COMPARATIVE GENOMICS OF APE *PLASMODIUM* PARASITES REVEALS KEY EVOLUTIONARY EVENTS LEADING TO HUMAN MALARIA

Sesh A. Sundararaman

A DISSERTATION

in

Cell and Molecular Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

Beatrice H. Hahn
Professor of Medicine

Co-Supervisor of Dissertation

Frederic D. Bushman
Professor of Microbiology

Graduate Group Chairperson

Daniel S. Kessler
Associate Professor of Cell and Developmental Biology

Dissertation Committee

Robert W. Doms, Chair and Pathologist in Chief, Children's Hospital of Philadelphia

Dustin Brisson, Associate Professor of Biology

David S. Roos, E. Otis Kendall Professor of Biology

Sarah A. Tishkoff, David and Lyn Silfen University Professor

COMPARATIVE GENOMICS OF APE *PLASMODIUM* PARASITES REVEALS KEY EVOLUTIONARY EVENTS LEADING TO HUMAN MALARIA

Sesh A. Sundararaman

DEDICATION

Dedicated to the memories of Wiley A. Green and Raghavachari Desikan

ACKNOWLEDGMENT

I would like to thank Beatrice Hahn and Rick Bushman for their support and guidance over the last five years. They gave me the freedom to explore and learn while keeping me focused and refining my scientific thinking. I thank my undergraduate advisors Claire Ting and Manuel Llinàs starting me on this journey. I also thank my committee members, Bob Doms, David Roos, Sarah Tishkoff, and Dustin Brisson for their support of both my research and my development as a scientist.

This work could not have been completed without the help of many others in both the Hahn and Bushman labs and beyond. Kyle Bittinger, Nirav Malani, Scott Sherrill-Mix, and Aubrey Bailey taught me the fundamentals of coding in R, perl and python. Weimin Liu, Yingying Li, Gerald Learn, Shilpa Iyer, Fred Bibollet-Ruche, introduced me to the techniques that they have refined in their time in the Hahn lab. I was lucky to have so many collaborators including Dorothy Loy, Ahidjo Ayouba, Jordan Malenke, Katharina Shaw, Erik Clarke, Stephanie Seifert, Dan Larremore, and Will Proto. Lindsey Plenderleith and Paul Sharp were a constant source of guidance. Lindsey was also a driving force in the assembly and analysis of the ape *Laverania* genomes. I also thank the Bushman, Hahn, Shaw, and Bar labs, especially Ben Scheinfeld, Tuhina Srivastava, Frances Male, Ranjit Warrier, Ronnie Russell, Eunlim Kim, Hannah Barbian, Marcus Gondim, Nick Parrish, Erica Parrish, Patricia Crystal, Shivani Sethi, Arwa Abbas, Brendan Kelly, Casey Hofstaeder, Young Hwang, Andrew Smith, Ted Kreider, Abby Lauder, Alice Laughlin, Alexandra Bryson, and Anatoly Dryga.

I especially thank my family and friends for their constant support throughout my life. Without the interest in science and learning instilled in me by my parents, I would not have achieved half of what I have today. Finally, Charlie, my constant canine companion. He is the best dog that I could ever wish for.

ABSTRACT

COMPARATIVE GENOMICS OF APE *PLASMODIUM* PARASITES REVEALS KEY EVOLUTIONARY EVENTS LEADING TO HUMAN MALARIA

African great apes are infected with at least six species of *P. falciparum*-like parasites, including the direct ancestor of *P. falciparum*. Comparative studies of these parasites and *P. falciparum* (collectively termed the *Laverania* subgenus) will provide insight into the evolutionary origins of *P. falciparum* and identify genetic features that influence host tropism. Here we show that ape *Laverania* parasites do not sperve as a recurrent source of human malaria and use novel enrichment techniques to derive near full-length genomes of close and distant relatives of *P. falciparum*. Using a combination of single template amplification and deep sequencing, we observe no evidence of ape *Laverania* infections in forest dwelling humans in Cameroon. This result supports previous findings that ape *Laverania* parasites are host specific and have successfully colonized humans only once. To understand the determinants of host specificity and identify genetic characteristics unique to *P. falciparum*, we develop a novel method for selective enrichment of *Plasmodium* DNA from sub-microscopically infected whole blood samples. We use this technique to enrich for *Laverania* genomic DNA from chimpanzee blood samples and assemble near full length genomes for both close (*P. reichenowi*) and distant (*P. gaboni*) relatives of *P. falciparum*. Comparative analyses of these genomes to *P. falciparum* identify features that are conserved across the *Laverania* subgenus, including the expansion of the FIKK kinases and the presence of *var*-like multigene families in all *Laverania* species. Our analyses also identify genetic features that are unique to *P. falciparum*, such as a very low within-species diversity and a complex

evolutionary history of the essential invasion genes *RH5* and *CyRPA*. This dissertation lays the groundwork for future comparative analyses of the *Laverania* subgenus including population genomic analyses of ape parasites and comparisons of *P. falciparum* to its ancestor, *P. praefalciparum*.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 – Malaria Parasites of the Great Apes

1.1 The Epidemiology and Biology of Malaria

Malaria, caused by parasites of the genus *Plasmodium*, is responsible for hundreds of millions of cases and over 550,000 deaths each year (*1*). An estimated 3.3 billion people across the globe are at risk of infection (*1*). The disease has a disproportion burden on young children who account for over 78% of total malaria related deaths (*1*). While the majority of malaria related deaths are caused by *P. falciparum* (*2*), the predominant malaria species in sub-Saharan Africa, studies suggest that *P. vivax*, the predominant species outside of Africa, can also cause severe disease (*3*). Recently, a third parasite has emerged as a public health threat. *P. knowlesi*, a zoonotic parasite of macaque origin, is now known to cause of hundreds of severe malaria cases each year (*4*, *5*).

The lifecycle of *Plasmodium* parasites is complex, involving both a vertebrate and mosquito host. The parasites, injected into the host by a mosquito vector, initially infect hepatocytes before being released into the blood stream. There they infect erythrocytes to initiate a cyclic blood stage infection. During this stage, some parasites differentiate into male and female gametocytes, which are taken up by a mosquito vector. They undergo sexual reproduction in the vector's midgut and develop into an ookinete, which migrates through the midgut wall and becomes an oocyst. Finally, the oocyst develops into numerous sporozoites. These travel to the mosquito salivary glands to be inoculated into a new vertebrate host.

While malaria infections in humans have been recognized and treated for thousands of years, efforts to decrease the burden of disease have only been partially successful. *P. falciparum* has shown the capacity to become resistant to all known antimalarial compounds, including the current first line treatment, artemisinin (*6*, *7*).

Recent vaccine trials have shown some promise, however the current best candidate yielded only a 55% decrease in clinical malaria during the first year post vaccination (*8*). Thus, novel drug and vaccine development is essential to future malaria control efforts.

The identification of drug and vaccine targets in malaria has been slowed by a number of factors. The *Plasmodium* proteome is highly divergent from all other eukaryotes sequenced to date. Of the 5,777 genes in the 22.6 megabase pair *P. falciparum* genome, 60% do not share sufficient homology to genes in other organisms to allow for functional assignment (*9*). Moreover, genetic manipulation of *Plasmodium* species is difficult and time consuming. Successful transfection of *P. falciparum* typically requires months, and knockdown studies are hindered by the absence of the RNAi pathway (*10*). Studies of *P. vivax* are nearly impossible, as the parasite cannot be cultured *in vitro* (*11*). While recent developments in *Plasmodium* genetics have improved the ability to study these parasites, more than 35% of genes of the *P. falciparum* genome remain annotated as unknown function (PlasmoDB).

1.2 *Plasmodium* Species in African Great Apes

Plasmodium species in our closest ape relatives were first identified in the early 1900s. In 1920, Edward Reichenow described both *P. vivax*-like and *P. falciparum*-like parasites in chimpanzees and gorillas (*12*). While it was initially unclear whether these parasites represented novel species or were the same as those in humans, they were eventually categorized as separate species with chimpanzee and gorilla *P. vivax*-like parasites denoted *P. schwetzi* and chimpanzee *P. falciparum*-like parasites denoted *P. reichenowi* (the *P. falciparum*-like parasites from gorillas were rarely studied outside of the original description by Reichenow). Epidemiologic studies showed that these parasites were

widespread in Africa, identifying them in apes from Liberia and Sierra Leone in the northwest to the lower Republic of Congo in the south and Lake Edward in the east (*13*).

Little data exists on the clinical course of *P. schwetzi* and *P. reichenowi* infections in great apes. Apes naturally infected with *P. schwetzi* tended to have low parasitemia, and while the parasite burden increased in splenectomized chimpanzees, there was still no evidence for clinical symptoms in the few studies that were performed (*13*). Even less data exists for *P. reichenowi*, although splenectomized chimpanzees with high parasitemia infections (360,000-500,000 parasites/cu. mm blood) did have periodic fevers of up to 103 °F (*14*, *15*). While these results may indicate that *P. reichenowi* and *P. schwetzi* are less likely to cause serious clinical disease in their natural hosts, it is also possible that the infected chimpanzees had some level of naturally acquired immunity, similar to that observed in older humans in malaria endemic areas (*16*), and therefore suffered fewer symptoms than would an infant or naïve host.

Given the close evolutionary relationship between chimpanzees and humans, and the resemblance of *P. reichenowi* and *P. schwetzi* to *P. falciparum* and *P. vivax*, attempts to infect humans with great ape *Plasmodium* species were made on multiple occasions. *P. schwetzi* was successfully transferred to humans, both by mosquitoes and inoculation with infected chimpanzee blood, and produced symptomatic infections (*13*). Importantly, while *P. schwetzi* successfully infected multiple volunteers of European descent, it did not yield infections in the one African American tested. In humans, *P. vivax* requires the Duffy antigen receptor for chemokines (DARC) to invade red blood cells (*17*). A vast majority of people of West African descent carry a mutation in the DARC promoter (the Duffy negative phenotype) which abrogates DARC expression on red blood cells, rendering them resistant to *P. vivax* invasion (*18*). It is possible that *P. schwetzi*, like *P. vivax*, requires DARC for red blood cell invasion and was therefore

unable to infect this African American volunteer who likely lacked the receptor on his/her red blood cells.

Unlike *P. schwetzi*, *P. reichenowi* appears very host specific. Multiple attempts to infect humans with the parasite, either by intravenous or subcutaneous injection, failed (*13*). Interestingly, at least some strains of *P. falciparum* can produce transient parasitemia in intact (non-splenectomized) chimpanzees and persistent parasitemia after splenectomy (*13*). Together, these data suggest that, while *P. falciparum*-like parasites are more host specific than those related to *P. vivax*, this specificity may not be complete and is dependent on either the host, the specific parasite strain used, or a combination of the two.

1.3 The Origin of Human Malaria

While numerous studies of great ape *Plasmodium* species were carried out between the 1920s and 1970s, the parasites were all but forgotten for the next 40 years. Only a single isolate of *P. reichenowi* has been maintained, through cryopreservation and passage in captive splenectomized chimpanzees at the CDC. Phylogenetic analyses that included sequences from this isolate confirmed its close relationship to *P. falciparum* and showed that the two formed a clade that is distinct from all other known *Plasmodium* species (*19*). Given these data, researchers hypothesized that *P. falciparum* and *P. reichenowi* represented sister lineages that had diverged and co-evolved with their respective hosts.

Historically, two separate hypotheses were proposed for the origin of *P. vivax*. The first, supported by morphological similarities between *P. vivax* and multiple Asian monkey parasites, proposed that human *P. vivax* was initially transmitted to humans from Asian monkeys after early humans had migrated out of Africa (*20*). The second

suggested that *P. vivax* had originated in Africa, where long-term selective pressures by the parasite had led to the emergence of fixation of the Duffy negative phenotype in West African populations (*21*, *22*). Later molecular characterizations and phylogenetic analyses of human *P. vivax* indicated that the parasite falls within the radiation of Asian monkey *Plasmodium* species (*23*). This suggests that *P. vivax* emerged during or after the divergence of the Asian monkey parasites. As Asian monkey *Plasmodium* species are thought to have radiated along with their hosts in Asia (*24*), these phylogenetic analyses supported the hypothesis that *P. vivax* had originated in Asia.

In 2009, Prugnolle and colleagues showed *Plasmodium* DNA could be detected in the fecal samples of wild living chimpanzees and gorillas (*25*). This discovery brought renewed life to the studies of great ape *Plasmodium*. As fecal detection is non-invasive, it could be applied to wild living great ape populations where blood sampling would be both impractical and unethical. Studies found that *P. falciparum*-like parasites were widespread among all four chimpanzee subspecies and western gorillas (*25*-*31*). These parasites were common in wild apes, with prevalence rates, estimated from feces, surpassing 50% at some field sites (*26*). Parasite mitochondrial sequences also appeared to cluster into multiple distinct lineages, suggesting that more than one species of *P. falciparum*-like parasite might be present in wild ape populations (*25*-*31*).

The identification of widespread *Plasmodium* infections in both chimpanzees and gorillas brought renewed interest to the question of the origins of human malaria. Analyzing over 1,100 mitochondrial, nuclear, and apicoplast sequences from wild living apes, our lab identified at least 9 distinct *Plasmodium* lineages, six of which were closely related to *P. falciparum* (*26*). Of these six lineages, three were found only in chimpanzees and the other three only in gorillas. Phylogenetic analyses of these species and *P. falciparum* showed that all extant strains of *P. falciparum* formed a single

monophyletic lineage that emerged from within the radiation of a single gorilla parasite species (*26*). These results indicated that *P. falciparum* emerged from gorillas, and that all extant *P. falciparum* strains may have originated from a single cross-species transmission event (*26*).

As the six species topology of these *P. falciparum*-like parasites was observed at all analyzed loci, including those from apicoplast, mitochondrial and nuclear genes, they have tentatively been classified as distinct species (*32*). The gorilla parasites, in order of increasing evolutionary distance to *P. falciparum*, are termed *P. praefalciparum*, *P. blacklocki*, and *P. adleri*, while the chimpanzee parasites are termed *P. reichenowi*, *P. billcollinsi*, and *P. gaboni*. In recognition of their dissimilarity to other *Plasmodium* species (*30*, *32*), we refer to these six ape species and *P. falciparum* as the *Laverania* subgenus, a term originally suggested by Bray (*33*).

While phylogenetic analyses showed that the predominant global lineages of *P. falciparum* were derived from a single cross-species transmission event from gorillas, we could not rule out the possibility of local transmission between apes and humans. This is the case for HIV-1, which also originated in African great apes. While 99% of HIV-1 infections are derived from a single lineage (group M), the virus has been transmitted from apes to humans at least three more times, giving rise to groups N, O, and P (*34*). Given the close phylogenetic relationship and likely morphological similarity between ape *Laverania* parasites and *P. falciparum*, zoonotic infections would be unlikely to be detected in humans. In chapter 2 of this dissertation, we develop new methods to screen for ape *Laverania* parasites in humans, providing the first evidence that ape *Laverania* parasites are not a significant source of human malaria infections.

1.4 Elucidating the Steps to the Emergence of *P. falciparum*

The zoonotic origin of *P. falciparum* and lack of additional cross-species transmissions of ape *Laverania* parasites suggests that *P. falciparum* is uniquely suited to infect humans. While targeted sequencing is useful for understanding the evolutionary history of an organism, the identification of genetic features that are unique to *P. falciparum* requires genome level analyses. The identification of six distinct *P. falciparum*-like species in great apes, none of which are recurrently transmitted to humans, provides a unique opportunity for comparative genomics. However, these studies have been hindered by a lack of samples suitable for whole genome sequencing.

Next generation sequencing technologies have increased the pace of whole genome sequencing by providing a rapid, high throughput, and relatively cheap method for shotgun sequencing. In spite of this, cost effective next generation sequencing of larger genomes still requires samples that are enriched in the DNA of the organism of interest (*35*, *36*). A sample consisting of only 1% *Plasmodium* DNA requires 100 fold more sequencing to achieve the same depth of genome coverage as a sample containing 100% *Plasmodium* DNA. This increase in sequencing requirements increases both the upfront cost and downstream analysis time.

Previous genomic studies of *Plasmodium* have used a variety of methods to enrich for *Plasmodium* DNA prior to high throughput sequencing. These include short term or long term *in vitro* culture, purification of infected erythrocytes from fresh blood, selective digestion of non-*Plasmodium* DNA, and *Plasmodium* DNA capture (*35*-*38*). While we were able to obtain *Laverania* positive chimpanzee blood samples from a sanctuary in Cameroon, we found the majority of these methods to be impractical given the source, storage conditions, and parasitemia of the available samples. In chapter 3 we develop a novel method for the enrichment of *Plasmodium* DNA, termed selective

whole genome amplification (SWGA) (*39*), from samples containing as little as 0.00081% *Plasmodium* DNA. Applying this method to three chimpanzee blood samples, we generate near full-length genomes of chimpanzee parasites that represent both close (*P. reichenowi*) and distant (*P. gaboni*) relatives of *P. falciparum*. Using these genomes we identify genetic characteristics that are shared across the *Laverania* subgenus, but also features that are unique to the ancestry of *P. falciparum*. These findings are discussed in chapters 3 and 4.

Selective whole genome amplification is an effective method for enrichment of target DNA from contaminating background DNA. The method itself is simple, requiring only a set of short primers and the phi29 DNA polymerase. SWGA relies on differences in short DNA motif frequency between the target and background genomes, and thus should be easily adapted to many situations (*39*). Prior to this dissertation, however, the identification of SWGA primer sets was time consuming and required a large amount of user input (*39*). In chapter 5 we present a program, *swga*, that automates the process of designing primer sets for selective whole genome amplification. The program also calculates a number of statistics for each primer set, allowing sets to be compared *in silico*. *swga* is built on a modular architecture. Additional modules can be easily added to the program as we improve our understanding of what defines a good SWGA primer set. By increasing the speed and ease of SWGA primer design, *swga* will make selective whole genome amplification accessible to a larger number of genetics and genomics researchers.

CHAPTER 2 - *Plasmodium falciparum*-like Parasites Infecting Wild African Apes Do Not Represent a Recurrent Source of Human Malaria

Sesh A. Sundararaman, Weimin Liu, Brandon F. Keele, Gerald H. Learn, Kyle Bittinger, Fatima Mouacha, Steve Ahuka-Mundeke, Magnus Manske, Scott Sherrill-Mix, Yingying Li, Jordan A. Malenke, Eric Delaporte, Christian Laurent, Eitel Mpoudi Ngole, Dominic P. Kwiatkowski, George M. Shaw, Julian C. Rayner, Martine Peeters, Paul M. Sharp, Frederic D. Bushman, Beatrice H. Hahn

Originally published in PNAS 110 (75):7020-7025.

Supplemental data are available at

http://www.pnas.org/content/suppl/2013/04/05/1305201110.DCSupplemental/sapp.pdf

Weimin Liu, Frederic D. Bushman, Beatrice H. Hahn, and I designed the study. I led the computational analysis of 454 sequencing data with assistance from Kyle Bittinger and Scott Sherrill-Mix. Weimin Liu, Brandon F. Keele, Gerald H. Learn, Paul M. Sharp, Beatrice H. Hahn, and I developed and tested the 454 amplification and sequencing protocol. Weimin Liu, Fatima Mouacha, Steve Ahuka-Mundeke, Yingying Li, and Jordan A. Malenke generated and analyzed PCR and SGA data. Magnus Manske, Eric Delaporte, Christian Laurent, Eitel Mpoudi Ngole, Dominic P. Kwiatkowski and Julian C. Rayner provided reagents and analytic tools. Weimin Liu, George M. Shaw, Julian C. Rayner, Martine Peeters, Paul M. Sharp, Frederic D. Bushman, and I wrote the paper with contributions from all authors.

2.1 Abstract

Wild-living chimpanzees and gorillas harbor a multitude of *Plasmodium* spp., including six of the subgenus *Laverania*, one of which is the progenitor of *P. falciparum*. Despite the magnitude of this natural reservoir, it is unknown whether apes represent a recurrent source of human infections. Here, we used *Plasmodium* species-specific PCR, single genome amplification (SGA) and 454 sequencing to screen humans from remote areas of Cameroon for ape *Laverania* infections. Among 1,403 blood samples, we found 1,000 to be positive for *Plasmodium* mitochondrial (mt)DNA, all of which contained human parasites as determined by sequencing and/or restriction enzyme digestion. To exclude low abundance infections, we subjected 514 samples to 454 sequencing, targeting a region of the mtDNA genome that distinguishes ape from human *Laverania* species. Developing algorithms capable of differentiating rare *Plasmodium* variants from 454 sequencing error, we identified mono- and mixed-species infections with *P. falciparum*, *P. malariae* and/or *P. ovale*. However, none of the human samples contained ape *Laverania* parasites, including the gorilla precursor of *P. falciparum*. To characterize further the diversity of *P. falciparum* in Cameroon, we used SGA to amplify 3.4 kb mtDNA fragments from 229 infected humans. Phylogenetic analysis identified 62 new variants, all of which clustered with extant *P. falciparum* strains, providing further evidence that *P. falciparum* emerged following a single gorilla-to-human transmission. Thus, unlike *P. knowlesi* infected macaques in Southeast Asia, African apes harboring *Laverania* parasites do not serve as a recurrent source of human malaria, a finding of import to ongoing control and eradication measures.

2.2 Introduction

Malaria is one of the most devastating infectious diseases of humans worldwide, with hundreds of millions of cases of clinical illness and over 650,000 deaths occurring annually (*40*). Given this enormous health burden, efforts to control and potentially eradicate this disease have become an urgent public health priority (*41*, *42*). Effective control and elimination measures require a clear understanding of the parasites, vectors as well as human and environmental factors that sustain malaria transmission. This includes a systematic evaluation of potential zoonotic reservoirs and the risk that they may pose for humans. Recently, close genetic relatives of the human malaria parasites *P. falciparum, P. ovale, P. malariae* and *P. vivax* have been identified in wild-living apes in sub-Saharan Africa (*25-27*, *29*, *30*). These parasites have been tentatively classified on the basis of their sequence relationships into a number of different species, six of which were closely related to human *P. falciparum* and placed into a separate *Plasmodium* subgenus, termed *Laverania* (*26*, *30-32*). Of these six *Laverania* species, *P. reichenowi*, *P. gaboni*, and *P. billcollinsi* were identified only in chimpanzees, while *P. adleri*, *P. blacklocki*, and *P. praefalciparum* were only found in gorillas. Moreover, *P. praefalciparum* was shown to be the immediate precursor of human *P. falciparum* (*26*). Although the *Anopheles* vectors that transmit these ape parasites have not yet been identified, the fact that a large fraction of wild-living apes is endemically infected has raised concerns that they might represent a source of recurring human infection (*25*, *32*, *43*, *44*).

In this study, we tested humans who live in remote rural areas of southern Cameroon for evidence of zoonotic *Plasmodium* infections. We specifically screened for *Laverania* infections, since these are the most abundant and widespread in resident ape populations, and since one of them, *P. praefalciparum*, has crossed the species barrier

from gorillas to humans already once (*26*). Moreover, *Laverania* parasites have been studied extensively at the molecular level, with numerous mitochondrial, apicoplast and nuclear sequences available for analyses. To detect zoonotic infections, we developed a new ape *Plasmodium* species-specific diagnostic PCR, used 454 ultra deep sequencing to determine whether humans harbored ape parasites at low abundance, and employed single genome amplification to characterize the genetic diversity of human *P. falciparum* in southern Cameroon. Our study is the first to systematically search for *Plasmodium* zoonoses in west central Africa, thus providing new insight into the host range of human and great ape parasites.

2.3 Results

Molecular Characterization of Human *Plasmodium* Infections in Rural Cameroon

Cameroon is an area of high malaria endemicity, with nearly 100% of clinical cases believed to be caused by *P. falciparum* (*40*). However, few of these infections have been molecularly characterized and the extent of parasite diversity, both at the intra- and inter-species level, is largely unknown. Studying the epidemiology and natural history of human immunodeficiency virus type 1 (HIV-1) infections in sub-Saharan Africa, we previously collected large numbers of buffy coat samples from humans native to rural Cameroon (*45*). These samples, which represent thin layers of leukocytes and platelets on the surface of sedimented erythrocytes, frequently contain *Plasmodium* DNA, since parasite infected red blood cells tend to accumulate immediately below the leukocyte layer (*46*). All samples were obtained from individuals living in close proximity to the habitat of *Laverania* infected apes (Fig. 2-1), thus providing a unique opportunity to search for zoonotic *Plasmodium* infections.

To characterize the *Plasmodium* species that commonly infect humans in rural

Cameroon, we first selected 318 buffy coat specimens from inhabitants of seven remote villages (*SI Appendix*, Table S1). These study sites were selected because of the high *Laverania* prevalence rates in chimpanzee and gorilla populations in adjacent forest regions (Fig. 2-1A). All sampled individuals lived in close proximity to ape habitats (Fig. 2-1B), and included forest dwellers, hunters, members of local pygmy tribes and individuals who lived at logging concessions. Given their lifestyles, we reasoned that at least some of the study subjects were exposed to ape *Plasmodium* infected *Anopheles* mosquitoes. To examine whether such exposures had resulted in parasite transmission, we screened buffy coat DNA for ape parasites by conventional PCR. Using primers previously shown to amplify ape *Laverania* parasites with high sensitivity and specificity (*26*), we targeted a 939 bp region (*cytb* gene) of the *Plasmodium* mitochondrial (mt) DNA genome (Fig. 2-2). This analysis identified 194 of the 318 blood samples to be PCR positive (61%), all of which contained human parasites as determined by direct sequencing: 181 samples contained *P. falciparum*, 12 samples contained *P. ovale*, and one sample contained *P. malariae* as the predominant *Plasmodium* species (*SI Appendix*, Table S1). From this experiment, we concluded that zoonotic *Laverania* infections, if they indeed occurred, were rare and unlikely to represent single species infections.

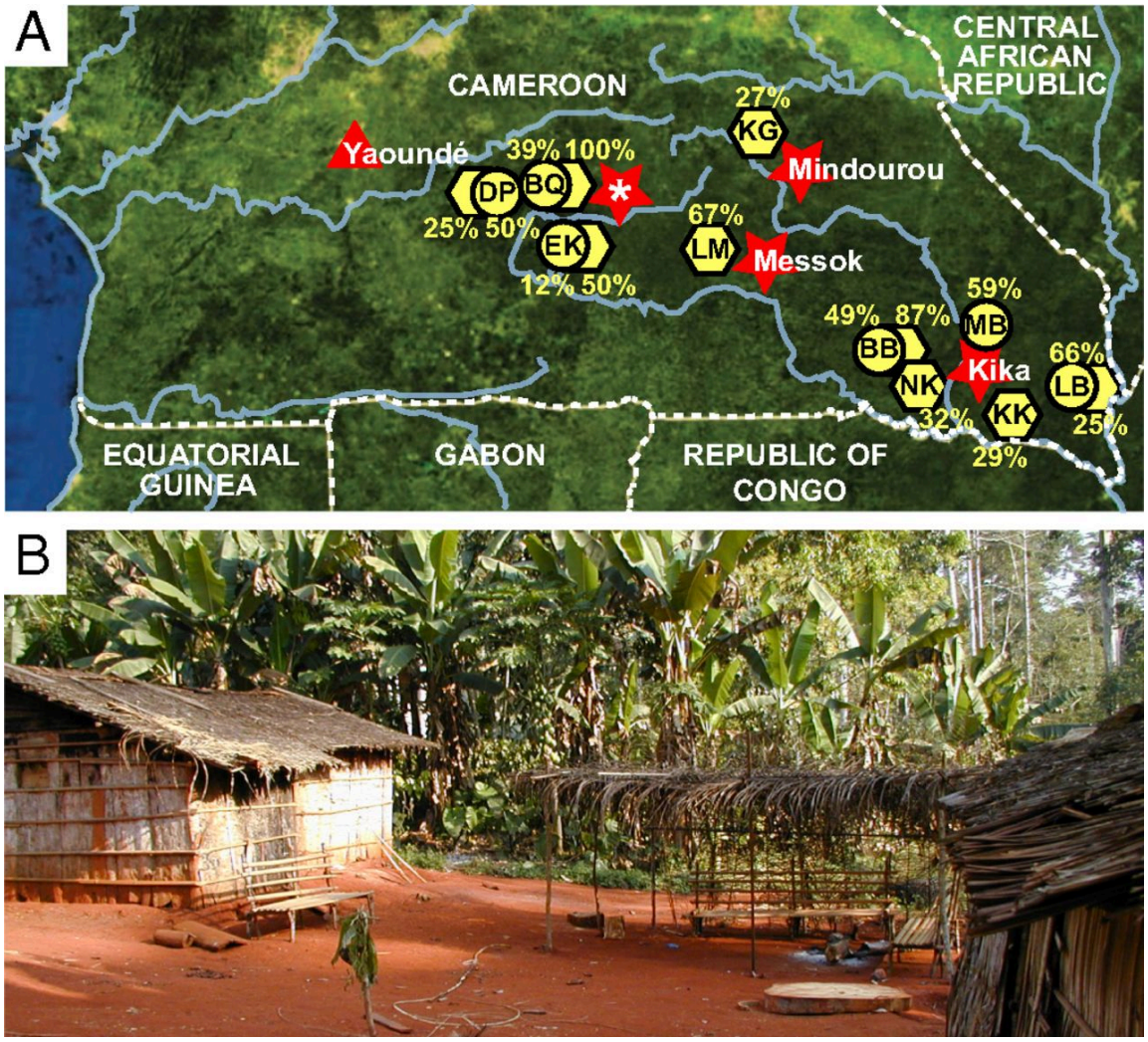

Figure 2-1. Screening of humans in rural Cameroon for zoonotic *Plasmodium* infections.

(A) Location of human study sites (red stars). Eight rural villages were selected for molecular epidemiological studies because of their proximity to wild-living chimpanzee (yellow circles) and gorilla (yellow hexagons) populations known to harbor *Plasmodium* infections at high prevalence rates. Previously estimated infection rates (*26*) are shown for the most proximal field sites (denoted by a two-letter code). Country borders, major rivers and the capital city of Yaoundé (red triangle) are also shown. A red star with asterisk highlights the location of five closely spaced villages (Mboumo, Eboumetoum,

Aviation, Nkonzu, and Kompia). **(B)** View of one rural village, depicting the close proximity of human residences and sleeping quarters to the surrounding forest (photograph credit, Bernadette Abela).

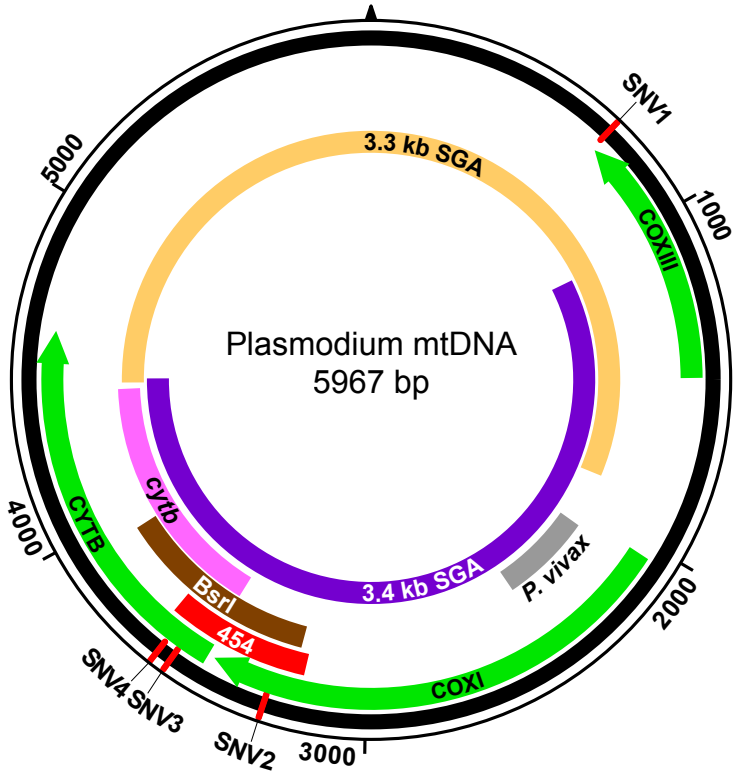

Figure 2-2. Schematic representation of the *Plasmodium* mitochondrial genome. DNA fragments amplified for diagnostic purposes (*cytb,* BsrI, *P. vivax*), 454 deep sequencing (454), and single genome amplification (mtDNA-3.3 kb; mtDNA-3.4kb) are shown in relation to cytochrome b (*cytb*), cytochrome c oxidase subunit I (*coxI*) and cytochrome c oxidase subunit III (*coxIII)* coding regions, respectively. The positions of four single nucleotide variants (SNV1-SNV4), which distinguish human *P. falciparum* from ape *Laverania* parasites, are shown in red.

A Diagnostic PCR Capable of Differentiating Human and Ape *Laverania* Species

Aligning several hundred ape and human *Plasmodium* mitochondrial genomes, we had previously noted four single nucleotide variants (SNVs) that distinguished all known *P. falciparum* strains from the six ape *Laverania* species (*26*). One of these (SNV4) comprised a BsrI restriction enzyme site (AC**T**GGN) that was present in 134 of 135 ape *Laverania* sequences, but was absent from all of 859 human *Plasmodium* sequences in the database (Fig. 2-2). To determine whether PCR amplification followed by BsrI cleavage could be used to screen human blood samples for ape *Laverania* infections, we designed primers for a ~700 bp DNA fragment that spanned the diagnostic SNV4 site (Figs. 2-2 and 2-3). Testing these primers on fecal samples from *Plasmodium* infected apes, we obtained PCR products that were all cleaved by BsrI and yielded the expected fragments for the respective ape *Plasmodium* species (Fig. 2-3A and B). In contrast, amplicons from human *P. falciparum*, *P. malariae* and *P. vivax* reference strains were not cleaved by BsrI, and although *P. ovale* amplicons were cleaved once, the resulting fragments were readily distinguishable from those of the ape parasites (not shown). BsrI cleavage products were also visible in mixtures of human and ape parasite DNAs, including in preparations that contained *P. falciparum* at a 10-fold excess (not shown). These data indicated that PCR amplification, followed by BsrI cleavage, represented a viable screening approach for zoonotic *Laverania* infections, even when ape and human parasites were present in mixed-species infections.

Using this *Plasmodium* species–specific PCR assay, we screened 1,165 buffy coat samples from villagers native to southeastern Cameroon (Fig. 2-1). For control, we also analyzed 85 samples from HIV-1 infected individuals in the capital city Yaoundé. Testing a total of 1,250 samples, we amplified BsrI-specific fragments from 872 of them (*SI Appendix*, Table S1), three of which were cleaved by BsrI (Fig. 2-3C). Two of these

samples (KI051 and EC1592) yielded PCR cleavage products consistent with *P. ovale* infection, which was confirmed by sequence analysis. The third sample (EC1041), from a child in Mboumo, yielded ape-specific PCR cleavage products of 395 bp and 316 bp, respectively (Fig. 2-3C). However, sequence analysis failed to confirm ape *Laverania* infection, identifying instead a *P. falciparum* variant that exhibited a single point mutation at the SNV4 site. This was confirmed after sequencing the entire mitochondrial genome of this variant, which contained the SNV4 point mutation, but lacked additional ape *Plasmodium* specific signatures. Thus, the Bsr1 diagnostic PCR had uncovered a rare *P. falciparum* variant whose mitochondrial sequence was identical to that of other *P. falciparum* strains, except for a single (ape-like) back mutation at the SNV4 site.

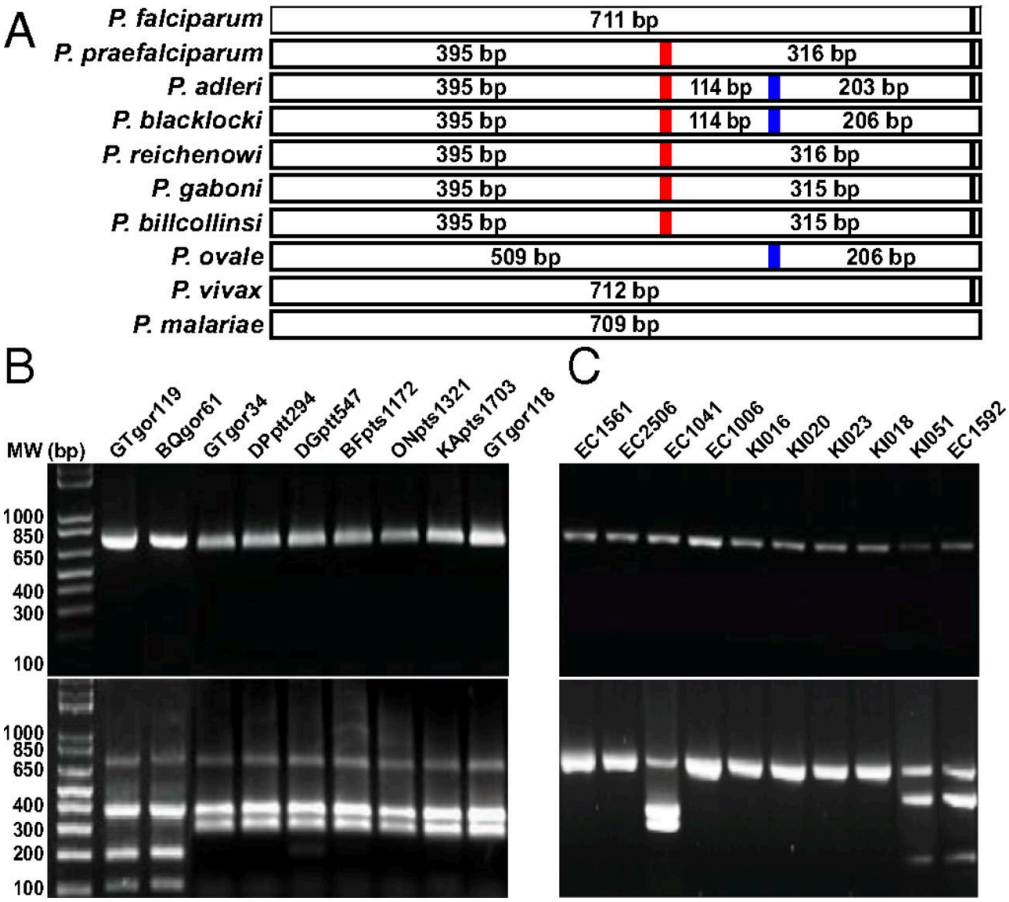

Figure 2-3. A *Plasmodium* species-specific PCR capable of differentiating human and ape *Laverania* parasites.

(A) Predicted BsrI cleavage products for different *Plasmodium* species infecting humans and apes. A red vertical line highlights a BsrI site unique to ape *Laverania* parasites. A second BsrI site found only in *P. adleri*, *P. blacklocki* and *P. ovale* is highlighted in blue. **(B)** Diagnostic PCR of ape *Laverania* infections. All *Laverania* positive ape fecal samples were PCR positive (upper panel) and yielded appropriately sized fragments upon BsrI cleavage (lower panel). **(C)** Diagnostic PCR of human *Plasmodium* infections. All *Plasmodium* positive human samples yielded appropriately sized amplicons (upper panel), with BsrI cleavage products observed for only three (see text for details).

Molecular Characterization of Human *Plasmodium* Infections by 454 Deep Sequencing

The combined results of the *cytb* and BsrI PCR screening studies indicated that the vast majority of humans in remote rural Cameroon harbored *P. falciparum,* which is known to reach very high blood titers. We thus reasoned that ape *Laverania* parasites -- if they were transmitted to humans -- would likely replicate less efficiently and represent only a minor fraction of the total parasite burden within an infected individual. To increase the likelihood of detecting such variants, we developed an ultra-deep sequencing approach, which is known to generate tens of thousands of sequences of the same genetic locus and can thus detect low abundance variants with great sensitivity (*47-49*). Specifically, we used the 454 GS FLX Titanium chemistry to sequence a 405 bp fragment of the *Plasmodium* mtDNA genome that included three of the four diagnostic SNVs (Fig. 2-2) and could thus be used to differentiate even the closest human and ape parasites (*SI Appendix*, Fig. S1).

To explore the utility of the 454 sequencing approach, we initially sequenced the combined DNA of 77 *Plasmodium* positive human buffy coat samples, which yielded 465,391 high quality reads (for details see *SI Appendix*, Supplemental Analysis). Each read was classified by determining its minimum edit distance to a large set of *Plasmodium* reference sequences (see Supplemental Analysis in *SI Appendix* and Table S2). This approach identified 458,676 reads (98.56%) to represent *P. falciparum,* 76 reads (0.02%) to represent *P. ovale,* and 6,266 reads (1.35%) to represent *P. malariae* (*SI Appendix*, Fig. S2 A-C), which was confirmed by phylogenetic analyses of select reads (*SI Appendix*, Fig. S3). A single read was classified as *P. praefalciparum*; however, closer inspection of its sequence revealed multiple indels that caused inactivating frameshift mutations as well as a substitution that was not found in any other

P. praefalciparum strain. Moreover, this read differed from the closest *P. praefalciparum* reference by 5 mutations (*SI Appendix*, Fig. S2D), but from the closest *P. falciparum* reference by 6 mutations, and contained only 2 of the 3 ape specific SNVs. We thus concluded that this read was erroneously classified as *P. praefalciparum* due to PCR and/or 454 process errors, and that there was no evidence of ape *Laverania* infection in any of the 77 *Plasmodium* infected individuals.

Identification of *Plasmodium* Multi-Species Infections by 454 Deep Sequencing

To extend the search for zoonotic *Laverania* infections, we selected an additional 438 samples for 454 sequencing, but improved the methodology. First, we inserted a 12-mer barcode into the sequencing primer to permit the computational sorting of individual samples (*50*). Second, we reversed the sequencing direction to increase the number of reads that covered at least two diagnostic SNVs *(SI Appendix,* Fig. S1). Third, we amplified samples using the lowest possible number of cycles to reduce PCR introduced errors (*SI Appendix*, Table S3). Finally, we included cloned fragments (3.4 kb) of the *P. falciparum*, *P. malariae*, and *P. ovale* mitochondrial genome (Fig. 2-2) as controls, which allowed us to perform a formal error calculation for each pyrosequencing run (for details see *SI Appendix*, Supplemental Analysis). The resulting pyrosequencing reads were sorted by sample and analyzed.

The identification of rare ape *Plasmodium* parasites necessitated a method that could differentiate true sequence changes from 454 sequencing error. We thus used a maximum likelihood based approach to determine which and how many different *Plasmodium* species were present in each barcoded human sample (for details see *SI Appendix*, Supplemental Analysis). For each sample, we generated pairwise alignments of all reads with all *Plasmodium* reference sequences and then applied a model for 454

sequencing error (*SI Appendix*, Table S4) to calculate the probability that a read was derived from a particular reference. Using this approach, we determined the *Plasmodium* species composition in all barcoded human samples. Of 437 samples, 349 contained only *P. falciparum*, one contained only *P. malariae*, and 4 contained only *P. ovale* sequences (Table 2-1). A further 61 samples contained both *P. falciparum* and *P. malariae*, 13 samples contained both *P. falciparum* and *P. ovale*, and 9 samples contained all three species (Table 2-1). Importantly, none of the human blood samples contained any of the six ape *Laverania* species, including *P. praefalciparum*. Moreover, none of the samples contained *P. vivax* sequences.

To be certain that our inability to find ape *Plasmodium* zoonoses was not due to technical limitations, we used the identical 454 methodology to amplify and deep sequence *Plasmodium* parasites from fecal samples of infected apes. Analysis of 37,644 filtered reads from two western lowland gorillas (*Gorilla gorilla gorilla*), three central chimpanzees (*Pan troglodytes troglodytes*) and one eastern chimpanzee (*Pan troglodytes schweinfurthii*) confirmed the presence of all 6 ape *Laverania* species as well as *P. vivax*-like parasites (*SI Appendix*, Fig. S4). We also characterized the proportion of humans who harbored multiple *P. falciparum* variants (*SI Appendix,* Supplemental Analysis). Multiple variant infections were detected in 10% of all subjects, with a maximum of four variants per person (*SI Appendix*, Table S5, Fig. S5). Importantly, minor variants could be identified at levels as low as 0.006% of the total parasite burden, thus providing direct evidence that our deep sequencing approach was capable of identifying very low abundance *Plasmodium* variants.

Table 2-1 Species composition of human *Plasmodium* infections in Cameroon as determined by 454 sequencing

Identified Specie (s)	Number of Samples
P. falciparum	349
P. falciparum and *P. malariae*	61
P. falciparum and *P. ovale*	13
P. falciparum, *P. malariae* and *P. ovale*	9
P. ovale	4
P. malariae	1
Total	**437**

[a] A breakdown of these *Plasmodium* species for each individual sample is shown in Figure S5

Genetic Diversity of *P. falciparum* in Rural Cameroon

Although there are over a hundred near-full-length *P. falciparum* mitochondrial DNA sequences in the database, little to nothing is known about the extent of genetic diversity of this parasite in central Africa. In particular, there are no molecularly characterized human strains from areas where wild-living apes are endemically infected with *Laverania* parasites. To characterize the *P. falciparum* variants prevalent in rural Cameroon, we selected *Plasmodium*-positive samples from seven different locations (*SI Appendix*, Table S1) and subjected them to single-genome amplification (SGA) targeting the region of the mitochondrial genome (3.4 kb) known to exhibit the greatest diversity between ape and human *Laverania* lineages (Fig. 2-2). We selected SGA rather than conventional PCR, since this method eliminates *Taq* polymerase-induced recombination as well as nucleotide misincorporations in finished sequences, and thus ensures an accurate representation of parasite variants as they exist *in vivo* (*26*, *51*). Sequencing between 1 and 8 SGA amplicons per sample, we generated a total of 684 half-genome mtDNA sequences. Phylogenetic analyses revealed that these represented 69 unique *P. falciparum* haplotypes, 62 of which had not previously been reported. Despite this

diversity, all variants grouped with previously identified *P. falciparum* sequences, forming a single well-supported clade within the radiation of *P. praefalciparum* from gorilla (Fig. 2-4). This was the case even after inclusion of a *P. falciparum* variant (EC1041, also see Fig. 2-3) that contained one of the three ape-specific SNVs at the BsrI cleavage site (Fig. 2-4). These results failed to uncover additional cross-species transmissions, including human-to-ape transfers, and thus confirmed that extant *P. falciparum* emerged in humans following a single introduction of a gorilla parasite.

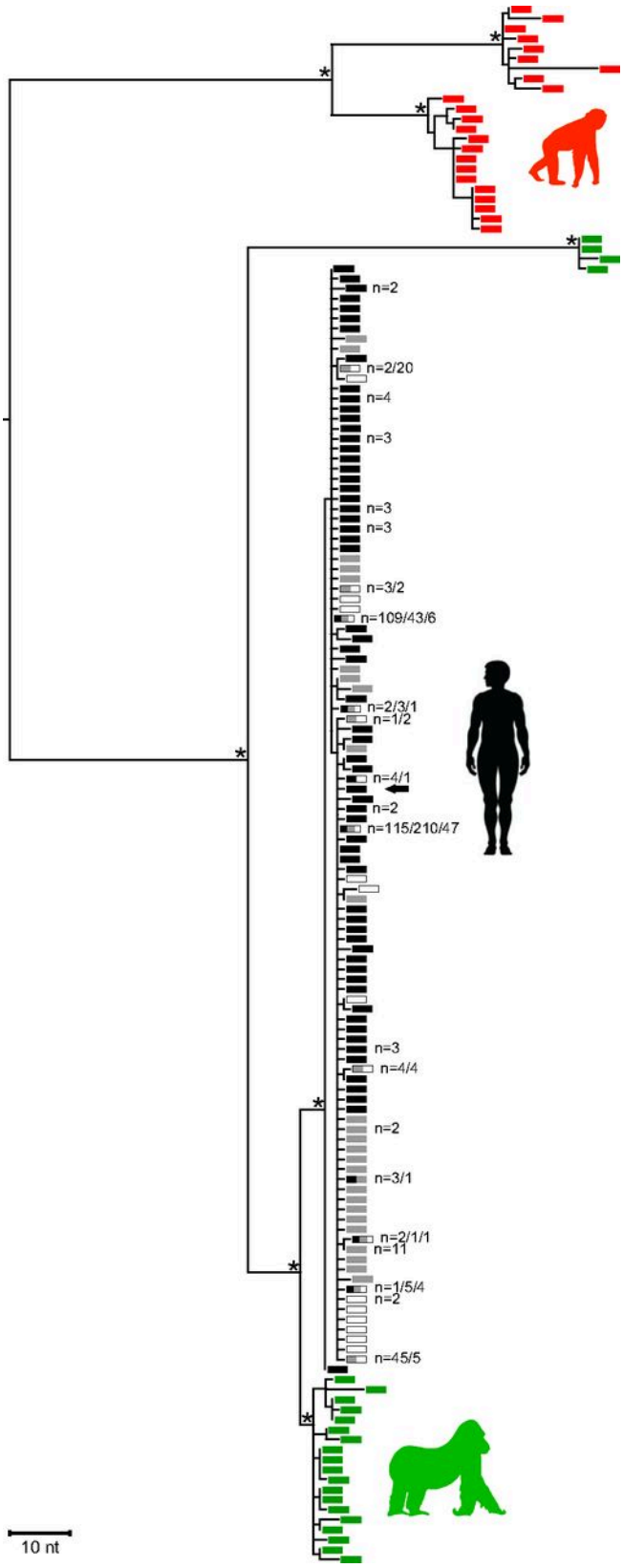
n=2
n=2/20
n=4
n=3
n=3
n=3
n=3/2
n=109/43/6
n=2/3/1
n=1/2
n=4/1
n=2
n=115/210/47
n=3
n=4/4
n=2
n=3/1
n=2/1/1
n=11
n=1/5/4
n=2
n=45/5
10 nt

Figure 2-4. Phylogeny of *P. falciparum* strains from rural Cameroon.

Newly derived *P. falciparum* sequences from humans living in Cameroon (black) are shown in relation to *P. falciparum* sequences from GenBank (white) and the Sanger Institute (grey), as well as to *P. praefalciparum* and *P. reichenowi* sequences from gorillas (green) and chimpanzees (red), respectively. The tree includes 684 new SGA-derived 3.4 kb mitochondrial sequences from 229 human samples, including one that contained a back mutation at the ape specific SNV4 site (highlighted by arrow; also see Fig. 2-3). Numbers at tips indicate the number of times that a sequence was found in Cameroon (black), the Sanger dataset (grey) or GenBank (white) (for sequences present in multiple datasets, numbers are listed in sequence). The tree was inferred using maximum-likelihood methods (*52*). Asterisks indicate posterior probabilities above 0.9. The scale bar represents ten nucleotide (nt) substitutions.

Absence of *P. vivax* in Humans Native to Southern Cameroon

Although the great majority of individuals in Cameroon are Duffy negative (*18*), it has been proposed that *P. vivax* persists in west central human populations at a very low frequency (*53*). Since deep sequencing failed to identify evidence of *P. vivax* infection in 515 individuals, we considered the possibility that the 454 primers were less efficient in amplifying *P. vivax* compared to the other *Plasmodium* species, and thus designed new *P. vivax* specific primers in the mtDNA *cox1* gene (Fig. 2-2). Using these to screen 558 *Plasmodium* positive human samples, we identified 47 that yielded a visible amplification product (*SI Appendix*, Table S6). However, none of these represented *P. vivax* or *P. vivax-like* infections as determined by sequence analysis of the corresponding amplicon. Instead, 37 of the PCR positive samples contained *P. malariae,* while the remaining 10 contained *P. ovale*. Moreover, sequence analysis of the Duffy promoter region from 90

human samples confirmed a Duffy negative phenotype in all of them (*SI Appendix*, Table S6). Thus, using both conventional PCR and 454 deep sequencing approaches, we found no evidence of *P. vivax* infections in individuals living in rural Cameroon.

2.4 Discussion

Chimpanzees and gorillas harbor at least ten different *Plasmodium* species, including six of the subgenus *Laverania* that are closely related to *P. falciparum* (*25-27*, *29*, *30*). The discovery of this previously unrecognized reservoir has prompted concerns that wild-living apes might constitute a source of recurrent human infection (*25*, *32*, *44*). In this study, we set out to examine this possibility for several reasons. First, the ape reservoir is substantial, both in terms of geographic distribution and complexity of *Plasmodium* species. Second, both western gorillas (*G. gorilla*) and common chimpanzees (*P. troglodytes*) are infected throughout their habitat, indicating widespread endemicity throughout west central and central Africa. Third, *Plasmodium* zoonoses can have significant public health impact. A case in point is *P. knowlesi*, a macaque parasite that has shown to cause hundreds of cases of human malaria every year (*4*, *5*). Finally, *Plasmodium* zoonoses have been misdiagnosed in the past: *P. knowlesi* was initially mistaken for *P. malariae*, ultimately requiring the development of molecular tools to facilitate its detection (*5*). Given that malaria infections in central Africa are rarely genetically (or even morphologically) characterized, we considered the possibility that ape *Plasmodium* zoonoses might have also been overlooked. To test this, we developed diagnostic PCR assays and next generation sequencing approaches that permitted the detection of rare *Plasmodium* variants, even when they occurred in the context of mixed-species infection with *P. falciparum*. Using these approaches to test 1,400 blood samples from individuals native to rural Cameroon, we failed to detect previously

unknown human *Plasmodium* infections (*SI Appendix*, Table S1). There was no evidence for zoonotic infection with any of the six ape-specific *Laverania* species or non-*Laverania* parasites identified only in wild apes (*26*). Instead, we detected *P. falciparum*, *P. ovale* and *P. malariae* in a large fraction of individuals, both as mono- and mixed species infections (Table 2-1; *SI Appendix*, Fig. S5). From these data, we conclude that ape *Laverania* zoonoses can be ruled out as an ongoing threat to public health in west central Africa.

Although we failed to find ape *Laverania* infections in humans, our data cannot exclude the possibility of very rare transmission events. Such events would require the screening of a much larger number of individuals from multiple forest environments. In this context, it is important to note that the *Anopheles* species that transmit *Laverania* parasites among wild apes have yet to be identified. It is conceivable that the ecology, distribution and feeding preferences of these vectors play a much greater role in determining the likelihood of zoonotic transmission than the mere geographic proximity of human habitations to infected ape populations. Nonetheless, it seems unlikely that the absence of ape *Laverania* infection in rural communities is solely due to a lack of human exposure. This is because even among endemically infected chimpanzees and gorillas, there is no evidence that *Laverania* parasites cross between the two ape species (*26*). This remarkable host specificity suggests a restriction at the parasite-host interface, which is supported by comparisons of *P. falciparum* and *P. reichenowi* gene sequences. It is known that the genes involved in erythrocyte invasion are evolving rapidly between *Laverania* parasites (*54*). Moreover, erythrocyte invasion of *P. falciparum* is absolutely dependent on the interaction of its PfRh5 ligand with the human Ok blood group antigen basigin (*55*). Human, chimpanzee and gorilla homologues of basigin are highly divergent, suggesting that ape *Laverania* species have to overcome significant adaptive

hurdles before they can spread efficiently in a different host.

Given these restrictions, the question arises how the gorilla precursor of *P. falciparum* managed to colonize humans. One possibility is that *P. praefalciparum* underwent a very specific mutation in a host-compatibility factor that changed its host preference from gorilla to human. Since *P. falciparum* emerged only once (Fig. 2-4), this mutation must have been difficult to generate and/or must have arisen under exceedingly favorable transmission conditions. Another possibility is that the immediate precursor of *P. falciparum* was the product of a rare recombination event. Regardless of the circumstances, it seems clear that the generation of an ape *Laverania* strain that is capable of spreading in humans is an extremely rare event, which may explain why we failed to detect such variants.

In addition to *Laverania* species, wild-living chimpanzees and gorillas also harbor *P. vivax*, *P. malariae*, and *P. ovale*-like infections. Since very few of these ape parasites have been molecularly characterized, it remains unknown whether they represent members of the same or different *Plasmodium* species as their human counterparts. Transmission studies conducted nearly a hundred years ago demonstrated that non-*Laverania* species cross between apes and humans more readily than *Laverania* species (*56*-*60*). Two species, *P. schwetzi* and *P. rodhaini* have been experimentally transmitted to humans in the past (*57*, *58*, *61*). As neither species has been molecularly characterized, it is unknown whether they represent the *P. vivax*, *P. malariae* and/or *P. ovale*-like infections that have more recently been identified in wild apes. It will thus be important to characterize more of these ape parasites to understand their evolutionary history and characterize their zoonotic potential.

The genetic diversity of *P. falciparum* in central Africa is largely unknown since only very few strains from this geographic region have been molecularly characterized. It

has thus been argued that this lack of sequence information has confounded previous evolutionary analyses and lead to erroneous conclusions concerning the origin of *P. falciparum* (*62*). In particular, it has been proposed that *P. praefalciparum* is more likely the result of a human-to-gorilla transmissions of *P. falciparum* than the other way around (*62*). To examine this possibility, we amplified mitochondrial half genome sequences from 229 *P. falciparum* positive samples collected in remote rural areas of Cameroon. We found that *P. falciparum* strains from rural Cameroon were indeed genetically more diverse than previously appreciated. Analyzing 684 single template-derived sequences, we identified 69 unique *P. falciparum* variants, 62 of which had not previously been reported. However, none of these variants changed previous conclusions concerning the evolutionary history of *P. falciparum*. Phylogenetic analysis revealed that all newly characterized variants grouped with previously reported *P. falciparum* strains, forming a well-supported clade within the *P. praefalciparum* radiation (Fig. 2-4). An additional 354 sequences of cosmopolitan *P. falciparum* strains from the Sanger Institute supported this conclusion (*35*), indicating that all human *P. falciparum* sequences coalesced to a single common ancestor. Thus, the addition of over 1,000 new *P. falciparum* sequences, including over 600 from individuals living near wild ape populations, confirmed that *P. falciparum* is of gorilla origin and emerged in humans following a single cross-species transmission event (*26*).

P. vivax is extremely rare in humans in west and central Africa due to the near fixation of the Duffy-negative phenotype which confers resistance to this parasite (*18*). However, a recent study reported *P. vivax* specific antibodies in 13% of humans living in Pointe-Noire, a city in the Republic of Congo, suggesting that *P. vivax* is maintained in a small fraction of Duffy positive individuals (*63*). To examine this possibility for rural Cameroon, we screened nearly 700 human blood samples for *P. vivax* mitochondrial

sequences. Using both PCR and 454 sequencing approaches, we failed to identify *P. vivax* infection in inhabitants from 7 different rural villages. Finally, all human samples tested were Duffy negative, suggesting that the fraction of Duffy positive individuals in rural areas of west central Africa is exceedingly low. These data differ from results of others who have reported the presence of *P. vivax* in Equatorial Guinea and Angola. In these studies, the *P. vivax* infected individuals were either Duffy positive (*64*) or diagnosed solely based on *P. vivax* specific PCR products without sequence verification (*65*). Given the frequency of off-target amplification even with *P. vivax* specific primers, any *P. vivax* diagnosis in central Africa should be confirmed by sequence analysis. Until this is done, the presence of *P. vivax* in rural west central Africa remains questionable.

2.5 Methods

Sample Collections

Human buffy coat samples (n=1,403) were selected from anonymized sample collections previously obtained for molecular epidemiological studies of HIV-1 in Cameroon(*45*). Fecal samples from wild-living apes known to contain *Laverania* and non-*Laverania* parasites served as positive controls (*26*).

Plasmodium Species Specific PCR

Human buffy coat samples were first screened for *Plasmodium cytb* sequences by conventional PCR as described (*26*). Positive samples were further characterized using a *Plasmodium* species specific PCR and BsrI digest approach (See *SI Appendix*, Supplemental Methods for more details).

Pyrosequencing

A 405 bp fragment of the mitochondrial genome that spanned three SNVs unique to ape *Laverania* parasites (Fig. 2-2) was amplified and sequenced on a Genome Sequencer FLX Titanium Series (Roche) (See *SI Appendix*, Supplemental Methods for more details).

Single Genome Amplification and Sequencing

To derive *Plasmodium* mitochondrial sequences without PCR induced substitutions and/or recombination, a 3.4 kb fragment of the *Plasmodium* mitochondrial genome was amplified and sequenced directly from a subset of *cytb* PCR positive samples (n = 229) as previously described (*26*).

Phylogenetic Analyses

SGA derived 3.4 kb mitochondrial sequences were aligned with human and simian reference sequences. Trees were inferred using Maximum-Likelihood and Bayesian methods (See *SI Appendix*, Supplemental Methods for more details).

***P. vivax* specific PCR**

Human samples were screened for *P. vivax* infections by nested PCR as described (*66*). Nested primers were specifically designed to avoid off-target amplification of *P. falciparum* or other *Laverania* species, and were shown to amplify ape *P. vivax*-like parasites as well as human *P. vivax* with high sensitivity and specificity (*66*) (see *SI Appendix* for primer sequences and amplification conditions).

Duffy Phenotype

Buffy coat DNA was extracted and used to amplify the Duffy promoter region by nested PCR. The Duffy phenotype was determined by direct sequencing (See *SI Appendix*, Supplemental Methods for more details).

GenBank Accession Numbers

All newly derived SGA sequences are available under GenBank accession numbers KC175306-KC175322 and KC203521-KC203587. The 454 pyrosequencing read data have been deposited in the National Center for Biotechnology Information Sequence Read Archive (SRA) under accession number SRP019191.

2.6 Acknowledgements

We thank Avelin Aghokeng, Eugenie Etam Ebong, Arrah Atem Tamba, Celine Montavon, Julius Chia, Nathalie Nkue, Mireille Mpoudi, Géraldine Manirakiza, Marie Bourgeois, Audrey Gleize, Justin Wadi, Anke Bourgeois for field work and logistics in Cameroon; The Ministry of Public Health (Republique Du Cameroun) for authorizations and logistical support. Aubrey Bailey and Nirav Malani for maintenance and organization of the 454 pyrosequencing datasets; Rohini Sinha for contributing various deep sequencing analysis tools; Xiaowen Zhang for assistance with laboratory techniques; Patricia Crystal for artwork and manuscript preparation; and Christian Hoffman, Robert W. Doms, Dustin C. Brisson, David S. Roos, Sarah A. Tishkoff for helpful discussions.

CHAPTER 3 - Cryptic chimpanzee *Plasmodium* parasites reveal key evolutionary events leading to human malaria

Sesh A. Sundararaman, Lindsey J. Plenderleith, Weimin Liu, Dorothy E. Loy, Gerald H. Learn, Yingying Li, Katharina S. Shaw, Ahidjo Ayouba, Martine Peeters, Sheri Speede, George M. Shaw, Frederic D. Bushman, Dustin Brisson, Julian C. Rayner, Paul M. Sharp and Beatrice H. Hahn.

Originally published in Nat Commun. 2016 Mar 22;7:11078

Online material available at

http://www.nature.com/ncomms/2016/160322/ncomms11078/extref/ncomms11078-s1.pdf

Lindsey J. Plenderleith, Dustin Brisson, Julian C. Rayner, George M. Shaw, Frederic D. Bushman, Paul M. Sharp, Beatrice H. Hahn, and I conceived and planned the study. I designed the selective whole genome amplification (SWGA) primers for *Plasmodium* and developed the SWGA method for *Plasmodium* infected great ape blood samples. Martine Peeters and Sheri Speede conducted or supervised the fieldwork. Dorothy E. Loy, Weimin Liu, Katharina S. Shaw and I performed selective whole genome amplification of ape *Laverania* parasites. Lindsey J. Plenderleith, Gerald H. Learn, and I constructed, curated, annotated and submitted the *Laverania* genome assemblies. Dorothy E. Loy, Weimin Liu, Yingying Li, Katharina S. Shaw, Ahidjo Ayouba, and I performed non-invasive ape *Plasmodium* testing and limiting dilution PCR. Gerald H. Learn, Lindsey J. Plenderleith, and Paul M. Sharp performed phylogenetic analyses. Lindsey J. Plenderleith, Weimin Liu, Gerald H. Learn, and I analyzed the data. Lindsey J. Plenderleith, George M. Shaw, Paul M. Sharp and Beatrice H. Hahn, and I wrote the manuscript with contributions from all authors.

3.1 Abstract

African apes harbor six *Plasmodium* species of the subgenus *Laverania*, one of which gave rise to human *P. falciparum* (*26*). Here, we used a novel phi29 polymerase-based selective amplification strategy (*39*) to sequence the genomes of three chimpanzee parasites, including one that is closely (*P. reichenowi*) and two that are distantly (*P. gaboni*) related to *P. falciparum*. Analysis of these sequences demonstrated a near-identical core genome (>4,600 orthologs), but also a 10-fold higher within-species diversity among the chimpanzee parasites, revealing a very recent origin of *P. falciparum* in humans. Surprisingly, genome-wide analyses uncovered a striking expansion and diversification of a multi-gene family (*67*) involved in erythrocyte remodeling, and showed that a region on chromosome 4, which encodes the essential invasion genes *RH5* and *CyRPA*, was horizontally transferred into a recent *P. falciparum* ancestor. These results provide a new paradigm for characterizing cryptic pathogen species (*68*) and reveal evolutionary events that likely predisposed the precursor of *P. falciparum* to successfully colonize humans.

3.2 Introduction

Plasmodium falciparum, the cause of malignant malaria in humans, evolved following a single cross-species transmission event involving a parasite that naturally infects western gorillas (*Gorilla gorilla*)(*26*). To elucidate key events that led to its emergence, we characterized the evolutionary history of related ape *Laverania* parasites. All *Plasmodium* reference genomes generated to date are derived from purified parasites grown to high titers in red blood cells (RBCs) *in vitro* or susceptible host species *in vivo* (*9*, *69*-*73*). Since blood samples from endangered chimpanzees and gorillas are not readily available, efforts to culture ape *Laverania* parasites, which are highly species specific *(26)*, have remained unsuccessful. To date, only a single genome of the chimpanzee parasite *P. reichenowi* has been sequenced, following extensive *in vivo* passage and amplification in experimentally infected, splenectomized chimpanzees (*69*, *74*). Since this method of parasite enrichment is neither ethical nor practical, we developed a strategy to selectively amplify and sequence near full-length *Plasmodium* genomes from subpatently-infected ape blood.

3.3 Results

Traditional whole genome amplification methods utilize the highly-processive phi29 polymerase and random primers to generate DNA fragments of up to 70kb in length, but amplify all templates within a sample with near-uniformity (*75*-*78*). Since microbial and host genomes differ in the frequency of common sequence motifs (*79*), we reasoned that it should be possible to design primers that would amplify pathogens selectively, even if they represented only a small fraction of the sample DNA. Testing this concept on *Wolbachia* infected fruit flies, we found that selective whole genome amplification (SWGA) generated sufficient quantities of bacterial genomes for next generation

sequencing (*39*). To extend this method to more complex eukaryotic pathogens, we tested whether SWGA could amplify the multi-chromosomal genomes of *Plasmodium* parasites from unprocessed human and ape blood samples. Searching for short (8-12 bp) sequences that are overrepresented in *P. falciparum* and *P. reichenowi*, but underrepresented in the genomes of their primate hosts (Online Methods), we identified 2,418 motifs that occur frequently in the parasite DNA (i.e., spaced on average less than 50,000 bp apart), but only rarely (i.e., spaced on average more than 500,000 bp apart) in human and chimpanzee DNA (Figs. 3-1A and B). We selected two sets of SWGA primers (Supplementary Fig. 3-1) based on their distribution across the parasite genomes and their DNA binding properties, and tested them on human DNA containing known quantities (0.001% to 5%) of *P. falciparum* DNA (Online Methods). These experiments showed that SWGA amplified *P. falciparum* genomes with remarkable breadth and selectivity over a wide range of concentrations, especially when results from independent amplifications were combined (Fig. 3-1C). Of ~2.5 million MiSeq reads derived from human DNA containing as little as 0.001% *P. falciparum* DNA (19 genome equivalents), ~1.7 million (70%) mapped to the *P. falciparum* genome, indicating a 70,000-fold enrichment of the parasite compared to the host DNA (Table 3-1). Read coverage was even across all 14 chromosomes, except for the sub-telomeres where low complexity sequence precludes accurate mapping (Supplementary Fig. 2). Stochastic amplification was seen only at the lowest (0.001%) parasite dilution (Supplementary Fig. 2). Thus, SWGA generated high-quality *Plasmodium* core genomes from samples containing large quantities of contaminating host DNA.

Table 3-1 Selective whole-genome amplification of *P. falciparum* from mixtures of human and parasite DNA.

Per cent parasite admixture*	Total DNA (ng)	*P. falciparum* DNA (ng)	*P. falciparum* genome copies	Total MiSeq reads	Reads mapping to the human genome (percent)	Reads mapping to the *P. falciparum* genome (percent)	Unmapped reads (percent)	Fold parasite enrichment
5 *P. falciparum*	50	2.5	96,507	3,157,170	9,156 (0.3)	2,968,254 (94.0)	179,760 (5.7)	19
				2,570,124	7,064 (0.3)	2,408,319 (93.7)	154,741 (6.0)	19
1 *P. falciparum*	50	0.5	19,301	3,412,530	34,936 (1.0)	3,152,623 (92.4)	224,971 (6.6)	92
				2,804,890	22,190 (0.8)	2,660,365 (94.8)	122,335 (4.4)	95
0.1 *P. falciparum*	50	0.05	1,930	3,422,726	43,108 (1.3)	3,174,382 (92.7)	205,236 (6.0)	930
				2,638,548	56,552 (2.1)	2,444,992 (92.7)	137,004 (5.2)	930
0.01 *P. falciparum*	50	0.005	193	3,917,388	332,468 (8.5)	3,390,560 (86.6)	194,360 (5.0)	8,700
				3,362,418	429,008 (12.8)	2,730,631 (81.2)	202,779 (6.0)	8,100
0.001 *P. falciparum*	50	0.0005	19	2,430,994	590,934 (24.3)	1,688,947 (69.5)	151,113 (6.2)	69,000
				2,635,560	613,656 (23.3)	1,832,657 (69.5)	189,247 (7.2)	70,000

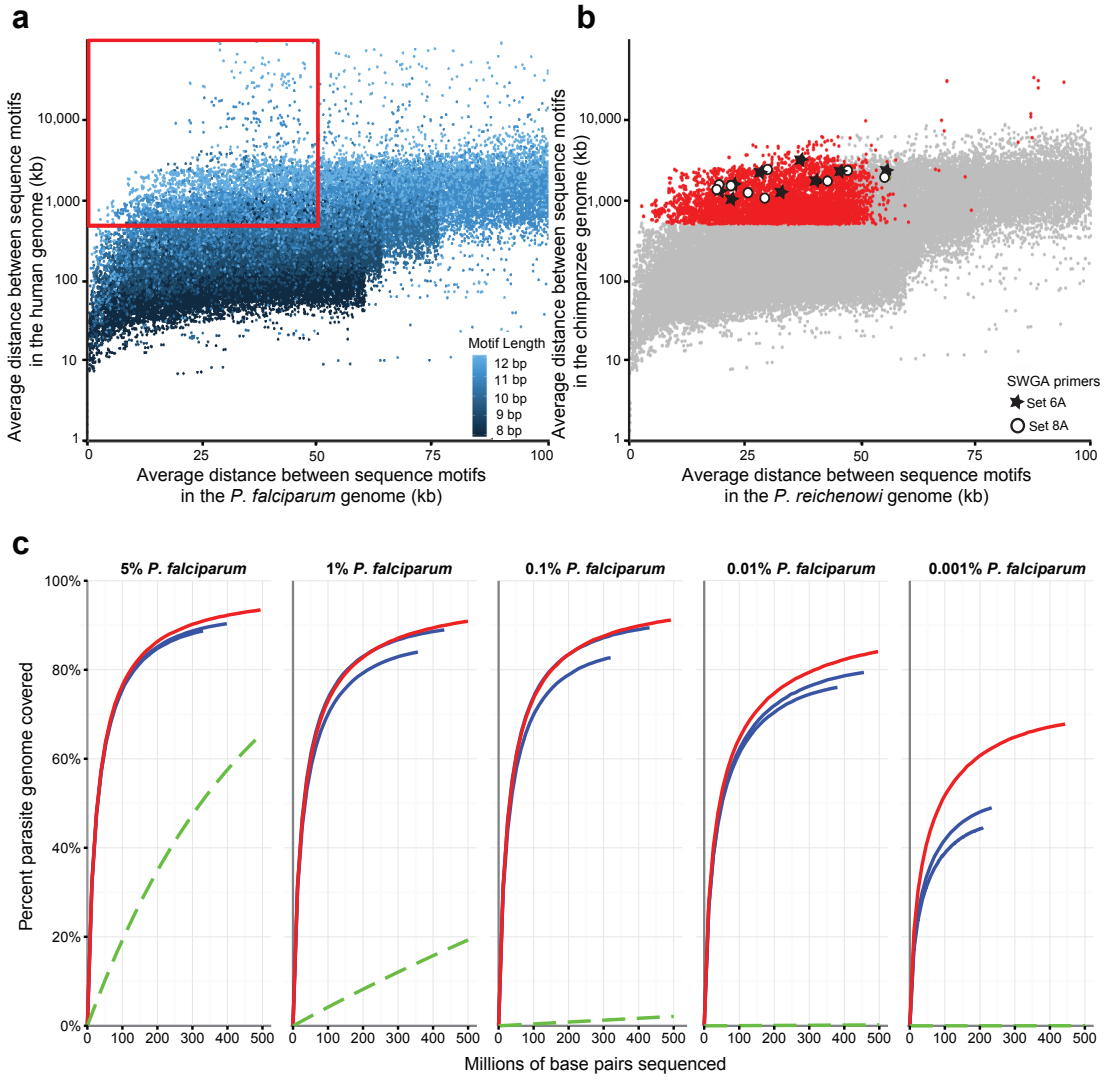

Figure 3-1. Selective whole genome amplification (SWGA) of *Plasmodium* parasites.

(A,B) Selection of SWGA primer sets. **(A)** The average distance (kb) between the 10,000 most frequent parasite motifs (color coded by length) is plotted for both the *P. falciparum (*Pf3D7) and human (GRCh37) genomes. The red box highlights motifs that are spaced (on average) less than 50,000 bp apart in the *P. falciparum*, but more than 500,000 bp apart in the human genome. **(B)** Average distances between the sequence motifs shown in a, but plotted for the *P. reichenowi* (PrCDC) and chimpanzee

(Pan_troglodytes-2.1.4) genomes. Red dots indicate all motifs that fall within the red box in a, with circles and stars denoting those selected for SWGA primer sets 6A and 8A, respectively (Supplementary Fig. 1). **(C)** Validation of the SWGA primer sets identified in A and B. Human genomic DNA containing known quantities of *P. falciparum (*5% to 0.001%) were subjected to consecutive rounds of SWGA, using primer set 6A in the first and primer set 8A in the second round, respectively. The number of total base pairs (in millions) sequenced is shown in relation to the percent coverage of the *P. falciparum* (Pf3D7) genome for five parasite concentrations. DNA mixtures were subjected to two independent amplifications, with individual and combined results shown in blue and red, respectively (the expected genome coverage without SWGA is shown in green).

We next used SWGA to amplify the genomes of three chimpanzee parasites, representing both close (*P. reichenowi*) and very distant (*P. gaboni*) relatives of *P. falciparum* (*26*). Whole blood samples were obtained from sanctuary chimpanzees (*Pan troglodytes*) during their annual health examination and tested for *Plasmodium* infection using conventional PCR. Parasite DNA positive blood samples were further characterized by limiting dilution (single template) PCR amplification of eight mitochondrial, apicoplast, and nuclear loci to determine their *Plasmodium* species composition (*26*, *80*). This analysis identified one sample (SY57) to contain almost exclusively (>99%) *P. reichenowi* and two others (SY75 and SY37) to contain only *P. gaboni* DNA (Supplementary Table 1). In each case, the parasites comprised only a miniscule fraction of the total blood DNA (0.0054%, 0.14% and 0.00081% for SY57, SY75 and SY37, respectively). To reduce the contaminating host DNA, we digested all samples with methylation dependent restriction enzymes (MspJI and FspEI) known to cleave ape, but not *Plasmodium*, genomic DNA (*38*), and then used the digestion

products for SWGA and Illumina sequencing (Online Methods). This approach yielded 27, 31 and 39 million MiSeq reads for samples SY57, SY75 and SY37, respectively, of which 89%, 73% and 61% mapped to *Plasmodium* sequences (Supplementary Table 1). Sequence coverage was even across all 14 chromosomes, with no evidence for selective sequence loss, including near the ends of some chromosomes (Supplementary Fig. 3). Reads from sample SY57 were mapped to the *P. reichenowi* PrCDC reference and shown to cover 96% of its genome (at a coverage depth of 5-fold or higher read). Since there is no published *P. gaboni* genome, reads from samples SY75 and SY37 were mapped to the *P. falciparum* Pf3D7 reference and shown to cover 79% and 69% of its genome (at ≥5x), respectively (Supplementary Table 1). This lower coverage was not due to a reduction in selective amplification, but the difficulty of mapping reads to a highly divergent reference sequence.

Using reference guided iterative assembly (Online Methods) (*81*), we generated draft genomes for PrSY57 and PgSY75, which contained 18.5 Mb and 18.9 Mb of chromosomal, as well as 1.4 and 0.8 Mb of sequence that could not be placed within the core genome, respectively (Table 3-2). Due to the very small quantities of parasite DNA, sequence coverage for PgSY37 was lower, yielding 15 Mb of chromosomal and 8 Mb of unplaced sequences. Syntenic annotation transfer and *ab initio* gene prediction identified 4,920, 4,962 and 4,179 full-length and partial protein coding genes in PrSY57, PgSY75 and PgSY37, respectively, which included 98.3%, 98.7% and 85.7% of the core genes in the respective reference sequences (Table 3-2). In genomic regions that were syntenic among all three species, there were only 5 *P. gaboni* genes that were missing in *P. falciparum* and/or *P. reichenowi*, and only 5 *P. reichenowi* and/or *P. falciparum* genes that were absent from *P. gaboni* (Supplementary Table 2, Supplementary Fig. 4). Of 76 pseudogenes identified in the three *Laverania* species, only 7, 14 and 9 were specific for

P. falciparum, P. reichenowi and *P. gaboni,* respectively (Supplementary Table 3). Thus, the core genome of ape and human *Laverania* parasites is highly conserved even among the most divergent species within the subgenus.

Table 3-2 Genome features of *P. gaboni* and *P. reichenowi*

	P. reichenowi	*P. gaboni*	*P. gaboni*
Genome ID	PrSY57	PgSY75	PgSY37
Chromosomal assembly (bp)[a]	18,853,635	18,465,530	15,330,638
Chromosomal contigs	1,012	331	n/a[g]
Unplaced assembly (bp)[b]	784,231	1,442,089	8,902,276
Unplaced contigs	741	809	14,793
Chromosomes	14	14	14
GC content (%)	18.6	18.3	17.1
Core protein-coding genes[c]	4670 (98.3%)	4689 (98.7%)	4071 (85.7%)
Full-length[d]	4359 (91.8%)	4382 (92.2%)	3295 (69.4%)
Partial[e]	311 (6.5%)	307 (6.5%)	776 (16.3%)
Subtelomeric protein-coding genes[c]	235 (23.8%)	222 (33.2%)	108 (16.2%)
Full-length[d]	182 (18.5%)	189 (28.3%)	72 (10.8%)
Partial[e]	53 (5.4%)	33 (4.9%)	36 (5.4%)
Other protein-coding genes[f]	15	51	0
Full-length[d]	14	44	n/a
Partial[e]	1	7	n/a
tRNA genes	42 (93.3%)	43 (95.6%)	32 (71.1%)
rRNA genes	8 (47.1%)	11 (43.3%)	2 (7.7%)
Full-length[d]	4 (23.5%)	10 (38.5%)	2 (7.7%)
Partial[e]	4 (23.5%)	1 (3.8%)	0
ncRNA genes	71 (75.5%)	67 (65.7%)	49 (48.0%)
Full-length[d]	66 (70.2%)	61 (59.8%)	40 (39.2%)
Partial[e]	5 (5.3%)	6 (5.9%)	9 (8.8%)
Apicoplast genes	45 (76.3%)	58 (85.3%)	n/a
Full-length[d]	27 (90.0%)	30 (100%)	n/a
Partial[e]	2 (6.7%)	0	n/a
tRNA genes	16 (59.3%)	26 (76.5%)	n/a
rRNA genes	0	2 (50%)	n/a

[a]Length of all contigs that could be placed in chromosomes, excluding gaps; bp, base pairs.
[b]Length of all contigs that could not be placed in chromosomes ('bin'), excluding gaps; bp, base pairs.
[c]Gene counts excluding splice variants, but including pseudogenes and partial genes; parentheses indicate the percentage of genes covered in the *Plasmodium* references Pf3D7 (PgSY75 and PgSY37) and PrCDC1 (PrSY57).
[d]Number includes all genes that comprise ≥90% of the lengths of their Pf3D7 or PrCDC orthologs/homologs as well as all genes that comprise ≥80% of the lengths of their Pf3D7 or PrCDC orthologs/homologs and contain no assembly gaps.
[e]All annotated coding sequences for which homologs could be identified by BLAST search, but did not contain a sufficiently long sequence to be considered full-length.
[f]Genes for which an ortholog could not be unambiguously identified in the reference genome.
[g]n/a, not available; the PgSY37 genome was generated by iteratively replacing the PgSY75 genome with PgSY37 reads and replacing the regions that lacked 5-fold coverage with Ns; reads not mapped to PgSY75 chromosomes were assembled *de novo* to generate 'unplaced contigs'.

Availability of new *P. reichenowi* and *P. gaboni* genomes allowed us to examine the within-species diversity among chimpanzee parasites. For comparison, the intra-species diversity of *P. falciparum* was calculated using published SNP data from 12 geographically diverse field isolates (see the 'Methods' section for details). Comparing more than 3,000 genes, we found that the two *P. gaboni* genomes (PgSY75 and PgSY37) differed at 0.4% of all coding sites, and 1.1% of fourfold degenerate (silent) sites. Similarly, the two *P. reichenowi* genomes (PrSY57 and PrCDC) differed at 0.3% of all coding sites, and 0.9% of fourfold degenerate sites (Table 3-3). Note that, for every gene, the divergence between the two *P. gaboni* sequences, or between the two *P. reichenowi* sequences, was lower than between these two species, consistent with the premise that *P. gaboni* and *P. reichenowi* are genetically isolated.

In contrast, 12 field isolates of *P. falciparum* selected from countries around the world differed on average at only 0.04% of all coding sites, and 0.08% of fourfold degenerate sites (Table 3-3). To ensure that the higher diversity among the ape parasites was not an artefact of the SWGA method, we amplified several nuclear loci that exhibited particularly high sequence diversity (three from *P. gaboni* and four from *P. reichenowi*) using limiting dilution PCR (Supplementary Table 4). The resulting sequences were identical to the SWGA-derived genomes except for two indels in difficult-to-assemble regions, which had been excluded from the diversity calculations, thus further validating the accuracy of the SWGA method (Supplementary Fig. 5). The distributions of diversity levels across genes were very similar in *P. reichenowi* and *P. gaboni* (Fig. 3-2). This was also the case when the within-species diversity for *P. reichenowi* or *P. gaboni* was compared with the maximum pairwise divergence obtained for each gene among the 12 *P. falciparum* field isolates (Supplementary Fig. 6). In all comparisons, the difference between ape and human parasites reflected significantly

higher diversity levels in genes distributed across the entire core genome (Sign test, Pr>Pf and Pg>Pf: $p < 2.2e-16$). Thus, for both chimpanzee parasite species, including two *P. gaboni* strains from the same location, the genetic diversity is about ten times higher than that seen among *P. falciparum* strains from different geographic regions across the globe. This reduced diversity in *P. falciparum* is consistent with a severe population bottleneck, which most likely occurred at the cross-species transmission from gorilla to human.

Table 3-3 Genome features of *P. gaboni* and *P. reichenowi*

Species[a]	N[b]	π[c]	π4[d]	Genes	π[c]	π4[d]	Genes
P. falciparum	12	0.00049	0.00081	4,818	0.00043	0.00079	3,111
P. reichenowi	2	0.00364	0.00899	4,439	0.00324	0.00876	3,111
P. gaboni	2	0.00407	0.01069	3,331	0.00381	0.01049	3,111

[a]Values represent weighted mean values across genes. Values at the right are for 3,111 genes available for all three species.
[b]N, number of strains.
[c]π, pairwise nucleotide diversity across all non-masked sites;
[d]π4, pairwise nucleotide diversity across non-masked 4-fold degenerate sites.

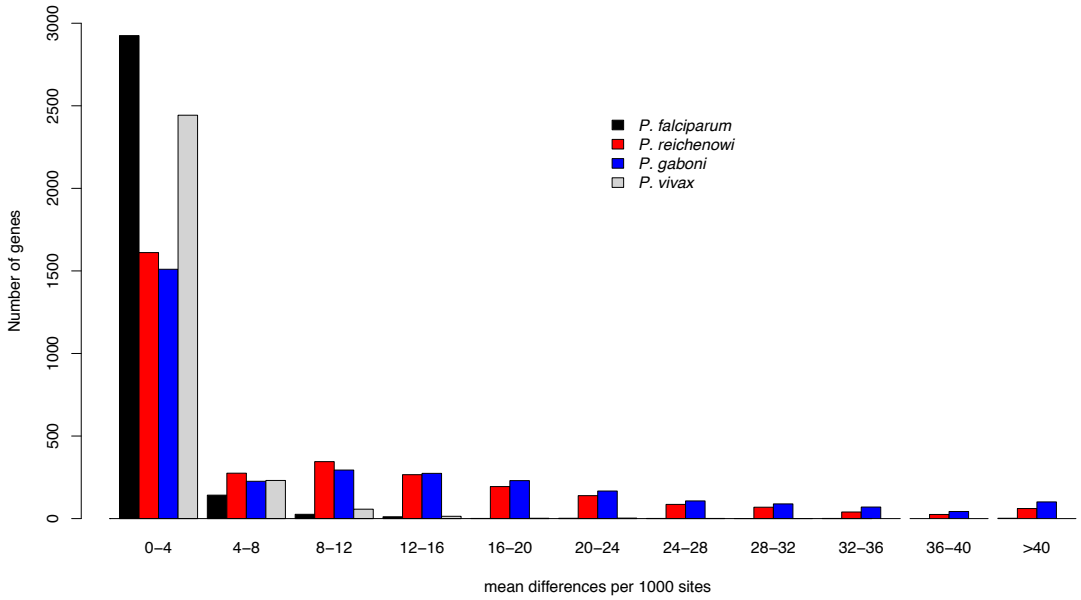

Figure 3-2. Sequence diversity within three *Laverania* species and *P. vivax*.

The average pairwise nucleotide sequence diversity is shown for 3,111 syntenic core genes at four-fold degenerate sites for 12 geographically diverse strains of *P. falciparum* (black), two strains of *P. reichenowi* (red), and two strains of *P. gaboni* (blue), respectively. The average pairwise diversity for 2,753 genes from five strains of *P. vivax* is included as a comparator. For *P. falciparum* strains other than 3D7, diversity information was obtained from SNP data (only datasets representing single parasite strains were used for analysis; see Online Methods for more detail). For *P. vivax*, diversity information was obtained from previously published SNP data (Neafsey et al., 2012)

It has long been suspected that *P. falciparum* has unusually low genetic diversity (*82*), although the underlying causes have been the subject of much debate (*83*). Genome-wide analysis of human and chimpanzee parasites now show that this low

diversity is not a general characteristic of *Laverania* parasites, and therefore not a consequence of their life cycle (*84*, *85*) nor an artifact of their very A+T-rich genomes. The expected neutral nucleotide diversity is dependent on the effective population size, which for a parasite is generally dependent on the population size of its host. While the effective population size of chimpanzees is higher (2-3x) than that of humans, numbers of chimpanzees seem unlikely to have been ten times larger than those of humans in the past (*86*). A simple explanation for the low diversity in *P. falciparum* is a very recent population bottleneck, reflecting the cross-species transmission of its gorilla precursor (*26*).

To gain insight into the host specificity of *Laverania* parasites, we examined members of multigene families known to function at the host-parasite interface. Many of these, including members of the *var*, *rif* and *stevor* families, could not be completely assembled because of their extreme variability and subtelomeric location, although *var*-like genes have been shown to be present in all ape *Laverania* species (*80*). One family of putative protein kinases, termed FIKK (after a conserved Phe-Ile-Lys-Lys motif in their amino acid sequence), was of particular interest because it expanded from a single gene present in all *Plasmodium* parasites to 20 genes in both *P. falciparum* and *P. reichenowi* (*67*, *69*). Remarkably, the new *P. gaboni* genome contained 21 *FIKK* genes, 20 of which represented clear syntenic orthologs of corresponding *P. falciparum* and *P. reichenowi* genes as demonstrated by phylogenetic analysis (Fig. 3-3A) and chromosomal location (Supplementary Table 5). The remaining *P. gaboni* gene on chromosome 9, termed *FIKK*9.15, did not have an ortholog in *P. falciparum* and *P. reichenowi* (Supplementary Fig. 7A), but was identified in the closest relative of *P. gaboni*, the gorilla parasite *P. adleri* (Supplementary Fig. 7B). These data indicate that the FIKK gene family underwent an unprecedented burst of gene duplications and rapid diversification

very early in *Laverania* evolution, followed by a period of greatly reduced divergence rates and near stasis of *FIKK* gene copy numbers after the radiation of extant *Laverania* species (Fig. 3-3A). Although their exact function remains to be determined, the *P. falciparum FIKK* genes are expressed at different time points during the erythrocytic cycle (*87*) (Fig. 3-3B), with all but the ancestral *FIKK8* believed to be exported into the host erythrocyte to contribute to the remodeling of its cytoskeleton and surface membrane structures (*67*, *88*-*90*). The slow rate of evolution of *FIKK8* (Fig. 3-3A) suggests that it retained its original, cytosolic function consistent with similar expression profiles of its ortholog in *P. vivax* (*87*, *91*). However, all other family members appear to have acquired novel, non-redundant and seemingly essential functions, since only very few have become pseudogenes in one or more *Laverania* species. For example, *FIKK7.2* and *FIKK14* are pseudogenes in *P. falciparum* (Fig. 3-3A), but lack the respective inactivating mutations in *P. praefalciparum* and *P. adleri* (Supplementary Fig. 8). Similarly, *FIKK9.5* is a pseudogene in *P. gaboni* and *P. reichenowi*, but is intact in *P. falciparum* (Fig. 3-3A) and possibly other *Laverania* species. Interestingly, other exported multigene families that have undergone lineage specific expansion in *P. falciparum* and *P. reichenowi* (*92*), including DNAJ and PHISTb genes, also have syntenic orthologs in *P. gaboni* (Supplementary Table 7). Thus, it seems likely that the rapid multiplication of the *FIKK* gene family, perhaps in concert with other members of the *Plasmodium* exportome (proteins exported from the parasite) (*92*), is at least partly responsible for the unique biology of *Laverania* parasites, including their ability to mediate red blood cell cytoadhesion, tissue sequestration and/or host immune escape (*93*, *94*).

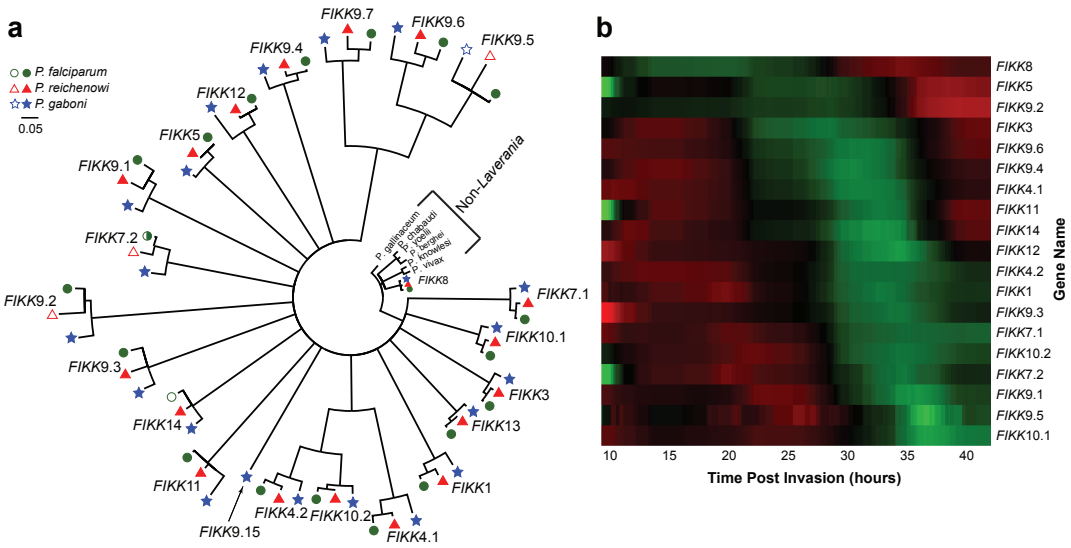

Figure 3-3. Expansion and diversification of the *FIKK* multigene family in the *Laverania* subgenus.

(A) Phylogeny of *FIKK* genes from *P. falciparum* (green), *P. reichenowi* (red), *P. gaboni* (blue), and non-*Laverania* species (black). *FIKK* genes are labeled according to their *P. falciparum* orthologs, with the new *P. gaboni* gene designated *FIKK*9.15. Open symbols indicate pseudogenes in all members of the species (*FIKK*7.2 is intact in some strains of *P. falciparum*). The tree was inferred using maximum-likelihood methods(*52*) using an alignment of first and second codon positions. Internal branches with bootstrap support of less than 70% are collapsed. The scale bar represents 0.05 substitutions per site. **(B)** Expression profiles of *P. falciparum FIKK* genes. Previously published microarray data (*87*) from 21 clonal *P. falciparum* strains were used to calculate the expression levels of 19 FIKK genes at different time points during the intra-erythrocytic lifecycle (data for the remaining *FIKK* gene were not available). Colors represent mRNA expression levels relative to a reference pool, with red, black and green indicating higher, equal, and lower expression levels than the reference pool, respectively. Genes were arranged to illustrate the sequential nature of *FIKK* gene expression.

To investigate whether any genes exhibit unusual patterns of divergence among *P. falciparum*, *P. reichenowi* and *P. gaboni*, we calculated inter-species distances for 4,500 syntenic orthologs (Supplementary Data). As expected from mitochondrial DNA, the pairwise distance between *P. falciparum* and *P. reichenowi* was about four-fold lower than the distance of either species to *P. gaboni*. However, there were four genes for which these relationships were reversed, that is the *P. falciparum*/*P. reichenowi* distance was about four-fold higher than the *P. falciparum*/*P. gaboni* distance (Supplementary Data). Remarkably, these four genes are all located on the same 8 kb segment of chromosome 4 (Fig. 3-4A) and include two essential invasion genes encoding the reticulocyte binding-like homologous protein 5 (RH5) and the cysteine-rich protective antigen (CyRPA) (*55*, *95*). To investigate this further, we amplified regions from both within and outside the 8 kb segment from additional ape *Laverania* species (Supplementary Table 8, Supplementary Fig. 9). Whereas a phylogeny derived from the *EBA165* gene (outside the 8 kb segment) was consistent with previous topologies (*26*), we found an unexpectedly close relationship of the *P. falciparum*/*P. praefalciparum* clade with the gorilla parasite *P. adleri* in trees based on the *RH5* and *CyRPA* genes (Fig. 3-4B; Supplementary Fig. 9). Mating between members of different *Laverania* species is unlikely to generate viable progeny, and so the discordant evolutionary history of this 8 kb region is most likely the result of horizontal gene transfer (HGT) from an ancestor of *P. adleri* to an ancestor of *P. praefalciparum*. Cultured erythrocyte-stage parasites of *P. falciparum* take up DNA spontaneously from their host cell cytoplasm (*96*) and infected red blood cells have been shown to communicate via exosome-like vesicles that are capable of delivering genes (*97*). Thus, this HGT most likely occurred during the blood stage infection of a gorilla harboring multiple *Laverania* species (*26*).

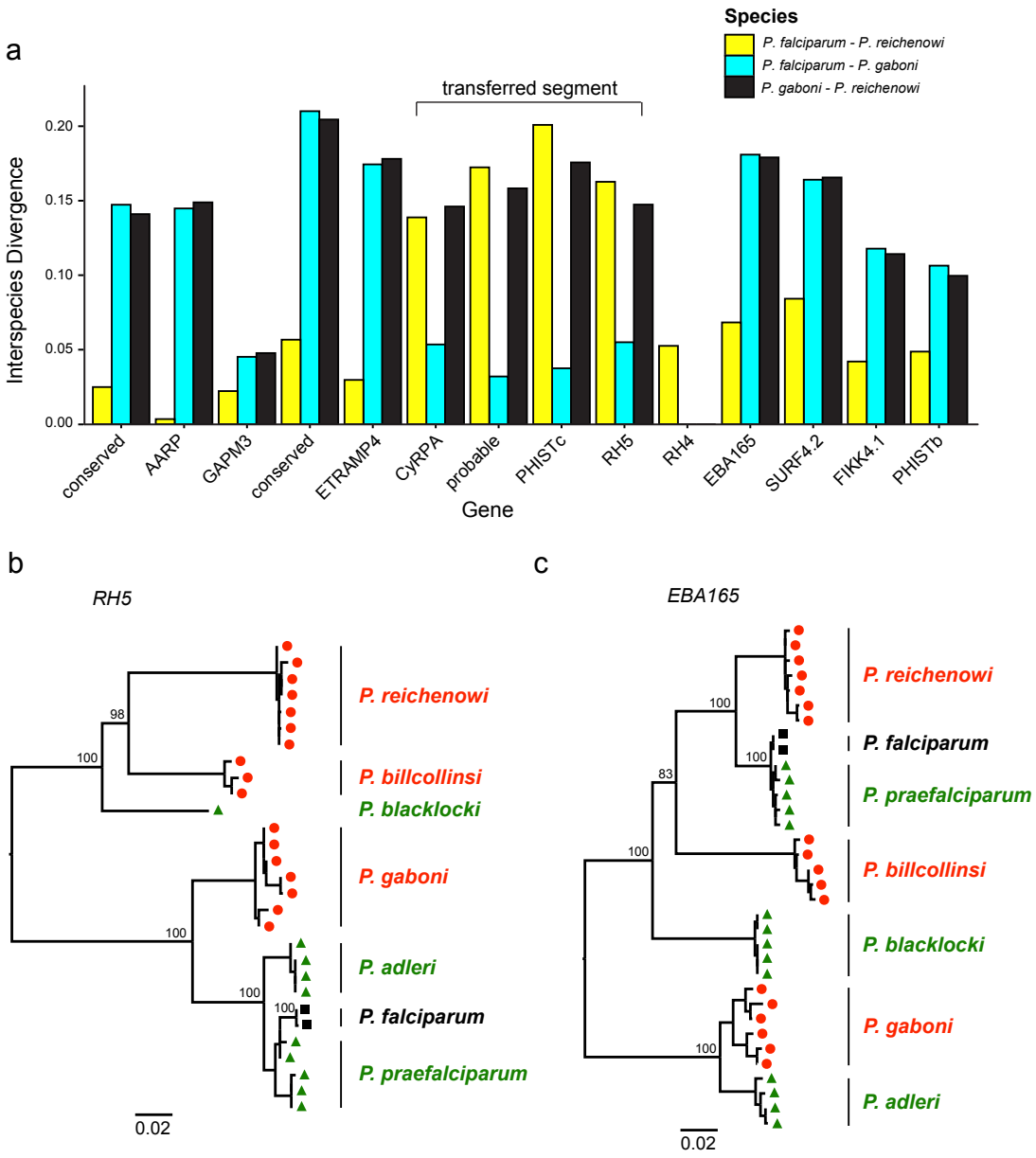

Figure 3-4. Horizontal gene transfer between two *Laverania* species includes two essential invasion genes.

(A) Identification of an 8 kb transferred segment on chromosome 4. Interspecies distances (color coded) are shown for syntenic orthologs of *P. falciparum*, *P. reichenowi* and *P. gaboni*. Four genes, including the essential invasion genes *CyRPA* and *RH5*, exhibit an unusually high *P. falciparum*/*P. reichenowi* (yellow) and an unusually low

P. falciparum/*P. gaboni* *(*aqua) distance, respectively. Genes are ordered by chromosomal location. Since *RH4* is absent from *P. gaboni* (see Supplementary Fig. 4), only the *P. falciparum*/*P. reichenowi* distance is shown. **(B,C)** Phylogenetic relationships of *Laverania RH5* and *EBA165* gene sequences, revealing an unexpectedly close relationship between the *P. praefalciparum/P. falciparum* and *P. adleri* lineages in *RH5*. *Laverania* parasites are colored coded according to their host species (chimpanzee, red; gorilla, green; human, black). Trees were inferred by maximum likelihood methods (*52*). Numbers at internal nodes represent bootstrap support values (*98*) (only numbers >80% are shown). The scale bars represent 0.02 substitutions per site (additional phylogenies are shown Supplementary Fig. 9).

Recent studies have shown that RH5, CyRPA and the RH5-interacting protein (RIPR) form a multiprotein complex that is attached to the merozoite surface via the CyRPA glycosylphosphatidylinositol (GPI) anchor (*95*). This adhesion complex ensures the proper positioning of RH5, which lacks a transmembrane domain, thus facilitating its binding to the erythrocyte receptor basigin, an obligate step in the erythrocyte invasion process (*99*). Given the essential nature of these interactions, the acquisition of "matching" *RH5* and *CyRPA* coding regions on both ends of a mosaic fragment seems unlikely to represent a chance event (Fig. 3-4A). Indeed, the initially transferred fragment may have been longer, but would have been reduced in size by successive recombination events eroding its edges until any further shortening was deleterious because it failed to conserve compatible RH5 and CyRPA proteins. Breakpoint analysis of the fragment boundaries provides support for this hypothesis (Supplementary Fig. 10).

3.4 Discussion

Although the origin of *P. falciparum* is now well established, nothing is known about the evolutionary and mechanistic processes that led to its emergence. Studying *Plasmodium* infections of great apes is challenging because of their endangered status, which precludes interventions that could cause or risk harm. Here, we describe a new approach that generates high quality *Plasmodium* genome sequences from samples that contain large amounts of contaminating host DNA. This provides an opportunity to characterize additional *Laverania* species, in particular the gorilla precursor of *P. falciparum*, by selectively amplifying parasite genomes from small quantities of unprocessed blood or even infected mosquito DNA (*100*, *101*). While the adaptive pathways required for the colonization of humans remain to be elucidated, it is tempting to speculate that the horizontal gene transfer of RH5, which represents a major *P. falciparum* host specificity determinant (*102*), conferred a fitness advantage that may have predisposed *P. praefalciparum* to infect humans. However, even if this was the case, HGT alone was clearly not sufficient since all characterized strains of *P. praefalciparum* carry this locus, yet there is evidence for only a single cross-species transmission event (*26*, *32*, *103*). Moreover, the HGT likely occurred long before the emergence of *P. falciparum* (Fig. 3-4B). Previous attempts to date the last common ancestor of *P. falciparum* have yielded estimates of up to several hundred thousand years ago (*29*, *104*), but all of these made assumptions concerning the *Plasmodium* molecular clock that can not be substantiated. In contrast, other data, including the timescale of the spread of *P. falciparum* resistance mutations in African populations (*105*), the evolutionary history of its main mosquito vector *Anopheles gambiae* (*106*), and the low probability of maintaining endemic *P. falciparum* infections in human hunter-gatherer populations (*107*, *108*) support a much more recent emergence or expansion of

P. falciparum, within the last 10,000 years. Our findings of a 10-fold lower within-species diversity in *P. falciparum* compared to *P. gaboni* and *P. reichenowi* is consistent with the latter estimate. Thus, the event(s) that promoted the emergence of *P. falciparum* in humans may have been associated with the transition from a hunter-gatherer to a more settled lifestyle, possibly involving a change in mosquito host preference, such as the ability to efficiently infect the main human vector *A. gambiae* (*109*).

3.5 Methods

Ape samples.

Blood samples (5-10 ml) were collected from sanctuary chimpanzees (*Pan troglodytes)* living in outside enclosures in close proximity to wild apes at the Sanaga Yong Chimpanzee Rescue Center (SY) in Cameroon (n=26) and the Tchimpounga Chimpanzee Rehabilitation Center (TC) (n=1) in the Republic of the Congo. Members of both the central (*P. t. troglodytes*) and the Nigeria-Cameroonian (*P. t. ellioti*) subspecies were sampled. Blood was obtained for veterinary purposes only or represented leftover specimens from yearly health examinations. None of the chimpanzee exhibited symptoms of malaria at the time of sampling. Most blood samples were preserved in RNAlater (1:1 vol/vol) without further processing, except for 8 samples, which were subjected to density gradient centrifugation in the field to enrich for red blood cells (RBC) (Supplementary Table 6). Briefly, blood was diluted in PBS (1:1 vol/vol), layered over Lymphoprep (Axis-Shield), and then centrifuged at 800xg for 20 minutes. After removal of the mononuclear cell layer, the purified erythrocytes were preserved in RNAlater (1:1 vol/vol). All samples were transported at ambient temperature and subsequently stored at -80 °C. Small quantities of blood were also obtained from two western gorillas (*Gorilla gorilla*) of unknown geographic origin (SA), who were killed by hunters and confiscated

by the anti-poaching program of the Cameroonian Ministry of Environment and Forestry. Blood was collected from around the inflicted wounds and frozen directly without preservation. Ape fecal samples (n=55) were selected from an existing bank of chimpanzee and western gorilla specimens previously shown to contain *Laverania* parasite DNA(*26*, *80*, *110*). These specimens were collected from non-habituated apes living in remote forest areas, with a two-letter-code indicating their field site of origin as previously reported (*26*, *110*). DNA was extracted from whole blood and RBCs using the QIAmp Blood DNA Mini Kit, the Puregene Core Blood Kit (Qiagen), or the NucliSENS miniMag extraction kit (Biomérieux). All samples were shipped in compliance with Convention on International Trade in Endangered Species of Wild Fauna and Flora regulations and country specific import and export permits.

***Laverania* species identification.**

The *Laverania* species composition of ape blood and fecal samples was determined by limiting dilution PCR (also termed single genome amplification) and phylogenetic analysis as previously described (*26*, *110*, *111*). Briefly, DNA was endpoint diluted such that fewer than 30% of PCR reactions yielded an amplification product (according to a Poisson distribution, a well yielding a PCR product at this dilution will contain only a single DNA template more than 83% of the time) (*111*). Amplification products were gel purified, and sequenced directly without interim cloning. Sequences containing double peaks, indicative of the presence of multiple templates or early PCR errors, were discarded. In addition to yielding an accurate representation of the *Plasmodium* species present in the sample, this approach generates sequences devoid of *Taq* polymerase induced misincorporations and recombination. Samples were analyzed at mitochondrial, nuclear, and apicoplast loci (Supplementary Tables 1 and 8), including portions of

cytochrome B (*cytB*), the erythrocyte binding antigens 165 and 175 (*EBA165*, *EBA175*), the gametocyte surface proteins P47 and P48/45 (*P47*, *P48/45*), the lactate dehydrogenase (*ldh*), the reticulocyte-binding protein homolog 5 (*RH5*), the cysteine-rich protective antigen (*CyRPA*), members of Phe-Ile-Lys-Lys (FIKK) containing protein kinase multigene family (*FIKK7.2, FIKK14* and *FIKK9.15)*, and the caseinolytic protein C (*clpC*) gene. Primers and PCR conditions have been described (*26*, *80*), except for those used for the amplification of *CyRPA* and *FIKK* genes. CyRPA gene fragments (461-792bp) were amplified using CyRPA_F1 (5'-TTTYATTTTTTCAAATTGTCTTAGTT-3') and CyRPA_R1 (5'-ATGTCTCGCCYTTGTCGTG-3') in the first round, and CyRPA_F2 (5'-GTCRTCATGTTTTYATAAGGACTG-3') and CyRPA_R2 (5'-CCATACATAAAATGTCATCCTTCTT-3') in the second round of PCR, or CyRPA5F1 (5'-AAGGACTGARTTRTCGTTYRTAAAG-3') and CyRPA5R1 (5'-AACKTYCCTCCATARCAACCT-3') in the first round, and CyRPA5iF2 (5'-TARTGTTCCTTGTRTTSGKGATAT-3') and CyRPA5iR2 (5'-ATCMCCYACATAAAAATGAAATGAC-3') in the second round of PCR. The FIKK7.2 fragment (637bp) was amplified using FIKK7.2_F993 (5'-AAGATTCCTATTARTGCATGGRTAAA-3') and FIKK7.2_R1782 (5'-ATGATGGATCAGAACGCTTCC-3') in the first round, and FIKK7.2_F1061 (5'-AAATGCTGAAAATTATGTTATGGAAG-3') and FIKK7.2_R1724 (5'-GATYCCCAACATATATTTATCAACTG-3') in the second round of PCR. The FIKK14 fragment (537bp) was amplified using FIKK14_F1280 (5'-TGAAATGTAGAAGTAGATTAGCAA-3') and FIKK14_R1965 (5'-GTGTTAAACCTGCTTCATGTAATCTT-3') in the first round, and FIKK14_F1321 (5'-ACTGTATATAATTGGACRTTAGGTAA- 3') and FIKK14_R1884 (5'-CTAAATCATCATCATCATCATCCATA-3') in the second round. Finally, the FIKK9.15

fragment (730-733bp) was amplified using PgSY75FIKK_F1 (5'-CGGATAGAGATGACGTTTCACA-3') and PgSY75FIKK_R1 (5'-AAGGCACATGCCTCCATAATA-3') in the first round, and PgSY75FIKK_F2 (5'-ACAGGAGATAATGGAGGAAATGTAG-3') and PgSY75FIKK_R2 (5'-CCTACCACGTTTACTAAGTCCAATA-3') in the second round of PCR. For each sample, multiple single template-derived amplicons were sequenced and their species origin identified by phylogenetic analysis (see GenBank accession numbers in Supplementary Table 8). This analysis permitted the identification of samples that represented single (or near single) *Laverania* species infections for selective whole genome amplification (Supplementary Table 1).

Laverania specific real time PCR.

To determine the amount of *Laverania* DNA within a blood or fecal sample, DNA was subjected to quantitative (q)PCR using a 7900HT Fast Real-Time PCR System and the Power SYBR Green qPCR kit (Life Technologies). *Laverania* specific forward (5'-ACATGCCACATGGAAAAGCTT-3') and reverse (5'-CTGGGGCCTTGGTAAATCCA-3') primers were used to amplify a 144 bp fragment of the nuclear *ldh* gene. PCR cycling conditions included 2 minutes at 50 °C, 10 minutes at 95 °C, and 40 cycles of 15 seconds at 95 °C and 1 minute at 60 °C. To estimate the number of genome copies per well, human genomic DNA containing known quantities of purified *P. falciparum* 3D7 DNA was used to generate a standard curve, which was included on all qPCR plates (Supplementary Table 1).

Design of SWGA primers.

In contrast to traditional phi29 whole genome amplification methods that use random

primers to amplify all DNA templates within a sample, selective whole genome amplification requires primers that bind frequently and evenly across the pathogen genome, but only rarely to the contaminating host DNA. To identify such primers, we used a sliding window to determine the frequency of all short sequence motifs (8-12 bp in length) in both a *P. falciparum* (3D7) and human (GRCh37) reference sequence and then calculated the average distance between their locations within these genomes (Supplementary Fig. 1). This approach identified 2,418 motifs that were spaced apart (on average) less than 50 kb in the *P. falciparum*, but more than 500 kb in the human genome (Fig. 4-1A). To select the best possible primers, motifs with a melting temperature (T_m) below 18 °C and above 30 °C were discarded because they were unlikely to properly anneal to the template DNA. Motifs that contained 4 or more contiguous self-complementary bases were also eliminated to avoid the formation of homodimers. Finally, motifs predicted to bind greater than 3 times to human mitochondrial DNA were eliminated, since this circular genome would be disproportionally targeted by phi29 for "rolling-circle" amplification (*75*). These criteria identified 149 potential SWGA primers.

In a previous study, we found that motifs that exhibited the highest target-to-nontarget binding ratios were able to mediate selective amplification of bacterial genomes from infected host DNA (*39*). However, it was unclear whether this criterion alone would be sufficient for more complex (multichromosomal) eukaryotic genomes. To design primers capable of amplifying all regions of the *Plasmodium* genome, we developed a metric that scored both selectivity and evenness of coverage. To score a set of primers, we divided the *P. falciparum* and human genomes into 10-kb non-overlapping segments and calculated the proportion of segments that contained at least one primer-binding site (Supplementary Fig. 1). Since our goal was to identify primer

binding sites in as many *P. falciparum* segments as possible, while minimizing segments containing the same binding site in the human genome, we defined our "set score" as the difference between the former and the latter (Equation 1). The complete *P. falciparum* and human genomes, including telomeric sequence, were used for the calculation.

$$Pf_p = \text{Proportion of 10-kb sites in } \textit{P. falciparum} \text{ genome containing at least 1 primer binding site}$$

$$Hu_p = \text{Proportion of 10-kb sites in the human genome containing at least 1 primer binding site}$$

$$\text{Set Score} = Pf_p - Hu_p$$

Equation 1: Scoring metric for SWGA primer sets

Starting with a set of 149 primers, there are a total of 1.2 x 10^{15} possible combinations of 10 or fewer primers. Since identifying the single best set would be computationally impossible, we used a heuristic approach to search for optimal primer combinations. Reasoning that heterodimer formation would reduce amplification efficiency, we divided the 149 primers into eight mutually exclusive groups, where no two primers contained 4 or more contiguous complementary bases. For each group, we first scored primers individually (using Equation 1), and selected the highest scoring primer. We then paired this primer with all other primers and identified the highest scoring pair. This process was repeated by iteratively adding primers until the set score no longer improved (Supplementary Fig. 1). Applying this approach to all primer groups generated eight high scoring sets. The two best sets (6A and 8A) were then tested using human genomic DNA containing known quantities of *P. falciparum* DNA (a primer design pipeline that is applicable to the genomes of other organisms is available upon request).

SWGA protocol and validation of primer sets.

SWGA was performed as described (*39*), essentially following previously published phi29 amplification protocols, but using primers designed to selectively amplify *Laverania* genomes (Fig. 4-1). Amplification conditions included a one-hour ramp down step (35 °C to 30 °C), followed by a 16 hour amplification step at 30 °C. Phi29 was then denatured for 10 min at 65 °C, and the SWGA product was stored at 4 °C. To validate the SWGA primers, genomic DNA extracted from cultured *P. falciparum* (3D7) parasites and human CD4+ T cells were mixed to generate human DNA preparations containing 5%, 1%, 0.1%, 0.01%, and 0.001% *P. falciparum* DNA. SWGA was performed in a volume of 50μl using 50 ng of DNA, 3.5 mM of each SWGA primer (set 6A), 1x phi29 buffer (New England Biolabs), 1 mM dNTPs, and 30 units of phi29 polymerase (New England Biolabs). 4 ul of the resulting SWGA product was then subjected to a second round of SWGA using the same amplification conditions, except for using a different set of primers (set 8A). Each of the human/Pf mixtures was amplified separately and purified using Agencourt AmpureXP beads (Beckman Coulter). 20 ng of the resulting SWGA products were used to generate short-insert libraries (Nextera Library Prep Kit) and sequenced on an Illumina MiSeq, yielding 150 bp paired reads. Enrichment was quantified by mapping paired reads first to the human and then to the *P. falciparum* 3D7 genome using SMALT 0.7.6 (https://www.sanger.ac.uk/resources/software/smalt/) and then calculating the percentage of reads that mapped to *P. falciparum* 3D7. This analysis showed that the SWGA method amplified *P. falciparum* with extraordinary selectivity, resulting in an up to 70,000-fold enrichment of parasite over host DNA (Table 3-1) while maintaining an even coverage of all chromosomes, except for sub-telomeric regions where extremely high AT-content precluded accurate mapping (Supplementary

Fig. 2).

To determine the efficiency of SWGA, we performed a rarefaction analysis, examining both the selectivity and evenness of amplification for different ratios of host/parasite DNA. For each human/Pf DNA mixture, subsets of reads were randomly selected and mapped to the *P. falciparum* (3D7) and human reference genomes (GRCh37) simultaneously. The percent of the *P. falciparum* genome with ≥1x coverage was then calculated and compared to the expected coverage of the same unamplified human/Pf mixture (Fig. 4-1C). This analysis showed that the SWGA approach was able to dramatically decrease the sequencing effort required to obtain broad coverage of the *P. falciparum* core genome, with little to no coverage loss when applied to samples containing <0.01% *P. falciparum (*Fig. 4-1C).

Selective amplification of *P. reichenowi* and *P. gaboni* genomes.

To amplify near-full-length *Laverania* parasite genomes from unprocessed ape blood, we selected one chimpanzee sample (SY57) that contained mostly (>99%) *P. reichenowi* and two others (SY75 and SY37) that contained exclusively *P. gaboni* DNA for SWGA analysis (Supplementary Table 1). Since these samples contained very little *Laverania* DNA (0.00081%-0.14%), we first digested them with methylation dependent restriction enzymes (MspJI and FspEI) to selectively cleave the contaminating host DNA (*38*). Briefly, 200ng - 1μg of total DNA were digested with FspEI (5U) and MspJI (5U) for 7 hours at 37 °C, after which the enzymes were heat inactivated. The digestion products were purified and subjected to two successive rounds of SWGA using the same conditions as described above. For each chimpanzee sample, SWGA was performed using multiple DNA replicates (Supplementary Table 1), with half being first amplified

with primer set 8A followed by primer set 6A, and the other half being first amplified with primer set 6A followed by primer set 8A. Amplification products were purified, pooled, and used to generate short-insert libraries (650 bp) using the Illumina TruSeq PCR-Free Library Preparation Kit (Supplementary Table 1). To facilitate subsequent genome assembly, we also generated long insert libraries (3 kb, 5 kb, 8 kb and 9 kb) for the *P. gaboni* sample SY75 (Illumina Nextera Mate Pair Sample Preparation Kit). All libraries were sequenced using the Illumina MiSeq and paired reads were first mapped to the chimpanzee reference genome (Pan_troglodytes-2.1.4) using SMALT. The remaining reads were then mapped to the *Plasmodium* genome, with Pf3D7 serving as the reference for SY75 and SY37, and PrCDC serving as the reference for SY57. Although SY75 (73%) and SY37 (61%) yielded fewer parasite-specific reads than SY57 (89%), this was not due to a reduced amplification selectivity, but reflected the difficulty of mapping *P. gaboni* reads to the much more divergent *P. falciparum* genome (Supplementary Table 1).

Assembly of *P. gaboni* and *P. reichenowi* draft genomes.

Draft genomes were generated for the *P. reichenowi* strain PrSY57 and the *P. gaboni* strain PgSY75 using reference guided *de novo* assembly with post-assembly genome improvements (*69*, *81*). First, working drafts of the PrSY57 and PgSY75 genomes were generated by iteratively mapping (non-chimpanzee) reads to the PrCDC and Pf3D7 references, respectively, using Geneious 6 (Biomatters Limited, http://www.geneious.com). This mapping process, which was repeated 10 times, resulted in a sequence that represented the read mapping consensus at all positions with ≥5 fold coverage. At positions with lower coverage, the sequence of the reference (Pf3D7 or PrCDC) was used instead. All reads were then re-mapped to this consensus

using two iterations. The resulting draft reference represented the mapping consensus at all positions with ≥5 fold coverage, with positions with <5 fold coverage denoted by "N"s.

Prior to *de novo* assembly, error correction was performed on short-insert libraries from each sample using String Graph Assembler (SGA 0.10.12) (*112*) as previously described (*69*). For the *P. gaboni* sample PgSY75, reads were also normalized using KHMER (*113*), which uses k-mer frequencies to estimate and normalize genome coverage in a reference-free manner, thus facilitating subsequent *de novo* assembly. This process yielded 11 million reads.

After mapping reads to the working draft reference using SMALT, a reference guided *de novo* assembly was generated using the Columbus extension to Velvet 1.1.06 (*114*). Assemblies were produced using a variety of k-mer lengths and coverage settings. Comparing these assemblies to the Pf3D7 and PrCDC references, we identified several tandem duplications, which upon visual inspection were judged to likely represent assembly errors. We thus changed the assembly parameters to minimize the number of these duplications. Specifically, we varied k-mer length, coverage cutoff, and minimum paired coverage, and analyzed the resulting assembly quality by comparing the length of contigs, maximum node length, total assembly length, and the number of tandem duplications compared to the reference genome.

For the *P. gaboni* sample PgSY75, contigs produced by Velvet Columbus were further scaffolded using long insert libraries with SSPACE 2.0 (*115*). Scaffolding was performed iteratively, first using the 3 kb library, then the 5 kb library, and finally the 8 kb and 9 kb libraries. Scaffolding was performed using default parameters, except for (i) a minimum number of mate pairs (-k) of 10 for the 3 kb library and 5 for the 5 kb, 8 kb and 9 kb libraries, respectively, (ii) a maximum ratio between the two best pairs (-a) of 0.6, (iii) a minimum required overlap (-n) of 60 bp, and (iv) a minimum contig size (-z) of 500.

Scaffolding was not performed for the *P. reichenowi* PrSY57 because long insert libraries were not generated for this sample.

To improve the quality of the draft references, contigs and scaffolds produced by Velvet Columbus and SSPACE were subjected to two iterations of post-assembly improvement using PAGIT v1 (*81*). Contigs were aligned against the respective reference genomes using ABACAS 1.3.1 and joined into a single ordered sequence separated by gaps ("N"s). The resulting ordering was compared to the reference genome using blastn to identify erroneously placed contigs. ABACAS parameters for minimum percent identity (-i) and minimum contig coverage (-v) were varied to maximize the total number of correctly placed contigs (e.g., -i 90 was used to minimize *P. gaboni* contamination in the *P. reichenowi* SY57 assembly). Contigs were then manually rearranged in the Artemis Comparison Tool (ACT) (*116*) to correct any remaining placement errors. Gaps between contigs were closed using gapfiller 1.10 (*117*) and IMAGE 2.4.1 (*118*). Since the closing of gaps also produced tandem duplications, parameters for gapfiller and IMAGE were varied to minimize the number of duplications and maximize the number of gaps closed.

Mapping paired reads to the improved draft genome identified several instances where Velvet or gap-closure produced erroneously assembled sequence. Since read coverage is often reduced on both sides of an assembly error, we calculated the mean read coverage for a 1,750 bp window surrounding these positions (the central 750 bp which were slightly larger than the library insert size were excluded from these calculation). We then broke the draft genome into contigs at positions where the coverage was either below five paired reads or 10% of the mean coverage of the 1,750 bp window, and repeated the process of contig ordering and gap closure using the broken contigs, varying the same parameters as before.

The ordered, gap-closed, draft genome produced by PAGIT was corrected using iCORN 2 (*119*), which corrects SNP and indel errors based on the read consensus. We ran iCORN iteratively until no additional corrections of the genome were required. The final output was designated version 0.1 of both the PgSY75 and PrSY57 draft chromosomal assemblies, with all additional edits made after manual inspection during gene annotation and subsequent analyses.

Generation of PgSY75 and PrSY57 unplaced read bins.

All contigs that could not be placed into chromosomal scaffolds of an assembly during the PAGIT process were put into an unplaced-read-bin (version 0.1). These "bins" were then expanded using *de novo* assemblies of (non-chimpanzee) reads that failed to map to both the chromosomal assembly and the v0.1 bin. This was done by mapping all reads from PgSY75 and PrSY57 to their respective draft assembly and bin using SMALT, and then performing *de novo* assembly of the remaining unplaced read pairs using SPAdes 3.1.1 (*120*). SPAdes was run using the default multicellular mode parameters, except for the k-mer length (-k) that was set to 21, 33, 55, and 77. For PgSY75, the resulting contigs were corrected using iCORN (*119*), using only unmapped read pairs from the previous step and added to the unplaced bin. For PrSY57, the combined unplaced bin contigs were screened for contaminating *P. gaboni* sequences by performing blastn searches to a combined database of PrCDC, Pf3D7 and PgSY75 chromosomes. Contigs were only retained if their best match was to a *P. reichenowi* contig, exhibited ≥90% identity, and had e-value $\leq 10^{-15}$. Duplicated contigs, which had been assembled erroneously due to the presence of inter-strain polymorphisms or sequencing error, were initially merged by running dipSPAdes (*121*) on the combined unplaced contig bins for each draft assembly, using haplocontig mode. Each unplaced

contig in the reduced bin was then compared to the chromosomal assembly and other unplaced bin contigs using blastn, and those that were >85% identical to chromosomal or bin contigs were aligned to their match, visually inspected, and either removed or used to improve the existing assembly. The resulting de-duplicated bin was combined with the v0.1 draft chromosomal assembly and designated the v0.1 draft genome.

Annotation of the PgSY75 and PrSY57 draft genomes.

Annotations were transferred to the PgSY75 and PrSY57 draft genomes from *P. falciparum* (Pf3D7) and *P. reichenowi* (PrCDC) reference genomes, respectively. Annotation transfer was performed using RATT (*122*) and corrected manually in ACT (*116*) using a blastn alignment to the corresponding reference. Genes in the draft genomes that were not present in Pf3D7 or PrCDC, or had been missed by RATT, were identified by *de novo* annotation in Augustus (*123*) using the *Plasmodium falciparum* species configuration. *De novo* annotations that overlapped transferred annotations were removed. The remaining *de novo* annotations were compared with their reference strains using blastn and tblastx to identify putative orthologs and homologs, and corrected by visual inspection. Annotations for which no homolog could be identified in the reference were compared individually with all available *Plasmodium* genomes, and deleted if no putative homolog could be found.

Generation and annotation of the PgSY37 draft genome.

Because the small amounts of *P. gaboni* DNA present in sample SY37 resulted in greater unevenness of whole genome amplification and sequence coverage, the PgSY37 draft genome was assembled by iteratively mapping the SWGA generated sequencing reads to the PgSY75 genome, using the same methods and parameters

described above. Unplaced reads were assembled using SPAdes (*120*) and placed into the PgSY37 unplaced read bin. The PgSY37 genome was annotated by strain level annotation transfer from the PgSY75 genome using RATT (*122*), and corrected by visual inspection.

Limitations of SWGA to generate *Plasmodium* genome sequences.

SWGA generated high quality *Plasmodium* genomes when samples were well preserved, comprised *Laverania* mono (or near mono) infections, and contained >0.0005% of parasite DNA. Although even a relatively low parasite load (e.g., 0.0054% in SY57) yielded broad coverage, coverage depth varied due to the process by which SWGA enriches for target DNA. Very low levels of parasite DNA (e.g., 0.00081% in SY37) led to stochastic peaks of amplification scattered across the genome. Moreover, partial sample degradation impeded SWGA. For example, two gorilla blood samples (SA3066 and SA3157), which were frozen without preservation, failed to yield genome sequences following SWGA and MiSeq sequencing, despite the presence of seemingly sufficient (albeit low level) quantities of parasite DNA (0.00073% and 0.00024%, respectively). Thus, SWGA requires not only a minimum amount of parasite DNA, but also high molecular weight *Plasmodium* genomes for efficient amplification. Nonetheless, the SWGA products of the two gorilla blood samples yielded nuclear gene fragments (*CyRPA, FIKK*) when subjected to conventional single template PCR, indicating selective amplification even of partially degraded parasite DNA.

Genes used in genome-wide analyses.

Syntenic orthologs in *P. falciparum* 3D7 and *P. reichenowi* CDC were identified by chromosomal alignment. After exclusion of (i) *var*, *rif*, and *stevor* gene families, (ii) genes

that were pseudogenes in at least one of these *Laverania* species (Supplementary Table 2), (iii) genes that had previously been suggested to be dimorphic in *P. falciparum* (*124*, *125*) (*msp1*, *msp2*, *msp3*, *msp6* and *EBA175*), and (vi) genes for which orthologs could not be identified in *P. gaboni* (Supplementary Data 1), the remaining sets of orthologs were used for genome-wide analyses. Subtelomeric regions, which were excluded from *P. falciparum* polymorphism data, were defined as regions at the ends of chromosomes that consisted primarily of genes previously annotated as subtelomeric or members of subtelomeric gene families, including *var*, *rif*, *stevor*, *PHIST*, *mc-2tm, hyp* gene families 1-17, *resa*, lysophospholipase, *DNAJ* and acyl-coA synthetase. Subtelomeric genes are identified in Supplementary Data.

Inter-species divergence.

The lengths of coding sequences from the annotated genomes were compared with their homologs or orthologs in the respective reference sequence (PrCDC for PrSY57, Pf3D7 for PgSY75 and PgSY37). Genes were only included in genome-wide analyses if they (i) were ≥90% of the length of the reference homolog/ortholog or (ii) were ≥80% of the length of the reference ortholog/homolog, but also lacked assembly gaps. Each coding region was translated and queried for amino acid repeats using tblastx. Repeated sequences were masked if they comprised at least 20 amino acids with at least 95% identity between repeat units. Low-complexity amino acid sequences were identified in translations using segmasker (NCBI BLAST+ package) using default settings, and masked in the corresponding nucleotide sequences. Masked nucleotide sequences were aligned using TranslatorX (*126*) and MUSCLE (*127*). After alignment, any position that was masked, or contained an assembly or alignment gap, was masked in all sequences. Pairwise inter-species genetic distances were calculated in R using the ape package

(*128*) with the TN93 model of DNA evolution. Genes with unusually high inter-species distances were manually inspected and corrected if necessary. If the best alignment required insertion of a gap not divisible by three, the gene was excluded from intra-species diversity analyses (since these required sequence translations). Inter-species distances were calculated using all available orthologs for Pf3D7, PrCDC, PrSY57, PgSY75 and PgSY37.

Intra-species diversity.

For *P. falciparum*, intra-species diversity was calculated using previously published parasite sequence datasets (Table 3-3) of geographically diverse field isolates collected in Bangladesh, Cambodia, DRC, Gambia, Ghana, Guinea, Laos, Myanmar, Nigeria, Thailand, Kenya, and Vietnam (Pf3k 1.0 pilot data release, http://www.malariagen.net/data/pf3k-1). For each country, three samples were chosen at random, reads were mapped to the 3D7 reference, and SNP variant calls were generated for all *P. falciparum* strains simultaneously using the GATK 3.1-1 UnifiedGenotyper after indel realignment (*129-131*). To differentiate true variants from sequencing or alignment artifacts, 354 variant calls were randomly selected and true variants identified by visual inspection of the alignments. The GATK values (QUAL, QD, ReadPosRankSum, Genotype Quality, FS, BaseQRankSum, MQRankSum) were then compared for each true and artifactual variant, and appropriate cutoffs were selected to minimize false variant calls. Using only SNPs from the core genome, the number of *P. falciparum* strains present in each sample was estimated using estMOI (*132*), with one likely mono-infection selected for each country (ERS174561, Bangladesh; ERS050887, Cambodia; ERS347597, DRC; ERS010044, Gambia; ERS157479, Ghana; ERS042044, Guinea; ERS174601, Laos, ERS143480, Myanmar; ERS199640, Nigeria; ERS224908,

Thailand; ERS143467, Vietnam). No Kenyan strain was selected since all available samples were likely to represent multi-strain infections. After exclusion of subtelomeric genes, alleles from polymorphic sites were extracted from variant call format (vcf) files using custom Perl scripts. Sites at which three or more samples had missing data (i.e. no genotype called) or where the majority genotype was represented by <80% of mapped reads, were excluded from the analysis; otherwise samples with missing data were assumed to have the reference allele. Intra-species diversity (π) was determined by calculating the mean number of differences per site for all pairwise combinations of 11 *P. falciparum* strains plus the 3D7 reference. Sites masked in 3D7 (see above) were excluded from intra-species diversity calculations.

For *P. gaboni* and *P. reichenowi*, intra-species diversity was calculated from the alignments used for inter-species genetic distance calculations, using the ape R package to count the proportion of non-masked sites that differed between the two strains available for each species (PrCDC and PrSY57 for *P. reichenowi*, PgSY75 and PgSY37 for *P. gaboni*).

Phylogenetic Analyses.

Nucleotide sequences used for phylogenetic analyses were aligned using CLUSTAL W (*133*), followed by manual correction when necessary. Regions that could not be unambiguously aligned were removed from further analyses. Maximum likelihood phylogenetic analyses were conducted using PhyML (*134*), with iterative model fitting (*52*) based on a class of evolutionary models selected using Modeltest (*135*). For the analyses of the *FIKK* orthologs, pseudogene nucleotide sequences were translated, with indels corrected and in-frame stops coded as “X”, and the deduced amino acid sequences were aligned using MUSCLE (*127*). Based on this alignment, the conserved

FIKK protein regions were identified. The corresponding nucleotide sequences were then codon aligned, guided by the amino acid alignment. To eliminate possible mutational saturation at third codon position sites, these were removed prior to phylogenetic analyses using PhyML.

3.6 Acknowledgements

We thank Richard Carter and Brian Charlesworth for helpful discussions, the staff of project PRESICA for fieldwork in Cameroon; the staff of the Sanaga Yong Rescue Center and the Tchimpounga Chimpanzee Rehabilitation Center for providing left-over blood samples from captive chimpanzees; the Cameroonian Ministries of Health, Forestry and Wildlife, and Scientific Research and Innovation for permission to collect samples in Cameroon; the Ministry of Forest Economy and Sustainable Development for permission to collect samples in the Republic of Congo; This work was supported by grants from the National Institutes of Health (R01 AI097137, R01 AI091595, R37 AI050529, T32 AI007532, P30 AI045008), the Burroughs Wellcome Fund (1012376), the Agence Nationale de Recherche sur le Sida (ANRS 12125/ 12182/12255), the Agence Nationale de Recherche (Programme Blanc, Sciences de la Vie, de la Santé et des Ecosystémes and ANR 11 BSV3 021 01, Projet PRIMAL), and the Wellcome Trust (098051).

CHAPTER 4 - Ape parasite origins of human malaria virulence genes

Daniel B. Larremore, Sesh A. Sundararaman, Weimin Liu, William R. Proto, Aaron Clauset, Dorothy E. Loy, Sheri Speede, Lindsey J. Plenderleith, Paul M. Sharp, Beatrice H. Hahn, Julian C. Rayner, and Caroline O. Buckee

Originally published in Nature Comms. 6:8368, 2015.

Supplemental data are available at

http://www.nature.com/ncomms/2015/151012/ncomms9368/extref/ncomms9368-s1.pdf

Daniel B. Larremore, William R. Proto, Caroline O. Buckee and Julian C. Rayner. conceived the study. Weimin Liu, William R. Proto, Dorothy E. Loy, Lindsey J. Plenderleith, Paul M. Sharp, Beatrice H. Hahn, Julian C. Rayner, and I characterized ape *Laverania* infections. Lindsey J. Plenderleith, Paul M. Sharp, Beatrice H. Hahn, and I amplified, sequenced, and analyzed near complete genomes of *P. gaboni*. I identified, curated, and analyzed *var*-like genes and DBL, ATS, and TM domains in the *P. gaboni* genomes. Daniel B. Larremore, Aaron Clauset, and Caroline O. Buckee performed network, phylogenetic, pairwise distance, k-mer and CP analyses. Daniel B. Larremore, Paul M. Sharp, Beatrice H. Hahn, Caroline O. Buckee, and I wrote the paper with contributions from all authors. Sheri Speede provided chimpanzee blood samples that were collected opportunistically during health screens or for specific veterinary purposes. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

4.1 Abstract

Antigens encoded by the *var* gene family are major virulence factors of the human malaria parasite *Plasmodium falciparum*, exhibiting enormous intra- and inter-strain diversity. Here we use network analysis to show that *var* architecture and mosaicism are conserved at multiple levels across the *Laverania* subgenus, based on *var*-like sequences from eight single-species and three multi-species *Plasmodium* infections of wild-living or sanctuary African apes. Using select whole-genome amplification, we also find evidence of multi-domain *var* structure and synteny in *Plasmodium gaboni*, one of the ape *Laverania* species most distantly related to *P. falciparum*, as well as a new class of Duffy Binding-Like (DBL) domains. These findings indicate that the modular genetic architecture and sequence diversity underlying *var*-mediated host-parasite interactions evolved prior to the radiation of the *Laverania* subgenus, long before the emergence of *P. falciparum.*

4.2 Introduction

Wild-living apes in Africa are naturally infected by at least six *Plasmodium* species that form a separate subgenus, termed *Laverania* (*12*, *25*, *27-32*, *110*, *136*). Three of these species, *P. reichenowi, P. gaboni*, and *P. billcollinsi*, have been found only in chimpanzees, while the other three, *P. adleri, P. blacklocki*, and *P. praefalciparum*, have been found only in gorillas (Fig. 4-1A). Zoonotic transfer has occurred at least once, when a gorilla parasite (*P. praefalciparum*) gave rise to human *P. falciparum*, which causes the vast majority of malaria-associated morbidity and mortality in humans (*32*, *110*).

A key component of *P. falciparum* virulence is the parasite's ability to cause infected erythrocytes to adhere to the vascular endothelium. This allows the parasite to escape elimination in the spleen but can also lead to vascular obstruction and inflammation, key components of severe pathological complications such as cerebral malaria (*137*, *138*). Cytoadherence is mediated by members of the *P. falciparum* Erythrocyte Membrane Protein 1 (PfEMP1) family, which contain between 3 and 8 different Duffy Binding Like (DBLα-ζ) and Cysteine-rich Interdomain Region (CIDRα-δ) domains and are expressed on the surface of infected erythrocytes, where they bind to endothelial receptors. Each *P. falciparum* genome encodes approximately 60 different PfEMP1 proteins, which are expressed from *var* genes, one at a time, by means of epigenetic regulation (*139*, *140*). Given their central role in *P. falciparum* pathogenesis, but absence from all other human *Plasmodium* species, the origins of *var* genes are of particular interest.

Three factors have limited our ability to investigate the evolutionary history of *var* genes. First, obtaining blood samples from *Laverania*-infected wild-living apes is not ethical. As a result, all ape derived *var* sequences analyzed to date come from a single

P. reichenowi parasite, called PrCDC, from a wild-born chimpanzee, who was found to be *Plasmodium* infected in captivity (*69*). Second, *P. falciparum var* genes are highly diverse (Fig 4-1B). Not only is there rapid recombination between genes within and across chromosomes, which shuffles gene content within genome repertoires during infection (*141*, *142*), but sexual reproduction in the mosquito vector also generates diversity via reassortment of chromosomes and conversion events (*143*). Thus, conventional phylogenetic approaches fail to resolve evolutionary relationships between *var* genes, requiring new and recombination-tolerant analysis techniques (*144-149*). Finally, the mosaicism and diversity generated by rapid recombination (*141*, *142*), combined with the fact that most *var* genes are subtelomeric, render the assembly of full-length *var* genes from shotgun sequenced parasite genomes extremely difficult (*35*, *150*).

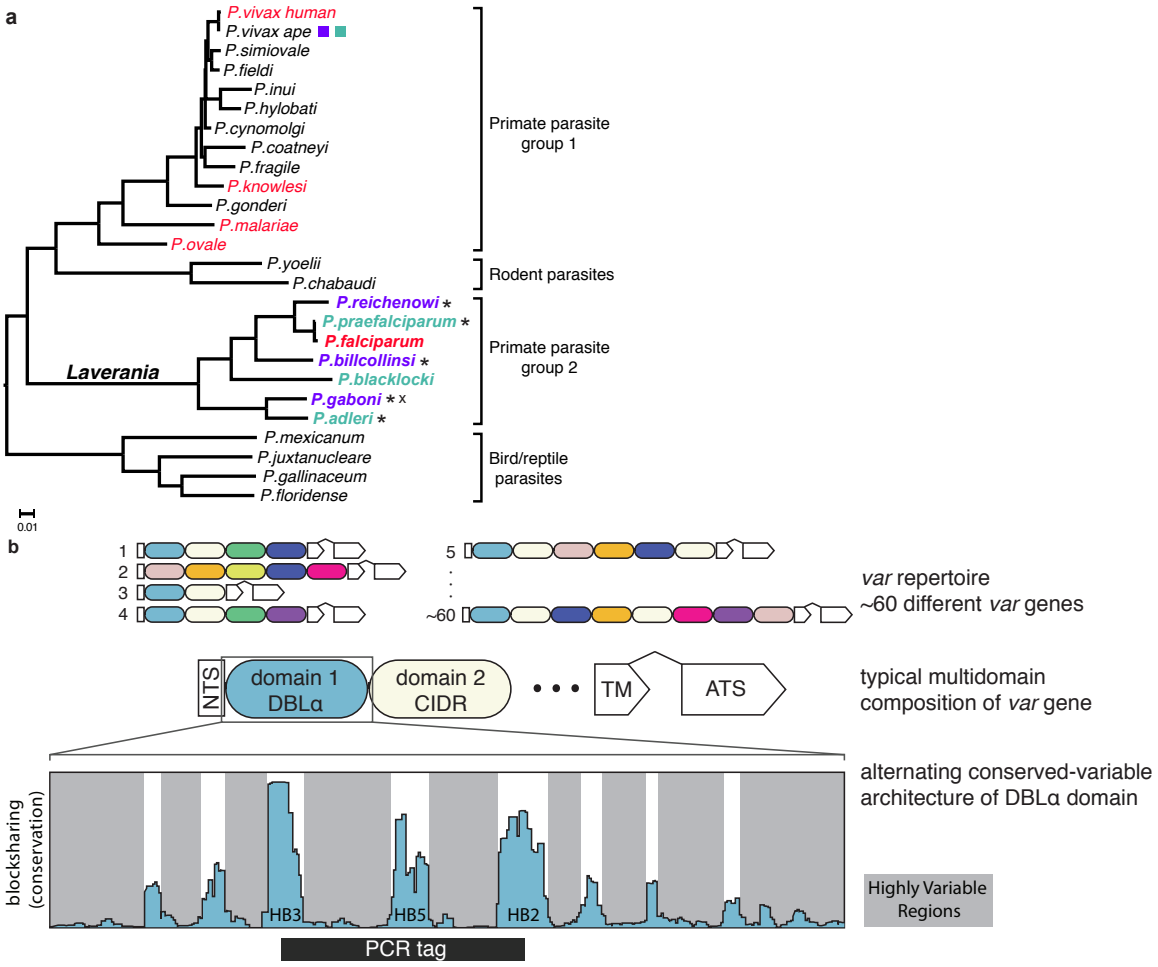

Figure 4-1. Characterization of *Laverania var* gene sequences.

(A) Phylogeny of *Plasmodium* species. The tree was constructed from mitochondrial sequences (2.4 kb spanning *cox*1 and *cyt*B). The scale bar indicates 0.01 substitutions per site. Colors indicate species infecting humans (red), chimpanzees (purple) and gorillas (aqua). Asterisks indicate successful PCR amplification of *var* sequences; a cross indicates identification of *var*-like genes in near full-length *P. gaboni* genomes. **(B)** Three-level schematic of modular *var* diversity, structure, and architecture. Colored ovals represent classes of DBL or CIDR domains. White boxes represent the N terminal segment (NTS), transmembrane (TM), and acidic terminal segment (ATS) domains; a wedge between TM and ATS domains indicates the intron that separates the two *var*

exons. Alternating conserved-variable architecture is illustrated using blocksharing (see Methods) between one representative DBLa domain (DD2var11) and other DBLa domains (*144*). A black bar indicates the location of the PCR amplified DBLa tag region, which spans three conserved homology blocks (HB3, HB5, and HB2) (*144*), 72 to 147 base pairs in length.

Here we overcome these impediments by generating 369 new *var* sequence fragments from five ape *Laverania* species, derived by PCR amplification from fecal and blood samples of naturally infected wild-living and sanctuary apes, respectively. We use network approaches and other recombination-tolerant methods to analyze these new sequences, together with 353 previously reported *var* gene sequences from one *P. reichenowi* and seven *P. falciparum* isolates (*69*, *144*). In addition, we identify and analyze partially assembled *var*-like sequences from otherwise near-full-length genomes of two *P. gaboni* parasites *(*SYpte37, SYptt75), one of the *Laverania* species most distantly related to *P. falciparum* (*39*, *151*). Analysis of these sequences reveals that several PfEMP1 domains, as well as the genetic structure and multi-domain architecture that are characteristic of *P. falciparum var* genes, are present across the *Laverania* subgenus. Thus, many *var* multi-gene family features predate the most recent common ancestor of extant *Laverania* species.

4.3 Results

Laverania species identification and sequence generation

To study *var* gene architecture in ape *Laverania* species, we first determined the *Plasmodium* species composition of eleven blood and fecal samples from sanctuary and wild-living apes using a limiting dilution PCR approach called single genome sequencing

(SGS) (*111*). To ensure amplification of single parasite templates, blood and fecal DNA was diluted such that fewer than 30% of all PCR reactions yielded an amplification product. Amplicons were sequenced directly without cloning into a plasmid vector and sequences containing ambiguous bases indicative of template mixtures were discarded. This approach eliminates *Taq* polymerase-induced recombination (template switching) and nucleotide misincorporations in finished sequences, and also ensures a proportional representation of plasmodial variants as they exist *in vivo* (see Methods for a more detailed description of SGS). Targeting eight different mitochondrial, apicoplast and nuclear loci and sequencing up to 174 different SGS amplicons per sample (Supplementary Table 1), we identified eight samples with single-species infections of *P. reichenowi* (C1), *P. gaboni* (C2), *P. billcollinsi* (C3), or *P. praefalciparum* (G1). Three additional fecal samples represented mixed-species infections of several gorilla or chimpanzee parasites, including one of unknown, non-*Laverania* species origin (Supplementary Table 1).

Given their enormous diversity, *var* homologs were amplified targeting a conserved region of the DBLa domain, termed the *var* gene "tag", using conventional PCR and previously reported primers (*152*, *153*) (see Methods and Supplementary Table 2). Amplicons were cloned, and multiple clones per sample were sequenced and grouped into unique haplotypes by phylogenetic analysis. The *var* gene tag is commonly analyzed because it is sufficiently conserved in two locations to allow reliable amplification, and is located within the DBLα domain, which, unlike other DBL domains, is present in almost all *var* genes (*145-147*, *152-154*). The DBLα tag consists of three conserved homology blocks (*144*) (HBs) interspersed with highly variable regions (HVRs) of diverse length and sequence content (Fig. 4-1B), an architecture that facilitates mosaicism (*146*). Standard sequence analysis techniques can not adequately

analyze these mosaic sequences (*144-149*) and we therefore used a network analysis method to characterize the evolutionary relationships between *Laverania var* fragments. Figure 4-2 illustrates this type of analysis, where each node represents a *var* DBL sequence tag and a link between two nodes represents a shared identical sequence mosaic element. Due to frequent recombination and the possibility that immune selection differs between adjacent HVRs, networks were constructed independently for each of the two HVRs, which in *P. falciparum* were shown to exhibit different community structures (*146*)**.** For each sample, only unique *var* tag haplotypes were included into the analysis (see Methods for a detailed description of network construction and statistical community detection).

Shared *var* mosaic structure in *P. reichenowi* and *P. falciparum*

We first examined the 37 new DBLα *var* tags from a *P. reichenowi* monoinfection detected by routine blood analysis in an asymptomatic sanctuary chimpanzee (SYptt15), who was housed in close proximity to the habitat of wild apes. It is well established that human *P. falciparum* and chimpanzee *P. reichenowi* are closely related sister taxa (*69*), and previous analyses of PrCDC *var* gene sequences indicated sequence homology with field and lab strains of *P. falciparum* (*69*, *145*, *147*, *148*, *155*). While early studies investigated shared polymorphisms in preliminary assemblies of a small subset of these genes (*145*), more recent studies analyzed the complete set of PrCDC DBLα domains, finding conserved gene regions between PrCDC and *P. falciparum* isolates 3D7 and HB3 (*148*) as well as the presence of *P. falciparum* homology blocks in PrCDC DBLα sequences (*155*). In contrast, we focused specifically on the most polymorphic HVR regions of *P. falciparum* and *P. reichenowi* DBLα homologs. Using a network community detection algorithm, a Bayesian *k*-mer analysis, and a pair-wise distance approach, we

found that *var* mosaics within the *P. falciparum-P. reichenowi* network do not cluster by parasite species (Fig. 4-2, Supplementary Fig. 1A,B), and that *var* genes from both species exhibit the same modular HVR architecture, i.e., a pattern of alternating regions of conservation and variability (Supplementary Fig. 1C). We have previously hypothesized that this genetic structure may allow for neighboring HVRs to respond independently to different selection pressures (*146*). Thus, our results confirm and extend previous findings that DBLα organization and capacity for diversification in response to immune selection were already present in the most recent common ancestor of *P. falciparum* and *P. reichenowi.*

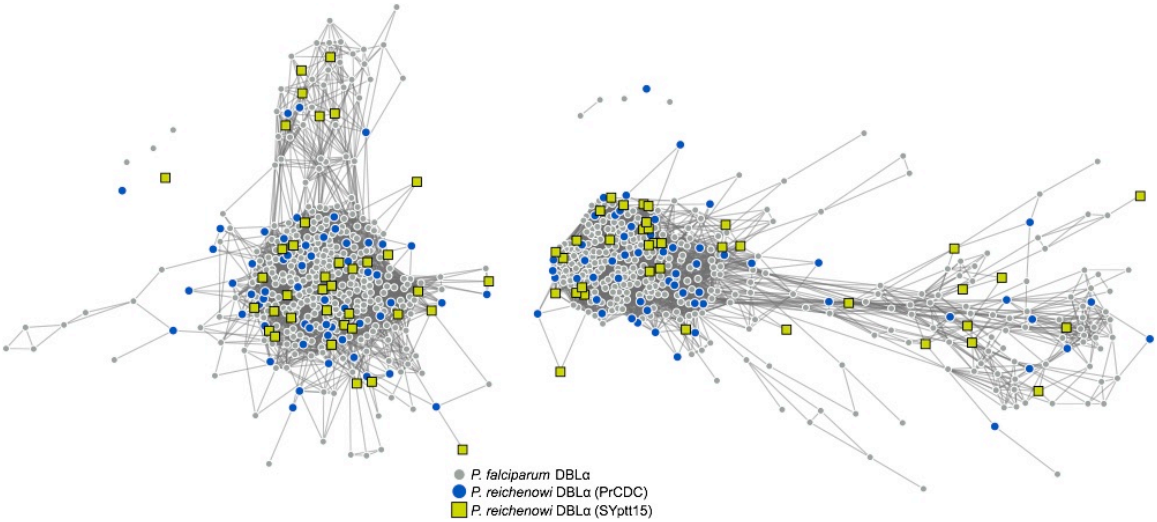

Figure 4-2. Networks of DBLα sequences from *P. reichenowi* and *P. falciparum*. Each node represents a DBLα HVR sequence and each link represents a shared amino acid substring of significant length (*146*). *Laverania* species and strain origin is indicated by node color and shape. Left and right networks correspond to left and right HVRs, respectively. *P. falciparum* and *P. reichenowi* sequences do not cluster by species or sample in either HVR. Link lengths and node placements are determined by a force-directed layout in order to better reveal structure, if it exists (see Methods). Additional analyses of these networks are shown in Supplementary Figure 1.

var DBLα tag structures predate the *Laverania* radiation

Having analyzed *var* tags from *P. falciparum* and *P. reichenowi*, we next examined parasite sequences from across the ape *Laverania* subgenus. Numerous identical mosaic elements in otherwise divergent sequences and a shared overall HVR architecture extended to the most divergent species (Fig. 4-3 and Supplementary Fig. 2). We were able to reconstruct highly connected networks for each HVR, indicating the presence of shared mosaic elements among the vast majority of tags from single-species parasite infections. Every *Laverania* *var* tag contained three conserved

sequence motifs separating two HVRs: in 86% of sequences, the three conserved motifs corresponded to three of the five most common *P. falciparum var* motifs (in order: HB3, HB5, HB2) (*144*), while in the remaining 14%, HB5 was intact in the middle of the tag and more divergent forms of HB3 and HB2 were encoded by the 5' and 3' end of the tag, respectively (Supplementary Fig. 3).

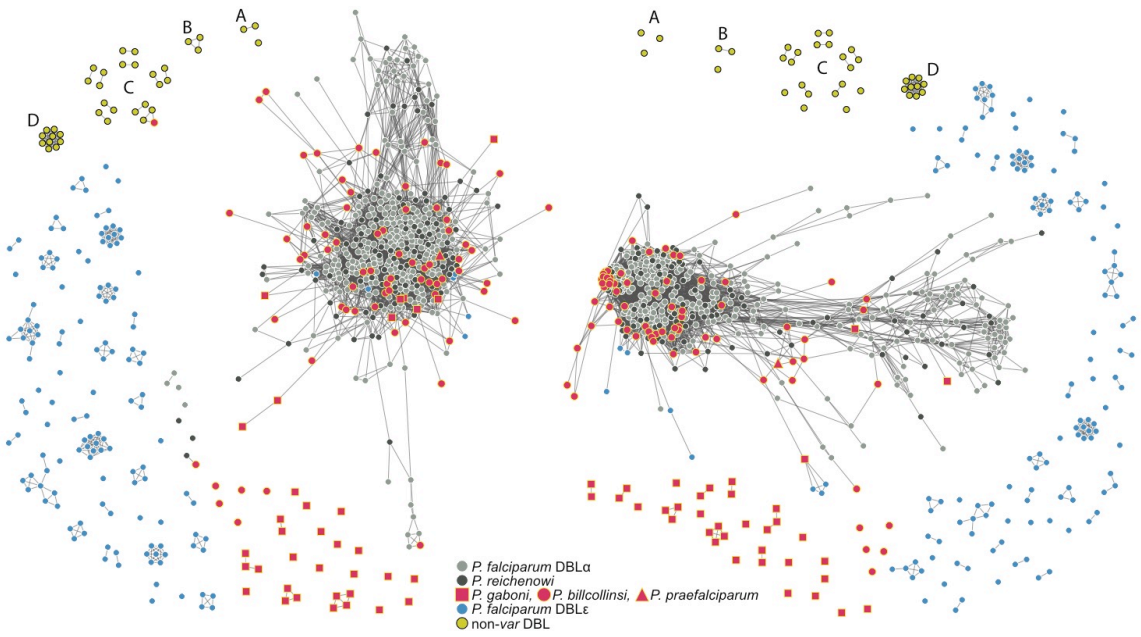

Figure 4-3. Networks of DBL sequences from *Laverania* single-species infections in the context of known DBLα and non-DBLα sequences.

Each node represents a DBL HVR sequence from a single-species infection and each link represents a shared amino acid substring of significant length. Note that for each sample, only unique *var* DBL haplotypes were included in the network analysis. Nodes with zero links indicate sequences that share no significant amino acid substrings with other sequences. Networks were built separately for each HVR, where mosaic diversity is highest (see Methods). Colors correspond to *Laverania* species as indicated; annotated yellow nodes correspond to (A) *dblsmsp*1 and (B) *dblmsp*2 from Pf3D7, PfIT, and PrCDC; (C) both DBL domains from *ebl1*, *eba140*, *eba165*, *eba175*, *eba181* of Pf3D7 and PfIT; (D) *P. vivax* Duffy Binding Proteins; see Supplementary Table 3 for a comprehensive list of non-DBLa sequences.

We confirmed that these tags were not derived from non-*var* DBL-containing genes by including tags from *P. falciparum* Erythrocyte Binding Antigen (*eba*) genes, *P. falciparum* and *P. reichenowi* DBL Merozoite Surface Protein 1 (*msp3.4*) and DBLMSP2 (*msp3.8*), and *P. vivax* Duffy Binding Proteins in our analysis (Supplementary Table 3). We also included *P. falciparum* DBLε tags to compare tags to *var*-derived, yet non-DBLα, sequences. As shown in Fig. 4-3, tags from single-species ape *Laverania* infections remained separated from both the non-*var* DBL tags and the *P. falciparum* DBLε tags, with a majority connected to one or both of the large connected components formed by the *P. falciparum* and known *P. reichenowi* tags. This majority included every new *P. reichenowi* and *P. praefalciparum* tag, and all but one *P. billcollinsi* tag. On the other hand, only 10 *P. gaboni* tags were connected to one or both large components, with the other 26 connected only to other *P. gaboni* tags in separate, small components. These smaller *P. gaboni* components did not share mosaic elements with DBLε and non-*var* DBL sequences, suggesting that they represented divergent, yet *var*-like, domains.

Laverania parasites contain ape-specific *var*-like DBL domains

We next investigated the relationships between sequences from all ape *Laverania* samples by conducting a network analysis that excluded *P. falciparum,* but included sequences from both mixed-species and single-species infections (Fig. 4-4). Sequences from *P. billcollinsi* and *P. praefalciparum* remained integrated within the large connected component that also included *P. reichenowi*, indicating conservation of mosaic elements within HVRs across these species. This finding is consistent with mtDNA (Fig. 4-1A), apicoplast, and nuclear phylogenies (*110*, *156*), which place *P. billcollinsi* and *P. praefalciparum* closer to *P. reichenowi*. In contrast, sequences from four single-species

infections of *P. gaboni*, which represent a much more distant *Laverania* species, exhibited much less shared sequence content in HVR networks. However, *P. gaboni* sequences appeared to fall into two subgroups based on tag length: (i) longer *P. gaboni* sequences (94-135 amino acids), which share mosaic elements with *P. reichenowi* and *P. billcollinsi* in 8 of 15 sequences in the Left HVR and 2 of 15 sequences in the Right HVR, and which we therefore term DBLα-like (red, unboxed in Fig. 4-4); and (ii) shorter *P. gaboni* sequences (72-85 amino acids), which remain disconnected from the *P. reichenowi-P.billcollinsi* component in 21 of 21 cases and which we therefore termed DBLx-like (red, boxed in Fig. 4-4). Thus, within the HVRs, longer *P. gaboni* DBLα-like sequences are partially overlapping with *P. reichenowi* and *P. billcollinsi*, while the shorter sequences appear to be distinct.

Although the DBLx tags fell outside the large connected component of the *P. reichenowi-P. billcollinsi* network group (Fig. 4-4, boxes), they were all amplified using standard *P. falciparum* DBLα primers, and they all exhibited the classical DBL architecture with fully intact HB5 motifs in the tag center. However, they were unrelated to other known DBL domain classes (Supplementary Fig. 4). All four single-species *P. gaboni* samples, as well as one *P. gaboni*-containing mixed-species sample, contained DBLx sequences. Based on polymorphisms in the HB3-like region, DBLx sequences formed two subgroups, which we refer to as DBLx1 and DBLx2 (Supplementary Fig. 3, Methods). DBLx sequences were not limited to chimpanzee parasites, as the mixed-species infection gorilla sample GTggg118, which contained both *P. praefalciparum* and *P. adleri*, also featured DBLx2 tags. The GTggg118 DBLx2 tags shared mosaic elements with both DBLx1 and DBLx2 tags from *P. gaboni*, while the GTggg118 DBLα-like tags were well-connected to the *P. billcollinsi-P. reichenowi* component (Fig. 4-4). We thus hypothesize that the GTggg118 DBLx2 tags derive from *P. adleri*, a closely

related sister taxon to *P. gaboni* (Fig. 4-1A), while the DBLα-like tags may be derived from either *P. adleri* or *P. praefalciparum*. Thus, it is likely that DBLx sequences represent new *var*-like DBL subdomains that are restricted to the C2/G2 branch of the *Laverania* subgenus (Fig. 4-1A).

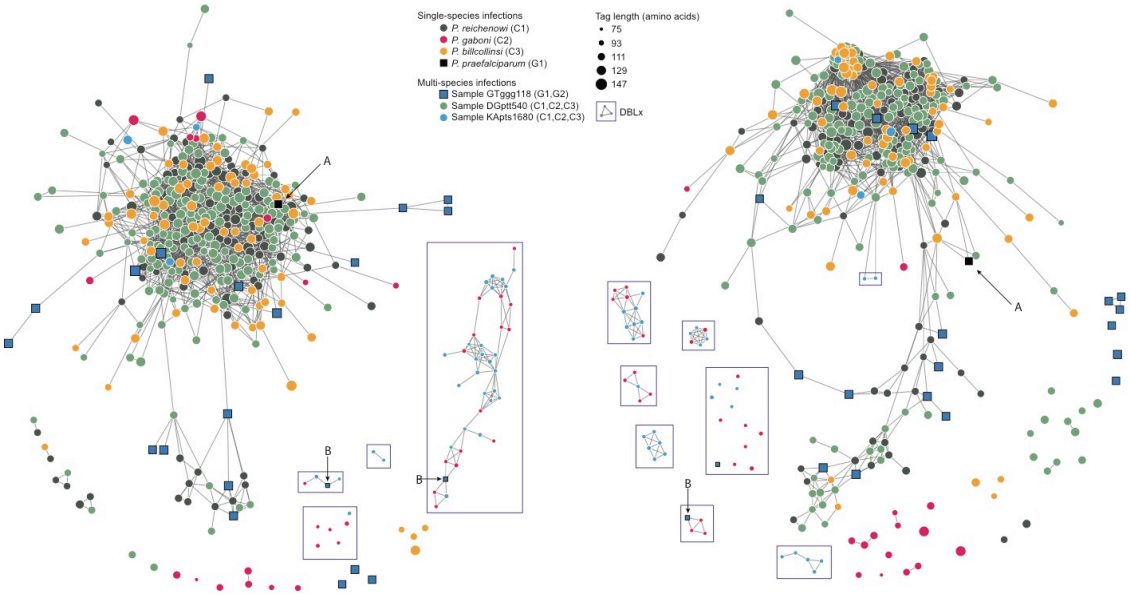

Figure 4-4. Networks of DBL sequences from single- and multi-species *Laverania* infections.

Each node represents a DBL HVR sequence and each link represents a shared amino acid substring of significant length. Note that for each sample only unique *var* DBL haplotypes were included in the network analysis. Nodes with zero links indicate sequences that share no significant amino acid substrings with other sequences. Networks were built separately for each HVR, where mosaic diversity is highest (see Methods). Circular nodes represent chimpanzee parasites and square nodes represent gorilla parasites. Node color corresponds to species and node size corresponds to tag length as indicated. DBLx sequences are enclosed in boxes.

var multi-domain structures predate the *Laverania* radiation

To confirm the presence of *var*-like genes in *P. gaboni*, we also examined near full-length parasite genomes and unplaced contigs, which were derived by select whole genome amplification (SWGA) (*39*) from two chimpanzee blood samples (SYpte37 and SYptt75). Three lines of evidence indicated that *var*-like genes, consisting of multiple DBL domains, were indeed present in this parasite species. First, we identified 55 *var*-like DBL domains in 40 different contigs, 14 and 2 of which were further classified using the VarDom server (*144*) as being related to *P. falciparum* DBLε and DBLζ domains, respectively (Table 1, Methods). None of the remaining DBL domains could be similarly subclassified, but four contigs featured exact nucleotide matches for DBLα-like tag sequences, providing a cross-validation between methods. Three contigs featured 3, 4, and 5 adjacent and non-identical DBL domains, a configuration unique to *vars*. An additional six contigs featured two adjacent DBL domains, but in these cases an *eba* gene origin could not formally be excluded (*157*).

Table 4-1 *Var* gene-like structures in *P. gaboni* whole-genome contigs

Sample	Total *var*-like DBLs identified	number of DBLs (number of contigs)					DBL classification[b]			*var*-like ATS
		1-DBL	2-DBL	3-DBL	4-DBL	5-DBL	DBLε	DBLζ	unclassified	
SYpte37	15	8 (8)	4 (2)	3 (1)	-	-	2	-	13	0
SYptt75	40	23 (23[a])	8 (4)	-	4 (1)	5 (1)	12	2	26	16[c]

[a]includes the three-exon single-DBL containing contig shown in Fig. 4-6.
[b]DBLa-δ domains according to the classification by Rask *et al*.,(*144*) were not identified. In addition, we found no evidence of DBLa-CIDR domain pairs.
[c]Includes three contigs with *var*-like DBL-TM^ATS multi domain (two-exon) structure

The finding of only nine contigs with *var*-like multi-DBL configurations in our *P. gaboni* genomic data is likely related to difficulties in assembling these sequences from short read data. *De novo* assembly is hindered by identical and near-identical motifs present in different DBL domains, which makes an accurate determination of the number and order of these domains in a given *var* gene difficult (*158*). In contrast, acidic terminal segment (ATS) domains, which are also a unique feature of *var* genes, lack these repeat structures, although they share some sequence motifs due to frequent recombination (*159*). We thus reasoned that ATS regions would more likely assemble into full length or near full-length domains and looked for these *var* signatures in the *P. gaboni* genomic sequences. Indeed, ATS domains were readily identified in 16 contigs derived from the *P. gaboni* SYptt75 genome. In *P. falciparum,* the ATS domain encodes the intracellular portion of the PfEMP1 protein, which is expressed from a separate exon (Fig. 4-1B). ATS domains are unique to *var* genes, except for a single copy non-*var* gene with an "ATS-like" domain on chromosome 1 (PF3D7_0113800) (*144*). Using the VarDom server to characterize the *P. gaboni* ATS domains, we identified seven of ten known major homology blocks (Fig. 4-5). These were very similar to *P. falciparum var* ATS homology blocks, but very different from the non-var "ATS-like" domains of PF3D7_0113800 and its *P. reichenowi*, and *P. gaboni* orthologs (Fig. 4-5, Supplementary Fig. 5), thus providing compelling evidence for the presence of bona fide *var* ATS domains in *P. gaboni*.

Finally, three of the ATS-containing contigs exhibited a longer two-exon *var* gene structure, with a DBL and trans-membrane (TM) domain in exon 1 and an ATS domain in exon 2. One of these contigs contained an additional open reading frame (ORF) downstream of the *var*-like exon 2, which was 88% identical in its nucleotide sequence to genes and intergenic flanking sequences in *P. falciparum (*PF3D7_0323800) and *P.*

reichenowi (PRCDC_0323100) on the same chromosome, respectively (the latter two shared 94% nucleotide sequence identity). Although the function of these orthologs is unknown, they are single-copy genes immediately adjacent to a *var* exon 2 pseudogene on chromosome 3 of both *P. falciparum* and *P. reichenowi* (Fig. 4-6). This synteny implies the existence of ancestral ORFs on chromosome 3, including a *var* gene that retained both exons in *P. gaboni*, but represents a single-exon pseudogene in *P. falciparum* and *P. reichenowi*. Thus, the presence of a two-exon *var* structure and synteny on chromosome 3 for three *Laverania* species, which span the root of the subgenus phylogeny, indicate that *var* genes evolved their extant two-exon and multi-domain structure prior to the radiation of this subgenus.

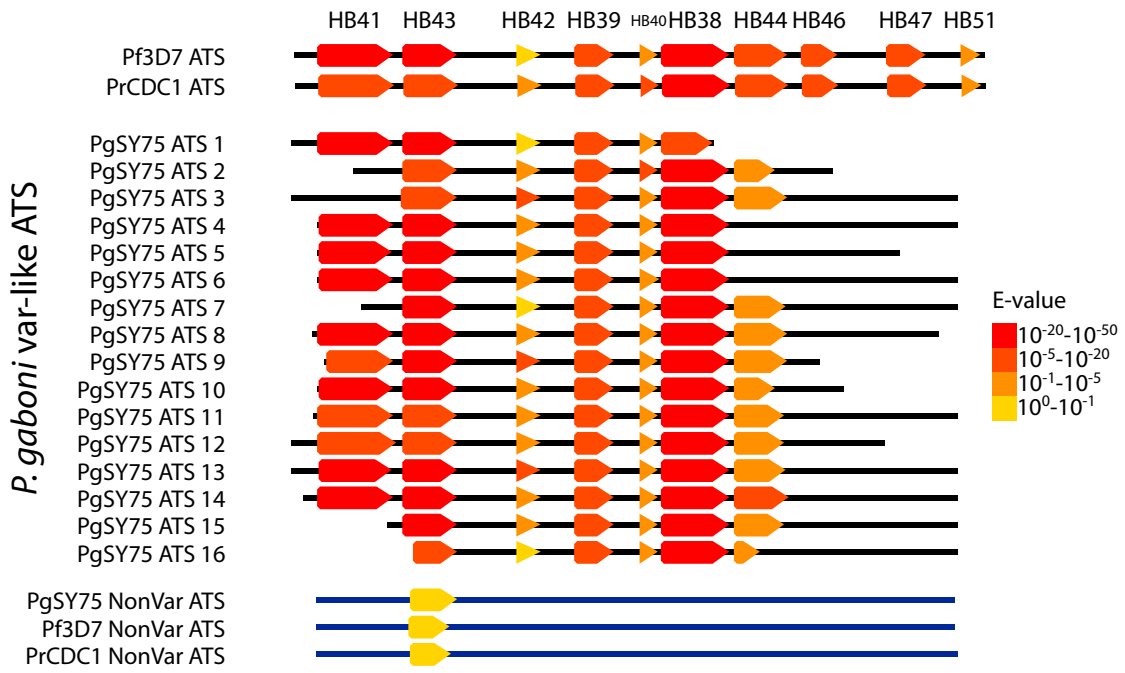

Figure 4-5. Conservation of *var* ATS domain homology block structure in *P. gaboni*.

The homology block (HB) structure of *var* ATS domains identified in 16 contigs of a near complete *P. gaboni* genome (PgSY75) are shown in relation to representative *P. falciparum* and *P. reichenowi var* ATS domains (Pf3D7 and PrCDC1, top) as well as a non-*var* "ATS-like" domain of the *P. falciparum* PF3D7_0113800 gene and its *P. reichenowi* and *P. gaboni* orthologues (bottom). HBs (arrows) were predicted by VarDom 1.0 and annotated in an alignment of all 20 sequences. Colors correspond to VarDom reported E-values, representing an estimate of the likelihood of observing such a match by random chance. Black lines indicate the relative length of each sequence.

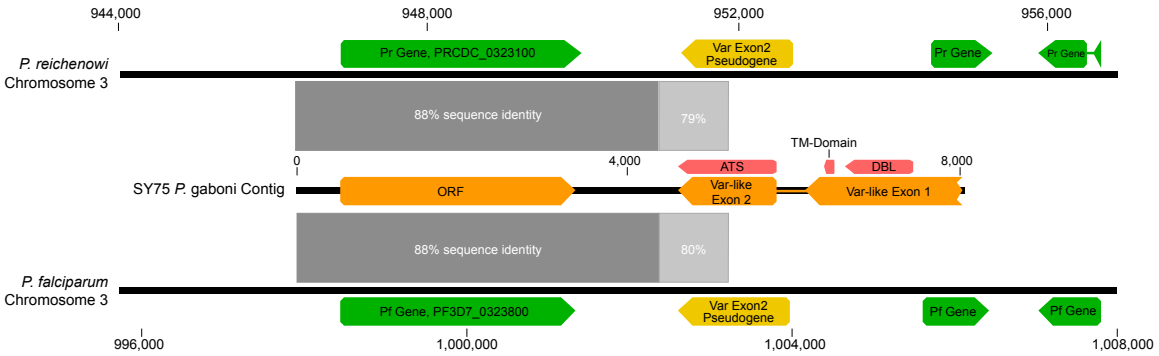

Figure 4-6. Shared synteny of *var*-like genes in *P. falciparum*, *P. reichenowi* and *P. gaboni*.

An open reading frame (ORF) located downstream of a predicted *var*-like gene in *P. gaboni* showed 88% sequence identity (dark grey bars) with a single copy gene present in both *P. falciparum* 3D7 (PF3D7_0323800) and *P. reichenowi* CDC1 (PRCDC_0323100). The *P. gaboni var*-like gene is syntenic with a *var* exon 2 pseudogene in both *P. falciparum* and *P. reichenowi*, suggesting that a *var* gene was present at this location in the ancestor of all three *Laverania* species.

Laverania var repertoire structure

It has previously been shown that *P. falciparum var* genes can be divided on the basis of DBLα domains into two main groups, classified by the number of cysteine residues in the tag region (*152*) which map to distinct community structures in network analyses (*146*). These two main groups can be further subdivided into a total of six Cys/PoLV (CP) groups based on the presence or absence of key amino acid residues (*152*, *160*). These cysteine-based classifications were found to be associated with different upstream promoter regions and clinical outcomes, and *var* repertoires in individual *P. falciparum* parasites appear to be stably structured with respect to these categories (*154*). We observed the same cysteine-based organization, both with respect to cysteine counts

and CP groups, in DBLα tags from *P. billcollinsi,* but not from *P. gaboni,* although in the latter case we identified far fewer DBLα-like motifs *(*Supplementary Fig. 6). Thus, cysteine-based organization of *var* gene repertoires extends to *P. billcollinsi,* but may not extend to *P. gaboni (*and by inference *P adleri*).

4.4 Discussion

Until recently, the only known close relative of *P. falciparum* was *P. reichenowi*, a *Laverania* parasite infecting chimpanzees. Over the past 5-6 years, five additional species within the *Laverania* subgenus have been described, each infecting either chimpanzees or gorillas. This *Laverania* species diversity provides an unprecedented opportunity to study the origins of genomic features that previously seemed unique to *P. falciparum,* such as the *var* gene family encoding erythrocyte membrane proteins. Here we show that various aspects of the multi-scale modularity of these loci can be recognized in diverse *Laverania* species, with the implication that a *var* or *var*-like gene family already existed in their last common ancestor. First, at the *var* gene repertoire level, we find genes with a characteristic two-exon structure, encoding multiple adjacent domains potentially capable of binding diverse endothelial markers. Like the constituent domains of the *P. falciparum*-encoded PfEMP1 proteins, the other *Laverania* DBL sequences can be subclassified into distinct groups, which may reflect differences in endothelial binding or other specificities. Second, at the domain architecture level, alternating conserved and hypervariable regions enable combinatorial diversity while presumably maintaining protein structure and binding functions. Finally, at the microscale level, some protein motifs within hypervariable regions are shared among even the most divergent *Laverania* species, despite the evidence of high frequency recombination within species. Thus, many key elements of the *var* multi-gene family

appear to have originated many (perhaps tens of) millions of years ago.

In *P. falciparum*, the *var*-encoded PfEMP1 proteins play a key role in pathogenesis by mediating the binding of infected red blood cells to specific host receptors in a wide range of tissues. Particular disease syndromes have been associated with individual DBL domains, two of which were present in *P. gaboni*. The first, DBLε, is found in the *var2csa* genes of *P. falciparum* (*144*) and *P. reichenowi*, which exist as only one or two *var* variants per genome and have been identified in every complete *var* repertoire analyzed to date. In *P. falciparum*, *var2csa* genes are responsible for placental binding, and the DBLε domain has thus been implicated in pregnancy-associated malaria (*161*). Similarly, we identified DBLζ in *P.gaboni*. Although there currently are no host receptors or disease syndromes that have been associated with this individual domain in *P. falciparum*, triplet combinations of DBLε and DBLζ domains have been linked to IgM-positive rosetting phenotypes (*162*). The presence of recognizable DBLε and DBLζ domains in the most divergent *Laverania* species suggests that DBL domain differentiation into subtypes represents an ancient host adaptation, and that DBLε and DBLζ may represent functionally constrained domains across the *Laverania* subgenus.

Beyond single *var* domains, the *var* repertoires of *P. falciparum* parasites can be divided into groups that have been associated with different clinical phenotypes, such as severe malarial anemia and cerebral malaria, using a cysteine-based classification of DBLα tags (*160*, *163*, *164*). These groups are represented in similar proportions across *P. falciparum* and *P. reichenowi* parasites, and our data suggest that this repertoire structure may also extend to *P. billcollinsi (*Supplementary Fig. 6); an insufficient number of DBLα-like tags precludes an extension of this classification to *P. gaboni* at the present time. Given their association with clinical disease in humans, the extent to which these

sequence features are also indicative of pathology in apes warrants further study.

Although we identified *var*-like features in species spanning the *Laverania* subgenus, we also found that certain signatures identified in *P. falciparum* and *P. reichenowi var* genes are absent from the more divergent parasite species. For example, we found no evidence of CIDR domains in either of the *P. gaboni* genomes, despite identifying numerous DBL domains (Table 1). Moreover, DBLα-like *P. gaboni* sequences were not sufficiently similar to *P. falciparum* DBLα domains to be confidently classified as such. Since the vast majority of *P. falciparum var* genes encode a DBLα-CIDR domain pair, the apparent absence of CIDR domains from *P. gaboni* is puzzling, especially in light of the role that CIDR domains are believed to play in host receptor binding (*165*). It will be important to determine whether *P. gaboni var*-like genes contain other domains with CIDR-like function or whether *P. gaboni* differs in its biology from other *Laverania* parasites. Second, the network analysis of PCR tags revealed new DBL domains that we termed DBLx because they are unlike the other six known *var* DBL domain classes shared by *P. falciparum* (*166*) and *P. reichenowi* (*69*) (Fig. 4-4, Supplementary Figs. 3 and 4). These DBLx tags, which were amplified using *P. falciparum* DBLα primers, are shorter than all other tags, and can be further subdivided into DBLx1 and DBLx2 subgroups based on differences in the highly conserved HB3-like region (Supplementary Fig. 3). Divergence from the *P. falciparum* "LARSFADIG" motif within this HB3-like region have also been reported for another partially characterized *P. gaboni* genome (*69*), but adjacent sequences were not analyzed, thus leaving their relationship with DBLx domains unknown. Finally, we identified multiple copies of *P. gaboni* ATS domains, which exhibit a *var*-like homology block structure that is very similar, but not identical, to *P. falciparum* and *P. reichenowi* ATS domains (Fig. 4-5, Table 1, Supplementary Fig. 5). Taken together, these data indicate that, while *var*-like

genes in *P. gaboni* (and possibly also *P. adleri*) share important structural similarities with those of *P. falciparum and P. reichenowi*, they also exhibit important differences, which may reflect differences in function and biology.

The presence of *var*-like genes throughout the *Laverania* subgenus suggests an ancient adaptation for antigenic variation, and potentially cytoadherence. However, while links exist between *var* expression and clinical disease in humans, the disease causing potential of *var*-like gene products in *Laverania* parasites infecting wild apes remains unknown. Nonetheless, there may be important parallels since recent field studies of habituated chimpanzees in the Tai Forest, Côte d'Ivoire revealed higher fecal parasite burdens in both young (*167*) and pregnant (*168*) individuals, similar to what has been described in humans. Given the role of the *var*-encoded PfEMP1 proteins to mediate endothelial binding in the presence of a vigorous host immune response, it is likely that *var* genes play a similar role in other *Laverania* species. However, the extent of *var* gene diversity, especially among the more divergent *Laverania* species that lack certain *P. falciparum*-specific DBL and CIDR domains, suggest potentially different biological solutions. Additional field studies of habituated ape populations will be necessary to establish the biological consequences of ape *Laverania* infections and the pathogenic potential of their *var*-like gene products.

4.5 Methods

Sample collection

Ape fecal samples were collected from wild-living central (*Pan troglodytes troglodytes*; DGptt540) and eastern (*P. t. schweinfurthii*; KApts1680) chimpanzees and western lowland gorillas (*Gorilla gorilla gorilla*; GTggg140, GTggg118) for previous molecular epidemiological studies of *Laverania* parasites (*26*). Samples were collected in RNAlater

(1:1 vol/vol), transported at ambient temperatures, and stored at -80°C. We also analyzed left-over blood samples from chimpanzees cared for at the Sanaga-Yong Rescue Centre (SYptt5, SYptt15, SYptt20, SYpte37, SYptt75, SYptt79, SYptt82), which were obtained in the context of routine health examinations or for specific veterinary purposes. Samples were shipped in compliance with Convention on International Trade in Endangered Species of Wild Fauna and Flora regulations and country-specific import and export permits. DNA was extracted from fecal and blood samples using the QIAamp Stool DNA Mini Kit and QIAamp Blood DNA Mini Kit (Qiagen, Valencia, CA) respectively, described in detail in (*110*).

Plasmodium species identification

The *Plasmodium* species composition in ape fecal and blood samples was determined by single genome sequencing (SGS) and phylogenetic analysis (*26*, *110*). Briefly, fecal and blood DNA was endpoint diluted in 96-well plates, and the dilution that yielded fewer than 30% wells with positive PCR reactions was used to generate between 2 and 174 different SGS sequences per sample (according to a Poisson distribution, the DNA dilution that yields PCR products in no more than 30% of wells contains one amplifiable template per positive PCR more than 83% of the time). Amplification products were gel purified, and sequenced directly without interim cloning. Sequences that contained double peaks as an indicator of more than one amplified template were discarded. Different genomic loci were amplified, including portions of mitochondrial (cytochrome B), nuclear (erythrocyte binding antigens *eba165* and *eba175*, 6-cysteine protein *p47* and *p48/45*, lactate dehydrogenase, reticulocyte-binding protein homolog 5), and apicoplast (caseinolytic protease C) genes. All relevant primers are provided in Supplementary Table 4. For each genomic region, up to 73 single template-derived

amplicons were sequenced and their species origin was identified by phylogenetic analysis (Supplementary Table 1). This analysis identified seven blood samples and one fecal sample to represent single-species infections of *P. reichenowi (*SYptt15, 46 SGS sequences), *P. gaboni (*SYptt5, 86 SGS sequences; SYpte37, 59 SGS sequences; SYptt75, 122 SGS sequences; SYptt82, 59 SGS sequences), *P. billcollinsi (*SYptt20, 174 SGS sequences; SYptt79, 16 SGS sequences), and *P. praefalciparum (*GTggg140; 2 SGS sequences), although many of these specimens contained multiple variants (haplotypes) of the respective species. Three other fecal samples (GTggg118, KApts1680, and DGptt540) contained more than one ape *Laverania* species, and one included an additional non-*Laverania* species of unknown origin (Supplementary Table 1).

PCR amplification of *var* genes

DBL domains were amplified, cloned and sequenced (see (*145*, *152*, *153*)) using conventional (rather than limiting dilution) PCR. Different primers sets, listed below, were used to amplify 2.5μl of fecal or blood derived DNA in a 25μl reaction volume, containing 0.5μl dNTPs (10mM of each dNTP), 10pmol of each primer, 2.5μl PCR buffer, 0.1μl BSA solution (50mg/ml), and 0.25μl expand long template enzyme mix (Expand Long Template PCR System, Roche). Most samples were subjected to single round amplification with previously published primers, including DBLα-5' (5'-GCACGAAGTTTTGCAGATATWGG-3') and DBLα-3' (5'-AARTCTTCKGCCCATTCCTCGAACCA-3')(*153*), or DBLαAF' (5' GCACGMAGTTTYGC -3') and DBLαBR (5'-GCCCATTCSTCGAACCA-3')(*152*). Only 3 samples were amplified with additional primers, including C1DBLαAF' (5'-GCACGVAGTTTTGC-3') and

C1DBLαBR (5'-GCCCATTCSTSGAACCA-3'), and C2DBLAF (5'-AARTAHAGTTTTGCTGATTTARG-3') and C2DBLAR (5'-TTCGGACCATTCGKCWAWCCA-3'), respectively, or by nested PCR. The C2DBLAF and C2DBLAR primers were designed to specifically amplify *P. gaboni* DBL tags using an alignment of select whole-genome amplification (SWGA) derived contigs of SYpte37. Cycling conditions included an initial denaturing step of 2 minutes at 94°C, followed by 35-60 cycles of denaturation (94°C, 10 sec), annealing (50-55°C, 30 sec), and elongation (68°C, 1min), followed by a final elongation step of 10 minutes at 68°C. Both single round and nested PCR derived amplicons were gel purified and subcloned into pGEM-T Easy (Promega) or PCR4 TOPO (Life Technologies) plasmid vectors. Positive clones were sequenced, and analyzed using SEQUENCHER (Gene Codes Corporation, Ann Arbor, MI) or Lasergene (DNASTAR) software.

Criteria of *var* gene sequence selection

Amplified *var* DBL sequences were inspected for primer sequences (which were removed from final sequences) and the presence of a single intact open reading frame (ORF); sequences lacking an intact ORF or identifiable 5' and 3' primer sequences were discarded. To remove *Taq* polymerase errors in cloned DBLα *var* tag sequences, a neighbor-joining tree was constructed for each sample and sequences differing by less than 3 nucleotides were condensed into a single consensus sequence. Thus, for each sample only unique DBLα var tag haplotypes were analyzed.

Network analysis

A short region of *var* gene sequence within the DBLα domain, which we refer to as a

"tag," comprises three conserved homology blocks (HB3, HB5, and HB2) separated by two highly variable regions (HVRs) (*144*). We identified HVRs using a sequence entropy approach, modifying a previously published procedure (*146*) to accommodate ape *Laverania* sequences. In order to extract highly variable sequence content for further study, we identified and removed the three conserved homology blocks from the 3' end, middle, and 5' end of each tag sequence. This was done by first aligning all sequences first to HB3 without inserting any gaps mid-sequence (step 1), i.e., we required that all sequences align at and only at HB3. Next, we calculated the Shannon entropy of the aligned sequences at each position (step 2) and scanned from HB3 toward the center of the tag to find the first position *p* at which entropy was greater than 2 bits (step 3) such that each subsequent position also had entropy greater than 2 bits. Finally we retained all sequences from *p* toward the center of the tag (step 4). Steps 1-4 were repeated for HB2, thus removing low entropy homology blocks from the ends of each sequence. Second, we removed conserved central sequence content, splitting the tag into two HVRs. We repeated steps 1 and 2 with HB5. We then scanned from HB5 toward each end of the tag, finding the first position *p* in each direction with entropy greater than 2 bits such that each subsequent position had entropy greater than 2 bits, and retained everything from *p* toward the end of the tag. All steps are shown graphically in Supplementary Fig. 7. The high entropy HVR between HB3 and HB5 is referred to as the Left HVR and the high entropy HVR between HB5 and HB2 is referred to as the Right HVR.

Two types of networks were created. First, networks of *var* sequences were generated by assigning each HVR sequence to a node and placing a link between two nodes when their corresponding sequences shared a block of length *L* or greater at the amino acid level. *L=7* for Left HVR and *L=6* for Right HVR, based on null model

calculations (*146*). Figures were produced using force-directed layouts in *webweb* software v3.1 (http://danlarremore.com/webweb). Second, bipartite networks of both *var* genes and their shared blocks were created by assigning each HVR sequence and each shared block of length *L* or greater to a node, and placing a link between a sequence node and a shared block node if the block is present in the sequence. These bipartite networks are related to the other type of network via one-mode projection. Community detection was performed using the biSBM method applied to bipartite networks of sequences and their shared amino acid substrings (*169*).

_k_-mer stackup analysis

Within an amino acid sequence, we refer to any contiguous substring of length *k* amino acids as a *k*-mer. All *k*-mers were extracted from all sequences, noting the starting position (normalized to the total length of the sequence). For Supplementary Fig. 2A, all *k*-mers from *P. falciparum* and *P. reichenowi* were sorted by their frequency of appearance, and stacked histograms of their starting positions were created with 50 bins. For Supplementary Fig. 2B, all *k*-mers from each of *P. falciparum, P. reichenowi, P. gaboni*, and *P. billcollinsi* were sorted by their presence across species, and stacked histograms of their starting positions (relative to the species indicated at the top of each plot) were created with 50 bins.

Bayesian _k_-mer analysis

A window of length *k* was scanned across each amino acid sequence from *P. falciparum* and *P. reichenowi* monoinfections, extracting all length *k* substrings. Some substrings appeared in sequences from both species, while others were species-specific. This analysis, derived and developed in detail in Supplementary Text 1, estimates the overlap

in populations of tag sequences using Bayesian statistics to correctly extrapolate the parameters of the conjugate prior distribution that characterizes the overlap from limited sample data (*170*).

For this analysis, we examine 296 DBLα tag sequences from *P. falciparum* and 94 from *P. reichenowi*. Each sequence is a string of amino acids, so from a sequences of length N, we can extract $N-k+1$ substrings (i.e. k-mers, or words) of length k. In what follows, we use k=7 for all examples. (Other values of k may be used, and results do not depend sensitively on moderate k; we tested $k \in [5, 15]$.) The 390 total sequences comprise 45731 words for k = 7, but some words appear in multiple sequences; the total number of unique words is 22431. This indicates that, on average, each word appears approximately 2 times across all 390 sequences. In fact the distribution is highly heterogeneous: 70% of words appear only once, 16% appear only twice, and 6% appear only three times, meaning that 92% of words appear in only 1 to 3 *of the total* 390 sequences. This heterogeneity, depicted in Supplementary Figure 8, makes it difficult to decide whether these two sets of sequences are drawn from distinct populations.

Some words are shared by both *P. falciparum* and *P. reichenowi* (8%), some are unique to *P. falciparum* (65%) and the rest unique to *P. reichenowi* (27%). If only 8% of (length 7) words are shared by both species, one might conclude that the populations of words are well separated. However, owing to the massive diversity of words in both species, and our small sample of sequence tags, this interpretation is incorrect. Instead of calculating the overlap between species for our data set, we wish to use this data to estimate the overlap for the global populations of *P. falciparum* and *P. reichenowi*.

Before the mathematical formulation, we advance the following helpful analogy, by imagining each word as a biased coin. Suppose we have a large bag of coins and

each coin has a biased probability of landing on heads. Further, imagine that the biases are not all the same, but are instead drawn from some distribution. We wish to estimate the distribution, so we take the coins, one by one, and flip them, writing down which coin was flipped and whether it lands on heads or tails each time. However, for 70% of the coins, we only get one flip. For 16% of the coins, we only get two flips, and for 6% of the coins we only get three flips, etc. When estimating the distribution of *p*, we must take into account the number of flips observed for each coin.

Given our small sample from the distribution, we wish to approximate the global distribution of values of p_i. This will tell us how much the populations overlap. Our data consist of f_i and r_i, the numbers of observations of word *i* in *P. falciparum* and *P. reichenowi* respectively. We model the assignment of each occurrence of word *i* to *P. reichenowi* as an independent Bernoulli trial, with parameter p_i. Let the set of p_i be Beta distributed with parameters α and β, where we use the Beta distribution because it is the conjugate prior of the Bernoulli distribution. Then, the likelihood of observing data $\{x_i\}$, given the parameters, is

$$L(\{x_i\}|\alpha,\beta) = \prod_{i=1}^{n}\left[\int_{p_i}\left(\prod_{j=1}^{f_i+r_i}\Pr(\text{word is from } P.reichenowi|p_i)\Pr\ (p_i|\alpha,\beta)\right)dp_i\right]$$

which may be integrated using beta functions *B* to get

$$L(\{x_i\}|\alpha,\beta) = \prod_{i=1}^{n}\frac{B\ (f_i+\alpha, r_i+\beta)}{B\ (\alpha,\beta)}$$

Taking a log yields

$$\log L\ (\{x_i\}|\alpha,\beta) = \sum_{i=1}^{n}\log\left(\frac{B\ (f_i+\alpha, r_i+\beta)}{B\ (\alpha,\beta)}\right)$$

This log-likelihood function is related to a solution to an analogous problem from the

domain of probabilistic competition dynamics (*170*) in which two teams were competing for points over the course of many competitions. We maximize it in MATLAB using the *fminsearch* function, using the observed f_i and r_i values.

Pairwise distance analysis

Protein sequences were aligned pair-wise using MUSCLE v3.8.1 (*127*) and Hamming distances (number of sites at which the two aligned sequences differ) were calculated neglecting gaps at both ends of the alignment to adjust for variable sequence lengths. Hamming distances were alternatively calculated by counting a contiguous block of gaps as a single difference, with no qualitative difference in results.

Blocksharing

We quantified sequence conservation from one particular sequence to an ensemble of others by scanning a window of length *k* across the particular sequence and computing the fraction of sequences in the ensemble containing each *k*-mer or block. This produces a measure of conservation between 0 and 1 in the frame of reference of the particular sequence; Figure 4-1B shows this blocksharing for the DBLα domain of *DD2var11* compared to the background of data (*144*); *k*=7.

CP group analysis

Var tag sequences can be classified according to the number of cysteine residues as well as sequence content at defined "positions of limited variability (PoLV)"(*152*). In the *var* sequence literature, these are referred to as Cys-PoLV groups, or simply CP groups. We identified CP groups with a MATLAB script according to the following definitions: Group 1: MFK* at PoLV, 2 cysteine residues; Group 2: *REY at PoLV2, 2 cysteine

residues; Group 3: 2 cysteine residues, not group 1 or 2; Group 4: 4 cysteine residues, not group 5; Group 5: *REY at PoLV2, 4 cysteine residues; Group 6: 1, 3, 5, or 6 cysteine residues. Histograms of cysteine counts and CP groups are shown in Supplementary Fig. 6.

P. gaboni select whole genome amplification (SWGA)

DNA was extracted from two chimpanzee blood samples (SYpte37, SYptt75) identified as *P. gaboni* single-species infections by single genome sequencing (Supplementary Table 1) and subjected to select whole *Plasmodium* genome amplification as described (*39*, *151*). Briefly, total DNA (100 ng – 1 ug) was digested using the methylation dependent restriction enzymes *Msp*JI and *Fsp*EI in multiple replicates. The digestion products were amplified using phi29 polymerase and one of two primer sets consisting of 10 primers (8-12 nt in length each) designed to bind frequently and broadly to the *P. falciparum* genome but only rarely to the chimpanzee genome (*151*). 50 ng of first round product was re-amplified in a second reaction using the second primer set. Replicates were pooled and a short insert library was constructed using the TruSeq DNA PCR-Free Sample Preparation Kit (Illumina) and sequenced using a MiSeq Reagent Kit V2 (500-cycles) (Illumina) to generate 250 bp paired end reads. Reads were mapped to the *P. falciparum* 3D7 reference genome using Geneious (Biomatters Limited, Auckland, New Zealand), and subjected to guided assembly using Velvet Columbus (*114*). For SYptt75, contigs produced by Velvet were aligned to the reference and the resulting core *P. gaboni* draft genome was iteratively corrected manually and using PAGIT v1.0 (*81*). All reads from SYptt75 and SYpte37 were mapped to this draft reference and reads that could not be mapped were assembled separately using Spades v3.1.1 (*120*).

Putative *var* gene identification *var* domain analysis

Due to the hypervariability of *var* sequences, *P. gaboni* reads did not map to *var* gene containing regions of the *P. falciparum* 3D7 reference genome, nor were *var* genes readily identified in the SYptt75 core *P. gaboni* genome. A search for contigs containing *var*-like sequence was therefore performed on unplaced SYptt75 and SYpte37 contigs (produced by either Velvet or Spades in a reference-independent manner). Specifically, tblastn and tblastx searches were performed using all *P. gaboni* unplaced contigs against a database of available full-length *P. falciparum* 3D7 and *P. reichenowi* CDC1 *var* genes. Genes and ORFs were identified in the top hits manually and by Augustus v2.5.5 gene prediction (*123*), and pblast searches using the resulting amino acid sequences were again performed against the translated *P. falciparum* and *P. reichenowi* *var* gene database. Hits were then submitted to the VarDom 1.0 server (http://www.cbs.dtu.dk/services/VarDom/) (*144*) for domain identification and classification.

The *P. gaboni* ortholog of PF3D7_0113800 was identified in the draft SY75 sequence by blast homology to PF3D7_0113800. Gene annotation was performed using RATT (*122*) with manual correction.

Neighbor-joining tree construction

Protein sequence tags were aligned using MUSCLE v3.8.1 (*127*) and the phylogeny were created using the neighbor joining (NJ) distance method, with Poisson distances, as implemented in Seaview 4.4.0 (*171*).

DBLx identification and classification

DBLx domains were identified as those tags that (i) were less than 90 residues in length,

and (ii) began with residues NI, DF, or DM. Those that began with residues NI were further classified as DBLx1 and those that began with DF or DM as DBLx2. 100% of DBLx sequences also featured a lysine residue (K) in the fourth position of the tag instead of the DBLα arginine (R). Sequence logos are shown in Supplementary Fig. 3.

4.6 Acknowledgements

This work was supported by grants from the National Institutes of Health (R21 GM100207, R01 AI091595, R37 AI050529, R01 AI058715, T32 AI007532, P30 AI045008) and the Wellcome Trust (grant #090851).

CHAPTER 5 - *swga*: Efficient Primer Set Design for Selective Whole Genome Amplification

Erik L. Clarke*, Sesh A. Sundararaman*, Stephanie N. Seifert, Frederic D. Bushman, Beatrice H. Hahn, Dustin Brisson.

Manuscript in preparation.

Stephanie N. Seifert, Dustin Brisson, and I developed the framework for the primer design algorithm. Erik L. Clarke developed the program with my input. Erik L. Clarke, Stephanie N. Seifert, Dustin Brisson, and I tested the program. I used the program to design primer sets, tested them along with pre-existing primer sets, and analyzed the data. Erik L. Clarke, Frederic D. Bushman, Beatrice H. Hahn, Dustin Brisson, and I wrote the manuscript with contributions from all authors.

5.1 Introduction

Population genomic analyses offer unprecedented capabilities to infer precise evolutionary, ecological, epidemiological, and molecular processes and address major outstanding questions in multiple microbiological fields (*172*, *173*). These analyses of many microbes are currently hindered by the practicalities of obtaining sufficient numbers of genomes for analysis. Laboratory culture is the classical method to isolate a target microbe from the other organisms and to obtain the appropriate samples for high-throughput sequencing, but many microbes cannot be cultured. Population genomics of most microbial species will require a practical method to collect sufficient amounts of target genomic DNA while limiting the amount of contaminating DNA (*36*).

Recently we developed selective whole genome amplification (SWGA), a culture-free technology that preferentially amplifies a specified target genome from total genomic extracts derived from an environmental sample (*39*). SWGA removes the need for ultra-deep sequencing or mechanical separation of target genomes from the background sample. This is accomplished by choosing primers that anneal to DNA sequence motifs that are common in the target genome but rare in the genomes of other species present in the environmental sample. A previous difficulty in implementing SWGA for many systems is a lack of automated, efficient, and reliable method to design effective primer sets. The original method uses a series of Perl scripts to identify primers that are common in the target genome and rare in the background sequences but considerable user input is required to design sets of primers to be used for selective amplification (*39*). Further, there is no way to quantitatively evaluate multiple primer sets prior to empirical testing which is expensive in both time and money.

Here we present *swga*, a program that addresses previous issues with SWGA primer design by computationally evaluating large numbers of primer sets on their

potential to selectively amplify the genome of a target species. The *swga* program identifies primers that should selectively amplify the target genome and rapidly evaluates primer sets based on user-defined metrics. Many aspects of the program's behavior, including evaluation criteria, can be modified to suit the needs of specific projects. The source code, download links, and documentation are hosted at https://www.github.com/eclarke/swga. The program is licensed under the GNU Public License (GPL) and is free to use and modify.

5.2 Methods

swga program overview

The *swga* program is designed to evaluate sets of primers on their potential to selectively amplify the genome of a target species even when it is very rare in the starting sample. The program can be divided into four steps (Figure 5-1): 1) counting binding sites for all primers in the target and background genomes to identify primers; 2) filtering primers based on parameters such as melting temperature and selectivity 3) ranking sets of primers on multiple metrics that can be user-defined; 4) outputs and visualization. Each part of the program can be run individually to optimize intermediate results or as an integrated unit.

Step 1: Primer identification. The first part of the *swga* program, *swga count*, - uses DSK(*174*) to identify all primers in the size range specified by parameters **min_size** and **max_size** that exist in the target and background sequences. All primers that are present in the target genome more times than specified by parameter **min_fg_bind**, and fewer times in the background sequences than specified by parameter **max_bg_bind** are saved in a local SQLite (*175*) database along with the number of times they appear in the target and background genome sequence files.

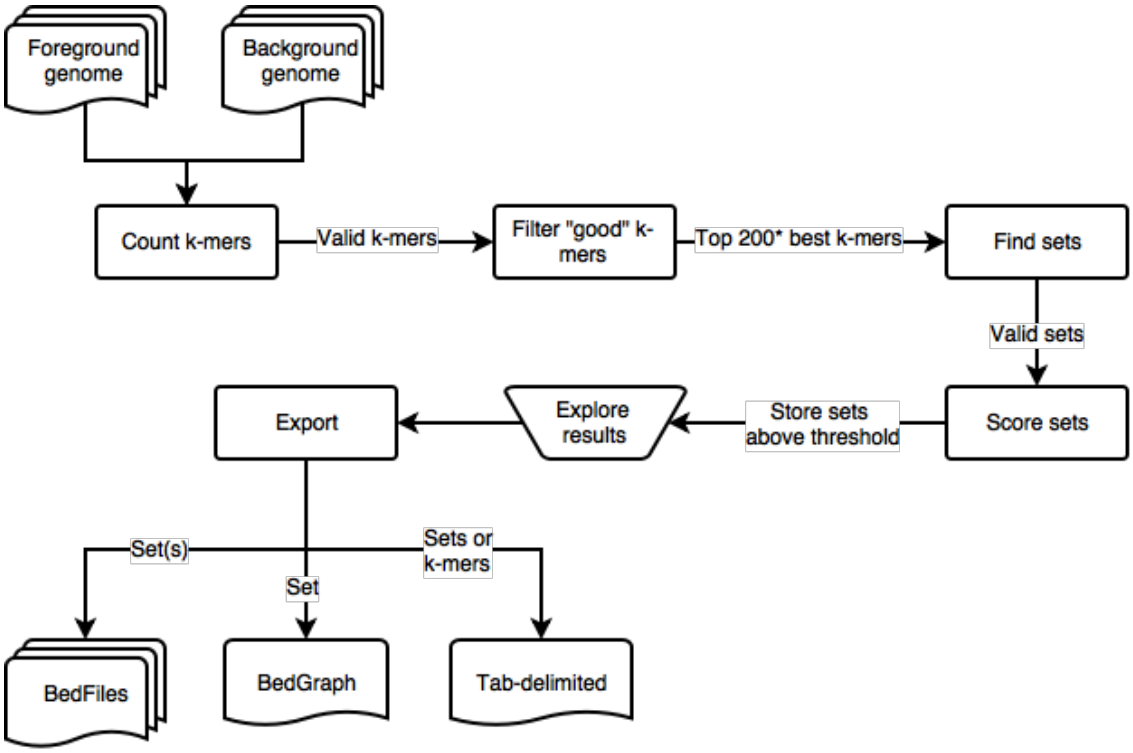

Figure 5-1. An overview of the *swga* workflow.

Step 2: Primer filters. The module *swga filter* ranks and filters primers by their selectivity to the target genome and melting temperatures. This selectivity is calculated by the ratio of binding sites in the foreground genome to binding sites in the background genome. The most selective primers, by default no more than 200 (parameter **max_primers**), that satisfy the calculated melting temperature parameters (**min_tm** and **max_tm**) are saved along with the locations of their binding sites in the target sequence.

Step 3: Primer set evaluation. Evaluation of all potential sets of primers is computationally intensive. The total number of possible primer combinations (*n* choose *k* combinations) is $2.4*10^{16}$ using the default parameters. However, not all of the possible combinations are compatible for SWGA. Primers are incompatible if they can form

heterodimers (i.e. they contain more consecutive complementary bases than specified by parameter **max_dimer_bp**) or if one is a subsequence of the other. *swga* plots the results of pairwise comparisons that establish mutual inter-compatibility as a compatibility graph, where each primer is a vertex and compatibility is denoted edges connecting primers. The problem of finding sets of entirely-compatible primers is thus reduced to finding collections of vertices that have edges to every other vertex in the collection (known as a *clique* in graph theory). We extend this by encoding the average distance between primer binding sites on the background genome as a "weight" on each vertex. This allows the program to prioritize cliques with higher total weights, representing sets of primers that rarely bind the background genome.

The module *swga find_sets* uses a modified version of the program *cliquer (*http://users.aalto.fi/~pat/cliquer.html) to evaluate cliques in the compatibility graph. The branch-and-bound algorithm in *cliquer* limits the search space to cliques within desired primer set size (**min_size** and **max_size**, weight (**min_bg_bind_dist**, and maximum distance between target binding sites (**max_fg_bind_dist**. For primer sets that satisfy these criteria, evaluative metrics including the average and maximum distance between primer binding sites in the target genome and the evenness of the primer binding sites in the target genome are calculated. Evenness is calculated on the Gini index (*176*), where indices near 0 represent evenly spaced binding sites, and indices near 1 represent uneven or "clumped" binding sites. These metrics are used in the default scoring algorithm (**fg_dist_mean * fg_dist_gini / bg_ratio**), which identifies primer sets that maximize selectivity and minimize the Gini index. Scoring algorithms can be altered by the user via the **score_expression** parameter.

Step 4: Outputs and visualization. Because the total number of sets of size *k* in the graph can be as high as (*n^k)*k^2*, and the wet-side procedure only needs one valid

set, the program by default stops after it has found 500 valid sets (**max_sets**). The saved primer sets can be evaluated by target selectivity of the primer set, the evenness of primer binding sites in the target (Gini index) (*176*), the average and maximum distance between binding sites in the target genome, and the average distance between binding sites in the background. Metrics can be added and removed by users to incorporate new criteria or alter the weight of the current criteria to the overall score. The user can query the sets that score highest by any of the calculated metrics and export the sets in a variety of common formats. Individual primers and primer sets can be exported in tab-delimited format for use in downstream applications, and genomic binding locations of the primers in a set can be exported in the common UCSC Genome Browser Bed and BedGraph track format (*177*).

Empirical primer set evaluation

To evaluate the efficacy of the *swga* program to identify primer sets successfully enrich a target genome from a complex genomic sample, we designed four primer sets to target *Wolbachia pipientis* from infected *Drosophila melanogaster*. As the ideal parameters for SWGA primer and primer set design have yet to be established, we designed sets with "standard" (15-45 C) or "high" (35-55 C) melting temperature (T_m) ranges and chose the set that either maximized the score or the set that minimized the Gini index within each T_m range. The remaining parameters included: a set size of 2-12 primers, a max consecutive bases for heterodimer and homodimer formation of 4, a minimum average distance between binding sites in the background genome of 30,000, and a maximum distance between any two binding sites in the target of 130,000. Instead of stopping at 500 sets (the default behavior), we let the program score ~1,000,000 sets in each T_m range before the maximum scoring and minimum Gini index sets were chosen. These

parameters were generally more permissive than the defaults in order to explore a broader range of possible sets.

Each primer set was tested on pooled genomic DNA extracted from 10 flies, which contained 4.7% *Wolbachia* DNA per ng. The pooled genomic DNA extract provided sufficient starting material to test each primer set in addition to the previously published SWGA primer set (*39*) in triplicate, while eliminating inter-fly variability in *Wolbachia* infection levels. The pooled genomic extract was digested with *NarI* at 37 C for 30 minutes to eliminate mitochondrial amplification as previously described prior to being aliquoted (40 ng per reaction) for SWGA. SWGA replicates were performed using the conditions previously described (*39*).

Amplified samples were purified using AmpureXP beads (Beckman Coulter), prepared for Illumina sequencing using a modified Nextera Library Preparation Kit protocol (*178*), and sequenced on an Illumina Miseq (150bp paired end). Adapter and primer sequences were removed from the resulting reads using Cutadapt (*179*). Trimmed reads were first mapped to the *Drosophila* genome using *smalt (*http://sourceforge.net/projects/smalt/) and the unmapped read pairs were then mapped to the *Wolbachia* genome. Breadth of coverage of the target genome (measured as percentage of the genome with at least 10x coverage) and sequencing rarefaction analyses were performed using R (*180*).

5.3 Results

Swga can rapidly identify target-biased primers, compatible primer sets, and evaluate primer sets on their likelihood to selectively amplify the specified target genome. The *swga* program was run four times using different parameter settings to characterize the

effects of primer melting temperature and the selectivity score on selective amplification and sequencing evenness.

Primer sets that satisfied the conditions of compatibility, minimum average distance between background binding sites, and maximum distance between any binding sites in the target were identified and evaluated from both the default and relax melting temperature runs. The primer sets were then ranked according to the total default score (see Methods) or according to the Gini index (*176*). We chose four primer sets to test selective amplification of *Wolbachia* from fruit flies: one set using **min_tm**=15C and max_tm=45C (Tm^-) that had the best total score (henceforth referred to as **Tm/Score**); one set using **min_tm**=15C and max_tm=45C that had the most even distribution of primers (**Tm/Gini**); one set using **min_tm**=35C and max_tm=55C (Tm^+) that had the best total score (**Tm^+/Score**); and one set using **min_tm**=35C and max_tm=55C that had the most even distribution of primers (**Tm^+/Gini**) (Table 5-1).

Table 5-1 Characteristics of primer sets chosen for selective whole genome amplification of *Wolbachia* from infected *Drosophila*

Characteristics	Score	Set Size	bg_ratio	fg_max_dist	fg_dist_mean	fg_dist_std	fg_dist_gini
Top Score; Standard Tm	0.03560	9	97839	31147	5327	7004	0.654
Top Gini; Standard Tm	0.05640	7	65271	34697	6853	6996	0.537
Top Gini; High Tm	0.00405	2	1732423	112818	13070	15515	0.537
Top Score; High Tm	0.00033	2	24253916	128209	12074	18011	0.660
Leichty *et al.*	0.01161	2	325172	112839	5305	10052	0.712

The efficacies of these four primer sets, as well as the previously published primer set (*39*), in selectively amplifying *Wolbachia* genomes from infected fruit flies were empirically evaluated using DNA extracted from a pool of 10 *D. melanogaster*. The proportion of sequencing reads that were derived from *Wolbachia* DNA was at least three times greater in all amplified samples than the sequencing reads from the unamplified genomic extract (Figure 5-2). The previously published primer set (*39*) was the least effective with 12.1%-27.7% of the sequencing reads mapping to the *Wolbachia* genome while 59.9%-81.4% mapped to *Drosophila*. The primer sets with higher melting temperatures (**min_tm**=35C and **max_tm**=55C) were considerably more effective at amplifying *Wolbachia* DNA than all other primer sets, with as much as 77.8% of the reads mapping to *Wolbachia*.

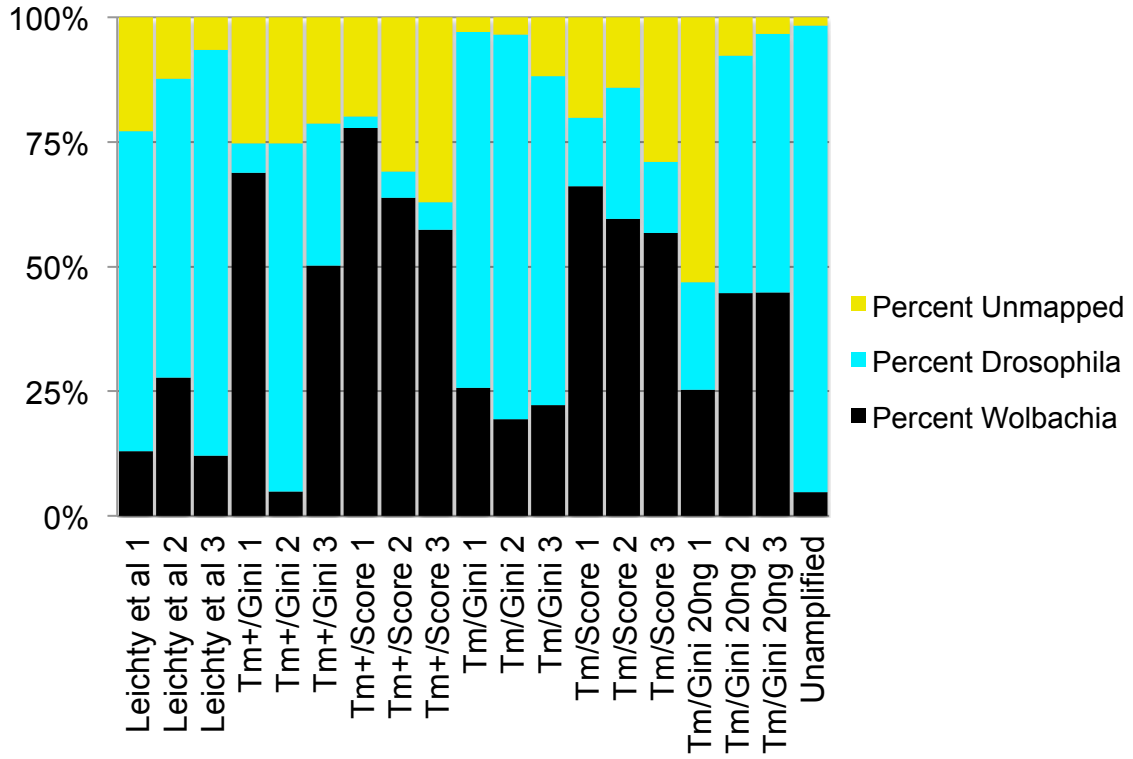

Figure 5-2. Percent of *Wolbachia*, *Drosophila*, and unmapped reads after SWGA with *swga* derived or previously published primer sets.

The percent of reads that mapped to *Wolbachia* (black), *Drosophila* (blue), or neither genome (gold), is shown for each primer set. Three replicates SWGA reactions (40 ng total DNA per reaction) were performed for each primer set and the results are presented separately for each. Additional triplicate SWGA reactions were performed using the Tm/Gini set and 20 ng total DNA per reaction.

The degree of enrichment of *Wolbachia* DNA after amplification did not necessarily correlate with a decrease sequencing effort needed to achieve broad coverage of the *Wolbachia* genome (Figure 5-3). For the Tm/Gini primer set, sequencing effort required to achieve at least 10x coverage of 90% of the genome was reduced 10 fold relative to the unamplified control. On the other hand, rarefaction analysis indicated that the higher melting temperature primer sets would never achieve 10x coverage of even 10% of the *Wolbachia* genome (Figure 5-3). The lack of correlation was primarily due to uneven amplification of across *Wolbachia* genome (Figure 5-4). Coverage analysis of the higher melting temperature primer sets identified substantial enrichment of a few short regions of the *Wolbachia* genome with little enrichment of the remaining sequence (Figure 5-4). By contrast, amplification with the Tm/Gini primer set yielded more even coverage across the genome, which was typically 10-100x better than that obtained from the unamplified control (Figure 5-4).

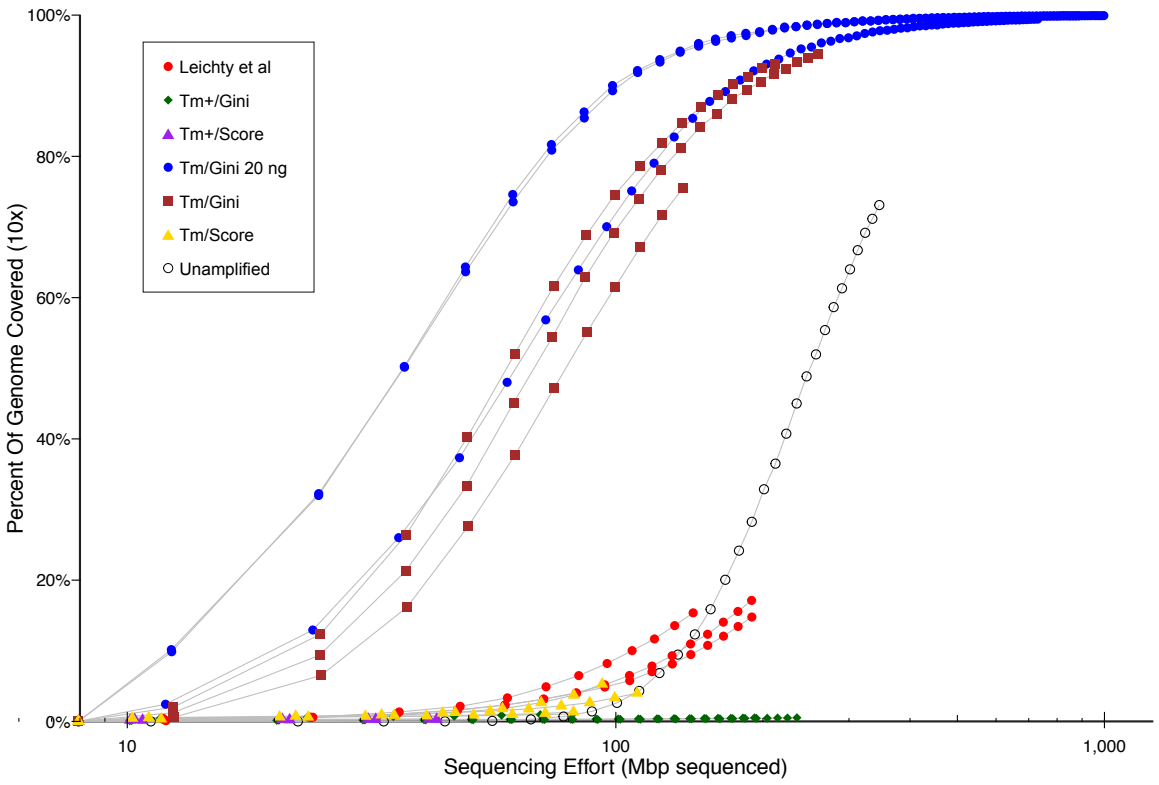

Figure 5-3. Relationship between sequencing effort and percent coverage of the *Wolbachia* genome (at 10x read depth) after SWGA.

The total number of base pairs (in millions) sequenced is shown relative to the percent of the *Wolbachia* genome covered at 10x read depth for the four *swga* derived primer sets and the set previously published by *Leichty* et al. SWGA reactions were performed in triplicate for each primer set using 40 ng of total DNA. An additional three reactions were performed for the Tm/Gini set using 20 ng of total DNA.

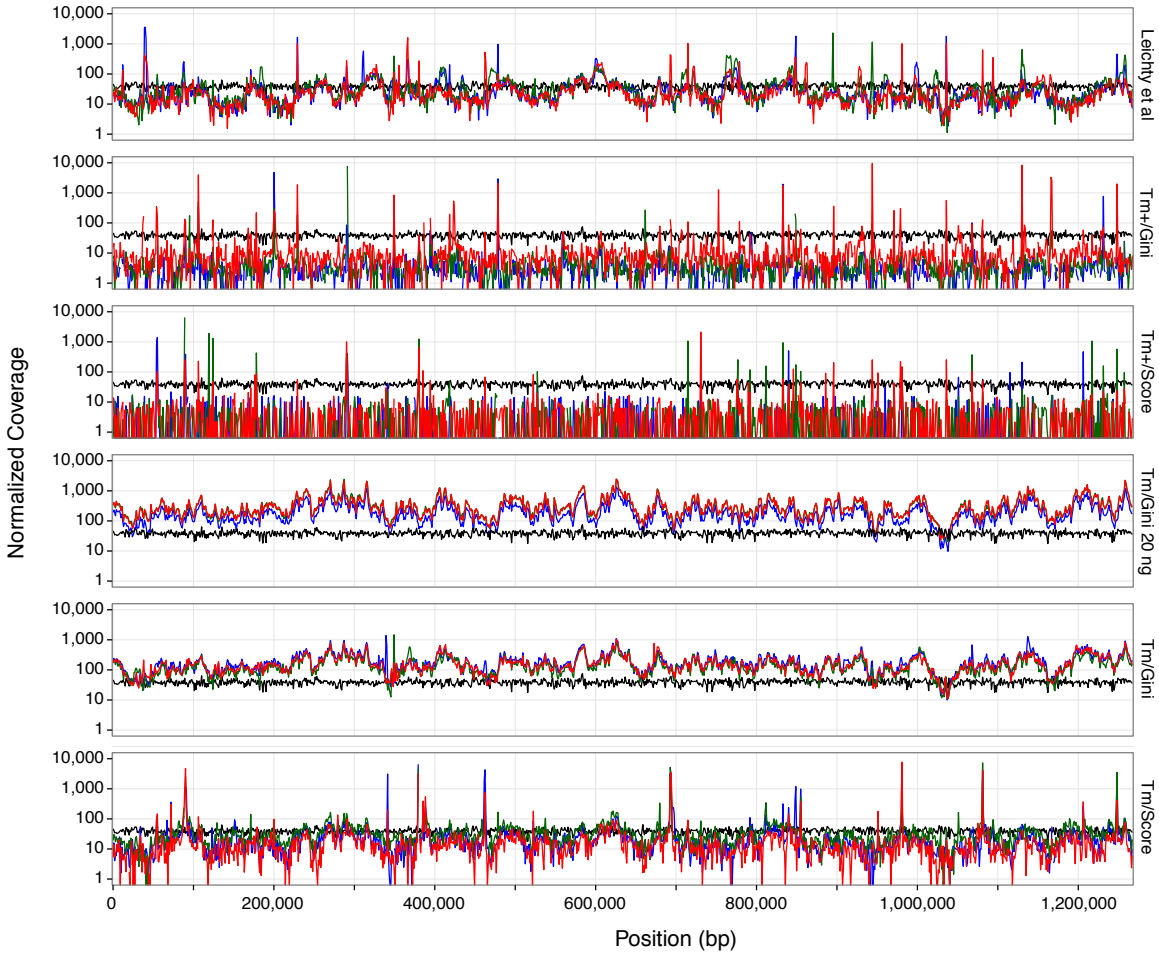

Figure 5-4. Normalized sequencing coverage across the *Wolbachia* genome after SWGA.

Normalized coverage (1,000,000,000 * Fold Coverage / Total bp Sequenced) of the *Wolbachia* genome is shown for each primer set. Three replicate SWGA amplifications are shown in red, blue, and green. The black line in each plot window represents coverage for an unamplified sample.

The evenness of sequence coverage was weakly associated with primer density among primer sets. For example, the primer sets chosen with standard melting temperatures had more than twice as many total binding sites in the *W. pipientis* genome than primer sets with high melting temperatures and the average sequencing coverage was several orders of magnitude greater (Table 5-1). Within each primer set, however, variation in primer density across the genome was not correlated with local sequence coverage (Figure 5-5). For example, both primer sets chosen with high melting temperatures had many 20,000 bp regions in which there were no primer binding sites but these regions did not have lower sequence coverage than the regions that had substantially higher densities of primer binding sites (>70, Figure 5-5).

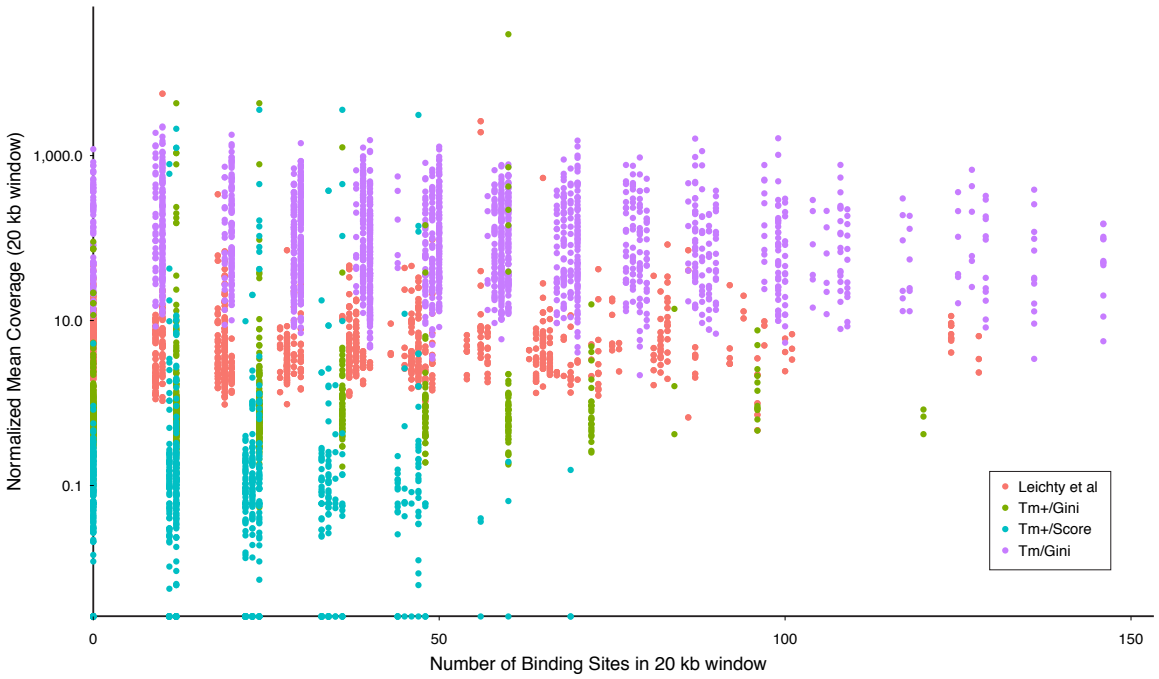

Figure 5-5. Relationship between primer binding site density and mean sequencing coverage after SWGA.

Mean normalized coverage within a 20,000 bp sliding window is shown relative to the number of binding sites within the same 20,000 bp window after SWGA. SWGA reactions were performed in triplicate using either *swga* primer sets or the primer set from *Leichty* et al. Coverage was calculated separately for each reaction and then combined to generate the plot.

5.4 Discussion

Microbial genomics is difficult when the genome of interest cannot be separated from contaminating DNA. In these situations, the majority of sequencing reads may be derived from contaminating DNA, increasing the sequencing effort required to get substantial coverage of the genome of interest. SWGA overcomes this by selectively enriching for a target genome from a complex DNA mixture (*39*), however this method requires the computationally difficult identification and validation of selective primer sets. *Swga* can identify and quantitatively evaluate millions of primers and primer sets to facilitate SWGA primer design. *Swga* quickly and efficiently identified sets for the amplification of *Wolbachia* from infected *Drosophila*. Moreover, at least one of the chosen sets performed significantly better than previously published (*39*), hand picked, primers.

The *swga* program identifies primers that are common in the target species and rare in the background and rapidly evaluates primer sets on their potential to amplify the target genome. Computational evaluation of thousands of primer sets provides a major advance over the previous implementation of SWGA in which the user hand-assembled a small number of primer sets (*39*). In the current version of *swga*, primer sets are evaluated on multiple criteria that are logically associated with phi29 amplification. The *swga* program can also be altered by the end user to add new evaluative criteria as correlations between amplification and primer set characteristics become available.

All of the primer sets chosen by the *swga* program and by human users (*39*) selectively amplified *W. pipientis* DNA from infected fruit flies, such that the proportion of sequencing reads that mapped to *Wolbachia* was at least 3 times greater than reads from the unamplified sample. The criteria used to choose the Tm/Score primer set were chosen to maximize the number of binding sites in the target genome and minimize the

number of binding sites in the background, while the evenness of primers across the target genome was not emphasized, which were based on the same principles used by Leichty and Brisson (2014). As may be expected, sequencing coverage after amplification with either of these primer sets was uneven (Figure 5-4). However, the decrease in sequencing effort required to yield broad coverage of the *Wolbachia* genome was substantially lower after amplification with the primers chosen by the *swga* program (Figure 5-1).

The *swga* program efficiently chooses primer sets to selectively amplify target microbial genomes without prior culture or molecular separation. However, it is unlikely that the currently implemented criteria are ideal for identifying the best primer sets to evenly amplify a target genome. The criteria used to choose primer sets could be improved by a better understanding of the biochemistry of the phi29 enzyme and by a careful evaluation of primer set characteristics against amplification and coverage evenness across many different primer sets. Sequence coverage and primer characteristics from both successful and unsuccessful SWGA amplifications could allow an empirical investigation of primer set characteristics that result in strong and even amplification. These criteria could be then incorporated into the *swga* program during future updates.

CHAPTER 6 – Summary and Future Directions

6.1 Summary

African great apes are infected with a plethora of *Plasmodium* species including the closest relatives and direct ancestors of *P. falciparum* (*26*, *110*). Studying these parasites can provide insights into the evolutionary history of this important pathogen and the genetic changes required to colonize humans. Chapter two of this dissertation shows that ape *Laverania* parasites do not recurrently infect humans, suggesting that there exists one or more blocks to cross-species transmission among these parasites. Chapters three and four describe the first comparative genomic analyses of both close and distant ape relatives of *P. falciparum*. These analyses identify features that are shared across the *Laverania* subgenus as well as some that are unique to the ancestry of *P. falciparum*. Chapter five provides a computational framework for efficient SWGA primer design that makes this genome enrichment strategy more accessible and easier to implement for any set of target and non-target genomes. Together, this dissertation offers the first genome wide insights into the *Laverania* subgenus and provides a foundation for future work. Here I propose future studies that may further elucidate the basis of *Laverania* host-specificity and the evolutionary steps that gave rise to *P. falciparum*.

6.2 Future Directions

Identifying barriers to cross species transmission

Chapter two of this dissertation describes the first large-scale screen for ape *Laverania* infections in humans in West Central Africa. While this study focused on populations within Cameroon, an additional study found no evidence of *Laverania* infections in forest-dwelling humans from Gabon (*181*). These data support the hypothesis that human infections with ape *Laverania* parasites are incredibly rare (*26*). Importantly, these studies focused solely on blood samples. It is possible that humans are exposed to ape *Laverania* parasites, but that these parasites cannot establish a blood stage infection and are therefore missed in blood based screens. Additional screening studies may allow us to determine if humans are exposed to ape *Laverania* parasites and if these parasites can establish pre-erythrocytic stage infections. Future studies should include of mosquito based screens, to determine whether ape *Laverania* infected mosquitoes feed on humans, and non-invasive pre-erythrocytic stage screens, to identify liver stage infections that fail to progress to the blood.

Mosquitoes are essential to the transmission of all *Plasmodium* parasites and may therefore play an important role in facilitating or preventing cross species transmission (*182*, *183*). While the major vectors of human malaria in West Africa are well characterized (*184-186*), little is known about the mosquitoes that transmit malaria between wild African apes. Mosquito studies can answer two important questions. First, blood meal screenings of mosquitoes caught near great apes populations will help pinpoint mosquito species that regularly feed on these species. Second, screens for *Plasmodium* infections in these mosquito populations will identify which species act as competent vectors for ape *Laverania* parasites. Once potential ape *Laverania* vectors have been identified, mosquito catches in human settlements, especially those that are

located within the home range of wild ape populations, will allow us to determine if certain mosquito species may act as vectors for cross species between apes and humans.

While the proposed mosquito studies will be useful in determining whether mosquitoes serve as a block to *Laverania* cross species transmission, it is important to recognize the technical challenges associated with mosquito collection. Traps typically rely on CO_2 or light to attract mosquitoes (*187*). CO_2 traps, while more effective (*187*), require a source of CO_2, which may be difficult to come by in more remote settings. The latter traps attract a wide variety of insects, and their contents must be sorted to remove unwanted species (DE Loy, unpublished). As it is likely that the proportion of both *Laverania* infected mosquitoes and ape blood fed mosquitoes is very small, these studies may require large numbers of mosquitoes to identify species that regularly feed on apes and those that serve as vectors for ape malaria. Some of these obstacles may be overcome by performing studies at chimpanzee sanctuaries. Sanctuaries provide a more controlled environment with fixed and accessible nesting sites that should allow the placement of more permanent traps. However, these sanctuaries may not be completely representative of natural ape habitats, and field studies will be necessary to ensure that these sanctuary based studies do not introduce unforeseen bias.

If humans are exposed to ape *Laverania* infected mosquitoes, it is possible that these parasites establish a liver stage infection but are unable to progress to the blood stage. Recent studies of the rodent parasite *P. yoelli* have shown that parasite DNA from pre-erythrocytic stage infections can be detected in fecal samples (*188*). If this holds true for the *Laverania* parasites, fecal PCR could be used as a non-invasive screening tool for liver stage infections in humans. If ape *Laverania* parasite DNA can be identified in

human populations, it would indicate that humans are exposed to ape *Laverania* and potentially develop liver stage infections.

Screening for antibodies to ape *Laverania* antigens could provide another means of detecting liver or aborted blood stage infections. Protein microarrays are now commonly used in studies of *P. falciparum* to compare the antibody profiles of exposed and naïve individuals (*189*). A similar approach could be used to screen for *Laverania* exposure, using protein sequences from the *P. gaboni* and *P. reichenowi* genomes. If such a study were to be undertaken, it would be important to differentiate true *Laverania* exposure from cross-reactivity to the antigens of endemic human *Plasmodium* species. This could be achieved by comparing the antibody profiles of individuals living near infected ape habitats with those from other human malaria endemic regions. The fecal DNA and antibody-based studies described here would complement mosquito studies. They would provide evidence for human exposure to ape *Laverania*, indicating that at least some ape *Laverania* vectors feed on both humans and apes.

While the studies I have proposed here may identify specific stages at which cross species transmission is prevented, it is reasonable to expect that the lack of zoonotic *Laverania* infections is multifactorial. This appears to be the case for *P. falciparum* infections of chimpanzees. While these infections have never been detected in wild living apes, our lab and others have identified multiple instances of reverse zoonosis in sanctuary chimpanzees (unpublished observation). This is not unexpected. Sanctuary chimpanzees are likely exposed to *P. falciparum* much more frequently than wild populations. Even in sanctuaries, however *P. falciparum* infections represent a very small proportion of the total malaria burden. These data would therefore suggest that, while a lack of exposure to infected mosquitoes limits cross species transmission, species-specific interactions at later stages of parasite development are also important.

Population genetics of ape *Laverania* parasites

While this dissertation provides the first genomic level analysis of close and distant chimpanzee relatives of *P. falciparum*, additional ape *Laverania* population genomic studies will provide further insight into the evolutionary history of *P. falciparum*. Current ape *Laverania* genomic studies have focused solely on chimpanzee parasites (*69*). These studies cannot differentiate between evolutionary events that occurred during the emergence of *P. falciparum* and those that were present in *P. praefalciparum*. The identification of adaptive changes that allowed *P. falciparum* to colonize humans requires a direct comparison of *P. falciparum* to its gorilla ancestor. Blood samples from gorillas are extremely difficult to obtain due to the endangered status of these apes and the low numbers of gorillas in sanctuaries in West Africa. Thus, obtaining complete genomes of *P. praefalciparum* and other gorilla parasites will require the development of new selective enrichment strategies from non-invasively collected samples. One unexplored source of *Laverania* infected gorilla blood are sanguinivorus insects. While many such insects would not be susceptible to *Laverania* infection, those caught soon after feeding may still contain intact *Laverania* parasites in the blood meal. Experiments to determine if these blood meals can serve as sources of full length *Laverania* genomes are ongoing (*100*).

Genome-wide comparisons of *P. falciparum*, *P. reichenowi* and *P. gaboni* identified the first evidence of horizontal gene transfer between two *Plasmodium* species. Strikingly, the horizontally transferred segment contains two essential invasion genes, *RH5* and *CyRPA*, which define the 3' and 5' ends of the segment. The maintenance and complete fixation of *P. adleri* derived alleles of both of these genes in *P. praefalciparum* suggests that this horizontally transferred segment was selected for. It is also possible, however, that this region was fixed in *P. praefalciparum* due to random

drift. These two phenomenon may be differentiated through genomewide comparisons of multiple *P. praefalciparum* isolates. Fixation due to recent selection should produce signatures of a recent selective sweep, a reduction in nucleotide diversity in the genes surrounding the region under selection. Fixation due to random drift, on the other hand, would not be expected to yield this phenomenon. Evidence of selection for the transferred alleles would support our hypothesis that the *P. adleri* derived *RH5* and *CyRPA* provided a fitness advantage for *P. praefalciparum*, and that this HGT event may have predisposed *P. praefalciparum* to infect humans.

The identification of a horizontal gene transfer event between two distinct *Plasmodium* species suggests a potential novel mechanism for genetic innovation in these parasites. This gene transfer event may have occurred by one of two mechanisms: sexual recombination followed by successive backcrossing, or asexual DNA transfer between two parasites. DNA transfer between *Plasmodium* infected erythrocytes is known to occur in cultures of *P. falciparum*. Recent work by Regev-Rudzki *et al* has shown that exosome-like vesicles can transport genomic or plasmid DNA between infected erythrocytes and that transferred DNA can be expressed by the recipient parasite (*97*). It is therefore possible DNA could be exchanged between distinct *Plasmodium* species during co-infection. It seems less likely that this region was exchanged via sexual recombination and subsequent backcrossing. Horizontal gene transfer by this mechanism would require an F1 generation parasite that was viable despite having inherited genes from two divergent, non-recombining, species. Moreover, it would require that all the offspring of subsequent backcrosses remained viable. Nevertheless, neither mechanism can be ruled out, as we cannot estimate the probability or frequency of HGT from a single ancestral event. Future population genomic studies of *Plasmodium* species may elucidate this by searching for inter-

species horizontal gene transfer events, especially in populations where mixed infections are common.

In their comparison of *P. falciparum* to *P. reichenowi* CDC1, Otto *et al* emphasize the results of screens for adaptive selection such as the McDonald-Kreitman (MK) test (*69*). The MK test compares the ratio of non-synonymous to synonymous fixed differences (between species, Dn/Ds) to the ratio of non-synonymous to synonymous polymorphisms (within species, Pn/Ps) (*190*). An excess in non-synonymous fixed differences between species (Dn/Ds > Pn/Ps) is suggestive of adaptive evolution, while equal ratios of Dn/Ds and Pn/Ps are suggestive of neutral evolution (*190*). Surprisingly, when we applied genome-wide MK tests to the a set of global *P. falciparum* isolates and *P. reichenowi* SY57, few genes were found to be significant after controlling for multiple hypothesis testing, and no genes showed significant evidence of adaptive selection. Instead, we observed an excess of non-synonymous polymorphism within the global population of *P. falciparum*. It has previously been proposed that this excess in non-synonymous polymorphism is derived from the extreme A-T richness of the *P. falciparum* genome (*191*), although this has been disputed (*192*). Another possibility is that the excess in nonsynonmyous polymorphism is the result of a recent *P. falciparum* population bottleneck (*190*, *193*). Sequencing additional *P. gaboni* or *P. reichenowi* genomes will help elucidate the cause of the observed excess of non-synonymous polymorphism in *P. falciparum*. If the excess of non-synonymous polymorphism were due to high A-T content, we would expect to observe a similar excess in Pn in all *Laverania* species. If, on the other hand, this excess were related to the recent population bottleneck in *P. falciparum* we would not expect to see an excess in non-synonymous polymorphism in other *Laverania* species. Identifying the cause of this

excess may enable us to control for it in future analyses, increasing the power of evolutionary tests to detect adaptive selective in *P. falciparum*.

The comparisons of *P. falciparum var* genes with *var*-like genes from other *Laverania* species demonstrate the ancient origins of this gene family. Our current analyses are, however, limited in their reliance on *P. falciparum var* sequences to query the *Laverania* genomes and to guide the design of *var*-specific primers. It is therefore possible that additional *var*-like genes or *var*-gene domains, such as CIDR domains, were missed because they are divergent from those found in *P. falciparum*. Additional whole genome sequencing and assembly may help to elucidate by yielding more complete *var*-like gene sequences.

***In vitro* studies**

Comparative analyses of *P. falciparum* and ape *Laverania* parasites have identified specific genes that may be important for adaptation to humans. While *var* gene analyses across the *Laverania* subgenus indicate that the precursors of the *var* family existed in the *Laverania* ancestor, key differences exist between *P. falciparum var* genes and *var*-like genes in the more distantly related *Laverania* parasite, *P. gaboni.* One important distinction is the lack of CIDR domains in *P. gaboni*. These domains have been shown to be important for binding to host receptors (*194*, *195*), and their absence in *P. gaboni* may indicate that the binding and sequestration properties of *Laverania* parasites have continued to evolve since the radiation of the subgenus. While the current *P. gaboni* assembly lacks full length *var* genes, this can be remedied by additional sequencing using long read high throughput technologies (*196*, *197*). Binding studies, using *P. falciparum* parasites that express full length ape *Laverania var* genes, will help determine if *var* gene function is conserved across the subgenus and, if so, whether the

var genes of divergent ape *Laverania* bind the similar targets as their *P. falciparum* orthologues.

As the extant *P. falciparum RH5* allele was derived via horizontal gene transfer, binding studies to determine the species specificity of the RH5-basigin interaction are of particular interest. While previous studies have shown that the interaction between RH5 and basigin limits the host tropism of *P. falciparum* (*102*), it is unclear if this is the case for other *Laverania* parasites. Unfortunately, the fixation of the *P. adleri* derived *RH5* allele in *P. praefalciparum* precludes a direct comparison of the binding properties of pre- and post-HGT RH5. Comparisons can, however, be made to other *RH5* alleles. While *in vitro* binding assays can identify differences in the binding interactions of various RH5-basigin combinations (*102*), a more direct method of determining the effect of this interaction on invasion would be erythrocyte invasion assays using transgenic parasites expressing *Laverania* RH5. A lack of invasion by some or all *Laverania* RH5 transgenic strains would be strongly support the role of the RH5-basigin interaction in determining *Laverania* host specificity.

In summary, this dissertation provides a foundation for future studies of the ape *Laverania* subgenus. While these parasites are endemic and present at high levels in wild African apes (*26*), they are not a source of recurrent human infections. Comparative genomics of close and distant relatives of *P. falciparum* has identified features that are shared across the subgenus, as well as those that are unique to the ancestry of *P. falciparum*. Additional studies are required to further our understanding of the barriers to cross species transmission and to perform direct comparisons of *P. falciparum* to its closest ancestor, *P. praefalciparum*. These studies will not only expand our understanding of the evolutionary origins of *P. falciparum*, but may also identify

previously unknown host-parasite interactions that can serve as a basis for future therapeutic interventions.

BIBLIOGRAPHY

1. World Health Organization, *World Malaria Report 2014* (Geneva, Switzerland, 2014).

2. R. W. Snow, C. A. Guerra, A. M. Noor, H. Y. Myint, S. I. Hay, The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature*. **434**, 214–217 (2005).

3. Q. Bassat, P. L. Alonso, Defying malaria: Fathoming severe Plasmodium vivax disease. *Nat. Med.* **17**, 48–49 (2011).

4. J. Cox-Singh *et al.*, *Plasmodium knowlesi* Malaria in Humans Is Widely Distributed and Potentially Life Threatening. *Clin. Infect. Dis.* **46**, 165–171 (2008).

5. B. Singh *et al.*, A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *Lancet*. **363**, 1017–1024 (2004).

6. C. Wongsrichanalai, S. R. Meshnick, Declining artesunate-mefloquine efficacy against falciparum malaria on the Cambodia-Thailand border. *Emerg Infect Dis*. **14**, 716–719 (2008).

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

8. D. C. Kaslow, S. Biernaux, RTS,S: Toward a first landmark on the Malaria Vaccine Technology Roadmap. *Vaccine* (2015), doi:10.1016/j.vaccine.2015.09.061.

9. M. J. Gardner *et al.*, Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*. **419**, 498–511 (2002).

10. J. Baum *et al.*, Molecular genetics and comparative genomics reveal RNAi is not functional in malaria parasites. *Nucleic Acids Res.* **37**, gkp239–3798 (2009).

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

12. E. Reichenow, Ueber das vorkommen der malariaparasiten des menschen bei den afrikanischen menschenaffen. *Centralbl. f. Bakt. I. Abt. Orig*. **85**, 207–216 (1920).

13. G. R. Coatney, W. E. Collins, M. Warren, P. G. Contacos, *The primate malarias* (Division of Parasitic Disease, Atlanta, Georgia, ed. 1, 2003).

14. R. S. Bray, Studies on malaria in chimpanzees. I. The erythrocytic forms of *Plasmodium reichenowi. J. Parasitol.* **42**, 588–592 (1956).

15. P. C. Garnham, R. Lainson, A. E. Gunders, Some observations on malaria parasites in a chimpanzee, with particular reference to the persistence of Plasmodium reichenowi and Plasmodium vivax. *Ann Soc Belg Med Trop (1920)*. **36**, 811–821 (1956).

16. D. L. Doolan, C. Dobaño, J. K. Baird, Acquired immunity to malaria. *Clinical Microbiology Reviews*. **22**, 13–36– Table of Contents (2009).

17. L. H. Miller, S. J. Mason, D. F. Clyde, M. H. McGinniss, The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N. Engl. J. Med.* **295**, 302–304 (1976).

18. C. A. Guerra *et al.*, The international limits and population at risk of *Plasmodium vivax* transmission in 2009. *PLoS Negl Trop Dis*. **4**, e774 (2010).

19. A. A. Escalante, F. J. Ayala, Phylogeny of the malarial genus Plasmodium, derived from rRNA gene sequences. *PNAS*. **91**, 11373–11377 (1994).

20. P. C. C. Garnham, Malaria Parasites and other Haemosporidia. (1966).

21. R. Carter, K. N. Mendis, Evolutionary and historical aspects of the burden of malaria. *Clinical Microbiology Reviews*. **15**, 564–594 (2002).

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

23. A. A. Escalante *et al.*, A monkey's tale: the origin of *Plasmodium vivax* as a human malaria parasite. *Proc. Natl. Acad. Sci. USA*. **102**, 1980–1985 (2005).

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

25. F. Prugnolle *et al.*, African great apes are natural hosts of multiple related malaria species, including *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. USA*. **107**, 1458–1463 (2010).

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

27. M. Kaiser *et al.*, Wild chimpanzees infected with 5 *Plasmodium* species. *Emerg Infect Dis*. **16**, 1956–1959 (2010).

28. B. Ollomo *et al.*, A new malaria agent in African hominids. *PLoS Pathog.* **5**, e1000446 (2009).

29. S. M. Rich *et al.*, The origin of malignant malaria. *Proc. Natl. Acad. Sci. USA*. **106**, 14902–14907 (2009).

30. S. Krief *et al.*, On the diversity of malaria parasites in African apes and the origin

of *Plasmodium falciparum* from Bonobos. *PLoS Pathog.* **6**, e1000765 (2010).

31. L. Duval *et al.*, African apes as reservoirs of *Plasmodium falciparum* and the origin and diversification of the *Laverania* subgenus. *Proc. Natl. Acad. Sci. USA*. **107**, 10561–10566 (2010).

32. J. C. Rayner, W. Liu, M. Peeters, P. M. Sharp, B. H. Hahn, A plethora of *Plasmodium* species in wild apes: a source of human infection? *Trends Parasitol*. **27**, 222–229 (2011).

33. R. S. Bray, Studies on malaria in chimpanzees. VI. Laverania falciparum. *Am J Trop Med Hyg*. **7**, 20–24 (1958).

34. P. M. Sharp, B. H. Hahn, Origins of HIV and the AIDS Pandemic. *Cold Spring Harbor Perspectives in Medicine*.

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

36. S. Auburn *et al.*, An effective method to purify *Plasmodium falciparum* DNA directly from clinical blood samples for whole genome high-throughput sequencing. **6**, e22213 (2011).

37. A. T. Bright *et al.*, Whole genome sequencing analysis of Plasmodium vivax using whole genome capture. *BMC Genomics*. **13**, 262 (2012).

38. S. O. Oyola *et al.*, Efficient depletion of host DNA contamination in malaria clinical sequencing. *J. Clin. Microbiol.* **51**, 745–751 (2013).

39. A. R. Leichty, D. Brisson, Selective whole genome amplification for resequencing target microbial species from complex natural samples. *Genetics*. **198**, 473–481 (2014).

40. World Health Organization, *World Malaria Report 2011* (World Health Organization, Geneva, Switzerland, 2011; http://www.who.int/entity/malaria/world_malaria_report_2011/9789241564403_eng.pdf).

41. P. L. Alonso *et al.*, A research agenda to underpin malaria eradication. *PLoS Med.* **8**, e1000406 (2011).

42. malERA Consultative Group on Basic Science and Enabling Technologies *et al.*, A research agenda for malaria eradication: basic science and enabling technologies. *PLoS Med.* **8**, e1000399 (2011).

43. F. Prugnolle *et al.*, A Fresh Look at the Origin of *Plasmodium falciparum*, the Most Malignant Malaria Agent. *PLoS Pathog.* **7**, e1001283 (2011).

44. N. O. Verhulst, R. C. Smallegange, W. Takken, Mosquitoes as potential bridge

vectors of malaria parasites from non-human primates to humans. *Frontiers in physiology*. **3** (2012), doi:10.3389/fphys.2012.00197.

45. C. Laurent *et al.*, Commercial logging and HIV epidemic, rural Equatorial Africa. *Emerg Infect Dis*. **10**, 1953–1956 (2004).

46. A. Spielman *et al.*, Malaria diagnosis by direct observation of centrifuged samples of blood. *Am J Trop Med Hyg*. **39**, 337–342 (1988).

47. M. Margulies *et al.*, Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. **437**, 376–380 (2005).

48. R. Mukherjee *et al.*, Switching between raltegravir resistance pathways analyzed by deep sequencing. *AIDS* (2011), doi:10.1097/QAD.0b013e32834b34de.

49. C. Wang, Y. Mitsuya, B. Gharizadeh, M. Ronaghi, R. W. Shafer, Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res.* **17**, 1195–1201 (2007).

50. J. Binladen *et al.*, The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. **2**, e197 (2007).

51. J. F. Salazar-Gonzalez *et al.*, Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol*. **82**, 3952–3970 (2008).

52. S. Guindon, F. Delsuc, J. F. Dufayard, O. Gascuel, Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.* **537**, 113–137 (2009).

53. R. L. Culleton, P. E. Ferreira, Duffy Phenotype and *Plasmodium vivax* infections in Humans and Apes, Africa. *Emerg Infect Dis*. **18**, 1704–1705 (2012).

54. J. C. Rayner, C. S. Huber, M. R. Galinski, J. W. Barnwell, Rapid evolution of an erythrocyte invasion gene family: the *Plasmodium reichenowi* Reticulocyte Binding Like (RBL) genes. *Mol. Biochem. Parasitol.* **133**, 287–296 (2004).

55. C. Crosnier *et al.*, Basigin is a receptor essential for erythrocyte invasion by *Plasmodium falciparum*. *Nature*. **480**, 534–537 (2011).

56. G. R. Coatney, Simian malarias in man: facts, implications, and predictions. *Am J Trop Med Hyg*. **17**, 147–155 (1968).

57. J. Rodhain, R. Dellaert, Contribution à l'étude du *Pl. schwetzi* E. Brumpt. III. L'infection à *Plasmodium schwetzi* chez l'homme. *Ann. Soc. Belg. de Med. Trop.* **35**, 757–775 (1955).

58. P. G. Contacos *et al.*, Transmission of *Plasmodium schwetzi* from the chimpanzee to man by mosquito bite. *Am J Trop Med Hyg*. **19**, 190–195 (1970).

59. J. Rodhain, Contribution à l'étude des *Plasmodiums* des anthropoïdes africains. *Ann. Soc. Belg. de Med. Trop.* **28**, 39–49 (1948).

60. J. Rodhain, Les *plasmodiums* des anthropoïdes de l'Afrique centrale et leurs relations avec les *plasmodiums* humains. *Ann. Soc. Belg. de Med. Trop.* **20**, 489–505 (1940).

61. J. Rodhain, *Plasmadia* of Central African Apes and their Relationship to Human Plasmodia. *Ann. Soc. Belg. de Med. Trop.* **20**, 489–505 . (1940).

62. A. L. Hughes, F. Verra, Malaria parasite sequences from chimpanzee support the co-speciation hypothesis for the origin of virulent human malaria (*Plasmodium falciparum*). *Mol Phylogenet Evol*. **57**, 135–143 (2010).

63. R. Culleton *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).

64. J. M. Rubio *et al.*, Semi-nested, multiplex polymerase chain reaction for detection of human malaria parasites and evidence of *Plasmodium vivax* infection in Equatorial Guinea. *Am J Trop Med Hyg*. **60**, 183–187 (1999).

65. C. Mendes *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).

66. W. Liu *et al.*, Widespread Infection of Wild-Living Chimpanzees and Gorillas With *Plasmodium Vivax*-Like Parasites. *ASTMH 60th Annual Meeting* (2011).

67. A. G. Schneider, O. Mercereau-Puijalon, A new *Apicomplexa*-specific protein kinase family: multiple members in *Plasmodium falciparum*, all with an export signature. *BMC Genomics*. **6**, 30 (2005).

68. S. L. Perkins, Species concepts and malaria parasites: detecting a cryptic species of *Plasmodium*. *Proc. Biol. Sci.* **267**, 2345–2350 (2000).

69. T. D. Otto *et al.*, Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat. Commun.* **5**, 4754 (2014).

70. A. Pain *et al.*, The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature*. **455**, 799–803 (2008).

71. J. M. Carlton *et al.*, Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*. **455**, 757–763 (2008).

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

73. J. M. Carlton *et al.*, Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature*. **419**, 512–519 (2002).

74. W. E. Collins, J. C. Skinner, M. Pappaioanou, J. R. Broderson, P. Mehaffey, The sporogonic cycle of *Plasmodium reichenowi*. *J. Parasitol.* **72**, 292–298 (1986).

75. F. B. Dean, J. R. Nelson, T. L. Giesler, R. S. Lasken, Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11**, 1095–1099 (2001).

76. J. G. Paez *et al.*, Genome coverage and sequence fidelity of Phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.* **32**, e71–e71 (2004).

77. L. Blanco *et al.*, Highly efficient DNA synthesis by the phage Phi29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* **264**, 8935–8940 (1989).

78. H. Yokouchi *et al.*, Whole-metagenome amplification of a microbial community associated with scleractinian coral by multiple displacement amplification using Phi29 polymerase. *Environ. Microbiol.* **8**, 1155–1163 (2006).

79. P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertil, Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.* **16**, 1391–1399 (1999).

80. D. B. Larremore *et al.*, Ape origins of human malaria virulence genes. *Nat. Commun.*

81. M. T. Swain *et al.*, A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat. Protoc.* **7**, 1260–1284 (2012).

82. S. M. Rich, M. C. Licht, R. R. Hudson, F. J. Ayala, Malaria's Eve: Evidence of a recent population bottleneck throughout the world populations of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. USA*. **95**, 4425–4430 (1998).

83. D. L. Hartl, The origin of malaria: mixed messages from genetic diversity. *Nat. Rev. Microbiol.* **2**, 15–22 (2004).

84. H. H. Chang *et al.*, Malaria life cycle intensifies both natural selection and random genetic drift. *Proc. Natl. Acad. Sci. USA*. **110**, 20129–20134 (2013).

85. H. H. Chang, D. L. Hartl, Recurrent bottlenecks in the malaria life cycle obscure signals of positive selection. *Parasitology*. **142**, S98–S107 (2015).

86. J. Prado-Martinez *et al.*, Great ape genetic diversity and population history. *Nature*. **499**, 471–475 (2013).

87. N. Rovira-Graells *et al.*, Transcriptional variation in the malaria parasite

Plasmodium falciparum. *Genome Res.* **22**, 925–938 (2012).

88. L. M. Kats *et al.*, An exported kinase (FIKK4.2) that mediates virulence-associated changes in *Plasmodium falciparum*-infected red blood cells. *Int. J. Parasitol.* **44**, 319–328 (2014).

89. G. S. Brandt, S. Bailey, Dematin, a human erythrocyte cytoskeletal protein, is a substrate for a recombinant FIKK kinase from *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **191**, 20–23 (2013).

90. M. C. Nunes, M. Okada, C. Scheidig-Benatar, B. M. Cooke, A. Scherf, *Plasmodium falciparum* FIKK kinase members target distinct components of the erythrocyte membrane. *PLoS One*. **5**, e11747 (2010).

91. Z. Bozdech *et al.*, The transcriptome of *Plasmodium vivax* reveals divergence and diversity of transcriptional regulation in malaria parasites. *Proc. Natl. Acad. Sci. USA*. **105**, 16290–16295 (2008).

92. T. J. Sargeant *et al.*, Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biol.* **7**, R12 (2006).

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

94. A. G. Maier, B. M. Cooke, A. F. Cowman, L. Tilley, Malaria parasite proteins that remodel the host erythrocyte. *Nat. Rev. Microbiol.* **7**, 341–354 (2009).

95. K. S. Reddy *et al.*, Multiprotein complex between the GPI-anchored CyRPA with PfRH5 and PfRipr is crucial for *Plasmodium falciparum* erythrocyte invasion. *Proc. Natl. Acad. Sci. USA*. **112**, 1179–1184 (2015).

96. K. Deitsch, C. Driskill, T. Wellems, Transformation of malaria parasites by the spontaneous uptake and expression of DNA from human erythrocytes. *Nucleic Acids Res.* **29**, 850–853 (2001).

97. N. Regev-Rudzki *et al.*, Cell-cell communication between malaria-Infected red blood cells via exosome-like vesicles. *Cell*. **153**, 1120–1133 (2013).

98. J. Felsenstein, Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* (1985).

99. G. J. Wright, J. C. Rayner, *Plasmodium falciparum* erythrocyte invasion: combining function with immune evasion. *PLoS Pathog.* **10**, e1003943 (2014).

100. D. E. Loy *et al.*, (http://www.abstractsonline.com/Plan/ViewAbstract.aspx?sKey=3581d347-82ee-47c1-9ecb-94ff44dec031&cKey=e5f9157a-e194-44bc-a3b5-95bc224b2db7&mKey=52ae2426-7f12-4d2b-9404-c0d0b5a8eb5a).

101. C. Paupy *et al.*, *Anopheles moucheti* and *Anopheles vinckei* are candidate vectors of ape *Plasmodium* parasites, including *Plasmodium praefalciparum* in Gabon. *PLoS One*. **8**, e57294 (2013).

102. M. Wanaguru, W. Liu, B. H. Hahn, J. C. Rayner, G. J. Wright, RH5-Basigin interaction plays a major role in the host tropism of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. USA*. **110**, 20735–20740 (2013).

103. S. A. Sundararaman *et al.*, *Plasmodium falciparum*-like parasites infecting wild apes in southern Cameroon do not represent a recurrent source of human malaria. *Proc. Natl. Acad. Sci. USA*. **110**, 7020–7025 (2013).

104. A. L. Hughes, F. Verra, Very large long-term effective population size in the virulent human malaria parasite *Plasmodium falciparum*. *Proc. Biol. Sci.* **268**, 1855–1860 (2001).

105. P. W. Hedrick, Population genetics of malaria resistance in humans. *Heredity*. **107**, 283–304 (2011).

106. M. Coluzzi, The clay feet of the malaria giant and its African roots: hypotheses and inferences about origin, spread and control of *Plasmodium falciparum*. *Parassitologia*. **41**, 277–283 (1999).

107. F. B. Livingstone, Anthropological implications of sickle cell gene distribution in West Africa. *American Anthropologist*. **60**, 533–562 (1958).

108. R. Carter, K. Mendis, Evolutionary and Historical Aspects of the Burden of Malaria. *Clinical Microbiology Reviews*. **15**, 564 (2002).

109. A. Molina-Cruz *et al.*, The human malaria parasite *Pfs47* gene mediates evasion of the mosquito immune system. *Science*. **340**, 984–987 (2013).

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

111. W. Liu *et al.*, Single genome amplification and direct amplicon sequencing of *Plasmodium spp*. DNA from ape fecal specimens. *Nature Protocol Exchange*. **2010** (2010), doi:10.1038/nprot.2010.156.

112. J. T. Simpson, R. Durbin, Efficient construction of an assembly string graph using the FM-index. *Bioinformatics*. **26**, i367–73 (2010).

113. C. T. Brown, A. Howe, Q. Zhang, A. B. Pyrkosz, A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv* (2012).

114. D. R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* (2008).

115. M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, W. Pirovano, Scaffolding pre-

assembled contigs using SSPACE. *Bioinformatics*. **27**, 578–579 (2011).

116. T. Carver, S. R. Harris, M. Berriman, J. Parkhill, J. A. McQuillan, Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*. **28**, 464–469 (2012).

117. F. Nadalin, F. Vezzi, A. Policriti, GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* (2012).

118. I. J. Tsai, T. D. Otto, M. Berriman, Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* **11**, R41 (2010).

119. T. D. Otto, M. Sanders, M. Berriman, C. Newbold, Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics*. **26**, 1704–1707 (2010).

120. A. Bankevich *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

121. Y. Safonova, A. Bankevich, P. A. Pevzner, dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. *J. Comput. Biol.* **22**, 528–545 (2015).

122. T. D. Otto, G. P. Dillon, W. S. Degrave, M. Berriman, RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.* **39**, e57–e57 (2011).

123. M. Stanke, S. Waack, Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*. **19 Suppl 2**, ii215–25 (2003).

124. L. A. Ware *et al.*, Two alleles of the 175-kilodalton *Plasmodium falciparum* erythrocyte binding antigen. *Mol. Biochem. Parasitol.* **60**, 105–109 (1993).

125. S. W. Roy, M. U. Ferreira, D. L. Hartl, Evolution of allelic dimorphism in malarial surface antigens. *Heredity*. **100**, 103–110 (2008).

126. F. Abascal, R. Zardoya, M. J. Telford, TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–13 (2010).

127. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

128. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. **20**, 289–290 (2004).

129. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

130. M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

131. G. A. Van der Auwera *et al.*, From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*. **11**, 11.10.1–11.10.33 (2013).

132. S. A. Assefa *et al.*, estMOI: estimating multiplicity of infection using parasite deep sequencing data. *Bioinformatics*. **30**, 1292–1294 (2014).

133. M. A. Larkin *et al.*, Clustal W and Clustal X version 2.0. *Bioinformatics*. **23**, 2947–2948 (2007).

134. S. Guindon *et al.*, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

135. D. Darriba, G. L. Taboada, R. Doallo, D. Posada, jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods*. **9**, 772–772 (2012).

136. B. Blacklock, S. Adler, A Parasite resembling Plasinodium falciparum in a Chimpanzee. *Ann. Trop. Med. Parasit.* **16**, 99–106 . (1922).

137. C. Newbold, A. Craig, S. Kyes, A. Rowe, Cytoadherence, pathogenesis and the infected red cell surface in Plasmodium falciparum. *Int. J. Parasitol.* (1999).

138. L. Turner *et al.*, Severe malaria is associated with parasite binding to endothelial protein C receptor. *Nature*. **498**, 502–505 (2013).

139. L. Jiang *et al.*, PfSETvs methylation of histone H3K36 represses virulence genes in Plasmodium falciparum. *Nature* (2013).

140. J. D. Smith, J. A. Rowe, M. K. Higgins, T. Lavstsen, Malaria's deadly grip: cytoadhesion of Plasmodium falciparum-infected erythrocytes. *Cell Microbiol*. **15**, 1976–1983 (2013).

141. S. E. R. Bopp *et al.*, Mitotic Evolution of Plasmodium falciparum Shows a Stable Core Genome but Recombination in Antigen Families. *PLOS Genet*. **9**, e1003293 (2013).

142. A. Claessens *et al.*, Generation of Antigenic Diversity in Plasmodium falciparum by Structured Rearrangement of Var Genes During Mitosis. *PLOS Genet*. **10**, e1004812 (2014).

143. S. M. Kraemer, J. D. Smith, Evidence for the importance of genetic structuring to the structural and functional specialization of the Plasmodium falciparum var gene family. *Mol. Microbiol.* (2003), doi:10.1046/j.1365-2958.2003.03814.x/full.

144. T. S. Rask, D. A. Hansen, T. G. Theander, A. Gorm Pedersen, T. Lavstsen,

Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes--divide and conquer. *PLoS Comput. Biol.* **6**, e1000933 (2010).

145. P. C. Bull, C. O. Buckee, S. Kyes, M. M. Kortok, Plasmodium falciparum antigenic variation. Mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks. *Molecular …* (2008), doi:10.1111/j.1365-2958.2008.06248.x/pdf.

146. D. B. Larremore, A. Clauset, C. O. Buckee, A Network Approach to Analyzing Highly Recombinant Malaria Parasite Genes. *PLoS Comput. Biol.* **9**, e1003268 (2013).

147. A. R. Trimnell, S. M. Kraemer, S. Mukherjee, Global genetic diversity and evolution of var genes associated with placental and severe childhood malaria. *Molecular and …* (2006).

148. M. M. Zilversmit *et al.*, Hypervariable antigen genes in malaria have ancient roots. *BMC Evol. Biol.* **13**, 110 (2013).

149. J. Bockhorst, F. Lu, J. H. Janes, J. Keebler, Structural polymorphism and diversifying selection on the pregnancy malaria vaccine candidate VAR2CSA. *Molecular and …* (2007).

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

151. S. A. Sundararaman *et al.*, Selective whole genome amplification yields near full length genomes of chimpanzee *Plasmodium* parasites. *th International Conference of Parasitologists*.

152. P. C. Bull *et al.*, Plasmodium falciparum variant surface antigen expression patterns during malaria. *PLoS Pathog.* **1**, e26 (2005).

153. M. Kaestli, A. Cortes, M. Lagog, M. Ott, Longitudinal assessment of Plasmodium falciparum var gene transcription in naturally infected asymptomatic children in Papua New Guinea. *Journal of Infectious …* (2004).

154. G. M. Warimwe, G. Fegan, J. N. Musyoki, Prognostic indicators of life-threatening malaria are associated with distinct parasite variant antigen profiles. *Science translational …* (2012).

155. M. M. Rorick, T. S. Rask, E. B. Baskerville, K. P. Day, Homology blocks of Plasmodium falciparum var genes and clinically distinct forms of severe malaria in a local population. *Bmc …* (2013).

156. W. Liu, S. A. Sundararaman, B. H. Hahn, (2014).

157. J. C. Rayner, C. S. Huber, J. W. Barnwell, *Conservation and divergence in erythrocyte invasion ligands: Plasmodium reichenowi EBL genes* (Molecular and

biochemical …, 2004).

158. J. R. Miller, S. Koren, G. Sutton, Assembly algorithms for next-generation sequencing data. *Genomics*. **95**, 315–327 (2010).

159. A. Claessens, Y. Adams, A. Ghumra, (2012).

160. P. C. Bull, S. Kyes, C. O. Buckee, J. Montgomery, An approach to classifying sequence tags sampled from Plasmodium falciparum var genes. *Molecular and …* (2007).

161. B. Gamain, J. D. Smith, N. K. Viebig, J. Gysin, Pregnancy-associated malaria: parasite binding, natural immunity and vaccine development. *International journal for …* (2007).

162. A. Ghumra *et al.*, Induction of Strain-Transcending Antibodies Against Group A PfEMP1 Surface Antigens from Virulent Malaria Parasites. *PLoS Pathog.* **8**, e1002665 (2012).

163. G. M. Warimwe, T. M. Keane, G. Fegan, (2009).

164. H. M. Kyriacou, G. N. Stone, R. J. Challis, A. Raza, Differential var gene transcription in Plasmodium falciparum isolates from patients with cerebral malaria compared to hyperparasitaemia. *Molecular and …* (2006).

165. T. Lavstsen, L. Turner, F. Saguti, (2012).

166. J. D. Smith, G. Subramanian, B. Gamain, D. I. Baruch, Classification of adhesive domains in the Plasmodium falciparum erythrocyte membrane protein 1 family. *Molecular and …* (2000).

167. H. M. De Nys *et al.*, Age-related effects on malaria parasite infection in wild chimpanzees. *Biol. Lett.* **9**, 20121160–20121160 (2013).

168. H. M. De Nys, S. Calvignac-Spencer, C. Boesch, Malaria parasite detection increases during pregnancy in wild chimpanzees. *Malaria …* (2014).

169. D. B. Larremore, A. Clauset, A. Z. Jacobs, Efficiently inferring community structure in bipartite networks. *Phys. Rev. E*. **90**, 012805 (2014).

170. S. Merritt, A. Clauset, Environmental structure and competitive scoring advantages in team competitions. *Scientific reports* (2013).

171. M. Gouy, S. Guindon, O. Gascuel, SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* (2010).

172. J. C. C. Hume, E. J. Lyons, K. P. Day, Human migration, mosquitoes and the evolution of Plasmodium falciparum. *Trends Parasitol*. **19**, 144–149 (2003).

173. N. A. Rosenberg, M. Nordborg, Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*. **3**, 380–390 (2002).

174. G. Rizk, D. Lavenier, R. Chikhi, DSK: k-mer counting with very low memory usage. *Bioinformatics*. **29**, btt020–653 (2013).

175. M. Owens, G. Allen, *SQLite* (2010).

176. C. W. Gini, *Variability and Mutability, Contribution to The Study of Statistical Distribution and Relaitons* (Studi Economico-Giuricici della R, 1912).

177. R. M. Kuhn, D. Haussler, W. J. Kent, The UCSC genome browser and associated tools. *Brief. Bioinformatics*. **14**, 144–161 (2013).

178. S. Kryazhimskiy, D. P. Rice, E. R. Jerison, M. M. Desai, Microbial evolution. Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*. **344**, 1519–1522 (2014).

179. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* (2011).

180. R Core Team, "R: A Language and Environment for Statistical Computing" (Vienna, Austria, 2012).

181. L. Délicat-Loembet *et al.*, No evidence for ape Plasmodium infections in humans in Gabon. *PLoS One*. **10**, e0126933 (2015).

182. R. Ross, Notes on the Parasites of Mosquitoes found in India between 1895 and 1899. *Journal of Hygiene*. **6**, 101–108 (1906).

183. R. Ross, *The logical basis of the sanitary policy of mosquito reduction* (1905).

184. J. Hamon, J. P. Adam, A. Grjébine, [Observations on the distribution and behavior of anophelines in French Equatorial Africa, the Cameroons and West Africa]. *Bull. World Health Organ.* **15**, 549–591 (1956).

185. M. Coluzzi, Heterogeneities of the malaria vectorial system in tropical Africa and their significance in malaria epidemiology and control. *Bull. World Health Organ.* **62 Suppl**, 107–113 (1984).

186. D. Fontenille, L. Lochouarn, The complexity of the malaria vectorial system in Africa. *Parassitologia*. **41**, 267–271 (1999).

187. J. B. P. Lima, M. G. Rosa-Freitas, C. M. Rodovalho, F. Santos, R. Lourenço-de-Oliveira, Is there an efficient trap or collection method for sampling Anopheles darlingi and other malaria vectors that can describe the essential parameters affecting transmission dynamics as effectively as human landing catches? - A Review. *Mem Inst Oswaldo Cruz*. **109**, 685–705 (2014).

188. H. M. Abkallo *et al.*, DNA from pre-erythrocytic stage malaria parasites is detectable by PCR in the faeces and blood of hosts. *Int. J. Parasitol.* **44**, 467–473 (2014).

189. P. D. Crompton *et al.*, A prospective analysis of the Ab response to Plasmodium falciparum before and after a malaria season by protein microarray. *Proc. Natl. Acad. Sci. USA*. **107**, 6958–6963 (2010).

190. J. H. McDonald, M. Kreitman, Adaptive protein evolution at the Adh locus in Drosophila. *Nature* (1991).

191. T. Endo, K. Ikeo, T. Gojobori, Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* **13**, 685–690 (1996).

192. A. A. Escalante, A. A. Lal, F. J. Ayala, Genetic polymorphism and natural selection in the malaria parasite Plasmodium falciparum. *Genetics*. **149**, 189–202 (1998).

193. J. Parsch, Z. Zhang, J. F. Baines, The influence of demography and weak selection on the McDonald–Kreitman test: an empirical study in Drosophila. *Mol. Biol. Evol.* (2009).

194. J. Smith *et al.*, Analysis of adhesive domains from the A4VAR Plasmodium falciparum erythrocyte membrane protein-1 identifies a CD36 binding domain. *Mol. Biochem. Parasitol.* **97**, 133–148 (1998).

195. M. M. Klein *et al.*, The Cysteine-Rich Interdomain Region from the Highly Variable Plasmodium falciparum Erythrocyte Membrane Protein-1 Exhibits a Conserved Structure. *PLoS Pathog.* **4**, e1000147 (2008).

196. J. Eid *et al.*, Real-time DNA sequencing from single polymerase molecules. *Science*. **323**, 133–138 (2009).

197. S. Goodwin *et al.*, Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* (2015), doi:10.1101/gr.191395.115.