemiological information needed to provide sufficient granularity for analysis—i.e., spatial and temporal information that efficiently delineates transmission foci (hotspots) and the duration of epidemic outbreaks. In this study, for example, clonal group CL-05 is suggested to have been imported and introduced into Panama from the Guiana Shield, and later imported from Panama into Guatemala. This interpretation is supported by migrants in Guatemala reporting symptom onset at different times, evidence of recombination between CL-05 and clones that are autochthonous to Panama (suggesting clonal spread in that region), and the high genetic diversity in Venezuela—even at the local level—making it unlikely that two people would be infected with the same haplotype. However, none of these arguments could be made without the epidemiological information for each case.

This is the first large-scale genomic study of *P. vivax* in the LAC region. It is inspired by, and aims to complement, genomic surveillance efforts for this parasite carried out by the MalariaGEN *Plasmodium vivax* Genome Variation Project in other regions of the world.⁹ It also aims to serve as a foundation and to provide concrete examples of how genetic and genomic data can support the new MMS Americas network.

In Chapter 2, we presented two practical tools for generating genetic data: PvGTSeq and PvCRiSP. These tools aim to make molecular surveillance more cost-effective for public health

decision-making. In this work, we compared PvGTSeq with two similar methods designed to assign the source of infections at a global level, and we showed that PvGTSeq has higher resolution for discriminating genotypes in Latin America. In addition, both methods improve the characterization of low-parasite-density infections, which represent more than 70% of *P. vivax* infections in the region.

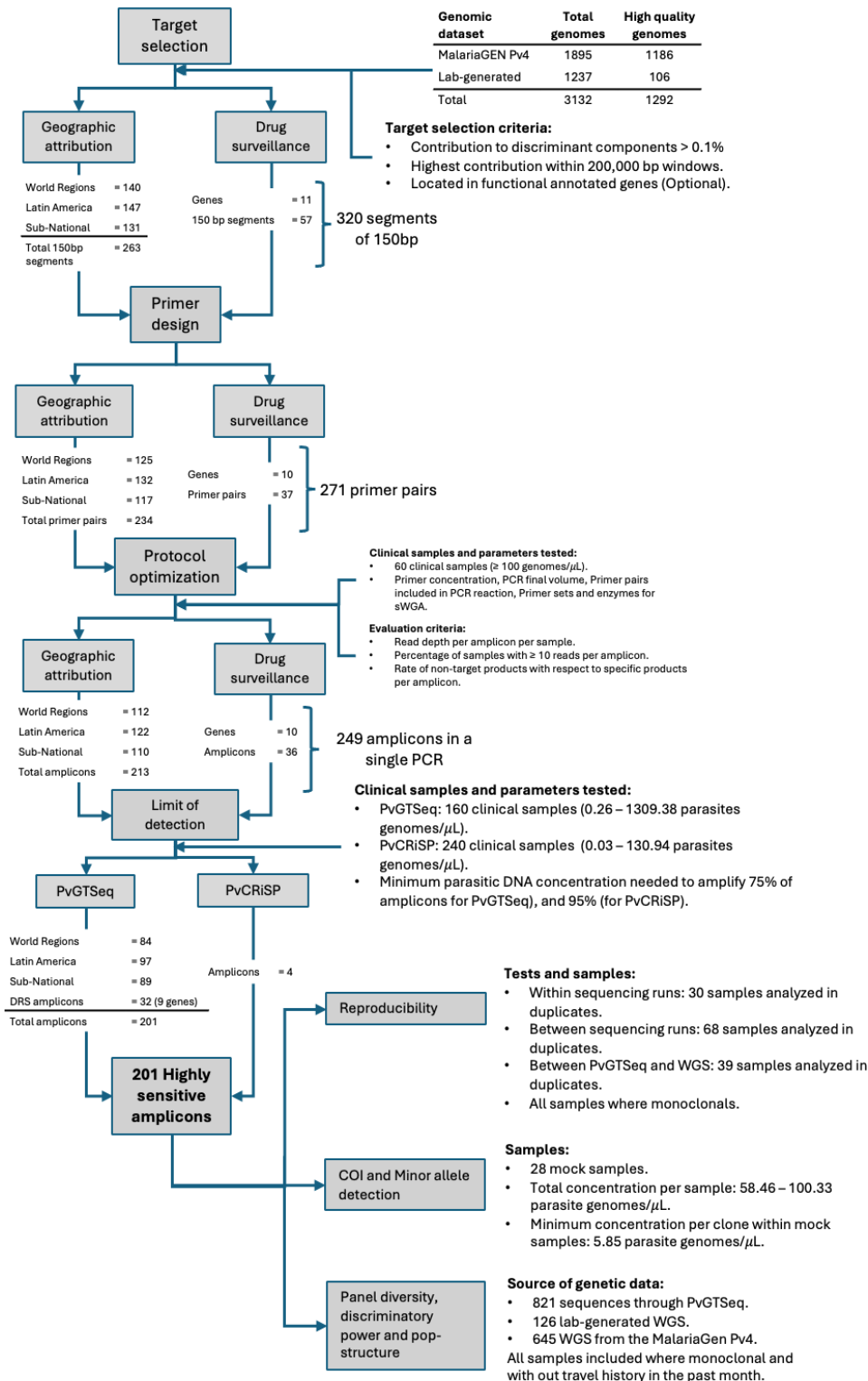
In Chapter 3, we present one practical use of *P. vivax* genomic surveillance: identifying the origin of infections. This is necessary in the current context of Mesoamerica, a region that aims to eliminate the disease but has seen a resurgence in recent years. Our results consistently indicate that the increase in *P. vivax* cases in Panama, Honduras, and Guatemala is driven mainly by local transmission of parasite populations that are autochthonous to the region, and that migrant populations increase diversity by introducing parasite groups from Colombia's Pacific coast and from the Guiana Shield. Moreover, these migrants are a vulnerable population that can acquire infections during transit from south to north. This information is essential for the ministries of health in these countries to intensify vector control and case management.

Finally, we conclude by discussing challenges posed by the epidemiological characteristics of Latin America for generating large-scale genomic data cost-effectively and producing correct interpretations. Taken together, the work presented in this thesis provides concrete evidence that molecular surveillance of *P. vivax* already has a solid foundation for implementation at scale in the region and can provide crucial information for the control and elimination of this disease.

4.1 References

1. MMS Américas: a regional network for malaria molecular surveillance. *Malar. J.* **25**, 35 (2026).
2. Buyon, L. E. *et al.* Population genomics of *Plasmodium vivax* in Panama to assess the risk of case importation on malaria elimination. *PLoS Negl. Trop. Dis.* **14**, e0008962 (2020).
3. Sutanto, E. *et al.* Genomics of *Plasmodium vivax* in Colombia reveals evidence of local bottle-necking and inter-country connectivity in the Americas. *Sci. Rep.* **13**, 19779 (2023).
4. Redmond, S. N. *et al.* De Novo Mutations Resolve Disease Transmission Pathways in Clonal Malaria. *Mol. Biol. Evol.* **35**, 1678–1689 (2018).
5. Camponovo, F., Buckee, C. O. & Taylor, A. R. Measurably recombining malaria parasites. *Trends Parasitol.* **39**, 17–25 (2023).
6. Achidi, E. A. *et al.* A global network for investigating the genomic epidemiology of malaria. *Nature* **456**, 732–737 (2008).
7. Olden, M., Holle, R., Heid, I. M. & Stark, K. IDGenerator: unique identifier generator for epidemiologic or clinical studies. *BMC Med. Res. Methodol.* **16**, 120 (2016).
8. CDC. Collecting Data. *Field Epidemiology Manual* <https://www.cdc.gov/field-epi-manual/php/chapters/collecting-data.html> (2026).
9. Adam, I. *et al.* An open dataset of *Plasmodium vivax* genome variation in 1,895 worldwide samples. *Wellcome Open Res.* **7**, 136 (2022).

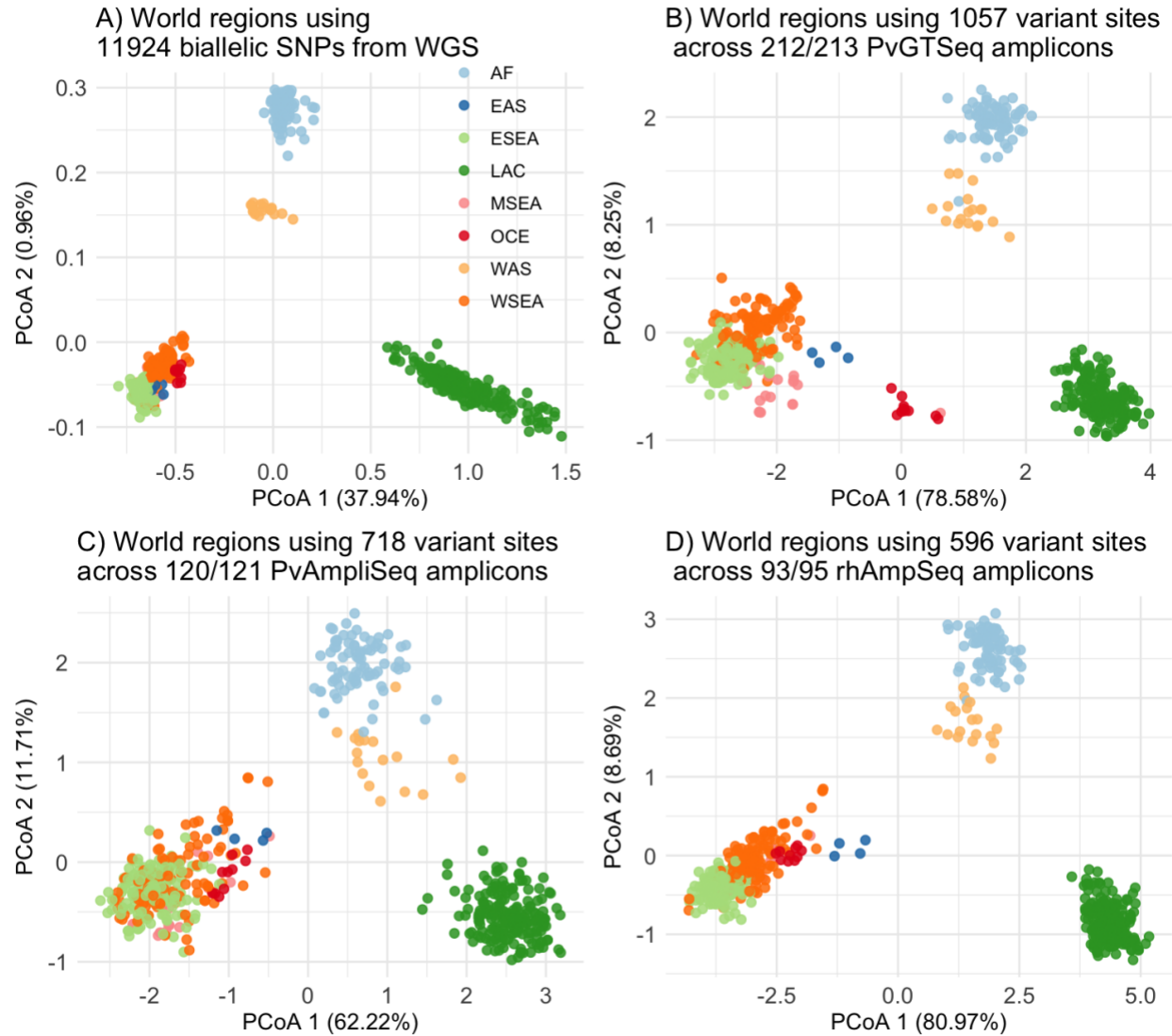
APENDIX



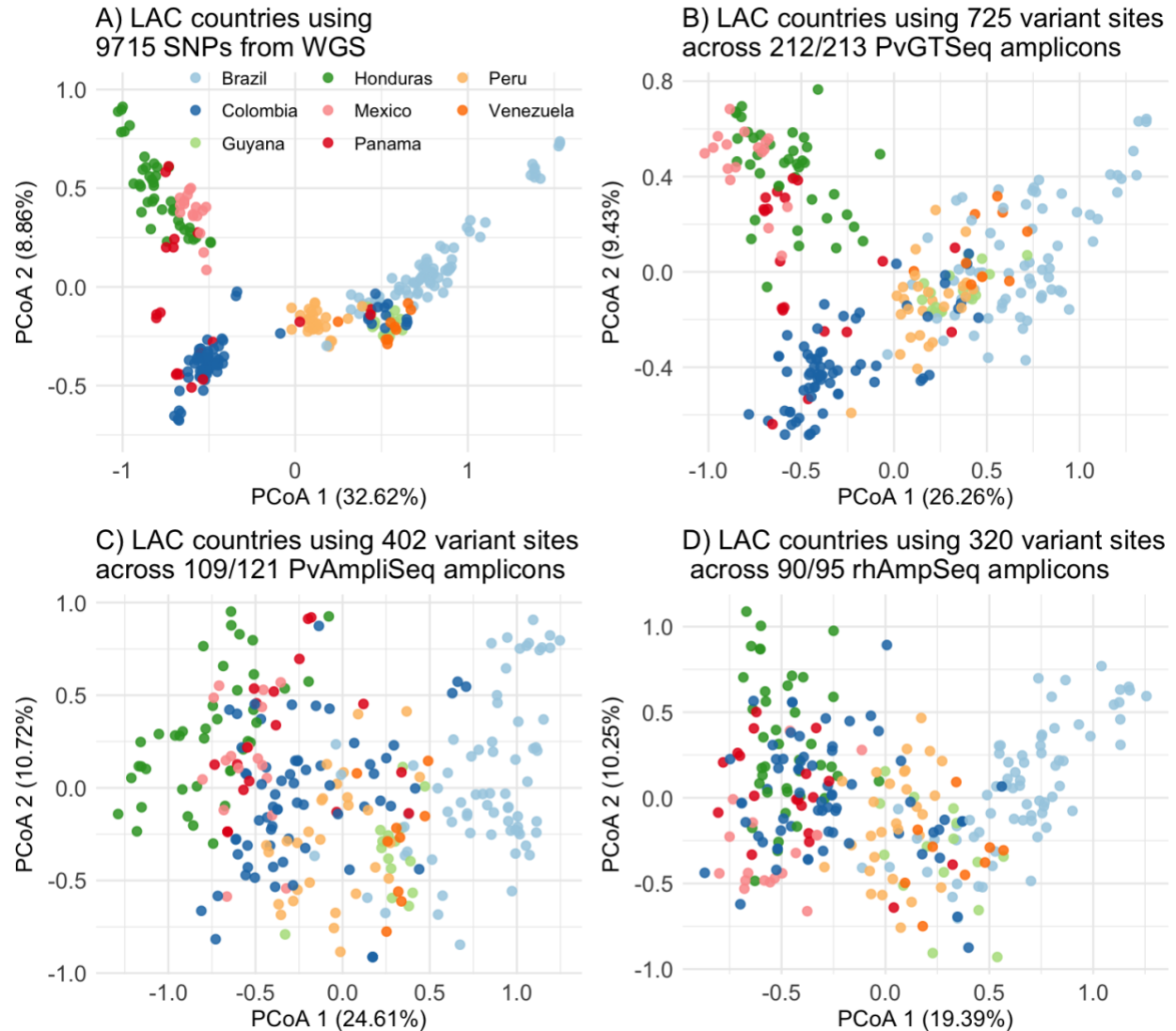
Supplementary figure 2.1: Flowchart of the selection, optimization and validation process for PvGTSeq and PvCRiSP. Each stage details the number of amplicons (segments or primer pairs) selected, the samples or genomic data used for the analysis, and the evaluation criteria applied.

Alignment position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	Pseudo-cigar string		
Reference position			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32		33	34
REFSEQ	.	.	A	A	C	T	A	C	A	G	T	G	G	A	.	.	.	.	.	.	.	C	T	T	C	G	T	C	C	T	G	A	.	.		.	
MHAP01	.	.	.	.	A	A	C	C	C	A	G	T	G	G	A	.	.	.	.	.	.	.	C	T	T	C	G	T	C	C	T	G	A	.		.	.
MHAP02	.	.	.	.	A	A	C	C	C	A	G	T	G	G	A	.	.	.	.	.	.	.	C	T	T	C	G	T	C	C	T	G	A	.		.	.
MHAP03	.	.	.	.	A	A	C	C	C	A	G	T	G	G	A	.	.	.	.	.	.	.	C	T	T	C	G	T	C	C	T	G	A	.	.	.	
MHAP04	.	.	.	.	A	A	C	C	C	A	G	T	G	G	A	.	.	.	.	.	.	.	C	T	T	C	G	T	C	C	T	G	A	.	.	.	
MHAP05	A	C	A	T	C	T	A	C	C	C	A	G	T	G	G	A	.	.	.	.	.	.	C	T	T	C	G	T	C	C	T	G	A	.	.	.	
MHAP06	A	G	A	A	C	T	A	C	C	C	A	G	T	G	G	A	.	.	.	.	.	.	C	T	T	C	G	T	C	C	T	G	A	.	.	.	
MHAP07	G	C	A	A	C	T	A	C	C	C	A	G	T	G	G	A	.	.	.	.	.	.	C	T	T	C	G	T	C	C	T	G	A	.	.	.	
																																		.	1D=A12I=ACGTCGT24I=GAA		
																																		.	179I=TCGTAGT10D=GGA14G17C24I=GAT		
																																		.	1D=AA3A9I=TCGTCGT10D=GGA14G17C24I=GTA		
																																		.	0I=AC2T12I=ACGT24I=TAA		
																																		.	0I=AG12D=ACTTCG23D=TG		
																																		.	0I=G12D=ACTTCG22G23D=TG		

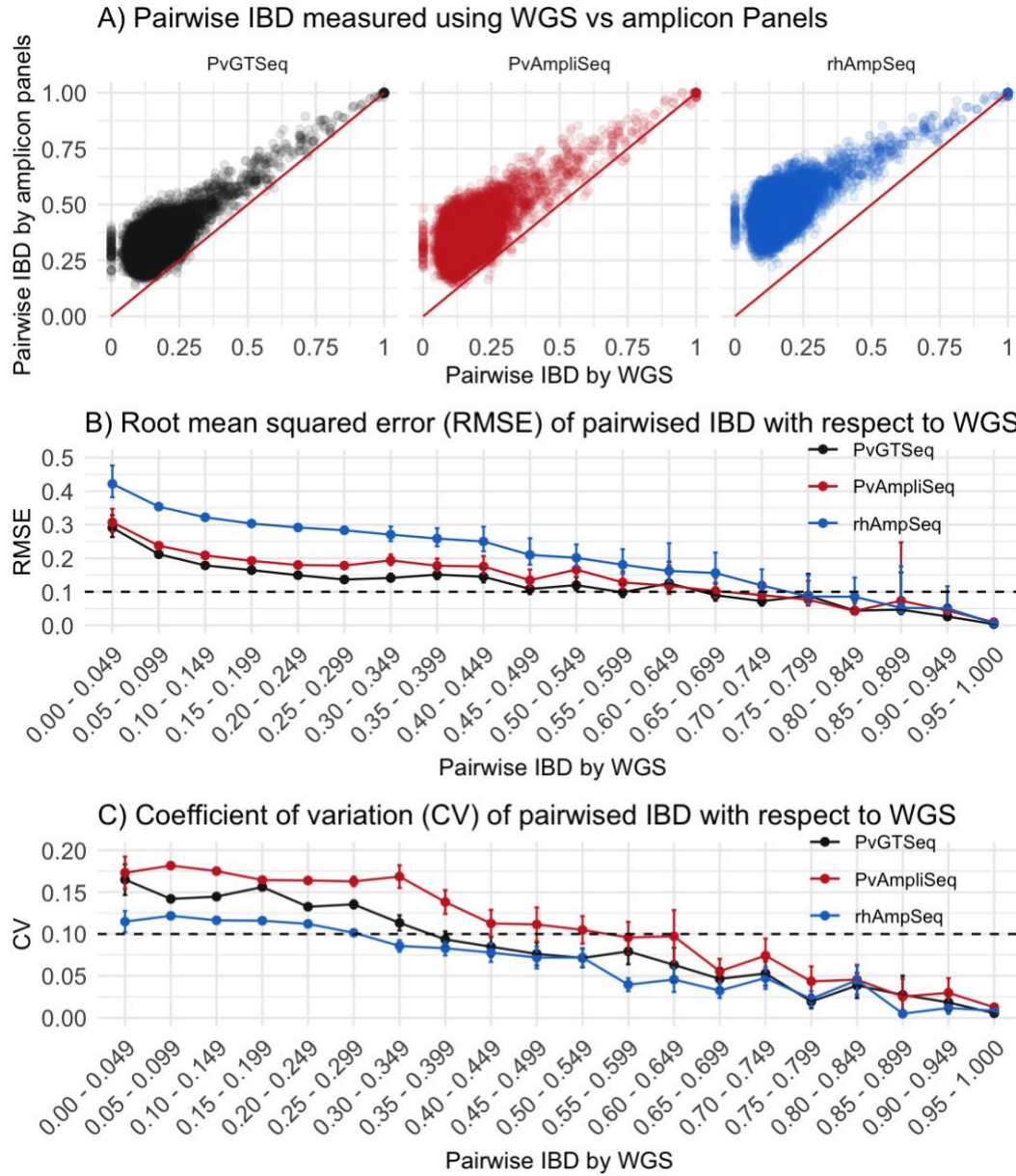
Supplementary figure 2.2: Examples of Pseudo-cigar representation of polymorphisms observed from multiple alignment of amplicon microhaplotypes with their reference sequence. In our analysis pipeline, after dada2 denoising of sequencing errors, all microhaplotypes are aligned to the reference sequence using MUSCLE (Multiple Sequence Comparison by Log-Expectation). Polymorphisms are summarized in Pseudo-cigar format following these rules: 1) All variants are annotated in ascending order. 2) For SNVs, we annotated the reference position followed by the substitute nucleotide. 3) For deletions, we annotated the starting position, "D=", then all deleted nucleotides (e.g., 23D=TG). 4) For insertions, we annotated the position before insertion, "I=", then the nucleotide at that position followed by inserted nucleotides. 5) If the nucleotide before an insertion is an SNV, we include this SNV within the insertion notation to avoid position duplication (see MHAP05). 6) When a deletion is followed by an insertion, the insertion is annotated first to maintain ascending order (see MHAP03 and MHAP04). 7) For insertions before position 1, we use position 0 and only annotate inserted nucleotides (see MHAP05-MHAP07). 8) If both reference and microhaplotype have deletions, but the microhaplotype deletion extends beyond the reference deletion, we consider it a single deletion (see MHAP06 and MHAP07). 9) If both, the reference and the microhaplotype, have deletions, but the reference deletion extends beyond the microhaplotype deletion, we consider it a single insertion (see MHAP05). 10) Microhaplotypes that are identical to the reference sequence are annotated using a period symbol "." (see MHAP01).



Supplementary figure 2.3: Principal coordinate analysis of world regions comparing whole genome sequencing (A), PvGTSeq (B), PvAmpliSeq (C) and rhAmpSeq (D). The analysis was done using genomic information from MalariaGEN Pv4, Brazilian genomes available from the European Nucleotide Archive (PRJEB56411, PRJEB44419, PRJEB36199) and 112 genomes from Colombia, Guyana, Honduras, Panama, Peru and Venezuela generated in our group. The analysis at world regions scale includes: Africa (AF), East Asia (EAS), Eastern Southeast Asia (ESEA), Latin America and Caribbean (LAC), Maritime South-East Asia (MSEA), Oceania (OCE), Western Asia (WAS), and Western South-East Asia (WSEA).



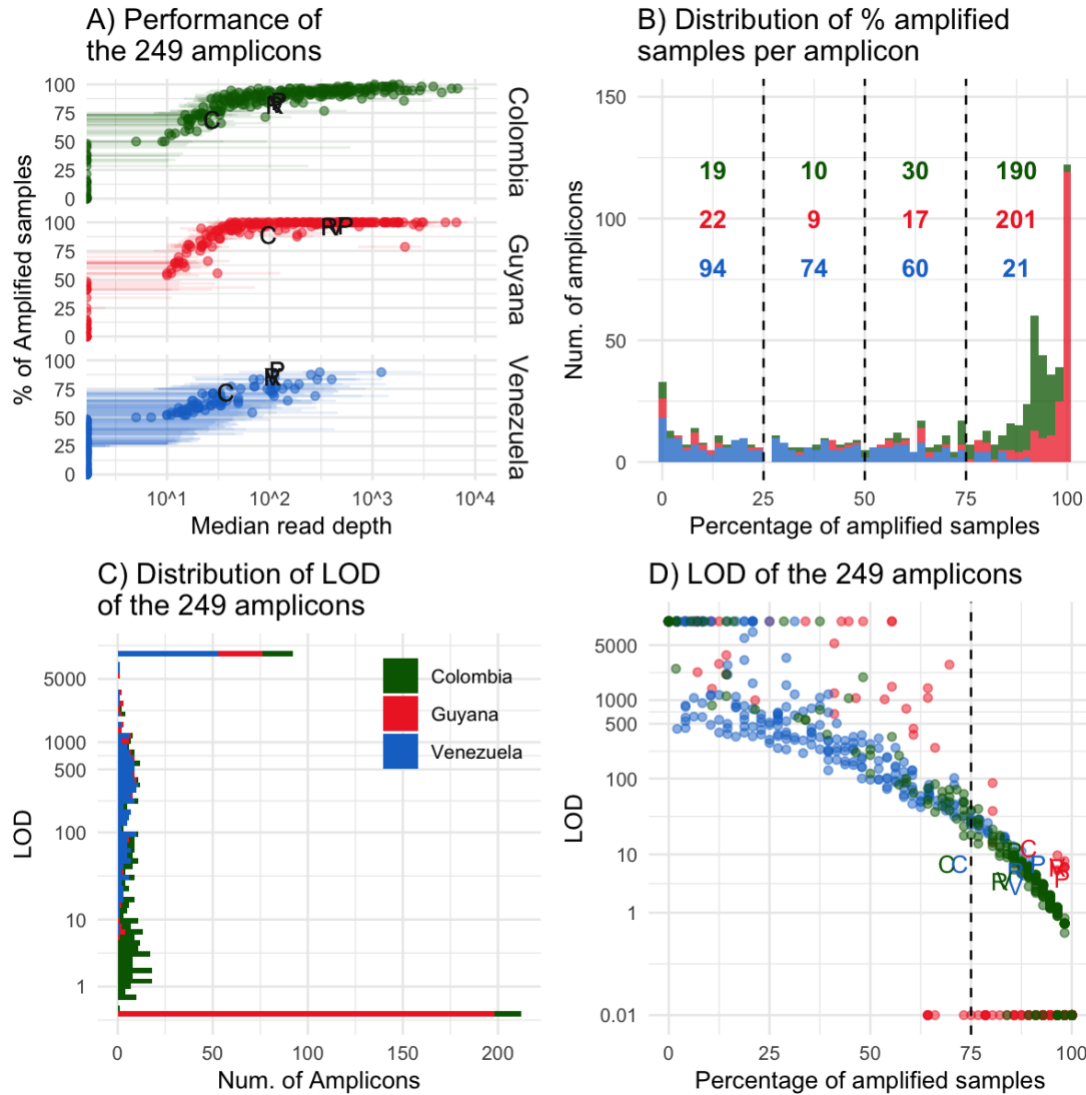
Supplementary figure 2.4: Principal coordinate analysis of Latin America and Caribbean *P. vivax* populations comparing whole genome sequencing (A), PvGTSeq (B), PvAmpliSeq (C) and rhAmpSeq (D). The analysis was done using genomic information from MalariaGEN Pv4, Brazilian genomes available from the European Nucleotide Archive (PRJEB56411, PRJEB44419, PRJEB36199), and 112 genomes from Colombia, Guyana, Honduras, Panama, Peru and Venezuela.



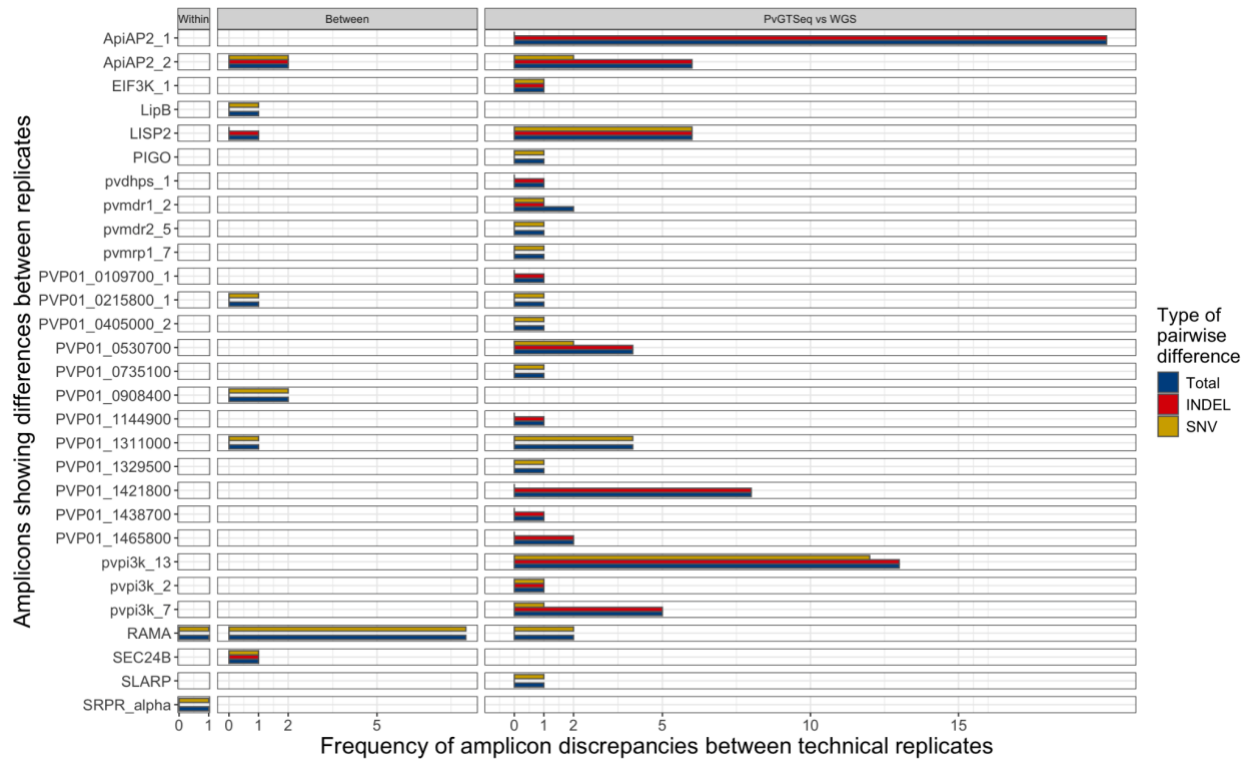
Supplementary figure 2.5: Deviation of IBD estimates from the amplicon panels relative to WGS. A) Scatter plot of IBD estimates for pairs of samples using WGS (x-axis) and the amplicon panels (y-axis). Each point represents a pairwise comparison, and the diagonal line indicates the expected value when IBD estimates from WGS and the panels are the same. B) Root mean squared error of IBD estimates from the amplicon panels relative to WGS across different IBD ranges. Confidence intervals were constructed assuming a chi-squared distribution. C) Coefficient of variation of IBD estimates from the amplicon panels relative to WGS across different IBD ranges. Confidence intervals were constructed assuming a T distribution.



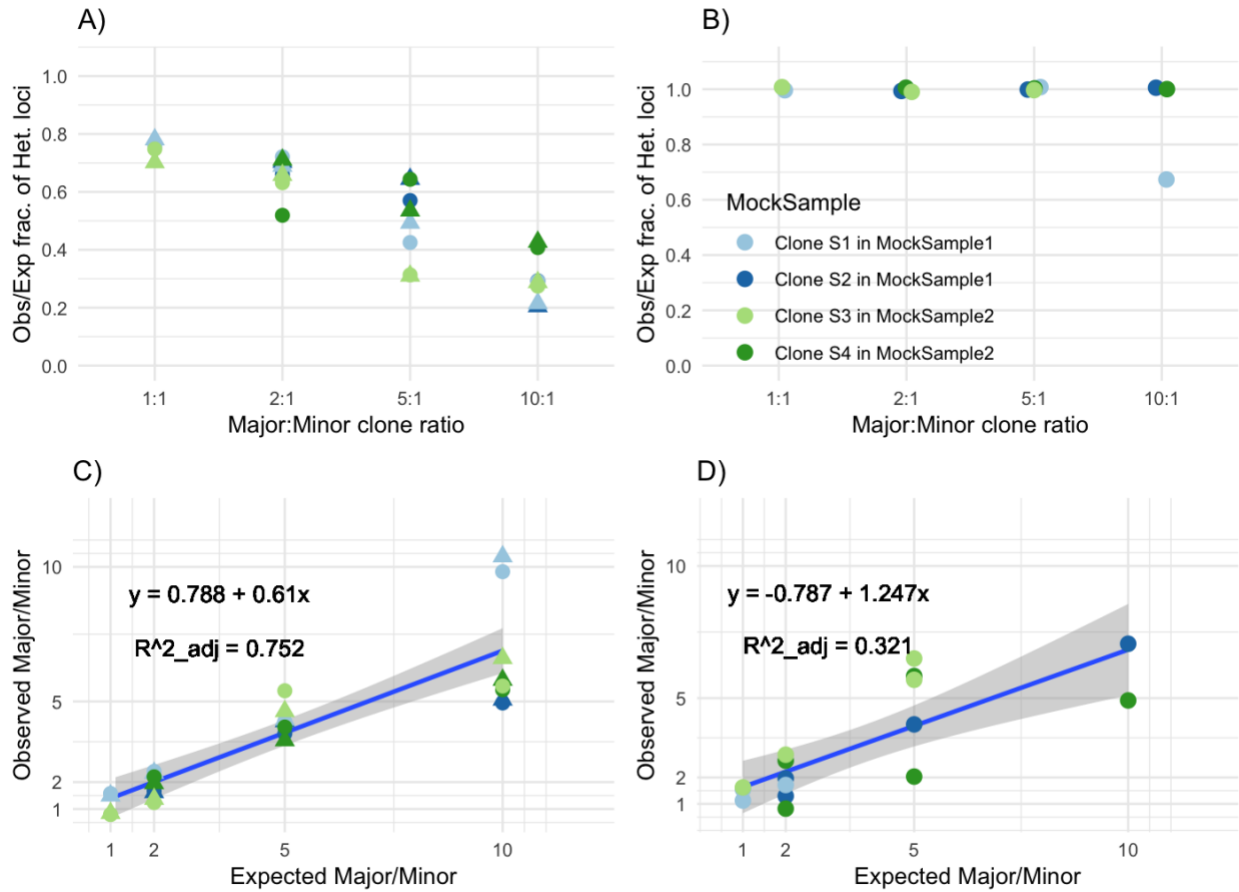
Supplementary figure 2.6: Bar plot showing the prevalence of haplotypes of genes that carry mutations associated with antimalarial resistance. y-axis shows the frequency in each population, x-axis shows the country in which the sample was collected, horizontal sections correspond to the analyzed gene, and vertical sections represent each world region.



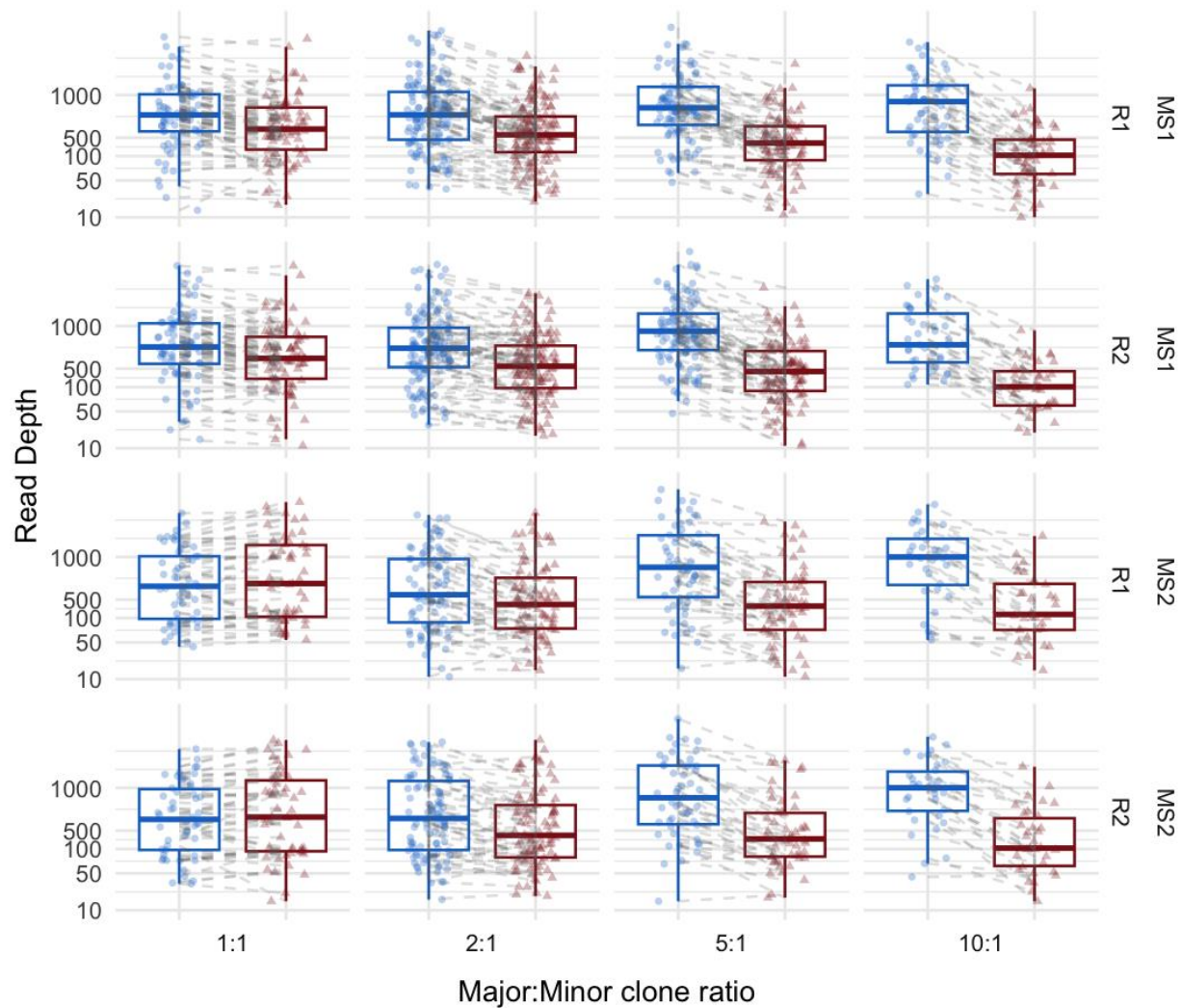
Supplementary figure 2.7: Performance of PvGTSeq and PvCRiSP amplicons. A) Scatter plot of the median read depth and the percentage of amplified samples of each amplicon on each of three populations: Colombia (Green), Guyana (Red) and Venezuela (Blue). Dots represent each of the 249 amplicons in PvGTSeq while the letters represent the 4 amplicons in PvCRiSP (CG2_related, RIPR, VSP11, and PIGM). B) Distribution of amplification rate of the amplicons in PvGTSeq and PvCRiSP in the three countries, as indicated by color. Numbers in figure B indicate the number of amplicons in each population that are in the 1st, 2nd, 3rd, and 4th quantile. C) Distribution of limit of detection (LOD) of each individual amplicon in PvGTSeq and PvCRiSP. D) Scatter plot of amplification rate (x-axis) and LOD (y-axis) of each individual amplicon (dots or letters) in three populations.



Supplementary figure 2.8: Contribution of each amplicon and the type of polymorphism to the discrepancies within and between PvGTSeq sequencing runs, and between PvGTSeq with respect to WGS. X-axis shows the number of times an amplicon (y-axis) showed a discrepancy between technical replicates in any of the 3 experiments (Vertical panels): Within and between sequencing runs of PvGTSeq, and between PvGTSeq and WGS. Colors indicate if the discrepancy in the amplicon was due to a single nucleotide variant (SNV, in gold), an insertion or deletion (INDEL, in red) or any of both (in blue).

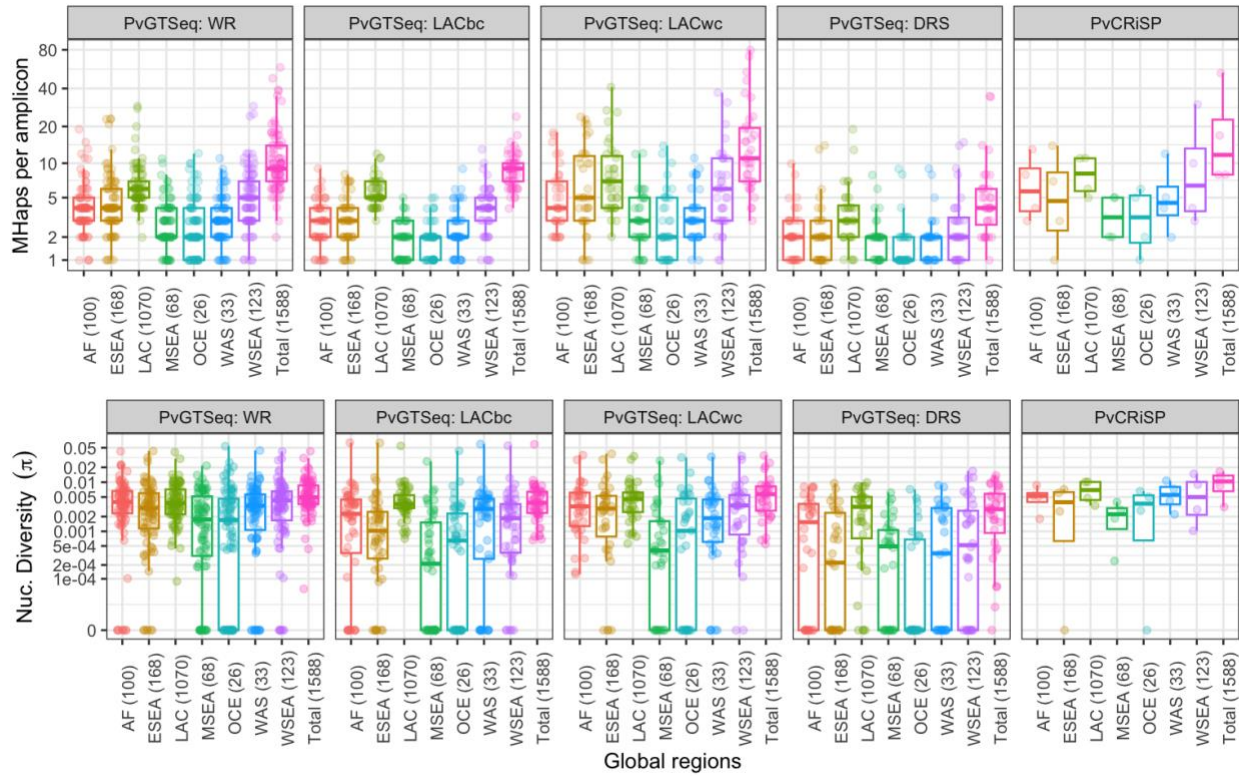


Supplementary figure 2.9: Detectability of major and minor clones in mock polyclonal infections. The top panels show the fraction of heterozygous loci detected relative to the ratio of major and minor clones in mock samples using PvGTSeq (A) and PvCRiSP (B). The bottom panels illustrate the correlation between the read depth ratio of major and minor clones for PvGTSeq (C) and PvCRiSP (D) compared to their DNA concentration ratios. MS1 and MS2 represent the two biological replicates, and shapes of the dots (circles and triangles) represent the 2 technical replicates.

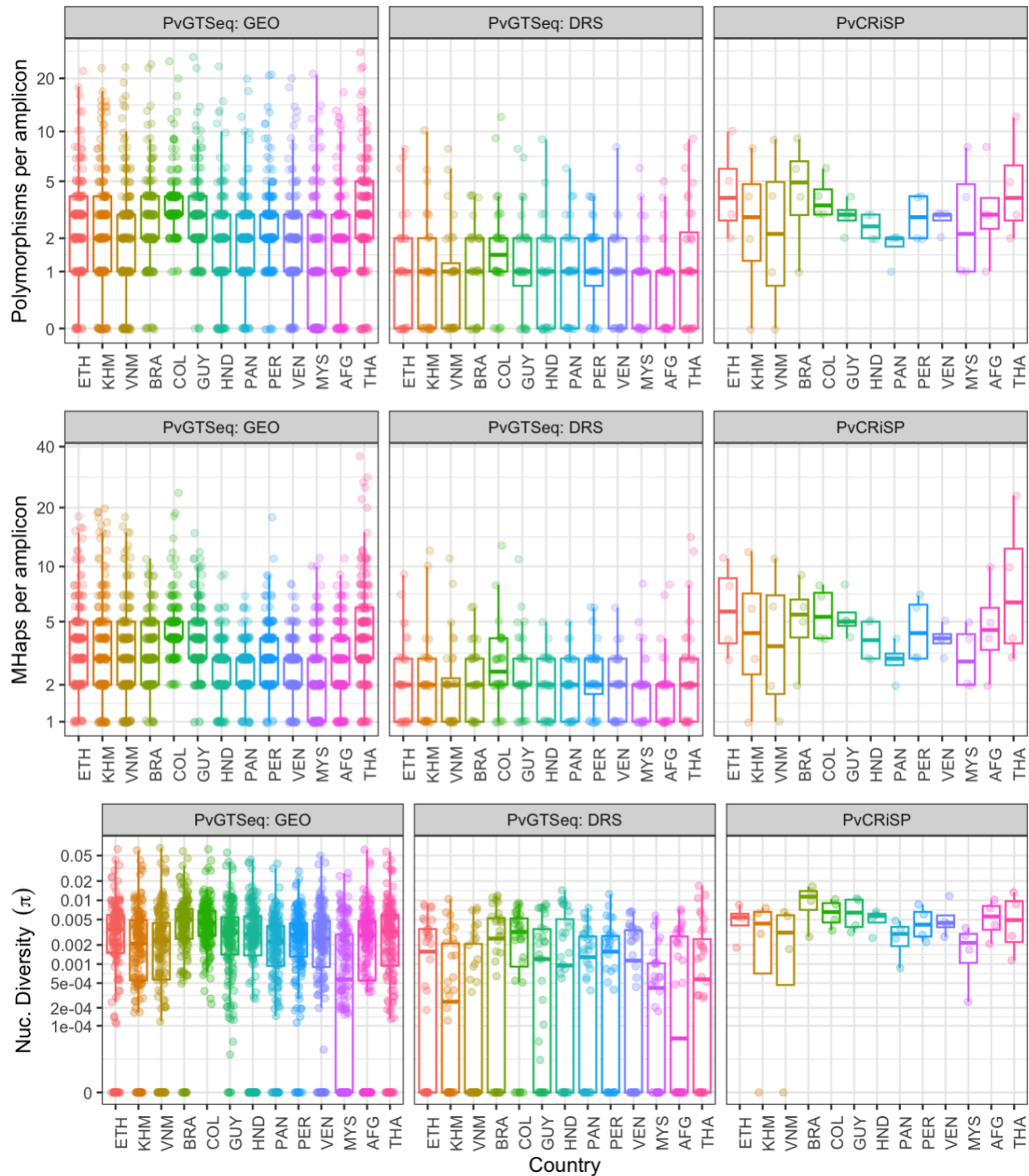


Supplementary figure 2.10: Read depth of major and minor clones within mock samples.

Each dot represents a private microhaplotype from major (blue) and minor (red) clones using the 201 amplicon from PvGTSeq. Horizontal panels represent each of the two combinations of mock samples (MS1 and MS2) and their technical replicates (R1 and R2), while the vertical panels represent the different ratios at which mock samples were generated.

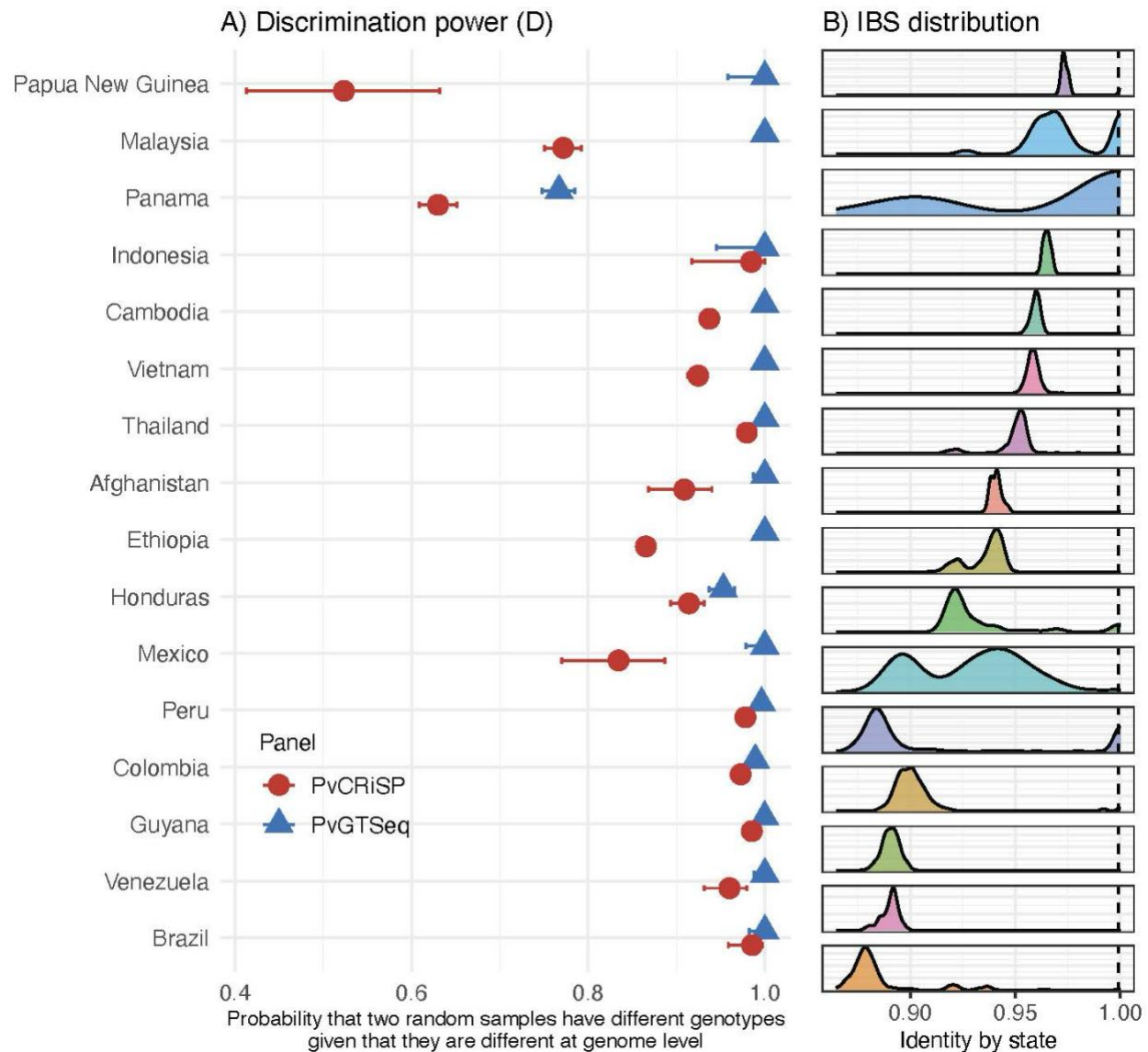


Supplementary figure 2.11: Diversity of PvGTSeq amplicons across world regions. Number of polymorphic sites, microhaplotypes (MHaps) alleles and nucleotide diversity (π) per amplicon use groups and across seven global regions: Africa (AF), Eastern South-East Asia (ESEA), Latin America and Caribbean (LAC), Maritime South-East Asia (MSEA), Oceania (OCE), Western Asia (WAS), and Western South-East Asia (WSEA). Numbers within parenthesis indicate the number of monoclonal clinical samples. Each dot represents an amplicon. Each dot represents an amplicon. Use case groups were defined as 81 amplicons for geographic differentiation between world regions (WR), 49 amplicons for geographic differentiation between countries in LAC (LACbc), 35 amplicons for geographic differentiation within countries in LAC (LACwc), 32 amplicons for drug resistance surveillance (DRS), and the four amplicons comprising PvCRiSP. Because there are amplicons for geographic differentiation that belong to multiple groups, to avoid duplication we assigned the amplicon to the highest geographical scale.

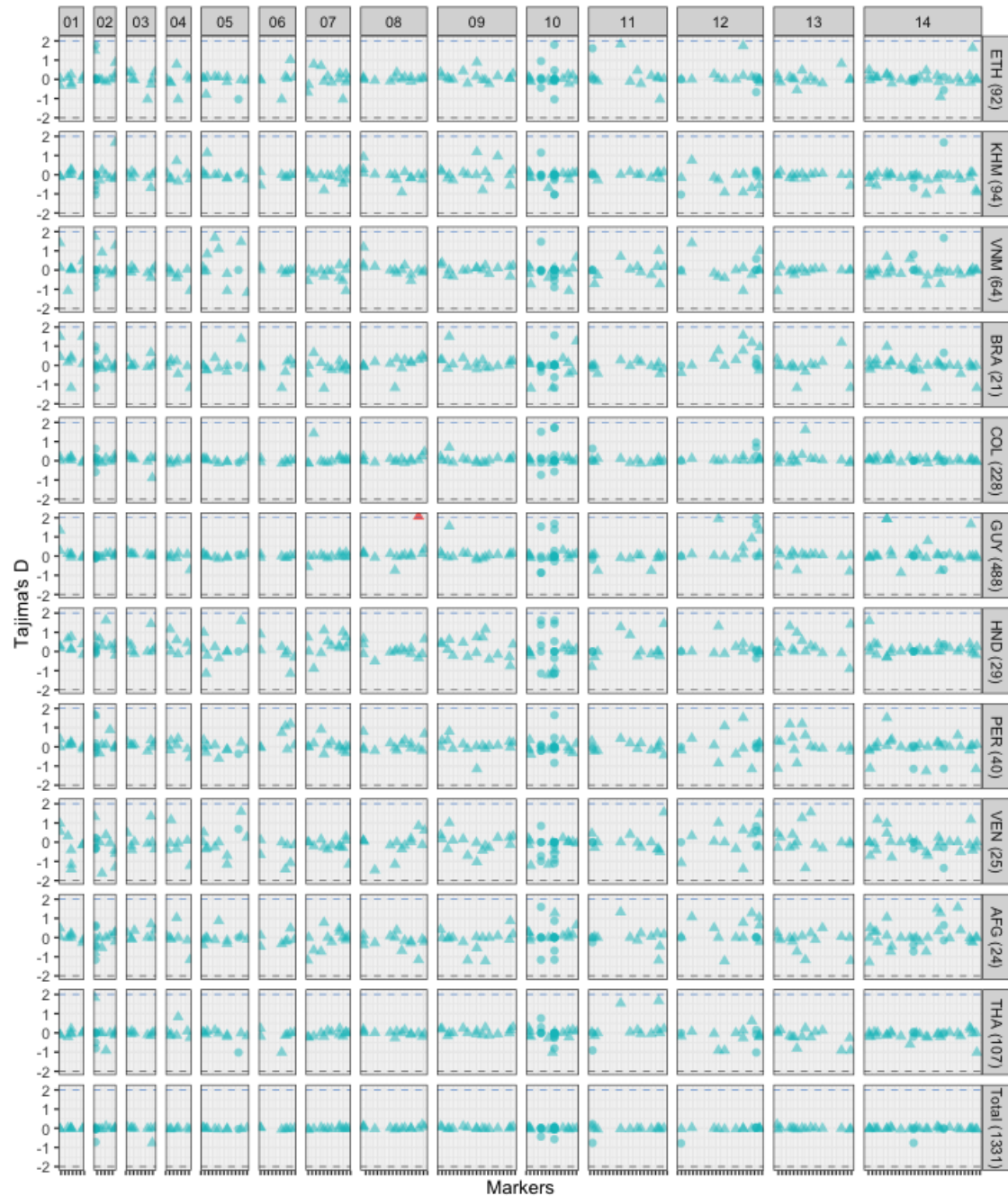


Supplementary figure 2.12: Diversity of PvGTSeq amplicons across countries. Number of polymorphic sites, microhaplotypes (MHaps) and nucleotide diversity (π) per amplicon use-case groups and across 13 countries: Ethiopia (ETH), Cambodia (KHM), Vietnam (VNM), Brazil (BRA), Colombia (COL), Guyana (GUY), Honduras (HND), Panama (PAN), Peru (PER), Venezuela (VEN), Malaysia (MYS), Afghanistan (AFG) and Thailand (THA). Each dot represents an amplicon. Use case-groups were defined as 169 amplicons for geographic

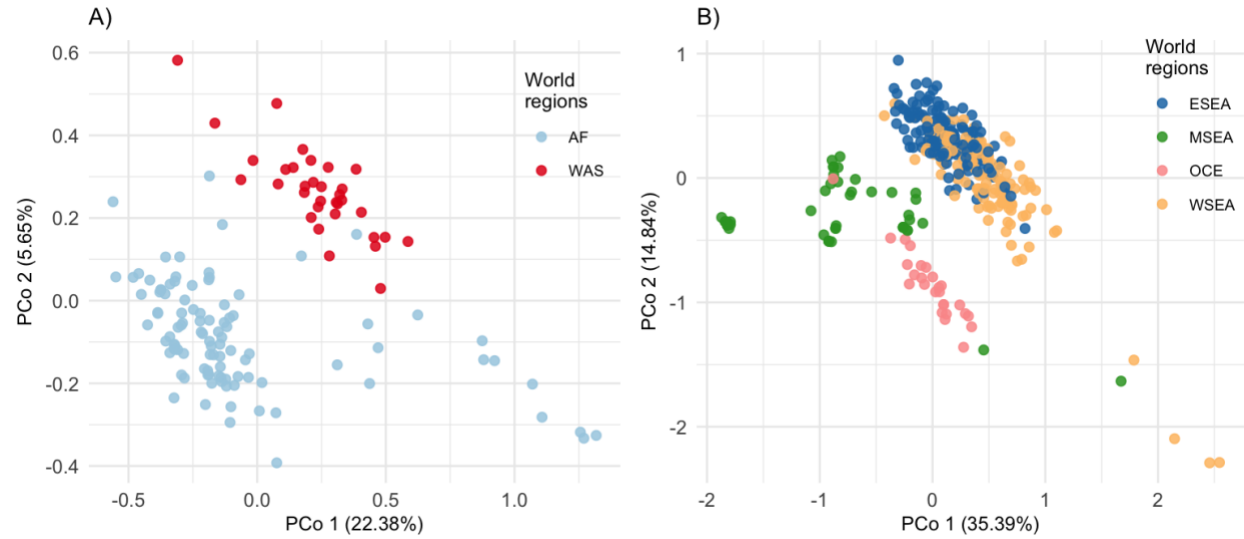
differentiation, 32 amplicons for drug resistance surveillance (DRS), and the four amplicons conforming PvCRiSP.



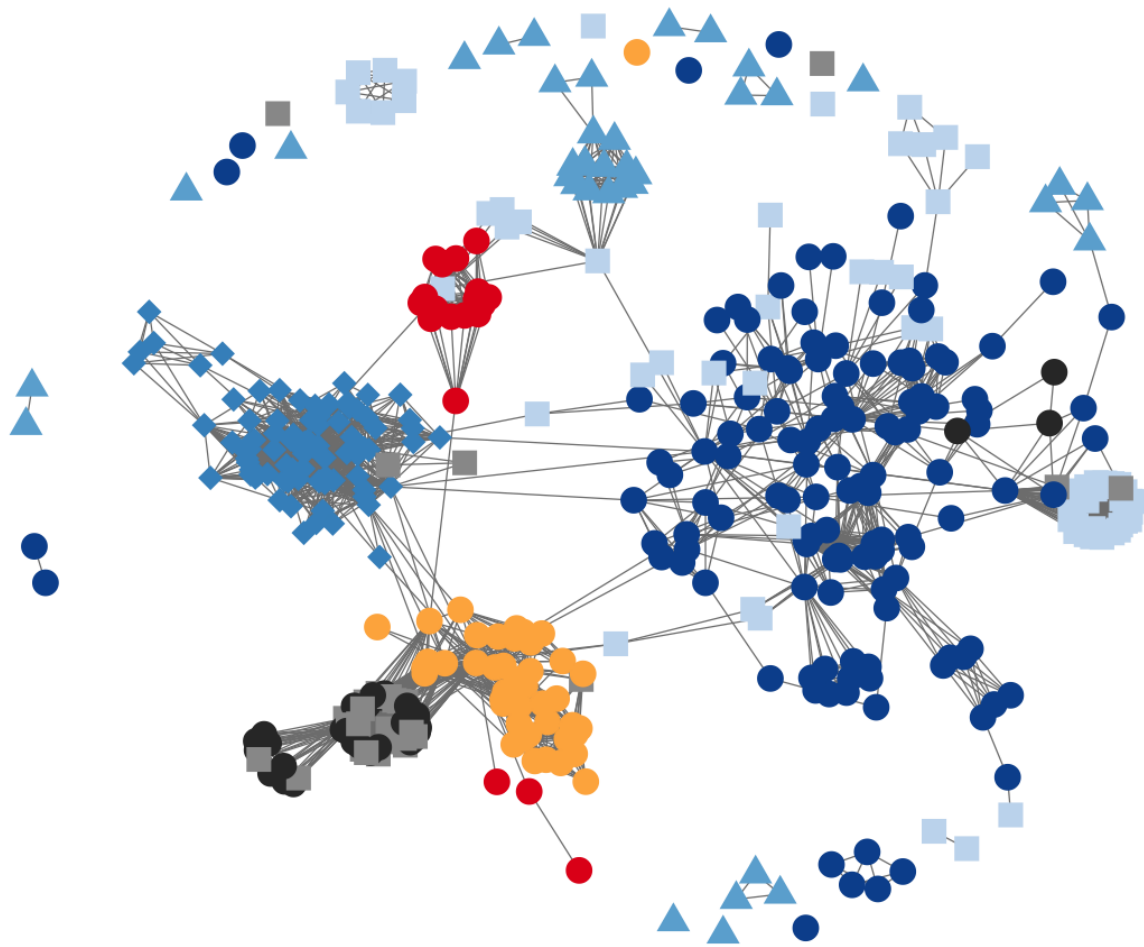
Supplementary figure 2.13: Panel discrimination power. The figure on the left shows the probability that two random samples differ from each other in at least one amplicon (x-axis) across global *P. vivax* populations (y-axis) using PvGTSeq (blue) and PvCRiSP (red). Figure on the right shows the distribution of identity by state within each population measured from whole genome sequencing data.



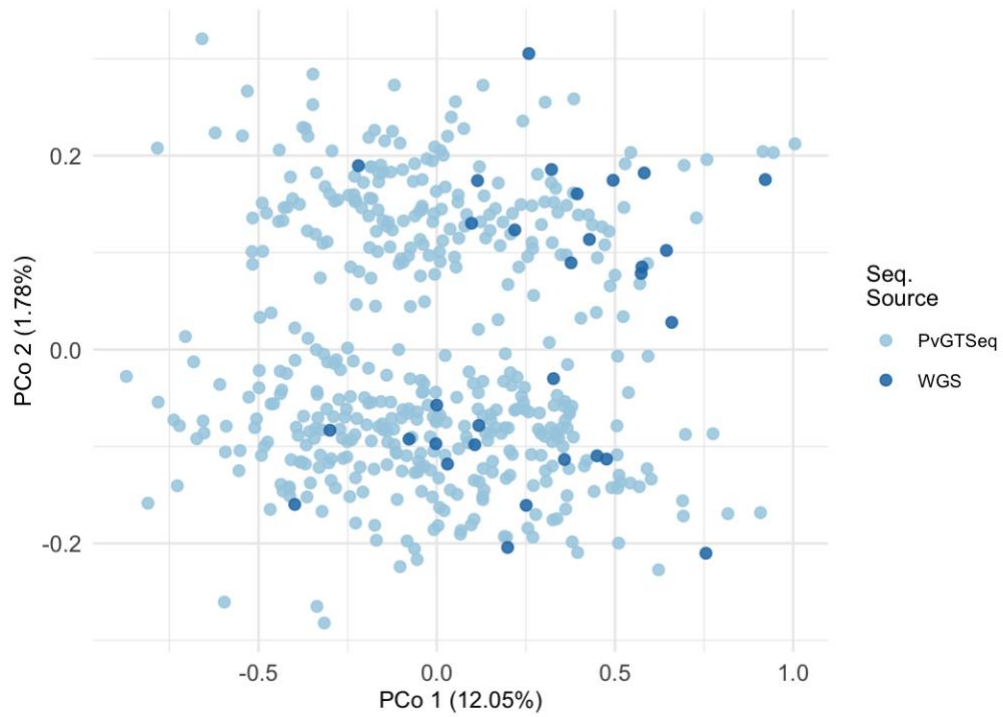
Supplementary figure 2.14: Scatter plots depicting Tajima's D values in each amplicon included in the PvGTSeq panel. Shapes of the dots represent whether the amplicon is used for geographic differentiation (triangles) or for drug resistance surveillance (circles) and dark gray color indicates where the amplicon is under strong selection signal (Tajima's D > 2 or < -2). The analyzed countries includes: Ethiopia (ETH), Cambodia (KHM), Vietnam (VNM), Brazil (BRA), Colombia (COL), Guyana (GUY), Honduras (HND), Peru (PER), Venezuela (VEN), Afghanistan (AFG) and Thailand (THA), and numbers within parenthesis indicates the number of non-clonal haplotypes used in the analysis.



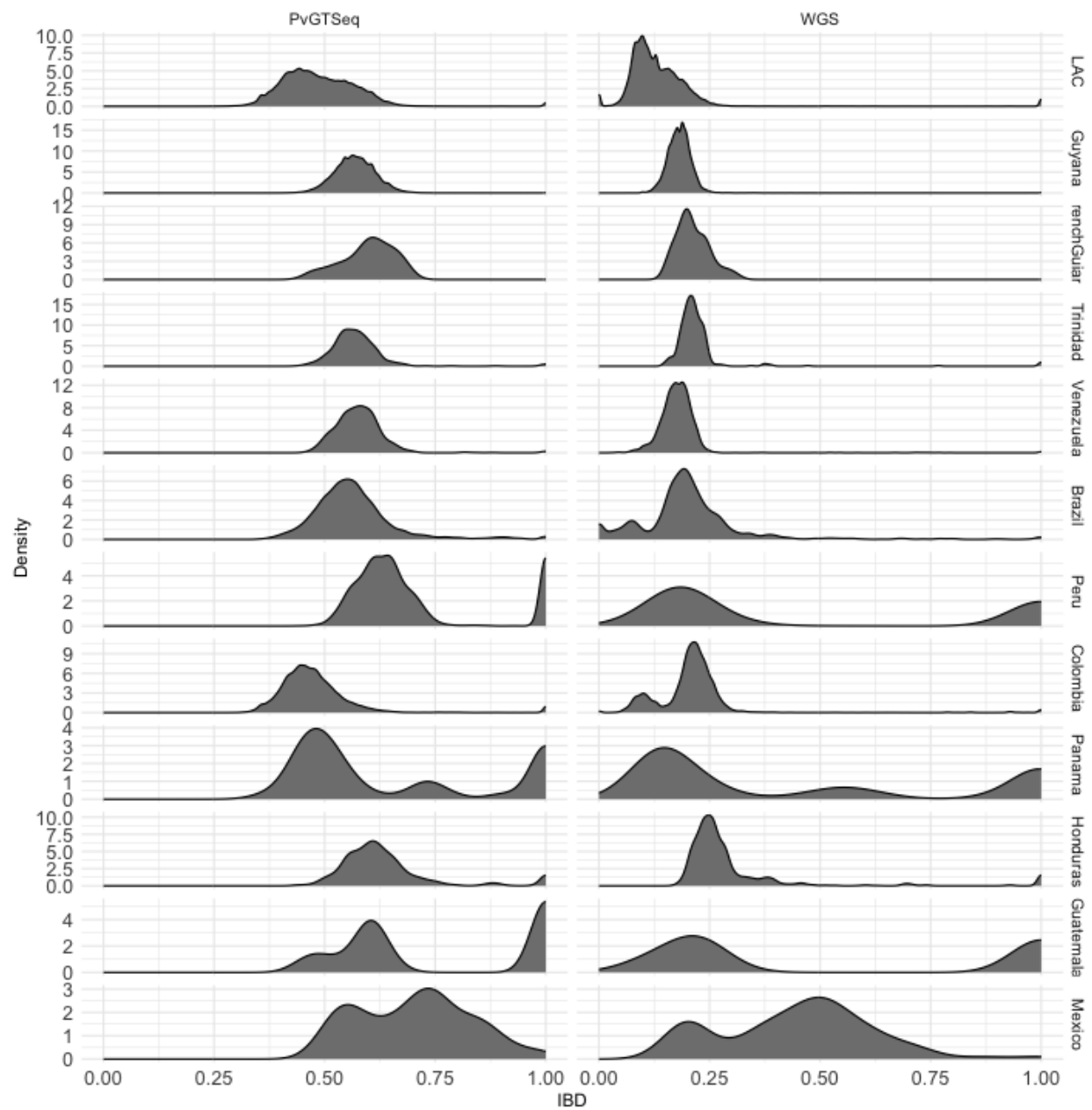
Supplementary figure 2.15: Population differentiation using PvGTSeq in other world regions. Principal coordinate analysis between Africa (AF) and Western Asia (WAS) in panel A, and between Eastern South-East Asia (ESEA), Maritime South-East Asia (MSEA), Oceania (OCE), and Western South-East Asia (WSEA) in panel B.



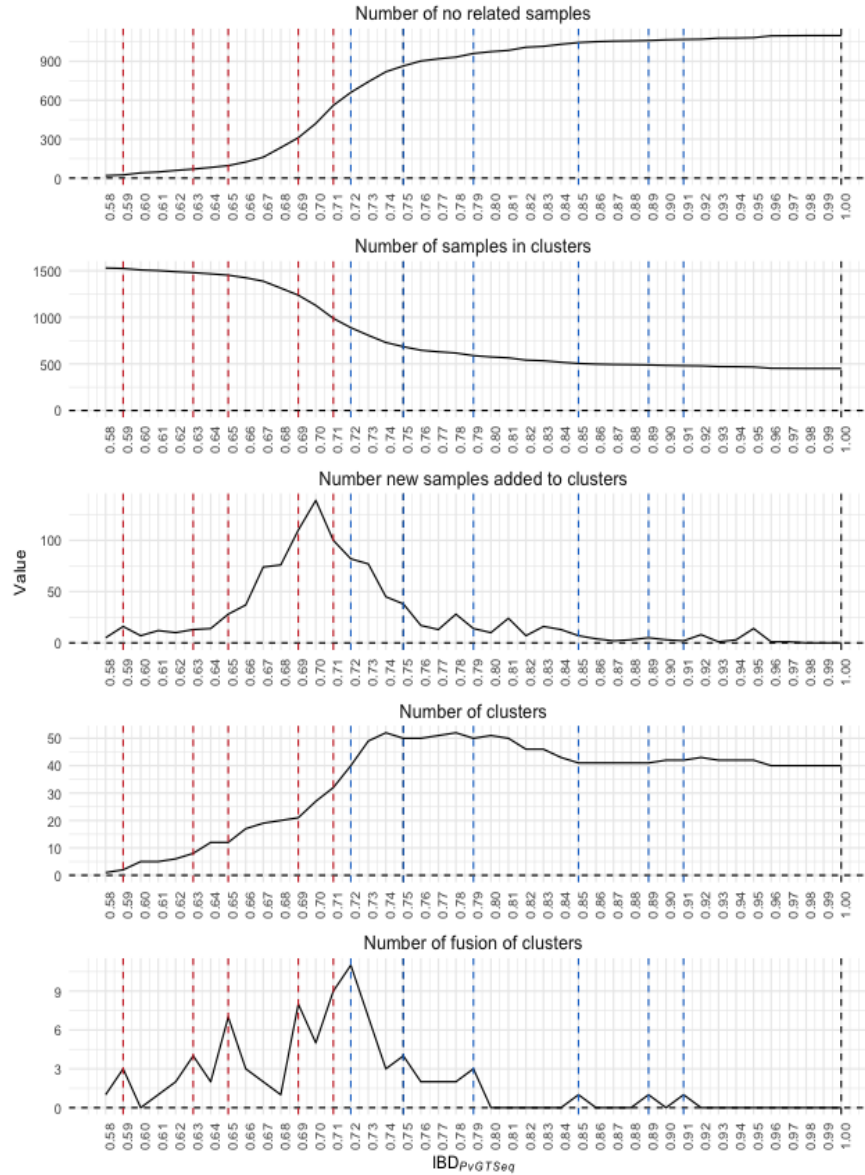
Supplementary figure 2.16: Network analysis in Mesoamerica and Colombia. Network graph showing genetic relationships (IBS) among samples from Colombia (Blues), Honduras (Gold), Mexico (Red) and Panama (Grey and Black). Each node represents a monoclonal sample, with edges indicating genetic relationships exceeding 0.6 IBS. Figures B, C and D share the same color scheme.



Supplementary figure 2.17: Harmonization of PvGTSeq and WGS data. Principal coordinate analysis of samples from Guyana generated through PvGTSeq (light blue dots) and WGS (blue dots).

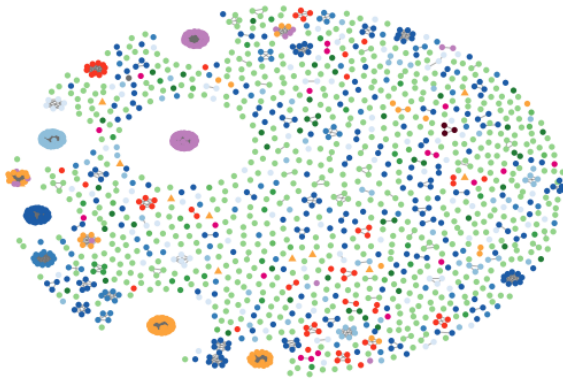


Supplementary figure 3.1: Pairwise IBD distribution using PvGTSeq and WGS within LAC countries.

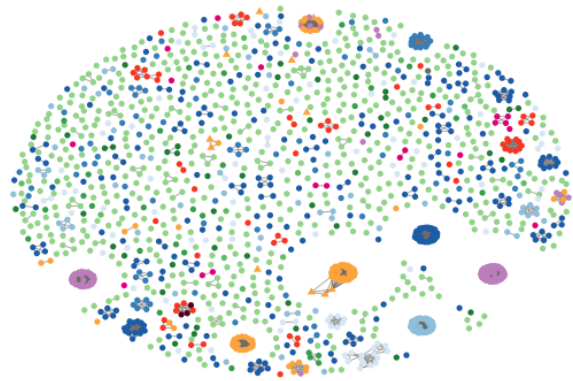


Supplementary figure 3.2: Scanning of critical $IBD_{PvGTseq}$ values that indicate cluster merges. The x-axis shows the IBD threshold analyzed, and the panels show the number of unrelated samples (not connected by branches in the network), the number of samples that belong to clusters, the number of new samples added to clusters relative to the previous IBD threshold, the number of clusters present at each threshold, and the number of merges between clusters at each threshold. Horizontal lines indicate critical $IBD_{PvGTseq}$ values.

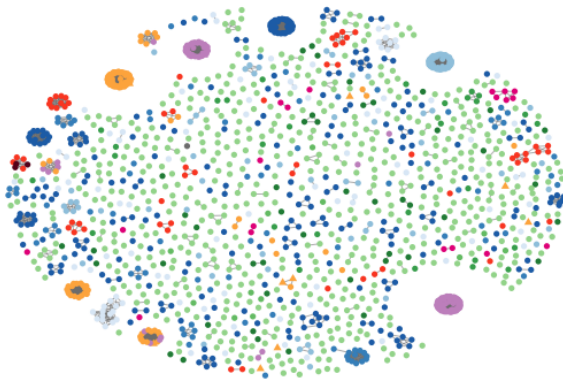
$IBD_{PvGTSeq} = 1$



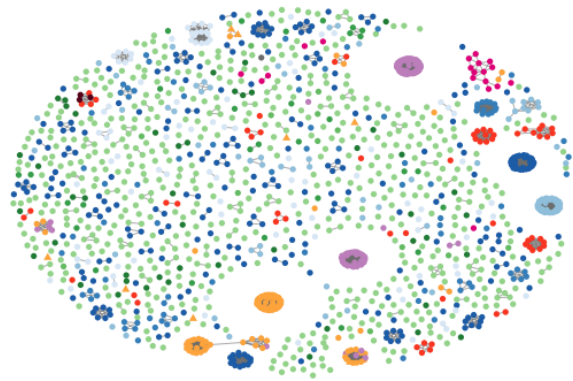
$IBD_{PvGTSeq} \geq 0.91$



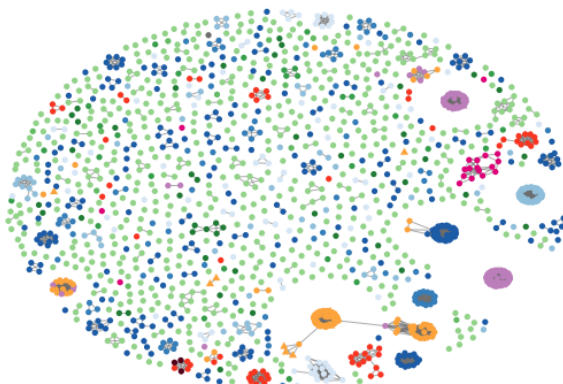
$IBD_{PvGTSeq} \geq 0.89$



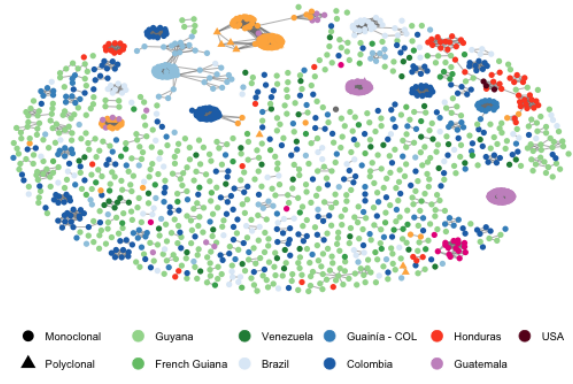
$IBD_{PvGTSeq} \geq 0.85$



$IBD_{PvGTSeq} \geq 0.79$

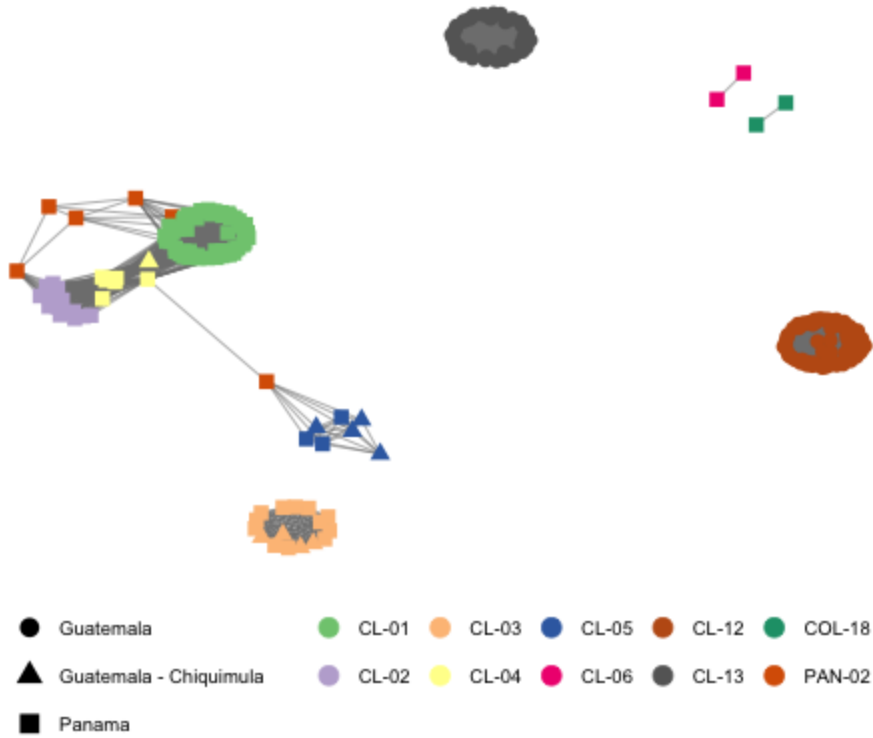


$IBD_{PvGTSeq} \geq 0.75$

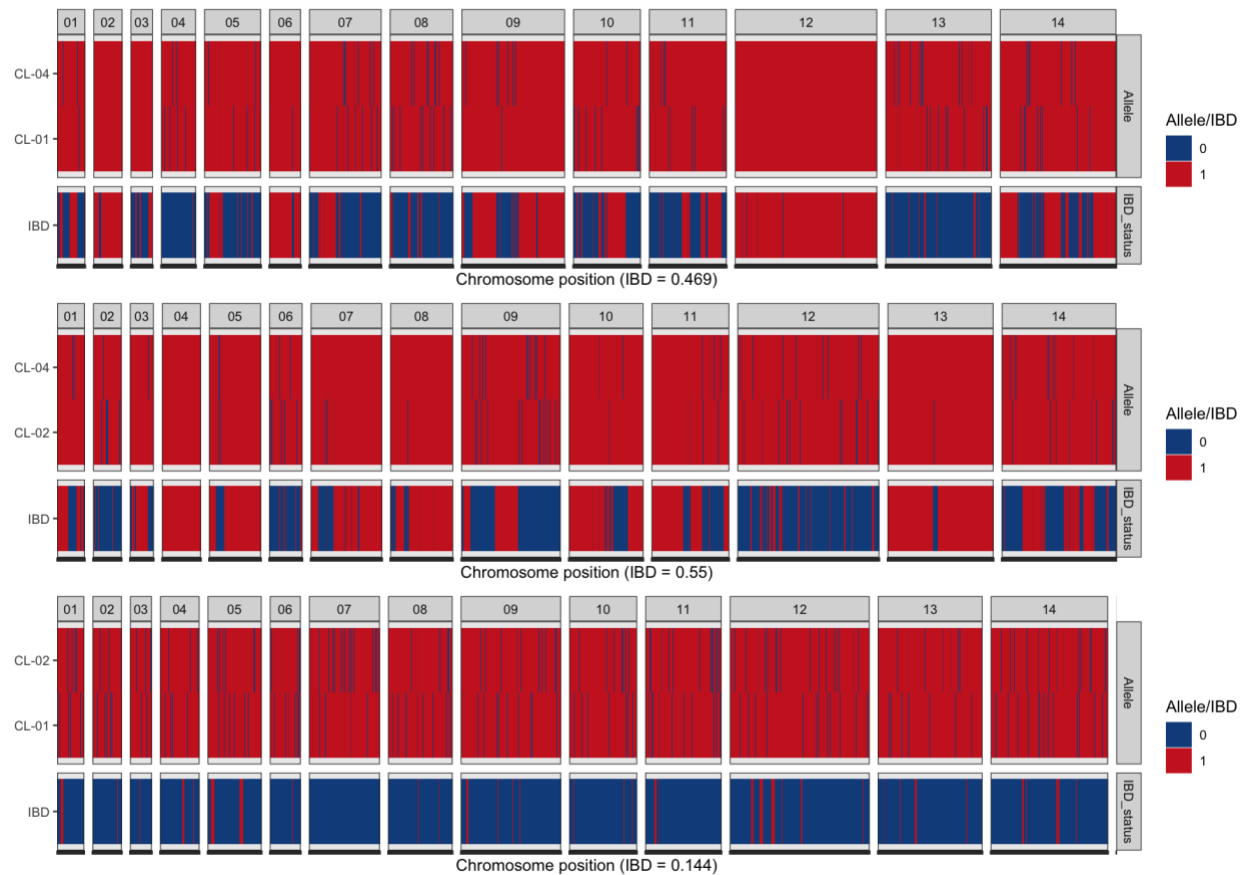


Supplementary figure 3.3: Identification of the appropriate $IBD_{PvGTSeq}$ value to define clonal or low-diversity groups using PvGTSeq. The critical IBD value was selected based on the threshold at which cluster fusion begins to be observed due to genetic recombination between members of two genetically different groups. Edges are drawn for pairwise comparisons above the specified $IBD_{PvGTSeq}$ threshold.

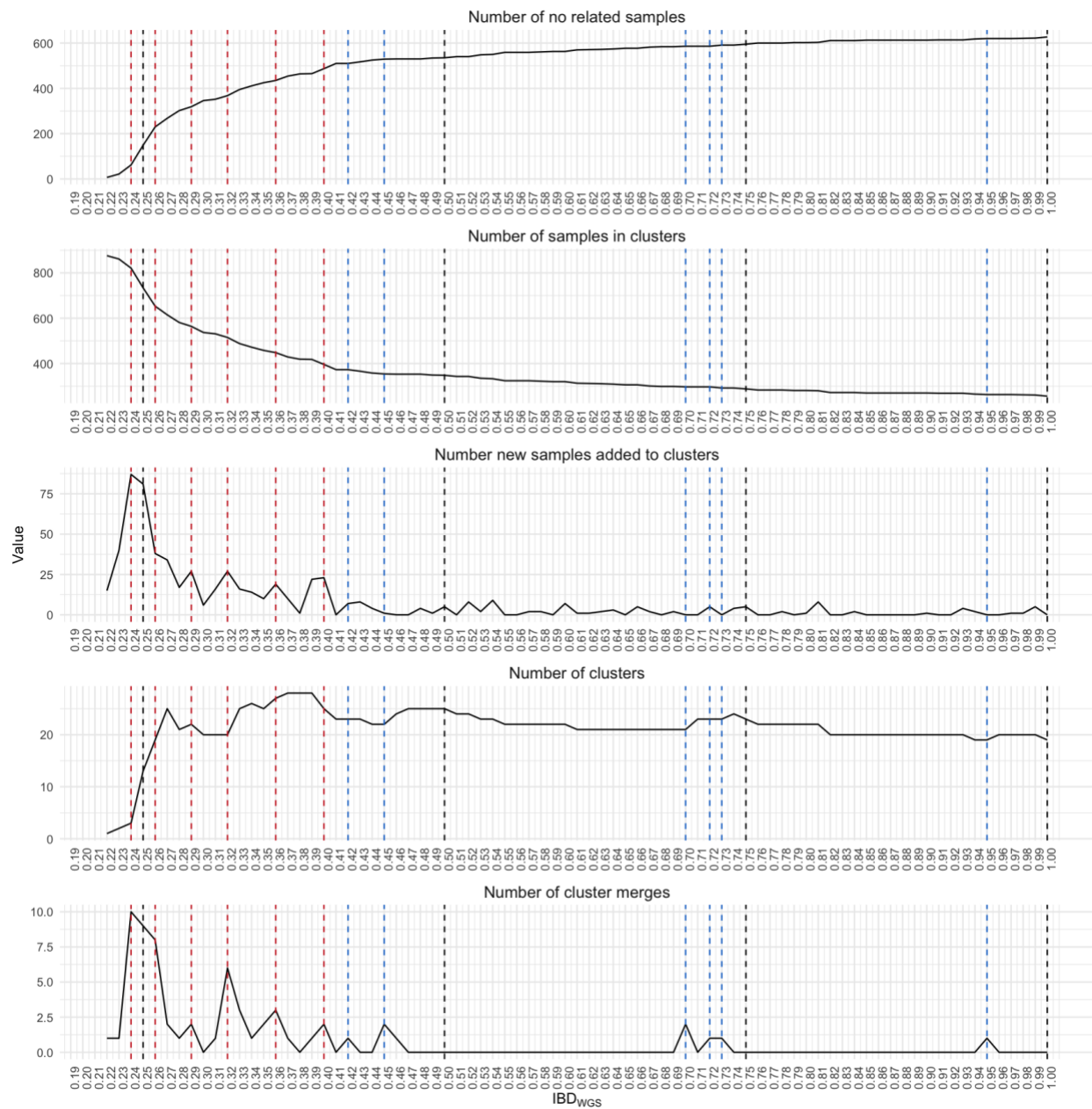
$IBD_{PvGTSeq} \geq 0.74$



Supplementary figure 3.4: Fusion of parental clonal groups and their recombinant progeny in Panama and Guatemala using PvGTSeq. Each node represents a sample; colors indicate clonal-group assignment ($IBD_{PvGTSeq} \geq 0.86$), and node shape indicates the country of collection. In Guatemala, we distinguish Chiquimula from the other study areas because it is not considered endemic, and cases are reported only in migrant individuals. Edges are drawn for pairwise comparisons with $IBD_{PvGTSeq} \geq 0.74$ (threshold to identify ecological groups).



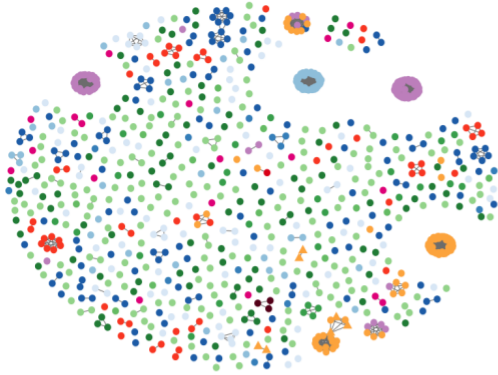
Supplementary figure 3.5: Origin of CL-04. Identification of long IBS and IBD segments between CL-04 and its two parental genetic groups, CL-01 and CL-02. Each column represents a position in the genome, and each row indicates a haplotype or the IBD status between haplotypes for each biallelic SNP in the genome. Haplotype alleles are represented by two colors: blue (0) and red (1). For ease of reading, when haplotypes were identical by state at a given position, the allele was shown in red in both haplotypes. Otherwise, one haplotype was shown in red and the other in blue. IBD at each position was estimated using the Viterbi function in the hidden Markov chain model (HMM), and a value of 1 (red) was assigned if the position was identical by descent and 0 (blue) if the position was not identical by descent. Note that identity by state and identity by descent are not necessarily reciprocal because the latter depends on allele frequency in the population and the status of neighboring positions. Panels in the bottom represent the genetic relatedness between the parental lineages CL-01 and CL-02.



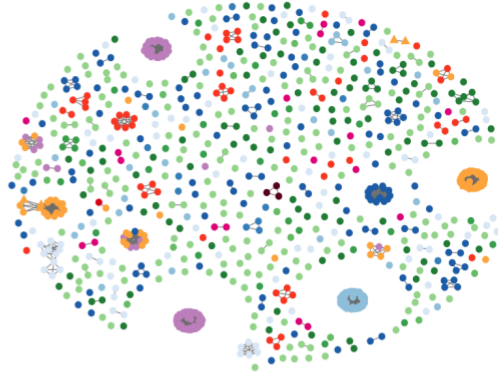
Supplementary figure 3.6: Scanning of critical IBD_{WGS} values that indicate cluster merges.

The x-axis shows the IBD threshold analyzed, and the panels show the number of unrelated samples (not connected by branches in the network), the number of samples that belong to clusters, the number of new samples added to clusters relative to the previous IBD threshold, the number of clusters present at each threshold, and the number of merges between clusters at each threshold. Horizontal lines indicate critical IBD_{WGS} values.

$IBD_{WGS} \geq 0.95$



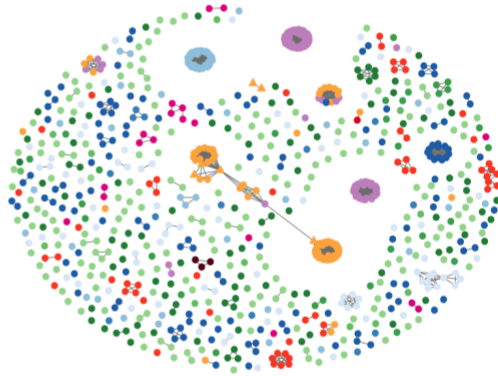
$IBD_{WGS} \geq 0.73$



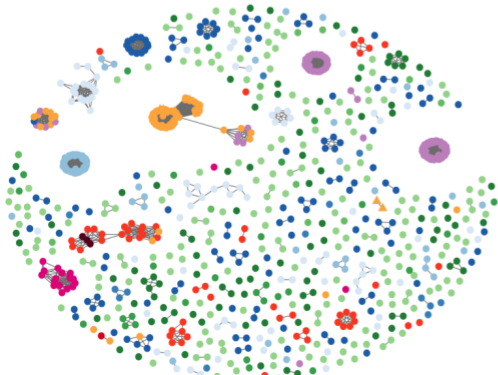
$IBD_{WGS} \geq 0.72$



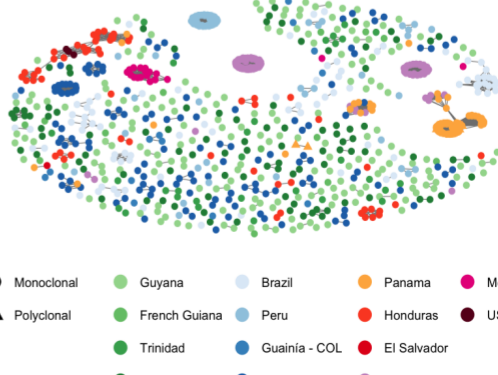
$IBD_{WGS} \geq 0.7$



$IBD_{WGS} \geq 0.45$

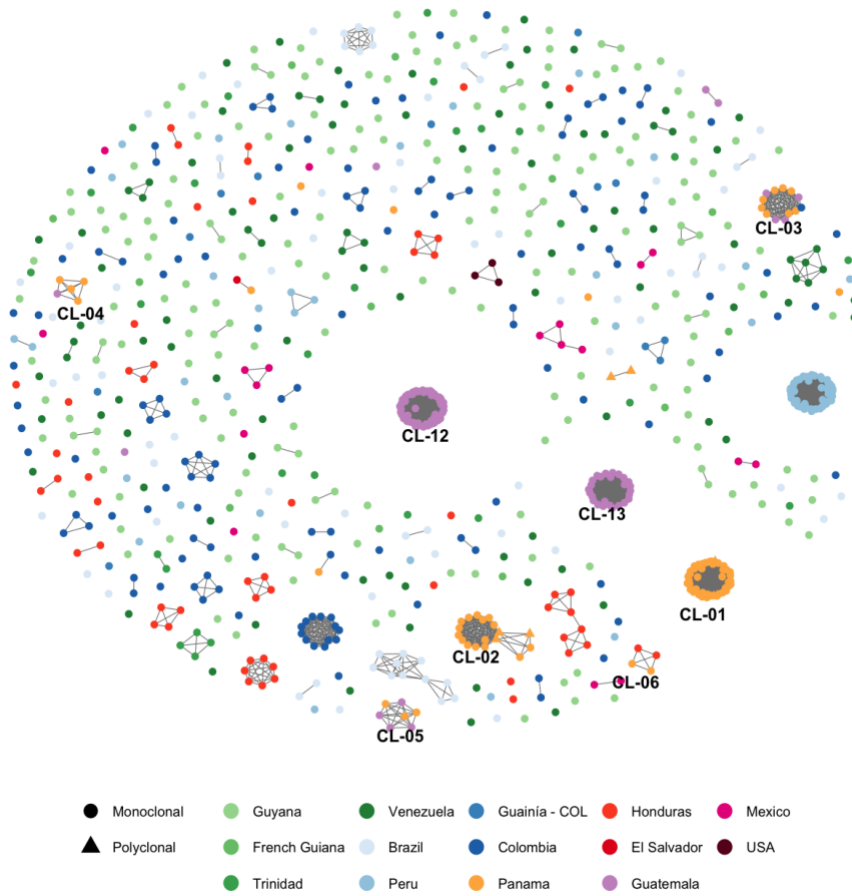


$IBD_{WGS} \geq 0.42$



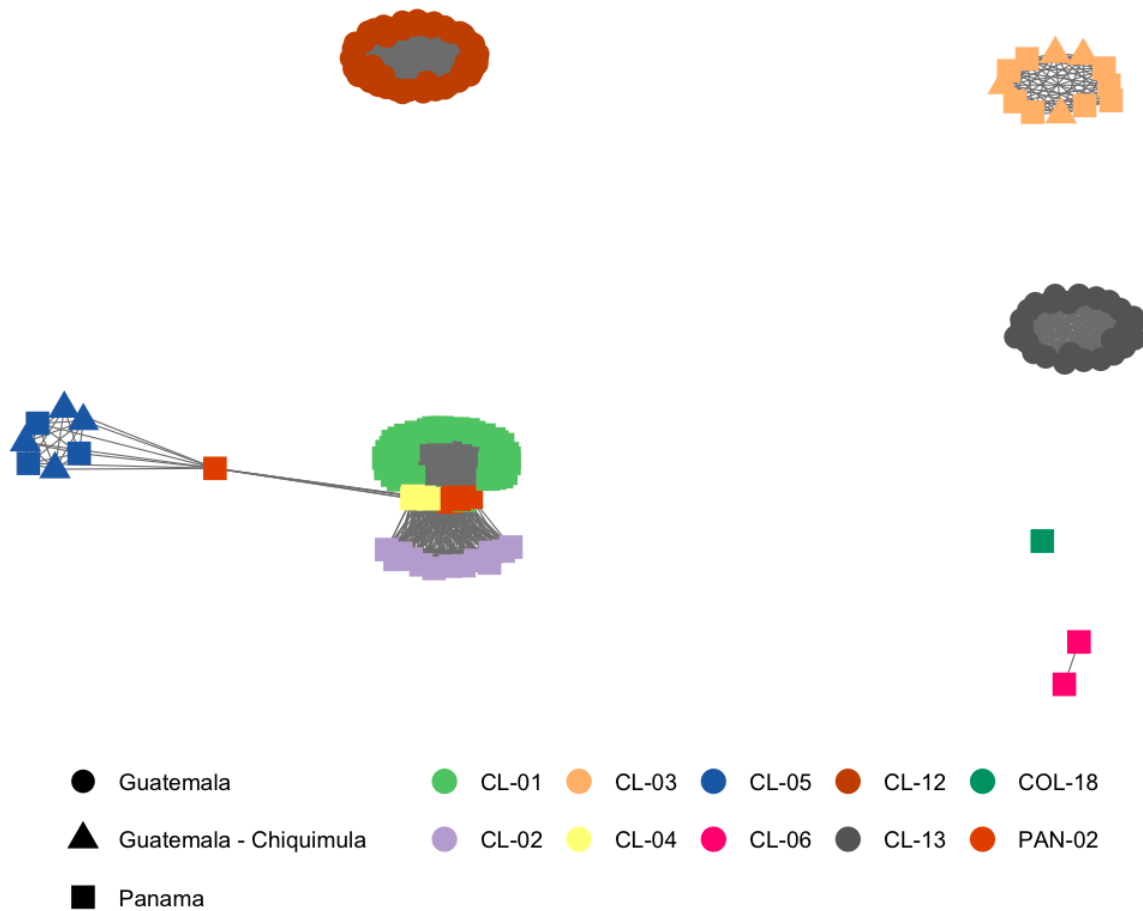
Supplementary figure 3.7: Identification of the appropriate IBD_{WGS} value to define clonal or low-diversity groups using WGS. The critical IBD value was selected based on the threshold at which cluster fusion begins to be observed due to genetic recombination between members of two genetically different groups. Edges are drawn for pairwise comparisons above the specified IBD_{WGS} threshold.

$IBD_{WGS} \geq 0.71$



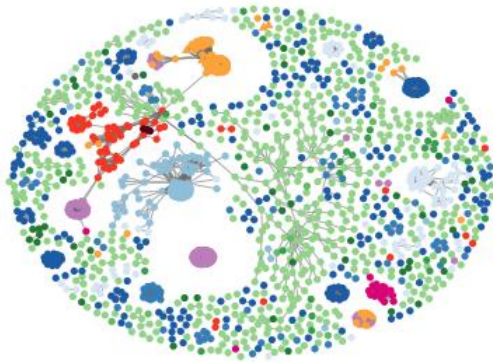
Supplementary figure 3.8: Identification of clonal and low-diversity groups using WGS genetic relatedness (IBD_{WGS}) in a network analysis. Each dot represents a genotyped sample. Colors are assigned based on the geographic origin of the sample (country). Shapes are based on the complexity of the sample infection. Polyclonal infections from Panama, Honduras, and Guatemala were only included in this analysis. Polyclonal infections from other countries were excluded. Branches represent pairwise IBD greater than 0.86. Clonal groups were defined as groups of four or more samples connected by branches. Labels above clusters are the labels used in the main text.

$$IBD_{WGS} \geq 0.37$$

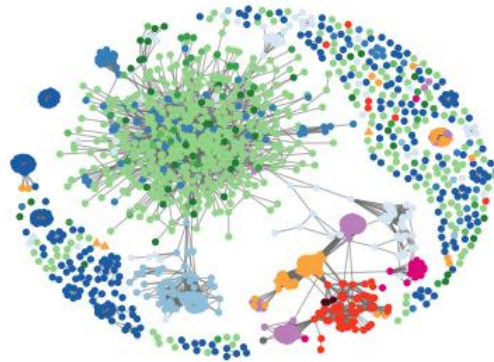


Supplementary figure 3.9: Fusion of parental clonal groups and their recombinant progeny in Panama and Guatemala using PvGTSeq. Each node represents a sample; colors indicate clonal-group assignment ($IBDPvGTSeq \geq 0.86$), and node shape indicates the country of collection. In Guatemala, we distinguish Chiquimula from the other study areas because it is not considered endemic, and cases are reported only in migrant individuals. Edges are drawn for pairwise comparisons with $IBD_{PvGTSeq} \geq 0.74$ (threshold to identify ecological groups).

$IBD_{PvGTSeq} \geq 0.72$



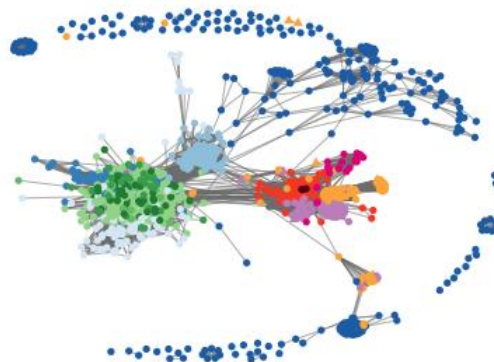
$IBD_{PvGTSeq} \geq 0.69$



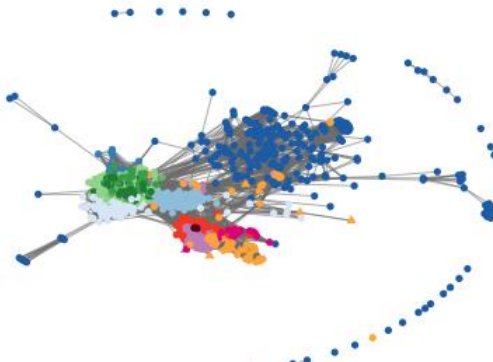
$IBD_{PvGTSeq} \geq 0.65$



$IBD_{PvGTSeq} \geq 0.63$

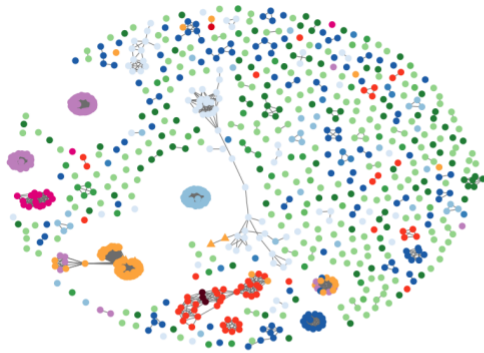


$IBD_{PvGTSeq} \geq 0.59$

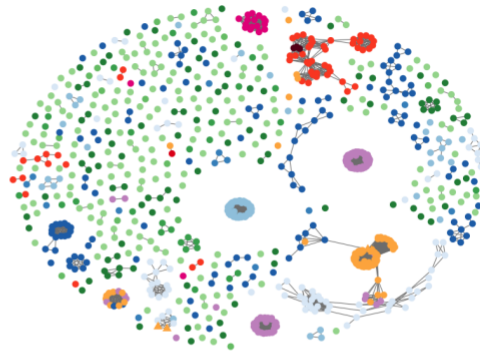


Supplementary figure 3.10: Identification of the appropriate IBD value to define ecological populations using PvGTSeq. The critical IBD value was selected based on the threshold at which cluster fusion begins to be observed due to genetic recombination between members of two genetically different groups. Edges are drawn for pairwise comparisons above the specified $IBD_{PvGTSeq}$ threshold.

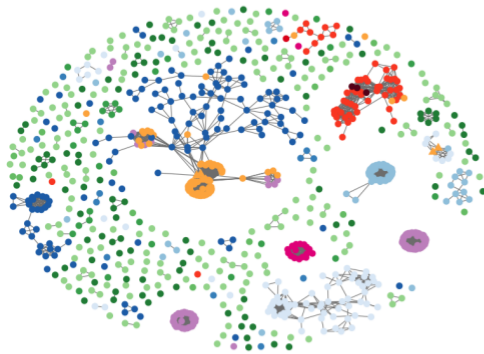
$IBD_{WGS} \geq 0.4$



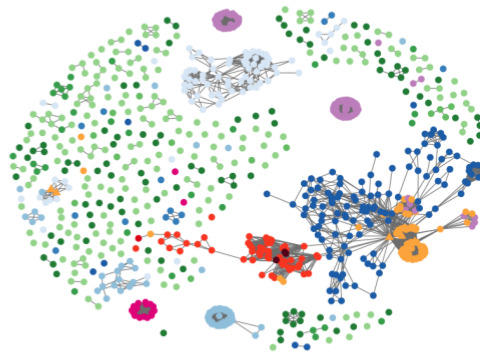
$IBD_{WGS} \geq 0.36$



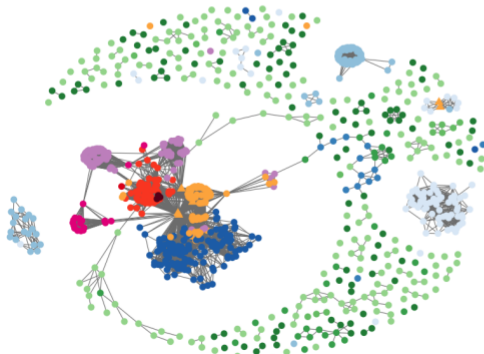
$IBD_{WGS} \geq 0.32$



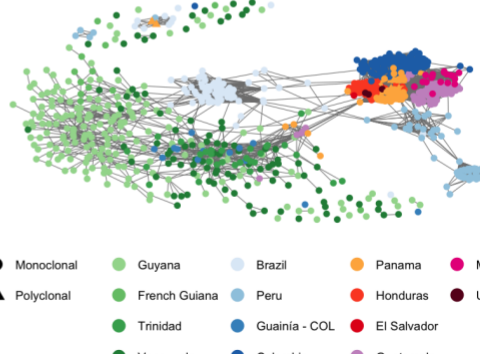
$IBD_{WGS} \geq 0.29$



$IBD_{WGS} \geq 0.26$

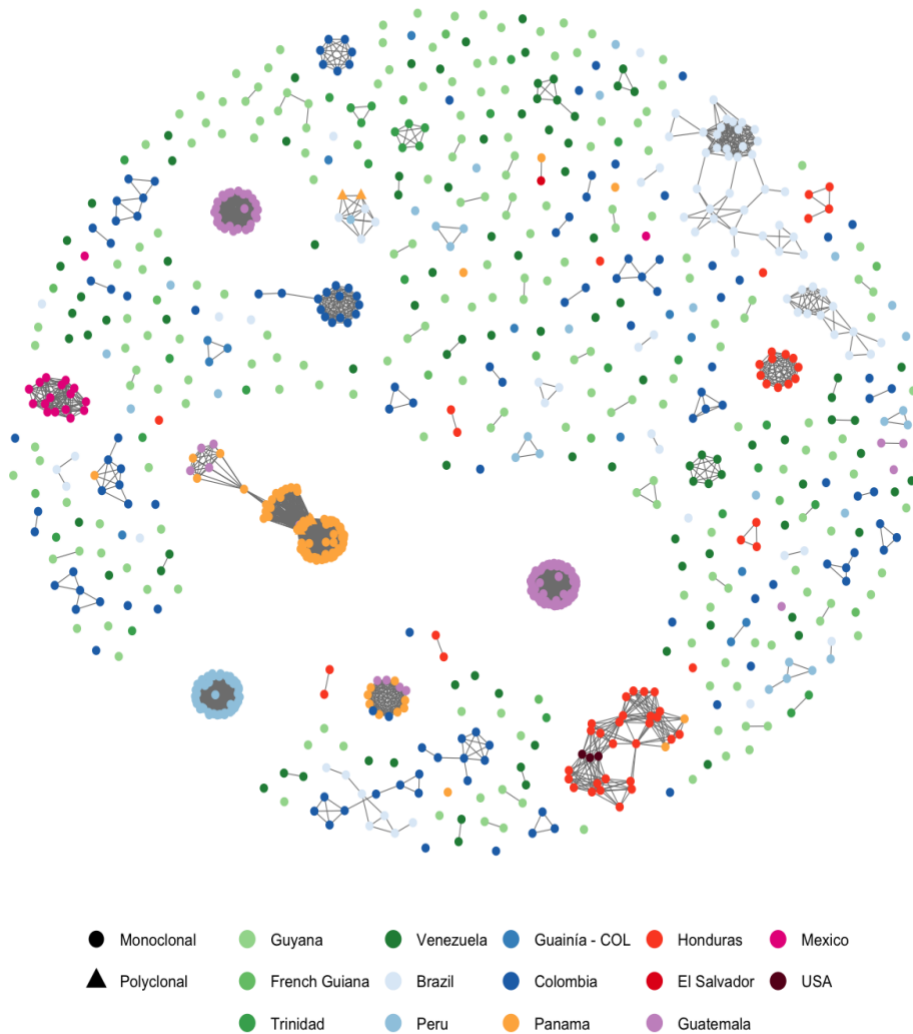


$IBD_{WGS} \geq 0.24$

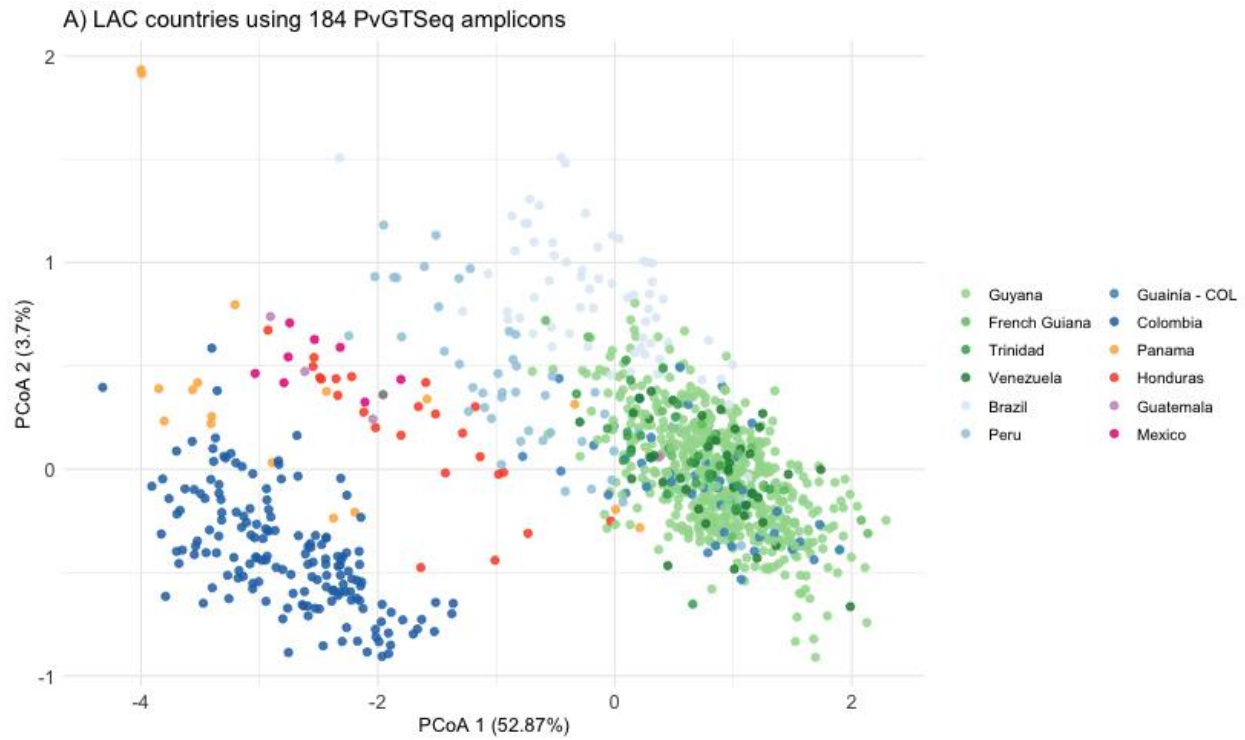


Supplementary figure 3.11: Identification of the appropriate IBD value to define ecological populations using WGS. The critical IBD value was selected based on the threshold at which cluster fusion begins to be observed due to genetic recombination between members of two genetically different groups. Edges are drawn for pairwise comparisons above the specified IBD_{WGS} threshold.

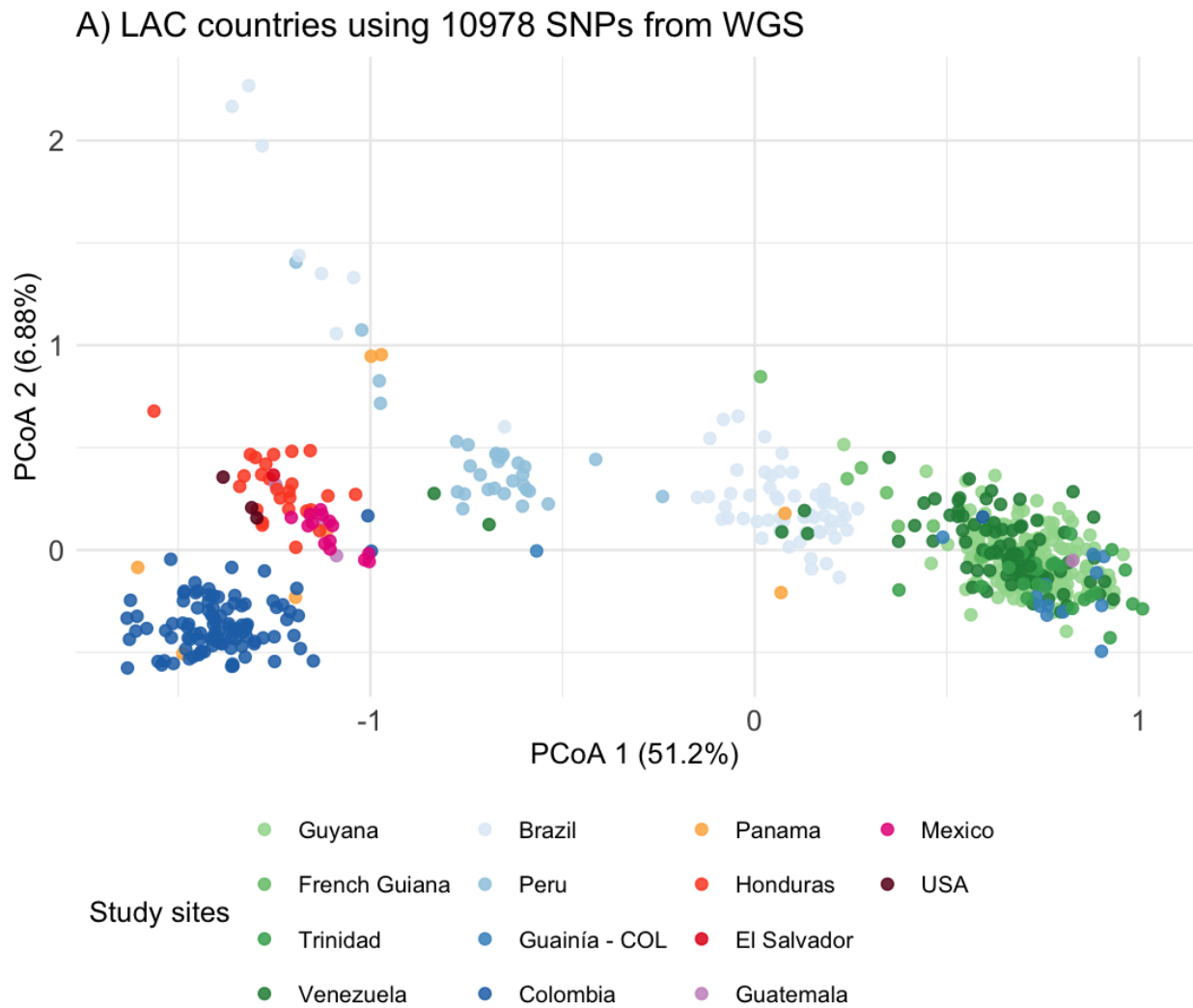
$$IBD_{WGS} \geq 0.37$$



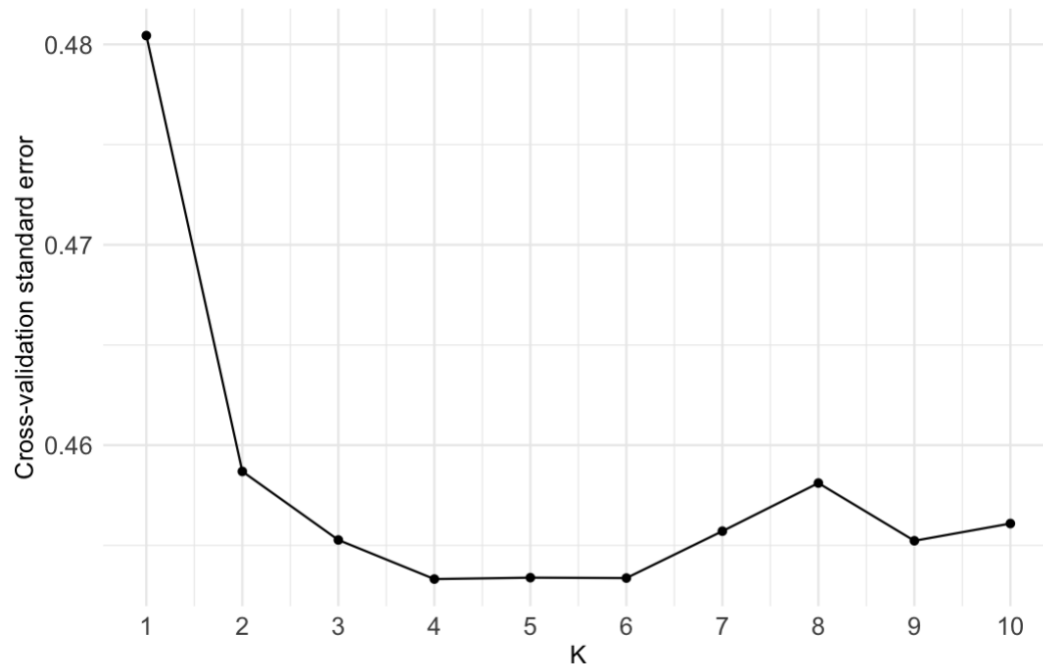
Supplementary figure 3.12: Identification of ecological groups using WGS genetic relatedness (IBD_{WGS}) in a network analysis. Each dot represents a genotyped sample. Colors are assigned based on the geographic origin of the sample (country). Shapes are based on the complexity of the sample infection. Polyclonal infections from Panama, Honduras, and Guatemala were only included in this analysis. Polyclonal infections from other countries were excluded. Branches represent pairwise IBD greater than 0.86. Clonal groups were defined as groups of four or more samples connected by branches. Labels above clusters are the labels used in the main text.



Supplementary figure 3.13: Principal coordinate analysis of LAC *P. vivax* populations using PvGTSeq markers.

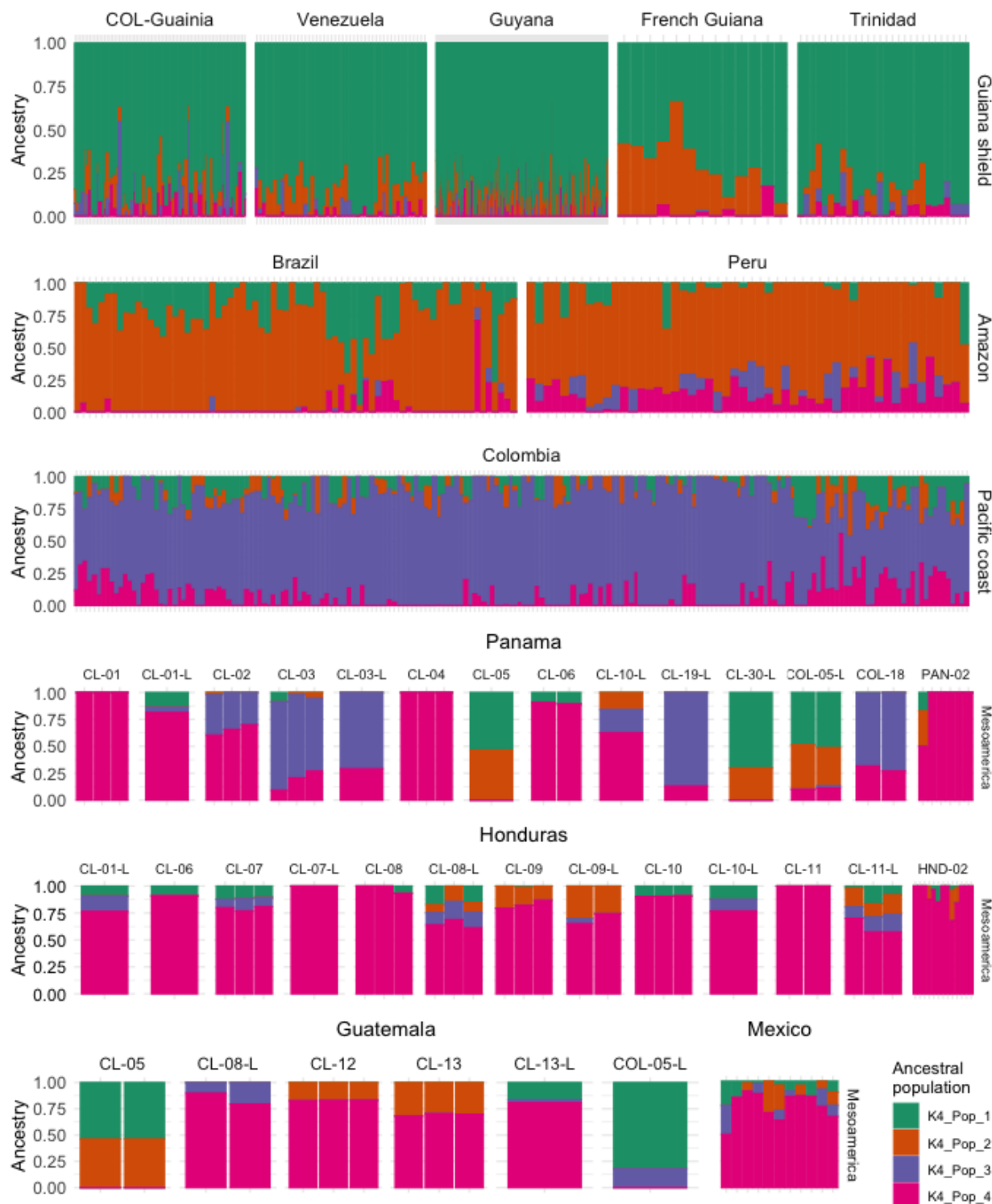


Supplementary figure 3.14: Principal coordinate analysis of LAC *P. vivax* populations using WGS.

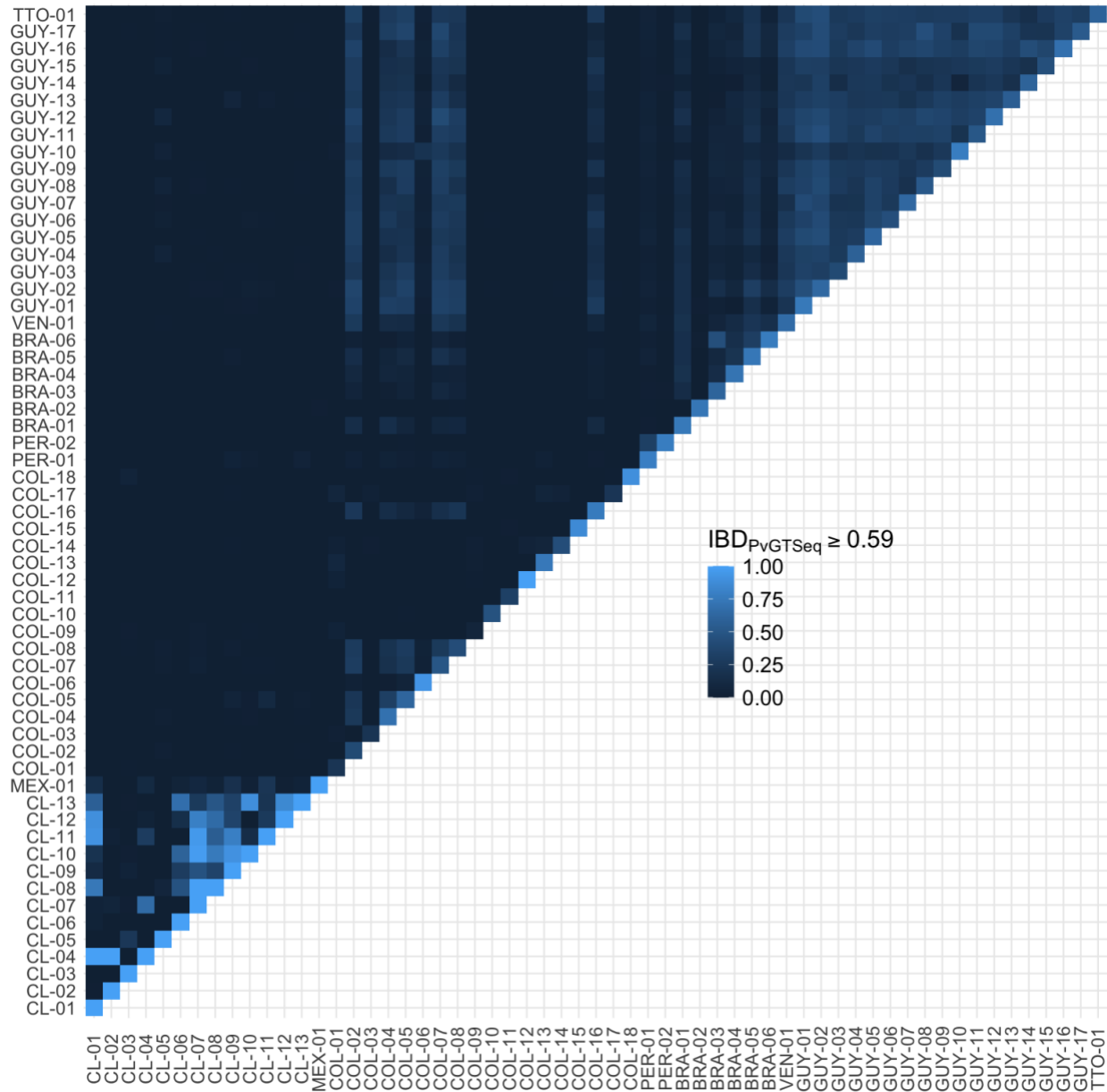


Supplementary figure 3.15: Possible number of ancestral populations in ADMIXTURE.

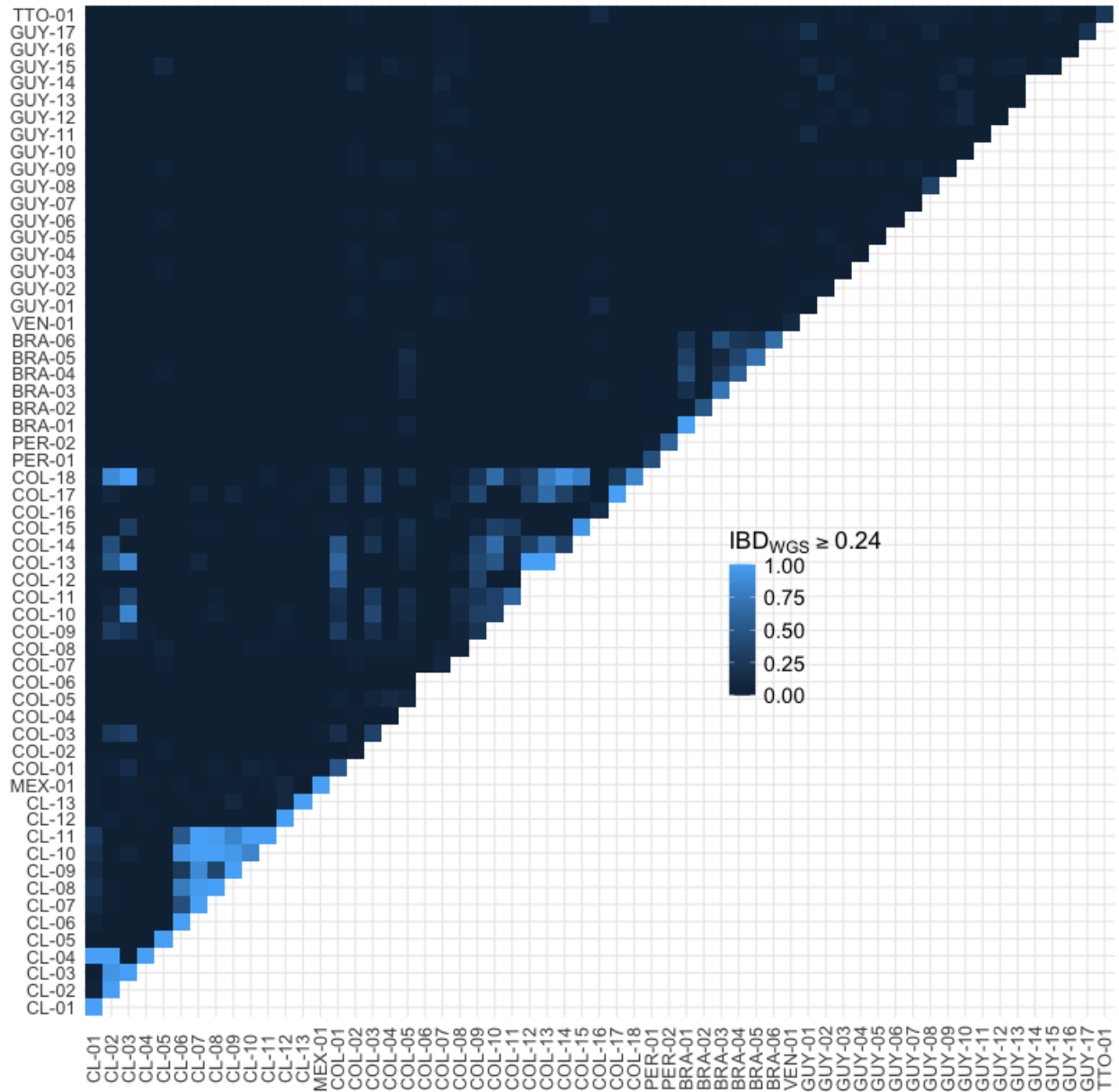
The most likely number of ancestral populations was determined by selecting the smallest number of populations that produced the lowest standard error across 10 cross-validations.



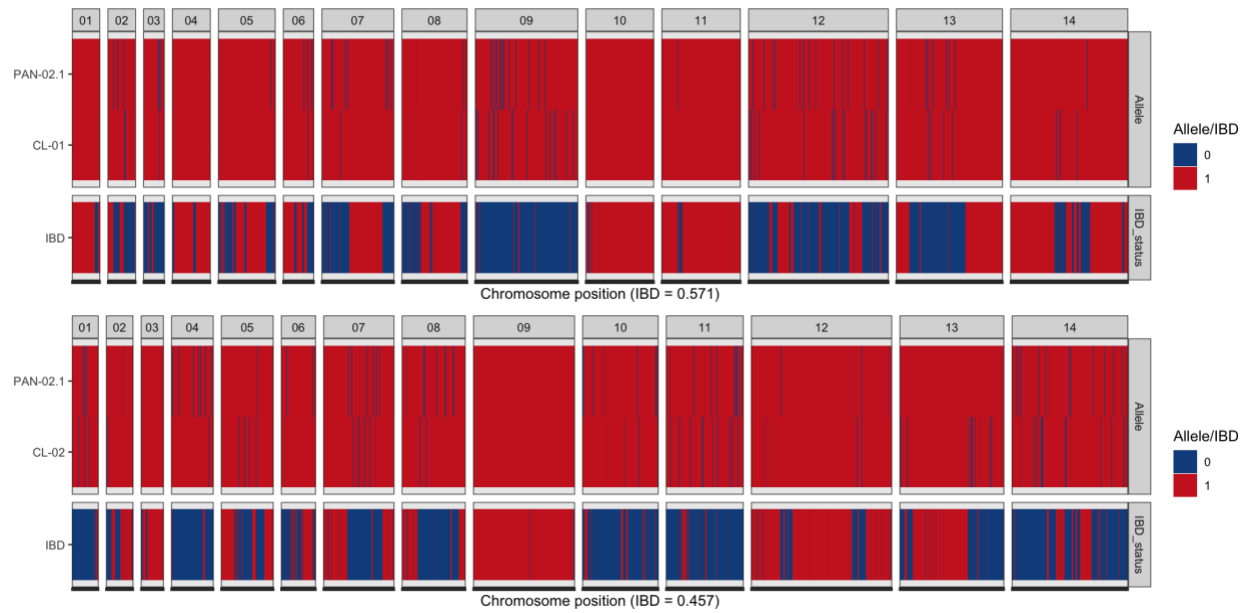
Supplementary figure 3.16: Pattern of population structure according to ADMIXTURE, assuming the presence of 4 ancestral populations. Each bar represents a non-clonal monoclonal sample. Bar colors represent the ancestral population, and the height of each bar represents the probability of ancestry of each sample in each ancestral population. Samples are grouped by geographic region and country of collection, and by the clonal or ecological groups assigned through network analysis (only for samples from Mesoamerica).



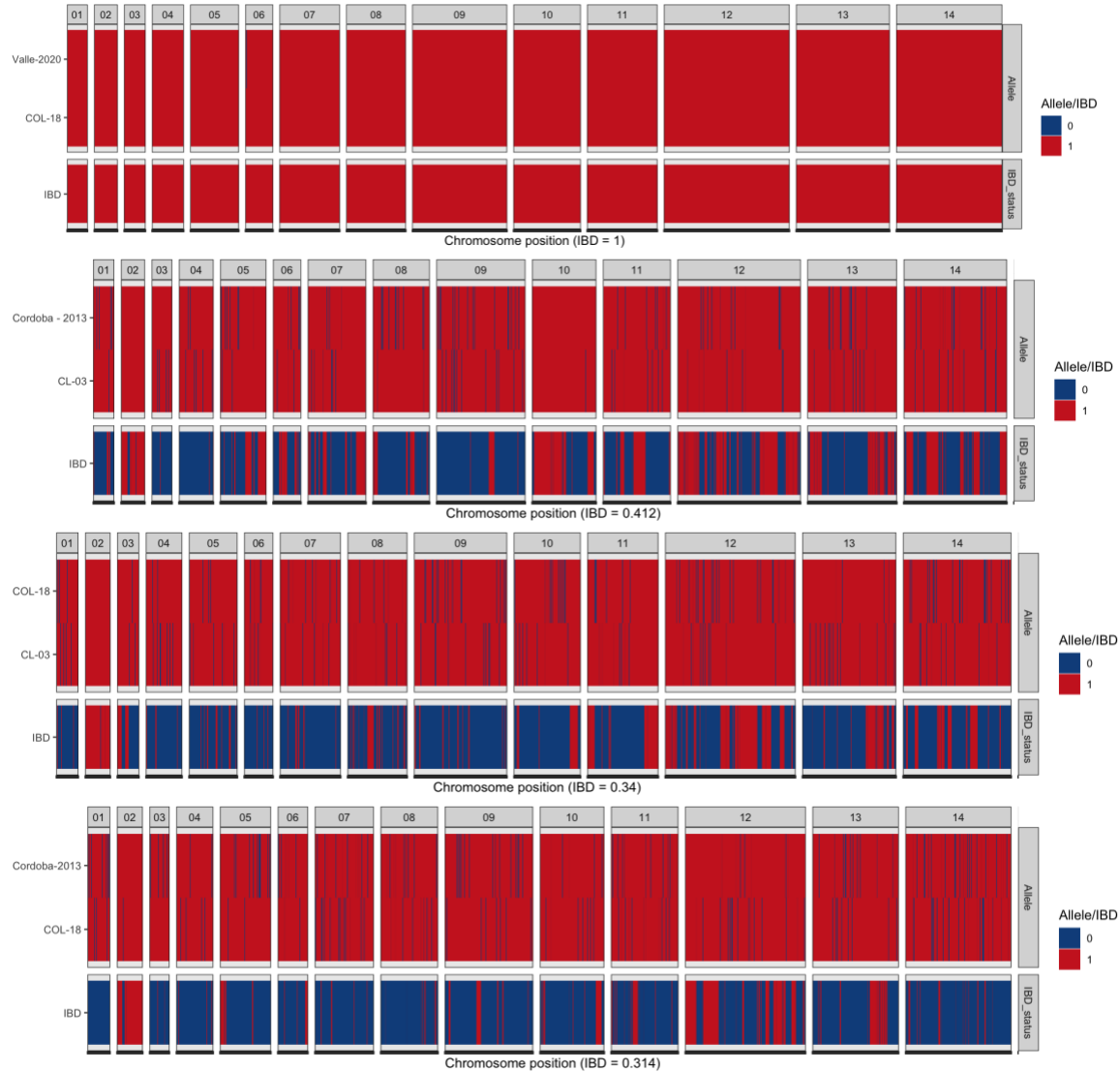
Supplementary figure 3.17: Heatmap of PvGTSeq pairwise genetic relatedness between clonal and ecological groups in LAC. Color indicates the fraction of pairwise comparisons with $IBD_{PvGTSeq} \geq 0.59$. We chose 0.59 because it is the critical value at which all samples are connected by at least one branch in the network.



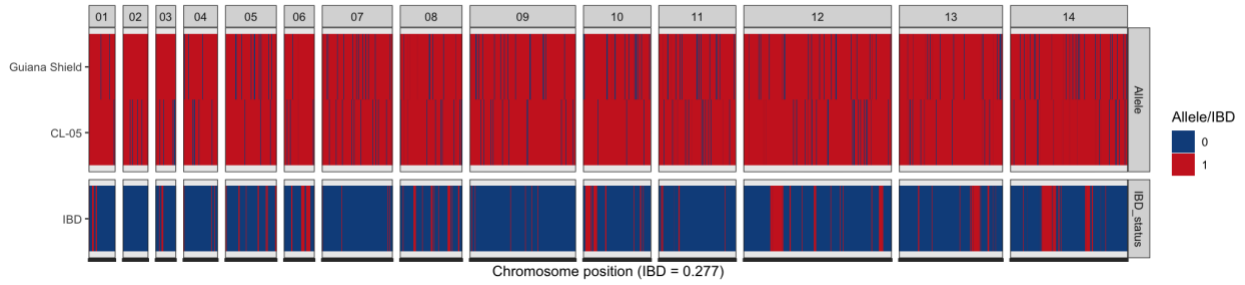
Supplementary figure 3.18: Heatmap of WGS pairwise genetic relatedness between clonal and ecological groups in LAC. Color indicates the fraction of pairwise comparisons with $IBD_{PvGTSseq} \geq 0.24$. We chose 0.24 because it is the critical value at which all samples are connected by at least one branch in the network.



Supplementary figure 3.19: Origin of PAN-02.1. Identification of long IBS and IBD segments between PAN-02.1. Each column represents a position in the genome, and each row indicates a haplotype or the IBD status between haplotypes for each biallelic SNP in the genome. Haplotype alleles are represented by two colors: blue (0) and red (1). For ease of reading, when haplotypes were identical by state at a given position, the allele was shown in red in both haplotypes. Otherwise, one haplotype was shown in red and the other in blue. IBD at each position was estimated using the Viterbi function in the hidden Markov chain model (HMM), and a value of 1 (red) was assigned if the position was identical by descent and 0 (blue) if the position was not identical by descent.



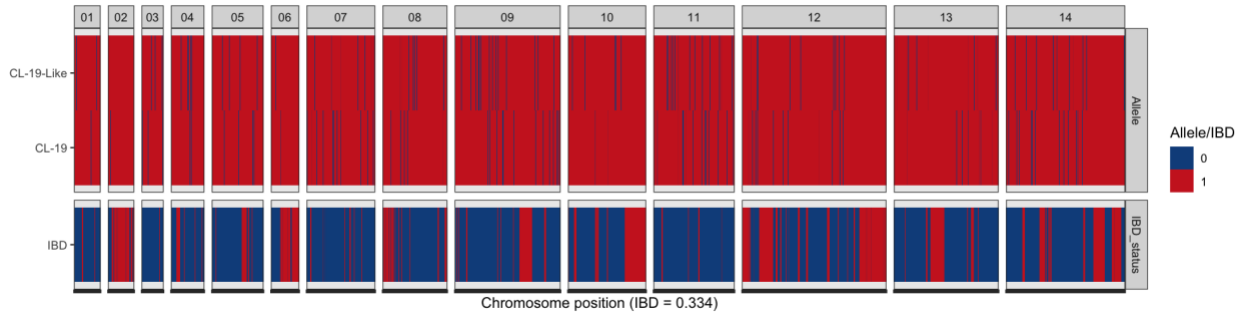
Supplementary figure 3.20: Origin of CL-03. Identification of long IBS and IBD segments between CL-03 and its closest genetic haplotypes in the dataset: a singleton sample from Buenaventura, Colombia, collected in 2020 (Valle-2020); a sample from Córdoba, Colombia, collected in 2013 (Cordoba-2013); and the ecological group COL-18 from Colombia, present in Buenaventura, Valle del Cauca, Colombia, in 2022. The bottom panel shows the relatedness between Valle-2020 and Cordoba-2013. Each column represents a position in the genome, and each row indicates a haplotype or the IBD status between haplotypes for each biallelic SNP in the genome. Haplotype alleles are represented by two colors: blue (0) and red (1). For ease of reading, when haplotypes were identical by state at a given position, the allele was shown in red in both haplotypes. Otherwise, one haplotype was shown in red and the other in blue. IBD at each position was estimated using the Viterbi function in the hidden Markov chain model (HMM), and a value of 1 (red) was assigned if the position was identical by descent and 0 (blue) if the position was not identical by descent.



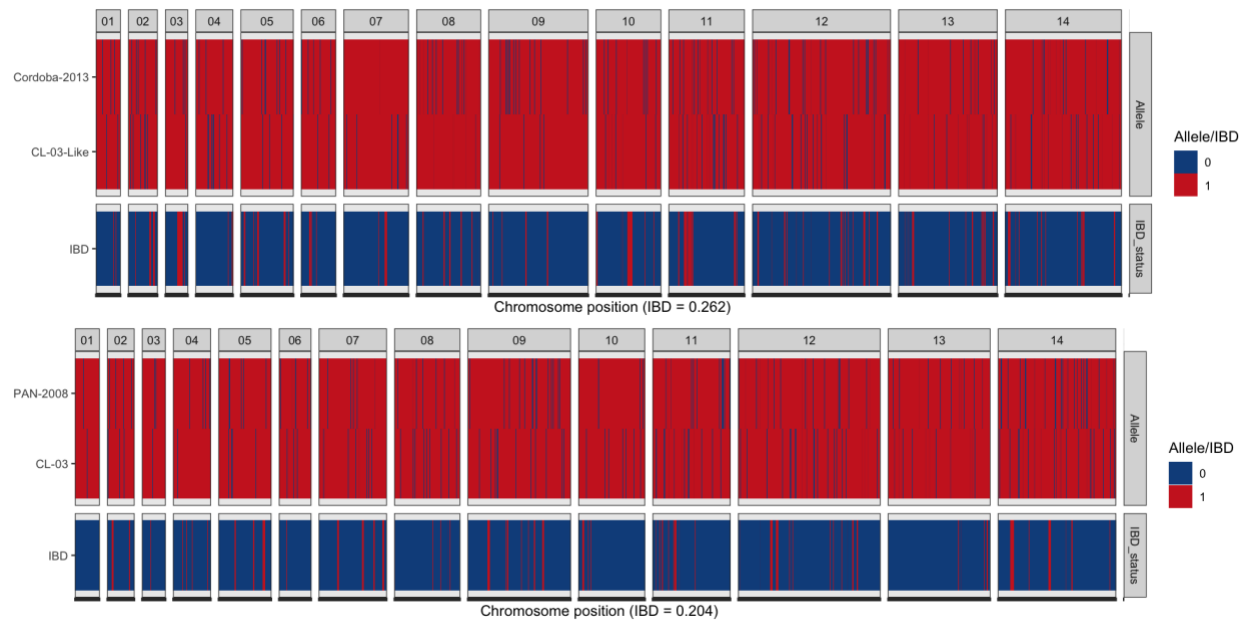
Supplementary figure 3.21: Origin of CL-05. Identification of long IBS and IBD segments between CL-05 and its closest genetic haplotype in the dataset, a sample from Trinidad and Tobago in the Guiana Shield. Each column represents a position in the genome, and each row indicates a haplotype or the IBD status between haplotypes for each biallelic SNP in the genome. Haplotype alleles are represented by two colors: blue (0) and red (1). For ease of reading, when haplotypes were identical by state at a given position, the allele was shown in red in both haplotypes. Otherwise, one haplotype was shown in red and the other in blue. IBD at each position was estimated using the Viterbi function in the hidden Markov chain model (HMM), and a value of 1 (red) was assigned if the position was identical by descent and 0 (blue) if the position was not identical by descent.



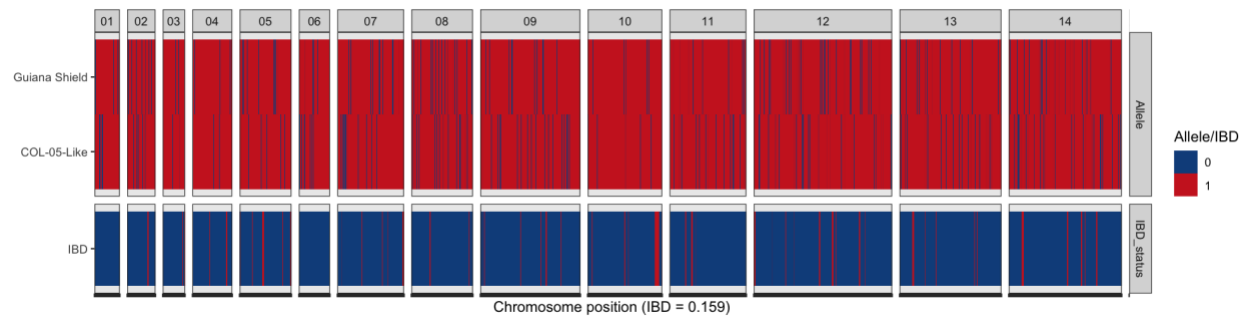
Supplementary figure 3.22: Origin of PAN-02.2. Identification of long IBS and IBD segments between PAN-02.2 and its two parental genetic groups, CL-04 and CL-05. Each column represents a position in the genome, and each row indicates a haplotype or the IBD status between haplotypes for each biallelic SNP in the genome. Haplotype alleles are represented by two colors: blue (0) and red (1). For ease of reading, when haplotypes were identical by state at a given position, the allele was shown in red in both haplotypes. Otherwise, one haplotype was shown in red and the other in blue. IBD at each position was estimated using the Viterbi function in the hidden Markov chain model (HMM), and a value of 1 (red) was assigned if the position was identical by descent and 0 (blue) if the position was not identical by descent. Note that identity by state and identity by descent are not necessarily reciprocal because the latter depends on allele frequency in the population and the status of neighboring positions. Panels in the bottom represent the genetic relatedness between the parental lineages CL-04 and CL-05.



Supplementary figure 3.23: Origin of CL-19-Like. Identification of long IBS and IBD segments between CL-19-Like and its closest clonal group in the dataset CL-19, a clone group from the Colombian Pacific coast. Each column represents a position in the genome, and each row indicates a haplotype or the IBD status between haplotypes for each biallelic SNP in the genome. Haplotype alleles are represented by two colors: blue (0) and red (1). For ease of reading, when haplotypes were identical by state at a given position, the allele was shown in red in both haplotypes. Otherwise, one haplotype was shown in red and the other in blue. IBD at each position was estimated using the Viterbi function in the hidden Markov chain model (HMM), and a value of 1 (red) was assigned if the position was identical by descent and 0 (blue) if the position was not identical by descent.

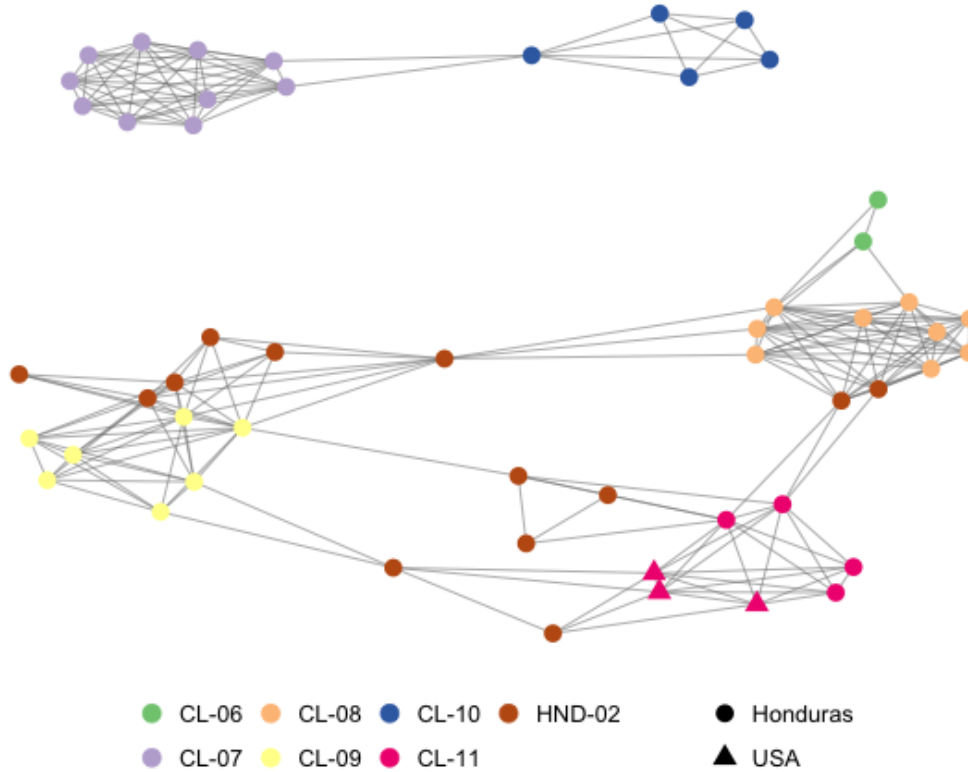


Supplementary figure 3.24: Origin of CL-03-Like. Identification of long IBS and IBD segments between CL-03-Like and its closest haplotype (a sample from Cordoba, Colombia in 2013) and the clonal group CL-03 in Panama. Each column represents a position in the genome, and each row indicates a haplotype or the IBD status between haplotypes for each biallelic SNP in the genome. Haplotype alleles are represented by two colors: blue (0) and red (1). For ease of reading, when haplotypes were identical by state at a given position, the allele was shown in red in both haplotypes. Otherwise, one haplotype was shown in red and the other in blue. IBD at each position was estimated using the Viterbi function in the hidden Markov chain model (HMM), and a value of 1 (red) was assigned if the position was identical by descent and 0 (blue) if the position was not identical by descent.



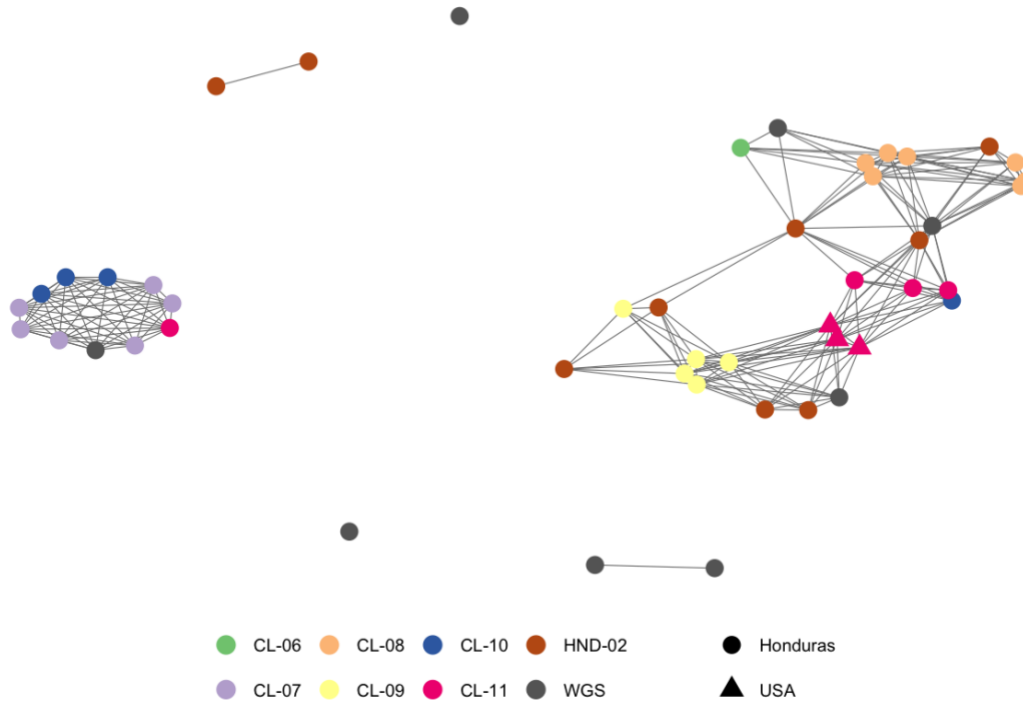
Supplementary figure 3.25: Origin of COL-05-Like in Panama. Identification of long IBS and IBD segments between COL-05-Like and its closest ecological group in the dataset COL-05, an ecological group from the region of Gainia in Colombia. Each column represents a position in the genome, and each row indicates a haplotype or the IBD status between haplotypes for each biallelic SNP in the genome. Haplotype alleles are represented by two colors: blue (0) and red (1). For ease of reading, when haplotypes were identical by state at a given position, the allele was shown in red in both haplotypes. Otherwise, one haplotype was shown in red and the other in blue. IBD at each position was estimated using the Viterbi function in the hidden Markov chain model (HMM), and a value of 1 (red) was assigned if the position was identical by descent and 0 (blue) if the position was not identical by descent.

$IBD_{PvGTSeq} \geq 0.74$

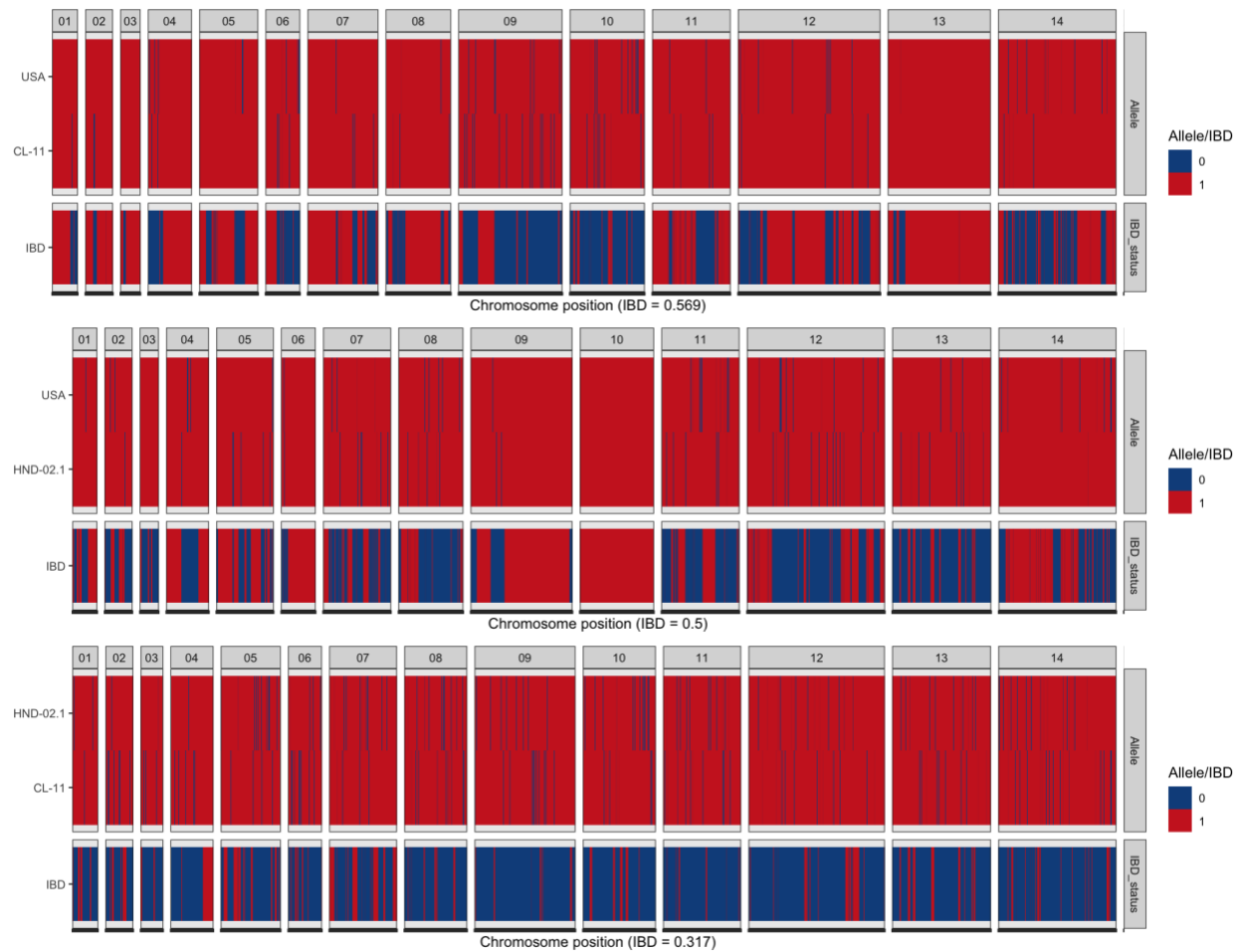


Supplementary figure 3.26: Fusion of parental clonal groups and their recombinant progeny in Honduras and USA using PvGTSeq. Each node represents a sample; colors indicate clonal-group assignment ($IBD_{PvGTSeq} \geq 0.86$), and node shape indicates the country of collection. In Guatemala, we distinguish Chiquimula from the other study areas because it is not considered endemic, and cases are reported only in migrant individuals. Edges are drawn for pairwise comparisons with $IBD_{PvGTSeq} \geq 0.74$ (threshold to identify ecological groups).

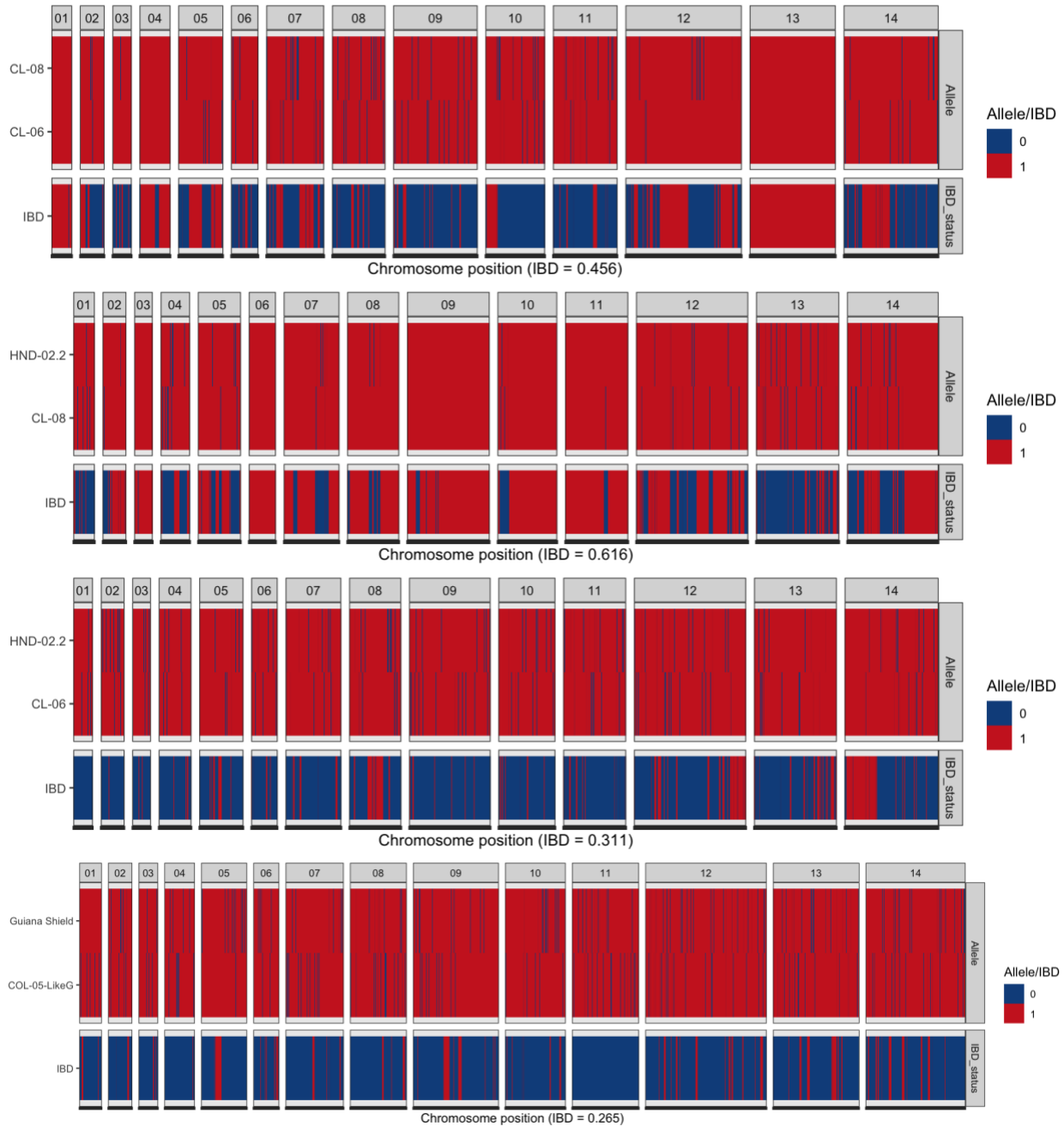
$IBD_{WGS} \geq 0.37$



Supplementary figure 3.27: Fusion of parental clonal groups and their recombinant progeny in Honduras and USA using WGS. Each node represents a sample; colors indicate clonal-group assignment ($IBD_{WGS} \geq 0.72$), and node shape indicates the country of collection. In Guatemala, we distinguish Chiquimula from the other study areas because it is not considered endemic, and cases are reported only in migrant individuals. Edges are drawn for pairwise comparisons with $IBD_{WGS} \geq 0.36$ (threshold to identify ecological groups).



Supplementary figure 3.28: Origin of the outbreak in USA. Identification of long IBS and IBD segments between the haplotype from USA and its closest clonal and ecological groups in Honduras (CL-11 and HND-02.1). Each column represents a position in the genome, and each row indicates a haplotype or the IBD status between haplotypes for each biallelic SNP in the genome. Haplotype alleles are represented by two colors: blue (0) and red (1). For ease of reading, when haplotypes were identical by state at a given position, the allele was shown in red in both haplotypes. Otherwise, one haplotype was shown in red and the other in blue. IBD at each position was estimated using the Viterbi function in the hidden Markov chain model (HMM), and a value of 1 (red) was assigned if the position was identical by descent and 0 (blue) if the position was not identical by descent.



Supplementary figure 3.29: Origin of CL-08 in Honduras. Identification of long IBS and IBD segments between CL-08 and its closest clonal and ecological groups in Honduras (CL-06 and HND-02.1). Each column represents a position in the genome, and each row indicates a haplotype or the IBD status between haplotypes for each biallelic SNP in the genome. Haplotype alleles are represented by two colors: blue (0) and red (1). For ease of reading, when haplotypes were identical by state at a given position, the allele was shown in red in both haplotypes. Otherwise, one haplotype was shown in red and the other in blue. IBD at each position was estimated using the Viterbi function in the hidden Markov chain model (HMM), and a value of 1 (red) was assigned if the position was identical by descent and 0 (blue) if the position was not identical by descent.