GENOMIC EPIDEMIOLOGY FOR MALARIA: NOVEL APPLICATION OF GEOSPATIAL METHODS, NEW GENOMIC MARKERS, AND POPULATION-LEVEL INSIGHTS

by

Alfred Hubbard

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2024

Approved by:

____

Dr. Daniel Janies

____

Dr. Elizabeth Cooper

____

Dr. Alex Dornburg

____

Dr. Jean-Claude Thill

ABSTRACT

ALFRED HUBBARD. Genomic Epidemiology for Malaria: Novel application of geospatial methods, new genomic markers, and population-level insights. (Under the direction of DR. DANIEL JANIES)

Genomic epidemiology is the use of genetic data to characterize and explain disease occurrence and transmission. Application of these methods to malaria has already yielded substantial benefits, such as identification and surveillance of drug resistance genotypes. However, the potential for genomic epidemiology to accelerate progress towards malaria eradication is far from fully realized. This dissertation demonstrates new applications of genomic information to questions that are impossible to address with conventional epidemiological data. First, the value of correlating genetic and environmental distances to understand the drivers of *Plasmodium falciparum* transmission is showcased with microsatellite data from 44 sites in Western Kenya. Second, the design and validation of a new panel of genetic markers, amplicons with multiple variant nucleotides on each short read that yield microhaplotypes, is presented for *P. vivax*, enabling sensitive, scalable characterization of within-host diversity in multi-strain infections. Finally, a similar panel of amplicons for *P. falciparum* is applied to samples from eight countries throughout Africa, yielding insights into continent scale transmission dynamics. The analysis of environmental drivers revealed the Winam Gulf of Lake Victoria as a barrier to malaria transmission, a conclusion that would be impossible to reach rigorously without this novel methodology. The new *P. vivax* panel yielded quality sequences and detected expected patterns of genetic relatedness, indicating this tool is ready for broad application. The *P. falciparum* microhaplotype analysis identified subtle patterns of genetic relatedness and surprisingly little relationship between within-host diversity and incidence, highlighting the potential of this technique but also a need for future work on the interpretation of the resulting data. This dissertation expands the scope of questions about malaria epidemiology that can be answered with genomic data and argues that routine application of these methods could accelerate progress towards malaria eradication.

DEDICATION

To my wife, Erin, for patient support at every juncture and for seeing my challenges with a clarity that I could not.

ACKNOWLEDGEMENTS

At UNC Charlotte, I am indebted to Dr. Daniel Janies, for being a constant source of sound guidance on both matters related to this dissertation and the broader scientific enterprise. I am also grateful to the other members of my committee, Dr. Elizabeth Cooper, Dr. Alex Dornburg, and Dr. Jean-Claude Thill, for their suggestions and their enthusiasm, both of which helped me greatly. Outside of our university, I received tremendous assistance from Drs. Elizabeth Hemming-Schroeder and Eugenia Lo, who provided the connections necessary to obtain my data as well as the lion's share of the technical guidance necessary to analyze it. Finally, this work rests on the shoulders of the people who collected, transported, and prepared the samples for sequencing. Some of their names I know, some I don't, but their collaboration is the reason an American who has never been to Africa can write a dissertation such as this.

The final year of my studies was supported by the UNC Charlotte Graduate Summer School Fellowship and the Lucille P. and Edward C. Giles Dissertation-Year Graduate Fellowship. This funding let me fully focus my attention on this project during a critical time, which was immensely beneficial.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABI Applied Biosystems

AICc Corrected Akaike information criterion

CADM Congruence among distance matrices

DBS Dried blood spot

DEM Digital elevation model

DNA Deoxyribonucleic acid

EIR Entomological inoculation rate

GWR Geographically-weighted regression

IBD Identity-by-descent

ICEMR International Center for Excellence for Malaria Research

IQR Inter-quartile range

LD Linkage disequilibrium

LST Land surface temperature

MOI Multiplicity of infection

NCBI National Center for Biotechnology Information

PCA Principal component analysis

PCoA Principal coordinates analysis

PCR Polymerase chain reaction

QC Quality control

qPCR Quantitative polymerase chain reaction

rRNA Ribosomal ribonucleic acid

SNP Single nucleotide polymorphism

SWGA Selective whole genome amplification

UTM Universal Transverse Mercator

WGS Whole genome sequence

CHAPTER 1: Introduction

Malaria is a deadly mosquito-borne disease responsible for tremendous burden in tropical countries. In 2022, there were an estimated 249 million cases of malaria spread across 85 countries, which led to 608,000 deaths (WHO, 2023). Local elimination of malaria has been achieved in many places, and global eradication is considered both possible and the only sustainable option to control the disease (Feachem et al., 2019). Substantial progress has been made towards the goal of global eradication, but that progress has stalled (WHO, 2023). This stall in progress is partially explained by the impact of the COVID-19 pandemic. However, the stall is also due to both declining effectiveness of the preeminent interventions used to combat malaria (WHO, 2023) and inadequate funding to deploy these measures (Feachem et al., 2019). Both of these challenges can be addressed with research, by developing new drugs and vector control technologies and by enhancing our understanding of malaria epidemiology to target limited control resources more efficiently. This second objective is the subject of this dissertation.

1.1 Malaria biology

To understand malaria epidemiology, it is first necessary to understand malaria biology. Malaria is caused by blood-borne parasites in the genus *Plasmodium* of the phylum Apicomplexa. There are several species of this parasite, six of which are known to cause disease in humans: *Plasmodium falciparum*, *P. vivax*, *P. knowlesi*, *P. malariae*, *P. ovale curtisi*, and *P. ovale wallikeri* (Ashley et al., 2018). Of these, *P. falciparum* and *P. vivax* are by far the most important in terms of cases and deaths, with *P. falciparum* causing the majority of both and *P. vivax* causing most of the remainder (WHO, 2023).

The biology of these two species is similar in many aspects, but there is a critical difference in their life cycles. In *P. falciparum*, when an infected mosquito feeds on a human, haploid sporozoites are transmitted to the bloodstream, where they invade and reproduce asexually in red blood cells. A subset of the parasites in this erythrocytic stage ultimately become male and female gametocytes. These gametocytes are transmitted back to the mosquito as part of a blood meal, and undergo sexual reproduction in the gut of the mosquito. The resulting diploid parasites pass through the gut wall and release haploid sporozoites, which travel to the mosquito salivary glands, allowing the cycle to repeat. All of these steps also occur in *P. vivax*, with one important addition: the sporozoites can also enter the human liver and remain dormant as hypnozoites. As explained below, these dormant hypnozoites have substantial epidemiological implications. (Ashley et al., 2018)

1.2 Malaria epidemiology: Concepts and importance

The primary goals of epidemiological surveillance for malaria are to measure the prevalence (proportion of the population with malaria), incidence (probability of occurrence of malaria in a population per unit time), and transmission (passage of the malaria parasite from one host to another) in a given area, and to characterize the drivers of these metrics. Incidence and prevalence can be calculated from clinical data relatively easily, but measuring transmission is considerably more difficult. The gold standard of malaria transmission metrics is the entomological inoculation rate (EIR), which is the number of infectious mosquito bites received by a person per unit time (Bousema et al., 2012; Koepfli & Mueller, 2017). Unfortunately, this information is laborious to obtain and cannot practically be gathered at scale (Bousema et al., 2012).

However, without an effective method to measure transmission, it is impossible to understand whether different regions are connected or isolated in terms of the transport of parasites via human and mosquito movement. This is important in low transmission settings, where reintroduction of malaria from returning travelers can be an important component of transmission (Wesolowski et al., 2018). Understanding transmission between locations also enables prediction of whether epidemiologically-relevant parasite traits with a genetic basis, such as resistance to therapeutic drugs or evasion of rapid diagnostic tests, are likely to spread (Dalmat et al., 2019).

In *P. vivax*, measuring transmission allows study of an additional phenomenon: relapse, caused by the dormant hypnozoites in the liver reactivating and entering the blood to cause disease (Ashley et al., 2018). This relapse can occur months or even years later (Ashley et al., 2018). Such a relapse does not constitute a transmission event and must be handled differently from a control perspective, creating an additional challenge and motivation for measuring transmission accurately in *vivax* malaria (Koepfli & Mueller, 2017).

The final reason measuring transmission matters is that it enables study of the drivers of transmission, which must be understood and perhaps manipulated to achieve reductions in malaria cases in a given area (Castro, 2017). As a vector-borne disease, there are numerous factors potentially relevant to malaria epidemiology: factors related to mosquito behavior and survival, such as temperature, humidity, availability of breeding habitat, and use of vector control measures; factors related to human exposure to mosquitoes, such as mobility patterns, housing quality, and sleeping under a bed net; and factors related to the likelihood that a human inoculated with malaria will develop disease, such as acquired immunity and use of prophylactic drugs (Sadoine et al., 2018). Of these groups, the first two are pertinent to transmission, and are discussed in greater detail in Chapter 2.

1.3 Malaria epidemiology: Genomic approaches

1.3.1 Applications and metrics

The importance of understanding transmission and the difficulty of measuring it directly has led to the application of a new sub-field of epidemiology to malaria: genomic epidemiology, or the practice of using genetic characteristics of the disease-causing agent to measure and explain prevalence, incidence, and transmission. Genetic relatedness between regions can be interpreted as a proxy for transmission between regions, and genetic diversity or genetic relatedness within a region can be interpreted as a proxy for transmission within a region (Wesolowski et al., 2018). Supposed drivers of transmission can be correlated with these metrics to assess their importance (see Chapter 2) and, in *P. vivax*, longitudinal genetic sampling can be used to distinguish relapse from reinfection with a new strain (Auburn et al., 2021).

In addition, there are some epidemiological questions that cannot even be formulated except in terms of genetic characteristics. The most noteworthy example in malaria is identifying the number and composition of genetically-distinct strains present within one host. This is a common occurrence with both *P. falciparum* and *P. vivax*, and the number of strains present in an infection, or the multiplicity of infection (MOI), is tied to the nature of the transmission environment (Lopez & Koepfli, 2021). Whether caused by a single bite from a mosquito that itself carried multiple strains (cotransmission) or by successive bites from different mosquitoes carrying different strains of malaria (superinfection), polyclonal infections (i.e., infections where MOI > 1) are more likely to occur in higher transmission environments (Lopez & Koepfli, 2021). For this reason, MOI and other measures of within-host genetic diversity have been suggested as useful metrics of transmission (Camponovo et al., 2023).

However, the interpretation of these metrics is not always straightforward. In *P. falciparum*, the theoretical expectation is that, as transmission increases, genetic diversity increases, MOI increases, and genetic relatedness decreases (Koepfli & Mueller, 2017). In reality, the correlations between these metrics and prevalence, while present, are weak (Lopez & Koepfli, 2021; Paschalidis et al., 2023), leading to concerns over how well they represent transmission (due to the laborious nature of such experiments, little data is available on the correlation between EIR-measured transmission and genetic metrics). In *P. vivax*, the possibility of any given infection being caused by relapse weakens the theoretical relationship between genetic metrics and transmission and, indeed, correlations between MOI and prevalence have been found to be even weaker for *vivax* malaria (Lopez & Koepfli, 2021). Therefore, further work is needed to clarify and validate the applicability of genetic metrics to measuring transmission (Dalmat et al., 2019).

1.3.2 Models, markers, and laboratory protocols

Setting aside the questions of how to interpret these genetic metrics, there are also numerous research questions around how these measures should, themselves, be measured. There are various statistical approaches to measuring genetic differentiation, various types of genetic data that may be used with these approaches, and various sample collection and preparation methods for obtaining such genetic data.

Regarding statistical measures, there has been a movement in recent years away from classic measures of genetic distance, such as Wright's F_{ST}, towards measures of genetic relatedness based on the concept of identity-by-descent (IBD; Thompson, 2013). If two alleles are the same in two individuals, they are identical-by-state. However, this could be due to chance. From an interpretation standpoint, it is much more useful to know if the alleles are the same because they were the same in the most recent common ancestor, in which case they are considered to be identical-by-descent (also abbreviated as IBD). Several models for estimating IBD relatedness have been published, which use population-level allele frequencies (Gerlovina et al., 2022) and, in some cases, hidden Markov models (Henden et al., 2018; Schaffner et al., 2018) to estimate the probability that shared genotypes are IBD.

Relatedness and other genetic measures have conventionally been estimated for malaria with three types of genetic data: microsatellites, single nucleotide polymorphisms (SNPs), and whole genome sequence (WGS). Microsatellite markers are segments of the genome where a short series of nucleotides is repeated many times. They are non-coding DNA, which makes them attractive for population genetics and genetic epidemiology because they are, in theory, selectively neutral, a common assumption of population genetics models and statistics. As the name implies, SNPs are single nucleotides that vary within a population. A single microsatellite or SNP genotype is not very informative from a population genetics standpoint, but when panels of multiple loci (typically 10-20 microsatellites or 25-200 SNPs) are genotyped for a set of samples they may be sufficient to identify differences in diversity and relatedness (Koepfli & Mueller, 2017). However, neither performs well at characterizing the diversity within a host (Argyropoulos et al., 2023; Koepfli & Mueller, 2017). WGS is exactly what it sounds like, and, while useful for identifying new variants and validating other genotyping methods, the cost of data production and storage is not scalable (Neafsey et al., 2021). The tradeoffs between these data types has led to recent interest in a fourth approach: microhaplotype loci, which are segments of the genome where multiple SNPs are close enough together that they can be genotyped on a single sequencing read. These markers have a similar production cost to SNPs and yet have much more power for relationship inference

(Baetscher et al., 2018). They are discussed in greater detail in Chapters 3 and 4.

Finally, the quality of sequencing data obtained can be substantially affected by sample collection and preparation techniques. The most scalable and convenient method for obtaining malaria samples is to collect small volumes of blood from finger or heel pricks and subsequently blot and dry them on filter paper (Bereczky et al., 2005; Färnert et al., 1999). These dried blood spot (DBS) samples do not require refrigeration in the short term and are easier to collect than the alternative: whole blood samples obtained from venous blood draws. However, host DNA is much more abundant in both types of sample than parasite DNA, especially for low density infections, which means some method for increasing the concentration of parasite DNA relative to host DNA is necessary to obtain quality parasite DNA sequences (Neafsey et al., 2021). This is possible for whole blood samples through leucocyte depletion with cellulose columns (Auburn et al., 2011), but until recently no solution to this problem existed for DBS samples.

This changed with the adaptation of selective whole genome amplification (SWGA) protocols to first *P. falciparum* (Oyola et al., 2016) and subsequently *P. vivax* (Cowell et al., 2017). These methods use a special set of primers and a specific DNA polymerase to greatly amplify the concentration of parasite DNA relative to host DNA. The primers in question are short mers designed to target motifs that are much more common in the parasite genome than in the host genome (Cowell et al., 2017; Oyola et al., 2016). The second ingredient is phi29 DNA polymerase, which tends to displace double-stranded DNA and thus preferentially opens up more primer binding sites in DNA template sequences that already have higher frequencies of primer binding (the parasite, in this case, thanks to the primers described above; Cowell et al., 2017). SWGA allows a high yield of parasite DNA to be obtained from DBS samples (Cowell et al., 2017; Oyola et al., 2016), enabling much more scalable application of WGS and other sequencing technologies where this is important, such as amplicon deep sequencing of microhaplotypes (see chapters 3 and 4).

1.4 Outline of remaining chapters

In this dissertation, the current state of genomic epidemiology in malaria described above is advanced in three major ways. In Chapter 2, a new epidemiological application for genetic relatedness between geographic locations is demonstrated: using environmental data layers and resistance surface modeling to gain insight into the drivers of malaria transmission. In Chapter 3, a panel of microhaplotype loci is developed and validated for *P. vivax*, providing the foundation for subsequent genomic epidemiology studies that take advantage of the deep information on within-host diversity furnished by this data type. Finally, in Chapter 4, an analogous panel of microhaplotype loci for *P. falciparum* is applied at an unprecedented geographic scale to elucidate the epidemiological utility of both this data type and the genetic metrics of transmission it supports. Chapter 5 summarizes the key outcomes of these three projects and the implications of those results for the future of genomic epidemiology in malaria.

CHAPTER 2: Implementing landscape genetics in molecular epidemiology to determine drivers of vector-borne disease: A malaria case study

Article Authors: Alfred Hubbard, Elizabeth Hemming-Schroeder, Maxwell Gesuge Machani, Yaw Afrane, Guiyun Yan, Eugenia Lo, Daniel Janies

2.1 Background

Progress towards malaria elimination has stalled (WHO, 2023), in part because an inadequate understanding of how the environment influences transmission has hampered epidemiological modeling and targeting of control measures (Rabinovich et al., 2017). The *Anopheles* mosquitoes that transmit malaria rely on favorable environmental conditions to feed and reproduce successfully, and human movement patterns that spread malaria are influenced by the available infrastructure. This means the environment plays a critical role in understanding and combating malaria transmission (Castro, 2017).

Several prior studies have sought to explain the environmental drivers of malaria transmission using methods such as geographically-weighted regression and Bayesian risk modeling (Canelas et al., 2016). The former is logical for demographic or socioeconomic drivers that are tied to the host, but it does not capture the full variability in environmental drivers. Malaria transmission and gene flow can occur across a wide range of geographic scales, and thus models that include the space between sample locations will yield greater insights. Risk maps inferred with Bayesian methods or kriging can be compared to environmental data layers for a more spatially distributed understanding of the drivers of transmission, but these methods typically do not incorporate varying levels of connectivity and gene flow between different locations.

In previous studies, we have addressed these limitations using resistance surface models that seek to explain genetic distances between populations in terms of environmental or resistance distances (Kepple et al., 2021; Lo, Hemming-Schroeder, et al., 2017; Lo, Lam, et al., 2017). These distances represent the difficulty of traveling between two locations in a manner that considers both geographic distance and the properties of the intervening landscape, allowing assessment of which landscape properties obstruct or enable gene flow. Although parasites do not directly traverse the landscape, the landscape indirectly influences parasite gene flow through the impact of the environment on the movements of parasite hosts (i.e., mosquitoes and humans). Thus, resistance surfaces are a promising tool for studying vector-borne disease (Hemming-Schroeder et al., 2018). Our prior work has begun to build a better understanding of the spatial determinants of malaria transmission, but small numbers of study sites have limited the scope of the conclusions.

In this study, we use resistance surface analysis to examine the drivers of malaria transmission in Western Kenya, a malaria endemic area with moderate to high levels of transmission. This is the first time these methods have been applied in malaria using more than 10 study sites. We also perform conventional population genetics analyses, such as Bayesian- and ordination-based clustering techniques, to contextualize our landscape genetics results and enable comparison with the existing body of literature. We show novel patterns and drivers of genetic differentiation, and thus, heterogeneity in transmission, and provide a rigorous demonstration of the utility of landscape genetics in the study of vector-borne diseases.

2.2 Materials and Methods

2.2.1 Scientific and ethical statement

Scientific and ethical clearance for sample collection and preparation was given by the institutional scientific and ethical review boards of the Kenya Medical Research Institute, Kenya and the University of California, Irvine, USA. Written informed consent/assent for study participation was obtained from all consenting heads of households, parents/guardians (for minors under age of 18), and each individual who was willing to participate in the study.

2.2.2 Study area

Malaria transmission in Western Kenya is moderate-to-high with contemporaneous prevalence estimates ranging from 40-60% in the lowlands near Lake Victoria (Okoyo et al., 2015; Zhou et al., 2016) to around 15% in the highlands (Zhou et al., 2016). There are pronounced gradients in elevation, rainfall, and temperature in our study region (Figure 2.1). Temperature and moisture are both crucial to mosquito survival and activity, and the reduced, seasonal malaria transmission observed in the highlands of Kenya is explained by lower temperature, rainfall, and humidity (Kenya NMCP et al., 2016).

2.2.3 Sample collection and genotyping

A total of 1,804 PCR-confirmed *P. falciparum* DNA samples collected in 2012 and 2013 across 44 sites in Western Kenya (Figure 2.2) were included. These samples were selected from 11,000 asymptomatic school children aged 3 to 12 years, as described in our earlier study (Lo et al., 2015). Roughly 50 μl of blood collected by finger prick was blotted on Whatman 3MM filter paper, from which *P. falciparum* DNA was extracted using the Saponin-Chelex method (Bereczky et al., 2005).

Eight single-copy microsatellite loci (Table 2.1) were genotyped for *P. falciparum*. Each PCR involved 2 μl of genomic DNA in 2 mM $MgCl_2$, 2 μM of each primer (forward primers were labeled with fluorescent dyes; Applied Biosystems, Foster City, CA), and 10 μl of 2xDreamTaq Green PCR Master Mix (Thermo Scientific, Waltham, MA). PCR cycling conditions were: 2 min, 94 °C; (30 sec, 94 °C; 40 sec, 58 °C; 50 sec, 72 °C) for 40 cycles; and 5 min, 72 °C. After amplification, the products were combined into three groups based on size and separated on an ABI 3730 sequencer. The allele sizes were recorded using two methods, depending on the sample: manual visualization using Peak Scanner and automated extraction in the Thermo Scientific Cloud Microsatellite Analysis Software. In both cases, a threshold of 300 relative fluorescent units was used for peak detection to filter out background noise. For each microsatellite, the dominant allele and any other alleles at least 33% of the dominant allele's height were scored. 574 samples were processed with both methods and 79% of the overlapping alleles were scored identically.

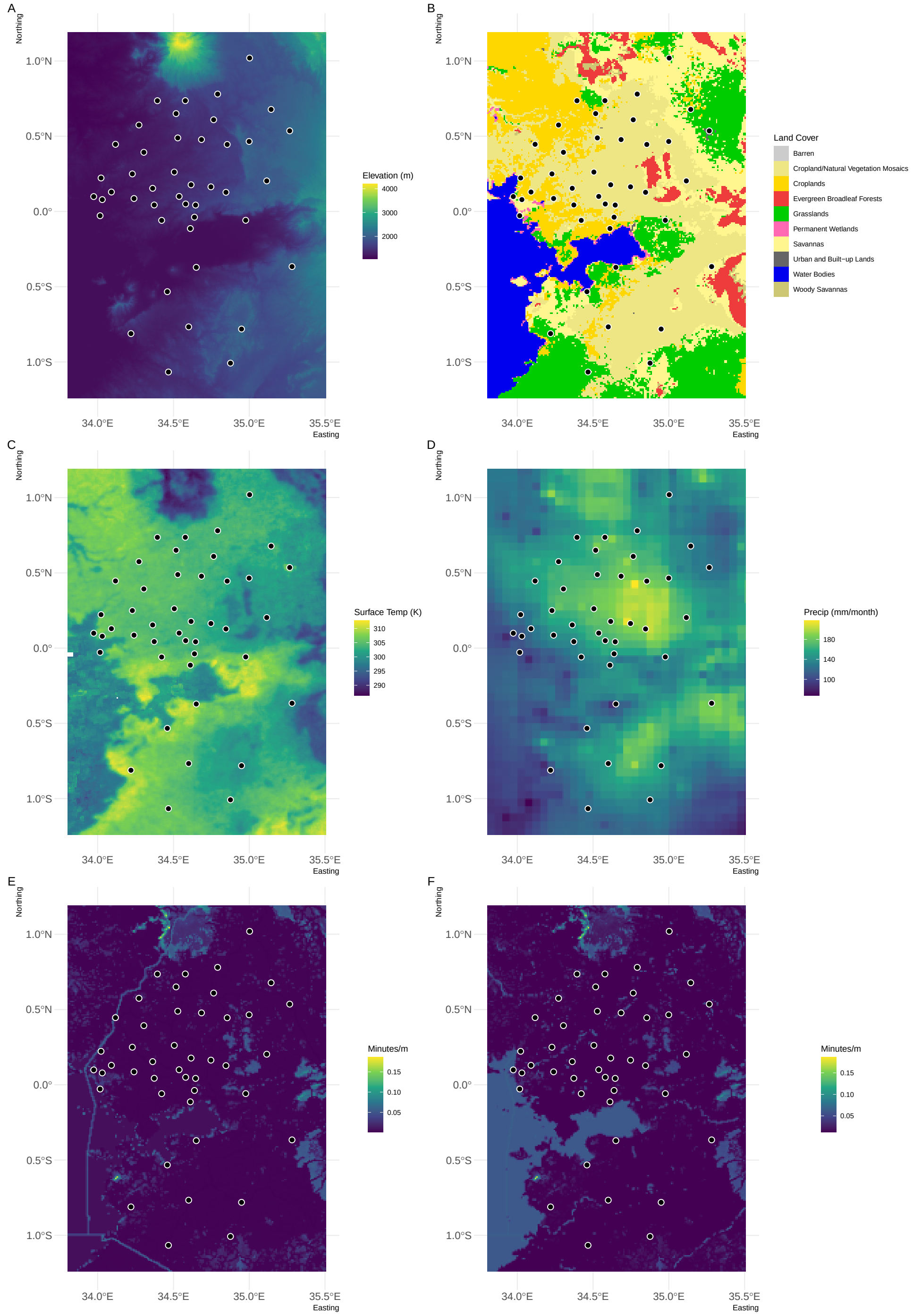


Figure 2.1: Spatial covariates used in our study: elevation, from the NASADEM product **(A)**; land cover, from the MCD12Q1 product **(B)**; land surface temperature (LST), from the MYD11A2 product **(C)**; precipitation, from the CHIRPS dataset **(D)**; and friction to human movement, modeled by Weiss et al. (2020) under assumptions of both access to motorized ground transport **(E)** and no access to such transport **(F)**. Study site locations are shown for context. These rasters were visualized with the `landscapetools` R package.

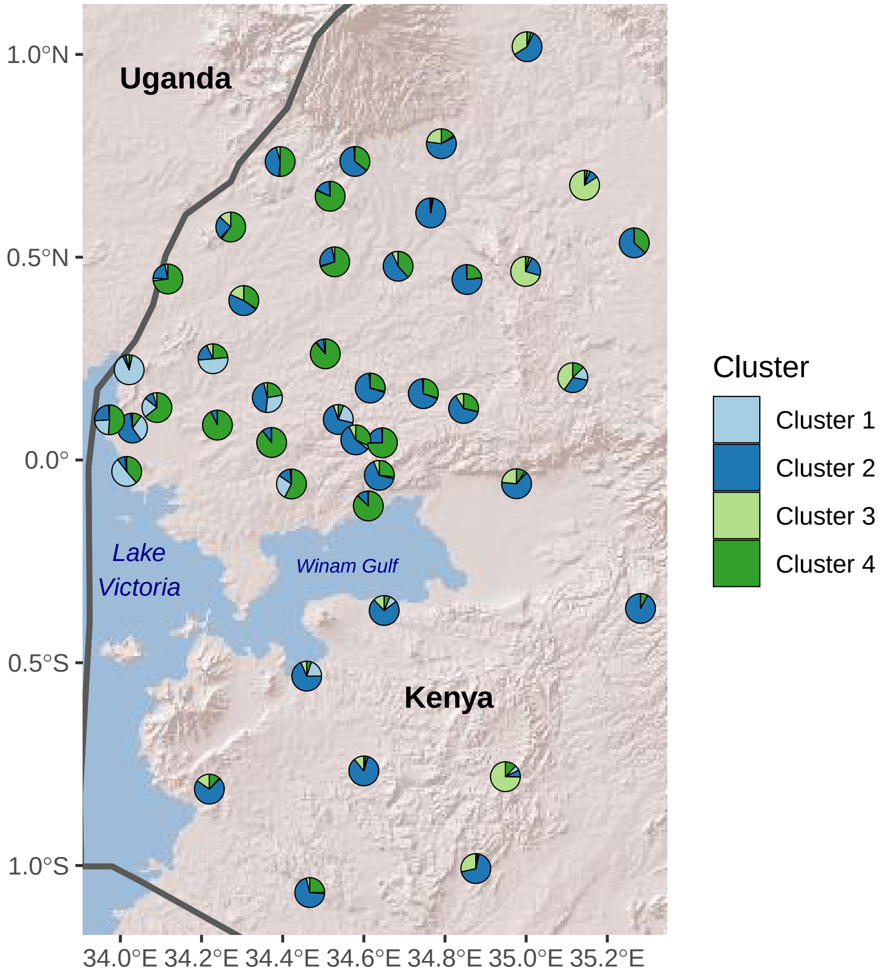


Figure 2.2: Map with pie charts showing average admixture coefficients (i.e., proportion membership in each cluster) for the samples from each study location, as estimated by `rmaverick`. The `scatterpie` R package (Yu, 2021) was used to create the pie charts. Country boundaries are included for context, obtained with the `rnaturalearth` R package (South, 2017). The background is Esri's World Shaded Relief layer (©2009 ESRI).

Table 2.1: Microsatellite markers used in this study.

Marker	Citation
TA1	(Anderson et al., 1999)
TAA87	(Anderson et al., 1999)
PfPK2	(Anderson et al., 1999)
TAA109	(Anderson et al., 1999)
TAA42	(Anderson et al., 1999)
TAA81	(Anderson et al., 1999)
poly2	(Su et al., 1999)
9735	(Su et al., 1999)

The automated method was considered more accurate and was used for alleles scored with both methods.

Samples were filtered by both the number of successfully-scored loci per sample and the number of samples per study site. Only samples with at least six successfully-scored loci were included in further analyses. This ensured that every pair of samples would have at least four overlapping loci. The *P. falciparum* samples were grouped into populations according to the clinic where they were collected (i.e., the geographic location). Of these populations, only those with at least five samples were used in further analyses. 81.3% of samples passed both filters. The study site locations are provided in Appendix 2A.

2.2.4 Population structure

Linkage disequilibrium (LD) was estimated by computing the $\bar{r}_d$ statistic, which approximates the popular index of association but does not increase with the number of loci (Agapow & Burt, 2001). This was computed with the `poppr` R package (Kamvar et al., 2014), both with and without clonal correction. Pairwise LD was also estimated for each unique pair of loci. Missing data values were ignored for these computations.

Genetic clustering of samples was first assessed with principal component analysis (PCA), a method which transforms the input matrix into a set of orthogonal components ranked in descending order of the variance they explain. Visualizing combinations of the highly-ranked components allows one to identify the number of clusters present in the data. In this case, the microsatellite genotypes were converted into binary format, meaning one column per locus/allele combination, one row per sample, and a value of 0 or 1 based on whether that allele was present. This format was chosen because it allows flexible representation of samples with multiple clones. Missing data values were replaced with the mean frequency of the allele in question, as this allows all samples to be used in PCA without biasing the results. PCA was then performed on this binary matrix using the R programming language (R Core Team, 2021) and visualized with the aid of the `GGally` R package (Schloerke et al., 2021).

PCA shows the overall pattern of clustering in a genetic dataset, but PCA cannot be used to estimate the degree of membership a given individual has in a given population. In addition, genetic data will likely not conform to the linearity assumption of PCA. For this reason, PCA was only used to identify the probable number of clusters present in the data. Another program, `rmaverick`, was used to assign individuals to clusters and estimate admixture coefficients for each individual (Verity, 2018). `rmaverick` is a Bayesian method that, similar to the popular program `STRUCTURE` (Pritchard et al., 2000), seeks to find the population groups that are not in Hardy-Weinberg or linkage equilibrium. The admixture coefficients for each individual represent the proportion of membership that individual has in each cluster. Unlike alternatives like `STRUCTURE`, `rmaverick` uses a Metropolis-coupled Markov chain Monte Carlo technique, in which parameter values are simultaneously estimated in different "chains" and information is periodically passed between chains (Verity & Nichols, 2016). This information passing improves mixing and can make it more likely that the model will converge.

Based on evaluation of different parameter configurations, `rmaverick` was run with 10,000 burn-in iterations, 2,000 sampling iterations, 500 rungs, a GTI power of 3, and the admixture model. The burn-in iterations are an initial period in which the model is run without saving results to avoid bias from the initial conditions, whereas the sampling iterations are the portion of the model run in which results are saved. The rungs are the number of Metropolis-coupled chains used and the GTI power controls the distribution of these chains. Mono- and biclonal infections were incorporated by running `rmaverick` with mixed ploidy and repeating the allele when only one was present for a given locus. Samples with more than two clones were discarded for this analysis. The `pophelper` R package (Francis, 2016) was used to assist in visualization of results.

2.2.5 Spatial patterns of relatedness

Genetic relatedness between samples was estimated as the proportion of shared alleles, treating each polyclonal infection as a single "subpopulation," given that no existing analytical method can separate individual parasite haplotypes for infections that have multiple alleles at more than one locus. Each infection was treated as a subpopulation when calculating genetic relatedness, similar to the approach of Wesolowski et al. (2018). Thus, the value we obtained was the proportion of shared alleles between samples or infections, but not necessarily between individual parasite clones. The specific algorithm used to compute the proportion of shared alleles was that employed in the R package `PopGenReport` (Adamack & Gruber, 2014), re-implemented to work with our data. Missing reads were treated as the absence of any alleles for the locus in question and did not affect the calculation.

This individual-based measure was then aggregated to the population level by taking the fraction of individual relationships that passed a relatedness threshold of 0.15 and converted to genetic distance by subtracting from 1. In other words, the relatedness for all of the pairs of individuals corresponding to each pair of populations was compared to this threshold and the population-based measure of relatedness was taken to be the fraction of individual pairs that exceed the threshold. This fraction of "highly-related" individuals should be more sensitive to recent demographic events than other metrics such as average relatedness (Taylor et al., 2017). After sensitivity testing using thresholds between 0.1 and 0.3, we selected a threshold of 0.15 because this did not lead to saturation at either the lower or upper bounds (i.e., clumping of pairs near a genetic distance of 0 or 1; Figure 2.3). This threshold is considerably lower than that used by Taylor et al. (2017), but they used SNPs, rather than microsatellite data. The relatedness expected by chance alone is much higher in the former case. The distribution of the final relatedness values is shown in Figure 2.4.

To identify clustering of genetic information in geographic space, an ordination technique named `MEMGENE` was used (Galpern et al., 2014). `MEMGENE` extends PCA to isolate the spatial portion of the

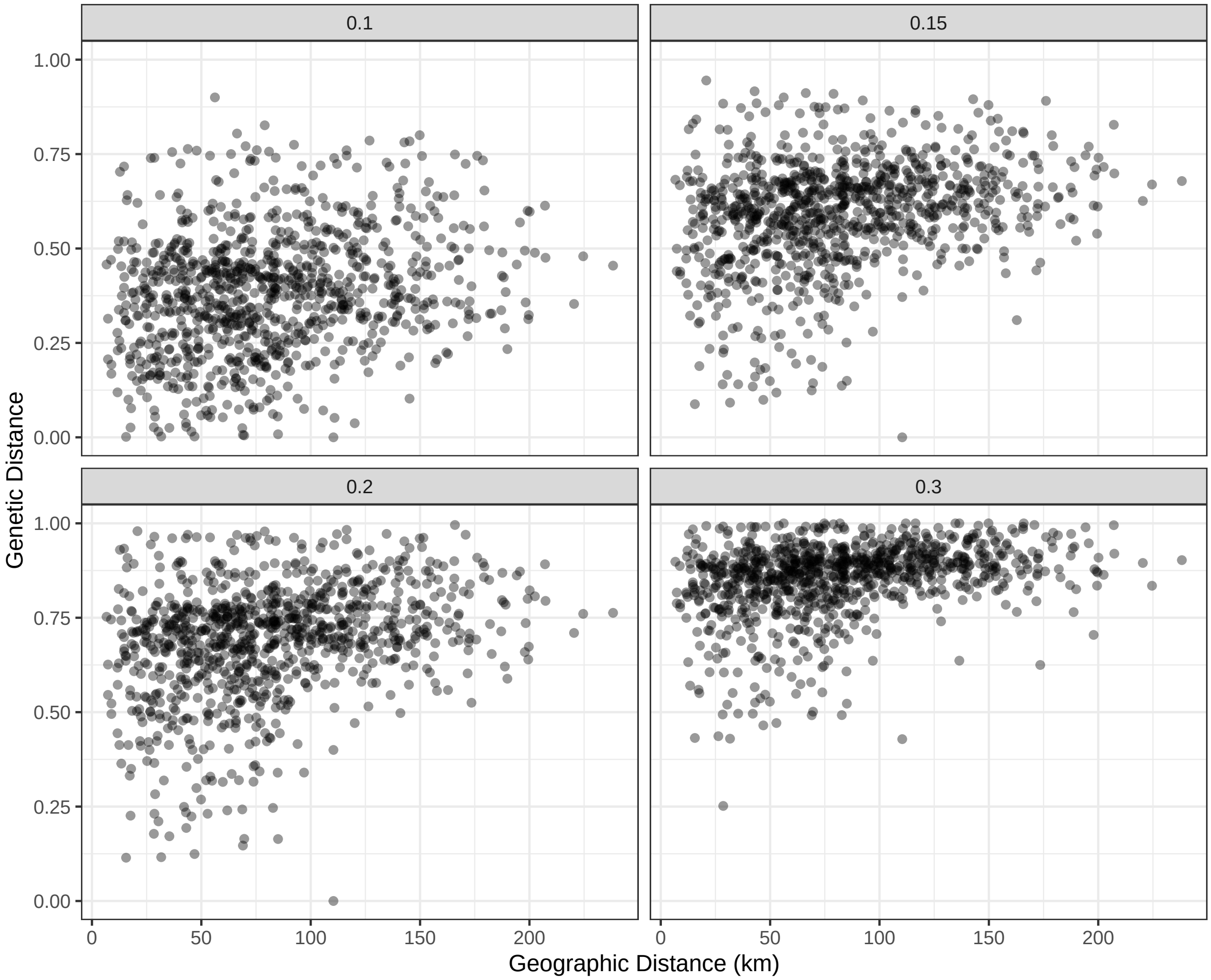


Figure 2.3: Scatterplots of geographic distance versus genetic distance, faceted by the threshold used to define highly-related sample pairs. Points are translucent to aid interpretation. In order to preserve the most information, it is desirable to avoid clustering at either the lower or upper limit of genetic distance. The plot for 0.3 exhibits this type of pattern most clearly, but it is also present in the plots for thresholds of 0.1 and 0.2. Therefore, 0.15 was selected as the best threshold. This plot shows the version of the data corresponding to a minimum of 5 samples per study site, but the version created with a minimum of 15 samples per study site shows a similar pattern.

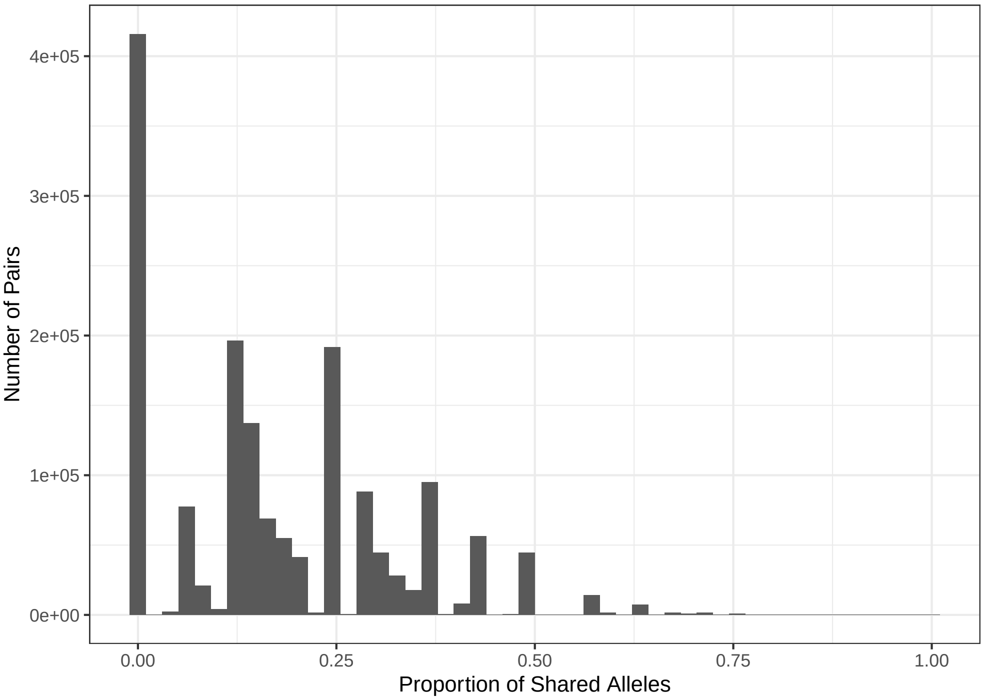


Figure 2.4: Histogram of the proportion of alleles shared between each pair of individuals.

variance in a matrix of genetic distances. `MEMGENE` does this by performing PCoA on the matrix of geographic distances to find components that represent the geographic patterns among study sites, regressing these components against the genetic distances, and then performing a second PCoA on the regression predictions. The components of this second PCoA are the `MEMGENE` variables, and each one can be thought of similarly to the components from standard PCA, except they only represent the portion of genetic variance that can be explained by geographic patterns. Visualizing these `MEMGENE` variables at each study location can show spatial clusters of related populations and point to possible barriers to transmission. This analysis was performed with the matrix of population-level genetic distances described above.

To evaluate the hypothesis of isolation-by-distance, the study site coordinates were reprojected into planar space and geographic and genetic distances were compared using both a Mantel correlogram and the test for congruence among distance matrices (CADM). Coordinate reprojection converts the elliptical coordinate space that describes the Earth's curved surface into a two-dimensional space that is amenable to analysis. The map projection selected for this study was the World Geodetic System 84-based coordinate system for UTM zone 36 N. Once reprojected, Euclidean distance was calculated between all pairs of study sites. Mantel correlograms (Oden & Sokal, 1986) were computed using the R package `vegan` (Oksanen et al., 2020) using these Euclidean distances and the genetic distances described above. The R package `ape` (Paradis & Schliep, 2019) was used to test for CADM.

2.2.6 Landscape genetics

Environmental variables that either have been previously associated with *P. falciparum* transmission or that influence host and/or vector movement were selected for inclusion in resistance surface analysis (Table 2.2). To evaluate a potential barrier effect from Lake Victoria, a binary layer was created that represented grid cells belonging to Lake Victoria with a 1 and other grid cells with a 0. The rainfall and land surface temperature (LST) data, which are available at sub-annual frequency, were aggregated for the entire year using a mean composite. Both 2012 and 2013 were used for these aggregations, but collection of the LST data did not begin until 2012-01-25, so the first part of January 2012 is not included. The land cover dataset was created at an annual time scale, and the 2012 version was selected for this study. The DEM and friction to human movement datasets are static, and therefore these considerations do not apply. All datasets were reprojected into the World Geodetic System 84-based UTM 36 N coordinate system with 1 km spatial resolution.

Using these environmental datasets and the matrix of population-based genetic distance, resistance surfaces were estimated to assess the degree to which each environmental variable explains the observed patterns of genetic relatedness. Resistance surfaces are gridded spatial layers in which the values of each cell represent the degree to which that space obstructs gene flow. By treating gene flow as a proxy for transmission, the surface as a whole represents areas that are more or less permissible to malaria transmission. We used the R package `ResistanceGA` (Peterman, 2018) to optimize resistance surfaces. Briefly, this process involves 1) finding the least cost path between every pair of locations through the current resistance surface; 2) fitting a mixed linear effects model that explains genetic distance in terms of this resistance distance; and 3) applying a transformation to the resistance surface to improve the fit. This process is iterated many times. In the first

Table 2.2: Spatial covariates included in resistance surface analysis.

Variable	Dataset	Resolution	Citation
Rainfall	CHIRPS	0.05 degree	(Funk et al., 2015)
LST	MYD11A2.061	1 km	(Wan et al., 2021)
Elevation	NASADEM	1 arc second	(NASA JPL, 2020)
Land cover	MCD12Q1	500 m	(Friedl & Sulla-Menashe, 2019)
Lake Victoria binary	Derived from MCD12Q1	500 m	NA
Human mobility	Friction to human movement (with access to motorized transport)	1 km	(Weiss et al., 2020)
Human mobility	Friction to human movement (without access to motorized transport)	1 km	(Weiss et al., 2020)

round, the resistance surfaces are simply the rescaled environmental inputs. The entire procedure is performed in the framework of a genetic algorithm that tests a certain number of mutations (transformations) per generation, chooses the most fit to carry to the next generation (based on the mixed linear effects model), and repeats until the change in fitness does not meet a certain threshold.

For each run, `ResistanceGA` merges the input layers into a single composite layer, which serves to incorporate multiple inputs without creating multicollinearity issues. `ResistanceGA` accomplishes this compositing by summing the surfaces after transformations have been applied, which ensures this operation is mathematically rational, even for categorical variables. All possible combinations of input layers, including each layer individually, were tested. Because multiple input layers are transformed into one variable prior to fitting the regression models, the multicollinearity issues in landscape genetics described by Prunier et al. (2015) are not a concern.

Each resistance surface was fit twice so that consistency between replicates could be assessed. After fitting the resistance surfaces, bootstrapping was performed to determine how robust the fit of each surface is to random subsets of the input samples.

2.2.7 Software pipeline

Unless otherwise noted, all analyses were performed using custom code written in the R (R Core Team, 2021) and Python (Python Software Foundation, 2022) programming languages. Throughout, the `adegenet` R package (Jombart, 2008) was used for reformatting genetic data, the Geospatial Data Abstraction Library (GDAL/OGR contributors, 2021) and the `raster` (Hijmans, 2022) and `sf` (Pebesma, 2018) R packages were used for spatial data processing, and the R `tidyverse` packages (Wickham et al., 2019) were used for general data manipulation and visualization. `knitr` (Xie, 2014) and R Markdown (Xie et al., 2018; Xie et al., 2020) were used to organize and document analyses. The entire workflow was automated with Snakemake (Mölder et al., 2021) and is available on Bitbucket at https://bitbucket.org/a-hubbard/hubbardetal_landgen_drivers_malaria/. This repository includes specifications of the exact versions of each package used.

2.3 Results

2.3.1 Population Structure

LD analysis shows weakly significant LD driven by a single loci pair (TA42 and 9735). This is consistent with weak population structure in a high transmission environment.

The PCA results suggested that the potential number of distinct genetic clusters is between three and six. The first four components of the PCA explain 5.9%, 3.8%, 3.6%, and 3.2% of the overall variance, respectively. Visualization of these components show two to three distinct clusters delineated by the first component and two weakly-separated clusters in the third component (Figure 2.5). The other components explained less than 3% of the overall variance, and thus were not analyzed in detail.

Based on these PCA results, `rmaverick` was run for K values from one to six. The admixture bar plots show considerable mixing overall but some structuring according to latitude (Figure 2.6a) and elevation (Figure 2.6b). To visualize these geographic patterns in more detail, pie charts depicting mean population admixture coefficients were visualized at each study location for a K of four (Figure 2.2). Although there is no "true" K, the model with four clusters was best supported

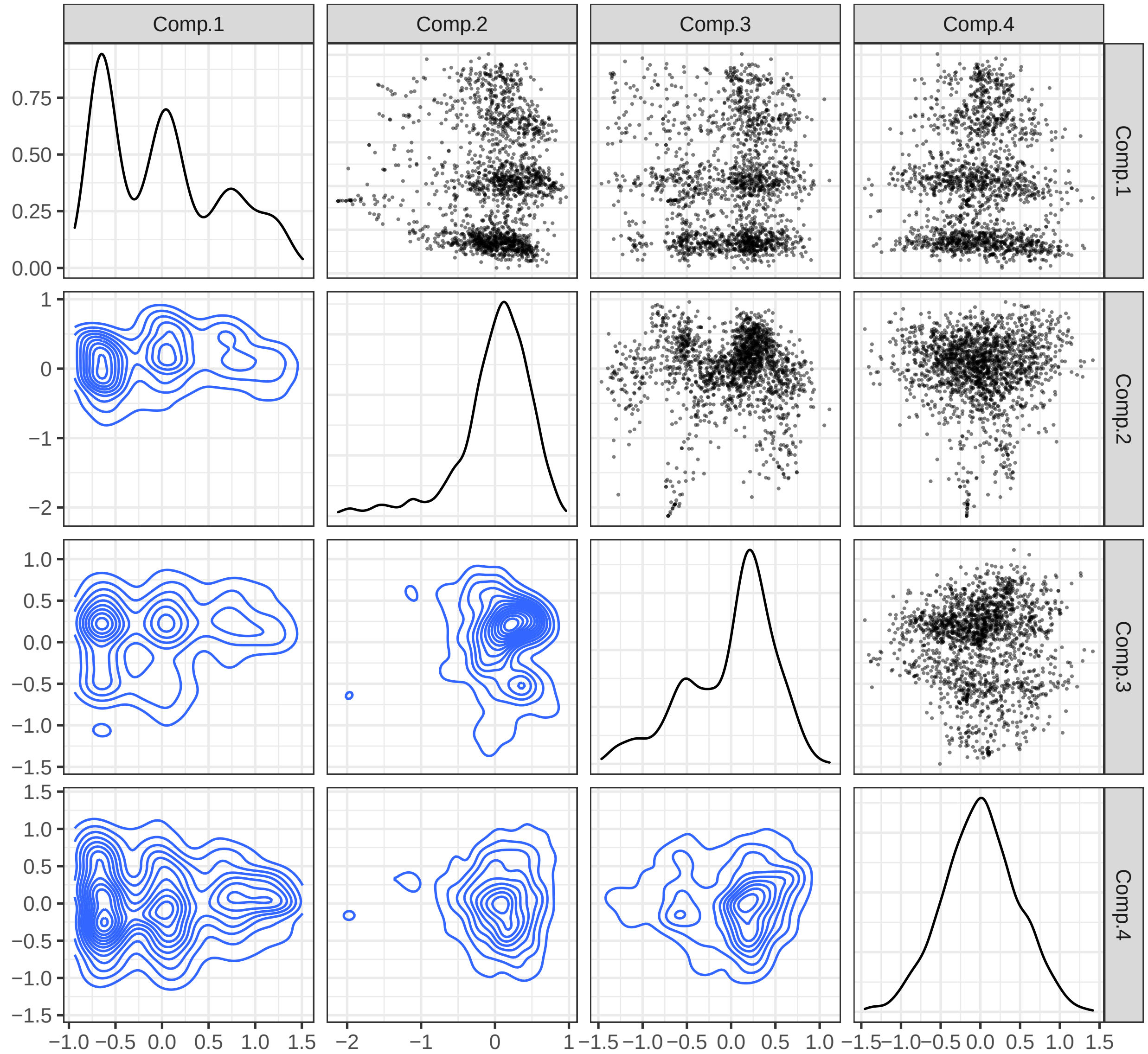


Figure 2.5: This figure depicts the first four components identified with PCA of the allele frequencies. The upper half of this grid contains simple scatter plots of the major components, while the lower half contains contour plots of the same information. Note that the axes for each plot are a unique combination of the two components in question, such that the scatter plot and contour plot for a given component pair visualize the same data but at a 90° angle to one another. Density plots of each component are along the diagonal. This figure was made with the help of the `GGally` R package (Schloerke et al., 2021).

by the posterior evidence (Figure 2.7) and so was a logical choice for further inspection. Inspection of Figure 2.2 indicates some differentiation on the basis of geography and elevation. Samples with substantial membership in cluster 1 primarily came from the western portion of the study region, near the border with Uganda. Cluster 4 has some overlap with this area, but encompasses a wider area covering all of the lowlands north of the Winam Gulf. Cluster 3 is primarily associated with samples from higher elevation sites in the eastern portion of the study area. Cluster 2 does not seem to be strongly tied to geographic factors.

2.3.2 Spatial Patterns of Relatedness

MEMGENE showed distinct spatial clusters north and south of Lake Victoria. This analysis revealed that 8.3% of the overall variance in genetic distances can be explained by spatial patterns. Of this fraction, the first component, or MEMGENE variable, explained 45.7% and showed a distinct spatial cluster of genetic similarity in the lowlands north of the Winam Gulf of Lake Victoria (Figure 2.8). The areas south and east of the Winam Gulf comprise a second cluster. Samples gathered near the Ugandan border, in the northwest of the study area, fall somewhere in between, but bear more similarity to samples from the south and east. The other components explained a considerably lower fraction of the spatial portion of the variance, and thus, a very low fraction of the overall variance, and so were not visualized. After the regression step in MEMGENE (see Materials and Methods), a redundancy analysis is performed to identify components that significantly improve fit. The results described above were found with a significance threshold of 0.05 in this step. When this threshold was lowered to 0.01, no significant components were found, suggesting the pattern described above is only weakly significant.

The Mantel correlogram clearly showed significant correlations between genetic and geographic distance over short distances (less than 70 km; Figure 2.9a, Table 2.3). This pattern of isolation-by-distance was corroborated by the test for CADM, which showed highly significant congruence between genetic and geographic distance matrices ($W = 0.642$; $p = 0.00045$). However, the correlogram shows this relationship disappears as geographic distance increases, until eventually significant negative correlation was found at higher geographic distances. This surprising result can

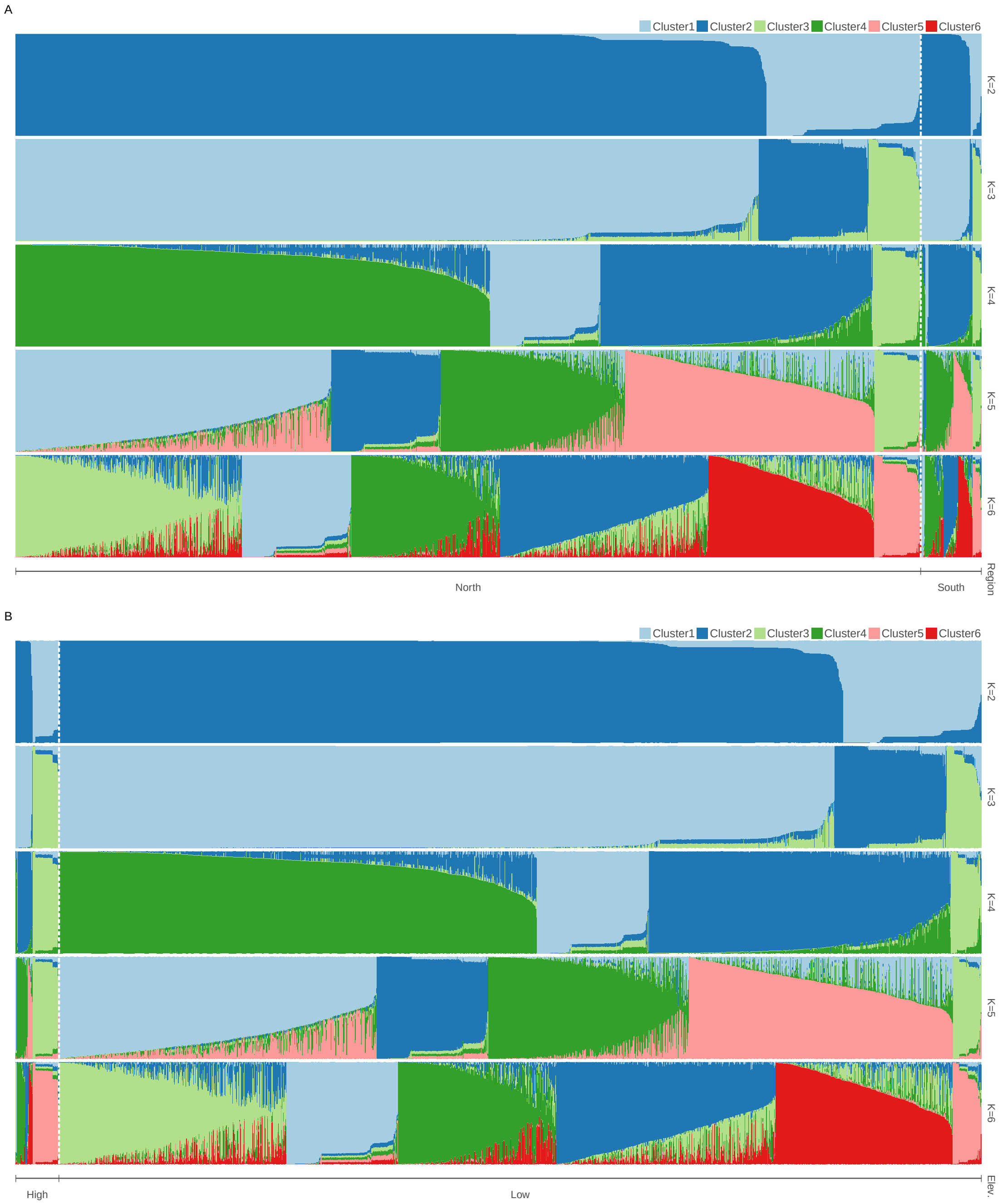


Figure 2.6: Bar plots showing the admixture coefficients estimated by `rmaverick` for each sample. Bars are sorted according to cluster values and two different groupings: geographic portion of the study area **(A)** and elevation **(B)**. "North" and "South" regions were defined relative to the Winam Gulf. "High" elevation was defined as greater than 1750 meters above sea level, whereas samples from sites below this cutoff were labeled as "Low." The 1750 m threshold was selected based on inspection of the distribution of elevations present in the data, with the intention of finding a natural break point between high and low elevation sites. These plots were created with the `pophelper` R package (Francis, 2016).

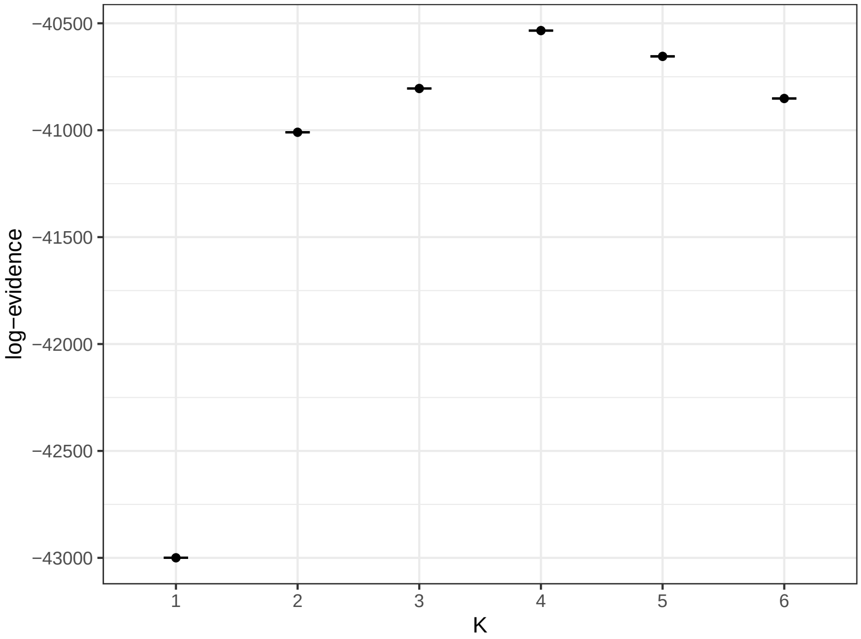


Figure 2.7: Plot of log evidence of the model for each value of K, as estimated by `rmaverick`. Intervals of 95% credibility are indicated by error bars for each K value.

be understood by studying the `MEMGENE` map discussed above. While the majority of the sites in the second cluster, corresponding to negative `MEMGENE` values, are south and east of the Winam Gulf, several of the sites near the Ugandan border north of the Gulf also belong to this cluster (Figure 2.8). Many of these sites are between 90 and 130 km from the other locations belonging to this cluster, south and east of the Gulf, which corresponds to the distances where negative correlations are observed in the Mantel correlogram. This suggests that a process that is not well-represented by geographic distance alone is driving genetic similarity between these two locations.

Table 2.3: Mantel correlations, significance values, and number of samples for each distance class of the Mantel correlogram.

Distance Class (km)	No. Samps.	Mantel corr.	p
17.2072	166	0.12135	0.009
38.2569	318	0.14397	0.002
59.3066	368	0.08503	0.018
80.3563	332	0.04333	0.133
101.4061	244	-0.11172	0.005
122.4558	182	-0.11709	0.012

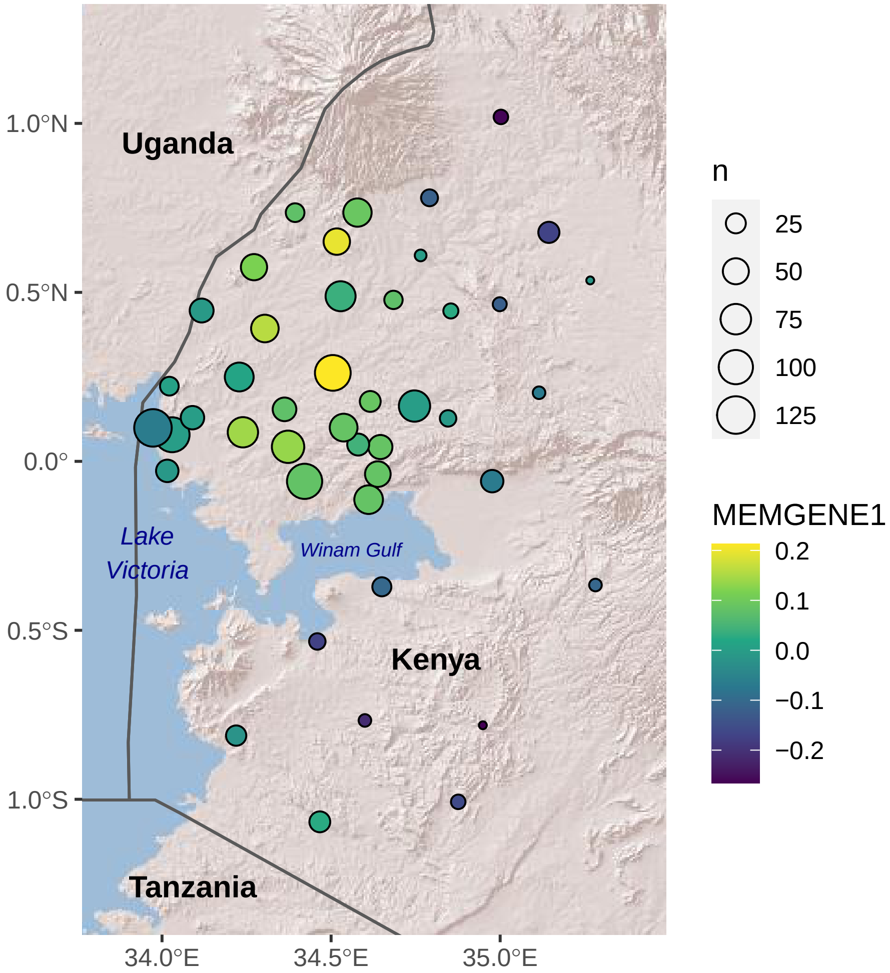


Figure 2.8: Map showing sample site locations with the point color scaled according to the first `MEMGENE` variable and the size representing the number of samples gathered at that location. Country boundaries are included for context, obtained with the `rnaturalearth` R package (South, 2017). The background is Esri's World Shaded Relief layer (©2009 ESRI).

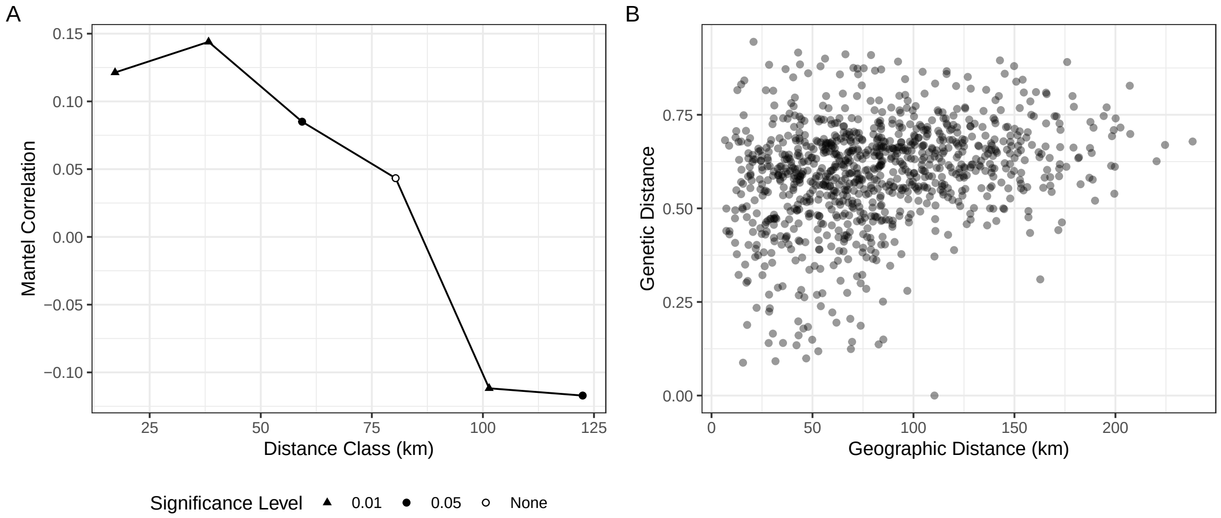


Figure 2.9: Mantel correlogram showing the correlation between genetic and geographic distance **(A)**. In this type of plot, the *x*-axis is a series of geographic distance classes and the *y*-axis is the correlation between genetic and geographic distance in samples separated by this distance class. Point shape indicates the significance level of each correlation. For reference, the distance-distance plot of geographic versus genetic distance is also included **(B)**. Note that only distance classes with adequate numbers of samples for analysis were included in the Mantel correlogram, which is why the *x* axes do not have the same extent.

2.3.3 Landscape Genetics

The resistance surfaces clearly show that Lake Victoria is acting as a barrier to gene flow, based on both the ranking of best-fitting surfaces and high resistance values over Lake Victoria. To rank the surfaces, the corrected Akaike information criterion (AICc) was used with all output surfaces generated by both replicates. This is displayed for the 10 best fitting surfaces in Table 2.4, along with the number of parameters and the conditional and marginal R^2. Generally speaking, the two replicates conducted for each set of inputs did not produce identical outputs. However, the differences in likelihood and AICc were always small (Appendix 2B), suggesting similar solutions and goodness-of-fit had been obtained between replicates. For the sake of comparing variables, the surface from the best-fitting replicate was selected for each set of inputs for display in Table 2.4 and visual inspection. Most of the best fitting surfaces contained the binary Lake Victoria layer. LST and friction to human movement without access to motorized transport were also consistently

Table 2.4: Fit statistics for top 10 best fitting resistance surfaces.

Surface	K	$AICc$	R^2_m	R^2_c
Lake Victoria and Land Surface Temperature	6	-1828.47	0.136461	0.495123
Lake Victoria	3	-1827.51	0.0674234	0.478118
Friction to Human Movement (w/o motorized)	4	-1826.23	0.0861454	0.48676
Elevation and Lake Victoria	6	-1823.91	0.130084	0.495179
Lake Victoria and Friction to Human Movement (w/o motorized)	6	-1822.71	0.0901474	0.491937
Land Surface Temperature	4	-1821.96	0.163029	0.515629
Land Surface Temperature and Friction to Human Movement (w/o motorized)	7	-1821.3	0.165838	0.518301
Lake Victoria and Precipitation	6	-1821.29	0.0501707	0.489653
Elevation and Friction to Human Movement (w/o motorized)	7	-1820.45	0.201408	0.53808
Lake Victoria and Friction to Human Movement (w/ motorized)	6	-1819.93	0.0672178	0.478748

present in the top ranking surfaces. The distance-only and null models did not rank particularly highly, indicating the best landscape resistance models explain patterns of gene flow that geographic distance alone cannot. All of these conclusions were corroborated by the bootstrapping results (Appendix 2C).

In the resistance surfaces themselves (Figure 2.10), pixels associated with Lake Victoria were assigned high resistance values in all of the best fitting surfaces, regardless of whether they included the binary Lake Victoria layer. However, in the highest ranked layer, Lake Victoria and LST, LST was weighted to contribute more to the final model (77%), indicating that variable explains a substantial amount of variance that Lake Victoria alone cannot. Low land surface temperature was associated with high resistance to gene flow, as was high friction to human movement without access to motorized transport. Of the other environmental covariates, high resistance to gene flow was associated with high elevation, high friction to human movement with access to motorized transport, low precipitation, and water bodies in the land cover layer (results not shown).

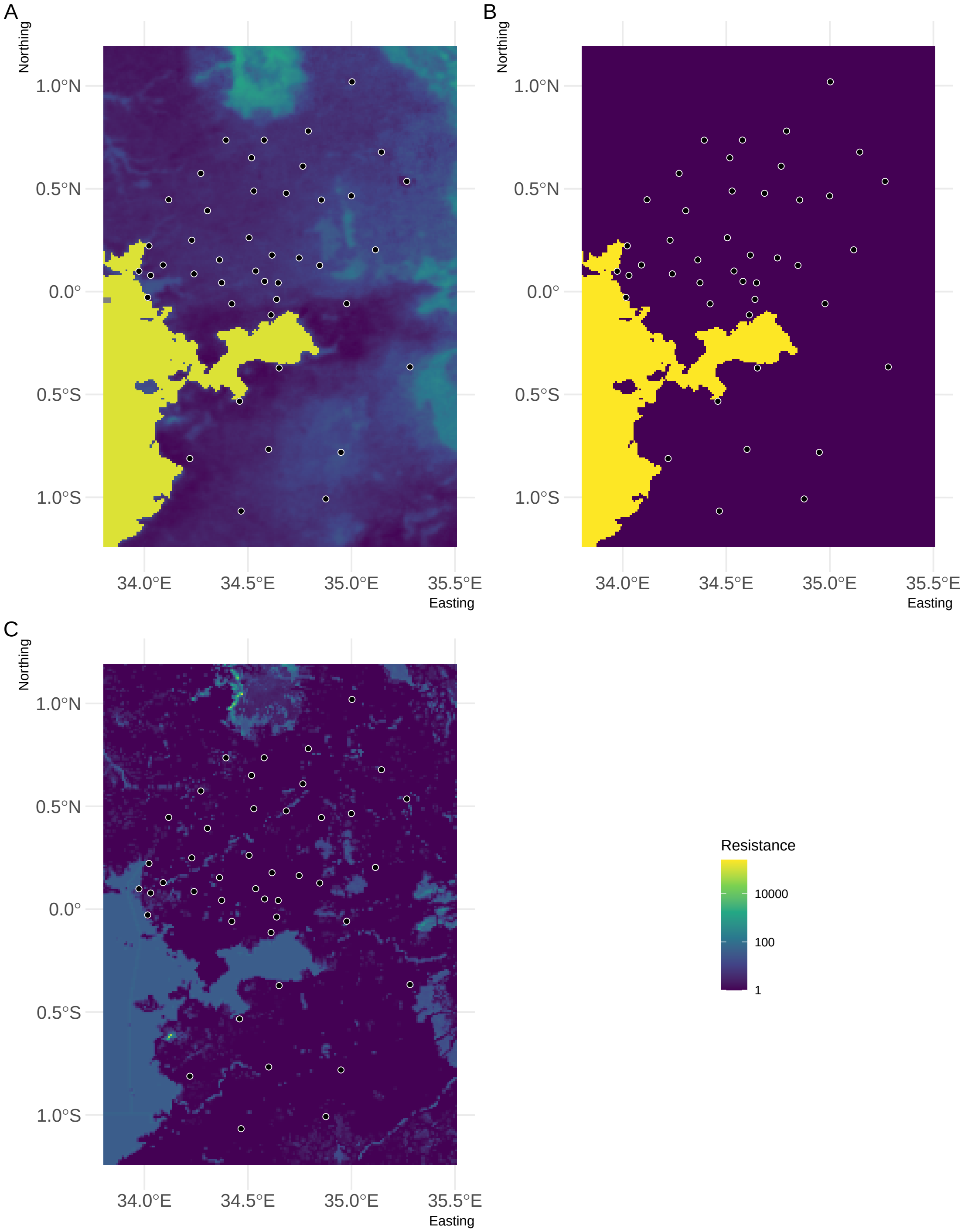


Figure 2.10: The three best-fitting resistance surfaces, ranked in order of fit: a composite surface of LST and the Lake Victoria binary layer **(A)**, a single surface modeled off the Lake Victoria binary layer **(B)**, and a single surface modeled off the friction to human movement dataset that does not assume access to motorized transport **(C)**. Resistance values are $\log_{10}$ transformed, and study site locations are shown for context. The surfaces were visualized with the `landscapetools` (Sciaini et al., 2018) and `patchwork` (Pedersen, 2020) R packages.

2.4 Discussion and Conclusion

The results presented in this study support an isolation-by-barrier hypothesis, where Lake Victoria acts as an obstacle to gene flow between the northern and southern parts of our study area. The `rmaverick` analysis gave the first indication of this conclusion, in that samples collected from north and south of the Winam Gulf of Lake Victoria tended to have membership in different genetic clusters. The pattern of spatial clustering in the first `MEMGENE` variable showed the same result, with one geographic cluster of genetic similarity in the lowlands north of Lake Victoria and the other encompassing the areas east and south of the lake. Finally, the resistance surface analysis suggested both that Lake Victoria was an important variable in dictating landscape resistance, as seen through the ranking of best-fitting surfaces, and that the lake is associated with a high resistance to gene flow, as evidenced by the surfaces themselves.

Other studies using data from a similar time period and region have by-and-large shown high gene flow (Nderu et al., 2019) leading to little genetic differentiation among parasite populations (Ingasia et al., 2016; Nderu et al., 2019; Nelson et al., 2019), although in one case this varied somewhat based on the genetic distance measure used (Nelson et al., 2019). A study conducted with more recent data (dating to 2018 and 2019) supported the same conclusion of little differentiation between populations (Onyango et al., 2021). Our results are qualitatively consistent with these findings, but cannot be compared quantitatively as these studies used different measures of genetic distance.

Our results are also consistent with previous investigations into the clustering of genetic relatedness in this area. Ingasia et al. (2016) showed with PCA that samples from Kisii, located in the highlands south of the Winam Gulf, clustered separately from samples collected from the lowlands north of the Gulf (Kisumu and Kombewa) and the highlands east of the gulf (Kericho). Omedo, Mogeni, Rockett, et al. (2017), using spatial scan statistics, identified a cluster of genetic similarity in part of the area north of the Winam Gulf, near the border with Uganda. Our findings indicate a distinct population north of the Winam Gulf, as well as some evidence of differentiation between highlands and lowlands. However, we also found a handful of sites near the Ugandan border that did not cluster with the rest of the sites north of the Gulf, but were more similar to samples collected south and east of Lake Victoria. These sites may correspond to the cluster found by Omedo, Mogeni, Rockett, et al. (2017) and are separated from the rest of Kenya by the Nzoia River, possibly explaining why they are distinct from the remainder of the lowlands north of the Gulf. The similarity with sites to the south and east of the Lake is less intuitive but may be explained by patterns of long distance human movement, which Wesolowski et al. (2012) found to be common in the Lake Victoria region. Another study in this area, Omedo, Mogeni, Bousema, et al. (2017), did not find any clustering from PCA, but they were focused on a small subset of our study region (Rachuonyo South).

Previous studies investigating isolation-by-distance in western Kenya have yielded inconsistent results. Qualitative assessments of isolation-by-distance have found none (Ingasia et al., 2016; Nelson et al., 2019), but more formal tests have shown some significant correlations between genetic and geographic distance at or below distances of 20 km between sites (Omedo, Mogeni, Bousema, et al., 2017; Omedo, Mogeni, Rockett, et al., 2017). This has some similarity to our findings, in that we discovered weakly significant correlations between geographic and genetic distance in distance classes at or below 60 km. Taken together, this suggests that some isolation-by-distance has occurred in *P. falciparum* populations in this region, but it is only noticeable at relatively short distances (0-60 km, depending on the study and methods) and between certain locations.

In terms of isolation-by-barrier and isolation-by-resistance, few studies have been performed on malaria, but those that do exist for our study region did not identify Lake Victoria as a barrier to gene flow. Ingasia et al. (2016) informally described isolation between highland and lowland populations, while Omedo, Mogeni, Rockett, et al. (2017) looked for a barrier more formally and found nothing. The first result is not inconsistent with our own. Ingasia et al. (2016) had relatively few study sites, with only one each north and south of the Winam Gulf. They may not have had the spatial coverage to identify a barrier effect from Lake Victoria, and their finding of isolation between highlands and lowlands is supported to some extent by our own clustering analyses. The contrasting conclusions on the presence of a barrier in our study and Omedo, Mogeni, Rockett, et al. (2017) may also come down to methodological differences. That study used a regression framework in which each 10x10 km pixel in the study region was treated as a separate variable and barrier effects were assessed for pixels separating site pairs where the straight line connecting the two sites passed through the pixel in question. In contrast, `ResistanceGA` fits values of high or low resistance to environmental covariates in their entirety (Peterman, 2018), rather than fitting different values in different parts of the study region. This makes our approach more suited to assessing the effect of environmental features holistically, throughout the study region, whereas the Omedo, Mogeni, Rockett, et al. (2017) method would be better suited to identifying barriers associated with small, specific geographic features similar in size to the 10x10 km pixels used in their model. In combination, then, these findings suggest that mixing is fairly homogeneous in the land areas of this study region, but that Lake Victoria, when considered as a single unit, does act as a barrier to gene flow between the northern and southern sides of the Winam Gulf.

Previous work on human movement in this area suggests relatively frequent travel within the Lake Victoria region (Blanford et al., 2015; Wesolowski et al., 2012). These studies did not clearly show Lake Victoria to be a barrier to movement, but they were intended for regional analyses and lacked the resolution to address this question in detail. One study, based on mobile phone data, does seem to indicate less connectivity with populations near the Ugandan border than in the rest of the Lake Victoria region (Wesolowski et al., 2012). This is consistent with the results of our clustering analyses and may explain the negative correlations revealed with the Mantel correlograms, but again, the resolution of that human movement study was such that we cannot be certain of this explanation.

In terms of vector populations, most research has considered Kenya as a whole and focused on the differentiation between western and coastal Kenya (e.g., Ogola et al., 2019). Of the groups that studied western Kenya in particular, the results have been inconsistent. In one case, significant population structuring was found in *Anopheles gambiae s. l.* in the Lake Victoria region and, while other landscape factors were found to be more important, a landscape genetics analysis did show

Lake Victoria to be an area of relatively low gene flow (Hemming-Schroeder et al., 2020). On the other hand, a more recent study done with the same species identified little genetic differentiation in this region and no apparent barrier in Lake Victoria, although no formal landscape genetics analysis was done (Onyango et al., 2022). However, only four study sites were used in this case, only one of which was south of the Winam Gulf, so it is likely this work lacked the spatial coverage to address these questions in detail. From this, we believe it likely that a large part of the barrier effect we discovered from Lake Victoria in *P. falciparum* genetics can be explained by the difficulty mosquitoes have traversing this large body of water.

Taken as a whole, these results indicate low overall genetic differentiation in the Lake Victoria region, but with some separation of populations north and south of the lake that is explained by the presence of the lake as a geographic barrier to gene flow. The resistance surface results suggest that both host and vector factors are important determinants of transmission, as friction to human movement and temperature, which will disproportionately affect mosquitoes, were both in the highest ranking surfaces.

This work is the most spatially comprehensive landscape genetics study done in malaria to date, and we have identified landscape impacts on gene flow, specifically a barrier effect from Lake Victoria, that have not been documented previously. However, this study does have certain limitations. First, while polygenomic microsatellites are relatively informative genetic markers, our study only used eight. Subsequent studies conducted with more genomic depth would be useful to confirm our findings. On a related subject, while our study has large sample sizes overall, some of the study locations only have a handful of samples, in particular south and east of Lake Victoria (Figure 2.8), which may bias our spatial analyses in those areas. In addition, the two separate genotyping methods used in our dataset do represent a source of inconsistency. We have characterized the level of agreement between the two methods, but we did not attempt to repeat genotyping due to the age of the samples. Also, many of the environmental inputs used are proxies for the true variable of interest (e.g., LST as a proxy for near-surface air temperature). The data products selected are all well-correlated with those variables, but as more direct measures become available it will be important to repeat this and other landscape genetics analyses to confirm their findings. On the subject of spatial covariates, no information on spatial coverage of malaria control measures was included, despite the importance of these factors in driving population structure. There is no quality data on sub-national spatial heterogeneity of coverage for most interventions, and for the one exception, insecticide-treated nets, little spatial heterogeneity was observed in our study area (Bertozzi-Villa et al., 2021). For this reason, these data were not included in our analyses. Finally, our conclusions rest on the assumption that the environmental data we used is reflective of the state of the landscape that is relevant to its impact on gene flow. In other words, we have implicitly assumed that the environment in 2012 and 2013, contemporaneous with sample collection, has the greatest impact on gene flow. In reality, the state of the environment prior to sample collection likely has had some impact. The relationship between contemporaneous and historical environmental variation and gene flow is a topic that deserves further research to characterize the lag time corresponding to the strongest correlation.

The results of this study have implications in a few different areas. For the sake of planning interventions, the populations around Lake Victoria are sufficiently connected that blanket control measures remain appropriate. However, it is probable that if interventions further reduce transmission in this area, these populations will become more distinct and interventions conducted in one area will have less impact on the other. In this situation, it would be recommended to consider populations north and south of the Winam Gulf as separate entities when targeting interventions.

For the modeling community, our results indicate that geographic distance is a poor proxy for transmission and that both vector and host factors can be important drivers of transmission at a moderate spatial scale. The first is an important finding because geographic distance is frequently used as a proxy for connectivity in models (Lee et al., 2021). More work is required to identify the best alternatives, but measures that represent heterogeneous patterns of transmission are necessary. In terms of the drivers of transmission, our study was performed at a reasonably coarse spatial scale, making it a surprise that an environmental variable that primarily impacts vector activity, LST, proved to be one of the most important explanatory variables. Further research is recommended to better understand how spatial scale impacts the drivers of transmission in vector-borne diseases, especially with respect to which scales are primarily governed by vector or host factors.

Finally, and most broadly, this study has demonstrated that landscape genetics analysis of vector-borne disease, when conducted with a large number of spatial locations, is capable of revealing and explaining barriers to gene flow in fairly high transmission settings that lack strong population structure. In future molecular epidemiology studies, we recommend that sensitive methods, such as MEMGENE (Galpern et al., 2014), first be used to characterize spatial heterogeneity in genetic variation. If significant variation is discovered, we recommend the use of landscape genetics methods, in particular resistance surface analysis, to explain the drivers of this structure. Doing so will extend and contextualize the results of traditional population genetics analyses, and thus yield more insights into the spatial determinants of transmission.

2.5 References

Adamack, A. T., & Gruber, B. (2014). PopGenReport: Simplifying basic population genetic analyses in R. *Methods in Ecology and Evolution*, *5*(4), 384–387. https://doi.org/10.1111/2041-210X.12158

Agapow, P.-M., & Burt, A. (2001). Indices of multilocus linkage disequilibrium. *Molecular Ecology Notes*, *1*(1-2), 101–102. https://doi.org/10.1046/j.1471-8278.2000.00014.x

Anderson, T. J. C., Su, X.-Z., Bockarie, M., Lagog, M., & Day, K. P. (1999). Twelve microsatellite markers for characterization of Plasmodium falciparum from finger-prick blood samples. *Parasitology*, *119*(2), 113–125. https://doi.org/10.1017/S0031182099004552

Bereczky, S., Mårtensson, A., Gil, J. P., & Färnert, A. (2005). Short report: Rapid DNA extraction from archive blood spots on filter paper for genotyping of Plasmodium falciparum. *The American Journal of Tropical Medicine and Hygiene*, *72*(3), 249–251. https://doi.org/10.4269/ajtmh.2005.72.249

Bertozzi-Villa, A., Bever, C. A., Koenker, H., Weiss, D. J., Vargas-Ruiz, C., Nandi, A. K., Gibson, H. S., Harris, J., Battle, K. E., Rumisha, S. F., Keddie, S., Amratia, P., Arambepola, R., Cameron, E., Chestnutt, E. G., Collins, E. L., Millar, J., Mishra, S., Rozier, J., … Bhatt, S. (2021). Maps and metrics of insecticide-treated net access, use, and nets-per-capita in Africa from 2000-2020. *Nature Communications*, *12*(1), 3589. https://doi.org/10.1038/s41467-021-23707-7

Blanford, J. I., Huang, Z., Savelyev, A., & MacEachren, A. M. (2015). Geo-Located Tweets. Enhancing Mobility Maps and Capturing Cross-Border Movement. *PloS One*, *10*(6), e0129202. https://doi.org/10.1371/journal.pone.0129202

Canelas, T., Castillo-Salgado, C., & Ribeiro, H. (2016). Systematized Literature Review on Spatial Analysis of Environmental Risk Factors of Malaria Transmission. *Advances in Infectious Diseases*, *6*(2), 52–62. https://doi.org/10.4236/aid.2016.62008

Castro, M. C. (2017). Malaria Transmission and Prospects for Malaria Eradication: The Role of the Environment. *Cold Spring Harbor Perspectives in Medicine*, *7*(10). https://doi.org/10.1101/cshperspect.a025601

Francis, R. M. (2016). Pophelper: An R package and web app to analyse and visualize population structure. *Molecular Ecology Resources*, *17*(1), 27–32. https://doi.org/10.1111/1755-0998.12509

Friedl, M. A., & Sulla-Menashe, D. (2019). MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006. *NASA EOSDIS Land Processes DAAC*. Retrieved March 16, 2020, from https://doi.org/10.5067/MODIS/MCD12Q1.006

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations–a new environmental record for monitoring extremes. *Scientific Data*, *2*, 150066. https://doi.org/10.1038/sdata.2015.66

Galpern, P., Peres-Neto, P. R., Polfus, J., & Manseau, M. (2014). MEMGENE: Spatial pattern detection in genetic distance data. *Methods in Ecology and Evolution*, *5*(10), 1116–1120. https://doi.org/10.1111/2041-210X.12240

GDAL/OGR contributors. (2021). *GDAL/OGR Geospatial Data Abstraction software Library*. https://gdal.org

Hemming-Schroeder, E., Lo, E., Salazar, C., Puente, S., & Yan, G. (2018). Landscape Genetics: A Toolbox for Studying Vector-Borne Diseases. *Frontiers in Ecology and Evolution*, *6*, 1–11. https://doi.org/10.3389/fevo.2018.00021

Hemming-Schroeder, E., Zhong, D., Machani, M., Nguyen, H., Thong, S., Kahindi, S., Mbogo, C., Atieli, H., Githeko, A., Lehmann, T., Kazura, J. W., & Yan, G. (2020). Ecological drivers of genetic connectivity for African malaria vectors Anopheles gambiae and An. arabiensis. *Scientific Reports*, *10*(1), 19946. https://doi.org/10.1038/s41598-020-76248-2

Hijmans, R. J. (2022). *Raster: Geographic Data Analysis and Modeling*. https://CRAN.R-project.org/package=raster

Ingasia, L. A., Cheruiyot, J., Okoth, S. A., Andagalu, B., & Kamau, E. (2016). Genetic variability and population structure of Plasmodium falciparum parasite populations from different malaria ecological regions of Kenya. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *39*, 372–380. https://doi.org/10.1016/j.meegid.2015.10.013

Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, *2*, e281. https://doi.org/10.7717/peerj.281

Kenya NMCP, KNBS, & ICF International. (2016). *Kenya Malaria Indicator Survey 2015*. NMCP, KNBS, and ICF International. Nairobi, Kenya. http://dhsprogram.com/pubs/pdf/MIS22/MIS22.pdf

Kepple, D., Hubbard, A., Ali, M. M., Abargero, B. R., Lopez, K., Pestana, K., Janies, D. A., Yan, G., Hamid, M. M., Yewhalaw, D., & Lo, E. (2021). Plasmodium vivax From Duffy-Negative and Duffy-Positive Individuals Share Similar Gene Pools in East Africa. *The Journal of Infectious Diseases*, *224*(8), 1422–1431. https://doi.org/10.1093/infdis/jiab063

Lee, S. A., Jarvis, C. I., Edmunds, W. J., Economou, T., & Lowe, R. (2021). Spatial connectivity in mosquito-borne disease models: A systematic review of methods and assumptions. *Journal of the Royal Society, Interface*, *18*(178), 20210096. https://doi.org/10.1098/rsif.2021.0096

Lo, E., Hemming-Schroeder, E., Yewhalaw, D., Nguyen, J., Kebede, E., Zemene, E., Getachew, S., Tushune, K., Zhong, D., Zhou, G., Petros, B., & Yan, G. (2017). Transmission dynamics of co-endemic Plasmodium vivax and P. falciparum in Ethiopia and prevalence of antimalarial resistant genotypes. *PLOS Neglected Tropical Diseases*, *11*(7), e0005806. https://doi.org/10.1371/journal.pntd.0005806

Lo, E., Lam, N., Hemming-Schroeder, E., Nguyen, J., Zhou, G., Lee, M. C., Yang, Z., Cui, L., & Yan, G. (2017). Frequent Spread of Plasmodium vivax Malaria Maintains High Genetic

Diversity at the Myanmar-China Border, Without Distance and Landscape Barriers. *The Journal of infectious diseases*, *216*(10), 1254–1263. https://doi.org/10.1093/infdis/jix106

Lo, E., Zhou, G., Oo, W., Afrane, Y., Githeko, A., & Yan, G. (2015). Low parasitemia in submicroscopic infections significantly impacts malaria diagnostic sensitivity in the highlands of Western Kenya. *PloS One*, *10*(3), e0121763. https://doi.org/10.1371/journal.pone.0121763

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. (10:33). https://doi.org/10.12688/f1000research.29032.2

NASA JPL. (2020). NASADEM Merged DEM Global 1 arc second V001. *NASA EOSDIS Land Processes DAAC*. https://doi.org/10.5067/MEaSUREs/NASADEM/NASADEM_HGT.001

Nderu, D., Kimani, F., Karanja, E., Thiong'o, K., Akinyi, M., Too, E., Chege, W., Nambati, E., Wangai, L. N., Meyer, C. G., & Velavan, T. P. (2019). Genetic diversity and population structure of Plasmodium falciparum in Kenyan-Ugandan border areas. *Tropical Medicine & International Health*, *24*(5), 647–656. https://doi.org/10.1111/tmi.13223

Nelson, C. S., Sumner, K. M., Freedman, E., Saelens, J. W., Obala, A. A., Mangeni, J. N., Taylor, S. M., & O'Meara, W. P. (2019). High-resolution micro-epidemiology of parasite spatial and temporal dynamics in a high malaria transmission setting in Kenya. *Nature Communications*, *10*(1), 5615. https://doi.org/10.1038/s41467-019-13578-4

Oden, N. L., & Sokal, R. R. (1986). Directional Autocorrelation: An Extension of Spatial Correlograms to Two Dimensions. *Systematic Zoology*, *35*(4), 608–617. https://doi.org/10.2307/2413120

Ogola, E. O., Odero, J. O., Mwangangi, J. M., Masiga, D. K., & Tchouassi, D. P. (2019). Population genetics of Anopheles funestus, the African malaria vector, Kenya. *Parasites & Vectors*, *12*(1), 15. https://doi.org/10.1186/s13071-018-3252-3

Okoyo, C., Mwandawiro, C., Kihara, J., Simiyu, E., Gitonga, C. W., Noor, A. M., Njenga, S. M., & Snow, R. W. (2015). Comparing insecticide-treated bed net use to Plasmodium falciparum infection among schoolchildren living near Lake Victoria, Kenya. *Malaria Journal*, *14*, 515. https://doi.org/10.1186/s12936-015-1031-6

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2020). *Vegan: Community Ecology Package*. Retrieved October 29, 2021, from https://CRAN.R-project.org/package=vegan

Omedo, I., Mogeni, P., Bousema, T., Rockett, K., Amambua-Ngwa, A., Oyier, I., Stevenson, J. C., Baidjoe, A. Y., de Villiers, E. P., Fegan, G., Ross, A., Hubbart, C., Jeffreys, A., Williams, T. N., Kwiatkowski, D., & Bejon, P. (2017). Micro-epidemiological structuring of Plasmodium falciparum parasite populations in regions with varying transmission intensities in Africa. *Wellcome Open Research*, *2*, 10. https://doi.org/10.12688/wellcomeopenres.10784.1

Omedo, I., Mogeni, P., Rockett, K., Kamau, A., Hubbart, C., Jeffreys, A., Ochola-Oyier, L. I., de Villiers, E. P., Gitonga, C. W., Noor, A. M., Snow, R. W., Kwiatkowski, D., & Bejon, P. (2017). Geographic-genetic analysis of Plasmodium falciparum parasite populations from surveys of primary school children in Western Kenya. *Wellcome Open Research*, *2*, 29. https://doi.org/10.12688/wellcomeopenres.11228.2

Onyango, S. A., Ochwedo, K. O., Machani, M. G., Olumeh, J. O., Debrah, I., Omondi, C. J., Ogolla, S. O., Lee, M.-C., Zhou, G., Kokwaro, E., Kazura, J. W., Afrane, Y. A., Githeko, A. K., Zhong, D., & Yan, G. (2022). Molecular characterization and genotype distribution of thioester-containing protein 1 gene in Anopheles gambiae mosquitoes in western Kenya. *Malaria Journal*, *21*(1), 235. https://doi.org/10.1186/s12936-022-04256-w

Onyango, S. A., Ochwedo, K. O., Machani, M. G., Omondi, C. J., Debrah, I., Ogolla, S. O., Lee, M.-C., Zhou, G., Kokwaro, E., Kazura, J. W., Afrane, Y. A., Githeko, A. K., Zhong, D., & Yan, G. (2021). Genetic diversity and population structure of the human malaria parasite Plasmodium falciparum surface protein Pfs47 in isolates from the lowlands in Western Kenya. *PloS One*, *16*(11), e0260434. https://doi.org/10.1371/journal.pone.0260434

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Pedersen, T. L. (2020). *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork

Peterman, W. E. (2018). ResistanceGA: An R package for the optimization of resistance surfaces using genetic algorithms. *Methods in Ecology and Evolution*, *9*(6), 1638–1647. https://doi.org/10.1111/2041-210X.12984

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Prunier, J. G., Colyn, M., Legendre, X., Nimon, K. F., & Flamand, M. C. (2015). Multicollinearity in spatial genetics: Separating the wheat from the chaff using commonality analyses. *Molecular Ecology*, *24*(2), 263–283. https://doi.org/10.1111/mec.13029

Python Software Foundation. (2022). *Python programming language*. https://www.python.org/

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria. https://www.R-project.org/

Rabinovich, R. N., Drakeley, C., Djimde, A. A., Hall, B. F., Hay, S. I., Hemingway, J., Kaslow, D. C., Noor, A., Okumu, F., Steketee, R., Tanner, M., Wells, T. N. C., Whittaker, M. A., Winzeler, E. A., Wirth, D. F., Whitfield, K., & Alonso, P. L. (2017). malERA: An updated research agenda for malaria elimination and eradication. *PLoS medicine*, *14*(11), e1002456. https://doi.org/10.1371/journal.pmed.1002456

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Crowley, J. (2021). *GGally: Extension to 'ggplot2'*. https://CRAN.R-project.org/package=GGally

Sciaini, M., Fritsch, M., Scherer, C., & Simpkins, C. E. (2018). NLMR and landscapetools: An integrated environment for simulating and modifying neutral landscape models in R. *Methods in Ecology and Evolution*, *9*(11), 2240–2248. https://doi.org/10.1111/2041-210X.13076

South, A. (2017). *Rnaturalearth: World Map Data from Natural Earth*. https://CRAN.R-project.org/package=rnaturalearth

Su, X., Ferdig, M. T., Huang, Y., Huynh, C. Q., Liu, A., You, J., Wootton, J. C., & Wellems, T. E. (1999). A genetic map and recombination parameters of the human malaria parasite Plasmodium falciparum. *Science (New York, N.Y.)*, *286*(5443), 1351–1353. https://doi.org/10.1126/science.286.5443.1351

Taylor, A. R., Schaffner, S. F., Cerqueira, G. C., Nkhoma, S. C., Anderson, T. J. C., Sriprawat, K., Phyo, A. P., Nosten, F., Neafsey, D. E., & Buckee, C. O. (2017). Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent. *PLOS Genetics*, *13*(10), e1007065. https://doi.org/10.1371/journal.pgen.1007065

Verity, R. (2018). *Rmaverick: Analysis of population structure*. https://bobverity.github.io/rmaverick/

Verity, R., & Nichols, R. A. (2016). Estimating the Number of Subpopulations (K) in Structured Populations. *Genetics*, *203*(4), 1827–1839. https://doi.org/10.1534/genetics.115.180992

Wan, Z., Hook, S., & Hulley, G. (2021). MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V061. *NASA EOSDIS Land Processes DAAC*. https://doi.org/10.5067/MODIS/MOD11A2.061

Weiss, D. J., Nelson, A., Vargas-Ruiz, C. A., Gligorić, K., Bavadekar, S., Gabrilovich, E., Bertozzi-Villa, A., Rozier, J., Gibson, H. S., Shekel, T., Kamath, C., Lieber, A., Schulman, K., Shao, Y., Qarkaxhija, V., Nandi, A. K., Keddie, S. H., Rumisha, S., Amratia, P., . . . Gething, P. W. (2020). Global maps of travel time to healthcare facilities. *Nature Medicine*, *26*(12), 1835–1838. https://doi.org/10.1038/s41591-020-1059-1

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, *338*(6104), 267–270. https://doi.org/10.1126/science.1223467

Wesolowski, A., Taylor, A. R., Chang, H.-H., Verity, R., Tessema, S., Bailey, J. A., Alex Perkins, T., Neafsey, D. E., Greenhouse, B., & Buckee, C. O. (2018). Mapping malaria by combining parasite genomic and epidemiologic data. *BMC medicine*, *16*(1), 190. https://doi.org/10.1186/s12916-018-1181-9

WHO. (2023). *World Malaria Report 2023*. World Health Organization. Geneva. https://www.who.int/publications/i/item/9789240086173

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Xie, Y. (2014). Knitr: A Comprehensive Tool for Reproducible Research in R. *Implementing Reproducible Research*. Chapman and Hall/CRC.

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC. Retrieved September 6, 2022, from https://bookdown.org/yihui/rmarkdown/

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R Markdown Cookbook*. Chapman and Hall/CRC. Retrieved September 6, 2022, from https://bookdown.org/yihui/rmarkdown-cookbook/

Yu, G. (2021). *Scatterpie: Scatter Pie Plot*. https://CRAN.R-project.org/package=scatterpie

Zhou, Z., Mitchell, R. M., Kariuki, S., Odero, C., Otieno, P., Otieno, K., Onyona, P., Were, V., Wiegand, R. E., Gimnig, J. E., Walker, E. D., Desai, M., & Shi, Y. P. (2016). Assessment of submicroscopic infections and gametocyte carriage of Plasmodium falciparum during peak malaria transmission season in a community-based cross-sectional survey in western Kenya, 2012. *Malaria Journal*, *15*(1), 421. https://doi.org/10.1186/s12936-016-1482-4

2.6 Appendices

2.6.1 Appendix 2A: Study site names, coordinates, and sample sizes

Site Name	Site ID	Long.	Lat.	No. Samples
Chulaimbo	CH	34.6382	-0.0378	48
Yala	YA	34.5371	0.0998	59
Simenya	SI	34.3623	0.1539	39
Sega	SE	34.2283	0.2496	65
Busia	BU	34.1171	0.4463	40
lPali	PA	34.5802	0.0497	32
Shitsitswi	ST	34.505	0.2618	111
Kabula	KA	34.5281	0.4886	72
Mayanja	MY	34.5169	0.6502	52
Chwele	CW	34.578	0.7362	62
Ng'iya (Ngia)	NG	34.3727	0.0431	87
Boro	BO	34.2392	0.086	74
Ruambwa	RW	34.09	0.1293	38
Port Victoria (Bunyala)	VI	33.9732	0.099	125
Sio Port	SP	34.0217	0.2225	21
Kanyawegi	KW	34.611	-0.1133	64
Akala	AK	34.4216	-0.0592	107
Osieko	OS	34.0155	-0.0282	34
Emutete	EM	34.6457	0.0424	41
Iguhu	IG	34.7467	0.1636	80
Malava	MA	34.8542	0.4451	12
Webuye	WE	34.7648	0.6093	7
Kamukuywa	KM	34.7911	0.7798	16

Site Name	Site ID	Long.	Lat.	No. Samples
Miwani	MW	34.9764	-0.0586	34
Kaimosi	KS	34.846	0.1272	15
Kapsabet	KP	35.115	0.203	8
Eldoret	EL	35.2664	0.5354	5
Soy	SO	35.1441	0.6776	29
Kitale	KT	35.0023	1.0191	11
Mwihila	MH	34.6158	0.1773	28
Lugulu	LU	34.3042	0.3931	58
Amukura	AR	34.272	0.5746	51
Chemasiri	CM	34.3936	0.7359	21
Sikubale	SB	34.6845	0.4777	20
Sarora	SR	34.9988	0.4649	10
Kericho	KR	35.2818	-0.366	8
Kendu Bay	KB	34.6501	-0.3712	22
Homa Bay	HB	34.4593	-0.533	15
Luanda	LD	34.2192	-0.8115	26
Paulo	PL	34.6	-0.7667	8
Keroka	KK	34.9489	-0.781	5
Migori	MI	34.4667	-1.0667	28
Kilgoris	KG	34.876	-1.0076	11
Mukhobola	MU	34.0298	0.0787	105

2.6.2 Appendix 2B: Resistance surface fitting results

Surface	Avg. AICc	Rep. Num.
Lake Victoria	-1009.1	2
Friction to Human Movement (w/o motorized)	-1007.02	2
Lake Victoria and Land Surface Temperature	-1006.28	2
Lake Victoria	-1005.41	1
Land Surface Temperature	-1004.61	2
Elevation and Lake Victoria	-1003.81	2
Friction to Human Movement (w/o motorized)	-1003.7	1
Lake Victoria and Land Surface Temperature	-1002.68	1
Lake Victoria and Friction to Human Movement (w/o motorized)	-1002.25	2
Land Surface Temperature and Friction to Human Movement (w/o motorized)	-1001.88	2
Distance	-1001.48	2
Land Surface Temperature	-1001.34	1
Elevation and Lake Victoria	-1001.23	1
Lake Victoria and Precipitation	-1000.97	2
Lake Victoria and Friction to Human Movement (w/ motorized)	-1000.72	2
Elevation and Friction to Human Movement (w/o motorized)	-1000.37	2
Lake Victoria and Friction to Human Movement (w/o motorized)	-998.56	1
Elevation	-998.484	2
Distance	-998.442	1
Land Surface Temperature and Precipitation	-998.277	2

Surface	Avg. AICc	Rep. Num.
Friction to Human Movement (w/ motorized) and Friction to Human Movement (w/o motorized)	-998.168	2
Friction to Human Movement (w/ motorized)	-997.935	2
Precipitation and Friction to Human Movement (w/o motorized)	-997.226	2
Lake Victoria and Friction to Human Movement (w/ motorized)	-997.157	1
Lake Victoria and Precipitation	-997.122	1
Precipitation	-996.789	2
Land Surface Temperature and Friction to Human Movement (w/o motorized)	-996.604	1
Elevation and Friction to Human Movement (w/o motorized)	-996.285	1
Elevation and Land Surface Temperature	-995.541	2
Friction to Human Movement (w/ motorized)	-994.961	1
Land Surface Temperature and Precipitation	-994.425	1
Land Surface Temperature and Friction to Human Movement (w/ motorized)	-994.106	2
Precipitation and Friction to Human Movement (w/o motorized)	-993.88	1
Elevation and Precipitation	-993.707	2
Precipitation	-993.547	1
Elevation and Friction to Human Movement (w/ motorized)	-993.509	2
Friction to Human Movement (w/ motorized) and Friction to Human Movement (w/o motorized)	-993.465	1
Elevation	-993.413	1
Elevation and Land Surface Temperature	-992.485	1
Land Surface Temperature and Friction to Human Movement (w/ motorized)	-992.055	1

Surface	Avg. AICc	Rep. Num.
Precipitation and Friction to Human Movement (w/ motorized)	-991.209	2
Elevation and Precipitation	-990.655	1
Elevation and Friction to Human Movement (w/ motorized)	-990.126	1
Precipitation and Friction to Human Movement (w/ motorized)	-987.679	1
Land Cover	-985.917	2
Land Cover	-980.849	1
Land Cover and Lake Victoria	-975.82	2
Land Cover and Friction to Human Movement (w/o motorized)	-972.567	2
Land Cover and Lake Victoria	-971.934	1
Elevation and Land Cover	-969.947	2
Land Cover and Precipitation	-968.967	2
Land Cover and Friction to Human Movement (w/ motorized)	-968.842	2
Land Cover and Friction to Human Movement (w/o motorized)	-968.524	1
Land Cover and Land Surface Temperature	-967.923	2
Elevation and Land Cover	-966.131	1
Land Cover and Friction to Human Movement (w/ motorized)	-966.102	1
Land Cover and Land Surface Temperature	-964.706	1
Land Cover and Precipitation	-964.568	1

2.6.3 Appendix 2C: Resistance surface bootstrapping results

Surface	Avg. AICc	Rep. Num.
Lake Victoria and Land Surface Temperature	-1828.48	1
Lake Victoria and Land Surface Temperature	-1828.47	2
Lake Victoria	-1827.51	1
Lake Victoria	-1827.51	2
Friction to Human Movement (w/o motorized)	-1826.92	1
Friction to Human Movement (w/o motorized)	-1826.23	2
Elevation and Lake Victoria	-1825.64	1
Land Surface Temperature and Friction to Human Movement (w/o motorized)	-1823.91	2
Elevation and Lake Victoria	-1823.91	2
Lake Victoria and Friction to Human Movement (w/o motorized)	-1822.71	1
Lake Victoria and Friction to Human Movement (w/o motorized)	-1822.71	2
Land Surface Temperature	-1822.26	2
Land Surface Temperature	-1821.96	1
Lake Victoria and Precipitation	-1821.54	2
Elevation and Friction to Human Movement (w/o motorized)	-1821.3	2
Land Surface Temperature and Friction to Human Movement (w/o motorized)	-1821.3	1
Lake Victoria and Precipitation	-1821.29	1
Elevation and Friction to Human Movement (w/o motorized)	-1820.45	1
Lake Victoria and Friction to Human Movement (w/ motorized)	-1820.32	1
Lake Victoria and Friction to Human Movement (w/ motorized)	-1819.93	2

Surface	Avg. AICc	Rep. Num.
Friction to Human Movement (w/ motorized) and Friction to Human Movement (w/o motorized)	-1818.71	2
Precipitation and Friction to Human Movement (w/o motorized)	-1818.4	1
Precipitation and Friction to Human Movement (w/o motorized)	-1818.3	2
Land Surface Temperature and Precipitation	-1817.73	2
Friction to Human Movement (w/ motorized) and Friction to Human Movement (w/o motorized)	-1816.87	1
Land Surface Temperature and Precipitation	-1816.7	1
Elevation and Land Surface Temperature	-1814.15	2
Elevation and Land Surface Temperature	-1814.14	1
Friction to Human Movement (w/ motorized)	-1813.4	1
Friction to Human Movement (w/ motorized)	-1813.27	2
Distance	-1813.18	1
Distance	-1813.18	2
Land Surface Temperature and Friction to Human Movement (w/ motorized)	-1813.18	1
Land Surface Temperature and Friction to Human Movement (w/ motorized)	-1811.76	2
Elevation	-1811.28	2
Elevation and Friction to Human Movement (w/ motorized)	-1810.88	2
Elevation and Friction to Human Movement (w/ motorized)	-1810.54	1
Elevation and Precipitation	-1809.53	1
Elevation and Precipitation	-1809.49	2
Land Cover	-1809.24	2
Precipitation	-1808.71	2

Surface	Avg. AICc	Rep. Num.
Precipitation	-1808.52	1
Elevation	-1808.45	1
Precipitation and Friction to Human Movement (w/ motorized)	-1808.4	2
Precipitation and Friction to Human Movement (w/ motorized)	-1807.83	1
Land Cover	-1807.51	1
Land Cover and Friction to Human Movement (w/o motorized)	-1803.18	2
Land Cover and Friction to Human Movement (w/o motorized)	-1802.76	1
Land Cover and Lake Victoria	-1802.19	1
Land Cover and Lake Victoria	-1802.05	2
Land Cover and Friction to Human Movement (w/ motorized)	-1799.06	1
Elevation and Land Cover	-1798.35	2
Elevation and Land Cover	-1798.24	1
Land Cover and Friction to Human Movement (w/ motorized)	-1796.8	2
Land Cover and Precipitation	-1796.79	1
Land Cover and Precipitation	-1796.73	2
Land Cover and Land Surface Temperature	-1796.36	1
Land Cover and Land Surface Temperature	-1795.1	2
Null	-1786.87	1
Null	-1786.87	2

CHAPTER 3: Development of a globally-applicable, highly-multiplexed microhaplotype amplicon panel for *Plasmodium vivax*

Article Authors: Alfred Hubbard, Edwin Solares, Lauren Bradley, Brook Jeang, Delenasaw Yewhalaw, Daniel Janies, Eugenia Lo, Guiyun Yan, Elizabeth Hemming-Schroeder

3.1 Background

Plasmodium vivax is the most widely geographically distributed malaria parasite, and there is increasing evidence that *P. vivax* is circulating in all regions of Africa (Twohig et al., 2019). In addition to causing considerable morbidity, including anemia, malnutrition, and poor school performance in early childhood, *P. vivax* can cause severe and life-threatening malaria (Anstey et al., 2012). In part because *P. vivax* has historically been considered benign or non-fatal, *P. vivax* malaria remains understudied in comparison to *P. falciparum* malaria. Improving our understanding of *P. vivax* epidemiology is a key step to planning effective antimalarial interventions and improving malaria control.

Population genomics is one powerful approach for gaining insights into malaria epidemiology and informing control programs. Malaria genomic information can be used to target control resources to areas of high transmission and evaluate the effectiveness of antimalarial interventions with genetic indicators of transmission intensity (Neafsey et al., 2021). However, the capacity of population genomics for investigating malaria epidemiology is limited by technical constraints and costs. For example, classical biallelic SNP assays have low sensitivity to detect multiple parasite strains and parasite diversity within a host (Koepfli & Mueller, 2017). On the other hand, whole genome sequencing (WGS) provides high resolution data, but the data is costly to produce (Tessema et al., 2022) and store (Neafsey et al., 2021). One cost-effective method for obtaining moderately high-resolution genomic data with high sensitivity to detecting within-host parasite diversity is targeted deep sequencing of genetically diverse and informative amplicons (Koepfli & Mueller, 2017; Tessema et al., 2022).

Furthermore, data generated from genotyping-by-sequencing can be used to assess multiallelic microhaplotypes, genetic loci of less than 300 nucleotides, defined by two or more proximate SNPs. The major advantage of this approach is that kinship analysis based on microhaplotype markers has been shown to provide higher power for relationship inference than biallelic SNPs (Baetscher et al., 2018), particularly in the case of infections that are polyclonal, meaning they have multiple, genetically-distinct parasite strains (Tessema et al., 2022). For researchers that study eukaryotic parasite epidemiology, highly-multiplexed amplicon sequencing panels of polymorphic microhaplotype markers along with advances in analytical methods present a promising avenue to accurately assessing Plasmodium genetic relatedness and transmission patterns (LaVerriere et al., 2022; Tessema et al., 2022). Indeed, recent efforts have been made to develop such panels for the more studied and deadly Plasmodium spp., *P. falciparum* (LaVerriere et al., 2022; Tessema et al., 2022).

Unfortunately, no comparable panel with a large number of high-diversity microhaplotype loci, suitable for population genetics, has been published for *P. vivax*. One panel that includes 11 microhaplotype markers has been released, but these are all for putative drug resistance genes (Kattenberg et al., 2022). While genotyping such genes is also of great interest for malaria control purposes, these genes may be under selection and thus are not suitable for population genetics analyses that can reveal information about patterns of transmission. At least one other panel is also under development (Siegel et al., 2023), but until it is published its utility for *P. vivax* genomic epidemiology will remain unclear.

To support a full range of objectives in genomic epidemiology studies of *P. vivax*, a microhaplotype marker panel would ideally have the following capabilities: sensitively detect multiple parasite strains in an infection; maintain quality coverage in low-parasitemia infections; generate sequence data on specific genes or points of interest, such as candidate markers for drug resistance; and enable effective relationship inference in areas with low or high population structuring.

In this study, we design and implement a highly-multiplexed amplicon sequencing panel of *P. vivax* microhaplotypes that satisfies these criteria.

3.2 Materials and Methods

3.2.1 Panel design

Whole genome sequence data for 198 *P. vivax* isolates from eight countries (Cambodia, the China-Myanmar border region, Colombia, Ethiopia, Madagascar, Malaysia, Panama, and Peru) were downloaded from NCBI. Sequences were aligned by bwa (Li & Durbin, 2009) in conjunction with SAMtools (Danecek et al., 2021). Alignments were removed using BCFTools (Danecek et al., 2021) if they showed multiple mappings, mappings across chromosomes, had mean coverage > 4x, or quality values < 20. SNPs and indels were called using GATK (Van der Auwera & O'Connor, 2020) in conjunction with Picard-tools ("Picard Toolkit", 2019) and VCF-tools (Danecek et al., 2011). Final SNPs and indels were called using the HaplotypeCaller and GenotypeGVCT algorithms.

To identify candidate markers for our amplicon sequencing panels, we used a sliding window method adapted from the approach used by Tessema et al. (2022) to design a microhaplotype panel for *P. falciparum*. We divided the genome into 200bp sliding windows every 100bp, yielding a total of 242,135 windows, using the sliding.window.transform function in the PopGenome R package (Pfeifer et al., 2014). This initial set of windows was then filtered based on presence of tandem repeats, presence of indels, and genetic diversity. Tandem repeats in the genome were identified using Tandem Repeats Finder (Benson, 1999). Windows were removed that contained tandem repeats > 40 bp, dinucleotide repeats > 8 bp, homopolymer repeats > 8 bp, or trinucleotide repeats > 12 bp. Second, windows containing any insertion or deletion from variant calling were removed. Third, within-country nucleotide diversity (π) was calculated for the remaining windows in PopGenome, and windows that were monomorphic ($\pi = 0$) in isolates from > 25% of the countries were excluded prior to further analysis, as these windows would not be informative for fine-scale genomic analyses in certain study regions. This filtering process led to 2,498 candidate windows remaining for potential inclusion in the panel.

We proceeded to evaluate and compare candidate windows based on polymorphism and genetic structuring. Specifically, we evaluated windows for mean within-country nucleotide diversity (π) and averaged fixation index (F_{ST}) for each country against all other individuals. Both values were calculated in PopGenome using the F_ST.stats function. Windows were then ranked by their F_{ST} and π values. To achieve a relatively even distribution of loci across the genome, for each chromosome the 10 windows with the highest π values, the window with the highest mean F_{ST} value, and the window with the highest π that also was in the highest 8% of F_{ST} values were selected. After that, the remaining windows were ranked overall (i.e., all chromosomes pooled together), and the 72 windows with the highest π value, the 28 windows with the highest F_{ST} value, and the 10 windows with the highest π that were also in the highest 8% of F_{ST} values were selected. At this point, windows were removed from consideration if there was insufficient availability of conserved regions outside of the target window for primer design and replaced with the window having the next highest value.

The final set of candidate windows which were selected for primer design consisted of 278 targets with the number of targets per chromosome ranging from 16 to 25. The minimum value for targets selected for π was 0.0024 and for F_{ST} was 0.50. These minimum values were among the top 55% and 93% of values, respectively, of the original 2,498 candidate windows.

Our goal was to generate a panel with approximately 100 targets, but we selected an abundance of potential targets expecting fall-out from primer incompatibilities during primer design and uneven amplification and/or sequencing coverage during assay development. The 278 targets were submitted to GTseek LLC for primer design with the goal of designing primers that generate minimal crosstalk during multiplexed PCR reactions. Primers were successfully designed for 179 of the 278 targets. After small test sequencing runs, we further removed primers for targets that did not amplify successfully, yielding a reduced set of 76 targets.

In addition to the targets selected for their π and/or F_{ST} values, eight additional loci of interest were added to the panel, bringing the final panel size to 84. The genes containing these additional loci are *PVDBP*, *PVCRT*, *PVDHFR*, *PVDHPS* (two loci), *PVK12* (two loci), and *PVMDR1*.

Primers for these targets were also designed by GTSeek LLC to minimize crosstalk among primers during amplification.

3.2.2 Panel evaluation with field samples

3.2.2.1 Sample collection

To evaluate the panel in terms of sequencing yield and ability to obtain haplotypes, we used six dried blood spot (DBS) samples previously collected as part of an ongoing Sub-Saharan Africa International Center for Excellence for Malaria Research (ICEMR) project. Some of these samples were obtained with passive case detection and some came from mass blood surveys. The samples were collected from three ongoing field sites in the Oromo region of Southwest Ethiopia; the Arjo-Didessa sugarcane farm, the Gambella region, and Jimma town.

In addition, four whole blood samples collected from Agaro Health Center, which is near Jimma, Ethiopia, were used to evaluate panel performance with DNA extracts derived from whole blood.

3.2.2.2 Library preparation

For DBS samples, two different extraction methods were tested: the Qiagen QIAamp DNA Investigator Kit (from now on referred to as the "kit") and the Saponin-Chelex method (Bereczky et al., 2005). In addition, for all samples, three different enrichment strategies were tested: no enrichment, a targeted nested PCR prior to the GT-Seq PCR (from here forward referred to as targeted pre-amplification), and selective whole genome amplification (SWGA) prior to the first PCR in the GT-Seq protocol.

When SWGA was used, we performed the reaction and bead cleanup according to the established protocol (Cowell et al., 2017), prior to the GT-Seq PCR. When targeted pre-amplification was performed, nested reaction mixtures and amplification were conducted following the same conditions as PCR 1 of the GT-Seq protocol, prior to the GT-Seq PCR. Following enrichment, the protocol consists of an initial PCR with primers for amplicon targets containing adapter sequences, an enzymatic cleanup of PCR products, a second PCR for normalization and incorporation of dual indexing tags (Nate's plates), and bead size selection. Libraries were sequenced on an Illumina

Miseq with a 500 cycle kit in paired-end mode.

3.2.2.3 Sequence analysis

We used the `SeekDeep` pipeline (Hathaway et al., 2018) to demultiplex, filter reads, and estimate haplotype frequencies. First, to join and extract reads by locus, we used the extractorPairedEnd function with default filtering and quality parameters and additional primer specifications (minOverlap: 6, primercoverage: 0.7, primerWithinStart: 20). Next, the qluster function with default parameters for Illumina data was used to create haplotypes with relative abundances. Final filtering was done using the processsClusters function with parameters that allow for a few low-quality mismatches and no indels, recommended by the developer for Illumina data (parameters: strictErrors, illumina, fracCutOff 0.02).

To evaluate the different laboratory protocols, the percentage of on-target reads was calculated by dividing the total number of reads that were matched to panel loci by `SeekDeep` for each sample by the total number of reads for that sample in the raw data.

3.2.3 Panel evaluation with MalariaGEN data

3.2.3.1 Data preparation

To assess the utility of the panel at the population level, whole genome sequences from the MalariaGEN Pv4 project (MalariaGEN et al., 2022) for the years 2015 through 2016 were downloaded for comparison with our results. These are the two most recent years in the dataset that have a large number of samples. Note that we intentionally chose to use a different set of sequences for the evaluation of the panel than was used for panel design, to avoid biasing our results.

Variants and samples that failed the QC process of the MalariaGEN authors were removed. Samples with an F_{WS} below 0.95 were also removed, as these are considered to be polyclonal (Auburn et al., 2012). Finally, samples from longitudinal studies and returning travelers were removed. This yielded a final analysis set of 185 samples, from 10 different countries and 19 distinct sites.

The remaining variants were subset to the genomic regions represented in the panel using `bcftools view` (Danecek et al., 2021). Haplotype sequences were obtained for each sample and locus using `bcftools consensus` (Danecek et al., 2021), using the PvP01 reference genome (Auburn et al., 2016).

3.2.3.2 Population genetics

Haplotype sequences for each locus were aligned using the MUSCLE algorithm with the R package `msa` (Bodenhofer et al., 2015), after which selection was assessed with Tajima's *D* (Tajima, 1989), calculated with the R package `pegas` (Paradis, 2010). Selection was assessed separately for each population in the dataset, using the population definitions provided by the MalariaGEN authors. Pairwise linkage disequilibrium (LD) between all pairs of loci that share a chromosome was estimated with the $\bar{r}_d$ statistic (Agapow & Burt, 2001), calculated with the `poppr` R package (Kamvar et al., 2014). The *p*-value thresholds for both the tests of selection and pairwise LD were corrected for multiple testing using the Bonferroni method. Loci under significant selection or in significant LD with other loci, at the 0.05 level, were filtered out before proceeding with the analyses described below.

Nei's expected heterozygosity (Nei, 1978) was estimated for each locus with more than one allele using the `poppr` R package (Kamvar et al., 2014). This metric was computed separately for each country and for each site in Cambodia and Vietnam, to facilitate comparison with the relatedness analysis (see below). In each case, *t*-tests were performed between each pair of groups to identify significant differences. The Bonferroni method was used to correct for multiple testing. At the country level, countries with fewer than 15 samples were removed to avoid introducing bias from low sample size.

Identity-by-descent (IBD) between samples was estimated using the R package `Dcifer` (Gerlovina et al., 2022). This tool uses population-level allele frequencies for each individual to estimate whether observed sharing of genotypes between sample pairs is because of sharing in the most recent common ancestor (in which case they are said to be identical-by-descent) or due to chance alone. Significance is assessed using a likelihood ratio approach.

Relatedness values estimated with `Dcifer` were analyzed in two ways: at the country level for the entire dataset, and at the site level for Cambodia and Vietnam. In both cases, the mean relatedness of all constituent sample pairs was computed for each pair of countries or sites. In addition, the fraction of highly-related pairs was computed for the site level comparison, using a relatedness threshold of 0.25, as metrics based on the number of highly-related pairs are more sensitive to recent gene flow (Taylor et al., 2017). As with the expected heterozygosity analysis, countries with fewer than 15 samples were removed.

3.3 Results

3.3.1 Panel characteristics

As intended, the genome windows selected for inclusion in the final panel all have high nucleotide diversity (Figure 3.1A) and/or high F_{ST} (Figure 3.1B), making them informative for genomic epidemiology analyses.

3.3.2 Panel evaluation with field samples

Among the various DNA preservation, extraction, and enrichment methods tested, the results point to two different strategies, depending on parasite density. From the DBS samples, SWGA more consistently produced high percentages of on-target reads than the other enrichment methods (Figure 3.2A), but targeted pre-amplification led to more consistent amplification of each marker (Figure 3.4A; Figure 3.3). The Chelex- and kit-based extraction methods were relatively comparable in terms of on-target reads (Figure 3.2A) and marker amplification (Figure 3.4A). High percentages of on-target reads were obtained from whole blood samples for both SWGA and targeted pre-amplification (Figure 3.2B), but many loci did not amplify well from these samples, regardless of enrichment method (Figure 3.4B). From these results, it can be concluded that with this panel whole blood samples are not worth the additional effort and cost, and there is little difference in performance between Chelex-based and kit-based extraction. However, there are tradeoffs between SWGA and targeted pre-amplification, and it is not apparent which is best.

The picture becomes more clear when the relationship with parasite density is considered (Figure 3.5). It would seem that the differences observed above between targeted pre-amplification and

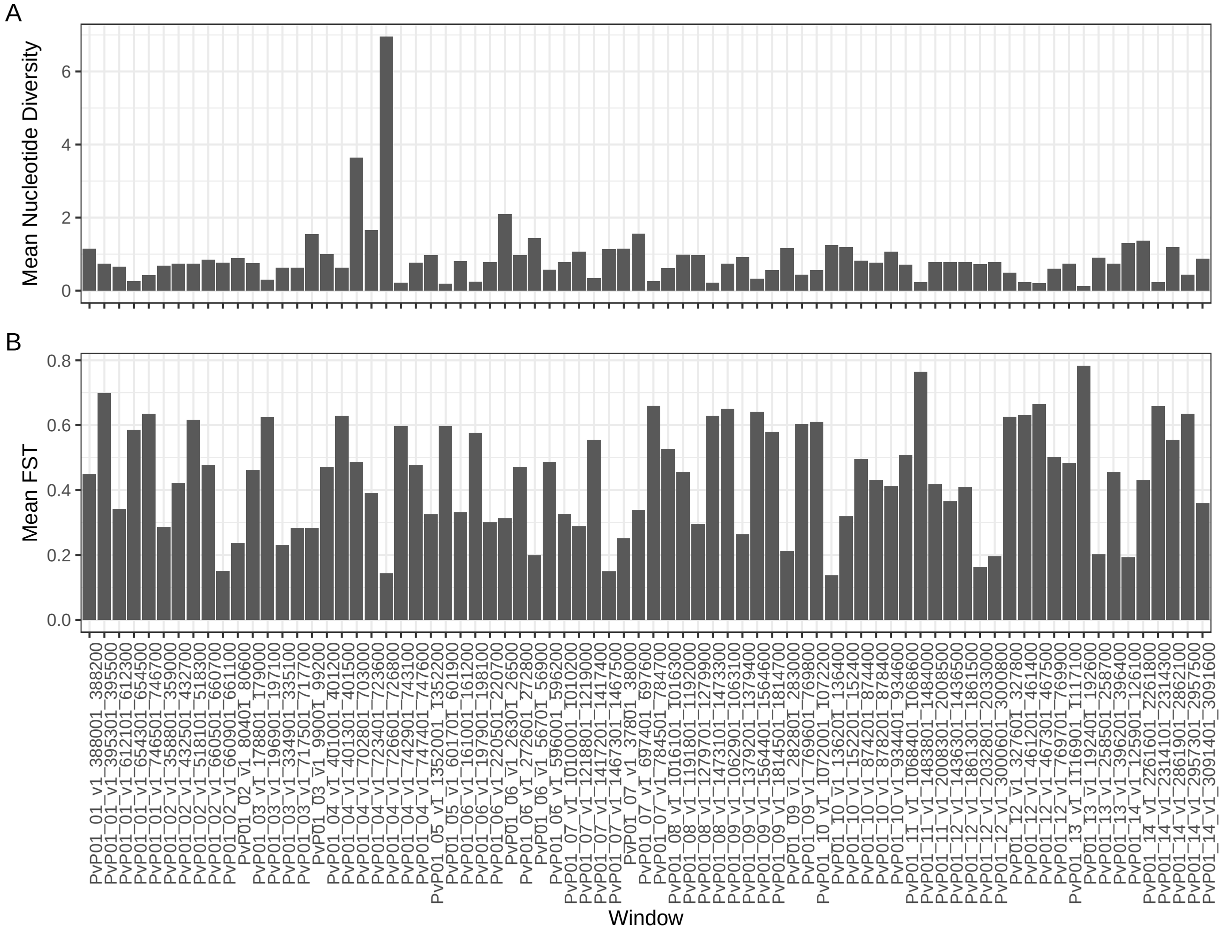


Figure 3.1: Mean nucleotide diversity **(A)** and mean F_{ST} **(B)** for the genome windows selected for inclusion in the final panel.

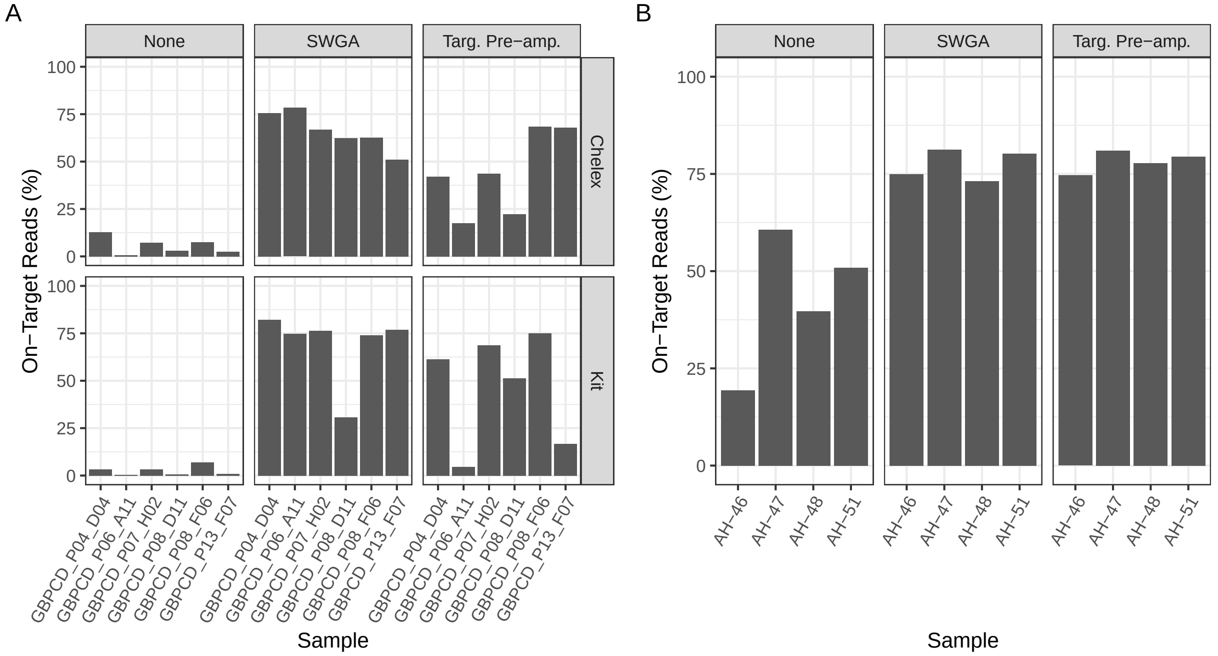


Figure 3.2: Bar plots indicating the percentage of on-target reads for the DBS samples **(A)** and the whole blood samples **(B)**. There is one plot for each DNA enrichment method and, in the case of the DBS samples, the extraction method that was tested.

SWGA are due at least in part to the effect of parasite density. Targeted pre-amplification performs better in samples with a high parasite load, whereas the performance of SWGA is unaffected by parasite density. These results indicate that different methods are preferable depending on parasite load. If parasitemia is high, using targeted pre-amplification will lead to more consistent amplification across the loci in the panel. However, if parasitemia is low, SWGA is recommended instead.

3.3.3 Panel evaluation with MalariaGEN data

In the analysis of Pv4 samples, no pairs of loci from the same chromosome were identified as being in significant LD, after the significance threshold of 0.05 was Bonferroni-corrected. However, two loci had a negative Tajima's *D* in one or more populations at the 0.05 significance level, after Bonferroni correction. These loci, *Pvcrt_o.10k.indel* and *PvP01_10_v1_1072001_1072200*, were removed from the dataset prior to genetic diversity and relatedness analysis.

In terms of genetic diversity, Colombia and Indonesia had a lower overall expected heterozygos-

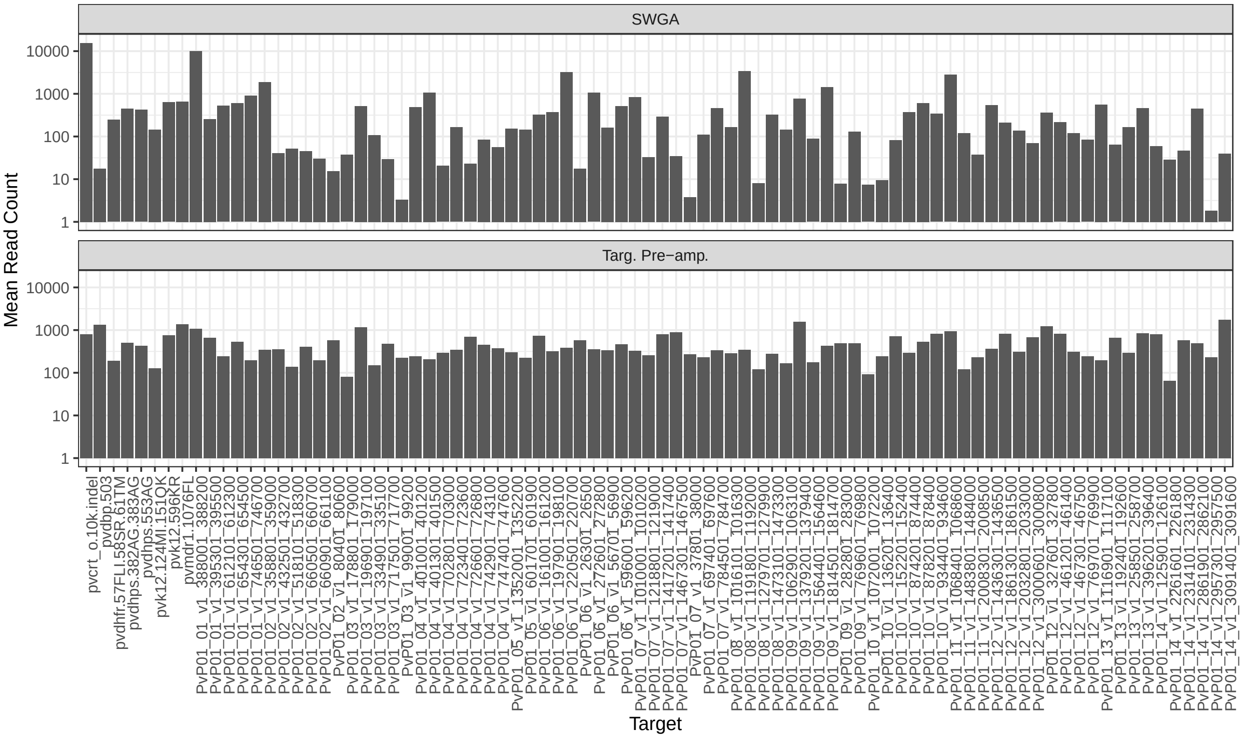


Figure 3.3: Mean read counts across all DBS samples for each locus, obtained with the kit DNA extraction method and the SWGA (top) and targeted pre-amplification (bottom) DNA enrichment methods. In both cases, the *y*-axis is $\log_{10}$ transformed.

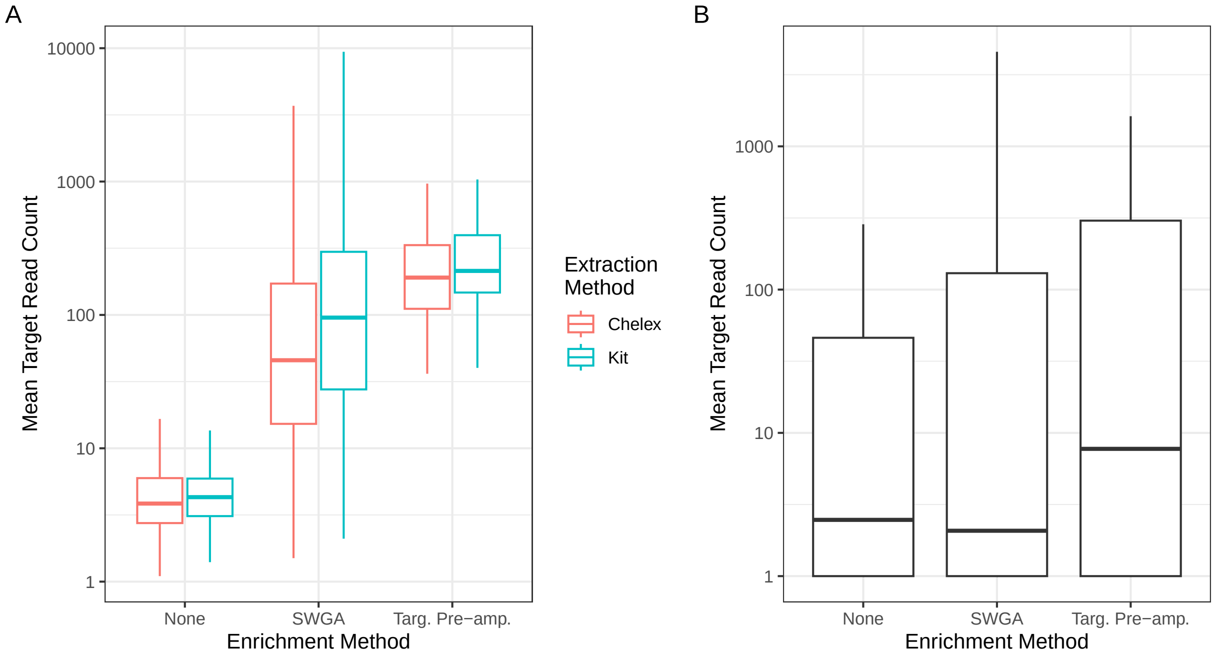


Figure 3.4: Box plots (with Tukey-style whiskers that extend to a maximum of 1.5 * IQR) showing the mean read count across all samples for each locus, separated according to DNA enrichment and extraction method. The DBS samples are shown in **(A)** and the whole blood samples in **(B)**. In both cases, the *y*-axis is log_{10} transformed.

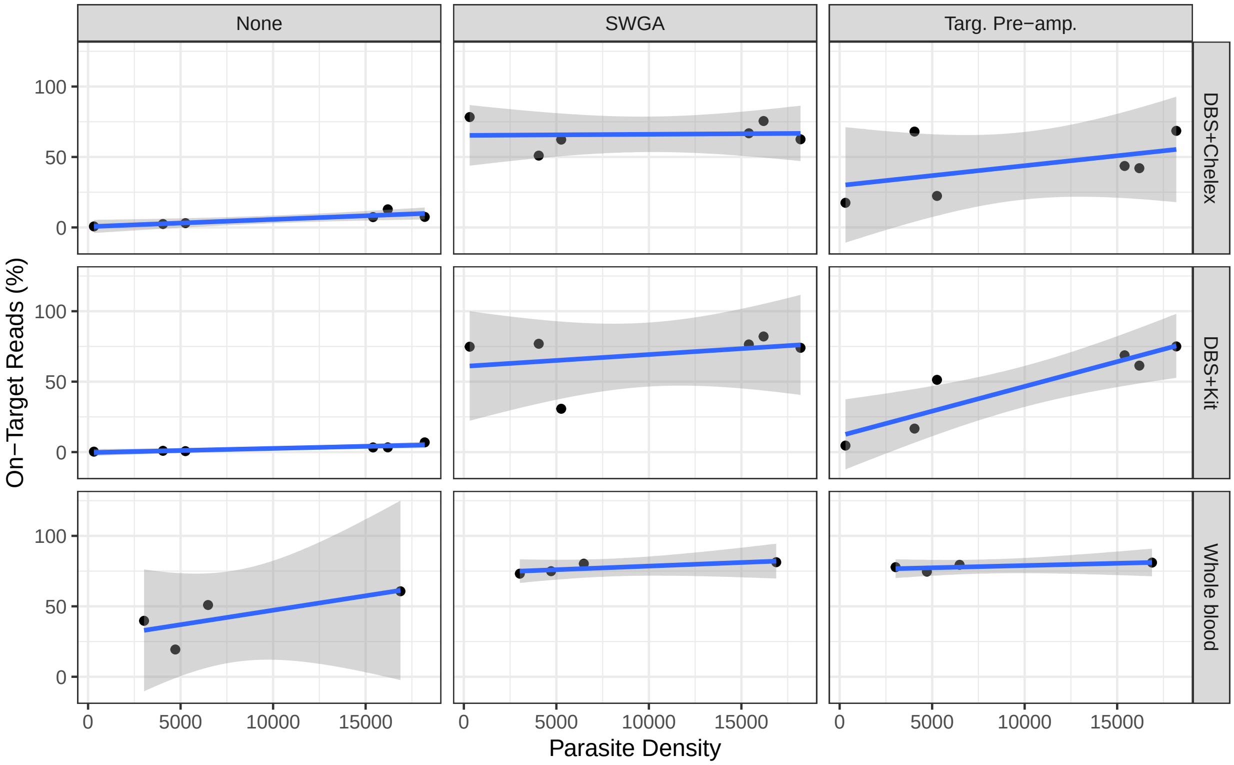


Figure 3.5: These scatterplots show the relationship between the percentage of on-target reads and parasite density, for each combination of enrichment protocol (none, SWGA, and targeted pre-amplification) and DNA preservation/extraction method (DBS+Chelex, DBS+Kit, and whole blood). In each case, a linear regression between parasite density and the percentage of on-target reads is displayed, with confidence intervals, to aid interpretation.

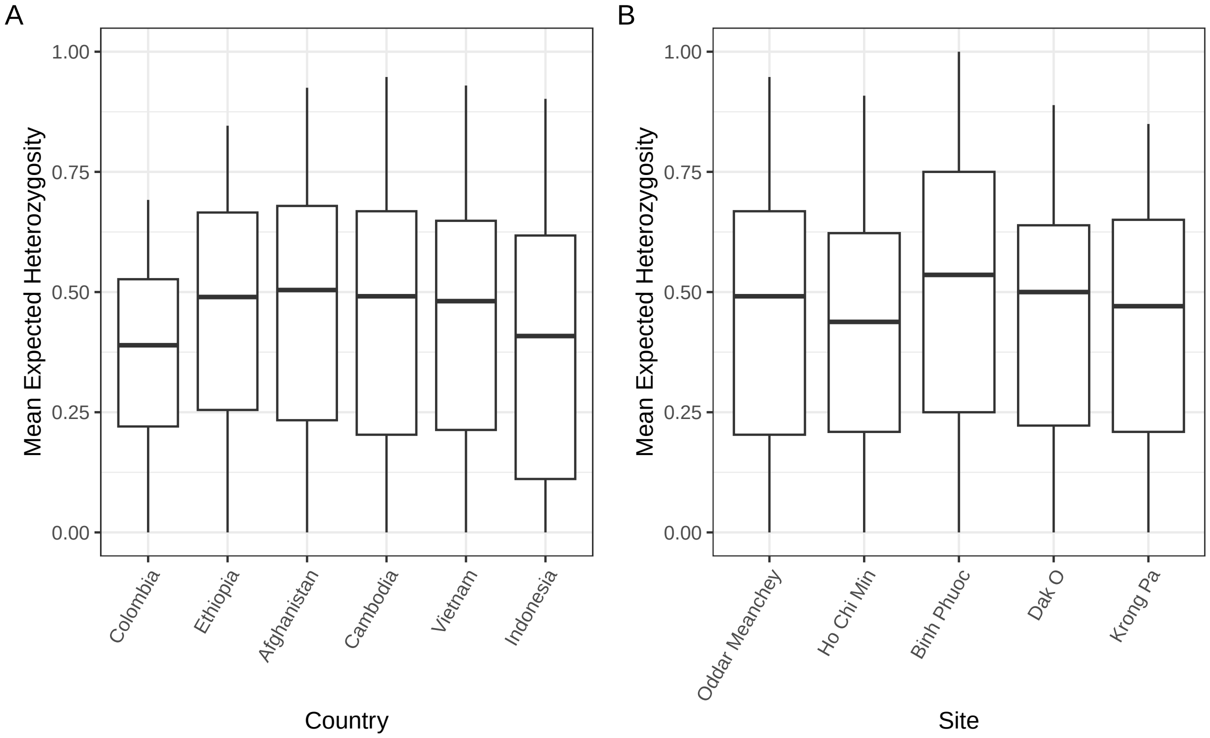


Figure 3.6: Distribution of mean expected heterozygosities for each marker, separated by country **(A)** and site **(B)**. The whiskers are Tukey-style and extend to a maximum of 1.5 * IQR.

ity than the other countries (Figure 3.6A), and Ho Chi Min had the lowest expected heterozygosity of any of the sites in Cambodia and Vietnam (Figure 3.6B). However, none of these differences were significant after applying the Bonferroni correction for multiple testing.

Most sample pairs have a relatedness of zero (Figure 3.7A). Among the other pairs, there are a few clonal samples (relatedness of one) and most of the rest have a relatedness below 0.25 (the relatedness of half-siblings; Figure 3.7B). As expected, mean IBD-based relatedness tends to be higher within countries than between countries (Figure 3.8). Also, relatedness between countries clearly shows regional divisions, particularly with Southwest Asia/East Africa and Southeast Asia (Figure 3.8). Gene flow appears to be particularly high between Cambodia and Vietnam, as the relatedness between these two countries is comparable to the relatedness within those countries.

To demonstrate the panel's ability to distinguish patterns of genetic connectivity within a region, the results for Cambodia and Vietnam are examined in detail. At the site level, mean relatedness of

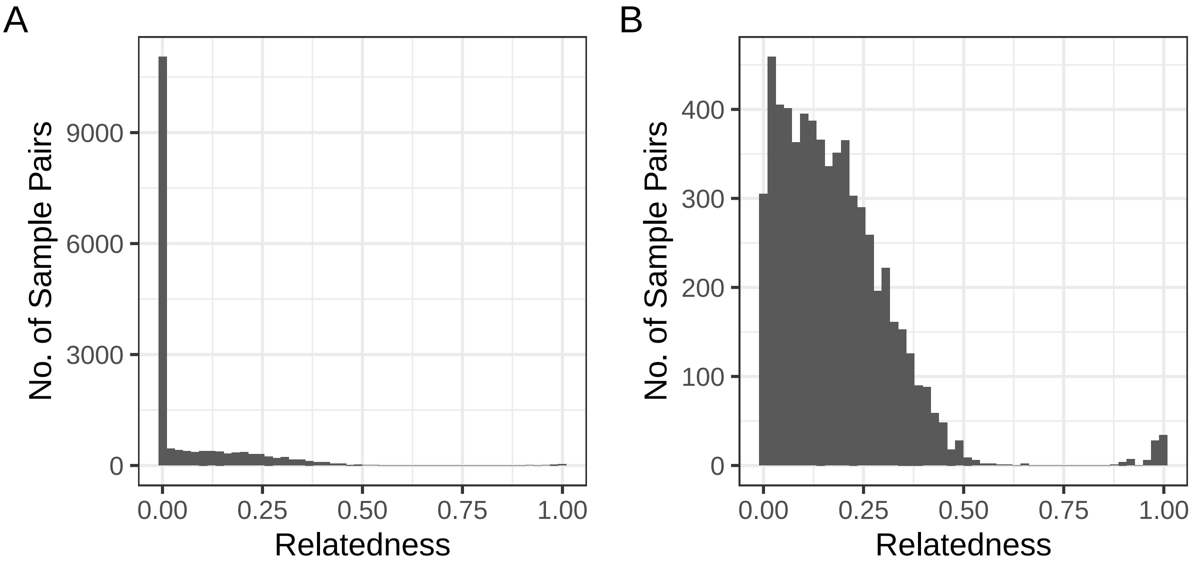


Figure 3.7: Histograms displaying relatedness estimated by `Dcifer` for **(A)** all sample pairs and **(B)** sample pairs with relatedness greater than zero.

constituent sample pairs provides some ability to distinguish gene flow between pairs of sites (Figure 3.9A), but the distinctions between sites become more clear when the fraction of highly-related sample pairs is used instead (Figure 3.9B). This is consistent with the theoretical expectation that metrics based on highly-related sample pairs will perform better as analysis moves from the global to the local level, as these metrics are better equipped to capture recent gene flow (Taylor et al., 2017). However, even when the fraction of highly-related sample pairs is considered, the patterns of genetic relatedness in this region are complex. When visualized in geographic space, it becomes apparent that a simple pattern of isolation-by-distance does not explain genetic relatedness of *P. vivax* in this region (Figure 3.10). Instead, it can be seen that Ho Chi Min and Dak O have comparatively high relatedness to the other sites, implying that these locations may be hubs of malaria transmission in the region.

3.4 Discussion and Conclusion

The microhaplotype marker panel designed and evaluated in this study shows substantial promise for enhancing genomic epidemiology in *P. vivax*. After several filters, we obtained a final panel

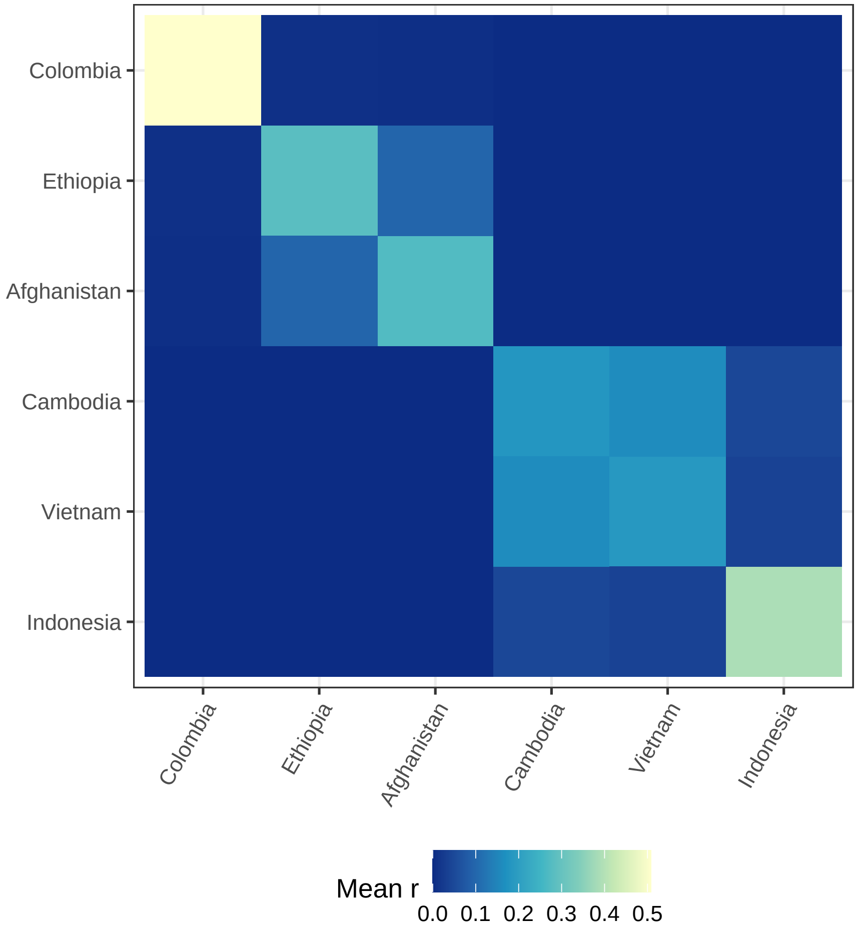


Figure 3.8: Heatmap showing the mean relatedness of all constituent sample pairs in each pair of countries in the MalariaGEN data. Color swatches along the diagonal indicate within-country relatedness. Countries with fewer than 15 samples have been removed.

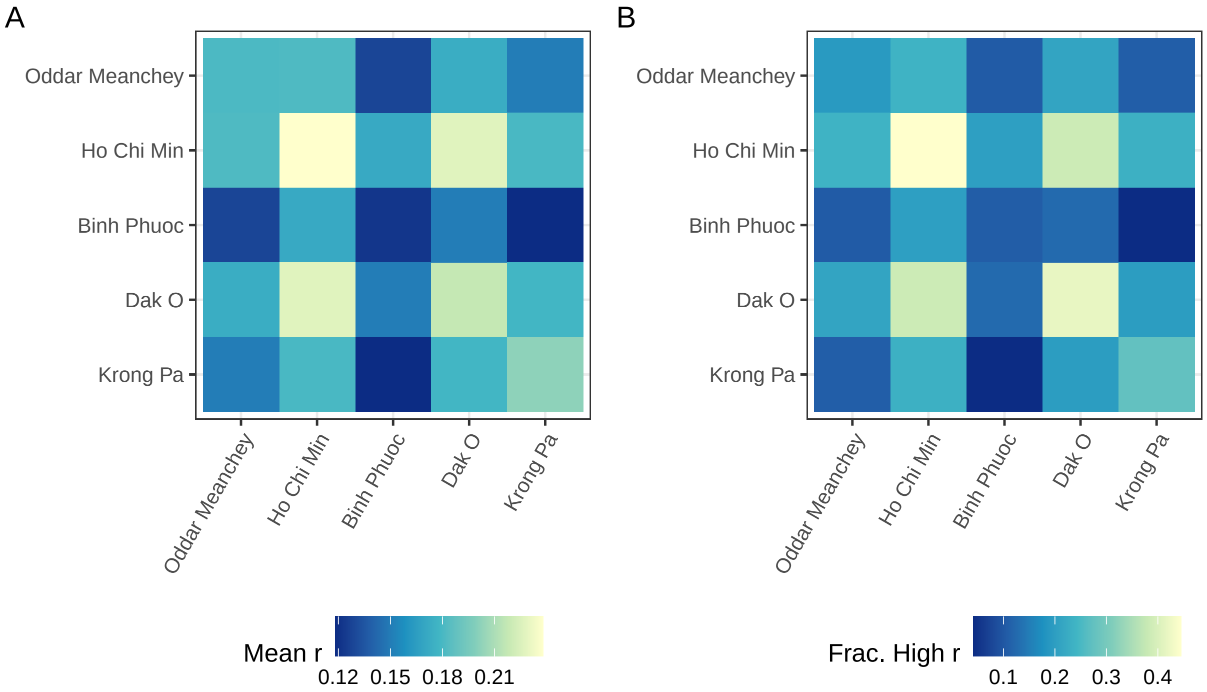


Figure 3.9: Heatmaps showing the relatedness within and between sites in Cambodia and Vietnam. **(A)** shows the mean relatedness of all constituent sample pairs and **(B)** gives the fraction of highly-related sample pairs corresponding to each pair of sites. Color swatches along the diagonal indicate within-site relatedness.

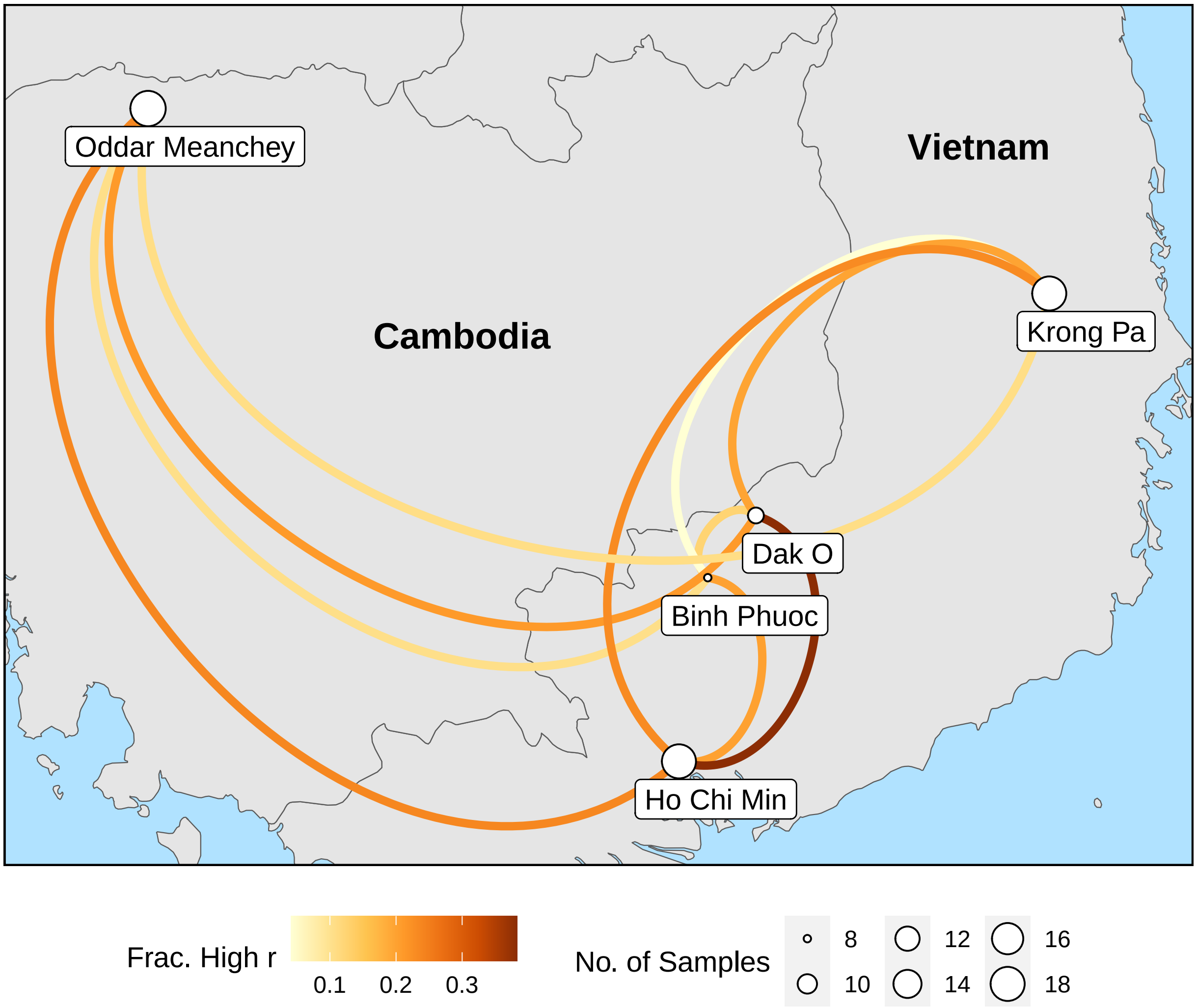


Figure 3.10: Network showing between-site relatedness (visualized as the color of the links) for samples from Cambodia and Vietnam. Between-site relatedness was calculated as the fraction of highly-related sample pairs. Node size is scaled according to the number of samples from that site.

with 84 loci total, 76 of which are designed for population genetics and 8 of which are genes of epidemiological interest (e.g., potential drug resistance markers). Our evaluation with field samples shows that the panel works well with DBS samples, and that the most consistent amplification across loci is achieved when our targeted pre-amplification enrichment method is used. However, SWGA may be advisable instead if parasitemia is low. We did not detect differences between Chelex- and kit-based extraction, and so do not make a recommendation in this area. The population genetics analysis with MalariaGEN data demonstrates that the panel not only distinguishes between geographic regions, but also can identify substantial within-region variation in genetic relatedness, as evidenced by the analysis of sites in Cambodia and Vietnam.

Currently, there is one other panel including microhaplotype loci for *P. vivax* that has been published (Kattenberg et al., 2022), and at least one other under development (Siegel et al., 2023). Most of the loci in the Kattenberg et al. (2022) panel are SNPs, with only a few amplicons with multiple variants, all of which are putative drug resistance genes. The panel described in Siegel et al. (2023) does have 100 microhaplotype loci selected for high diversity, but at the time of writing only *in silico* panel design results have been published - the panel has not been applied to actual field samples. Once this panel, and perhaps others, is published with accompanying laboratory protocols, it will be possible to conduct a full comparison of panel performance for different research questions and in different geographic settings.

This study has certain limitations. First, it would be helpful to perform an evaluation with a second, larger set of field samples to provide a case study of the panel's utility in genomic epidemiology with "real" microhaplotypes obtained with amplicon deep sequencing, rather than those reconstructed from WGS data. This is particularly important as we were forced to filter all polyclonal samples out of the WGS dataset, due to the unreliability of phasing haplotypes from this data format. It could also be useful to perform simulations to understand the panel's effectiveness in a broad range of transmission environments. Once more *P. vivax* microhaplotype marker panels are published, it is advisable to perform these analyses for this and other available panels, as suggested above. Finally, the relatedness patterns observed in the WGS-based microhaplotypes may be confounded by the fact that the data was aggregated from multiple contributing studies, each with its own sample collection strategy. This could be particularly true in Southeast Asia, where the Ho Chi Min and Oddar Meanchey samples came from one study (1128-PV-MULTI-GSK) and Binh Phuoc, Dak O, and Krong Pa came from another (1157-PV-MULTI-PRICE; MalariaGEN et al., 2022).

The panel described in this paper is the only one of its kind for *P. vivax*: a large panel of microhaplotype loci selected for high diversity, ideal for genomic epidemiology. As described in the Background section, such a panel constitutes an important step forward. Microhaplotype panels offer considerably more discriminatory power than SNP panels of the same size (Baetscher et al., 2018), while costing substantially less than whole genome analyses (Tessema et al., 2022). In addition, they provide an enhanced ability to not only identify polyclonal infections but to accurately estimate the actual MOI of a sample, which may be an important metric for understanding *vivax* epidemiology in some regions. Given the growing need for cost-effective yet powerful tools to measure the complexities of *P. vivax* malaria epidemiology, this panel has the potential to dramatically enhance the surveillance of *vivax* malaria and thus accelerate progress towards elimination.

3.5 References

Agapow, P.-M., & Burt, A. (2001). Indices of multilocus linkage disequilibrium. *Molecular Ecology Notes*, *1*(1-2), 101–102. https://doi.org/10.1046/j.1471-8278.2000.00014.x

Anstey, N. M., Douglas, N. M., Poespoprodjo, J. R., & Price, R. N. (2012). Plasmodium vivax: Clinical spectrum, risk factors and pathogenesis. *Advances in Parasitology* (pp. 151–201).

Auburn, S., Böhme, U., Steinbiss, S., Trimarsanto, H., Hostetler, J., Sanders, M., Gao, Q., Nosten, F., Newbold, C. I., Berriman, M., Price, R. N., & Otto, T. D. (2016). A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. *Wellcome Open Research*, *1*, 4. https://doi.org/10.12688/wellcomeopenres.9876.1

Auburn, S., Campino, S., Miotto, O., Djimde, A. A., Zongo, I., Manske, M., Maslen, G., Mangano, V., Alcock, D., MacInnis, B., Rockett, K. A., Clark, T. G., Doumbo, O. K., Ouédraogo, J. B., & Kwiatkowski, D. P. (2012). Characterization of within-host Plasmodium falciparum diversity using next-generation sequence data. *PloS One*, *7*(2), e32891. https://doi.org/10.1371/journal.pone.0032891

Baetscher, D. S., Clemento, A. J., Ng, T. C., Anderson, E. C., & Garza, J. C. (2018). Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. *Molecular Ecology Resources*, *18*(2), 296–305. https://doi.org/10.1111/1755-0998.12737

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Bereczky, S., Mårtensson, A., Gil, J. P., & Färnert, A. (2005). Short report: Rapid DNA extraction from archive blood spots on filter paper for genotyping of Plasmodium falciparum. *The American Journal of Tropical Medicine and Hygiene*, *72*(3), 249–251. https://doi.org/10.4269/ajtmh.2005.72.249

Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C., & Hochreiter, S. (2015). Msa: An R package for multiple sequence alignment. *Bioinformatics (Oxford, England)*, *31*(24), 3997–3999. https://doi.org/10.1093/bioinformatics/btv494

Cowell, A. N., Loy, D. E., Sundararaman, S. A., Valdivia, H., Fisch, K., Lescano, A. G., Baldeviano, G. C., Durand, S., Gerbasi, V., Sutherland, C. J., Nolder, D., Vinetz, J. M., Hahn, B. H., & Winzeler, E. A. (2017). Selective Whole-Genome Amplification Is a Robust Method That Enables Scalable Whole-Genome Sequencing of Plasmodium vivax from Unprocessed Clinical Samples. *mBio*, *8*(1), e02257–16. https://doi.org/10.1128/mBio.02257-16

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Gerlovina, I., Gerlovin, B., Rodríguez-Barraquer, I., & Greenhouse, B. (2022). Dcifer: An IBD-based method to calculate genetic distance between polyclonal infections. *Genetics*, *222*(2). https://doi.org/10.1093/genetics/iyac126

Hathaway, N. J., Parobek, C. M., Juliano, J. J., & Bailey, J. A. (2018). SeekDeep: Single-base resolution de novo clustering for amplicon deep sequencing. *Nucleic Acids Research*, *46*(4). https://doi.org/10.1093/nar/gkx1201

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, *2*, e281. https://doi.org/10.7717/peerj.281

Kattenberg, J. H., Nguyen, H. V., Nguyen, H. L., Sauve, E., Nguyen, N. T. H., Chopo-Pizarro, A., Trimarsanto, H., Monsieurs, P., Guetens, P., Nguyen, X. X., Esbroeck, M. V., Auburn, S., Nguyen, B. T. H., & Rosanas-Urgell, A. (2022). Novel highly-multiplexed AmpliSeq targeted assay for Plasmodium vivax genetic surveillance use cases at multiple geographical scales. *Frontiers in Cellular and Infection Microbiology*, *12*, 953187. https://doi.org/10.3389/fcimb.2022.953187

Koepfli, C., & Mueller, I. (2017). Malaria Epidemiology at the Clone Level. *Trends in Parasitology*, *33*(12), 974–985. https://doi.org/10.1016/j.pt.2017.08.013

LaVerriere, E., Schwabl, P., Carrasquilla, M., Taylor, A. R., Johnson, Z. M., Shieh, M., Panchal, R., Straub, T. J., Kuzma, R., Watson, S., Buckee, C. O., Andrade, C. M., Portugal, S., Crompton, P. D., Traore, B., Rayner, J. C., Corredor, V., James, K., Cox, H., … Neafsey, D. E. (2022). Design and implementation of multiplexed amplicon sequencing panels to serve genomic epidemiology of infectious disease: A malaria case study. *Molecular Ecology Resources*, *22*(6), 2285–2303. https://doi.org/10.1111/1755-0998.13622

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

MalariaGEN, Adam, I., Alam, M. S., Alemu, S., Amaratunga, C., Amato, R., Andrianaranjaka, V., Anstey, N. M., Aseffa, A., Ashley, E., Assefa, A., Auburn, S., Barber, B. E., Barry, A., Batista Pereira, D., Cao, J., Chau, N. H., Chotivanich, K., Chu, C., … Yilma, D. (2022). An open dataset of Plasmodium vivax genome variation in 1,895 worldwide samples. *Wellcome Open Research*, *7*, 136. https://doi.org/10.12688/wellcomeopenres.17795.1

Neafsey, D. E., Taylor, A. R., & MacInnis, B. L. (2021). Advances and opportunities in malaria population genomics. *Nature Reviews. Genetics*, (22), 502–517. https://doi.org/10.1038/s41576-021-00349-5

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, *89*(3), 583–590. https://doi.org/10.1093/genetics/89.3.583

Paradis, E. (2010). Pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics (Oxford, England)*, *26*(3), 419–420. https://doi.org/10.1093/bioinformatics/btp696

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient Swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, *31*(7), 1929–1936. https://doi.org/10.1093/molbev/msu136

Picard toolkit. (2019). *Broad Institute*. https://broadinstitute.github.io/picard/

Siegel, S. V., Amato, R., Trimarsanto, H., Sutanto, E., Kleinecke, M., Murie, K., Whitton, G., Taylor, A. R., Watson, J. A., Imwong, M., Assefa, A., Rahim, A. G., Chau, N. H., Hien, T. T., Green, J. A., Koh, G., White, N. J., Day, N., Kwiatkowski, D. P., ... Auburn, S. (2023, March 16). *Lineage-informative microhaplotypes for spatio-temporal surveillance of Plasmodium vivax malaria parasites*. 36993192. https://doi.org/10.1101/2023.03.13.23287179

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595. https://doi.org/10.1093/genetics/123.3.585

Taylor, A. R., Schaffner, S. F., Cerqueira, G. C., Nkhoma, S. C., Anderson, T. J. C., Sriprawat, K., Phyo, A. P., Nosten, F., Neafsey, D. E., & Buckee, C. O. (2017). Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent. *PLOS Genetics*, *13*(10), e1007065. https://doi.org/10.1371/journal.pgen.1007065

Tessema, S. K., Hathaway, N. J., Teyssier, N. B., Murphy, M., Chen, A., Aydemir, O., Duarte, E. M., Simone, W., Colborn, J., Saute, F., Crawford, E., Aide, P., Bailey, J. A., & Greenhouse, B. (2022). Sensitive, Highly Multiplexed Sequencing of Microhaplotypes From the Plasmodium falciparum Heterozygome. *The Journal of Infectious Diseases*, *225*(7), 1227–1237. https://doi.org/10.1093/infdis/jiaa527

Twohig, K. A., Pfeffer, D. A., Baird, J. K., Price, R. N., Zimmerman, P. A., Hay, S. I., Gething, P. W., Battle, K. E., & Howes, R. E. (2019). Growing evidence of Plasmodium vivax across malaria-endemic Africa. *PLoS neglected tropical diseases*, *13*(1), e0007140. https://doi.org/10.1371/journal.pntd.0007140

Van der Auwera, G., & O'Connor, B. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media.

CHAPTER 4: *Plasmodium falciparum* genomic epidemiology in Africa with microhaplotypes

Article Authors: Alfred Hubbard, Cheikh C. Dieng, Elizabeth Hemming-Schroeder, Daniel Janies, Eugenia Lo

4.1 Background

Understanding patterns of transmission is critical to achieving timely malaria eradication. This is true because financial and material resources must be deployed efficiently and because measuring changes in transmission is essential to understanding whether the actions of national malaria control programs are having the desired effect (Feachem et al., 2019). In Africa, where the primary cause of malaria is *Plasmodium falciparum*, this is especially true, because most malaria programs lack the funds to apply control measures everywhere they are needed.

It is becoming increasingly common to use genetic relatedness as a proxy for malaria transmission. This approach is particularly valuable in vector-borne diseases, such as malaria, due to the difficulty of measuring transmission directly (Neafsey et al., 2021). In low transmission environments, high genetic relatedness may be used to infer infections that are close to each other in the (unobserved) transmission chain, whereas in higher transmission environments overall genetic diversity (both within and between hosts) can be a proxy for the level of transmission (Wesolowski et al., 2018). At larger scales, relatedness can show the level of genetic connectivity, and therefore transmission, between different geographies (e.g., Taylor et al., 2020).

Previous work has found a divide between *P. falciparum* populations in East and West Africa (Amambua-Ngwa et al., 2019; Tonkin-Hill et al., 2021; Verity et al., 2020), with additional substructuring within East African populations (Amambua-Ngwa et al., 2019). This is believed to be the historical background established by ancient population movements (Amambua-Ngwa et al., 2019). Selection for drug resistance and transmission reductions from measures deployed by control programs may now be altering this structure (Verity et al., 2020), and it is expected that populations will become increasingly fragmented as control measures achieve their desired effect.

However, in high transmission environments, which includes many African countries, population structure tends to be weak and genetic relatedness generally is quite low (Wesolowski et al., 2018). In such conditions, achieving sufficient discriminatory power to identify patterns of relatedness, and therefore transmission, is challenging. SNP (single-nucleotide polymorphism) panels have been shown to be ineffective in such environments (Argyropoulos et al., 2023), and so more powerful alternatives are required. One option is whole genome sequence (WGS) data, which is effective, but costly (Noviyanti et al., 2020). Also, it is limited in its ability to accurately estimate the number of parasite strains present in a host (also known as multiplicity of infection, or MOI), which can be a valuable measure of transmission in endemic contexts (Camponovo et al., 2023).

Recent advancements in sequencing technology have led to a new generation of microhaplotype markers, genomic regions with multiple, linked SNPs that can be sequenced together in a single amplicon. These panels of microhaplotype loci offer superior discriminatory power compared to SNP panels of similar size (Baetscher et al., 2018), while costing considerably less than WGS. Moreover, the amplicon deep sequencing techniques employed for genotyping these markers yield sufficient read depth for accurate estimates of MOI (LaVerriere et al., 2022).

In this study, we successfully genotype 180 samples from eight countries in Africa using the microhaplotype marker panel described in LaVerriere et al. (2022). We analyze the patterns of genetic differentiation revealed by this data for insights into malaria transmission in the region, drawing comparisons to previous studies performed with other genotyping methods. Finally, we discuss the additional benefits of this genotyping technique over its predecessors.

4.2 Materials and Methods

4.2.1 Sample collection and molecular screening

Between 2016 and 2017, a cross-sectional study was conducted in hospitals and health facilities across Nigeria, Sudan, Botswana, Ethiopia, Kenya, and Cameroon. Additionally, samples from Senegal were collected in 2019. Locality names and coordinates are given in Table 4.1. The

Table 4.1: Locality names and coordinates.

Locality	Country	Lat.	Long.
Casamance	Senegal	12.83	-15.81
Seikwa	Ghana	7.723	-2.517
South Nigeria	Nigeria	7.229	6.557
Dschang	Cameroon	5.448	10.057
Gaborone	Botswana	-24.638	25.905
Khartoum	Sudan	15.561	32.548
Homa Bay	Kenya	-0.531	34.459
Asosa	Ethiopia	10.066	34.548

distribution of samples was as follows: 50 from Botswana, 34 from Nigeria, 50 from Senegal, 50 from Sudan, 29 from Cameroon, 99 from Ghana, 72 from Ethiopia, and 41 from Kenya. In total, 425 dried blood spots (DBS) were collected from suspected *P. falciparum* patients across these locations, which vary substantially in transmission intensity (Figure 4.1).

Diagnosis of *P. falciparum* was conducted through microscopic examination of Giemsa-stained thin and thick blood films and/or rapid diagnosis test (SD Bioline, Standard Diagnostics Inc., South Korea). Demographic and clinical data, including age, gender, ethnicity, and medical history, were recorded via questionnaire. Patients infected with other *Plasmodium* species (*P. vivax*, *P. malariae*, *P. ovale*, and/or mixed infections) were excluded from the study. Parasite DNA was extracted from dried blood spots by the Saponin-Chelex method. The final extracted volume was 200μl and parasite DNA amount was estimated using the SYBR Green qPCR detection method with species-specific primers that targeted the 18S rRNA genes. Amplification was conducted in a 20μL reaction mixture containing 2μL of genomic DNA, 10μL 2xSYBR Green qPCR Master Mix (Thermo Scientific, USA), and 0.5μM primer. Reaction was performed in CFX96 TouchTM Real-Time PCR Detection System (Bio-Rad), with an initial denaturation at 95°C for 3 min, followed by 45 cycles at 94°C for 30 sec, 55°C for 30 sec, and 68°C for 1 min with a final 95°C for 10 sec. This was then followed by a melting curve step of temperature that ranged from 65°C to 95°C with 0.5°C increment to determine the melting temperature of each amplified product. Melting curve analyses were performed for each amplified sample to confirm specific amplifications of

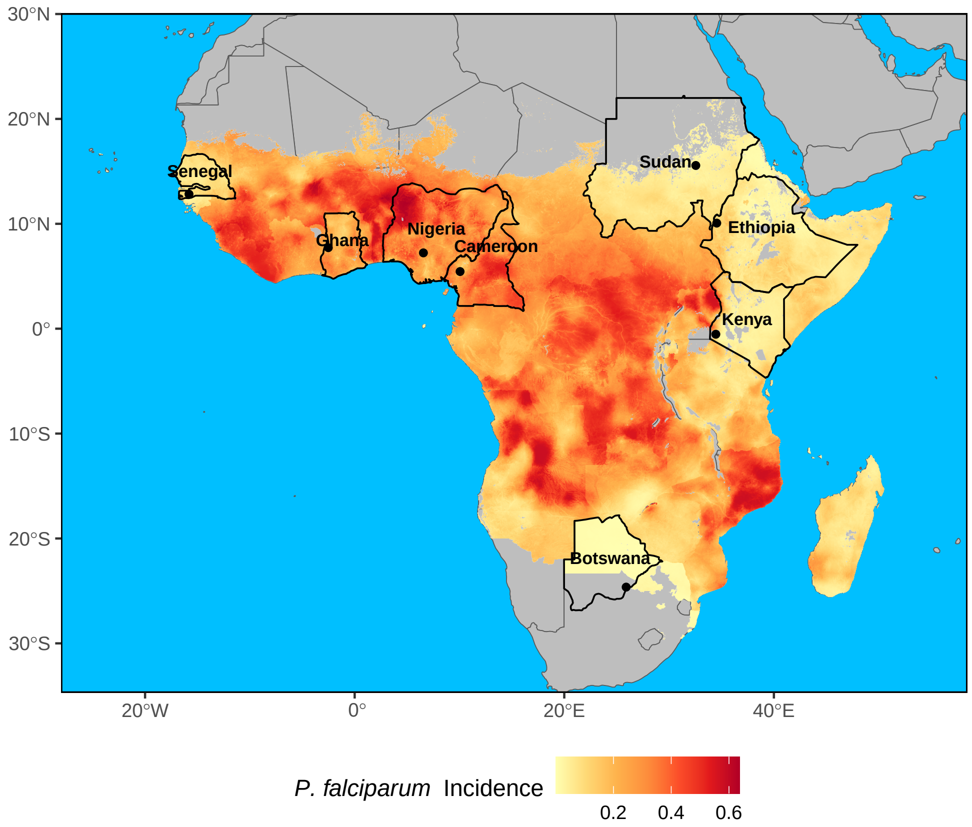


Figure 4.1: The eight countries from which samples were obtained. Modeled incidence of *P. falciparum* (Weiss et al., 2019) is displayed in the background for context.

the target sequence. A cut-off threshold of 0.02 fluorescence units that robustly represented the threshold cycle at the log-linear phase of the amplification and above the background noise was set to determine *Ct* value for each assay. Samples yielding *Ct* values higher than 40 (as indicated in the negative controls) were considered negative for *Plasmodium* species.

Sequencing libraries were prepared for samples ascertained to be positive for *P. falciparum*, using the panel and protocol described in LaVerriere et al. (2022) with the addition of three primer pairs corresponding to the second, third, and fourth domains of the kelch13 (K13) protein. Selective whole genome amplification (SWGA) was performed according to the protocol described in Oyola et al. (2016). Libraries were sequenced on an Illumina MiSeq in paired-end mode.

4.2.2 Amplicon analysis

The de-multiplexed reads were processed with the amplicon pipeline published with LaVerriere et al. (2022), which uses the Divisive Amplicon Denoising Algorithm (`DADA2`) software (Callahan et al., 2016) to extract microhaplotypes. The same parameters and settings described by LaVerriere et al. (2022) were used.

4.2.3 Population genetics

Genetic diversity was estimated for each locus with Nei's expected heterozygosity (*He*), using the `poppr` R package (Kamvar et al., 2014). MOI was estimated for each sample as the maximum number of unique haplotypes found across all loci. Both metrics were calculated separately for each country, and paired *t*-tests were performed between each pair of countries to identify significant differences. The Bonferroni correction for multiple hypothesis testing was applied to both sets of tests.

Selection was assessed for each locus with Tajima's *D* (Tajima, 1989), estimated with the `pegas` R package (Paradis, 2010). Samples were divided into populations for this analysis (East Africa, Southern Africa, and West Africa), and selection was tested separately in each population. Note that Cameroon was grouped with West Africa, rather than treated as a separate, Central Africa population, because A) this is consistent with the MalariaGEN population definitions (MalariaGEN et al., 2023) and B) only 21 samples were successfully genotyped for Cameroon. The Bonferroni correction for multiple hypothesis testing was applied to the *p*-values from the Tajima tests. Loci under significant selection were removed prior to downstream population genetics analysis.

Linkage disequilibrium (LD) was estimated among all pairs of loci from the same chromosome with $\bar{r}_d$ (Agapow & Burt, 2001). This was accomplished with the `poppr` R package (Kamvar et al., 2014). Once again, the Bonferroni correction for multiple hypothesis testing was used, and loci in significant LD with other loci were removed.

A principle component analysis (PCA) was performed on a binary matrix of the genotype data, where rows are samples and columns are haplotypes. PCA was performed with the R programming language (R Core Team, 2021).

Finally, genetic relatedness was assessed using `Dcifer`, a tool that estimates relatedness between samples using allele frequencies and the concept of identity-by-descent (Gerlovina et al., 2022). The population-level allele frequencies and estimated MOI for each individual are used to estimate whether observed sharing of genotypes between sample pairs is because of sharing in the most recent common ancestor (in which case they are said to be identical-by-descent) or due to chance alone. MOI was calculated as described above.

The sample-level relatedness estimates from `Dcifer` were aggregated to the country-level by computing the fraction of constituent pairs that are highly related. A relatedness threshold of 0.25 was used to define highly-related pairs, which corresponds to the half-sibling level of relatedness in humans (the biological interpretation is less straightforward in malaria, as it undergoes both sexual and asexual reproduction). The fraction of highly-related infections is expected to be a better measure of recent transmission, as opposed to the historical background, than mean relatedness (Taylor et al., 2017).

4.3 Results

4.3.1 Sequencing results

Inspection of the read counts obtained for each locus and sample revealed that some samples and loci did not amplify well (Appendix 4A). In addition, *P. falciparum* DNA corresponding to five panel markers was found in the negative controls. These five loci were removed for subsequent analyses, as were samples with fewer than 10 loci and loci with fewer than 20 samples. After these filters were applied, there were 184 samples and 43 loci remaining. Each country still has at least 12 samples in this reduced dataset.

4.3.2 Population genetics

No pairs of loci were found to be in significant LD after Bonferroni correction of the p=0.05 significance threshold. However, after Bonferroni correction of the Tajima test results, two markers were found to have negative values of Tajima's D at the 0.05 significance level: *PF11_0373*, in East Africa, and *PF3D7_0401900*, in all three populations. These loci were filtered out of the dataset prior to performing downstream analyses.

Genetic diversity is moderately high in all countries (Figure 4.2), and none of the countries have significantly different diversity after Bonferroni correction of the p=0.05 significance threshold. MOI is also fairly high across the board (Figure 4.3). Strikingly, not a single monoclonal infection was present in the dataset. Once again, none of the countries have significant differences in MOI once the p=0.05 threshold was Bonferroni corrected.

The first component of the PCA suggests the presence of three distinct genetic clusters, but these do not appear to have any correspondence with geography (Figure 4.4).

Relatedness between sample pairs was fairly low, with most pairs having a relatedness of zero (Figure 4.5A), and most of the remainder with a relatedness below 0.25, the half-sibling level (Figure 4.5B). A few sample pairs appear to be clones, as evidenced by their relatedness of one (Figure 4.5B). At the country level, the fraction of highly-related sample pairs appears to be somewhat elevated among Ghana, Nigeria, and Cameroon, relative to other country pairs, and Kenya and Sudan, relative to other country pairs. Senegal, Botswana, and Ethiopia appear to be comparatively isolated from the other countries (Figures 4.6 and 4.7). Within-country relatedness did not substantially exceed between-country relatedness in most cases, indicating a high degree of mixing and gene flow between countries.

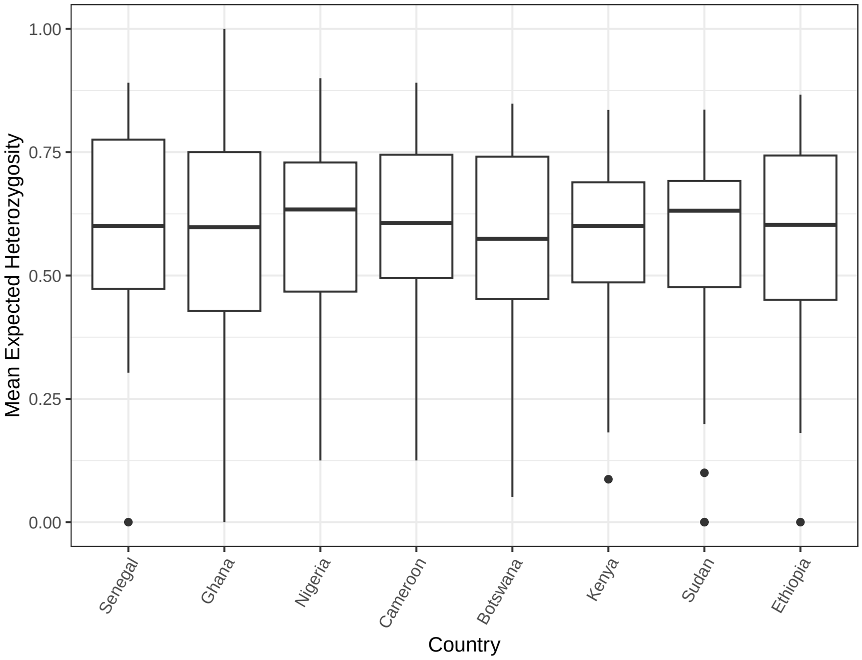


Figure 4.2: Box plots showing the mean expected heterozygosities across all samples for each marker in the panel. Each box corresponds to the samples from one country.

4.4 Discussion and Conclusion

The results of this study point to two main conclusions. First, the observations of high diversity and low relatedness across all countries indicate that *P. falciparum* populations in Africa are relatively panmictic. There are some nuances to this general pattern, which are discussed more below, but overall little population structure was found. Second, the lack of significant differences in both genetic diversity and MOI between countries, despite notable differences in falciparum malaria incidence, call into question the utility of these statistics as measures of malaria transmission. These two conclusions suggest that additional work clarifying the added value of microhaplotype marker panels in genomic epidemiology would be useful to control programs confronted with the decision of whether to employ such panels in routine surveillance.

Previous population genetics studies of *P. falciparum* in Africa have detected modest genetic differentiation between East, Central, and West Africa (Amambua-Ngwa et al., 2019; Tonkin-Hill et al., 2021; Verity et al., 2020). Our relatedness results are largely consistent with this picture, as we measured lower relatedness between populations from East and West Africa than within

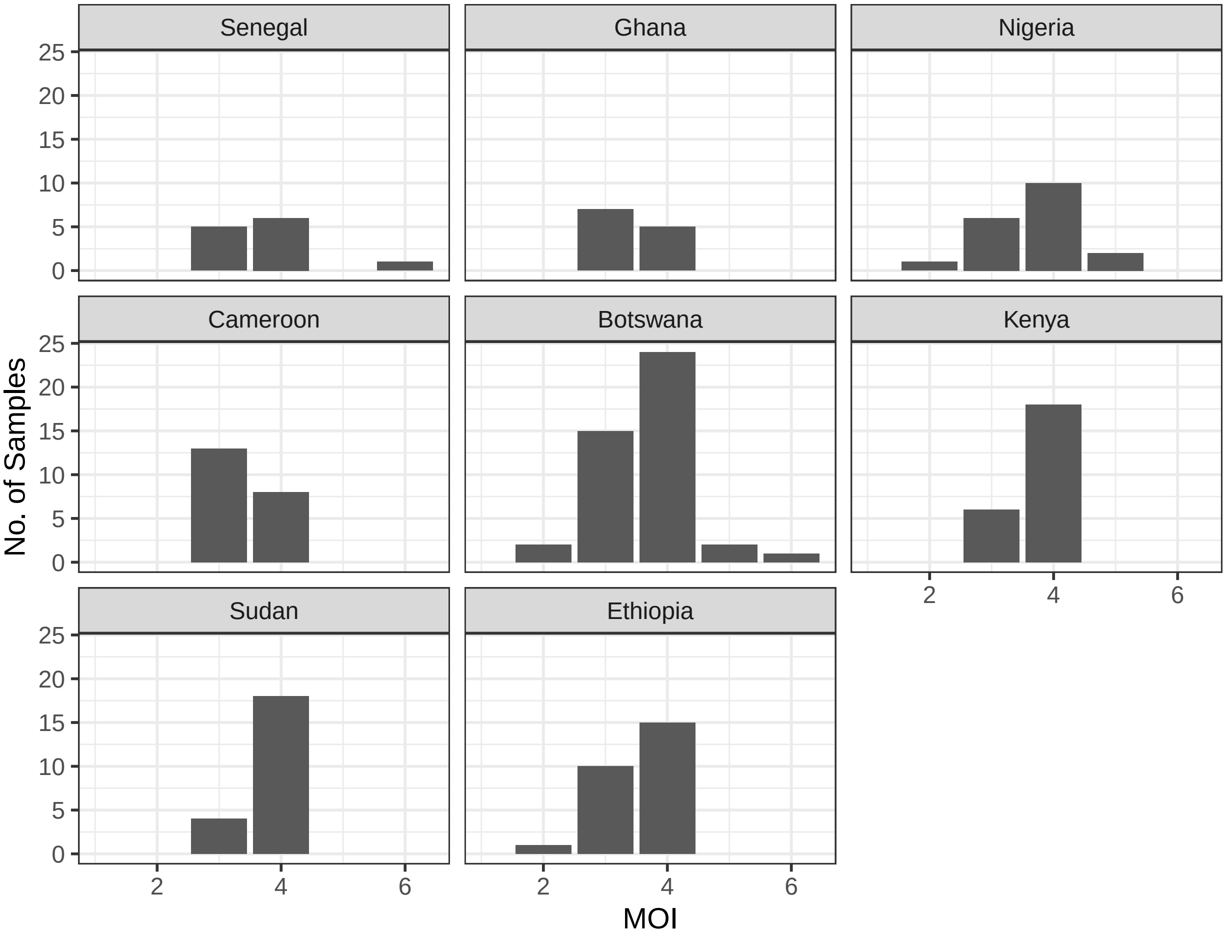


Figure 4.3: Bar plots displaying the distribution of MOI values for samples from each country.

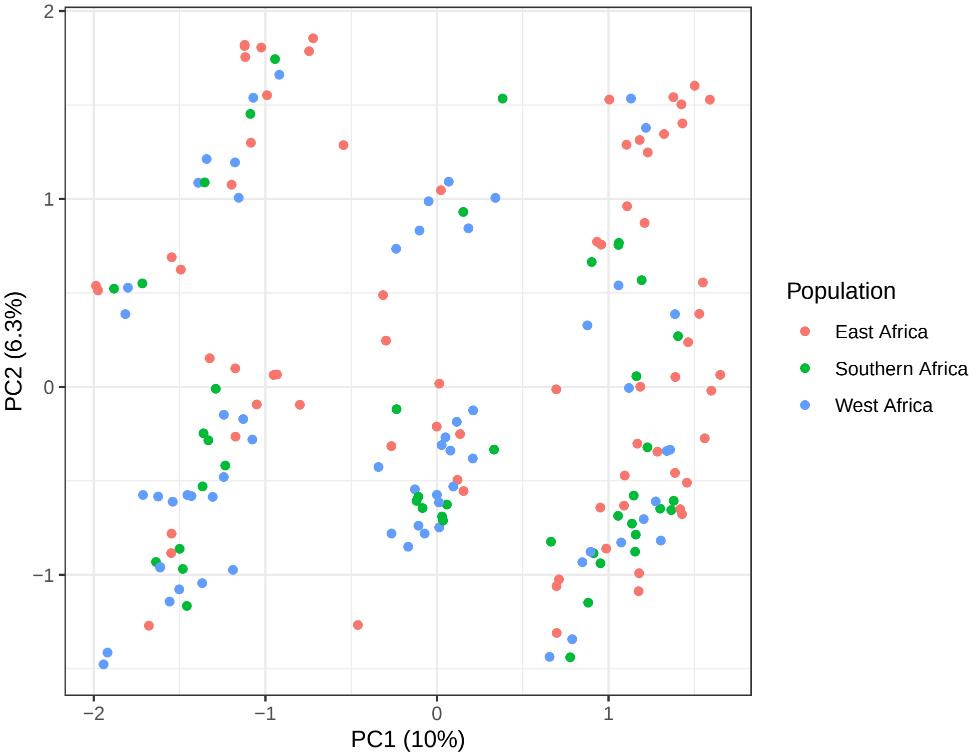


Figure 4.4: Scatterplot showing the distribution of samples in the first two principal components, with points colored according to population. The proportion of total variance explained by each component is noted in the axis label. Note that, as with the selection analysis, Cameroon was treated as part of the West Africa population - see the Methods section for justification.

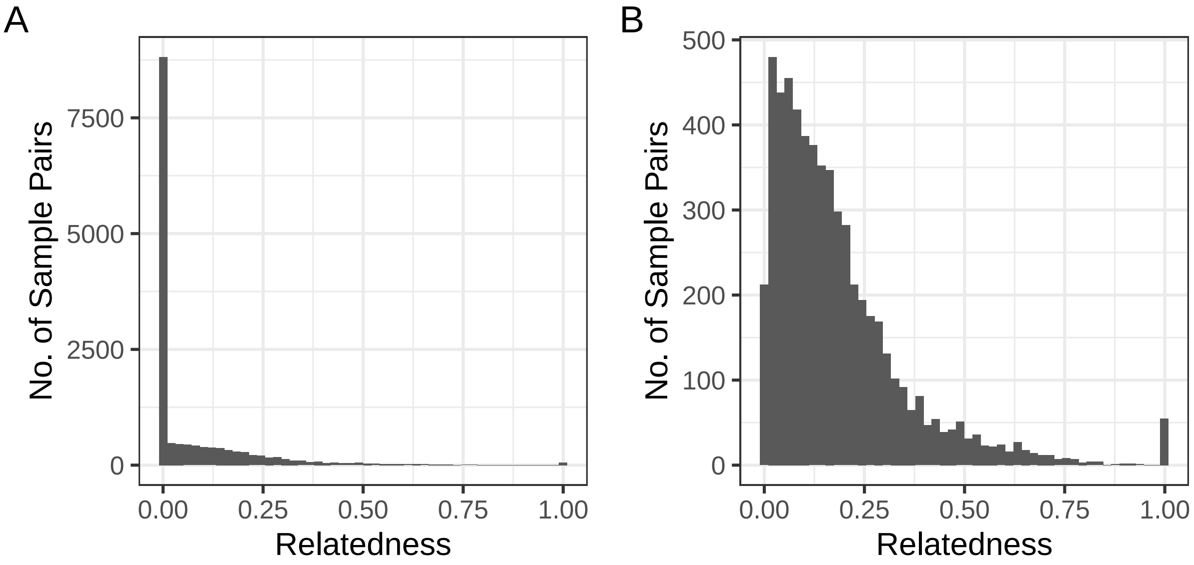


Figure 4.5: Histograms showing the distribution of relatedness values for **(A)** all sample pairs and **(B)** the sample pairs with relatedness greater than 0.

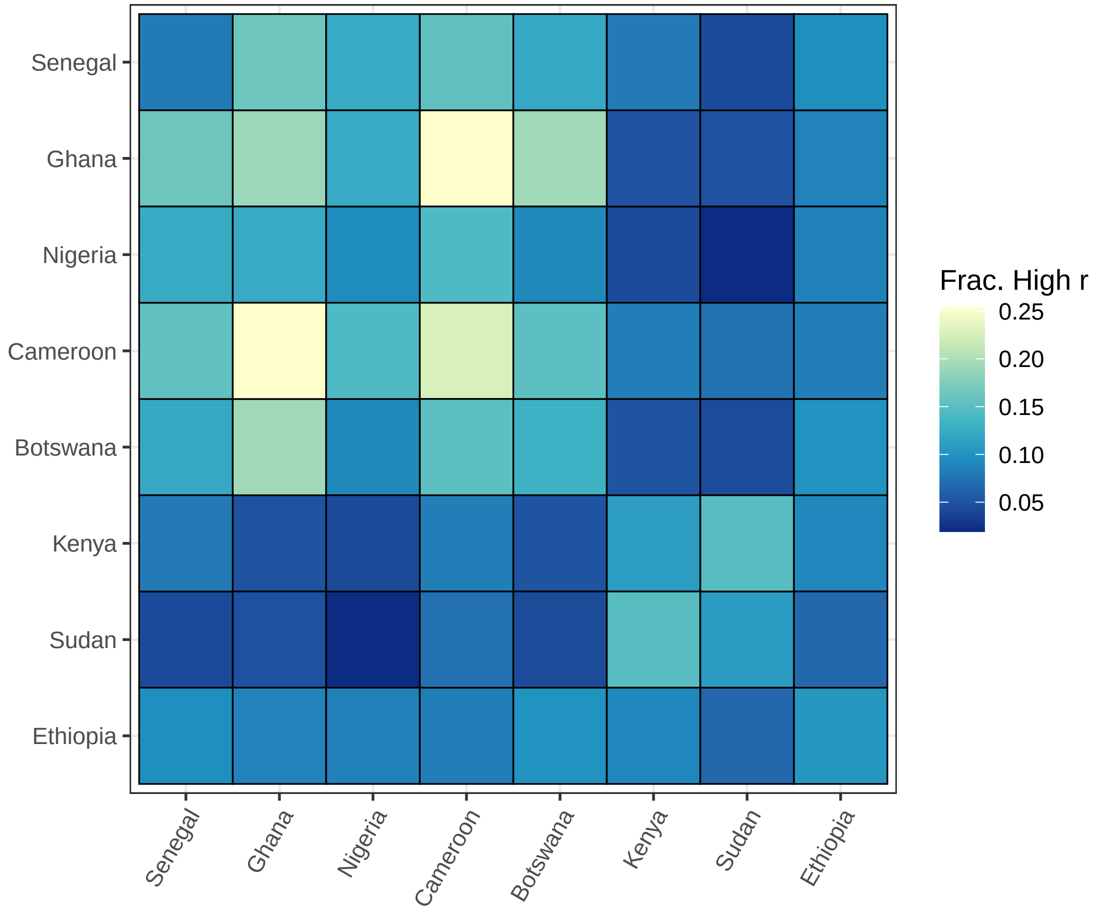


Figure 4.6: Heatmap displaying the fraction of highly related sample pairs corresponding to each pair of countries. Color swatches along the diagonal indicate within-country relatedness.

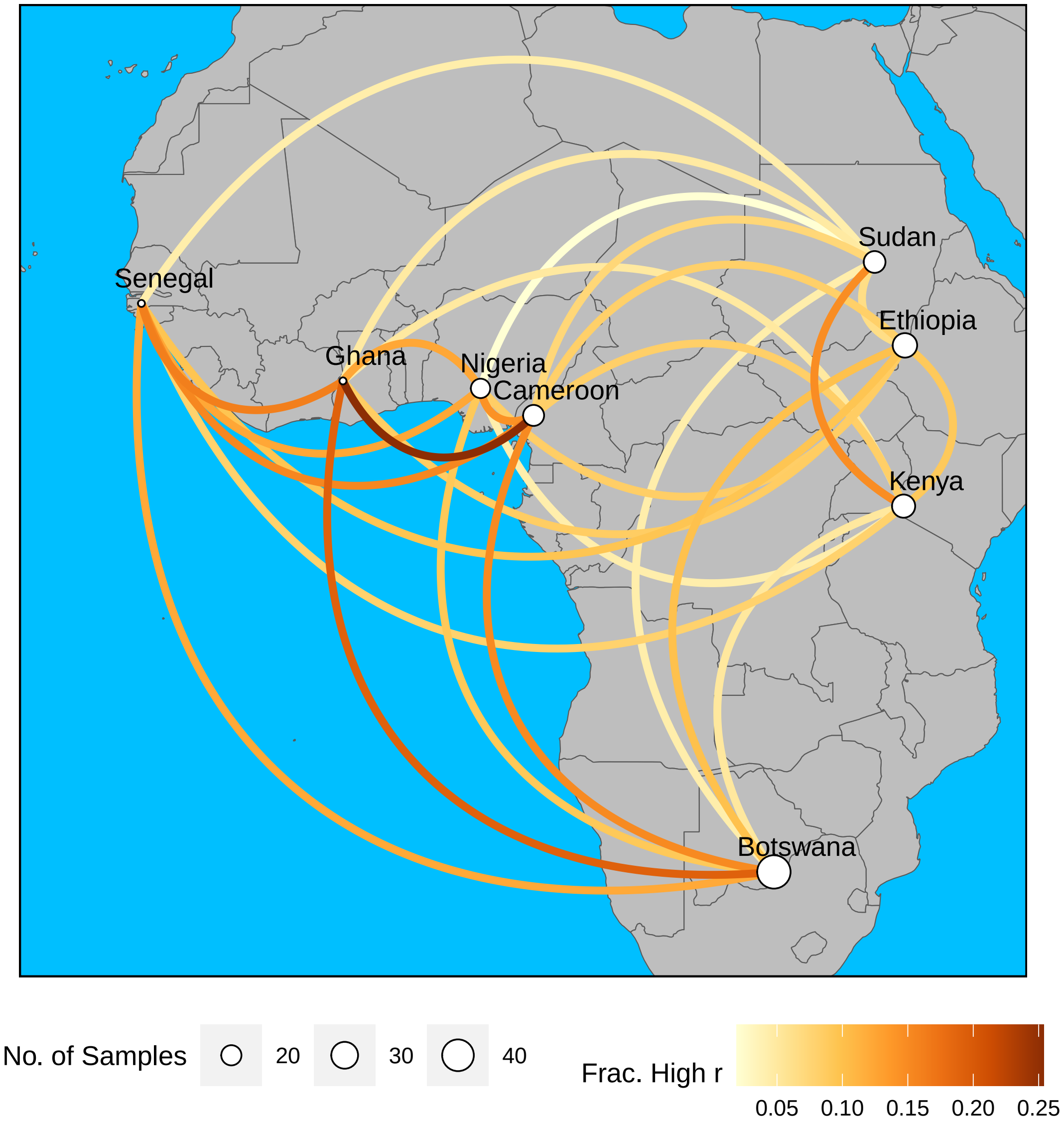


Figure 4.7: Network with edge colors scaled according to the fraction of highly related sample pairs shared between each pair of countries. Node size is scaled according to the number of samples from each country.

those regions. Also, Amambua-Ngwa et al. (2019) described Ethiopia as a distinct subpopulation, relative even to other countries in East Africa. They attribute this separation to landscape and vector factors, rather than differences in human populations and/or movements. This is once again consistent with our relatedness results, as Ethiopia has a notably lower relatedness with Kenya than Sudan has with Kenya.

However, the genetic differentiation identified in these studies was relatively small, especially compared to the distinctions between populations of *P. falciparum* from different continents (MalariaGEN et al., 2023). This may explain why our principal component analysis did not identify the same east-west gradient of genetic differentiation. This, coupled with the fact that relatedness is generally low and genetic diversity is generally high, indicates that despite some regional differences African populations of *P. falciparum* are relatively panmictic.

Our second main conclusion, that MOI and diversity are not highly-informative metrics of transmission in Africa, is also supported by previous work. Koepfli and Mueller (2017) reviewed the literature on malaria transmission metrics and concluded that while there is some relationship between diversity and transmission, changes in diversity can be "minimal despite substantial differences in transmission intensity." Similarly, prior investigations of the relationship between MOI and prevalence have only found weak correlations (Lopez & Koepfli, 2021; Paschalidis et al., 2023). However, there has yet to be equally rigorous work on the relationship between MOI and transmission, as measured by the EIR. Also, it has been argued that the real value of MOI is that it can detect changes in transmission intensity over short periods of time (Camponovo et al., 2023), and this idea has been supported with simulation results (Watson et al., 2021). Longitudinal studies that assess the relationship between within-host diversity metrics, such as MOI, and malaria transmission, rather than prevalence, would help clarify this claim. In the meantime, the evidence presented here and elsewhere suggests that both genetic diversity and MOI have some relationship to underlying malaria transmission dynamics, but the correlation may be low and the interpretation of these metrics from a control perspective is therefore unclear.

These results have implications for future applications of microhaplotype panels, such as the one employed in this study. While these panels undeniably provide richer, more detailed information on the minority clones present in an infection than other genotyping technologies, our results suggest the utility of this information is situational. We did not identify patterns of genetic differentiation that had not already been described with more conventional methods, and the MOI estimated from our data does not have obvious bearing on the local transmission environment. However, this is the first study applying such a panel at this scale. Additional investigation into the added value of this genotyping technology in different transmission environments and for different research questions will help clarify its utility.

This work has certain limitations. The most prominent of these is the negative impact that adding K13 primers to the LaVerriere et al. (2022) panel had on our sequencing yield. Redoing sequencing without these primers might substantially improve yield and strengthen the power of our analyses. Second, MOI is essentially within-host richness, and it may not be the most informative or useful metric of within-host diversity relevant to microhaplotype data. An alternative that incorporates both the richness and the relatedness of the strains circulating within a host, $eMOI$, has been proposed (Murphy & Greenhouse, 2023). Once published, applying this metric to microhaplotype data such as that presented in this study might yield more useful insights into transmission dynamics. Also, the number of geographic locations included in this study is insufficient to perform a statistically rigorous evaluation of the relationship between diversity, relatedness, MOI, and incidence. Finally, the geographic locations that were included were selected based on existing collaborations and the feasibility of collecting data, rather than a statistically-based sample design. Follow-on studies with a larger number of geographic locations that were selected to be statistically-representative of a wide variety of transmission environments could shed further light on these questions.

We have performed the first application of the LaVerriere et al. (2022) amplicon panel at the continental scale. While the results overall point to a panmictic population, IBD-based relatedness analyses did permit us to identify more subtle geographic patterns that are consistent with previous work. More surprisingly, our results do not support the idea that diversity and MOI are informative metrics of transmission. This result has major implications for the genomic epidemiology field, and must be investigated in more detail before large investments are made in the surveillance of genetic metrics of transmission intensity. Therefore, while this study shows that microhaplotype markers are a promising tool for genomic epidemiology of *P. falciparum*, additional work is necessary to characterize the specific situations and questions best suited to their application.

4.5 References

Agapow, P.-M., & Burt, A. (2001). Indices of multilocus linkage disequilibrium. *Molecular Ecology Notes*, *1*(1-2), 101–102. https://doi.org/10.1046/j.1471-8278.2000.00014.x

Amambua-Ngwa, A., Amenga-Etego, L., Kamau, E., Amato, R., Ghansah, A., Golassa, L., Randrianarivelojosia, M., Ishengoma, D., Apinjoh, T., Maïga-Ascofaré, O., Andagalu, B., Yavo, W., Bouyou-Akotet, M., Kolapo, O., Mane, K., Worwui, A., Jeffries, D., Simpson, V., D'Alessandro, U., ... Djimde, A. A. (2019). Major subpopulations of Plasmodium falciparum in sub-Saharan Africa. *Science*, *365*(6455), 813–816. https://doi.org/10.1126/science.aav5427

Argyropoulos, D. C., Tan, M. H., Adobor, C., Mensah, B., Labbé, F., Tiedje, K. E., Koram, K. A., Ghansah, A., & Day, K. P. (2023). Performance of SNP barcodes to determine genetic diversity and population structure of Plasmodium falciparum in Africa. *Frontiers in Genetics*, *14*, 1071896. https://doi.org/10.3389/fgene.2023.1071896

Baetscher, D. S., Clemento, A. J., Ng, T. C., Anderson, E. C., & Garza, J. C. (2018). Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. *Molecular Ecology Resources*, *18*(2), 296–305. https://doi.org/10.1111/1755-0998.12737

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Camponovo, F., Buckee, C. O., & Taylor, A. R. (2023). Measurably recombining malaria parasites. *Trends in Parasitology*, *39*(1), 17–25. https://doi.org/10.1016/j.pt.2022.11.002

Feachem, R. G. A., Chen, I., Akbari, O., Bertozzi-Villa, A., Bhatt, S., Binka, F., Boni, M. F., Buckee, C., Dieleman, J., Dondorp, A., Eapen, A., Sekhri Feachem, N., Filler, S., Gething, P., Gosling, R., Haakenstad, A., Harvard, K., Hatefi, A., Jamison, D., ... Mpanju-Shumbusho, W. (2019). Malaria eradication within a generation: Ambitious, achievable, and necessary. *The Lancet*, *394*(10203), 1056–1112. https://doi.org/10.1016/S0140-6736(19)31139-0

Gerlovina, I., Gerlovin, B., Rodríguez-Barraquer, I., & Greenhouse, B. (2022). Dcifer: An IBD-based method to calculate genetic distance between polyclonal infections. *Genetics*, *222*(2). https://doi.org/10.1093/genetics/iyac126

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, *2*, e281. https://doi.org/10.7717/peerj.281

Koepfli, C., & Mueller, I. (2017). Malaria Epidemiology at the Clone Level. *Trends in Parasitology*, *33*(12), 974–985. https://doi.org/10.1016/j.pt.2017.08.013

LaVerriere, E., Schwabl, P., Carrasquilla, M., Taylor, A. R., Johnson, Z. M., Shieh, M., Panchal, R., Straub, T. J., Kuzma, R., Watson, S., Buckee, C. O., Andrade, C. M., Portugal, S., Crompton, P. D., Traore, B., Rayner, J. C., Corredor, V., James, K., Cox, H., … Neafsey, D. E. (2022). Design and implementation of multiplexed amplicon sequencing panels to serve genomic epidemiology of infectious disease: A malaria case study. *Molecular Ecology Resources*, *22*(6), 2285–2303. https://doi.org/10.1111/1755-0998.13622

Lopez, L., & Koepfli, C. (2021). Systematic review of Plasmodium falciparum and Plasmodium vivax polyclonal infections: Impact of prevalence, study population characteristics, and laboratory procedures. *PloS One*, *16*(6), e0249382. https://doi.org/10.1371/journal.pone.0249382

MalariaGEN, Abdel Hamid, M. M., Abdelraheem, M. H., Acheampong, D. O., Ahouidi, A., Ali, M., Almagro-Garcia, J., Amambua-Ngwa, A., Amaratunga, C., Amenga-Etego, L., Andagalu, B., Anderson, T., Andrianaranjaka, V., Aniebo, I., Aninagyei, E., Ansah, F., Ansah, P. O., Apinjoh, T., Arnaldo, P., … van der Pluijm, R. W. (2023). Pf7: An open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples. *Wellcome Open Research*, *8*, 22. https://doi.org/10.12688/wellcomeopenres.18681.1

Murphy, M., & Greenhouse, B. (2023, October 5). *MOIRE: A software package for the estimation of allele frequencies and effective multiplicity of infection from polyallelic data*. https://doi.org/10.1101/2023.10.03.560769

Neafsey, D. E., Taylor, A. R., & MacInnis, B. L. (2021). Advances and opportunities in malaria population genomics. *Nature Reviews. Genetics*, (22), 502–517. https://doi.org/10.1038/s41576-021-00349-5

Noviyanti, R., Miotto, O., Barry, A., Marfurt, J., Siegel, S., Thuy-Nhien, N., Quang, H. H., Anggraeni, N. D., Laihad, F., Liu, Y., Sumiwi, M. E., Trimarsanto, H., Coutrier, F., Fadila, N., Ghanchi, N., Johora, F. T., Puspitasari, A. M., Tavul, L., Trianty, L., ... Auburn, S. (2020). Implementing parasite genotyping into national surveillance frameworks: Feedback from control programmes and researchers in the Asia-Pacific region. *Malaria Journal*, *19*(1), 271. https://doi.org/10.1186/s12936-020-03330-5

Oyola, S. O., Ariani, C. V., Hamilton, W. L., Kekre, M., Amenga-Etego, L. N., Ghansah, A., Rutledge, G. G., Redmond, S., Manske, M., Jyothi, D., Jacob, C. G., Otto, T. D., Rockett, K., Newbold, C. I., Berriman, M., & Kwiatkowski, D. P. (2016). Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. *Malaria Journal*, *15*(1), 597. https://doi.org/10.1186/s12936-016-1641-7

Paradis, E. (2010). Pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics (Oxford, England)*, *26*(3), 419–420. https://doi.org/10.1093/bioinformatics/btp696

Paschalidis, A., Watson, O. J., Aydemir, O., Verity, R., & Bailey, J. A. (2023). Coiaf: Directly estimating complexity of infection with allele frequencies. *PLoS computational biology*, *19*(6), e1010247. https://doi.org/10.1371/journal.pcbi.1010247

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria. https://www.R-project.org/

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595. https://doi.org/10.1093/genetics/123.3.585

Taylor, A. R., Echeverry, D. F., Anderson, T. J. C., Neafsey, D. E., & Buckee, C. O. (2020). Identity-by-descent with uncertainty characterises connectivity of Plasmodium falciparum populations on the Colombian-Pacific coast. *PLoS genetics*, *16*(11), e1009101. https://doi.org/10.1371/journal.pgen.1009101

Taylor, A. R., Schaffner, S. F., Cerqueira, G. C., Nkhoma, S. C., Anderson, T. J. C., Sriprawat, K., Phyo, A. P., Nosten, F., Neafsey, D. E., & Buckee, C. O. (2017). Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent. *PLOS Genetics*, *13*(10), e1007065. https://doi.org/10.1371/journal.pgen.1007065

Tonkin-Hill, G., Ruybal-Pesántez, S., Tiedje, K. E., Rougeron, V., Duffy, M. F., Zakeri, S., Pumpaibool, T., Harnyuttanakorn, P., Branch, O. H., Ruiz-Mesía, L., Rask, T. S., Prugnolle, F., Papenfuss, A. T., Chan, Y.-B., & Day, K. P. (2021). Evolutionary analyses of the major variant surface antigen-encoding genes reveal population structure of Plasmodium falciparum within and between continents. *PLoS genetics*, *17*(2), e1009269. https://doi.org/10.1371/journal.pgen.1009269

Verity, R., Aydemir, O., Brazeau, N. F., Watson, O. J., Hathaway, N. J., Mwandagalirwa, M. K., Marsh, P. W., Thwai, K., Fulton, T., Denton, M., Morgan, A. P., Parr, J. B., Tumwebaze, P. K., Conrad, M., Rosenthal, P. J., Ishengoma, D. S., Ngondi, J., Gutman, J., Mulenga, M., … Juliano, J. J. (2020). The impact of antimalarial resistance on the genetic structure of Plasmodium falciparum in the DRC. *Nature Communications*, *11*(1), 2107. https://doi.org/10.1038/s41467-020-15779-8

Watson, O. J., Okell, L. C., Hellewell, J., Slater, H. C., Unwin, H. J. T., Omedo, I., Bejon, P., Snow, R. W., Noor, A. M., Rockett, K., Hubbart, C., Nankabirwa, J. I., Greenhouse, B., Chang, H.-H., Ghani, A. C., & Verity, R. (2021). Evaluating the Performance of Malaria Genetics for Inferring Changes in Transmission Intensity Using Transmission Modeling. *Molecular Biology and Evolution*, *38*(1), 274–289. https://doi.org/10.1093/molbev/msaa225

Weiss, D. J., Lucas, T. C. D., Nguyen, M., Nandi, A. K., Bisanzio, D., Battle, K. E., Cameron, E., Twohig, K. A., Pfeffer, D. A., Rozier, J. A., Gibson, H. S., Rao, P. C., Casey, D., Bertozzi-Villa, A., Collins, E. L., Dalrymple, U., Gray, N., Harris, J. R., Howes, R. E., … Gething, P. W. (2019). Mapping the global prevalence, incidence, and mortality of Plasmodium falciparum, 2000–17: A spatial and temporal modelling study. *The Lancet*, *394*(10195), 322–331. https://doi.org/10.1016/S0140-6736(19)31097-9

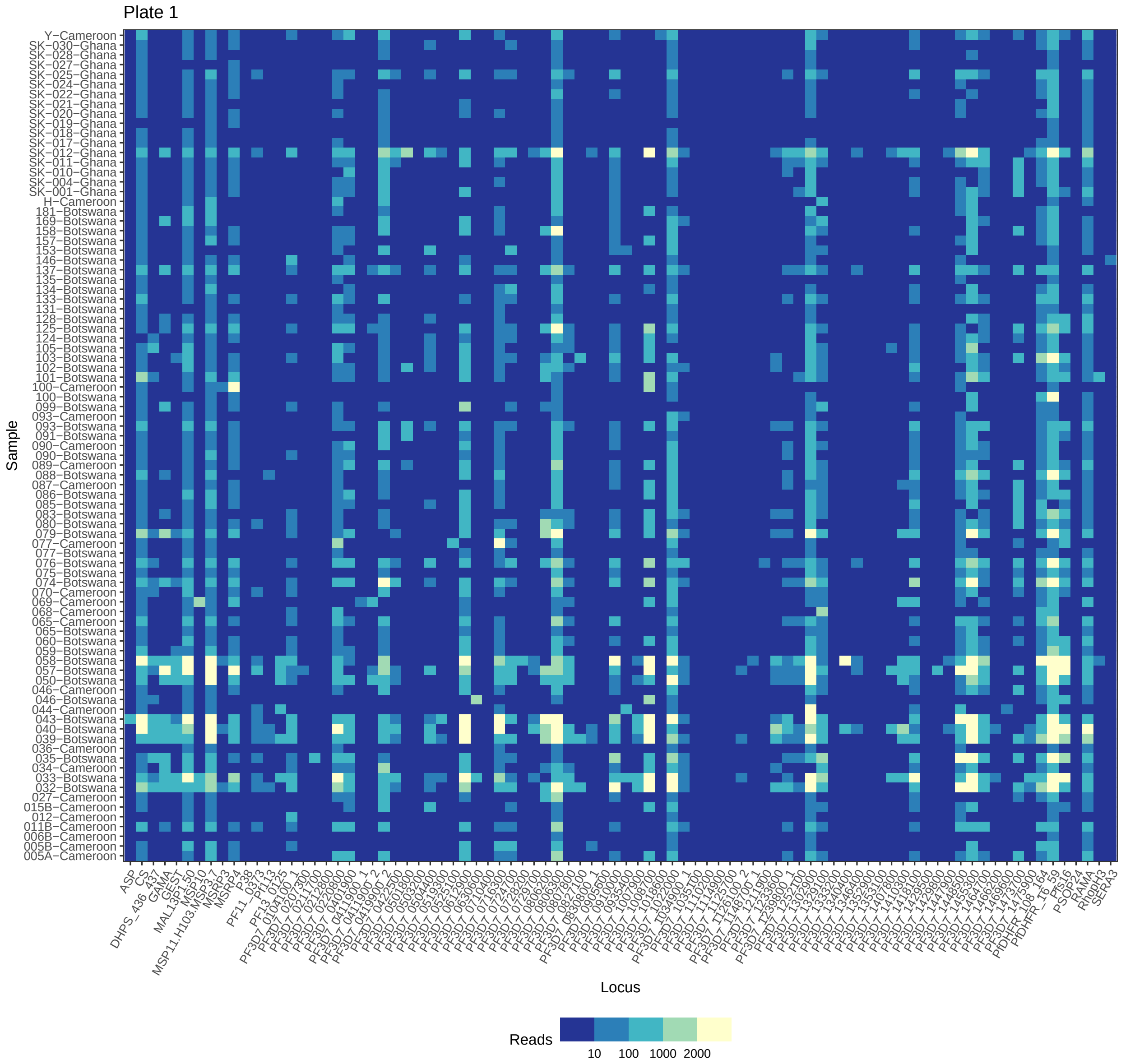
Wesolowski, A., Taylor, A. R., Chang, H.-H., Verity, R., Tessema, S., Bailey, J. A., Alex Perkins, T., Neafsey, D. E., Greenhouse, B., & Buckee, C. O. (2018). Mapping malaria by combining parasite genomic and epidemiologic data. *BMC medicine*, *16*(1), 190. https://doi.org/10.1186/s12916-018-1181-9

4.6 Appendices

4.6.1 Appendix 4A: Read counts for each sample and locus

The heatmaps below show the read count obtained for each locus and sample, divided into different plots according to library plate. The read counts were taken from the CIGAR sequence table produced by `ASV_to_CIGAR.py` in the Broad Institute's malaria amplicon pipeline.

Note that plate 2 had very low yield across the board and was not processed with the malaria amplicon pipeline.



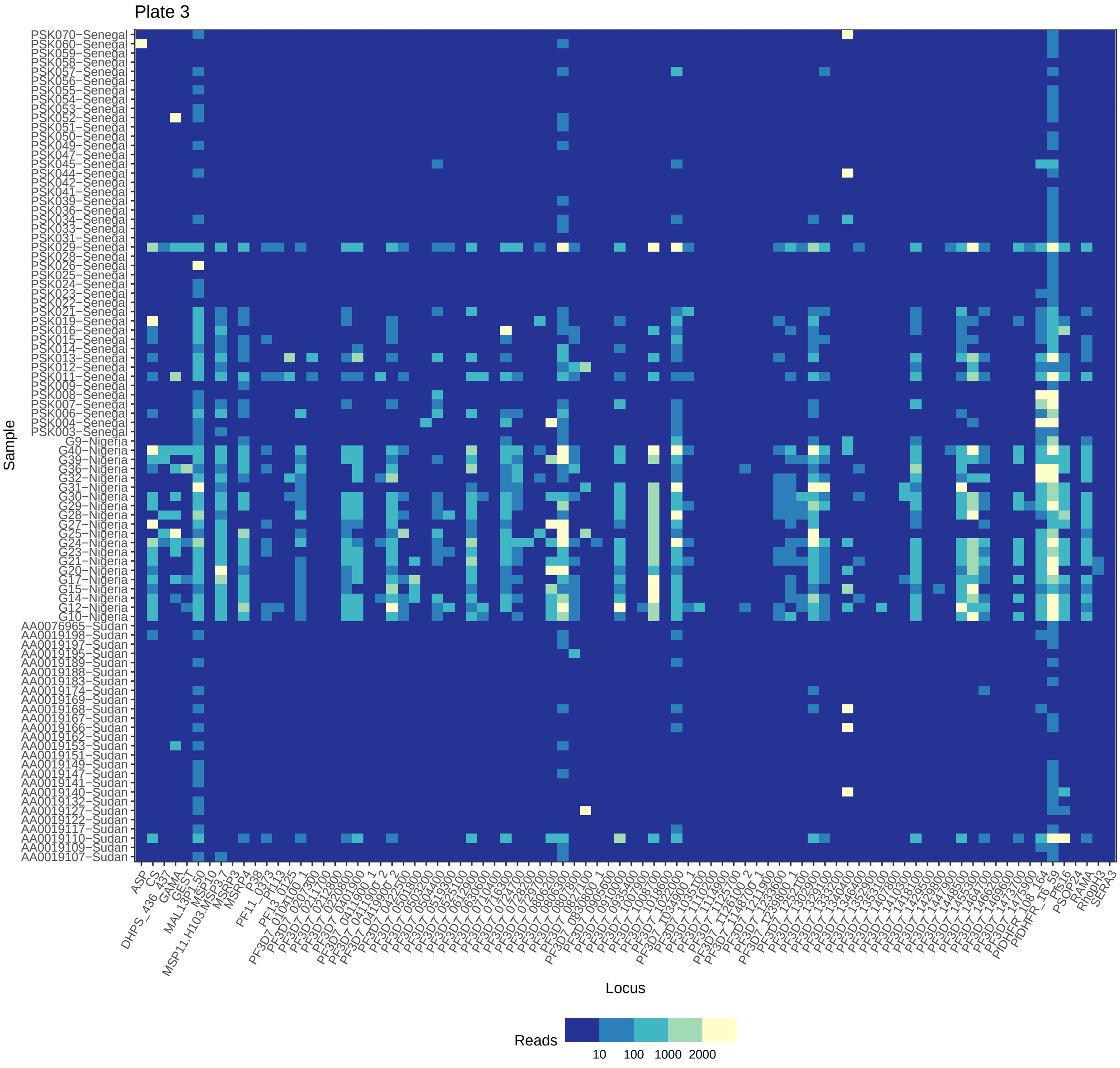


Plate 3
Sample
PSK070−Senegal
PSK060−Senegal
PSK059−Senegal
PSK058−Senegal
PSK057−Senegal
PSK056−Senegal
PSK055−Senegal
PSK054−Senegal
PSK053−Senegal
PSK052−Senegal
PSK051−Senegal
PSK050−Senegal
PSK049−Senegal
PSK047−Senegal
PSK045−Senegal
PSK044−Senegal
PSK042−Senegal
PSK041−Senegal
PSK039−Senegal
PSK036−Senegal
PSK034−Senegal
PSK033−Senegal
PSK031−Senegal
PSK029−Senegal
PSK028−Senegal
PSK026−Senegal
PSK025−Senegal
PSK024−Senegal
PSK023−Senegal
PSK022−Senegal
PSK021−Senegal
PSK019−Senegal
PSK016−Senegal
PSK015−Senegal
PSK014−Senegal
PSK013−Senegal
PSK012−Senegal
PSK011−Senegal
PSK009−Senegal
PSK008−Senegal
PSK007−Senegal
PSK006−Senegal
PSK004−Senegal
PSK003−Senegal
G9−Nigeria
G40−Nigeria
G39−Nigeria
G36−Nigeria
G32−Nigeria
G31−Nigeria
G30−Nigeria
G29−Nigeria
G28−Nigeria
G27−Nigeria
G25−Nigeria
G24−Nigeria
G23−Nigeria
G21−Nigeria
G20−Nigeria
G17−Nigeria
G15−Nigeria
G14−Nigeria
G12−Nigeria
G10−Nigeria
AA0076965−Sudan
AA0019198−Sudan
AA0019197−Sudan
AA0019195−Sudan
AA0019189−Sudan
AA0019188−Sudan
AA0019183−Sudan
AA0019174−Sudan
AA0019169−Sudan
AA0019168−Sudan
AA0019167−Sudan
AA0019166−Sudan
AA0019162−Sudan
AA0019153−Sudan
AA0019151−Sudan
AA0019149−Sudan
AA0019147−Sudan
AA0019141−Sudan
AA0019140−Sudan
AA0019132−Sudan
AA0019127−Sudan
AA0019122−Sudan
AA0019117−Sudan
AA0019110−Sudan
AA0019109−Sudan
AA0019107−Sudan
ASP
CS
DHPS_436_437
GAMA
GEST
MAL13P1.50
MSP10
MSP11.H103.MSP3.7
MSRP3
MSRP4
P38
PF11_0373
Pf113
PF13_0125
PF3D7_0104100_1
PF3D7_0207300
PF3D7_0211700
PF3D7_0212800
PF3D7_0220800
PF3D7_0401900
PF3D7_0411900_1
PF3D7_0419900_2
PF3D7_0422500
PF3D7_0501800
PF3D7_0503200
PF3D7_0504400
PF3D7_0519300
PF3D7_0525100
PF3D7_0612900
PF3D7_0630600
PF3D7_0710400
PF3D7_0716300
PF3D7_0724700
PF3D7_0728200
PF3D7_0729700
PF3D7_0806200
PF3D7_0806300
PF3D7_0807800
PF3D7_0827100
PF3D7_0830800_1
PF3D7_0905600
PF3D7_0910000
PF3D7_0935400
PF3D7_1007900
PF3D7_1008700
PF3D7_1018600
PF3D7_1022000
PF3D7_1034900_1
PF3D7_1035100
PF3D7_1110200
PF3D7_1114900
PF3D7_1125700
PF3D7_1126100_2
PF3D7_1148700_1
PF3D7_1211900
PF3D7_1233600
PF3D7_1239800_1
PF3D7_1252100
PF3D7_1302900
PF3D7_1329100
PF3D7_1333200
PF3D7_1340400
PF3D7_1346400
PF3D7_1352900
PF3D7_1353100
PF3D7_1407800
PF3D7_1410300
PF3D7_1418100
PF3D7_1429500
PF3D7_1429800
PF3D7_1447900
PF3D7_1448500
PF3D7_1455300
PF3D7_1464700
PF3D7_1466200
PF3D7_1469600
PF3D7_1473700
PF3D7_1475900
PfDHFR_108_164
PfDHFR_T6_59
Pfs16
PSOP24
RAMA
RhopH3
SERA3
Locus
Reads
10
100
1000
2000

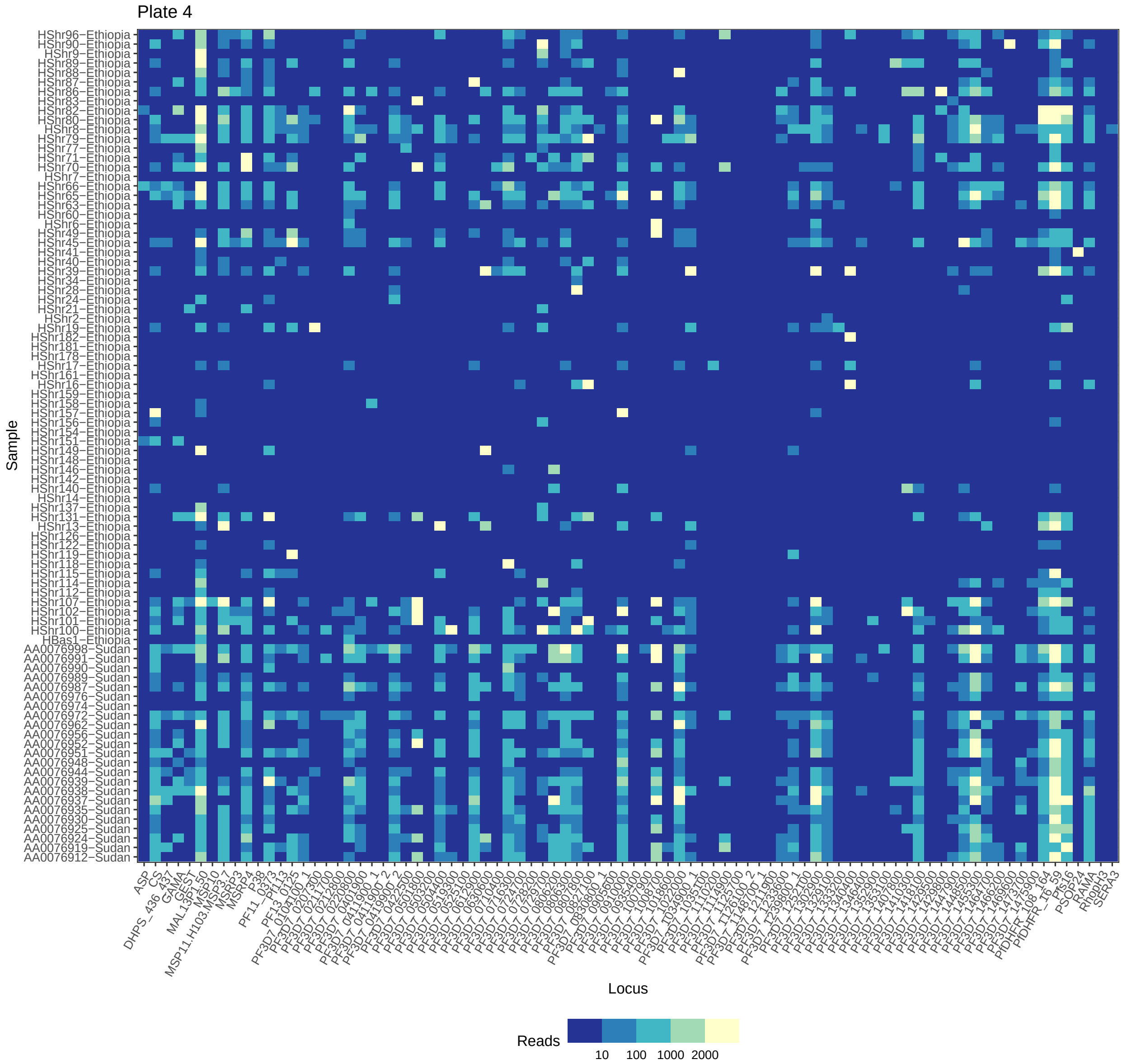


Plate 4
Sample
HShr96−Ethiopia
HShr90−Ethiopia
HShr9−Ethiopia
HShr89−Ethiopia
HShr88−Ethiopia
HShr87−Ethiopia
HShr86−Ethiopia
HShr83−Ethiopia
HShr82−Ethiopia
HShr80−Ethiopia
HShr8−Ethiopia
HShr79−Ethiopia
HShr77−Ethiopia
HShr71−Ethiopia
HShr70−Ethiopia
HShr7−Ethiopia
HShr66−Ethiopia
HShr65−Ethiopia
HShr63−Ethiopia
HShr60−Ethiopia
HShr6−Ethiopia
HShr49−Ethiopia
HShr45−Ethiopia
HShr41−Ethiopia
HShr40−Ethiopia
HShr39−Ethiopia
HShr34−Ethiopia
HShr28−Ethiopia
HShr24−Ethiopia
HShr21−Ethiopia
HShr2−Ethiopia
HShr19−Ethiopia
HShr182−Ethiopia
HShr181−Ethiopia
HShr178−Ethiopia
HShr17−Ethiopia
HShr161−Ethiopia
HShr16−Ethiopia
HShr159−Ethiopia
HShr158−Ethiopia
HShr157−Ethiopia
HShr156−Ethiopia
HShr154−Ethiopia
HShr151−Ethiopia
HShr149−Ethiopia
HShr148−Ethiopia
HShr146−Ethiopia
HShr142−Ethiopia
HShr140−Ethiopia
HShr14−Ethiopia
HShr137−Ethiopia
HShr131−Ethiopia
HShr13−Ethiopia
HShr126−Ethiopia
HShr122−Ethiopia
HShr119−Ethiopia
HShr118−Ethiopia
HShr115−Ethiopia
HShr114−Ethiopia
HShr112−Ethiopia
HShr107−Ethiopia
HShr102−Ethiopia
HShr101−Ethiopia
HShr100−Ethiopia
HBas1−Ethiopia
AA0076998−Sudan
AA0076991−Sudan
AA0076990−Sudan
AA0076989−Sudan
AA0076987−Sudan
AA0076976−Sudan
AA0076974−Sudan
AA0076972−Sudan
AA0076962−Sudan
AA0076956−Sudan
AA0076952−Sudan
AA0076951−Sudan
AA0076948−Sudan
AA0076944−Sudan
AA0076939−Sudan
AA0076938−Sudan
AA0076937−Sudan
AA0076935−Sudan
AA0076930−Sudan
AA0076925−Sudan
AA0076924−Sudan
AA0076919−Sudan
AA0076912−Sudan
ASP
CS
DHPS_436_437
GAMA
GEST
MAL13P1.50
MSP10
MSP11.H103.MSP3.7
MSRP3
MSRP4
P38
PF11_0373
Pf113
PF13_0125
PF3D7_0104100_1
PF3D7_0207300
PF3D7_0211700
PF3D7_0212800
PF3D7_0220800
PF3D7_0401900
PF3D7_0411900_1
PF3D7_0411900_2
PF3D7_0422500
PF3D7_0501800
PF3D7_0503200
PF3D7_0504400
PF3D7_0519300
PF3D7_0525100
PF3D7_0612900
PF3D7_0630600
PF3D7_0710400
PF3D7_0716300
PF3D7_0724700
PF3D7_0728200
PF3D7_0729700
PF3D7_0806200
PF3D7_0806300
PF3D7_0807800
PF3D7_0827100_1
PF3D7_0830800_1
PF3D7_0905600
PF3D7_0910000
PF3D7_0935400
PF3D7_1007900
PF3D7_1008700
PF3D7_1018600
PF3D7_1022000
PF3D7_1034900_1
PF3D7_1035100
PF3D7_1110200
PF3D7_1114900
PF3D7_1125700
PF3D7_1126100_2
PF3D7_1148700_1
PF3D7_1211900
PF3D7_1233600
PF3D7_1239800_1
PF3D7_1252100
PF3D7_1302900
PF3D7_1329100
PF3D7_1333200
PF3D7_1340400
PF3D7_1346400
PF3D7_1352900
PF3D7_1353100
PF3D7_1407800
PF3D7_1410300
PF3D7_1418100
PF3D7_1429500
PF3D7_1429800
PF3D7_1447900
PF3D7_1448500
PF3D7_1455300
PF3D7_1464700
PF3D7_1466200
PF3D7_1469600
PF3D7_1473700
PF3D7_1475900
PfDHFR_108_164
PfDHFR_T6_59
Pfs16
PSOP24
RAMA
RhopH3
SERA3
Locus
Reads
10
100
1000
2000

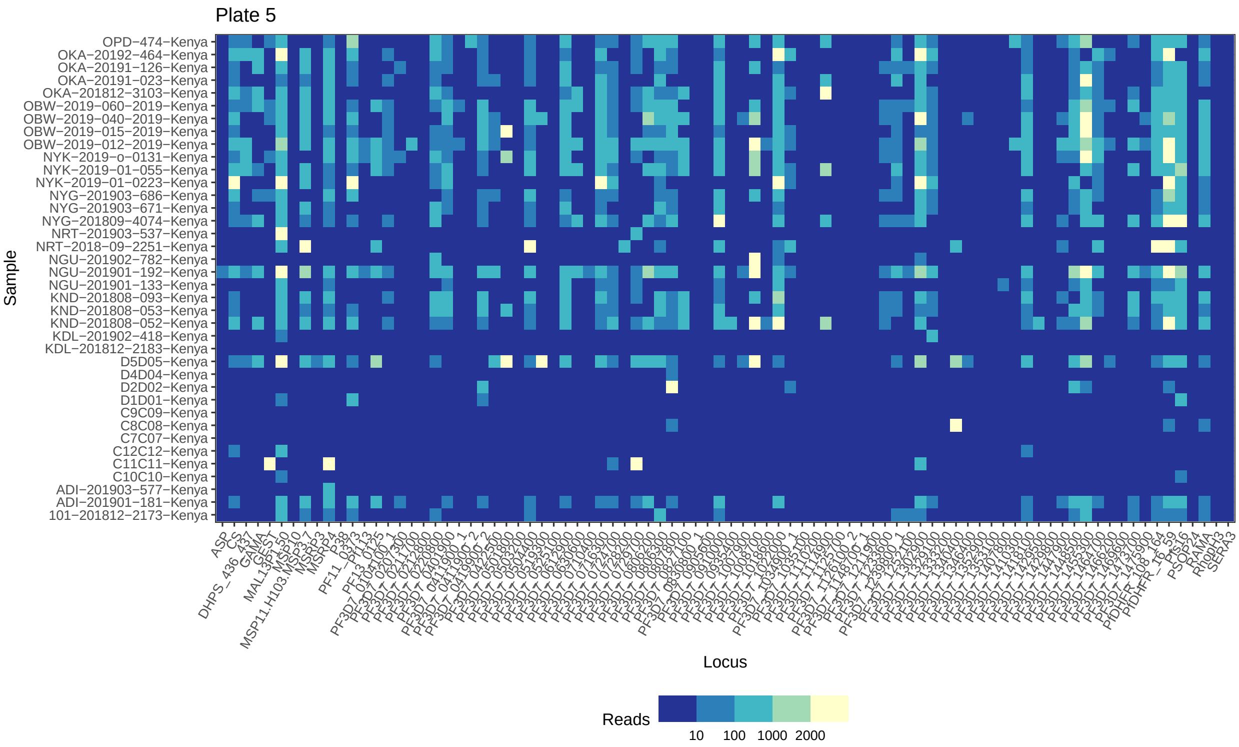


Plate 5
Sample
OPD−474−Kenya
OKA−20192−464−Kenya
OKA−20191−126−Kenya
OKA−20191−023−Kenya
OKA−201812−3103−Kenya
OBW−2019−060−2019−Kenya
OBW−2019−040−2019−Kenya
OBW−2019−015−2019−Kenya
OBW−2019−012−2019−Kenya
NYK−2019−o−0131−Kenya
NYK−2019−01−055−Kenya
NYK−2019−01−0223−Kenya
NYG−201903−686−Kenya
NYG−201903−671−Kenya
NYG−201809−4074−Kenya
NRT−201903−537−Kenya
NRT−2018−09−2251−Kenya
NGU−201902−782−Kenya
NGU−201901−192−Kenya
NGU−201901−133−Kenya
KND−201808−093−Kenya
KND−201808−053−Kenya
KND−201808−052−Kenya
KDL−201902−418−Kenya
KDL−201812−2183−Kenya
D5D05−Kenya
D4D04−Kenya
D2D02−Kenya
D1D01−Kenya
C9C09−Kenya
C8C08−Kenya
C7C07−Kenya
C12C12−Kenya
C11C11−Kenya
C10C10−Kenya
ADI−201903−577−Kenya
ADI−201901−181−Kenya
101−201812−2173−Kenya
ASP
CS
DHPS_436_437
GAMA
GEST
MAL13P1.50
MSP10
MSP11.H103.MSP3.7
MSRP3
MSRP4
P38
PF11_0373
Pf113
PF13_0125
PF3D7_0104100_1
PF3D7_0207300
PF3D7_0211700
PF3D7_0212800
PF3D7_0220800
PF3D7_0401900
PF3D7_0411900_1
PF3D7_0411900_2
PF3D7_0422500
PF3D7_0501800
PF3D7_0503200
PF3D7_0504400
PF3D7_0519300
PF3D7_0525100
PF3D7_0612900
PF3D7_0630600
PF3D7_0710400
PF3D7_0716300
PF3D7_0724700
PF3D7_0728200
PF3D7_0729700
PF3D7_0806200
PF3D7_0806300
PF3D7_0807800
PF3D7_0827100
PF3D7_0830800_1
PF3D7_0905600
PF3D7_0910000
PF3D7_0935400
PF3D7_1007900
PF3D7_1008700
PF3D7_1018600
PF3D7_1022000
PF3D7_1034900_1
PF3D7_1035100
PF3D7_1110200
PF3D7_1114900
PF3D7_1125700
PF3D7_1126100_2
PF3D7_1148700_1
PF3D7_1211900
PF3D7_1233600
PF3D7_1239800_1
PF3D7_1252100
PF3D7_1302900
PF3D7_1329100
PF3D7_1333200
PF3D7_1340400
PF3D7_1346400
PF3D7_1352900
PF3D7_1353100
PF3D7_1407800
PF3D7_1410300
PF3D7_1418100
PF3D7_1429500
PF3D7_1429800
PF3D7_1447900
PF3D7_1448500
PF3D7_1455300
PF3D7_1464700
PF3D7_1466200
PF3D7_1469600
PF3D7_1473700
PF3D7_1475900
PfDHFR_108_164
PfDHFR_16_59
Pfs16
PSOP24
RAMA
RhopH3
SERA3
Locus
Reads
10
100
1000
2000

CHAPTER 5: Conclusions

The key outcomes of this dissertation can be summarized as follows: 1) resistance surface modeling is a promising way to investigate drivers of malaria transmission, and is recommended for complex environments where barriers to transmission may not be obvious; 2) the malaria field now has access to a panel of microhaplotype loci for *Plasmodium vivax*, which enables robust characterization of within-host diversity at the scale necessary for routine surveillance; and 3) microhaplotype data for *P. falciparum* has been shown to perform well for genetic relatedness analysis in Africa, but further work is needed to evaluate how the genetic metrics obtained from this and other data relate to transmission.

In Chapter 2, resistance surface modeling and other landscape genetics methods were applied to microsatellite data from 44 locations in Western Kenya. The results clearly identified the Winam Gulf of Lake Victoria as a barrier to gene flow. While this conclusion might seem obvious after glancing at a map of the study area, it is not apparent *a priori* that this factor would consistently outrank other environmental factors that vary throughout the area, such as elevation. Resistance surface analysis makes it possible to quantitatively compare and rank the impact of various purported drivers of transmission. Furthermore, it does so in a manner that explicitly incorporates the state of these drivers between sample locations, which sets it apart from other methods, such as geographically-weighted regression (GWR), that only consider the state of drivers in the immediate vicinity of study sites. Indeed, it would have been impossible to arrive at the conclusion described above with a method like GWR, because one obviously cannot obtain malaria samples from the Gulf itself. This case study demonstrates that resistance surface modeling is a more rigorous and informative approach to investigating the drivers of malaria transmission between locations than any of the methods that are now commonly applied.

Chapters 3 and 4 advance the application of amplicon sequencing of microhaplotypes in malaria genomic epidemiology. Chapter 3 describes the design and validation of a first-of-its-kind microhaplotype panel for *P. vivax*, and Chapter 4 demonstrates and discusses the application of an analogous panel for *P. falciparum*. This genotyping technology is a promising method to gather data on within-host diversity and minority infections at a scale necessary for genomic surveillance, and the work presented in Chapter 3 lays the groundwork for this activity to begin in *P. vivax*. However, the results of Chapter 4 suggest that further theoretical and experimental work will be necessary before the full potential of this technology may be realized. While the panel performed capably in detecting (subtle) patterns of genetic relatedness in Africa, the results of this study also highlighted the fact that the relationship between genetic metrics and transmission itself remains poorly understood. This does not mean that these new genotyping techniques will not be useful. On the contrary, microhaplotype loci and the more sensitive estimates of MOI they enable will likely be necessary to perform the types of studies needed to disentangle the complex interactions of genetic diversity, relatedness, and MOI with prevalence, incidence, and transmission. As these interactions are clarified, the utility of genetic metrics as proxies for transmission will finally be understood at the level of detail necessary for genomic epidemiology to become a reliable and robust tool in the malaria surveillance toolkit.

Prior to the completion of this dissertation, resistance surface modeling and microhaplotype panels were new and relatively untested technologies in the malaria field. They were considered promising, but the feasibility of applying them at scale had not been evaluated. By doing just that, this dissertation makes a major contribution to our understanding of when and whether these methods can yield valuable information about malaria transmission. When quantitative comparison of transmission drivers is valuable, resistance surface modeling can fill the gap in a spatially-rigorous manner. When characterization of minor alleles and powerful relatedness inference are needed at a scalable cost, amplicon sequencing of microhaplotypes can now satisfy these goals for both *P. falciparum* and *P. vivax*. This dissertation raises at least as many questions as it answers, particularly with respect to the interpretation of genetic metrics of transmission, but this is another form of progress, as it helps focus future research on the important problems.

Genomic epidemiology for malaria is coming of age. It has already been used extensively to monitor the spread of drug resistance and other genetic traits of interest, as well as map the major populations of malaria and understand their historic spread. This dissertation helps expand the scope of the field to address additional important questions: what drives transmission between locations, and how can within-host diversity be best measured and interpreted? Further work is required in both areas, but the methods and tools presented in this set of studies have already shown great promise. One can only hope they will become standard practice in malaria epidemiology and thus accelerate progress towards global eradication of this deadly disease.

5.1 General References

Argyropoulos, D. C., Tan, M. H., Adobor, C., Mensah, B., Labbé, F., Tiedje, K. E., Koram, K. A., Ghansah, A., & Day, K. P. (2023). Performance of SNP barcodes to determine genetic diversity and population structure of Plasmodium falciparum in Africa. *Frontiers in Genetics*, *14*, 1071896. https://doi.org/10.3389/fgene.2023.1071896

Ashley, E. A., Pyae Phyo, A., & Woodrow, C. J. (2018). Malaria. *The Lancet*, *391*(10130), 1608–1621. https://doi.org/10.1016/S0140-6736(18)30324-6

Auburn, S., Campino, S., Clark, T. G., Djimde, A. A., Zongo, I., Pinches, R., Manske, M., Mangano, V., Alcock, D., Anastasi, E., Maslen, G., MacInnis, B., Rockett, K., Modiano, D., Newbold, C. I., Doumbo, O. K., Ouédraogo, J. B., & Kwiatkowski, D. P. (2011). An Effective Method to Purify Plasmodium falciparum DNA Directly from Clinical Blood Samples for Whole Genome High-Throughput Sequencing. *PLOS ONE*, *6*(7), e22213. https://doi.org/10.1371/journal.pone.0022213

Auburn, S., Cheng, Q., Marfurt, J., & Price, R. N. (2021). The changing epidemiology of Plasmodium vivax: Insights from conventional and novel surveillance tools. *PLoS medicine*, *18*(4), e1003560. https://doi.org/10.1371/journal.pmed.1003560

Baetscher, D. S., Clemento, A. J., Ng, T. C., Anderson, E. C., & Garza, J. C. (2018). Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. *Molecular Ecology Resources*, *18*(2), 296–305. https://doi.org/10.1111/1755-0998.12737

Bereczky, S., Mårtensson, A., Gil, J. P., & Färnert, A. (2005). Short report: Rapid DNA extraction from archive blood spots on filter paper for genotyping of Plasmodium falciparum. *The American Journal of Tropical Medicine and Hygiene*, *72*(3), 249–251. https://doi.org/10.4269/ajtmh.2005.72.249

Bousema, T., Griffin, J. T., Sauerwein, R. W., Smith, D. L., Churcher, T. S., Takken, W., Ghani, A., Drakeley, C., & Gosling, R. (2012). Hitting Hotspots: Spatial Targeting of Malaria for Control and Elimination. *PLOS Medicine*, *9*(1), e1001165. https://doi.org/10.1371/journal.pmed.1001165

Camponovo, F., Buckee, C. O., & Taylor, A. R. (2023). Measurably recombining malaria parasites. *Trends in Parasitology*, *39*(1), 17–25. https://doi.org/10.1016/j.pt.2022.11.002

Castro, M. C. (2017). Malaria Transmission and Prospects for Malaria Eradication: The Role of the Environment. *Cold Spring Harbor Perspectives in Medicine*, *7*(10). https://doi.org/10.1101/cshperspect.a025601

Cowell, A. N., Loy, D. E., Sundararaman, S. A., Valdivia, H., Fisch, K., Lescano, A. G., Baldeviano, G. C., Durand, S., Gerbasi, V., Sutherland, C. J., Nolder, D., Vinetz, J. M., Hahn, B. H., & Winzeler, E. A. (2017). Selective Whole-Genome Amplification Is a Robust Method That Enables Scalable Whole-Genome Sequencing of Plasmodium vivax from Unprocessed Clinical Samples. *mBio*, *8*(1), e02257–16. https://doi.org/10.1128/mBio.02257-16

Dalmat, R., Naughton, B., Kwan-Gett, T. S., Slyker, J., & Stuckey, E. M. (2019). Use cases for genetic epidemiology in malaria elimination. *Malaria Journal*, *18*(1), 163. https://doi.org/10.1186/s12936-019-2784-0

Färnert, A., Arez, A. P., Correia, A. T., Björkman, A., Snounou, G., & do Rosário, V. (1999). Sampling and storage of blood and the detection of malaria parasites by polymerase chain reaction. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *93*(1), 50–53. https://doi.org/10.1016/s0035-9203(99)90177-3

Feachem, R. G. A., Chen, I., Akbari, O., Bertozzi-Villa, A., Bhatt, S., Binka, F., Boni, M. F., Buckee, C., Dieleman, J., Dondorp, A., Eapen, A., Sekhri Feachem, N., Filler, S., Gething, P., Gosling, R., Haakenstad, A., Harvard, K., Hatefi, A., Jamison, D., . . . Mpanju-Shumbusho, W. (2019). Malaria eradication within a generation: Ambitious, achievable, and necessary. *The Lancet*, *394*(10203), 1056–1112. https://doi.org/10.1016/S0140-6736(19)31139-0

Gerlovina, I., Gerlovin, B., Rodríguez-Barraquer, I., & Greenhouse, B. (2022). Dcifer: An IBD-based method to calculate genetic distance between polyclonal infections. *Genetics*, *222*(2). https://doi.org/10.1093/genetics/iyac126

Henden, L., Lee, S., Mueller, I., Barry, A., & Bahlo, M. (2018). Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLoS genetics*, *14*(5), e1007279. https://doi.org/10.1371/journal.pgen.1007279

Koepfli, C., & Mueller, I. (2017). Malaria Epidemiology at the Clone Level. *Trends in Parasitology*, *33*(12), 974–985. https://doi.org/10.1016/j.pt.2017.08.013

Lopez, L., & Koepfli, C. (2021). Systematic review of Plasmodium falciparum and Plasmodium vivax polyclonal infections: Impact of prevalence, study population characteristics, and laboratory procedures. *PloS One*, *16*(6), e0249382. https://doi.org/10.1371/journal.pone.0249382

Neafsey, D. E., Taylor, A. R., & MacInnis, B. L. (2021). Advances and opportunities in malaria population genomics. *Nature Reviews. Genetics*, (22), 502–517. https://doi.org/10.1038/s41576-021-00349-5

Oyola, S. O., Ariani, C. V., Hamilton, W. L., Kekre, M., Amenga-Etego, L. N., Ghansah, A., Rutledge, G. G., Redmond, S., Manske, M., Jyothi, D., Jacob, C. G., Otto, T. D., Rockett, K., Newbold, C. I., Berriman, M., & Kwiatkowski, D. P. (2016). Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. *Malaria Journal*, *15*(1), 597. https://doi.org/10.1186/s12936-016-1641-7

Paschalidis, A., Watson, O. J., Aydemir, O., Verity, R., & Bailey, J. A. (2023). Coiaf: Directly estimating complexity of infection with allele frequencies. *PLoS computational biology*, *19*(6), e1010247. https://doi.org/10.1371/journal.pcbi.1010247

Sadoine, M. L., Smargiassi, A., Ridde, V., Tusting, L. S., & Zinszer, K. (2018). The associations between malaria, interventions, and the environment: A systematic review and meta-analysis. *Malaria Journal*, *17*(1), 73. https://doi.org/10.1186/s12936-018-2220-x

Schaffner, S. F., Taylor, A. R., Wong, W., Wirth, D. F., & Neafsey, D. E. (2018). hmmIBD: Software to infer pairwise identity by descent between haploid genotypes. *Malaria Journal*, *17*(1), 196. https://doi.org/10.1186/s12936-018-2349-7

Thompson, E. A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, *194*(2), 301–326. https://doi.org/10.1534/genetics.112.148825

Wesolowski, A., Taylor, A. R., Chang, H.-H., Verity, R., Tessema, S., Bailey, J. A., Alex Perkins, T., Neafsey, D. E., Greenhouse, B., & Buckee, C. O. (2018). Mapping malaria by combining parasite genomic and epidemiologic data. *BMC medicine*, *16*(1), 190. https://doi.org/10.1186/s12916-018-1181-9

WHO. (2023). *World Malaria Report 2023*. World Health Organization. Geneva. https://www.who.int/publications/i/item/9789240086173