THE MOLECULAR, SPATIAL, AND GENETIC EPIDEMIOLOGY OF MALARIA IN THE DEMOCRATIC REPUBLIC OF THE CONGO

Nicholas Ford Brazeau

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Epidemiology in the Gillings School of Global Public Health.

Chapel Hill
2020

Approved by:

Steven R. Meshnick

Jonathan J. Juliano

Robert J. Verity

Jessie K. Edwards

Michael E. Emch

Emily W. Gower

ABSTRACT

Nicholas Ford Brazeau: The Molecular, Spatial, and Genetic Epidemiology of Malaria in the Democratic Republic of the Congo
(Under the direction of Steven R. Meshnick)

In 2018, the Democratic Republic of the Congo (DRC), accounted for nearly 12% of the global malaria case burden and 11% of the global malaria death toll. In my dissertation, I explore the molecular, spatial, and genetic epidemiology of malaria in the DRC and provide novel insights that will help inform malaria control policy in this high burden country.

In the first aim of my dissertation, I investigate the recent finding that *Plasmodium vivax* transmission is occurring among Duffy-negative host in sub-Saharan Africa. Using data from approximately 18,000 adults, I found a 2.97% prevalence of *P. vivax* infections across the DRC. Nearly all infections were among Duffy-negative adults (486/489). Infections were not associated with typical risk-factors and were not geographically clustered. Mitochondrial genomes suggested that DRC *P. vivax* is an older clade with isolates from South America as its most recent common ancestor. Although *P. vivax* is more prevalent than previously expected, *P. vivax* in the DRC appears to be innocuous given its relatively flat distribution across space, lack of association with expected malaria risk factors, and potentially ancestral lineage. As a result, the first aim of my dissertation helps to provide public health officials with the information needed to form strategies for *P. vivax* in sub-Saharan Africa.

In the second aim of my dissertation, I used 1,111 *P. falciparum* isolates genotyped at nearly 1,800 loci from across the DRC to analyze the decay of genetic and spatial relatedness across three measures of space: (1) greater-circle distance, (2) road distance, and (3) river

distance. I found that road distance best explained the genetic relatedness in the DRC under a classic isolation by distance model. In addition, I found evidence that suggests that highly related pairs in the DRC are more frequently connected between urban and rural settings. These results suggest that human movement may be driving falciparum parasite dispersion across the DRC. Characterization of how *P. falciparum* parasites are migrating in the DRC can direct policymakers where antimalarial interventions may be most effective.

ACKNOWLEDGEMENTS

Most individuals are lucky if they have one great mentor; I have been fortunate to have three.

First, and foremost, I would like to thank my dissertation chair and PhD advisor, Steve Meshnick, who is one of the most brilliant, kind, and all-around wonderful mentors imaginable. Over six years ago, Steve took me in as a naïve budding scientist (fresh from studying wild chimpanzees in Kanyawara forest) and made me believe that I could do research at any level. I cannot thank you enough for fostering this belief in myself, developing me as a scientist (and person), and for always making me feel like you had my best interest at heart.

To Jon Juliano, thank you for being an exceptional researcher, mentor, and person. You have always made lab feel like a research-home, and it was been a privilege to sit in your space and learn from you all these years. Your versatility coupled with your humility constantly amaze me, and I can't thank you enough for the innumerable lessons (in both science and life) you have taught me.

To Bob Verity, words can hardly express my gratitude for your time and commitment to teaching me programming, Bayesian statistics, population genetics, and a number of other subjects too long to list. I am truly grateful for the opportunity to study under your expertise, and I cannot thank you enough for your out-the-box thinking, patience, and willingness to foster my passion for research.

A huge amount of thanks and credit also needs to go to the rest of my committee: Jess Edwards, who has provided innumerable insights on causality and what it means to be modern-

day epidemiologist; Mike Emch, who has provided expertise in spatial analyses; and Emily Gower, who has provided invaluable insights and critiques.

I would also like to thank the members of the infectious disease epidemiology and ecology laboratory (IDEEL) who have made science fun! It has truly been a home away from home. In particular, thanks to Molly Deutsch-Feldman for her in-depth epidemiology knowledge, enthusiasm, and friendship. I would also like to sincerely thank Cedar Mitchell and Jolly Thwai for making these projects a possibility. A special thanks to Jonathan Parr for offering up advice at opportune moments and for his enthusiasm for science, medicine, and life. Additionally, I am thankful for the opportunity to have worked under the tutelage of several of the senior IDEEL lab members: Jessica Lin, Ross Boyce, and Jeff Bailey – thank you!

I would like to thank the UNC Epidemiology department, the numerous faculty members, and peers that I have had the privilege to learn from and who have shaped my growth as an epidemiologist. I cannot imagine a better department or group of people.

Finally, and most importantly, I would like to thank my family (especially my mom, dad, and sister), friends, and partner, Katelyn. Without their support and love, none of this would have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACT	artemisinin combination therapy
CI	confidence interval
CI	credible interval
COI	complexity of infection
Comp. Score	composite score
CQ	chloroquine
crt	chloroquine resistance transporter
DAG	directed acyclic graph
DARC	Duffy antigen/receptor chemokine
DBP	Duffy-binding protein
DHS	Demographic Health Survey
DIC	deviance information criterion
DRC	Democratic Republic of the Congo
F_{st}	fixation index
G6PD	glucose-6-phosphatase dehydrogenase
GEE	generalized estimating equations
HBHI	high burden, high impact
HIV	human immunodeficiency virus
Hospital Dist.	Distance to a hospital
HRM	high resolution melt
HRP2	histidine rich protein 2
IBD	identity by descent

IBS	identity by state
IPW	inverse probability weights
ITN	insecticide treated net
km	kilometer
LD	linkage disequilibrium
LDH	lactate dehydrogenase
m	meters
mdr1	multidrug resistance 1
MLE	maximum likelihood estimation
mm	millimeters
MRCA	most recent common ancestor
MSM	marginal structural model
mtDNA	mitochondrial genome
N_e	effective population size
NHA	non-human apes
OSRM	Open Street Routing Machine
P. falciparum	*Plasmodium falciparum*
P. vivax	*Plasmodium vivax*
pOR	prevalence odds ratios
qPCR	quantitative polymerase chain reaction
RDT	rapid diagnostic test
Rur.	rural
SD	standard deviation

SSA	sub-Saharan Africa
Trad.	traditional
Water Dist.	Distance to water
WGS	whole genome sequencing
WHO	World Health Organization

CHAPTER ONE: SPECIFIC AIMS

Aim 1: Identifying the Risk, Distribution, and Origin of P. vivax in the Democratic Republic of the Congo

In 2017, *Plasmodium vivax* was estimated to cause 14.3 million cases globally, with the majority of infections occurring outside of sub-Saharan Africa [1]. However, recent evidence has shown that *P. vivax* is prevalent across the sub-Saharan Africa region, overturning the consensus of its absence from the region. Despite growing concern of *P. vivax* in the sub-Saharan region, no studies have systematically determined the prevalence, distribution, or clinical relevance of these infections to date [2–6]. Similarly, no studies have identified the source of sub-Saharan *P. vivax*, which will be informative to determine the history of these infections in the region. Although the resurgence of *P. vivax* within sub-Saharan Africa has the potential to undermine malaria elimination campaigns, its clinical burden and relevance need to be determined prior to allocation of resources. Proper allocation of resources is becoming increasingly crucial for malaria elimination campaigns as malaria incidence is rebounding globally [7]. This prompts the need to evaluate the infectious burden of *P. vivax* by determining where infections are occurring, who is at risk for infection, and the likely origin of these infections.

To fill this critical gap in knowledge, I screened the 2013-2014 Demographic Health Survey (DHS) from the Democratic Republic of the Congo (DRC) for *P. vivax*. The 2013-2014 DRC DHS was a nationally representative survey of approximately 18,000 adults with over 500 demographic and behavioral covariates, geographical information, and a dried blood spot for each participant [8].

Using this rich resource, I aimed to do the following:

<u>Aim 1.1</u>: Determine the environmental, behavioral, genetic, and spatial risk factors associated with *P. vivax infection*.

Rationale. Although many of the risk factors associated with *P. vivax* infection are expected to be similar to *P. falciparum, P. vivax* is unique in its life cycle (i.e. hypnozoites) and its propensity to be infectious prior to the presentation of clinical symptoms (i.e. shortened intrinsic period) [9–12]. Therefore, identifying areas of transmission (**Aim 1.2**) and risk factors that predict *P. vivax* infection will inform public health officials on <u>*who*</u> is being infected.

Hypothesis. I hypothesize that important risk factors will include urbanicity and wealth -- but to a lesser extent than *P. falciparum*, due to the shorter *P. vivax* intrinsic period.

<u>Aim 1.2:</u> Characterize the national prevalence and geographical distribution of *P. vivax*.

Rationale. Previous reports on *P. vivax* prevalence and burden have only been conducted in a few field sites through convenience sampling and have lacked robust spatial sampling [4]. As a result, there is a critical gap in our understanding of *P. vivax* prevalence across sub-Saharan Africa. Maps are needed to inform public health officials <u>*where*</u> disease is occurring.

Hypothesis. I hypothesize that the unique biological characteristics of *P. vivax* will lead to increased local transmission, as relapses from dormant infections will cause "hotspots," or local infection clusters.

<u>Aim 1.3:</u> Using mitochondrial genomes, determine if *P. vivax* infections represent local transmission or importation from outside of the sub-Saharan African region.

Rationale. To date, no studies have attempted to determine the origin of *P. vivax* infections among Duffy-negative hosts in sub-Saharan Africa. Differentiating between imported cases and cases of local transmission is critical for informing public health interventions and can be readily accomplished using measures of genetic differentiation and population structure.

Hypothesis. I hypothesize that the DRC *P. vivax* population is an ancient population that has lingered in the country among various non-human ape and human hosts and has not been recently imported into the country.

This aim will provide the first population-based study to quantify the prevalence of *P. vivax*, to identify the epidemiological risk factors associated with *P. vivax* infection, and to determine the origin of these infections in a sub-Saharan African country. Collectively, this novel combination of epidemiological and population genetics data will provide public health officials and policymakers with the critical information they need to differentiate if *P. vivax* is a reemerging infection or simply an innocuous threat in sub-Saharan Africa.

Aim 2: Tracing the Genetic Relatedness of Plasmodium falciparum in the Democratic Republic of the Congo across Space

Decreasing sequencing costs and advancing methods have allowed genomic data to become a standard tool in infectious disease epidemiology for tracking outbreaks and inferring transmission patterns. In the study of *P. falciparum* malaria, these genomic tools have been combined with epidemiological data to identify importation of drug-resistance, identify transmission networks, and track parasite migration [13–24]. Leveraging these tools and approaches to identify areas where parasites are highly connected can help guide malaria control efforts.

Patterns of identity by descent (IBD) have been recognized as an informative measure of *P. falciparum* parasite connectedness, particularly, spatial connectedness [21–25]. However, many of the previous studies linking IBD and geographic space have focused on isolates originating from a few sites or being sourced from multiple countries. As a result, these studies have largely been limited to estimating patterns of local transmission or global transmission without any connection between these two dynamics.

To fill this critical gap in knowledge, I used a spatially robust dataset that included 351 sites and 1,111 samples across the Democratic Republic of the Congo (DRC) to identify patterns of genetic and spatial connectedness among *P. falciparum* parasites. The DRC is an ideal location to study *P. falciparum* genetic-transmission dynamics: it is the second largest country in sub-Saharan Africa, it is a bridge linking East and West Africa *P. falciparum* genetic diversity, and it has a high burden of *P. falciparum* that exhibits spatial heterogeneity in prevalence [26–28].

From this spatially rich dataset, I aim to do the following:

<u>Aim 2.1</u>: Identify patterns of parasite dispersion.

Rationale. Parasites may be dispersed by: (1) mosquito movement, which can be approximated by greater-circle distances, or (2) human travel, which can be approximated by road- and river-distances. Differentiating between these two scenarios has important implications for malaria control efforts. For example, if parasites are more commonly spread to areas by mosquitoes, resources may be better allocated to vector control. In contrast, if parasites dispersion is dominated by human movements, control programs may want to prioritize importation surveillance and target major crossroads.

Hypothesis. I hypothesize that there will be a strong spatial signal in the road-network model that approximates malaria dispersion by human migration. This would suggest that road-distance, or human travel, may drive parasite connectedness events more than mosquito dispersion.

<u>Aim 2.2</u>: Determine the spatial, ecological, and malaria-interventional predictors of identity by descent among DRC parasites.

Rationale. Identifying the covariates that are associated with IBD can inform malaria control programs on which interventions most decrease parasite genetic diversity. Decreasing genetic diversity is associated with a lower effective population size and potential culling of the malaria parasite. In order to avoid autocorrelation in the pairwise IBD-measure, I will aggregate IBD

within each province and between provinces as the outcomes of interest. Given that intervention planning is typically implemented on the province level, modeling province-level effects is likely to be most informative for malaria policymakers in the DRC.

Hypothesis. I hypothesize that there will be few non-spatial predictors of IBD in the DRC given its high-transmission setting. However, I predict that there will be a strong signal between IBD and prevalence that will reflect the differing levels of transmission intensity in the region.

In this aim, I will use a spatially-rich genomic dataset to determine the connectedness of *P. falciparum* infections across the DRC. By identifying regions of high parasite connectedness, I can provide targeted feedback for intervention planning and intervention efforts. Maps of spatial and genetic connectedness differ from traditional incidence or prevalence maps, as they provide a picture of how *P. falciparum* infections may be arising instead of simply where. Through these means, public health officials will be able to target the root of *P. falciparum* infections instead of chasing infections across the country.

REFERENCES

1. Battle, K. E. *et al.* Mapping the global endemicity and clinical burden of Plasmodium vivax, 2000–17: a spatial and temporal modelling study. *Lancet* **394**, 332–343 (2019).

2. Zimmerman, P. A. Plasmodium vivax Infection in Duffy-Negative People in Africa. *Am. J. Trop. Med. Hyg.* **97**, 636–638 (2017).

3. Mendes, C. *et al.* Duffy negative antigen is no longer a barrier to Plasmodium vivax--molecular evidences from the African West Coast (Angola and Equatorial Guinea). *PLoS Negl. Trop. Dis.* **5**, e1192 (2011).

4. Twohig, K. A. *et al.* Growing evidence of Plasmodium vivax across malaria-endemic Africa. *PLoS Negl. Trop. Dis.* **13**, e0007140 (2019).

5. Tournamille, C., Colin, Y., Cartron, J. P. & Van Kim, C. L. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy–negative individuals. *Nat. Genet.* **10**, 224–228 (1995).

6. Miller, L. H., Mason, S. J. & Clyde, D. F. The resistance factor to Plasmodium vivax in blacks: the Duffy-blood-group genotype, FyFy. *New England Journal* (1976).

7. World Health Organization. *World Malaria Report 2018*. (World Health Organization, 2019).

8. Ministère du Plan et Suivi de la Mise en œuvre de la Révolution de la Modernité (MPSMRM), Ministère de la Santé Publique (MSP) et ICF International, 2014. Enquête Démographique et de Santé en République Démocratique du Congo 2013-2014.

9. White, N. J. Determinants of relapse periodicity in Plasmodium vivax malaria. *Malar. J.* **10**, 297 (2011).

10. Adams, J. H. & Mueller, I. The Biology of Plasmodium vivax. *Cold Spring Harb. Perspect. Med.* **7**, (2017).

11. Mueller, I. *et al.* Key gaps in the knowledge of Plasmodium vivax, a neglected human malaria parasite. *Lancet Infect. Dis.* **9**, 555–566 (2009).

12. McCarthy, J. S. *et al.* Experimentally induced blood-stage Plasmodium vivax infection in healthy volunteers. *J. Infect. Dis.* **208**, 1688–1694 (2013).

13. Patel, J. C. *et al.* Genetic Evidence of Importation of Drug-Resistant Plasmodium falciparum to Guatemala from the Democratic Republic of the Congo. *Emerg. Infect. Dis.* **20**, 932–940 (2014).

14. Redmond, S. N. *et al.* De Novo Mutations Resolve Disease Transmission Pathways in

Clonal Malaria. *Mol. Biol. Evol.* **35**, 1678–1689 (2018).

15. Chang, H.-H. *et al.* Mapping imported malaria in Bangladesh using parasite genetic and human mobility data. *Elife* **8**, (2019).

16. Tessema, S. K. *et al.* Applying next-generation sequencing to track falciparum malaria in sub-Saharan Africa. *Malar. J.* **18**, 268 (2019).

17. Omedo, I. *et al.* Micro-epidemiological structuring of Plasmodium falciparum parasite populations in regions with varying transmission intensities in Africa. *Wellcome Open Res* **2**, 10 (2017).

18. Wesolowski, A. *et al.* Mapping malaria by combining parasite genomic and epidemiologic data. *BMC Med.* **16**, 190 (2018).

19. Tessema, S. *et al.* Using parasite genetic and human mobility data to infer local and cross-border malaria connectivity in Southern Africa. *Elife* **8**, (2019).

20. Daniels, R. F. *et al.* Modeling malaria genomics reveals transmission decline and rebound in Senegal. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7067–7072 (2015).

21. Daniels, R. *et al.* Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. *PLoS One* **8**, e60780 (2013).

22. Zhu, S. J. *et al.* The origins and relatedness structure of mixed infections vary with local prevalence of P. falciparum malaria. *Elife* **8**, (2019).

23. Taylor, A. R. *et al.* Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent. *PLoS Genet.* **13**, e1007065 (2017).

24. Shetty, A. C. *et al.* Genomic structure and diversity of Plasmodium falciparum in Southeast Asia reveal recent parasite migration patterns. *Nat. Commun.* **10**, 2665 (2019).

25. Taylor, A. R., Jacob, P. E., Neafsey, D. E. & Buckee, C. O. Estimating Relatedness Between Malaria Parasites. *Genetics* **212**, 1337–1351 (2019).

26. Verity, R. J., Aydemir, O., Brazeau, N. F. & Watson, O. J. The Impact of Antimalarial Resistance on the Genetic Structure of Plasmodium falciparum in the DRC. *bioRxiv* (2019).

27. Taylor, S. M. *et al.* Molecular malaria epidemiology: mapping and burden estimates for the Democratic Republic of the Congo, 2007. *PLoS One* **6**, e16420 (2011).

28. Molly Deutsch-Feldman, Nicholas F. Brazeau, Jonathan B. Parr, Kyaw L. Thwai, Jérémie Muwonga, Melchior Kashamuka, Antoinette K. Tshefu, Jessie K. Edwards, Robert Verity, Michael Emch, Emily W. Gower, Jonathan J. Juliano, Steven R. Meshnick. Spatial and epidemiological drivers of P. falciparum malaria among adults in the Democratic Republic

of the Congo.

CHAPTER TWO: INTRODUCTION

Overview and Global Burden of Malaria

There are five protozoan parasites that cause clinical malaria among human hosts, all within the *Plasmodium* genus: *P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae*, and *P. knowlesi*. Of these five species, *P. falciparum* and *P. vivax* account for the vast majority of the global malaria burden [1]. In 2017, *P. falciparum* accounted for approximately 193.9 million cases of malaria, while *P. vivax* accounted for 14.3 million cases [2,3]. Although the global burden has decreased over the past decade, recent evidence suggests that case-reduction rates have plateaued and malaria burden may be increasing globally [1]. This plateau represents an impasse in malaria elimination efforts, as concerns of *Plasmodium* drug-resistance, donor-fatigue, change in vector behavior, and vector insecticide-resistance are mounting [1].

Although current literature is evolving, it is generally believed that: (1) the burden of malaria is highest in sub-Saharan Africa, where *P. vivax* is absent; (2) *P. falciparum* is much more deadly than *P. vivax*; and (3) *P. vivax* is not associated with strong-seasonal outbreaks or epidemics like *P. falciparum*. These factors have long led to *P. vivax* being considered a "benign malaria" and has led to the neglect of *P. vivax* as a global burden, when compared to *P. falciparum* [4–6].

Malaria Parasite Life-Cycle

Malaria is a vector-borne disease transmitted by the *Anopheles sp.* mosquito. Specifically, inoculation occurs when an infected female anopheline mosquito takes a blood-meal from a susceptible host and releases sporozoites into the bloodstream. These sporozoites then migrate to

the liver and infect hepatocytes (**Figure 1**: Liver Stage). After a 5-8 day incubation period within the liver, the schizont hepatocyte erupts and spills merozoites into the peripheral circulation (**Figure 1**: Blood Stage) [7,8]. From the peripheral circulation, merozoites invade red blood cells (i.e. *erythrocytes*, *reticulocytes*) through a series of interactions between merozoite surface invasion ligands and host-cell receptor that are both malaria-species and host-species dependent (i.e. *host tropism*) [9]. Predominantly, malaria-species invasion ligands are contained within two protein families: the erythrocyte binding-like proteins and reticulocyte binding-like proteins [9,10]. These protein-families are thought to dictate much of the host specificity of the *Plasmodium* genus [9].

Following invasion of the erythrocyte, the merozoite enters a ring stage and starts to consume the different components of the cell while also altering the cellular cytoskeleton to facilitate the importation of nutrients [7]. After a short incubation period, the infected erythrocyte erupts and releases merozoites and a subset of sexual-stage gametocytes into the peripheral circulation [7,8]. The released merozoites will then infect new susceptible erythrocytes and propagate the blood stage cycle.

Separately, gametocytes will remain in the peripheral circulation with the goal of inoculating a female *Anopheles* mosquito during a blood-meal. After the gametocytes have successfully inoculated the *Anopheles* mosquito, a "male" gamete and a "female" gamete fuse to form a zygote. At this point, the zygote undergoes meiosis and produces an ookinete [7,8]. Through a not well-understood process, the ookinete invades the mosquito midgut cellular wall and travels through the endolymph to the mosquito salivary glands [11]. Once in the mosquito salivary gland, the female *Anopheles* mosquito is infectious and can propagate the spread of malaria parasites.

Globally, there are approximately 41 dominant/competent vector species that can transmit malaria [12–14]. The main vector depends on the country with some differences between *Plasmodium* species competences (reviewed in Sinka *et al.* 2010a*,* Sinka *et al.* 2010b, Sinka *et al.* 2011).

P. vivax follows the typical malaria parasite life cycle described above with one additional stage: the hypnozoite stage (**Figure 1**: Hypnozoite Stage). After sporozoites have migrated to the liver and infected hepatocytes, a subset of sporozoites differentiate into hepatic schizonts (described above) and hypnozoites, respectively. Hypnozoites are a dormant stage of the parasite that can reemerge weeks to months after the primary infection has occurred, termed a "relapsing" infection [10,15]. Although still debated, the periodicity of *P. vivax* relapses appears to correspond with climate -- temperate zones: relapse periodicity appears to be on the order of magnitude of months; tropical zones: relapse periodicity appears to be on the order of magnitude of weeks [15].

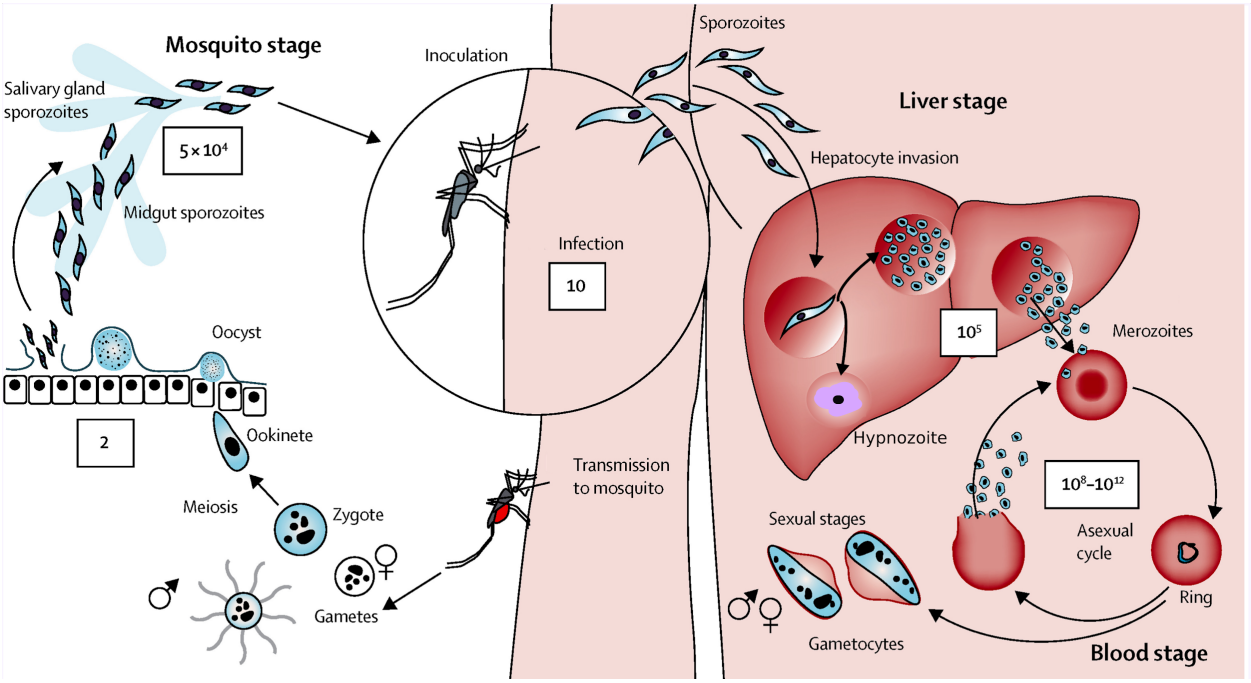


Figure 1.1 – Malaria Life Cycle: Displayed is the malaria life cycle showing inoculation by an anopheline mosquito (vector) and subsequent liver and blood stage infections (figure adapted from White et al. 2014). For each stage, the number of infectious agents is displayed for *P. falciparum*, while the number of *P. vivax* infectious agents are expected to one order of magnitude lower (Baird, 2016). Hypnozoites (purple) are a unique feature of the *P. vivax* life cycle, which allow parasites to remain dormant for weeks to months.

Diagnosis of Malaria

Malaria typically presents with a low- to high-grade fever, malaise, anorexia, myalgias, arthralgias, gastrointestinal symptoms (i.e. vomiting, diarrhea, abdominal pain), and other nonspecific symptoms. In addition, splenomegaly is often appreciated on physical exam. Laboratory results typically show increased inflammation and signs of infection (i.e. elevated erythrocyte sedimentation rate, C-reactive protein, thrombocytopenia, elevated white blood cell count with a predominating lymphocytic pattern) and elevated lactate dehydrogenase levels with low hemoglobin (resulting from the destruction of erythrocytes). Some patients may also present with elevated liver function tests, elevated ferritin levels, and an altered basic metabolic panel, typically indicating acidosis.

Diagnosis of malaria can be made with microscopy (gold standard), a rapid diagnostic test (RDTs), or with a polymerase-chain reaction (PCR) assay. Although PCR is likely the most sensitive and specific method for detecting malaria, it requires specialized laboratory conditions, reagents, and training. As such, sensitivities and specificities will range widely depending on resource-availability, primer-sets used, and laboratory conditions.

In resource limited settings, the diagnosis of *P. falciparum* has historically relied on microscopy, and more recently RDTs. *P. falciparum* RDTs typically target the *P. falciparum* histidine-rich protein 2 (*pfHRP2*) antigen, which has been shown to be heat-stable relative to other antigen targets [16]. However, recent reports in the DRC and elsewhere in Africa have identified parasites with *pfHRP2*-gene deletions that are undetectable by traditional RDTs [16–19]. These "stealth parasites" greatly complicate the diagnosis of *P. falciparum* in resource limited settings [19].

The key difference between diagnosing *P. falciparum* versus *P. vivax* infections is the blood-stage parasite density, such that a *P. vivax* infection tends to have one order of magnitude fewer parasites than a *P. falciparum* infection [20]. This systematically lower parasitemia in *P. vivax* makes diagnosis more challenging. In a systematic review from the Cochrane Organization, RDTs designed to specifically detect *P. vivax* lactate dehydrogenase (*pvLDH*) performed well, missing only 5% of infections relative to microscopy [21]. However, RDTs designed to detect falciparum and non-falciparum malaria (*pfHRP2/panLDH)* had lower sensitivity for *P. vivax* infections, ranging from 78-89% [21]. As a result, in regions where *P. falciparum* and *P. vivax* are co-endemic, the type of RDT used will have an impact on *P. vivax* diagnostic capacity. Despite varying levels of sensitivity, both vivax-specific and non-falciparum malaria RDTs had high specificity for *P. vivax*, ranging from 98-99% [21].

Finally, there is no diagnostic test available to detect hypnozoites. This inability to detect hypnozoites has been a long-standing barrier to *P. vivax* elimination [22,23]. Recently, White *et al.* 2014 demonstrated that the hypnozoite reservoir can propagate infections and maintain a steady-state of infections if untreated [22]. Put more simply, *P. vivax* elimination is likely impossible without treatment of hypnozoite parasites.

Treatment and Management of Malaria

Treatment of *P. falciparum* is primarily focused on eliminating liver-stage parasites [24]. Due to widespread resistance of chloroquine (CQs) and antifolate-based agents, artemisinin-combination therapies (ACTs) are the first-line treatment for *P. falciparum* malaria infections (recommendation since 2010) [24]. ACTs are the last antimalarial drug-class that does not have near ubiquitous drug-resistance. However, artemisinin monotherapy drug-resistance has emerged in Southeast Asia, likely dating to 2008 [24–26]. In addition, ACT partner drug resistance (i.e. quinoline-based agents, antifolate/sulfur-based agents, and artemisinin-derivatives) continues to arise and proliferate across Southeast Asia [24–29]. As a result of this growing antimalarial resistance in Southeast Asia, trials for artemisinin-triple therapies have been initiated.

To date, artemisinin-resistance has not been identified at appreciable levels in sub-Saharan Africa [30–32]. However, as in Southeast Asia, ACT partner drug resistance is prevalent in sub-Saharan Africa [24–26,28,29]. Partner drug resistance in sub-Saharan Africa appears to be mainly due to mutations in the multidrug resistance 1 (*pfmdrl*) locus (namely N76Y, Y184F, D1246Y) [24]. These mutations in the *pfmdrl* locus provide some resistance to partner drugs (e.g. lumefantrine, mefloquine), but selectional pressures at the *pfmdrl* locus appear to be variable and with potential balancing selection between various resistance phenotypes (or other biological functions) [24]. Separately, CQ resistance is essentially fixed across Africa due to the CVIET

haplotype at the *pfcrt* locus, which also offers some resistance against other quinoline- and artemisinin-based agents [24,33,34]. Given that ACT are the first-line treatment in Africa, importation or the emergence of ACT resistance in Africa could wreak havoc on malaria control and elimination programs.

Treatment of *P. vivax* is separated into two categories: (1) active parasites (merozoites, schizonts, and gametocytes), and (2) dormant parasites (hypnozoites) based on the antimalarials needed to access each respective compartment. In order for radical cure to be achieved, both compartments must be addressed, which acquires the administrated of at least two different classes of medications. For the active infection, CQ (drug class: 4-aminoquinoline), is the first-line antimalarial [35]. However, in regions with high-CQ resistance, ACTs are indicated [35]. In order to eliminate the hypnozoite reservoir, either primaquine or tafenoquine (drug class: 8-aminoquinoline) must be used [35,36]. However, both primaquine and tafenoquine are contraindicated in patients with glucose-6-phosphatase dehydrogenase (G6PD) deficiency, as the drugs have the potential to cause hemolytic anemia in these patients.

In terms of *P. vivax* drug-resistance, both CQ, antifolate/sulfur-based agents, and ACT partner-drug resistance (e.g. mefloquine) have been identified; however, orthologous mutations that confer artemisinin-resistance in *P. falciparum* have not yet been recorded [37,38]. Similarly, although there have been reports of *P. vivax* primaquine resistance, many of these putatively-resistant cases have been attributed to inappropriate administration of the drug, and primaquine appears to remain highly efficacious as a radical cure [39].

Within-host Plasmodium *Interactions*

There is evidence that *P. falciparum* and *P. vivax* compete within the host and within-host interactions influence infectiousness and transmissibility [40]. Bruce *et al.* 2000 suggested that

there is a density-dependent relationship between *Plasmodium* species, such that the emergence and growth of one parasite line has the potential to inhibit lower-density coinfections. This is important in the context of *P. falciparum*, which is associated with higher parasitemias than *P. vivax*, and thereby may suppress *P. vivax* intra-host expansion [40]. This potential density-dependent relationship is often observed in cases where *P. vivax* recurrences appear to be triggered by co-infections of *P. falciparum* or other infectious diseases [41–43]. Specifically, Lin *et al.* 2011 showed that when patients were treated with ACTs, baseline *P. falciparum* gametocyte carriage was predictive of risk of *P. vivax* infection (or recurrence) during a 28-day follow-up period. This would suggest that *P. falciparum* parasites were inhibiting *P. vivax* growth and once removed by the blood-stage ACT treatment, allowed the *P. vivax* minor clones to expand from hypnozoites or merozoites that survived treatment. However, in Lin *et al.* 2011, patients were not isolated (or removed to a vivax free-zone) and sequencing was not performed. As a result, it was not possible to determine if follow-up infections were a new infection, recrudescence infection, or a relapse infection [43]. Collectively, this suggests that there may be evidence for a within-host interaction -- or intra-host competition -- between *P. falciparum* and *P. vivax* but the evidence is currently incomplete. If an interaction does exist, the elimination of *P. falciparum* -- without the concurrent elimination of *P. vivax* -- may open a niche for *P. vivax* to propagate.

Aim 1: Identifying the Risk, Distribution, and Origin of P. vivax in the Democratic Republic of the Congo

Unique Features of the Plasmodium vivax Life Cycle and Biology

The *P. vivax* parasite has several key features that differentiate it from other types of malaria, including: (1) merozoites preferentially infect reticulocytes (i.e. young erythrocytes); (2) merozoite incubation within the reticulocytes typically lasts for 24-48 hours (i.e. “tertian malaria”); (3) gametocyte production occurs early in the infection, typically before presentation

of symptoms at the onset of the blood-stage; (4) gametocyte infectivity appears to be high, as it takes few gametocytes to infect an anopheline mosquito; (5) the hypnozoite stage [5,44–49]. The duration of a single *P. vivax* life-cycle within the human host -- from inoculation to gametocyte production -- is estimated at approximately 7 days, although the point of infectivity may be later [45]. As a result of this short intrinsic period, individuals are likely to be infectious before seeking treatment. In addition, the hypnozoite reservoir and subsequent reactivation of *P. vivax* parasites weeks to months (or even years) after the primary infection can perpetuate transmission [15]. The implications of the unique features of *P. vivax* biology for vivax malaria control and elimination efforts are discussed below (section: ***Plasmodium vivax Obstacles and Relevance to Public Health***).

Plasmodium vivax Invasion

The invasion pathway of *P. vivax* is paradoxical, such that there appears to be a degree of host promiscuity, as strains can readily infect chimpanzees, gorillas, and human hosts, but there is a high degree of cellular tropism within the human host [9,50–52]. As discussed above, within the host, *P. vivax* parasites preferentially infect reticulocytes. Reticulocyte invasion is mediated through the *P. vivax* Duffy-binding protein (*pvDBP*) ligand, which is part of the erythrocyte/Duffy binding protein family [53]. The *pvDBP* ligand attaches to the host Duffy antigen/receptor chemokine (DARC; also known as the atypical chemokine receptor 1) and creates a junction-formation as the first-step in the cellular invasion process. This junction-formation essentially acts as a bridge for the parasite to invade the cell [9,50,54–58]. In addition to this DARC-parasite interaction, *P. vivax* also appears to use a series of *P. vivax* reticulocyte binding proteins, including *PvRBP1a*, *PvRBP1b*, *PvRBP2a*, and *PvRBP2b* to invade reticulocytes [53,59,60]. The exact mechanism of reticulocyte invasion is still poorly understood but the repertoire of

antigens needed for cellular invasion may help to explain *P. vivax* reticulocyte selectivity [53,59,60]. Regardless of the exact mechanism, the DARC gene and expression of DARC on the surface of reticulocytes has long been considered a key-criterion for *P. vivax* infection.

The Duffy-Negative Phenotype

The DARC blood group system was first discovered in 1950 following a hemolytic transfusion reaction in a patient with hemophilia [61]. Following this discovery, the underlying codominant phenotype of the DARC blood group was resolved as: $Fy^{a,b}$, $Fy^{a-,b}$, $Fy^{a,b-}$, and $Fy^{a-,b-}$ [62]. In particular, the $Fy^{a-,b-}$ phenotype results from a single-point mutation in the GATA-1 transcription factor (-33 T:C) of the FY*B gene that abrogates expression of the DARC antigen [54]. The absence of the DARC antigen on the RBC surface is commonly referred to as the Duffy-negative phenotype ($Fy^{a-,b-}$). This phenotype is extremely common among individuals of African descent (**Figure 1.2**) and was shown to provide resilience to *P. vivax* infection [55,56].

However, despite the evidence that the Duffy-negative phenotype provides resilience to *P. vivax* infection, several recent studies have shown that contemporary *P. vivax* strains can infect Duffy-negative individuals [50,63–75]. Much recent work has tried to identify the alternative invasion pathway for *P. vivax* in Duffy-negative hosts, with two predominating hypotheses: (1) Expansion of the *pvDBP1* ligand and increased copy-number variation, which may provide low-affinity binding to Duffy-like antigens on host reticulocytes [50]; (2) Erythrocyte binding protein (*EBP2*) as an alternative ligand for invasion of Duffy-null erythrocytes [76,77]. However, both hypotheses have largely been shown to be incorrect [9,53,78,79]. More recent evidence suggests that *P. vivax* tryptophan-rich antigen (*pv-fam-a*) and merozoite surface protein 3 (*pvMSP3*) families may be involved in Duffy-negative invasion, but the study consisted of only a few monkeys with limited differential RNA-sequencing results [80]. As a result, the mechanism of how *P. vivax* is

infecting Duffy-negative cells remains unknown (for a review on potential ligands, see Gunalan *et al.* 2018, Table 2).

Plasmodium vivax Obstacles and Relevance to Public Health

Of the numerous unique features of *P. vivax* biology, two features thwart intervention efforts most: (1) the hypnozoite reservoir and 21) the production of gametocytes before symptoms (shortened intrinsic period). The hypnozoite reservoir and subsequent reactivation of *P. vivax* parasites weeks to months after a primary infection has significant potential to perpetuate transmission. The potential reactivation of hypnozoites emphasizes the need of radical cure with 8-aminoquinoline medications (e.g. primaquine or tafenoquine) in order to stop transmission. However, 8-aminoquinoline medications are contraindicated in patients with G6PD-deficiency (discussed above). G6PD-deficiency is prevalent in regions of high-malaria endemicity, as it provides some resilience against malaria infection (**Figure 1.2**) [81]. As a result, a paradoxical situation arises as a mutation that was historically beneficial for conferring a malara-resilient phenotype is now interfering with contemporary medical practices to eliminate the hypnozoite reservoir. As noted above, gametocytes are often produced before or at the onset of symptoms, which means individuals infected with *P. vivax* are often infectious before they seek treatment. As a result, isolation or quarantine efforts based on *P. vivax* clinical presentation are relatively futile.

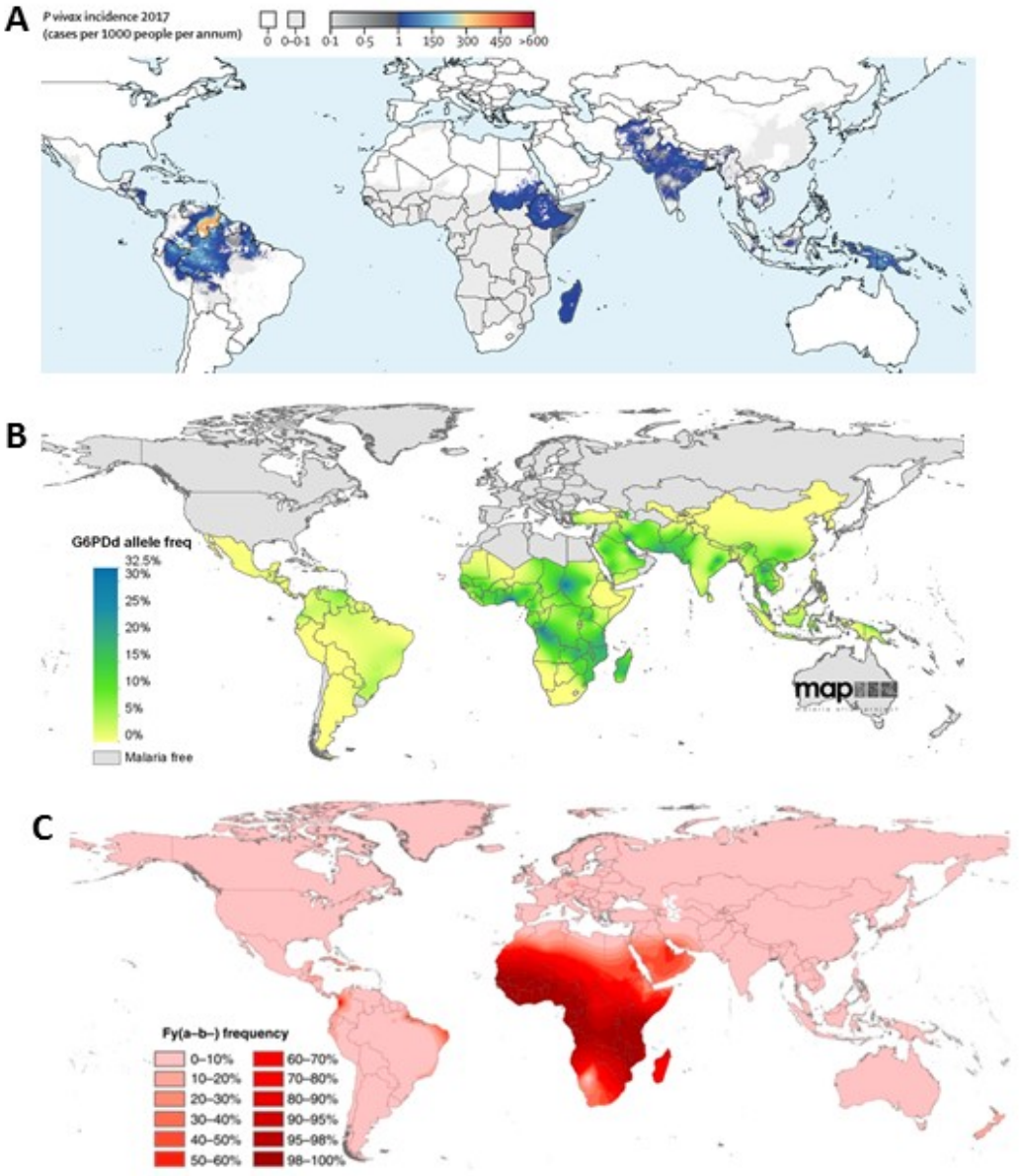


Figure 1.2 - Global Distribution of *P. vivax* Incidence, Glucose-6-Phosphate-Dehydrogenase deficiency allele frequency, and the Duffy-Negative allele frequency: (A) Over 14 million infections of *P. vivax* were predicted across the globe in 2017. However, much of sub-Saharan Africa was predicted to have a very low incidence to no incidence of disease (figure reproduced from Battle *et al.* 2019); (B) The allele conferring glucose-6-phosphate dehydrogenase (G6PD) deficiency is highly prevalent in sub-Saharan Africa. This high prevalence of the G6PD deficient phenotype poses a paradox for *P. vivax* prevalence and treatment, as 8-aminoquinoline drugs needed for radical cure of *P. vivax* are contraindicated in G6PD deficient individuals (figure reproduced from Howes et al. 2012); (C) Prevalence of the Duffy-negative phenotype ($Fy^{a-,b-}$ phenotype) is highest in regions of sub-Saharan Africa (figure reproduced from Howes et al. 2011).

Plasmodium vivax in Sub-Saharan Africa

The origin, age, and evolutionary history of human *P. vivax* has been debated for a number of years, with two predominant hypotheses: (1) a Southeast Asian origin, likely resulting from a Southeast Asian macaque zoonotic transmission event [82,83] or (2) a sub-Saharan African origin, likely resulting from a non-human ape zoonotic transmission event [51,52,84–87]. Recent evidence suggests that the latter hypothesis is more likely, with a zoonotic transmission from an ancestral *Pan troglodytes* (chimpanzee) lineage to a hominid in sub-Saharan Africa [51,87]. Following this zoonotic transmission, it is hypothesized that there was an "Out-of-Africa" migration event, where a specific lineage of *P. vivax* escaped Africa, spread to Europe, and then seeded Asia and the Americas [51,84,85,87–90].

This "Out-of-Africa" migration of *P. vivax* is supported by human genomics, as the origin of the Duffy-negative mutation (GATA-1, -33 T:C) was dated to approximately 42 kya [91–93]. The Duffy-negative mutation was then shown to have undergone a hard sweep with the allele reaching near fixation in the Africa population at approximately 33 kya [91–93]. This sweep of the Duffy-negative phenotype has one of the highest -- if not highest -- selection coefficients in the human genome [93]. This suggests that this mutation was extremely advantageous and that at one point, *P. vivax* may have exerted a strong effect on the African population's fitness. However, it must be noted that this *P. vivax* "Out-of-Africa" hypothesis is still highly debated and not dated well within the context of human evolution [51,87,94].

Regardless of the origin of *P. vivax*, contemporary strains appear to segregate into 2-5 genomic sub-populations (hereafter referred to as *demes*) depending on the type of genetic sequencing used [37,90,95–97]. When comparing global isolates with whole-genome sequencing, two large demes emerge: (1) an American deme and (2) an Asian Deme [37,95–97]. Among these demes,

finer population-structuring has been recognized with potential introgression and admixture between deme subpopulations and even demes as a whole [37,95–97]. However, these whole-genome sequencing studies all lack robust sampling of *P. vivax* isolates from sub-Saharan Africa.

Similarly, when comparing global isolates with mitochondrial Sanger-sequencing, several demes emerge with population substructuring in Central America, South America, the Atlantic Forest (related to *P. simium*), Africa, Melanasia, Southeast Asia, and East Asia [90]. However, there are few African sequences and most are sourced from travelers returning to Western countries with *P. vivax* [86,90]. As a result, there is a very limited data on *P. vivax* sequences from Africa, particularly sub-Saharan Africa. This prompts the critical need for *P. vivax* sequences from the region to resolve *P. vivax* evolutionary history and contemporary transmission dynamics.

Reports of Plasmodium vivax in Sub-Saharan Africa

Recent reports have documented *P. vivax* infections among Duffy-negative hosts throughout western and central Africa. Among these studies, *P. vivax* has been reported with a prevalence of up to 15.5%, produced symptomatic cases, been identified in circulating mosquitoes, and been identified within red blood cells on standard microscopy [50,63–75,98]. Additionally, a recent meta-analysis of *P. vivax* in Africa combining vector, community, clinical, and traveler surveys demonstrated that the distribution of *P. vivax* in Sub-Saharan Africa may be widespread and not necessarily concentrated in countries with a large Duffy-positive population [98]. Together these results suggest ongoing, active transmission. Given the resilience of *P. vivax* to control and elimination efforts, the presence of *P. vivax* in sub-Saharan Africa is concerning for malaria elimination programs and raises several questions about its reemergence.

Despite these concerns, no studies have systematically evaluated the burden, spatial distribution, or risk factors associated with *P. vivax* in sub-Saharan Africa to date**.** Previous reports on *P. vivax* in the region have only been conducted at a few field sites through convenience sampling and have lacked robust epidemiological analyses [50,63–75]. Convenience sampling may result in bias that is nonrandom and may be difficult to account for in any statistical model (e.g. selection bias can cause bias in any direction). In addition, a study population created through convenience sampling will lack generalizability [99]. As a result, additional analyses with robust study designs are needed to estimate the burden of *P. vivax* across sub-Saharan Africa.

Similarly, no studies have attempted to determine the origins of these infections, which may represent imported cases or an independent sub-Saharan African lineage. Genetic sequencing can be leveraged to identify the origin of the DRC *P. vivax* population by comparing it to other publicly available global isolates.

Aim 1 Summary

P. vivax has the potential to cause significant disease and morbidity if its reemergence is allowed to become a resurgence in sub-Saharan Africa. However, to date, no studies have systematically determined the level of threat that *P. vivax* poses in the region. Identifying the clinical burden of *P. vivax* in Sub-Saharan Africa is critical, as *P. vivax* is considered more difficult to diagnose, treat, and eliminate than *P. falciparum* [20,100,101]. As a result, differentiating *P. vivax* in sub-Saharan Africa as an innocuous threat versus an imposing threat is a public health priority.

Aim 2: Tracing the Genetic Relatedness of *Plasmodium falciparum* in the Democratic Republic of the Congo across Space

P. falciparum Genomics

It is widely accepted that *P. falciparum* emerged as a human pathogen due to a zoonotic transmission between a human host in Africa and a gorilla host within the past 10,000 years [84,102,103]. This relatively recent transmission is consistent with the lack of diversity seen in the *P. falciparum* genome relative to the *P. vivax* genome, as zoonotic jumps typically cause extreme bottlenecks [84,102–104]. Despite a lack of global diversity, *P. falciparum* appears to be separated into three relatively isolated subpopulations, or demes: (1) an American deme; (2) an Asian deme; and (3) an African deme [32,105]. These three demes correlate with regions of high *P. falciparum* prevalence (**Figure 1.3**) [2].

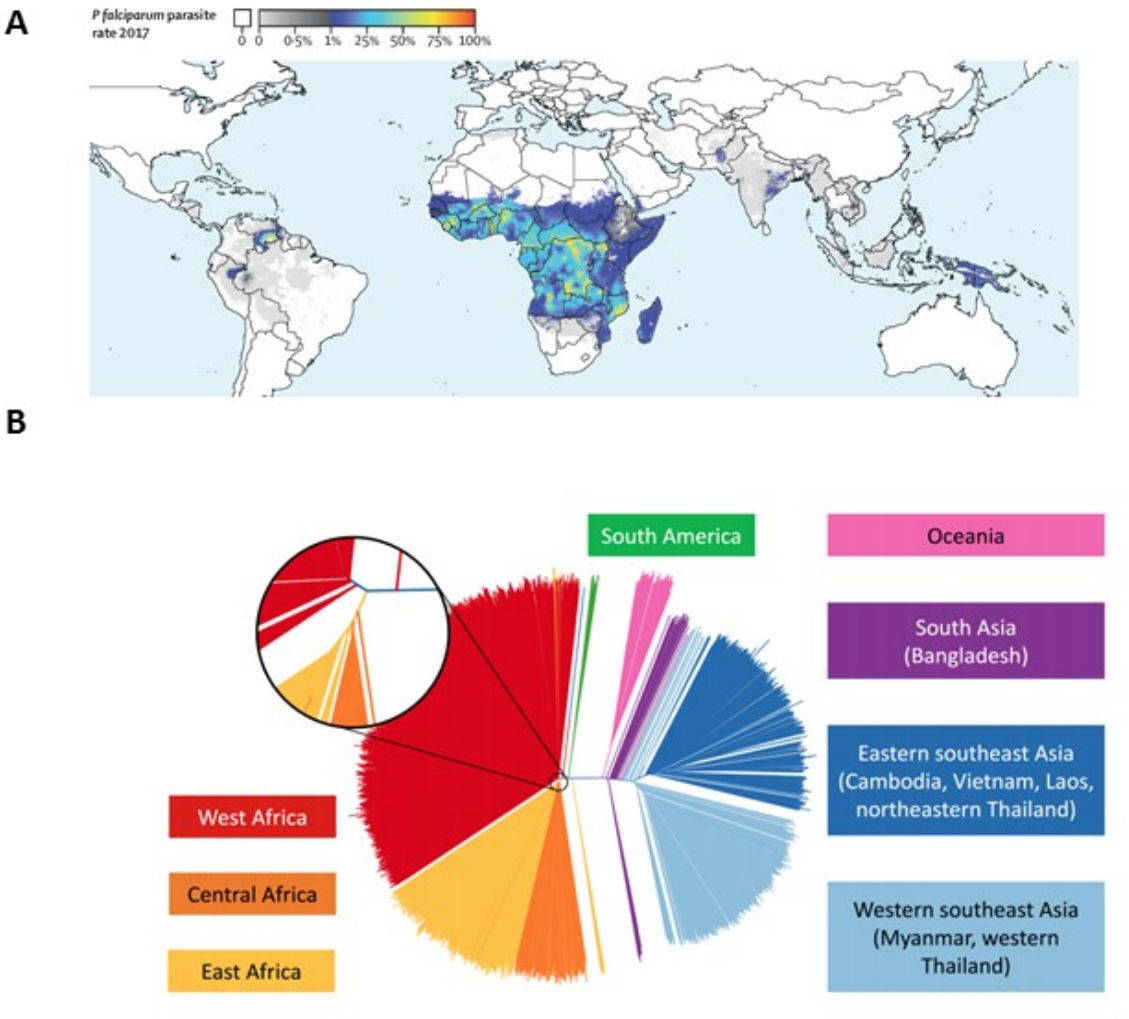


Figure 1.3 - Incidence and Population Structure of Global *P. falciparum*: (A) In 2017, there were 193.9 million predicted *P. falciparum* infections across the globe, with the vast majority occurring in children under ten-years of age in sub-Saharan Africa (Figure reproduced from Weiss *et al.* 2019). (B) *P. falciparum* global population structuring is correlated with regions of transmission (Figure reproduced from Pearson & Amato *et al.* 2019).

Within these three global populations, within-deme population substructuring has largely been driven by selection for drug-resistance mutations [27,106–108]. For example, in Asia, population substructuring has resulted from selection for resistance to artemisinin and ACT partner drugs [27,106–108]. In contrast, in Africa, population substructuring within Africa appears to be driven by resistance to other antimalarial drugs, including CQ, sulfonamides, antifolates, and ACT partner

drugs [109]. These differing patterns may be due to the lack of artemisinin-resistance in Africa to date [30–32]. Additional population substructuring in Africa appears to be due to geographical barriers and nonrandom mating among human hosts (identified by linguistic groups) [109,110]. Although there is a degree of deme-isolation in Africa, admixture appears to be frequent, potentially due to contemporary travel or historical migrations of Bantu populations [109,110].

Identity by Descent and Identity by State

Identity by descent (IBD) is the process of inheriting segments of DNA from a common ancestor. Segments of DNA are inherited due to meiotic recombination during successive generations, resulting in an expectation that any two haploid individuals separated by *m* meiosis will have 2^{-m} proportion of their genome be identical by descent (**Figure 1.3**) [111]. Given this straightforward pattern of inheritance, IBD has long been a centerpiece of population genetic studies [111–114].

It is worth noting, that historically IBD has been considered under two different frameworks. In the first framework presented by Gustave Malécot, IBD is defined as the proportion of the genome (or segments) that are inherited by a common ancestor (i.e. identical by descent) and have not been broken by mutation [114,115]. Separately, IBD has been defined as the probability that two individuals reach a common ancestor at some time in the past, termed the most recent common ancestor (MRCA) [114,116–120]. As a result, the latter definition assumes that shared IBD segments and IBD block lengths have accumulated mutations at a rate proportional to the time (i.e. branch length) to the MRCA [114,116–120]. IBD segments that contain multiple mutations are often "deep" events that represent historic recombination and relatedness events. In this dissertation, I will focus on recent IBD events and Malécot's definition of IBD, under the assumption that recent transmission is more relevant for public health.

Identity by state (IBS) -- a similar concept to IBD -- is defined as the number of shared loci between pairs of individuals (also termed runs of homozygosity in some fields) [120]. Given that alleles only need to be identical to considered IBS, IBS include alleles that are IBD as well as alleles that are identical due to mutation, other IBD events, drift/draft, or chance alone (i.e. IBS reflects the population allele frequencies). As a result, depending on the number of alternative alleles at a given loci and the frequencies of those alleles in a population, IBS can be relatively uninformative in determining the relatedness in a population [121]. Given that IBS is not rooted in population genetic theory -- and is not a true estimator of relatedness -- it's utility is largely rooted in its simplicity to calculate [121]. As a result, I elected to not use IBS in my dissertation.

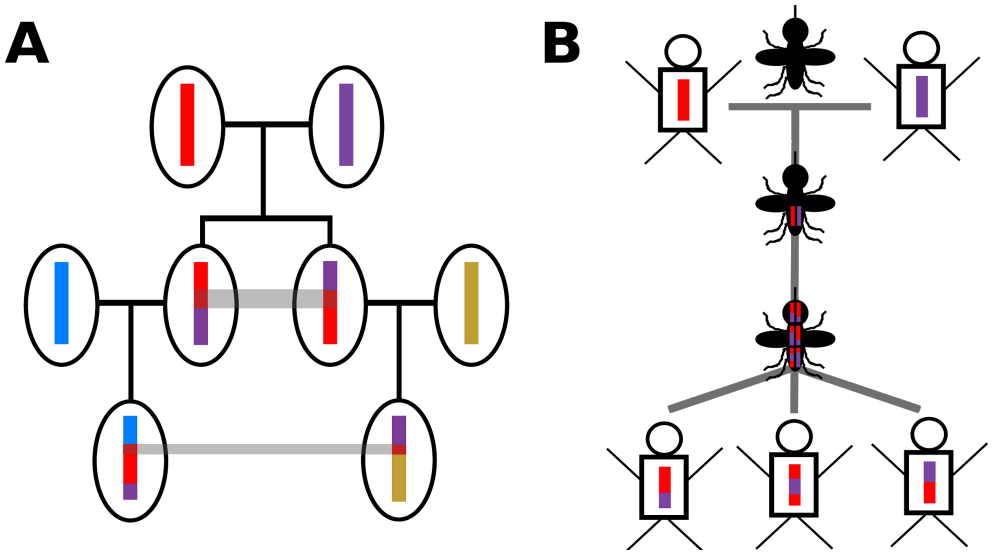


Figure 1.4 - Schematic of Identity by Descent in Malaria: (A) Identity by descent is the process of inheriting segments of DNA from previous generations. These segments can be tracked to determine genetic relatedness among probands. (B) DNA segments are mixed by meiotic recombination. Within the malaria life cycle, meiotic recombination occurs in the mosquito midgut. In the schematic above, a single mosquito bites two infectious hosts with two distinct parasite haplotypes (red; purple). Parasites then undergo recombination to form progeny with new haplotypes that are a mix of their parental haplotypes.

P. falciparum Genetics and IBD

Most infectious agents do not undergo recombination and have a relatively high per-generation mutation rate. This lack of recombination coupled with a high mutation rate allows for direct application of the coalescent and phylogenetics, as mutations are expected to accumulate with respect to the divergence time from the MRCA [114]. This technique of "divergence dating" is widely applied in viral and bacterial studies and is often coupled with epidemiological data for phylodynamic analysis [122,123]. Phylodynamic approaches have been used to model the evolution of HIV, ebola, influenza, and several other infectious diseases and are, arguably, the gold-standard for outbreak investigation and disease surveillance [124–127].

However, phylodynamic methods have remained elusive in the field of malaria genetic epidemiology due to recombination, a relatively low nucleotide mutation rate, and the phenomenon of complexity of infection (COI) [128]. COI is defined as a single host being inoculated with more than one distinct clone or malaria haplotype: a monoclonal infection versus a polyclonal infection. Polyclonal infections arise through two distinct processes: (1) superinfection or (2) cotransmissoin. In the case of superinfection, a single individual receives multiple infectious bites from different mosquitoes that harbor distinct parasites (**Figure 1.5**). In contrast, cotransmission occurs when multiple distinct haplotypes harbored within the mosquito midgut are transferred to the host during one infectious bite (**Figure 1.5**). Superinfections are thought to predominate in high-transmission settings, while cotransmissions are thought to predominate in low-transmission settings [129,130]. However, this assumption of cotransmission relating to low-transmission settings has recently been challenged, as cotransmission events may be common in high transmission settings [131]. Regardless of infection dynamics, polyclonal

infections that are the result of cotransmission are expected to be highly related and frequently meiotic siblings (**Figure 1.5**) [130].

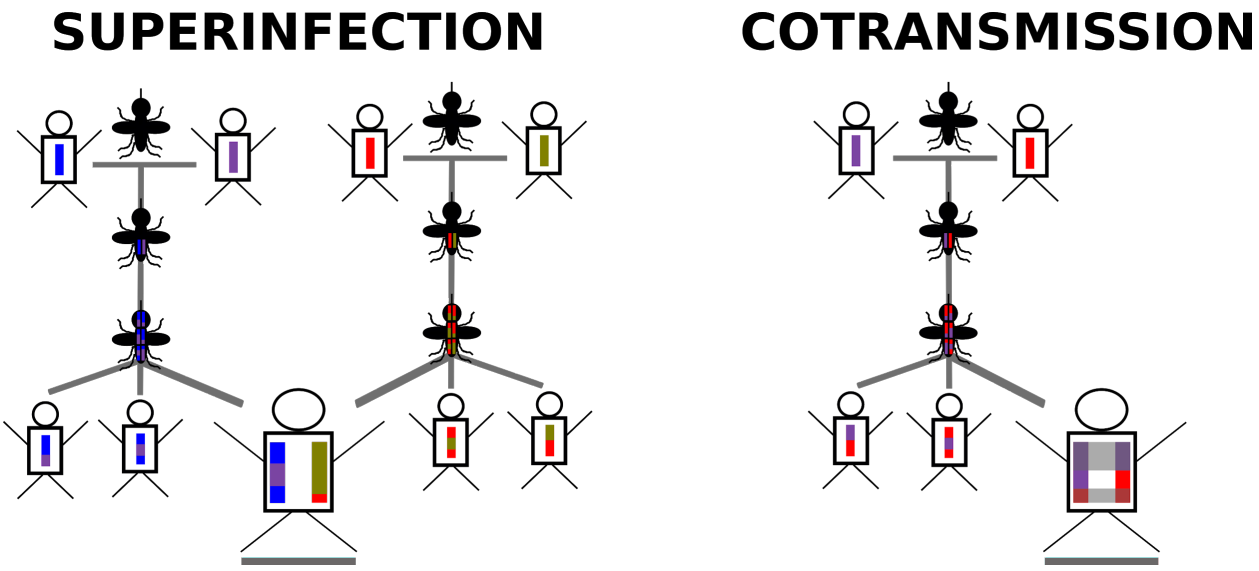


Figure 1.5 - Polyclonality in Malaria: Polyclonality, or multiple malaria haplotype clones within a host arise through two processes: (1) superinfection (left) and (2) cotransmission (right). Superinfection occurs when a host receives more than one infectious bite from mosquitoes with different clones. Cotransmission occurs when multiple distinct clones within the mosquito vector midgut are transferred to the host during a single innocuous bite. Given that parasites undergo recombination in the mosquito midgut, clones that are transferred by cotransmission are expected to be highly related and frequently meiotic siblings. At the top of the figure, parent haplotypes (blue, purple, red) are broken down by recombination to form progeny with new distinct haplotypes that then go on to infect new hosts.

Inferring genetic relatedness is a typical goal in any malaria genetic epidemiology study, as relatedness estimates can be used to identify migration patterns, population demography, and several other applications [121,132]. Recently, there has been an explosion in using IBD measures to quantify genetic relatedness between malaria parasites [30,109,129,130,133–142]. This recent attraction to IBD as a measure of genetic relatedness among malaria parasite is in part due to its ability to capture transmission dynamics and spatial processes [129,137]. For example, recent work has shown that IBD decays at an exponential as prevalence increases (**Figure 1.6**) [129,133]. These results

recapitulate the expected inverse relationship between linkage disequilibrium (LD) and the population level recombination rate, and have been observed in real data among *P. falciparum* parasites (e.g. lower LD in Africa than Southeast Asia) [143–145].

Similarly, recent work has shown that IBD among malaria parasites decays exponentially over geographical space (isolation by distance) and was able to capture spatial patterns that were missed by Wright's F_{st} (**Figure 1.7**) [115,137,146]. IBD is likely a superior estimator of isolation by distance, as space is considered as a continuum, while Wright's F_{st} assumes discrete populations and discrete space (see Wright, 1949 in reference to "neighborhoods").

Collectively, these results demonstrate the power of IBD to capture spatial and transmission-based processes that are critical in characterizing malaria transmission dynamics. IBD, as an estimator of genetic relatedness, can be leveraged to answer questions pertinent to malaria control efforts, most notably through elucidating the connectedness of malaria parasites.

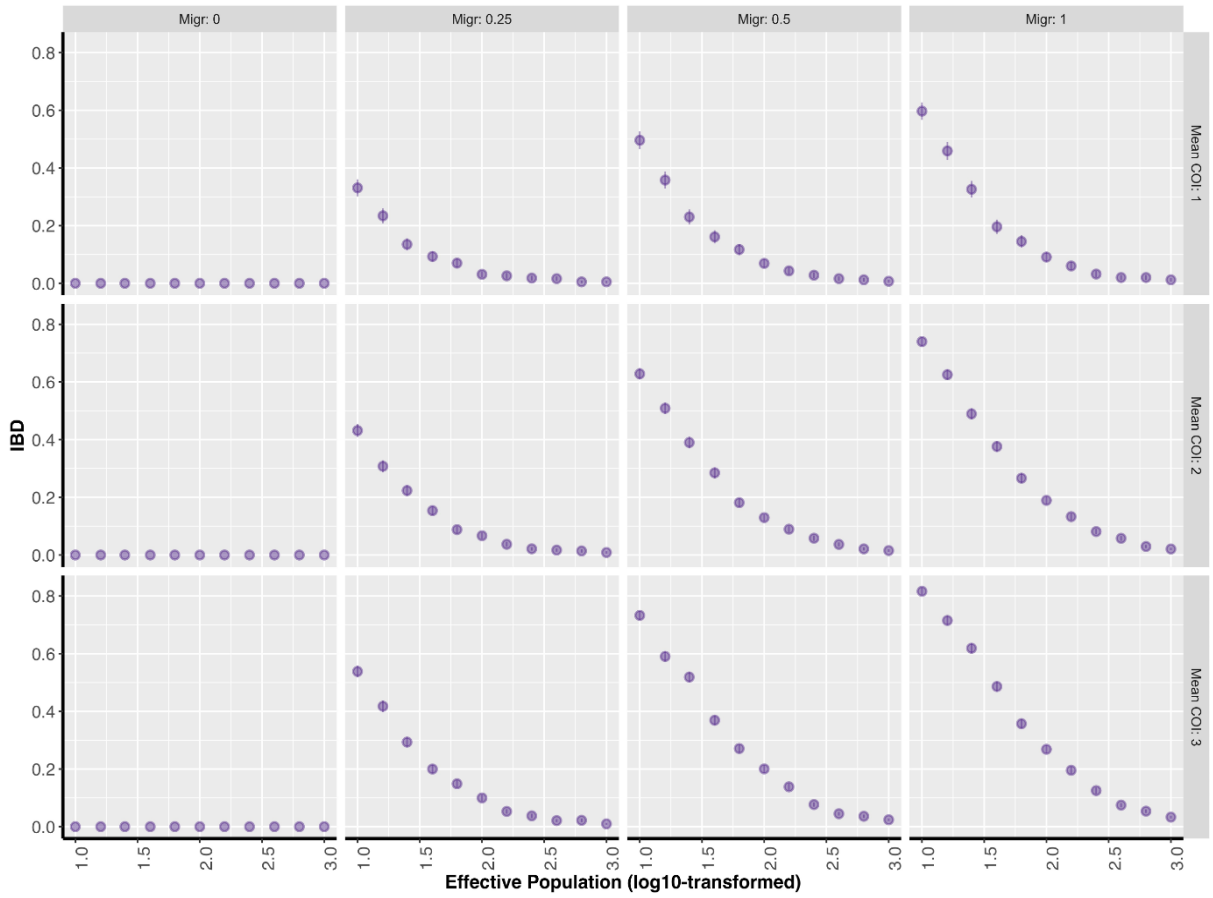


Figure 1.6 - IBD and Transmission Dynamics: Pairwise IBD estimates were generated under a structured Wright-Fisher model. The full mathematical formulation of the model can be found in the Supplementary Materials of Verity, Aydemic, Brazeau *et al.* 2019. As a brief lay summary, the authors assume that each individual host can be represented by a deme, or a subpopulation within a large population. The authors then allow a number of parasites (that reside within the host population) to mate at random with the previous generation of parasites and produce a large number of parasite progeny. During mating, genetic recombination has the potential to occur based on the length of the genome and the recombination rate. Progeny are then allowed to migrate to a new host or stay in the same host at a rate dependent on the number of individuals in the population. Progeny are then culled down to a smaller number of parasites per host by drawing from a Poisson distribution with lambda set to the mean complexity of infection (COI). From the simulations, the proportion of pairs of samples that have any segments of IBD decreases as the effective population size increases. As populations move increasingly towards panmixia (Migration: 1) -- which approximates an exclusively superinfection setting -- the rate of pairwise IBD decay decreases (with the exception of migration set to zero). Finally, as COI increases, the decay in pairwise IBD slows.

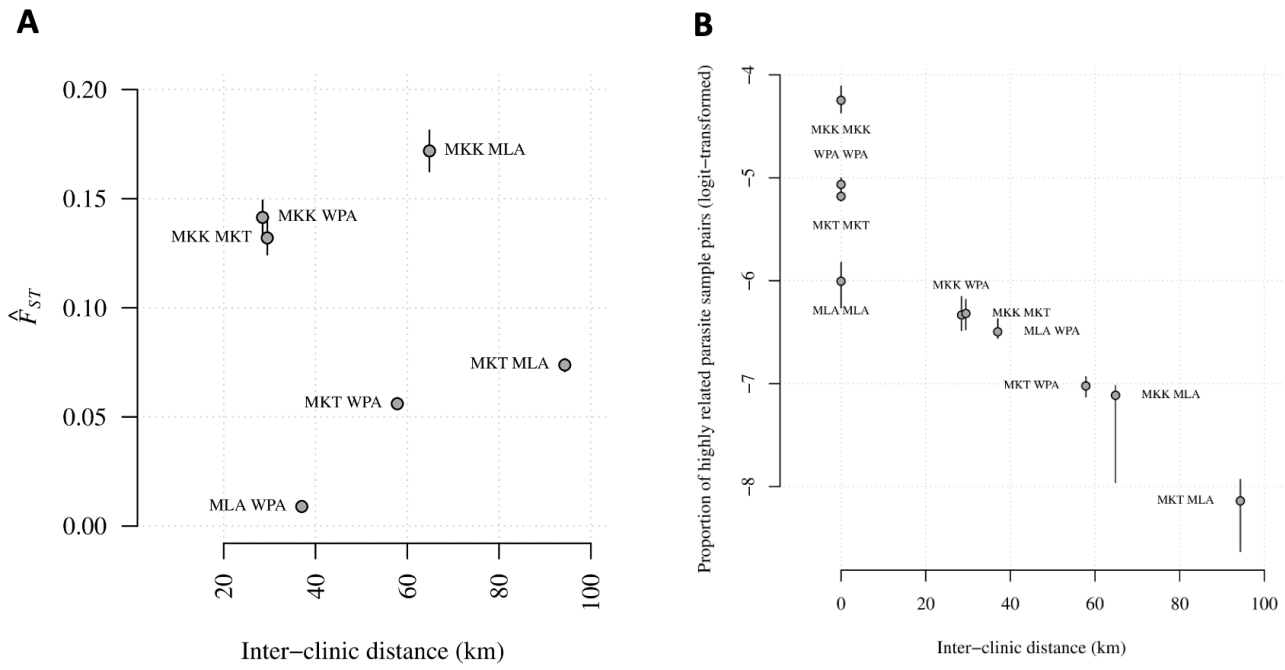


Figure 1.7 - IBD and Spatial Distance: As physical distance increases, genetic relatedness is expected to decrease exponentially, a phenomenon termed isolation by distance. Among malaria parasites, measures of identity by descent (**A**) appear to capture this process of isolation by distance while measures of Wright's F_{st} may not (**B**). Differences between these two estimators of genetic relatedness likely are rooted in differing assumptions of geographical space.

Leveraging IBD in the DRC

Despite increased interventions in the DRC, case-burden and mortality have remained relatively stagnant over the past five years [1,147,148]. Moreover, the World Health Organization (WHO) recently recognized the DRC as a high impact, high burden country that was not on track to meet the goals for malaria control by 2030 laid in the Global Technical Strategy [1,149]. As a result, there is a critical need for new methods and approaches to address the burden and stalled progress of elimination in the DRC.

Recently, Verity, Aydemir, Brazeau *et al.* 2019 analyzed over 1,100 *P. falciparum* isolates sequenced at approximately 1,800 loci from the DRC. In this study, the authors found that the DRC acts as a watershed region, or bridge, between West and East African falciparum

parasites [30]. When focusing in on the DRC, the authors found that allele frequencies among falciparum followed North-South and East-West spatial clines [30]. The East-West spatial clines appeared to be strongly associated with differences in drug resistance mutations among East versus West parasites, particularly at the *pfcrt* and *dhps* loci [30]. These allele frequency clines likely indicate events at least 6-12 generations in the past based on the extended haplotype heterozygosity segment lengths [150]. In addition to these "historic" relationships, they also found three samples with pairwise IBD along at least 90% of the genome [30]. Based on the visualization of pairs terminal points, the authors hypothesized that these highly related pairs may be due to travel along the Congo river [30].

In this dissertation, I will further investigate this pattern of IBD in the DRC and determine the connectedness of *P. falciparum* infections across the DRC. By identifying regions of high parasite connectedness, I can provide targeted feedback for intervention planning and intervention efforts. Maps of spatial and genetic connectedness differ from traditional incidence or prevalence maps, as they provide a picture of how *P. falciparum* infections may be arising instead of simply where. Targeting the source of *P. falciparum* may be the spark needed to help alleviate the stalled burden of *P. falciparum* malaria in the DRC.

Aim 2 Summary

The case-burden and mortality due to *P. falciparum* in the DRC has remained relatively constant over the past five years despite numerous intervention roll-outs and campaigns [1,147,148]. New methodological approaches are needed to address this stagnated progress. Recent work has shown that IBD captures both the spatial processes and transmission dynamics of *P. falciparum* malaria [129,133,137]. By leveraging a spatially rich-dataset of *P. falciparum* genetics from the DRC,

I analyzed the connectedness of *P. falciparum* infections across the DRC and identified hubs of relatedness that may be prime sites for intervention targeting.

REFERENCES

1. World Health Organization. World Malaria Report 2019. 232.

2. Weiss, D. J. *et al.* Mapping the global prevalence, incidence, and mortality of Plasmodium falciparum, 2000–17: a spatial and temporal modelling study. *Lancet* **394**, 322–331 (2019).

3. Battle, K. E. *et al.* Mapping the global endemicity and clinical burden of Plasmodium vivax, 2000–17: a spatial and temporal modelling study. *Lancet* **394**, 332–343 (2019).

4. Gething, P. W. *et al.* A long neglected world malaria map: Plasmodium vivax endemicity in 2010. *PLoS Negl. Trop. Dis.* **6**, e1814 (2012).

5. Price, R. N. *et al.* Vivax malaria: neglected and not benign. *Am. J. Trop. Med. Hyg.* **77**, 79–87 (2007).

6. Mueller, I. *et al.* Key gaps in the knowledge of Plasmodium vivax, a neglected human malaria parasite. *Lancet Infect. Dis.* **9**, 555–566 (2009).

7. White, N. J. *et al.* Malaria. *Lancet* **383**, 723–735 (2014).

8. Crutcher, J. M. & Hoffman, S. L. Malaria. in *Medical Microbiology* (ed. Baron, S.) (University of Texas Medical Branch at Galveston, 2011).

9. Scully, E. J., Kanjee, U. & Duraisingh, M. T. Molecular interactions governing host-specificity of blood stage malaria parasites. *Curr. Opin. Microbiol.* **40**, 21–31 (2017).

10. Miller, L. H., Baruch, D. I., Marsh, K. & Doumbo, O. K. The pathogenic basis of malaria. *Nature* **415**, 673–679 (2002).

11. Baton, L. A. & Ranford-Cartwright, L. C. How do malaria ookinetes cross the mosquito midgut wall? *Trends Parasitol.* **21**, 22–28 (2005).

12. Sinka, M. E. *et al.* The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. *Parasit. Vectors* **3**, 117 (2010).

13. Sinka, M. E. *et al.* The dominant Anopheles vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic précis. *Parasit. Vectors* **3**, 72 (2010).

14. Sinka, M. E. *et al.* The dominant Anopheles vectors of human malaria in the Asia-Pacific region: occurrence data, distribution maps and bionomic précis. *Parasit. Vectors* **4**, 89 (2011).

15. White, N. J. Determinants of relapse periodicity in Plasmodium vivax malaria. *Malar. J.* **10**, 297 (2011).

16. Organization, W. H. & Others. *Response plan to pfhrp2 gene deletions*. https://apps.who.int/iris/bitstream/handle/10665/325528/WHO-CDS-GMP-2019.02-eng.pdf (2019).

17. Watson, O. J. *et al.* Modelling the drivers of the spread of Plasmodium falciparum hrp2 gene deletions in sub-Saharan Africa. *Elife* **6**, (2017).

18. Watson, O. J. *et al.* Impact of seasonal variations in Plasmodium falciparum malaria transmission on the surveillance of pfhrp2 gene deletions. *Elife* **8**, (2019).

19. Parr, J. B. *et al.* Pfhrp2-Deleted Plasmodium falciparum Parasites in the Democratic Republic of the Congo: A National Cross-sectional Survey. *J. Infect. Dis.* (2016) doi:10.1093/infdis/jiw538.

20. Baird, J. K., Valecha, N., Duparc, S., White, N. J. & Price, R. N. Diagnosis and Treatment of Plasmodium vivax Malaria. *Am. J. Trop. Med. Hyg.* **95**, 35–51 (2016).

21. Abba, K. *et al.* Rapid diagnostic tests for diagnosing uncomplicated non-falciparum or Plasmodium vivax malaria in endemic countries. *Cochrane Database Syst. Rev.* CD011431 (2014) doi:10.1002/14651858.CD011431.

22. White, M. T. *et al.* Modelling the contribution of the hypnozoite reservoir to Plasmodium vivax transmission. *Elife* **3**, (2014).

23. Wells, T. N. C., Burrows, J. N. & Baird, J. K. Targeting the hypnozoite reservoir of Plasmodium vivax: the hidden obstacle to malaria elimination. *Trends Parasitol.* **26**, 145–151 (2010).

24. Haldar, K., Bhattacharjee, S. & Safeukui, I. Drug resistance in Plasmodium. *Nat. Rev. Microbiol.* **16**, 156–170 (2018).

25. Dondorp, A. M. *et al.* Artemisinin resistance in Plasmodium falciparum malaria. *N. Engl. J. Med.* **361**, 455–467 (2009).

26. Organization, W. H. & Others. *Artemisinin resistance and artemisinin-based combination therapy efficacy: status report*. https://apps.who.int/iris/bitstream/handle/10665/274362/WHO-CDS-GMP-2018.18-eng.pdf (2018).

27. Parobek, C. M. *et al.* Partner-Drug Resistance and Population Substructuring of Artemisinin-Resistant Plasmodium falciparum in Cambodia. *Genome Biol. Evol.* **9**, 1673–1686 (2017).

28. Lin, J. T., Juliano, J. J. & Wongsrichanalai, C. Drug-Resistant Malaria: The Era of ACT. *Curr. Infect. Dis. Rep.* **12**, 165–173 (2010).

29. White, N. J. The treatment of malaria. *N. Engl. J. Med.* **335**, 800–806 (1996).

30. Verity, R. J., Aydemir, O., Brazeau, N. F. & Watson, O. J. The Impact of Antimalarial Resistance on the Genetic Structure of Plasmodium falciparum in the DRC. *bioRxiv* (2019).

31. Taylor, S. M. *et al.* Absence of putative artemisinin resistance mutations among Plasmodium falciparum in Sub-Saharan Africa: a molecular epidemiologic study. *J. Infect. Dis.* **211**, 680–688 (2015).

32. MalariaGEN Plasmodium falciparum Community Project. Genomic epidemiology of artemisinin resistant malaria. *Elife* **5**, (2016).

33. Ecker, A., Lehane, A. M., Clain, J. & Fidock, D. A. PfCRT and its role in antimalarial drug resistance. *Trends Parasitol.* **28**, 504–514 (2012).

34. Sidhu, A. B. S., Verdier-Pinard, D. & Fidock, D. A. Chloroquine resistance in Plasmodium falciparum malaria parasites conferred by pfcrt mutations. *Science* **298**, 210–213 (2002).

35. World Health Organization. *Guidelines for the Treatment of Malaria. Third Edition*. (World Health Organization, 2015).

36. NEJM Journal Watch: Summaries of and commentary on original medical and scientific articles from key medical journals. https://www.jwatch.org/na47230/2018/08/14/tafenoquine-krintafel-approved-prevention-relapse-vivax.

37. Hupalo, D. N. *et al.* Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax. *Nat. Genet.* (2016) doi:10.1038/ng.3588.

38. Brazeau, N. F. *et al.* Longitudinal Pooled Deep Sequencing of the Plasmodium vivax K12 Kelch Gene in Cambodia Reveals a Lack of Selection by Artemisinin. *Am. J. Trop. Med. Hyg.* **95**, 1409–1412 (2016).

39. Fernando, D., Rodrigo, C. & Rajapakse, S. Primaquine in vivax malaria: an update and review on management issues. *Malar. J.* **10**, 351 (2011).

40. Bruce, M. C. *et al.* Cross-species interactions between malaria parasites in humans. *Science* **287**, 845–848 (2000).

41. Douglas, N. M. *et al.* Plasmodium vivax recurrence following falciparum and mixed species malaria: risk factors and effect of antimalarial kinetics. *Clin. Infect. Dis.* **52**, 612–620 (2011).

42. Shanks, G. D. & White, N. J. The activation of vivax malaria hypnozoites by infectious diseases. *Lancet Infect. Dis.* **13**, 900–906 (2013).

43. Lin, J. T. *et al.* Plasmodium falciparum gametocyte carriage is associated with subsequent Plasmodium vivax relapse after treatment. *PLoS One* **6**, e18716 (2011).

44. Boyd, M. F., Muench, H. & Stratman-Thomas, W. K. The Occurrence of Gametocytes of Plasmodium Vivax during the Primary Attack 1. *Am. J. Trop. Med. Hyg.* **s1-16**, 133–138 (1936).

45. Vallejo, A. F., García, J., Amado-Garavito, A. B., Arévalo-Herrera, M. & Herrera, S. Plasmodium vivax gametocyte infectivity in sub-microscopic infections. *Malar. J.* **15**, 48 (2016).

46. Jeffery, G. M. The infection of mosquitoes by Plasmodium vivax (Chesson strain) during the early primary parasitemias. *Am. J. Trop. Med. Hyg.* **1**, 612–617 (1952).

47. Boyd, M. F. & Kitchen, S. F. On the Infectiousness of Patients Infected with Plasmodium Vivax and Plasmodium Falciparum 1. *Am. J. Trop. Med. Hyg.* **s1-17**, 253–262 (1937).

48. Ellis McKenzie, F. *et al.* Gametocytemia in Plasmodium Vivax and Plasmodium Falciparum Infections. *J. Parasitol.* **92**, 1281 (2006).

49. Olliaro, P. L. *et al.* Implications of Plasmodium vivax Biology for Control, Elimination, and Research. *Am. J. Trop. Med. Hyg.* **95**, 4–14 (2016).

50. Gunalan, K. *et al.* Role of Plasmodium vivax Duffy-binding protein 1 in invasion of Duffy-null Africans. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6271–6276 (2016).

51. Liu, W. *et al.* African origin of the malaria parasite Plasmodium vivax. *Nat. Commun.* **5**, 3346 (2014).

52. Prugnolle, F. *et al.* Diversity, host switching and evolution of Plasmodium vivax infecting African great apes. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 8123–8128 (2013).

53. Gunalan, K., Niangaly, A., Thera, M. A., Doumbo, O. K. & Miller, L. H. Plasmodium vivax Infections of Duffy-Negative Erythrocytes: Historically Undetected or a Recent Adaptation? *Trends Parasitol.* **34**, 420–429 (2018).

54. Tournamille, C., Colin, Y., Cartron, J. P. & Van Kim, C. L. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy–negative individuals. *Nat. Genet.* **10**, 224–228 (1995).

55. Miller, L. H., Mason, S. J. & Clyde, D. F. The resistance factor to Plasmodium vivax in blacks: the Duffy-blood-group genotype, FyFy. *New England Journal* (1976).

56. Miller, L. H., Mason, S. J., Dvorak, J. A., McGinniss, M. H. & Rothman, I. K. Erythrocyte receptors for (Plasmodium knowlesi) malaria: Duffy blood group determinants. *Science* **189**, 561–563 (1975).

57. Miller, L. H., McAuliffe, F. M. & Mason, S. J. Erythrocyte receptors for malaria merozoites. *Am. J. Trop. Med. Hyg.* **26**, 204–208 (1977).

58. Miller, L. H., Aikawa, M., Johnson, J. G. & Shiroishi, T. Interaction between cytochalasin B-treated malarial parasites and erythrocytes. Attachment and junction formation. *J. Exp. Med.* **149**, 172–184 (1979).

59. França, C. T. *et al.* Plasmodium vivax Reticulocyte Binding Proteins Are Key Targets of Naturally Acquired Immunity in Young Papua New Guinean Children. *PLoS Negl. Trop. Dis.* **10**, e0005014 (2016).

60. Gupta, S. *et al.* Targeting a Reticulocyte Binding Protein and Duffy Binding Protein to Inhibit Reticulocyte Invasion by Plasmodium vivax. *Sci. Rep.* **8**, 10511 (2018).

61. Cutbush, M. & Mollison, P. L. The Duffy blood group system. *Heredity* **4**, 383–389 (1950).

62. Meny, G. M. The Duffy blood group system: a review. *Immunohematology* **26**, 51–56 (2010).

63. Mendes, C. *et al.* Duffy negative antigen is no longer a barrier to Plasmodium vivax--molecular evidences from the African West Coast (Angola and Equatorial Guinea). *PLoS Negl. Trop. Dis.* **5**, e1192 (2011).

64. Poirier, P. *et al.* The hide and seek of Plasmodium vivax in West Africa: report from a large-scale study in Beninese asymptomatic subjects. *Malar. J.* **15**, 570 (2016).

65. Motshoge, T. *et al.* Molecular evidence of high rates of asymptomatic P. vivax infection and very low P. falciparum malaria in Botswana. *BMC Infect. Dis.* **16**, 520 (2016).

66. Ngassa Mbenda, H. G. & Das, A. Molecular evidence of Plasmodium vivax mono and mixed malaria parasite infections in Duffy-negative native Cameroonians. *PLoS One* **9**, e103262 (2014).

67. Russo, G. *et al.* Molecular evidence of Plasmodium vivax infection in Duffy negative symptomatic individuals from Dschang, West Cameroon. *Malar. J.* **16**, 74 (2017).

68. Woldearegai, T. G., Kremsner, P. G., Kun, J. F. J. & Mordmüller, B. Plasmodium vivax malaria in Duffy-negative individuals from Ethiopia. *Trans. R. Soc. Trop. Med. Hyg.* **107**, 328–331 (2013).

69. Ryan, J. R. *et al.* Evidence for transmission of Plasmodium vivax among a duffy antigen negative population in Western Kenya. *Am. J. Trop. Med. Hyg.* **75**, 575–581 (2006).

70. Ménard, D. *et al.* Plasmodium vivax clinical malaria is commonly observed in Duffy-negative Malagasy people. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5967–5971 (2010).

71. Niangaly, A. *et al.* Plasmodium vivax Infections over 3 Years in Duffy Blood Group Negative Malians in Bandiagara, Mali. *Am. J. Trop. Med. Hyg.* **97**, 744–752 (2017).

72. Wurtz, N. *et al.* Vivax malaria in Mauritania includes infection of a Duffy-negative individual. *Malar. J.* **10**, 336 (2011).

73. Abdelraheem, M. H., Albsheer, M. M. A., Mohamed, H. S., Amin, M. & Mahdi Abdel Hamid, M. Transmission of Plasmodium vivax in Duffy-negative individuals in central Sudan. *Trans. R. Soc. Trop. Med. Hyg.* **110**, 258–260 (2016).

74. Culleton, R. *et al.* Evidence for the transmission of Plasmodium vivax in the Republic of the Congo, West Central Africa. *J. Infect. Dis.* **200**, 1465–1469 (2009).

75. Zimmerman, P. A. Plasmodium vivax Infection in Duffy-Negative People in Africa. *Am. J. Trop. Med. Hyg.* **97**, 636–638 (2017).

76. Hester, J. *et al.* De novo assembly of a field isolate genome reveals novel Plasmodium vivax erythrocyte invasion genes. *PLoS Negl. Trop. Dis.* **7**, e2569 (2013).

77. Menard, D. *et al.* Whole genome sequencing of field isolates reveals a common duplication of the Duffy binding protein gene in Malagasy Plasmodium vivax strains. *PLoS Negl. Trop. Dis.* **7**, e2489 (2013).

78. Ntumngia, F. B. *et al.* A Novel Erythrocyte Binding Protein of Plasmodium vivax Suggests an Alternate Invasion Pathway into Duffy-Positive Reticulocytes. *MBio* **7**, (2016).

79. Hostetler, J. B. *et al.* Independent Origin and Global Distribution of Distinct Plasmodium vivax Duffy Binding Protein Gene Duplications. *PLoS Negl. Trop. Dis.* **10**, e0005091 (2016).

80. Gunalan, K. *et al.* Transcriptome profiling of Plasmodium vivax in Saimiri monkeys identifies potential ligands for invasion. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 7053–7061 (2019).

81. Howes, R. E. *et al.* G6PD deficiency prevalence and estimates of affected populations in malaria endemic countries: a geostatistical model-based map. *PLoS Med.* **9**, e1001339 (2012).

82. Taylor, J. E. *et al.* The evolutionary history of Plasmodium vivax as inferred from mitochondrial genomes: parasite genetic diversity in the Americas. *Mol. Biol. Evol.* **30**, 2050–2064 (2013).

83. Cornejo, O. E. & Escalante, A. A. The origin and age of Plasmodium vivax. *Trends Parasitol.* **22**, 558–563 (2006).

84. Loy, D. E. *et al.* Out of Africa: origins and evolution of the human malaria parasites

Plasmodium falciparum and Plasmodium vivax. *Int. J. Parasitol.* (2016) doi:10.1016/j.ijpara.2016.05.008.

85. Jongwutiwes, S. *et al.* Mitochondrial Genome Sequences Support Ancient Population Expansion in Plasmodium vivax. *Mol. Biol. Evol.* **22**, 1733–1739 (2005).

86. Culleton, R. *et al.* The origins of African Plasmodium vivax; insights from mitochondrial genome sequencing. *PLoS One* **6**, e29137 (2011).

87. Loy, D. E. *et al.* Evolutionary history of human Plasmodium vivax revealed by genome-wide analyses of related ape parasites. *Proceedings of the National Academy of Sciences* **115**, E8450–E8459 (2018).

88. Gelabert, P. *et al.* Mitochondrial DNA from the eradicated European Plasmodium vivax and P. falciparum from 70-year-old slides from the Ebro Delta in Spain. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11495–11500 (2016).

89. Carter, R. Speculations on the origins of Plasmodium vivax malaria. *Trends Parasitol.* **19**, 214–219 (2003).

90. Rodrigues, P. T. *et al.* Human migration and the spread of malaria parasites to the New World. *Sci. Rep.* **8**, 1993 (2018).

91. Hamblin, M. T., Thompson, E. E. & Di Rienzo, A. Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* **70**, 369–383 (2002).

92. Hamblin, M. T. & Di Rienzo, A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* **66**, 1669–1679 (2000).

93. McManus, K. F. *et al.* Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. *PLoS Genet.* **13**, e1006560 (2017).

94. Gilabert, A. *et al.* Plasmodium vivax-like genome sequences shed new insights into Plasmodium vivax biology and evolution. *PLoS Biol.* **16**, e2006035 (2018).

95. Pearson, R. D. *et al.* Genomic analysis of local variation and recent evolution in Plasmodium vivax. *Nat. Genet.* **48**, 959–964 (2016).

96. Cowell, A. N., Valdivia, H. O., Bishop, D. K. & Winzeler, E. A. Exploration of Plasmodium vivax transmission dynamics and recurrent infections in the Peruvian Amazon using whole genome sequencing. *Genome Med.* **10**, 52 (2018).

97. Auburn, S. *et al.* Genomic analysis of a pre-elimination Malaysian Plasmodium vivax population reveals selective pressures and changing transmission dynamics. *Nat. Commun.* **9**, 2585 (2018).

98. Twohig, K. A. *et al.* Growing evidence of Plasmodium vivax across malaria-endemic Africa. *PLoS Negl. Trop. Dis.* **13**, e0007140 (2019).

99. Rothman, K., Greenland, S. & Lash, T. L. Modern Epidemiology, 3rd Edition: (2008).

100. Kevin Baird, J. Malaria caused by Plasmodium vivax: recurrent, difficult to treat, disabling, and threatening to life--the infectious bite preempts these hazards. *Pathog. Glob. Health* **107**, 475–479 (2013).

101. Gruenberg, M. *et al.* Plasmodium vivax molecular diagnostics in community surveys: pitfalls and solutions. *Malar. J.* **17**, 55 (2018).

102. Sundararaman, S. A. *et al.* Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria. *Nat. Commun.* **7**, 11078 (2016).

103. Liu, W. *et al.* Origin of the human malaria parasite Plasmodium falciparum in gorillas. *Nature* **467**, 420–425 (2010).

104. Neafsey, D. E. *et al.* The malaria parasite Plasmodium vivax exhibits greater genetic diversity than Plasmodium falciparum. *Nat. Genet.* **44**, 1046–1050 (2012).

105. Manske, M. *et al.* Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature* **487**, 375–379 (2012).

106. Zhu, L. *et al.* The origins of malaria artemisinin resistance defined by a genetic and transcriptomic background. *Nat. Commun.* **9**, 5158 (2018).

107. Miotto, O. *et al.* Genetic architecture of artemisinin-resistant Plasmodium falciparum. *Nat. Genet.* **47**, 226–234 (2015).

108. Amato, R. *et al.* Origins of the current outbreak of multidrug-resistant malaria in southeast Asia: a retrospective genetic study. *Lancet Infect. Dis.* **18**, 337–345 (2018).

109. Amambua-Ngwa, A. *et al.* Major subpopulations of Plasmodium falciparum in sub-Saharan Africa. *Science* **365**, 813–816 (2019).

110. Busby, G. B. J. *et al.* Admixture into and within sub-Saharan Africa. *eLife* vol. 5 (2016).

111. Browning, S. R. & Browning, B. L. Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics* vol. 46 617–633 (2012).

112. Thompson, E. A. Correlations between relatives: From Mendelian theory to complete genome sequence. *Genet. Epidemiol.* **43**, 577–591 (2019).

113. Weir, B. S., Anderson, A. D. & Hepler, A. B. Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* **7**, 771–780 (2006).

114. Wakeley, J. *Coalescent theory: an introduction*. (2009).

115. Malécot, G. *The Mathematics of Heredity*. (W. H. Freeman, 1970).

116. Palamara, P. F. Population genetics of identity by descent. *arXiv [q-bio.PE]* (2014).

117. Gillespie, J. H. *Population Genetics: A Concise Guide*. (JHU Press, 2004).

118. Hartl, D. L. & Others. *A primer of population genetics*. (Sinauer Associates, Inc., 1988).

119. Hartl, D. L., Clark, A. G. & Clark, A. G. *Principles of population genetics*. vol. 116 (Sinauer associates Sunderland, MA, 1997).

120. Hahn, M. W. *Molecular Population Genetics*. (Oxford University Press, 2018).

121. Taylor, A. R., Jacob, P. E., Neafsey, D. E. & Buckee, C. O. Estimating Relatedness Between Malaria Parasites. *Genetics* **212**, 1337–1351 (2019).

122. Grenfell, B. T. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science* vol. 303 327–332 (2004).

123. Kühnert, D., Wu, C.-H. & Drummond, A. J. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infect. Genet. Evol.* **11**, 1825–1841 (2011).

124. Koelle, K., Cobey, S., Grenfell, B. & Pascual, M. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science* **314**, 1898–1903 (2006).

125. Hughes, G. J. *et al.* Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. *PLoS Pathog.* **5**, e1000590 (2009).

126. Gray, R. R. *et al.* Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. *AIDS* vol. 23 F9–F17 (2009).

127. Volz, E. & Pond, S. Phylodynamic analysis of ebola virus in the 2014 sierra leone epidemic. *PLoS Curr.* **6**, (2014).

128. Miles, A. *et al.* Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. *Genome Research* vol. 26 1288–1299 (2016).

129. Watson, O. J., Okell, L. C., Joel, H., Slater, H. & Unwin, H. J. T. Evaluating the performance of malaria genomics for inferring changes in transmission intensity using transmission modelling. *bioRxiv* (2019).

130. Wong, W., Wenger, E. A., Hartl, D. L. & Wirth, D. F. Modeling the genetic relatedness of Plasmodium falciparum parasites following meiotic recombination and cotransmission.

PLoS Comput. Biol. **14**, e1005923 (2018).

131. Nkhoma, S. C. *et al.* Co-transmission of Related Malaria Parasite Lineages Shapes Within-Host Parasite Diversity. *Cell Host Microbe* (2019) doi:10.1016/j.chom.2019.12.001.

132. Wesolowski, A. *et al.* Mapping malaria by combining parasite genomic and epidemiologic data. *BMC Med.* **16**, 190 (2018).

133. Zhu, S. J. *et al.* The origins and relatedness structure of mixed infections vary with local prevalence of P. falciparum malaria. *Elife* **8**, (2019).

134. Daniels, R. *et al.* Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. *PLoS One* **8**, e60780 (2013).

135. Omedo, I. *et al.* Micro-epidemiological structuring of Plasmodium falciparum parasite populations in regions with varying transmission intensities in Africa. *Wellcome Open Res* **2**, 10 (2017).

136. Daniels, R. F. *et al.* Modeling malaria genomics reveals transmission decline and rebound in Senegal. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7067–7072 (2015).

137. Taylor, A. R. *et al.* Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent. *PLoS Genet.* **13**, e1007065 (2017).

138. Andrew P Morgan, Nicholas F Brazeau, Billy Ngasala, Lwidiko Edward, Madeline Denton, Ulrika Morris, Ozkan Aydemir, Jeffrey A. Bailey, Jonathan Parr, Andreas Mårtensson, Anders Bjorkman, Jonathan J Juliano. Falciparum malaria from coastal Tanzania and Zanzibar remains highly connected despite effective control efforts on the archipelago.

139. Schaffner, S. F., Taylor, A. R., Wong, W., Wirth, D. F. & Neafsey, D. E. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *bioRxiv* 188078 (2017) doi:10.1101/188078.

140. Henden, L., Lee, S., Mueller, I., Barry, A. & Bahlo, M. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLoS Genet.* **14**, e1007279 (2018).

141. Shetty, A. C. *et al.* Genomic structure and diversity of Plasmodium falciparum in Southeast Asia reveal recent parasite migration patterns. *Nat. Commun.* **10**, 2665 (2019).

142. Miotto, O. *et al.* Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia. *Nat. Genet.* **45**, 648–655 (2013).

143. Mu, J. *et al.* Recombination hotspots and population structure in Plasmodium falciparum. *PLoS Biol.* **3**, e335 (2005).

144. Neafsey, D. E. *et al.* Genome-wide SNP genotyping highlights the role of natural selection in Plasmodium falciparum population divergence. *Genome Biol.* **9**, R171 (2008).

145. Volkman, S. K. *et al.* A genome-wide map of diversity in Plasmodium falciparum. *Nat. Genet.* **39**, 113–119 (2007).

146. Wright, S. The Genetical Structure of Populations. *Ann. Eugen.* **15**, 323–354 (1949).

147. Taylor, S. M. *et al.* Molecular malaria epidemiology: mapping and burden estimates for the Democratic Republic of the Congo, 2007. *PLoS One* **6**, e16420 (2011).

148. Molly Deutsch-Feldman, Nicholas F. Brazeau, Jonathan B. Parr, Kyaw L. Thwai, Jérémie Muwonga, Melchior Kashamuka, Antoinette K. Tshefu, Jessie K. Edwards, Robert Verity, Michael Emch, Emily W. Gower, Jonathan J. Juliano, Steven R. Meshnick. Spatial and epidemiological drivers of P. falciparum malaria among adults in the Democratic Republic of the Congo.

149. World Health Organization. *Global Technical Strategy for Malaria 2016-2030*. (World Health Organization, 2015).

150. Speed, D. & Balding, D. J. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics* vol. 16 33–44 (2015).

CHAPTER THREE: IDENTIFYING THE RISK, DISTRIBUTION, AND ORIGIN OF P. VIVAX IN THE DEMOCRATIC REPUBLIC OF THE CONGO

Introduction

Plasmodium vivax is the most prevalent malaria-causing parasite outside of Africa, accounting for approximately 14.3 million cases in 2017 [1]. The relative absence of *P. vivax* in Africa has long been attributed to the high prevalence of the Duffy-negative phenotype throughout most of sub-Saharan Africa (SSA) [2–4]. However, recent evidence has demonstrated that *P. vivax* infections are occurring throughout SSA among Duffy-negative hosts [5]. Although these *P. vivax* infections have been associated with clinical cases, the distribution and extent of asymptomatic versus symptomatic disease in SSA remains unclear [1,5,6].

Despite growing concern, no studies have systematically evaluated the burden, risk factors, spatial distribution, or origins of these SSA *P. vivax* infections. This lack of research is problematic as resources have begun to be directed towards diagnosing and addressing SSA *P. vivax*. While the return of *P. vivax* to SSA has the potential to undermine years of malaria control and elimination efforts, its threat-status has not yet been characterized. To address this critical gap in knowledge, I used samples from the Democratic Republic of the Congo (DRC) 2013-14 Demographic Health Survey (DHS) to screen a nationally representative population of over 17,000 adults for *P. vivax*. Surveys from the DHS program are community based and are expected to contain mostly healthy, asymptomatic participants. The DRC is situated in the center of Africa and is the second-largest country in SSA. Moreover, previous work has indicated that the DRC is a watershed region that appears to link East and West Africa malaria [7,8]. As a result, findings from the DRC may be generalizable to much of SSA.

Using this nationally representative survey, I provide the first national level estimate of *P. vivax* prevalence, associated risk factors, and the geographical distribution of cases in the SSA region. In addition, I use mitochondrial genomes to identify the potential origin of these infections. By coupling a nationally-representative, spatially-rich dataset with cutting edge spatial statistics, novel machine learning techniques, and genomics, I advance efforts to uncover the hidden distribution and history of *P. vivax* in SSA.

Methods

Study Participants & Malaria Detection

I studied men and women aged 15 - 59 years and 15 - 49 years, respectively, that were surveyed in the 2013-2014 DRC DHS. Each participant answered an extensive questionnaire and provided a dried blood spot (DBS) for HIV and other biomarker screening. Spatial and ecological data were collected for each sampling cluster (**Appendix 3.1**). DNA was extracted from each DBS using Chelex-100 (Bio-Rad, Hercules, CA) and Saponin and then screened all participants for *P. falciparum* using quantitative PCR (qPCR) targeting the *P. falciparum* lactate dehydrogenase gene as previously described [9]. In addition, samples were screened for *P. vivax* using qPCR targeting the 18S ribosomal RNA gene [10]. Samples that screened positive by 18S-qPCR underwent reflex confirmatory screening using a nested-PCR assay targeting 18S rRNA (**Appendix 3.1**) [11]. To ensure the quality of DNA extraction, I excluded samples that failed to amplify human-beta-tubulin from analysis. Finally, participants were excluded if they had missing data or were not a part of the DHS sampling schematic (**Appendix 3.1 Figure 2**) [12]. This study reanalyzes previously published *P. falciparum* data (sample size differences are due to different inclusion criteria) [9].

Duffy Genotyping

Host Duffy antigen/chemokine receptor (DARC) genotype was determined using a previously validated High-Resolution Melt (HRM) assay [13]. Genotypes that could not be definitely resolved by HRM were reconciled by Sanger sequencing [6]. In addition, HRM results were validated by sequencing approximately 10% samples (**Appendix 3.1**).

Risk Factor Modeling

P .vivax risk factors were identified from a comprehensive literature search and previous work from the 2013-2014 DRC DHS identifying *P. falciparum* risk factors [9]. Risk factors were derived from the DHS questionnaires and other open-data sources (**Appendix 3.1**). All continuous risk factors were standardized in order to promote model stability and ease of comparability. For dichotomized risk factors, the *a priori* protective level was selected as the referent level (e.g. HIV-negative) or the largest group if a protective level was not obvious (e.g. female for biological sex).

For each risk factor, confounding covariates were identified using a directed acyclic diagram (DAG) built from an *a priori* causal framework of covariate and outcome relationships (**Appendix 3.1 Figure 3**). I then used inverse-probability weighting (IPW) to obtain marginal structural models and control for confounding between the risk factors and outcome of interest, malaria. IPWs were calculated with a super learning algorithm, which uses a loss-based approach with V-fold cross-validation to maximize predictions from an ensemble of candidate algorithms [14]. I extended the standard super learning algorithm to account for spatial dependence among observations using spatial cross validation (**Appendix 3.1**) [15]. The super learner algorithm was selected for IPW calculations to account for known issues and biases of functional form in fitting the exposure dose-response curve [16]. Using the IPWs, I performed weighted regression using

generalized estimating equations (GEE) with a logit-link function and binomial variance. IPWs and DHS sampling weights were accounted for in the GEE under the assumption that the distribution of the sampling was independent of the distribution of confounding covariates, which allows for weights to be considered jointly, $w_f = w_s * w_{iptw}$.

In addition, I considered several alternative explanations for the pattern of *P. vivax* infections. These alternative explanations included: (1) interactions between non-human ape (NHA) ranges and *P. vivax* cluster-level prevalences using permutation tests; (2) within-host interactions of *P. vivax* and *P. falciparum* using a multinomial likelihood-based model that assumes independent infection acquisition; and (3) *post-hoc* power calculations (**Appendix 3.1**).

P. vivax Prevalence Maps

I considered spatial autocorrelation with Moran's I using a province adjacency matrix as well as a matrix of greater-circle distances between clusters.[17] Greater circle distances were calculated using a geodesic approach [18]. Significance was evaluated using a permutation test with 100,000 iterations and a one-sided p-value.

To determine the spatial distribution of *P. vivax*, I fit two types of Bayesian mixed spatial models: (1) a province-level areal model and (2) a cluster-level point process model. Province-based spatial models are important for intervention-planning, as most interventions in the DRC are implemented at the province-level. However, cluster-level models with Gaussian processes may be more representative of the intrinsic malaria distribution under the assumption of a continuous, and potentially heterogeneous, spatial process. Both sets of spatial models were fit with generalized linear mixed models using the logistic link function and a binomial error distribution with a spatial random effect (**Appendix 3.1**). For each of the respective spatial-

levels, I fit an intercept-only model and a model with all significant risk factors. Gelman's deviance information criterion (DIC) was used to assess model fit [19].

P. vivax Mitochondrial Genomics

Three DRC *P. vivax* samples among children previously identified by the University of North Carolina Infectious Disease Epidemiology and Ecology group from the 2013-2014 DRC DHS were considered to have the highest quality DNA and were prepared for sequencing [20]. All analyses were subsetted to the mitochondrial genome (mtDNA) due to lack of coverage in the nuclear genome. Nucleotide variants were identified among all samples and used to create consensus haplotypes (**Appendix 3.1**). Using these three DRC isolates and 685 globally sourced sequences, I created subpopulations based on geographical K-means clustering. Genetic distance measures, phylogenetic trees, and genetic summary statistics were generated to explore population diversity and differentiation (**Appendix 3.1**).

Results

Study Population and Molecular Validation

Among the 17,972 samples successfully shipped to the University of North Carolina for processing, 17,934 (99.79%) were linked to the 2013-2014 DRC DHS survey. Of these 17,934 samples, 169 samples failed to amplify human beta-tubulin, 1,402 individuals had missing geospatial data, and 484 individuals were classified as *de facto* (visitors rather than household members) and were excluded from analysis. In total, the final dataset consisted of 15,574 individuals across 489 clusters (**Appendix 3.1 Figure 2**).

The *P. vivax* qPCR assay achieved an analytical sensitivity of 94% and analytical specificity of 100% (zero false positive calls) when at least 1.25×10^{-7} ng/µL of 18S target (approximately 6 genomes/µL) was present. No off-target amplification was observed when the qPCR assay was challenged with highly concentrated DNA template from other *Plasmodium*

species (**Appendix 3.1 Figure 1**). *P. vivax* infection was confirmed by a separate, nested-PCR in 534 of 579 (93.6%) of qPCR-positive samples, with strong agreement between the initial and reflex confirmatory assays (Cohen's $\kappa = 0.80$, $p < 0.05$). All samples selected for Duffy-Genotyping validation had concordant HRM-qPCR and Sanger sequencing results, except for one sample that failed genotyping (**Appendix 3.1**).

Prevalence of P. vivax among Adults in the DRC

I restricted the prevalence estimates to 467 *P. vivax* infections that were confirmed by both qPCR and reflex nested-PCR ($n_{weighted}$: 459.18, 95% $CI_{weighted}$: 346.54, 571.82) and were among the 15,574 adults included in the study ($n_{weighted}$: 15,490.20, 95% $CI_{weighted}$: 14,060.60, 16,919.80). The national weighted prevalence of *P. vivax* among adults was 2.96% (95% $CI_{weighted}$: 2.28, 3.65%), with cluster point-prevalences ranging from 0 - 46.15% (**Figure 1**). Most clusters only contained a single *P. vivax* infection, although the weighted count of infections ranged from 0 - 30.63 infections per cluster. In contrast, I identified 5,179 *P. falciparum* infections ($n_{weighted}$: 4,651.94, 95% $CI_{weighted}$: 4,121.93, 5,181.94) accounting for a weighted national prevalence of 30.03% (95% $CI_{weighted}$: 27.87, 32.19%). Overall, there were 174 ($n_{weighted}$: 145.29, 95% $CI_{weighted}$: 108.11, 182.48) *P. falciparum* - *P. vivax* coinfections.

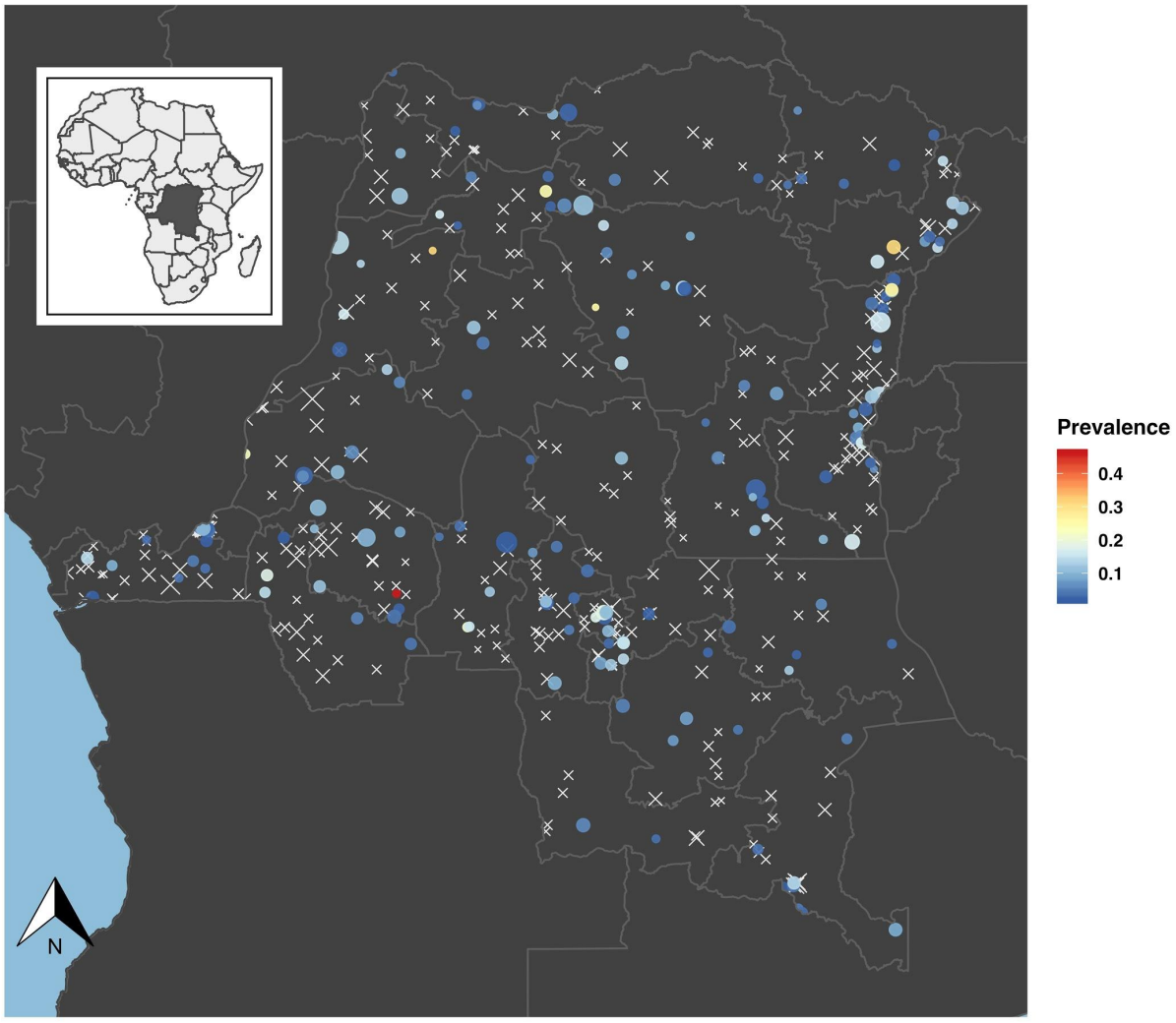


Figure 3.1 - The Distribution of *P. vivax* Infections across the Democratic Republic of the Congo: For clusters with *P. vivax* infections, the prevalence is indicated by a blue-red spectrum, while the size of the point indicates the size of the cluster denominator. Clusters with no *P. vivax* infections are represented with white X-marks. *P. vivax* infections appeared to be diffusely spread throughout the country with cluster prevalences ranging from 0 - 46.15%.

Risk Factors

Among the 579 qPCR-positive samples, only three had a putative Duffy-positive phenotype, all of whom were heterozygous at the loci associated with Duffy-negative phenotype (-33T:T/C). These three hosts ($n_{weighted}$: 1.61, 95% $CI_{weighted}$: -1.11, 4.34) were a part of the final study cohort, which led to an overall putative Duffy-positive phenotype frequency of 0.64% ($Prevalence_{weighted}$: 0.35%, 95% $CI_{weighted}$: 0.21, 0.58%) among those individuals infected with *P.*

vivax included in this study. From the cross-species interference model that assumes independent acquisition of infections, I failed to find any significant interactions between *P. falciparum*-*P. vivax* co-infections ($p > 0.05$; **Appendix 3.1 Figure 5**). Similarly, I did not find an association between NHA habitats and *P. vivax* cluster prevalence ($p > 0.05$; **Appendix 3.1 Figure 6**). Although baseline characteristics differed by infection status, the differences appeared to be most pronounced between *P. falciparum* infections and uninfected-individuals rather than *P. vivax* infections and uninfected-individuals (**Table 1**).

Covariate	*P. vivax* Infection	*P. falciparum* Infection	Uninfected
$N_{weighted}$	459.2	4651.9	10524.4
Vivax-Falciparum Coinfection	145.3		-
Urbanicity (Rural, %)	195.8 (42.6)	1573.2 (33.8)	4445.3 (42.2)
Lagged Precipitation (mm, SD)	131.9 (25.8)	137.8 (25.0)	138.3 (24.7)
Lagged Temperature (C, SD)	29.7 (3.2)	30.8 (2.7)	30.5 (3.6)
Altitude (m, SD)	771.5 (489.1)	617.6 (310.4)	749.6 (473.8)
Distance to Water (m, SD)	6695.7 (7745.5)	7762.3 (8096.5)	6364.5 (7074.8)
Distance to Nearest Public Hospital (m, SD)	224.2 (48.8)	2709.7 (58.3)	5338.4 (50.72)
HIV (Positive, %)	8.1 (1.8)	26.8 (0.6)	118.0 (1.1)
Sex (Male, %)	235.9 (51.4)	2435.4 (52.4)	4790.0 (45.5)
Age (years, SD)	29.4 (11.4)	28.2 (10.8)	30.4 (11.0)
Farmer (Farmer, %)	275.5 (60.0)	2541.7 (54.6)	5527.7 (52.5)
Housing Materials (Traditional, %)	262.1 (57.1)	3127.1 (67.2)	5555.2 (52.8)
Wealth (Comp. Score, SD)	0.1 (1.0)	-0.1 (1.0)	0.3 (1.2)
Education (Lower, %)	181.8 (39.6)	1975.9 (42.5)	4063.9 (38.6)
Number of Household Members (N, SD)	6.6 (3.1)	6.7 (3.1)	6.8 (3.3)
ITN Use (No, %)	213.8 (46.6)	2611.9 (56.2)	5301.4 (50.4)

Table 3.1 - Baseline Distributions of Identified Risk Factors among Individuals with *P. vivax* Infections, *P. falciparum* Infections, and those that are Uninfected: Risk factor distributions appeared to differ by infection status, with more noticeable differences between *P. falciparum* infections and uninfected individuals. For dichotomized risk factors, the counts and percentages for each category are provided. For continuous risk factors, the mean and standard deviation (SD) are provided. *Abbreviations*: N - number of individuals, mm - millimeters, m - meters, Comp. Score - composite score, ITN - insecticide-treated net.

In order to formally assess the risk-factor prevalence odds ratios (pORs) among *P. vivax* and *P. falciparum*, I adjusted for confounding using IPW. IPWs that were calculated with the

spatially cross-validated super learner algorithm appeared to be stable with approximately log-normal distributions (**Appendix 3.1 Figure 8**). Additionally, for most covariates, IPW resulted in a considerable decrease in the average correlation among baseline covariates as compared to unadjusted baseline correlations (mean fold-reduction: 3.14, range: 0.85 - 7.63; **Appendix 3.1 Figure 9**).

When *P. vivax* was considered as the outcome of interest, higher levels of precipitation were found to reduce prevalence (IPW-pOR: 0.79, 95% CI: 0.63, 0.99) while being a farmer appeared to increase prevalence (IPW-pOR: 1.42, 95% CI: 1.08, 1.89). In contrast, when considering *P. falciparum* infections as the outcome of interest, several risk factors were associated with prevalence: an urban setting reduced prevalence (IPW-pOR: 0.70, 95% CI: 0.54, 0.89), lack of insecticide-treated net (ITN) use increased prevalence (IPW-pOR: 1.23, 95% CI: 1.07, 1.42), increasing altitude reduced prevalence (IPW-pOR: 0.73, 95% CI: 0.65, 0.82), temperature increased prevalence (IPW-pOR: 1.41, 95% CI: 1.05, 1.90), lower levels of education increased prevalence (IPW-pOR: 1.44, 95% CI: 1.25, 1.67), higher levels of wealth reduced prevalence (IPW-pOR: 0.82, 95% CI: 0.73, 0.92), older age reduced prevalence (IPW-pOR: 0.81, 95% CI: 0.77, 0.86), while being male increased prevalence (IPW-pOR: 1.31, 95% CI: 1.20, 1.43).

Based on the *post hoc* power calculations for *P. vivax*, I was able to detect harmful pOR estimates of at least 1.54, 1.36, 1.29 with at least 80% power when the exposure probability was 10%, 25%, and 50%, respectively. In contrast, for *P. falciparum*, I was able to detect harmful pOR estimates of at least 1.18, 1.12, 1.10 with at least 80% power when the exposure probability was 10%, 25%, and 50%, respectively (**Appendix 3.1 Figure 12**).

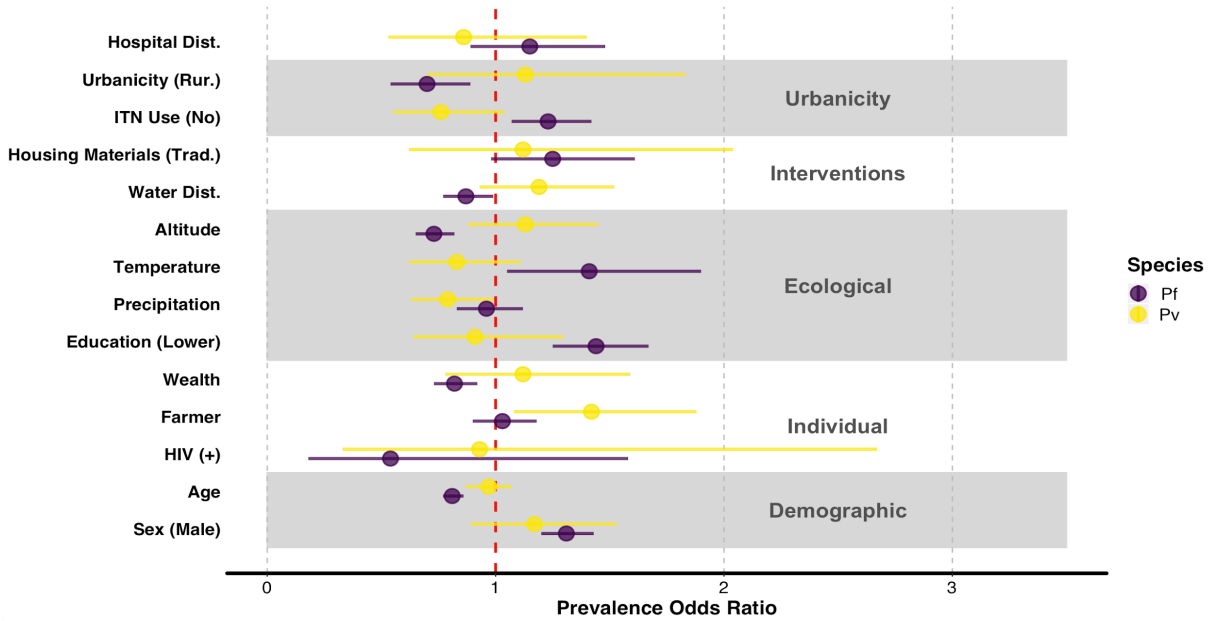


Figure 3.2- Inverse Probability Treatment Weight Adjusted Prevalence Odds Ratios for Expected Malaria Risk Factors: The inverse probability weight adjusted prevalence odds ratios (IPW-pORs) demonstrated a lack of risk factors for *P. vivax* infection, as all risk factors contain the null estimate (red line) with the exception of precipitation and farming. In contrast, numerous risk factors were associated with *P. falciparum*, including living in a rural area, ITN use, altitude, education, wealth, age, and biological sex. For both species, hospital distance, traditional housing materials, distance to water, and HIV-status were not significant risk factors. The unadjusted pORs effect estimates and confidence intervals as well as the IPW-pORs are provided in **Appendix 3.1 Table 6** for reference. *Abbreviations*: Hospital Dist. – Distance to a hospital, Water Dist. – Distance to water, Rur. - rural, Trad. - traditional, ITN - insecticide-treated net.

Spatial Distribution of P. vivax

When considering spatial autocorrelation, I found that the province-level showed a slight signal of structure for *P. vivax* prevalence (Moran's I: 0.16; $p = 0.05$), but this structure did not hold at the cluster-level (Moran's I: 0.02; $p > 0.05$). Among the *P. vivax* province-level models considered, I found that the best fitting model contained the precipitation, night light intensity, and farming covariates (**Appendix 3.1 Table 7**). Means-fitted province prevalences ranged from 1.24 - 7.61% (**Figure 3A**). Standard errors for the province prevalence estimates ranged from 4.46×10^{-3} - 2.30×10^{-2} (**Appendix 3.1 Figure 11A**).

Similarly, when I modeled the spatial distribution of *P. vivax* at the cluster-level, I found that the best fitting model contained the precipitation, night light intensity, and farming covariates (**Appendix 3.1 Table 7**). Based on the model predictions, *P. vivax* fitted prevalence ranged from 0.50 - 11.20% across the DRC (**Figure 3B**). The standard errors around the prevalence predictions ranged from 1.62×10^{-8} - 2.30×10^{-6} (**Appendix 3.1 Figure 11B**). Most *P. vivax* prevalence predictions were less than the observed national prevalence (19,903/20,000; 99.52%).

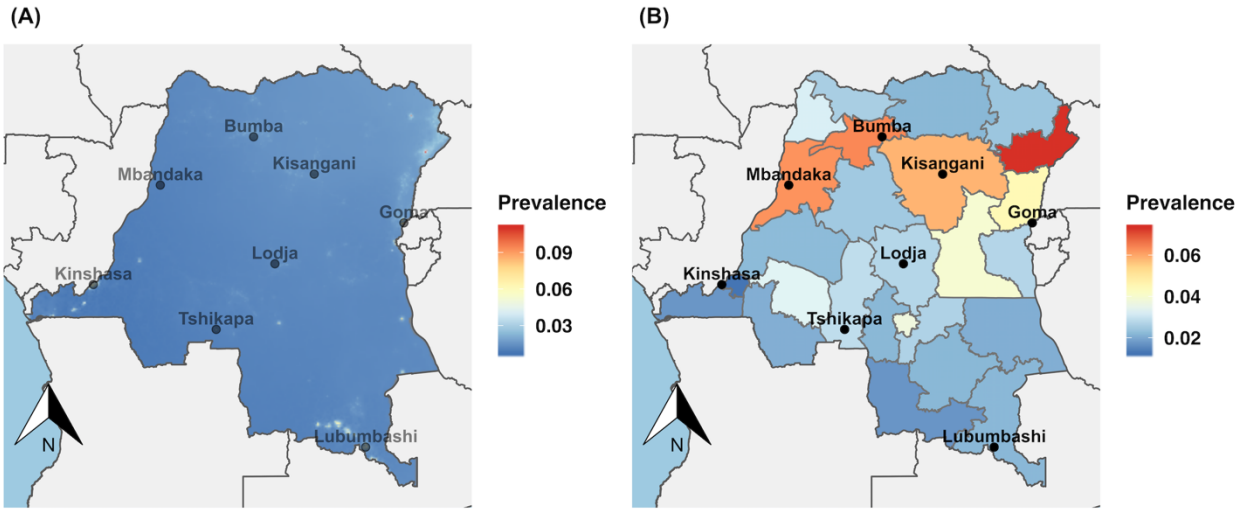


Figure 3.3 - Spatial Model Posterior Means: Shown are the means of the posterior prevalence distribution for the cluster-level (left) and province-level (right) models. At the province-level, *P. vivax* infections appeared to be more common in the north. However, this process was not recapitulated at the cluster-level, where very local transmission appeared to dominate with a few focal regions of high prevalence amidst a relatively uniform background of *P. vivax* prevalence.

P. vivax Diversity, Differentiation, and Phylogeography

The three sequenced *P. vivax* DRC isolates had high-quality coverage in ≥98.0% of the mtDNA genome, with an average mtDNA base-depth of 40.85. Among the 636 publicly available Illumina sequenced *P. vivax* isolates that passed QC-thresholds, I detected 57 unique mitochondrial haplotypes (**Appendix 3.1 Figure 13)**. Among the haplotypes, I identified 65

biallelic sites and 1 polyallelic site, with most mutations occurring in the non-protein coding regions (Ti:Tv = 1.2). Overall, the NHA and the Asian *P. vivax* population demonstrated the greatest within-population nucleotide and haplotype diversity, while there was limited within-population diversity among the isolates from the DRC (**Appendix 3.1 Table 9**). Based on between-population measures of nucleotide diversity, the DRC samples were most similar to samples from the Americas. However, when considering pairwise measures of F_{st} between populations, the DRC population appeared to be relatively isolated from other populations (**Appendix 3.1 Table 10**). When considering the evolutionary relationship of the DRC samples with samples sourced from across the globe, I found that the DRC samples formed a separate monophyletic clade (Bootstrap Support: 62.0%). The DRC monophyletic clade had a most recent common ancestor (MRCA) with a clade that contained a subset of samples from Peru (**Figure 4**; Bootstrap Support 11.8%). Although the DRC haplotype appeared to be most closely related to haplotypes circulating in the Americas, the DRC haplotype was similar to haplotypes from the Asian and Oceanic populations (**Figure 5**).

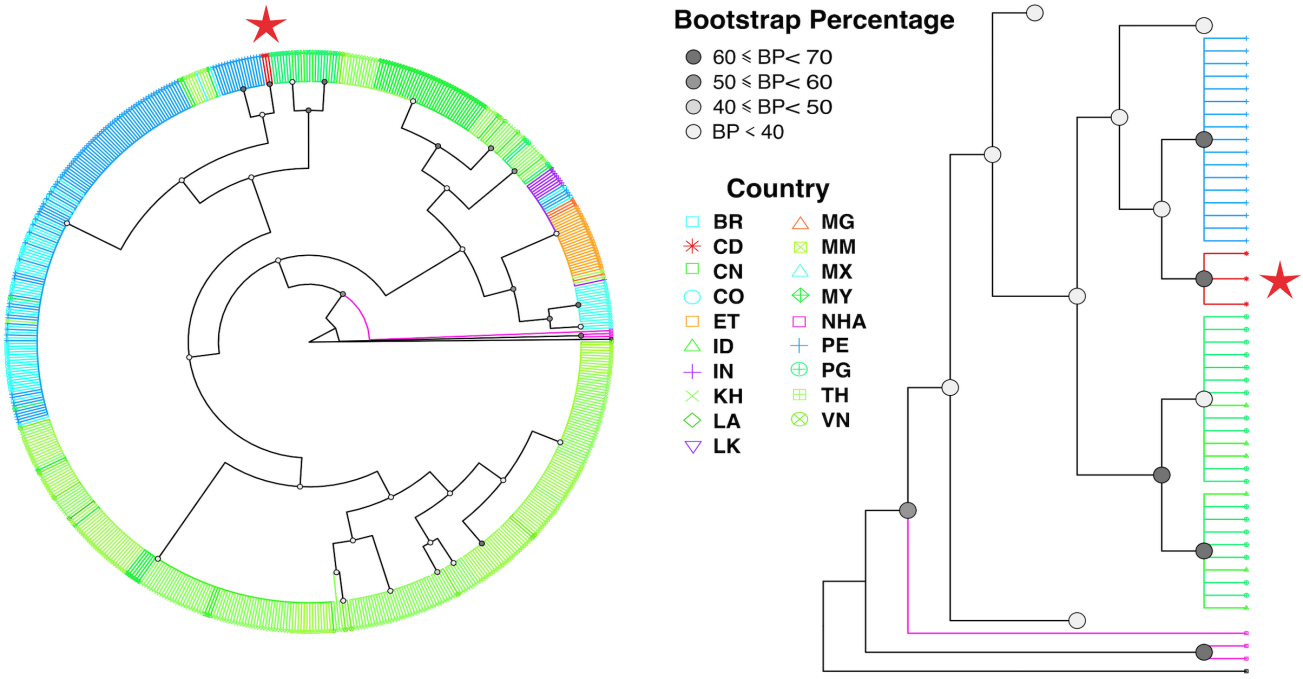


Figure 3.4 - Phylogenetic Tree of *P. vivax* Global Isolates: Comparison of DRC *P. vivax* with 636 globally sourced *P. vivax* isolates showed that *P. vivax* shared a most recent common ancestor with samples from Peru. Although the DRC-Peru node did not have strong support, collapsing the node still results in the same conclusion of *P. vivax* sharing a most recent common ancestor with isolates from South America. The full phylogenetic tree is provided (left), with various clades not along the DRC ancestry collapsed for a focused view of the DRC clade (right). Isolates from the Americas are colored in shades of blue and included Brazil (BR), Colombia (CO), Mexico (MX), and Peru (PE). Asian countries are indicated in shades of green and included China (CN), Indonesia (ID), Cambodia (KH), Laos (LA), Myanmar (MM), Malaysia, Papua New Guinea, Thailand (TH), and Vietnam (VN). India (IN) and Sri Lanka (LK) are indicated in shades of purple, while Ethiopia (ET) and Madagascar (MG) are indicated in shades of orange. Finally, the Democratic Republic of the Congo (CD) is shown in red and non-human apes (NHA) are shown in magenta. *P. cynomologi* was set as the tree root and is indicated in black.

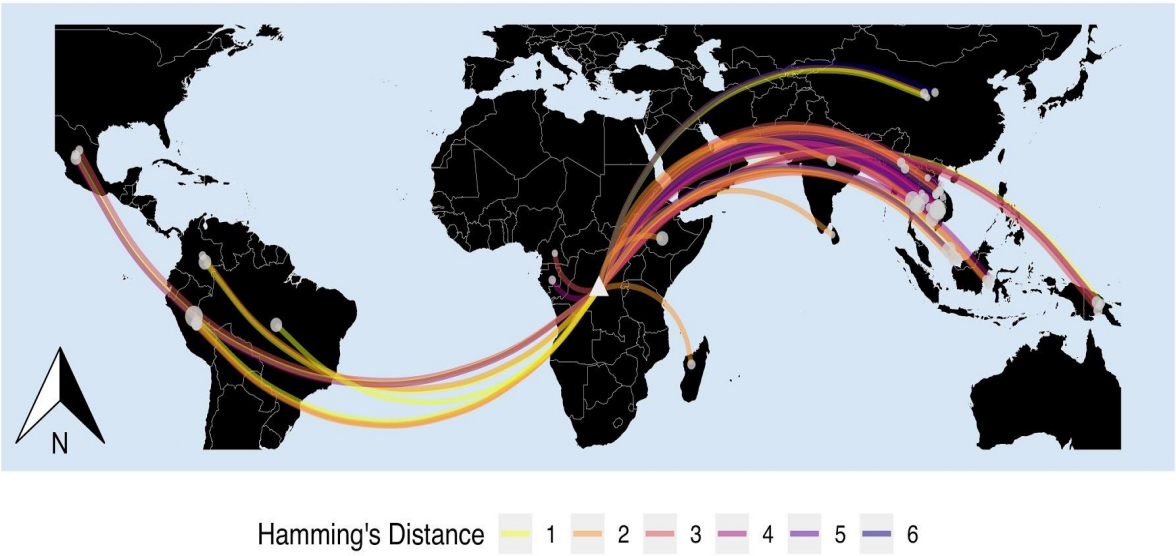


Figure 3.5 - Haplotype Genetic Distances among Global Isolates with respect to the DRC. Although DRC samples may share a most recent common ancestor with isolates from South America, DRC haplotypes contain elements from both South America, Asia, and the Oceanic regions. Given that DRC haplotypes are more similar to haplotypes in South America and Asia than in Africa, this suggests that the DRC *P. vivax* may be an ancestral strain that potentially seeded those regions. All unique haplotypes with respect to country of origin are provided for comparison and context (**Appendix 3.1 Figure 13**).

Discussion

P. vivax infections among adults in the DRC are more common than previously realized. From the spatially robust dataset across the DRC, I detected 467 *P. vivax* infections corresponding to a national prevalence of 2.96% (95% $CI_{weighted}$: 2.28, 3.65%). Among those infected, nearly all were Duffy-negative (576/579, 99.48%).

Malaria risk-factors typically associated with *P. falciparum* infection, such as ITN use and wealth, were not associated with *P. vivax* infection. Instead, only precipitation and farming were identified as *P. vivax* risk factors. This relationship between *P. vivax* prevalence and precipitation has been previously described, although the underlying effect is likely complicated by other ecological factors, such as vector habitats, seasonality, altitude, and temperature [21]. Similarly, increased prevalence of malaria has previously been attributed to agriculture and

farming in the DRC [22]. Overall, the *P. vivax* malaria risk factors differed greatly from risk factors found for *P. falciparum* in this study using the same methodological approach. This contrast between *P. vivax* and *P. falciparum* risk factors may be the result of the shortened intrinsic period in *P. vivax* or hypnozoite infections being resilient to typical antimalarial interventions [23]. *P. vivax* infections were found throughout the entire country with a few focal regions of relatively high prevalence (**Figure 3.2**). The highest prevalence of *P. vivax* was found in the Ituri province. This may be due to cross-border migration with South Sudan and Uganda, which border countries that are endemic for *P. vivax* (*P. vivax* infections have been reported in both countries).[1,5] However, the sources of infection in the other provinces are not clear. Microheterogeneity in *P. vivax* prevalence has previously been reported in the Amazon and was found to be associated with human movement [24]. Future *P. vivax* epidemiological studies in the DRC should consider human mobility data, particularly with respect to Kinshasa and regions along the eastern border where interactions with Duffy-positive immigrants may be more frequent.

Although there appears to be small-scale heterogeneity of *P. vivax* in the DRC, more than half of predicted prevalences were less than one-percent and 99.95% of predicated prevalences were less than the observed national average. These localized regions of prevalence, or "hotspots," contrast the broad spatial distribution of *P. falciparum* infections previously observed in the 2007 and 2013 DRC DHS.[9,25] As a result, I suggest that *P. vivax* has been unable to gain a foothold in the region and is persisting rather than spreading.

The relatively large differences in the DRC *P. vivax* and the NHA mitochondrial genomes likely negates recent zoonotic transmission as the source of DRC *P. vivax*. I identified a MRCA between the DRC samples and a subset of samples from Peru during phylogenetic

analysis. Although the node support for the DRC-Peru relationship was weak (11%), collapsing the node does not alter the conclusion that the DRC MRCA was most similar to extant South American parasites. This finding that Africa may have seeded American *P. vivax* has recently gained traction based on analysis of a historical sample originating from the Ebro Delta in Spain, circa 1944 [26–28]. Using this historical sample, the authors demonstrated that now extinct European *P. vivax* was closely related to extant *P. vivax* from the Americas, potentially dating to the European colonial expansion during the 15th century [26,27]. When I included the Erbo-1944 sample in my comparisons, I found that the mitochondrial haplotype differed by only a single base-pair from the DRC haplotypes. As a result, I hypothesize that DRC *P. vivax* may have migrated from Africa to Europe prior to being transported to the New World on the wave of European expansion.

However, the history of DRC *P. vivax* is not straightforward. The haplotypes differed from samples collected in Asia, Oceania, and the Americas by only a single base pair. This close relationship may indicate ongoing or historical mixing with Asian and Oceanic *P. vivax*, an idea supported by the genetic measures of population differentiation. In addition, the Erbo-1944 consensus haplotype also matched haplotypes from the Americas, Asia, and Oceania populations. These similarities may have arisen due to waves of historical introgression and panmixia among *P. vivax* globally or may be an artifact of my conservative approach to variant filtering. Consistent with previous reports, I found relatively few informative sites in the *P. vivax* mitochondria [28]. Despite this low variation, the mitochondria is a non-recombining region with putatively neutral SNPs that is ideal for phylogenetic analysis to resolve ancestry [28].

The DRC is a critical region for the study of malaria in SSA due to its size, central location, and evidence that bridges East and West Africa malaria [7]. These characteristics allow

the DRC to serve as a microcosm of the region [7,8]. The main limitations of this study are the cross-sectional design, which limits inference of effects with a temporal component (e.g. seasonality) and restricts the study population largely to asymptomatic individuals, and the small number of high-quality DRC mitochondrial sequences generated. Future efforts will require more biological material than DBS to improve the likelihood of successful genomic sequencing. Whole genomes from the DRC would provide more insights on the demographic history of *P. vivax* in SSA and putative regions of selection for adaptation to the Duffy-negative host.

Until recently, *P. vivax,* was an unrecognized cause of disease in SSA. This study provides the first systematic and nationally representative survey of *P. vivax* in a SSA country not considered endemic for the disease. I demonstrated that *P. vivax* is circulating at prevalences higher than previously thought, despite a high frequency of Duffy-negativity [1]. However, *P. vivax* infections were not associated with classic malaria risk factors, were spread diffusely throughout the country, and may represent an old lineage. These findings suggest that *P. vivax* may have been circulating in SSA as an innocuous, chronic infection that was overlooked in past studies due to frequently sub-microscopic or low parasitemia infections. This hypothesis is consistent with previous work that suggests *P. vivax* infections among Duffy-negative individuals are frequently mild and asymptomatic compared with Duffy-positive individuals [5,6]. Finally, emerging research suggests that genotypically Duffy-negative hosts express the Duffy antigen among erythroid progenitors in the bone marrow and that *P. vivax* gametocytes are able to mature and proliferate in the bone marrow of non-human primate animal models [29,30]. Collectively, this suggests that *P. vivax* in Sub-Saharan Africa may be persisting as low parasitemic, asymptomatic, or relatively innocuous infections, by hiding in the bone marrow of Duffy-negative hosts. While the malaria community should remain mindful of *P. vivax* in SSA,

its distribution and low prevalence support continued investments targeting *P. falciparum* as likely having the greatest impact on malaria elimination, morbidity, or mortality.

REFERENCES

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

2. Miller LH, Mason SJ, Clyde DF. The resistance factor to Plasmodium vivax in blacks: the Duffy-blood-group genotype, FyFy. *New England Journal* 1976. http://www.nejm.org/doi/full/10.1056/nejm197608052950602.

3. Tournamille C, Colin Y, Cartron JP, Van Kim CL. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy–negative individuals. *Nat Genet* 1995; **10**: 224–8.

4. Howes RE, Patil AP, Piel FB, *et al.* The global distribution of the Duffy blood group. *Nat Commun* 2011; **2**: 266.

5. Twohig KA, Pfeffer DA, Baird JK, *et al.* Growing evidence of Plasmodium vivax across malaria-endemic Africa. *PLoS Negl Trop Dis* 2019; **13**: e0007140.

6. Ménard D, Barnadas C, Bouchier C, *et al.* Plasmodium vivax clinical malaria is commonly observed in Duffy-negative Malagasy people. *Proc Natl Acad Sci U S A* 2010; **107**: 5967–71.

7. Verity RJ, Aydemir O, Brazeau NF, Watson OJ. The Impact of Antimalarial Resistance on the Genetic Structure of Plasmodium falciparum in the DRC. *bioRxiv* 2019. https://www.biorxiv.org/content/10.1101/656561v1.abstract.

8. Taylor SM, Antonia AL, Parobek CM, *et al.* Plasmodium falciparum sulfadoxine resistance is geographically and genetically clustered within the DR Congo. *Sci Rep* 2013; **3**: 1165.

9. Molly Deutsch-Feldman, Nicholas F. Brazeau, Jonathan B. Parr, Kyaw L. Thwai, Jérémie Muwonga, Melchior Kashamuka, Antoinette K. Tshefu, Jessie K. Edwards, Robert Verity, Michael Emch, Emily W. Gower, Jonathan J. Juliano, Steven R. Meshnick. Spatial and epidemiological drivers of P. falciparum malaria among adults in the Democratic Republic of the Congo. .

10. Srisutham S, Saralamba N, Malleret B, Rénia L, Dondorp AM, Imwong M. Four human Plasmodium species quantification using droplet digital PCR. *PLoS One* 2017; **12**: e0175771.

11. Snounou G, Singh B. Nested PCR analysis of Plasmodium parasites. *Methods Mol Med* 2002; **72**: 189–203.

12. Croft TN, Marshall AMJ, Allen CK, Others. Guide to DHS statistics. *Rockville, Maryland, USA: ICF* 2018.

13. Tanaka M, Takahahi J, Hirayama F, Tani Y. High-resolution melting analysis for genotyping Duffy, Kidd and Diego blood group antigens. *Leg Med* 2011; **13**: 1–6.

14. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol* 2007; **6**: Article25.

15. Brenning A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In: 2012 IEEE International Geoscience and Remote Sensing Symposium. 2012: 5372–5.

16. Pirracchio R, Petersen ML, van der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *Am J Epidemiol* 2015; **181**: 108–19.

17. Moran PAP. Notes on continuous stochastic phenomena. *Biometrika* 1950; **37**: 17–23.

18. Karney CFF. Algorithms for geodesics. *J Geodesy* 2013; **87**: 43–55.

19. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A. Bayesian data analysis. 2013. https://www.taylorfrancis.com/books/9780429113079.

20. Brazeau NF, Whitesell AN, Doctor SM, Keeler C, Mwandagalirwa MK, Tshefu AK, Likwela JL, Juliano JJ, Meshnick SR. Plasmodium vivax Infections in Duffy-Negative Individuals in the Democratic Republic of the Congo. *Am J Trop Med Hyg* 2018; published online Nov. DOI:10.4269/ajtmh.18-0277.

21. Chowell G, Munayco CV, Escalante AA, McKenzie FE. The spatial and temporal patterns of falciparum and vivax malaria in Perú: 1994–2006. *Malar J* 2009; **8**: 142.

22. Janko MM, Irish SR, Reich BJ, *et al.* The links between agriculture, Anopheles mosquitoes, and malaria risk in children younger than 5 years in the Democratic Republic of the Congo: a population-based, cross-sectional, spatial study. *The Lancet Planetary Health* 2018; **2**: e74–82.

23. Olliaro PL, Barnwell JW, Barry A, *et al.* Implications of Plasmodium vivax Biology for Control, Elimination, and Research. *Am J Trop Med Hyg* 2016; **95**: 4–14.

24. Carrasco-Escobar G, Gamboa D, Castro MC, *et al.* Micro-epidemiology and spatial heterogeneity of P. vivax parasitaemia in riverine communities of the Peruvian Amazon: A multilevel analysis. *Sci Rep* 2017; **7**: 8082.

25. Taylor SM, Messina JP, Hand CC, *et al.* Molecular Malaria Epidemiology: Mapping and Burden Estimates for the Democratic Republic of the Congo, 2007. *PLoS One* 2011; **6**: e16420.

26. Gelabert P, Sandoval-Velasco M, Olalde I, *et al.* Mitochondrial DNA from the eradicated

European Plasmodium vivax and P. falciparum from 70-year-old slides from the Ebro Delta in Spain. *Proc Natl Acad Sci U S A* 2016; **113**: 11495–500.

27. van Dorp L, Gelabert P, Rieux A, de Manuel M. Plasmodium vivax Malaria viewed through the lens of an eradicated European strain. *bioRxiv* 2019. https://www.biorxiv.org/content/10.1101/736702v1.abstract.

28. Rodrigues PT, Valdivia HO, de Oliveira TC, *et al.* Human migration and the spread of malaria parasites to the New World. *Sci Rep* 2018; **8**: 1993.

29. Dechavanne C, Dechavanne S, Metral S, *et al.* Duffy Antigen Expression in Erythroid Bone Marrow Precursor Cells of Genotypically Duffy Negative Individuals. bioRxiv. 2018; : 508481.

30. Obaldia N 3rd, Meibalan E, Sa JM, *et al.* Bone Marrow Is a Major Parasite Reservoir in Plasmodium vivax Infection. *MBio* 2018; **9**. DOI:10.1128/mBio.00625-18.

CHAPTER FOUR: TRACING THE GENETIC RELATEDNESS OF PLASMODIUM FALCIPARUM IN THE DEMOCRATIC REPUBLIC OF THE CONGO ACROSS SPACE

Introduction

Using genetic relatedness to infer transmission chains, migration patterns, and population demographic histories are fundamental goals of infectious disease genetic epidemiology. For many infectious diseases, relatedness can be estimated from coalescent methods, which allows for phylodynamic modeling and inference of transmission dynamics. However, malaria pathogens undergo recombination, exhibit low nucleotide mutation rates, and can be polyclonal infections -- all factors that violate classic coalescent assumptions [1]. Instead, there has been a resurgence in using identity by descent (IBD) methods to quantify genetic relatedness among malaria parasites [2–15].

Identity by descent (IBD) is the process of inheriting segments of DNA from a common ancestor through meiotic recombination. Under a classic Wright-Fisher population, two haploid individuals are expected to share 2^{-G} proportion of their genome by IBD, where G is the number of generations that separate the pair [16–19]. As part of the malaria life-cycle, recombination among parasites occurs within the mosquito midgut prior to host inoculation [20]. Although generation intervals vary widely depending on treatment, seasonality, and transmission intensity, *P. falciparum* generation times are assumed to be approximately 1-3 months long [21,22]. As a result, IBD is an ephemeral signal that captures recent relatedness, which is relevant for public health.

IBD reflects the interplay between effective population sizes (N_e) and geographical space and is expected to decrease exponentially as both entities increase. IBD is expected to decrease

exponentially as the N_e increases due to a lower chance of inbreeding [23]. Previous research has shown that as *P. falciparum* prevalence increases, pairwise IBD decays exponentially [14]. This suggests that although the census size and N_e are not necessarily the same, *P. falciparum* transmission intensity is proportional to N_e. Similarly, IBD is expected to decay exponentially as the spatial distance between individuals increases, a process termed isolation by distance [24,25]. This process of isolation by distance has previously been identified in *P. falciparum* parasites along the Thailand-Myanmar border and the Democratic Republic of the Congo (DRC) [6,8]. In addition, IBD was shown to capture spatial process at a much higher resolution than traditional statistics of genetic differentiation (i.e. F-statistics) [6].

Identifying and quantifying perturbations to expectations of isolation by distance are likely informative for malaria control efforts, as they may indicate reservoirs of malaria transmission, importation events, or heterogeneity in transmission over geographical space. This latter point is critical, as previous work quantity *P. falciparum* isolation by distance has focused on greater-circle distances and not considered other spatial distances as measures of connectedness. Other geographical distances, such as road distance or distance along rivers, will approximate migration of parasites through human movement, while greater-circle distances approximate migration by mosquito movement.

In this study, I used a spatially robust dataset of 1,111 samples from 351 geographic clusters to explore how geographic distance affect patterns of IBD in the DRC. The DRC is an ideal location to analyze patterns of isolation by distance in *P. falciparum*, as it exhibits a large degree of spatial heterogeneity in prevalence and has previously been shown to be a bridge between West and East Africa parasite genetic diversity [8,26]. As a result, I was able to capture regions of low-transmission and high-transmission as well as genetic demes with varying levels

of admixture. I found that parasite dispersion appeared to be driven by human movement and not necessarily mosquito movement. In addition, I identified cities as potential genetic hubs and discuss their importance for malaria control. By combining these spatial and genetic approaches, I provide a picture of how *P. falciparum* infections may be arising in the DRC instead of simply where.

Materials and Methods

Parasite Genetic Data and Genetic Calculations

This study utilizes the genetic data and samples from the DRC previously analyzed in Verity *et al.* 2019. In brief, dried blood spots from adults and children in the 2013-2014 Demographic Health Survey (DHS) in the DRC underwent DNA-extraction using Chelex-100 (Bio-Rad, Hercules, CA) and Saponin [27,28]. Samples with cycle-threshold values of less than 30 from a *P. falciparum* lactate-dehydrogenase quantitative PCR reaction were identified for sequencing using molecular inversion probes [8,29]. Sequences then underwent alignment, variant calling, and filtering as previously described. Briefly, samples were genotyped at 1,890 sites across the genome ("genome-wide panel"), and sites were filtered based on Phred-scaled quality score of <20, low-coverage and lack of genetic segregation. Variants were then limited to biallelic loci. Separately, samples were excluded if more than half of the loci were determined to be low-coverage [8,29].

From these filtered variants, I evaluated the autocorrelation among loci using Pearon's correlation coefficient on the within-sample allele frequencies. I then calculated pairwise IBD between all samples using a maximum-likelihood estimator that has been previously described [8]. The estimator is based on the classic definition of IBD proposed by Malécot, where pairs of DNA segments not broken by mutation are followed back to a common ancestor [19,30].

Measures of Geographic Distance

I calculated pairwise greater-circle distance between clusters using the greater-circle distance in the R package, `sf` [31]. Greater-circle distance is essentially the shortest euclidean distance along a curved surface.

In order to build road networks for the DRC, I downloaded geographical data from Geofabrik (https://www.geofabrik.de/data/download.html; accessed August 23, 2019) and formatted it for the Open Source Routing Machine API (**Figure 1B**) [32]. I then calculated pairwise road distances between clusters using the `orsm` R-package. Among the 351 clusters considered, one cluster (ID: 469) could not be resolved by `osrm`. I imputed the road distances for the 469 cluster based on cluster 313 (nearest neighbor of cluster 469) road distances with an offset of 2,000 meters: the approximate greater-circle distance between the two clusters. Given that only distances along roads was considered, clusters that do not directly lie on a road are snapped to the nearest road. This snapping effect will result in shorter path estimates if any roads were not in the database or were unmarked (i.e. dirt paths).

In order to calculate distances along rivers between clusters, I created a river network using waterway lines downloaded from the Humanitarian OpenStreetMap Team database for the DRC (https://data.humdata.org/dataset/hotosm_cod_waterways; accessed October 30, 2019) and DIVA-GIS (http://www.diva-gis.org/gdata; accessed January 3, 2020). Using the GRASS (v7.4) and QGIS (v3.8) programs, I merged the two data sources and fixed topology of the river network with the `v.clean` suite (`rmsa`, `break`, `rmdup`, `rmline`, `rmdangle`, `snap` (0.05)). I then simplified the river network using the `generalize` function and the Douglas-Peuker algorithm [33]. Finally, I limited the simplified river network to the single largest connected component, thereby removing islands using the `shp2graph` R-package (**Figure 1C**). From this

newly created river network, I calculated the pairwise river distance between clusters by snapping clusters to the nearest node and calculating the shortest edge lengths -- mimicking the road network behavior described above. For clusters that snapped to the same river vertex, I included a 5,000 meter offset (approximate minimum between clusters snapped to different vertices).

Feature Engineering for IBD Parametric Models

I identified potential predictors of genetic relatedness among *P. falciparum* parasites from a comprehensive literature review. I then downloaded the raster for each available predictor from open source venues (**Table 4.1**).

Temperature data from the MYD11C3 (v6) product was downloaded from the Level-1 and Atmosphere Archive & Distribution System (LAADS) Distributed Active Archive Center (Goddard Space Flight Center, Greenbelt, MA; accessed September 20, 2019) [34]. Similarly, I downloaded precipitation data from the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) server using an R-wrapper package, (GitHub: `environmentalinformatics-marburg/heavyRain`) for the CD2013 study period [35]. Both temperature and precipitation data were downloaded at a 0.05° x 0.05° spatial resolution. Given the 2013-2014 Demographic Health Survey in the DRC was conducted over six months, I calculated the mean temperature and precipitation across the study period [36].

For the remaining rasters, I aggregated raster cells to a 0.05° x 0.05° spatial resolution by taking the mean values within cells. Covariates encoded as proportions were logit-transformed back to the real line. Due to extreme collinearity among the night light intensity, population density, and travel time covariates, I created an overall “urbanicity” factor score. The urbanicity factor score for each geographic location was calculated by performing a principal component

analysis on the night light intensity, population density estimates, and travel times rasters and extracting the eigenvalues of the first principal component. The upper bound of the urban factor score was set to the 99.9th percentile of the original distribution to truncate outliers. Each covariate was aggregated at the province-level by taking the mean value with respect to province boundaries. All province-aggregated covariates were then standardized (mean-centered and scaled by the standard deviation). Province-level models were selected because most intervention-planning and intervention-implementation occurs at the province-level in the DRC.

Covariate	Source	Manipulations	Year	Citation
Elevation	Elevatr	Standardize	-	[37]
Precipitation	CHIRPS	Standardize	2013-2014	[35]
Temperature	LAADS	Standardize	2013-2014	[34]
Urbanicity	WorldPop	Standardize	2013	[38,*]
	Travel Time (MAP)	Standardize	2015	[39]
	Friction Surface (MAP)	Standardize	2015	
	NOAA (VIIRS)	Zero-truncated Standardize	2015	[40,41]
Cropland	ESA	Binary	2013	[42]
Falciparum Parasite Rate	MAP	Logit transform, Standardize	2015	[43]
Net Use	MAP	Logit transform, Standardize	2015	
House	MAP	Logit transform, Standardize	2015	[44]

Table 4.1 - Risk Factors Covariate Source and Transformations: Covariates were downloaded from several open source platforms, including: the Malaria Atlas Project (MAP), the European Space Agency and Copernicus Atmosphere Monitoring Services (ESA), the Earth Observations Group at National Oceanic and Atmospheric Administration/National Centers for Environmental Information (NOAA), the Level-1 and Atmosphere Archive & Distribution System Distributed Active Archive Center (LAADS), WorldPop (*https://www.worldpop.org/), and Amazon Web Services Open Data Terrain Tiles via the `elevatr` R package. For each spatial raster, the year of data collection is indicated. Covariates were then curated and transformed for model fitting depending on the functional form of the data.

Feature Engineering for Non-Parametric Tests

Based on my *a priori* assumptions, I explored the relationship between IBD and urbanicity and IBD and prevalence at the cluster level, respectively. Urbanicity was recoded as a binary measure (urban vs. rural) using a 95th percentile cutoff in my previously created urban factor score. To account for the offset used by the DHS to anonymize clusters, I extracted raster cells within 2 km (urban clusters) or 10 km (rural clusters) of each cluster location [45,46]. Clusters that had at least 50% of raster cells coded as urban in their catchment area were considered as urban. A new urban-rural variable was created in light of previous evidence indicating that the DHS coding of urbanicity may be biased in the DRC [47]. Similarly, cluster level prevalence was calculated as the mean prevalence within the DHS catchment area, using the Malaria Atlas Project parasite-rate raster (**Table 4.1**). The catchment area for one cluster (ID: 54) was increased from 2 km to 6 km due to issues of missing data.

Permutation Tests and Edge Density

Correlations among the genetic relatedness measures and the geographic distances were assessed using a Mantel's test with 10,000 iterations [48]. A Mantel's test is essentially a permutation approach, where each permuted iteration accounts for the autocorrelation among cells in a distance adjacency matrix.

To determine if urbanicity (binary, cluster-level covariate) differed between highly related pairs, I performed permutation testing with 10,000 iterations. For each permutation, I drew 86 cluster urban/rural observations (maximum number of cluster observations between 43 pairs under an independence assumption) using the distribution of urbanicity found among the 351 observed clusters. To form the null distribution, I then calculated the proportion of simulated

urban clusters from each iteration. The null distribution was then compared to the observed proportion of urban clusters among the highly related pairs.

For each sample that was a part of a highly related pair, I calculated the weighted edge-density, or the mean of the IBD measures between pairs. This approach approximates the degree centrality measure in network analysis [49]. Correlations between cluster parasite rate and sample edge densities were calculated using the Szekely-Rizzo-Bakirov distance correlation test with 10,000 permutations using the `energy` R package [50–52]. The Szekely-Rizzo-Bakirov distance correlation test measure multivariate dependence through euclidean distances and thus can capture nonlinear patterns [50–52].

Spatial Distance and Genetic Relatedness Likelihood

In order to determine which geographic and genetic distance best approximated the isolation by distance framework, I created a likelihood function based on Malécot's original formulation of genetic isolation by distance [30,53]. For this likelihood, I made several simplifying assumptions: (1) pairwise-IBD between sample x and sample y, F_{xy} is known and fixed; (2) F_{xy} is stationary and at equilibrium, such that $F_{xy}^{t-1} = F_{xy}^{t}$; (3) the mutation rate in the population is negligible. Furthermore, let samples reside within non−overlapping regions of interest, or clusters, where clusters are indexed by $n \in \mathbb{Z}_{\geq 0}$. Given that generational relatedness is at an equilibrium, time is considered as a binary: present and past. Let i represent a given cluster in the present and u represent the same cluster in the past. Further, let $\mathbf{v}$ represent all other clusters in the population, indexed by $v \in \mathbf{v}$. The probability of migration between clusters is exponentially distributed with respect to the geographic distance, such that $m_{ij} \sim e^{d_{ij}}$. I assume that d_{ij} can be derived from the differences in geographic location $S(.)$ between the two clusters: $d_{ij} = S(i) - S(j)$ and is scaled by $E(distance)$. Finally, let the cluster-level inbreeding

coefficient for each sample pair be represented by $\rho(i,j)$ and made up of two parts, such that $\rho(i,j) = \rho(u,u) + \rho(u,v)$. Given that f_{xy} is known and fixed, the mean pairwise IBD between two clusters can be calculated as $\bar{f_{i,j}}$. Using these model components the probability of cluster relatedness given its distance to all other clusters can be caluclated: $\Pr(\rho_{ij}|d_{ij})$.

The model components can be calculated as:

$$\rho(u,u) = m_{i,u} m_{j,u} \bar{f_{ii}}$$

$$\rho(u,v) = m_{i,u} m_{j,v} \bar{f_{ij}}$$

From these probabilities, I can calculate the likelihood of the distance data as:

$$L(d_{ij}|\rho_{ij}) = \prod_{i=1}^{I} \prod_{j \neq i}^{J} \Pr(\rho_{ij}|d_{ij}),\ i \in n$$

Bayesian Spatial Generalized Linear Mixed Models Predictors

Predictors of pairwise-IBD aggregated at the province level were assessed using generalized linear models within a Bayesian framework with two outcomes of interest: (1) within-province mean pairwise IBD and (2) between-province mean pairwise IBD. For within-province models, I fit a generalized linear mixed model with spatially correlated random effects. Following the framework described in Lee 2017, I assume that there are P total provinces that are non-overlapping regions that are indexed by $k \in 1, \ldots, P$, such that:

$$IBD_k|\mu_k \sim N(\mu_k, \nu^2)$$

$$\mu_k \sim \beta_k^T x_{predictors,k} + \phi_k + \epsilon$$

I modeled spatial autocorrelation, ϕ_k, using a conditional autoregressive (CAR) prior. The CAR prior followed the formulation provided in Leroux *et al.* 2000, where dependence is

assumed to consist of two joint random effects: (1) dependence among observations, and (2) spatial dependence among observations. Model priors were as follows:

$$\beta \sim MVN(\mu_\beta,\ \Sigma_\beta)$$

$$\phi_k|\phi_{-k}, W, \tau^2, \rho \sim N(\frac{\rho\sum_{i=1}^{K} w_{ki}\phi_i}{\rho\sum_{i=1}^{K} w_{ki}+1-\rho},\ \frac{\tau^2}{\rho\sum_{i=1}^{K} w_{ki}+1-\rho})$$

$$\tau^2 \sim IG(1,\ 0.01)$$

$$\rho \sim U(0,\ 1)$$

$$\nu^2 \sim IG(1,\ 0.01)$$

The multivariate Gaussian mean prior μ_β vector consisted of zeroes and the diagonal elements of the covariance matrix, σ_β, were set to 50,000 [54]. I allowed the ρ parameter to vary under the model in order to fit the spatial process that was most consistent with the data [54,55]. For context, when $\rho = 1$, the CAR term assumes all spatial autocorrelation is modeled by the random effects (i.e. the Intrinsic CAR or Besag model). In contrast, when $\rho = 0$, the CAR term assumes no spatial autocorrelation. The adjacency matrix, W, was a neighborhood matrix, where cells were scaled distances between province centroids [56]. Scaled distances were expected to follow an exponential distribution, with $\lambda = E(distance)$.

For models with between-province mean pairwise IBD as the outcome of interest, generalized linear models without random effects were considered. Provinces are now indexed by $i \in 1, \ldots, P.$ and $j \in 1, \ldots, P.$ Each model contained a covariate for the difference in distance between province centroids and an indicator value if the province pairwise-comparison was from the same province (i.e. when $i = j$). The covariate matrix consisted of pairwise squared-differences among the observed predictors. As such, the model was formulated as:

$$IBD_{ij}|\mu_{ij} \sim N(\mu_{ij}, \nu^2)$$

$$\mu_{ij} \sim \beta_{ij}^T x_{predictors,ij} + \beta(\Delta Distance_{ij}) + \beta(I(Prov_{ij})) + \epsilon$$

$$\beta \sim MVN(\mu_\beta, \Sigma_\beta)$$

$$\nu^2 \sim IG(1, 0.01)$$

As above, the covariates were assumed to have a multivariate Gaussian mean prior, where μ_β was a series of zeros and the diagonal elements of the covariance matrix, σ_β, were set to 50,000 [54]. Although the province-province indicator parameter is binomial, it was assumed to be Gaussian for computational efficiency. This province-province indicator parameter is included to account for the variance in the sample size and observation differences among provinces.

All models were first fit with 1,000,000 sample iterations and a 10,000 iteration burn-in. For each model, the minimum effective size of each parameter was assessed to determine how well the posterior was sampled. An *a priori* cutoff of 100,000 for the minimum effective size for all parameters was set as the minimum cutoff to indicate model convergence. Models were then compared using the Deviance Information Criterion (DIC) [57]. Once the best models were identified for each respective distance category, a final set of models was considered with 10,000,000 sample iterations and a 10,000 iteration burn-in. For these final models, I reported the posterior median, 2.5th percentile, and 97.5th percentile values (95% credible interval) for each model parameter.

Results

Summary Statistics

Using previously published data, I analyzed 1,111/2,039 *P. falciparum* isolates at 1,079/1,890 loci across 351/492 geographic clusters across the DRC (**Figure 4.1**) [8]. The number of isolates sequenced per cluster ranged from 1-11. Genetic autocorrelation among loci was low (mean ρ = 0.004, range: 0 - 0.017 with respect to the fourteen nuclear chromosomes; **Appendix 4.1 Figure 1**).

Among the 1,111 samples considered, the mean IBD estimate was 0.021 (range: 0 - 0.999). As expected, most pairwise comparisons had IBD estimates of zero (n=325,967/616,605). However, 62 sample-pairs had an IBD estimate of at least 0.5 (mean: 0.725, range: 0.502 - 0.999). Under a Wright-Fisher model with outgroup mating, these pairs are expected to be at least meiotic siblings and represent transmission events that are separated by only a single generation (hereafter referred to as "highly related pairs").

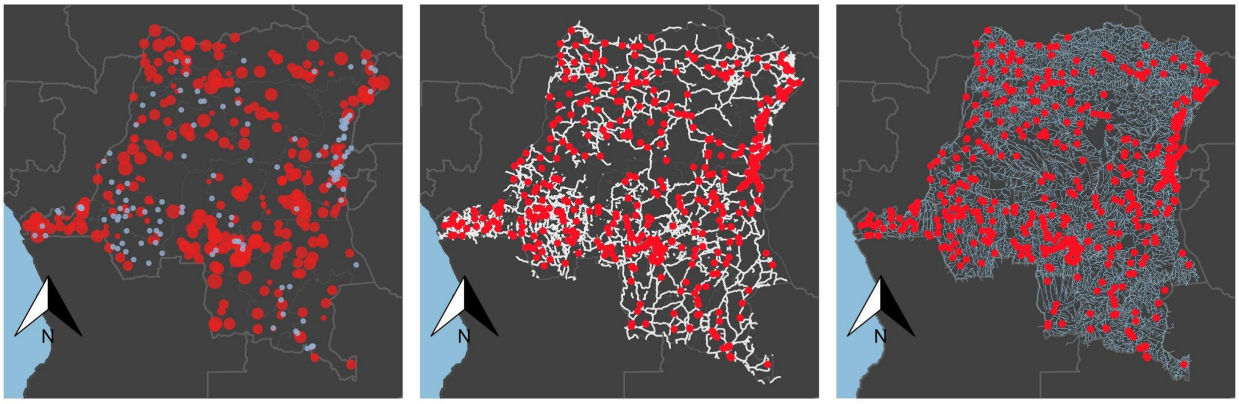


Figure 4.1 - Sampling Locations, Road & River Spatial Network: The 351 sampling locations are shown (red) alongside clusters from the 2013-2014 DRC DHS that were not included in this study (blue). Clusters sizes are scaled by the number of isolates sequenced from each sampling location (***Left***). Cluster locations (red) are displayed over the primary road network used to calculate road-distance between clusters. For aesthetics, only roads that were coded as "primary", "secondary", "tertiary", "motorway", "trunk", or "road" in the Geofabrik dataset were plotted. Additional roads were calculated in the shortest-path calculations (***Center***). Cluster locations (red) are displayed over the pruned river network that was used to calculate river-distance between clusters. For aesthetics, the river network plot only includes those rivers classified as "permanent" or "rivers" by DIVA-GIS and OSM, respectfully (***Right***).

Genetic Relatedness versus Geographic Distance

In order to determine how genetic relatedness varied with space, I compared measures of IBD across three geographic distances: greater-circle distance, road distance, and river distance. I first used Mantel tests to measure the correlation between the pairwise IBD relatedness measures

and the pairwise spatial distances. Mantel tests were statistically significant when considering each spatial category (**Table 4.2**).

Genetic Relatedness	Geographic Distance Category	Expected Disperser	Mantel-Test p-value	Log Likelihood
IBD	Greater-Circle	Mosquito	<0.01	-3,821,114,081.484
IBD	Road	Human	<0.01	-3,821,108,379.258
IBD	River	Human	<0.01	-3,821,112,409.554

Table 4.2 - Genetic-Geographic Statistics: Mantel test p-values comparing the correlation between genetic and geographic distances. For each of the respective geographical distances, 10,000 permutations were simulated and results were evaluated with a two-sided p-value. Log likelihoods were calculated using a non-parametric approach based on the classic isolation-by-distance model. Parameters are consistent between the distance-categories allowing for the direct comparison of log likelihoods. The road distance model demonstrated the best fit under the isolation by distance model framework.

To further explore the relationship between measures of genetic relatedness and space, I calculated the likelihood for spatial distance versus IBD using an isolation by distance framework. Although measures of pairwise IBD demonstrated a strong signal of isolation by distance across all three levels of spatial distance (**Figure 4.2**), I found that road distance provided the best fit (log-likelihood: -743205.502, **Table 4.2**). This finding is consistent with the observed nearly monotonic trend of decay in pairwise IBD as road distance increased. In contrast, greater-circle distance and river-distance had tails of higher than expected relatedness at far distances (**Figure 4.2**).

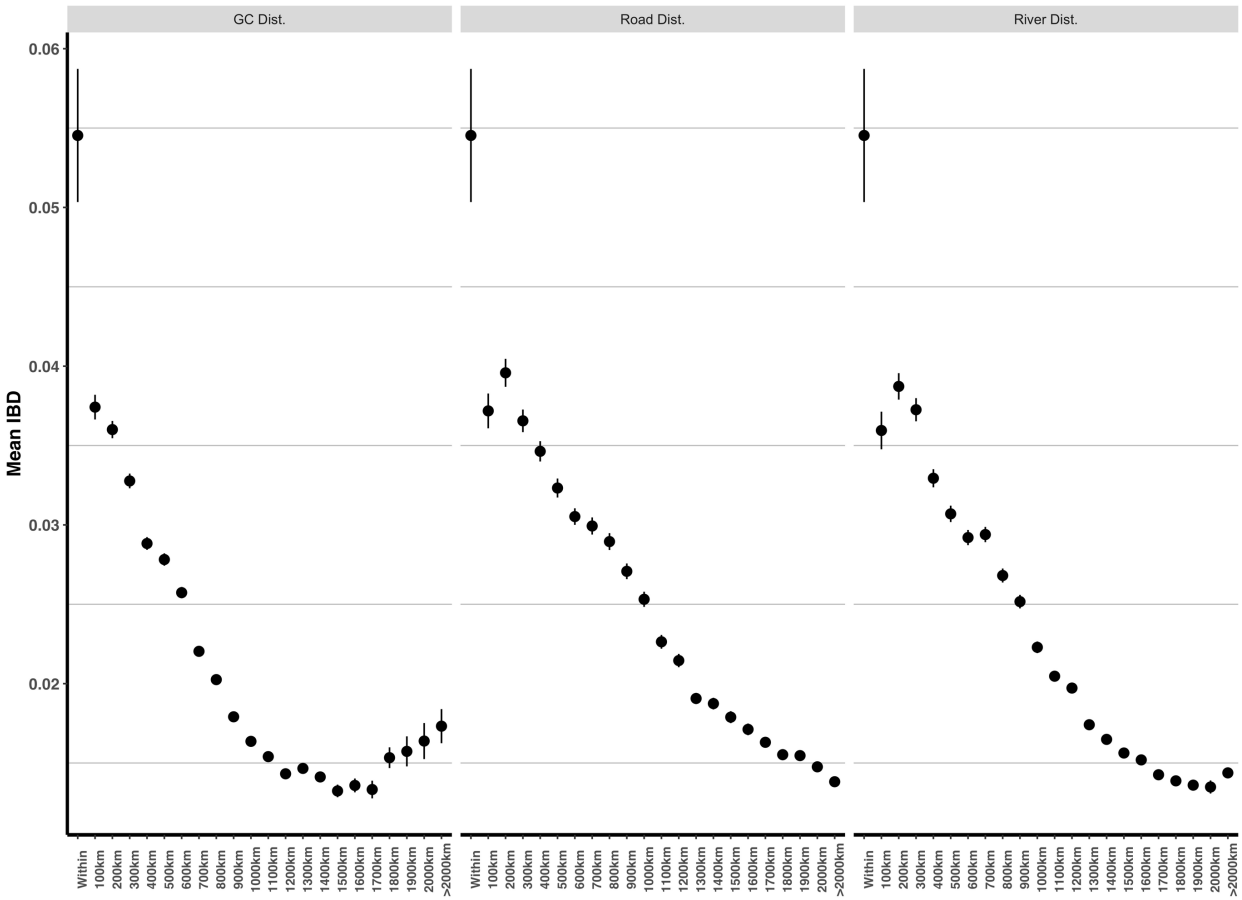


Figure 4.2 - Mean IBD Across Three Spatial Distances: The distribution of mean IBD with respect to the different distance metrics considered: greater-circle (***Left***), road (***Center***), river (***Right***). Distances were categorized following the distributions used in Verity, Aydemir, Brazeau *et al.* 2019.

Province Aggregation & Predictors of IBD

Mean within-province IBD ranged from 0.024 in Kasai to 0.090 in Nord-Ubangi (**Appendix 4.1 Figure 2)**. Among the 512 models that I considered with mean within-province IBD as the outcome of interest, the minimum effective sample size for any parameter was approximately, 127,575, 125,717, and 127,521 for greater-circle, road, and river distance, respectively. The DIC among all 512 models had a small range, indicating that additional covariates did not greatly improve model fit (range: -96.810, -85.0855). Based on the DICs, the best fitting models for greater-circle, road, and river distance were the same and included a

prevalence and precipitation covariate (**Table 4.3**). For all models, the precipitation and prevalence parameter effect estimates were negatively associated with the increasing levels of IBD. However, the magnitude of each effect estimate was small (**Table 4.3**). Among the three models, the estimation of autocorrelation was the same ($\rho = 0.683$).

Overall, between province IBD did not appear to differ greatly from the DRC-wide measure of mean pairwise IBD (mean: 0.021, range: 0.007 - 0.090; **Appendix 4.1 Figure 3**). When considering mean between-province IBD as the outcome of interest, the minimum effective sample size for a given parameter among the 512 model evaluated was 937,868, 935,889, and 93,4704 for greater-circle, road, and river distance, respectively. As above, the DIC range among the models was small, which indicates that covariates did not greatly improve model fit (range: -2,283.413, -2,251.783). Among the between-province models, the best fitting model included a housing, urbanicity, and precipitation covariate for all spatial frameworks considered (**Table 4.3**). In all cases, the housing, urbanicity, and precipitation parameter effect estimates were negatively associated with IBD. This suggests that a larger difference in covariates between province results in less between-province relatedness.

Outcome	Distance Category	Parameter	Median	2.50%	97.50%	Effective Size
Within Province	Greater Circle	Int.	0.059	0.043	0.075	10,000,000
		Precip.	-0.014	-0.033	0.006	5,575,000
		Par. Rate	-0.017	-0.036	0.002	5,721,569
		ν^2	0.002	0.001	0.003	5,730,592
		τ^2	0.003	0.001	0.008	2,142,403
		ρ	0.683	0.222	0.965	1,514,903
	Road	Int.	0.059	0.043	0.075	10,002,764
		Precip.	-0.014	-0.033	0.005	5,713,636
		Par. Rate	-0.017	-0.036	0.002	5,801,656
		ν^2	0.002	0.001	0.003	5,718,237
		τ^2	0.003	0.001	0.008	2,134,699
		ρ	0.683	0.221	0.966	1,511,499
	River	Int.	0.059	0.043	0.075	9,990,267
		Precip.	-0.014	-0.033	0.006	5,629,140
		Par. Rate	-0.017	-0.036	0.002	5,771,800
		ν^2	0.002	0.001	0.003	5,672,075
		τ^2	0.003	0.001	0.008	2,136,912
		ρ	0.683	0.223	0.966	1,521,384

Between Province	Greater Circle	Int.	0.017	0.015	0.019	10,000,000
		Housing	-0.001	-0.002	0.000	10,000,000
		Precip.	-0.002	-0.003	-0.001	10,000,000
		Urban	-0.002	-0.003	0.000	10,000,000
		Prov.	0.006	0.001	0.012	10,000,000
		Dist.	-0.008	-0.011	-0.006	10,000,000
		ν^2	0.000	0.000	0.000	9,662,049
	Road	Int.	0.023	0.021	0.024	10,000,000
		Housing	-0.001	-0.002	0.000	10,020,694
		Precip.	-0.002	-0.003	-0.001	10,000,000
		Urban	-0.002	-0.003	0.000	10,000,000
		Prov.	0.007	0.001	0.012	10,000,000
		Dist.	-0.004	-0.006	-0.003	10,000,000
		ν^2	0.000	0.000	0.000	9,669,512
	River	Int.	0.021	0.020	0.023	9,966,331
		Housing	-0.001	-0.002	0.000	10,000,000
		Precip.	-0.002	-0.003	0.000	10,007,827
		Urban	-0.002	-0.003	0.000	10,000,000
		Prov.	0.005	-0.001	0.010	10,000,000
		Dist.	-0.006	-0.008	-0.005	10,000,979
		ν^2	0.000	0.000	0.000	9,644,796

Table 4.3 - Bayesian Spatial Generalized Linear Mixed Models: The relationship between province-level IBD and hypothesized predictors of IBD was modeled using spatial generalized linear mixed models. Models were evaluated using within-province mean pairwise-IBD and between-province mean pairwise-IBD as outcomes of interest. Parameter effect estimates are provided for precipitation (Precip.), parasite rate (Par. Rate), mean urbanicity (Urban), housing quality (Housing), the province-level indicator (Prov.), between-province distance (Prov.), and the scaling parameter for the variance of the outcome (Gaussian distribution). The covariate effective sample size is also provided.

Highly Related Samples

I identified 62 highly related pairs across 58/351 (16.5%) clusters. Among the 62 highly related pairs, 19 pairs were from the same cluster, indicating likely local transmission events (**Figure 4.3**). Of the highly related pairs that had only a single connection (n=25), most were from different clusters (n=13; **Figure 4.4A**). Similarly, most highly related pairs were from different clusters (n=30) among the 37 pairs that had more than one connection (**Figure 4.4B**). One set of pairs in cluster 284 formed a quadrad, where all four samples had at least 0.721 of the genome IBD but were from four separate households (**Figure 4.4B**). Both putative local and long-range transmission events among the highly related samples appeared to be relatively dispersed across the DRC (minimum geographic distance: 25.12 km, **Appendix 4.1 Table 3**).

Two samples violated the expectation of transitivity among pairwise IBD estimates and have a dyadic and triadic relationship with network clusters that are unconnected. Both of these IBD estimates hovered around 0.50 and are likely error due to the IBD-MLE calculation and not true violations of transitivity.

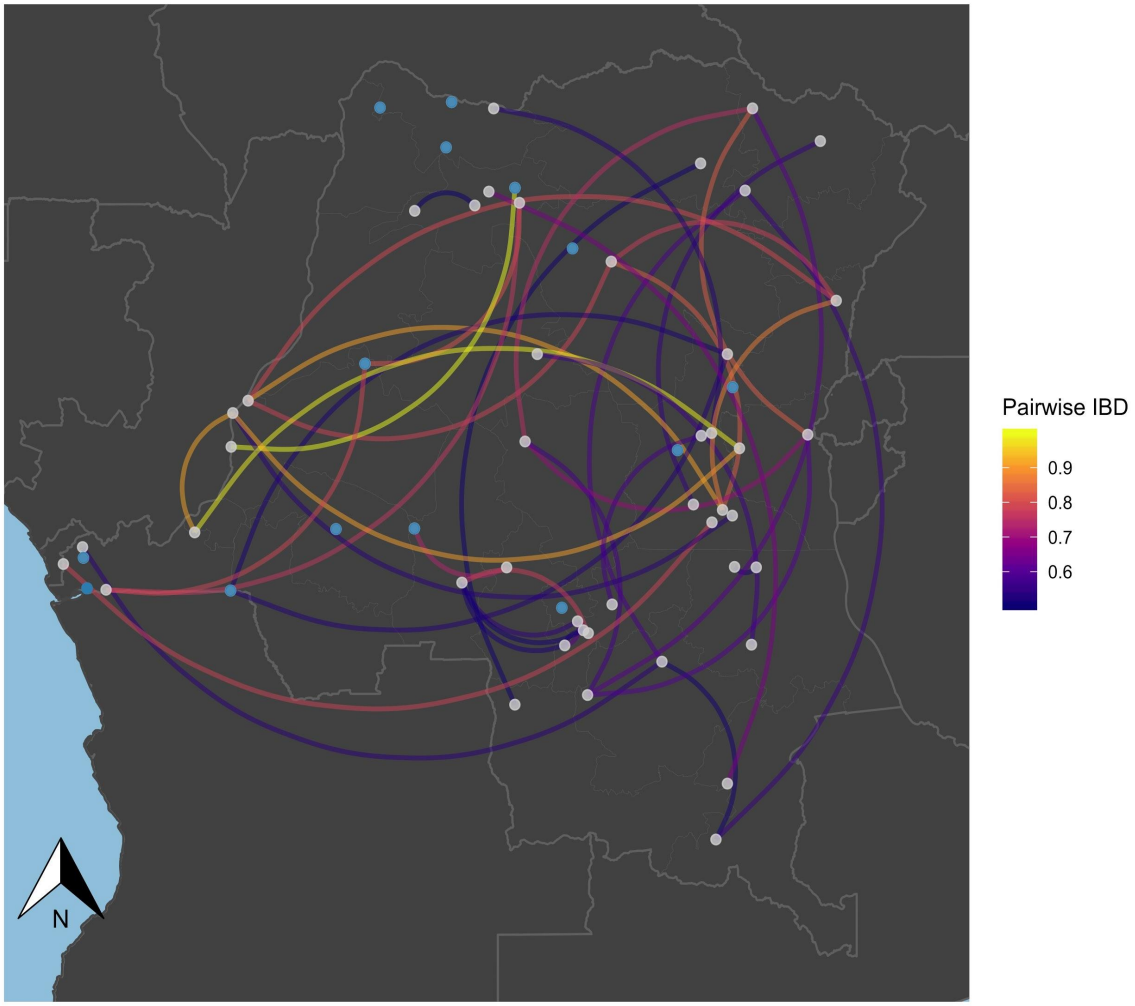


Figure 4.3 - Putative Within and Between Cluster Transmission Among Highly Related Pairs across the DRC: Highly related pairs are mapped across the DRC with the color of the edges corresponding to the pairwise IBD. Clusters that contain at least one highly related pair with both samples originating and terminating in the same cluster are marked in blue. These pairs likely represent local transmission events. The majority of pairs were between clusters, where the minimum distance between clusters was 25.12 km, which exceeds the maximum flight distance of an anopheline mosquito (Kaufmann & Briegel 2004). This suggests that human movement between cluster may be driving these connections.

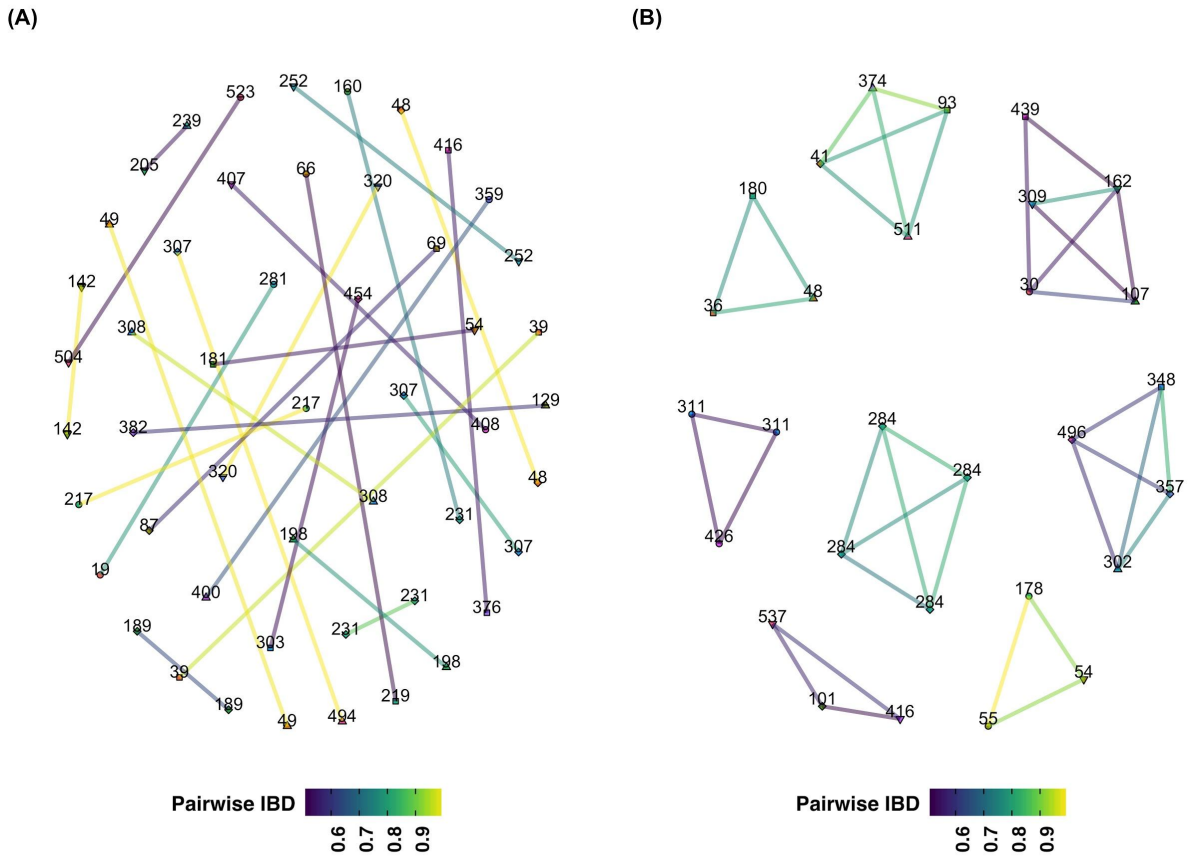


Figure 4.4 - Pairwise IBD Networks among Highly Related Samples: (A) Highly related pairs with one connection are displayed with cluster identifiers as the nodes. Overall, most highly related pairs with only a single connection were split between originating from the same cluster (n=12) or different clusters (n=13) (B) In contrast, among those highly related pairs with more than one connection, most stretched across multiple cluster locations (n=30). This indicates likely long-range transmission events. Most long-range transmission events exceed mosquito flight distances and were likely dispersed by human movement.

Given the finding that road networks were most consistent with the genetic isolation-by-distance pattern, I examined if urbanicity was associated with highly related pairs. Among the 351 clusters sampled in this study, I identified 130 urban clusters. Among the highly related pairs, I identified 14 unique urban clusters among the 48 clusters that had at least one sample within a highly related pair. Using the permutation test, I found that there were fewer urban clusters than would be expected under an assumption of independence (**Appendix 4.1 Figure 4**).

However, this result was not statistically significant ($p > 0.05$). Overall, I identified 18 pairs that had both nodes in rural clusters, 22 pairs with urban-rural nodes, and 3 pairs with urban-urban nodes. Visualization of the urban and highly related pairs relationship suggests that the edges of highly-related pairs were more densely concentrated in urban areas across the DRC (**Figure 4.5**).

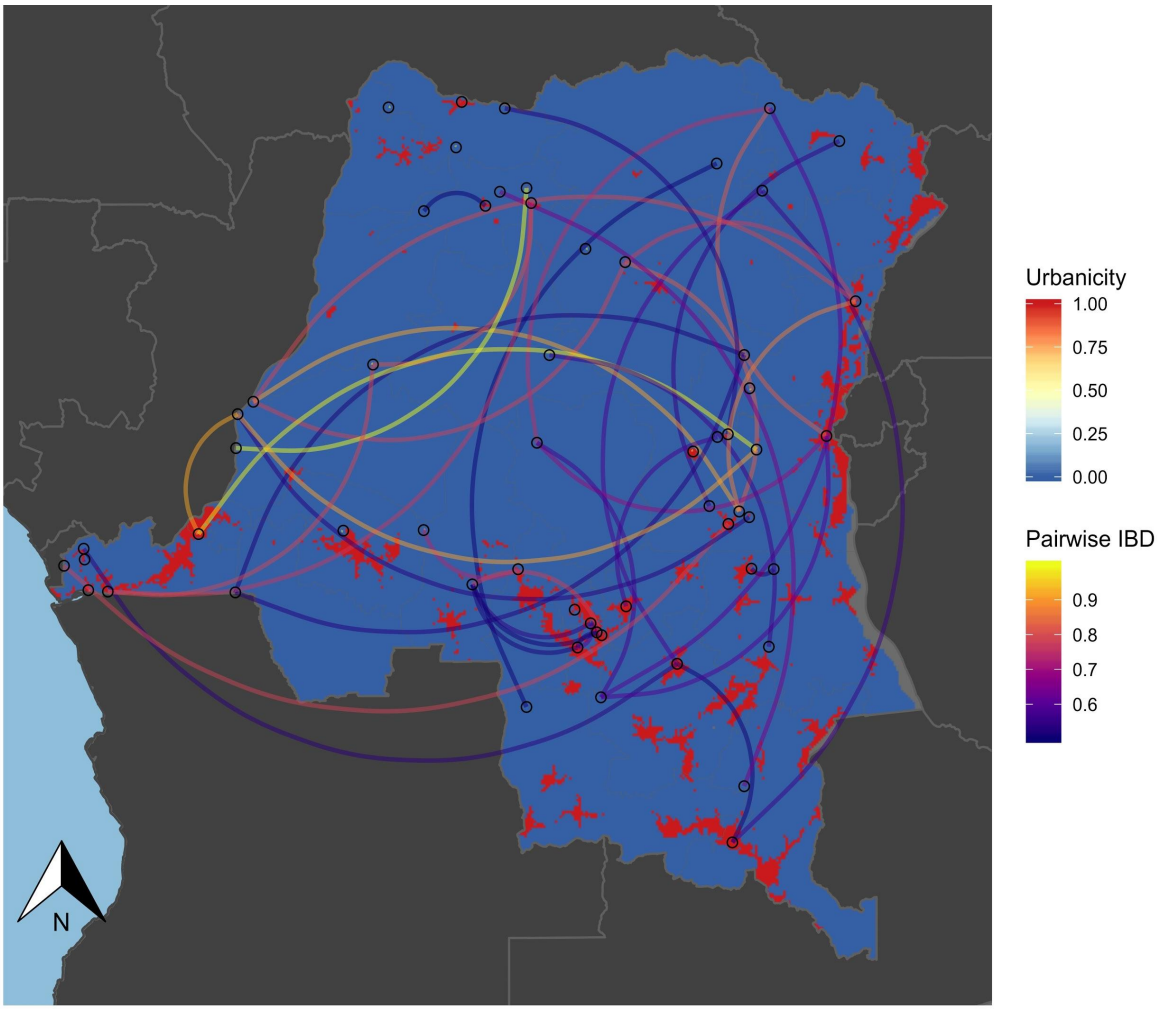


Figure 4.5 - Between Cluster Highly Related Samples and Urbanicity across the DRC: The binary urbanicity raster is overlain with the highly related pairs from different clusters. Urbanicity is considered as a binary with urban areas marked in red while rural areas are marked in blue. Overall, the highly related pairs show a high number urban-rural connections, which suggests that cities may be acting as hubs for parasite transmission.

I did not find evidence for a relationship between edge-density and cluster level parasite when all samples in highly related pairs were considered (distance-correlation: 0.282, $p > 0.05$).

Similarly, I did not find an association between edge density and parasite rate when evaluating only urban samples (distance-correlation: 0.340, $p > 0.05$) or rural samples (distance-correlation: 0.297, $p > 0.05$), respectively.

Discussion

In this study I used pairwise measures of IBD and different measures of geographical space to analyze patterns of genetic relatedness across the DRC. Given the interaction between IBD, geographic distance, and transmission intensity, measures of IBD can be used to better understand *P. falciparum* gene-flow and transmission across space. I showed that pairwise road distance was most consistent with the classical isolation by distance model among the DRC samples. In addition, I demonstrated that the minimum spatial distance between highly related pairs that were not from the same cluster was 25.12 km. This minimum exceeds the expected maximum flight pattern of an anopheline mosquito by at least 10-15 km [58]. This suggests that migration of *P. falciparum* parasites may be largely driven by human movement instead of mosquito ranges. This finding is consistent with several previous studies that have suggested that human movement is a major contributor to the dispersion of *P. falciparum* across large geographic regions [6,59].

Analysis of highly related pairs can offer insights into contemporary infection dynamics, as they represent very recent transmission events. *P. falciparum* parasites with pairwise-IBD measures of 0.5 or greater is consistent with shared ancestry in the previous generation, which is expected to be within the last 1-3 months [21,22]. Here I found evidence that suggested that highly related pairs were more frequently between urban and rural settings, although this result was not statistically significant. However, this indicates that cities and urban areas may be acting as hubs for genetic relatedness across the DRC.

From the province-aggregated models, I found that precipitation and parasite rate were negatively associated with within-province IBD. This suggests that as precipitation and parasite rates increase, within-province relatedness decreases. Previous work has shown that precipitation was negatively associated with *P. falciparum* prevalence in the DRC [60]. This negative relationship between *P. falciparum* prevalence and precipitation likely reflects vector dynamics, where too much rainfall may "wash away" vector breeding sites [61,62]. Assuming that prevalence is proportional to the effective population size, these findings fit with the expected pattern of IBD decreasing in larger populations [3,14,63]. When considering the between-province IBD models, I found that larger differences in mean urbanicity, housing quality, and precipitation resulted in less relatedness between provinces. This finding suggests that provinces with similar sociodemographics and weather patterns are more likely to share parasites. Although a distance covariate was considered in the between-province models, these findings likely reflect regional similarities and isolation by distance dynamics. In all models, the parameter effect estimates were small. This lack of signal may be partly due to the low variation in the within- and between-province IBD measures.

A major limitation of this study was that all genotypes were coerced to be monoclonal despite an overall mean complexity of infection of 2.23 among the samples [8]. Previous work has shown that this coercion results in a downward bias when calculating IBD with the MLE approach used in this study, such that relatedness between samples is underestimated as complexity of infection increases (see Supplementary Materials 2, Verity, Aydemir, Brazeau *et al.* 2019) [8]. As a result, my findings are likely conservative. Future work leveraging the information encoded in polyconal infections may be able to resolve finer levels of IBD and detect more nuanced transmission patterns.

A second limitation of the study was the few loci (n=1,079) that I considered for IBD measurements. Overall, the autocorrelation among the genomic positions was very low, which suggests limited linkage-disequilibrium for detecting recombination blocks. Given that I primarily focused on highly related pairs to infer connectedness between regions and that highly-related pairs are expected to share at least half of their genome, few markers are needed to infer relatedness [8,64]. However, future studies attempting to identify relatedness in more minute detail may require additional loci, depending on the study site transmission intensity and extent of the past queried. Finally, measures of IBD as a binary relationship between pairs of genomes in a pedigree framework is potentially underpowered [65]. Instead, methods characterizing the most recent common ancestors for each locus may be needed to truly resolve detailed measures of genetic relatedness.

A final limitation of this study is the curation of the river network used to calculate river distances between cluster pairs. Given that this included a component of manual-cleaning and editing, a degree of arbitrariness and misclassification bias is likely introduced into the dataset. Unfortunately, without an open-source API like Open Source Routing Machine, these are inherent limitations when working with this type of data.

To the best of my knowledge, this is the first study to use multiple measures of geographic space to explore patterns of *P. falciparum* IBD in a high burden region. Despite the high incidence in the DRC, I was still able to detect signals of isolation by distance due to human movement. Given the potential mixing between urban and rural regions, I hypothesize that infected individuals from high transmission rural areas may be importing parasites into low transmission urban areas, allowing for the long-range mixing of parasites [67]. Future control efforts may benefit from further characterizing and identifying these hubs for increased

interventions. Despite IBD and genetic relatedness measures having several limitations in high-transmission settings, relatedness can provide policymakers with an idea of where and how parasites are migrating across a country. In addition, identifying patterns of gene flow and migration patterns among *P. falciparum* parasite can help characterize how drug-resistance is likely to spread through a given region. Identifying these paths is critical for drug-resistance control efforts with respect to the imminent threat of artemisinin resistance being imported into the DRC [68,69]. Although there are current limitations in inferring relatedness among *P. falciparum* parasites in high-burden settings, these methods can help to inform public health officials on the likely migration patterns of parasites and warrant application.

REFERENCES

1. Miles, A. *et al.* Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. *Genome Research* vol. 26 1288–1299 (2016).

2. Zhu, S. J. *et al.* The origins and relatedness structure of mixed infections vary with local prevalence of P. falciparum malaria. *Elife* **8**, (2019).

3. Daniels, R. *et al.* Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. *PLoS One* **8**, e60780 (2013).

4. Omedo, I. *et al.* Micro-epidemiological structuring of Plasmodium falciparum parasite populations in regions with varying transmission intensities in Africa. *Wellcome Open Res* **2**, 10 (2017).

5. Daniels, R. F. *et al.* Modeling malaria genomics reveals transmission decline and rebound in Senegal. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7067–7072 (2015).

6. Taylor, A. R. *et al.* Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent. *PLoS Genet.* **13**, e1007065 (2017).

7. Andrew P Morgan, Nicholas F Brazeau, Billy Ngasala, Lwidiko Edward, Madeline Denton, Ulrika Morris, Ozkan Aydemir, Jeffrey A. Bailey, Jonathan Parr, Andreas Mårtensson, Anders Bjorkman, Jonathan J Juliano. Falciparum malaria from coastal Tanzania and Zanzibar remains highly connected despite effective control efforts on the archipelago.

8. Verity, R. J., Aydemir, O., Brazeau, N. F. & Watson, O. J. The Impact of Antimalarial Resistance on the Genetic Structure of Plasmodium falciparum in the DRC. *bioRxiv* (2019).

9. Wong, W., Wenger, E. A., Hartl, D. L. & Wirth, D. F. Modeling the genetic relatedness of Plasmodium falciparum parasites following meiotic recombination and cotransmission. *PLoS Comput. Biol.* **14**, e1005923 (2018).

10. Schaffner, S. F., Taylor, A. R., Wong, W., Wirth, D. F. & Neafsey, D. E. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *bioRxiv* 188078 (2017) doi:10.1101/188078.

11. Henden, L., Lee, S., Mueller, I., Barry, A. & Bahlo, M. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLoS Genet.* **14**, e1007279 (2018).

12. Shetty, A. C. *et al.* Genomic structure and diversity of Plasmodium falciparum in Southeast Asia reveal recent parasite migration patterns. *Nat. Commun.* **10**, 2665 (2019).

13. Amambua-Ngwa, A. *et al.* Major subpopulations of Plasmodium falciparum in sub-Saharan Africa. *Science* **365**, 813–816 (2019).

14. Watson, O. J., Okell, L. C., Joel, H., Slater, H. & Unwin, H. J. T. Evaluating the performance of malaria genomics for inferring changes in transmission intensity using transmission modelling. *bioRxiv* (2019).

15. Miotto, O. *et al.* Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia. *Nat. Genet.* **45**, 648–655 (2013).

16. Browning, S. R. & Browning, B. L. Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics* vol. 46 617–633 (2012).

17. Thompson, E. A. Correlations between relatives: From Mendelian theory to complete genome sequence. *Genet. Epidemiol.* **43**, 577–591 (2019).

18. Weir, B. S., Anderson, A. D. & Hepler, A. B. Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* **7**, 771–780 (2006).

19. Wakeley, J. *Coalescent theory: an introduction*. (2009).

20. White, N. J. *et al.* Malaria. *Lancet* **383**, 723–735 (2014).

21. Huber, J. H., Johnston, G. L., Greenhouse, B., Smith, D. L. & Perkins, T. A. Quantitative, model-based estimates of variability in the generation and serial intervals of Plasmodium falciparum malaria. *Malar. J.* **15**, 490 (2016).

22. Churcher, T. S. *et al.* Public health. Measuring the path toward malaria elimination. *Science* **344**, 1230–1232 (2014).

23. Thompson, E. A. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* **194**, 301–326 (2013).

24. Wright, S. Isolation by Distance. *Genetics* **28**, 114–138 (1943).

25. Rousset, F. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219–1228 (1997).

26. Taylor, S. M. *et al.* Plasmodium falciparum sulfadoxine resistance is geographically and genetically clustered within the DR Congo. *Sci. Rep.* **3**, 1165 (2013).

27. Deutsch-Feldman, M. *et al.* The changing landscape of Plasmodium falciparum drug resistance in the Democratic Republic of Congo. *BMC Infect. Dis.* **19**, 872 (2019).

28. Plowe, C. V., Djimde, A., Bouare, M., Doumbo, O. & Wellems, T. E. Pyrimethamine and proguanil resistance-conferring mutations in Plasmodium falciparum dihydrofolate reductase: polymerase chain reaction methods for surveillance in Africa. *Am. J. Trop. Med. Hyg.* **52**, 565–568 (1995).

29. Aydemir, O. *et al.* Drug-Resistance and Population Structure of Plasmodium falciparum Across the Democratic Republic of Congo Using High-Throughput Molecular Inversion Probes. *J. Infect. Dis.* **218**, 946–955 (2018).

30. Malécot, G. *The Mathematics of Heredity*. (W. H. Freeman, 1970).

31. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* vol. 10 439–446 (2018).

32. Luxen, D. & Vetter, C. Real-time routing with OpenStreetMap data. in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* 513–516 (ACM, 2011).

33. Douglas, D. H. & Peucker, T. K. Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization* **10**, 112–122 (1973).

34. Wan, Z., Hook, S., Hulley, G. MYD11C3 MODIS/Aqua Land Surface Temperature/Emissivity Monthly L3 Global 0.05Deg CMG V006 [Data set]. NASA EOSDIS Land Processes DAAC. (2015).

35. Funk, C. *et al.* The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data* **2**, 150066 (2015).

36. Nicholas F. Brazeau, Cedar L. Mitchell, Andrew P. Morgan, Molly Deutsch-Feldman, Oliver John Watson, Kyaw L. Thwai, Andreea Waltmann, Michael Emch, Valerie Gartner, Ben Redelings, Greg Wray, Melchior K. Mwandagalirwa, Antoinette K. Tshefu, Joris L. Likwela, Jonathan B. Parr, Jessica Edwards, Robert Verity, Steven R. Meshnick, Jonathan J. Juliano. The Epidemiology of P. vivax Among Adults in the Democratic Republic of the Congo: A Nationally-Representative, Cross-Sectional Survey.

37. Hollister, J. & Tarak Shah. elevatr: Access Elevation Data from Various APIs. (2017).

38. Linard, C., Gilbert, M., Snow, R. W., Noor, A. M. & Tatem, A. J. Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS One* **7**, e31743 (2012).

39. Weiss, D. J. *et al.* A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* vol. 553 333–336 (2018).

40. Mills, S., Weiss, S. & Liang, C. VIIRS day/night band (DNB) stray light characterization and correction. in *Earth Observing Systems XVIII* vol. 8866 88661P (International Society for Optics and Photonics, 2013).

41. Elvidge, C. D., Baugh, K., Zhizhin, M., Hsu, F. C. & Ghosh, T. VIIRS night-time lights. *Int. J. Remote Sens.* **38**, 5860–5879 (2017).

42. Santoro M., Kirches G., Wevers J., Boettcher M., Brockmann C., Lamarche C., Bontemps S., Moreau I., Defourny P. ESA Land Cover CCI, Climate Change Initiative (ESA LC CCI) data: LC 2.0 via Centre for Environmental Data Analysis. (2019).

43. Bhatt, S. *et al.* The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature* **526**, 207–211 (2015).

44. Tusting, L. S. *et al.* Mapping changes in housing in sub-Saharan Africa from 2000 to 2015. *Nature* vol. 568 391–394 (2019).

45. Croft, T. N., Marshall, A. M. J., Allen, C. K. & Others. Guide to DHS statistics. *Rockville, Maryland, USA: ICF* (2018).

46. Mayala, B., Fish, T. D., Eitelberg, D. & Dontamsetti, T. *The DHS Program Geospatial Covariate Datasets Manual*. (2018).

47. Marivoet, W. & De Herdt, T. Tracing Down Real Socio-Economic Trends From Household Data With Erratic Sampling Frames: The Case of the Democratic Republic of the Congo. *J. Asian Afr. Stud.* **53**, 532–552 (2018).

48. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209–220 (1967).

49. Newman, M. *Networks: An Introduction*. (OUP Oxford, 2010).

50. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **35**, 2769–2794 (2007).

51. Székely, G. J. & Rizzo, M. L. Brownian distance covariance. *Ann. Appl. Stat.* **3**, 1236–1265 (2009).

52. Rizzo, M. & Szekely, G. energy: E-Statistics: Multivariate Inference via the Energy of Data. (2019).

53. Malécot, G. [Not Available]. *Ann. Genet. Sel. Anim.* **5**, 333–361 (1973).

54. Lee, D. CARBayes version 4.6: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors. *University of Glasgow, Glasgow* (2017).

55. Leroux, B. G., Lei, X. & Breslow, N. Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence. in *Statistical Models in Epidemiology, the Environment, and Clinical Trials* 179–191 (Springer New York, 2000).

56. Bivand, R. S., Pebesma, E. & Gómez-Rubio, V. *Applied Spatial Data Analysis with R*. (Springer, New York, NY, 2013).

57. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* vol. 64 583–639 (2002).

58. Kaufmann, C. & Briegel, H. Flight performance of the malaria vectors Anopheles gambiae and Anopheles atroparvus. *J. Vector Ecol.* **29**, 140–153 (2004).

59. Wesolowski, A. *et al.* Quantifying the impact of human mobility on malaria. *Science* **338**, 267–270 (2012).

60. Janko, M. M. *et al.* The links between agriculture, Anopheles mosquitoes, and malaria risk in children younger than 5 years in the Democratic Republic of the Congo: a population-based, cross-sectional, spatial study. *Lancet Planet Health* **2**, e74–e82 (2018).

61. Briët, O. J. T., Vounatsou, P. & Amerasinghe, P. H. Malaria seasonality and rainfall seasonality in Sri Lanka are correlated in space. *Geospatial health* vol. 2 183 (2008).

62. Kipruto, E. K. *et al.* Effect of climatic variability on malaria trends in Baringo County, Kenya. *Malaria Journal* vol. 16 (2017).

63. Wesolowski, A. *et al.* Mapping malaria by combining parasite genomic and epidemiologic data. *BMC Med.* **16**, 190 (2018).

64. Taylor, A. R., Jacob, P. E., Neafsey, D. E. & Buckee, C. O. Estimating Relatedness Between Malaria Parasites. *Genetics* **212**, 1337–1351 (2019).

65. Speed, D. & Balding, D. J. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics* vol. 16 33–44 (2015).

66. Tessema, S. *et al.* Using parasite genetic and human mobility data to infer local and cross-border malaria connectivity in Southern Africa. *eLife* vol. 8 (2019).

67. Molly Deutsch-Feldman, Nicholas F. Brazeau, Jonathan B. Parr, Kyaw L. Thwai, Jérémie Muwonga, Melchior Kashamuka, Antoinette K. Tshefu, Jessie K. Edwards, Robert Verity, Michael Emch, Emily W. Gower, Jonathan J. Juliano, Steven R. Meshnick. Spatial and epidemiological drivers of P. falciparum malaria among adults in the Democratic Republic of the Congo.

68. Rosenthal, P. J. Artemisinin Resistance Outside of Southeast Asia. *Am. J. Trop. Med. Hyg.* **99**, 1357–1359 (2018).

69. Lu, F. *et al.* Emergence of Indigenous Artemisinin-Resistant Plasmodium falciparum in Africa. *N. Engl. J. Med.* **376**, 991–993 (2017).

CHAPTER FIVE: DISCUSSION

Summary of Results

Aim 1

***Plasmodium vivax* infections among adults in the Democratic Republic of the Congo (DRC) are more prevalent than previously realized.** In Aim 1 of my dissertation, I detected 467 *P. vivax* infections among 15,574 adults from the 2013-2014 Demographic Health Survey in the DRC. Overall, this resulted in a national prevalence of 2.96% (95% $CI_{weighted}$: 2.28, 3.65%). Among those adults that were infected with *P. vivax*, nearly all were Duffy-negative (576/579, 99.48%). Among the fourteen malaria risk factors considered, only higher-levels of precipitation were found to reduce *P. vivax* prevalence while being a farmer appeared to increase *P. vivax* prevalence.

When including these covariates in the spatial prediction models, the prevalence of *P. vivax* mapped relatively uniformly across the DRC with the exception of a few focal locations of relatively high prevalence (prevalence range: 0.50 - 11.20%). These "hotspots" are consistent with previous reports of spatial microheterogeneity in *P. vivax* transmission [1]. However, despite the presence of a few hotspots, *P. vivax* prevalence mostly ranged from 0.5 - 1.5% across the country, suggesting widespread low-level prevalence. Finally, when considering the evolutionary history of the DRC *P. vivax* infections, I found that the DRC samples shared a most recent common ancestor (MRCA) with a subset of samples from Peru. This MRCA with samples from South America suggests that the DRC samples may be part of an ancient lineage. However, the ancestry of these DRC *P. vivax* samples is not straightforward, as there were several limitations

to my phylogenetic analyses. As a result, I concluded that there was sufficient evidence to state that the DRC *P. vivax* infections were not the result of recent zoonotic transmission but further sequencing was needed to resolve the recent origins of these samples. Given the relatively flat prevalence of *P. vivax* across the DRC, the few associated risk factors, and the indication of a potentially ancestral lineage, I hypothesized that much of the burden of *P. vivax* in the DRC is clinically innocuous.

Aim 2

Based on measures of genetic relatedness, dispersion of *Plasmodium falciparum* infections in the DRC is likely dominated by human movement versus mosquito movement. In Aim 2 of my dissertation, I used measures of identity by descent (IBD) to analyze patterns of genetic isolation by distance based on three different measures of geographic distance: (1) greater-circle distance, (2) road distance, and (3) river distance. Greater-circle distance is expected to represent dispersion by mosquitoes, while road and river distance are expected to represent dispersion by humans. From the 1,111 *P. falciparum* isolates that were sequenced from the 2013-2014 Demographic Health Survey in the DRC, I found that road distance best explained the expected pattern of isolation by distance. This suggests that migration of *P. falciparum* parasites may be largely driven by human movement along roads instead of dispersion by mosquitoes. This indication of human movement being a main contributor of *P. falciparum* parasite dispersion was recapitulated by analyzing highly related pairs of isolates. When comparing highly related pairs of *P. falciparum* isolates, I found that pairs were more frequently between urban and rural areas (not statistically significant). Given that there are fair fewer urban areas than rural areas in the DRC, I hypothesized that urban areas were acting as hubs for genetic relatedness across the DRC. This hypothesis of urban regions serving as hubs of

genetic relatedness in the DRC is consistent with my finding that road distances best explained *P. falciparum* parasite dispersion. Overall, human movement -- particularly between urban and rural areas -- may be promoting genetic diversity by continually supplying parasites from high-transmission settings, which are predominantly in rural regions in the DRC, to low-transmission settings, which are typically cities in the DRC [2].

Context and Importance

Until recently, *P. vivax,* was an unrecognized cause of disease in sub-Saharan Africa. Vivax malaria is more difficult to diagnose, treat, and eliminate than other malarias due to its complex life-cycle and complicated treatment algorithm [3–5]. Although my results show it is much more common than previously thought, I have hypothesized that these infections are relatively innocuous and do not pose a substantial threat to malaria control efforts in the DRC (at this time).

Although controversial to call an infectious disease a relatively innocuous threat, the *P. vivax* data from the DRC indicates this use of language and a conservative approach. These "controversial calls" are needed if resources are to be allocated properly in an effort to maximize malaria control, as progress towards malaria elimination have plateaued in recent years [6]. This plateau in malaria elimination progress is occurring despite increases in long-lasting insecticidal net use and access to care [6]. Taken together, this slowing of case-reductions, despite increased intervention uptake, suggests that the cost-effectiveness of preventing each case of malaria is decreasing. Although costs are expected to be highest at the end of elimination campaigns, many malaria-endemic countries are far from this "endgame" consideration. For example, the DRC -- the main focus of my dissertation -- recorded approximately 27-million cases and 40,000 deaths due to malaria in 2018 [6]. In 2017, malaria cases and deaths were approximately 27-million and 40,000, respectively [6]. These numbers reflect the overall pattern that case-burden and mortality

have remained relatively stagnant in many countries with high prevalences despite intervention uptake [2,6,7].

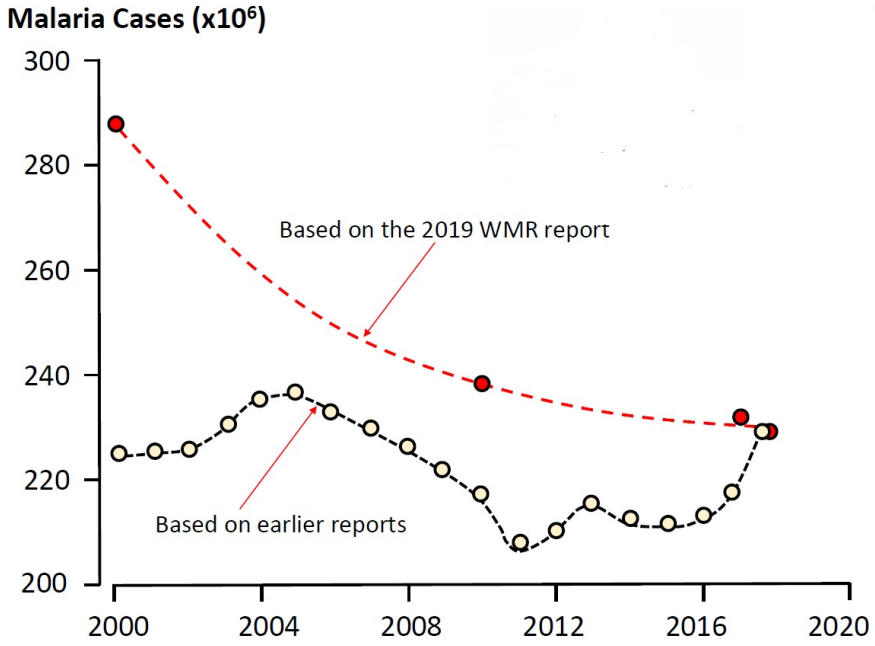


Figure 5.1 - Global Malaria Incidence: Based on the 2019 World Malaria Report from the World Health Organization (WHO), malaria incidence has declined from 2000 - 2018 (red line). However, in recent years, the rate of decline has shrunk dramatically. When considering previous WHO reports, global malaria incidence is no longer decreasing monotonically, and instead shows a complex pattern (black line). Although discrepancies between the two curves may be due to numerous factors, both curves suggest global malaria elimination progress has stalled. This figure was publicly posted by Ric Price (@ricprice99) on Twitter on December 5, 2019.

In response to stagnated progress in many high burden countries, the World Health Organization (WHO) announced a new strategy in 2019: High Burden, High Impact (HBHI) [6]. HBHI targets eleven countries with the highest burden of malaria with the intent of bringing these "countries back on track to achieve the 2025 [Global Technical Strategy] milestones" (DRC is ranked second in global malaria burden) [6]. In order to accomplish the goals of HBHI, new methods and approaches are needed.

Although molecular epidemiology has long been used to track drug-resistance among malaria parasites, recently, its capacity for identifying patterns of gene flow and parasite connectedness has been increasingly recognized [8–21]. Using various measures of genetic relatedness in combination with other epidemiological data sources, researchers can reconstruct likely "sources," or identify locations where parasites likely originated, and "sinks," where parasites reside. These migration models would provide public health officials with critical information, as interventions could then target the ultimate pool (i.e. the source) of the infectious reservoir instead of chasing the proximal pool (i.e. sinks or prevalence maps). Similar techniques have been used in epidemic modeling targeting "super-spreaders" to optimize intervention-distribution during outbreaks [22–24]. In this dissertation, I showed that falciparum malaria in the DRC is likely being spread by human movements along roads with an emphasis on cities as potential reservoirs. These data suggest that targeting malaria in cities -- despite an overall lower *P. falciparum* prevalence in urban areas versus rural areas -- would be a more effective strategy for control in the DRC. These findings are likely critical to reignite malaria elimination progress in HBHI countries, where current interventions are faltering if the goal of malaria eradication is to be realized by 2040 [25].

Future Work

If the status of *P. vivax* in sub-Saharan Africa shifts and *P. vivax* becomes a more virulent infection, malaria elimination efforts will be greatly complicated [25,26]. As a result, continued efforts to understand *P. vivax* in sub-Saharan Africa are necessary. In my thesis, I have only provided a small glimpse -- limited to only three mitochondrial genomes -- of the potential genetic diversity of sub-Saharan *P. vivax*. Future work should focus on collecting whole blood or higher parasitemic samples with a greater likelihood of whole genome sequencing success. With whole genomes, researchers would be able to identify regions of positive selection and putative

sites associated with erythrocyte/reticulocyte-invasion among Duffy-negative hosts. Whole genomes would also provide enough variant sites for demographic modeling, which could be combined with coalescent simulations and infectious disease mathematical models (i.e. SIR models) to predict the burden of *P. vivax* across Sub-Saharan Africa forward-in-time [27,28]. Demographic models would also allow for inference on the evolutionary origins and the effective population size of *P. vivax* in the distant past [29,30]. Additionally, a more complicated landscape genetics study could be undertaken that considers within-country phylogeography, genetic corridors, and migration patterns with programs such as *EEMS* [31].

From an epidemiological perspective, continued surveillance of *P. vivax* in sub-Saharan Africa is indicated to assess if disease prevalence is increasing. Future studies should consider incorporating travel data in efforts to resolve potential small-scale heterogeneity (as observed in my dissertation). Additionally, longitudinal sampling to assess for fluctuations in *P. vivax* incidence in the region due to seasonality is needed. Further work to incorporate mathematical models specific to *P. vivax* in sub-Saharan Africa may also identify populations at risk, shed light on transmission dynamics, and identify optimal intervention strategies [32,33].

Although IBD has proven to be extremely useful for characterizing malaria transmission dynamics, gene flow, and other population demographics, studies to date have largely been limited to monoclonal infections or methods that coerce polyclonal samples to monoclonal genotypes [12,14,18,34]. Previous work has shown that the coercion of polyclonal infections to monoclonal genotypes greatly reduces the power to detect pairwise IBD (see Supplementary Materials 2, Verity, Aydemir, Brazeau *et al.* 2019). However, characterization of gene-flow and migration patterns in high-transmission regions is likely to greatly aid in malaria elimination efforts (discussed above). Additionally, the characterization of gene-flow in sub-Saharan Africa

is urgent due to the emerging threat of importation of artemisinin-resistance mutations from Southeast Asia [35,36]. As a result, new methods are needed to account for polyclonality in IBD calculations. IBD calculations may also benefit by extending the definition of relatedness from a binary-measure at each locus (IBD Yes/No) to measuring time to the most recent common ancestor at each locus [37]. These latter methods are possible using the coalescent with recombination and ancestral recombination graph theory. Additionally, future work is needed to define what constitutes a "source" and "sink" with respect to malaria endemicity [38]. Identification and targeting of malaria parasite sources should cause genomic bottlenecking and potential collapse of the malaria population.

Conclusions

Although *Plasmodium vivax* is more prevalent in the DRC than previously recognized, these infections are likely innocuous and do not warrant urgent public health action. Additional surveillance of *P. vivax* in sub-Saharan Africa is needed, but current control measures aimed at reducing the burden of *P. falciparum* in the region should continue to mitigate the threat of *P. vivax*.

The dispersion of *P. falciparum* parasites in the DRC appears to be largely driven by human movement -- particularly along roads -- and not mosquito movement. Characterization of genetics hubs and gene-flow in high-transmission settings is crucial for identifying optimal intervention sites and routes to slow malaria transmission and the spread of emerging drug-resistance. Although counterintuitive, the targeting of urban settings in the DRC may reduce malaria prevalence country-wide.

REFERENCES

1. Carrasco-Escobar, G. *et al.* Micro-epidemiology and spatial heterogeneity of P. vivax parasitaemia in riverine communities of the Peruvian Amazon: A multilevel analysis. *Sci. Rep.* **7**, 8082 (2017).

2. Molly Deutsch-Feldman, Nicholas F. Brazeau, Jonathan B. Parr, Kyaw L. Thwai, Jérémie Muwonga, Melchior Kashamuka, Antoinette K. Tshefu, Jessie K. Edwards, Robert Verity, Michael Emch, Emily W. Gower, Jonathan J. Juliano, Steven R. Meshnick. Spatial and epidemiological drivers of P. falciparum malaria among adults in the Democratic Republic of the Congo.

3. Kevin Baird, J. Malaria caused by Plasmodium vivax: recurrent, difficult to treat, disabling, and threatening to life--the infectious bite preempts these hazards. *Pathog. Glob. Health* **107**, 475–479 (2013).

4. Baird, J. K., Valecha, N., Duparc, S., White, N. J. & Price, R. N. Diagnosis and Treatment of Plasmodium vivax Malaria. *Am. J. Trop. Med. Hyg.* **95**, 35–51 (2016).

5. Gruenberg, M. *et al.* Plasmodium vivax molecular diagnostics in community surveys: pitfalls and solutions. *Malar. J.* **17**, 55 (2018).

6. World Health Organization. World Malaria Report 2019. 232.

7. Taylor, S. M. *et al.* Molecular malaria epidemiology: mapping and burden estimates for the Democratic Republic of the Congo, 2007. *PLoS One* **6**, e16420 (2011).

8. Zhu, S. J. *et al.* The origins and relatedness structure of mixed infections vary with local prevalence of P. falciparum malaria. *Elife* **8**, (2019).

9. Daniels, R. *et al.* Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. *PLoS One* **8**, e60780 (2013).

10. Omedo, I. *et al.* Micro-epidemiological structuring of Plasmodium falciparum parasite populations in regions with varying transmission intensities in Africa. *Wellcome Open Res* **2**, 10 (2017).

11. Daniels, R. F. *et al.* Modeling malaria genomics reveals transmission decline and rebound in Senegal. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7067–7072 (2015).

12. Taylor, A. R. *et al.* Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent. *PLoS Genet.* **13**, e1007065 (2017).

13. Andrew P Morgan, Nicholas F Brazeau, Billy Ngasala, Lwidiko Edward, Madeline Denton, Ulrika Morris, Ozkan Aydemir, Jeffrey A. Bailey, Jonathan Parr, Andreas Mårtensson, Anders Bjorkman, Jonathan J Juliano. Falciparum malaria from coastal Tanzania and

Zanzibar remains highly connected despite effective control efforts on the archipelago.

14. Verity, R. J., Aydemir, O., Brazeau, N. F. & Watson, O. J. The Impact of Antimalarial Resistance on the Genetic Structure of Plasmodium falciparum in the DRC. *bioRxiv* (2019).

15. Wong, W., Wenger, E. A., Hartl, D. L. & Wirth, D. F. Modeling the genetic relatedness of Plasmodium falciparum parasites following meiotic recombination and cotransmission. *PLoS Comput. Biol.* **14**, e1005923 (2018).

16. Schaffner, S. F., Taylor, A. R., Wong, W., Wirth, D. F. & Neafsey, D. E. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *bioRxiv* 188078 (2017) doi:10.1101/188078.

17. Henden, L., Lee, S., Mueller, I., Barry, A. & Bahlo, M. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLoS Genet.* **14**, e1007279 (2018).

18. Shetty, A. C. *et al.* Genomic structure and diversity of Plasmodium falciparum in Southeast Asia reveal recent parasite migration patterns. *Nat. Commun.* **10**, 2665 (2019).

19. Amambua-Ngwa, A. *et al.* Major subpopulations of Plasmodium falciparum in sub-Saharan Africa. *Science* **365**, 813–816 (2019).

20. Watson, O. J., Okell, L. C., Joel, H., Slater, H. & Unwin, H. J. T. Evaluating the performance of malaria genomics for inferring changes in transmission intensity using transmission modelling. *bioRxiv* (2019).

21. Miotto, O. *et al.* Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia. *Nat. Genet.* **45**, 648–655 (2013).

22. Christakis, N. A. & Fowler, J. H. Social network sensors for early detection of contagious outbreaks. *PLoS One* **5**, e12948 (2010).

23. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).

24. Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nat. Phys.* **6**, 888–893 (2010).

25. Gates, B. & Chambers, R. From aspiration to action. *What will it take to end malaria. Available at* (2015).

26. Olliaro, P. L. *et al.* Implications of Plasmodium vivax Biology for Control, Elimination, and Research. *Am. J. Trop. Med. Hyg.* **95**, 4–14 (2016).

27. Chang, H.-H. *et al.* Genomic sequencing of Plasmodium falciparum malaria parasites from

Senegal reveals the demographic history of the population. *Mol. Biol. Evol.* **29**, 3427–3439 (2012).

28. Lessler, J., Azman, A. S., Kate Grabowski, M., Salje, H. & Rodriguez-Barraquer, I. Trends in the Mechanistic and Dynamic Modeling of Infectious Diseases. *Curr Epidemiol Rep* **3**, 212–222 (2016).

29. Liu, W. *et al.* African origin of the malaria parasite Plasmodium vivax. *Nat. Commun.* **5**, 3346 (2014).

30. Loy, D. E. *et al.* Out of Africa: origins and evolution of the human malaria parasites Plasmodium falciparum and Plasmodium vivax. *Int. J. Parasitol.* (2016) doi:10.1016/j.ijpara.2016.05.008.

31. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).

32. White, M. T. *et al.* Mathematical modelling of the impact of expanding levels of malaria control interventions on Plasmodium vivax. *Nat. Commun.* **9**, 3300 (2018).

33. White, M. T. *et al.* Modelling the contribution of the hypnozoite reservoir to Plasmodium vivax transmission. *Elife* **3**, (2014).

34. Tessema, S. *et al.* Using parasite genetic and human mobility data to infer local and cross-border malaria connectivity in Southern Africa. *eLife* vol. 8 (2019).

35. Rosenthal, P. J. Artemisinin Resistance Outside of Southeast Asia. *Am. J. Trop. Med. Hyg.* **99**, 1357–1359 (2018).

36. Lu, F. *et al.* Emergence of Indigenous Artemisinin-Resistant Plasmodium falciparum in Africa. *N. Engl. J. Med.* **376**, 991–993 (2017).

37. Speed, D. & Balding, D. J. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics* vol. 16 33–44 (2015).

38. Dias, P. C. Sources and sinks in population biology. *Trends Ecol. Evol.* **11**, 326–330 (1996).

APPENDIX 3.1: IDENTIFYING THE RISK, DISTRIBUTION, AND ORIGIN OF P. VIVAX IN THE DEMOCRATIC REPUBLIC OF THE CONGO

Molecular Diagnostics

P. vivax Infection Detection

DNA was extracted from dried blood spots using Chelex-100 (Bio-Rad, Hercules, CA) and Saponin as previously described [1,2]. *P. vivax* infections were detected using a two-stage approach that combined a TaqMan quantitative PCR (qPCR) assay targeting the 18S rRNA gene and a confirmatory nested-PCR assay [3,4]. A two-step approach was utilized to increase specificity and limit potential false positives given the range of cycle-threshold (CT) values considered. The qPCR recipe was as follows: 6 µL FastStart Universal Probe Master Mix (Roche Diagnostics, Indianapolis, IN), 0.24 µL of forward primer (20 µM), 0.24 µL of reverse primer (20 µM), 0.24 µL of probe (10 µM) and 3.28 µL of molecular grade water (**Appendix 3.1 Table 1**). Reactions were ran on a QuantStudio 6 Flex Real-Time PCR System (ThermoFisher Scientific, Waltham, MA, USA) using the following thermocycler conditions: 50 °C for 2-minutes, 95 °C for 10-minutes, followed by 45 cycles of annealing at 95 °C for 15-seconds and denaturing at 60 °C for 1-minute. All bulk qPCR reactions included two replicates of positive controls (serial dilutions from 4,550 parasites/µL (10^{-4} ng/µL) to 4.55 parasites/µL (10^{-7} ng/µL), assuming 6 copies of 18S parasite), and four negative template controls [5,6].

The nested confirmatory PCR assay involved two steps: (1) amplification of a general region of the *Plasmodium* 18S gene (outer reaction); (2) amplification of a *P. vivax* specific region (inner reaction). For both the outer- and inner-reaction, the PCR recipe was as follows: 12.5 µL HotStarTaq Master Mix (Qiagen©, Venlo, Netherlands), 0.5µL (20uM primers) of the forward and reverse primer, and 6.5 µL of molecular grade water (25 µL final reaction volume; **Appendix 3.1 Table 2**). Reactions were performed on a BioRad T100 Thermal Cycler (Applied

Biosystems, Foster City, CA, USA) using the following conditions: 95°C for 15-minutes, followed by 35 cycles of 94°C for 1-minute, 50°C for 1-minute (inner)/62°C for 1-minute (outer), and 72°C for 1-minute, with a final extension of 72 °C for 10-minutes. For the inner reaction, product from the outer reaction was used as the template (no cleaning was performed between reactions). The product of the inner reaction was then visualized with gel electrophoresis to confirm the presence of *P. vivax* DNA (expected band size was 121 base-pairs). For each confirmatory PCR reaction, two reviewers independently assessed the gel for the presence/absence of a band. A confirmed infection was only considered when the reviewers were in agreement. Among the 579/17,972 qPCR-positive samples, the inter-observer agreement between the absence/presence of a PCR band was high (Agreement: 564/579, Cohen's $\kappa = 0.80$, $p < 0.05$).

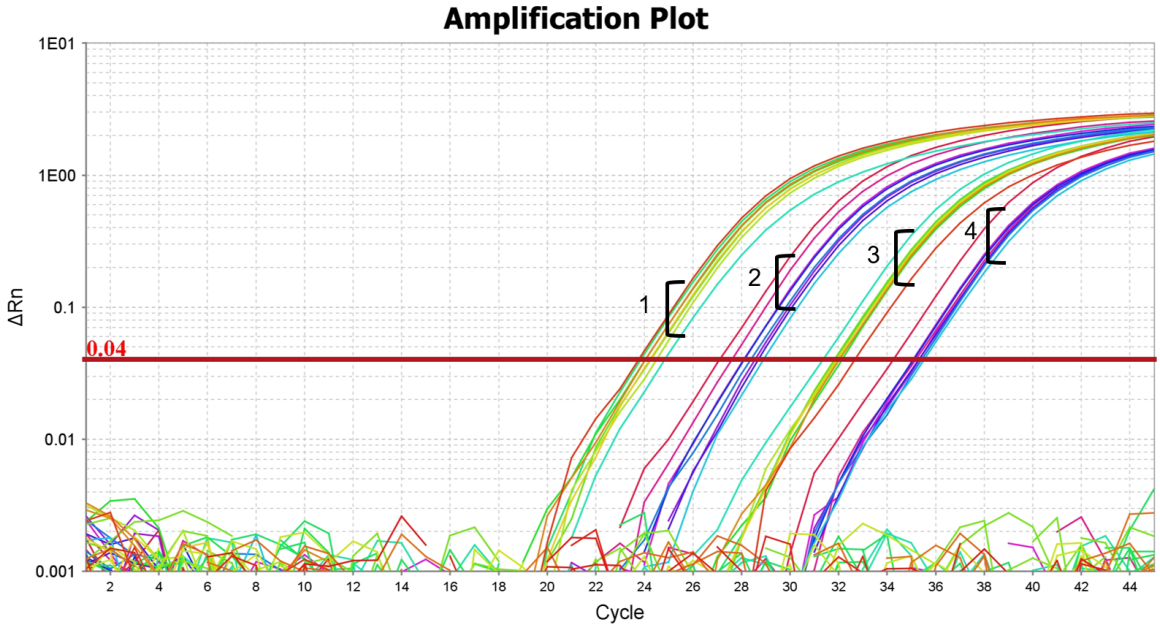
To test the specificity and sensitivity of the qPCR approach, the qPCR assay was challenged with high concentrations ($1x10^{-4}$ ng/µL) of *P. falciparum* (MRA-177, BEI Resources), *P. ovale* (MRA-180, BEI Resources), and *P. malariae* (MRA-179, BEI Resources) 18S plasmid DNA. For each species, 22 replicates were performed. For all reactions, no off-target amplification was appreciated (**Appendix 3.1 Figure 1**). In addition, *P. vivax* plasmid was serially diluted from $1x10^{-6}$ ng/µL to $0.03125x10^{-6}$ ng/µL to detect the assay lower limit of detection. The qPCR assay was able to detect *P. vivax* parasites with a sensitivity of approximately 90% (20/22) when the concentration of 18S plasmid was at least $1.25x10^{-7}$ ng/µL. Below this concentration, the assay was less reliable (**Appendix 3.1 Figure 1**).

Assay	Primer	Sequence	Ref.
Diagnostic qPCR	PvForward	5'-ACGCTTCTAGATTAATCCACATAACT	3
	PvReverse	5'-ATTTACTCAAAGTAACAAGGACTTCCAAGC	
	Pv-probe (FAM-IowaBlack)	5'-TTCGTATCG/ZEN/ACTTTGTGCGCATTTTGC	

Appendix 3.1 Table 1 - *P. vivax* qPCR Assay Primers: The primers and citations used for the qPCR assay. Adaptations to the probe from the original publication are indicated.

Assay	Primer	Sequence	Ref.
Confirmatory PCR	Plu1	5'-TCAAAGATTAAGCCATGCAAGTGA	4
	Plu5	5'- CCTGTTGTTGCCTTAAACTCC	
	rVivi1	5'-CGCTTCTAGCTTAATCCACATAACTGATAC	
	rVivi2	5'-ACTTCCAAGCCGAAGCAAAGAAAGTCCTTA	

Appendix 3.1 Table 2 - *P. vivax* Confirmatory PCR Reaction Primers: The primers used for the inner- and outer-reactions in the confirmatory PCR reaction are listed. The original reference for the reactions is also provided.



Appendix 3.1 Figure 1 - *P. vivax* qPCR Challenge: The qPCR assay was challenged with 22 replicates of highly concentrated DNA from three non-vivax 18S targets. Among these 66 replicates, no off-target amplification was appreciated. In addition, the lower limit of detection of the assay for *P. vivax* 18S was determined as approximately 1.25×10^{-4} ng/µL (section 3).

Duffy-Genotype

For each sample that was positive by qPCR, I used a previously validated high-resolution melt (HRM) assay to genotype the GATA-1 transcription factor (-33 T:C) point mutation that has been previously shown to silence Duffy Antigen/Chemokine Receptor (DARC) expression [7,8]. Each HRM reaction contained a final concentration of 1x MeltDoctor HRM Master Mix (Applied Biosystems, Foster City, CA, USA), 0.3 µM forward primer (DARCf), 0.3 µM reverse primer (DARCr), 100 pg of template DNA in a final volume of 20 µM (**Appendix 3.1 Table 3**). Reactions were performed using the following thermocycler conditions: denaturation at 95°C for 10 minutes, followed by 45 cycles of 95°C for 15 seconds, 60°C for 1 minute, 95°C for 10 seconds, 60°C for 1 minute, 95°C for 15 seconds, and 60°C for 15 seconds on a QuantStudio 6 Flex Real-Time PCR System (ThermoFisher Scientific, Waltham, MA, USA). Each HRM plate

contained a DARC-positive (-33 C:C), DARC-negative (-33 T:C), and a non-template control which were used to call HRM results on each plate independently.

Samples that could not be definitively determined by HRM and a 10% random subset of *P. vivax* qPCR-positive samples underwent confirmatory Sanger sequencing genotyping at Eton Bioscience (Research Triangle, NC). PCR products were generated from a previously validated assay [9]. Final reactions contained 0.25 µL of FastStart High Fidelity Taq (Enzyme Blend; Roche, Indianapolis, IN), 2.5 µL of 10x FastStart High Fidelity reaction buffer with 18 mM $MgCl_2$, 0.36 µM forward primer, 0.36 µM reverse primer, 250 µM dNTPs and 3 µL of template DNA in a volume of 25 µL. Reactions were amplified using the following thermocycler conditions: denaturation at 94°C for 15 minutes followed by 40 cycles of 94°C for 30 seconds, annealing at 58°C for 30 seconds, extension at 72°C for 90 seconds, and a final extension at 72°C for 10 minute on a BioRad T100 Thermal Cycler (Applied Biosystems, Foster City, CA, USA). PCR products and Sanger sequences were also generated for a DARC-positive control (-33 C:C) and DARC-negative control (-33 T:C).

For each sample, forward and reverse sequences were analyzed using Geneious 10.1.3 (Biomatters Limited, Auckland, New Zealand). First, the 5' and 3' ends of each sequence was trimmed using Geneious `Trim Ends` tool with a 0.05 error probability limit. For each sample, forward and reverse sequences were then *de novo* assembled using the Geneious `Assembler` tool with the sensitivity flag set to "Highest Sensitivity/Slow". Of the 51 randomly samples sequenced, one sample was unable to be assembled due to low sequencing quality. Of the 17 samples that underwent confirmatory sequencing, all samples were assembled. The mapped sequences were then visually assessed for the DARC (-33 T:C) point mutation. Duffy-Genotypes

by Sanger sequencing were concordant with the HRM results among the remaining 50/51 samples selected for validation.

Assay	Primer	Sequence	Ref.
HRM Genotyping	DARCf	5'-CGTGGGGTAAGGCTTCCTGA	7
	DARCr	5'-CTGTGCAGACAGTTCCCCAT	
Confirmatory PCR	ESf	5'-GTGGGGTAAGGCTTCCTGAT	9
	ESr	5'-CAAACAGCAGGGGAAATGAG	

Appendix 3.1 Table 3 - DARC-Genotyping Primers: All samples that were positive by qPCR underwent genotyping at the Duffy Antigen/Chemokine Receptor (DARC) promoter region using High Resolution Melt (HRM) analysis. A subset of randomly selected samples and those samples that could not be absolutely confirmed by HRM underwent confirmatory Sanger sequencing of a GATA-1 transcription factor amplicon that contained the region of interest.

Epidemiological Analyses

Study Population and Data Sources

In the DRC, the DHS aims to create a nationally representative survey using a two-stage stratified cluster sampling design [10]. In the first stage, clusters, or enumeration area, are selected with a known and fixed probability. During the second stage, within each cluster, a subset of households are selected. Finally, among those adults residing in selected households, a subset are consented for HIV and other biomarker testing. To control for this sampling scheme, the DHS weights each individual with an inverse probability weights of selection, hereafter, sampling weights [10].

The Democratic Republic of the Congo (DRC) 2013-2014 DHS-VI survey was conducted from August 2013 - September 2013 and November 2013 - February 2014.

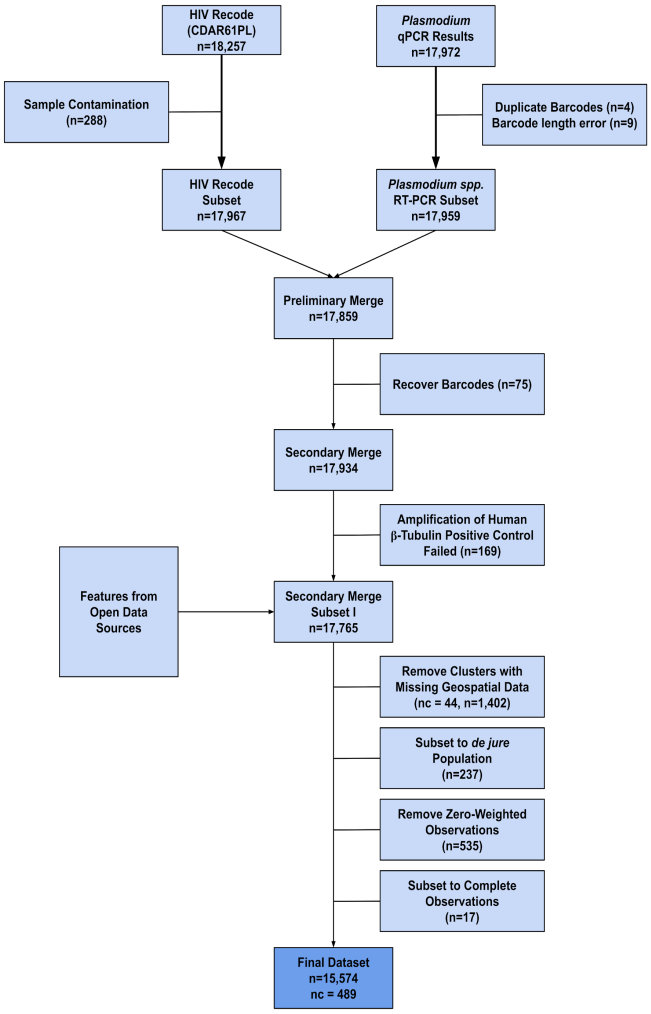
Specifically, DHS surveyors screened Kinshasa and surrounding areas from August 2013 - September 2013 and then subsequently administered the survey across the rest of the country from November 2013 - February 2014.

For each household, DHS surveyors acquired informed consent and administered a substantial questionnaire to all individuals that had slept in the household the night prior to the interview [10]. Individuals that permanently reside in the household are classified as *de jure* while individuals that were coincidentally in the household the night preceding the interview were classified as *de facto* [10]. Given that household variables were considered as potential malaria risk factors, I subsetted observations to the *de jure* population, as *de facto* individuals' homes may differ substantially from the home that they were visiting [10].

Among those adults that agreed to undergo HIV and other biomarker testing, a dried blood spot (DBS) was taken. DBS were then punched into 96-well plates and associated barcodes were manually recorded in a spreadsheet in the DRC. The 96-well plates were then sent to the University of North Carolina-Chapel Hill (UNC) for malaria testing.

In total, 17,959/17,972 samples with properly formatted barcodes were screened by qPCR at UNC. These samples were then linked to the DHS HIV (AR) recode excluding the 288 samples that were contaminated during shipment from the DRC to UNC. On the initial merge, 17,859/17,959 samples were successfully linked. In order to recover more samples, I allowed for a one-character mismatch between the manually recorded DBS barcode and the DHS barcode among those samples that did not have a match in the preliminary merge. Using this strategy, I successfully recovered an additional 75 samples accounting for the total of 17,934/17,959 samples that were screened by qPCR. Among these 17,934 samples, 169 samples failed to amplify human beta-tubulin, which was used as a within-sample positive control, and thus, were

excluded from the study population [1]. Of these 17,765 samples, 1,402 were missing geospatial data (44 clusters), 237 individuals were not *de jure* household members, 535 have sampling-weights set to zero, and 22 had missing risk-factor covariate information and were excluded from the study. As a result, the total study population consisted of 15,571 individuals **Appendix 3.1 Figure 3**).



Appendix 3.1 Figure 2 - Flow Chart of Study Participants that were Included in the Study: Of the 18,257 Demographic Health Survey (DHS) records that had a dried blood spot, 15,574 were included in the final study population. Dried blood spots were lost due to contamination during transport or barcode errors. A small portion (n=169) samples failed to amplify the human beta-tubulin gene used as a positive control. *Abbreviations*: Quantitative polymerase chain reaction - qPCR.

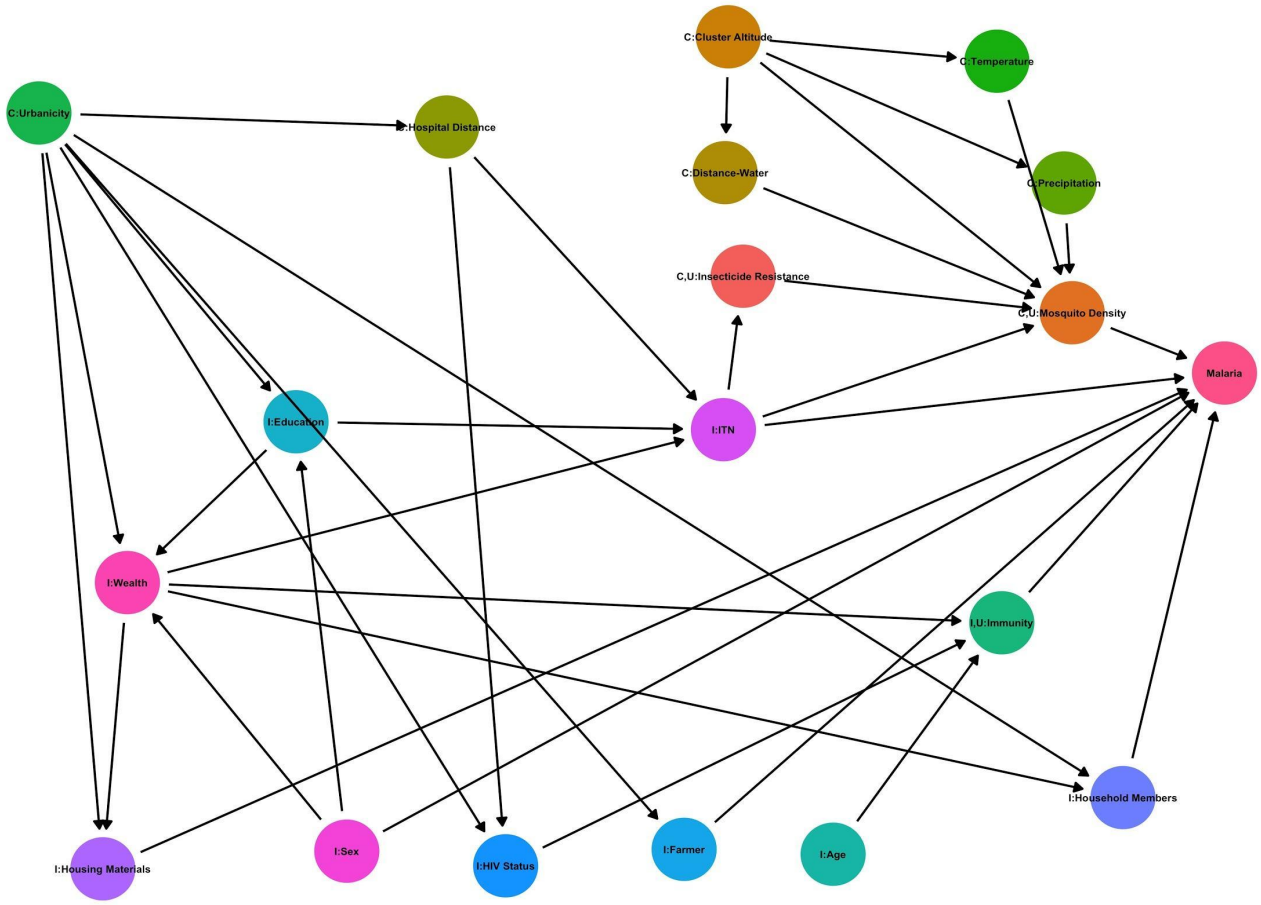
Covariate Feature Engineering

From the DHS questionnaires, I used data from the household members recode (PR), the HIV testing recode (AR), the geospatial covariate (GC) dataset, and the geographical dataset (GE) [10]. Data from the CD2013 was downloaded using the `rDHS` package [11].

In addition to the data provided by the DHS, I downloaded data from several open sources, including: (1) waterways lines and polygon shape-files for the DRC from the Humanitarian OpenStreetMap Team database (https://data.humdata.org/dataset/hotosm_cod_waterways; accessed October 30, 2019); (2) Locations of public hospitals within sub-Saharan Africa (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/JTL9VY; Accessed October 30, 2019) [12]; and (3) non-human ape (NHA) territories from the International Union for Conservation of Nature Red List database (https://www.iucnredlist.org/; accessed January 21, 2019). Temperature data for the 2013-2014 DRC DHS study period was downloaded from the Level-1 and Atmosphere Archive & Distribution System (LAADS) Distributed Active Archive Center (Goddard Space Flight Center, Greenbelt, MA). Specifically, I downloaded monthly layers of land surface temperature and emissivity data from the MYD11C3 (v6) product with a 0.05° x 0.05° spatial resolution (accessed September 20, 2019) [13]. Monthly precipitation data with a 0.05° x 0.05° spatial resolution was downloaded from the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) server using an R-wrapper package (GitHub: `environmentalinformatics-marburg/heavyRain`) for the CD2013 study period [14]. OpenStreetMap extracts from Geofabrik (https://www.geofabrik.de/data/download.html) for Africa (accessed August 23, 2019) were downloaded and used as input into the Open Source Routing Machine(`ORSM`) tool [14]. Finally, additional map features included: (1) ocean spatial

polygons from Natural Earth (naturalearthdata.com); (2) Geographical base-map layers from the Database of Global Administrative Areas (http://www.gadm.org/); (3) Country geographies from the R-package, `rnaturalearth` [15].

Prior to analysis, I identified risk factors for *P. vivax* and *P. falciparum* from a comprehensive literature review [1,16–18]. The relationships among risk factors and my outcome of interest, malaria (i.e. either *P. vivax* or *P. falciparum*) was modeled using a directed acyclic graph (DAG) with the `daggity` graphical user interface and R-package (**Appendix 3.1 Figure 3**) [19]. As a result, not all risk factors identified were measured and included in the analysis. Although anemia and anti-malarial use were considered to be *a priori* risk factors, both were determined to have cyclic relationships with the outcome of interest, malaria, and were excluded (i.e. anemia and antimalarial use could not be resolved by the DAG).



Appendix 3.1 Figure 3 - Malaria Risk Factor Directed Acyclic Diagram: Risk factors were identified from an extensive literature search. Similarly, causal relationship among risk factors were based on previous literature and putative associations. Based on the directed acyclic diagram (DAG), I expected urbanicity, cluster altitude, age, and biological sex to all be unconfounded in expectation (no ancestor nodes).

The majority of risk factors were abstracted from the DHS recodes and kept in their original form with the exception of standardizing continuous variables. Dichotomized variables were set to have an *a priori* protective referent level. Housing type was coded as either "traditional" or "modern" based on a composite score of floor, wall, and roof type as previously outlined by Tustings *et al.* 2017. I also considered any house that had a metal roof as "modern", given recent findings that metal roofs alone appear to be protective against malaria [20].

Given that the DHS wealth variable accounts for housing type in its calculation, I recreated the wealth variable in order to avoid issues of collinearity and non-independence between the wealth and housing covariates [10,18,21]. The wealth covariate was recreated using the factor-score approach based on the instructions by Rustein 2015 and Tustings *et al.* 2017. Wealth factor scores were then considered as a continuous covariate in order to smooth over issues of positivity in wealth and residual confounding.

I defined insecticide treated net (ITN) usage based on the definition outlined in Tustings *et al.* 2017, which limits the ITN classification to long-lasting insecticidal nets less than or equal to three-years-old at the time of the survey, convention ITNs that were less than or equal to one-year-old at the time of the survey, or any net that was retreated within a year of the survey. All other net-usage was coded as "no net" alongside those individuals that reported not using a net the night prior to the survey [10].

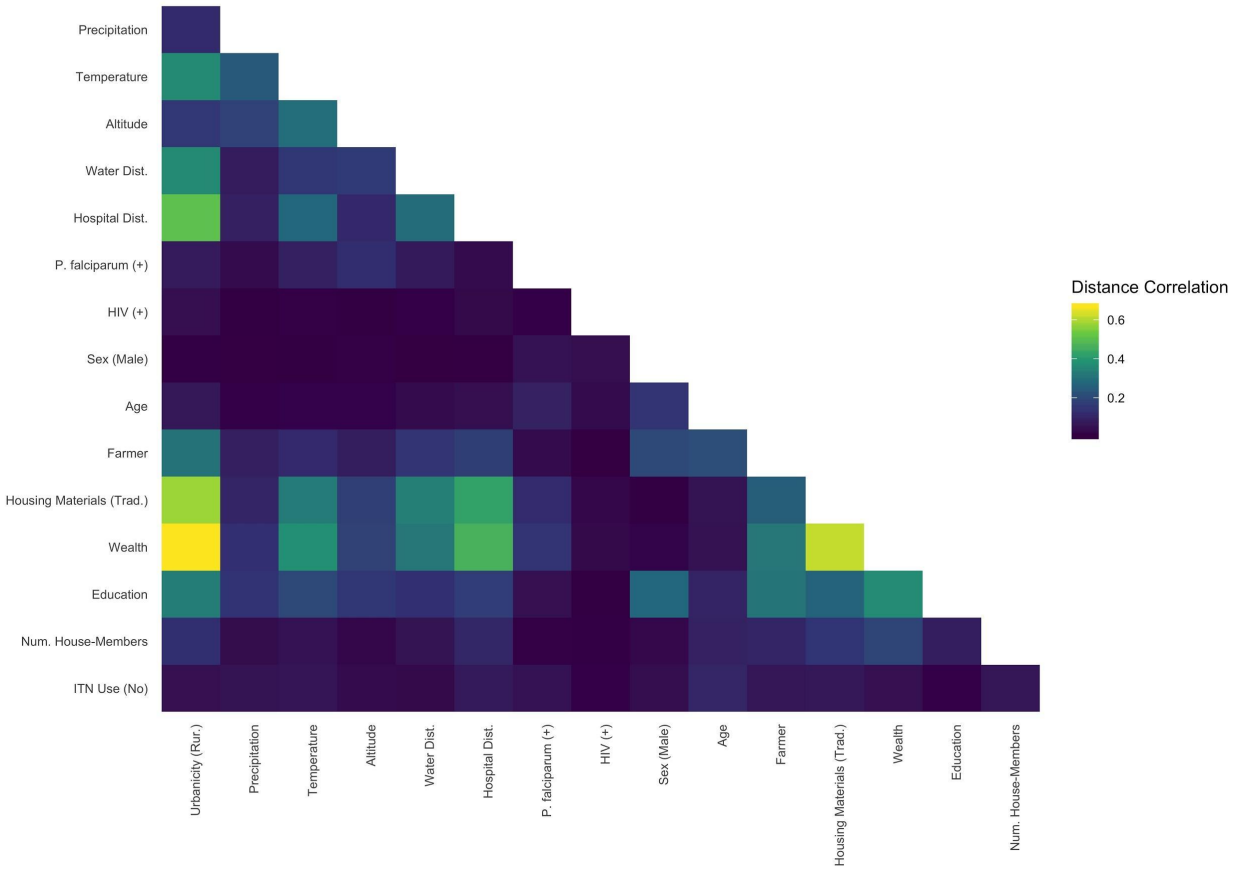
The distance from a hospital covariate was coded as the average duration of travel in minutes between a respective cluster and all public hospitals within the cluster's catchment area. A catchment area was defined as a circle with a 100 km radius with the cluster's location as the centroid. Catchment areas were considered in order to better approximate overall cluster accessibility to health-sites, which may otherwise be biased if a cluster is close to a single hospital but far from all others. If all hospitals were farther than 100 km from a given cluster, the minimum duration between the cluster and all hospitals was considered in place of the catchment area. Travel times were calculated using the Open Street Routing Machine (OSRM) tool [14]. Among the 489 clusters considered, one cluster (469) could not be resolved by OSRM. As a result, the hospital distance for cluster 469 was considered as the average duration among its five nearest-neighbor clusters. Clusters were then coded as "near" or "far" from a public hospitals if

they were within 120 minutes of average travel time or not, respectively [12]. Distance to water was measured as the minimum greater circle distance between a cluster and a body of water that was either labeled as a "river" or "lake" by the OpenStreetMap water-type (Humanitarian OpenStreetMap Team database). Greater circle distances were measured using the R `sf` package [22,23].

Given that the 2013-2014 DRC DHS was conducted in two phases, with the first phase contained to Kinshasa and surrounding areas during months that coincided with the dry-season, while the remaining areas were surveyed during months mostly coinciding with the rainy-season, I elected to take the average monthly temperature and monthly precipitation across the six-months included in the study. Although previous studies have shown that lagging precipitation and temperature can improve predictions of malaria transmission in some cases, I felt that I was unable to lag the weather covariates without introducing spatial confounding [24–28]. As a result, for each cluster in a given study-period month, I first took the average amount of precipitation or daytime temperature among all raster squares within 2 km or 10 km radius of the cluster. The 2 versus 10 km boundary depended on the cluster's designation as urban or rural designation, respectively. This approximates the offsets of geographical coordinates applied by the DHS for each cluster [10,29]. I then aggregated these catchment-area averages for each month into a final study period average. Among the 489 clusters considered, four urban clusters (200, 225, 271, 419) had missing values for temperature and/or precipitation. For these four clusters, the radius was extended to 6km and precipitation and temperature means were calculated as described above.

Correlations among risk factors were evaluated using the Szekely-Rizzo-Bakirov distance correlation with the `energy` R package [30–32]. Based on the covariate-pairwise correlations, I

determined that covariate collinearity was manageable and no covariates needed to be excluded from the analysis (**Appendix 3.1 Figure 4**).



Appendix 3.1 Figure 4 - Covariate Collinearity: The correlation between each pair of covariates were explored for potential bias due to extreme collinearity. Although there were strong correlations that were consistent with *a priori* expectations (e.g. wealth and urbanicity), these correlations did not appear to be completely dependent. As a result, all covariate were kept in the analysis.

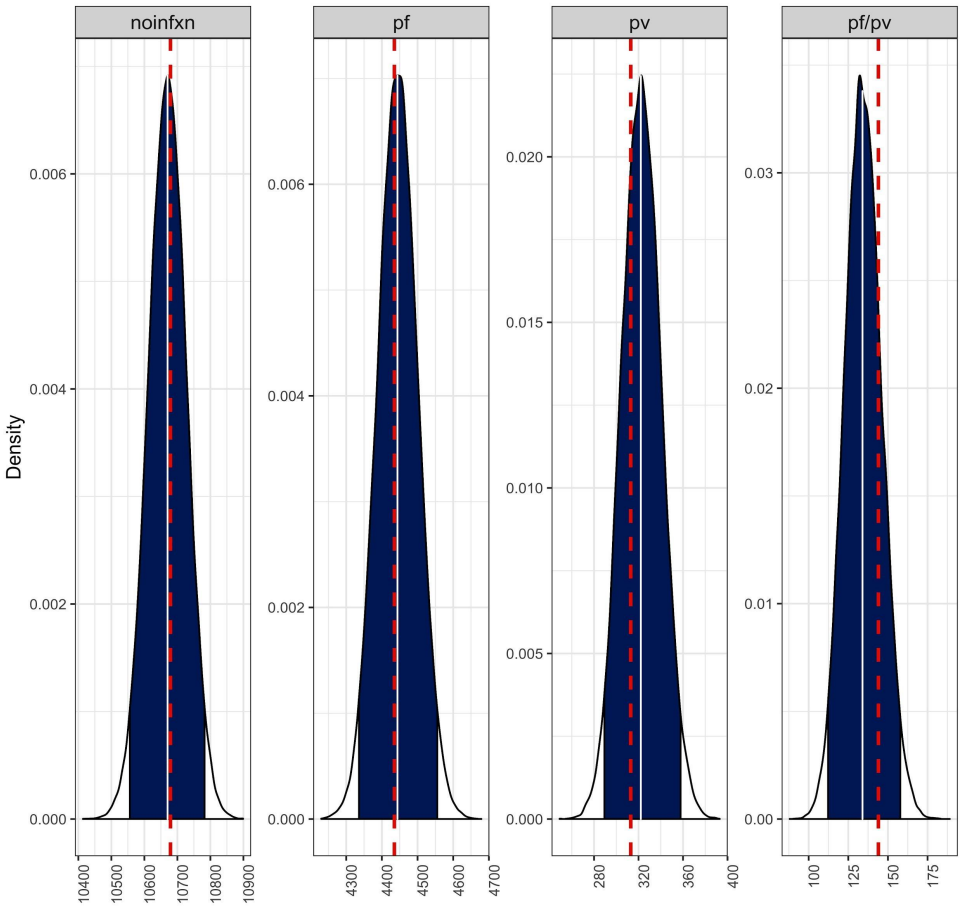
Species Interactions

Interactions between *P. vivax* and *P. falciparum* were examined using an extended version of the independent acquisition of infection model put forth by Akala & Watson *et al.* 2019 to account for individuals that were not infected but still considered in the study population. As in the previous model, I used the observed frequency of each parasite species to fit the expected frequencies of monospecies and cospecies infections using a multinomial likelihood.

An additional category -- uninfected -- was added as a parameter to the multinomial model to account for the case when no successful infectious bites occurred. As a result, the unobserved sequence of species, Y that can be passed to a host is now modeled as:

$$Y = \begin{cases} y_1, y_2, , y_k \in S, & \text{if } k > 0 \\ 0, & \text{if } k = 0 \end{cases}$$

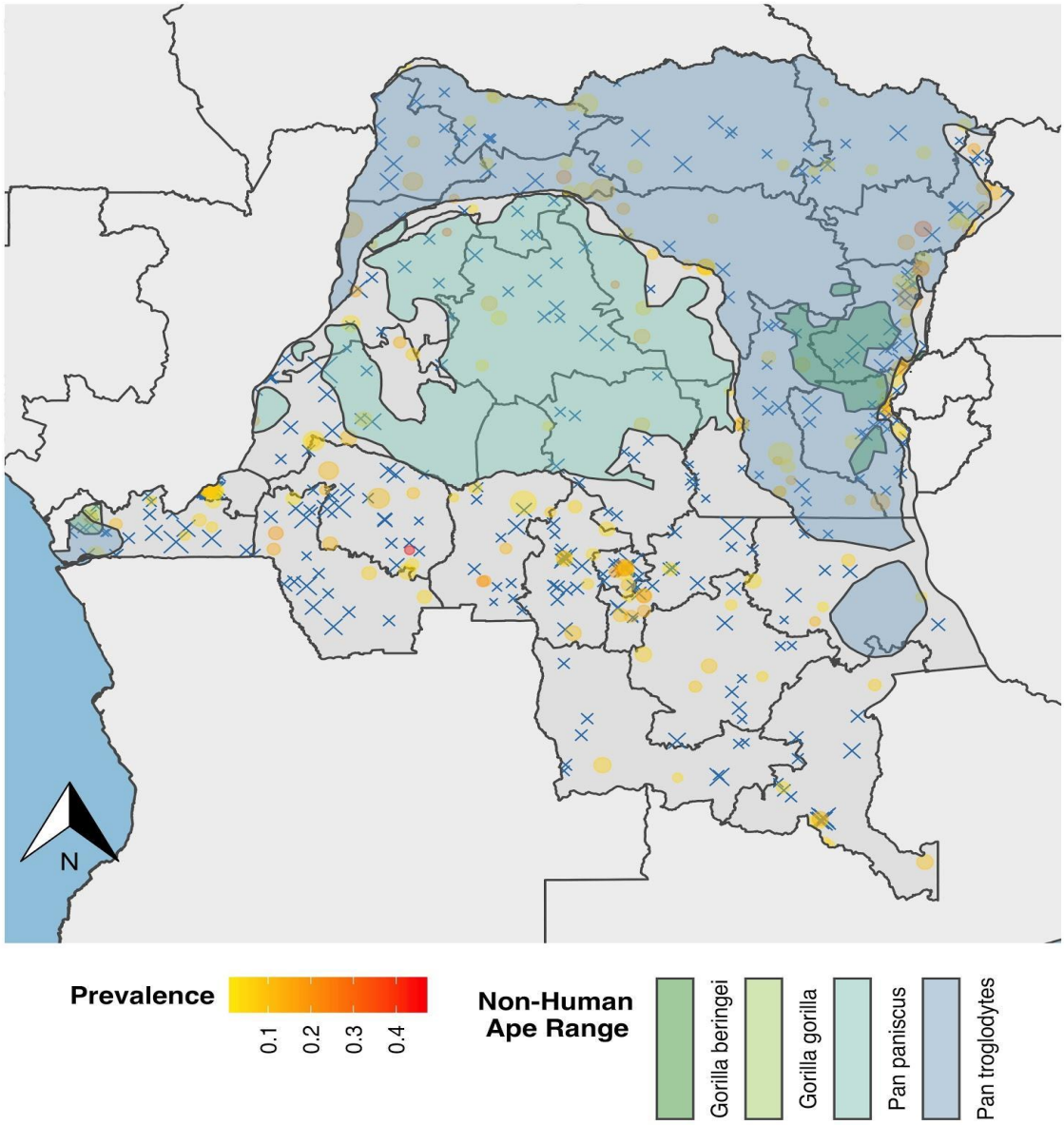
Where S was previously defined as the set of *Plasmodium* species of interest and k as the number of infectious bites a host received [33]. Otherwise, the model was unchanged. For the *P. vivax*-*P. falciparum* model, I considered μ as a Poisson distribution and drew 50,000 bootstrap iterations to form the expected infection compositions. Expected infection compositions were then compared against the observed mono- and co-infection data. Overall, mono-infection and co-infection compositions were consistent with the expectation of independent acquisition of parasites, as the observed data fell within the simulated data (**Appendix 3.1 Figure 5**).



Appendix 3.1 Figure 5 - Composition of *P. vivax* and *P. falciparum* Co-infections: The expected versus observed composition of *P. vivax* and *P. falciparum* infections were explored using a multinomial probability likelihood model. The plot shows the expected distribution for individuals without infection ("noinfxn"), *P. falciparum* infections ("pf"), *P. vivax* infections ("pv"), and *P. falciparum*- *P. falciparum* coinfections ("pf/pv"). The blue shading indicates the 95% bootstrapped interval and the red-dotted line indicates the observed number of cases for each infection category.

Interactions between NHA territories and *P. vivax* prevalence were assessed using a permutation test with 10,000 iterations. Null distributions for the permutation test were calculated by drawing n_{ape} clusters at random, where n_{ape} was the number of 2013-2014 DRC DHS clusters that overlapped with NHA territories. I then calculated the prevalence of *P. vivax* infections among the selected clusters. I considered NHA territories for (1) *Pan troglodytes* and *Gorilla sp.* and (2) *Pan troglodytes*, *Pan paniscus*, and *Gorilla sp.*, separately, as *P. paniscus*

(bonobos) have only recently been shown to harbor *P. vivax*-like parasites at a single field-site (TL2) [34]. In contrast, *Pan troglodytes* (chimpanzees) and *Gorilla sp.* have previously been shown to harbor *P. vivax*-like parasites at various prevalences across the DRC [35]. From the permutation tests, NHA territories and *P. vivax* prevalence were not associated ($p > 0.05$). This lack of an association is also evident when visualizing a map of NHA territories and cluster level *P.* vivax prevalences (**Appendix 3.1 Figure 6**).



Appendix 3.1 Figure 6 - *P. vivax* and Non-Human Ape Distributions: Based on visual inspection, *P. vivax* prevalence did not appear to be associated with non-human ape habitat distribution. This lack of a P. vivax - non-human ape association was recapitulated with permutation testing. Clusters with *P. vivax* infections are shaded on a yellow-red spectrum with respect to the cluster-level prevalence. Clusters without *P. vivax* infections are indicated by blue X-ticks. Finally, the distribution of each non-human ape habitat is indicated in shades of green for the *Gorilla* genus and blue for the *Pan* genus.

Inverse Probability Weights and Prevalence Odds Ratios

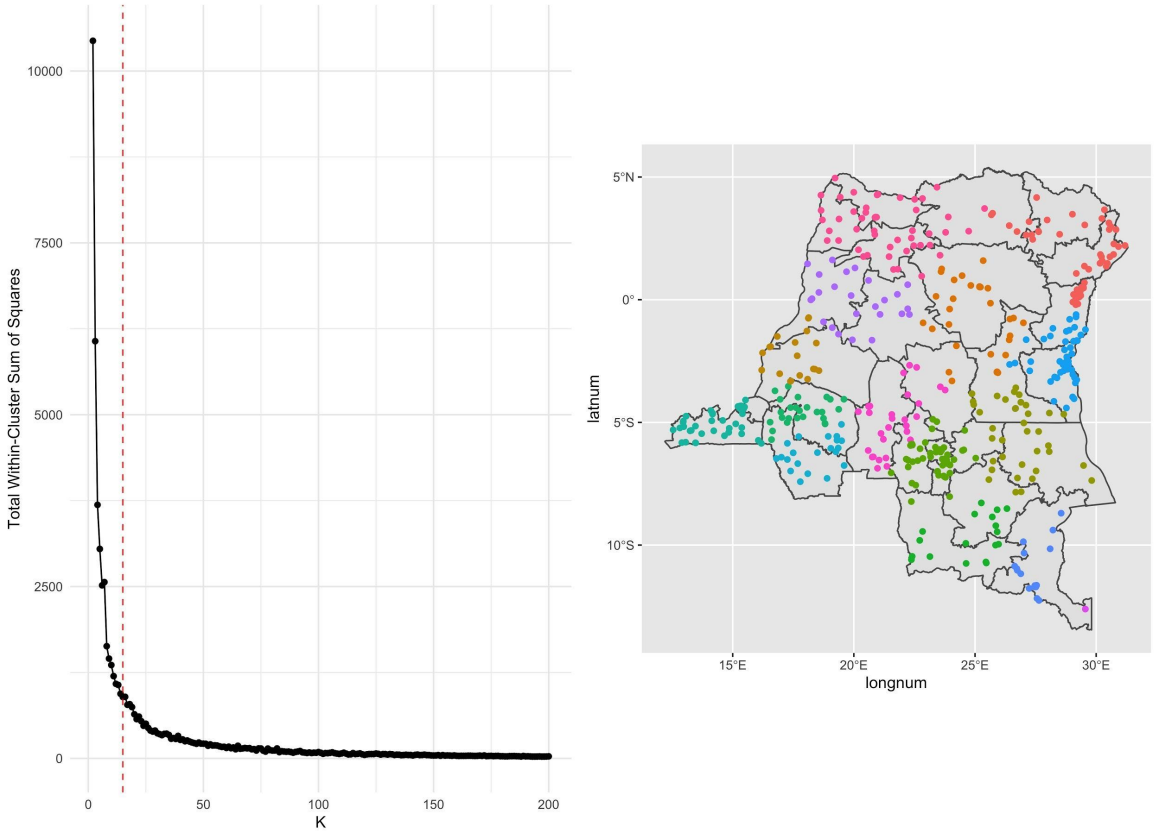
The average effect of each risk-factor, A, on the binary outcome of interest Y (i.e. malaria infection), was estimated using marginal structural models (MSMs): $g(P(Y|A=a)) = \beta_0 + \beta_{1a}$, where $g(.)$ is a logit link for our prevalence odds ratio effect estimates [36–39]. For each MSM, I adjusted for confounders, L, using inverse probability weights (IPWs) [36–39]. IPWs were modeled as $w_i = \frac{1}{f_{A|L}(A|L)}$ for each individual, i in the study population, N. Each weight was stabilized by the marginal mean of the risk factor, such that final weights were: $w_i = \frac{f(A)}{f_{A|L}(A|L)} \text{for} i \in N$. In the case of a binary risk factor, $f_{A|L}(A|L)$ was a probability mass function with each level of A representing the predictive probability of receiving a risk factor given a sequence of confounders. Similar, in the case of a continuous treatment, $f_{A|L}(A|L)$ was a probability density function with each level of A representing the predictive probability of receiving a dose of the risk factor given a sequence of confounders. In the continuous setting, I assumed that $f(A)$ and $f_{A|L}(A|L)$ followed normal distributions and could be estimated with a standard normal density [36,38,40].

IPWs were calculated using the super learner algorithm with spatial cross-validation [41–44]. I used a diverse set of candidate algorithms, as the super learner is expected to asymptomatically outperform any individual candidate algorithm as the number of candidate algorithms becomes polynomial in sample size (**Appendix 3.1 Table 4**) [41–43]. In some cases, if IPWs appeared to be unstable, I subsetted the candidate algorithm library to either logistic or linear regression, depending on the outcome type (**Appendix 3.1 Table 5**).

Base Learner	R-package, Function	Relevant Hyperparameters	Justification
Generalized Linear Regression*	stat, lm/glm [45]	-	-
Cross-Validated L1/L2 Regularized Regression (x3)	glmnet, cvglmnet [46]	α: 1 α: 0.5 α: 0	Shrinkage of covariates based on fit
Boosted Generalized Additive Modeling	mboost, gamboost [47]	-	Non-linearity in covariates
K-Nearest Neighbor	kknn, kknn [48]	k: 7 Kernel: optimal	Interactions, Non-linearity in Covariates
Single Vector Machines	e1071, svm [49]	Cost: 1 Kernel: radial	Interactions, Non-linearity in Covariates
Neural Net	nnet, nnet [50]	Hidden Layers: 1 Units in Hidden Layer: 3	Interactions, Non-linearity in Covariates
Random Forest	ranger, ranger [51]	Number of Trees: 500 Variables at Node split: $\sqrt{p}$	Interactions, Non-linearity in Covariates

Appendix 3.1 Table 4 - Base Learners used in the Super Learner Algorithm: Various base learners were inputted into the super learner algorithm. The super learner algorithm is an ensemble based method that optimizes the predictions of base learners using a loss-based approach that minimizes the prediction error. A diverse suite of base learners was selected to account for various non-linear effects as well as interactions among covariates.

Folds for cross-validation were based on K-means clustering of geographical coordinates to account for potential spatial autocorrelation among observations [44]. I selected a K of 15, as it was the inflection point that appeared to minimize the within-cluster sum of squares while avoiding overfitting (**Appendix 3.1 Figure 7**).

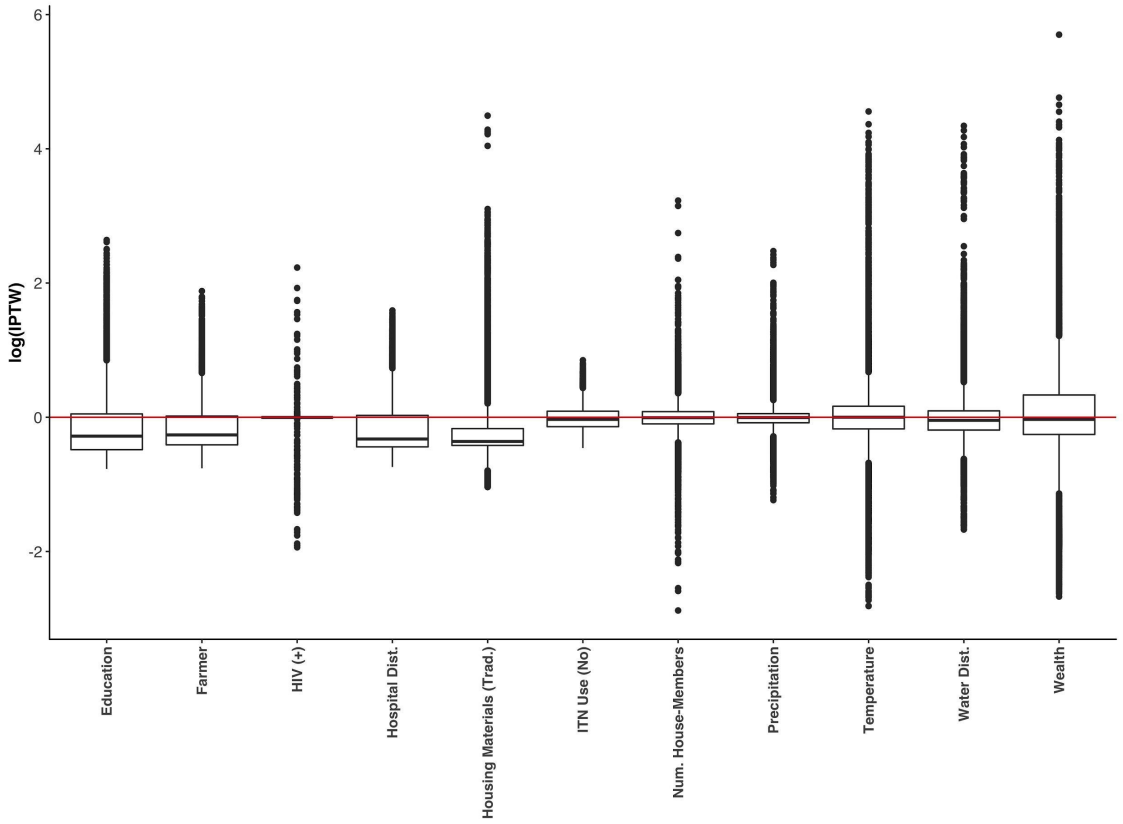


Appendix 3.1 Figure 7 - Spatial Cross-Validation K-Clusters: The DRC was partitioned into K-clusters for spatial cross-validation. Based on the geographical K-means total within-cluster sum-of-squares, 15 clusters appeared to be a reasonable inflection point that did not overfit the data but still captured natural geographic partitions in the DRC (left). The fifteen partitions are mapped to show the geographical partition (right).

All machine-learning models were built and analyzed using the `mlr` package, which provides a machine-learning infrastructure within the R-environment [52]. The super learner algorithm was selected for IPW calculations to account for issues of functional form and non-linearity that can bias predictions [53] . I assumed that a single iteration of the super learner algorithm was adequate to predict the IPWs. For each risk factor, I considered all descendants and ancestors of the risk factor and the outcome that were not on the causal pathway as predictors in the IPW-model to account for any "backdoor" paths not considered in the DAG, (i.e. the IPW adjustment set).[36] For risk-factors that were unconfounded in expectation (i.e.

biological sex, age, urbanicity, and altitude), no adjustment set was considered. Weights were incorporated with the R `survey` package and base R functions [54].

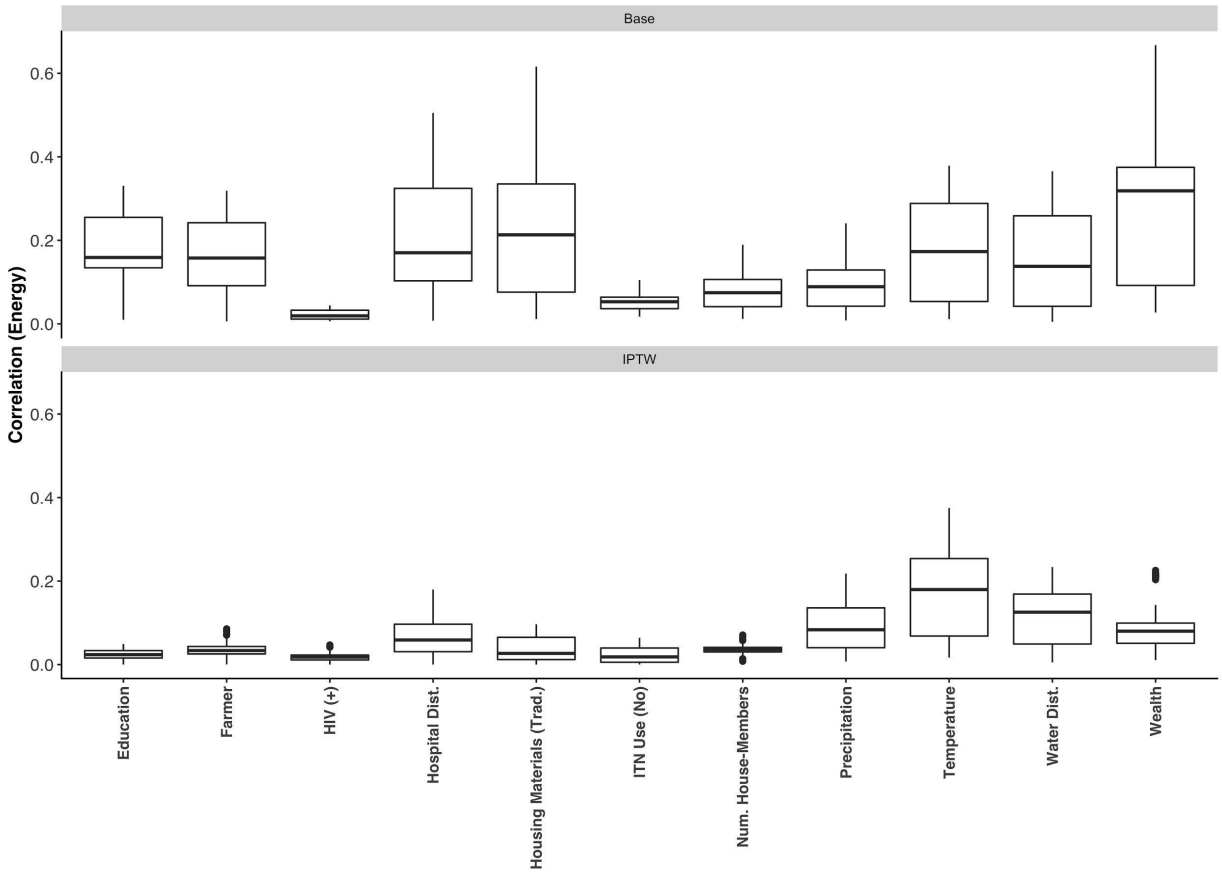
Overall stability of the IPWs were assessed visually and were determined to have log-transformed standard normal distribution (**Appendix 3.1 Figure 8**). IPW distributions that are not definitively centered may suffer from lingering issues of structural positivity or may be correctly identifying multimodal distributions in risk-factor distributions.



Appendix 3.1 Figure 8 - Distribution of Inverse Probability Weights: For each covariate, the distribution of weights for the 15,571 individuals included in the study are shown. Distributions have been log-transformed and appear to be approximately normally distributed.

The effects of the IPW on baseline risk-factor associations (i.e. putative confounding) were assessed using Szekely-Rizzo-Bakirov distance correlations for each risk-factor pair [30–32]. Given that a weight option is not specified in the Szekely-Rizzo-Bakirov distance correlation

calculation, I applied my IPWs by sampling observations according to their IPWs. To account for variability in sampling, I created 100 IPW-pseudopopulations for each risk-factor pair. The distribution of pairwise distance correlations for the risk factors was then plotted and compared with no weights applied and with IPWs applied (**Appendix 3.1 Figure 9**).



Appendix 3.1 Figure 9 - Correlation among Covariates at Baseline and After Application of Inverse Probability Weights: A classic measure of confounding is baseline correlations among covariates, or the unequal distribution of covariates among different treatment classes. Shown for each covariate are the measures of pairwise covariate correlation at baseline (top) and after inverse probability weights (IPWs) have been considered (bottom). Baseline covariates show a large degree of correlation -- potentially indicating confounding -- while, for the most part, covariates with IPWs applied show a considerable reduction in pairwise covariate correlations (mean fold-reduction: 3.14, range: 0.85 - 7.63). Interesting, temperature appeared to still have somewhat high pairwise correlations even after applying IPWs. *Abbreviations*: Hospital Dist. – Distance to hospital, Trad. – traditional, ITN – insecticide treated net, Num. – number, Water Dist – Distance to water.

Covariate	Cross-Validated Risk Coefficient	Base Learner
Precipitation	1	regr
Temperature	0.16	Simple Linear Regression
Temperature	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Temperature	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Temperature	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Temperature	0.19	Support Vector Machines (libsvm)
Temperature	0.12	K-Nearest-Neighbor regression
Temperature	0.51	Gradient Boosting with Smooth Components
Temperature	0.03	Neural Network
Temperature	0	Random Forests
Water Dist.	0.10	Simple Linear Regression
Water Dist.	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Water Dist.	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Water Dist.	0.31	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Water Dist.	0	Support Vector Machines (libsvm)
Water Dist.	0	K-Nearest-Neighbor regression
Water Dist.	0	Gradient Boosting with Smooth Components
Water Dist.	0.20	Neural Network
Water Dist.	0.39	Random Forests

HIV (+)	1	regr
Farmer	0	Logistic Regression
Farmer	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Farmer	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Farmer	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Farmer	0.56	Gradient boosting with smooth components
Farmer	0.23	Support Vector Machines (libsvm)
Farmer	0.01	k-Nearest Neighbor
Farmer	0.18	Neural Network
Farmer	0.02	Random Forests
Wealth	0	Simple Linear Regression
Wealth	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Wealth	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Wealth	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Wealth	0.02	Support Vector Machines (libsvm)
Wealth	0.41	K-Nearest-Neighbor regression
Wealth	0.57	Gradient Boosting with Smooth Components
Wealth	0	Neural Network
Wealth	0	Random Forests
Education	1	regr
Housing Materials (Trad.)	1	regr

ITN Use (No)	1	regr
Hospital Dist.	0	Logistic Regression
Hospital Dist.	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Hospital Dist.	0	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Hospital Dist.	0.98	GLM with Lasso or Elasticnet Regularization (Cross Validated Lambda)
Hospital Dist.	0	Gradient boosting with smooth components
Hospital Dist.	0	Support Vector Machines (libsvm)
Hospital Dist.	0.02	k-Nearest Neighbor
Hospital Dist.	0	Neural Network
Hospital Dist.	0	Random Forests

Appendix 3.1 Table 5 - Cross-Validated Risk and Contribution of Base Learners for each Covariate: Given that the Super Learner algorithm optimizes the contribution of individual base learners, not all base learners are included in the final predictions for each covariate. In some instances, Super Learner predictions result in unstable weights. As a result, I culled the base learner library to either a linear or logistic regression algorithm for continuous and dichotomous covariates, respectively (indicated by a 1 in the Cross-Validated Risk Coefficient column and "regr" in the Base Learner column). *Abbreviations*: Hospital Dist. – Distance to hospital, Trad. – traditional, ITN – insecticide treated net, Num. – number, Water Dist – Distance to water.

Risk Factor	Species	IPTW-pOR	IPTW-pOR, L95	IPTW-pOR, U95	pOR	pOR, L95	pOR, U95
Age	Pv	0.97	0.87	1.07	0.97	0.87	1.07
Altitude	Pv	1.13	0.88	1.45	1.13	0.89	1.44
Education (Lower)	Pv	0.91	0.64	1.3	0.99	0.74	1.34
Farmer	Pv	1.42	1.08	1.88	1.32	1	1.75
HIV (+)	Pv	0.93	0.33	2.67	1.86	0.76	4.54
Hospital Dist.	Pv	0.86	0.53	1.4	0.86	0.53	1.38
Housing Materials (Trad.)	Pv	1.12	0.62	2.04	1	0.64	1.57
ITN Use (No)	Pv	0.76	0.55	1.04	0.8	0.58	1.09
Precipitation	Pv	0.79	0.63	0.99	0.78	0.63	0.97
Sex (Male)	Pv	1.17	0.89	1.53	1.17	0.89	1.54
Temperature	Pv	0.83	0.62	1.11	0.78	0.62	0.97
Urbanicity (Rur.)	Pv	1.13	0.7	1.83	1.13	0.7	1.82
Water Dist.	Pv	1.19	0.93	1.52	0.97	0.79	1.19
Wealth	Pv	1.12	0.78	1.59	0.93	0.81	1.07
Age	Pf	0.81	0.77	0.86	0.81	0.77	0.86
Altitude	Pf	0.73	0.65	0.82	0.73	0.66	0.8
Education (Lower)	Pf	1.44	1.25	1.67	1.18	1.02	1.35
Farmer	Pf	1.03	0.9	1.18	1.08	0.94	1.24
HIV (+)	Pf	0.54	0.18	1.58	0.5	0.26	0.93
Hospital Dist.	Pf	1.15	0.89	1.48	1.37	1.1	1.7
Housing Materials (Trad.)	Pf	1.25	0.98	1.61	1.84	1.54	2.19
ITN Use (No)	Pf	1.23	1.07	1.42	1.27	1.11	1.45
Precipitation	Pf	0.96	0.83	1.12	0.99	0.87	1.12
Sex (Male)	Pf	1.31	1.2	1.43	1.31	1.2	1.43
Temperature	Pf	1.41	1.05	1.9	1.07	0.97	1.19
Urbanicity (Rur.)	Pf	0.7	0.54	0.89	0.7	0.56	0.86
Water Dist.	Pf	0.87	0.77	0.99	1.12	0.99	1.28
Wealth	Pf	0.82	0.73	0.92	0.75	0.69	0.81

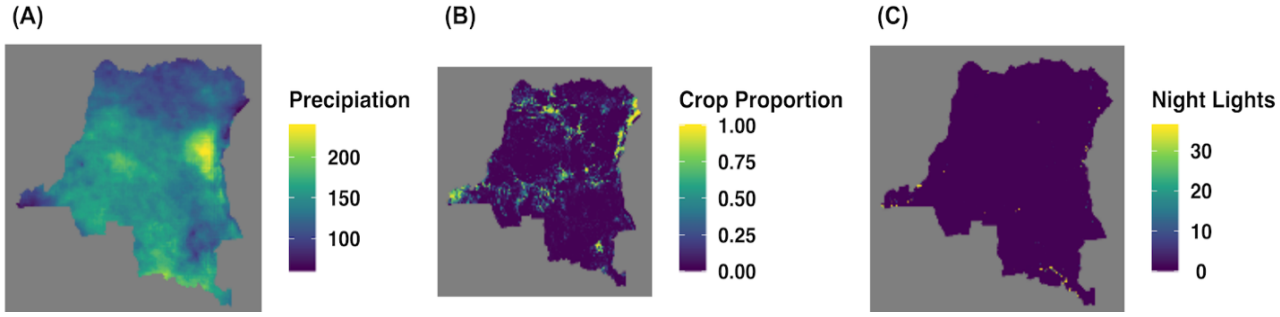
Appendix 3.1 Table 6 - Inverse Probability Weight (IPW) Adjusted and Unadjusted Prevalence Odds Ratios for the Malaria Risk Factors: Inverse probability weight (IPW) adjusted and unadjusted prevalences odd ratios (pOR) risk factor effect estimates for *P. vivax* (Pv) and *P. falciparum* (Pf) are provided with corresponding 95% confidence intervals. IPW adjustments were performed using the super learner algorithm. Unadjusted estimates are modeled using generalized estimating equations with a logit-link and binomial variance accounting for the DHS sample-weights. These bivariate association models are essentially two-by-two tables weighted for the 2013-2014 Demographic Health Survey in the Democratic Republic of the Congo sampling scheme. In instances where the adjusted and unadjusted estimates are the same (age, biological sex, urbanicity, and altitude), the risk factor was expected to be unconfounded at baseline and IPWs were not considered (**Appendix 3.1Figure 3**). *Abbreviations*: Hospital Dist. – Distance to hospital, Trad. – traditional, ITN – insecticide treated net, Rur. - rural, Num. – number, Water Dist – Distance to water.

Spatial and Raster Feature Engineering

In order to incorporate the risk factor covariate information into the spatial models, I downloaded spatial raster data for significant risk factors identified by the MSMs. The precipitation raster was used from above, with the surface consisting of mean values over the study period. To account for the risk factor associated with farming, I downloaded a raster of light intensity and land coverage for the DRC. Specifically, I used the 2015 annual night-light composite vcm-orm-ntl version raster (https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html, accessed Nov 8, 2019), which provides an average night-light intensity for each point in the DRC at a 15 arcsecond resolution [55,56]. In addition, the vcm-orm-ntl version has been pre-processed to exclude outliers and spurious measurements due to fires or cloud coverage [55,56]. The 2015 annual night-light composite raster was selected as rasters for 2013 and 2014 were not available. Land coverage in the DRC was accessed through the Land Cover Climate Change Initiative (CCI) Climate Research Data Package from the European Space Agency Climate Change Initiative, which provides yearly land coverage maps at 300 x 300 meter resolution for 1992-2015 (https://maps.elie.ucl.ac.be/CCI/, accessed Nov 8, 2019). Specifically, I used the 2013 land coverage raster and reclassified raster points as a binary of cropland or not-cropland based on the CCI classifications (values 10, 20, 30, 40, Yes; all others, No; **Appendix 3.1 Figure 10**).

From these raster surfaces, I then aggregated raster points to fit within the DHS cluster design and DHS province boundaries. Specifically, for each cluster and covariate of interest, I took the mean value from all raster squares within a 2 km or 10 km radius with respect to the cluster urban/rural designation [10,29]. For each province, all raster cells within the province boundary were aggregated and summarized as a mean value.

As described above, precipitation values were standardized. Similarly, cropland proportion was transformed onto the real-line using a logit-transformation and was then standardized. Given that most points in the DRC had no measured light-intensity throughout the year, night-light standardization was performed under a zero-truncated framework (i.e. standardization did not include zeros). Standardization was performed in favor of model stability.



Appendix 3.1 Figure 10 - Spatial Raster Covariates: Spatial covariates that were associated with *P. vivax infection* by the risk factor analysis were included in the spatial prediction prevalence models, and included: precipitation (A) and farming (B, C). Farming was captured through the proportion of crops (B) at each raster cell as well as with the night-light intensity (C) in a raster cell across the DRC.

Bayesian Mixed Spatial Prediction Models

Prevalence maps were fit as mixed generalized linear models with spatially correlated random effects in a Bayesian framework. I modeled prevalence at two different levels: (1) Province-level using the `CARBayes` R-package and (2) Cluster-level using the `PrevMap` R-package [57,58]. DHS sampling weights were accounted for by rounding the number of cases, Y_i, to the nearest whole individual in order to conform with the binomial error distribution of the model. For the province-level models, there are i total survey regions, such that $i \in \geq \mathbb{Z}_1$, and survey regions are defined as non-overlapping areal units with defined boundaries: $S = S_1, \ldots, S_i$. For the cluster model, the survey region is the DHS second-level enumeration

area, which is a collection of households aggregated at a single set of GPS coordinates (i.e. clusters) [10]. In total, there are there are j total clusters, where $j \in \geq \mathbb{Z}_1$ and clusters are indexed as: $C = C_1, \ldots, C_j$. Risk factors that were identified as significant were included as linear predictors, β_i . As a result, the model can be specified as:

$$Y_i|n_i, p_i \sim \text{Binomial}(n_i, p_i)$$
$$logit(p_i) \sim \beta X_i^\intercal + S(a_i) + Z_{a_i}$$

Following Lee 2017, for the province-level model, the spatial ($S(a_i)$) and non-spatial (Z_{a_i}) random effects were modeled using a random effect, ϕ with the conditional-autoregressive prior proposed in Leroux *et al.* 2000 (hereafter referred to as the Leroux CAR model). Specifically,

$$\beta \sim \text{MultivariateNormal}(\mu_\beta,\ \Sigma_\beta)$$
$$\phi_k|\phi_{-k}, W, \tau^2, \rho \sim \text{Normal}(\frac{\rho \sum_{i=1}^{K} w_{ki}\phi_i}{\rho \sum_{i=1}^{K} w_{ki} + 1 - \rho},\ \frac{\tau^2}{\rho \sum_{i=1}^{K} w_{ki} + 1 - \rho})$$
$$\tau^2 \sim \text{InverseGamma}(1,\ 0.01)$$
$$\rho \sim Uniform(0,\ 1)$$

The adjacency matrix, W, was a simple neighborhood matrix, where border sharing was indicated as a binary [59]. Models with the ρ parameter fixed at one assume complete spatial autocorrelation among the random effects (i.e. the Intrinsic CAR or Besag model), while models the ρ parameter fixed at zero assume independence [60–62]. By allowing ρ to vary under the model, as specified above, I can fit this spatial autocorrelation process [57,60]. Finally, I set the multivariate Gaussian mean prior μ_β as a vector of zeros and the diagonal elements of the covariance matrix, σ_β, to 50,000 [57].

For the cluster model, the survey region is the DHS second-level enumeration area, which is a collection of households aggregated at a single set of GPS coordinates (i.e. clusters) [10]. In total, there are there are j total clusters, where $j \in \geq \mathbb{Z}_1$ and clusters (i.e. sampling locations) are indexed as: $C = c_1, \ldots, c_j$. As a result, the model was specified as:

$$logit(p_i) \sim \beta X_i^{\intercal} + S(c_i) + Z_{c_i}$$

Following the model presented in Giorgi and Diggle 2017, the spatial random effect, S_i, was modeled as a stationary isotropic Gaussian process with variance σ^2 and a Matérn covariance function, $\rho(d; \phi, \kappa)$. Here, d is the distance between any two clusters, c_i, $c_{\bar{i}}$. Based on an exploratory analysis of the κ that maximized the log-likelihood of our logit-transformed prevalence data, we fixed κ at 1. The remainder of the model was specified using diffuse priors:

$$\beta|\sigma^2 \sim \text{Normal}(0, \sigma^2)$$
$$log(\tau^2) \sim \text{Normal}(0,\ 25)$$
$$\phi \sim \text{Uniform}(0,\ 50)$$
$$log(\sigma^2) \sim \text{Normal}(0,\ 50)$$

Each model was first evaluated with four diagnostic chains using 1,000 burn-in iterations and 10,000 sample iterations. Chains were then visually assessed for convergence and appropriate mixing patterns. A final long chain with 10,000 burn-in iterations and 100,000 sampling iterations was then considered for each model. Chains were again visually assessed and all parameters were required to have an effective sample size of at least 500.

Predictions were performed using the fitted values for the province level. Similarly, for the cluster level, predictions were made out-of-sample using the fitted covariates under the assumption of a multivariate Gaussian distribution as previously described in Giorgi and Diggle 2017. Covariate observations for predictions were taken from the precipitation, crop-proportion, and night light intensity rasters described above. For the crop-proportion and night light intensity raster, I aggregated the rasters to a 0.05° x 0.05° resolution by taking the mean and sum of raster cells, respectively (a 0.05° x 0.05° resolution was selected as this was the least precise spatial resolution among the covariates). For each of the prediction sampling locations, the covariate matrix was calculated by taking the mean value for each raster cell within a six km radius (mean of DHS maximum offset) [10,29]. Given that I was performing interpolation, any value in the covariate prediction matrix that exceeded the observed maximum in the fitted covariate matrix was truncated (i.e. the observed maximum for each covariate served as an upper bound among the predictions to avoid extrapolation).

Predictions were then calculated for each of the 100,000 sampling iterations. For the sake of computational burden, I subsetted the approximately 160,000 potential prediction sampling locations in the DRC that would need to be estimated at 100,000 sampling iterations (16 billion estimates) to 20,000 randomly selected sampling locations. Local interpolation was performed using inverse distance weighting and an inverse power parameter of two with the R `gstat` package [63,64].

Level	Model	Parameter	Mean	Median	2.5% CI	97.5% CI	Effective N	DICg
Province	Intercept	Intercept	-3.59	-3.58	-3.72	-3.45	27,459	-57.4
		τ^2	0.02	0.46	0.20	1.13	13,731	
		ρ	0.22	0.31	0.02	0.87	11,257	
	Covariate	Intercept	-0.23	-3.59	-3.71	-3.47	9,920	-52.1
		Precip.	0.45	0.02	-0.28	0.35	808	
		Crop Prop.	0.29	0.21	-0.07	0.52	1,232	
		Night Light	-3.58	-0.23	-0.48	0.02	1,960	
		τ^2	0.52	0.39	0.15	1.07	4,847	
		ρ	0.35	0.24	0.01	0.83	8,447	
Cluster	Intercept	Intercept	-2.66	-2.64	-6.49	1.20	92,453	1331975.7
		σ^2	7.57	5.30	0.78	20.60	4,543	
		ϕ	37.21	38.91	19.55	50.00	18,353	
		τ^2	3.16	3.11	2.24	4.21	1,844	
	Covariate	Intercept	-2.66	-2.64	-5.37	-0.11	71,797	157849.2
		Precip.	-0.04	-0.04	-0.32	0.24	19,276	
		Crop Prop.	0.06	0.06	-0.22	0.34	18,305	
		Night Light	-0.08	-0.08	-0.55	0.40	19,862	
		σ^2	3.59	2.78	0.64	8.80	6,757	
		ϕ	35.44	36.98	16.82	50.00	13,206	
		τ^2	3.22	3.18	2.23	4.25	1,949	

Appendix 3.1 Table 7 - Spatial Model Parameter Estimates and Fits: The mean, median, and 95% credible interval (CI) summary statistics are provided for each parameter with respect to the models evaluated. The fit of each model was calculated using Gelman's deviance information criteria and compared at the province-level and cluster-level, respectively. Overall, the best fitting province-level and cluster-level models included a precipitation, crop, and night light intensity covariate. For reference, the posterior ϕ values for each province are also provided (**Appendix 3.1 Table 8**).

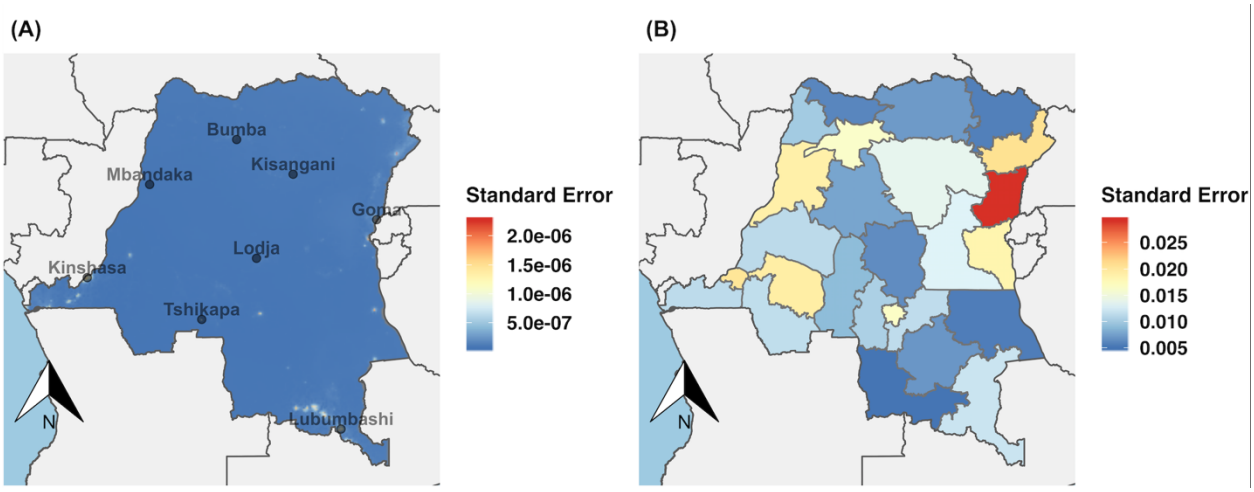
Model	Province	Mean	Median	2.5% CI	97.5% CI	Effective N
Intercept	Bas-Uele	-0.27	-0.26	-0.92	0.30	51,421
	Equateur	0.92	0.91	0.51	1.35	5,159
	Haut-Katanga	-0.32	-0.32	-0.80	0.14	20,460
	Haut-Lomami	-0.24	-0.23	-0.84	0.31	60,477
	Haut-Uele	-0.20	-0.19	-0.96	0.45	51,822
	Ituri	1.03	1.03	0.59	1.47	2,529
	Kasai	-0.06	-0.05	-0.56	0.42	58,972
	Kasai-Central	-0.36	-0.35	-0.85	0.10	46,118
	Kasai-Oriental	0.30	0.30	-0.15	0.74	6,079
	Kinshasa	-0.81	-0.80	-1.24	-0.40	11,908
	Kongo-Central	-0.69	-0.68	-1.26	-0.18	25,429
	Kwango	-0.38	-0.38	-0.85	0.06	32,753
	Kwilu	0.15	0.15	-0.26	0.58	5,662
	Lomami	-0.01	0.00	-0.44	0.40	43,101

	Lualaba	-0.50	-0.48	-1.22	0.13	58,352
	Mai-Ndombe	-0.26	-0.26	-0.73	0.17	37,980
	Maniema	0.35	0.35	-0.07	0.76	24,416
	Mongala	0.83	0.84	0.42	1.25	15,558
	Nord-Kivu	0.45	0.47	-0.02	0.82	1,445
	Nord-Ubangi	-0.18	-0.17	-0.89	0.45	61,027
	Sankuru	-0.13	-0.12	-0.72	0.40	62,775
	Sud-Kivu	-0.07	-0.06	-0.52	0.37	3,666
	Sud-Ubangi	0.09	0.09	-0.46	0.58	34,199
	Tanganyika	-0.31	-0.30	-0.96	0.28	65,937
	Tshopo	0.80	0.80	0.39	1.21	26,513
	Tshuapa	-0.13	-0.12	-0.70	0.37	52,301
Covariate	Bas-Uele	-0.23	-0.23	-0.99	0.49	2,408
	Equateur	1.08	1.08	0.59	1.59	2,218
	Haut-Katanga	-0.28	-0.28	-0.76	0.16	13,191

	Haut-Lomami	-0.01	-0.01	-0.69	0.67	4,683
	Haut-Uele	-0.22	-0.21	-1.02	0.50	5,160
	Ituri	0.71	0.70	0.09	1.38	898
	Kasai	-0.15	-0.14	-0.69	0.37	4,648
	Kasai-Central	-0.44	-0.43	-0.97	0.04	5,819
	Kasai-Oriental	0.21	0.21	-0.21	0.62	5,105
	Kinshasa	-0.10	-0.10	-1.17	1.01	1,925
	Kongo-Central	-0.86	-0.84	-1.54	-0.23	2,879
	Kwango	-0.35	-0.35	-0.85	0.12	5,531
	Kwilu	-0.01	-0.01	-0.47	0.44	1,862
	Lomami	0.21	0.20	-0.33	0.80	2,675
	Lualaba	-0.30	-0.29	-1.07	0.41	7,447
	Mai-Ndombe	-0.13	-0.12	-0.64	0.36	4,720
	Maniema	0.25	0.25	-0.21	0.72	3,230
	Mongala	0.56	0.56	0.05	1.12	2,077

	Nord-Kivu	0.28	0.28	-0.27	0.79	893
	Nord-Ubangi	-0.26	-0.26	-1.03	0.46	3,612
	Sankuru	-0.20	-0.18	-0.81	0.34	8,963
	Sud-Kivu	-0.17	-0.16	-0.81	0.45	1,154
	Sud-Ubangi	-0.06	-0.05	-0.65	0.52	3,050
	Tanganyika	-0.18	-0.17	-0.86	0.45	12,109
	Tshopo	0.72	0.72	0.28	1.18	3,767
	Tshuapa	-0.08	-0.07	-0.65	0.44	18,725

Appendix 3.1 Table 8 - Summary of the ϕ posterior for each province: The mean, median, and 95% credible interval for ϕ posterior was calculated with respect to the province. These are provided as reference.



Appendix 3.1 Figure 11 - Spatial Model Standard Errors: The standard errors of the posterior prevalence distribution for the final cluster-level (left) and province-level (right) model. For the cluster-level model, the standard error range was small (range: 1.62 x 10^{-8}, 2.30 x 10^{-6}). Standard errors were highest where the prevalence estimates were greatest, indicating a degree of uncertainty that coincides with higher covariates values (Appendix 3.1 Figure 10). The province-level models also exhibited a small standard error range (range: 5.0 x 10^{-3}, 3.43 x 10^{-2}). Standard errors at the province-level appeared to be greatest along the Eastern and Western borders.

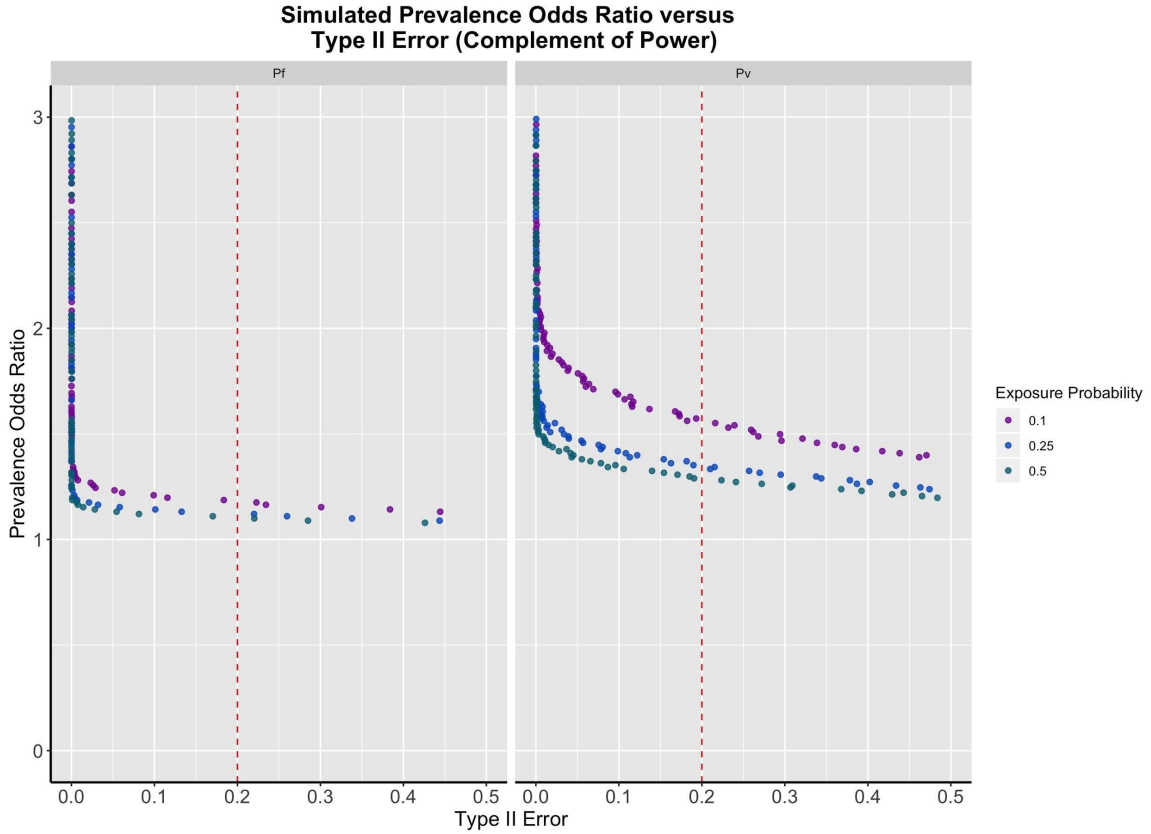
post-hoc Power Calculations

Power calculations were simulated from a population of 15,490 individuals (the weighted N from the study population), where the probability of exposure, $\Pr(E)$ was varied at 10%, 25%, and 50% within the population. For the *P. vivax* models, the overall prevalence of the outcome, *O*, was set at 3% but was varied in the unexposed group from 0.01 - 3.0% (O_U). In contrast, for *P. falciparum* models, the overall prevalence of the outcome was set at 30% and was varied in the unexposed group from 1.0 - 30.0%. ORs were simulated under the following framework:

$$A_i \sim Bernoulli(\Pr(E)) for i \in 1:N$$
$$O_E = 2 * O - O_U$$

$$D_i = \begin{cases} Bernoulli(\Pr(O_E)), & \text{if } A_1 = 1\ , \\ Bernoulli(\Pr(O_U)), & \text{if}_i = 0\ . \end{cases}$$

As a result, from the exposure status, A_i, and disease status, D_i, I can calculate the simulated OR using the standard generalized linear model function with a logit-link in R. Power was calculated as the number of iterations that the parameter estimate α was less than 0.05 with respect to each OR.



Appendix 3.1 Figure 12 - Power Calculations for *P. vivax* and *P. falciparum*: I performed *a posteriori* power calculations to determine the minimum detectable risk factor at varying levels of exposure with 80% power given the prevalence of *P. vivax* and *P. falciparum* identified in the study. At the lowest exposure probability (lowest expected power), I could detect a harmful prevalence odds ratio of approximately 1.54 and 1.18 for *P. vivax* ("Pv") and *P. falciparum* ("Pf"), respectively.

Population Genetics

Hybrid Selection and Next Generation Sequencing

Samples from the DRC were amplified using the Illustra Genomic Phi V2 DNA Amplification Kit (GE Healthcare Life Sciences, Pittsburgh, PA) and prepared for sequencing using the NEBNext Ultra DNA Library Prep Kit for Illumina (New England BioLabs Inc., Ipswich, MA). Amplified libraries were then enriched using custom MYbaits targeting the *P. vivax* genome (version 3.0; MYcroarray: The Oligo Library Company, Ann Arbor, MI). Enriched genomes were sequenced on MiSeq 150 base-pair paired-end and HiSeq2500 125 base-pair paired-end chemistry (Illumina, San Diego, CA).

Publicly Available Whole Genome Sequences

I downloaded 684 publicly available Illumina paired-end *P. vivax* or *P. vivax*-like whole genome sequences from across the globe from the European Nucleotide Archive (**Appendix 3.2**) [65–78]. In addition, I downloaded Illumina single-end sequences for a single isolate that was recovered from a microscopy slide dating to Spain, 1944 [79]. *P. cynomologi* Illumina paired-end sequences were downloaded for both the M- and B-strains (Accessions: DRS000258, ERS001838, ERS023609) [80,81].

Alignment, Quality Control, and Variant Discovery

Reads were aligned to the *P. vivax* P01 reference genome (ftp://ftp.sanger.ac.uk/pub/project/pathogens/gff3/CURRENT/PvivaxP01.genome.fasta.gz) with `bwa mem` (v0.7) after undergoing adaptor-trimming with `cutadapt` (v1.16) [82]. Alignments were then deduplicated and mate-tags were added using `samblaster` (v0.1.24). The quality of the alignments were assessed using the Genome Analysis Toolkit (GATK) `CallableLoci` tool (v3.8-0). I defined a "callable" loci as sites with greater than or equal to five high-quality reads (MQ >= 10, BQ >= 20). Upon inspection of the DRC isolates, I found that genomic coverage

was sparse and only the mitochondria passed quality-thresholds. As a result, all further analyses were limited to the mitochondria. I then performed short variant discovery using GATK `HaplotypeCaller` followed by joint genotyping across all *P. vivax* samples with GATK `GenotypeGVCFs` (v4.0.3) [83].

Variant Filtering and Consensus Haplotypes

Samples were excluded from downstream processing if less than 95% of their mitochondrial genome was callable (24/685 samples). Loci were then filtered using the GATK "hard filtering" approach, following previously established guidelines for both single nucleotide variants (SNVs) and insertion-deletions (INDELs) [84]. Specifically, I filtered loci with a quality-depth of less than two (QD < 2), position bias (ReadPosRankSum < -8.0 for SNV, ReadPosRankSum < -20.0 for INDELs), strand bias (FS > 60 for SNV, FS > 200 for INDEL, SOR > 3 for SNV, SOR > 10 for INDELs), and low mapping-quality (MQ < 35, MQSR < -12.5) using the GATK `VariantFiltration` and `SelectVariants` tools (v4.0.3).

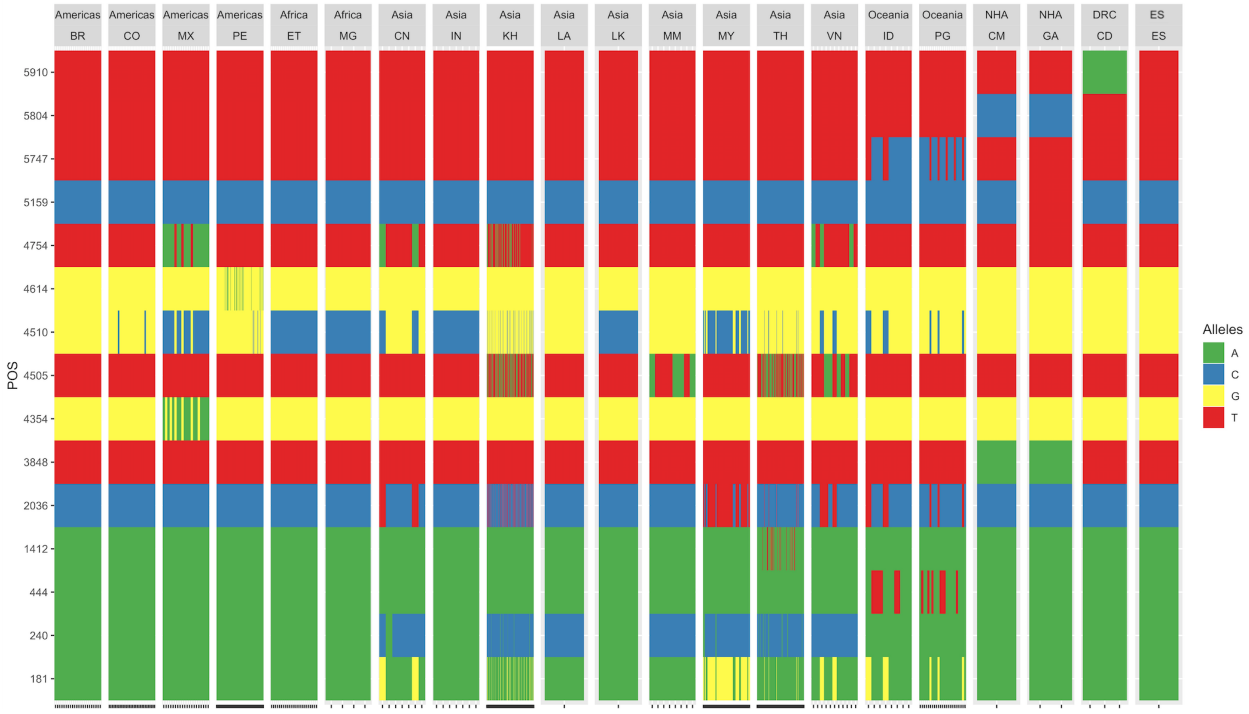
Following hard-filtering, I performed post-processing of loci and samples using the `vcfR` package and other custom scripts (GitHub: IDEELResearch/vcfRmanip) [85]. Passed loci were first limited to SNVs and sites that encoded a deletion as an alternative allele were excluded (i.e. `*` in the ALT category). Samples with more than 20% of SNV genotyped as heterozygous were excluded under an assumption of heteroplasmy. I then imputed the genotype of missing loci based on the sample's within-country allele frequency. Two samples that were the only isolate from their country of origin, ERS347479 (Laos) and ERS040109 (Sri Lanka), were combined into Thailand and India for imputation, respectively. Following imputation, heterozygous sites were recoded as the major allele. Finally, I removed alleles within a country if the within-country

allele frequency was less than $\frac{n}{n-1} * 0.1$. In a large population, this expression simplifies to removing alleles that are at less than 10% frequency within a country.

From the resulting stringently filtered genotype calls, I created a consensus haplotype for each sample using the P01 mitochondrial sequence as a backbone. Two samples -- both a part of the *P. vivax*-like Clade 2 from Gilabert *et al.* 2018 -- were found to have a higher-order of diversity than expected (ERS333076, ERS352725) and were subsequently excluded from further analysis.

In total, 636/685 samples passed quality thresholds and were included in analyses. The Ebro-1944 sample was originally excluded at the callable loci stage (3,148/5,989 bases callable) but was later recovered for visual comparison (**Appendix 3.1 Figure 13**).

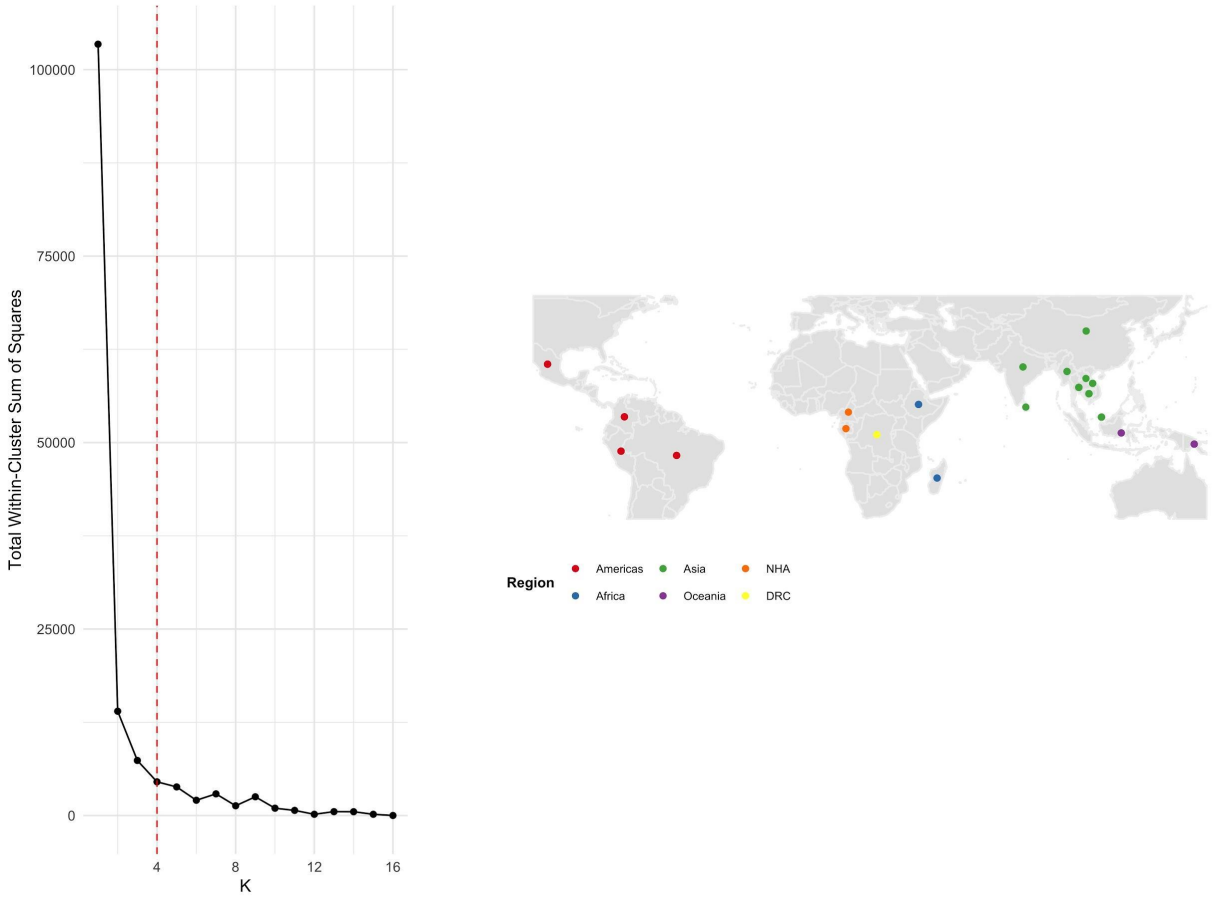
Separately, the *P. cynomolgi* samples also underwent variant discovery, joint genotyping, and hard-filtering as described above. The resulting hard-filtered variants among the three *P. cynomolgi* isolates were then processed by recoding heterozygous alleles as homozygous based on the major allele. Variants were then limited to SNVs and for each variant site, the most common allele among the three isolates was selected. Using these consensus SNVs, I then used the *P. vivax* P01 backbone to create a *P. cynomolgi* consensus haplotype using the `SeqinR` and `Biostrings` packages [86,87].



Appendix 3.1 Figure 13 - Consensus Haplotypes: Haplotypes are shown for each isolate that passed quality-control (QC) threshold with the exception of the sample from Spain (ES) dating to 1944 in the Ebro region (Ebro-1944). As described above, the Ebro-1944 sample did not initially pass QC thresholds but was later recovered for visual comparison. *Abbreviations*: DRC – Democratic Republic of the Congo, NHA – non-human apes, Brazil (BR), Colombia (CO), Mexico (MX), Peru (PE), China (CN), Indonesia (ID), Cambodia (KH), Laos (LA), Myanmar (MM), Malaysia, Papua New Guinea, Thailand (TH), and Vietnam (VN). India (IN), Sri Lanka (LK), Ethiopia (ET), Madagascar (MG), Democratic Republic of the Congo (CD), Cameroon (CM), and Gabon (GA).

Population Genetic Statistics and Phylogenetics

Isolates were first divided into global regions using geographic K-means clustering. I selected K to be four based on minimizing the within-cluster sum of squares while avoiding overfitting. Samples from the DRC and NHA samples were also designated separate clusters (**Appendix 3.1 Figure 14**).



Appendix 3.1 Figure 14 - Spatial Cross-Validation K-Clusters: Countries with *P. vivax* isolates included in the study were partitioned into K-groups for diversity and population-structure measures. Based on the geographical K-means total within-cluster sum-of-squares, four sub-populations appeared to be a reasonable balance between minimizing the total within-cluster sum of squares while avoiding overfitting the data (left). The DRC samples and non-human ape samples were included as separate populations based on the overall study question and prior assumptions (right). *Abbreviations*: DRC – Democratic Republic of the Congo, NHA – non-human apes.

To explore patterns of diversity among the global isolates, I first measured within-region nucleotide and haplotype using the R-package, `PopGenome` (**Appendix 3.1 Table 8**) [88–91]. I then evaluated the degree population differentiation among parasite using measures of between-region nucleotide and haplotype diversity as well as pairwise measures of Hudson's F_{st} (**Appendix 3.1 Table 9**) [88,89,92,93]. Population differentiation was also calculated using a

Hamming's distance between consensus haplotypes with the `ape` R-package [94]. Haplotype differences were then mapped and visualized directly for the DRC (**Figure 3.5**).

Population	Nucleotide Diversity	Haplotype Diversity
Americas	0.7	0.38
Africa	0	0
Asia	1.8	0.77
Oceania	1.65	0.68
NHA	0.67	0.67
DRC	0	0

Appendix 3.1 Table 8 - Within Population Measures of Diversity: For each population, the within-population nucleotide diversity and haplotype diversity were evaluated. Overall, there was little within population diversity among samples from Africa as a whole. This lack of diversity may be an effect of the sample size. *Abbreviations*: DRC – Democratic Republic of the Congo, NHA – non-human apes.

Pop1	Pop2	Between Haplotype Diversity	Between Nucleotide Diversity	Hudson's Fst
Africa	Americas	0.97	1.14	0.81
Asia	Americas	0.7	1.5	0.18
Asia	Africa	0.97	1.74	0.61
Oceania	Americas	0.95	1.93	0.44
Oceania	Africa	1	2.26	0.66
Oceania	Asia	0.95	2.56	0.23
NHA	Americas	1	3.05	0.48
NHA	Africa	1	3.67	0.67
NHA	Asia	1	3.86	0.28
NHA	Oceania	1	4.25	0.32
DRC	Americas	1	1.39	0.81
DRC	Africa	1	2	1
DRC	Asia	1	2.19	0.62
DRC	Oceania	1	2.58	0.66
DRC	NHA	1	3.67	0.67
Global F_{st}		-	-	0.81

Appendix 3.1 Table 9 - Between Population Measures of Diversity and Population Structure: Pairwise comparisons were made for each population (Pop1 versus Pop2) with respect to genetic diversity and population differntiation. Overall, the DRC differed from samples from the Americas the least. However, based on Hudon's F_{st} this similarity was ancestral and did not represent recent mixing. Instead, the DRC samples appeared to be relatively isolated based on Hudson's F_{st}. Overall lack of haplotype sharing is likely -- in part -- due to small sample sizes. *Abbreviations*: DRC – Democratic Republic of the Congo, NHA – non-human apes.

Evolutionary relationships among the isolates were explored using phylogenetic analysis. I first identified the mutational model that best fit the observed data by comparing the Jukes-Cantor versus the General Time Reverse substitution model (GTR + $\gamma(4)$) using maximum likelihood estimation with the `ape` and `phangorn` R-packages [94–97]. For both substitution models, the tree topology, base frequencies, rate matrix, and gamma rate parameters were simultaneously optimized while finding the maximum likelihood. Model fit was compared using AIC, with the GTR model demonstrating a lower AIC and a better model fit. I then performed 1,000 bootstrap iterations of my phylogenetic tree under the GTR model. The phylogenetic tree with the bootstrapped node support was then plotted using the R-package `ggtree`. Finally, I set *P. cynomologi* as the outgroup to orient the tree.

REFERENCES

1. Molly Deutsch-Feldman, Nicholas F. Brazeau, Jonathan B. Parr, Kyaw L. Thwai, Jérémie Muwonga, Melchior Kashamuka, Antoinette K. Tshefu, Jessie K. Edwards, Robert Verity, Michael Emch, Emily W. Gower, Jonathan J. Juliano, Steven R. Meshnick. Spatial and epidemiological drivers of P. falciparum malaria among adults in the Democratic Republic of the Congo.

2. Plowe, C. V., Djimde, A., Bouare, M., Doumbo, O. & Wellems, T. E. Pyrimethamine and proguanil resistance-conferring mutations in Plasmodium falciparum dihydrofolate reductase: polymerase chain reaction methods for surveillance in Africa. *Am. J. Trop. Med. Hyg.* **52**, 565–568 (1995).

3. Srisutham, S. *et al.* Four human Plasmodium species quantification using droplet digital PCR. *PLoS One* **12**, e0175771 (2017).

4. Snounou, G. & Singh, B. Nested PCR analysis of Plasmodium parasites. *Methods Mol. Med.* **72**, 189–203 (2002).

5. Mercereau-Puijalon, O., Barale, J.-C. & Bischoff, E. Three multigene families in Plasmodium parasites: facts and questions. *Int. J. Parasitol.* **32**, 1323–1344 (2002).

6. Gruenberg, M. *et al.* Plasmodium vivax molecular diagnostics in community surveys: pitfalls and solutions. *Malar. J.* **17**, 55 (2018).

7. Tanaka, M., Takahahi, J., Hirayama, F. & Tani, Y. High-resolution melting analysis for genotyping Duffy, Kidd and Diego blood group antigens. *Leg. Med.* **13**, 1–6 (2011).

8. Tournamille, C., Colin, Y., Cartron, J. P. & Van Kim, C. L. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy–negative individuals. *Nat. Genet.* **10**, 224–228 (1995).

9. Ménard, D. *et al.* Plasmodium vivax clinical malaria is commonly observed in Duffy-negative Malagasy people. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5967–5971 (2010).

10. Croft, T. N., Marshall, A. M. J., Allen, C. K. & Others. Guide to DHS statistics. *Rockville, Maryland, USA: ICF* (2018).

11. API Client and Dataset Management for the Demographic and Health Survey (DHS) Data [R package rdhs version 0.6.3].

12. Ouma, P., Okiro, E. A. & Snow, R. W. Sub-Saharan Public Hospitals Geo-coded database. (2017) doi:10.7910/DVN/JTL9VY.

13. Wan, Z., Hook, S., Hulley, G. MYD11C3 MODIS/Aqua Land Surface Temperature/Emissivity Monthly L3 Global 0.05Deg CMG V006 [Data set]. NASA

EOSDIS Land Processes DAAC. (2015).

14. Luxen, D. & Vetter, C. Real-time routing with OpenStreetMap data. in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* 513–516 (ACM, 2011).

15. South, A. rnaturalearth: World Map Data from Natural Earth. (2017).

16. Taylor, S. M. *et al.* Molecular Malaria Epidemiology: Mapping and Burden Estimates for the Democratic Republic of the Congo, 2007. *PLoS One* **6**, e16420 (2011).

17. Millar, J. *et al.* Detecting local risk factors for residual malaria in northern Ghana using Bayesian model averaging. *Malar. J.* **17**, 343 (2018).

18. Tusting, L. S. *et al.* Housing Improvements and Malaria Risk in Sub-Saharan Africa: A Multi-Country Analysis of Survey Data. *PLoS Med.* **14**, e1002234 (2017).

19. Textor, J., van der Zander, B., Gilthorpe, M. S., Liskiewicz, M. & Ellison, G. T. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *Int. J. Epidemiol.* **45**, 1887–1894 (2016).

20. Lindsay, S. W. *et al.* Reduced mosquito survival in metal-roof houses may contribute to a decline in malaria transmission in sub-Saharan Africa. *Sci. Rep.* **9**, 7770 (2019).

21. Rutstein, S. O. Steps to constructing the new DHS Wealth Index. *Rockville, MD: ICF International* (2015).

22. Karney, C. F. F. Algorithms for geodesics. *J. Geodesy* **87**, 43–55 (2013).

23. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* vol. 10 439–446 (2018).

24. Darkoh, E. L., Larbi, J. A. & Lawer, E. A. A Weather-Based Prediction Model of Malaria Prevalence in Amenfi West District, Ghana. *Malar. Res. Treat.* **2017**, 7820454 (2017).

25. Ferrão, J. L., Mendes, J. M. & Painho, M. Modelling the influence of climate on malaria occurrence in Chimoio Municipality, Mozambique. *Parasit. Vectors* **10**, 260 (2017).

26. Wangdi, K. *et al.* Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: A case study in endemic districts of Bhutan. *Malar. J.* **9**, 251 (2010).

27. Nkurunziza, H., Gebhardt, A. & Pilz, J. Bayesian modelling of the effect of climate on malaria in Burundi. *Malar. J.* **9**, 114 (2010).

28. Janko, M. M. *et al.* The links between agriculture, Anopheles mosquitoes, and malaria risk

in children younger than 5 years in the Democratic Republic of the Congo: a population-based, cross-sectional, spatial study. *The Lancet Planetary Health* **2**, e74–e82 (2018).

29. Mayala, B., Fish, T. D., Eitelberg, D. & Dontamsetti, T. *The DHS Program Geospatial Covariate Datasets Manual*. (2018).

30. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **35**, 2769–2794 (2007).

31. Székely, G. J. & Rizzo, M. L. Brownian distance covariance. *Ann. Appl. Stat.* **3**, 1236–1265 (2009).

32. Rizzo, M. & Szekely, G. energy: E-Statistics: Multivariate Inference via the Energy of Data. (2019).

33. Hoseah M. Akala, Oliver Watson, Kenneth K. Mitei, Dennis W. Juma, Robert Verity, Luiser A. Ingasia, Benjamin O. Opot, Raphael O. Okoth, Gladys C. Chemwor, Jackline Juma1, Edwin W. Mwakio, Nicholas Brazeau, Agnes Cheruiyot, Redemptah Yeda, Maureen N. Maraka, Charles Okello, David P. Kateete, Ben Andagalu, Bernhards R. Ogutu, Matthew L. Brown, Jim Ray Managbanag, Edwin Kamau. Characterising Plasmodium inter-species interactions during a period of increasing prevalence of Plasmodium ovale.

34. Liu, W. *et al.* Wild bonobos host geographically restricted malaria parasites including a putative new Laverania species. *Nat. Commun.* **8**, 1635 (2017).

35. Liu, W. *et al.* African origin of the malaria parasite Plasmodium vivax. *Nat. Commun.* **5**, 3346 (2014).

36. Hernán MA, R. J. M. *Causal Inference*. (Boca Raton: Chapman & Hall/CRC).

37. Hernán, M. A. & Robins, J. M. Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health* **60**, 578–586 (2006).

38. Robins, J. M., Hernán, M. A. & Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–560 (2000).

39. Cole, S. R. & Hernán, M. A. Constructing inverse probability weights for marginal structural models. *Am. J. Epidemiol.* **168**, 656–664 (2008).

40. Zhu, Y., Coffman, D. L. & Ghosh, D. A Boosting Algorithm for Estimating Generalized Propensity Scores with Continuous Treatments. *Journal of Causal Inference* vol. 3 (2015).

41. van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super learner. *Stat. Appl. Genet. Mol. Biol.* **6**, Article25 (2007).

42. Polley, E. C. & van der Laan, M. J. Super Learner In Prediction. (2010).

43. Gruber, S., Logan, R. W., Jarrín, I., Monge, S. & Hernán, M. A. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Stat. Med.* **34**, 106–117 (2015).

44. Brenning, A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. in *2012 IEEE International Geoscience and Remote Sensing Symposium* 5372–5375 (2012).

45. R Core Team. R: A Language and Environment for Statistical Computing. (2019).

46. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1 (2010).

47. Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M. & Hofner, B. mboost: Model-Based Boosting. (2018).

48. Schliep, K. & Hechenbichler, K. kknn: Weighted k-Nearest Neighbors. (2016).

49. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. (2019).

50. Venables, W. N. & Ripley, B. D. Modern Applied Statistics with S. (2002).

51. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* vol. 77 1–17 (2017).

52. Bischl, B. *et al.* mlr: Machine Learning in R. *J. Mach. Learn. Res.* **17**, 5938–5942 (2016).

53. Pirracchio, R., Petersen, M. L. & van der Laan, M. Improving propensity score estimators' robustness to model misspecification using super learner. *Am. J. Epidemiol.* **181**, 108–119 (2015).

54. Lumley, T. & Others. Analysis of complex survey samples. *J. Stat. Softw.* **9**, 1–19 (2004).

55. Mills, S., Weiss, S. & Liang, C. VIIRS day/night band (DNB) stray light characterization and correction. in *Earth Observing Systems XVIII* vol. 8866 88661P (International Society for Optics and Photonics, 2013).

56. Elvidge, C. D., Baugh, K., Zhizhin, M., Hsu, F. C. & Ghosh, T. VIIRS night-time lights. *Int. J. Remote Sens.* **38**, 5860–5879 (2017).

57. Lee, D. CARBayes version 4.6: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors. *University of Glasgow, Glasgow* (2017).

58. Giorgi, E., Diggle, P. J. & Others. PrevMap: an R package for prevalence mapping. *J. Off.*

Stat. (2017).

59. Bivand, R. S., Pebesma, E. & Gómez-Rubio, V. *Applied Spatial Data Analysis with R*. (Springer, New York, NY, 2013).

60. Leroux, B. G., Lei, X. & Breslow, N. Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence. in *Statistical Models in Epidemiology, the Environment, and Clinical Trials* 179–191 (Springer New York, 2000).

61. Besag, J., York, J. & Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Stat. Math.* **43**, 1–20 (1991).

62. Lee, D. CARBayes: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors.

63. Pebesma, E. J. Multivariable geostatistics in S: the gstat package. *Comput. Geosci.* **30**, 683–691 (2004).

64. Gräler, B., Pebesma, E. & Heuvelink, G. Spatio-Temporal Interpolation using gstat. The R Journal, 8 (1), 204--218. (2016).

65. Parobek, C. M. *et al.* Selective sweep suggests transcriptional regulation may underlie Plasmodium vivax resilience to malaria control measures in Cambodia. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E8096–E8105 (2016).

66. Hupalo, D. N. *et al.* Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax. *Nat. Genet.* **48**, 953–958 (2016).

67. Loy, D. E. *et al.* Evolutionary history of human Plasmodium vivax revealed by genome-wide analyses of related ape parasites. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E8450–E8459 (2018).

68. Shen, H.-M., Chen, S.-B., Wang, Y. & Chen, J.-H. Whole-genome sequencing of a Plasmodium vivax isolate from the China-Myanmar border area. *Mem. Inst. Oswaldo Cruz* **110**, 814–816 (2015).

69. Shen, H.-M. *et al.* Genome-wide scans for the identification of Plasmodium vivax genes under positive selection. *Malar. J.* **16**, 238 (2017).

70. Gilabert, A. *et al.* Plasmodium vivax-like genome sequences shed new insights into Plasmodium vivax biology and evolution. *PLoS Biol.* **16**, e2006035 (2018).

71. Popovici, J. *et al.* Genomic Analyses Reveal the Common Occurrence and Complexity of Plasmodium vivax Relapses in Cambodia. *MBio* **9**, (2018).

72. Pearson, R. D. *et al.* Genomic analysis of local variation and recent evolution in

Plasmodium vivax. *Nat. Genet.* **48**, 959–964 (2016).

73. Auburn, S. *et al.* Genomic analysis of a pre-elimination Malaysian Plasmodium vivax population reveals selective pressures and changing transmission dynamics. *Nat. Commun.* **9**, 2585 (2018).

74. Cowell, A. N., Valdivia, H. O., Bishop, D. K. & Winzeler, E. A. Exploration of Plasmodium vivax transmission dynamics and recurrent infections in the Peruvian Amazon using whole genome sequencing. *Genome Med.* **10**, 52 (2018).

75. Menard, D. *et al.* Whole genome sequencing of field isolates reveals a common duplication of the Duffy binding protein gene in Malagasy Plasmodium vivax strains. *PLoS Negl. Trop. Dis.* **7**, e2489 (2013).

76. Chan, E. R. *et al.* Whole genome sequencing of field isolates provides robust characterization of genetic diversity in Plasmodium vivax. *PLoS Negl. Trop. Dis.* **6**, e1811 (2012).

77. Sanguinetti, L., Toti, S., Reguzzi, V., Bagnoli, F. & Donati, C. A novel computational method identifies intra- and inter-species recombination events in Staphylococcus aureus and Streptococcus pneumoniae. *PLoS Comput. Biol.* **8**, e1002668 (2012).

78. Auburn, S. *et al.* Genomic Analysis of Plasmodium vivax in Southern Ethiopia Reveals Selective Pressures in Multiple Parasite Mechanisms. *J. Infect. Dis.* **220**, 1738–1749 (2019).

79. Gelabert, P. *et al.* Mitochondrial DNA from the eradicated European Plasmodium vivax and P. falciparum from 70-year-old slides from the Ebro Delta in Spain. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11495–11500 (2016).

80. Pasini, E. M. *et al.* An improvedPlasmodium cynomolgigenome assembly reveals an unexpected methyltransferase gene expansion. *Wellcome Open Res* **2**, 42 (2017).

81. Tachibana, S.-I. *et al.* Plasmodium cynomolgi genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. *Nat. Genet.* **44**, 1051–1055 (2012).

82. Auburn, S. *et al.* A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. *Wellcome Open Res* **1**, 4 (2016).

83. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2017) doi:10.1101/201178.

84. Shultz, A. Whole-genome resquencing for population genomics (Fastq to VCF). *Harvard FAS Informatics* https://informatics.fas.harvard.edu/whole-genome-resquencing-for-population-genomics-fastq-to-vcf.html#variantcalling (2018).

85. Knaus, B. J. & Grünwald, N. J. vcfr: a package to manipulate and visualize variant call

format data in R. *Mol. Ecol. Resour.* **17**, 44–53 (2017).

86. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. (2019).

87. Charif, D. & Lobry, J. R. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. in *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations* (eds. Bastolla, U., Porto, M., Roman, H. E. & Vendruscolo, M.) 207–232 (Springer Berlin Heidelberg, 2007).

88. Pfeifer, B., Wittelsbuerger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution* vol. 31 1929–1936 (2014).

89. Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).

90. Nei, M. Molecular Evolutionary Genetics. (1987) doi:10.7312/nei-92038.

91. Wakeley, J. The variance of pairwise nucleotide differences in two populations with migration. *Theor. Popul. Biol.* **49**, 39–57 (1996).

92. Hudson, R. R. A new statistic for detecting genetic differentiation. *Genetics* **155**, 2011–2014 (2000).

93. Verity, R. & Nichols, R. A. What is genetic differentiation, and how should we measure it—GST, D, neither or both? *Mol. Ecol.* (2014).

94. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* vol. 35 526–528 (2018).

95. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).

96. Jukes, T. H. & Cantor, C. R. Evolution of Protein Molecules. *Mammalian Protein Metabolism* 21–132 (1969) doi:10.1016/b978-1-4832-3211-9.50009-7.

97. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences* **17**, 57–86 (1986).

APPENDIX 3.2: NEXT GENERATION SEQUENCES USED IN THIS DISSERTATION

Accession	PMCID/PMID	Host
SRS1061002	PMC5167194	*Homo sapiens*
SRS1061003	PMC5167194	*Homo sapiens*
SRS1061040	PMC5167194	*Homo sapiens*
SRS1061084	PMC5167194	*Homo sapiens*
SRS1061134	PMC5167194	*Homo sapiens*
SRS1061155	PMC5167194	*Homo sapiens*
SRS1061166	PMC5167194	*Homo sapiens*
SRS1061211	PMC5167194	*Homo sapiens*
SRS1061258	PMC5167194	*Homo sapiens*
SRS693907	PMC5347536	*Homo sapiens*
SRS693915	PMC5347536	*Homo sapiens*
SRS696215	PMC5347536	*Homo sapiens*
SRS694268	PMC5347536	*Homo sapiens*
SRS693978	PMC5347536	*Homo sapiens*
SRS693582	PMC5347536	*Homo sapiens*
SRS693927	PMC5347536	*Homo sapiens*
SRS693491	PMC5347536	*Homo sapiens*
SRS693551	PMC5347536	*Homo sapiens*
SRS693939	PMC5347536	*Homo sapiens*
SRS693916	PMC5347536	*Homo sapiens*
SRS693976	PMC5347536	*Homo sapiens*
SRS693917	PMC5347536	*Homo sapiens*
SRS693940	PMC5347536	*Homo sapiens*
SRS693928	PMC5347536	*Homo sapiens*
SRS693934	PMC5347536	*Homo sapiens*
SRS693910	PMC5347536	*Homo sapiens*
SRS693578	PMC5347536	*Homo sapiens*
SRS693922	PMC5347536	*Homo sapiens*
SRS693278	PMC5347536	*Homo sapiens*
SRS3371819	PMC6130405	*Pan troglodytes*
SRS3371817	PMC6130405	*Pan troglodytes*
SRS3371818	PMC6130405	*Pan troglodytes*
SRS3371815	PMC6130405	*Pan troglodytes*
SRS3371816	PMC6130405	*Pan troglodytes*
SRS3371814	PMC6130405	*Gorilla gorilla*
SRS941624	PMC4667588	*Homo sapiens*

SRS1566636	PMC5461743	*Homo sapiens*
SRS1566640	PMC5461743	*Homo sapiens*
SRS1566641	PMC5461743	*Homo sapiens*
SRS1566642	PMC5461743	*Homo sapiens*
SRS1566602	PMC5461743	*Homo sapiens*
SRS694262	PMC5347536	*Homo sapiens*
SRS693274	PMC5347536	*Homo sapiens*
SRS694239	PMC5347536	*Homo sapiens*
SRS694233	PMC5347536	*Homo sapiens*
SRS694234	PMC5347536	*Homo sapiens*
SRS694230	PMC5347536	*Homo sapiens*
SRS693584	PMC5347536	*Homo sapiens*
SRS694257	PMC5347536	*Homo sapiens*
SRS694264	PMC5347536	*Homo sapiens*
SRS693897	PMC5347536	*Homo sapiens*
SRS693580	PMC5347536	*Homo sapiens*
SRS693296	PMC5347536	*Homo sapiens*
SRS694265	PMC5347536	*Homo sapiens*
SRS694263	PMC5347536	*Homo sapiens*
SRS693489	PMC5347536	*Homo sapiens*
SRS694235	PMC5347536	*Homo sapiens*
SRS694241	PMC5347536	*Homo sapiens*
SRS693267	PMC5347536	*Homo sapiens*
SRS693950	PMC5347536	*Homo sapiens*
SRS693947	PMC5347536	*Homo sapiens*
SRS694231	PMC5347536	*Homo sapiens*
SRS693938	PMC5347536	*Homo sapiens*
SRS693575	PMC5347536	*Homo sapiens*
SRS693442	PMC5347536	*Homo sapiens*
SRS694258	PMC5347536	*Homo sapiens*
SRS694251	PMC5347536	*Homo sapiens*
SRS693577	PMC5347536	*Homo sapiens*
SRS693951	PMC5347536	*Homo sapiens*
SRS694242	PMC5347536	*Homo sapiens*
SRS694247	PMC5347536	*Homo sapiens*
ERS1452911	PMC6130868	*Pan troglodytes*
ERS333071	PMC6130868	*Pan troglodytes*
ERS333073	PMC6130868	*Pan troglodytes*
ERS333077	PMC6130868	*Pan troglodytes*

ERS333070	PMC6130868	*Pan troglodytes*
ERS352729	PMC6130868	*Pan troglodytes*
ERS333076	PMC6130868	*Pan troglodytes*
ERS434568	PMC6130868	*Pan troglodytes*
ERS333055	PMC6130868	*Anopheles gambiae*
ERS352726	PMC6130868	*Pan troglodytes*
ERS352725	PMC6130868	*Pan troglodytes*
SRS696222	PMC5347536	*Homo sapiens*
SRS805922	PMC5347536	*Homo sapiens*
SRS807702	PMC5347536	*Homo sapiens*
SRS807711	PMC5347536	*Homo sapiens*
SRS807712	PMC5347536	*Homo sapiens*
SRS805942	PMC5347536	*Homo sapiens*
SRS805943	PMC5347536	*Homo sapiens*
SRS807544	PMC5347536	*Homo sapiens*
SRS807701	PMC5347536	*Homo sapiens*
SRS2746073	PMC5784252	*Homo sapiens*
SRS2746025	PMC5784252	*Homo sapiens*
SRS2745931	PMC5784252	*Homo sapiens*
SRS2745827	PMC5784252	*Homo sapiens*
SRS2745833	PMC5784252	*Homo sapiens*
SRS2745857	PMC5784252	*Homo sapiens*
SRS2745842	PMC5784252	*Homo sapiens*
SRS2745858	PMC5784252	*Homo sapiens*
SRS2745937	PMC5784252	*Homo sapiens*
SRS2745959	PMC5784252	*Homo sapiens*
SRS2745846	PMC5784252	*Homo sapiens*
SRS2745815	PMC5784252	*Homo sapiens*
SRS2745933	PMC5784252	*Homo sapiens*
SRS2746083	PMC5784252	*Homo sapiens*
SRS2746099	PMC5784252	*Homo sapiens*
SRS2746066	PMC5784252	*Homo sapiens*
SRS2745934	PMC5784252	*Homo sapiens*
SRS2746116	PMC5784252	*Homo sapiens*
SRS2746061	PMC5784252	*Homo sapiens*
SRS2746071	PMC5784252	*Homo sapiens*
SRS2746197	PMC5784252	*Homo sapiens*
SRS2745879	PMC5784252	*Homo sapiens*
SRS2745892	PMC5784252	*Homo sapiens*

SRS2745886	PMC5784252	*Homo sapiens*
SRS2745999	PMC5784252	*Homo sapiens*
SRS2746067	PMC5784252	*Homo sapiens*
SRS2746210	PMC5784252	*Homo sapiens*
SRS2746138	PMC5784252	*Homo sapiens*
SRS2746090	PMC5784252	*Homo sapiens*
SRS2746088	PMC5784252	*Homo sapiens*
SRS2745935	PMC5784252	*Homo sapiens*
SRS2745942	PMC5784252	*Homo sapiens*
SRS2745839	PMC5784252	*Homo sapiens*
SRS2745840	PMC5784252	*Homo sapiens*
SRS2745835	PMC5784252	*Homo sapiens*
SRS2745922	PMC5784252	*Homo sapiens*
SRS2745838	PMC5784252	*Homo sapiens*
SRS2746207	PMC5784252	*Homo sapiens*
SRS2746069	PMC5784252	*Homo sapiens*
SRS2745932	PMC5784252	*Homo sapiens*
SRS2745986	PMC5784252	*Homo sapiens*
SRS2746070	PMC5784252	*Homo sapiens*
SRS2746255	PMC5784252	*Homo sapiens*
SRS2745973	PMC5784252	*Homo sapiens*
SRS2746064	PMC5784252	*Homo sapiens*
SRS2746209	PMC5784252	*Homo sapiens*
SRS2746213	PMC5784252	*Homo sapiens*
SRS2745866	PMC5784252	*Homo sapiens*
SRS2746214	PMC5784252	*Homo sapiens*
SRS2745868	PMC5784252	*Homo sapiens*
SRS2746216	PMC5784252	*Homo sapiens*
SRS2745998	PMC5784252	*Homo sapiens*
SRS2746215	PMC5784252	*Homo sapiens*
SRS2745883	PMC5784252	*Homo sapiens*
SRS2746205	PMC5784252	*Homo sapiens*
SRS2746206	PMC5784252	*Homo sapiens*
SRS2746112	PMC5784252	*Homo sapiens*
SRS2746208	PMC5784252	*Homo sapiens*
SRS2746065	PMC5784252	*Homo sapiens*
SRS2746109	PMC5784252	*Homo sapiens*
SRS2746068	PMC5784252	*Homo sapiens*
SRS2746043	PMC5784252	*Homo sapiens*

SRS1061259	PMC5167194	*Homo sapiens*
SRS1061004	PMC5167194	*Homo sapiens*
SRS1061005	PMC5167194	*Homo sapiens*
SRS1061008	PMC5167194	*Homo sapiens*
SRS1061007	PMC5167194	*Homo sapiens*
SRS1061011	PMC5167194	*Homo sapiens*
SRS1061014	PMC5167194	*Homo sapiens*
SRS1061030	PMC5167194	*Homo sapiens*
SRS1061034	PMC5167194	*Homo sapiens*
SRS693900	PMC5347536	*Homo sapiens*
SRS694248	PMC5347536	*Homo sapiens*
SRS694266	PMC5347536	*Homo sapiens*
SRS693902	PMC5347536	*Homo sapiens*
SRS694259	PMC5347536	*Homo sapiens*
SRS694255	PMC5347536	*Homo sapiens*
SRS693265	PMC5347536	*Homo sapiens*
SRS693908	PMC5347536	*Homo sapiens*
SRS694246	PMC5347536	*Homo sapiens*
SRS694244	PMC5347536	*Homo sapiens*
SRS693462	PMC5347536	*Homo sapiens*
SRS694245	PMC5347536	*Homo sapiens*
SRS694256	PMC5347536	*Homo sapiens*
SRS693273	PMC5347536	*Homo sapiens*
SRS694229	PMC5347536	*Homo sapiens*
SRS693576	PMC5347536	*Homo sapiens*
SRS694237	PMC5347536	*Homo sapiens*
SRS693463	PMC5347536	*Homo sapiens*
SRS694267	PMC5347536	*Homo sapiens*
SRS694260	PMC5347536	*Homo sapiens*
SRS694232	PMC5347536	*Homo sapiens*
SRS694243	PMC5347536	*Homo sapiens*
SRS694254	PMC5347536	*Homo sapiens*
SRS694249	PMC5347536	*Homo sapiens*
SRS694227	PMC5347536	*Homo sapiens*
SRS694261	PMC5347536	*Homo sapiens*
SRS694236	PMC5347536	*Homo sapiens*
SRS693271	PMC5347536	*Homo sapiens*
SRS1061036	PMC5167194	*Homo sapiens*
SRS1061038	PMC5167194	*Homo sapiens*

SRS1061042	PMC5167194	*Homo sapiens*
SRS1061044	PMC5167194	*Homo sapiens*
SRS1061045	PMC5167194	*Homo sapiens*
SRS1061046	PMC5167194	*Homo sapiens*
SRS1061049	PMC5167194	*Homo sapiens*
SRS1061078	PMC5167194	*Homo sapiens*
SRS1061080	PMC5167194	*Homo sapiens*
SRS1061081	PMC5167194	*Homo sapiens*
SRS1061082	PMC5167194	*Homo sapiens*
SRS1061083	PMC5167194	*Homo sapiens*
SRS1061085	PMC5167194	*Homo sapiens*
SRS1061086	PMC5167194	*Homo sapiens*
SRS1061088	PMC5167194	*Homo sapiens*
SRS1061090	PMC5167194	*Homo sapiens*
SRS1061091	PMC5167194	*Homo sapiens*
SRS1061094	PMC5167194	*Homo sapiens*
SRS1061098	PMC5167194	*Homo sapiens*
SRS1061119	PMC5167194	*Homo sapiens*
SRS1061125	PMC5167194	*Homo sapiens*
SRS1061128	PMC5167194	*Homo sapiens*
SRS1061129	PMC5167194	*Homo sapiens*
SRS1061135	PMC5167194	*Homo sapiens*
SRS1061136	PMC5167194	*Homo sapiens*
SRS1061142	PMC5167194	*Homo sapiens*
SRS1061143	PMC5167194	*Homo sapiens*
SRS1061147	PMC5167194	*Homo sapiens*
SRS1061151	PMC5167194	*Homo sapiens*
SRS1061152	PMC5167194	*Homo sapiens*
SRS1061153	PMC5167194	*Homo sapiens*
SRS1061154	PMC5167194	*Homo sapiens*
SRS1061156	PMC5167194	*Homo sapiens*
SRS1061157	PMC5167194	*Homo sapiens*
SRS1061158	PMC5167194	*Homo sapiens*
SRS1061159	PMC5167194	*Homo sapiens*
SRS1061160	PMC5167194	*Homo sapiens*
SRS1061161	PMC5167194	*Homo sapiens*
SRS1061162	PMC5167194	*Homo sapiens*
SRS1061163	PMC5167194	*Homo sapiens*
SRS1061164	PMC5167194	*Homo sapiens*

SRS1061165	PMC5167194	*Homo sapiens*
SRS1061191	PMC5167194	*Homo sapiens*
SRS1061192	PMC5167194	*Homo sapiens*
SRS1061193	PMC5167194	*Homo sapiens*
SRS1061194	PMC5167194	*Homo sapiens*
SRS1061197	PMC5167194	*Homo sapiens*
SRS1061196	PMC5167194	*Homo sapiens*
SRS1061199	PMC5167194	*Homo sapiens*
SRS1061202	PMC5167194	*Homo sapiens*
SRS1061200	PMC5167194	*Homo sapiens*
SRS1061212	PMC5167194	*Homo sapiens*
SRS1061214	PMC5167194	*Homo sapiens*
SRS1061226	PMC5167194	*Homo sapiens*
SRS1061231	PMC5167194	*Homo sapiens*
SRS1061232	PMC5167194	*Homo sapiens*
SRS1061234	PMC5167194	*Homo sapiens*
SRS1061238	PMC5167194	*Homo sapiens*
SRS1061252	PMC5167194	*Homo sapiens*
SRS1061253	PMC5167194	*Homo sapiens*
ERS055892	PMC4966634	*Homo sapiens*
ERS055878	PMC4966634	*Homo sapiens*
ERS055895	PMC4966634	*Homo sapiens*
ERS055889	PMC4966634	*Homo sapiens*
ERS055896	PMC4966634	*Homo sapiens*
ERS055885	PMC4966634	*Homo sapiens*
ERS055881	PMC4966634	*Homo sapiens*
ERS055887	PMC4966634	*Homo sapiens*
ERS055888	PMC4966634	*Homo sapiens*
ERS055877	PMC4966634	*Homo sapiens*
ERS055882	PMC4966634	*Homo sapiens*
ERS055884	PMC4966634	*Homo sapiens*
ERS055893	PMC4966634	*Homo sapiens*
ERS055886	PMC4966634	*Homo sapiens*
ERS055883	PMC4966634	*Homo sapiens*
ERS055891	PMC4966634	*Homo sapiens*
ERS055890	PMC4966634	*Homo sapiens*
ERS055880	PMC4966634	*Homo sapiens*
ERS055894	PMC4966634	*Homo sapiens*
ERS055879	PMC4966634	*Homo sapiens*

ERS142861	PMC4966634	*Homo sapiens*
ERS241405	PMC4966634	*Homo sapiens*
ERS241409	PMC4966634	*Homo sapiens*
ERS241412	PMC4966634	*Homo sapiens*
ERS241415	PMC4966634	*Homo sapiens*
ERS241418	PMC4966634	*Homo sapiens*
ERS241421	PMC4966634	*Homo sapiens*
ERS241424	PMC4966634	*Homo sapiens*
ERS241427	PMC4966634	*Homo sapiens*
ERS241430	PMC4966634	*Homo sapiens*
ERS241433	PMC4966634	*Homo sapiens*
ERS241436	PMC4966634	*Homo sapiens*
ERS241439	PMC4966634	*Homo sapiens*
ERS241406	PMC4966634	*Homo sapiens*
ERS241410	PMC4966634	*Homo sapiens*
ERS241413	PMC4966634	*Homo sapiens*
ERS241416	PMC4966634	*Homo sapiens*
ERS241419	PMC4966634	*Homo sapiens*
ERS241422	PMC4966634	*Homo sapiens*
ERS241425	PMC4966634	*Homo sapiens*
ERS241428	PMC4966634	*Homo sapiens*
ERS241431	PMC4966634	*Homo sapiens*
ERS241434	PMC4966634	*Homo sapiens*
ERS241437	PMC4966634	*Homo sapiens*
ERS241440	PMC6030216	*Homo sapiens*
ERS241407	PMC4966634	*Homo sapiens*
ERS241411	PMC4966634	*Homo sapiens*
ERS241414	PMC4966634	*Homo sapiens*
ERS241417	PMC4966634	*Homo sapiens*
ERS241423	PMC6030216	*Homo sapiens*
ERS241426	PMC6030216	*Homo sapiens*
ERS241429	PMC4966634	*Homo sapiens*
ERS241432	PMC4966634	*Homo sapiens*
ERS347497	PMC4966634	*Homo sapiens*
ERS347698	PMC4966634	*Homo sapiens*
ERS347704	PMC4966634	*Homo sapiens*
ERS403521	PMC4966634	*Homo sapiens*
ERS403526	PMC4966634	*Homo sapiens*
ERS403530	PMC4966634	*Homo sapiens*

ERS403534	PMC6030216	*Homo sapiens*
ERS403538	PMC4966634	*Homo sapiens*
ERS403542	PMC4966634	*Homo sapiens*
ERS403550	PMC4966634	*Homo sapiens*
ERS403554	PMC4966634	*Homo sapiens*
ERS503258	PMC6030216	*Homo sapiens*
ERS403562	PMC6030216	*Homo sapiens*
ERS403517	PMC4966634	*Homo sapiens*
ERS403522	PMC4966634	*Homo sapiens*
ERS403527	PMC4966634	*Homo sapiens*
ERS403531	PMC4966634	*Homo sapiens*
ERS403535	PMC4966634	*Homo sapiens*
ERS403539	PMC4966634	*Homo sapiens*
ERS403543	PMC4966634	*Homo sapiens*
ERS403547	PMC4966634	*Homo sapiens*
ERS403551	PMC4966634	*Homo sapiens*
ERS403555	PMC4966634	*Homo sapiens*
ERS403559	PMC4966634	*Homo sapiens*
ERS403563	PMC4966634	*Homo sapiens*
ERS403518	PMC4966634	*Homo sapiens*
ERS403523	PMC4966634	*Homo sapiens*
ERS403528	PMC4966634	*Homo sapiens*
ERS403532	PMC4966634	*Homo sapiens*
ERS403536	PMC4966634	*Homo sapiens*
ERS403540	PMC4966634	*Homo sapiens*
ERS403544	PMC4966634	*Homo sapiens*
ERS403548	PMC4966634	*Homo sapiens*
ERS403552	PMC4966634	*Homo sapiens*
ERS403556	PMC4966634	*Homo sapiens*
ERS403560	PMC4966634	*Homo sapiens*
ERS403564	PMC4966634	*Homo sapiens*
ERS403519	PMC4966634	*Homo sapiens*
ERS403524	PMC4966634	*Homo sapiens*
ERS403529	PMC4966634	*Homo sapiens*
ERS403533	PMC4966634	*Homo sapiens*
ERS403537	PMC4966634	*Homo sapiens*
ERS403541	PMC4966634	*Homo sapiens*
ERS403545	PMC4966634	*Homo sapiens*
ERS403549	PMC4966634	*Homo sapiens*

ERS564517	PMC6030216	*Homo sapiens*
ERS564518	PMC6030216	*Homo sapiens*
ERS564519	PMC6030216	*Homo sapiens*
ERS564520	PMC6030216	*Homo sapiens*
ERS980423	PMC6030216	*Homo sapiens*
ERS980426	PMC6030216	*Homo sapiens*
ERS980427	PMC6030216	*Homo sapiens*
ERS980428	PMC6030216	*Homo sapiens*
ERS980429	PMC6030216	*Homo sapiens*
ERS980431	PMC6030216	*Homo sapiens*
SRS2909995	PMC6032790	*Homo sapiens*
SRS2909996	PMC6032790	*Homo sapiens*
SRS1885934	PMC6032790	*Homo sapiens*
SRS1885924	PMC6032790	*Homo sapiens*
SRS1885936	PMC6032790	*Homo sapiens*
SRS1885920	PMC6032790	*Homo sapiens*
SRS1885929	PMC6032790	*Homo sapiens*
SRS1885928	PMC6032790	*Homo sapiens*
SRS1885938	PMC6032790	*Homo sapiens*
SRS1885931	PMC6032790	*Homo sapiens*
SRS1885941	PMC6032790	*Homo sapiens*
SRS1885925	PMC6032790	*Homo sapiens*
SRS1885932	PMC6032790	*Homo sapiens*
SRS1885922	PMC6032790	*Homo sapiens*
SRS1885930	PMC6032790	*Homo sapiens*
SRS1885933	PMC6032790	*Homo sapiens*
SRS1885923	PMC6032790	*Homo sapiens*
SRS1885921	PMC6032790	*Homo sapiens*
SRS1885937	PMC6032790	*Homo sapiens*
SRS1885939	PMC6032790	*Homo sapiens*
SRS1885935	PMC6032790	*Homo sapiens*
SRS1885927	PMC6032790	*Homo sapiens*
SRS2910041	PMC6032790	*Homo sapiens*
SRS2909997	PMC6032790	*Homo sapiens*
SRS2910026	PMC6032790	*Homo sapiens*
SRS2910023	PMC6032790	*Homo sapiens*
SRS2909991	PMC6032790	*Homo sapiens*
SRS2909989	PMC6032790	*Homo sapiens*
SRS2909999	PMC6032790	*Homo sapiens*

SRS2910011	PMC6032790	*Homo sapiens*
SRS2910013	PMC6032790	*Homo sapiens*
SRS2910007	PMC6032790	*Homo sapiens*
SRS2910016	PMC6032790	*Homo sapiens*
SRS2910009	PMC6032790	*Homo sapiens*
SRS2910002	PMC6032790	*Homo sapiens*
SRS2910049	PMC6032790	*Homo sapiens*
SRS2910051	PMC6032790	*Homo sapiens*
SRS2910045	PMC6032790	*Homo sapiens*
SRS2910057	PMC6032790	*Homo sapiens*
SRS2910046	PMC6032790	*Homo sapiens*
SRS2910053	PMC6032790	*Homo sapiens*
SRS2910052	PMC6032790	*Homo sapiens*
SRS2910021	PMC6032790	*Homo sapiens*
SRS2910019	PMC6032790	*Homo sapiens*
SRS2910027	PMC6032790	*Homo sapiens*
SRS2910017	PMC6032790	*Homo sapiens*
SRS2910014	PMC6032790	*Homo sapiens*
SRS2910024	PMC6032790	*Homo sapiens*
SRS2910022	PMC6032790	*Homo sapiens*
SRS2910031	PMC6032790	*Homo sapiens*
SRS2910036	PMC6032790	*Homo sapiens*
SRS2910035	PMC6032790	*Homo sapiens*
SRS2910034	PMC6032790	*Homo sapiens*
SRS2910054	PMC6032790	*Homo sapiens*
SRS2910032	PMC6032790	*Homo sapiens*
SRS2910042	PMC6032790	*Homo sapiens*
SRS2910030	PMC6032790	*Homo sapiens*
SRS2910029	PMC6032790	*Homo sapiens*
SRS2910044	PMC6032790	*Homo sapiens*
SRS2910043	PMC6032790	*Homo sapiens*
SRS2910008	PMC6032790	*Homo sapiens*
SRS2910000	PMC6032790	*Homo sapiens*
SRS2910015	PMC6032790	*Homo sapiens*
SRS2910001	PMC6032790	*Homo sapiens*
SRS2910040	PMC6032790	*Homo sapiens*
SRS2910028	PMC6032790	*Homo sapiens*
SRS2910004	PMC6032790	*Homo sapiens*
SRS2910005	PMC6032790	*Homo sapiens*

SRS2910033	PMC6032790	*Homo sapiens*
SRS2910025	PMC6032790	*Homo sapiens*
SRS2910038	PMC6032790	*Homo sapiens*
SRS2910037	PMC6032790	*Homo sapiens*
SRS2910055	PMC6032790	*Homo sapiens*
SRS2910039	PMC6032790	*Homo sapiens*
SRS2910056	PMC6032790	*Homo sapiens*
SRS1885926	PMC6032790	*Homo sapiens*
SRS1885940	PMC6032790	*Homo sapiens*
SRS693277	PMC5347536	*Homo sapiens*
SRS694228	PMC5347536	*Homo sapiens*
SRS819436	PMC5347536	*Homo sapiens*
SRS819437	PMC5347536	*Homo sapiens*
SRS819479	PMC5347536	*Homo sapiens*
SRS696218	PMC5347536	*Homo sapiens*
SRS693979	PMC5347536	*Homo sapiens*
SRS693263	PMC5347536	*Homo sapiens*
SRS693264	PMC5347536	*Homo sapiens*
SRS693270	PMC5347536	*Homo sapiens*
SRS696214	PMC5347536	*Homo sapiens*
SRS693490	PMC5347536	*Homo sapiens*
SRS693569	PMC5347536	*Homo sapiens*
SRS696223	PMC5347536	*Homo sapiens*
SRS693272	PMC5347536	*Homo sapiens*
SRS696221	PMC5347536	*Homo sapiens*
SRS819493	PMC5347536	*Homo sapiens*
SRS694240	PMC5347536	*Homo sapiens*
SRS693933	PMC5347536	*Homo sapiens*
SRS819579	PMC5347536	*Homo sapiens*
SRS696220	PMC5347536	*Homo sapiens*
SRS693911	PMC5347536	*Homo sapiens*
SRS693268	PMC5347536	*Homo sapiens*
SRS693953	PMC5347536	*Homo sapiens*
SRS693923	PMC5347536	*Homo sapiens*
SRS693941	PMC5347536	*Homo sapiens*
SRS693949	PMC5347536	*Homo sapiens*
SRS693276	PMC5347536	*Homo sapiens*
SRS693977	PMC5347536	*Homo sapiens*
SRS693972	PMC5347536	*Homo sapiens*

SRS693307	PMC5347536	*Homo sapiens*
SRS693476	PMC5347536	*Homo sapiens*
SRS696216	PMC5347536	*Homo sapiens*
SRS696213	PMC5347536	*Homo sapiens*
SRS693354	PMC5347536	*Homo sapiens*
SRS693581	PMC5347536	*Homo sapiens*
SRS693925	PMC5347536	*Homo sapiens*
SRS696217	PMC5347536	*Homo sapiens*
SRS693926	PMC5347536	*Homo sapiens*
SRS693364	PMC5347536	*Homo sapiens*
SRS693345	PMC5347536	*Homo sapiens*
SRS693952	PMC5347536	*Homo sapiens*
SRS693355	PMC5347536	*Homo sapiens*
SRS693921	PMC5347536	*Homo sapiens*
SRS693281	PMC5347536	*Homo sapiens*
SRS693948	PMC5347536	*Homo sapiens*
SRS693942	PMC5347536	*Homo sapiens*
SRS819643	PMC5347536	*Homo sapiens*
SRS693957	PMC5347536	*Homo sapiens*
SRS819716	PMC5347536	*Homo sapiens*
SRS693905	PMC5347536	*Homo sapiens*
SRS693544	PMC5347536	*Homo sapiens*
SRS819740	PMC5347536	*Homo sapiens*
SRS694250	PMC5347536	*Homo sapiens*
SRS696219	PMC5347536	*Homo sapiens*
SRS693975	PMC5347536	*Homo sapiens*
SRS693903	PMC5347536	*Homo sapiens*
SRS693408	PMC5347536	*Homo sapiens*
SRS819741	PMC5347536	*Homo sapiens*
SRS693946	PMC5347536	*Homo sapiens*
SRS693974	PMC5347536	*Homo sapiens*
SRS693574	PMC5347536	*Homo sapiens*
SRS819715	PMC5347536	*Homo sapiens*
ERS040108	PMC4966634	*Homo sapiens*
ERS010154	PMC4966634	*Homo sapiens*
ERS123116	PMC4966634	*Homo sapiens*
ERS164689	PMC4966634	*Homo sapiens*
ERS164670	PMC4966634	*Homo sapiens*
ERS164692	PMC4966634	*Homo sapiens*

ERS014174	PMC4966634	*Homo sapiens*
ERS014175	PMC4966634	*Homo sapiens*
ERS014176	PMC4966634	*Homo sapiens*
ERS040110	PMC4966634	*Homo sapiens*
ERS164662	PMC4966634	*Homo sapiens*
ERS164674	PMC4966634	*Homo sapiens*
ERS014177	PMC4966634	*Homo sapiens*
ERS014179	PMC4966634	*Homo sapiens*
ERS055897	PMC4966634	*Homo sapiens*
ERS123117	PMC4966634	*Homo sapiens*
ERS164691	PMC4966634	*Homo sapiens*
ERS123119	PMC4966634	*Homo sapiens*
ERS025405	PMC4966634	*Homo sapiens*
ERS071835	PMC4966634	*Homo sapiens*
ERS174573	PMC4966634	*Homo sapiens*
ERS336352	PMC4966634	*Homo sapiens*
ERS224881	PMC4966634	*Homo sapiens*
ERS224901	PMC4966634	*Homo sapiens*
ERS338595	PMC4966634	*Homo sapiens*
ERS338597	PMC4966634	*Homo sapiens*
ERS444642	PMC4966634	*Homo sapiens*
ERS338598	PMC4966634	*Homo sapiens*
ERS338599	PMC4966634	*Homo sapiens*
ERS338601	PMC4966634	*Homo sapiens*
ERS338602	PMC4966634	*Homo sapiens*
ERS338603	PMC4966634	*Homo sapiens*
ERS338605	PMC4966634	*Homo sapiens*
ERS338607	PMC4966634	*Homo sapiens*
ERS338610	PMC4966634	*Homo sapiens*
ERS338611	PMC4966634	*Homo sapiens*
ERS338612	PMC4966634	*Homo sapiens*
ERS338617	PMC4966634	*Homo sapiens*
ERS338618	PMC4966634	*Homo sapiens*
ERS338619	PMC4966634	*Homo sapiens*
ERS338621	PMC4966634	*Homo sapiens*
ERS017708	PMC6030216	*Homo sapiens*
ERS040112	PMC6030216	*Homo sapiens*
ERS040113	PMC4966634	*Homo sapiens*
ERS241435	PMC4966634	*Homo sapiens*

ERS241438	PMC4966634	*Homo sapiens*
ERS241441	PMC4966634	*Homo sapiens*
ERS403565	PMC4966634	*Homo sapiens*
ERS666267	PMC6030216	*Homo sapiens*
ERS150822	PMC4966634	*Homo sapiens*
ERS150862	PMC4966634	*Homo sapiens*
ERS012699	PMC4966634	*Homo sapiens*
ERS164683	PMC4966634	*Homo sapiens*
ERS150819	PMC4966634	*Homo sapiens*
ERS164688	PMC4966634	*Homo sapiens*
ERS012057	PMC4966634	*Homo sapiens*
ERS404153	PMC4966634	*Homo sapiens*
ERS404132	PMC4966634	*Homo sapiens*
ERS404155	PMC4966634	*Homo sapiens*
ERS404141	PMC4966634	*Homo sapiens*
ERS010042	PMC4966634	*Homo sapiens*
ERS123115	PMC4966634	*Homo sapiens*
ERS010153	PMC4966634	*Homo sapiens*
ERS040109	PMC4966634	*Homo sapiens*
ERS013104	PMC4966634	*Homo sapiens*
ERS012702	PMC4966634	*Homo sapiens*
ERS164634	PMC4966634	*Homo sapiens*
ERS012693	PMC4966634	*Homo sapiens*
ERS012695	PMC4966634	*Homo sapiens*
ERS012692	PMC4966634	*Homo sapiens*
ERS012694	PMC4966634	*Homo sapiens*
ERS164639	PMC4966634	*Homo sapiens*
ERS164684	PMC4966634	*Homo sapiens*
ERS164678	PMC4966634	*Homo sapiens*
ERS336392	PMC4966634	*Homo sapiens*
ERS143422	PMC4966634	*Homo sapiens*
ERS143517	PMC4966634	*Homo sapiens*
ERS174558	PMC4966634	*Homo sapiens*
ERS040115	PMC4966634	*Homo sapiens*
ERS040116	PMC4966634	*Homo sapiens*
ERS040117	PMC4966634	*Homo sapiens*
ERS123120	PMC4966634	*Homo sapiens*
ERS123124	PMC4966634	*Homo sapiens*
ERS403520	PMC4966634	*Homo sapiens*

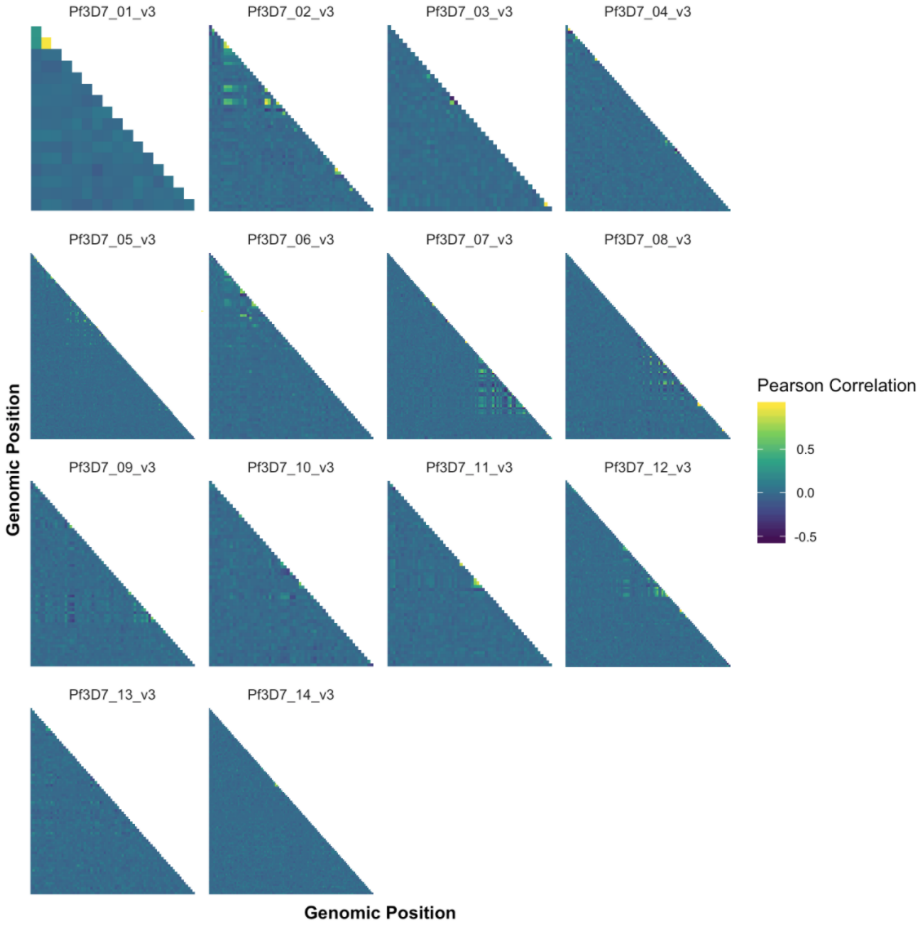
ERS403525	PMC4966634	*Homo sapiens*
ERS792599	PMC6030216	*Homo sapiens*
ERS792600	PMC6030216	*Homo sapiens*
ERS792602	PMC6030216	*Homo sapiens*
ERS792603	PMC6030216	*Homo sapiens*
ERS792610	PMC6030216	*Homo sapiens*
ERS792611	PMC6030216	*Homo sapiens*
ERS792618	PMC6030216	*Homo sapiens*
ERS792620	PMC6030216	*Homo sapiens*
ERS792621	PMC6030216	*Homo sapiens*
ERS792624	PMC6030216	*Homo sapiens*
ERS792626	PMC6030216	*Homo sapiens*
ERS792627	PMC6030216	*Homo sapiens*
ERS792629	PMC6030216	*Homo sapiens*
ERS792630	PMC6030216	*Homo sapiens*
ERS792631	PMC6030216	*Homo sapiens*
ERS792632	PMC6030216	*Homo sapiens*
ERS792633	PMC6030216	*Homo sapiens*
ERS792634	PMC6030216	*Homo sapiens*
ERS792636	PMC6030216	*Homo sapiens*
ERS792637	PMC6030216	*Homo sapiens*
ERS792643	PMC6030216	*Homo sapiens*
ERS792644	PMC6030216	*Homo sapiens*
ERS792648	PMC6030216	*Homo sapiens*
ERS792649	PMC6030216	*Homo sapiens*
ERS792650	PMC6030216	*Homo sapiens*
ERS792651	PMC6030216	*Homo sapiens*
ERS792652	PMC6030216	*Homo sapiens*
ERS989898	PMC6030216	*Homo sapiens*
ERS989900	PMC6030216	*Homo sapiens*
ERS989872	PMC6030216	*Homo sapiens*
ERS989873	PMC6030216	*Homo sapiens*
ERS989833	PMC6030216	*Homo sapiens*
ERS989874	PMC6030216	*Homo sapiens*
ERS989875	PMC6030216	*Homo sapiens*
ERS989876	PMC6030216	*Homo sapiens*
ERS989878	PMC6030216	*Homo sapiens*
ERS989881	PMC6030216	*Homo sapiens*
ERS989882	PMC6030216	*Homo sapiens*

ERS989883	PMC6030216	*Homo sapiens*
ERS989902	PMC6030216	*Homo sapiens*
ERS989903	PMC6030216	*Homo sapiens*
ERS989919	PMC6030216	*Homo sapiens*
ERS989897	PMC6030216	*Homo sapiens*
ERS989917	PMC6030216	*Homo sapiens*
ERS989907	PMC6030216	*Homo sapiens*
ERS989908	PMC6030216	*Homo sapiens*
ERS347714	PMC4966634	*Homo sapiens*
ERS174628	PMC4966634	*Homo sapiens*
ERS347479	PMC4966634	*Homo sapiens*
ERS386739	PMC4966634	*Homo sapiens*
ERS403553	PMC4966634	*Homo sapiens*
SRS693583	PMC5347536	*Homo sapiens*
SRS693904	PMC5347536	*Homo sapiens*
SRS694226	PMC5347536	*Homo sapiens*
SRS693407	PMC5347536	*Homo sapiens*
SRS694253	PMC5347536	*Homo sapiens*
SRS693980	PMC5347536	*Homo sapiens*
SRS693269	PMC5347536	*Homo sapiens*
SRS693945	PMC5347536	*Homo sapiens*
SRS693468	PMC5347536	*Homo sapiens*
SRS693573	PMC5347536	*Homo sapiens*
SRS693954	PMC5347536	*Homo sapiens*
SRS693579	PMC5347536	*Homo sapiens*
SRS693360	PMC5347536	*Homo sapiens*
SRS693970	PMC5347536	*Homo sapiens*
SRS693971	PMC5347536	*Homo sapiens*
SRS693956	PMC5347536	*Homo sapiens*
SRS693944	PMC5347536	*Homo sapiens*
SRS693912	PMC5347536	*Homo sapiens*
SRS693909	PMC5347536	*Homo sapiens*
SRS693570	PMC5347536	*Homo sapiens*
SRS1607662	PMC5068322	*Homo sapiens*
SRS417747	PMC3836732	*Homo sapiens*
SRS363193	PMC3435244	*Homo sapiens*
SRS363192	PMC3435244	*Homo sapiens*
SRS363171	PMC3435244	*Homo sapiens*
SRS363191	PMC3435244	*Homo sapiens*

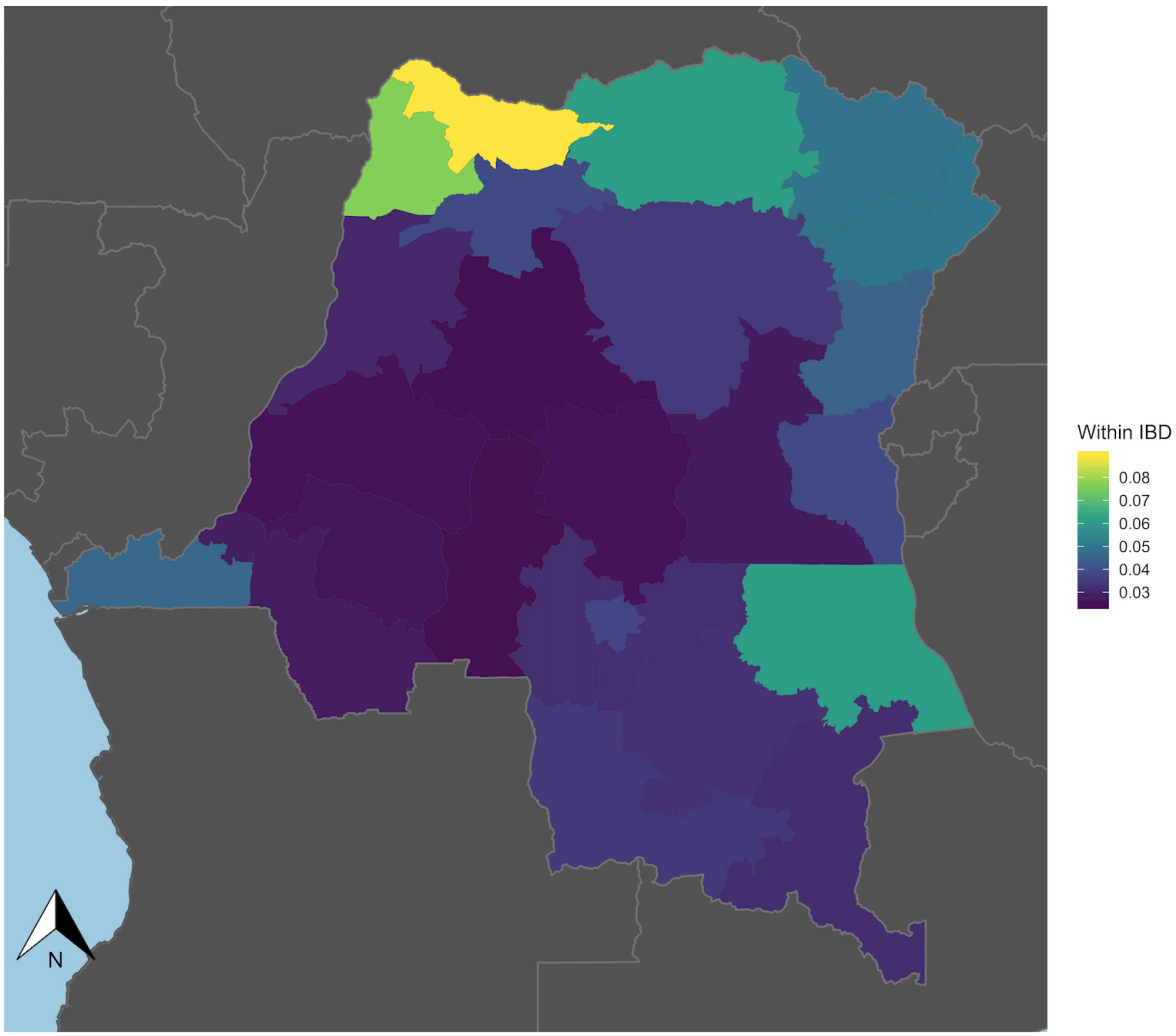
SRS363190	PMC3435244	*Homo sapiens*
ERS564521	PMID: 30668735	*Homo sapiens*
ERS564522	PMID: 30668735	*Homo sapiens*
ERS564523	PMID: 30668735	*Homo sapiens*
ERS564524	PMID: 30668735	*Homo sapiens*
ERS666222	PMID: 30668735	*Homo sapiens*
ERS666223	PMID: 30668735	*Homo sapiens*
ERS666224	PMID: 30668735	*Homo sapiens*
ERS666225	PMID: 30668735	*Homo sapiens*
ERS666226	PMID: 30668735	*Homo sapiens*
ERS666227	PMID: 30668735	*Homo sapiens*
ERS666229	PMID: 30668735	*Homo sapiens*
ERS666236	PMID: 30668735	*Homo sapiens*
ERS666238	PMID: 30668735	*Homo sapiens*
ERS666239	PMID: 30668735	*Homo sapiens*
ERS666242	PMID: 30668735	*Homo sapiens*
ERS666243	PMID: 30668735	*Homo sapiens*
ERS666244	PMID: 30668735	*Homo sapiens*
ERS666246	PMID: 30668735	*Homo sapiens*
ERS666248	PMID: 30668735	*Homo sapiens*
ERS666253	PMID: 30668735	*Homo sapiens*
ERS666255	PMID: 30668735	*Homo sapiens*
ERS666260	PMID: 30668735	*Homo sapiens*
ERS666262	PMID: 30668735	*Homo sapiens*
ERS666264	PMID: 30668735	*Homo sapiens*
DRS000258	PMC3759362	*P. cynomolgi* (Lab)
ERS001838	PMC5500898	*P. cynomolgi* (Lab)
ERS023609	PMC5500898	*P. cynomolgi* (Lab)

Appendix 3.2 Table 1 – Publicly Available Next-Generation Sequences Used in this Dissertation: Sequencing data was downloaded from the European Nucleotide Agency. The study citation (PMCID or PMC code) is provided for each isolate as well as the host (*Homo sapiens*: Homo sapiens, *Pan troglodytes*: Pan_troglodytes, *Gorilla gorilla*: Gorilla_gorilla). In addition, the *P. cynomolgi* lab strains are also indicated (*P. cynomolgi* Lab).

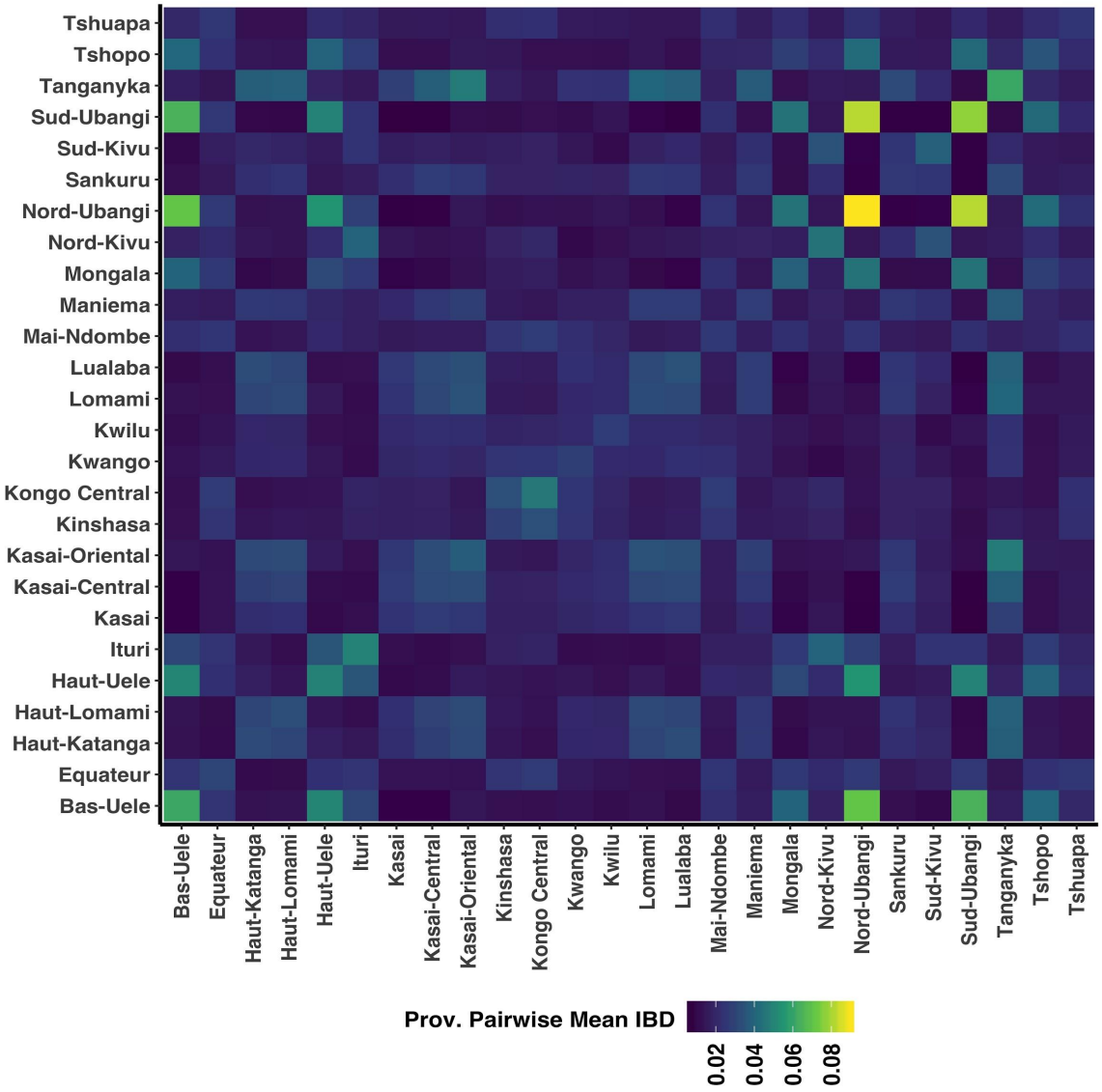
APPENDIX 4.1: SUPPLEMENT TO TRACING THE GENETIC RELATEDNESS OF PLASMODIUM FALCIPARUM IN THE DEMOCRATIC REPUBLIC OF THE CONGO ACROSS SPACE



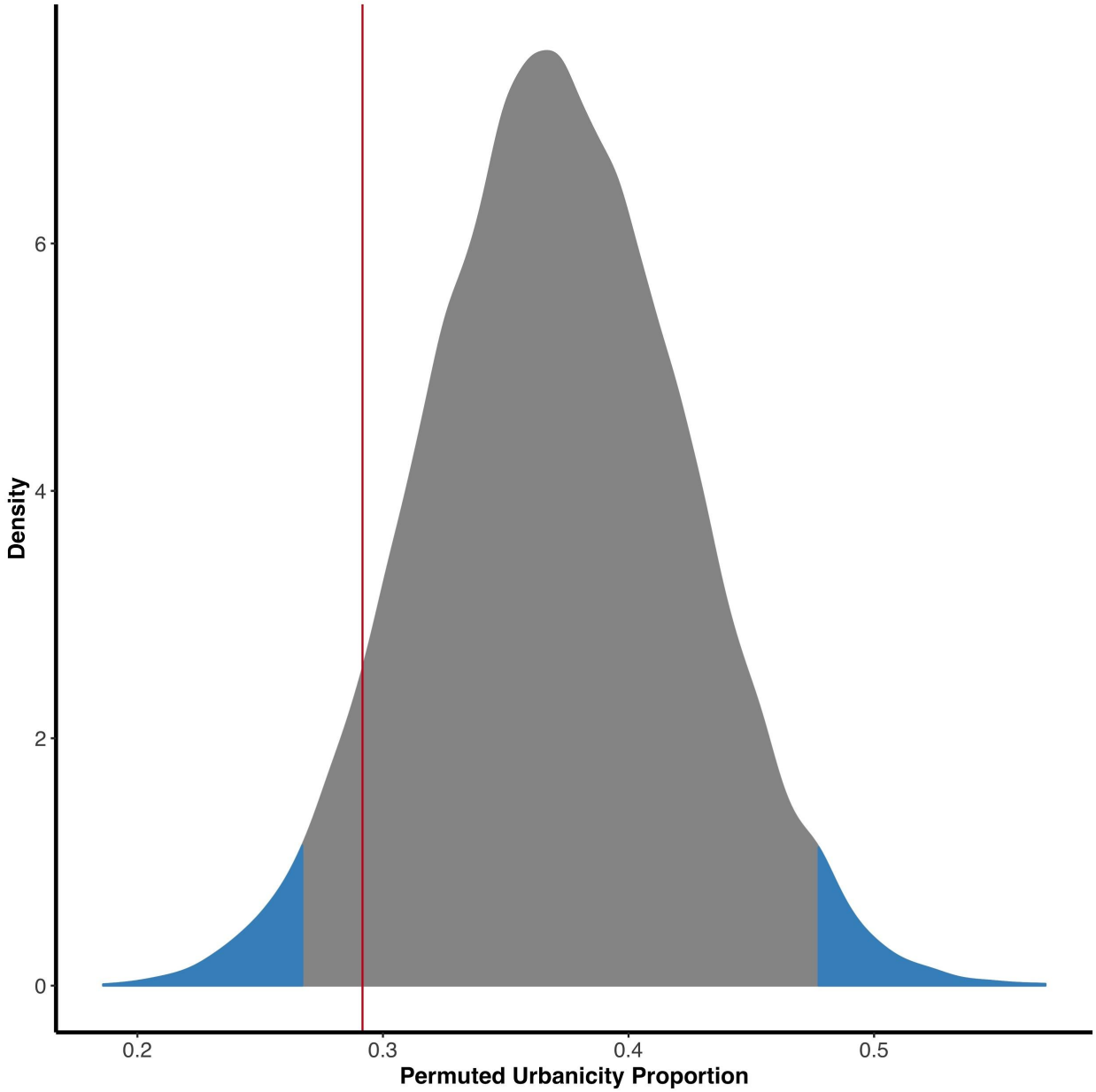
Appendix 4.1 Figure 1 - Genetic Autocorrelation among Loci: For each locus, the genetic autocorrelation is shown across the fourteen nuclear chromosomes. Most sites exhibit no autocorrelation, which suggests that the vast majority of the loci are relatively independent.



Appendix 4.1 Figure 2 - Mean Within-Province IBD: Within-province IBD is indicated on a purple-yellow spectrum with little spatial structure that can be visualized. Overall, the within-province IBD variance was low.



Appendix 4.1 Figure 3 - Between province pairwise IBD: The pairwise mean IBD between-province is indicated by a purple-yellow spectrum. Overall, between-province IBD was low and did not exhibit any strong spatial patterns.



Appendix 4.1 Figure 4 - Urbanicity Permutation Test Distribution: Shown is the distribution of the 10,000 iterations of permuted urbanicity proportions from the recoded DRC data (130/351 urban clusters). Values less than the 2.75th percentile and greater than the 97.5th percentile are shaded in blue. The observed proportion of urban clusters among the highly related pairs is indicated by a red line. Although not statistically significant, there appears to be fewer urban samples among the highly related pairs than is expected under complete independence. This suggests that urban areas are connected more frequently than would be expected (i.e. fewer are needed to make the same number of pairs).

Dist. Cat.	N	Min.	25th Perc.	Median	Mean	Std. Dev.	75th Perc.	Max.
GC	43	25.12	377.05	813.43	802.3	476.25	1274.6	1568.7
Road	43	51.31	693.28	1364.73	1385.87	817.54	2108.19	3019.48
River	43	32.88	581.9	1195.04	1149.78	638.05	1745.53	2097.45

Appendix 4.1 Table 1 - Between Cluster Highly Related Pairs Pairwise Geographical Distances: The distribution of geographic distances (km) among the highly related pairs between clusters is summarized as the number of comparisons (N), the minimum distance (Min.), the 25th and 75th percentile, median, mean, standard deviation (Std. Dev.), and maximum (Max.) for each distance category.