Using whole genome sequence data to study genomic diversity and develop molecular barcodes to profile *Plasmodium* malaria parasites

Ernest Díez Benavente

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of London

August 2018

Department of Pathogen Molecular Biology

Faculty of Infectious and Tropical Diseases

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

No funding received.

Research group affiliation(s): Professor Taane G Clark and Associate Professor Susana Campino

I, Ernest Diez Benavente, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in this thesis.

Signed _________________ Date _________________

Abstract

Malaria is a major threat to human health, causing over 300 million clinical cases and approximately ~500,000 deaths per year. Countries attempting malaria elimination are increasingly concerned with identifying pockets of transmission and outbreaks arising from imported cases, and there is a need to establish molecular barcodes for implementation in the field. The genetic diversity and non-recombining properties of mitochondrial and apicoplast sequence can be powerfully exploited for geographic genetic profiling of *P. falciparum* malaria at an inter-continent level. However, this approach provides limited insights for assessing drug resistance, intra-regional geographical differentiation, and ignores malaria caused by other *Plasmodium spp*. (*P. vivax and P. knowlesi*). To overcome these limitations, this project proposes to study the genomic diversity found in the nuclear and organellar genomes of the Plasmodium species causing human malaria and establish robust ways to create SNP barcodes. In this study, an assessment of the current libraries of genomic sequence data across the species *P. falciparum*, *P. vivax* and *P. knowlesi* was performed and using a range of bioinformatics approaches the genetic diversity in the different populations was assessed. For this, a new high-quality reference for the A1-H.1 *P. knowlesi* strain was generated and its methylome was characterized. Using this reference, the first evidence of genetic exchange events between the three subpopulations of *P. knowlesi* was found in Malaysia. Furthermore, a study of the structural and genetic diversity found in the hypervariable vaccine candidate *var2csa* gene in *P. falciparum* and its potential geographical signal associated with Malaria in Pregnancy (MiP) were performed. Finally, we accomplished a genetic diversity study of global *P. vivax* isolates and the insights obtained from this analysis allowed the development of a 71 SNP barcode to predict the geographical origin of P. vivax

isolates. The obtained barcode were tested using prospectively and retrospectively collected datasets, particularly from endemic settings with complex mixed infections and near-elimination settings. The identification of SNP barcodes using this methodology can inform future rapid diagnostics and promote the application of field-based sequencing.

Acknowledgements

I would like to thank my supervisors Taane Clark and Susana Campino for the opportunity to work on a PhD and for all the support along the way; a big thank you to my advisors Colin Sutherland and Cally Roper for the support and the interesting discussions.

Big thanks go to members of the "hub": Jody Phelan, Matt Ravenhall, Yaa Oppong, Ben Sobkowiak, Amy Ibrahim, Pepita Gomez Gonzalez, Matt Higgins, Dan Ward, James Charleston "the brief" and Neneh Sallah.

Many thanks to colleagues who have helped me during my PhD in many different ways: Francesc Coll, Sammy Assefa, Rob Moon, Martin Hibberd, Chris Drakeley, Arnab Pain, David Baker, Kevin Tetteh, Julius Hafalla.

To all the friends supporting me and making my time in LSHTM as great as it gets: Tom Hall, Stephanie Nofal, Inke Lubis, Lotus van den Hoogen, Will Stone, Tate Oulton, David McLeod, Jackie Cook, Sonal Shah, Lindsey Wu, Bronner Gonçalves, Nuno Sepulveda, Felipe Cia and many others.

A mi familia por todo: Mama, Papa, Alvar, Iris.

Fernanda Costa for clearing my thinking and being there when I needed her.

Additional Publications

I contributed to other manuscripts which were not part of my PhD:

(1) ***Benavente, E.D.**, F. Coll, N. Furnham, R. McNerney, J.R. Glynn, S. Campino, A. Pain, F.R. Mohareb, and T.G. Clark. 2015. "PhyTB: Phylogenetic Tree Visualisation and Sample Positioning for M. Tuberculosis." BMC Bioinformatics 16 (1). doi:10.1186/s12859-015-0603-3.*

(2) *Ravenhall, M., **E.D. Benavente**, M. Mipando, A.T.R. Jensen, C.J. Sutherland, C. Roper, N. Sepúlveda, et al. 2016. "Characterizing the Impact of Sustained Sulfadoxine/Pyrimethamine Use upon the Plasmodium Falciparum Population in Malawi." Malaria Journal 15 (1). doi:10.1186/s12936-016-1634-6.*

(3) *Ansari, H.R., T.J. Templeton, A.K. Subudhi, A. Ramaprasad, J. Tang, F. Lu, R. Naeem, , Y. Hashish, M.C. Oguike, **E.D. Benavente**, et al. 2016. "Genome-Scale Comparison of Expanded Gene Families in Plasmodium Ovale Wallikeri and Plasmodium Ovale Curtisi with Plasmodium Malariae and with Other Plasmodium Species." International Journal for Parasitology 46 (11). doi:10.1016/j.ijpara.2016.05.009.*

(4) *Campino, S., **E.D. Benavente**, S. Assefa, E. Thompson, L.G. Drought, C.J. Taylor, Z. Gorvett, et al. 2016. "Genomic Variation in Two Gametocyte Non-Producing Plasmodium Falciparum Clonal Lines." Malaria Journal 15 (1). doi:10.1186/s12936-016-1254-1.*

(5) *Walk, J., I.J. Reuling, M.C. Behet, L. Meerstein-Kessel, W. Graumans, G. van Gemert, R. Siebelink-Stoter, M. van de Vegte-Bolmer, T. Janssen, K. Teelen, J.H.W. de Wilt, Q. de Mast, A.J. van der Ven, **E.D. Benavente** et al. 2017. "Modest Heterologous Protection after Plasmodium Falciparum Sporozoite Immunization: A Double-Blind Randomized Controlled Clinical Trial." BMC Medicine 15 (1): 168. doi:10.1186/s12916-017-0923-4.*

(6) *Gomes, A.R., M. Ravenhall, **E.D. Benavente**, A. Talman, C. Sutherland, C. Roper, T.G. Clark, and S. Campino. 2017. "Corrigendum to 'Genetic Diversity of next Generation Antimalarial Targets: A Baseline for Drug Resistance Surveillance Programmes' (International Journal for Parasitology: Drugs and Drug Resistance (2017) 7(2) (174–180) (S2211320717300131) (10.1016/j.Ijpd." International Journal for Parasitology: Drugs and Drug Resistance 7 (3). doi:10.1016/j.ijpddr.2017.11.001.*

(7) *Trimarsanto, H., **E.D. Benavente**, R. Noviyanti, R.A.S. Utami, L. Trianty, Z. Pava, S. Getachew, et al. 2017. "VivaxGEN: An Open Access Platform for Comparative Analysis of Short Tandem Repeat Genotyping Data in Plasmodium Vivax Populations." PLoS Neglected Tropical Diseases 11 (3). doi:10.1371/journal.pntd.0005465.*

(8) *Leon, L.J., R. Doyle**, E.D. Benavente**, T.G. Clark, N. Klein, P. Stanier, and G.E. Moore. 2018. "Enrichment of Clinically Relevant Organisms in Spontaneous Preterm Delivered Placenta and Reagent Contamination across All Clinical Groups in a Large UK Pregnancy Cohort." Applied and Environmental Microbiology, May. United States. doi:10.1128/AEM.00483-18.*

(9) *Sepulveda, N., J. Phelan, **E.D. Benavente**, S. Campino, T.G. Clark, H. Hopkins, C. Sutherland, C.J. Drakeley, and K.B. Beshir. 2018. "Global Analysis of Plasmodium Falciparum Histidine-Rich Protein-2 (Pfhrp2) and Pfhrp3 Gene Deletions Using Whole-Genome Sequencing Data and Meta-Analysis." Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases 62 (May). Netherlands: 211–19. doi:10.1016/j.meegid.2018.04.039.*

(10) *Cook, S., S. Malyutina, A.V. Kudryavtsev, M. Averina, N. Bobrova, S. Boytsov, S. Brage, T. G. Clark, **E.D. Benavente**, et al. 2018. "Know Your Heart: Rationale, Design and Conduct of a Cross-Sectional Study of Cardiovascular Structure, Function and Risk*

Factors in 4500 Men and Women Aged 35-69 Years from Two Russian Cities, 2015-18 [Version 1; Referees: 1 Approved, 1 Approved with Reservations].” Wellcome Open Research 3 (67). doi:10.12688/wellcomeopenres.14619.1.

(11) *Auburn, S.,* ***E.D. Benavente****, O. Miotto, R.D. Pearson, R. Amato, M.J. Grigg, B.E. Barber, et al. 2018. “Genomic Analysis of a Pre-Elimination Malaysian Plasmodium Vivax Population Reveals Selective Pressures and Changing Transmission Dynamics.” Nature Communications 9 (1). doi:10.1038/s41467-018-04965-4.*

Table of Contents

List of Abbreviations

ACT Artemisinin-based Combination Therapy

DBL Duffy Binding Like

DBP Duffy Binding Protein

CDC Centre for Disease Control

GTS Global Technical Strategy for Malaria 2016–2030

InDel Insertion or Deletion

kbp kilo-base pairs

M Macaca

Mf-Pk Macaca fascicularis Plasmodium knowlesi genotype

Mn-Pk Macaca nemestrina Plasmodium knowlesi genotype

NBP Normocyte Binding Protein

P Plasmodium

RDT Rapid Diagnostic Test

RBC(s) Red Blood Cell(s)

RBP Reticulocyte Binding Protein

SEA Southeast Asia

SNPs Single Nucleotide Polymorphisms

sp. species

ssp. sub-species

SP Sulfadoxine-Pyrimethamine

SWGA Selective Whole Genome Amplification

WGS Whole Genome Sequencing

WHO World Health Organization

Chapter 1
Introduction

1. Malaria in the global context

The malaria parasites

Malaria is a protozoan parasitic mosquito-borne disease that can affect humans and other animals. The disease is caused by infections with *Plasmodium* genera parasites. There are currently 5 species of *Plasmodium* that can naturally cause human infections of malaria: *P. falciparum*, *P. vivax*, *P. knowlesi*, *P. malariae*, *P. ovale* (*ssp. curtisi* and *ssp. wallikeri*). Infections with *P. falciparum* and *P. vivax*, cause the greatest disease burden [1].

The *Plasmodium* parasite has a complex life cycle that involves several stages split into two differentiated host- and vector-driven larger subdivisions (see Figure 1). The *P. falciparum* host life cycle starts with an infective bite of a female *Anopheles* mosquito to a human host. During the mosquito blood meal, the parasite enters the bloodstream as a sporozoite and infects the liver cells where it develops into a schizont, which contains merozoites. The merozoites are released into the bloodstream and recognize a range of receptors in the red blood cells to invade and develop into a new schizont, which ruptures and starts the erythrocytic cycle again. The disease symptoms appear at this blood stage and range from fever, tiredness, nausea, and headaches to more severe symptoms such as severe malarial anaemia, cerebral malaria, respiratory distress and death. The symptoms usually develop between 10 to 15 days after the bite of an infected female mosquito [1]. During the erythrocytic cycle, some of the parasites develop into the sexual forms, either male or female gametocytes. The mosquito takes the gametocytes during a blood meal starting the life cycle in the vector. The female and male gametocytes generate an ookinete that matures in the gut into an oocyst full of sporozoites, which migrate into the salivary glands from where it will infect the next host when taking the blood meal [2].

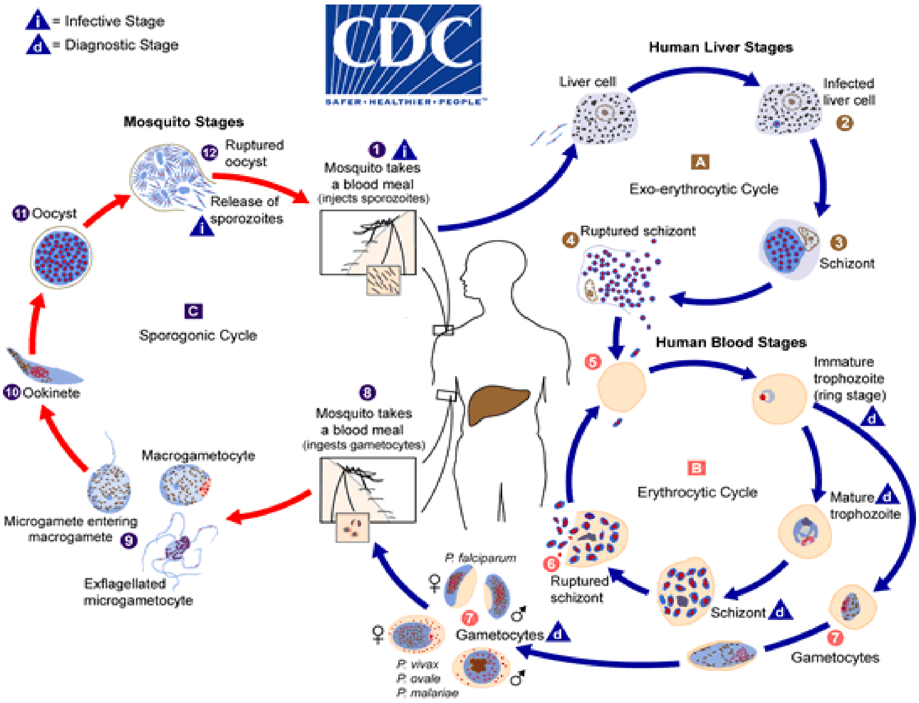


Figure 1. Malaria life cycle (https://www.cdc.gov/malaria/about/biology/)

Burden of disease

It is estimated that nearly half of the world's population was at risk of malaria in 2016 [1], with 91 countries considered endemic **(Figure 2),** 212 million estimated cases and 429,000 estimated deaths per year. The malaria global burden is unevenly distributed geographically, with the sub-Saharan Africa region accounting for the majority of the cases (90%) and deaths (92%), with a high mortality burden in children under 5 years of age (70%). Several regions in Southeast Asia (SEA), South America and the Middle East are also at high risk [1]**,** with 7% of malaria cases occurring in SEA.

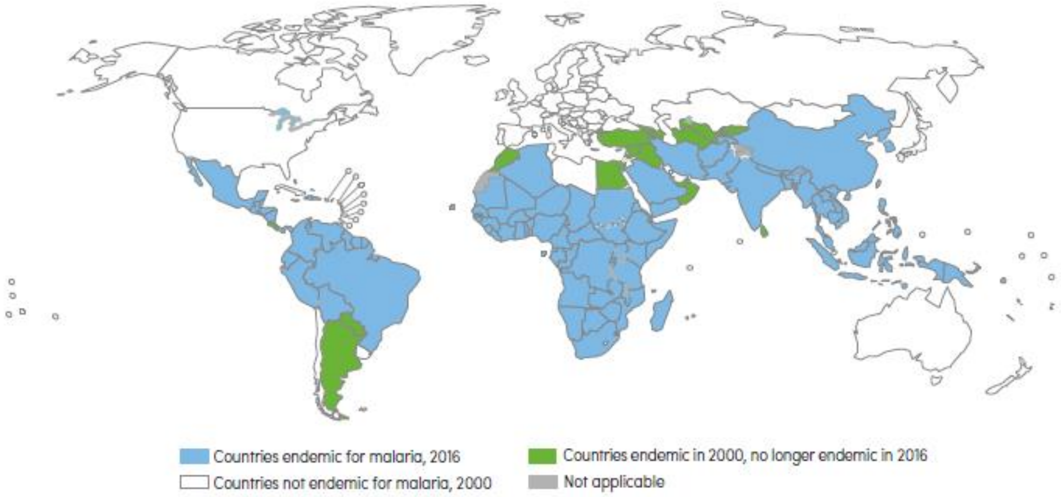

Countries endemic for malaria, 2016
Countries not endemic for malaria, 2000
Countries endemic in 2000, no longer endemic in 2016
Not applicable

Figure 2. Malaria-endemic countries in 2000 and 2016. Countries with 3 consecutive years with no cases are considered to have eliminated malaria [1].

Almost all deaths (99%) are caused by *P. falciparum* parasites, with the remainder due to *P. vivax* (3,100 deaths; 84% outside Africa) [1]**.** Similarly, *P. falciparum* causes the vast majority of cases worldwide, followed by *P. vivax* (4% of the total). Nevertheless, excluding the African continent, *P. vivax* is responsible for 41% of the cases [1]**,** which constitutes an important challenge for the rest of endemic regions.

P. knowlesi malaria is a newly emergent zoonotic disease, where the parasite primarily infects macaques *Macaca fascicularis* (*Mf*, long-tailed) an*d M. nemestrina (Mn, pig-tailed).* The geographic distributions of the primary macaque hosts and *Anopheles* vectors of this species are localized in SEA, leading to human populations at risk in that region [3], resulting in an increasing number of human cases been reported in Borneo Malaysia since 2004 [4].

The factors driving the geographical distribution of the burden of the disease include climatic conditions, such as high temperatures and humidity, which determine the distribution of the *Anopheles* vector together with the host-related factors that either prevent or facilitate the transmission of the disease. Human-related disease can be driven by socioeconomic (e.g. poverty, lack of access to health care) or biological factors (e.g. protective genetic characteristics such as sickle cell mutations or the absence of Duffy-binding receptors in the red-blood cells) amongst others [5]**.**

Despite the still high number of cases, malaria incidence has reduced by 40% between 2000 and 2015 (see **Figure 3** for Africa), due mainly to the roll-out of insecticide-treated mosquito nets, the use of new antimalarial drugs such as the Artemisinin-combination therapies, and the use of indoors residual spraying [1]. The target set by the *Global Technical Strategy for Malaria 2016–2030* is to reduce the mortality and incidence rates by 90% in 2030 compared to 2015 and eliminate malaria in at least 35 countries [1]**.** One of the main challenges to achieve these targets is the presence of antimalarial drug resistance across several populations of the different *Plasmodium* species. Drug resistance occurs when a population of parasites, under the pressure of a drug, undergoes rapid selection for variants in their genome that confer some advantage in the survival or transmission of the disease in the presence of the drug [6]**.** Some of the most commonly used drugs have seen their effectiveness challenged by such resistance in two of the main malaria causing species, *P. falciparum* and *P. vivax* [7].

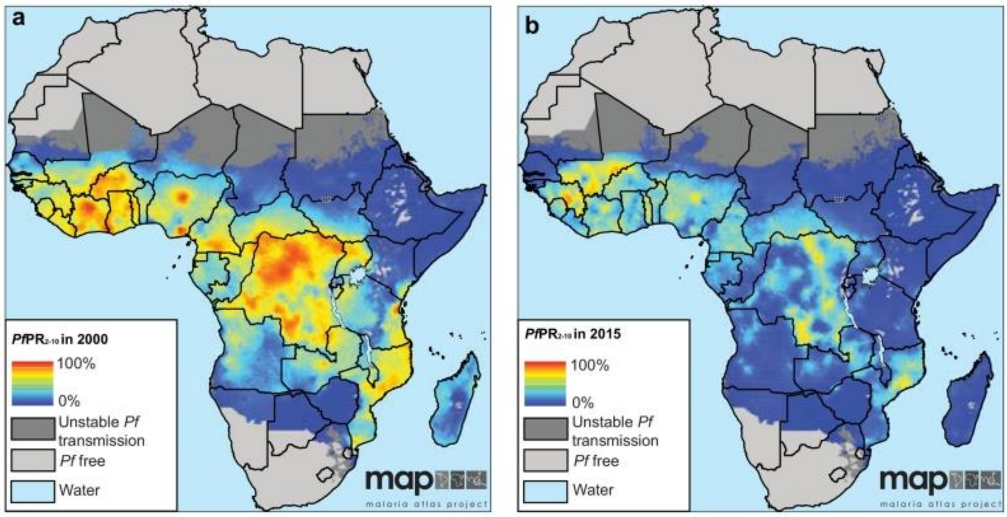


Figure 3. Change in infection prevalence between 2000 and 2015 in the African region (PfPR, Plasmodium falciparum prevalence) , figure taken from [8].

Drug resistance

P. falciparum is highly resistant to chloroquine in almost all the endemic malaria regions spreading from SEA regions in the late 1950s [1]. Resistance to sulfadoxine-pyrimethamine (SP), which was introduced in the early 1940s as an antimalarial, appeared as early as 10 years after its introduction and developed in a surprisingly rapid manner leading to current widespread SP ineffectiveness [9]. In the case of mefloquine, resistance has been reported since the mid-1980s in the areas where it has been widely used, mainly in Cambodia, Thailand and Vietnam [7]. This has pushed the Artemisinin-based Combination Therapies (ACTs) to be used as the front-line drug for treatment of *P. falciparum* malaria. However, artemisinin resistance in the form of the slow clearance of parasites has been reported in western Cambodia first and has since either spread, emerged independently or both to other regions in SEA [10]**.**

Regarding *P. vivax,* drug resistance is not well studied nor understood; resistance to chloroquine has been reported in Indonesia and Papua New Guinea as well as closely located regions [8]. There is also a larger spread of SP resistance in *P. vivax* [11–16]***.***

In order to help tackle these challenges, one of the approaches taken by the scientific community has been to study and characterize the genome of the *Plasmodium* parasites causing the disease. An approach in which the parasites are isolated from the blood collected directly from patients infected with malaria, DNA is extracted and sequenced using high throughput sequencing technology, such as the Illumina HiSeq or MiSeq devices [17]. Understanding the complexity and variability of the *Plasmodium* genomes can provide insights into the biology underlying the disease and inform the processes of drug and vaccine development [18]**,** help guide control interventions[19] and understand the dynamics and mechanisms of drug resistance in the parasites [20]**.**

Plasmodium genome

A number of *Plasmodium* reference genomes have been published: *P. falciparum* (3D7 [21], in 2002), *P. vivax* (sal1, [22]**,** 2008), *P. knowlesi* (H strain, [23]**,** 2008), and P. *malariae, P. ovale ssp. curtisi and P. ovale ssp wallikeri i*n 2016 [24]**.** These genomes vary in size and GC content (see **Table 1**), and consist of nuclear (>23Mb) and organellar (mitochondria ~6kb, apicoplast ~35kb) genomes. Whilst nuclear genomes undergo recombination, organellar genomes are comparatively highly stable and conserved over time, and are uniparentally co-inherited from one of the gametocytes.

Table 1. Comparison of the human malaria Plasmodium species

Plasmodium species	*Main host*	*Febrile Episodes (h)*	*Life cycle peculiarities*	*Genome*	*Genome length (Mb)*	*GC content (%)*
P. falciparum	Human	~48	No dormant stage	3D7 [21]	23.3	19.4
P. vivax	Human	48	Liver dormant stage (hypnozoite) causes relapses	Salvador-I [22]	26.8	42.3
P. knowlesi	Macaque (*Macaca spp.*)	24	Zoonotic Disease	H Strain [23]	23.5	37.5
P. ovale ssp. curtisi	Human	48	Relapse can happen up to 4 years after infection	Poc1 (Nigeria I) [24]	34.5	28.5
P. ovale ssp. wallikeri	Human	48		Pow1 (Gabon) [24]	35.2	28.9
P. malariae	Human	72		Pm (Uganda I) [24]	31.9	24.7

Additional “short” sequences or “read” data are being increasingly generated by new high throughput sequencing technologies [25], this data characteristically covers the genome totally to high depth. As a result large whole genome sequencing data across *Plasmodium spp.* is being placed into the public domain (see **Table 2**). The study of data sourced from endemic field isolates has provided great insight into the structure and distribution of the parasite populations and intra- and inter-population genomic diversity [17,26–29]**.**

However, the use of WGS has been discarded as a ubiquitous solution that can suit all the possible scenarios given its high cost and the challenge of obtaining the amount and quality of parasite DNA required. Several methods have been used to study and characterize the genomic diversity of parasite isolates collected in the field at a lower cost, such as the use of microsatellites, a series of tandem repeated sequences which length varies across different isolates [30,31]**.**

Another approach adopted more recently, is to analyze the previously collected WGS data to generate single nucleotide polymorphisms (SNP) barcodes that can easily characterize and classify geographic subpopulations of *Plasmodium* isolates within the same species [32,33], including the use of robust organellar barcodes [34]**.**

Table 2. Available data and prospective data generated

Plasmodium species	*Illumina Data available*	*PacBio long read available data*	*Illumina Data Analyzed in this thesis*	*PacBio Data Analyzed in this thesis*	*Data being generated*
P. falciparum	3407 isolates	21 lines	3407 isolates	21 lines	-
P. vivax	697 isolates	-	697 isolates	-	~30 isolates
P. knowlesi	81 isolates	2 lines	81 isolates	2 lines	~20 isolates
P. ovale ssp. curtisi	4 isolates	-	4 isolates	-	-
P. ovale ssp. wallikeri	3 isolates	-	3 isolates	-	-
P. malariae	1 isolate	1 isolate	1 isolate	-	-

2. *Plasmodium falciparum*

Given the disproportionately high disease burden caused by *P. falciparum* (*Pf*)*,* it is the most studied species, including from a genetic point of view. The parasite has a genome of 23.3 Mb in length arranged in 14 chromosomes, 1 mitochondrion genome, present in approximately 30 copies per parasite and 2 copies of an apicoplast genome (**Table 1)** [21]**.**

The *P. falciparum* genome has a low GC content (19.4%), which makes the sequencing of this parasite a challenge using the available technologies [21]. Of special relevance is the *var* gene family, a family of hypervariable genes that are alternatively expressed and presented in the surface of the infected Red Blood Cells (RBCs). These proteins help cytoadherence of the parasite and have a main role in helping to overcome the host immune system [35]**.** One of the genes in this family, the *var2csa* gene, has been associated with placental malaria and is thought to be the responsible for the parasite binding to the receptor chondroitin sulfate A in the placenta [36]**.**

3. *Plasmodium vivax*

Plasmodium vivax is the second most virulent malaria-causing parasite. Its geographic distribution includes Asia, Central and South America, the Middle East, Oceania and some parts of Africa (e.g. Ethiopia), being more than 2.8 billion people are at risk of infection [2]**.** Given its low mortality rates and the current inability to culture the parasite *in vitro,* it is considered a neglected tropical disease compared *to P. falciparum* [1].

Some of the control interventions aimed at reducing *P. falciparum* incidence rates are not as effective when applied to *P. vivax* given its very particular biology, including the hypnozoite, a dormant stage that can persist for a long time in the human liver [37] **(Table 1)**. This stage of the parasite can then relapse after a period of time that ranges from several months to years in some cases [38]**.** Thus, *P. vivax* has become the most prevalent malaria parasite in some of the countries where *P. falciparum* control measures have been rolled out and its presence and transmission has been effectively reduced [39,40]. Control and future elimination of the *P. vivax* malaria requires additional efforts to effectively combat the dormant liver stage. Genomic research is an approach with the potential to contribute greatly to our understanding of the basic biology and evolution trends of *P. vivax*; thus, supporting the future interventions during development and surveillance stages.

The sequencing and characterization of the *P. vivax* genome (Sal-1 [22]) primed population genetic studies, the first of them based on microsatellite data [41–43]. The development of appropriate DNA isolation and amplification of clinical samples involving high parasitaemias has led to whole genome sequencing. These studies have shed light into the high degree of polymorphism of this parasite compared to *P. falciparum* [27,28,44,45]. Even on a small spatial scale, *P. vivax* populations present high genomic diversity, this variability may be due to host genomics, vector species and environmental factors [28,44,45].

The advances in WGS technologies and ability to multiplex many samples has opened up the opportunity to obtain a data-driven picture of the genomic epidemiology and genetic diversity of *P. vivax* populations. This includes identifying structural variation that define *P. vivax* populations, which can assist the development of appropriate geographical barcodes for this species, as previously implemented for *P. falciparum* [33,46].

4. *Plasmodium knowlesi*

This parasite species is now recognized as a substantial cause of malaria in humans, and infections are now known to be widespread across SEA [47] including cases of travellers from outside the region [48]. The clinical symptoms of disease range from asymptomatic carriage to high levels of parasitaemia, which in severe cases, can lead to death [49,50]. The deforestation and encroachment on the wild macaque habitats as a consequence of human population growth can increase the chances of human-macaque contact [51].

The analysis of WGS data from *P. knowlesi* isolates from human infections in Sarikei in Malaysia Borneo demonstrated existing dimorphism over at least 50% of the *P. knowlesi* genome [52]. Microsatellite diversity in parasites from both macaque and human infections across Borneo and Peninsular Malaysia supported this observation, and found a strong association of the two distinct genome dimorphs with adaptations to either of the two *Macaca* host, but with no evidence of a complete primate host susceptibility barrier [53]. Most recently, a new geographical genetic cluster has been identified from Peninsular Malaysia [29,54–56]. Genomic diversity in *P. knowlesi* is likely to be driven by host- and geography-related factors, as well as the contemporary effect of the human alteration in host and vector distributions during the ecological transition that is taking place in the region [57].

Several *Anopheles* species from the Leuchosphyrus group are able to transmit *P. knowlesi* malaria. Some examples are *A. latens* and *A. balbacensis* in Borneo, Malaysia [58–60], *A. cracens* and *A. hackeri* in the Peninsula in Malaysia [61] and *A. dirus* in Vietnam [62].

It is required to undertake in-depth studies that help elucidate the impact of such diverse host and vector patterns in the evolution of the *P. knowlesi* genome. These studies will help understand how the different geographic and host-related subpopulations of this parasite are structured. Such knowledge would be crucial in order to design suitable SNP barcodes for this highly variable species.

Furthermore, there is a need to further investigate the structure and characteristics of complex gene families such as the *SICAvar* genes, which are expressed in the surface of the infected RBCs and are responsible for the antigenic variation and immune evasion in the host, a highly similar function is carried out by the *var* genes in the *P. falciparum* parasites.

5. Other Plasmodium species

There are two other main species causing human malaria, *P. ovale* and *P. malariae*. These species have been less studied given their low prevalence and mild disease symptoms, although there have been reports of rare severe cases [63]. The two subspecies of *P. ovale* are morphologically indistinguishable, but analysis of their genomic sequences revealed a dimorphism between the two [64]. They show a sympatric distribution throughout the tropics in Africa, Asia and Oceania although its prevalence is thought to be highly underestimated [63]. This species shares with *P. vivax* the capability to form hypnozoites, a dormant liver stage that can relapse up to 3 years after infection [65].

P. malariae is a *Plasmodium* causing quartian human malaria and is associated with the production of immune complexes located in the patient kidneys, which can develop into a nephrotic syndrome [66].

The use of WGS to study these parasites is still embryonic, but the sequencing of the genomes for these two species [24] has opened up the opportunity to inform the biology of their populations. Their genomes have shown the presence of expanded *pir* and *surfin* gene families, which increases the length of their genomes up to a size of ~35 Mb [24]**.** The most recent sequencing study considering co-infections of these parasites with other *Plasmodium* species, sheds light on the evolutionary relationships of species within the *Plasmodium* lineage, estimating the divergence of the 2 *P. ovale* subspecies at 20.3 million years ago [67]**.** These species are normally found in mixed infections together with other *Plasmodium* species and therefore it is of great interest to identify highly sensitive genetic markers to detect the presence of these parasites, and therefore their actual prevalence in prospective studies.

6. The need for a barcode of *Plasmodium* and attempts to date

The use of SNP barcodes for *Plasmodium* parasites has been made possible by advances in whole genome sequencing. The first attempt in *P. falciparum* used 24 nuclear SNPs to identify and track the isolates of a population [46]. The underlying SNPS within this barcode were derived from the genome sequences of long-term adapted laboratory lines [46]. The 24-SNP barcode was then applied in an endemic region in Senegal, analysing samples over a 7-year period [68] to infer transmission intensity**.** This work demonstrated the potential that genomic barcodes in combination with epidemiological methods have to elucidate transmission intensity [68]**.** Nevertheless, the use of isolates from a very limited region to generate the initial barcode, can potentially underestimate the genomic variability present in the overall population. A recent study showed that the prediction power of the 24 SNP barcode for geographical determination was poor compared to one formed of 23 SNPs in the mitochondria and apicoplast organellar genomes, which predicted the continental origin of samples with 92% accuracy [33].

A recently published barcode formed of 105 highly frequent SNPs was developed to estimate the complexity of infection of samples and therefore, like the 24 SNP barcode, infer transmission intensity [69]**.** However, the markers were selected using a much broader panel of data, thereby proving greater utility across malaria-endemic countries [69].

These studies demonstrated how barcodes can potentially provide insight into the intensity of transmission, identify the geographical origin of the field isolates, and inform dynamics of the diversity in a parasite population. They could also identify potential imported outbreaks as well as mixed infections involving *Plasmodium spp*. Other efforts have attempted to generate similar barcodes for species such as *P. vivax* [32,70] but they were also based on limited datasets, and therefore can have poor predictive power when used for more geographically distinct datasets.

The current diagnostic for the different *Plasmodium sp.* relies on microscopy, which although cheap and easy to implement, is prone to misdiagnosis. A routine PCR approach used in reference laboratories, which is based on the amplification of the small subunit ribosomal RNA (ssrRNA), can identify the presence/non-presence of the *Plasmodium* species [72]. However, the creation of SNP-based barcodes could potentially make this identification cheaper and more informative.

As technology advances, the implementation of high throughput sequencing of specific regions of the genome for high number of samples is becoming cheaper and more rapid to implement, by using pooled PCR assays which can be multiplexed in the Illumina sequencing machine using sequence barcodes for each sample [73]**.** The identification of informative regions in the Plasmodium genomes gains relevance under this scenario. Therefore, in combination with the use of new technologies such as the Oxford Nanopore MiniOn, could potentially make the use of barcodes to predict origin as well as infer transmission patterns possible in real time [74]**.**

Data available

In my work, I sought to understand the genetic diversity across *Plasmodium* species and construct intra-species and regional molecular barcodes that could be implemented in the field. The majority of data available **(Table 2**) has come from Illumina sequencing technologies, including MiSeq or the HiSeq platforms [75]**.** I use raw sequence data that has been previously published and publicly available [17,27–29,52], and is complemented by in-house data generated through our collaborations.

To tackle these questions, in **Chapter 2** I generated a new high-quality reference for the A1-H.1 *P. knowlesi* strain and studied epigenetic methylation patterns. Using this new reference, in **Chapters 3** and **4** we provided evidence for the first time of both recent and non-recent genetic introgression events occurring between the three subpopulations of *P. knowlesi* found in both Borneo and Peninsular Malaysia. In **Chapter 5**, I attempted to use the sequencing diversity found in the hypervariable vaccine candidate *var2csa* gene in *P. falciparum* in order to identify geography specific signatures, while performing a structural and genetic diversity analysis of the gene. Finally, in **Chapters 6** and **7** I performed a genetic diversity study of global *P. vivax* isolates from both South East Asia and South America where prevalence of *P. vivax* is high, I then used the insights obtained from the analysis and complementary isolate data to develop a 71 SNP barcode to predict the geographical origin of *P. vivax* isolates. This barcode was assessed using prospectively and retrospectively collected datasets, particularly from endemic settings with complex mixed infections and near-elimination settings in **Chapter 7**.

REFERENCES

1. WHO. *World Malaria Report 2017*. (2017).

2. Gazzinelli, R. T., Kalantari, P., Fitzgerald, K. A. & Golenbock, D. T. Innate sensing of malaria parasites. *Nat Rev Immunol* **14,** 744–757 (2014).

3. Moyes, C. L. *et al.* Predicting the geographical distributions of the macaque hosts and mosquito vectors of Plasmodium knowlesi malaria in forested and non-forested areas. *Parasit. Vectors* **9,** 242 (2016).

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

5. Hay, S. I., Guerra, C. A., Tatem, A. J., Noor, A. M. & Snow, R. W. The global distribution and population at risk of malaria: past, present, and future. *Lancet Infect. Dis.* **4,** 327–336 (2004).

6. Shah, N. K. *et al.* Antimalarial drug resistance of Plasmodium falciparum in India: changes over time and space. *Lancet Infect. Dis.* **11,** 57–64 (2011).

7. White, N. J. Antimalarial drug resistance. *J. Clin. Invest.* **113,** 1084–1092 (2004).

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

9. Ravenhall, M. *et al.* Characterizing the impact of sustained sulfadoxine/pyrimethamine use upon the Plasmodium falciparum population in Malawi. *Malar. J.* **15,** (2016).

10. Ashley, E. A. *et al.* Spread of Artemisinin Resistance in Plasmodium falciparum Malaria. *N. Engl. J. Med.* **371,** 411–423 (2014).

11. Asih, P. B. *et al.* Distribution of Plasmodium vivax pvdhfr and pvdhps alleles and their

association with sulfadoxine–pyrimethamine treatment outcomes in Indonesia. *Malar J.* **14,** (2015).

12. Hastings, M. D. *et al.* Dihydrofolate reductase mutations in Plasmodium vivax from Indonesia and therapeutic response to sulfadoxine plus pyrimethamine. *J Infect Dis* **189,** (2004).

13. Tjitra, E. *et al.* Multidrug-Resistant Plasmodium vivax Associated with Severe and Fatal Malaria: A Prospective Study in Papua, Indonesia. *PLOS Med.* **5,** e128 (2008).

14. Imwong, M. *et al.* Association of genetic mutations in Plasmodium vivax dhfr with resistance to sulfadoxine–pyrimethamine: geographical and clinical correlates. *Antimicrob Agents Chemother* **45,** (2001).

15. Thongdee, P. *et al.* Genetic polymorphisms in Plasmodium vivax dihydrofolate reductase and dihydropteroate synthase in isolates from the Philippines, Bangladesh, and Nepal. *Korean J Parasitol* **53,** (2015).

16. Korsinczky, M. *et al.* Sulfadoxine resistance in Plasmodium vivax is associated with a specific amino acid in dihydropteroate synthase at the putative sulfadoxine-binding site. *Antimicrob. Agents Chemother.* **48,** 2214–2222 (2004).

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

18. Conway, D. J. Paths to a malaria vaccine illuminated by parasite genomics. *Trends Genet.* **31,** 97–107 (2015).

19. Gunawardena, S. & Karunaweera, N. D. Advances in genetics and genomics: use and limitations in achieving malaria elimination goals. *Pathog. Glob. Health* **109,** 123–141 (2015).

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

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

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

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

24. Ansari, H. R. *et al.* Genome-scale comparison of expanded gene families in Plasmodium ovale wallikeri and Plasmodium ovale curtisi with Plasmodium malariae and with other Plasmodium species. *Int. J. Parasitol.* **46,** (2016).

25. Le Roch, K. G., Chung, D.-W. D. & Ponts, N. Genomics and Integrated Systems Biology in Plasmodium falciparum: A Path to Malaria Control and Eradication. *Parasite Immunol.* **34,** 50–60 (2012).

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

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

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

29. Assefa, S. *et al.* Population genomic structure and adaptation in the zoonotic malaria parasite Plasmodium knowlesi. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 13027–13032 (2015).

30. Nabet, C. *et al.* Genetic diversity of Plasmodium falciparum in human malaria cases in Mali. *Malar. J.* **15,** 353 (2016).

31. Pava, Z. *et al.* Genetic micro-epidemiology of malaria in Papua Indonesia: Extensive P. vivax

diversity and a distinct subpopulation of asymptomatic P. falciparum infections. *PLoS One* **12,** e0177445 (2017).

32. Baniecki, M. L. *et al.* Development of a Single Nucleotide Polymorphism Barcode to Genotype Plasmodium vivax Infections. *PLoS Negl. Trop. Dis.* **9,** e0003539 (2015).

33. Preston, M. D. *et al.* A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. *Nat. Commun.* **5,** 4052 (2014).

34. Creasey, A. M. *et al.* Uniparental inheritance of the mitochondrial gene cytochrome b in Plasmodium falciparum. *Curr. Genet.* **23,** 360–364 (1993).

35. Flick, K. & Chen, Q. var genes, PfEMP1 and the human host. *Mol. Biochem. Parasitol.* **134,** 3–9 (2004).

36. Clausen, T. M. *et al.* Structural and functional insight into how the Plasmodium falciparum VAR2CSA protein mediates binding to chondroitin sulfate A in placental malaria. *J. Biol. Chem.* **287,** 23332–23345 (2012).

37. Hulden, L. & Hulden, L. Activation of the hypnozoite: a part of Plasmodium vivax life cycle and survival. *Malar. J.* **10,** 90 (2011).

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

39. Cotter, C. *et al.* The changing epidemiology of malaria elimination: new strategies for new challenges. *Lancet (London, England)* **382,** 900–911 (2013).

40. Sattabongkot, J., Tsuboi, T., Zollner, G. E., Sirichaisinthop, J. & Cui, L. Plasmodium vivax transmission: chances for control? *Trends Parasitol.* **20,** 192–198 (2004).

41. Abdullah, N. R. *et al.* Plasmodium vivax population structure and transmission dynamics in Sabah Malaysia. *PLoS One* **8,** e82553 (2013).

42. Wangchuk, S. *et al.* Where chloroquine still works: the genetic make-up and susceptibility of Plasmodium vivax to chloroquine plus primaquine in Bhutan. *Malar. J.* **15,** 277 (2016).

43. Waltmann, A. *et al.* Increasingly inbred and fragmented populations of Plasmodium vivax associated with the eastward decline in malaria transmission across the Southwest Pacific. *PLoS Negl. Trop. Dis.* **12,** e0006146 (2018).

44. Arnott, A., Barry, A. E. & Reeder, J. C. Understanding the population genetics of Plasmodium vivax is essential for malaria control and elimination. *Malar. J.* **11,** 14 (2012).

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

46. Daniels, R. *et al.* A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking. *Malar. J.* **7,** 223 (2008).

47. Kantele, A. & Jokiranta, T. S. Review of cases with the emerging fifth human malaria parasite, Plasmodium knowlesi. *Clin. Infect. Dis.* **52,** 1356–1362 (2011).

48. Müller, M. & Schlagenhauf, P. <em>Plasmodium knowlesi</em> in travellers, update 2014. *Int. J. Infect. Dis.* **22,** 55–64 (2017).

49. Singh, B. & Daneshvar, C. Human infections and detection of Plasmodium knowlesi. *Clin. Microbiol. Rev.* **26,** 165–184 (2013).

50. Lubis, I. N. D. *et al.* Contribution of Plasmodium knowlesi to Multispecies Human Malaria Infections in North Sumatera, Indonesia. *J. Infect. Dis.* **215,** 1148–1155 (2017).

51. Lee, K.-S. *et al.* Plasmodium knowlesi: Reservoir Hosts and Tracking the Emergence in Humans and Macaques. *PLOS Pathog.* **7,** e1002015 (2011).

52. Pinheiro, M. M. *et al.* <italic>Plasmodium knowlesi</italic> Genome Sequences from Clinical Isolates Reveal Extensive Genomic Dimorphism. *PLoS One* **10,** e0121303 (2015).

53. Divis, P. C. S. *et al.* Admixture in Humans of Two Divergent <italic>Plasmodium knowlesi</italic> Populations Associated with Different Macaque Host Species. *PLoS Pathog* **11,** e1004888 (2015).

54. Ahmed, M. A., Fong, M. Y., Lau, Y. L. & Yusof, R. Clustering and genetic differentiation of the normocyte binding protein (nbpxa) of Plasmodium knowlesi clinical isolates from Peninsular Malaysia and Malaysia Borneo. *Malar. J.* **15,** 241 (2016).

55. Divis, P. C. S. *et al.* Three Divergent Subpopulations of the Malaria Parasite Plasmodium knowlesi. *Emerg. Infect. Dis.* **23,** 616–624 (2017).

56. Fornace, K. M. *et al.* Association between Landscape Factors and Spatial Patterns of Plasmodium knowlesi Infections in Sabah, Malaysia. *Emerg. Infect. Dis.* **22,** 201–208 (2016).

57. Yusof, R. *et al.* Phylogeographic Evidence for 2 Genetically Distinct Zoonotic Plasmodium knowlesi Parasites, Malaysia. *Emerg. Infect. Dis.* **22,** 1371–1380 (2016).

58. Vythilingam, I. *et al.* Natural transmission of Plasmodium knowlesi to humans by Anopheles latens in Sarawak, Malaysia. *Trans. R. Soc. Trop. Med. Hyg.* **100,** 1087–1088 (2006).

59. Tan, C. H., Vythilingam, I., Matusop, A., Chan, S. T. & Singh, B. Bionomics of Anopheles latens in Kapit, Sarawak, Malaysian Borneo in relation to the transmission of zoonotic simian malaria parasite Plasmodium knowlesi. *Malar. J.* **7,** 52 (2008).

60. Brant, H. L. *et al.* Vertical stratification of adult mosquitoes (Diptera: Culicidae) within a tropical rainforest in Sabah, Malaysia. *Malar. J.* **15,** 370 (2016).

61. Vythilingam, I. *et al.* Plasmodium knowlesi in humans, macaques and mosquitoes in peninsular Malaysia. *Parasit. Vectors* **1,** 26 (2008).

62. Moyes, C. L. *et al.* Defining the Geographical Range of the Plasmodium knowlesi Reservoir. *PLoS Negl. Trop. Dis.* **8,** e2780 (2014).

63. Sutherland, C. J. Persistent Parasitism: The Adaptive Biology of Malariae and Ovale Malaria. *Trends Parasitol.* **32,** 808–819 (2016).

64. Tachibana, M., Tsuboi, T., Kaneko, O., Khuntirat, B. & Torii, M. Two types of Plasmodium ovale defined by SSU rRNA have distinct sequences for ookinete surface proteins. *Mol. Biochem. Parasitol.* **122,** 223–226 (2002).

65. Collins, W. E. & Jeffery, G. M. Plasmodium ovale: Parasite and Disease. *Clin. Microbiol. Rev.* **18,** 570–581 (2005).

66. Collins, W. E. & Jeffery, G. M. Plasmodium malariae: Parasite and Disease. *Clin. Microbiol. Rev.* **20,** 579–592 (2007).

67. Rutledge, G. G. *et al.* Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. *Nature* **542,** 101–104 (2017).

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

69. Chang, H.-H. *et al.* THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. *PLOS Comput. Biol.* **13,** e1005348 (2017).

70. Rodrigues, P. T. *et al.* Using mitochondrial genome sequences to track the origin of imported plasmodium vivax infections diagnosed in the United States. *Am. J. Trop. Med. Hyg.* **90,** 1102–1108 (2014).

71. Benavente, E. D. *et al.* Genomic variation in Plasmodium vivax malaria reveals regions under selective pressure. *PLoS One* **12,** (2017).

72. Snounou, G., Viriyakosol, S., Jarra, W., Thaithong, S. & Brown, K. N. Identification of the four human malaria parasite species in field samples by the polymerase chain reaction and

detection of a high prevalence of mixed infections. *Mol. Biochem. Parasitol.* **58,** 283–292 (1993).

73. Nag, S. *et al.* High throughput resistance profiling of Plasmodium falciparum infections based on custom dual indexing and Illumina next generation sequencing-technology. *Sci. Rep.* **7,** (2017).

74. Lu, H., Giordano, F. & Ning, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics. Proteomics Bioinformatics* **14,** 265–279 (2016).

75. Oyola, S. O. *et al.* Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. *BMC Genomics* **13,** 1 (2012).

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED <u>FOR EACH</u> RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Ernest Diez Benavente
Principal Supervisor	Taane Clark & Susana Campino
Thesis Title	**Using whole genome sequence data to study genomic diversity and develop molecular barcodes to profile Plasmodium malaria parasites**

<u>If the Research Paper has previously been published please complete Section B, if not please move to Section C</u>

<u>SECTION B – Paper already published</u>

Where was the work published?	International Journal of Parasitology		
When was the work published?	December 2017		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

<u>SECTION C – Prepared for publication, but not yet published</u>

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	

Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I received the raw data from our collaborators. I used the previously created SMRT analysis suit to analyse the raw data and assembled the new genome reference, which was then annotated using the Companion system. I performed the QC analysis, as well as the interpretation of the data assembled. I later created custom R scripts to process the information in order to obtain information related to coverage, SNPs and other genomic information. I performed an analysis of the methylated patterns across the data. The figures presented in this work have all been generated using scripts written by myself or publicly available software (circos). I wrote the first draft of the manuscript and circulated to co-authors. Once the comments were received I gathered them and made the relevant changes on the article manuscript. I then performed the submission of the work to the International Journal of Parasitology journal and once comments from reviewers arrived, I made the corresponding changes and corrections.

Student Signature: ______________________________ **Date:** ______________

Supervisor Signature: ______________________________ **Date:** ______________

Chapter 2
A reference genome and methylome for the *Plasmodium knowlesi* A1-H.1 line

International Journal for Parasitology 48 (2018) 191–196

Contents lists available at ScienceDirect

International Journal for Parasitology

journal homepage: www.elsevier.com/locate/ijpara

Succinctus

A reference genome and methylome for the *Plasmodium knowlesi* A1-H.1 line

Ernest Diez Benavente[a], Paola Florez de Sessions[b], Robert W. Moon[a], Munira Grainger[c], Anthony A. Holder[c], Michael J. Blackman[a,c], Cally Roper[a], Christopher J. Drakeley[a], Arnab Pain[d], Colin J. Sutherland[a], Martin L. Hibberd[a,b], Susana Campino[a], Taane G. Clark[a,e,*]

[a] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom
[b] Genomics Institute Singapore, Singapore
[c] The Francis Crick Institute, 1 Midland Road, London NW1 1AT, United Kingdom
[d] King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
[e] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

ARTICLE INFO

Article history:
Received 15 April 2017
Received in revised form 10 August 2017
Accepted 26 September 2017
Available online 16 December 2017

Keywords:
Plasmodium knowlesi
Reference genome
PkA1-H.1
Sequencing
Malaria
Methylation

ABSTRACT

Plasmodium knowlesi, a common parasite of macaques, is recognised as a significant cause of human malaria in Malaysia. The *P. knowlesi* A1H1 line has been adapted to continuous culture in human erythrocytes, successfully providing an in vitro model to study the parasite. We have assembled a reference genome for the PkA1-H.1 line using PacBio long read combined with Illumina short read sequence data. Compared with the H-strain reference, the new reference has improved genome coverage and a novel description of methylation sites. The PkA1-H.1 reference will enhance the capabilities of the in vitro model to improve the understanding of *P. knowlesi* infection in humans.

Plasmodium knowlesi, a common malaria parasite of long-tailed *Macaca fascicularis* and pig-tailed *Macaca nemestrina* macaques in southeastern Asia, is now recognised as a significant cause of human malaria. Clinical outcomes range from high parasitaemia to severe complications including death (Singh and Daneshvar, 2013). Early cases were largely misdiagnosed as *Plasmodium malariae*, a morphologically similar but distantly-related species (Lee et al., 2009). Although sporadic human cases had been described in the 1960s, the public health importance of *P. knowlesi* was first understood with the reporting of a substantial focus of human infections in Malaysian Borneo in 2004 (Singh et al., 2004). Molecular detection has confirmed that human cases of *P. knowlesi* infection are relatively common in eastern Malaysia and occur in other southeastern Asian countries including The Philippines, Indonesia, Vietnam and Myanmar (Kantele and Jokiranta, 2011). The geographical distribution of *P. knowlesi* appears to be constrained by the range of its natural macaque hosts and the Leucosphyrus mosquito group vector (Moyes et al., 2014). There is no evidence of significant human-to-human transmission of *P. knowlesi* (Millar and Singh, 2015). Compared with other *Plasmodium* spp., the field of *P. knowlesi* genomics has been understudied, with only one reference genome available. The sequencing of the *P. knowlesi* reference "H-Pk1 (A+) clone" (PKNH, 14 chromosomes, 23.5 Mb, 5188 genes, 37.5% GC content) (Pain et al., 2008) has provided insights into novel genomic features, including highly variable *kir* and *sicavar* protein families, but is incomplete. Genetic diversity among *P. knowlesi* isolates is high compared with other members of the genus. Genome variation in this species exhibits dimorphism (Pinheiro et al., 2015), and there is evidence this may be driven by partitioning between the two distinct macaque hosts, *M. fasciularis* and *M. nemestrina* (Assefa et al., 2015).

Recently, the *P. knowlesi* A1.H1 (PkA1-H.1) line was the first to be successfully adapted to continuous culture in human erythrocytes, providing an in vitro model suitable for genetic modification (Moon et al., 2013). To support in vitro studies with this human-adapted clonal line we have assembled a new reference genome for the PkA1-H.1 line using PacBio (Pacific Biosciences Inc., USA) RS-II long read and Illumina HiSeq short read sequence data. An advantage of PacBio RS-II SMRT cell sequencing is the potential

* Corresponding author at: Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, United Kingdom.
E-mail address: taane.clark@lshtm.ac.uk (T.G. Clark).

https://doi.org/10.1016/j.ijpara.2017.09.008
0020-7519/© 2017 The Authors. Published by Elsevier Ltd on behalf of Australian Society for Parasitology.

to identify modifications on individual nucleotide bases, and thereby provide insights into methylation sites (Morgan et al., 2016). Cytosine and adenine DNA methylation is an epigenetic mark in most eukaryotic cells that regulates numerous processes including gene expression and stress responses. Genome-wide analysis of DNA methylation in *Plasmodium falciparum* has mapped the positions of methylated cytosines. This work has identified a single functional DNA methyltransferase, *PfDNMT* (*PF3D7_0727300*), which may mediate these genomic modifications (Ponts et al., 2013) and is thus a potential target for anti-malarial drugs. Analyses have revealed that the malaria genome is asymmetrically methylated, in which only one DNA strand is methylated, and this could regulate virulence gene expression and transcription elongation (Ponts et al., 2013). Using PacBio RS-II data, we describe for the first known time, over 40 000 potential modified bases in the *P. knowlesi* genome, including ~5% that were specifically pinpointed as 6-methyladenine modifications, which recently have been shown to have a role in epigenetic regulation of gene expression in other eukaryotic organisms (Greer et al., 2016). Both the PkA1.H-1 reference genome and the associated methylation data are available to support future in vitro and in vivo studies of *P. knowlesi* parasites, and assist the continuing development of anti-malarial drugs and vaccines.

DNA of high quality and molecular weight was purified from phenol chloroform extraction of magnetic activated cell sorted column enriched PkA1-H.1 schizonts (Bates et al., 2010). The DNA (20 µg, A260/280:1.98) was used to prepare 20 kb insert libraries, which were sequenced using nine SMRT cells on the PacBio RS-II device (Pacific Biosciences Inc.) at the Genome Institute of Singapore. The sequencing yielded a total of 365,956 reads with a mean length of 9645 bp. These data were complemented by raw sequences from the Illumina HiSeq2500 platform (500 bp fragment, 150 bp paired end reads), performed at King Abdullah University of Science and Technology, Kingdom of Saudi Arabia, and yielded in excess of 5 million reads. PacBio data were available for the A1-H.1 clone, but were insufficient in themselves to lead to a complete genome i.e. single contig chromosomes (Moon et al., 2016). The de novo assembly of the reads using the HGAP3 software pipeline within the SMRT Portal (Koren et al., 2012) yielded a total of 111 contigs with a theoretical ~145-fold coverage. These contigs were then used as a reference to map with the *bwa-mem* aligner (Li and Durbin, 2009) almost 5 million PkA1-H.1 Illumina 150 bp reads, leading to 7120 single nucleotide polymorphisms (SNPs) and insertion and deletion (indel) corrections. The high quality corrected contigs were ordered with Abacas software (http://abacas.sourceforge.net) using the current *Plasmodium knowlesi H strain* (PKNH) reference (version 2.0, www.genedb.org), and manually checked to remove possible errors. This led to 14 complete chromosomes (number of contigs: range 1 –5) and mitochondrial (two contigs, copy number seven-fold), and apicoplast (one contig, 1.8 copies) genomes. The PkA1-H.1 genome was then annotated using the Companion webserver (https://companion.sanger.ac.uk).

The resulting PkA1-H.1 reference covers 98.2% of the PKNH (v2) reference and 100% of a smaller draft assembly for this line (Moon et al., 2016), and contains only 42 gaps. Compared with the PKNH (v2), PkA1-H.1 has 3993 SNP and 19,936 indel differences. The new genome improved characterisation of *kir* and *sicavar* genes, leading to a total genome length of 24.4 Mb (chromosome size range: 726,979–3,301,832). To assess the performance of the PkA1-H.1 reference, we aligned whole genome sequence data from 60 published human *P. knowlesi* isolates (Assefa et al., 2015; Pinheiro et al., 2015). Raw Illumina sequence data were downloaded for 60 *P. knowlesi* isolates (Assefa et al., 2015; Pinheiro et al., 2015) (Supplementary Table S1). The samples were aligned against the PkA1-H.1 reference using *bwa-mem* (Li and Durbin, 2009) and SNPs were called using the *Samtools* software suite (Li, 2011), from which a set of high quality SNPs was filtered using previously described methods (Samad et al., 2015; Campino et al., 2016). The alignment process yielded an average coverage of ~143-fold across 99% of the reference genome, and 1,632,024 high quality SNPs (one every 15 bp) were characterised (Fig. 1). These coverage statistics were marginally superior to mapping to PKNH, which yielded average coverage of ~140-fold across 96% of the reference genome (Supplementary Table S1). The improvement includes an increase in coverage in the *kir* and *sicavar* genes, and the closing of several gaps in these genes that exist in the PKNH reference.

The 38-fold sequence coverage obtained during the new sequencing was in excess of the recommended 25-fold for the 6-methyladenine (m6A) and 4-methylcytosine (m4C) modification detection (Morgan et al., 2016). The *RS_Modification_and_Motif_analysis.1* program within the SMRT Analysis Portal was used to perform a methylation analysis (Morgan et al., 2016). Overall, we have identified 41,508 potential modified bases (*P* values <0.01) in the *P. knowlesi* genome (by chromosome: mean 2965; range 1089–5823) (Supplementary Table S2). Of the total, 7231 modifications (~17%) were identified from an independent unanalysed dataset (Moon et al., 2016). Approximately five percent (2218) of the modifications were specifically pinpointed as m6A modifications, a type that has been recently confirmed to play a role in epigenetic regulation of eukaryotic organisms such as *Caenorhabditis elegans* (Greer et al., 2016). Furthermore, 3646 (~9%) modified bases were classified as m4C methylation events. The proportion of the total adenosine and cytosine bases that are modified is 0.11 and 0.25, respectively. These fractions are lower than those reported in a study of *P. falciparum* for 5-methylcytosine (m5C), which estimated that two-thirds of the cytosine bases were methylated (Ponts et al., 2013). Analysis of the distribution of the methylation sites across the genome revealed that 45.3% of the m6A sites are located within gene boundaries. However, the m4C modifications were distributed more evenly, with 50.1% being intragenic (m4C versus m6A, $P < 0.0004$). We calculated the number of modified bases over a 10 kb sliding window, revealing a stable distribution over the different chromosomes (modified bases/kb: mean 1.6; range 1.38–1.81) (Fig. 1). In order to identify genomic regions (islands) with an accumulation of modified bases, the fold change of modified bases per window was compared with the chromosome average. This analysis revealed 85 10 kb regions that present at least an increase of two-fold change over the chromosome average in either of the two PacBio datasets analysed (Table 1, Supplementary Table S3). These regions included four methyltransferase genes, and several loci involved in the ribosome constitution and with DNA/RNA manipulation functionality. Further, the *reticulocyte binding protein NBPXa* gene (*PKNH_1472300*) was also identified, which has been demonstrated to be essential for the infection of human red blood cells in this *P. knowlesi* line (Moon et al., 2016).

The dominant cause of malaria in Malaysia is now *P. knowlesi*. To assist with disease control, a deeper understanding of the biology of this neglected parasite is required, and genomics has the potential to provide useful insights. The use of state-of-the-art technology such as PacBio RS-II SMRT cell sequencing has assisted with the biological understanding of constantly changing parasite populations (Ahmed et al., 2016). Using sequence data from PacBio RS-II and Illumina HiSeq2500 technologies, we have constructed a reference genome for the PkA1-H.1 line, which is the first known *P. knowlesi* line to be successfully adapted to continuous culture in human erythrocytes. The PkA1-H.1 reference is complete to a chromosome level, base-corrected, fully annotated, and spans over 98% of the PKNH "H strain" reference, including the closure of some of its gaps in highly variable *kir* and *sicavar* genes. The completeness of the PkA1-H.1 genome improved alignment of sequences and variant detection from published *P. knowlesi* clinical strains, when

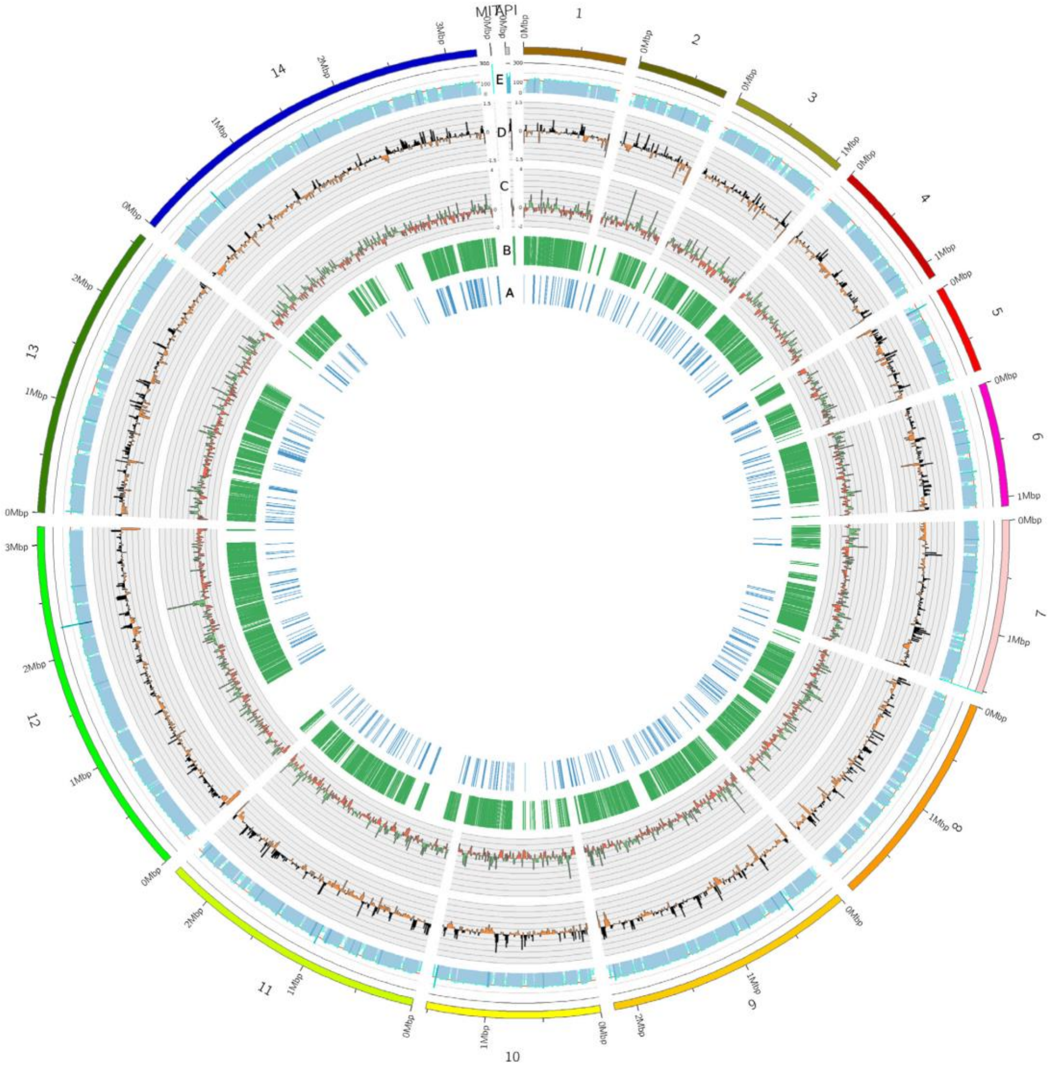


Fig. 1. Analyses using the assembled *Plasmodium knowlesi* A1-H.1 (PkA1-H.1) genome. (A) The 3993 single nucleotide polymorphism differences with PKNH *Plasmodium knowlesi* H strain version 2. (B) The 19,936 insertion and deletion (indel) differences with PKNH version 2. (C) The distribution of the methylated bases per window. (D) Average coverage from mapping 60 publically available samples, using a non-overlapping 10 kb sliding window. (E) The single nucleotide polymorphism density as the fold change of number of single nucleotide polymorphism positions compared with the chromosome average.

compared with using the PKNH genome reference (Supplementary Table S1). The long-read sequencing technology combined with Illumina paired sequences effectively resolved gaps in the genome caused by low complexity regions and large multigene families such as the *sicavar* genes. A more complete genome would assist population genomic studies of genes critical to host-pathogen interactions and virulence. This could include the use of field isolates from across southeastern Asia to investigate host-parasite population structure and to develop molecular barcodes for surveillance (Preston et al., 2014).

The use of the PacBio RS-II technology allowed us to describe for the first time the distribution of methylated bases, particularly m4C and m6A modifications, with single base resolution in the *P. knowlesi* genome. Whilst m4C modifications are usually confined to prokaryotes, m6A modifications have been previously described as having a role as a transcription regulator in eukaryotes such as *C. elegans* (Greer et al., 2016). The contribution of m6A modifications to epigenetic control of gene expression could be investigated by integration of genomics and methylation with RNAseq transcriptome studies. Concordance of modifications from the same *P. knowlesi* lines identified in PacBio sequence data obtained in an earlier study (Moon et al., 2016) was modest (~20%), but this may be expected as the DNA extracted for sequencing was not derived from synchronised identical parasite cultures. The regions in the genome with the highest density of modified bases included four methyltransferase genes (*PKNH_0103500, PKNH_0211900, PKNH_0305100* and *PKNH_1416400*), which could suggest a role of epigenetic modifications in the regulation of methylation path-

Table 1
Regions with putative methylation sites in *Plasmodium knowlesi* line A1-H.1.

Chr.	Window	Fold-change[a]	Gene IDs (*PKNH_*)[b]
1	90001–100000	2.39, 1.08	0101800, 0101900, 0102000, 0102100
1	220001–230000	2.92, 1.40	**0103500, 0103600, 0103700**
1	300001–310000	2.16, 1.68	**0105200, 0105700**
1	410001–420000	2.22, 1.01	0108300, 0108400
1	810001–820000	2.34, 1.25	0117200, 0117400
2	480001–490000	2.42, 1.29	0210600, 0210700, 0210800
2	540001–550000	0.75, 2.00	0211700, 0211800, 0211900, 0212000
3	230001–240000	2.30, 1.26	0304200, 0304300, 0304400
3	260001–270000	2.24, 1.02	0304900, 0305000, 0305100, 0305200
3	310001–320000	2.66, 1.55	0306400, 0306500
3	460001–470000	2.54, 1.66	0310200, 0310300, 0310400
3	750001–760000	2.96, 1.81	**0315000, 0315100, 0315200**
4	160001–170000	2.39, 0.73	0404100, 0404200
4	170001–180000	2.39, 0.81	0404300, 0404400
4	780001–790000	2.33, 1.07	0418100, 0418200, 0418300, 0418400
4	1010001–1020000	0.69, 2.17	0422400, 0422500, 0422600
5	110001–120000	2.10, 1.49	0502300
5	230001–240000	1.09, 2.01	0505400, 0505500, 0505600
5	570001–580000	2.03, 1.28	0512300, 0512400, 0512500
5	600001–610000	2.32, 1.07	0512700, 0512800
5	710001–720000	1.02, 2.13	0514600, 0514700, 0514800, 0514900, 0515000, 0515100
6	220001–230000	2.66, 0.86	0605300, 0605400, 0605500
6	570001–580000	2.01, 0.98	0612900, 0613000
7	280001–290000	2.98, 1.99	0705200, 0705300, 0705400
7	1180001–1190000	2.36, 1.09	0726600, 0726700, 0726800, 0726900, 0727000
8	240001–250000	2.25, 1.17	0804700, 0804800, 0804900
8	350001–360000	0.81, 2.55	0807600
8	500001–510000	2.06, 1.56	0811300, 0811400, 0811500
8	600001–610000	2.44, 1.29	0813500, 0813600, 0813700
8	1090001–1100000	2.19, 1.00	**0823500, 0823600**
8	1400001–1410000	2.19, 1.59	0830400, 0830500, 0830600, 0830700
8	1680001–1690000	2.75, 1.70	0837400, 0837500, 0837600
9	270001–280000	2.21, 1.45	0905300, 0905400
9	720001–730000	2.21, 0.83	0915800, 0915900, 0916000
9	830001–840000	2.10, 0.90	0918700, 0918800, 0918900, 0919000
9	1440001–1450000	1.16, 2.04	0932400
9	1660001–1670000	2.10, 0.95	0937000
9	1960001-1970000	2.10, 1.38	0943000, 0943100
9	2010001–2020000	2.71, 1.16	0944300
9	2110001–2120000	2.55, 1.36	0946100
9	2130001–2140000	2.44, 0.49	0946200
10	540001–550000	2.15, 1.06	1011700, 1011800
10	1090001–1100000	2.38, 1.29	1023800, 1023900, 1024000
10	1390001–1400000	2.38, 1.52	1030600, 1030700
10	1410001–1420000	2.26, 1.06	1031200, 1031300, 1031400, 1031500
10	1420001–1430000	2.21, 1.60	1031500, 1031600, 1031700, 1031800
11	300001–310000	2.30, 1.59	1107200, 1107300, 1107400
11	320001–330000	2.24, 1.21	1107700, 1107800, 1107900, 1108000
11	460001–470000	1.19, 2.12	1110700, 1110800
11	500001–510000	2.30, 1.64	1111500, 1111600
11	680001–690000	2.17, 0.94	1114700, 1114800
11	1700001–1710000	2.24, 1.02	1137000, 1137100
11	1710001–1720000	2.04, 1.49	1137200, 1137300, 1137400, 1137500, 1137600
11	2120001–2130000	2.77, 0.73	1144900, 1145000
12	700001–710000	2.23, 1.71	1215700, 1215800, 1215900
12	760001–770000	2.00, 1.06	1217000, 1217200
12	920001–930000	2.45, 1.70	1219600, 1219700, 1219800, 1219900
12	1050001–1060000	2.17, 0.91	1223800
12	1130001–1140000	2.73, 1.06	1225300, 1225400, 1225500, 1225600
12	1730001–1740000	2.39, 1.05	1239200, 1239300, 1239400
12	1750001–1760000	2.06, 0.87	1239600, 1239700
12	1760001–1770000	2.06, 1.49	1239700, 1239800
12	1810001–1820000	2.00, 0.91	1241000
12	2200001–2210000	2.34, 1.02	1248600, 1248700, 1248800
12	2230001–2240000	2.28, 0.87	1249300, 1249400, 1249500
12	2240001–2250000	2.23, 1.19	1249500, 1249600
12	2630001–2640000	2.00, 1.31	1257800, 1257900, 1258000
13	350001–360000	2.36, 1.32	1307500, 1307600, 1307700, 1307800, 1307900
13	400001–410000	2.01, 0.89	1308600, 1308700
13	1100001–1110000	2.54, 0.57	1324500, 1324600
13	1450001–1460000	2.12, 1.06	1331100, 1331200, 1331300
13	1650001–1660000	2.89, 0.72	1336600, 1336900, 1337000
13	1860001–1870000	2.18, 1.57	1341400, 1341500
13	1920001–1930000	2.12, 0.71	1342600

Table 1 (*continued*)

Chr.	Window	Fold-change[a]	Gene IDs (*PKNH_*)[b]
13	2250001–2260000	2.60, 1.38	1349600, 1349700
13	2340001–2350000	1.18, 2.10	1351200, 1351300, 1351400
13	2380001–2390000	2.89, 1.43	**1352100, 1352200, 1352300**
13	2400001–2410000	2.24, 0.99	1352500, 1352600, 1352800
13	2420001–2430000	2.30, 1.28	1353100, 1353200, 1353300
14	740001–750000	2.25, 1.18	1416300, 1416400, 1416500
14	900001–910000	2.19, 1.26	1420200, 1420300, 1420400
14	2000001–2010000	1.68, 0.72	1445800, 1445900, 1446000
14	2830001–2840000	2.19, 1.23	1463300
14	3020001–3030000	2.19, 0.80	1467900, 1468000, 1468100, 1468200
14	3200001–3210000	2.64, 1.04	**1472300, 1472400**

Chr, chromosome number; PKNH, *Plasmodium knowlesi* H strain.

[a] Deviation from chromosomal mean: this study, data analysed from Moon et al. (2016).

[b] Bold indicates methylation related annotation.

ways. We also observed a wide range of ribosomal proteins and genes involved in manipulation of the genetic material. Further, the *reticulocyte binding protein NBPXa* gene (*PKNH_1472300*), which is essential for the infection of human erythrocytes in the A1-H.1 line, was one of the genes showing highest density of modifications in the genome. The modifications associated with the *P. knowlesi* orthologue for the *P. falciparum* mediator *PfDNMT* (*PKNH_0211900*) were not detected, but are thought to create m5C modifications. It is not possible to detect m5C modifications described in other *Plasmodium* spp., due to limitations in sequencing platform coverage. At least 250× coverage would be required to confidently detect m5C modifications, although some uncharacterised modifications might refer to m5C methylation events (Morgan et al., 2016).

In summary, we have provided a genomic reference and methylation data for the PkA1-H.1 line for researchers to undertake biological and clinical research into *P. knowlesi* and other malaria parasites, potentially assisting with the design of anti-malarial drugs and vaccines, and diagnostic tools.

The reference genome and methylome are available from pathogenseq.lshtm.ac.uk/knowlesi_1/. The underlying raw sequence data is available from the European Nucleotide Archive (accession numbers: PRJEB19298, ERS763679).

Acknowledgements

TGC is funded by the Medical Research Council (MRC), UK (Grant no. MR/K000551/1, MR/M01360X/1, MR/N010469/1). SC and CR are funded by the MRC (Grant no. MR/M01360X/1). RWM is supported by an MRC Career Development Award jointly funded by the MRC and UK Department for International Development. This work was supported in part by the Francis Crick Institute, UK which receives its core funding from Cancer Research, UK (FC001097, FC001043), the MRC (FC001097, FC001043), and the Wellcome Trust, UK (FC001097, FC001043) We gratefully acknowledge the Scientific Computing Group for data management and computer infrastructure at Genome Institute of Singapore for their help. Bioinformatic analysis was performed on the MRC, UK, funded eMedLab computing resource.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ijpara.2017.09.008.

References

Ahmed, M.A., Fong, M.Y., Lau, Y.L., Yusof, R., 2016. Clustering and genetic differentiation of the normocyte binding protein (nbpxa) of *Plasmodium knowlesi* clinical isolates from Peninsular Malaysia and Malaysia Borneo. Malar. J. 15, 241.

Assefa, S., Lim, C., Preston, M.D., Duffy, C.W., Nair, M.B., Adroub, S.A., Kadir, K.A., Goldberg, J.M., Neafsey, D.E., Divis, P., Clark, T.G., Duraisingh, M.T., Conway, D.J., Pain, A., Singh, B., 2015. Population genomic structure and adaptation in the zoonotic malaria parasite *Plasmodium knowlesi*. Proc. Nat. Acad. Sci. U.S.A. 112, 13027–13032.

Bates, A.H., Mu, J., Jiang, H., Fairhurst, R.M., Su, X.Z., 2010. Use of magnetically purified *Plasmodium falciparum* parasites improves the accuracy of erythrocyte invasion assays. Exp. Parasitol. 126, 278–280.

Campino, S., Benavente, E.D., Assefa, S., Thompson, E., Drought, L.G., Taylor, C.J., Gorvett, Z., Carret, C.K., Flueck, C., Ivens, A.C., Kwiatkowski, D.P., Alano, P., Baker, D.A., Clark, T.G., 2016. Genomic variation in two gametocyte non-producing *Plasmodium falciparum* clonal lines. Malar. J. 15, 229.

Greer, E.L., Blanco, M.A., Gu, L., Sendinc, E., Liu, J., Aristizábal-Corrales, D., Hsu, C.H., Aravind, L., He, C., Shi, Y., 2016. DNA methylation on N6-adenine in *C. elegans*. Cell 161, 868–878.

Kantele, A., Jokiranta, T.S., 2011. Review of cases with the emerging fifth human malaria parasite, *Plasmodium knowlesi*. Clin. Infect. Dis. 52, 1356–1362.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., Adam, M.P., 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat. Biotechnol. 30, 693–700.

Lee, K.S., Cox-Singh, J., Singh, B., 2009. Morphological features and differential counts of *Plasmodium knowlesi* parasites in naturally acquired human infections. Malar. J. 8, 73.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25, 1754–1760.

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993.

Millar, S., Singh, B., 2015. Human infections with *Plasmodium knowlesi*—zoonotic malaria. Clin. Microbiol. Infect. 21, 640–648.

Moon, R.W., Sharaf, H., Hastings, C.H., Ho, Y.S., Nair, M.B., Rchiad, Z., Nair, M.B., Rchiad, Z., Knuepfer, E., Ramaprasad, A., Mohring, F., Amir, A., Yusuf, N.A., Hall, J., Almond, N., Lau, Y.L., Pain, A., Blackman, M.J., Holder, A.A., 2016. Normocyte-binding protein required for human erythrocyte invasion by the zoonotic malaria parasite *Plasmodium knowlesi*. Proc. Natl. Acad. Sci. U.S.A. 113, 7231–7236.

Moon, R.W., Hall, J., Rangkuti, F., Ho, Y.S., Almond, N., Mitchell, G.H., Pain, A., Holder, A.A., Blackman, M.J., 2013. Adaptation of the genetically tractable malaria pathogen *Plasmodium knowlesi* to continuous culture in human erythrocytes. Proc. Natl. Acad. Sci. U.S.A. 110, 531–536.

Morgan, R.D., Luyten, Y.A., Johnson, S.A., Clough, E.M., Clark, T.A., Roberts, R.J., 2016. Novel m4C modification in type I restriction-modification systems. Nucleic Acids Res. 44, 9413–9425.

Moyes, C.L., Henry, A.J., Golding, N., Huang, Z., Singh, B., Baird, J.K., Newton, P.N., Huffman, M., Duda, K.A., Drakeley, C.J., Elyazar, I.R., Anstey, N.M., Chen, Q., Zommers, Z., Bhatt, S., Gething, P.W., Hay, S.I., 2014. Defining the geographical range of the *Plasmodium knowlesi* reservoir. PLoS Negl. Trop. Dis. 8, e2780.

Pain, A., Böhme, U., Berry, A.E., Mungall, K., Finn, R.D., Jackson, A.P., Mourier, T., Mistry, J., Pasini, E.M., Pasini, E.M., Aslett, M.A., Balasubrammaniam, S., Borgwardt, K., Brooks, K., Carret, C., Carver, T.J., Cherevach, I., Chillingworth, T., Clark, T.G., Galinski, M.R., Hall, N., Harper, D., Harris, D., Hauser, H., Ivens, A., Janssen, C.S., Keane, T., Larke, N., Lapp, S., Marti, M., Moule, S., Meyer, I.M., Ormond, D., Peters, N., Sanders, M., Sanders, S., Sargeant, T.J., Simmonds, M., Smith, F., Squares, R., Thurston, S., Tivey, A.R., Walker, D., White, B., Zuiderwijk, E., Churcher, C., Quail, M.A., Cowman, A.F., Turner, C.M., Rajandream, M.A., Kocken, C.H., Thomas, A.W., Newbold, C.I., Barrell, B.G., Berriman, M., 2008. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature 455, 799–803.

Pinheiro, M.M., Ahmed, M.A., Millar, S.B., Sanderson, T., Otto, T.D., Lu, W.C., Krishna, S., Rayner, J.C., Cox-Singh, J., 2015. *Plasmodium knowlesi* genome sequences from clinical isolates reveal extensive genomic dimorphism. PLoS One 10, e0121303.

Ponts, N., Fu, L., Harris, E.Y., Zhang, J., Chung, D.-W.D., Cervantes, M.C., Prudhomme, J., Atanasova-Penichon, V., Zehraoui, E., Bunnik, E.M., Rodrigues, E.M., Lonardi, S., Hicks, G.R., Wang, Y., Le Roch, K.G., 2013. Genome-wide mapping of DNA methylation in the human malaria parasite *Plasmodium falciparum*. Cell Host Microbe 14, 696–706.

Preston, M.D., Campino, S., Assefa, S.A., Echeverry, D.F., Ocholla, H., Amambua-Ngwa, A., Stewart, L.B., Conway, D.J., Borrmann, S., Michon, P., Zongo, I., Ouédraogo, J.B., Djimde, A.A., Doumbo, O.K., Nosten, F., Pain, A., Bousema, T., Drakeley, C.J., Fairhurst, R.M., Sutherland, C.J., Roper, C., Clark, T.G., 2014. A barcode of organellar genome polymorphisms identifies the geographic origin of *Plasmodium falciparum* strains. Nat. Commun. 5, 4052.

Samad, H., Coll, F., Preston, M.D., Ocholla, H., Fairhurst, R.M., Clark, T.G., 2015. Imputation-based population genetics analysis of *Plasmodium falciparum* malaria parasites. PLoS Genet. 11, e1005131.

Singh, B., Kim Sung, L., Matusop, A., Radhakrishnan, A., Shamsul, S.S., Cox-Singh, J., Thomas, A., Conway, D.J., 2004. A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. Lancet 363, 1017–1024.

Singh, B., Daneshvar, C., 2013. Human infections and detection of *Plasmodium knowlesi*. Clin. Microbial. Rev. 26, 165–184.

Supplementary Information

Supplementary Table S1. Illumina sequenced *Plasmodium knowlesi* samples used in this study to support new reference generation.

Sample	Total	PkA1-H.1	PkA1-H.1 Fraction genome	PkA1-H.1	PKNH Mapped	PKNH Fraction genome	PKNH Coverage
	Reads	Mapped	covered	Coverage	Reads (%)	covered	Mean
		Reads (%)		Mean			
ERR274221	46353062	96.9	0.99	174.1	97.0	0.96	175.5
ERR274222	60189967	97.7	0.99	228.7	97.8	0.96	230.7
ERR274224	42350437	97.0	0.98	161.5	97.0	0.95	162.8
ERR274225	52883838	97.2	0.98	201.6	97.2	0.95	203.2
ERR366425	6391836	97.1	0.95	36.0	97.2	0.92	36.3
ERR366426	6270503	94.8	0.95	34.1	94.8	0.92	34.3
ERR985372	30620879	66.0	0.98	79.7	66.0	0.95	80.3
ERR985374	55120863	96.4	0.99	207.5	80.5	0.95	102.6
ERR985375	66747737	22.7	0.98	55.4	96.5	0.96	209.2
ERR985376	50924729	79.3	0.99	157.7	22.7	0.95	55.8
ERR985377	71277238	95.3	0.99	248.7	79.4	0.96	159.1
ERR985378	47365570	95.6	0.99	177.7	95.4	0.96	250.7
ERR985379	40897898	95.5	0.99	153.6	95.7	0.96	179.1
ERR985380	61989913	97.2	1.00	234.8	95.6	0.96	154.8
ERR985381	51234398	94.9	0.99	189.9	97.3	0.97	236.8
ERR985382	48838502	49.6	0.98	89.3	94.9	0.96	191.5
ERR985383	40307405	85.8	0.99	126.8	49.6	0.95	90.0
ERR985384	52753853	88.0	0.99	167.0	85.8	0.96	127.7
ERR985385	93728041	60.6	1.00	213.1	88.0	0.96	171.4
ERR985386	52908664	22.3	0.98	45.3	60.6	0.97	214.9
ERR985387	48092075	91.8	0.99	173.0	22.3	0.95	45.6
ERR985388	68600487	92.2	0.99	247.4	91.8	0.96	174.4
ERR985394	26219163	95.2	0.98	99.0	92.3	0.96	249.4
ERR985395	45997535	86.1	1.00	154.6	88.6	0.95	108.1
ERR985396	31849384	60.9	0.99	76.4	85.7	0.95	101.0
ERR985397	40463921	92.0	1.00	146.3	91.5	0.95	121.6
ERR985398	33387794	61.2	0.98	80.5	88.4	0.95	115.4
ERR985399	34420982	97.0	0.99	131.2	73.2	0.95	75.4
ERR985400	36366112	84.6	0.99	121.4	95.3	0.95	99.8
ERR985401	36045889	95.3	0.99	135.9	86.1	0.97	155.9
ERR985402	56171218	92.5	0.99	202.0	60.9	0.96	76.9
ERR985403	52517284	95.4	0.99	193.6	92.1	0.96	147.5
ERR985404	52157592	94.6	1.00	190.2	61.3	0.95	81.1
ERR985410	32160549	97.1	0.99	122.9	97.0	0.96	132.2
ERR985411	54179835	80.4	0.98	161.6	84.6	0.95	122.3
ERR985412	66023442	44.2	0.99	109.4	95.4	0.95	137.0
ERR985414	65300398	93.9	0.99	242.3	92.6	0.96	203.6
ERR985415	60931327	92.5	0.99	223.7	95.4	0.96	195.2
ERR985416	78064720	92.6	0.99	286.4	94.7	0.96	191.7
ERR985417	41500510	95.7	0.99	157.1	89.9	0.96	109.1
ERR985418	54311715	89.9	0.98	191.6	85.3	0.95	119.8
ERR985419	57630655	81.2	0.99	183.6	87.7	0.95	115.2
ERR985373	42713162	78.6	0.98	119.3	88.5	0.95	127.1
ERR985389	41469453	86.3	0.99	126.4	82.9	0.96	115.8
ERR985390	41673914	83.2	0.99	117.2	97.1	0.96	123.9
ERR985391	40562077	90.6	0.99	141.0	80.4	0.95	162.9
ERR985392	42267002	87.0	0.99	132.1	44.2	0.96	110.2
ERR985393	34929904	71.1	0.98	88.5	77.7	0.95	102.9
ERR985405	38834680	88.7	1.00	128.1	93.9	0.95	244.2
ERR985406	47157966	83.6	0.99	138.9	92.5	0.96	225.6
ERR985407	43617554	86.2	0.99	135.2	92.7	0.96	289.0
ERR985408	47394270	86.8	0.99	148.1	95.8	0.96	158.4
ERR985409	50949905	80.5	0.99	134.1	89.9	0.95	193.1
ERR985413	51040091	74.5	0.98	119.8	81.3	0.96	185.0
SRR2221468	22542689	94.5	0.99	83.2	94.6	0.96	83.8
SRR2222335	23371486	96.1	1.00	89.6	96.2	0.99	90.0
SRR2225467	19455031	79.3	1.00	60.9	79.4	0.99	61.4
SRR2225571	22264926	74.6	0.98	64.0	74.6	0.95	64.5
SRR2225573	25538996	95.2	0.99	94.6	95.2	0.96	95.4
SRR3135172	21164226	90.1	1.00	74.6	90.1	0.99	75.6

PKNH, *Plasmodium knowlesi* **H** strain

Supplementary Table 2 from this article has the following title: "**Supplementary Table S2.** Methylated sites identified in Plasmodium knowlesi data from our study, and their presence in the re-analysed clone described in Moon et al. (2016)." and can be found following this link (https://ars.els-cdn.com/content/image/1-s2.0-S0020751917303466-mmc2.xlsx). It will not be printed here given it is composed of 40,000 rows and was uploaded to the journal as an EXCEL table.

Supplementary Table S3. Methylated sites identified in *Plasmodium knowlesi i*n our study and the re-analysed clone data from Moon et al. (2016). Highlighted in bold the relevant genes with functions related to ribosomal function, methylation, DNA interactions and invasion.

Chromosome	Position start window	Position end window	Modifications (this study)	Modifications (Moon et al., 2016)	Deviation from chromosomal mean (this study)	Deviation from chromosomal mean (Moon et al., 2016)	Gene IDs	Products
PKNH_01_v2	90001	100000	41	73	2.39	1.08	PKNH_0101800 /PKNH_0101900 /PKNH_0102000 /PKNH_0102100	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_01_v2	220001	230000	50	94	2.92	1.4	**PKNH_0103500** /PKNH_0103600 /PKNH_0103700	**methyltransferase, putative**/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_01_v2	300001	310000	37	113	2.16	1.68	**PKNH_0105200** /PKNH_0105700	**DNA mismatch repair protein MSH2**, putative/conserved Plasmodium protein, unknown function
PKNH_01_v2	410001	420000	38	68	2.22	1.01	PKNH_0108300 /PKNH_0108400	lysophospholipase, putative/lysophospholipase, putative

PKNH_01_v2	810001	820000	40	84	2.34	1.25	PKNH_0117200 /PKNH_0117400	sentrin-specific protease 2, putative (SENP2)/nucleolar protein 10, putative (NOL10)
PKNH_02_v2	480001	490000	42	80	2.42	1.29	PKNH_0210600 /PKNH_0210700 /PKNH_0210800	vacuolar protein sorting-associated protein VTA1, putative/aspartate--tRNA ligase, putative/conserved Plasmodium protein, unknown function
PKNH_02_v2	540001	550000	13	124	0.75	2	PKNH_0211700 /PKNH_0211800 /**PKNH_0211900** /PKNH_0212000	conserved Plasmodium protein, unknown function/cysteine desulfurase, putative (NFS)/**DNA (cytosine-5)-methyltransferase, putative (DNMT)**/proteasome subunit alpha type-5, putative
PKNH_03_v2	230001	240000	38	79	2.30	1.26	PKNH_0304200 /PKNH_0304300 /PKNH_0304400	V-type proton ATPase subunit B, putative/sexual stage-specific protein precursor, putative/cytosolic glyoxalase **II**, putative (cGlo**II**)
PKNH_03_v2	260001	270000	37	64	2.24	1.02	PKNH_0304900 /PKNH_0305000 /**PKNH_0305100** /PKNH_0305200	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/**methyltransferase, putative/**peptidyl-tRNA hydrolase 2, putative (PTH2)
PKNH_03_v2	310001	320000	44	97	2.66	1.55	PKNH_0306400 /PKNH_0306500	conserved Plasmodium protein, unknown function/flap

								endonuclease 1, putative (FEN1)
PKNH_03_v2	460001	470000	42	104	2.54	1.66	PKNH_0310200 /PKNH_0310300 /PKNH_0310400	conserved Plasmodium protein, unknown function/ribosomal protein, putative/conserved Plasmodium protein, unknown function
PKNH_03_v2	750001	760000	49	113	2.96	1.81	**PKNH_0315000** **/PKNH_0315100** /PKNH_0315200	**60S ribosomal protein L11a, putative/40S ribosomal protein S10**, putative/conserved Plasmodium protein, unknown function
PKNH_04_v2	160001	170000	38	52	2.39	0.73	PKNH_0404100 /PKNH_0404200	conserved Plasmodium protein, unknown function/autophagy-related protein 11, putative (ATG11)
PKNH_04_v2	170001	180000	38	58	2.39	0.81	PKNH_0404300 /PKNH_0404400	MtN3/saliva family, putative/conserved Plasmodium protein, unknown function
PKNH_04_v2	780001	790000	37	76	2.33	1.07	PKNH_0418100 /PKNH_0418200 /PKNH_0418300 /PKNH_0418400	repetitive organellar protein, putative (ROPE)/palmitoyltransferase, putative (DHHC12)/conserved Plasmodium protein, unknown function/octaprenyl pyrophosphate synthase, putative (OPP)

PKNH_04_v2	1010001	1020000	11	155	0.69	2.17	PKNH_0422400 /PKNH_0422500 /PKNH_0422600	GTPase Era, putative (ERA)/conserved Plasmodium protein, unknown function/glucose-6-phosphate isomerase
PKNH_05_v2	110001	120000	29	86	2.10	1.49	PKNH_0502300	rhoptry neck protein 5, putative (RON5)
PKNH_05_v2	230001	240000	15	116	1.09	2.01	PKNH_0505400 /PKNH_0505500 /PKNH_0505600	**50S ribosomal protein L10**, putative/ubiquitin carboxyl-terminal hydrolase 13, putative (USP13)/chromosome associated protein, putative
PKNH_05_v2	570001	580000	28	74	2.03	1.28	PKNH_0512300 /PKNH_0512400 **/PKNH_0512500**	ran binding protein 1, putative/apical merozoite protein, putative (Pk34)/**60S ribosomal protein L7ae/L30e, putative**
PKNH_05_v2	600001	610000	32	62	2.32	1.07	PKNH_0512700 /PKNH_0512800	zinc finger protein, putative/serine/threonine protein kinase RIO2, putative (RIO2)
PKNH_05_v2	710001	720000	14	123	1.02	2.13	PKNH_0514600 /PKNH_0514700 /PKNH_0514800 /PKNH_0514900 /PKNH_0515000 /PKNH_0515100	eukaryotic initiation factor, putative/serpentine receptor, putative (SR12)/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/AP-4 complex subunit sigma, putative/conserved

								Plasmodium protein, unknown function
PKNH_06_v2	220001	230000	45	57	2.66	0.86	PKNH_0605300 /PKNH_0605400 /PKNH_0605500	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_06_v2	570001	580000	34	65	2.01	0.98	PKNH_0612900 /PKNH_0613000	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_07_v2	280001	290000	53	142	2.98	1.99	PKNH_0705200 /PKNH_0705300 /PKNH_0705400	ATP-dependent protease ATPase subunit ClpY, putative (ClpY)/conserved Plasmodium protein, unknown function/translation initiation factor SUI1, putative
PKNH_07_v2	1180001	1190000	42	78	2.36	1.09	PKNH_0726600 /PKNH_0726700 /PKNH_0726800 /PKNH_0726900 /PKNH_0727000	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function

PKNH_08_v2	240001	250000	36	75	2.25	1.17	PKNH_0804700 /PKNH_0804800 **/PKNH_0804900**	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function**/CCR4-NOT transcription complex subunit 5, putative (NOT5)**
PKNH_08_v2	350001	360000	13	164	0.81	2.55	PKNH_0807600	zinc finger C-x8-C-x5-C-x3-H type, putative
PKNH_08_v2	500001	510000	33	100	2.06	1.56	PKNH_0811300 /PKNH_0811400 /**PKNH_0811500**	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/**DNA repair protein RAD23, putative**
PKNH_08_v2	600001	610000	39	83	2.44	1.29	PKNH_0813500 /PKNH_0813600 /PKNH_0813700	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/initiation factor 2 subunit family, putative
PKNH_08_v2	1090001	1100000	35	64	2.19	1	**PKNH_0823500** /PKNH_0823600	**DNA-directed RNA polymerase II subunit RPB1, putative (RPB1)**/KIR protein
PKNH_08_v2	1400001	1410000	35	102	2.19	1.59	PKNH_0830400 /PKNH_0830500 /PKNH_0830600 /PKNH_0830700	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/conserved protein, unknown function/aspartyl protease, putative

PKNH_08_v2	1680001	1690000	44	109	2.75	1.7	PKNH_0837400 /PKNH_0837500 /PKNH_0837600	conserved Plasmodium membrane protein, unknown function/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_09_v2	270001	280000	40	101	2.21	1.45	PKNH_0905300 /PKNH_0905400	pescadillo homolog, putative (PES)/transcription factor with AP2 domain(s), putative (ApiAP2)
PKNH_09_v2	720001	730000	40	58	2.21	0.83	PKNH_0915800 /PKNH_0915900 /PKNH_0916000	AP-1 complex subunit sigma, putative/heat shock protein 90, putative/insulinase, putative
PKNH_09_v2	830001	840000	38	63	2.10	0.9	PKNH_0918700 /PKNH_0918800 /PKNH_0918900 /PKNH_0919000	palmitoyltransferase, putative (DHHC3)/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/tyrosine kinase-like protein, putative (TKL2)
PKNH_09_v2	1440001	1450000	21	142	1.16	2.04	PKNH_0932400	conserved Plasmodium protein, unknown function
PKNH_09_v2	1660001	1670000	38	66	2.10	0.95	PKNH_0937000	WD repeat-containing protein, putative
PKNH_09_v2	1960001	1970000	38	96	2.10	1.38	PKNH_0943000 /PKNH_0943100	coatomer subunit gamma, putative (SEC21)/serine/threonine protein kinase, putative

PKNH_09_v2	2010001	2020000	49	81	2.71	1.16	PKNH_0944300	leucine-rich repeat protein (LRR10)
PKNH_09_v2	2110001	2120000	46	95	2.55	1.36	PKNH_0946100	conserved Plasmodium protein, unknown function
PKNH_09_v2	2130001	2140000	44	34	2.44	0.49	PKNH_0946200	conserved Plasmodium protein, unknown function
PKNH_10_v2	540001	550000	37	74	2.15	1.06	PKNH_1011700 /PKNH_1011800	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_10_v2	1090001	1100000	41	90	2.38	1.29	PKNH_1023800 /PKNH_1023900 /PKNH_1024000	phosphatidylinositol 4-kinase, putative (PI4K)/conserved Plasmodium protein, unknown function/asparagine--tRNA ligase, putative
PKNH_10_v2	1390001	1400000	41	106	2.38	1.52	PKNH_1030600 /PKNH_1030700	merozoite surface protein 3, putative/conserved Plasmodium protein, unknown function
PKNH_10_v2	1410001	1420000	39	74	2.26	1.06	PKNH_1031200 /PKNH_1031300 /PKNH_1031400 /PKNH_1031500	hypothetical protein, conserved in Apicomplexan species/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/merozoite surface protein 8, putative (MSP8)
PKNH_10_v2	1420001	1430000	38	112	2.21	1.6	PKNH_1031500 /PKNH_1031600	merozoite surface protein 8, putative (MSP8)/conserved Plasmodium protein, unknown

							/PKNH_1031700 /PKNH_1031800	function/conserved Plasmodium protein, unknown function/HCNGP-like protein, putative
PKNH_11_v2	300001	310000	35	105	2.30	1.59	PKNH_1107200 /PKNH_1107300 /PKNH_1107400	6-cysteine protein (P92)/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_11_v2	320001	330000	34	80	2.24	1.21	PKNH_1107700 /PKNH_1107800 /PKNH_1107900 /**PKNH_1108000**	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/polyubiquitin binding protein, putative (DOA1)/**50S ribosomal protein L9, mitochondrial, putative**
PKNH_11_v2	460001	470000	18	140	1.19	2.12	PKNH_1110700 /PKNH_1110800	E3 SUMO-protein ligase PIAS, putative (PIAS)/guanylyl cyclase, putative
PKNH_11_v2	500001	510000	35	108	2.30	1.64	PKNH_1111500 /PKNH_1111600	ADP-ribosylation factor, putative/conserved Plasmodium protein, unknown function
PKNH_11_v2	680001	690000	33	62	2.17	0.94	PKNH_1114700 /PKNH_1114800	serine/threonine protein kinase, putative (ARK3)/DNAJ like protein, putative
PKNH_11_v2	1700001	1710000	34	67	2.24	1.02	PKNH_1137000 /PKNH_1137100	conserved Plasmodium protein, unknown

								function/nucleolar GTP-binding protein 1, putative
PKNH_11_v2	1710001	1720000	31	98	2.04	1.49	PKNH_1137200 /PKNH_1137300 /PKNH_1137400 /PKNH_1137500 /PKNH_1137600	6-cysteine protein (P12p)/6-cysteine protein (P12)/PP-loop family protein, putative/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_11_v2	2120001	2130000	42	48	2.77	0.73	PKNH_1144900 /PKNH_1145000	nucleoside diphosphate kinase, putative/cyclin dependent kinase binding protein, putative
PKNH_12_v2	700001	710000	40	124	2.23	1.71	PKNH_1215700 /PKNH_1215800 /PKNH_1215900	dynein beta chain, putative/aspartyl protease, putative/conserved Plasmodium protein, unknown function
PKNH_12_v2	760001	770000	36	77	2.00	1.06	PKNH_1217000 /PKNH_1217200	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_12_v2	920001	930000	44	123	2.45	1.7	PKNH_1219600 /PKNH_1219700 /PKNH_1219800 /PKNH_1219900	conserved Plasmodium protein, unknown function/thioredoxin-like protein (TLP1)/valine--tRNA ligase, putative/conserved Plasmodium protein, unknown function

PKNH_12_v2	1050001	1060000	39	66	2.17	0.91	PKNH_1223800	knowpain-1 (KP1)
PKNH_12_v2	1130001	1140000	49	77	2.73	1.06	PKNH_1225300 /PKNH_1225400 /PKNH_1225500 /PKNH_1225600	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_12_v2	1730001	1740000	43	76	2.39	1.05	PKNH_1239200 /PKNH_1239300 /PKNH_1239400	conserved Plasmodium protein, unknown function/serine/threonine protein kinase, putative (SRPK2)/protein transport protein SEC7, putative (SEC7)
PKNH_12_v2	1750001	1760000	37	63	2.06	0.87	PKNH_1239600 /PKNH_1239700	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_12_v2	1760001	1770000	37	108	2.06	1.49	PKNH_1239700 /PKNH_1239800	conserved Plasmodium protein, unknown function/protein prenyltransferase alpha subunit, putative
PKNH_12_v2	1810001	1820000	36	66	2.00	0.91	PKNH_1241000	serine/threonine protein kinase, putative

PKNH_12_v2	2200001	2210000	42	74	2.34	1.02	PKNH_1248600 /PKNH_1248700 **/PKNH_1248800**	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/**U3 small nucleolar RNA-associated protein 15, putative (UTP15)**
PKNH_12_v2	2230001	2240000	41	63	2.28	0.87	PKNH_1249300 /PKNH_1249400 **/PKNH_1249500**	inner membrane complex protein 1f, putative (IMC1f)/glycerol kinase, putative/**60S ribosomal protein L17, putative**
PKNH_12_v2	2240001	2250000	40	86	2.23	1.19	PKNH_1249500 /PKNH_1249600	**60S ribosomal protein L17**, putative/conserved Plasmodium protein, unknown function
PKNH_12_v2	2630001	2640000	36	95	2.00	1.31	PKNH_1257800 /PKNH_1257900**/ PKNH_1258000**	UDP-N-acetylglucosamine pyrophosphorylase, putative/conserved Plasmodium protein, unknown function/**DNA repair protein RAD5, putative (RAD5)**
PKNH_13_v2	350001	360000	40	95	2.36	1.32	PKNH_1307500 /PKNH_1307600 /PKNH_1307700 /PKNH_1307800 **/PKNH_1307900**	tRNA delta(2)-isopentenylpyrophosphate transferase, putative (MiaA)/blood-stage antigen 41-3, putative/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/**50S ribosomal protein L29, putative**

PKNH_13_v2	400001	410000	34	64	2.01	0.89	PKNH_1308600 /PKNH_1308700	conserved protein, unknown function/zinc finger protein, putative
PKNH_13_v2	1100001	1110000	43	41	2.54	0.57	PKNH_1324500 /PKNH_1324600	conserved Plasmodium protein, unknown function/Plasmodium exported protein (PHIST), unknown function
PKNH_13_v2	1450001	1460000	36	76	2.12	1.06	PKNH_1331100 /**PKNH_1331200** /PKNH_1331300	CorA-like Mg2+ transporter protein, putative/**DNA mismatch repair protein MSH2,** putative/conserved Plasmodium protein, unknown function
PKNH_13_v2	1650001	1660000	49	52	2.89	0.72	PKNH_1336600 /PKNH_1336900 /PKNH_1337000	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_13_v2	1860001	1870000	37	113	2.18	1.57	PKNH_1341400 /PKNH_1341500	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_13_v2	1920001	1930000	36	51	2.12	0.71	PKNH_1342600	conserved Plasmodium protein, unknown function
PKNH_13_v2	2250001	2260000	44	99	2.60	1.38	PKNH_1349600 /PKNH_1349700	conserved Plasmodium protein, unknown function/conserved

								Plasmodium protein, unknown function
PKNH_13_v2	2340001	2350000	20	151	1.18	2.1	PKNH_1351200/PKNH_1351300/PKNH_1351400	radical SAM protein, putative/glideosome associated protein with multiple membrane spans 3, putative (GAPM3)/vacuolar protein sorting-associated protein 29, putative (VPS29)
PKNH_13_v2	2380001	2390000	49	103	2.89	1.43	**PKNH_1352100/PKNH_1352200/PKNH_1352300**	**RNA-binding protein, putative/RNA-binding protein, putative/ribosome biogenesis protein BOP1, putative (BOP1)**
PKNH_13_v2	2400001	2410000	38	71	2.24	0.99	PKNH_1352500/PKNH_1352600/PKNH_1352800	ribonucleotide reductase small subunit, putative/COBW domain-containing protein 1, putative (CBWD1)/conserved Plasmodium protein, unknown function
PKNH_13_v2	2420001	2430000	39	92	2.30	1.28	PKNH_1353100/PKNH_1353200/PKNH_1353300	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_14_v2	740001	750000	40	85	2.25	1.18	PKNH_1416300/**PKNH_1416400**/PKNH_1416500	CDP-diacylglycerol--inositol 3-phosphatidyltransferase (PIS)/**tRNA (adenine(58)-N(1))-methyltransferase catalytic subunit TRM61, putative**

							(GCD14)/transcription factor MYB1, putative (MYB1)
PKNH_14_v2	900001	910000	39	91	2.19	1.26	PKNH_1420200 /PKNH_1420300 /PKNH_1420400	conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function
PKNH_14_v2	2000001	2010000	30	52	1.68	0.72	PKNH_1445800 /PKNH_1445900 /PKNH_1446000	conserved Plasmodium protein, unknown function/RAP protein, putative/proliferating cell nuclear antigen 2, putative (PCNA2)
PKNH_14_v2	2830001	2840000	39	89	2.19	1.23	PKNH_1463300	glucose inhibited division protein a homologue, putative
PKNH_14_v2	3020001	3030000	39	58	2.19	0.8	PKNH_1467900 /PKNH_1468000 /PKNH_1468100 /PKNH_1468200	RNA-binding protein, putative/conserved Plasmodium protein, unknown function/conserved Plasmodium protein, unknown function/bax inhibitor 1, putative
PKNH_14_v2	3200001	3210000	47	75	2.64	1.04	**PKNH_1472300 /**PKNH_1472400	**reticulocyte binding protein (NBPXa)**/tryptophan-rich antigen

Reference

Moon, R. W., Sharaf, H., Hastings, C. H., Ho, Y. S., Nair, M. B., Rchiad, Z., Knuepfer, E., Ramaprasad, A., Mohring, F., Amir, A., Yusuf, N. A., Hall, J., Almond, N., Lau, Y. L., Pain, A., Blackman, M. J., Holder, A. A., 2016. Normocyte-binding protein required for human erythrocyte invasion by the zoonotic malaria parasite *Plasmodium knowlesi*. Proc Natl Acad Sci U S A, 113, 7231–7236.

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Ernest Diez Benavente
Principal Supervisor	Taane Clark & Susana Campino
Thesis Title	**Using whole genome sequence data to study genomic diversity and develop molecular barcodes to profile Plasmodium malaria parasites**

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	PLOS Genetics		
When was the work published?	September 2017		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	

Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I downloaded the raw data from public repositories. I created the pipeline in which I ran the samples through and performed the QC analysis, as well as the interpretation analysis. I later created custom R and perl scripts to process the data in order to obtain information related to coverage, SNPs and other genomic information. The figures presented in this work have all been generated using scripts written by myself or publicly available software adapted for this purpose. I wrote the first draft of the manuscript and circulated to co-authors. Once the comments were received I gathered them and made the relevant changes on the article manuscript. I then performed the submission of the work to the PLOS Genetics journal and once comments from reviewers arrived, I made the corresponding changes and corrections.

Student Signature: ____ **Date:** __________________

Supervisor Signature: ____ **Date:** __________________

Chapter 3
Analysis of nuclear and organellar genomes of *Plasmodium knowlesi* in humans reveals ancient population structure and recent recombination among host-specific subpopulations

PLOS | GENETICS

RESEARCH ARTICLE

Analysis of nuclear and organellar genomes of *Plasmodium knowlesi* in humans reveals ancient population structure and recent recombination among host-specific subpopulations

Ernest Diez Benavente[1], Paola Florez de Sessions[2☯], Robert W. Moon[1☯], Anthony A. Holder[3], Michael J. Blackman[1,3], Cally Roper[1], Christopher J. Drakeley[1], Arnab Pain[4], Colin J. Sutherland[1], Martin L. Hibberd[1,2], Susana Campino[1‡], Taane G. Clark[1,5‡]*

1 Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom, **2** Genome Institute of Singapore, Biopolis, Singapore, **3** The Francis Crick Institute, London, United Kingdom, **4** King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia, **5** Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

☯ These authors contributed equally to this work.
‡ These authors are joint senior authors.
* taane.clark@lshtm.ac.uk

Abstract

The macaque parasite *Plasmodium knowlesi* is a significant concern in Malaysia where cases of human infection are increasing. Parasites infecting humans originate from genetically distinct subpopulations associated with the long-tailed (*Macaca fascicularis (Mf))* or pig-tailed macaques (*Macaca nemestrina (Mn))*. We used a new high-quality reference genome to re-evaluate previously described subpopulations among human and macaque isolates from Malaysian-Borneo and Peninsular-Malaysia. Nuclear genomes were dimorphic, as expected, but new evidence of chromosomal-segment exchanges between subpopulations was found. A large segment on chromosome 8 originating from the *Mn* subpopulation and containing genes encoding proteins expressed in mosquito-borne parasite stages, was found in *Mf* genotypes. By contrast, non-recombining organelle genomes partitioned into 3 deeply branched lineages, unlinked with nuclear genomic dimorphism. Subpopulations which diverged in isolation have re-connected, possibly due to deforestation and disruption of wild macaque habitats. The resulting genomic mosaics reveal traits selected by host-vector-parasite interactions in a setting of ecological transition.

Author summary

Plasmodium knowlesi, a common malaria parasite of long-tailed and pig-tailed macaques, is now recognized as a significant cause of human malaria, accounting for up to 70% of malaria cases in certain areas in Southeast Asia including Malaysian Borneo. Rapid

Citation: Diez Benavente E, Florez de Sessions P, Moon RW, Holder AA, Blackman MJ, Roper C, et al. (2017) Analysis of nuclear and organellar genomes of *Plasmodium knowlesi* in humans reveals ancient population structure and recent recombination among host-specific subpopulations. PLoS Genet 13(9): e1007008. https://doi.org/10.1371/journal.pgen.1007008

Editor: Giorgio Sirugo, Ospedale San Pietro Fatebenefratelli, ITALY

Received: June 16, 2017

Accepted: September 7, 2017

Published: September 18, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: TGC is funded by the Medical Research Council UK (Grant no. MR/K000551/1, MR/M01360X/1, MR/N010469/1, MC_PC_15103). SC and CR are funded by the Medical Research Council UK (Grant no. MR/M01360X/1). RWM is supported by an MRC Career Development Award jointly funded by the UK MRC and UK Department for International Development. This work was supported in part by the Francis Crick Institute that receives its core funding from Cancer Research UK (FC001097, FC001043), the UK Medical Research Council (FC001097, FC001043), and the Wellcome Trust (FC001097, FC001043). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

human population growth, deforestation and encroachment on wild macaque habitats potentially increase contact with humans and drive up the prevalence of human *Plasmodium knowlesi* infections. Appropriate molecular tools and sampling are needed to assist surveillance by malaria control programmes, and to understand the genetics underpinning *Plasmodium knowlesi* transmission and switching of hosts from macaques to humans. We report a comprehensive analysis of the largest assembled set of *Plasmodium knowlesi* genome sequences from Malaysia. It reveals genetic regions that have been recently exchanged between long-tailed and pig-tailed macaques, which contain genes with signals indicative of rapid contemporary ecological change, including deforestation. Additional analyses partition *Plasmodium knowlesi* infections in Borneo into 3 deeply branched lineages of ancient origin, which founded the two divergent populations associated with long-tailed and pig-tailed macaques and a third, highly diverse population, on the Peninsular mainland. Overall, the complex *Plasmodium* parasite evolution observed and likelihood of further host transitions are potential challenges to malaria control in Malaysia.

Introduction

Plasmodium knowlesi, a common malaria parasite of long-tailed *Macaca fascicularis (Mf)* and pig-tailed *M. nemestrina (Mn)* macaques in Southeast Asia, is now recognized as a significant cause of human malaria. A cluster of human *P. knowlesi* cases were reported from Malaysian Borneo in 2004 [1], but now human infections are known to be widespread in Southeast Asia [2,3], and have been reported in travellers from outside the region [2,4]. Clinical symptoms range from asymptomatic carriage to high parasitaemia with severe complications including death [5,6]. As rapid human population growth, deforestation and encroachment on remaining wild macaque habitats potentially increases contact with humans [7], in Southeast Asian countries *P. knowlesi* is now coming to the attention of national malaria control and elimination programmes that have hitherto focused on *P. vivax* and *P. falciparum* [2].

P. knowlesi commonly displays multi-clonality in humans and macaques, and analysis of microsatellite markers, *csp*, *18S rRNA*, and *mtDNA* sequences indicates no systematic differences between human and macaque isolates from Malaysian Borneo [8]. Whole genome-level genetic diversity among *P. knowlesi* from human infections in Sarikei in Sarawak demonstrates substantial dimorphism extending over at least 50% of the genome [9]. This finding is supported by analysis of microsatellite diversity in parasites from *Mf*, *Mn* and human infections across Peninsular and Borneo Malaysia [10]. It also provides evidence that the two distinct genome dimorphs reflect adaptation to either of the two host macaque species, although no evidence of a complete barrier in primate host susceptibility was found [10]. A third genome cluster has been described from geographically distinct Peninsular Malaysia [11, 12, 13, 14].

Studies of *mtDNA* have revealed that ancestral *P. knowlesi* predates the settlement of *Homo sapiens* in Southeast Asia, the evolutionary emergence of *P. falciparum* and *P. vivax*, and underwent population expansion 30–40 thousand years ago [8]. Diversity at the genomic level is thus likely to reflect host- and geography-related partitioning during this expansion, as well as additional recent complexity due to contemporary changes in host and vector distributions during ongoing ecological transition in the region [15]. Several *Anopheles* species, all from the Leuchosphyrus group, are capable of transmitting *P. knowlesi* malaria, including *A. latens* and *A. balbacensis* in Malaysian Borneo [16, 17, 18], *A. hackeri* and *A. cracens* in Peninsular Malaysia [19] and *A. dirus* in southern Vietnam [20]. It is thus likely that patterns of genome diversity in natural populations of *P. knowlesi* reflect partitioning among both Dipteran and primate hosts occurring on varying time-scales through the evolutionary history of the species. Such partitioning can plausibly prevent or reduce panmictic genetic exchange.

Genomic studies of *P. knowlesi* to date have considered nuclear gene diversity and dimorphism among naturally-infected human hosts, and macaque-derived laboratory-maintained isolates from the 1960s [10, 12]. However, these studies did not consider non-nuclear organellar genomes in the mitochondrion and apicoplast of malaria parasites, which are non-recombinant and uniparentally inherited, and can provide evidence of genome evolution on a longer timescale [21]**.** Recombination barriers among insect and primate hosts may have less impact on sequence diversity in the organellar genomes of *P. knowlesi*. Utilising a new *P. knowlesi* reference genome generated using long-read technology [22] we performed a new analysis of all available nuclear and non-nuclear genome sequences. Patterns of polymorphisms were analysed to identify evolutionary signals of both recent and ancient events associated with the partitioning of the di- or tri-morphic genomes previously reported.

Results

Sequence data reveals multiplicity of infection

Raw short-read sequence data from all available *P. knowlesi* isolates (**S1 Fig**) were mapped to a new reference genome [22] from the human-adapted *P. knowlesi* line A1-H.1 genome [23], yielding an average coverage of ~120-fold across 99% of the reference genome (**S1 Table**), and 1,632,024 high quality SNPs. The high density of point mutations (1 every 15bp) in *P. knowlesi* compared to *P. vivax* and *P. falciparum* has been previously noted [10]**.** Seven macaque-derived isolates were found to have high multiplicity of infection (S2 Fig), and were excluded, leaving an analysis set of 60 isolates.

Population structure analysis reveals new natural genetic exchange

SNP-based neighbour-joining tree analysis revealed three subpopulation groups that coincide with isolates presenting the *Mf*-associated *P. knowlesi* genotype (*Mf-Pk*, Borneo Malaysia, Cluster 1), the *Mn*-associated *P. knowlesi* genotype (*Mn-Pk*, Borneo Malaysia, Cluster 2) [**10, 11, 12, 14**], and older Peninsular Malaysia strains (Cluster 3) (Fig 1A). Within Cluster 1 we observed two geographic sub-groups that coincide with Kapit and Betong regions in Malaysian Borneo. The samples from Sarikei region (DIM prefix), geographically located equidistant between Kapit and Betong, fall into either cluster (S3 Fig). Overall, the regional clusters from Kapit and Betong were more genetically similar to each other (mean fixation index F_{ST} 0.03, **S4 Fig**) than were the host-associated clusters (Cluster 1 vs. 2, mean F_{ST} 0.21). However, a significant chromosomal anomaly was identified that differentiated the Kapit and Betong *Mf-Pk* subgroups; this occurred in a multi-gene region on chromosome 8 (~500 SNPs with F_{ST} values >0.4; **Fig 1B; S4 Fig**).

Signatures of introgression events in chromosome 8

To explore the anomaly in chromosome 8, individual haplotypes and neighbour-joining trees were constructed across several loci (Fig 1C **and** Fig 1D) revealing two very distinct patterns. The first pattern was observed in the chromosomal sections with low genetic diversity between the two *Mf-Pk* regional clusters ($F_{ST} < 0.2$, **Fig 1B**). The tree structure for these genomic regions (**Fig 1D**, 1st tree) mimics that of the genome-wide tree in **Fig 1A.** Strong haplotype differentiation between the host-associated Clusters 1 (*Mf-Pk)* and 2 *(Mn-Pk)* was confirmed in the SNP-based profiles (**Fig 1C**, 1st column).

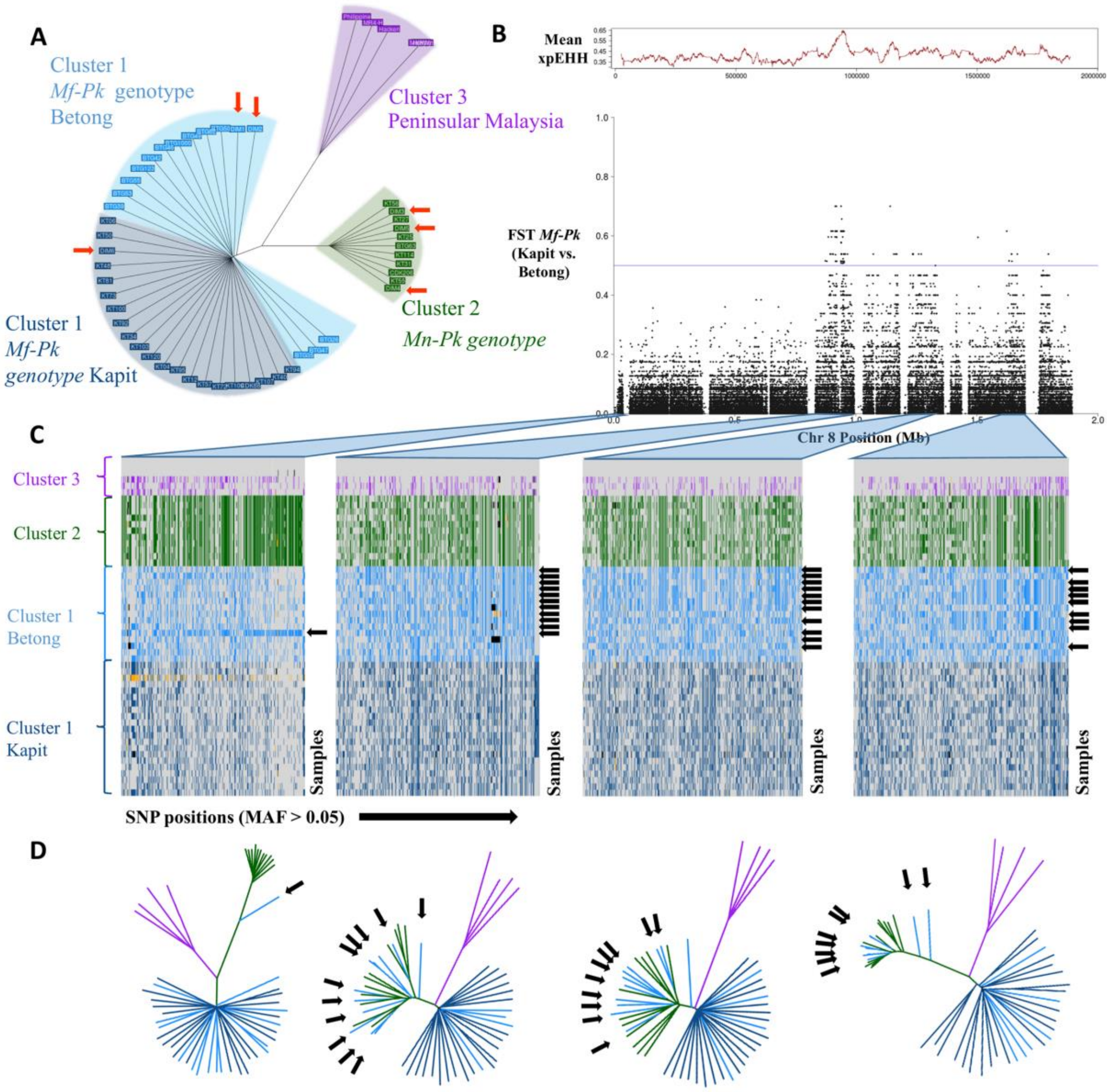


Fig 1. Whole genome population structure and evidence of genetic exchange in chromosome 8. A) Neighbour joining tree constructed using 1,632,024 genome-wide SNPs across the 60 *P. knowlesi (Pk)* samples. The tree shows two levels of resolution involved in the clustering of genotypes. The first level differentiates Peninsular Malaysia samples (Cluster 3, purple) from the Malaysian-Borneo host-related *Pk* genotypes (Cluster 1, *M. fascicularis* macaques (*Mf-Pk*), blue; Cluster 2, *M. nemestrina* macaques (*Mn-Pk*), green). The second level differentiates within Cluster 1, where *Mf-Pk* genotypes fall in subgroups from Betong (light blue) and from Kapit (dark blue). Samples from Sarikei have been highlighted using orange arrows. **B)** Allele frequency differences between Betong and Kapit regional subgroups of the *Mf-Pk* genotype in chromosome 8 SNPs using the population differentiation measure $F_{ST.}$ There is high differentiation (F_{ST} > 0.4) in several regions across chromosome 8 (0.85-1Mb, 1.2Mb-1.35Mb, and 1.6–1.7Mb), and these signals overlap with strong evidence of recent positive selection, measured by the average XP-EHH calculated in 1kbp windows (red trace above). **C)** Haplotype plots for all samples (y-axis) at common SNP positions (MAF >5%, x-axis) highlighting the regions with abnormally high F_{ST} values (0.85-1Mb, 1.2Mb-1.35Mb, and 1.6–1.7Mb), as well as the low *Fst* region spanning from 0.1 to 0.2Mb for comparison. The black arrows indicate samples with the *Mf-Pk* genotype from Betong

present with a *Mn-Pk* Cluster 2-like haplotype. These patterns are indicative of genetic exchange between the *Mf-Pk* and *Mn-Pk* genotype clusters, which is supported by the neighbour joining trees included in **D**). Missing calls are coloured in black and mixed calls are coloured in yellow. **D)** Neighbour joining trees constructed using SNPs in each of the regions in **C**). The trees show clear clustering of *Mf-Pk* Betong samples with the *Mn-Pk* genotype cluster in the genetic regions of abnormal F_{ST} (2nd, 3rd and 4th trees) compared to the 1st tree where only sample DIM2 presents introgression.

https://doi.org/10.1371/journal.pgen.1007008.g001

A second pattern was observed in regions of chromosome 8 with distinct genetic differentiation between Kapit and Betong subgroups ($F_{ST} > 0.4$). Many *Mf-Pk* Betong subgroup isolates presented segments almost identical to chromosome 8 sequences of the *Mn-Pk* genotype from Cluster 2 (Fig 1D**,** 2nd, 3rd and 4th trees). This exchange is supported by the SNP-based haplotype patterns, where a distinct haplotype in the Betong samples is Cluster 2-like (Fig 1C, 2nd, 3rd and 4th columns, black arrows), suggesting the introgression of large chromosomal regions (up to 200Kb) between *Mf-Pk* (Cluster 1) and *Mn-Pk* (Cluster 2). This is consistent with a very recent event of natural genetic exchange between these subgroups of *P. knowlesi* recently isolated from human infections. The high frequency of the new haplotype (73%) in the Betong subgroup suggests that it is under (recent) strong selection pressure in this region. The presence of differences in extended haplotype homozygosity between the recombinant and non-recombinant regional *Mf-Pk* subpopulations provides additional evidence of recent positive selection (XP-EHH peak, $P<0.0001$) in a region of increased population differentiation ($F_{ST} >$ 0.4, Fig 1B).

The functional nature of genes in chromosome 8 involved in these putative introgression events was investigated ($F_{ST} > 0.4$, Table 1), and found to include loci that are important in the vector component of the *Plasmodium* life cycle. For example, *cap380* (*PKNH_0820800*, 101 SNPs with $F_{ST} > 0.4$) encodes a protein expressed in the external capsule of the oocyst. This gene is essential in the maturation from ookinete into oocyst in *P. berghei*, and is assumed to assist in evasion of mosquito immune mechanisms [24]**.** Another gene, *PKNH_0826900* (19 SNPs) encodes for the circumsporozoite- and TRAP-related protein (CTRP), which has an established role in ookinete motility *in P. berghei* and is essential for binding to and invading the mosquito midgut [25]**.** Further, homologues of *PKNH_0826400* (21 SNPs) display increased transcription levels in ookinete and gametocyte V sexual stages in both *P. falciparum* [26] and *P. berghei* [27] compared to the asexual ring stage (fold change of at least 2). The transcriptomic profiles of these strongly selected genes are shown in S5 Fig.

Genome-wide evidence of genetic exchange events in *P. knowlesi*

By applying a combination of neighbour joining trees and SNP diversity analysis across 50 Kbp windows, we identified that 33/60 isolates show clear evidence of genetic exchange between Clusters 1 and 2 (S2 Table). Regions involved in exchange (recombination) (137/494 regions, 86% contained an ookinete related gene) showed evidence of enrichment for ookinete-expressed genes compared to other (non-recombinant) chromosome regions (357/494 regions, 77% contained an ookinete related gene) (Chi Square P = 0.03). One such region in chromosome 12 included the *Pf47-like (PKH_120710)* gene, where the orthologue in *P. falciparum* is a known mediator of the evasion of the mosquito immune system [28]. Furthermore, it has been shown that a change in haplotype in this gene in a *P. falciparum* isolate is sufficient to make it compatible to a different mosquito species [28]. Nearly half (45%) of isolates from Betong presented with a recombinant profile in *PKH_120710*.

In general, the genetic exchanges generated differing levels of mosaicism in each population and among individual isolates across all chromosomes (S6 Fig). One isolate from Sarikei with the *Mf* genome dimorph type (DIM2) appeared to harbour *Mn*-type introgressed sequences in 8% of the genome, occurring across 6 chromosomes (6, 7, 8, 9, 11 and 12), including an almost

Table 1. Genes located within the chromosome 8 regions of genetic exchange and transcriptional changes.

No. SNPs*	Gene name *(PKNH_)*	Product	*P.falcip* Ortholog *(PF3D7_)*	*P. berghei ortholog (PBANKA_)*	*P.falcip* (Ring vs. Ookinete)*	*P.berghei* (Ring vs. Ookinete)**
218	Non-genic	-	-	-	-	-
101	***0820800***	**oocyst capsule protein (Cap380)**	***0320400***	***1218100***	**Yes**	**Yes**
21	***0826400***	**conserved Plasmodium membrane protein**	***0315700***	***0413500***	**Yes**	**Yes**
19	***0826900***	**circumsporozoite- & TRAP-related (CTRP)**	***0315200***	***0412900***	**Yes**	**Yes**
15	*0819600*	N-acetylglucosaminephosphotransferase	*0321200*	*1217300*	Yes	No
11	*0820300*	nicotinamidase	*0320500*	*1218000*	Yes	No
9	*0828500*	conserved Plasmodium protein	*0313700*	*0411400*	Yes	No
7	*0819500*	conserved Plasmodium protein	*0321300*	*1217200*	No	No
7	*0820200*	conserved Plasmodium protein	*0320600*	*1217900*	Yes	No
7	*0828400*	conserved Plasmodium protein	*0313800*	*0411500*	Yes	No
7	*0837200*	conserved Plasmodium protein	*0305500*	*0404000*	No	No
6	*0822900*	conserved Plasmodium protein	*0318500*	*0806700*	Yes	No
5	*0836500*	activator of Hsp90 ATPase (AHA1)	*0306200*	*0404600*	Yes	No
5	*0839000*	inner membrane complex protein 1e	*0304100*	*0402700*	Yes	No
4	*0819700*	conserved Plasmodium protein	*0321100*	*1217400*	Yes	No
4	*0820100*	signal peptidase complex subunit 2 (SPC2)	*0320700*	*1217800*	Yes	No
4	*0823100*	conserved Plasmodium protein	*0318300*	*0806900*	No	No
4	*0836000*	membrane magnesium transporter	*0306700*	*0405100*	No	No
4	*0839100*	inner membrane complex protein 1a	*0304000*	*0402600*	No	Yes
3	*0820900*	T-complex protein 1 subunit epsilon (CCT5)	*0320300*	*1218200*	No	No
3	*0828300*	conserved Plasmodium protein,	*0313900*	*0411600*	Yes	No
3	*0838400*	conserved Plasmodium protein,	-	-	No	No
2	*0821000*	CPW-WPC family protein	*0320200*	*1218300*	Yes	No
2	*0821500*	ABC transporter I family member 1 (ABCI3)	*0319700*	*1218800*	No	No
2	*0822800*	cleavage and polyadenylation factor	*0318600*	*0806600*	No	No
2	*0824500*	conserved Plasmodium protein	*0317300*	*0807900*	No	No
2	*0825900*	conserved Plasmodium protein	*0316200*	*0414000*	Yes	Yes
2	*0827400*	zinc finger protein	*0314700*	*0412400*	Yes	Yes
2	*0828700*	conserved Plasmodium protein	*0313500*	*0411200*	Yes	No
2	*0837600*	conserved Plasmodium protein	*0305100*	*0403600*	Yes	No
2	*0838500*	circumsporozoite (CS) protein (CSP)	*0304600*	*0403200*	No	Yes
2	*0838800*	conserved Plasmodium protein	*0304300*	*0402900*	No	Yes
2	*0839200*	phosphatidylethanolamine-binding protein	*0303900*	*0402500*	Yes	No
1	*0819100*	conserved Plasmodium protein	*0321700*	*1216800*	Yes	No
1	*0819200*	ATP-dependent RNA helicase	*0321600*	*1216900*	Yes	No
1	*0819400*	protein kinase, putative	*0321400*	*1217100*	Yes	No
1	*0819900*	histone H2A variant, putative (H2A.Z)	*0320900*	*1217600*	No	No
1	*0820000*	ATP-dependent RNA helicase DDX6 (DOZI)	*0320800*	*1217700*	Yes	No
1	*0821200*	protein phosphatase inhibitor 2	*0320000*	*1218500*	Yes	No
1	*0821400*	conserved Plasmodium protein	*0319800*	*1218700*	No	Yes
1	*0821600*	elongation factor 1	-	*1218900*	No	No

(*Continued*)

Table 1. (*Continued*)

No. SNPs*	Gene name (PKNH_)	Product	*P.falcip* Ortholog (PF3D7_)	*P. berghei ortholog* (PBANKA_)	*P.falcip* (Ring vs. Ookinete)*	*P.berghei* (Ring vs. Ookinete)**
1	*0821700*	RNA-binding protein	*0319500*	*0805800*	No	No
1	*0823000*	conserved Plasmodium protein	*0318400*	*0806800*	Yes	Yes
1	*0823500*	DNA-directed RNA polymerase II subunit	*1329000*	*0807000*	No	No
1	*0824700*	6-cysteine protein	*0317100*	*0808100*	Yes	Yes
1	*0824900*	E3 ubiquitin-protein ligase, putative	*0316900*	*0808300*	No	No
1	*0825200*	formate-nitrite transporter, putative (FNT)	*0316600*	*0414400*	No	No
1	*0826700*	conserved Plasmodium protein	*0315400*	*0413200*	Yes	No
1	*0827000*	eukaryotic translation initiation factor 4E	*0315100*	*0412800*	No	No
1	*0827300*	conserved Plasmodium protein,	*0314800*	*0412500*	No	Yes
1	*0836300*	FAD-dependent glycerol-3-phosphate	*0306400*	*0404800*	Yes	No
1	*0836600*	conserved Plasmodium protein	*0306100*	*0404500*	No	No
1	*0838900*	EH (Eps15 homology) protein	*0304200*	*0402800*	Yes	No
1	*0839300*	IBR domain protein	*0303800*	*0402400*	No	No

* Cells in green (with "Yes") imply that the *P. falciparum* orthologue (Column 4) of the *P. knowlesi* gene (Column 2) has at least a two-fold change difference in the transcriptional signals from *P. falciparum* when comparing Ring vs. Ookinete stages

** similar to *, but refers to the *P. berghei* orthologue (Column 5) having at least a two-fold change difference in the transcriptional signals; **bolded genes** have >10 SNPs with $F_{ST} > 0.4$.

https://doi.org/10.1371/journal.pgen.1007008.t001

complete *Mn*-type chromosome 8. Of the 33 samples with evidence of exchanges, 13 were from the Betong region, 14 from Kapit and 6 from Sarikei, which indicates that the events are not geographically restricted. Although, the majority of genetic exchange events involve the integration of *Mn*-type motifs into *Mf*-type genomes, introgression in the opposite direction was also observed, but on a smaller scale and at lower frequency.

Organellar genomes also reflect genetic exchange events

The mitochondrial and apicoplast genomes of each *P. knowlesi* isolate was interrogated for signals of evolutionary history over longer time-scales, as in previous studies [21, 29, 30]. Combining the mitochondrial sequence data from the 60 *P. knowlesi* isolates from this study together with 54 previously published mitochondrial sequences including human and both *Mn* and *Mf* samples [9], we generated a phylogenetic tree (Fig 2). This tree shows four clades (shown in purple, red, blue and green). To interpret these clades, they were cross-referenced to the previously defined 3 nuclear genotypes (Clusters 1 to 3) and the host contributing the sample (human, macaque-type). The red and purple clades possess similar mitochondrial haplotypes as highlighted by their inter-cluster average F_{ST} (red vs. purple: average $F_{ST} = 0.16$), which is lower than comparisons including the other two clusters (red or purple vs. blue or green: average $F_{ST} > 0.18$). The purple clade consists of cultured isolates from Peninsular Malaysia, and is associated with the Peninsular nuclear genotype (Cluster 3). The red and green clades each contain a mixture of Borneo Malaysia samples from both humans and macaques with nuclear genotypes from Clusters 1 and 2. The green clade also includes the only sequence sourced from a *M. nemestrina* host. The blue clade contains samples from humans and macaques, all with Cluster 1 nuclear genotypes. The divergence of these mitochondrial clades from their common ancestor was estimated to be 72k years ago, and younger than the previous the estimate of 257k but within error [8]. Furthermore, the presence of monkey-derived sequences spread across the tree seems to indicate that none of the mitochondrial

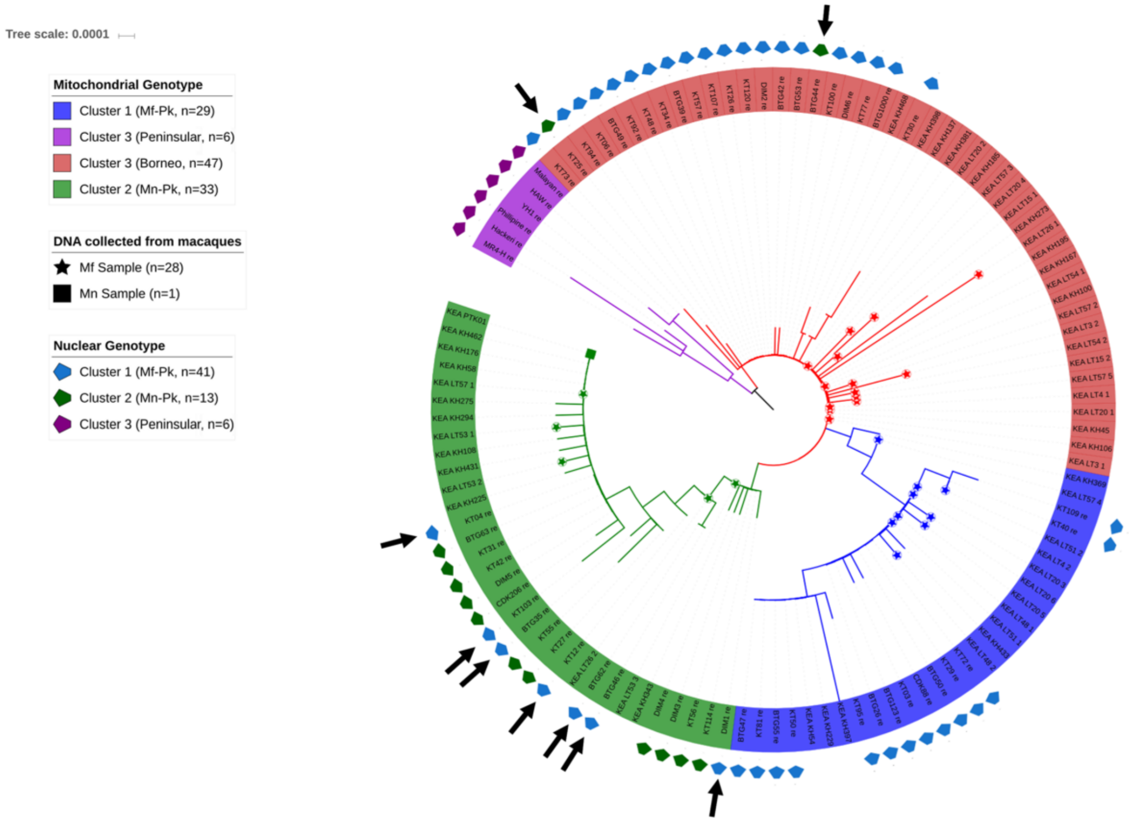


Fig 2. Phylogenetic tree constructed from *P. knowlesi* mitochondrial sequences for the 60 whole genome sequenced samples and 54 published others [6] sourced from human, *M. nemestrina (Mn)* and *M. fascicularis (Mf)* samples. The mitochondrial genotype groups defined here are cross-referenced to the nuclear genotypes in Fig 1A (pentagons in the outer ring, missing pentagons relate to the 54 samples with only mitochondrial sequence data [6]). Samples sourced from the different macaques are highlighted in the tree branches. The tree shows three main subpopulations: (i) two clades including Peninsular Malaysia (Peninsular nuclear genotype, Cluster 3, purple) and Borneo Malaysia (mix of *Mf-Pk* and *Mn-Pk* nuclear genotypes, Cluster 1 and 2, red) presenting a very similar mitochondrial haplotype; (ii) the majority of the samples with a *Mn-Pk* nuclear genotype together with the only sequence obtained from a *Mn* sample (Cluster 2, green); (iii) samples with a *Mf-Pk* nuclear genotype (Cluster 1, blue). These clusters are consistent with microsatellite-based trees [12]. The presence of monkey samples spread throughout the tree indicates that none of the mitochondrial genotypes groups are human-specific, consistent with microsatellite-based analysis [9]. Black arrows indicate the presence of samples with mismatched nuclear and mitochondrial subtypes.

https://doi.org/10.1371/journal.pgen.1007008.g002

genotypic groups found is human-specific as all have also been observed in macaques, also consistent with previous findings [9].

Using the common SNPs (280/425 with MAF > 5%: apicoplast 252, mitochondria 28 SNPs) in the 60 isolates with the sequence data we confirmed that the organellar genomes are co-inherited (mean pairwise organellar linkage disequilibrium D' = 0.99). SNP-based haplotype profile analysis (S7A Fig) revealed clustering that is consistent with the three main clusters seen in Fig 2. Similarly, a phylogenetic tree constructed using only apicoplast SNPs (S7B Fig) is congruent with the mitochondrial based tree (Fig 2). The presence of mismatched nuclear and organellar type genomes in two of the three clusters (**black arrows in Fig 2**) and the presence of such mismatched samples with little or no evidence of nuclear genome recombination suggests ancient genetic exchange events between distinct lineages. The nuclear footprints of such exchanges are likely to have been broken down by recombination over time. We observed a significant incongruence between the robust phylogenetic tree topologies based on organellar and nuclear genome SNPs (Shimodaira-Hasegawa test P = 0.001; Templeton test P = 0.003) (Fig 2). These results from organellar and nuclear genomes, in a small but geographically diverse set of *P. knowlesi*, indicate that there have been several genetic exchanges between the host-associated clusters in Malaysian Borneo.

Discussion

P. knowlesi is now the major cause of malaria in Malaysian Borneo, but the biology of the parasite [15, 22, 23], host and vector interactions, and disease distribution and epidemiology [19, 31, 32] are not well understood. The availability of a new high-quality reference sequence and a more robust approach to MOI were used to re-evaluate the previously described peninsular and macaque-associated subpopulations of *P. knowlesi* parasites. We report two major new findings. First, clear evidence of natural genetic exchanges between the divergent *Mf*- and *Mn*-associated subpopulations of *P. knowlesi*, including a major segment of introgression on chromosome 8, is presented. Second, the presence of haplotype sub-divisions in the organellar genomes that do not map onto the subpopulations implied by nuclear genome analysis indicate that exchange events have previously occurred in non-recent history. A similar multi-tiered pattern of evolution among nuclear and organellar genomes has been found in *Trypanosoma cruzi*, an unrelated protozoan parasite with a mammalian host-insect vector life cycle [29, 30].

Unexpectedly, observed mosaicism and population differentiation signals were not encountered equally across the *P. knowlesi* nuclear genome, but were particularly prominent on chromosome 8, with genes expressed in mosquito stages over-represented. For example, the majority (73%) of *Mf*-associated isolates from Betong harboured the *Mn*-associated allele of the oocyst-expressed *cap380* gene, which differs at 101 positions from the allele found in the *Mf*-associated cluster. This is essential for ookinete to oocyst maturation and therefore for the transmission of the parasite during the vector stage [24, 25]; here, we identify signals of recent selective pressure on this locus (Fig 1B). Other vector-related genes were identified within the introgressed segment, and point towards strong evolutionary selection pressure on the parasites driven by the transmitting *Anopheles* vector species. Such effects have been found in *P. falciparum* [28] and *P. vivax* genomes [33], and highlight the importance of understanding the distribution of the different *Anopheles* vector species, their host feeding preferences, and their interactions with the parasite in highly dynamic and complex environments such as the ecological niche of *P. knowlesi*.

Nearly 80% of Malaysian Borneo has undergone deforestation or agricultural expansion, which have driven habitat modification affecting both macaque and *Anopheles* host species, and the proximity of humans to both [8, 31]. Furthermore, studies have predicted that *Mn* predominantly inhabits forested areas while *Mf* reside in more cosmopolitan areas, which include croplands, vegetation mosaics, rubber plantations and forested areas [8, 34]. The main genomic exchange event on chromosome 8 involves essential vector-related genes and is pin-pointed geographically to the Betong area. This region has undergone significant forest degradation due to expansion of industrial plantations in the recent years [35]. These types of environmental changes have been previously related to alterations in the vector species distribution in Malaysia, leading to malaria epidemics [36]. Environmental changes also affect macaque habitats, and increase the opportunities for human-macaque interaction [31], but selection events highlighted in this study seem to primarily reflect adaptation of the parasite to changes in mosquito distribution or to recent changes in the vectorial capacity of the existing vectors. The depth, breadth and spread of the genetic exchanges observed in three different areas (Betong, Kapit and Sarikei) in Sarawak highlight the potential importance of these events for parasite adaptation in both vertebrate and invertebrate species.

Although, the level of genetic diversity between *Mf*- and *Mn*-associated *P. knowlesi* has some similarity to that observed between *P. ovale curtisi* and *P. o. wallikeri*, now considered separate species [37], the evidence of recombination and genetic exchanges observed in this study precludes species designation, as reproductive isolation is not complete. Nevertheless, better understanding of *P. knowlesi* population structure could aid future studies across the regions where human populations have been identified at risk of infection including both symptomatic and asymptomatic cases [4, 38, 39]. This would assist with characterising and tracking subpopulations and genetic exchanges, and provide a flexible framework for better understanding *P. knowlesi* diversity across the region.

Our work has provided insight into *Plasmodium* parasite evolution. It has been suggested that malaria parasites have survived using either adaptive radiation where host switching plays a key role [40], or alternatively adaptation to complex historical and geographical environments leading to speciation [41]. *Plasmodium* species in non-human natural conditions in the absence of drug selection pressure have a wide range of possible hosts [41, 42]. The *P. knowlesi* data has shown that geographical or ecological isolation of the different hosts over an extended time can generate subgroups of parasites with substantial genetic differentiation, but capable of recombining when in contact [12, 30, 31]. This pattern has a major impact on the parasite genome, as illustrated by the profound chromosome mosaicism observed among our study isolates. Our data suggest that the broad host specificity of some of the *Plasmodium* species are important drivers of parasite genomic diversity. In *P. knowlesi* this means that genetic divergence is enabled not only by long-term geographic isolation, as is the case between Peninsular and Bornean isolates, but also via the isolation afforded by extended transmission cycles within different primate hosts. The genetic trimorphism suggests that the separate macaque hosts provides sufficient genetic isolation to allow for host specific adaptations to occur, even within relatively small geographic areas. Furthermore, the possibility of recombination between partially differentiated parasite genomes increases opportunities for new adaptation, including further host transitions, and can only make malaria control more difficult. Genome-level studies on *P. knowlesi* isolates from *Mf* and *Mn* across the parasite's geographic range are now needed to test the generalizability of this remarkable conclusion.

Materials and methods

P. knowlesi sequence data

Raw sequence data were downloaded for 48 isolates from Kapit and Betong in Malaysian Borneo [11], 6 isolates from Sairikei in Malaysian Borneo (S1 Fig) [9] and 6 long-time isolated lines, maintained in rhesus monkeys sourced originally from Peninsular Malaysia and Philippines [11]. The sequence data accession numbers can be found in S1 Table. The samples were aligned against the new reference for the human-adapted line A1-H.1 (pathogenseq.lshtm.ac.uk/knowlesi_1, accession number ERZ389239, [22]) using *bwa-mem* [43] and SNPs were called using the *Samtools* suite [44], and filtered for high quality SNPs using previously described methods [45, 46]. In particular, the SNP calling pipeline generated a total of 2,020,452 SNP positions, which were reduced to 1,632,024 high quality SNPs after removing those in non-unique regions, and in low quality and coverage positions. Samples were individually assessed for detecting multiplicity of infection (MOI) using: (i) *estMOI* [47] software, and (ii) quantifying the number of positions with mixed genotypes (if more than one allele at a specific position have been found in at least 20% of the reads [46]). The measures led to correlated results ($r^2 = 0.8$), which highlighted the robustness of these two methods. Samples were classified into three subcategories: (i) single infections (> = 98% genome showing no evidence of MOI and < = 1/10,000 SNP positions with mixed genotypes), (ii) low MOI (>85% genome showing no evidence of MOI and < = 4/10,000 SNPs positions with mixed genotypes); (iii) high MOI (<85% genome showing no evidence of MOI, and > 4/10,000 SNPs positions with mixed genotypes). Samples with high MOI were removed from subsequent analyses.

Population genetics analysis

For comparisons between populations, we first applied the principal component analysis (PCA) and neighbourhood joining tree clustering based on a matrix of pairwise identity by state values calculated from the SNPs. We used the ranked F_{ST} statistics to identify the informative polymorphism driving the clustering observed in the PCA [48]**.** Finally, we created haplotype plots using only SNP positions with MAF > 0.05 over all the populations, and displayed each sample as a row to allow closer inspection of the chromosome regions where interesting recombination events are observed. The XP-EHH metric [49] implemented within the *rehh* R package was used to assess evidence of recent relative positive selection between regional clusters from Kapit and Betong. The results were smoothed by calculating means in 1 Kbp windows, where windows overlapped by 250bp. The *raXML* software (v.8.0.3, 1000 bootstrap samples) was used to construct robust phylogenetic trees (90% bootstrap values > 95) for nuclear and organellar SNPs. Estimates of divergence times for subpopulations was based on a Bayesian Markov Chain Monte Carlo (MCMC) (BEAST, v.1.8.1) approach applied to mitochondrial sequences, with identical parameters settings to those described elsewhere [8]**.** The Shimodaira-Hasegawa [50] and the Templeton [51] tests were used to detect incongruence between the tree topologies.

Identification of introgressed regions in the different chromosomes

In order to identify regions that have undergone introgression we calculated the pairwise SNP diversity (π) of each sample against all the Borneo samples using a 50 Kbp sliding window. This window size was sufficient to include the required number of SNPs for the robust identification of introgression events. The average π in the *M. nemestrina* associated (*Mn-Pk*) and *M. fascicularis* associated (*Mf-Pk*) clusters was calculated, leading to two diversity values for each sample (Mf_π and Mn_π) and thereby a measure of genetic distance to the average of the two clusters. For *Mf* samples, an increase in the Mf_π and a decrease in Mn_π would mean the sample is more similar to the *Mn-Pk* cluster than the average; vice versa for the *Mf* samples. In order to avoid the identification of spurious events, we applied a threshold of a 0.001 increase in the deviation from the original cluster.

Characterization of genes under strong selection after recombination

For *P. knowlesi* genes of interest, orthologues in *P. falciparum* and *P. berghei* genomes were identified using *PlasmoDB* (plasmodb.org). Gene expression data (including from the RNAseq platform) for these genes across different stages of the life cycle of the parasite were considered [26, 27]. In particular, we compared the average of the asexual blood stages and the sexual ookinete stage, highlighting the genes upregulated with a two-fold change (P<0.000001), for *P. falciparum* [26] and *P. berghei* [27].

Supporting information

S1 Table. Study samples. * Multiplicity of infection (MOI) is % of genome presenting multiplicity of infection; **Group established by whole Genome PCA: *Mf M. fascicularis*, *Mn M. nemestrina*, *Penin. Peninsular;* Rh mac Rhesus macaque, *** evidence of genetic exchange (ExΔ)
(DOCX)

S2 Table. 50 Kb regions in the *P. knowlesi* genome that present genetic exchanges in the full set of samples.
(XLSX)

S1 Fig. Geographical source of the *P. knowlesi* isolates: Betong (n = 14), Kapit (n = 33) and Sarikei (n = 6).
(TIFF)

S2 Fig. Evaluation of multiplicity of infection (MOI) using mixed genotype calls (x-axis) and the estMOI read-pair haplotype counting approach [45] (y-axis) reveals seven highly non-clonal samples.
(TIFF)

S3 Fig. Principal components analysis of the *M. fascicularis P. knowlesi* genotype group (*Mf-Pk*, Cluster 1) confirms that the subgroups from Kapit and Betong are separated. The *Mf-Pk* Sarikei samples (DIM code in orange) cluster with either one of the two groups, which is consistent with the geographic location of Sarikei as an equidistant region between Kapit and Betong. There is increased diversity of Betong samples compared to the Kapit samples.
(TIFF)

S4 Fig. Genome-wide differences in allele frequencies (measured using the fixation index (F_{ST})) between *M. fascicularis P. knowlesi* genotype groups (*Mf-Pk)* from Kapit and Betong. The comparison shows clear abnormalities in several genomic regions in chromosome 8 shown to be a result of genetic exchange with the *Mn-Pk* genotype.
(TIFF)

S5 Fig. Transcriptomic profiles for the orthologues of the introgressed genes under selection pressure. The transcriptomic profiles of the orthologues in *P. falciparum* [26] and *P. berghei* [27] for the three genes found to be under strong selection pressure were extracted from PlasmoDB (http://plasmodb.org/plasmo/), including the percentile and the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) plots. These included data for 5 *P. berghei* stages (4-hour Ring, 16-hour Trophozoite, 22-hour Schizont, Gametocyte and Ookinete) and 7 *P. falciparum* stages (Ring, early Trophozoite, late Trophozoite, Schizont, Gametocyte stage II, Gametocyte stage V and Ookinete), and showing a clear increased expression in mosquito related stages, particularly the ookinete stage.
(TIFF)

S6 Fig. Genome distribution of introgression events for each chromosome estimated using SNP diversity in 50Kb sliding windows. (Left panel) location of introgressions from *M. nemestrina P. knowlesi* (*Mn-Pk*) genotype into *M. fasciscularis P. knowlesi* (*Mf-Pk*) genotypes, a dashed shaded region has been added where at least 1 gene related with the ookinete life stage of the parasite has been identified based on gene expression for the orthologue genes in *P. berghei* and/or *P. falciparum.* **(Right panel)** location of introgressions from *Mf-Pk* genotype into *Mn-Pk* genotypes.
(TIFF)

S7 Fig. Analysis of organellar mitochondria (MIT) and apicoplast (Api) SNPs confirms clustering into three core haplotype groups a) Haplotype plot for the 36 samples with sufficient coverage across the organellar genomes. Three clearly defined clusters are present. The first cluster represents the mitochondrial genotype found in the Peninsular strains (purple, n = 5) and a set of 10 samples with a highly related haplotype with the smallest inter-cluster average F_{ST} (average F_{ST} = 0.16) from Borneo Malaysia (represented in red in Fig 2**).** The second cluster (green in Fig 2) includes the majority of *M. nemestrina P. knowlesi (Mn-Pk)* nuclear genotype isolates. The third cluster (blue in Fig 2) consists only of isolates with *Mf-Pk* nuclear genotypes. The presence of samples in the other two clusters with mismatched nuclear and organellar genomes indicates that these two subpopulations have undergone genetic exchange. **b) Phylogenetic tree generated using 362 apicoplast SNPs.** The tree shows a very similar pattern of clustering to Fig 2**.**
(TIFF)

Acknowledgments

Sequence data analysis was performed on the MRC UK eMedLab computing resource.

Author Contributions

Conceptualization: Ernest Diez Benavente, Susana Campino, Taane G. Clark.

Data curation: Ernest Diez Benavente, Paola Florez de Sessions.

Formal analysis: Ernest Diez Benavente.

Funding acquisition: Martin L. Hibberd, Taane G. Clark.

Investigation: Susana Campino, Taane G. Clark.

Methodology: Paola Florez de Sessions.

Project administration: Susana Campino, Taane G. Clark.

Resources: Paola Florez de Sessions, Robert W. Moon, Anthony A. Holder, Michael J. Blackman, Cally Roper, Christopher J. Drakeley, Colin J. Sutherland, Martin L. Hibberd.

Supervision: Martin L. Hibberd, Susana Campino, Taane G. Clark.

Visualization: Ernest Diez Benavente.

Writing – original draft: Ernest Diez Benavente, Taane G. Clark.

Writing – review & editing: Cally Roper, Arnab Pain, Colin J. Sutherland, Susana Campino, Taane G. Clark.

References

1. Singh B, Kim Sung L, Matusop A, Radhakrishnan A, Shamsul SS, Cox-Singh J, et al. A large focus of naturally acquired Plasmodium knowlesi infections in human beings. Lancet 2004; 363, 1017–1024. https://doi.org/10.1016/S0140-6736(04)15836-4 PMID: 15051281

2. Kantele A, Jokiranta TS. Review of cases with the emerging fifth human malaria parasite, Plasmodium knowlesi. Clin Infect Dis 2011; 52, 1356–1362. https://doi.org/10.1093/cid/cir180 PMID: 21596677

3. Putaporntip C, Hongsrimuang T, Seethamchai S, Kobasa T, Limkittikul K, Cui L, et al. Differential Prevalence of Plasmodium Infections and Cryptic Plasmodium Knowlesi Malaria in Humans in Thailand. J Infect Dis 2009; 199: 1143–50. https://doi.org/10.1086/597414 PMID: 19284284

4. Muller M, Schlagenhauf P. Plasmodium knowlesi in travellers, update 2014. Int J infect Dis 2014; 22, 55–64.
5. Singh B, Daneshvar C. Human infections and detection of Plasmodium knowlesi. Clinical microbiology reviews 2013; 26, 165–184. https://doi.org/10.1128/CMR.00079-12 PMID: 23554413
6. Lubis IND, Wijaya H, Lubis M, Lubis CP, Divis PCS, Beshir KB, et. al. Contribution of Plasmodium knowlesi to Multispecies Human Malaria Infections in North Sumatera, Indonesia. J Infect Dis 2017; 215(7), 1148–1155. https://doi.org/10.1093/infdis/jix091 PMID: 28201638
7. Imai N, White MT, Ghani AC, Drakeley CJ. Transmission and Control of Plasmodium Knowlesi: A Mathematical Modelling Study. PLOS Negl Trop Dis 2014; 8 e2978. https://doi.org/10.1371/journal.pntd.0002978 PMID: 25058400
8. Lee KS, Divis PC, Zakaria SK, Matusop A, Julin RA, Conway DJ, et al. Plasmodium knowlesi: reservoir hosts and tracking the emergence in humans and macaques. PLoS Pathog 2011; 7, e1002015. https://doi.org/10.1371/journal.ppat.1002015 PMID: 21490952
9. Pinheiro MM, Ahmed MA, Millar SB, Sanderson T, Otto TD, Lu WC, et al. Plasmodium knowlesi Genome Sequences from Clinical Isolates Reveal Extensive Genomic Dimorphism. PLoS ONE 2015; 10(4), e0121303. https://doi.org/10.1371/journal.pone.0121303 PMID: 25830531
10. Divis PC, Singh B, Anderios F, Hisam S, Matusop A, Kocken CH, et al. Admixture in Humans of Two Divergent Plasmodium knowlesi Populations Associated with Different Macaque Host Species. PLoS Pathog 2015; 11(5), e1004888. https://doi.org/10.1371/journal.ppat.1004888 PMID: 26020959
11. Assefa S, Lim C, Preston MD, Duffy CW, Nair MB, Adroub SA, et al. Population genomic structure and adaptation in the zoonotic malaria parasite Plasmodium knowlesi. Proc National Academy Sci U.S.A 2015; 112(42), 13027–13032.
12. Ahmed MA, Fong MY, Lau YL., Yusof R. Clustering and genetic differentiation of the normocyte binding protein (nbpxa) of Plasmodium knowlesi clinical isolates from Peninsular Malaysia and Malaysia Borneo. Malaria J 2016; 15, 241.
13. Divis PC, Lin LC, Rovie-Ryan JJ, Kadir KA, Anderios F, Hisam S, et al. Three Divergent Subpopulations of the Malaria Parasite Plasmodium knowlesi. Emerging Infectious Diseases 2017; 23(4), 616–624. https://doi.org/10.3201/eid2304.161738 PMID: 28322705
14. Fornace KM, Abidin TR, Alexander N, Brock P, Grigg MJ, Murphy A, et al. Association between landscape factors and spatial patterns of Plasmodium knowlesi Infections in Sabah, Malaysia. Emerg Infect Dis. 2016; 22, 201–208. https://doi.org/10.3201/eid2202.150656 PMID: 26812373
15. Yusof R, Ahmed MA, Jelip J, Ngian HU, Mustakim S, Hussin HM, et al. Phylogeographic Evidence for 2 Genetically Distinct Zoonotic Plasmodium knowlesi Parasites, Malaysia. Emerging Infectious Diseases 2016; 22(8), 1371–1380. https://doi.org/10.3201/eid2208.151885 PMID: 27433965
16. Vythilingam I, Tan CH, Asmad M, Chan ST, Lee KS, Singh B. Natural transmission of Plasmodium knowlesi to humans by Anopheles latens in Sarawak, Malaysia. Trans Roy Soc Tropl Med Hyg 2006; 100(11), 1087–1088.
17. Tan CH, Vythilingam I, Matusop A, Chan ST, Singh B. Bionomics of Anopheles latens in Kapit, Sarawak, Malaysian Borneo in relation to the transmission of zoonotic simian malaria parasite Plasmodium knowlesi. Malaria J. 2008; 7(1), 52.
18. Brant HL, Ewers RM, Vythilingam I, Drakeley C, Benedick S, Mumford JD. D. Vertical stratification of adult mosquitoes (Diptera: Culicidae) within a tropical rainforest in Sabah, Malaysia. Malaria J. 2016; 15(1), 370.
19. Vythilingam I, Noorazian YM, Huat TC, Jiram AI, Yusri YM, Azahari AH, et al. Plasmodium knowlesi in humans, macaques and mosquitoes in peninsular Malaysia. Parasit Vectors 2008; 1(1):26. https://doi.org/10.1186/1756-3305-1-26 PMID: 18710577
20. Moyes CL, Henry AJ, Golding N, Huang Z, Singh B, Baird JK, et al. Defining the Geographical Range of the Plasmodium knowlesi Reservoir. PLoS Negl Trop Dis 2014; 8: e2780. https://doi.org/10.1371/journal.pntd.0002780 PMID: 24676231
21. Preston MD, Campino S, Assefa SA, Echeverry DF, Ocholla H, Amambua-Ngwa A, et al. A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. Nature Comm2014; 5, 4052. PMID: 24923250
22. Benavente ED, de Sessions PF, Moon RW, Grainger M, Holder AA, Blackman MJ, et al. A reference genome and methylome for the Plasmodium knowlesi malaria A1-H.1 line. Int J Parasit. In press. https://doi.org/10.1645/12-11.1
23. Moon RW, Sharaf H, Hastings CH, Ho YS, Nair MB, Rchiad Z, et al. A. Normocyte-binding protein required for human erythrocyte invasion by the zoonotic malaria parasite Plasmodium knowlesi. Proc Natl Acad Sci U S A 2016; 113: 7231–6. https://doi.org/10.1073/pnas.1522469113 PMID: 27303038

24. Srinivasan P, Fujioka H, Jacobs-Lorena M. PbCap380, a novel oocyst capsule protein, is essential for malaria parasite survival in the mosquito. Cellular Microbiology 2008; 10(6), 1304–1312. https://doi.org/10.1111/j.1462-5822.2008.01127.x PMID: 18248630

25. Dessens JT, Beetsma AL, Dimopoulos G, Wengelnik K, Crisanti A, Kafatos FC, et al. CTRP is essential for mosquito infection by malaria ookinetes. The EMBO Journal 1999; 18(22), 6221–6227. https://doi.org/10.1093/emboj/18.22.6221 PMID: 10562534

26. López-Barragán MJ, Lemieux J, Quiñones M, Williamson KC, Molina-Cruz A, Cui K et al. Directional gene expression and antisense transcripts in sexual and asexual stages of Plasmodium falciparum. BMC Genomics 2011; 12(1), 587.

27. López-Barragán MJ, Lemieux J, Quiñones M, Williamson KC, Molina-Cruz A, Cui K, et al. A comprehensive evaluation of rodent malaria parasite genomes and gene expression. BMC Biology 2014; 12 (1), 86.

28. Molina-Cruz A, Garver LS, Alabaster A, Bangiolo L, Haile A, Winikor J, et al. The human malaria parasite Pfs47 gene mediates evasion of the mosquito immune system. Science 2013; 340(6135):984–7. https://doi.org/10.1126/science.1235264 PMID: 23661646

29. Messenger LA, Llewellyn MS, Bhattacharyya T, Franzén O, Lewis MD, Ramírez JD, et al. Multiple mitochondrial introgression events and heteroplasmy in Trypanosoma cruzi revealed by Maxicircle MLST and Next Generation Sequencing. PLoS Negl Trop Dis 2012; 6(4), e1584. https://doi.org/10.1371/journal.pntd.0001584 PMID: 22506081

30. Messenger LA, Miles MA. Evidence and importance of genetic exchange among field populations of Trypanosoma cruzi. Acta Tropica 2015; 151, 150–155. https://doi.org/10.1016/j.actatropica.2015.05.007 PMID: 26188331

31. Brock PM, Fornace KM, Parmiter M, Cox J, Drakeley CJ, Ferguson HM, et al. Plasmodium knowlesi transmission: integrating quantitative approaches from epidemiology and ecology to understand malaria as a zoonosis. Parasitology 2016; 143(4), 389–400. https://doi.org/10.1017/S0031182015001821 PMID: 26817785

32. Vythilingam I, Wong ML, Wan-Yussof WS. Current status of Plasmodium knowlesi vectors: a public health concern? Parasitology 2016; 1–9.

33. Diez Benavente E, Ward Z, Chan W, Mohareb FR, Sutherland CJ, Roper C, et al. Genomic variation in Plasmodium vivax malaria reveals regions under selective pressure. PLOS ONE 2017; 12(5), e0177134. https://doi.org/10.1371/journal.pone.0177134 PMID: 28493919

34. Moyes CL, Shearer FM, Huang Z, Wiebe A, Gibson HS, Nijman V, et al. Predicting the geographical distributions of the macaque hosts and mosquito vectors of Plasmodium knowlesi malaria in forested and non-forested areas. Parasites & Vectors 2016; 9, 242.

35. Miettinen J, Shi C, Liew SC. Land cover distribution in the peatlands of Peninsular Malaysia, Sumatra and Borneo in 2015 with changes since 1990. Global Ecology and Conservation 2016; 6, 67–78.

36. Yasuoka J, Levins R. Impact of deforestation and agricultural development on anopheline ecology and malaria epidemiology. Am J Trop Med Hyg 2007; 76(3), 450–460. PMID: 17360867

37. Ansari HR, et al. Genome-scale comparison of expanded gene families in Plasmodium ovale wallikeri and Plasmodium ovale curtisi with Plasmodium malariae and with other Plasmodium species. Int J Parasitol 2016; 46(11):685–96. https://doi.org/10.1016/j.ijpara.2016.05.009 PMID: 27392654

38. Ansari HR, Templeton TJ, Subudhi AK, Ramaprasad A, Tang J, Lu F, et al. Estimating Geographical Variation in the Risk of Zoonotic Plasmodium knowlesi Infection in Countries Eliminating Malaria. PLOS Negl Trop Dis 2016; 10(8), e0004915. https://doi.org/10.1371/journal.pntd.0004915 PMID: 27494405

39. Fornace KM, Nuin NA, Betson M, Grigg MJ, William T, Anstey NM, et al. Asymptomatic and Submicroscopic Carriage of Plasmodium knowlesi Malaria in household and community members of clinical cases in Sabah, Malaysia. J Infect Dis 2016; 213(5), 784–787. https://doi.org/10.1093/infdis/jiv475 PMID: 26433222

40. Hayakawa T, Culleton R, Otani H, Horii T, Tanabe K. Big bang in the evolution of extant malaria parasites. Mol Biol Evol 2008; 25(10), 2233–2239. https://doi.org/10.1093/molbev/msn171 PMID: 18687771

41. Muehlenbein MP, Pacheco MA, Taylor JE, Prall SP, Ambu L, Nathan S, et al. Accelerated diversification of nonhuman primate malarias in Southeast Asia: adaptive radiation or geographic speciation? Molecular Biology and Evolution 2015; 32(2), 422–439. https://doi.org/10.1093/molbev/msu310 PMID: 25389206

42. Sutherland CJ. Persistent Parasitism: The Adaptive Biology of Malariae and Ovale Malaria. Trends in Parasitology 2016; 32(10), 808–819. https://doi.org/10.1016/j.pt.2016.07.001 PMID: 27480365

43. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

44. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011; 27(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509 PMID: 21903627

45. Campino S, Benavente ED, Assefa S, Thompson E, Drought LG, Taylor CJ, et al. Genomic variation in two gametocyte non-producing Plasmodium falciparum clonal lines. Malaria J 2016; 15(1), 229. PMID: 27098483

46. Samad H, Coll F, Preston MD, Ocholla H, Fairhurst RM, Clark TG. Imputation-based population genetics analysis of Plasmodium falciparum malaria parasites. PLoS Genet. 2015; 11(4):e1005131. https://doi.org/10.1371/journal.pgen.1005131 PMID: 25928499

47. Assefa SA, Preston MD, Campino S, Ocholla H, Sutherland CJ, Clark TG. estMOI: estimating multiplicity of infection using parasite deep sequencing data. Bioinformatics 2014; 30(9), 1292–1294. https://doi.org/10.1093/bioinformatics/btu005 PMID: 24443379

48. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting FST. Nat Rev Genet 2009; 10(9), 639–650. https://doi.org/10.1038/nrg2611 PMID: 19687804

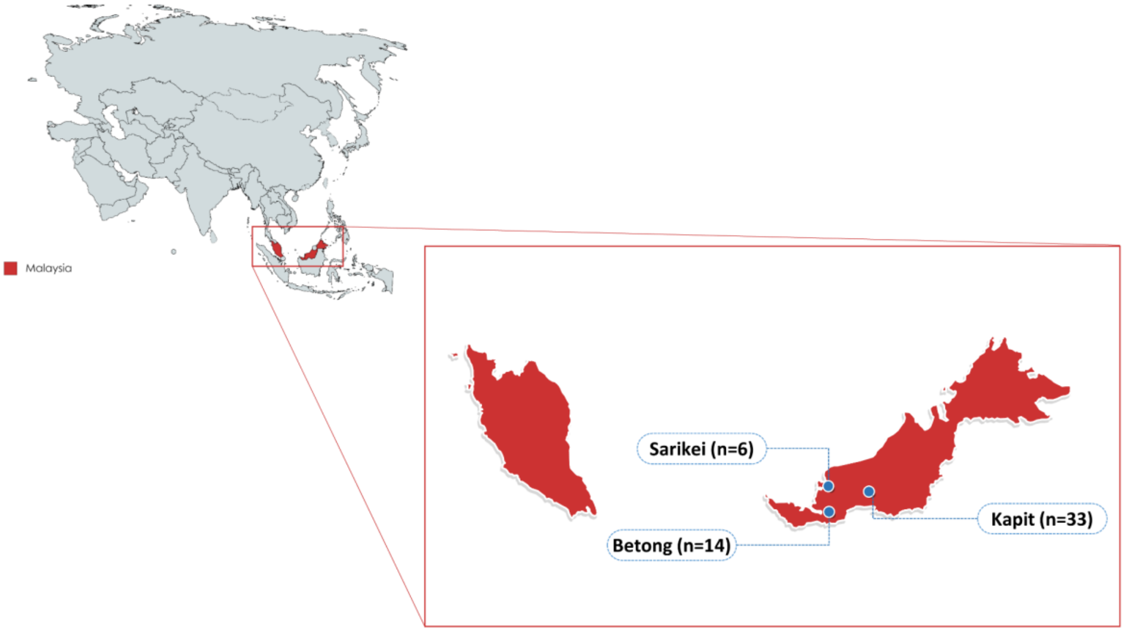
49. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature 2007; 449(7164), 913–918. https://doi.org/10.1038/nature06250 PMID: 17943131

50. Shimodaira H, Hasegawa H. Multiple comparisons of loglikelihoods with applications to phylogenetic inference. Mol Biol Evol 1999; 16, 1114.

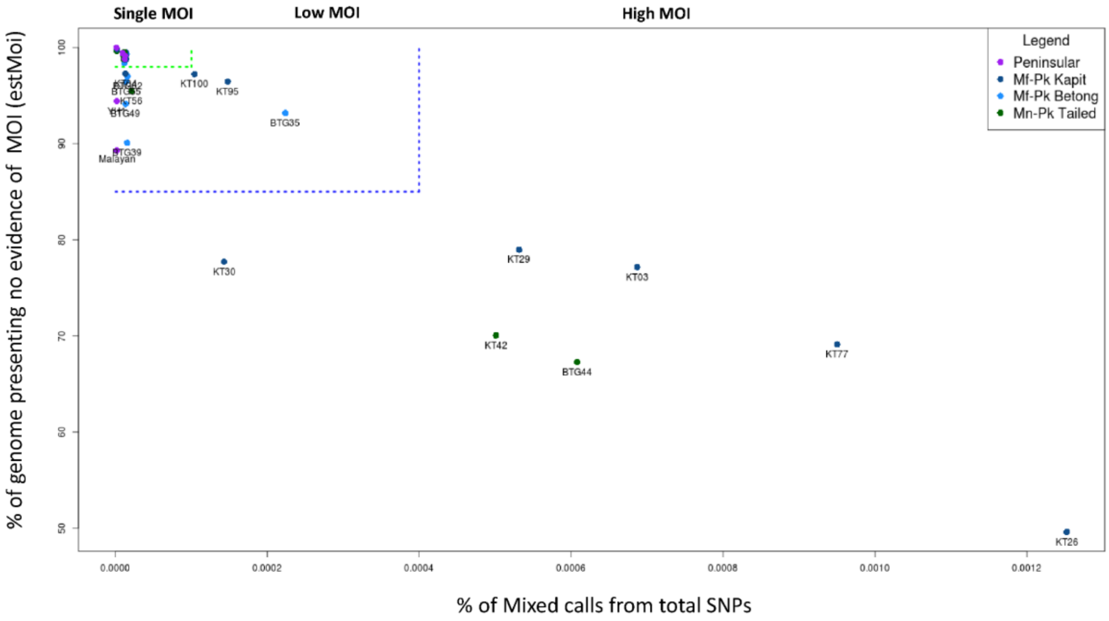
51. Templeton AR. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution 1983; 37, 221–244. https://doi.org/10.1111/j.1558-5646.1983.tb05533.x PMID: 28568373

Supplementary Information

Supplementary Figure 1: Geographical source of the *P. knowlesi* isolates: Betong (n = 14), Kapit (n = 33) and Sarikei (n = 6).

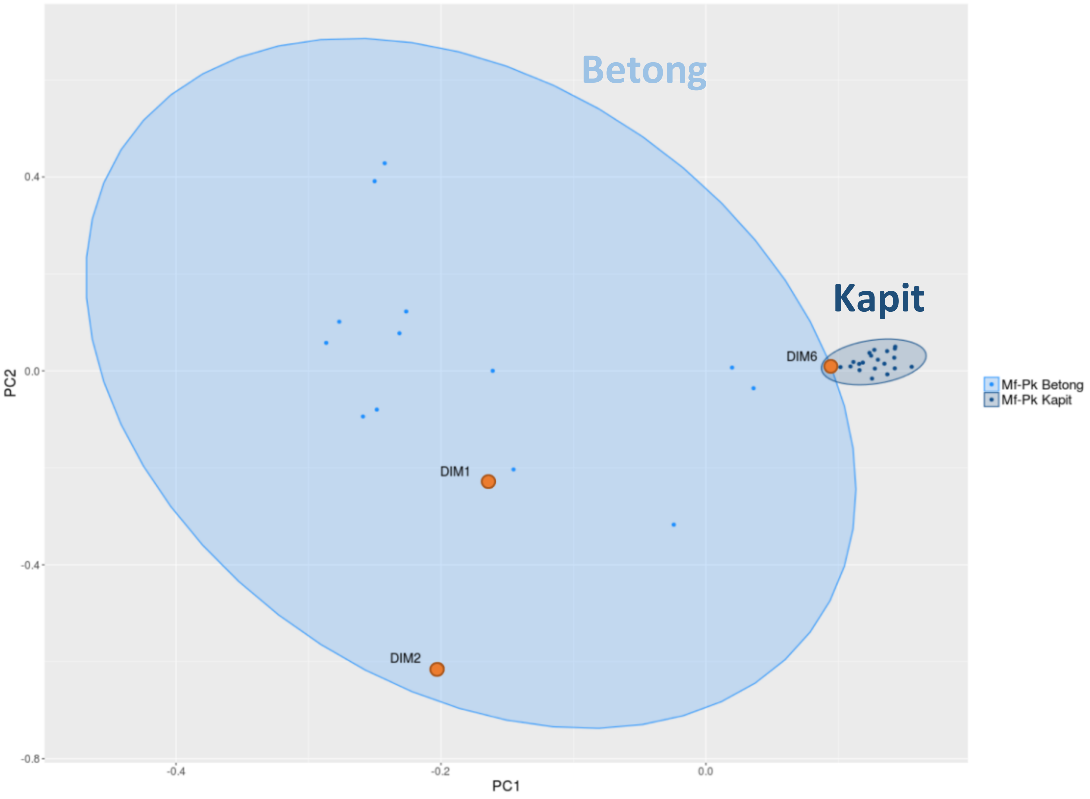


Supplementary Figure 2: Evaluation of multiplicity of infection (MOI) using mixed genotype calls (x-axis) and the estMOI read-pair haplotype counting approach [45] (y-axis) reveals seven highly non-clonal samples.



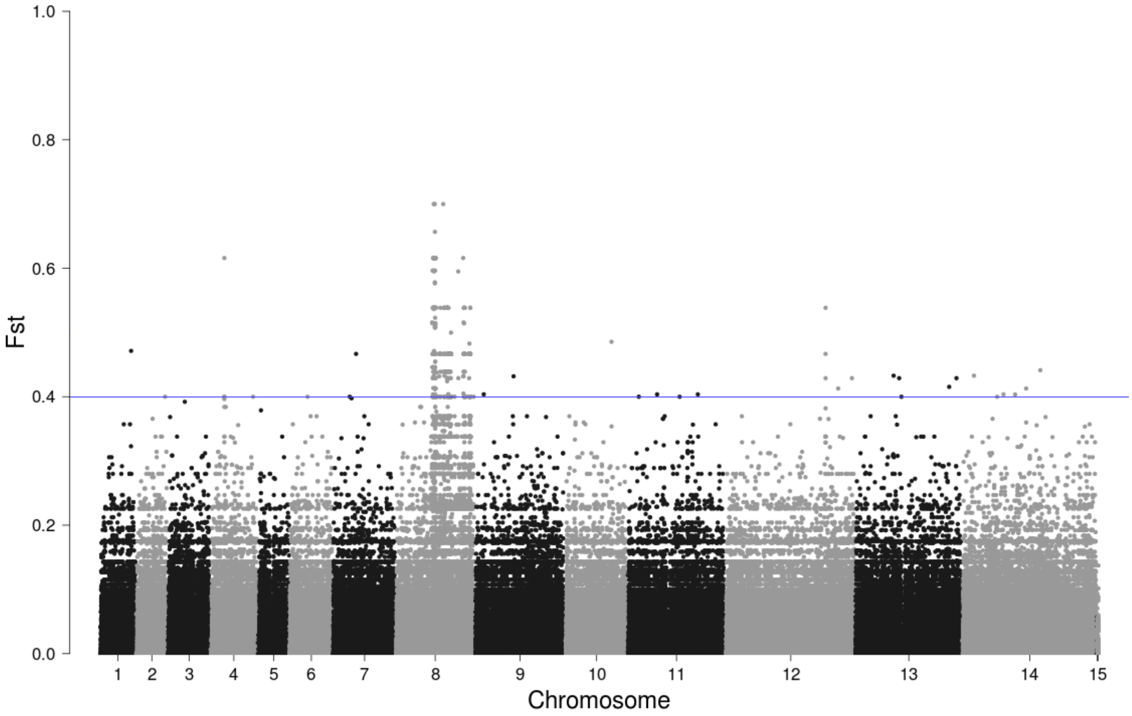
Supplementary Figure 3: Principal components analysis of the *M. fascicularis P. knowlesi* genotype group (*Mf-Pk*, Cluster 1) confirms that the subgroups from Kapit and Betong are separated.

The *Mf-Pk* Sarikei samples (DIM code in orange) cluster with either one of the two groups, which is consistent with the geographic location of Sarikei as an equidistant region between Kapit and Betong. There is increased diversity of Betong samples compared to the Kapit samples.



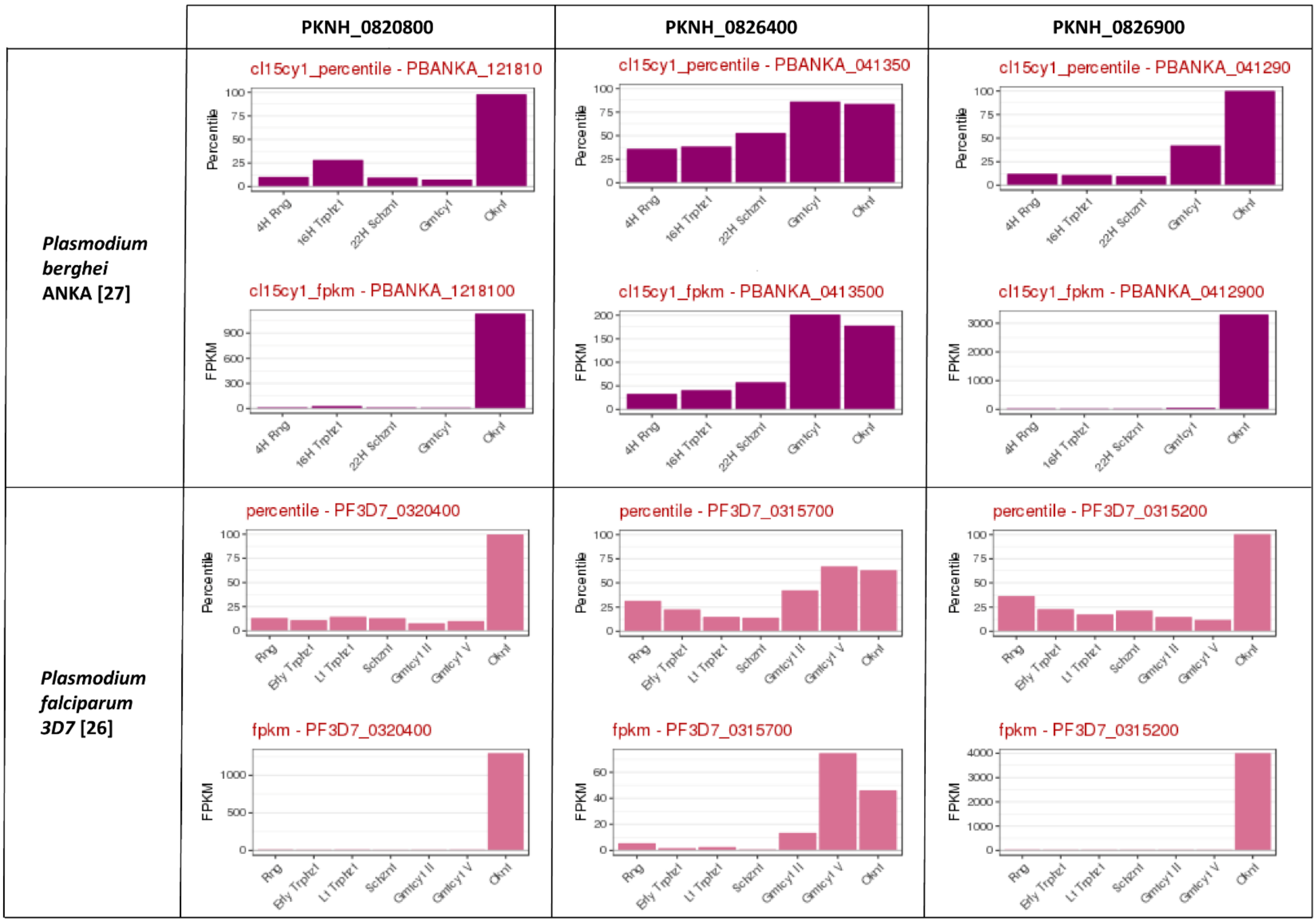
Supplementary Figure 4: Genome-wide differences in allele frequencies (measured using the fixation index (FST)) between *M. fascicularis P. knowlesi* genotype groups (*Mf-Pk*) from Kapit and Betong.

The comparison shows clear abnormalities in several genomic regions in chromosome 8 shown to be a result of genetic exchange with the *Mn-Pk* genotype.



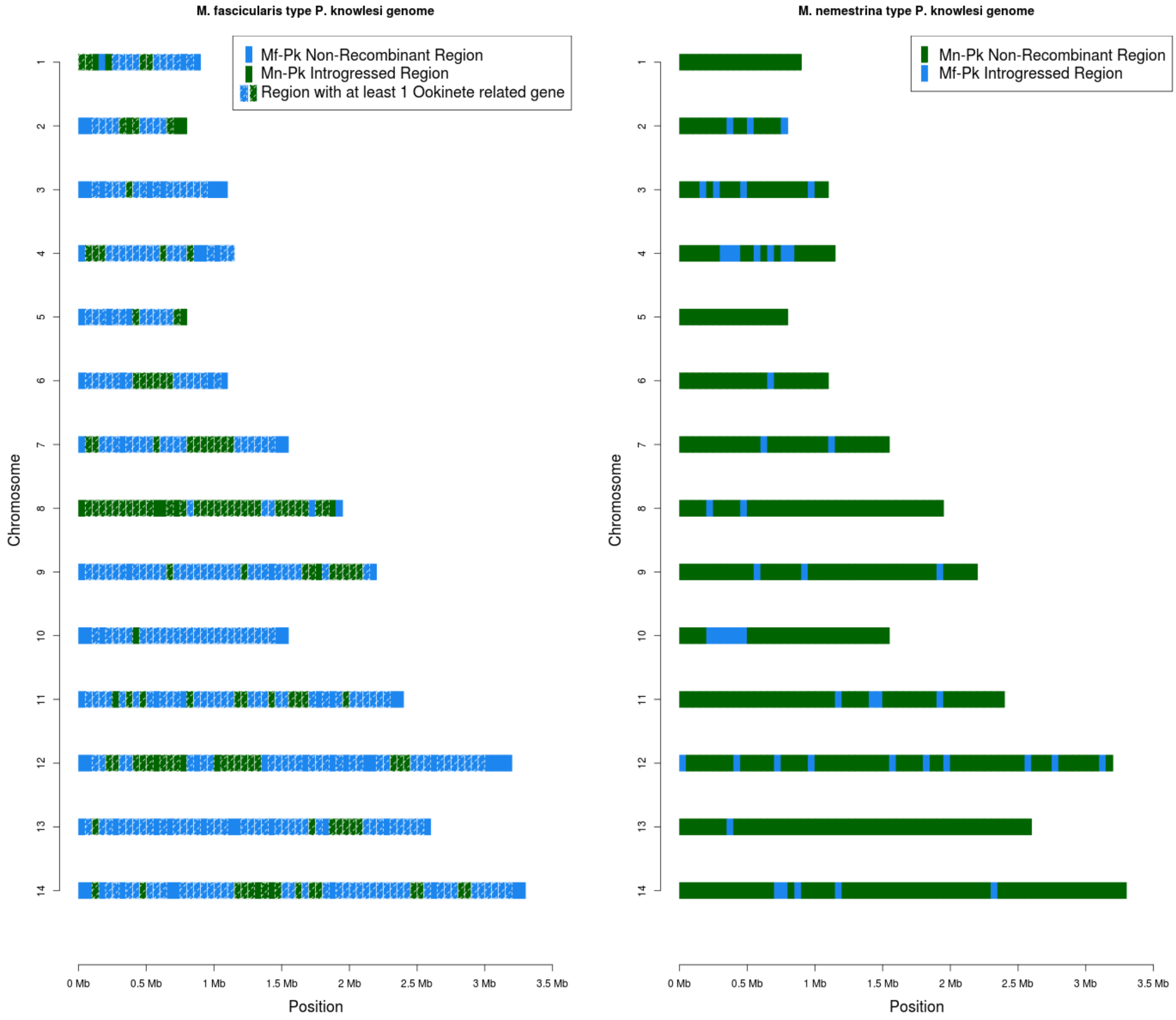
Supplementary Figure 5: Transcriptomic profiles for the orthologues of the introgressed genes under selection pressure.

The transcriptomic profiles of the orthologues in *P. falciparum* [26] and *P. berghei* [27] for the three genes found to be under strong selection pressure were extracted from PlasmoDB (http://plasmodb.org/plasmo/), including the percentile and the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) plots. These included data for 5 *P. berghei* stages (4-hour Ring, 16-hour Trophozoite, 22-hour Schizont, Gametocyte and Ookinete) and 7 *P. falciparum* stages (Ring, early Trophozoite, late Trophozoite, Schizont, Gametocyte stage II, Gametocyte stage V and Ookinete), and showing a clear increased expression in mosquito related stages, particularly the ookinete stage.

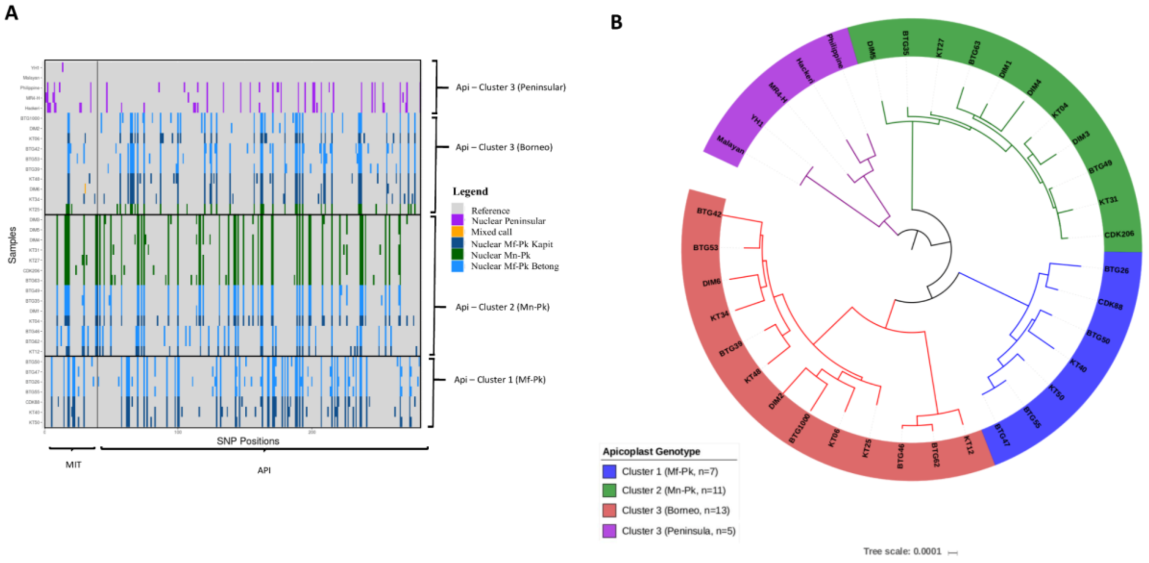


Supplementary Figure 6: Genome distribution of introgression events for each chromosome estimated using SNP diversity in 50Kb sliding windows.

(Left panel) location of introgressions from *M. nemestrina P. knowlesi (Mn-Pk*) genotype into *M. fasciscularis P. knowlesi (Mf-Pk)* genotypes, a dashed shaded region has been added where at least 1 gene related with the ookinete life stage of the parasite has been identified based on gene expression for the orthologue genes in *P. berghei* and/or *P. falciparum*. (Right panel) location of introgressions from *Mf-Pk* genotype into *Mn-Pk* genotypes.



Supplementary Figure 7: Analysis of organellar mitochondria (MIT) and apicoplast (Api) SNPs confirms clustering into three core haplotype groups a) Haplotype plot for the 36 samples with sufficient coverage across the organellar genomes. Three clearly defined clusters are present. The first cluster represents the mitochondrial genotype found in the Peninsular strains (purple, n = 5) and a set of 10 samples with a highly related haplotype with the smallest inter-cluster average FST (average FST = 0.16) from Borneo Malaysia (represented in red in Fig 2). The second cluster (green in Fig 2) includes the majority of *M. nemestrina P. knowlesi (Mn-Pk)* nuclear genotype isolates. The third cluster (blue in Fig 2) consists only of isolates with *Mf-Pk* nuclear genotypes. The presence of samples in the other two clusters with mismatched nuclear and organellar genomes indicates that these two subpopulations have undergone genetic exchange. b) Phylogenetic tree generated using 362 apicoplast SNPs. The tree shows a very similar pattern of clustering to Fig 2.



Supplementary Table 1: Study samples.

Sample	Code	Area	Host	MOI*	Group **	Total Reads	Mapped Reads %	Genome covered	Cover. Mean	ExΔ ***	Year
ERR274221	DIM1	Sarikei	Human	0.8	*Mf*	46353062	96.9	0.99	174.1	Yes	2012/3
ERR274222	DIM2	Sarikei	Human	0.7	*Mf*	60189967	97.7	0.99	228.7	Yes	2012/3
ERR274224	DIM3	Sarikei	Human	0.6	*Mn*	42350437	97.0	0.98	161.5	Yes	2012/3
ERR274225	DIM4	Sarikei	Human	0.5	*Mn*	52883838	97.2	0.98	201.6	Yes	2012/3
ERR366425	DIM5	Sarikei	Human	0.3	*Mn*	6391836	97.1	0.95	36.0	Yes	2012/3
ERR366426	DIM6	Sarikei	Human	0.2	*Mf*	6270503	94.8	0.95	34.1	Yes	2012/3
ERR985372	BTG1000	Betong	Human	0.8	*Mf*	30620879	66.0	0.98	79.7	Yes	2012/3
ERR985374	BTG26	Betong	Human	1.6	*Mf*	55120863	96.4	0.99	207.5	Yes	2012/3
ERR985375	BTG35	Betong	Human	6.8	*Mf*	66747737	22.7	0.98	55.4	Yes	2012/3
ERR985376	BTG39	Betong	Human	9.9	*Mf*	50924729	79.3	0.99	157.7	Yes	2012/3
ERR985377	BTG42	Betong	Human	1.0	*Mf*	71277238	95.3	0.99	248.7	Yes	2012/3
ERR985378	BTG46	Betong	Human	1.4	*Mf*	47365570	95.6	0.99	177.7	Yes	2012/3
ERR985379	BTG47	Betong	Human	1.2	*Mf*	40897898	95.5	0.99	153.6	Yes	2012/3
ERR985380	BTG49	Betong	Human	5.9	*Mf*	61989913	97.2	1	234.8	Yes	2012/3
ERR985381	BTG50	Betong	Human	1.4	*Mf*	51234398	94.9	0.99	189.9	Yes	2012/3
ERR985382	BTG53	Betong	Human	1.3	*Mf*	48838502	49.6	0.98	89.3	Yes	2012/3
ERR985383	BTG55	Betong	Human	3.6	*Mf*	40307405	85.8	0.99	126.8	Yes	2012/3
ERR985384	BTG62	Betong	Human	3.0	*Mf*	52753853	88.0	0.99	167.0	Yes	2012/3
ERR985385	CDK88	Kapit	Human	1.1	*Mf*	93728041	60.6	1	213.1	No	2012/3
ERR985386	KT03	Kapit	Human	22.8	*Mf*	52908664	22.3	0.98	45.3	No	2012/3
ERR985387	KT04	Kapit	Human	2.7	*Mf*	48092075	91.8	0.99	173.0	No	2012/3
ERR985388	KT06	Kapit	Human	1.0	*Mf*	68600487	92.2	0.99	247.4	No	2012/3
ERR985394	KT12	Kapit	Human	1.0	*Mf*	26219163	95.2	0.98	99.0	No	2012/3
ERR985395	KT26	Kapit	Human	50.4	*Mf*	45997535	86.1	1	154.6	No	2012/3
ERR985396	KT29	Kapit	Human	21.0	*Mf*	31849384	60.9	0.99	76.4	No	2012/3
ERR985397	KT30	Kapit	Human	22.3	*Mf*	40463921	92.0	1	146.3	No	2012/3
ERR985398	KT34	Kapit	Human	1.2	*Mf*	33387794	61.2	0.98	80.5	Yes	2012/3
ERR985399	KT40	Kapit	Human	1.3	*Mf*	34420982	97.0	0.99	131.2	No	2012/3
ERR985400	KT48	Kapit	Human	1.0	*Mf*	36366112	84.6	0.99	121.4	No	2012/3

ERR985401	KT50	Kapit	Human	1.0	*Mf*	36045889	95.3	0.99	135.9	Yes	2012/3
ERR985402	KT57	Kapit	Human	0.7	*Mf*	56171218	92.5	0.99	202.0	No	2012/3
ERR985403	KT72	Kapit	Human	0.6	*Mf*	52517284	95.4	0.99	193.6	No	2012/3
ERR985404	KT73	Kapit	Human	1.2	*Mf*	52157592	94.6	1	190.2	No	2012/3
ERR985410	BTG44	Betong	Human	32.7	*Mn*	32160549	97.1	0.99	122.9	No	2012/3
ERR985411	BTG63	Betong	Human	0.8	*Mn*	54179835	80.4	0.98	161.6	Yes	2012/3
ERR985412	CDK206	Kapit	Human	0.8	*Mn*	66023442	44.2	0.99	109.4	Yes	2012/3
ERR985414	KT25	Kapit	Human	0.8	*Mn*	65300398	93.9	0.99	242.3	Yes	2012/3
ERR985415	KT27	Kapit	Human	0.9	*Mn*	60931327	92.5	0.99	223.7	Yes	2012/3
ERR985416	KT31	Kapit	Human	1.2	*Mn*	78064720	92.6	0.99	286.4	Yes	2012/3
ERR985417	KT42	Kapit	Human	29.9	*Mn*	41500510	95.7	0.99	157.1	No	2012/3
ERR985418	KT55	Kapit	Human	0.7	*Mn*	54311715	89.9	0.98	191.6	Yes	2012/3
ERR985419	KT56	Kapit	Human	4.5	*Mn*	57630655	81.2	0.99	183.6	Yes	2012/3
ERR985373	BTG123	Betong	Human	0.7	*Mf*	42713162	78.6	0.98	119.3	Yes	2012/3
ERR985389	KT100	Kapit	Human	2.8	*Mf*	41469453	86.3	0.99	126.4	Yes	2012/3
ERR985390	KT103	Kapit	Human	0.5	*Mf*	41673914	83.2	0.99	117.2	No	2012/3
ERR985391	KT107	Kapit	Human	0.6	*Mf*	40562077	90.6	0.99	141.0	No	2012/3
ERR985392	KT109	Kapit	Human	0.6	*Mf*	42267002	87.0	0.99	132.1	Yes	2012/3
ERR985393	KT120	Kapit	Human	0.5	*Mf*	34929904	71.1	0.98	88.5	No	2012/3
ERR985405	KT77	Kapit	Human	30.9	*Mf*	38834680	88.7	1	128.1	No	2012/3
ERR985406	KT81	Kapit	Human	0.5	*Mf*	47157966	83.6	0.99	138.9	No	2012/3
ERR985407	KT92	Kapit	Human	0.8	*Mf*	43617554	86.2	0.99	135.2	Yes	2012/3
ERR985408	KT94	Kapit	Human	0.6	*Mf*	47394270	86.8	0.99	148.1	Yes	2012/3
ERR985409	KT95	Kapit	Human	3.5	*Mf*	50949905	80.5	0.99	134.1	Yes	2012/3
ERR985413	KT114	Kapit	Human	0.5	*Mn*	51040091	74.5	0.98	119.8	No	2012/3
SRR2221468	Hackeri	Penin.	Rh mac	1.2	*Penin.*	22542689	94.5	0.99	83.2	No	1960s
SRR2222335	H(AW)	Penin	Rh mac	0	*Penin.*	23371486	96.1	1	89.6	No	1960s
SRR2225467	Malayan	Penin.	Rh mac	10.7	*Penin.*	19455031	79.3	1	60.9	No	1960s
SRR2225571	MR4-H	Penin.	Rh mac	0.6	*Penin.*	22264926	74.6	0.98	64.0	No	1960s
SRR2225573	Philipp.	Penin.	Rh mac	0.8	*Penin.*	25538996	95.2	0.99	94.6	No	1960s
SRR3135172	YH1	Penin.	Rh mac	5.6	*Penin.*	21164226	90.1	1	74.6	No	1960s

* Multiplicity of infection (MOI) is % of genome presenting multiplicity of infection; **Group established by whole Genome PCA: Mf M. fascicularis, Mn M. nemestrina, Penin. Peninsular; Rh mac Rhesus macaque, *** evidence of genetic exchange (ExΔ)

Supplementary Table 2 from this article has the following title: "**Supplementary Table 2.** 50 Kb regions in the *P. knowlesi* genome that present genetic exchanges in the full set of samples" and can be found following this link (https://ndownloader.figshare.com/files/9430912). It will not be printed here given it is composed of 495 rows and was uploaded to the journal as an EXCEL table.

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED <u>FOR EACH</u> RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Ernest Diez Benavente
Principal Supervisor	Taane Clark & Susana Campino
Thesis Title	**Using whole genome sequence data to study genomic diversity and develop molecular barcodes to profile Plasmodium malaria parasites**

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	NOT APPLICABLE		
When was the work published?	NOT APPLICABLE		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**No**	Was the work subject to academic peer review?	**No**

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Scientific Reports
Please list the paper's authors in the intended authorship order:	Ernest Diez Benavente, Ana Rita Gomes, Jeremy Ryan De Silva, Matthew Grigg, Harriet Walker, Bridget E. Barber, Timothy William, Tsin Wen Yeo, Paola Florez de Sessions, Abhinay Ramaprasad, Amy Ibrahim, James Charleston, Martin L. Hibberd, Arnab Pain, Robert W. Moon, Sarah Auburn, Lau Yee Ling, Nicholas M. Anstey, Taane G. Clark, Susana Campino
Stage of publication	**Undergoing revision**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I downloaded the raw data from public repositories and collected from collaborators. I created the pipeline in which I ran the samples through and performed the QC analysis, as well as the interpretation analysis. I later created custom R and perl scripts to process the data in order to obtain information related to coverage, SNPs and other genomic information as well as performed de novo assembly. The figures presented in this work have all been generated using scripts written by myself or publicly available software adapted for this purpose. I wrote the first draft of the manuscript and circulated to co-authors. Once the comments were received I gathered them and made the relevant changes on the article manuscript. Then submitted the paper to the Scientific Reports journal.

Student Signature: ______________________________________ **Date:** __________

Supervisor Signature: ______________________________________ **Date:** __________

Chapter 4
Whole genome sequencing of amplified *Plasmodium knowlesi* DNA from unprocessed blood reveals genomic exchange events between Malaysian Peninsular and Borneo subpopulations

Whole genome sequencing of amplified *Plasmodium knowlesi* DNA from unprocessed blood reveals genomic exchange events between Malaysian Peninsular and Borneo subpopulations

Ernest Diez Benavente [1], Ana Rita Gomes [1,2], Jeremy Ryan De Silva[3], Matthew Grigg[4], Harriet Walker [1], Bridget E. Barber [4,5], Timothy William [5,6,7], Tsin Wen Yeo [4], Paola Florez de Sessions [8], Abhinay Ramaprasad [9], Amy Ibrahim [1], James Charleston [1], Martin L. Hibberd [1,8], Arnab Pain [9], Robert W. Moon [1], Sarah Auburn [4,], Lau Yee Ling [3], Nicholas M. Anstey [4], Taane G. Clark [1,10,*], Susana Campino [1,*]

[1] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom

[2] Centre Hospitalier Universitaire de Montpellier, Montpellier, France

[3] University of Malaya, Kuala Lumpur, Malaysia

[4] Global and Tropical Health Division, Menzies School of Health Research and Charles Darwin University, Darwin, Northern Territory, Australia

[5]Infectious Diseases Society Sabah-Menzies School of Health Research Clinical Research Unit, 88300 Kota Kinabalu, Sabah, Malaysia

[6]Clinical Research Centre, Queen Elizabeth Hospital, 88300 Kota Kinabalu, Sabah, Malaysia

[7]Jesselton Medical Centre, 88300 Kota Kinabalu, Sabah, Malaysia

[8] Genomics Institute, Singapore

[9] King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia

[10] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

* Joint corresponding authors

Prof. Taane Clark (taane.clark@lshtm.ac.uk)

Dr. Susana Campino (Susana.campino@lshtm.ac.uk)

ABSTRACT

Plasmodium knowlesi, a zoonotic malaria parasite, is the most common cause of human malaria in Malaysia. Genetic analysis has shown that the parasites are divided into three subpopulations according to their geographic origin (Peninsular or Borneo) and, in Borneo, their macaque host (*Macaca fascicularis* or *M. nemestrina*). Recent evidence suggests that genomic exchange events have occurred between the two host-specific subpopulations in Borneo, potentially reflecting adaption to host-vector-parasite interactions in a setting of ecological transition. However, the picture is less clear in Peninsular strains, which have been neglected in genomic studies. One difficulty is that genome sequencing requires sufficient high-quality biological material, but *P. knowlesi* infected individuals tend to present with low parasitaemias leading to samples with high human DNA contamination. Here, using a parasite selective whole genome amplification approach on unprocessed blood samples with high concentrations of human DNA, we were able to analyse recent *P. knowlesi* genomes sourced from both Peninsular Malaysia and Borneo. The high-quality data generated provides evidence that recombination events are also found in the Peninsular Malaysia parasite subpopulation. These parasites have acquired fragments of the *M. nemestrina* associated subpopulation genotype, including the *DBPβ* and *NBPXa* genes, which encode proteins involved in erythrocyte invasion. Furthermore, comprehensive characterization of the invasion genes revealed that *NBPXb* has also been exchanged within the monkey host-associated subpopulations of Malaysian Borneo. Our work provides strong evidence that exchange events are far more ubiquitous than expected and should be taken into consideration when studying the highly complex *P. knowlesi* population genomic structure.

Word count: 251

INTRODUCTION

Plasmodium knowlesi, a common malaria parasite of long-tailed (*Macaca fascicularis)* and pig-tailed (*M. nemestrina*) macaques, is now recognized as a significant cause of human malaria. Reported across all countries of Southeast Asia, *P. knowlesi* is now the predominant cause of malaria in Malaysia [1–4]. Severe disease occurs in 6-9% of clinical presentations and fatalities are well-described [1,5,6]. Rapid human population growth, deforestation and encroachment on wild macaque habitats are thought to increase contact with humans and drive up the incidence of human *P. knowlesi* infections [10,11]. Regional elimination efforts have targeted *P. falciparum* and *P. vivax* transmission, with significant progress demonstrated by the near-elimination of these *Plasmodium* species from areas such as Malaysian Borneo [3,4,9]. However, due to the difficulties in reducing the monkey parasite reservoir, it is unclear if similar control approaches are able to limit the risk for humans acquiring *P. knowlesi* malaria [12,13].

Appropriate molecular tools and sampling are needed to assist surveillance of *P. knowlesi* by malaria control programs, and to understand its genetic diversity and transmission. *P. knowlesi* genomics could lead to biological insights informing control measures. Advances in whole genome sequencing (WGS) technologies have led to the characterization of single nucleotide polymorphisms (SNPs) across *P. falciparum* and *P. vivax*, with an improved understanding of their population structure and diversity, as well as loci underpinning drug resistance (e.g. [14–19]). For *P. knowlesi* WGS studies, sample sizes have been small (n<70), but revealed that this parasite species is more polymorphic than *P. falciparum* and that three main subpopulations exist based on geographical source (Peninsular-Malaysia vs. Malaysian Borneo) and, within Malaysian Borneo, different hosts (*M. nemestrina* [*Mn-Pk*] and *M. fascicularis* [*Mf-Pk*] macaques, and humans) [20–22]. These studies have also provided evidence that *P. knowlesi* nuclear genomes are not genetically isolated, but also have chromosomal-segment exchanges between subpopulations [20–22]. This observation points to subpopulations that have diverged in isolation and then re-connected, possibly due to deforestation and disruption of wild macaque habitats. The resulting genomic mosaics reveal traits selected by host-vector-parasite interactions in a setting of ecological transition. However, despite these insights, *P. knowlesi* isolates from both monkeys and humans in Peninsular Malaysia are under-represented in analyses, and the genetic diversity in that region is less clear.

One roadblock to large-scale genomic studies of clinical *P. knowlesi* parasites is that the majority of infections have low parasitaemias, leading to samples with high levels of human compared to parasite DNA. Until now the WGS data for *Plasmodium* parasites has been obtained from venous blood of clinical cases that were filtered to remove human leukocytes, and therefore reduce human DNA "contamination". However, this approach does not always yield sufficient parasite DNA for WGS. Recently, a selective whole genome amplification (SWGA) strategy has been used to sequence *P. falciparum* and *P. vivax* genomes from non-filtered blood and from dried blood spots of clinical samples [23–25]. The SWGA method uses oligonucleotides that preferentially bind with high frequency to the target DNA, but less frequently to the "contaminating" genome [26]. The high fidelity Phi29 polymerase is then used to amplify long segments of DNA. Here we developed an SWGA approach for *P. knowlesi*, and sequenced 26 isolates across Malaysia, including both Peninsular and Borneo, revealing new insights into the population structure and evolution of this parasite.

RESULTS

Selective whole genome amplification of *P. knowlesi* parasite DNA from clinical samples

We performed SWGA on *P. knowlesi* DNA samples obtained directly from human blood using six selected primers that specifically amplified this parasite genome and bind less frequently to the human genome (See **Materials and Methods**). The primer set had a mean binding frequency of at least once every 4,826bp to the *P. knowlesi* genome, much higher than the once every 40,307bp to the human genome. Binding sites sufficiently near each other, as obtained with the primer set for *P. knowlesi*, enable the branching and displacement actions of the Phi29 polymerase and increase the success of the genome amplification[27].

For 10 samples, we performed WGS on both the non-amplified and the SWGA DNA. Both sets of samples sequenced at a similar depth, and we observed a significant increase (mean 7.7-fold more) in the proportion of reads that mapped to the *P. knowlesi* A1-H.1 reference genome after DNA amplification (**Table 1, s**howing only non-pooled sequencing results). As a result, amplified samples have higher genome coverage (mean 6.8-fold greater) and a much higher number of callable SNPs (mean 182-fold greater, with an average of 14,078 SNPs for no SWGA vs. 115,995 SNPs for SWGA) (**Table 1**). Gene regions presented with higher coverage (% of genes with coverage > 5-fold: average for no SWGA 5.3% vs. SWGA 43.3%) than intergenic regions (% of genes with coverage > 5-fold: average for no SWGA 4.1% vs. SWGA 30.4%). DNA from a further 16 clinical isolates underwent SWGA and WGS. A trend towards improved coverage in samples with higher parasitaemias was observed (R^2= 0.6, **Figure 1**), with superior results for samples with ≥5,000 parasites/µl, consistent with data from *P. vivax* and *P. falciparum* isolates [24,25]. For samples with <5,000 parasites/µl results are more variable and do not correlate with an increase in parasitaemia. For these low-parasitaemia samples the genome coverage (> 5-fold) ranged from 6 to 43% after amplification, and represent an average increase of 78% in coverage compared to non-amplified samples, and an average of 66,143 callable SNPs over 2,908 SNPs for non-amplified samples (**Table 1**). For samples with lower parasite densities, increased sequencing and merging of the resulting reads can lead to improved genome coverage, as shown for *P. vivax* [25]. Evidence of mixed infections (multiclonality) was detected in 2 SWGA samples, demonstrating that the method can amplify more than one clone present in an infection, as was observed for *P. vivax* amplified samples [25].

P. knowlesi genetic variation and clustering of isolates

The sequence data for the 26 new *P. knowlesi* isolates and 156 previously sequenced samples (see **Materials and Methods**) were mapped to the A1-H.1 reference genome [28]. The new genomes include recently collected Peninsular Malaysian isolates (n=5) and clinical isolates from Sabah, Malaysia Borneo (n=21). From the resulting alignments 1,741,056 high quality SNPs were identified across the 14 chromosomes. Out of the 156 samples, 12 isolates with high levels of Multiplicity of Infection (see **Supplementary figure 1**) and 41 isolates with low genome coverage (<30% of genome covered) were excluded; leaving 103 (20 SWGA) isolates for further analysis (see **Supplementary table 1**). A neighbour-joining tree was constructed using the SNP data (**Figure 2**) and revealed 3 predominant clusters, consistent with recent findings [20,22]. In particular, these clusters relate to the specific geographic Peninsular-Malaysia subpopulation (purple), and Borneo macaque *Mn-Pk* (green) and *Mf-Pk* (blue) associated subpopulations. Furthermore, the tree showed a consistent positioning for SWGA isolates: 4 SWGA Peninsular isolates (red branches) clustered within the Peninsular Malaysia clade (purple), and of the 16 SWGA Sabah isolates, 2 and 14 clustered within the *Mn-Pk* (green) and *Mf-Pk* (blue) Borneo clades, respectively. This result demonstrates that the SWGA method can amplify all known sub-populations of *P. knowlesi*.

***Genomic exchange events in P. knowlesi* isolates from Peninsular Malaysia**

It has been shown that the subpopulations of *P. knowlesi* in Malaysian Borneo, although presenting a strong genetic differentiation, are not genetically isolated. In particular, we have identified genomic exchanges predominantly between the *Mf-Pk* and *Mn-Pk* clusters [20]. We sought to investigate whether these events are also found in the clinical isolates from Peninsular Malaysia, by estimating SNP nucleotide diversity (*SNP π*) across the genome in sliding 50kb windows. In two isolates (P137 and P050), we identified several regions with an exceptional increase in similarity with the *Mn-Pk* cluster and reduced similarity with the Peninsular Malaysia cluster **(Figure 3**). Analysis of the haplotypes for each individual isolate confirmed exchange events. The analysis of the individual sequence haplotypes of genes in the identified regions showed mis-clustering in a neighbour-joining tree when compared to the whole genome clustering patterns. These genes are represented in **Supplementary Table 2**. All the events observed were associated with introgression from *Mn-Pk* cluster haplotypes into the Peninsular Malaysia genomes and spanned mostly subtelomeric regions in chromosomes 1, 2, 7, 9, 10, 11, 12, 13 and 14. A high proportion of *Plasmodium* exported proteins with unknown function were found to be affected by the exchange as well as tryptophan-rich antigens and lysophospholipases, genes associated with parasite invasion (*Normocyte Binding Protein Xa (*NBPXa), *Duffy Binding Protein beta (DBPβ)*) [29], and a cytoadherence linked asexual protein gene (*PKNH_1401300*). These results could indicate that the exchange events found in Peninsular Malaysia may driven by host-related factors in the erythrocytic stages of the parasite life cycle.

Following the discovery of *P. knowlesi* invasion-related genes in the exchanged regions we performed a comprehensive analysis of the genetic diversity harboured by the five reticulocyte binding like (*RBL*) and Duffy binding protein like (*DBP*) genes involved in erythrocyte invasion: (A) *DBPα, (B) NBPXb, (C) DBPβ, (D) DBPγ and (E) NBPXa*. For each individual isolate, we compared its “invasion” haplotypes **(Figure 4, left)** to its position on neighbour-joining trees, with its expected clustering based on WGS data (**Figure 4, right**). There was a strong genetic divergence of the sequences from the different clusters for each of the 5 genes, with the Peninsular Malaysia cluster presenting with marginally greater nucleotide diversity (Peninsular Malaysia:

mean= 1.88x10^{-5}; range=1.63x10^{-5} – 2.17x10^{-5}) in all 5 genes when compared to the other two clusters (*Mf-Pk*: mean=1.85x10^{-5}, range=1.15x10^{-5} - 2.71x10^{-5}; *Mn-Pk*: mean= 1.01x10^{-5}, range=0.78x10^{-5} - 1.42x10^{-5}).

For all the isolates*, DBPα* and *DBPγ* gene sequence-based clusters matched those from whole genome data. For *DBPβ*, the overall clustering pattern was still present, but one isolate from Peninsular Malaysia (P050; **Figure 4D, right; red star**) had clear evidence of genomic exchanges with the *Mn-Pk* cluster. The longer branch length of the P050 isolate in the tree, where the introgression is found, reveals a stronger genetic difference compared to the other *Mn-Pk* DBPβ haplotypes, indicating that this exchange event could be non-recent. This observation is confirmed by the partial similarity of the P050 haplotype with the *Mn-Pk* haplotypic patterns (**Figure 4D, left**). There was also evidence of genomic exchanges events in the *NBPXa,* which had the lowest overall genetic diversity of all RBL/DBP genes (mean=1.20x10^{-5}). For example, the Peninsular Malaysia isolate P137 appears to have introgressed from the *Mn-Pk* cluster, and its partial similarity to the current *NBPXa* haplotypes of the *Mn-Pk*'s clade suggests it could also be a non-recent event. Finally, *NBPXb* gene had low diversity, with intra-cluster genetic distances being smaller than those found in the *DBP* genes. Several introgression events were identified, where 9 out of 33 (27%) of the *Mn-Pk* cluster isolates presented with *Mf-Pk* type haplotypes (**Figure 4B, right; red stars**). Most of these isolates clustered together and are separated from the *Mf-Pk* samples, which could be a reflection of a unique non-recent introgression event. The isolate KT233 positioned in the "*Mn-Pk"* group using all SNPs, has a much similar *NBPXb* haplotype to those found in the *Mf-Pk* samples, reflecting a more recent exchange event.

DISCUSSION

The SWGA method produced reliable sequence data for parasite isolates obtained from unprocessed blood belonging to the three currently known subpopulations of *P. knowlesi*. This method is cost-effective, does not require sample processing at the time of collection, requires low quantities of input DNA, and is easy to implement. Importantly, the approach permits the genomic analysis of isolates that would otherwise be impossible to investigate, as demonstrated by the poor WGS results of the non-SWGA DNA when compared with their respective SWGA samples. A neighbour-joining tree based on SNP data revealed 3 predominant clusters, consistent with recent findings, and the positioning of the SWGA isolates confirmed their origin. The 4 recent Peninsular Malaysia isolates clustered closely with the long-term maintained samples originating from different regions in Peninsular Malaysia and the Philippines. Of the 16 isolates originating from Sabah, Malaysia, 2 belonged to the *Mn-Pk* associated cluster, and 14 belonged to the *Mf-Pk* associated cluster. This finding is consistent with the higher proportions of samples circulating in humans belonging to the *Mf-Pk* cluster [30] and confirms the presence of both Borneo monkey host-related subpopulations in Sabah.

Previous population genetics studies on *P. knowlesi* subpopulations among human and macaque isolates from Malaysian Borneo provided evidence that chromosomal-segment exchanges between subpopulations have occurred recently [20]. This observation could be indicative of subpopulations that diverged in isolation and have re-connected, possibly due to deforestation and disruption of wild macaque habitats. Up until now these introgressions had only been observed in parasites from Malaysian Borneo [10], but the inclusion of recent Peninsular Malaysia isolates allowed us to scan for more widespread events. We identified regions that presented evidence of introgression, in particular, with exceptionally high average SNP diversity when compared to the Peninsular isolates and exceptionally low diversity when compared to *Mn-Pk* or *Mf-Pk* isolates. This approach highlighted the presence of such events in 2 Peninsular samples, where the identified regions were found to be introgressions from *Mn-Pk* type haplotypes and spanned several chromosomes. This work shows that genomic exchange events are more widespread than previously thought, and they also affect the geographic subpopulation of Peninsular Malaysia. This finding does not contradict the previously observed distribution of *P. knowlesi* subpopulations, and is consistent with microsatellite analyses that have identified traces of Borneo-associated clusters in regions of Peninsular Malaysia [30,31].

Regions identified with genomic exchange events were enriched with large multi-gene families coding for *Plasmodium* exported proteins and tryptophan-rich antigens, as well as loci associated with erythrocyte invasion in *P. knowlesi*. These findings contrast with our previous results found in previous studies in Borneo [20], where the genes involved in introgression events were enriched by mosquito-stage related genes and could suggest that there are different factors driving the exchanges between geographical regions. We found that all known RBL/DBP invasion genes (*DBP* α, β and γ; *NBP Xa* and *Xb*) are highly differentiated and cluster the isolates into 3 subpopulations. It is known that both NBPXa and NBPXb bind to RBCs and NBPXa is known to be essential for the invasion of human RBCs, furthermore DBPα in *P. knowlesi* is the adhesin binding the Duffy antigen/chemokine receptor (DARC) with the role of the two paralogues (DBP β and γ) still unclear. This clustering is consistent with host-related factors being one of the main drivers for their genetic differentiation; although the Peninsular subpopulation is assumed to be a geographic subdivision rather than a host-associated cluster. Across all five loci the *Mf-Pk* group is the most genetically diverse. The DBPα and DBPγ did not present any genetic exchanged pattern. For the *NBPXb* gene the subpopulations were not as strongly differentiated as in other genes, and some *Mn-Pk* isolates (including from Sabah) presented with introgression from the *Mf-Pk* subpopulation. This introgression event could be related to the adaptation of the parasites to the different hosts, as it has been shown that the two Borneo subpopulations can be found in both species of macaques [21,22,30]. Furthermore, the events observed here could involve yet another subpopulation of Peninsular *Mn-Pk* type of *P. knowlesi* as the level of sampling currently is very low and the haplotypes do show a degree of differentiation with the other *Mn-Pk* haplotypes. Other genes such as *DBP*β and *NBPXa* presented introgression events from *Mn-Pk* into the Peninsular subpopulation. *NBPXa* has the lowest level of genetic diversity, which suggests that this gene is highly conserved across subpopulations. This is an important finding because *NBPXa* is required for the invasion of human red blood cells (RBCs) *in vitro* [29] and it has been shown that different haplotypes in *DBP*α region II have different binding affinities to the DARC receptor in human RBCs [32]. Therefore, these genetic exchanges affecting genes involved in invasion may reflect an adaptation to a new host and may confer improved binding and increase invasion efficiency. It will be important to investigate whether changes in the haplotypes of other genes involved in human RBC invasion affect the ability of the parasite to invade and multiply in human cells. Genetic interactions between invasion genes may assist parasites with adapting more efficiently to humans and facilitate transmission, which could hamper malaria elimination efforts.

Overall, by establishing an effective SWGA strategy for *P. knowlesi*, it will be possible to perform much needed large-scale WGS studies of the parasite genomic diversity across Asia, as well as investigate important fundamental biology such as the genetics underlying mechanisms involved red blood cell invasion.

MATERIALS AND METHODS

Sample collection and preparation

For this project we use *P. knowlesi* DNA samples from Sabah in Malaysian Borneo (n=21) (provided by the Menzies School of Health Research) and from Peninsular Malaysia (n=5) (provided by the University of Malaya). Samples from Sabah were obtained from patients enrolled as part of clinical malaria studies conducted from 2010 to 2014 [6,33]. Ethical approval for these studies was obtained from the Ministry of Health, Malaysia, and Menzies School of Health Research, Australia. Samples from Peninsular Malaysia were collected from patients admitted to University Malaya Medical Centre (UMMC), Kuala Lumpur, from July 2008 to December 2014[34]. Ethics was approved by the University of Malaya Medical Centre Medical Ethics Committee (MEC Ref. No: 817.18). All DNA samples were quantified using the Qubit Fluorometer using the dsDNA high sensitivity method (Invitrogen). Confirmation of *P. knowlesi* monoinfection was determined by PCR, as described previously [7]. The relative amount of parasite DNA and human DNA in each sample was determined using a qPCR protocol using primers and probes specific for each species [35–37]. Pure human and *P. knowlesi* standards (10 points) were included to determine the relative concentration (ng/ul) of each organism's DNA in a sample.

Primer design for selective whole genome amplification

The *swga* program (www.github.com/eclarke/swga) was used to identify primer sets that preferentially amplify the *P. knowlesi* genome, providing as input the new A1-H.1 reference for the *P. knowlesi* human-adapted line A1-H.1 [28] and the established human reference human_g1k_v37 (ftp://ftp.1000genomes.ebi.ac.uk). The 10 best sets consist of combinations of 4 to 6 oligonucleotides each, with several overlapping primers and contained two that were present in all sets. The set with the lowest Gini index and perfectly even binding across the genome consists of the following 6 primers: 5'-ATAATC*G*T-3', 5'-ATTATC*G*T-3', 5'-CGAAAT*A*G-3', 5'-CGATAA*A*G-3', 5'-GAATAA*C*G-3' and 5'-TCGTAA*T*A-3'; where asterisks represent phosphorothioate bonds to prevent primer degradation by the exonuclease activity of the Phi29 polymerase

Selective whole genome amplification

Selective whole genome amplification (SWGA) was performed according to the published protocols [24]. All SWGA reactions were carried out in the UV Cabinet for PCR Operations (UV-B-AR, Grant-Bio) to minimize contamination. SWGA reactions were performed containing a maximum of 50ng of total input genomic DNA (and a minimum of 5ng), 5µl of 10x Phi29 DNA Polymerase Reaction Buffer (New England BioLabs), 0.5µl of Purified 100x BSA (New England BioLabs), 0.5µl of 250µM Primer mix of Pkset1, 5µl 10mM dNTP (Roche), 30 units Phi29 DNA Polymerase (New England BioLabs) and Nuclease-Free Water (Ambion, The RNA Company) to reach a final reaction volume of 50µl. The reaction was carried out on a thermocycler with the following step-down program: 5 minutes at 35°C, 10 minutes at 34°C, 15 minutes at 33°C, 20 minutes at 32°C, 25 min 31°C, 16 hours at 30°C and 10 minutes at 65°C. The SWGA samples were diluted 1:1 with EB buffer (Qiagen) and the reaction was purified using the AMPure XP beads (Beckman-Coulter), using a sample to beads ratio of 1:1 according to the protocol.

Whole-genome sequencing, bioinformatics analysis and population genetics

The SWGA products and unamplified DNA were sequenced on an Illumina MiSeq or HiSeq4000. DNA and SWGA Libraries for MiSeq were prepared using the QIAseq FX DNA Library Kit (Qiagen) as per manufacturer's instructions. A 20-minute fragmentation step was optimized for *Plasmodium* samples. For HiSeq runs, libraries were prepared using the NEB Next Ultra DNA Library Prep Kit for Illumina (from New England BioLabs Inc., E7370). All samples were run using 150 bp paired-end reads. The raw sequence data for the isolates was aligned against the new reference for the human-adapted line A1-H.1 (no regions were excluded for analysis)[28] using *bwa-mem* software [38]. WGS data from an extra 156 publicly available samples were also used for analysis (sourced from [21,22,39], where sequencing accession numbers are listed). SNPs were called using the *Samtools* software suite [40], and those of high quality were retained using previously described methods[20]. Samples with high levels of multiplicity of infection (detecting using estMOI [41]) or statistical outliers established using a principal component analysis (PCA) were removed. For comparisons between populations, we applied the PCA and neighbour-joining tree clustering approaches using a matrix of pairwise identity by state values calculated from the SNPs. Nucleotide diversity (π) was calculated using in-house R scripts. R-base scripts were used to calculate linear regressions for the *SNP* π plots.

REFERENCES

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

2. Shearer, F. M. *et al.* Estimating Geographical Variation in the Risk of Zoonotic Plasmodium knowlesi Infection in Countries Eliminating Malaria. *PLoS Negl. Trop. Dis.* **10,** e0004915 (2016).

3. William, T. *et al.* Changing epidemiology of malaria in Sabah, Malaysia: increasing incidence of Plasmodium knowlesi. *Malar. J.* **13,** 390 (2014).

4. World Health Organization Malaria Policy Advisory Commitee. *Outcomes from the Evidence Review Group on Plasmodium knowlesi*. (2017).

5. Daneshvar, C. *et al.* Clinical and Laboratory Features of Human Plasmodium knowlesi Infection. *Clin. Infect. Dis.* **49,** 852–860 (2009).

6. Grigg, M. J. *et al.* Age-Related Clinical Spectrum of Plasmodium knowlesi Malaria and Predictors of Severity. *Clin. Infect. Dis.* **67,** 350–359 (2018).

7. Lubis, I. N. D. *et al.* Contribution of Plasmodium knowlesi to Multispecies Human

Malaria Infections in North Sumatera, Indonesia. *J. Infect. Dis.* **215,** 1148–1155 (2017).

8. Herdiana, H. *et al.* Two clusters of Plasmodium knowlesi cases in a malaria elimination area, Sabang Municipality, Aceh, Indonesia. *Malar. J.* **17,** 186 (2018).

9. Rajahram, G. S. *et al.* Falling Plasmodium knowlesi Malaria Death Rate among Adults despite Rising Incidence, Sabah, Malaysia, 2010-2014. *Emerg. Infect. Dis.* **22,** 41–48 (2016).

10. Brock, P. M. *et al.* Plasmodium knowlesi transmission: integrating quantitative approaches from epidemiology and ecology to understand malaria as a zoonosis. *Parasitology* **143,** 389–400 (2016).

11. Fornace, K. M. *et al.* Association between Landscape Factors and Spatial Patterns of Plasmodium knowlesi Infections in Sabah, Malaysia. *Emerg. Infect. Dis.* **22,** 201–208 (2016).

12. Grigg, M. J. *et al.* Individual-level factors associated with the risk of acquiring human Plasmodium knowlesi malaria in Malaysia: a case-control study. *Lancet. Planet. Heal.* **1,** e97–e104 (2017).

13. Imai, N., White, M. T., Ghani, A. C. & Drakeley, C. J. Transmission and Control of Plasmodium knowlesi: A Mathematical Modelling Study. *PLoS Negl. Trop. Dis.* **8,** e2978 (2014).

14. Pearson RD, Amato R, Auburn S, Miotto O, Almagro-Garcia J, Amaratunga, et al. Genomic analysis of local variation and recent evolution in Plasmodium vivax. *Nat Genet* **in pess,** 959–964 (2016).

15. Gomes, A. R. *et al.* Genetic diversity of next generation antimalarial targets: A baseline for drug resistance surveillance programmes. *Int. J. Parasitol. Drugs Drug Resist.* **7,** 174–

180 (2017).

16. Ravenhall, M. *et al.* Characterizing the impact of sustained sulfadoxine/pyrimethamine use upon the Plasmodium falciparum population in Malawi. *Malar. J.* **15,** 575 (2016).

17. Diez Benavente, E. *et al.* Genomic variation in Plasmodium vivax malaria reveals regions under selective pressure. *PLoS One* **12,** e0177134 (2017).

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

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

20. Benavente, E. D. *et al.* Analysis of nuclear and organellar genomes of Plasmodium knowlesi in humans reveals ancient population structure and recent recombination among host-specific subpopulations. *{PLOS} Genet.* **13,** e1007008 (2017).

21. Pinheiro, M. M. *et al.* Plasmodium knowlesi Genome Sequences from Clinical Isolates Reveal Extensive Genomic Dimorphism. *PLoS One* **10,** e0121303 (2015).

22. Assefa, S. *et al.* Population genomic structure and adaptation in the zoonotic malaria parasite Plasmodium knowlesi. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 13027–13032 (2015).

23. Sundararaman, S. A. *et al.* Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading. *Nat. Commun.* **7,** 1–14 (2016).

24. Oyola, S. O. *et al.* Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. *Malar. J.* **15,** 597 (2016).

25. Cowell, A. N. *et al.* Selective Whole-Genome Amplification Is a Robust Method That Enables Scalable Whole-Genome Sequencing of Plasmodium vivax from Unprocessed

Clinical Samples. *MBio* **8,** (2017).

26. Leichty, A. R. & Brisson, D. Selective whole genome amplification for resequencing target microbial species from complex natural samples. *Genetics* **198,** 473–81 (2014).

27. Clarke, E. L. *et al.* Swga: a Primer Design Toolkit for Selective Whole Genome Amplification. *Bioinformatics* 1–7 (2017). doi:10.1093/bioinformatics/btx118

28. Benavente, E. D. *et al.* A reference genome and methylome for the Plasmodium knowlesi A1-H.1 line. *Int. J. Parasitol.* (2017). doi:10.1016/J.IJPARA.2017.09.008

29. Moon, R. W. *et al.* Normocyte-binding protein required for human erythrocyte invasion by the zoonotic malaria parasite Plasmodium knowlesi. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 7231–7236 (2016).

30. Divis, P. C. S. *et al.* Admixture in Humans of Two Divergent Plasmodium knowlesi Populations Associated with Different Macaque Host Species. *PLOS Pathog.* **11,** e1004888 (2015).

31. Divis, P. C. S. *et al.* Three Divergent Subpopulations of the Malaria Parasite Plasmodium knowlesi. *Emerg. Infect. Dis.* **23,** 616–624 (2017).

32. Lim, K. L., Amir, A., Lau, Y. L. & Fong, M. Y. The Duffy binding protein (PkDBPαII) of Plasmodium knowlesi from Peninsular Malaysia and Malaysian Borneo show different binding activity level to human erythrocytes. *Malar. J.* **16,** 331 (2017).

33. Barber, B. E. *et al.* A Prospective Comparative Study of Knowlesi, Falciparum, and Vivax Malaria in Sabah, Malaysia: High Proportion With Severe Disease From Plasmodium Knowlesi and Plasmodium Vivax But No Mortality With Early Referral and Artesunate Therapy. *Clin. Infect. Dis.* **56,** 383–397 (2013).

34. Fong, M. Y., Wong, S. S., Silva, J. R. De & Lau, Y. L. Genetic polymorphism in domain I of the apical membrane antigen-1 among Plasmodium knowlesi clinical isolates from Peninsular Malaysia. *Acta Trop.* **152,** 145–150 (2015).

35. Auburn, S. *et al.* An Effective Method to Purify Plasmodium falciparum DNA Directly from Clinical Blood Samples for Whole Genome High-Throughput Sequencing. *PLoS One* **6,** e22213 (2011).

36. Divis, P. C., Shokoples, S. E., Singh, B. & Yanow, S. K. A TaqMan real-time PCR assay for the detection and quantitation of Plasmodium knowlesi. *Malar. J.* **9,** 344 (2010).

37. Reller, M. E., Chen, W. H., Dalton, J., Lichay, M. A. & Dumler, J. S. Multiplex 5' nuclease quantitative real-time PCR for clinical diagnosis of malaria and species-level identification and epidemiologic evaluation of malaria-causing parasites, including Plasmodium knowlesi. *J. Clin. Microbiol.* **51,** 2931–8 (2013).

38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

39. Divis, P. C. S., Duffy, C. W., Kadir, K. A., Singh, B. & Conway, D. J. Genome-wide mosaicism in divergence between zoonotic malaria parasite subpopulations with separate sympatric transmission cycles. *Mol. Ecol.* **27,** 860–870 (2018).

40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

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

ACKNOWLEDGMENTS

We thank Don Van Schalkwyk for providing some biological materials from the *P. knowlesi* reference strain. AI is funded by a UK Medical Research Council LiD PhD studentship. RWM is supported by an MRC Career Development Award (MR/M021157/1) jointly funded by the UK Medical Research Council and UK Department for International Development. TGC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no. BB/R013063/1). SC is funded by Medical Research Council UK grants (MR/M01360X/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no. BB/R013063/1). The Sabah studies were funded by the Malaysian Ministry of Health (grant number BP00500420) and the Australian National Health and Medical Research Council (grant numbers 1037304, 1132975 and 1045156) and fellowships (1042072 and 1135820 to NMA, 1088738 to BEB, and 1138860 to MJG). We gratefully acknowledge the Scientific Computing Group for data management and compute infrastructure at Genome Institute of Singapore for their help. The MRC eMedLab computing resource was used for bioinformatics and statistical analysis.

AUTHOR CONTRIBUTIONS

SC and TGC conceived and directed the project. LYL, TW and NA coordinated sample collection. ARG, SC, JRDS, MG, AR, AI, TY, SA, AP, and RWM undertook sample collection, processing and DNA extraction. PFdS, MLH, AP and SC coordinated sequencing. EDB performed bioinformatic and statistical analyses under the supervision of TGC and SC. TGC, SC and EDB interpreted results. EDB, TGC and SC wrote the first draft of the manuscript. All authors commented and edited on various versions of the draft manuscript and approved the final manuscript. EDB, TGC and SC compiled the final manuscript.

DISCLOSURE DECLARATION

The authors declare that they have no conflicts of interest.

Table 1

Comparison of whole genome sequencing before and after parasite enrichment using SWGA

(Table on next page)

Sample ID	Parasitemia p/µl* (%)	Sample type	Reads aligned to *P. knowlesi* reference (%)*	% genome with coverage>5	% genes with coverage>5	% intergenic regions with coverage>5	Mean coverage	Total N SNPs
1	320 (0.006%)	No SWGA	2.45	0.64	0.80	0.52	1.71	1,797
		SWGA	12.11	19.08	23.36	16.51	11.94	59,031
2	539 (0.01%)	No SWGA	0.81	0.04	0.03	0.04	1.28	15
		SWGA	3.66	6.49	7.54	5.97	4.38	14,746
3	851 (0.017%)	No SWGA	3.15	3.88	4.69	3.34	2.25	11,974
		SWGA	27.95	43.32	53.40	37.14	17.52	143,483
4	1581 (0.03%)	No SWGA	1.14	0.08	0.08	0.09	1.34	127
		SWGA	20.37	11.79	14.59	10.05	8.96	34,314
5	3554 (0.07%)	No SWGA	1.22	0.32	0.35	0.30	1.58	628
		SWGA	6.31	28.01	35.21	23.32	6.51	79,139
6	5300 (0.1%)	No SWGA	2.31	3.38	4.34	2.66	2.18	10,479
		SWGA	17.10	48.26	59.66	41.19	15.71	159,652
7	5875 (0.11%)	No SWGA	1.87	0.28	0.31	0.26	1.55	609
		SWGA	20.26	54.74	66.95	47.41	18.72	179,304
8	10634 (0.2%)	No SWGA	2.34	2.67	3.55	2.00	2.10	8,147
		SWGA	22.33	60.50	73.96	52.40	23.61	208,202
9	ND	No SWGA	10.59	2.51	2.58	2.56	2.14	2,197
		SWGA	44.54	32.03	38.01	28.87	8.97	119,157
10	26368 (0.5%)	No SWGA	11.01	31.36	36.22	28.77	4.05	104,805
		SWGA	42.21	48.26	59.89	40.99	18.04	162,920

*These results are from single runs, and not pooled samples (average of a total of 2 billion bp sequenced per sample (human and parasite)).

FIGURES

Figure 1

Correlation of parasitaemia (%) and genome coverage (> 5 reads) in amplified DNA samples.

Parasitaemia data was available for 13 amplified samples used in this study. An increase in coverage is observed with samples with higher parasitaemias (R-squared = 0.6).

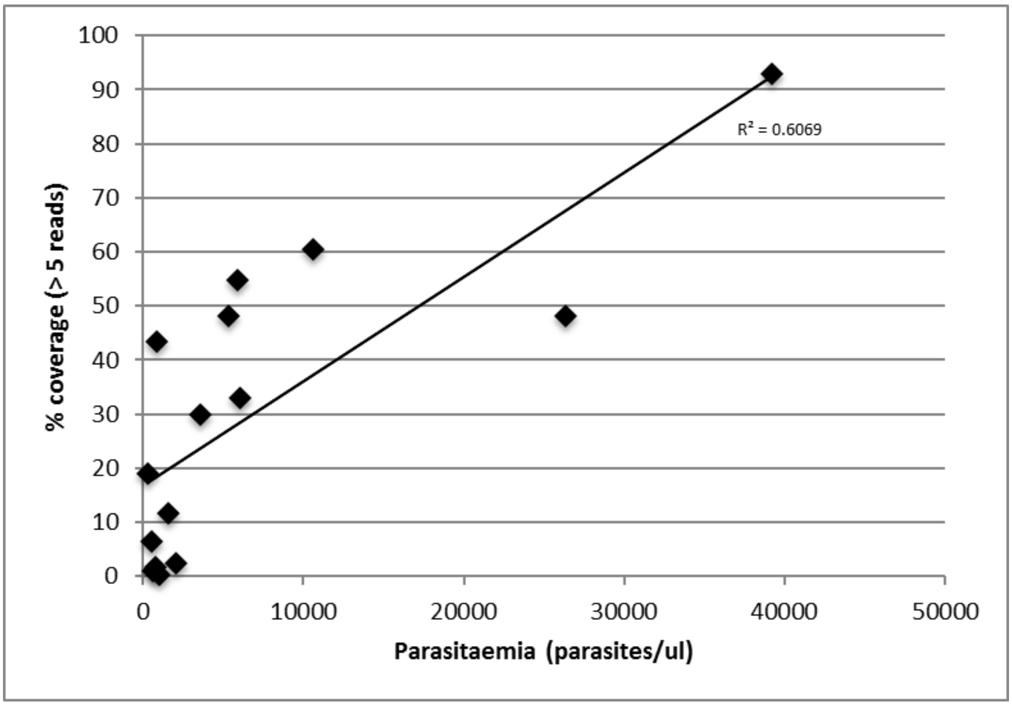


Figure 2

Neighbour-Joining tree for 103 *P. knowlesi* isolates shows expected grouping into three clusters.

The tree shows the expected split into three different clusters associated with: (i) Peninsular Malaysia (purple in tips), (ii) Malaysian Borneo *Macaca nemestrina* (*Mn-Pk*, green) and (iii) Malaysian Borneo *M. fascicularis* (*Mf-Pk*, blue). The tree also shows the correct positioning of the 4 newly generated Peninsular isolates (in red) within the Peninsular cluster, and the clustering of the 16 new Malaysian Borneo isolates from Sabah (in orange) within either of the *Mf-Pk or Mn-Pk* associated clusters.

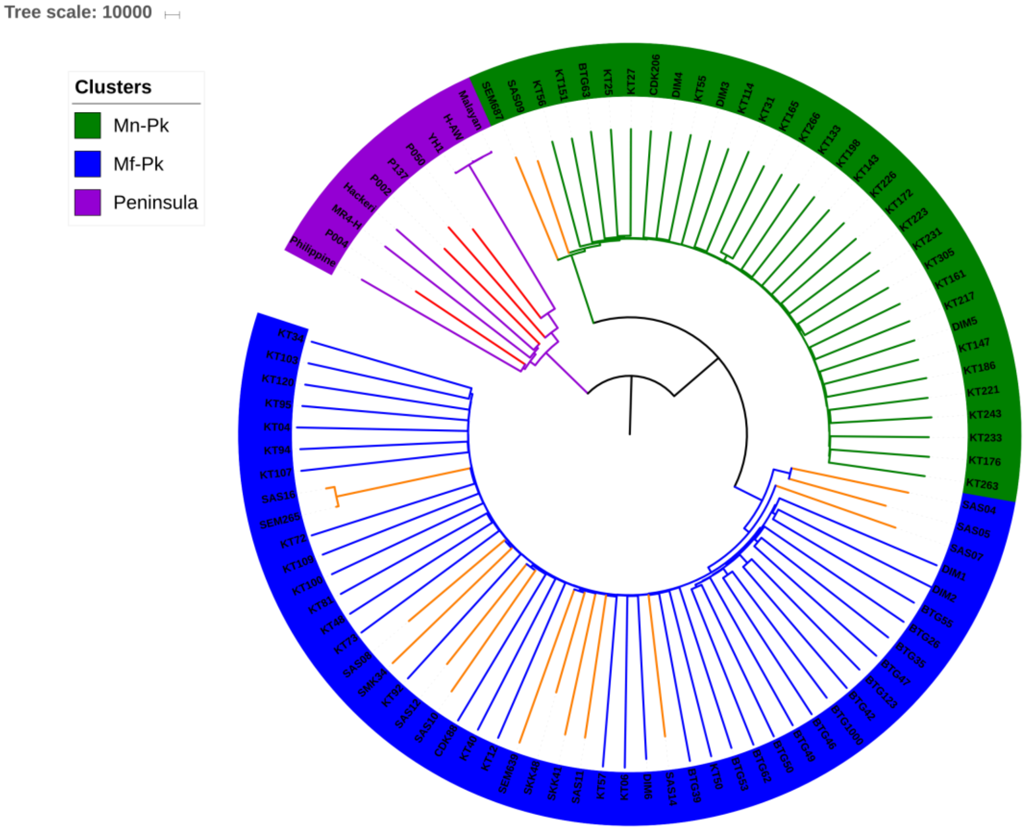
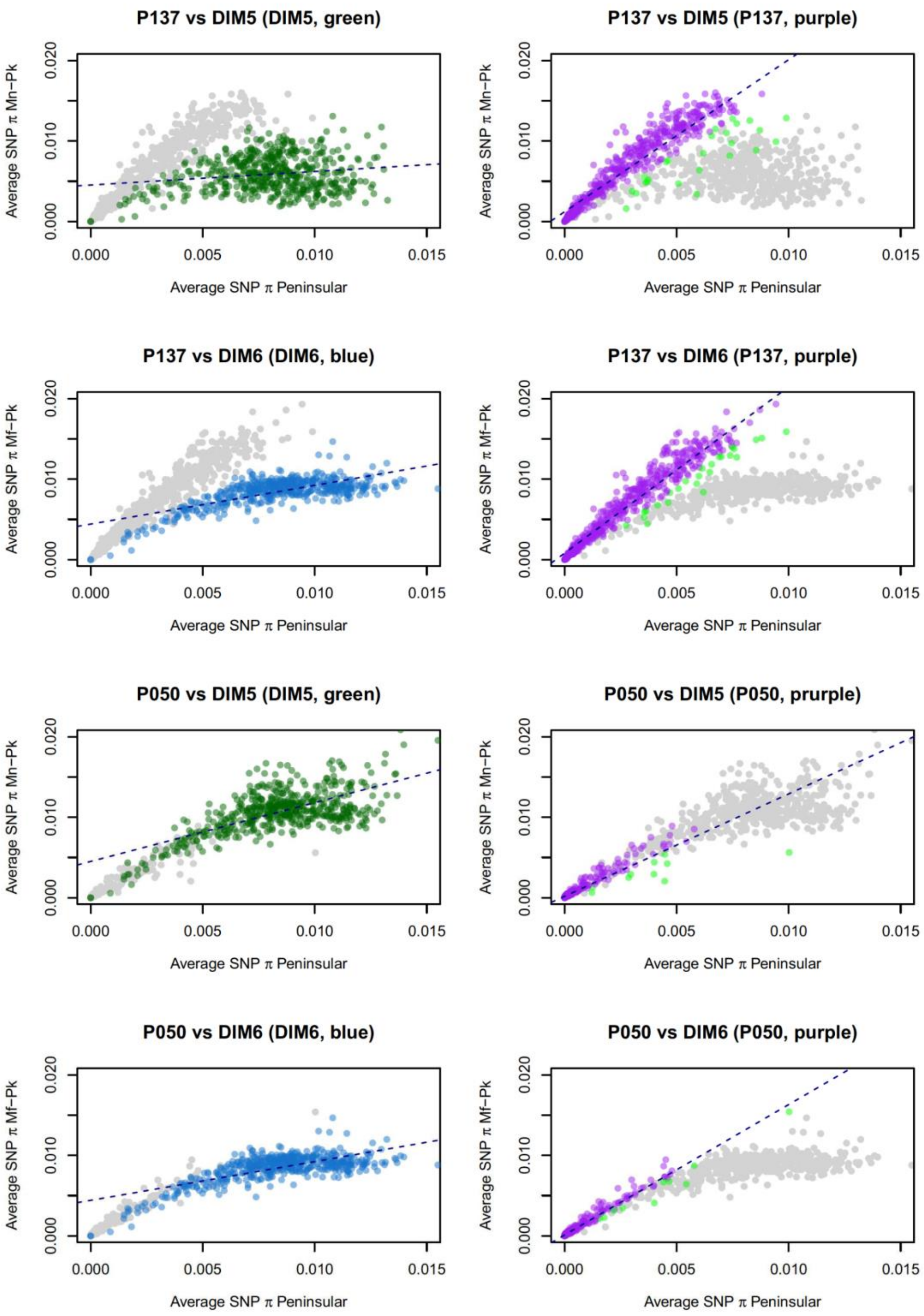


Figure 3

***P. knowlesi* isolates from Peninsular Malaysia (P137 and P050) present introgression events from *Mn-Pk* sub-population**. Peninsular isolate P137 was compared to DIM5 (**top two panels**) as a representative of the *Mn-Pk* cluster, and DIM6 (**second row panels**) of the *Mf-Pk* cluster. Isolate P050 was compared to the same isolates in the bottom 4 panels. On the top most left panel each green dot represents a 50 kbp section of the DIM5 genome. Its position on the X-axis is defined by the average SNP *π* obtained by comparing its sequence in a pairwise manner to the same genome fragment in each isolate in the *Mn-Pk* cluster, and in the Y-axis the average SNP π is compared to the same fragment of the Peninsular isolates. This average SNP π defines the similarity of each dot to the different clusters. The top most right panel represents the same data as the top left most panel with the P137 50Kb fragments highlighted in purple for clarity. The same analysis was conducted in the second row of panels but using an *Mf-Pk* cluster isolate and using the average SNP *π* to *Mf-Pk* as the X-axis. The dashed line represents the linear regression for the coloured dots in each plot, and the regions of interest were identified (in light green in the right panels) by finding the fragments of the Peninsular isolate genomes that presented low similarity to the Peninsular cluster and high similarity to either the *Mn-Pk* (green) or *Mf-Pk* (blue). This approach accounts for the highest residuals. After further filtering by exploration of individual genes, the results were reported in **Supplementary Table 2**.

(Figure on next page)



P137 vs DIM5 (DIM5, green)

P137 vs DIM5 (P137, purple)

P137 vs DIM6 (DIM6, blue)

P137 vs DIM6 (P137, purple)

P050 vs DIM5 (DIM5, green)

P050 vs DIM5 (P050, prurple)

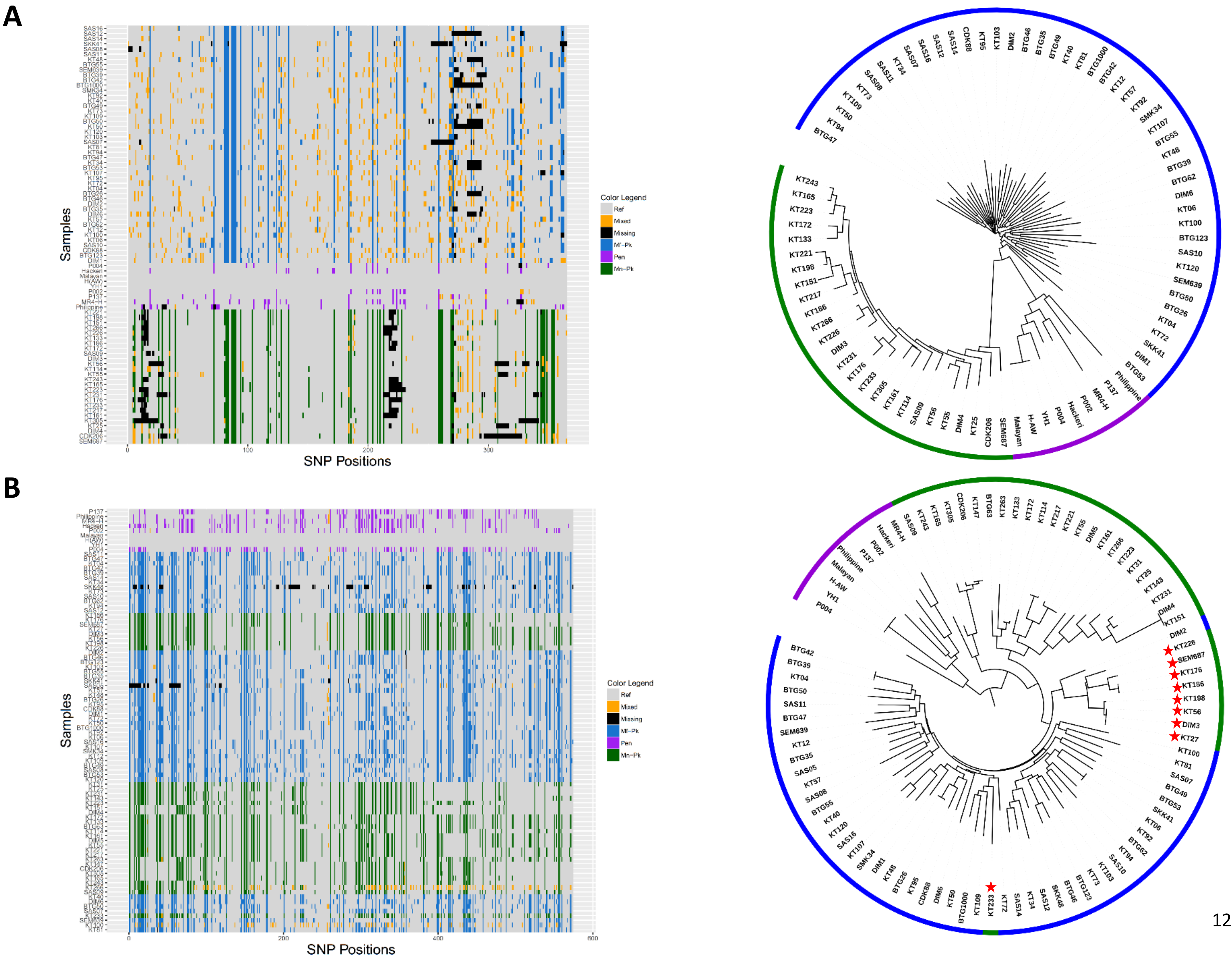
P050 vs DIM6 (DIM6, blue)

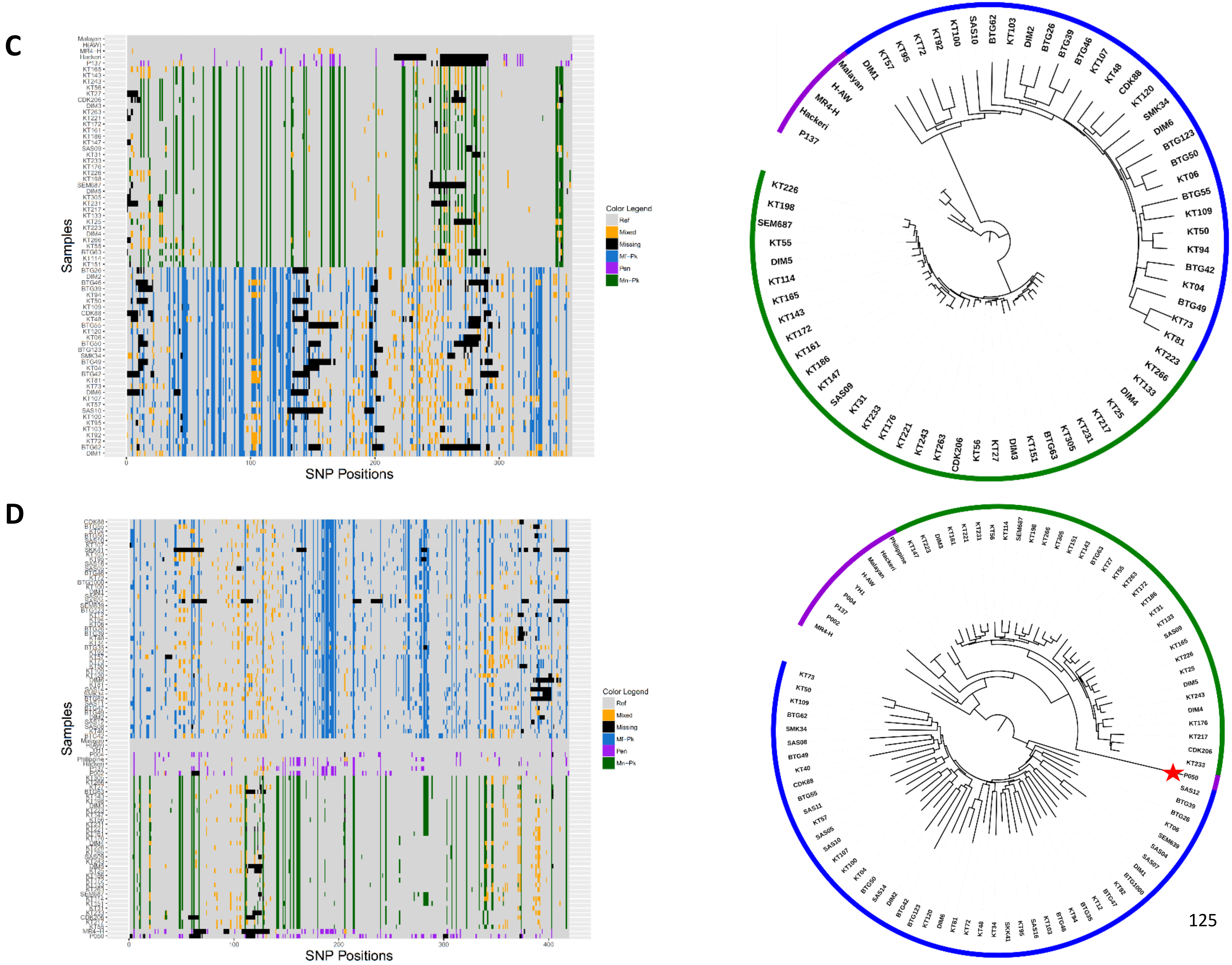
P050 vs DIM6 (P050, purple)

Figure 4

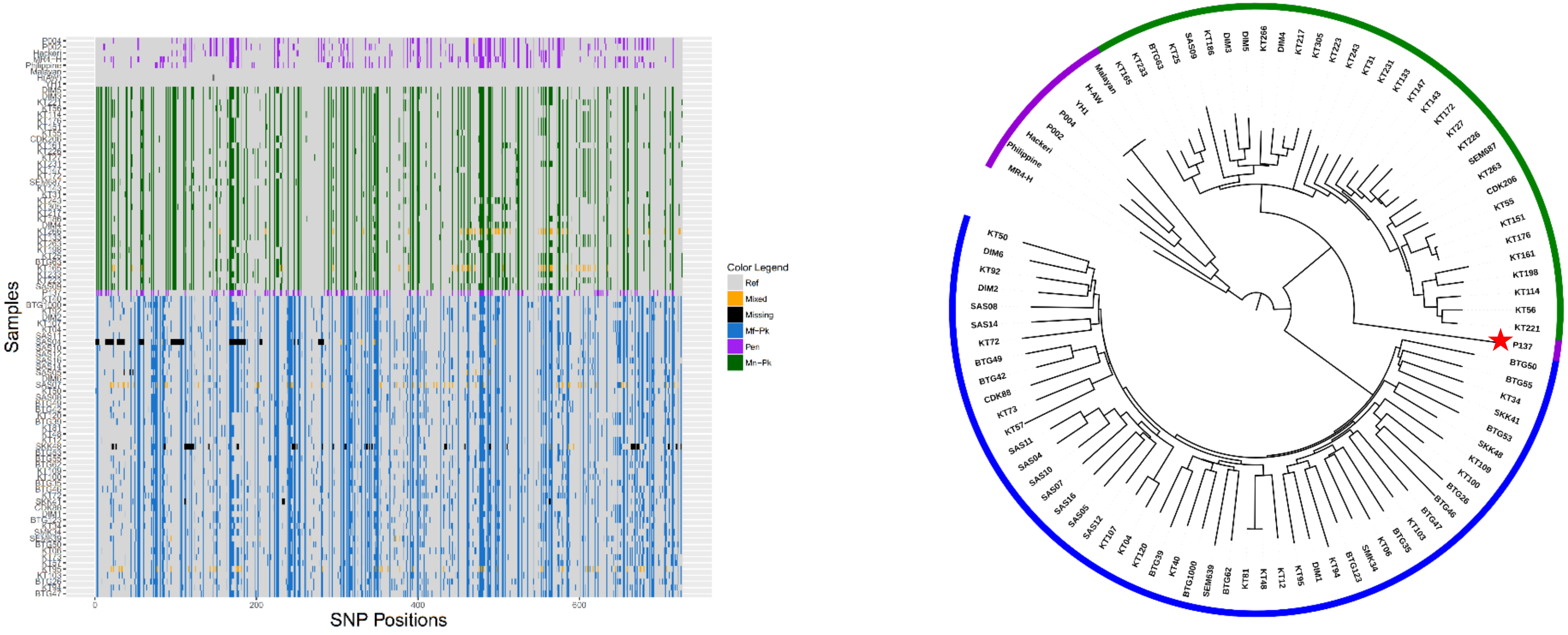
Haplotype plot and neighbour-joining tree for 5 invasion genes ((A) *DBPα, (B) NBPXb, (C) DBPβ, (D) DBPγ and (E) NBPXa*) provides insight into population dynamics of the different haplotypes. A strong genetic divergence of the sequences from the different clusters was found for each of the 5 genes and the Peninsular cluster presented the highest diversity in all 5 genes (A-E). In the case of *DBPα* (A, right) and *DBPβ* (C, right) gene sequences clustered in concordance with the whole genome pattern. In the case of *DBPγ* (D, right), the sample P050 showed clear evidence of having been introgressed from the *Mn-Pk* cluster (green, highlighted with a red star). This can also be confirmed by the partial similarity of the haplotype for sample P050 with the *Mn-Pk* (D, left). For the *NBPXa* gene (E), an introgression event can be seen from Mn-Pk cluster for the isolate P137. The *NBPXb* gene (B, right) presented a fairly distinct pattern of diversity. The clusters have a small genetic distance between themselves, making the separation between them less obvious and 9 out of 33 (27%) of the *Mn-Pk* cluster isolates presented with *Mf-Pk* type haplotypes (highlighted with red stars).

(Figure on pages 123 [A and B], 124 [C and D] and 125 [E])


A
Samples
SNP Positions
0
100
200
300
Color Legend
Ref
Mixed
Missing
Mf-Pk
Pen
Mn-Pk
BTG47
KT94
KT50
KT109
KT73
SAS08
SAS11
KT34
SAS07
SAS16
SAS12
SAS14
CDK88
KT95
KT103
DIM2
BTG46
BTG35
BTG49
KT40
KT81
BTG1000
BTG42
KT12
KT57
KT92
SMK34
KT107
BTG55
KT48
BTG39
BTG62
DIM6
KT06
KT100
BTG123
SAS10
KT120
SEM639
BTG50
BTG26
KT04
KT72
SKK41
DIM1
BTG53
Philippine
P137
MR4-H
P002
Hackeri
P004
YH1
H-AW
Malayan
SEM687
CDK206
KT25
DIM4
KT55
KT56
SAS09
KT114
KT161
KT305
KT233
KT176
KT231
DIM3
KT226
KT266
KT186
KT217
KT151
KT198
KT221
KT133
KT172
KT223
KT165
KT243
B
Samples
SNP Positions
0
200
400
600
Color Legend
Ref
Mixed
Missing
Mf-Pk
Pen
Mn-Pk
P004
YH1
H-AW
Malayan
Philippine
P137
P002
Hackeri
MR4-H
SAS09
KT243
KT165
KT305
CDK206
KT147
BTG63
KT263
KT133
KT172
KT114
KT217
KT221
KT55
DIM5
KT161
KT266
KT223
KT31
KT25
KT143
KT231
DIM4
KT151
DIM2
KT226
SEM687
KT176
KT186
KT198
KT56
DIM3
KT27
KT100
KT81
SAS07
BTG49
BTG53
SKK41
KT06
KT92
BTG62
KT94
SAS10
KT103
KT73
BTG123
BTG46
SKK48
SAS12
KT34
SAS14
KT72
KT233
KT109
BTG1000
KT50
DIM6
CDK88
KT95
BTG26
KT48
DIM1
SMK34
KT107
SAS16
KT120
KT40
BTG55
SAS08
KT57
SAS05
BTG35
KT12
SEM639
BTG47
SAS11
BTG50
KT04
BTG39
BTG42


C
Samples
SNP Positions
0
100
200
300
Color Legend
Ref
Mixed
Missing
Mf−Pk
Pen
Mn−Pk
Malayan
H(AW)
MR4_H
Hackeri
P137
KT165
KT143
KT243
KT56
KT27
CDK206
DIM3
KT263
KT221
KT172
KT161
KT186
KT147
SAS09
KT31
KT233
KT176
KT226
KT198
SEM687
DIM5
KT305
KT231
KT217
KT133
KT25
KT223
DIM4
KT266
KT55
BTG63
KT114
KT151
BTG26
DIM2
BTG46
BTG39
KT94
KT50
KT109
CDK88
KT48
BTG55
KT120
KT06
BTG50
BTG123
SMK34
BTG49
KT04
BTG42
KT81
KT73
DIM6
KT107
KT57
SAS10
KT100
KT95
KT103
KT92
KT72
BTG62
DIM1
Malayan
H-AW
MR4-H
Hackeri
P137
KT226
KT198
SEM687
KT55
DIM5
KT114
KT165
KT143
KT172
KT161
KT186
KT147
SAS09
KT31
KT233
KT176
KT221
KT243
KT263
CDK206
KT56
KT27
DIM3
KT151
BTG63
KT305
KT231
KT217
KT25
DIM4
KT133
KT266
KT223
KT81
KT73
BTG49
KT04
BTG42
KT94
KT50
KT109
BTG55
KT06
BTG50
BTG123
DIM6
SMK34
KT120
CDK88
KT48
KT107
BTG46
BTG39
BTG26
DIM2
KT103
BTG62
SAS10
KT100
KT92
KT72
KT95
KT57
DIM1
D
Samples
SNP Positions
0
100
200
300
400
Color Legend
Ref
Mixed
Missing
Mf−Pk
Pen
Mn−Pk
Philippine
Hackeri
Malayan
H-AW
YH1
P004
P137
P002
MR4-H
KT73
KT50
KT109
BTG62
SMK34
SAS08
BTG49
KT40
CDK88
BTG55
SAS11
KT57
SAS05
SAS10
KT107
KT100
KT04
BTG50
SAS14
DIM2
BTG42
BTG123
KT120
DIM6
KT81
KT72
KT48
KT34
SKK41
KT95
SAS16
KT103
BTG46
KT94
BTG35
KT12
BTG47
KT92
BTG1000
DIM1
SAS07
SAS04
SEM639
KT06
BTG26
BTG39
SAS12
P050
KT233
CDK206
KT217
KT176
DIM4
KT243
DIM5
KT25
KT226
KT165
SAS09
KT133
KT31
KT186
KT172
KT263
KT55
KT27
BTG63
KT143
KT151
KT305
KT266
KT198
SEM687
KT114
KT56
KT231
KT221
KT161
DIM3
KT223
KT147

E



Samples
SNP Positions
0
200
400
600
Color Legend
Ref
Mixed
Missing
Mf-Pk
Pen
Mn-Pk
KT165
KT233
BTG63
KT25
SAS09
KT186
DIM3
DIM5
KT266
DIM4
KT217
KT305
KT223
KT243
KT31
KT231
KT133
KT147
KT143
KT172
KT27
KT226
SEM687
KT263
CDK206
KT55
KT151
KT176
KT161
KT198
KT114
KT56
KT221
P137
BTG50
BTG55
KT34
SKK41
BTG53
SKK48
KT109
KT100
BTG26
BTG46
BTG47
KT103
BTG35
KT06
SMK34
BTG123
KT94
DIM1
KT95
KT112
KT148
KT81
BTG62
SEM639
BTG1000
KT40
BTG39
KT120
KT04
KT107
SAS12
SAS05
SAS16
SAS07
SAS10
SAS04
SAS11
KT57
KT73
CDK88
BTG42
BTG49
KT72
SAS14
SAS08
DIM2
KT92
DIM6
KT50
MR4-H
Philippine
Hackeri
P002
P004
YH1
H-AW
Malayan

Supplementary table 1

The samples analysed

Sample	**Code**	**Area**	**estMOI****	**Group***
ERR274221	DIM1	Sarikei	0.8	*Mf-Pk*
ERR274222	DIM2	Sarikei	0.7	*Mf-Pk*
ERR274224	DIM3	Sarikei	0.6	*Mn-Pk*
ERR274225	DIM4	Sarikei	0.5	*Mn-Pk*
ERR366425	DIM5	Sarikei	0.3	*Mn-Pk*
ERR366426	DIM6	Sarikei	0.2	*Mf-Pk*
ERR985372	BTG1000	Betong	0.8	*Mf-Pk*
ERR985373	BTG123	Betong	0.7	*Mf-Pk*
ERR985374	BTG26	Betong	1.6	*Mf-Pk*
ERR985375	BTG35	Betong	6.8	*Mf-Pk*
ERR985376	BTG39	Betong	9.9	*Mf-Pk*
ERR985377	BTG42	Betong	1.0	*Mf-Pk*
ERR985378	BTG46	Betong	1.5	*Mf-Pk*
ERR985379	BTG47	Betong	1.2	*Mf-Pk*
ERR985380	BTG49	Betong	5.9	*Mf-Pk*
ERR985381	BTG50	Betong	1.4	*Mf-Pk*
ERR985382	BTG53	Betong	1.3	*Mf-Pk*
ERR985383	BTG55	Betong	3.6	*Mf-Pk*
ERR985384	BTG62	Betong	3.0	*Mf-Pk*
ERR985385	CDK88	Kapit	1.1	*Mf-Pk*
ERR985386	KT03	Kapit	22.8	*Mf-Pk*
ERR985387	KT04	Kapit	2.7	*Mf-Pk*
ERR985388	KT06	Kapit	1.0	*Mf-Pk*
ERR985389	KT100	Kapit	2.8	*Mf-Pk*
ERR985390	KT103	Kapit	0.5	*Mf-Pk*

ERR985391	KT107	Kapit	0.6	*Mf-Pk*
ERR985392	KT109	Kapit	0.6	*Mf-Pk*
ERR985393	KT120	Kapit	0.5	*Mf-Pk*
ERR985394	KT12	Kapit	1.0	*Mf-Pk*
ERR985395	KT26	Kapit	50.4	*Mf-Pk*
ERR985396	KT29	Kapit	21.0	*Mf-Pk*
ERR985397	KT30	Kapit	22.3	*Mf-Pk*
ERR985398	KT34	Kapit	1.2	*Mf-Pk*
ERR985399	KT40	Kapit	1.3	*Mf-Pk*
ERR985400	KT48	Kapit	1.0	*Mf-Pk*
ERR985401	KT50	Kapit	1.0	*Mf-Pk*
ERR985402	KT57	Kapit	0.7	*Mf-Pk*
ERR985403	KT72	Kapit	0.6	*Mf-Pk*
ERR985404	KT73	Kapit	1.2	*Mf-Pk*
ERR985405	KT77	Kapit	30.9	*Mf-Pk*
ERR985406	KT81	Kapit	0.5	*Mf-Pk*
ERR985407	KT92	Kapit	0.8	*Mf-Pk*
ERR985408	KT94	Kapit	0.6	*Mf-Pk*
ERR985409	KT95	Kapit	3.5	*Mf-Pk*
ERR985410	BTG44	Betong	32.7	*Mn-Pk*
ERR985411	BTG63	Betong	0.8	*Mn-Pk*
ERR985412	CDK206	Kapit	0.8	*Mn-Pk*
ERR985413	KT114	Kapit	0.5	*Mn-Pk*
ERR985414	KT25	Kapit	0.8	*Mn-Pk*
ERR985415	KT27	Kapit	0.9	*Mn-Pk*
ERR985416	KT31	Kapit	1.2	*Mn-Pk*
ERR985417	KT42	Kapit	29.9	*Mn-Pk*
ERR985418	KT55	Kapit	0.7	*Mn-Pk*
ERR985419	KT56	Kapit	4.5	*Mn-Pk*

PKAP_KK41	SKK41	Sabah	0.3	*Mf-Pk*
PKAP_KK48	SKK48	Sabah	3.0	*Mf-Pk*
PKAP_MK34	SMK34	Sabah	0.8	*Mf-Pk*
PKAP_QEM265	SEM265	Sabah	0.0	*Mf-Pk*
PKAP_QEM639	SEM639	Sabah	9.0	*Mf-Pk*
PKAP_QEM687	SEM687	Sabah	0.6	*Mn-Pk*
PKAS04_com	SAS04	Sabah	0.0	*Mf-Pk*
PKAS05_com	SAS05	Sabah	1.8	*Mf-Pk*
PKAS07_com	SAS07	Sabah	7.5	*Mf-Pk*
PKAS08_com	SAS08	Sabah	0.3	*Mf-Pk*
PKAS09_com	SAS09	Sabah	0.5	*Mn-Pk*
PKAS10_com	SAS10	Sabah	0.6	*Mf-Pk*
PKAS11_com	SAS11	Sabah	0.5	*Mf-Pk*
PKAS12_com	SAS12	Sabah	0.3	*Mf-Pk*
PKAS13_com	SAS13	Sabah	18.0	*Mf-Pk*
PKAS14_com	SAS14	Sabah	0.5	*Mf-Pk*
PKAS15_com	SAS15	Sabah	20.9	*Mf-Pk*
PKAS16_com	SAS16	Sabah	0.5	*Mf-Pk*
SRR2221468	Hackeri	Clinic	1.2	Peninsular
SRR2222335	H(AW)	Clinic	0.0	Peninsular
SRR2225467	Malayan	Clinic	10.7	Peninsular
SRR2225571	MR4-H	Clinic	0.6	Peninsular
SRR2225573	Philippine	Clinic	0.8	Peninsular
SRR3135172	YH1	Clinic	5.6	Peninsular
swga002d	P002	Peninsular	0.8	Peninsular
swga004	P004	Peninsular	3.2	Peninsular
ERR2214837	KT133	Kapit	0.3	*Mn-Pk*
ERR2214838	KT143	Kapit	0.2	*Mn-Pk*
ERR2214839	KT147	Kapit	0.3	*Mn-Pk*

ERR2214840	KT151	Kapit	0.3	*Mn-Pk*
ERR2214841	KT161	Kapit	0.3	*Mn-Pk*
ERR2214842	KT165	Kapit	14.6	*Mn-Pk*
ERR2214843	KT172	Kapit	0.2	*Mn-Pk*
ERR2214844	KT176	Kapit	0.3	*Mn-Pk*
ERR2214845	KT186	Kapit	1.0	*Mn-Pk*
ERR2214846	KT198	Kapit	0.3	*Mn-Pk*
ERR2214847	KT217	Kapit	0.4	*Mn-Pk*
ERR2214848	KT221	Kapit	0.2	*Mn-Pk*
ERR2214849	KT223	Kapit	0.3	*Mn-Pk*
ERR2214850	KT224	Kapit	19.0	*Mn-Pk*
ERR2214851	KT226	Kapit	0.6	*Mn-Pk*
ERR2214852	KT231	Kapit	0.2	*Mn-Pk*
ERR2214853	KT233	Kapit	0.2	*Mn-Pk*
ERR2214854	KT243	Kapit	0.4	*Mn-Pk*
ERR2214855	KT263	Kapit	0.4	*Mn-Pk*
ERR2214856	KT266	Kapit	13.6	*Mn-Pk*
ERR2214857	KT305	Kapit	0.7	*Mn-Pk*
swga_0137_f	P137	Peninsular	0.4	Peninsular
swga_050_f	P050	Peninsular	0.1	Peninsular

* Borneo Malaysia (*M. nemestrina* (*Mn-Pk*) and *M. fascicularis* (*Mf-Pk*) macaques and humans)

****** Percentage of the genome that shows evidence of MOI >1 based on estMOI software [41]

Supplementary table 2

Genetic regions with evidence of introgression* for the P050 and P137 samples

Chr	Start	End	Samples	Genes affected
1	1	50000	P050, P137	*PKNH_0100500 (hypothetical protein, conserved), PKNH_0100600 (CPPUF**), PKNH_0100700 (CPPUF)*
2	750000	787646	P050	*PKNH_0211700 (CPPUF), PKNH_0211800 (NFS), PKNH_0211900 (DNMT)*
7	1	50000	P050	*PKNH_0700700 (CPPUF), PKNH_0700800 (RER1), PKNH_0700900 (RAB7)*
7	1250000	1300000	P137	*PKNH_0728200 (CPPUF), PKNH_0728300 (CPPUF), PKNH_0728400 (CPPUF), PKNH_0728800 (MSP1P), PKNH_0728900 (MSP1), PKNH_0729000 (CPPUF), PKNH_0729100 (diacylglycerol kinase, putative), PKNH_0729200 (CYP72), PKNH_0729300 (CPPUF)*
7	1450000	1500000	P050, P137	*PKNH_0734500 (PEPUF***), PKNH_0734600 (CPPUF), PKNH_0734700 (ETRAMP), PKNH_0734800 (Plasmodium exported protei, PHIST), PKNH_0734900 (PEPUF), PKNH_0735000 (lysophospholipase, putative)*
9	1	50000	P137	*PKNH_0900100 (hypothetical protein), PKNH_0900200 (PEPUF), PKNH_0900300 (hypothetical protein)*
9	50000	100000	P050, P137	*PKNH_0900400 (hypothetical protein), PKNH_0900500 (PEPUF), PKNH_0900600 (CPPUF)*
10	1450000	1500000	P137	*PKNH_1032400 (SBP1), PKNH_1032500 (tryptophan-rich antigen, putative), PKNH_1032600 (PEPUF), PKNH_1032800 (PEPUF)*
11	2300000	2350000	P050, P137	*PKNH_1149000 (CPPUF), PKNH_1149100 (PEPUF), PKNH_1149200 (PEPUF), PKNH_1149400 (PEPUF), PKNH_1149600 (hypothetical protein)*
12	2050000	2100000	P050, P137	*PKNH_1246200 (CPPUF), PKNH_1246500 (PEPUF), PKNH_1246600 (PEPUF)*
12	2100000	2150000	P050, P137	*PKNH_1246800 (PEPUF), PKNH_1246900 (PEPUF), PKNH_1247000 (PEPUF)*
12	2150000	2200000	P050	*PKNH_1247800 (PEPUF), PKNH_1247900 (CPPUF)*
13	1	50000	P050, P137	*PKNH_1300400 (Plasmodium exported protein, PHIST), PKNH_1300500 (tryptophan-rich antigen, putative), PKNH_1300600 (tryptophan-rich antigen, putative),*

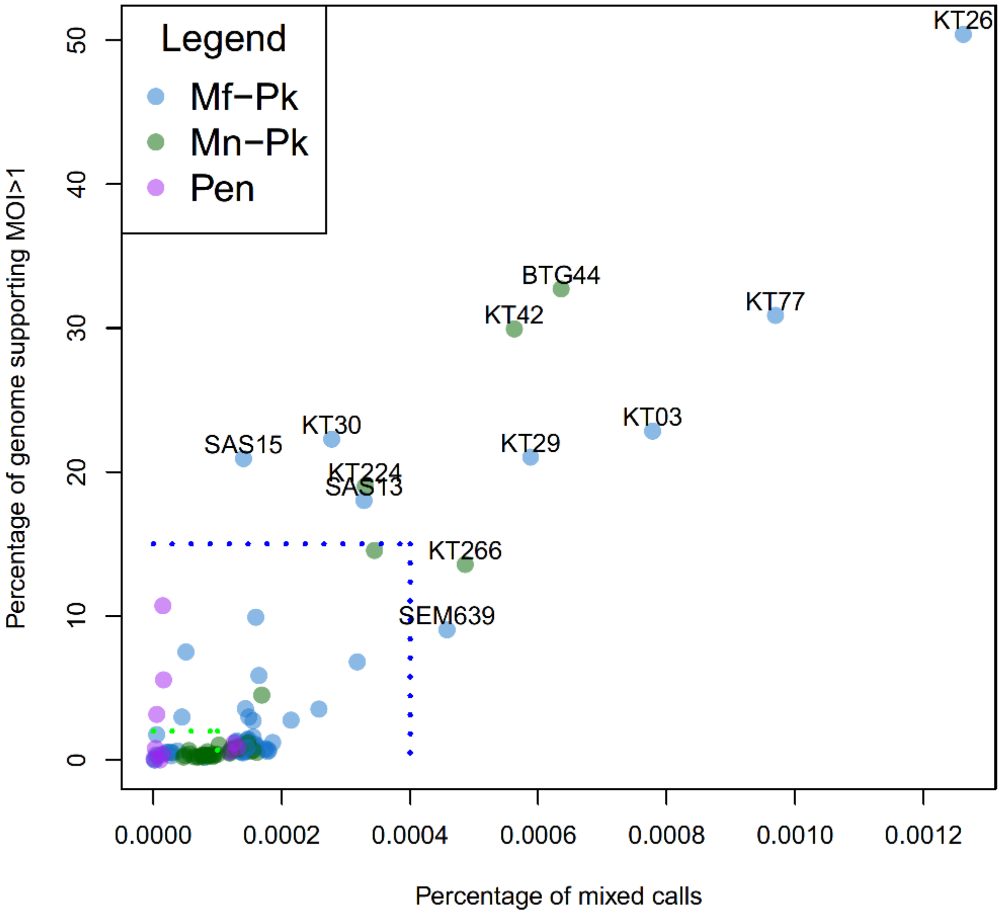
				PKNH_1300700 (tryptophan/threonine-rich antigen, putative), PKNH_1300900 (PEPUF)
13	50000	100000	P050, P137	*PKNH_1301000 (tryptophan-rich antigen, putative), PKNH_1301100 (lysophospholipase, putative), PKNH_1301300 (PPUF), PKNH_1301800 (CPPUF), PKNH_1302100 (mitochondrial phosphate carrier protein, putative)*
13	1200000	1250000	P050, P137	*PKNH_1325500 (CPPUF), PKNH_1325600 (PEPUF), PKNH_1325700 (KAHRP), PKNH_1325800 (PEPUF), PKNH_1325900 (PEPUF), PKNH_1326000 (Plasmodium exported protein, PHIST), PKNH_1326100 (PEPUF), PKNH_1326200 (PEPUF)*
14	1	50000	P050	*PKNH_1400800 (DBPbeta), PKNH_1401000 (PEPUF)*
14	50000	100000	P050	*PKNH_1401100 (PEPUF), PKNH_1401200 (PEPUF), PKNH_1401300 (cytoadherence linked asexual protein, putative), PKNH_1401400 (PEPUF), PKNH_1401500 (lysophospholipase, putative), PKNH_1401600 (PEPUF), PKNH_1401800 (PEPUF), PKNH_1401900 (CPPUF), PKNH_1402000 (PPUF), PKNH_1402100 (PEPUF)*
14	100000	150000	P050	*PKNH_1402200 (PEPUF), PKNH_1402300 (GAP), PKNH_1402400 (ETRAMP), PKNH_1402500 (hypothetical protein), PKNH_1402600 (CPMPUF****)*
14	800000	850000	P137	*PKNH_1417800 (MCM4), PKNH_1417900 (ApiAP2)*
14	3150000	3200000	P050, P137	*PKNH_1472100 (RON3), PKNH_1472200 (CPPUF), PKNH_1472300 (NBPXa)*
14	3200000	3250000	P137	*PKNH_1472400 (tryptophan-rich antigen, putative), PKNH_1472600 (PEPUF), PKNH_1472700 (PEPUF), PKNH_1472800 (Plasmodium exported protei, PHIST), PKNH_1472900 (PEPUF)*

* Cluster Origin of Introgression is Mn-Pk; ** CPPUF: Conserved Plasmodium protein, unknown function; *** PEPUF: Plasmodium exported protein, unknown function; **** CPMPUF: Conserved Plasmodium membrane protein, unknown function

Supplementary figure 1

Estimation of multiplicity of infection (MOI) identifies isolates presenting clear evidence of presenting several clones.

MOI was assessed using overall number of heterozygous calls, as well as the fraction of genome supporting multiplicity >1 obtained using estMOI software [41]. This data supports the ability of the SWGA to identify mixed infections and suggests a non-bias towards a specific clone, as samples SAS15 and SAS13 were SWGA amplified and yet have been classified as presenting MOI >1 by both methods.



Pen = Peninsular; Borneo Malaysia (*M. nemestrina* (Mn-Pk) and *M. fascicularis* (Mf-Pk) macaques and humans)

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Ernest Diez Benavente
Principal Supervisor	Taane Clark & Susana Campino
Thesis Title	**Using whole genome sequence data to study genomic diversity and develop molecular barcodes to profile Plasmodium malaria parasites**

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	Scientific Reports		
When was the work published?	18th October 2018		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	NOT APPLICABLE
Please list the paper's authors in the intended authorship order:	NOT APPLICABLE
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I downloaded the raw data from public repositories and collected from collaborators. I created the pipeline in which I ran the samples through and performed the QC analysis, as well as the interpretation analysis. I later created custom R and perl scripts to process the data in order to obtain information related to coverage, SNPs and other genomic information as well as performed de novo assembly. The figures presented in this work have all been generated using scripts written by myself or publicly available software adapted for this purpose. I wrote the first draft of the manuscript and circulated to co-authors. Once the comments were received I gathered them and made the relevant changes on the article manuscript. I then submitted the work to the Scientific Reports journal, and responded to Review comments accordingly.

Student Signature: ______________________________ **Date:** _________

Supervisor Signature: ______________________________ **Date:** _________

Chapter 5
Global genetic and structural diversity of var2csa in *Plasmodium falciparum* with implications for malaria in pregnancy and vaccine development

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Global genetic diversity of *var2csa* in *Plasmodium falciparum* with implications for malaria in pregnancy and vaccine development

Received: 31 July 2018
Accepted: 6 October 2018
Published online: 18 October 2018

Ernest Diez Benavente[1], Damilola R. Oresegun[2], Paola Florez de Sessions[3], Eloise M. Walker[1], Cally Roper[1], Jamille G. Dombrowski[4], Rodrigo M. de Souza[4,5], Claudio R. F. Marinho[4], Colin J. Sutherland[1], Martin L. Hibberd[1,3], Fady Mohareb[2], David A. Baker[1], Taane G. Clark[1,6] & Susana Campino[1]

Malaria infection during pregnancy, caused by the sequestering of *Plasmodium falciparum* parasites in the placenta, leads to high infant mortality and maternal morbidity. The parasite-placenta adherence mechanism is mediated by the VAR2CSA protein, a target for natural occurring immunity. Currently, vaccine development is based on its ID1-DBL2Xb domain however little is known about the global genetic diversity of the encoding *var2csa* gene, which could influence vaccine efficacy. In a comprehensive analysis of the *var2csa gene* in >2,000 *P. falciparum* field isolates across 23 countries, we found that *var2csa* is duplicated in high prevalence (>25%), African and Oceanian populations harbour a much higher diversity than other regions, and that insertions/deletions are abundant leading to an underestimation of the diversity of the locus. Further, ID1-DBL2Xb haplotypes associated with adverse birth outcomes are present globally, and African-specific haplotypes exist, which should be incorporated into vaccine design.

Malaria infection during pregnancy, caused by *Plasmodium falciparum* parasites, is a major public health burden in tropical areas of Africa and South East Asia, being responsible for substantial maternal and infant morbidity and mortality, including increased adverse outcomes such as miscarriage, maternal anaemia and low birth weight. There is an estimated 200,000 infant and 10,000 maternal deaths per year caused by placental malaria (PM)[1]. The increased susceptibility to *P. falciparum* infection during pregnancy, regardless of previously acquired malaria immunity, has been attributed to the sequestration of infected erythrocytes in the placenta[2,3]. *P. falciparum* blood stage parasites accumulate in the placenta by adhering to chondroitin sulfate A (CSA)[4]. The interaction between infected erythrocytes and the placental syncytium receptor is mediated by the parasite protein VAR2CSA[5]. The extracellular region of this ~350 kDa cysteine–rich transmembrane protein is formed by six Duffy-binding-like (DBL) domains and binds with high specificity to the CSA receptor[6,7]. The protein is encoded by the *var2csa* gene (~10 Kbp in length), which is the most conserved member of the *var* family that encodes the variable antigen *P. falciparum* erythrocyte membrane protein-1 (PfEMP1)[8]. VAR2CSA is preferentially expressed by placental parasites[9]. Primigravid women with a lack of immunity against these subpopulation of parasites are the most affected, but become less susceptible in subsequent pregnancies[5,10,11].

[1]Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom. [2]School of Water, Energy and Environment, Applied Bioinformatics, Cranfield University, Cranfield, United Kingdom. [3]Genomics Institute of Singapore, Biopolis, Singapore. [4]Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil. [5]Multidisciplinary Center, Federal University of Acre, Acre, Brazil. [6]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom. Correspondence and requests for materials should be addressed to T.G.C. (email: taane.clark@lshtm.ac.uk) or S.C. (email: Susana.campino@lshtm.ac.uk)

The full length *var2csa* gene has been identified on chromosome 12, but additional loci have also been found on other chromosomes (e.g. 1, 5–9) both in laboratory and field isolates[12]. However, links to phenotypic advantage have not yet been established[12–14]. Within a single genome, multiple *var2csa* gene copies are not necessarily identical[12], and a high proportion of parasites from infected pregnant women have been found to possess multi-*var2csa* variants, suggesting that having multiple copies may be advantageous during disease progression[14]. Parasites with multiple *var2csa* copies may persist longer during pregnancy by having an increased capacity for antigenic variation and evasion of the maternal immune response[12,14].

Intermittent preventive treatment in pregnancy (IPTp) has contributed to a reduction in PM burden[1], but currently recommended anti-malarial drugs are threatened by high levels of parasite resistance. The development of a placental malaria vaccine is based on naturally occurring immunity, and two VAR2CSA-based candidates are currently in clinical trials[15,16]. Several studies have compared levels of antibody in the sera of primigravid and multigravid women that recognise specific domains of VAR2CSA, including N-terminal VAR2CSA fragments that have high binding affinity for CSA[15–17]. The purpose of any vaccine for PM is to induce immunity in nulligravid women that would confer protection against PM during subsequent infection, and similarly should boost the immunity acquired by multigravid in endemic areas. It remains to be established whether individual VAR2CSA immunogens are able to induce PM protective immunity analogous to that of naturally-acquired immunity.

Studies looking at *var2csa* genetic diversity in field isolates sourced from pregnant infected women, have found that parasites cluster into five different clades based on the CSA-minimal binding site sequence (ID1-DBL2Xb)[18,19]. A clade identified as 3D7-like was associated with the delivery of infants with lower birthweight[18], suggesting that VAR2CSA diversity affects pathogenicity and, by inference, antigenicity. Thus, the intra- and inter-population genetic diversity of the *var2csa* gene is likely to affect the efficacy of any vaccine developed based on the VAR2CSA protein, highlighting the need for genetic diversity studies[20]. Therefore, to systematically evaluate the magnitude of this variation, we assess the genetic diversity and structure of the *var2csa*, and estimate copy number profiles, across more than 2,000 *P. falciparum* field isolates and laboratory strains spanning 23 countries. We find strong evidence that isolates circulating amongst African populations harbour a much higher diversity in the gene compared to other regions, in both nucleotide variants and structural variants, as well as a significantly higher prevalence of parasites encoding two or more different copies. We report for the first time the global distribution of the different ID1-DBL2Xb clades associated with adverse birth outcomes and an unexpectedly high structural variability of the DBL2x domain across the different populations.

Results

Var2csa gene copy numbers in laboratory strains. We investigated the number of *var2csa* gene copies present in 21 genomes from laboratory cultured strains sequenced using the PacBio RS-II long-read sequencing platform. These were Dd2 (IndoChina); KH01 and KH02 (Cambodia); D10 (Papua New Guinea); T9/96 and K1 (Thailand); 7G8 and IT (Brazil); HB3 (x2) (Honduras); 3D7 and NF54 (Africa); GA01 (Gabon); GB4 (Ghana); GN01 (Guinea); CD01 (Congo); KE01 (Kenya); SD01 (Sudan); TG01 (Togo); SN01 (Senegal) and ML01 (Mali)[21]. For each strain, contigs for the *var2csa* gene were extracted from assemblies of high-quality reads, and aligned to the 3D7 reference, known to have only one copy of *var2csa*, to generate a phylogenetic tree (S1 Fig.). Almost all strains have single copies of the *var2csa* gene, except HB3 (confirming[13]), D10 and KH01, which have 2 copies. The HB3 copies are closely related and were more similar to each other (93% similarity) than the D10 (87% similarity) and KH01 (88%) pairs. Isolates TG01 and ML01 isolates presented with evidence of multiplicity of infection (MOI>1) and the extra *var2csa* gene copies observed are thought to belong to the different clones in the sample. Investigation of the sequences found 3,401 unique mutations, spanning 2,632 polymorphic sites from a total of 8,011 sites (excluding gaps and without missing data). Also, 601 unique InDels were found, of which 439 were overlapping, leaving a set of 162 non-overlapping InDels.

Global structural analysis of _var2csa_ gene extra copies. We sought to determine the number of *var2csa* gene copies present in *P. falciparum* field isolates (n = 3,125; 23 countries; including from the Pf3k project (https://www.malariagen.net/projects/pf3k)) and laboratory strains (n = 5) with Illumina sequencing data in the public domain. The analytical pipeline is summarised in S2 Fig. After quality control filtering, a total of 2,099 (67.2%) field isolates with non-high multiplicity of infection (clonal for >70% of genome), low numbers of heterozygous single nucleotide polymorphism (SNPs) (<0.015% of total SNPs) and high genome-wide coverage (>30-fold) were retained (S3 Fig.). By comparing the coverage of the larger N-terminal of *var2csa* exon 2 to the average read coverage across the rest of the resident chromosome (see S4 Fig.), we confirmed the inferred copy numbers for 5 laboratory strains (HB3 and D10 have 2 copies; 3D7, 7G8 and GB4 have 1 copy). The presence of extra and different *var2csa* gene copies manifests itself in heterozygous or mixed genotype signatures. Therefore we compared the estimated number of *var2csa* copies to the proportion of mixed calls present in the gene. There was a clear increase in heterozygous calls in the gene for the samples with additional copies, and the mean proportion of heterozygous calls for samples presenting 1 copy is close to zero (S5 Fig.). This approach confirmed the presence of similar and different *var2csa* copies in HB3 (93% similarity) and D10 (87%), respectively. In the field isolates (n = 2,099) we found geographical differences in the frequency of extra copies, where Oceania (21/24, 88%) was highest, followed by African populations (West Africa (172/489, 35%); East Africa (120/409, 29%)), and the lowest frequency was in South East Asia (235/1108, 21%) (test of proportions P < 0.001) (Table S1).

Characterisation of the _var2csa_ gene and genetic diversity in field isolates. Isolates with a single *var2csa* copy number and low numbers of heterozygous SNPs (<2%) in *var2csa* (S6 Fig.) were identified (n = 1,647), and raw Illumina reads *de-novo* assembled using *velvet* software[22]. A robust multi-software pipeline (see Materials and Methods) led to a high quality multi-alignment of 1,249 sequences, spanning more than 7Kb of the N-terminal end of the *var2csa* gene (Table 1, Fig. 1A and Table S1). The length of non-missing sequence

Population	**n**	**Number sites without missing data**	**Number of Haplotypes**	**Haplotype diversity (*Hd*)**	**Nucleotide diversity (π)**	**Average no. of nucleotide differences (pairwise comparison)**
Burkina Faso	**7**	—	—	—	—	—
Cameroon	**49**	**5317**	**44**	**0.995**	**0.0828**	**440**
Gambia	**31**	**6700**	**21**	**0.963**	**0.0795**	**533**
Ghana	**116**	**6037**	**105**	**0.998**	**0.0798**	**482**
Guinea	**33**	**6553**	**32**	**0.998**	**0.0866**	**567**
Mali	**15**	**6792**	**15**	**1**	**0.0873**	**593**
Nigeria	**3**	—	—	—	—	—
DRC	**77**	**6484**	**71**	**0.998**	**0.0815**	**530**
Kenya	**21**	**6784**	**19**	**0.990**	**0.0821**	**557**
Malawi	**95**	**5953**	**63**	**0.992**	**0.0819**	**488**
Tanzania	**30**	**6578**	**30**	**1**	**0.0892**	**587**
Uganda	**4**	—	—	—	—	—
Madagascar	**10**	**6899**	**10**	**1**	**0.0976**	**673**
Bangladesh	11	6830	10	0.982	0.0896	612
Cambodia	*360*	*4828*	*46*	*0.911*	*0.0696*	*336*
Laos	*40*	*6776*	*29*	*0.979*	*0.0786*	*533*
Myanmar	*72*	*6784*	*27*	*0.949*	*0.0836*	*567*
Thailand	*184*	*6651*	*47*	*0.968*	*0.0809*	*538*
Vietnam	*68*	*6712*	*30*	*0.893*	*0.0794*	*533*
PNG	3	—	—	—	—	—
Brazil	3	—	—	—	—	—
Colombia	10	7045	6	0.889	0.0775	546
Peru	6	—	—	—	—	—

Table 1. Summary of the *Plasmodium falciparum var2csa* assembled 7Kb fragment and its diversity by country using the field isolates with little evidence of mixed infections (n = 1,249). bolded are African countries, italicised are South East Asian countries, DRC = Democratic Republic of Congo; PNG = Papua New Guinea.

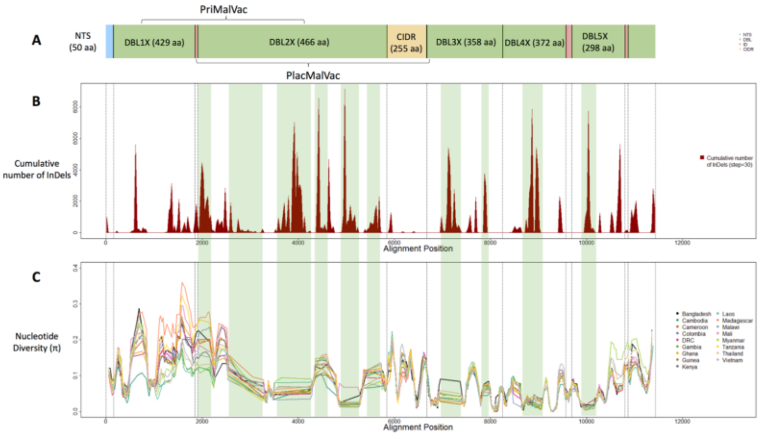


Figure 1. *Plasmodium falciparum var2csa* diversity across a 7 kb region covering the five DBL domains, and 1,249 field isolates. (**A**) Schematic structure of the *var2csa* gene including the N-terminal sequence (NTS, blue), 5 Duffy binding like-Domains (DBL, green), 3 Inter-Domains (ID, red) and the Cysteine-Rich Inter-Domain (CIDR, yellow); the lengths in amino acids of the 3D7 reference are presented in parentheses. (**B**) Accumulation of unique insertions and deletions (InDels) across the *var2csa* gene. (**C**) Distribution of nucleotide diversity (π) across the gene and by population. Regions of abnormal ("flat") nucleotide diversity are highlighted in green (**B** and **C**).

Domain	Length in amino acids in 3D7 reference	Region in alignment	% of InDel positions	Mean length	Standard Deviation (length)	Positions without missing data	Variant Sites	Variant Sites (%)	No. Haplotypes (n = 1,249)	Haplotype Diversity *(Hd)*	Nucleotide diversity *(π)*
NTS	50	1–153	3	151.2	1.3	132	48	36	285	0.977	0.105
DBL1X	429	154–1852	36	1300.0	7.8	856	391	46	436	0.986	0.115
ID1	**14**	**1853–1907**	**98**	**42.6**	**3.9**	**56**	**1**	**2**	**2**	**0.003**	**0.003**
DBL2X	**466**	**1908–5844**	**61**	**1428.4**	**39.7**	**657**	**188**	**29**	**437**	**0.986**	**0.071**
CIDR	**255**	**5845–6670**	**13**	**769.0**	**1.6**	**636**	**207**	**33**	**436**	**0.986**	**0.103**
DBL3X	358	6671–8253	30	1102.2	21.6	890	172	19	449	0.986	0.044
DBL4X	372	8254–9568	27	1103.9	14.9	808	121	15	450	0.986	0.032
ID2	40	9569–9691	0	122.0	0	122	25	20	133	0.958	0.045
DBL5X	298	9692–10799	32	885.3	37.6	668	202	30	414	0.986	0.063
ID3	20	10800–10863	13	58.8	3.3	56	32	57	56	0.919	0.193

Table 2. Diversity statistics across the different domains of the *var2csa* gene in 1,249 isolates of *P. falciparum*. InDel Insertions and deletions; Bolded is the CSA minal binding-domain.

varied per population (median 6,700 bp; inter-quartile range: 6484–6784 bp; Table 1). The assembly pipeline was validated on the GB4, 3D7 and 7G8 Illumina data. When the resulting >7Kb contigs were compared to the Pacbio and capillary sequencing long-read assemblies[23], there was 100% match for all strains (S7 Fig.).

Across the 1,249 samples, we identified 1,387 polymorphic SNPs. Haplotype diversity *(Hd)* was high and invariant between the majority of the different Duffy-Binding-Like (DBL) domains (*Hd* = 0.986), consistent with previous work[19]. There was some evidence of higher haplotype diversity in African populations (mean *Hd* = 0.993) compared to South East Asian populations (mean *Hd* = 0.940) (Table 1, T-test P = 0.03). The nucleotide diversity (π) was more variable across domains, with higher values towards the N-terminus of the protein (Table 2). The DBL2x region of the CSA minimal binding domain had the lowest nucleotide diversity. The diversity trends observed were consistent across populations (Fig. 1C). Consistent with the haplotype diversity result, the overall nucleotide diversity in African populations (mean $\pi = 0.085$) was marginally greater than for South East Asian parasites (mean $\pi = 0.078$), but not statistically significant (T-test P = 0.06).

The presence of insertions and deletions. The *de-novo* assembly of the N-terminal *var2csa* gene fragment, encoding the 5 extra-cellular DBL domains, enabled the study of insertions and deletions (InDels). InDels concentrated around specific regions in the gene (Fig. 1B), where peaks in density coincide with regions of flat nucleotide diversity (highlighted in green in Fig. 1C). The presence of high density, low frequency InDels in the domains leads to an underestimation of both the nucleotide diversity and variation in their length (S8 Fig.). The DBL2X domain has the highest density of InDels (Table 2), with sequence lengths across samples (430 to 550 amino acids) twice those of the 3D7 reference, and greater diversity than the other domains. It is unclear how the diversity created by these InDels might affect the structure of the protein and therefore, both its binding affinity to the CSA receptor during pregnancy and recognition by vaccine-generated antibodies. However, in extreme cases the same DBL domain differs by more than 120 amino acids in length.

Population structure analysis of the ID1-DBL2Xb region. The ID1-DBL2Xb region is the CSA-minimal binding domain and several studies have aimed to characterize its protein structure[18,19]. By combining the published ID1-DBL2Xb sequences (n = 124) with those extracted from the *de-novo* assembled sequences (n = 1,249), we assessed whether the variants led to clustering of *P. falciparum* parasites (Neighbour-Joining tree (Fig. 2A); principal component analysis (Fig. 2B,C)). Four clades were observed, where Clades 1 and 2 have been previously identified as the 3D7-like and FCR3-like[18], respectively (Fig. 2D, Table S2). All geographical regions were represented in Clades 1, 2 and 4, but Clade 3 contains only parasites from Africa (S9 Fig.). The presence of Clade 1 (3D7-like) has been associated with low birthweight in African populations[18], and had highest representation in West African populations (41.7%) (East Africa 27.5%; South East Asia 23.5%). A rarefaction curve analysis of the haplotype diversity in the ID1-DBL2Xb region (S10A,B Fig.) revealed much greater diversity and more unique haplotypes in African parasites when compared to South East Asian populations. This observation is confirmed by the neighbourhood joining tree analysis (S10C–E Fig.), where the individual African populations have much longer distance at the tips than the South East Asian populations.

The open reading frame (ORF) element upstream of the *var2csa* gene is highly conserved. The ORF region of the *var2csa* plays a role in regulating expression of the gene[24]. Analysis of the complete sequences (n = 1,249; S2 Data) revealed that these elements harbour little diversity and are highly conserved across geographical regions and populations (mean $\pi = 0.02$) (Table S3). The clustering of *var2csa* genes into four clades is not complemented by a similar pattern in the ORF regulatory region, which seems to indicate that the gene expression regulatory function of this region might be conserved across clades.

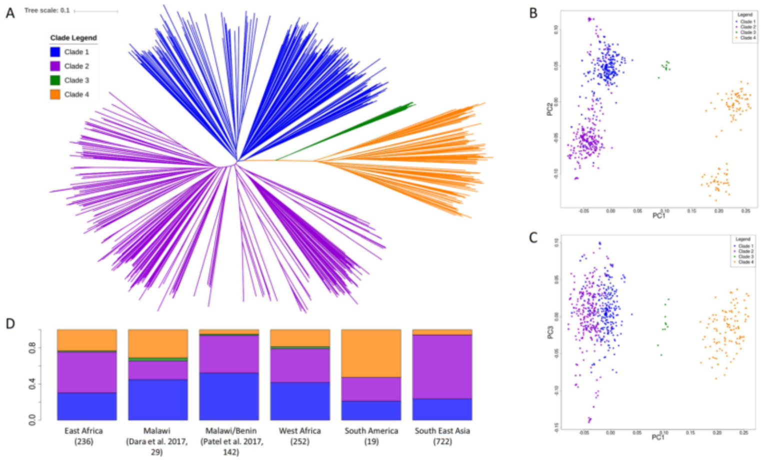


Figure 2. Population structure using the *ID1-DBL2Xb* protein sequences (**A**) Four distinct clades are identified, with some overlap to the clades found in[18], where Clade 1 is 3D7-like and Clade 2 is FCR3-like. The PCA analysis (**B** and **C**) supports the separation of these clades and reveals the proximity of Clades 1 and 2. (**D**) The distribution of clades across the different regions and previous studies, with three of the clades present across all the populations (Clades 1, 2 and 4); 3D7-like clade associated with adverse outcome in pregnancy is in West Africa (41.2%), East Africa (27.5%), South East Asia (23.5%) and South America (20%); Clade 3 is present in African parasite populations and Clade 4 is predominantly in African populations.

Discussion

The VAR2CSA protein is a vaccine candidate for PM[25] and is the basis of two vaccines currently in Phase I clinical trials[15,16]. It is therefore essential to understand the genetic and structural diversity of *var2csa* in natural populations of *P. falciparum* parasites, to predict the impact of potential vaccines, and improve the understanding of the mechanisms by which the malaria parasites sequester in the placenta. It has been suggested that parasite populations with multiple copies of the *var2csa* gene persist longer during pregnancy, hypothesising that this could be due to the ability of these parasites to generate a wider diversity of antigenic variation[14]. Across global field samples (n = 2,099), all populations had evidence of extra copies (27% on average, broadly consistent with[14]), but the prevalence was higher in African and Papua New Guinean parasites. For the African populations this could be due to the higher immune pressure that this gene might be under in higher transmission settings. Whilst, for PNG it could be due to a founder effect arising from the parasite population's mixed Asian and African ancestry[26]. It is possible the degree of copy number variation in the population-based field isolates sourced predominantly from children with malaria could be different to that found in pregnant mothers. However, the *var2csa* sequences from pregnant mothers included in our analysis overlapped with the sequences and resulting tree clusters observed in the field isolates.

By applying a pipeline validated using Pacbio sequenced strains, 7 kb fragments encoding the VAR2CSA extra-cellular domains were assembled across 1,249 field isolates with little evidence of mixed infections. The diversity was highest towards the N-terminus of the protein, as seen previously[19] and, in general, higher in African parasite populations compared to South East Asian ones. This result is consistent with African genomes being older and with the higher transmission rates in that continent currently. InDels have been a neglected source of variation and diversity in *var2csa* studies. The presence of a high number of low frequency InDels, which had the highest density in regions with flat nucleotide diversity, revealed that the level of diversity was underestimated. Comparing across the 5 VAR2CSA domains, the highest density of InDels was found in the DBL2X (part of CSA minimal binding domain), leading to the greatest variability in sequence length, where up to an additional 120 amino acid insertions were present. The impact of small in-frame InDels and short frame shifts on the protein structure of the DBL domain and their binding capabilities are not clear, and structural modelling approaches are difficult to scale-up to the levels of variation observed. A large number of the SNPs and InDels could lead to important changes in the amino acid sequence while conserving the overall gene structure. However, this should be fully explored to understand any impact on antibody binding affinity and CSA-binding in placental malaria. Further phenotypic characterization is required to evaluate the contribution of the diversity observed in the gene at both SNP and InDel level to parasite sequestration in the placenta, which could provide insight into the mechanisms by which the infection causes the associated adverse outcomes during pregnancy.

A recent study in pregnant women in Malawi and Benin identified 5 clades of the ID1-DBL2Xb domain of the *var2csa* gene and found an association between the infection of parasites harbouring the 3D7-like sequence and low birthweight[18]. Our work suggests that there are four main ID1-DBL2Xb domain clades, including a 3D7-like (Clade 1) and FCR3-like (Clade 2). Two of the previously reported clades appear to be too homogenous to be separated in our much larger dataset. Three of the four remaining clades (1, 2 and 4) were present across all the regions, including Clade 1, which was found in West Africa (41.2%), East Africa (27.5%), South East Asia (23.5%) and South America (21.1%), consistent with previous findings in Malawi (44.8%)[19] and Malawi and Benin (52.1%)[18]. Clade 3 isolates appear to be rare (<1% overall) and almost exclusively present in African parasites. Previous studies[18] have focused on the effect of the major clades have on pregnancy outcomes, but these data highlight the presence of Clade 4 isolates across very diverse regions very diverse regions globally, therefore stressing the need to further investigate the relationship between the prevalence of the different parasite clades and the pregnancy outcomes. Countries outside of sub-Saharan Africa such as the South East Asian region where the Clade 1 (3D7-like) has been reported for the first time, are of particular interest given that there is no evidence on the impact that parasites harbouring Clade 1 VAR2CSA protein have on pregnancy in these host populations. Furthermore, studies of the potential impact of the less prevalent Clades 3 and 4 on adverse outcomes in pregnancy are needed.

An analysis of the ID1-DBL2Xb domain haplotypes revealed higher diversity across African populations compared to South East Asian populations. A higher proportion of unique haplotypes in African parasites is likely to be a reflection of the higher transmission intensity. These diversity patterns suggest that introduction of a vaccine based on one single haplotype or a few heterologous haplotypes might be of greater benefit in South East Asian populations, although it is in Africa that the vaccine is most needed. Our work suggests that it would be advisable to consider the four clades of related *var2csa* haplotypes when testing the efficacy of the vaccine and, if possible, including heterologous haplotypes for the African specific clades found in this study in future vaccine design.

Overall, our study reveals the genetic diversity of the *var2csa* gene across more than 23 countries, demonstrating that SNPs and InDels are frequent at this locus, and generate considerable haplotype diversity, especially in African parasites. We also report a high frequency of multiple *var2csa* gene copies in the genome of field isolates. Further molecular and association studies are essential to understand the effect of VAR2CSA variants and extra gene copies on parasite sequestration in the placenta, the outcome of pregnancy and vaccine efficacy. Our findings can be used to support the assessment and development of new preventive tools against placental malaria, including the design of new vaccines that are robust to regional genetic diversity.

Materials and Methods

Samples, sequence data and processing. High-quality and high molecular weight DNA (20 µg) was purified from laboratory strains (D10, K1, HB3, T9/96 and NF54) using the Qiagen Genomic tip 20/G. Sequencing data for D10 (Papua New Guinea), T9/96 (Thailand), HB3 (Honduras), K1 (Thailand), NF54 (3D7 Parental line) strains were generated on Pacific Biosciences (PacBio) RS-II long read technology at the Genome Institute Singapore and were complemented by similar data for 7G8 and IT (Brazil), GB4 (Ghana), GN01 (Guinea), CD01 (Congo), Dd2 (IndoChina), KE01 (Kenya), KH01 and KH02 (Cambodia), GA01 (Gabon), SD01 (Sudan), TG01 (Togo), SN01 (Senegal) and ML01 (Mali) in the public domain (ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/falciparum/PF3K/PilotReferenceGenomes/GenomeSequence/Version1/). High quality sequencing reads (range: 33,855–66,185 reads) were assembled using Hierarchical Genome Assembly Process HGAP3 and corrected using Illumina reads when available, implemented in the SMRT Portal software suite and following previously described methods[27]. Overlaps between the start and end of large contigs were found using *Mummer* software[28] and removed using in-house scripts.

Raw sequencing data from Illumina data was available for previously published *P. falciparum* strains (3D7, HB3, D10, 7G8 and GB4)[29,30] and isolates from East Africa (Kenya (n = 38), Malawi (353), Tanzania (63), Uganda (12), Madagascar (18)), West Africa (Burkina Faso (48), Gambia (63), Ghana (443), Guinea (116), Mali (55), Nigeria (6), Cameroon (127)), Central Africa (Democratic Republic of Congo (DRC) (232)), South America (Colombia (15), Peru (9), Brazil (3)), South Asia (Bangladesh (53)) and South East Asia and Oceania (Cambodia (649), Laos (112), Myanmar (134), Papua New Guinea (26), Thailand (326), Vietnam (199))[31]. Public accession numbers for raw sequence data analysed are contained in SRA studies ERP000190 and ERP000199, as well as being accessible from the Pf3k project website (https://www.malariagen.net/projects/pf3k). All Illumina short reads were mapped to the 3D7 reference genome (version 3.0) using *bwa-mem* (version 0.7.17)[32]. SNPs and small InDels were called from the alignment *bam* files using *samtools* and *bcf/vcftools* (version 1.5) with default settings[33]. Only those variants with quality scores in excess of 30 (indicating an error rate less than 1 per 1000 bp) and with minimum coverage of 10 were retained[31]. The pipeline is summarised in S2 Fig. In total, the dataset contains 1,649 isolates and 1,513,940 high quality SNPs; 47.3% within genes and 5.2% have a minor allele frequency greater than 1%. A principal component analysis (not shown) based on pairwise SNP differences between isolates did not reveal any geographic outliers. We used the proportion of heterozygous calls per sample (>0.015%) as well as the fraction of genome indicating Multiplicity Of Infection (MOI)>1 obtained using estMOI (>30%)[34] as described previously[35] to remove samples with MOI>1 (S3 Fig.).

Characterisation of the *var2csa* gene and copy number. The Pacbio sequencing contigs that contained the *var2csa* gene were identified, and mapped to the 3D7 reference genome, allowing the direct assessment of copy numbers. Similarly, the *var2csa* gene was characterised from contigs constructed through *de novo* assembly of Illumina short reads using *velvet software*[22]. The resulting contigs obtained were aligned and assembly errors were corrected manually using *Aliview*[36]. Once corrected the sequences were trimmed from the N-Terminal to form a 7 kb fragment which was then translated using *OrfFinder*[37] and sequences presenting more than one ORF were excluded from further analysis. The approach was validated by comparing the contigs obtained from PacBio and Illumina platforms, using 3D7, 7G8 and GB4 samples with both sets of data (S8 Fig.). Illumina data were also used to infer copy numbers. In particular, *Delly* software (version 0.7.7)[38] was used to calculate genomic coverage from the alignment bam files, and the *Control-FREEC tool* (Version 10.6)[39] was used to estimate copy number based on GC content corrected ratios of coverage using a sliding window of 500 bp and a step size of 100 bp[39]. We also estimated copy number based on the ratio of the average *var2csa* gene coverage against the average gene coverage in the rest of the genome, optimised and tested using the HB3, D10, GB4, 3D7, and 7G8 strains.

Genetic diversity within and across populations. The selection of the samples for genetic diversity analysis (n = 1,649) focused on those with the presence of a single copy of the *var2csa* gene based on coverage (S4 Fig.) and a maximum of 2% of bases in the *var2csa* gene presenting heterozygous calls in the gene (S7 Fig.). We then screened the contigs assembled using Blast alignment[37] to the 3D7 reference of the *var2csa* gene and retrieved 1,317 isolate sequences with a hit that contained a contig of more than 8 kb in length. We aligned the sequences using *Mafft*[40] with default parameters, reverse-complemented when needed and extracted the region corresponding to the *var2csa* gene and μORF region using *AliView*[36]. The alignment was manually curated to remove poly-Ns added by the assembly software as well as obvious assembly errors. The final set of sequences was then translated using *orfinder*[37] and a further 72 sequences presenting fragmented translation were excluded. The final dataset consisted of 1,249 sequences spanning 7Kb of the N-terminal end of the *var2csa* gene (Table 1, S1 Data).

The *DnaSP* (version 5)[41] and the *ape* and *agenet*[42] R libraries were used to compute population genetic parameters at the region and country levels, including nucleotide diversity (π), the average number of nucleotide differences per site between any two given sequences and the haplotype diversity (Hd), which is the probability that two randomly sampled haplotypes are different. We also calculated distance matrices using R package *seqinr*[43] in order to generate a Principal Component Analysis (PCA) of the nucleotide and protein sequences and the corresponding Neighbour-Joining trees. A *k-means* approach was used in R to obtain the clusters observed in the neighbour joining trees. The *vegan* R package was used to calculate the rarefaction curves using the ID1-DBL2Xb region haplotypes by region and country.

Data Availability

Public accession numbers for raw sequence data analysed are contained in SRA studies ERP000190 and ERP000199, as well as being accessible from the Pf3k project website (https://www.malariagen.net/projects/pf3k). Data was complemented by PacBio data in the public domain (ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/falciparum/PF3K/PilotReferenceGenomes/GenomeSequence/Version1/).

References

1. WHO. *World Malaria Report 2017* (2017).
2. Bray, R. S. & Sinden, R. E. The sequestration of Plasmodium falciparum infected erythrocytes in the placenta. *Trans. R. Soc. Trop. Med. Hyg.* **73**, 716–719 (1979).
3. Nunes, M. C. & Scherf, A. Plasmodium falciparum during pregnancy: a puzzling parasite tissue adhesion tropism. *Parasitology* **134**, 1863–1869 (2007).
4. Duffy, P. E. & Fried, M. Malaria during pregnancy: parasites, antibodies and chondroitin sulphate A. *Biochem. Soc. Trans.* **27**, 478–482 (1999).
5. Salanti, A. *et al*. Evidence for the Involvement of VAR2CSA in Pregnancy-associated Malaria. *J. Exp. Med.* **200**, 1197–1203 (2004).
6. Srivastava, A. *et al*. Full-length extracellular region of the *var2CSA* variant of PfEMP1 is required for specific, high-affinity binding to CSA. *Proc. Natl. Acad. Sci. USA* **107**, 4884–4889 (2010).
7. Khunrae, P. *et al*. Full-Length Recombinant Plasmodium falciparum VAR2CSA Binds Specifically to CSPG and Induces Potent Parasite Adhesion-Blocking Antibodies. *J. Mol. Biol.* **397**, 826–834 (2010).
8. Salanti, A. *et al*. Selective upregulation of a single distinctly structured var gene in chondroitin sulphate A-adhering Plasmodium falciparum involved in pregnancy-associated malaria. *Mol. Microbiol.* **49**, 179–191 (2003).
9. Tuikue Ndam, N. G. *et al*. High level of *var2csa* transcription by Plasmodium falciparum isolated from the placenta. *J. Infect. Dis.* **192**, 331–335 (2005).
10. Fried, M., Nosten, F., Brockman, A., Brabin, B. J. & Duffy, P. E. Maternal antibodies block malaria. *Nature* **395**, 851–852 (1998).
11. Staalsoe, T. *et al*. Variant surface antigen-specific IgG and protection against clinical consequences of pregnancy-associated Plasmodium falciparum malaria. *Lancet (London, England)* **363**, 283–289 (2004).
12. Sander, A. F. *et al*. Multiple *var2csa*-type PfEMP1 genes located at different chromosomal loci occur in many Plasmodium falciparum isolates. *PLoS One* **4**, e6667 (2009).
13. Brolin, K. J. M. *et al*. Simultaneous transcription of duplicated *var2csa* gene copies in individual Plasmodium falciparum parasites. *Genome Biol.* **10**, R117–R117 (2009).
14. Sander, A. F. *et al*. Positive selection of Plasmodium falciparum parasites with multiple *var2csa*-type PfEMP1 genes during the course of infection in pregnant women. *J. Infect. Dis.* **203**, 1679–1685 (2011).
15. European Vaccine Initiative. http://www.euvaccine.eu/portfolio/project-index/placmalvac.
16. European Vaccine Initiative. http://www.euvaccine.eu/portfolio/project-index/primalvac.
17. Kane, E. G. & Taylor-Robinson, A. W. Prospects and Pitfalls of Pregnancy-Associated Malaria Vaccination Based on the Natural Immune Response to Plasmodium falciparum VAR2CSA-Expressing Parasites. *Malar. Res. Treat.* **2011**, 764845 (2011).
18. Patel, J. C. *et al*. Increased risk of low birth weight in women with placental malaria associated with P. falciparum VAR2CSA clade. *Sci. Rep.* **7**, 7768 (2017).
19. Dara, A. *et al*. A new method for sequencing the hypervariable Plasmodium falciparum gene *var2csa* from clinical samples. *Malar. J.* **16**, 343 (2017).
20. Rogerson, S. J. *et al*. Burden, pathology, and costs of malaria in pregnancy: new developments for an old problem. *Lancet Infect. Dis.* **18**, e107–e118 (2018).
21. Otto, T. D. *et al*. Long read assemblies of geographically dispersed Plasmodium falciparum isolates reveal highly structured subtelomeres. *Wellcome open Res.* **3**, 52 (2018).
22. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18** (2008).
23. Trimnell, A. R. *et al*. Global genetic diversity and evolution of var genes associated with placental and severe childhood malaria. *Mol. Biochem. Parasitol.* **148**, 169–180 (2006).
24. Bancells, C. & Deitsch, K. W. A molecular switch in the efficiency of translation reinitiation controls expression of *var2csa*, a gene implicated in pregnancy-associated malaria. *Mol. Microbiol.* **90**, 472–488 (2013).
25. Tutterrow, Y. L. *et al*. High Avidity Antibodies to Full-Length VAR2CSA Correlate with Absence of Placental Malaria. *PLoS One* **7**, e40049 (2012).

26. Preston, M. D. *et al.* A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. *Nat. Commun.* **5**, 4052 (2014).
27. Benavente, E. D. *et al.* A reference genome and methylome for the Plasmodium knowlesi A1-H.1 line. *Int. J. Parasitol.* https://doi.org/10.1016/j.ijpara.2017.09.008 (2017).
28. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12–R12 (2004).
29. Shears, M. J. *et al.* Characterization of the Plasmodium falciparum and P. berghei glycerol 3-phosphate acyltransferase involved in FASII fatty acid utilization in the malaria parasite apicoplast. *Cell. Microbiol.* **19**, e12633 (2017).
30. Miles, A. *et al.* Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. *Genome Res.* **26**, 1288–1299 (2016).
31. Ravenhall, M. *et al.* Characterizing the impact of sustained sulfadoxine/pyrimethamine use upon the Plasmodium falciparum population in Malawi. *Malar. J.* **15** (2016).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
33. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
34. Assefa, S. A. *et al.* estMOI: estimating multiplicity of infection using parasite deep sequencing data. *Bioinformatics* **30**, 1292–1294 (2014).
35. Diez Benavente, E. *et al.* Analysis of nuclear and organellar genomes of Plasmodium knowlesi in humans reveals ancient population structure and recent recombination among host-specific subpopulations. *PLOS Genet.* **13**, e1007008 (2017).
36. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).
37. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **39**, D38–51 (2011).
38. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28** (2012).
39. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2012).
40. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
41. Librado, P. & Rozas, J. DnaSPv5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452 (2009).
42. Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).
43. Charif, D., Thioulouse, J., Lobry, J. R. & Perriere, G. Online synonymous codon usage analyses with the ade4 and seqinR packages. *Bioinformatics* **21**, 545–547 (2005).

Acknowledgements

Sequence data analysis was performed on the MRC UK eMedLab computing resource. TGC is supported by the Medical Research Council UK (Grant no. MR/K000551/1, MR/M01360X/1, MR/N010469/1, MR/R020973/1) and BBSRC (BB/R013063/1). SC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1) and BBSRC (BB/R013063/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

E.D.B., T.G.C. and S.C. conceived and designed the study; E.M.W. and D.A.B. cultured malaria parasites and E.M.W., C.R., J.G.D., R.M.d.S., C.R.F.M., C.J.S. and D.A.B. contributed parasite DNA for sequencing; P.F.d.S., M.L.H. and S.C. coordinated the sequencing of samples; E.D.B. and D.R.O. performed the statistical analysis, under supervision of F.M., T.G.C. and S.C.; E.D.B., T.G.C. and S.C. wrote the first draft of the manuscript, and the final version included edits from all authors. The final manuscript was read and approved by all authors.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-33767-3.

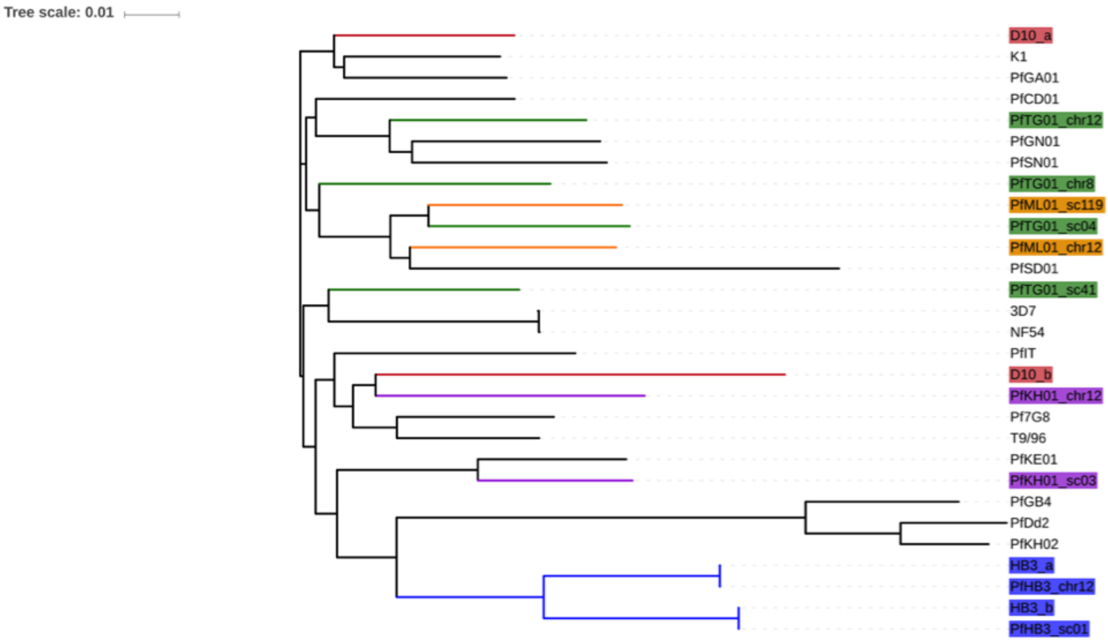
Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

SUPPLEMENTARY MATERIAL

S1 figure

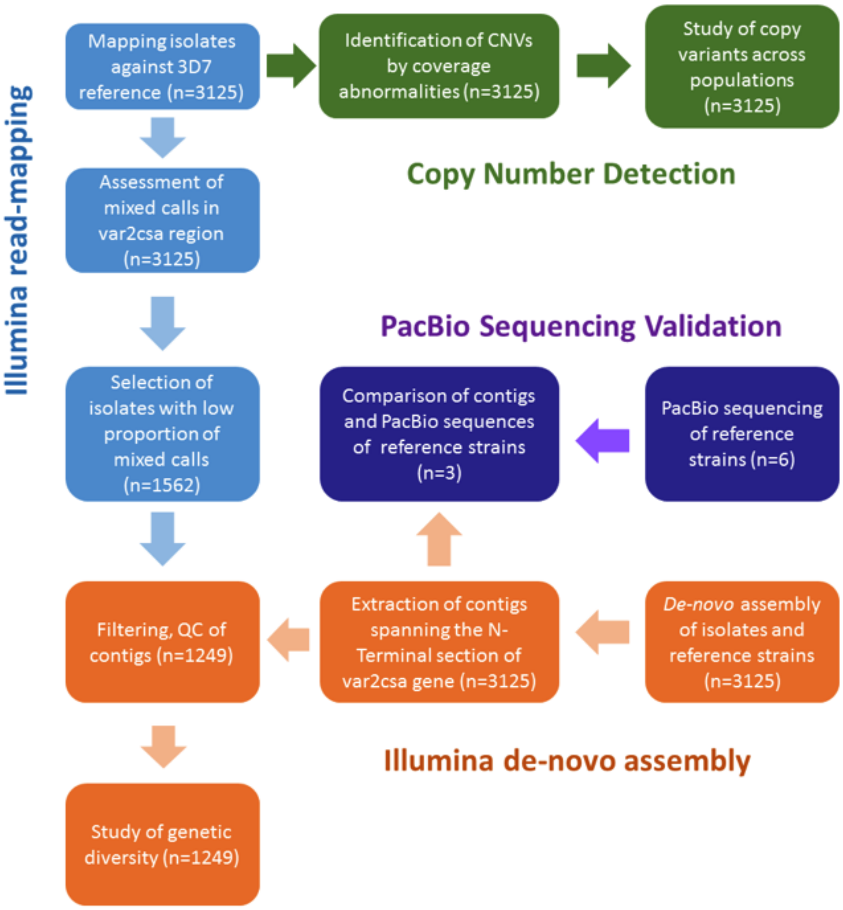
Phylogenetic tree for *P. falciparum* laboratory strains constructed using the *var2csa* alignments reveals two near identical gene copies for HB3 and distinct copies for D10 and KH01. Two sequences of each HB3 copy were obtained from two different PacBio runs and were found to be identical, highlighting the robustness of the approach. No significant clustering by geography was found for other samples, as expected from the patterns shown by the short read data.



D10 (Papua New Guinea), T9/96 (Thailand), HB3 (Honduras), K1 (Thailand), and NF54 (3D7 Parental line), GN01 (Guinea), CD01 (Congo), Dd2 (IndoChina), KE01 (Kenya), KH01 and KH02 (Cambodia), GA01 (Gabon), IT (Brazil), SD01 (Sudan), TG01 (Togo), SN01 (Senegal) and ML01 (Mali); samples with extra copies are colour coded.

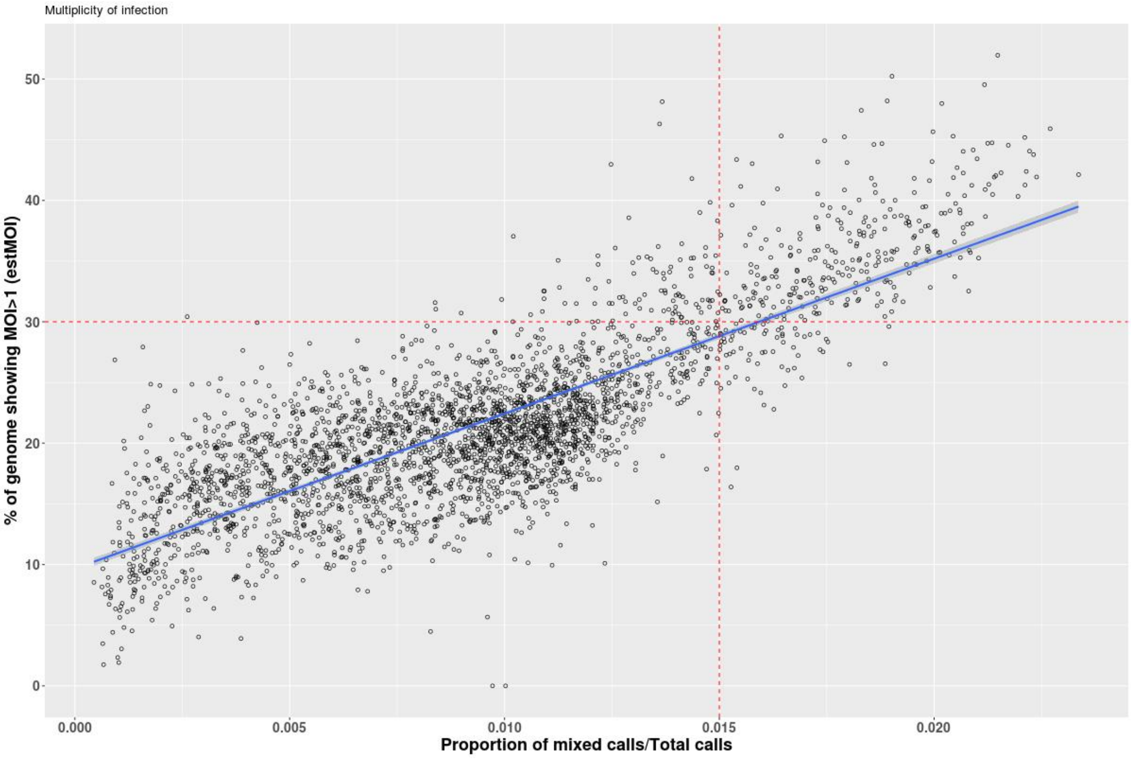
S2 figure

Summary of the analytical and bioinformatics pipelines used


Illumina read-mapping
Mapping isolates against 3D7 reference (n=3125)
Identification of CNVs by coverage abnormalities (n=3125)
Study of copy variants across populations (n=3125)
Copy Number Detection
Assessment of mixed calls in var2csa region (n=3125)
PacBio Sequencing Validation
Selection of isolates with low proportion of mixed calls (n=1562)
Comparison of contigs and PacBio sequences of reference strains (n=3)
PacBio sequencing of reference strains (n=6)
Filtering, QC of contigs (n=1249)
Extraction of contigs spanning the N-Terminal section of var2csa gene (n=3125)
De-novo assembly of isolates and reference strains (n=3125)
Illumina de-novo assembly
Study of genetic diversity (n=1249)

S3 figure

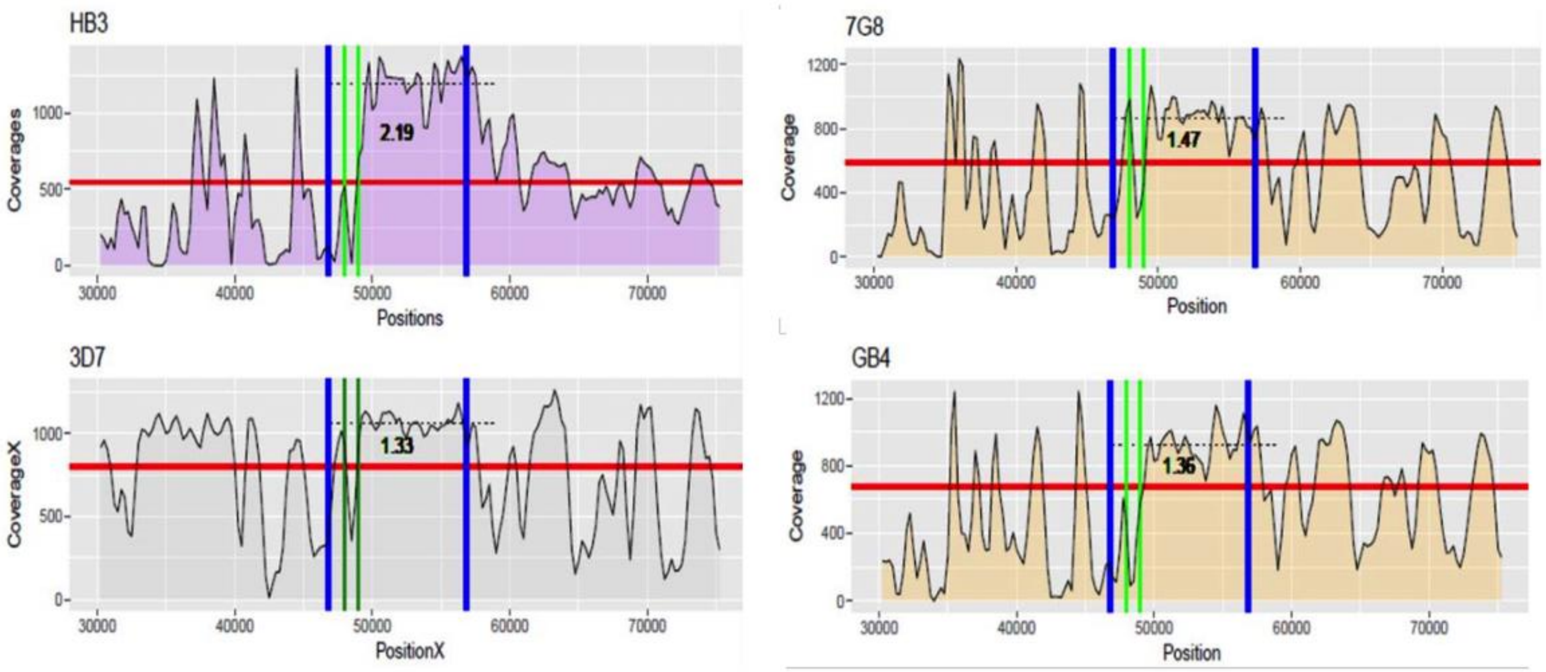
The filtered *P. falciparum* isolates based on MOI and heterozygous calls* (n=2,099; with >70% of their genome with multiplicity of infection of 1 and <1.5% of heterozygous SNP calls)



*All samples have >30-fold genome-wide coverage

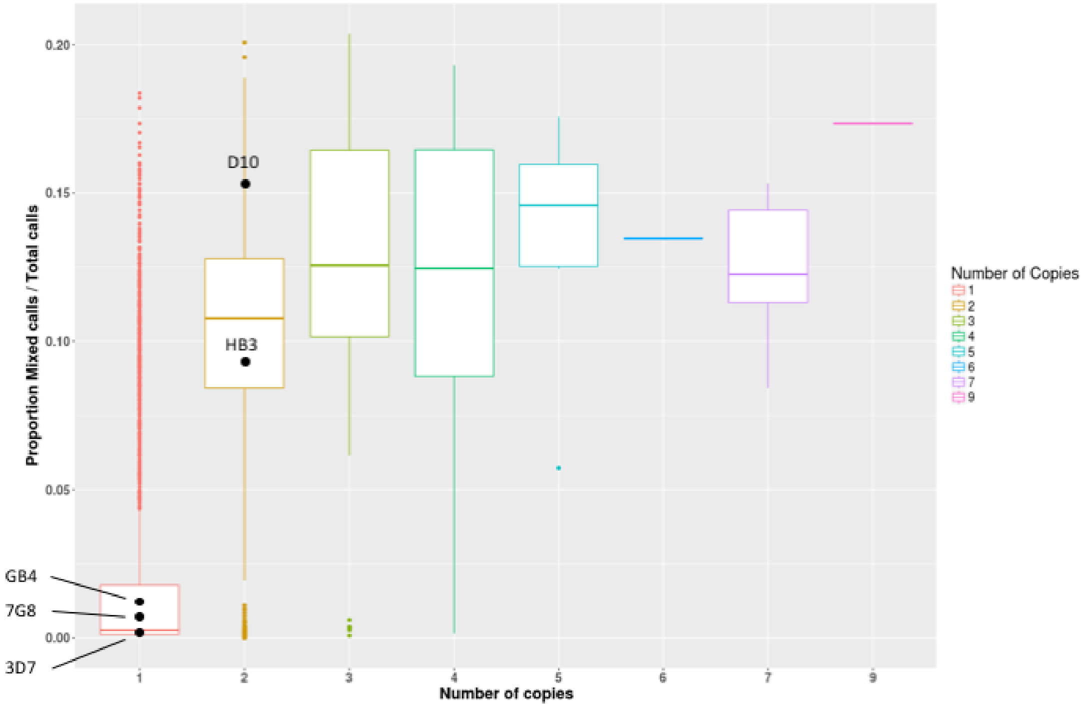
S4 figure

Estimating copy number in the *var2csa* gene using coverage in 4 *P. falciparum* laboratory strains for which Paired End data was available, (no D10 paired end data available). Red line indicates the chromosome average, and the dashed line represents the average coverage in the long exon 2 of the var2csa gene. Gene boundaries are in clue vertical lines, and intron boundaries are represented by green vertical lines in each plot.



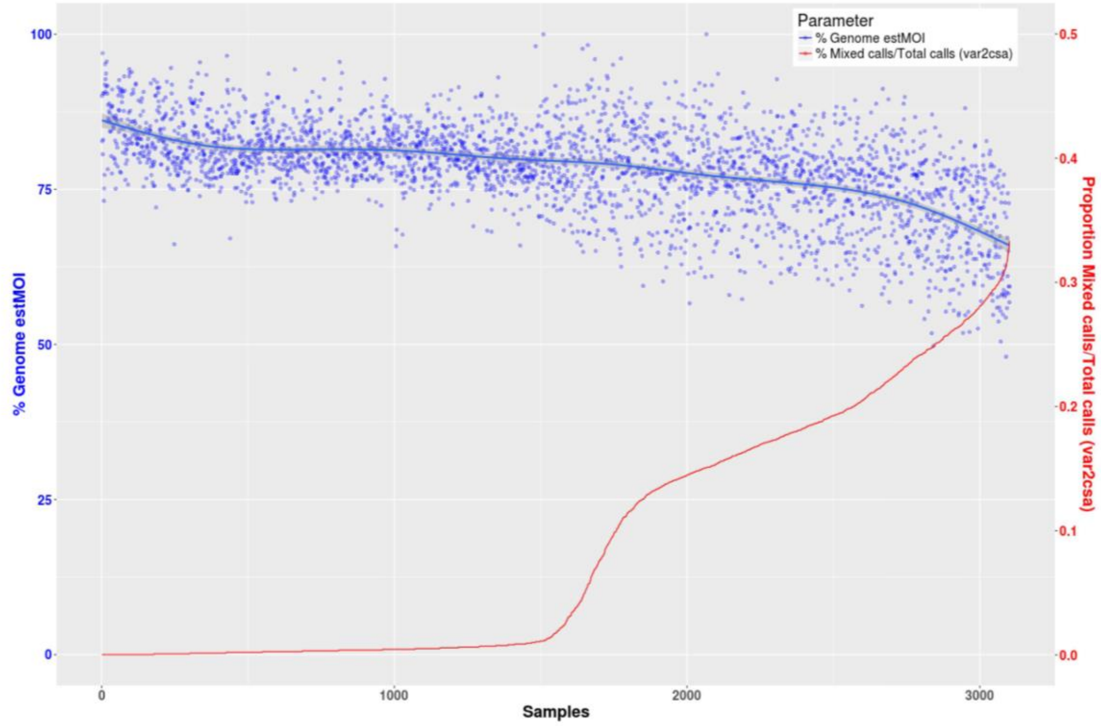
S5 figure

Var2csa **copy number distributions in** ***P. falciparum*** **and mixed SNP calls across the samples, including 5 laboratory strains with Illumina data available.** A clear trend of increased mixed calls (y-axis) is found when extra copies of the gene are identified using coverage fold increase (x-axis).



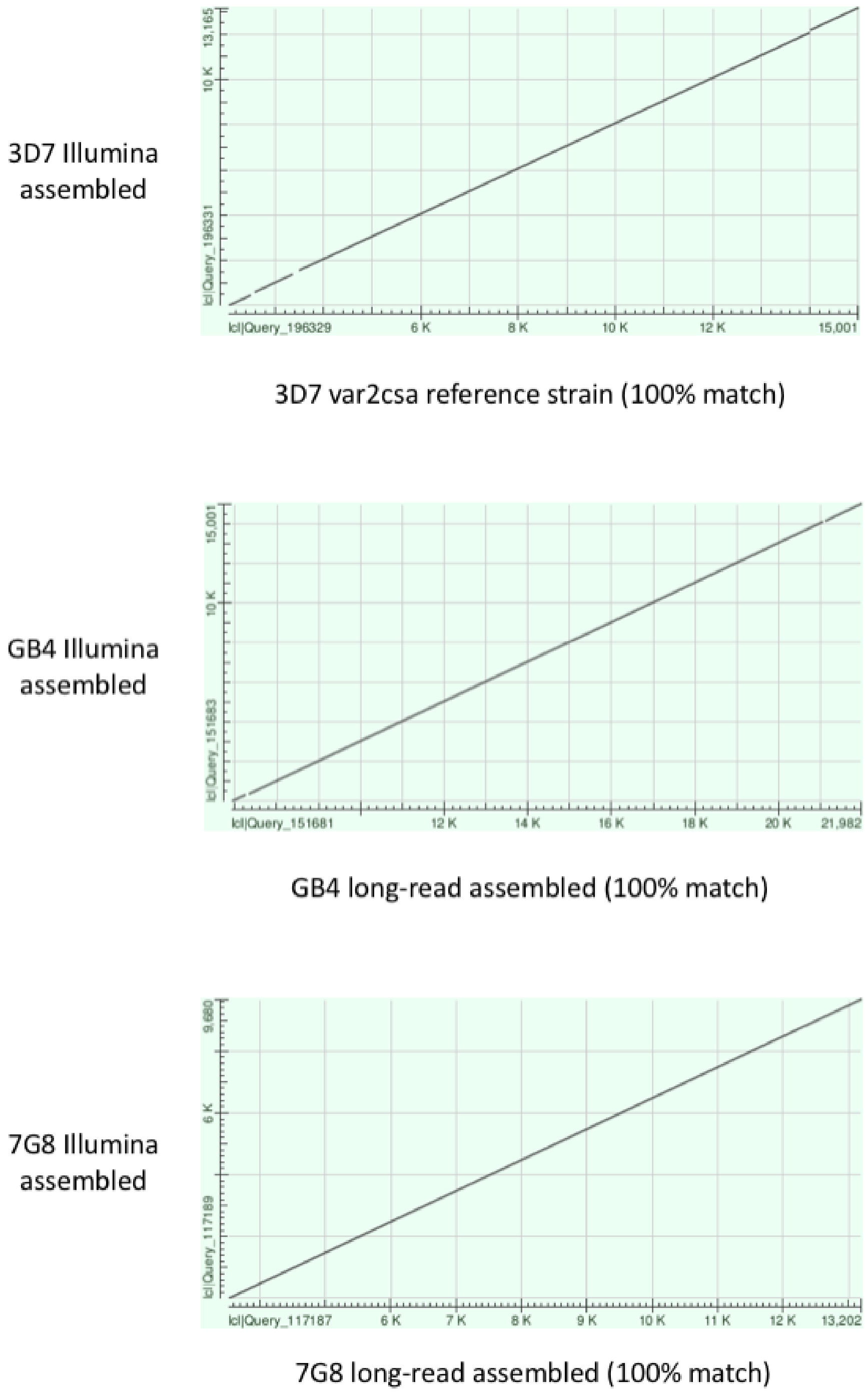
S6 figure

The multiplicity of infection (MOI*) ^{32} **in the genome and the proportion of mixed calls in the *var2csa* gene in *P. falciparum.*** Samples sorted based on proportion of mixed calls in the *var2csa* gene (red axis). A decrease in the average percentage of genome supporting MOI of 1 is observed for isolates with increased proportion of mixed calls in the var2csa gene. Supporting the presence of mixed calls in the gene due to MOI >1 rather than by extra copies of the gene. The threshold used for downstream de-novo assembly was 5% of mixed calls/total number of calls.



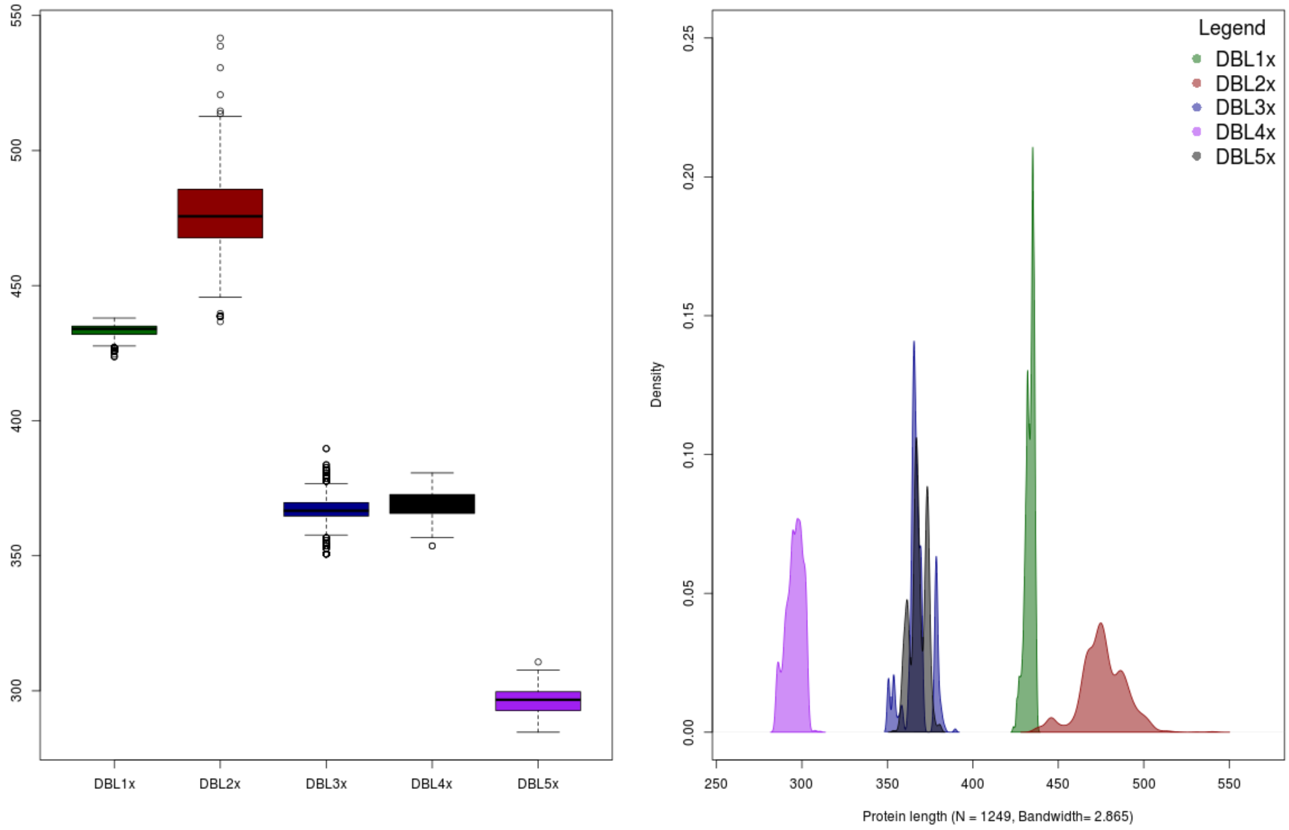
S7 figure

Near perfect matching of the assembled *P. falciparum* strain *var2csa* sequences from our pipeline compared to the long-read gold standard. Dot plot comparing the Illumina assembly for each laboratory strain and its PacBio generated sequence.



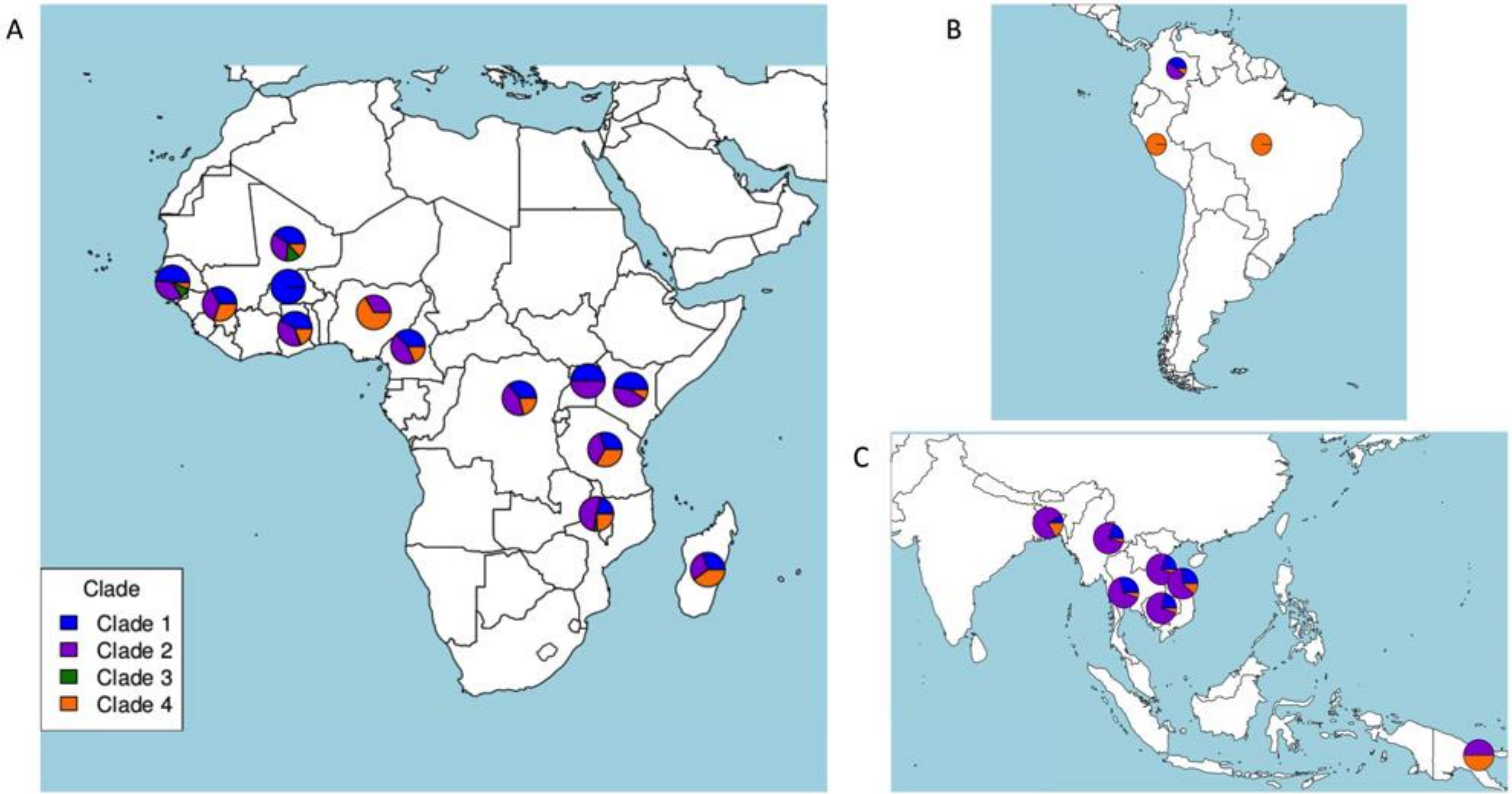
S8 figure

Distribution of the lengths of the different DBL domains in the *var2csa* gene of *P. falciparum* across the field samples reveals increased length distribution for DBL2X domain (n=1,249). A wider range can be found in the distribution of the lengths for the DBL2X domain of the var2csa gene of the isolates compared to the other DBL domains. This a result of the presence of multiple rare insertions and deletions in this region.



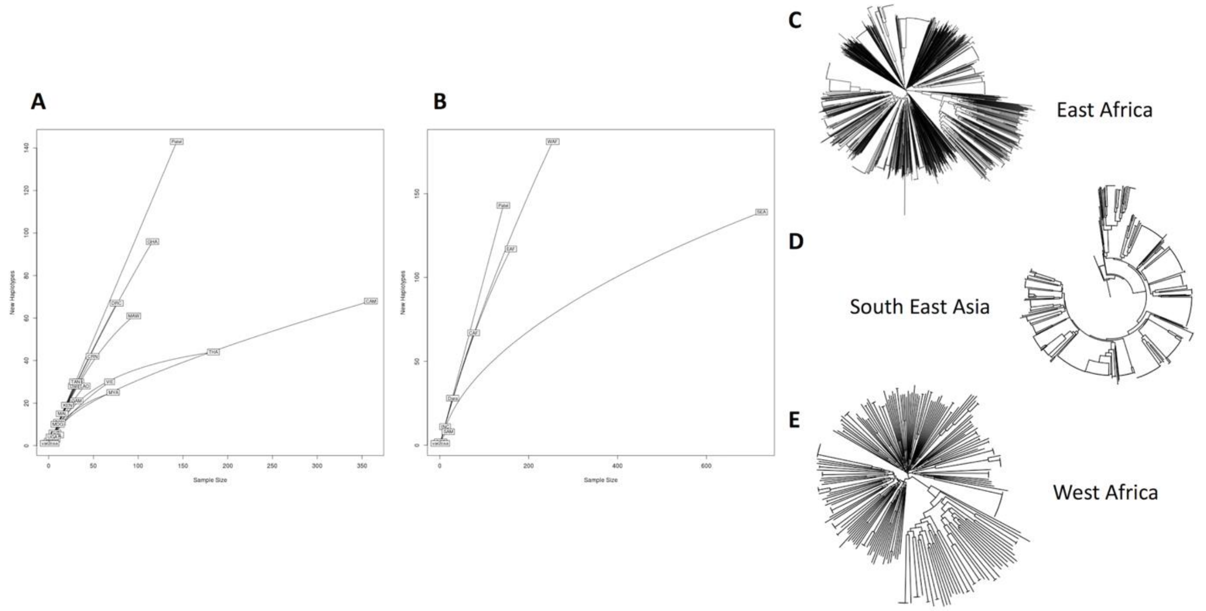
S9 figure

Global map of the distribution of the four clades based on the ID1-DBL2Xb region of the *var2csa* gene in *Plasmodium falciparum*



S10 figure

A rarefaction curve analysis of the haplotype diversity in the ID1-DBL2Xb region of the var2csa gene in *Plasmodium falciparum* (A; country, B; region) reveals higher diversity in African populations compared to South East Asian populations, which is supported by the neighbourhood joining trees for these regions across geography (C, D, E)



S1 table

Distribution of P. falciparum isolates with extra copies of the var2csa gene per country

	Total	Single copy		Multiple Copies	
Country	**n**	**n**	**%**	**n**	**%**
Burkina Faso	27	14	51.9	13	48.1
Cameroon	80	58	72.5	22	27.5
Gambia	47	35	74.5	12	25.5
Ghana	222	138	62.2	84	37.8
Guinea	78	51	65.4	27	34.6
Mali	29	17	58.6	12	41.4
Nigeria	6	4	66.7	2	33.3
West Africa	**489**	**317**	**64.8**	**172**	**35.2**
DRC	136	96	70.6	40	29.4
Kenya	26	18	69.2	8	30.8
Malawi	182	123	67.6	59	32.4
Tanzania	43	36	83.7	7	16.3
Uganda	5	4	80.0	1	20.0
Madagascar	17	12	70.6	5	29.4
East Africa	**409**	**289**	**70.7**	**120**	**29.3**
Bangladesh	**36**	**18**	**50.0**	**18**	**50.0**
Cambodia	492	428	87.0	64	13.0
Laos	88	50	56.8	38	43.2
Myanmar	109	89	81.7	20	18.3
Thailand	264	208	78.8	56	21.2
Vietnam	155	98	63.2	57	36.8
South East Asia	**1108**	**873**	**78.8**	**235**	**21.2**
PNG (Oceania)	**24**	**3**	**12.5**	**21**	**87.5**
Brazil	9	9	100.0	0	0.0
Colombia	15	14	93.3	1	6.7
Peru	9	9	100.0	0	0.0
South America	**33**	**32**	**97.0**	**1**	**3.0**
Overall	***2099***	***1532***	***73.0***	***567***	***27.0***

DRC Democratic Republic of Congo; PNG Papua New Guinea

S2 table

The frequency of the different protein sequence clades based on the ID1-DBL2Xb region of the var2csa genes in P. falciparum, across countries (n=1,373)

	Clade 1 (3D7-like)		**Clade 2 (FCR3-like)**		**Clade 3**		**Clade 4**	
Country	**n**	**%**	**n**	**%**	**n**	**%**	**N**	**%**
Burkina Faso	**7**	**100.0**	**0**	**0.0**	**0**	**0.0**	**0**	**0.0**
Cameroon	**7**	**100.0**	**0**	**0.0**	**0**	**0.0**	**0**	**0.0**
Gambia	**15**	**48.4**	**11**	**35.5**	**3**	**9.7**	**2**	**6.5**
Ghana	**48**	**41.7**	**44**	**38.3**	**1**	**0.9**	**22**	**19.1**
Guinea	**11**	**33.3**	**12**	**36.4**	**0**	**0.0**	**10**	**30.3**
Mali	**6**	**40.0**	**5**	**33.3**	**2**	**13.3**	**2**	**13.3**
Nigeria	**0**	**0.0**	**1**	**33.3**	**0**	**0.0**	**2**	**66.7**
West Africa	***94***	***44.5***	***73***	***34.6***	***6***	***2.8***	***38***	***18.0***
DRC	**27**	**35.5**	**33**	**43.4**	**0**	**0.0**	**16**	**21.1**
Kenya	**10**	**47.6**	**9**	**42.9**	**0**	**0.0**	**2**	**9.5**
Malawi	**20**	**21.1**	**49**	**51.6**	**3**	**3.2**	**23**	**24.2**
Malawi*	**13**	**44.8**	**6**	**20.7**	**1**	**3.4**	**9**	**31.0**
Malawi & Benin**	**74**	**52.1**	**59**	**41.5**	**2**	**1.4**	**7**	**4.9**
Tanzania	**9**	**30.0**	**11**	**36.7**	**0**	**0.0**	**10**	**33.3**
Uganda	**2**	**50.0**	**2**	**50.0**	**0**	**0.0**	**0**	**0.0**
Madagascar	**3**	**30.0**	**3**	**30.0**	**0**	**0.0**	**4**	**40.0**
East Africa	***158***	***38.8***	***172***	***42.3***	***6***	***1.5***	***71***	***17.4***
Bangladesh	1	8.3	9	75.0	0	0.0	2	16.7
Cambodia	76	21.1	266	73.9	0	0.0	18	5.0
Laos	8	20.0	30	75.0	0	0.0	2	5.0
Myanmar	14	19.4	55	76.4	0	0.0	3	4.2
Thailand	54	29.7	118	64.8	0	0.0	10	5.5
Vietnam	18	26.5	42	61.8	0	0.0	8	11.8

South East Asia	*170*	*23.5*	*511*	*70.8*	*0*	*0*	*41*	*5.7*
PNG	0	0.0	1	50.0	0	0.0	1	50.0
Brazil	0	0.0	0	0.0	0	0.0	3	100.0
Colombia	4	40.0	5	50.0	0	0.0	1	10.0
Peru	0	0.0	0	0.0	0	0.0	6	100.0
South America	*4*	*21.1*	*5*	*26.3*	*0*	*0.0*	*10*	*52.6*
Overall	*427*	*31.1*	*771*	*56.2*	*12*	*0.9*	*163*	*11.9*

*Dara et al. 2017; ** Patel el at 2017 (n=124); DRC Democratic Republic of Congo; PNG Papua New Guinea; Bolded denotes African Populations and Shaded denotes South East Asian Populations

S3 table

Diversity statistics by country in the open reading frame region (359 bp) upstream of the *var2csa* gene in 1,245 *P. falciparum* isolates

Population	n	No. of Haplotypes	Haplotype diversity (*Hd*)	Nucleotide Diversity (*π*)	Average no. of nucleotide differences	No. variable sites (n=360)
Cameroon	**46**	**28**	**0.974**	**0.020**	**7.21**	**31**
Gambia	**27**	**14**	**0.940**	**0.020**	**7.35**	**26**
Ghana	**111**	**53**	**0.973**	**0.021**	**7.42**	**38**
Guinea	**38**	**21**	**0.929**	**0.020**	**7.37**	**32**
Mali	**12**	**10**	**0.970**	**0.013**	**4.52**	**12**
DRC	**75**	**48**	**0.98**	**0.025**	**9.00**	**35**
Kenya	**16**	**12**	**0.958**	**0.017**	**6.26**	**19**
Malawi	**103**	**42**	**0.970**	**0.024**	**8.65**	**38**
Tanzania	**30**	**20**	**0.966**	**0.018**	**6.57**	**31**
Madagascar	**10**	**8**	**0.956**	**0.017**	**6.20**	**15**
Bangladesh	12	10	0.970	0.031	11.11	30
Cambodia	365	20	0.851	0.021	7.71	33
Laos	42	18	0.911	0.030	10.76	37
Myanmar	71	17	0.833	0.020	7.08	29
Thailand	180	30	0.934	0.028	10.05	40
Vietnam	72	23	0.867	0.025	8.89	38
Colombia	10	5	0.822	0.016	5.78	16
Overall	***1245***	***163***	***0.952***	***0.025***	***8.86***	***50***

Bolded denotes African countries and shaded denotes South East Asian countries; DRC Democratic Republic of Congo

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Ernest Diez Benavente
Principal Supervisor	Taane Clark & Susana Campino
Thesis Title	**Using whole genome sequence data to study genomic diversity and develop molecular barcodes to profile Plasmodium malaria parasites**

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	PLOS One		
When was the work published?	May 2017		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I extracted the raw data from public repositories. I created the analysis pipeline and ran the samples through this pipeline, I performed the QC analysis, as well as the interpretation analysis. I later created custom R and perl scripts to process the information in order to obtain information related to coverage, SNPs and other genomic information (including Copy Number Variants (CNVs). The figures presented in this work have all been generated using scripts written by myself. I wrote the first draft of the manuscript and circulated to co-authors. Once the comments were received I gathered them and made the relevant changes on the article manuscript. I then performed the submission of the work to PLOS One journal and once comments from reviewers arrived, I made the corresponding changes and corrections.

Student Signature: ____________________________________ **Date:** ________________

Supervisor Signature: ____________________________________ **Date:** ________________

Chapter 6
Genomic variation in *Plasmodium vivax* malaria reveals regions under selective pressure

RESEARCH ARTICLE

Genomic variation in *Plasmodium vivax* malaria reveals regions under selective pressure

Ernest Diez Benavente[1], Zoe Ward[1,2], Wilson Chan[1,3], Fady R. Mohareb[3], Colin J. Sutherland[1], Cally Roper[1], Susana Campino[1‡], Taane G. Clark[1‡]*

1 London School of Hygiene and Tropical Medicine, Keppel Street, London, United Kingdom, **2** The Bioinformatics Group, School of Water Energy and Environment, Cranfield University, Cranfield, Bedfordshire, United Kingdom, **3** Department of Pathology & Laboratory Medicine, Diagnostic & Scientific Centre, Faculty of Medicine, University of Calgary, Calgary, Alberta, Canada

‡ These authors are joint last authors on this work.
* taane.clark@lshtm.ac.uk

Citation: Diez Benavente E, Ward Z, Chan W, Mohareb FR, Sutherland CJ, Roper C, et al. (2017) Genomic variation in *Plasmodium vivax* malaria reveals regions under selective pressure. PLoS ONE 12(5): e0177134. https://doi.org/10.1371/journal.pone.0177134

Editor: Georges Snounou, Université Pierre et Marie Curie, FRANCE

Received: January 17, 2017

Accepted: April 21, 2017

Published: May 11, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: TGC receives funding from the Medical Research Council UK (grant numbers MR/K000551/1, MR/M01360X/1, MR/N010469/1, MC_PC_15103). CR and SC are funded by the Medical Research Council UK (grant no. MR/M01360X/1). The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Abstract

Background

Although *Plasmodium vivax* contributes to almost half of all malaria cases outside Africa, it has been relatively neglected compared to the more deadly *P. falciparum*. It is known that *P. vivax* populations possess high genetic diversity, differing geographically potentially due to different vector species, host genetics and environmental factors.

Results

We analysed the high-quality genomic data for 46 *P. vivax* isolates spanning 10 countries across 4 continents. Using population genetic methods we identified hotspots of selection pressure, including the previously reported *MRP1* and *DHPS* genes, both putative drug resistance loci. Extra copies and deletions in the promoter region of another drug resistance candidate, *MDR1* gene, and duplications in the Duffy binding protein gene (*PvDBP*) potentially involved in erythrocyte invasion, were also identified. For surveillance applications, continental-informative markers were found in putative drug resistance loci, and we show that organellar polymorphisms could classify *P. vivax* populations across continents and differentiate between *Plasmodia spp*.

Conclusions

This study has shown that genomic diversity that lies within and between *P. vivax* populations can be used to elucidate potential drug resistance and invasion mechanisms, as well as facilitate the molecular barcoding of the parasite for surveillance applications.

Background

The *Plasmodium vivax* malaria parasite is the second most virulent malaria species after *P. falciparum*. Geographically, it is found throughout Asia, South and Central America, Oceania,

Competing interests: The authors have declared that no competing interests exist.

Middle East and some parts of Africa, with nearly 2.85 billion people at risk of infection [1]. Although *P. vivax* contributes to almost half of all malaria cases outside Africa, as it kills infrequently and is not amenable to continuous *in vitro* culture, it has been relatively neglected compared to the more deadly *P. falciparum* [2]. However, as *P. vivax* drug-resistant strains emerge and spread and fatality rates increase, the need to implement better control and elimination strategies is becoming urgent. Many of the interventions used for controlling *P. falciparum* malaria are not as effective against *P. vivax*. Consequently, *P. vivax* has become the dominant malaria parasite in several countries where *P. falciparum* transmission has been successfully reduced. Hence, control and elimination of *P. vivax* malaria calls for additional interventions, notably against the dormant liver stage of the parasite. However, gaps in our knowledge of *P. vivax* epidemiology and biology may compromise its control. Genomic research can contribute greatly to enhancing our understanding of *P. vivax* basic biology and evolutionary history, supporting the development and surveillance of new interventions.

Since the first characterisation of the *P. vivax* genome sequence (Sal-1, [3]), several population genetic studies, based on microsatellite data and more recently using whole genomes, have shown that this parasite is more polymorphic than *P. falciparum* [4,5,6,7]. *P. vivax* populations harbour high genetic diversity, even on small spatial scales, and can differ extensively between locations due to vector species, host genetics and environmental factors [4,5,6]. Genetic variation enables the parasite to overcome host immune responses and antimalarial drugs to establish persistent infections and increase transmission. Genomic studies in natural populations of *P. vivax* can pinpoint genetic regions that are under selective pressure, including those associated with resistance to antimalarial drugs. Such studies can also contribute to the identification of vaccine targets. Moreover, global genomic studies can assist with identifying sets of polymorphism private to populations, allowing the monitoring of gene flow over space and time, and the tracking of imported infections. By developing a molecular barcode of individual parasites it will also be possible to distinguish recrudescent from re-infections.

Highly polymorphic microsatellites have been the preferred method of genetic analysis, revealing high levels of diversity and highlighting interesting genotypic patterns and geographical clustering across global populations [8, 9, 10, 11, 12]. The advancement of whole genome sequencing technologies has opened up opportunities to obtain a comprehensive picture of the epidemiology and structural variation of *P. vivax*. There is now the ability to perform genome-wide analysis of the various populations without the need for *in vitro* culture and overcoming difficulties with low parasitaemias and high human DNA contamination [13]. Recent studies using genome-wide SNPs highlighted that signals of natural selection suggest that *P. vivax* is evolving in response to antimalarial drugs and is adapting to regional differences in the human host and the mosquito vector [6,7]. Several other whole genome sequence studies have been published [13, 14, 15, 16], covering 10 countries. Using these and other data, we explore the genetic diversity within and between continents, identify signatures of drug pressure and molecular barcodes that could be useful for determining the source of infections and monitoring parasite populations.

Methods

Samples and sequence data

Publicly available whole genome sequence data for 74 *P. vivax* samples were gathered from multiple sources, and included reference strains (India VII, Mauritania X, North Korea II, Brazil I, Sal-1 from El Salvador (see [2])), field and clinical isolates (Cambodia (n = 3) [13], Thailand (n = 39) [13], Madagascar (n = 3) [2,17], Colombia (n = 8) [14] and Peru (n = 11) [7,15]) and clinical samples from travellers (to Papua Indonesia (n = 2) [13], India (n = 2) [13], and

Papua New Guinea (PNG, n = 6) [13]). All sequencing data for non-reference strains were generated using Illumina paired end technologies (read lengths ≥75bp). The raw sequence data were mapped to the Sal-1 reference genome (version 10.0) using *bwa-mem* with default parameters. SNPs (n = 447,232) were identified using the *samtools* software suite (samtools.sourceforge.net) with high quality scores (*phred* score >30, 1 error per 1 kbp). Genotypes were called using ratios of coverage, where the minimal heterozygous calls still present after filtering were converted to the majority genotype if the coverage ratio was 80:20 or greater [18,19]. SNPs were retained if they were biallelic, had low genotype missingness (<10%) and heterozygous (<0.4%) calls. SNPs in regions of extreme coverage and very low coverage were excluded, as well as in non-unique regions (using a k-mer approach with length of 54 bp) and highly polymorphic VIR genes. Two samples were found to have *P. vivax* and *P. falciparum* co-infections (ERR020124 and SRR828528), and were excluded from population genetic analysis. Isolates were retained if they had at least average 10-fold genome-wide coverage, and at most 10% missing genotype calls. The final high quality dataset consisted of 46 (62.2%) isolates (Thailand 22, Southeast Asia 24, South America 11; other 11; **S1 Table**) and 219,288 SNPs, and used as the basis of population genetic analyses. *FreeC* software (http://bioinfo-out.curie.fr/projects/freec/tutorial.html) was used to identify regions of the genome with a significant increase or decrease in read coverage identifying potential copy number variants (CNVs) after accounting for GC bias through coverage normalization. Regions identified as CNVs were inspected visually and assessed using *de novo* assembly methods [20].

Population genetics

Genetic diversity was estimated using the average pairwise nucleotide diversity (π) with the R package *"pegas"*. An in-house R script was used to compute the allele frequency-based Tajima's *D* test [21] to identify genes under balancing selection in the individual populations (values > 2.5; [18]), this method was chosen over the dN/dS approach given the latter being not fit for analysis on individual populations [22]. To detect signals of directional selection, the integrated Haplotype Score (*iHS*) approach [23] was applied to individual populations supported by a principal component analysis (PCA). This approach used the most frequent allele where mixed calls where found so the haplotype analysis will be based on the most abundant strain in each sample [7]**.** *P*-values for *iHS* were computed from standardised values based on a 2-tailed conversion from a Gaussian distribution [19]. The Salvador-I being the reference and oldest sample was used as ancestral haplotype. Multiplicity of infection was estimated using a novel method of counting the unique haplotypes formed by polymorphism on paired sequencing reads (*estMOI*, [24]). For comparisons between populations, we first applied PCA and neighbourhood joining tree clustering based on a matrix of pairwise identity by state values. These analyses were followed by applying the cross population long-range haplotype method (*XP-EHH* [25]**,** *Rsb* implementation [19]) and the population differentiation metric F_{ST} [26]. *P*-values for the *Rsb* estimates were calculated using a Gaussian approximation [19]. A significance threshold of $P < 0.001$ was established for both *iHS* and *Rsb* using bootstrap- and permutation-based simulation approaches [18,19]**.** We used the ranked F_{ST} statistics to identify the informative polymorphism for the barcoding of populations and driving the clustering observed in the PCA. Linkage disequilibrium (LD) was assessed in the two populations with the largest sample sizes (Thailand and South America) using the r^2 and *D'* metrics [27], calculated for pairs of SNPs with different physical separation up to 2 kbp using a sliding window approach. The SNPs were annotated and effects of variants on genes (such as amino acid changes) were predicted using *snpEFF* software [28]**.** The R statistical package was used to analyse SNP data, including implementation of selection analyses using the "rehh" library.

Results

Genetic polymorphisms

The genomic coverage in the nuclear genome was high (median 103-fold, range (30-5973-fold), and in keeping with multiple organellar copies, the mitochondria and apicoplast coverage was 30-fold and 1.8 fold greater than the nuclear coverage. The density of SNPs in the nuclear genome (219,288 SNPs, 1 every 99 bp) was greater than in the mitochondrial (23 SNPs, 1 every 165 bp) and apicoplast genomes (176 SNPs, 1 every 165 bp) (S2 Table). Although 60% of the annotated reference genome is coding (chromosomal range: 54%-64%), approximately half the SNPs in the isolates were found in genic regions (mean 48% per chromosome, range 43% to 52%) (Fig 1A). The proportion of non-synonymous sites is consistent with those found in other *Plasmodium* species, with 52% of coding SNP sites being non-synonymous in the nuclear genome, 36% in the mitochondrion and 56% in the apicoplast. The differences in these genomes suggest they may be subject to differential selective pressure [29]. The majority of SNPs are rare, with nearly half of the mutations (45%) being observed in single samples (Fig 1B) as seen in other *Plasmodium* populations [18]. There was some evidence of polyclonality in 22 samples (Cambodia 1/2, Colombia 5/5, Madagascar 2/2, PNG 2/5, Thailand 11/22).

Analysis of structural variants and copy number variants was limited to Thai, Cambodian and Madagascan isolates, which had high and uniform genomic coverage. CNVs were located

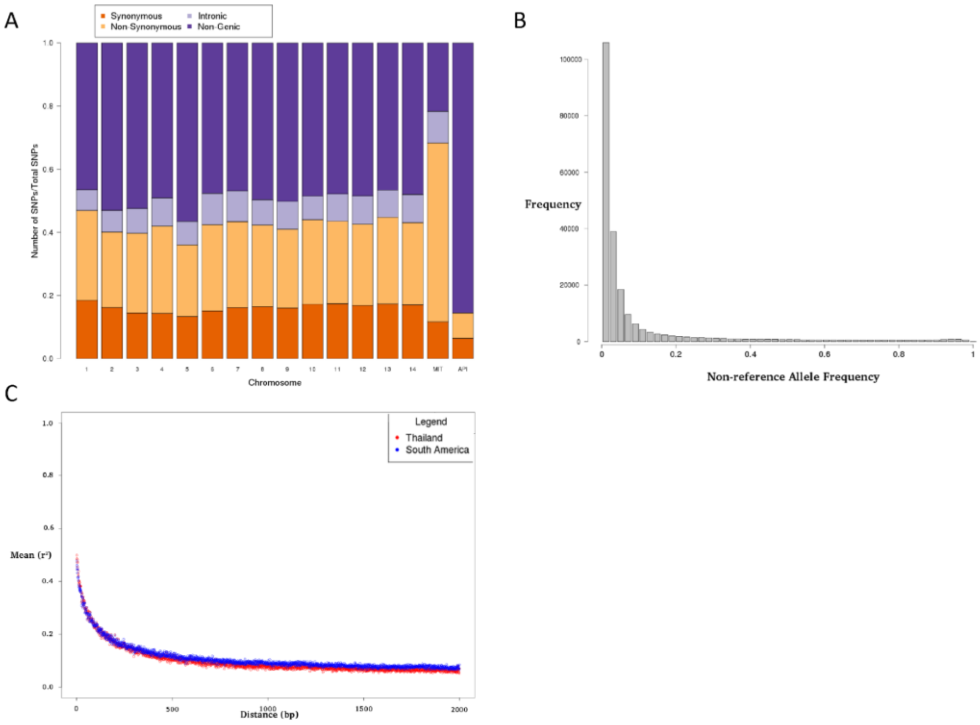


Fig 1. (a) SNP locations by annotation*. (b) Minor allele frequency spectrum indicates a predominance of rare alleles. (c) Linkage disequilibrium (r2) decays rapidly with physical genetic distance. * established using *snpEFF* software.

https://doi.org/10.1371/journal.pone.0177134.g001 less than 1 kbp distance from the *MDR1* gene (chromosome 10, *PVX_080100*) in Cambodian and Thai isolates (**S1 Fig**). Several *MDR1* variants have previously been reported, some considered putative chloroquine- and mefloquine-resistance alleles [14,30–35]. At the *MDR1* locus, we observed either a duplication of ~35kb (position 351kbp to 389kbp, n = 1, Thailand), a major deletion in the promoter region of the gene (n = 7, Thailand; n = 1 Cambodia), or a combination of both structural variants, including two copies, one with the deletion in the promoter and another copy with a complete promoter (n = 4, Thailand); as confirmed by the increase or decrease in coverage and accumulation of split reads in the regions where a break in the coverage occurs. The known duplication in the Duffy binding protein *PvDBP* (chr. 6: 974,000–982,000, *PVX_110810*) in Malagasy [36] was confirmed in one of the two Madagascan isolates (SRR828416). The *PvDBP* gene is potentially involved in erythrocyte invasion, and the duplication was also observed in thirteen Thai isolates. A further duplication was observed in *Pv-fam-e* (a RAD gene, chr. 5: 895,000–900,000, n = 8, Thailand), a gene linked to *P. vivax* selectivity for young erythrocytes and/or immune evasion [36].

Assessing genetic diversity, LD and positive directional selection

The average polymorphism (pair-wise mismatches measured by nucleotide diversity π) was calculated by gene and chromosome. There was little difference across the chromosomes with mean 11.1x10^{-4} (range 6.0 x 10^{-4} to 19.0x10^{-4}), which is consistent with other studies with similar sample size [14] as well as larger datasets when restricted to high quality SNPs (1.5x10^{-3}) [6, 7]. LD decays rapidly for non-rare polymorphism (minor allele frequency $\geq$ 5%) within a few hundred base pairs, and reaches a baseline within 500bp in South American and Thai nuclear genomes (**Fig 1C**). Like *P. falciparum*, there is a high correlation between non-rare SNPs (median *D'* 0.918, range 0.425–1) in the mitochondrial and apicoplast genomes supporting the inference that the organelles are co-inherited and supporting the view that these SNPs have potential utility for barcoding [29].

To examine evidence for signatures of positive natural selection we calculated the *iHS* metric in the Thailand and South America populations, informed by the population structure reported in **S3 Fig**. Five contiguous loci of strong positive directional selection were identified, including the *MRP1* gene (*PVX_097025*) and its promoter region in Thailand, and a region surrounding the *MRP2* gene (*PVX_097025*, *P. falciparum* homologue associated with primaquine and antifolate drug sensitivity [14]). Several surface proteins were identified in both populations, including the *MSP7* and *MSP3.1*, which are thought to be under selection pressure due to their role in erythrocyte invasion and strong vaccine candidates and have been identified before by other studies using sanger sequencing [29] (**Table 1**, **Table 2**, **S2 Fig**). In addition, some helicases showed strong signals of selection (*PVX_088190* and *PVX_111220*) which were also detected in the same study [36] reinforcing the method used. Furthermore, we identified in South America a proximal region of selection (chr14: 1,414,164–1,479,586) described elsewhere [7].

Allele frequency spectrum and balancing selection

The allele frequency spectrum of different classes of nucleotide sites all show an excess of rare alleles, with coding, non-synonymous, synonymous and intergenic sites more skewed than expected under a Wright-Fisher model of constant population size [18]. This observation could indicate a population expansion in the recent past, where as a population grows in size, the frequency of rare alleles also increases [18]. The Tajima's *D* method was applied to genes with at least five SNPs in the two main populations (Thailand 4,673 (91.0%) and South America 3,549 (70.0%) genes). The majority of Tajima's *D* values were negative (Thailand 90.2%;

Table 1. Regions under directional selective pressure in Thailand *.

Chr	Position / Range	Max *iHS*	Gene	Annotation
1	284000	5.083	*PVX_087910*	E3 ubiquitin-protein ligase, putative
1	379430	3.511	.	.
2	148413	4.319	.	**Promoter region MRP1**
2	158122	3.452	***PVX_097025***	**multidrug resistance-associated protein 1, MRP1**
4	576773	4.289	.	.
4	629852	3.483	*PVX_003770*	merozoite surface protein 5 (MSP5)
5	673939	3.566	*PVX_089575*	trafficking protein particle complex protein, TRAPPC2L
7	778719	3.986	.	.
7	1397181	5.829	*PVX_086903*	Plasmodium exported protein, unknown function
8	766604	3.709	*PVX_095055*	Rh5 interacting protein, putative (RIPR)
8	921104	3.451	*PVX_095235*	protein phosphatase inhibitor 2, putative
8	927191	3.489	*PVX_095245*	hypothetical protein, conserved
8	985454	3.778	*PVX_095305*	hypothetical protein, conserved
9	107123	6.186	*PVX_090925*	protein kinase domain containing protein
9	311594	3.590	.	.
9	526557	4.115	*PVX_091440*	hypothetical protein, conserved
10	1222646	5.499	*PVX_097715*	hypothetical protein
10	1225529	4.109	.	.
10	1261650	9.180	.	Promotor region MSP3.1
10	1261852	3.982	*PVX_097670*	merozoite surface protein 3 (MSP3.1)
11	926166	4.034	.	.
12	732115	3.531	*PVX_082735*	thrombospondin-related anonymous protein (TRAP)
12	734223–745860	4.901	*PVX_082730*	hypothetical protein, conserved
12	746536	4.319	.	.
12	751773	5.473	*PVX_082710*	hypothetical protein
12	752332	3.558	.	.
12	765929	6.849	.	.
12	766784	6.170	*PVX_082675*	merozoite surface protein 7 (MSP7)
12	864218	3.561	*PVX_082510*	hypothetical protein
12	865780	3.930	*PVX_082505*	CPW-WPC family protein, putative
12	1020235	5.042	.	.
12	2475528	4.740	.	.
12	2540841	3.633	*PVX_118270*	serine/threonine protein kinase, putative
12	2622092	3.446	*PVX_118345*	protein transport protein SEC7, (SEC7)
12	2638874	4.900	.	.
12	2671299	3.501	*PVX_118380*	GTP-binding protein, putative
12	2732268	3.486	*PVX_118460*	hypothetical protein, conserved
14	3028986	5.168	.	.

* |iHS| > 3.

https://doi.org/10.1371/journal.pone.0177134.t001

South America 64.4%), reinforcing the presence of an excess of low frequency and singleton polymorphisms, potentially due to population expansion in the recent past or purifying selection. For Thailand, we identified 398 (8.5%) genes with positive Tajima's *D* values, of which 14 were in excess of 2.5 and potentially under balancing selection (Table 3). Similarly, for South America, of the 1,260 (35.5%) values that were positive, 12 were in excess of 2.5 (Table 3). The loci under potential balancing selection in both populations encode proteins with predominantly roles surface proteins (e.g. MSPs) and antigens. The majority of the 26 genes identified

Table 2. Regions under directional selective pressure in South America *.

Chr	Position / Range	Max *iHS*	Gene	Annotation
1	490369	3.020	*PVX_088190*	helicase, putative
1	662401	3.050	*PVX_093585*	SF-assemblin, putative
2	244790–1	5.316	.	Promoter region PVX_081315
3	247791	3.813	*PVX_000860*	hypothetical protein, conserved
3	372679	4.109	*PVX_000695*	hypothetical protein, conserved
4	574831	3.029	*PVX_003830*	serine-repeat antigen 5 (SERA)
5	560637	3.649	*PVX_089445*	RAD protein (Pv-fam-e)
5	1046482	3.358	*PVX_090020*	hypothetical protein, conserved
6	627960–1	3.605	*PVX_111230*	hypothetical protein, conserved
7	437942–60	4.773	*PVX_099005*	cysteine repeat modular protein 1, CRMP1
7	527651	3.182	*PVX_099125*	pseudouridylate synthase, putative
7	1116251–2	5.712	*PVX_099915*	RNA-binding protein, putative
7	1214179	6.840	*PVX_087145*	nucleolar protein Nop52, putative
8	219359	3.180	*PVX_094405*	hypothetical protein, conserved
9	730034	3.611	*PVX_091700*	circumsporozoite-related antigen, EXP1
9	751857	3.149	.	Promoter region PVX_091715
9	829890	3.161	*PVX_091770*	calcium-dependent protein kinase 7, CDPK7
9	1042906–7	5.519	*PVX_092035*	6-phosphofructokinase, putative
9	1136873	3.110	*PVX_092160*	hypothetical protein, conserved
10	380535	3.108	*PVX_080110*	G10 protein, putative
10	1063432	3.592	*PVX_097895*	TBC domain containing protein
10	1257251	3.291	.	Promoter region MSP3.2
10	1260441–2	5.201	.	1 Kb from MSP3.2
11	822715	4.026	*PVX_114575*	transmembrane amino acid transporter protein
11	1973708	6.074	.	.
12	955488	3.652	*PVX_082400*	myosin C, putative
12	1317195	3.435	*PVX_116815*	hypothetical protein, conserved
13	215135	3.283	*PVX_084350*	hypothetical protein, conserved
13	611668	5.708	*PVX_084755*	hypothetical protein, conserved
13	856226–30	4.519	*PVX_085030*	aspartyl protease, putative
14	1275835	3.114	*PVX_123250*	aquaporin, putative (AQP2)
14	1665191	3.572	*PVX_123685*	histone-lysine N-methyltransferase, SET10
14	1875833	4.986	*PVX_123890*	hypothetical protein, conserved

* |iHS| > 3.

https://doi.org/10.1371/journal.pone.0177134.t002

in this study have had positive indices of balancing selection in previous studies [37, 38], or have orthologues in *P*. *falciparum* [18].

Population structure and evidence of differing directional selection in populations

Both a principal component and a neighbourhood joining tree analysis (Fig 2, S4 Fig) revealed clustering by continent, in keeping with similar *P*. *falciparum* analyses [18, 19]. The across population long-range haplotype method (*Rsb* implementation) was applied to compare Thailand to the South American population, to identify regions potentially under recent directional selection (Table 4). We detected several loci including at multidrug resistance-associated protein *MRP1 (PVX_097025)*, and the CCR4-associated factor 1 (*CAF1*, *PVX_123230*) located within 20kb of *DHPS* (associated with resistance to sulfadoxine [14]). Five non-synonymous mutations were identified in the *DHPS* gene (M616T, P553A, P383A, P382R, P382A), with evidence that the P383A has driven toward fixation across all geographical regions. Except for mutation in codon 616, all the others have been previously reported [39–42]. The *DHFR* gene, associated with resistance to pyrimethamine (part of the SP drug combination), exhibited elevated *Rsb* (>3). Seven non-synonymous mutations were identified, including the previously described S58R and S117N [42–44] that were fixed across populations, and F57I/L and T61M [44–46] that were absent from South America (S3 Table). No evidence was observed of a hard sweep around the *MDR1* copy number gene. However, nine non-synonymous SNP mutations were identified, five of which have been reported previously. These included the fixed alleles T958M and M908L, F1076L at high frequency across populations, and G698S and S513R absent from South America [41–44]. There was no evidence of a sweep around the *P. vivax* orthologues of the *falciparum* chloroquine related *CRT* (*pvcrt-o*, *PVX_087980*) or GTP cyclohydrolase I folate pathway (*GTPCH*, *PVX_123830*) genes. No common non-synonymous mutations were identified within the *CRT* gene, while 7 low frequency non-synonymous SNPs where identified in the *GTPCH* locus.

The *Rsb* analysis also revealed loci associated with the diversity of vectors, including the *P28* (*PVX_111180*) gene expressed in the surface of the ookinete stage during the mosquito part of the life cycle, *pv47* (*PVX_083240*) and *pv48/45* (*PVX_083235*) involved in the transmission of the parasite. There are continental-specific *pv47* and *pv48/45* SNPs (and haplotypes) as previously found [47, 48], consistent with the presence of different species of mosquito in each the regions [49], resembling a similar pattern found in *P. falciparum* [50].

Table 3. Genetic regions under potential balancing selection pressure in South America (SA) and Thailand (T)*.

Chr.	Gene Start	Gene End	Tajima's *D*	Gene**	Annotation	Population
1	521978	527387	3.265	*PVX_088235*	ferlin, putative	SA
3	19187	30715	14.166, 7.134	*PVX_001080*	hypothetical protein, conserved	SA, T
4	265018	267216	2.928	*PVX_002785*	ATP-dependent acyl-CoA synthetase	T
4	562755	566374	4.871, 6.085	*PVX_003840*	serine-repeat antigen 3 (SERA)	SA, T
4	567313	571093	4.944	*PVX_003835*	serine-repeat antigen 1 (SERA)	T
4	572172	575852	3.36	*PVX_003830*	serine-repeat antigen 5 (SERA)	T
4	596283	600192	4.224	*PVX_003805*	serine-repeat antigen (SERA), putative	SA
5	1297808	1301010	3.279	*PVX_090285*	Pvstp1, putative	T
5	1345372	1354047	15.907	*PVX_090325*	reticulocyte binding protein 2c (RBP2c)	T
5	1358748	1360820	5.762	*PVX_090330*	reticulocyte binding protein 2 (PvRBP-2)	T
7	1157742	1162997	5.593, 4.124	*PVX_099980*	merozoite surface protein 1 (MSP1)	SA, T
9	6424	7811	4.195	*PVX_090835*	hypothetical protein	T
10	22046	23460	2.925	*PVX_079700*	hypothetical protein, conserved	T
10	65101	69250	2.793	*PVX_079750*	hypothetical protein, conserved	T
10	1187639	1188909	5.206	*PVX_097760*	60S ribosomal protein L31, RPL31	SA
10	1218512	1221845	2.939,5.585	*PVX_097720*	merozoite surface protein 3 (MSP3.10)	SA, T
10	1272354	1274193	3.869	*PVX_097660*	4-diphosphocytidyl-2-C-methyl- kinase, IspE	SA
10	1306384	1308153	5.991	*PVX_097600*	hypothetical protein, conserved	SA
12	751041	752204	5.578	*PVX_082710*	hypothetical protein	SA
13	37121	59181	3.876	*PVX_084160*	dynein heavy chain, putative	SA
13	128618	131751	4.099	*PVX_084260*	hypothetical protein, conserved	SA
14	3044644	3046339	2.773	*PVX_101575*	hypothetical protein, conserved	T

* Tajima's *D* > 2.5

** at least 5 SNPs per gene.

https://doi.org/10.1371/journal.pone.0177134.t003

Towards molecular barcoding of *P. vivax*

The development of molecular barcode for *P. vivax* could ultimately assist with surveillance and disease control. Previous work [51] has described a 42 SNP barcode to classify geographically *P. vivax* across 7 countries. Across the 46 isolates analysed here, we found 3 SNPs in the barcode to be either non-segregating or not passing quality control filtering. Use of the remaining 39 SNPs led to imperfect clustering by continent (**S4 Table, S5 Fig**). Application of the F_{ST} population differentiation metric identified SNPs driving the observed differences between Thailand, South America and other populations (**S5 Table**). These SNPs occurred in drug resistance loci, including *MRP1 (PVX_097025)*, *DHPS* (*PVX_123230*) and *UBP1* (*PVX_081540*) (all $F_{ST} > 0.72$), and in close proximity (e.g. *PVX_089960* within 8kb of *DHFR)*. Population differentiation due to genetic diversity in drug resistant loci is also observed in *P. falciparum* [18,19].

Previous work has proposed the mitochondria and apicoplast organellar genomes as candidate regions for a barcode [29]**.** Genotyping of organellar markers would benefit from greater copy number and coverage as well as highly conserved sequences [29]. Eight markers across five apicoplast genes could differentiate Thai and Southeast Asian samples from the other isolates, and two non-genic markers were found to be exclusive to South America (all F_{ST}>0.7, **S6 Table**). No informative mitochondrial markers were identified (all F_{ST}<0.7). Further, as the

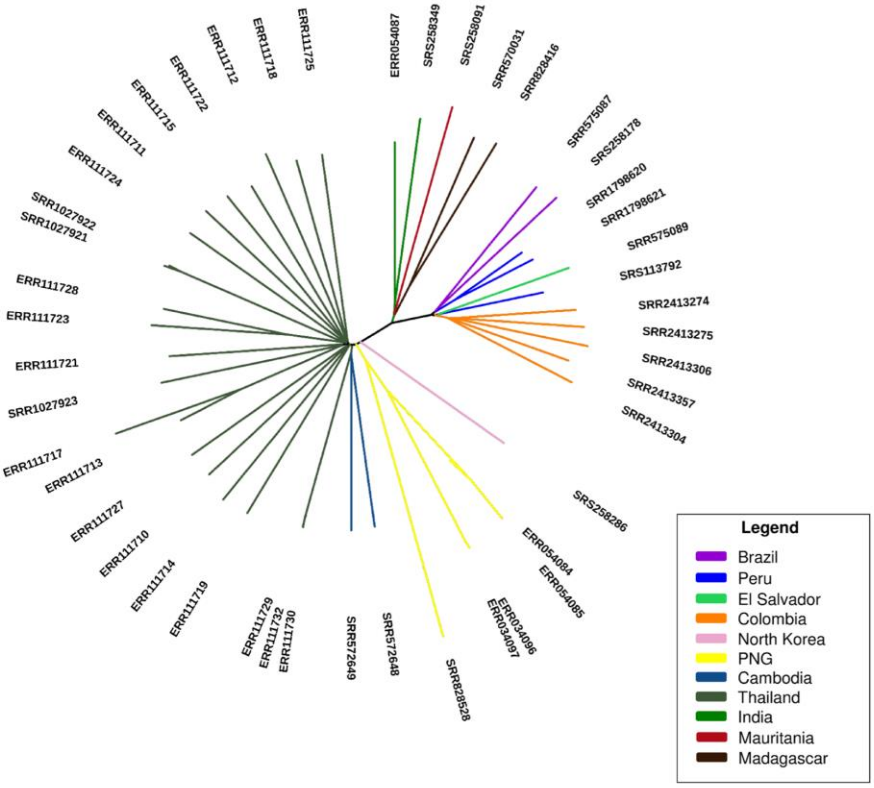


Fig 2. Population structure analysis based on 219,288 SNPs shows clustering by continent.

https://doi.org/10.1371/journal.pone.0177134.g002

Table 4. Regions under directional selective pressure between Thailand and South America *.

Chr	Position/Range	*Rsb*	Gene	Annotation
2	145708–151606	11.80	.	**Promoter region MRP1**
2	154067–158122	5.230	***PVX_097025***	**multidrug resistance-associated protein 1, MRP1**
2	175191–176803	3.717	*PVX_081215*	hypothetical protein, conserved
4	91457	5.879	*PVX_002550*	hypothetical protein, conserved
4	607568–607837	4.457	*PVX_003795*	serine-repeat antigen (SERA)
4	629831–630120	5.205	*PVX_003770*	merozoite surface protein 5
5	1132736	10.20	*PVX_090105*	holo-[acyl-carrier-protein] synthase, putative (ACPS)
5	964771	3.624	***PVX_089950***	**bifunctional dihydrofolate reductase-thymidylate synthase, DHFR-TS**
6	199049–199165	4.703	PVX_001850	hypothetical protein
6	605656–608119	3.788	PVX_111260	hypothetical protein, conserved
6	635433–635539	4.406	*PVX_111220*	RNA helicase, putative
6	661816	6.048	***PVX_111180***	**28 kDa ookinete surface protein, (P28)**
7	1396929–1396961	4.708	.	Promoter region PVX_086903
7	1397181	4.700	*PVX_086903*	Plasmodium exported protein, unknown function
8	219359–220251	5.257	*PVX_094405*	hypothetical protein, conserved
8	1417014–1417038	4.406	PVX_119515	hypothetical protein, conserved
8	1533222	3.214	*PVX_119360*	hypothetical protein
9	419318–419619	4.971	.	Promoter region PVX_091307
9	920056–920166	4.676	*PVX_091880*	hypothetical protein, conserved
9	1048990	3.304	*PVX_092040*	geranylgeranyl pyrophosphate synthase (GGPPS)
9	1229833	3.296	*PVX_092275*	apical membrane antigen 1 (AMA1)
10	1251585–1257251	7.094	.	Promoter region MSP3.2
10	1257754–1257815	6.617	*PVX_097675*	merozoite surface protein 3 (MSP3.2)
11	1517269	3.234	*PVX_113775*	6-cysteine protein (P12)
11	1223546–1223790	3.816	.	Promoter region PVX_114125
11	1383108–1383155	6.010	*PVX_113925*	hypothetical protein, conserved
12	286960	3.227	***PVX_083240***	**6-cysteine protein (P47)**
13	141889–142286	5.680	*PVX_084280*	hypothetical protein, conserved
13	620154–620261	5.922	.	Promoter region PVX_084770
13	731328–792522	5.375	*PVX_084860*	hypothetical protein, conserved
13	1034635–1034718	4.101	PVX_085235	hypothetical protein
13	1042774	5.406	*PVX_085245*	hypothetical protein, conserved
13	1553113	4.126	*PVX_085835*	hypothetical protein, conserved
14	1231525–1231528	6.056	***PVX_123205***	**CCR4-associated factor 1, (CAF1)**
14	1429874	3.903	*PVX_123415*	adrenodoxin-type ferredoxin, putative

* *Rsb* > 3; genes in bold refer to loci related with mosquito life stages of the parasite or drug-resistance.

https://doi.org/10.1371/journal.pone.0177134.t004

organelle genomes are known to be highly conserved between *Plasmodia* species, when comparing a set of *P. falciparum* geographical markers [26] to *P. vivax* sequences, we found evidence of positions close in the sequence. Two of the samples (ERR020124 and SRR828528) had a high density of mixed calls in the organellar genomes, in this case, a signature of *P. falciparum* overlaying onto *P. vivax* (S6 Fig). In general, this density signature is indicative of a co-infection of *P. vivax* with another *Plasmodium spp*. By comparing the sequencing reads to the *Plasmodium knowlesi* reference genome [52]**,** there was no evidence of any *vivax* and *knowlesi* co-infections. However, the presence of a unique triallelic SNP reinforces the potential for an organellar inter-plasmodia species barcode (S6 Fig).

Discussion

Several studies have previously described the genomic diversity of *P. vivax* populations using whole genome data, but with low sample sizes. Recently, two papers using a combined collection of over 400 isolates from 17 countries described major genomic diversity in *Plasmodium vivax* [6, 7]. Here we analysed a complementary collection of 46 high quality isolates spanning 10 countries across 4 continents in order to position them within the context of this new work. As expected we confirmed that *P. vivax* genomic diversity is greater compared to *P. falciparum*, and even at a relatively low sample size, the samples clustered geographically. We reveal a wider genomic distance between South American and Southeast Asian continents than observed between *P. falciparum* African and Southeast Asian populations [6, 18, 19], highlighted by the greater and more uniform distribution of SNPs with a high F_{ST} across the genome. Hotspots of selection pressure were identified, including the previously reported *MRP1*, *DHPS* [14] and other putative drug resistance genes, as well as several loci related with the mosquito stage of the parasite life cycle. The latter observation is consistent with recent work [6,7] and the presence of different *Anopheles* species across continents. We identified structural variants, including extra copies and deletions in the promoter region of the *MDR1* gene, a locus associated with multiple antimalarial drugs [14]**.** We also confirmed the duplication in the Duffy binding protein gene (*PvDBP*) in a Madagascan sample, and detected it in Thai isolates. This duplication has been found in parasites from several regions in Africa, South America and Asia [6,37]**.** Many of these locations are areas where Duffy-negative individuals make up >45% of the population. However other regions like Cambodia do not present Duffy-negative individuals [53]. It has been theorized that the duplication allows the parasite to infect Duffy negative individuals [53], however more research is needed in this area.

Microsatellite genotyping has been used previously to cluster geographically *P. vivax* isolates, and together with antigen genotyping identify mixed infections and extent of transmission, used as the basis of genetic epidemiology. In comparison, whole genome sequencing provides a higher specificity in the application of geographical clustering [51]**.** While other studies have focused on creating a barcode using the nuclear genome [51], we also considered organelle genomes (mitochondrion and apicoplast), which are more stable over time, do not undergo recombination and are co-inherited [29]. The analysis revealed organellar markers that are potentially Southeast Asian and South American specific, and others that highlighted the presence of multi-species mixed infections. The sequencing of large numbers of isolates, beyond currently published samples sizes, will be required to establish robust intra- and interspecies organellar-based barcode. Such large-scale datasets across multiple regions will also serve to identify the high genomic diversity that lies within and between *P. vivax* populations, which could be exploited for biological insights, including elucidating drug resistance and invasion mechanisms, and ultimately measures of disease control.

Conclusion

This study has shown that genomic diversity that lies within and between *P. vivax* populations can be used to elucidate potential drug resistance and invasion mechanisms, as well as facilitate the molecular barcoding of the parasite for surveillance applications.

Supporting information

S1 Fig. Structural variants located around the *MDR1* gene (chromosome 10) in the Thailand population; (i) a sample without a copy number variant or deletion (even coverage), **(ii)** a major deletion in the promoter region of the gene (n = 7); **(iii)** duplication of ~35kb (position 351kbp to 389kbp, n = 1); and (iv) a combination of both structural variants **(ii)** and **(iii)**,

including two copies, one with the deletion in the promoter and another copy with a complete promoter (n = 4, Thailand). The horizontal dashed line is average chromosomal coverage and the red outline encloses the promoter region of the *MDR1* gene.
(TIFF)

S2 Fig. Intra-population evidence of directional selective pressure (*iHS) a)** Thailand **b)** South America. * *iHS* integrated haplotype score; see Table 1 and Table 2 for a summary of the hits.
(TIFF)

S3 Fig. Principal component analysis based on 225k SNPs reveals strong clustering of isolates by continent.
(PNG)

S4 Fig. Identifying regions under directional selective pressure between Thailand and South America. Blue line: $|Rsb| > 3$ (P<0.003); Red line represents a human GWAS cut-off; see Table 4 for a summary of the hits.
(PNG)

S5 Fig. Principal component analysis based on the previously characterised 42 barcoding SNPs* does not reveal strong population clustering. * SNPs and genotypes are shown in S4 Table
(PNG)

S6 Fig. Signatures of a mixed species infection based on heterozygous calls in mitochondrial markers (positions: 3,736–3,935bp).
(PNG)

S1 Table. The 46 study isolates.
(DOCX)

S2 Table. The SNPs.
(DOCX)

S3 Table. Non-synonymous mutations in candidate genes.
(DOCX)

S4 Table. Previously characterised 42 barcoding SNPs* in the 46 study isolates.
(DOCX)

S5 Table. Sites of population differentiation between Thailand and South America.
(DOCX)

S6 Table. Population informative apicoplast variants.
(DOCX)

Acknowledgments

We thank Samuel Assefa, Francesc Coll, and Mark Preston for providing computing support. Sequence data were analysed using the MRC UK funded eMedlab computing resource.

Author Contributions

Conceptualization: EDB CJS CR SC TGC.

Data curation: EDB.

Formal analysis: EDB ZW WC SC TGC.

Funding acquisition: TGC CR.

Supervision: FRM SC TGC.

Writing – original draft: EDB SC TGC.

Writing – review & editing: EDB ZW WC FRM CJS CR SC TGC.

References

1. World Health Organization. World malaria report 2013. http://www.who.int/malaria/publications/world_malaria_report_2013/en/
2. Bright AT, Tewhey R, Abeles S, Chuquiyauri R, Llanos-Cuentas A, Ferreira MU, et al. Whole genome sequencing analysis of Plasmodium vivax using whole genome capture. BMC Genomics 2012; 13:262. https://doi.org/10.1186/1471-2164-13-262 PMID: 22721170
3. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature 2008; 455:757–63. https://doi.org/10.1038/nature07327 PMID: 18843361
4. Arnott A, Barry AE, Reeder JC. Understanding the population genetics of Plasmodium vivax is essential for malaria control and elimination. Malar J 2012; 11:14. https://doi.org/10.1186/1475-2875-11-14 PMID: 22233585
5. Neafsey DE, Galinsky K, Jiang RH, Young L, Sykes SM, Saif S, et al. The malaria parasite Plasmodium vivax exhibits greater genetic diversity than Plasmodium falciparum. Nat Genet 2012; 44:1046–50. https://doi.org/10.1038/ng.2373 PMID: 22863733
6. Hupalo DN, Luo Z, Melnikov A, Sutton PL, Rogov P, Escalante A, et al. Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax. Nat Genet 2016, 48(8), 953–958. https://doi.org/10.1038/ng.3588 PMID: 27348298
7. Pearson RD, Amato R, Auburn S, Miotto O, Almagro-Garcia J, Amaratunga C, et al. Genomic analysis of local variation and recent evolution in Plasmodium vivax. Nat Genet 2016; 48(8), 959–964. https://doi.org/10.1038/ng.3599 PMID: 27348299
8. Imwong M, Nair S, Pukrittayakamee S, Sudimack D, Williams JT, Mayxay M, et al. Contrasting genetic structure in Plasmodium vivax populations from Asia and South America. Int J Parasitol 2007; 37:1013–22. https://doi.org/10.1016/j.ijpara.2007.02.010 PMID: 17442318
9. Dharia NV, Bright AT, Westenberger SJ, Barnes SW, Batalov S, Kuhen K, et al. Whole-genome sequencing and microarray analysis of ex vivo Plasmodium vivax reveal selective pressure on putative drug resistance genes. Proc Natl Acad Sci USA 2010; 107:20045–50. https://doi.org/10.1073/pnas.1003776107 PMID: 21037109
10. Gunawardena S, Karunaweera ND, Ferreira MU, Phone-Kyaw M, Pollack RJ, Alifrangis M, et al. Geographic structure of Plasmodium vivax: microsatellite analysis of parasite populations from Sri Lanka, Myanmar, and Ethiopia. Am J Trop Med Hyg 2010; 82:235–42. https://doi.org/10.4269/ajtmh.2010.09-0588 PMID: 20133999
11. Koepfli C, Rodrigues PT, Antao T, Orjuela-Sánchez P, Van den Eede P, Gamboa D, et al. Plasmodium vivax Diversity and Population Structure across Four Continents. PLoS Negl Trop Dis 2015; 9: e0003872. https://doi.org/10.1371/journal.pntd.0003872 PMID: 26125189
12. Kim JY, Goo YK, Zo YG, Ji SY, Trimarsanto H, To S et al. Further Evidence of Increasing Diversity of Plasmodium vivax in the Republic of Korea in Recent Years. PLoS One 2016; 11:e0151514. https://doi.org/10.1371/journal.pone.0151514 PMID: 26990869
13. Auburn S, Marfurt J, Maslen G, Campino S, Ruano Rubio V, Manske M, et al. Effective preparation of Plasmodium vivax field isolates for high-throughput whole genome sequencing. PLoS One 2013; 8: e53160. https://doi.org/10.1371/journal.pone.0053160 PMID: 23308154
14. Winter DJ, Pacheco MA, Vallejo AF, Schwartz RS, Arevalo-Herrera M3, Herrera S, et al. Whole Genome Sequencing of Field Isolates Reveals Extensive Genetic Diversity in Plasmodium vivax from Colombia. PLoS Neg Trop Dis 2015; 9:e0004252.
15. Flannery EL, Wang T, Akbari A, Corey VC, Gunawan F, Bright AT, et al. Next-Generation Sequencing of Plasmodium vivax Patient Samples Shows Evidence of Direct Evolution in Drug-Resistance Genes. ACS Infect Dis 2015; 1:367–379. https://doi.org/10.1021/acsinfecdis.5b00049 PMID: 26719854

16. Shen HM, Chen SB, Wang Y, Chen JH. Whole-genome sequencing of a Plasmodium vivax isolate from the China-Myanmar border area. Mem Inst Oswaldo Cruz 2015; 110:814–6. https://doi.org/10.1590/0074-02760150216 PMID: 26517664

17. Parobek CM, Bailey JA, Hathaway NJ, Socheat D, Rogers WO, Juliano JJ. Differing Patterns of Selection and Geospatial Genetic Diversity within Two Leading Plasmodium vivax Candidate Vaccine Antigens. PLoS Negl Trop Dis 2014, 8: e2796. https://doi.org/10.1371/journal.pntd.0002796 PMID: 24743266

18. Ocholla H, Preston MD, Mipando M, Jensen AT, Campino S, MacInnis B, et al. Whole-genome scans provide evidence of adaptive evolution in Malawian Plasmodium falciparum isolates. J Infect Dis 2014; 210:1991–2000. https://doi.org/10.1093/infdis/jiu349 PMID: 24948693

19. Samad H, Coll F, Preston MD, Ocholla H, Fairhurst RM, Clark TG. Imputation-based population genetics analysis of Plasmodium falciparum malaria parasites. PLoS Genet 2015; 11:e1005131. https://doi.org/10.1371/journal.pgen.1005131 PMID: 25928499

20. Campino S, Diez Benavente E, Assefa S, Thompson E, Drought LG, Taylor CJ, et al. Genomic variation in two gametocyte non-producing Plasmodium falciparum clonal lines. Malar J 2016; 15:229. https://doi.org/10.1186/s12936-016-1254-1 PMID: 27098483

21. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989; 123:585–95. PMID: 2513255

22. Kryazhimskiy S, Plotkin JB. The Population Genetics of dN/dS. PLOS Genetics 2008; 4(12): e1000304. https://doi.org/10.1371/journal.pgen.1000304 PMID: 19081788

23. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006; 4:e72. https://doi.org/10.1371/journal.pbio.0040072 PMID: 16494531

24. Assefa S, Preston M, Campino S, Ocholla H, Sutherland CJ, Clark TG. estMOI: Estimating multiplicity of infection using parasite deep sequencing data. Bioinformatics 2014; 30:1292–4. https://doi.org/10.1093/bioinformatics/btu005 PMID: 24443379

25. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature 2007; 449:913–8. https://doi.org/10.1038/nature06250 PMID: 17943131

26. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting $F_{ST.}$ Nat Rev Genet 2009; 10:639–50. https://doi.org/10.1038/nrg2611 PMID: 19687804

27. Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor Appl Genet 1968; 38:226–31. https://doi.org/10.1007/BF01245622 PMID: 24442307

28. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. Fly 2012, 6(2), 80–92. https://doi.org/10.4161/fly.19695 PMID: 22728672

29. Preston MD, Campino S, Assefa SA, Thompson E, Drought LG, Taylor CJ, et al. A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. Nat Commun 2014, 5:4052. https://doi.org/10.1038/ncomms5052 PMID: 24923250

30. Brega S, Meslin B, de Monbrison F, Severini C, Gradoni L, Udomsangpetch R, et al. Identification of the Plasmodium vivax mdr-like gene (pvmdr1) and analysis of single-nucleotide polymorphisms among isolates from different areas of endemicity. J Infect Diseas 2005, 191(2), 272–277.

31. Schousboe ML, Ranjitkar S, Rajakaruna RS, Amerasinghe PH, Morales F, Pearce R, et al. Multiple Origins of Mutations in the mdr1 Gene—A Putative Marker of Chloroquine Resistance in P. vivax. PLoS Negl Trop Dis 2015, 9(11), e0004196. https://doi.org/10.1371/journal.pntd.0004196 PMID: 26539821

32. Barnadas C, Ratsimbasoa A, Tichit M, Bouchier C, Jahevitra M, Picot S, et al. Plasmodium vivax resistance to chloroquine in Madagascar: clinical efficacy and polymorphisms in pvmdr1 and pvcrt-o genes. Antimicrobial Agents and Chemo 2008, 52(12), 4233–4240.

33. Lu F, Lim CS, Nam DH, Kim K, Lin K, Kim TS, et al. Genetic polymorphism in pvmdr1 and pvcrt-o genes in relation to in vitro drug susceptibility of Plasmodium vivax isolates from malaria-endemic countries. Acta Tropica 2011, 117(2), 69–75. https://doi.org/10.1016/j.actatropica.2010.08.011 PMID: 20933490

34. Suwanarusk R, Russell B, Chavchich M, Chalfein F, Kenangalem E, Kosaisavee V, et al. Chloroquine Resistant Plasmodium vivax: In Vitro Characterisation and Association with Molecular Polymorphisms. PLoS ONE 2007, 2(10), e1089. https://doi.org/10.1371/journal.pone.0001089 PMID: 17971853

35. Imwong M, Pukrittayakamee S, Pongtavornpinyo W, Nakeesathit S, Nair S, Newton P, et al. Gene amplification of the multidrug resistance 1 gene of Plasmodium vivax isolates from Thailand, Laos, and Myanmar. Antimicrobial Agents and Chemotherapy 2008, 52(7), 2657–2659. https://doi.org/10.1128/AAC.01459-07 PMID: 18443118

36. Cornejo EO, Fisher D, and Escalante AA. Genome-Wide Patterns of Genetic Polymorphism and Signatures of Selection in Plasmodium vivax. Genome Biol Evol 2015; 7:106–119.

37. Menard D, Chan ER, Benedet C, Ratsimbasoa A, Kim S, Chim P, et al. Whole genome sequencing of field isolates reveals a common duplication of the Duffy binding protein gene in Malagasy Plasmodium vivax strains. PLoS Negl Trop Dis 2013; 7:e2489. https://doi.org/10.1371/journal.pntd.0002489 PMID: 24278487

38. Ord R, Polley S, Tami A, Sutherland CJ. High sequence diversity and evidence of balancing selection in the Pvmsp3α gene of Plasmodium vivax in the Venezuelan Amazon. Mol Biochem Parasitol 2005; 144:86–93. https://doi.org/10.1016/j.molbiopara.2005.08.005 PMID: 16159677

39. Triglia T, Cowman AF. Primary structure and expression of the dihydropteroate synthetase gene of Plasmodium falciparum. Proc National Academy Sci 1994, 91(15), 7149–7153.

40. Hawkins VN, Suzuki SM, Rungsihirunrat K, Hapuarachchi HC, Maestre A, Na-Bangchang K, Sibley CH. Assessment of the origins and spread of putative resistance-conferring mutations in Plasmodium vivax dihydropteroate synthase. Am J Trop Med Hyg 2009, 81(2), 348–355. PMID: 19635897

41. Imwong M, Sudimack D, Pukrittayakamee S, Osorio L, Carlton JM, Day NP, et al. Microsatellite variation, repeat array length, and population history of Plasmodium vivax. Mol Biol Evol 2006, 23(5):1016–8. https://doi.org/10.1093/molbev/msj116 PMID: 16507919

42. Hawkins VN, Joshi H, Rungsihirunrat K, Na-Bangchang K, Sibley CH. Antifolates can have a role in the treatment of Plasmodium vivax. Trends in Parasitology 2007, 23(5), 213–222. https://doi.org/10.1016/j.pt.2007.03.002 PMID: 17368986

43. Hawkins VN, Auliff A, Prajapati SK, Rungsihirunrat K, Hapuarachchi HC, Maestre A, et al. Multiple origins of resistance-conferring mutations in Plasmodium vivax dihydrofolate reductase. Malaria J 2008, 7, 72. https://doi.org/10.1186/1475-2875-7-72

44. Saralamba N, Nakeesathit S, Mayxay M, Newton PN, Osorio L, Kim JR, et al. Geographic distribution of amino acid mutations in DHFR and DHPS in Plasmodium vivax isolates from Lao PDR, India and Colombia. Malaria J 2016, 15(1), 484.

45. Rungsihirunrat K, Na-Bangchang K, Hawkins VN, Mungthin M, Sibley CH. Sensitivity to antifolates and genetic analysis of Plasmodium vivax isolates from Thailand. Am J Trop. Med Hyg 2007, 76(6), 1057–1065. PMID: 17556611

46. Lu F, Lim CS, Nam DH, Kim K, Lin K, Kim TS. Mutations in the antifolate-resistance-associated genes dihydrofolate reductase and dihydropteroate synthase in Plasmodium vivax isolates from malaria-endemic countries. Am J Trop Med Hyg 2010, 83.

47. Tachibana M, Suwanabun N, Kaneko O, Iriko H, Otsuki H, Sattabongkot J, et al. Plasmodium vivax gametocyte proteins, Pvs48/45 and Pvs47, induce transmission-reducing antibodies by DNA immunization. Vaccine 2015, 33(16), 1901–1908. https://doi.org/10.1016/j.vaccine.2015.03.008 PMID: 25765968

48. Vallejo AF, Martinez NL, Tobon A, Alger J, Lacerda MV, Kajava AV, et al. Global genetic diversity of the Plasmodium vivax transmission-blocking vaccine candidate Pvs48/45. Malaria J 2016, 15, 202.

49. Sinka ME, Bangs MJ, Manguin S, Chareonviriyaphap T, Patil AP, Temperley WH, et al. The dominant anopheles vectors of human malaria in the Asia-Pacific region: Occurrence data, distribution maps and bionomic précis. Parasit Vectors 2011; 4:89. https://doi.org/10.1186/1756-3305-4-89 PMID: 21612587

50. Molina-Cruz A, Garver LS, Alabaster A, Bangiolo L, Haile A, Winikor J, et al. The human malaria parasite Pfs47 gene mediates evasion of the mosquito immune system. Science 2013; 340:984–7. https://doi.org/10.1126/science.1235264 PMID: 23661646

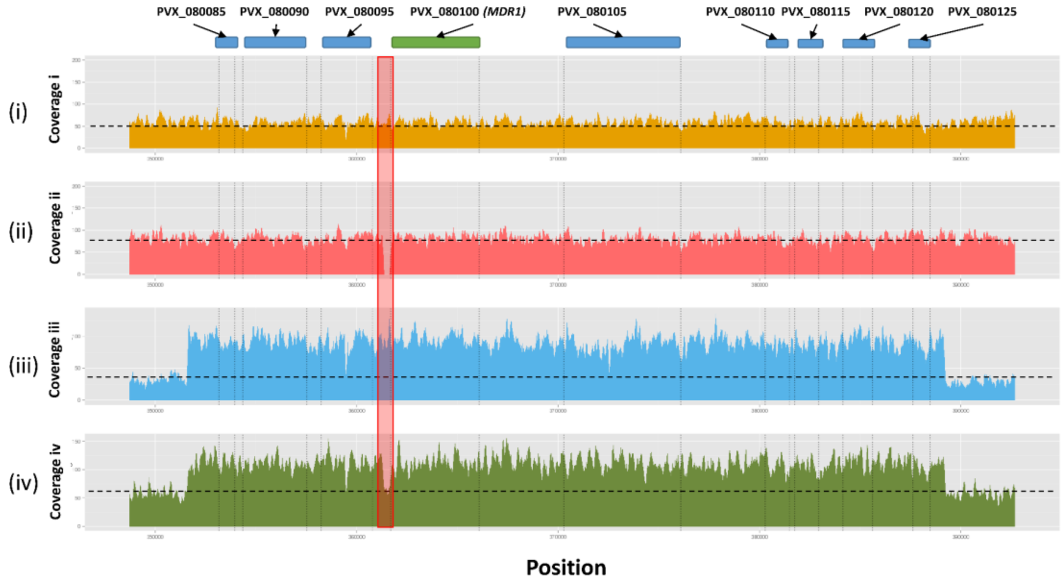
51. Baniecki ML, Faust AL, Schaffner SF, Park DJ, Galinsky K, Daniels RF et al. Development of a single nucleotide polymorphism barcode to genotype Plasmodium vivax infections. PLoS Negl Trop Dis 2015; 9:e0003539. https://doi.org/10.1371/journal.pntd.0003539 PMID: 25781890

52. Pain A, Böhme U, Berry AE, Mungall K, Finn RD, Jackson AP, et al. The genome of the simian and human malaria parasite Plasmodium knowlesi. Nature 2008; 455:799–803. https://doi.org/10.1038/nature07306 PMID: 18843368

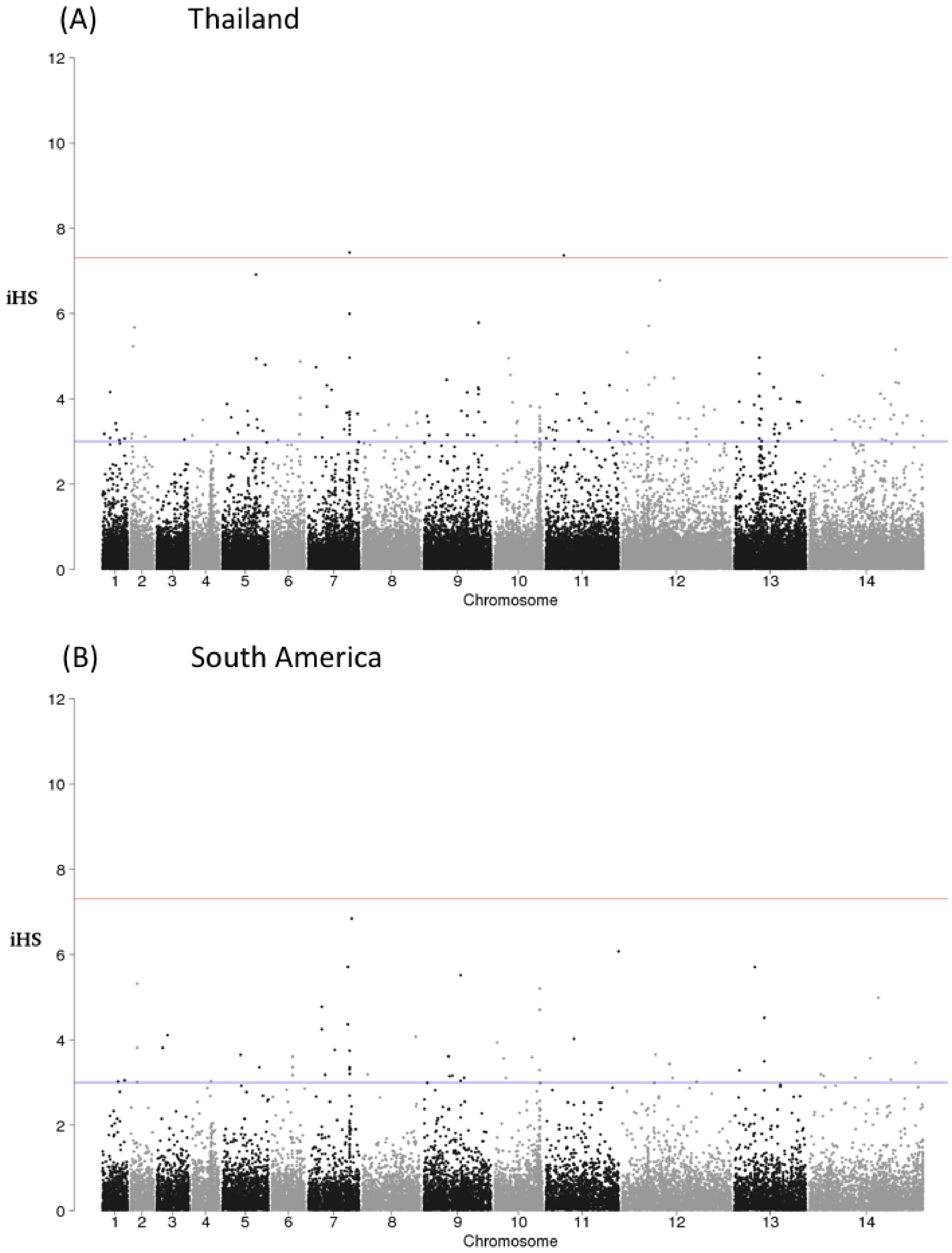
53. Hostetler JB, Lo E, Kanjee U, Amaratunga C, Suon S, Sreng S, et al. Independent Origin and Global Distribution of Distinct Plasmodium vivax Duffy Binding Protein Gene Duplications. PLOS Neglected Tropical Diseases 2016, 10(10), e0005091. https://doi.org/10.1371/journal.pntd.0005091 PMID: 27798646

Supplementary Information

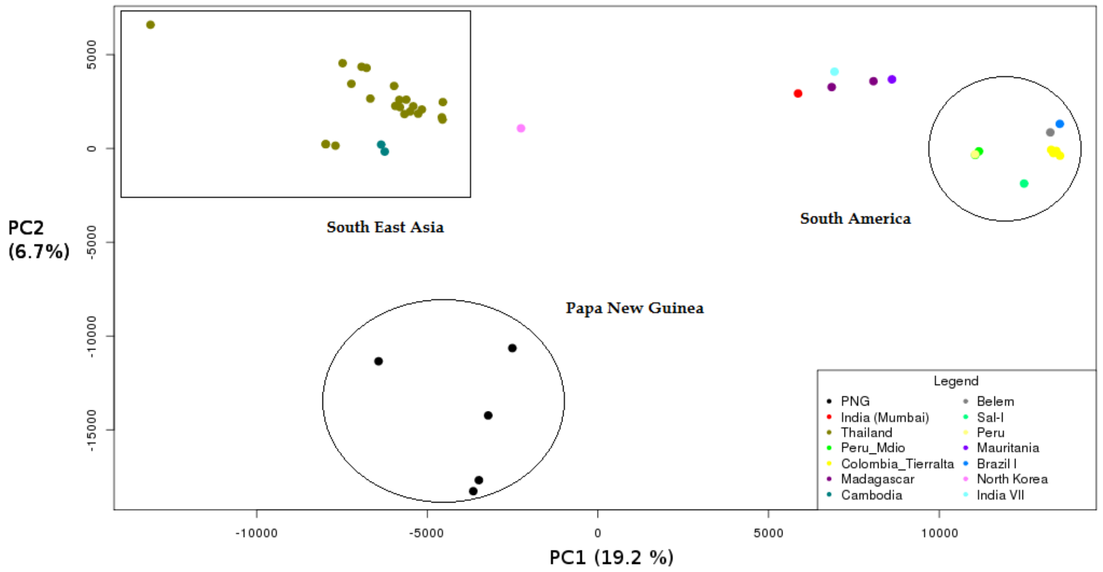
Supplementary Figure 1: Structural variants located around the MDR1 gene (chromosome 10) in the Thailand population; (i) a sample without a copy number variant or deletion (even coverage), (ii) a major deletion in the promoter region of the gene (n = 7); (iii) duplication of ~35kb (position 351kbp to 389kbp, n = 1); and (iv) a combination of both structural variants (ii) and (iii), including two copies, one with the deletion in the promoter and another copy with a complete promoter (n = 4, Thailand). The horizontal dashed line is average chromosomal coverage and the red outline encloses the promoter region of the MDR1 gene.



Supplementary Figure 2: Intra-population evidence of directional selective pressure (iHS*) a) Thailand b) South America. * iHS integrated haplotype score; see Table 1 and Table 2 for a summary of the hits.

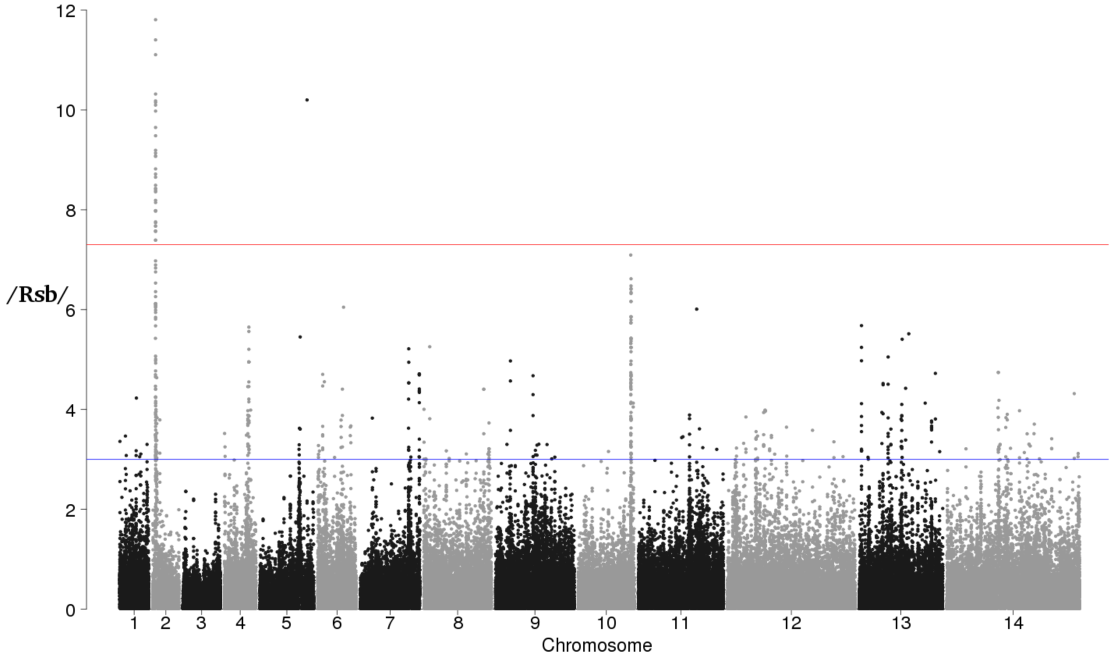


Supplementary figure 3: Principal component analysis based on 225k SNPs reveals strong clustering of isolates by continent.



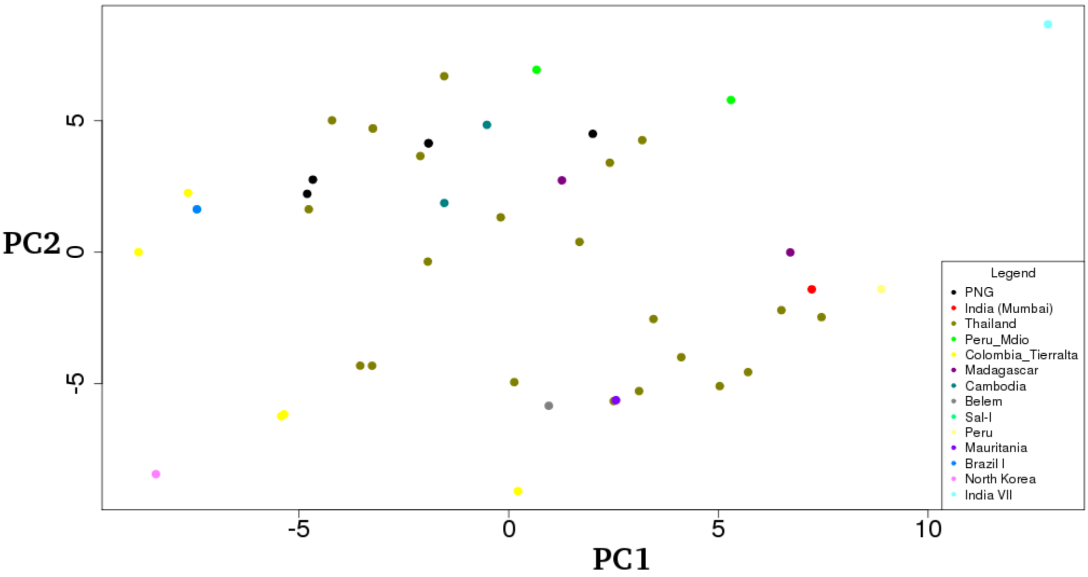
Supplementary Figure 4: Identifying regions under directional selective pressure between Thailand and South America.

Blue line: |Rsb| > 3 (P<0.003); Red line represents a human GWAS cut-off; see Table 4for a summary of the hits.

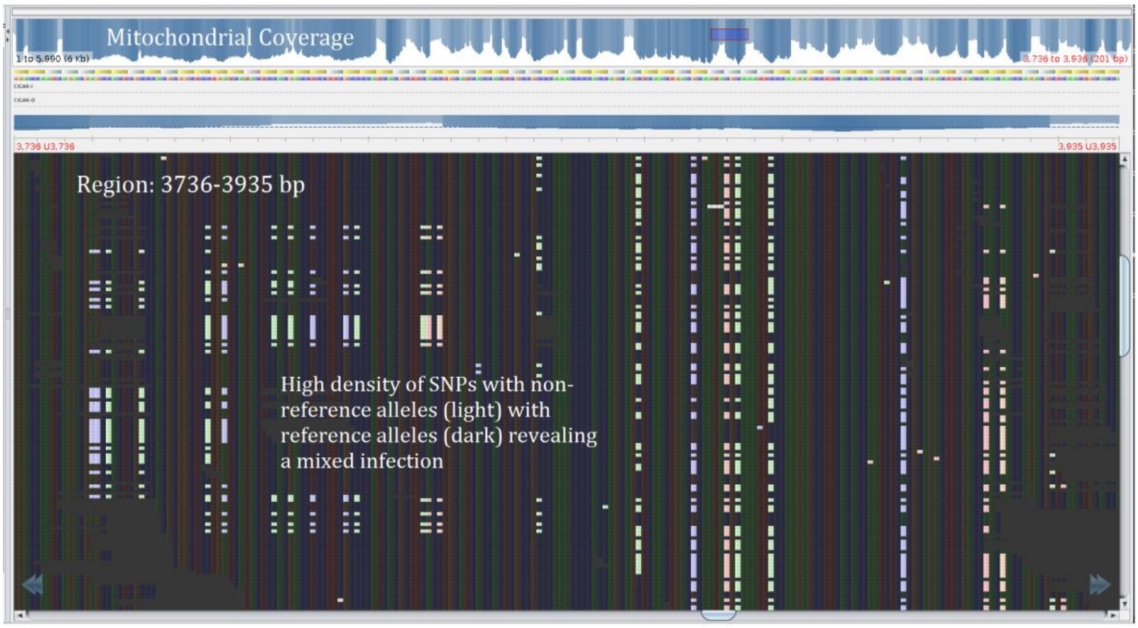


Supplementary Figure 5: Principal component analysis based on the previously characterised 42 barcoding SNPs* does not reveal strong population clustering.

* SNPs and genotypes are shown in S4 Table



Supplementary Figure 6: Signatures of a mixed species infection based on heterozygous calls in mitochondrial markers (positions: 3,736–3,935bp).



Region: 3736-3935 bp

High density of SNPs with non-reference alleles (light) with reference alleles (dark) revealing a mixed infection

	3750	3751	3753	3759	3771	3774	3783	3786	3790	3796	3798	3810	3811	3813	3831	3849	3859	3865	3867	3873	3897	3912	3915	3924	3926	3931
Pv	A	C	T	T	T	A	T	C	G	C	T	T	A	T	T	T	C	T	T	T	A	A	A	T	T	C
Pf	T	T	A	A	A	T	A	A	T	T	A	A	G	C	A	A	T	G	A	A	T	G	C	A	G	T
Pk	A	C	T	T	T	A	A	C	A	C	T	T	A	T	T	T	T	T	T	T	A	A	A	T	T	T

Supplementary Table 1: The 46 study isolates

Country	Sample	Mean coverage	SNPs
Cambodia*	SRR572648	359.4	29,791
Cambodia	SRR572649	1082.4	29,490
Thailand	ERR111710	59.8	29,326
Thailand*	ERR111711	121.2	30,494
Thailand*	ERR111712	93.0	31,933
Thailand*	ERR111713	86.7	29,977
Thailand	ERR111714	107.5	29,945
Thailand*	ERR111715	81.0	29,632
Thailand*	ERR111717	34.5	37,674
Thailand	ERR111718	94.2	30,324
Thailand	ERR111719	91.1	29,856
Thailand*	ERR111721	30.5	28,280
Thailand	ERR111722	50.2	29,370
Thailand*	ERR111723	42.1	31,034
Thailand	ERR111724	59.9	30,278
Thailand	ERR111725	62.8	29,966
Thailand	ERR111727	78.1	29,672
Thailand*	ERR111728	55.7	29,773
Thailand*	ERR111729	22.7	28,504
Thailand	ERR111730	86.0	30,105
Thailand	ERR111732	75.1	30,054
Thailand*	SRR1027921	63.0	30,600
Thailand	SRR1027922	104.9	30,180
Thailand*	SRR1027923	60.7	28,949
North Korea	SRS258286	2084.2	27,722
Brazil I	SRS258178	1173.8	23,100
Belem	SRR575087	629.0	22,491
Sal-I	SRR575089	31.5	108
Peru Mdio	SRR1798620	25.7	13,151
Peru Mdio	SRR1798621	32.8	14,580
Peru	SRS113792	33.0	14,675
Colombia Tierralta*	SRR2413274	14.9	13,367
Colombia Tierralta*	SRR2413275	24.2	19,812
Colombia Tierralta*	SRR2413304	37.0	21,339
Colombia Tierralta*	SRR2413306	48.6	20,973
Colombia Tierralta*	SRR2413357	15.2	15,559
Mauritania	SRS258091	1703.9	28,922
Madagascar*	SRR570031	456.6	28,077
Madagascar*	SRR828416	330.7	28,597
Papua New Guinea*	SRR828528	54.2	36,630
Papua New Guinea	ERR034096	136.8	30,097
Papua New Guinea	ERR034097	62.1	29,339

Papua New Guinea	ERR054084	14.6	18,760
Papua New Guinea*	ERR054085	54.9	30,004
India (Mumbai)	ERR054087	21.4	25,098
India VII	SRS258349	501.0	28,035

* evidence of polyclonality using the *estMOI* method [24]

Supplementary Table 2: The SNPs

Genome	Size (bp)	Total SNPs	Non-Coding	Coding	Non-Syn.	Syn.	Intronic
Nuclear	22,621,100	219,288	107,687 (49%)	111,601 (51%)	56,916 (51%)	35,812 (32%)	18,873 (17%)
Apicoplast	29,093	176	53 (30%)	123 (70%)	53 (43%)	70 (57%)	0 (0%)
Mito.	5,990	23	16 (70%)	7 (30%)	6 (86%)	1 (14%)	0 (0%)
Total	22,656,183	219,487	107,756 (49%)	111,731 (51%)	56,975 (51%)	35,883 (32%)	18,873 (17%)

Syn. Synonymous

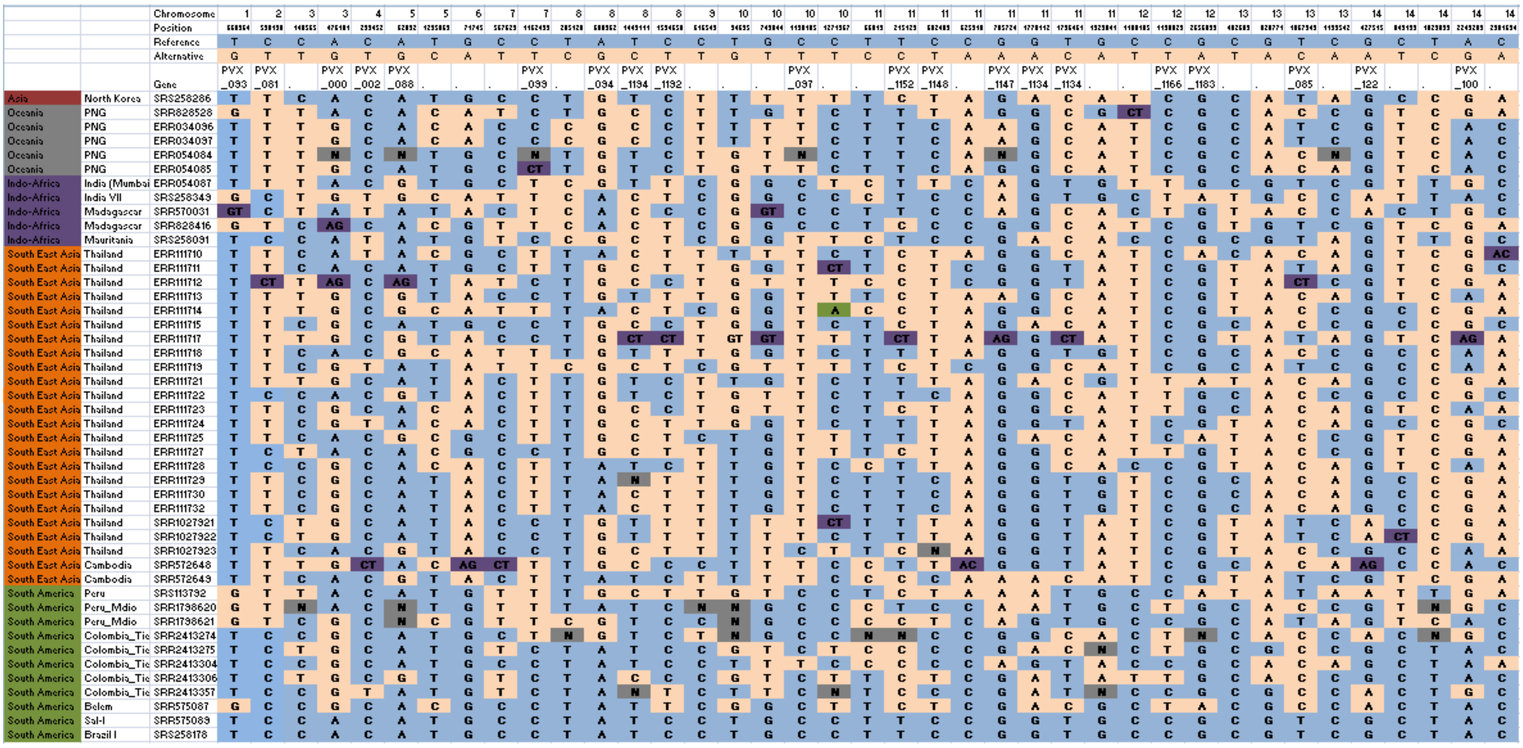
Supplementary Table 3: Non-synonymous mutations in candidate genes

Gene	Chr.	Position	Ref	Alt	Coding Change	Thailand (n=22)	S.America (n=11)	Others (n=13)
GTPCH	14	1825110	G	C	L38F	3	0	0
GTPCH	14	1825322	G	A	G109D	1	1	0
GTPCH	14	1825574	G	T	R193I	2	0	0
GTPCH	14	1825654	G	A	E220K	1	1	0
GTPCH	14	1826170	G	A	G392S	1	1	0
GTPCH	14	1826180	C	G	A395G	2	0	0
GTPCH	14	1826216	C	T	A407V	2	0	0
DHFR	5	964758	T	C,A	F57L,F57I	18	0	2
DHFR	5	964760	C	A	F57L	18	0	0
DHFR	5	964761	A	C	S58R	0	3	0
DHFR	5	964763	C	A,G	S58R,S58R	22	7	10
DHFR	5	964771	C	T	T61M	18	0	2
DHFR	5	964939	G	C,A	S117T,S117N	22	10	9
DHFR	5	965106	A	C	I173L	0	2	0
DHPS	14	1257156	A	G	M616T	0	0	2
DHPS	14	1257346	G	C	P553A	18	0	1
DHPS	14	1257856	G	C	P383A	22	8	6
DHPS	14	1257858	G	C	P382R	7	0	0
DHPS	14	1257859	G	C	P382A	0	5	0
MDR1	10	361917	C	G	K1393N	4	0	0
MDR1	10	362870	A	G	F1076L	14	2	12

MDR1	10	363169	T	A	Y976F	5	2	8
MDR1	10	363223	G	A	T958M	22	10	13
MDR1	10	363374	T	G	M908L	22	9	13
MDR1	10	363514	G	T	A861E	3	0	0
MDR1	10	364004	C	T	G698S	22	0	12
MDR1	10	364557	A	T	S513R	6	0	3
MDR1	10	364598	C	T	D500N	0	4	0
MDR1	10	365435	C	A	V221L	0	3	0
MRP1	2	154067	G	A	H1586Y	2	1	0
MRP1	2	154249	G	A	T1525I	0	2	0
MRP1	2	154391	C	T	V1478I	22	5	13
MRP1	2	154567	C	G	G1419A	1	5	3
MRP1	2	154646	A	C	Y1393D	22	8	13
MRP1	2	154979	A	T	L1282I	0	2	0
MRP1	2	155204	G	T	L1207I	21	0	1
MRP1	2	156107	G	C	Q906E	21	6	11
MRP1	2	158047	G	C	T259R	22	5	13
MRP1	2	158122	G	A	T234M	21	0	1
MRP1	2	158171	C	A	D218Y	0	0	2
MRP1	2	158444	T	C	I127V	0	0	2
MRP1	2	158717	T	G	K36Q	1	0	1
MRP1	2	158795	G	A	R10C	0	0	2
MRP2	14	2042200	G	C	H1960D	0	0	2
MRP2	14	2043096	C	G	G1661A	0	0	4
MRP2	14	2043285	C	T	S1598N	0	0	2
MRP2	14	2043838	G	A	H1414Y	20	0	11
MRP2	14	2043859	G	C	Q1407E	18	4	13
MRP2	14	2044225	A	G	Y1285H	2	0	0
MRP2	14	2044528	G	A	P1184S	0	3	0
MRP2	14	2044658	G	T	S1140R	4	0	0
MRP2	14	2044708	T	G	T1124P	1	0	1
MRP2	14	2044749	A	C	V1110G	0	2	4
MRP2	14	2044798	C	A	A1094S	0	5	0
MRP2	14	2045050	C	T	V1010M	22	10	13
MRP2	14	2045101	C	T	D993N	2	0	0
MRP2	14	2047069	A	G	W337R	4	0	1
MRP2	14	2047224	A	C	V285G	2	0	0
MRP2	14	2047233	C	A	R282M	22	10	11
MRP2	14	2047269	G	C	P270R	3	0	0
MRP2	14	2047816	C	G	E88Q	4	4	4
MRP2	14	2047893	C	T,A	C62Y,C62F	22	0	13
MRP2	14	2047961	C	T,G	K39N,Syn	6	0	7
P47	12	286327	C	A	F22L	22	1	12
P47	12	286331	T	C	F24L	22	0	12
P47	12	286340	A	G	K27E	22	1	12
P47	12	286431	G	C	S57T	3	0	4

P47	12	286446	G	A	S62N	7	0	1
P47	12	286949	G	A	V230I	15	0	4
P47	12	286960	G	T	M233I	21	0	10
P47	12	286970	T	A	F237I	2	0	1
P47	12	287046	T	C,A	I262T,I262K	13	1	11
P47	12	287078	A	G	I273V	8	0	2
P47	12	287080	A	G	I273M	0	0	2
P47	12	287379	C	T	A373V	16	1	6
P48/45	12	289467	G	A	E35K	1	0	2
P48/45	12	289995	C	A	H211N	22	3	13
P48/45	12	290114	A	C	K250N	22	4	13
P48/45	12	290367	G	T	D335Y	15	0	11
P48/45	12	290421	G	C	E353Q	0	3	0
P48/45	12	290490	G	A	A376T	14	0	10
P48/45	12	290617	A	G	K418R	22	1	12

Supplementary Table 4: Previously characterised 42 barcoding SNPs* in the 46 study isolates



		Chromosome	1	2	3	3	4	5	5	6	7	7	8	8	8	8	9	10	10	10	10	11	11	11	11	11	11	11	11	12	12	12	13	13	13	13	14	14	14	14	14
		Position	668964	538198	148565	476181	299452	62892	1295869	71745	567629	1162499	285128	688962	1449111	1594658	616549	94695	745844	1198185	1271967	66819	215129	682489	625918	785724	1778112	1796461	1929841	1188185	1198829	2656899	482689	828771	1867949	1139542	427515	849199	1829899	2249289	2981694
		Reference	T	C	C	A	C	A	T	G	C	C	T	A	T	C	C	T	G	C	C	T	T	C	C	G	G	T	G	C	C	G	C	G	T	C	G	C	T	A	C
		Alternative	G	T	T	G	T	G	C	A	T	T	C	G	C	T	T	G	T	T	T	C	C	T	A	A	A	C	A	T	T	A	T	A	C	A	A	T	C	G	A
		Gene	PVX_093	PVX_081	.	PVX_000	PVX_002	PVX_088	.	.	.	PVX_099	.	PVX_094	PVX_1194	PVX_1192	.	.	.	PVX_097	.	.	PVX_1152	PVX_1148	.	PVX_1147	PVX_1134	PVX_1134	.	.	PVX_1166	PVX_1183	.	.	PVX_085	.	PVX_122	.	.	PVX_100	.
Asia	North Korea	SRS258286	T	T	C	A	C	A	T	G	C	C	T	G	T	C	T	T	T	T	T	T	C	T	A	G	A	C	A	T	C	G	C	A	T	A	G	C	C	G	A
Oceania	PNG	SRR828528	G	T	T	A	C	A	C	A	T	C	T	G	C	C	T	T	G	T	C	T	T	T	A	G	G	C	G	CT	C	G	C	A	C	C	G	T	C	G	A
Oceania	PNG	ERR034096	T	T	T	G	C	A	C	A	C	C	C	G	C	C	T	T	T	T	C	T	T	C	A	A	G	C	A	T	C	G	C	A	T	C	G	T	C	A	C
Oceania	PNG	ERR034097	T	T	T	G	C	A	C	A	C	C	C	G	C	C	T	T	T	T	C	T	T	C	A	A	G	C	A	T	C	G	C	A	T	C	G	T	C	A	C
Oceania	PNG	ERR054084	T	T	T	N	C	N	T	G	C	N	T	G	T	C	T	G	T	N	C	T	T	C	A	N	G	C	A	T	C	G	C	A	C	N	G	T	C	A	C
Oceania	PNG	ERR054085	T	T	T	G	C	A	T	G	C	CT	T	G	T	C	T	G	T	T	C	T	T	C	A	G	G	C	A	T	C	G	C	A	C	A	G	T	C	A	C
Indo-Africa	India (Mumbai	ERR054087	T	T	T	A	C	G	T	G	C	T	C	G	T	T	C	G	G	C	T	C	T	T	C	A	G	T	G	T	T	G	C	G	T	C	G	T	T	G	C
Indo-Africa	India VII	SRS258349	G	C	T	G	T	G	C	A	T	T	C	A	C	T	C	G	G	C	C	C	T	C	C	A	G	T	G	C	T	A	T	G	C	C	A	T	T	A	C
Indo-Africa	Madagascar	SRR570031	GT	C	T	A	T	A	T	A	C	T	C	A	C	C	C	G	GT	C	C	T	T	C	C	A	G	C	A	C	T	G	T	A	C	C	A	C	T	G	C
Indo-Africa	Madagascar	SRR828416	G	T	C	AG	C	A	C	G	T	T	C	A	C	T	C	G	G	C	C	T	C	C	C	G	G	C	A	T	C	G	T	G	T	C	G	T	C	G	A
Indo-Africa	Mauritania	SRS258091	T	C	C	A	T	A	T	G	T	C	C	G	C	T	C	G	G	T	T	C	T	C	C	G	A	C	A	C	C	G	C	G	T	A	G	T	T	G	C
South East Asia	Thailand	ERR111710	T	T	C	A	T	A	C	G	C	T	T	A	C	T	T	T	T	T	C	T	C	T	A	G	G	C	A	T	C	A	C	A	C	A	G	T	C	G	AC
South East Asia	Thailand	ERR111711	T	T	C	A	C	A	T	G	C	T	T	G	C	T	T	G	G	T	CT	T	C	T	C	G	G	T	A	T	C	G	T	A	T	A	G	T	C	G	C
South East Asia	Thailand	ERR111712	T	CT	T	AG	C	AG	T	A	T	C	T	G	C	C	T	G	T	T	T	C	C	T	C	G	G	T	A	T	C	G	T	A	CT	C	G	T	C	G	A
South East Asia	Thailand	ERR111713	T	T	T	G	C	G	T	A	C	C	T	G	T	T	T	G	G	T	T	T	C	T	A	A	G	C	A	T	C	G	T	A	C	A	G	T	C	A	A
South East Asia	Thailand	ERR111714	T	T	T	G	C	G	C	A	T	T	T	A	C	T	C	G	G	T	A	C	C	T	A	G	G	C	A	T	C	G	T	A	C	C	G	C	C	G	A
South East Asia	Thailand	ERR111715	T	T	C	G	C	A	T	G	C	C	T	G	C	C	T	G	G	T	C	T	C	T	A	G	A	C	A	T	C	G	C	A	C	C	G	C	C	G	C
South East Asia	Thailand	ERR111717	T	T	T	G	C	G	T	A	C	C	T	G	CT	CT	T	GT	GT	T	T	T	CT	T	A	AG	G	CT	A	T	C	G	T	A	T	A	G	T	C	AG	A
South East Asia	Thailand	ERR111718	T	T	C	A	C	G	C	A	T	T	T	G	T	T	T	G	G	T	C	T	T	T	A	G	G	T	G	T	C	G	C	A	C	C	G	C	C	A	A
South East Asia	Thailand	ERR111719	T	T	C	G	T	A	T	A	T	T	C	G	C	T	C	G	T	T	T	T	C	T	C	G	G	C	A	T	C	G	C	A	T	C	G	C	C	A	A
South East Asia	Thailand	ERR111721	T	T	T	G	C	A	T	A	C	T	T	G	T	C	T	T	G	T	C	T	T	T	A	G	A	C	G	T	T	A	T	A	C	A	G	C	C	G	A
South East Asia	Thailand	ERR111722	T	C	C	A	C	G	T	A	C	T	T	G	T	C	T	G	T	T	C	T	T	C	A	G	G	C	A	T	T	G	C	A	C	C	G	C	C	G	C
South East Asia	Thailand	ERR111723	T	T	C	G	C	A	C	A	C	T	T	G	C	C	T	G	T	T	C	T	C	T	A	G	G	C	A	T	T	G	C	A	C	A	G	T	C	A	A
South East Asia	Thailand	ERR111724	T	T	C	G	T	A	C	A	C	T	T	G	C	T	T	G	G	T	C	T	T	T	A	G	G	T	A	T	C	G	T	A	C	A	G	C	C	G	C
South East Asia	Thailand	ERR111725	T	T	C	A	C	G	C	G	C	T	T	G	C	T	C	T	G	T	T	T	T	T	A	G	A	C	A	T	C	A	T	A	C	C	G	T	C	G	A
South East Asia	Thailand	ERR111727	T	C	T	A	C	A	C	G	C	C	T	G	C	T	T	T	G	T	T	T	C	T	A	G	G	C	A	T	T	G	T	A	C	C	G	T	C	G	A
South East Asia	Thailand	ERR111728	T	C	C	G	C	A	C	A	C	T	T	A	T	C	T	T	G	T	C	C	T	T	A	G	G	C	A	C	C	G	T	A	C	A	G	T	C	A	A
South East Asia	Thailand	ERR111729	T	T	C	G	C	A	T	A	C	T	T	A	N	T	T	T	G	T	C	T	T	C	A	G	G	T	G	T	C	G	C	A	C	A	G	C	C	G	A
South East Asia	Thailand	ERR111730	T	T	C	G	C	A	T	A	C	T	T	A	C	T	T	T	G	T	C	T	T	C	A	G	G	T	G	T	C	G	C	A	C	A	G	C	C	G	A
South East Asia	Thailand	ERR111732	T	T	C	G	C	A	T	A	C	T	T	A	C	T	T	T	G	T	C	T	T	C	A	G	G	T	G	T	C	G	C	A	C	A	G	C	C	G	A
South East Asia	Thailand	SRR1027921	T	C	T	G	C	A	T	A	C	C	T	G	T	T	T	T	T	T	CT	T	T	T	A	G	G	T	A	T	C	G	T	A	T	C	A	C	C	G	A
South East Asia	Thailand	SRR1027922	T	C	T	G	C	A	T	A	C	C	T	G	T	T	T	T	T	T	C	T	T	T	A	G	G	T	A	T	C	G	T	A	T	C	A	CT	C	G	A
South East Asia	Thailand	SRR1027923	T	T	C	A	C	G	T	A	C	C	T	G	C	T	T	T	T	C	T	T	C	N	A	G	G	T	A	T	C	G	T	A	C	C	G	C	C	A	A
South East Asia	Cambodia	SRR572648	T	T	T	G	CT	A	C	AG	CT	T	T	G	C	C	C	T	T	T	C	C	T	T	AC	G	G	T	A	T	C	G	C	A	C	A	AG	C	C	A	C
South East Asia	Cambodia	SRR572649	T	T	C	A	C	G	T	A	C	T	T	A	T	C	T	T	T	T	C	C	C	C	A	A	A	C	A	T	C	G	T	A	T	C	G	T	C	G	A
South America	Peru	SRS113792	G	T	T	A	C	A	T	G	T	T	T	G	C	T	T	G	T	C	C	T	C	T	A	A	A	T	G	C	C	A	T	A	T	A	A	T	T	G	A
South America	Peru_Mdio	SRR1798620	G	T	N	A	C	N	T	G	T	T	T	A	T	C	N	N	G	C	C	C	T	C	C	A	A	T	G	C	T	G	C	A	C	C	G	T	N	G	C
South America	Peru_Mdio	SRR1798621	G	T	C	G	C	N	C	G	T	T	C	G	T	C	C	N	G	C	C	C	C	T	C	A	G	T	G	C	C	G	C	A	T	A	G	T	C	A	C
South America	Colombia_Tic	SRR2413274	T	C	C	G	C	A	T	G	C	T	N	G	T	C	T	N	G	C	C	N	N	C	C	G	G	C	A	C	T	N	C	A	C	C	A	C	N	G	C
South America	Colombia_Tic	SRR2413275	T	C	T	G	C	A	T	G	T	C	T	A	T	C	C	G	T	C	T	C	C	C	C	G	A	C	N	C	T	G	C	G	C	C	G	C	T	A	C
South America	Colombia_Tic	SRR2413304	T	C	C	G	C	A	T	G	C	C	T	A	T	C	C	T	T	T	C	C	C	C	C	A	G	T	A	C	C	G	C	A	C	A	G	C	T	A	A
South America	Colombia_Tic	SRR2413306	T	C	T	G	C	G	T	G	T	C	T	A	C	C	C	G	T	C	T	T	C	T	C	G	A	T	A	T	T	G	C	A	C	C	G	C	T	A	C
South America	Colombia_Tic	SRR2413357	T	C	C	G	T	A	T	G	T	C	T	A	N	T	C	T	T	C	N	T	C	C	C	G	A	T	N	C	C	G	C	G	C	C	A	C	T	G	C
South America	Belem	SRR575087	G	C	C	G	C	A	C	G	C	C	T	A	T	T	C	G	G	C	T	T	C	T	C	G	A	C	G	C	T	A	C	G	C	C	A	C	T	A	C
South America	Sal-I	SRR575089	T	C	C	A	C	A	T	G	C	C	T	A	T	C	C	T	G	C	C	T	T	C	C	G	G	T	G	C	C	G	C	G	T	C	G	C	T	A	C
South America	Brazil I	SRS258178	T	C	C	A	C	A	T	G	C	C	T	A	T	C	C	T	G	C	C	T	T	C	C	G	G	T	G	C	C	G	C	G	T	C	G	C	T	A	C

* from reference [51]

Supplementary Table 5: Sites of population differentiation between Thailand and South America

Chr.	**Position**	**FST***	***Gene***	**Product**
1	8396	0.758	*PVX_087670*	hypothetical protein, conserved
1	137109	0.710	*PVX_087775*	E3 ubiquitin-protein ligase, putative
1	422094	0.721	*PVX_088090*	hypothetical protein, conserved
1	449291	0.710	*PVX_088130*	AP-3 complex subunit delta, putative
1	638670	0.668	*PVX_093565*	hypothetical protein, conserved
2	153642	0.853	***PVX_097025***	**multidrug resistance-associated protein 1, *MRP1***
2	174990	0.726	*PVX_081215*	hypothetical protein, conserved
2	299845	0.924	*PVX_081395*	serine/threonine protein kinase, putative
2	408852	0.716	*PVX_081525*	hypothetical protein, conserved
2	416092	0.710	***PVX_081540***	**ubiquitin carboxyl-terminal hydrolase 1, *UBP1***
3	64629	0.694	*PVX_001050*	SET domain protein, putative (*SET8*)
3	466211	0.849	*PVX_000565*	hypothetical protein, conserved
4	295743	0.716	*PVX_002815*	hypothetical protein, conserved
4	356929	0.924	*PVX_002905*	hypothetical protein, conserved
4	481817	0.788	*PVX_003935*	amine transporter, putative
4	618245	0.668	*PVX_003780*	hypothetical protein, conserved
4	715122	0.716	*PVX_003635*	hypothetical protein, conserved
4	812551	0.710	*PVX_003535*	Phist protein (*Pf-fam-b*)
5	388029	0.767	*PVX_089215*	serine/threonine protein kinase VPS15, putative (*VPS15*)
5	452949	0.664	*PVX_089310*	hypothetical protein, conserved
5	591010	0.709	*PVX_089480*	hypothetical protein, conserved
5	621120	0.726	*PVX_089510*	D13 protein, putative
5	675424	0.668	*PVX_089580*	chaperone protein ClpB1, putative (*ClpB1*)
5	972932	0.788	*PVX_089960*	hypothetical protein, conserved
6	65601	0.709	*PVX_001690*	Phist protein (*Pf-fam-b*)
6	171699	0.710	*PVX_001830*	hypothetical protein, conserved
6	246827	0.779	*PVX_001920*	hypothetical protein, conserved
6	268713	0.783	*PVX_001950*	hypothetical protein, conserved
6	420885	0.664	*PVX_111495*	hypothetical protein, conserved
6	521859	0.705	*PVX_111370*	hypothetical protein
6	575464	0.709	*PVX_111292*	conserved Plasmodium protein, unknown function
6	625419	0.779	*PVX_111230*	hypothetical protein, conserved
7	68940	0.713	*PVX_098625*	hypothetical protein, conserved
7	154973	0.709	*PVX_098710*	dynein heavy chain, putative
7	174254	0.674	*PVX_098712*	high molecular weight rhoptry protein 3, *RhopH3*
7	504372	0.788	*PVX_099100*	hypothetical protein, conserved
7	568345	0.709	*PVX_099180*	type II NADH:ubiquinone oxidoreductase, putative
7	726188	0.709	*PVX_099375*	hypothetical protein, conserved
7	1019659	0.726	*PVX_099750*	hypothetical protein, conserved
7	1038999	0.726	*PVX_099780*	replication termination factor, putative
7	1064692	0.730	*PVX_099820*	hypothetical protein, conserved
8	172625	0.924	*PVX_094350*	hypothetical protein, conserved
8	573474	0.705	*PVX_094855*	NLI interacting factor-like phosphatase, putative (*NIF4*)
8	730101	0.705	*PVX_095010*	ribonucleotide reductase small subunit, putative
8	787639	0.716	*PVX_095095*	hypothetical protein, conserved
8	829721	0.713	*PVX_095145*	hypothetical protein, conserved
8	1502402	0.710	*PVX_119390*	hypothetical protein, conserved
9	20038	0.664	*PVX_090840*	hypothetical protein
9	77676	0.694	*PVX_090895*	hypothetical protein, conserved
9	102780	0.716	*PVX_090925*	protein kinase domain containing protein

9	279221	0.713	*PVX_091136*	hypothetical protein, conserved
9	659985	0.710	*PVX_091630*	hypothetical protein, conserved
9	729619	0.709	*PVX_091700*	circumsporozoite-related antigen, putative (*EXP1*)
9	1154029	0.769	*PVX_092185*	hypothetical protein, conserved
9	1316937	0.713	*PVX_092370*	hypothetical protein, conserved
10	667466	0.664	*PVX_080425*	transporter, putative
10	671226	0.924	*PVX_080430*	phosducin-like protein, putative (*PhLP3*)
10	800750	0.850	*PVX_080615*	hypothetical protein, conserved
10	829185	0.708	*PVX_080660*	RNA pseudouridylate synthase, putative
10	973520	0.712	*PVX_098015*	soluble NSF attachment protein (*SNAP*), putative
10	975894	0.668	*PVX_098010*	hypothetical protein
10	1003754	0.779	*PVX_097980*	transcription factor IIb, putative
11	98032	0.924	*PVX_115400*	DNA repair endonuclease, putative (*ERCC4*)
11	565497	0.849	*PVX_114865*	T-complex protein 1 subunit delta, putative (*CCT4*)
11	920422	0.709	*PVX_114495*	acetyl-CoA synthetase, putative (*ACS*)
11	1505435	0.710	*PVX_113790*	hypothetical protein, conserved
11	1530963	0.726	*PVX_113750*	eukaryotic translation initiation factor protein
11	1686293	0.730	*PVX_113576*	sorting assembly machinery 50 kDa subunit, *SAM50*
11	1700734	0.674	*PVX_113560*	hypothetical protein, conserved
11	1912448	0.668	*PVX_113325*	mitochondrial chaperone *BCS1*, putative
11	1938368	0.775	*PVX_113285*	elongation factor G, putative
12	174109	0.668	*PVX_083360*	tyrosine kinase-like protein, putative (*TKL3*)
12	212484	0.779	*PVX_083315*	hypothetical protein, conserved
12	225457	0.712	*PVX_083310*	translation elongation factor, putative
12	339353	0.668	*PVX_083185*	isocitrate dehydrogenase [*NADP*], mitochondrial, *IDH*
12	537177	0.726	*PVX_082980*	GPI mannosyltransferase 3, putative (*GPI10*)
12	575061	0.705	*PVX_082937*	hypothetical protein, conserved
12	890022	0.776	*PVX_082470*	elongation factor Tu, mitochondrial precursor, putative
12	1173110	0.730	*PVX_116655*	hypothetical protein, conserved
12	1307254	0.788	*PVX_116805*	hypothetical protein, conserved
12	1403249	0.668	*PVX_116930*	conserved Plasmodium protein, unknown function
12	1528154	0.721	*PVX_117065*	TLD domain-containing protein
12	1588169	0.850	*PVX_117145*	transcription factor with AP2 domain(s), *ApiAP2*
12	1623826	0.790	*PVX_117175*	dynein beta chain, putative
12	1853264	0.664	*PVX_117405*	hypothetical protein, conserved
12	2167573	0.790	*PVX_117815*	hypothetical protein, conserved
12	2471252	0.701	*PVX_118175*	hypothetical protein, conserved
13	307322	0.783	*PVX_084445*	cysteine repeat modular protein 3, putative (*CRMP3*)
13	466345	0.674	*PVX_084600*	DnaJ domain containing protein
13	503529	0.664	*PVX_084640*	hypothetical protein
13	696401	0.775	*PVX_084840*	hypothetical protein, conserved
13	725430	0.715	*PVX_084860*	hypothetical protein, conserved
13	773964	0.783	*PVX_084925*	hypothetical protein, conserved
13	1012549	0.705	*PVX_085205*	ABC transporter G family member 2, putative (*ABCG2*)
13	1146760	0.783	*PVX_085390*	hypothetical protein, conserved
13	1209407	0.726	*PVX_085490*	glutathione reductase, putative
13	1673325	0.702	*PVX_085955*	cytidine diphosphate-diacylglycerol synthase, putative
13	1773064	0.709	*PVX_086035*	transcription factor with AP2 domain(s), *AP2-G2*
14	129079	0.712	*PVX_121945*	gametocyte associated protein, putative (*GAP*)
14	188124	0.668	*PVX_122015*	sodium/hydrogen exchanger 1, putative
14	918073	0.767	*PVX_122845*	hypothetical protein, conserved
14	1058168	0.924	*PVX_122995*	transporter, putative
14	1254772	0.668	*PVX_123225*	hypothetical protein, conserved

14	1256701	0.668	***PVX_123230***	**hydroxymethylpterin pyrophosphokinase-dihydropteroate synthetase, putative (DHPS)**
14	1263307	0.668	*PVX_123240*	DEAD/DEAH box helicase, putative
14	1296964	0.668	PVX_123283	JmjC domain containing protein (*JmjC1*)
14	1314634	0.836	*PVX_123300*	hypothetical protein, conserved
14	1416599	0.770	*PVX_123395*	*GPI* ethanolamine phosphate transferase 3, *PIGO*
14	1963668	0.713	*PVX_124005*	hypothetical protein, conserved
14	2432144	0.668	*PVX_100865*	cell cycle control protein, putative
14	2759960	0.779	*PVX_101265*	cyclin g-associated kinase, putative
14	2957235	0.721	*PVX_101500*	hypothetical protein
API	22771	0.726	***PVIV_000008700***	***sufB***

API apicoplast, all FST >0.68; bolded – known drug resistance loci

Supplementary Table 6: Population informative apicoplast variants

Chr.	**Pos**	**Ref. Allele**	**Alt. Allele**	***F_{ST}* Thailand vs. other**	***F_{ST}* SEA vs. other**	***F_{ST}* South America vs. other**	**Gene**
API	1,416	G	A	0.265	0.238	**1.000**	.
API	2,461	A	G	0.289	0.258	**1.000**	.
API	5,562	A	G	**0.729**	**0.620**	0.207	*RPS8*
API	8,308	C	T	**0.720**	**0.625**	0.186	*RPS7*
API	16,619	T	C	**0.729**	**0.628**	0.193	.
API	18,222	G	A	**0.729**	**0.634**	0.193	*rpoC*
API	20,024	T	C	**0.729**	**0.634**	0.193	*rpoB*
API	22,668	G	A	0.289	0.258	**1.000**	.
API	23,740	A	C	**0.729**	**0.634**	0.193	*sufB*
API	23,829	C	T	**0.729**	**0.634**	0.193	*sufB*
API	26,441	C	T	**0.729**	**0.634**	0.193	.

SEA Southeast Asia

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Ernest Diez Benavente
Principal Supervisor	Taane Clark & Susana Campino
Thesis Title	**Using whole genome sequence data to study genomic diversity and develop molecular barcodes to profile Plasmodium malaria parasites**

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	Not applicable		
When was the work published?	Not applicable		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Nature Communications
Please list the paper's authors in the intended authorship order:	Ernest Diez Benavente, Jody Phelan, Jamille G. Dombrowski, Claudio R. F. Marinho, Colin J. Sutherland, Cally Roper, Susana Campino, Taane G. Clark
Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I extracted the raw data from public repositories. I created the analysis pipeline and ran the samples through this pipeline, I performed the QC analysis, as well as the interpretation analysis. I later created custom R and perl scripts to process the information in order to obtain information related to coverage, SNPs and other genomic information. The figures presented in this work have all been generated using scripts written by myself. I wrote the first draft of the manuscript and circulated to co-authors. Once the comments were received I gathered them and made the relevant changes on the article manuscript.

Student Signature: ____________________________________ **Date:** __________

Supervisor Signature: ____________________________________ **Date:** __________

Chapter 7
SNP Barcoding of *Plasmodium vivax* for geographical origin prediction and transmission inference

SNP Barcoding of *Plasmodium vivax* for geographical origin prediction and transmission inference

Short title: SNP Barcoding of *Plasmodium vivax*

Ernest Diez Benavente [1], Jody Phelan [1], Jamille G. Dombrowski [2], Claudio R. F. Marinho [2], Colin J. Sutherland [1], Cally Roper [1], Susana Campino [1], Taane G. Clark[1,3,*]

[1] Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom

[2] Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil.

[3] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

* Corresponding author

Prof. Taane G Clark

Pathogen Molecular Biology Department,

Faculty of Infectious and Tropical Diseases

London School of Hygiene & Tropical Medicine, London, United Kingdom

Email: taane.clark@lshtm.ac.uk; Susana.campino@lshtm.ac.uk

ABSTRACT

Although *Plasmodium vivax* parasites are the predominant cause of malaria outside of sub-Saharan Africa, they are commonly neglected by elimination programmes. *P. vivax* is resilient and poses challenges through its ability to re-emerge, especially from dormancy in the human liver. With reports of growing drug-resistance and the increasing prevalence of life-threatening infections, there is an urgent need for tools that can inform elimination efforts. In order to halt transmission through effective intervention, malaria programmes need to identify reservoirs of infection and understand the dynamics of transmission and movement of parasite populations. The use of molecular epidemiology for tracking and studying parasite populations has been applied successfully to other malaria species such as *P. falciparum.* By assembling the largest set of *P. vivax* whole genome sequences (n=446) spanning 16 countries, and applying a machine learning approach, we created a 71 SNP barcode with high predictive ability to identify geographic origin (93.4%) and infer transmission. Further, by using *P. vivax* data from a low-transmission setting in Malaysia, as well as other validation datasets, we demonstrate the ability to outperform published genotyping methods and infer outbreak events. When rolled-out in new portable platforms, this SNP barcode and its future data-informed machine-learning based improvements, will be an invaluable tool to help elimination efforts of this resilient neglected *Plasmodium* species.

Word count: 213

Keywords:

Plasmodium vivax, SNP barcode, geographic origin, transmission

INTRODUCTION

Plasmodium vivax is the predominant cause of malaria outside of sub-Saharan Africa [1,2] with increasing reports of drug-resistance and severe complications that pose a threat to children and pregnant women [3–6]. Elimination efforts have led to reductions in the prevalence of the deadlier *P. falciparum* malaria, but areas of co-endemicity have seen a corresponding rise in *P. vivax* infections, which appear more resilient to control strategies[7]. In order to halt transmission, malaria programmes should focus on identifying the main reservoirs of infection and adapt control measures to efficiently tackle these. *P. vivax* has been observed in regions where malaria transmission had once been interrupted [8], therefore long-term monitoring is essential. Further, understanding the transmission dynamics of the parasite populations through assessment of genetic diversity has the potential to play a key role in guiding the elimination efforts, including by revealing transmission events and identifying potential foci of infection. In recent years, genomic studies have dissected the molecular dynamics of *P. vivax* populations in regions with stable transmission [9–13] . Other studies have used microsatellites to study trends in the population dynamics [14–17]. The availability of whole genome sequencing data can inform the design of "genetic barcodes" for inferring transmission and geographic source of infections, and when combined with affordable delivery systems, can facilitate *P. vivax* epidemiological and surveillance investigations.

By leveraging off whole genome sequencing for population characterisation [11,12,18,19], a number of SNP-based barcodes have been derived for *P. falciparum*. A barcode based on 24 SNPs in the nuclear genome ("24-SNP barcode") has been used to identify and track isolates from an endemic population in Senegal [20]. Although it was designed using only a limited set of long-term adapted laboratory lines [20], the barcode has demonstrated local changes in genetic diversity across a 7-year period [21], as well as the potential effectiveness of such tools in combination with epidemiological methods to elucidate transmission intensity in a malaria endemic region [21]. Nevertheless, the use of isolates from a very limited region to generate the barcode, can potentially underestimate the genomic variability present in other *Plasmodium* populations in the global setting, impeding the transportability of the 24-SNP barcode to other geographical settings. Further, it has been shown that the predictive power of the 24-SNP barcode for geographical determination is poor, especially when compared to one formed of 23 SNPs from the mitochondria and apicoplast organellar genomes, which predicted the continental origin of samples with 92% accuracy [22]. Another barcode formed of 105 highly frequent nuclear genome SNPs was developed to infer transmission intensity using a much broader panel of data, thereby proving greater utility across malaria-endemic countries [23]. Overall, these studies in *P. falciparum* have demonstrated that SNP barcodes can potentially provide insights into the intensity of transmission, identify the geographical origin of the field isolates, and inform the dynamics of the diversity in a parasite population including outbreak identification, an event which has been shown to be more likely in low-transmission settings [24].

For *P. vivax* there have been two attempts to generate SNP barcodes [25,26], including a "42-SNP barcode", for ascertaining the source of infection. However, these barcodes were based on small datasets with limited geographical coverage, and therefore have poor predictive power when used for more global applications [11–13,27]. With technological advancement in genomics, including the high throughout sequencing of candidate genomic regions and portable genotyping [28], the identification and integration of informative loci for *P. vivax* and other plasmodium parasites for inferring transmission and infection source, has the potential to revolutionise global malaria surveillance. Here, using whole genome sequencing data for 446 *P. vivax* isolates across 16 countries, we applied machine learning and SNP tagging approaches from human genome-wide association studies (GWAS) to create a 71 SNP barcode with the high predictive power for geographic origin (93.4% accuracy) and the ability to infer transmission. We demonstrate that the barcode outperforms alternative approaches, including microsatellite genotyping, for global geographical profiling and inferring outbreak events within a low-transmission setting in Malaysia. The "71-SNP" barcode has the potential to be an invaluable tool to help elimination efforts of this resilient neglected *Plasmodium* species.

RESULTS

SNPs, samples and population structure

We aligned raw sequence data from 867 isolates [9,11–13,29] to the PvP01_v1 (http://genedb.org) reference genome, and identified 1,522,046 SNPs. Isolates with high levels of missing genotype calls (>30%) and high multiplicity of Infection (>20% heterozygous genotypes) were excluded from analysis. The final dataset was formed of 446 isolates, spanning 16 countries and 6 regions (Africa 3; South Asia 4; Southeast Asia (SEA) 264; Papua New Guinea 26; South America 129; North America 20), and 741,074 high quality SNPs of which 86% were non-common (minor allele frequency (MAF) < 5%) (**Supplementary figure 1, left**), consistent with previous findings [27].

A principal component analysis (PCA) revealed that the isolates clustered broadly by regional groups: SEA (Thailand, Myanmar, Cambodia, Vietnam and Laos), South Asian/SEA (China), Americas (Peru, Colombia, Mexico and Brazil) and Oceanian/SEA (Papua New Guinea, Indonesia and Malaysia) and others located around the Arabian Sea (**Figure 1**). At a country level, the intra-border genomic distance between isolates was on average smaller (mean: 26,505 SNPs, range: 0 - 37,801 SNPS) compared to the inter-border distance (mean: 33,160 SNPs, range: 6,034 -39,676 SNPs), in line with previous studies [11,12]. The notable exception to the pattern was isolates sourced from neighbouring Thailand and Myanmar, supporting evidence they belong to a similar parasite population [11]. These patterns suggest the possibility of identifying markers that could potentially identify geographic origin to the country level in most cases. Intuitively, the highly skewed distribution of the MAF towards rare variants suggests that more common SNPs are those driving the population structure observed. To assert this observation, we split the dataset in three equally sized divisions based on MAF (tertiles: <0.2%, 0.2 – 1.3%, >1.3%; **Supplementary figure 1, right**), and for each constructed neighbour-joining trees and correlated the pair-wise genome distance with the estimate based on all 741k SNPs (**Figure 2**). These comparisons revealed the strong explanatory power harboured by those SNPs with the highest MAF (correlation 0.93 with the full 741,074 SNP set) compared to the subsets with lower MAF (correlations of 0.14 and 0.18). A MAF cut-off (> 0.3) was determined **(Supplementary figure 1, right)**, leading to a subset of 16,075 SNPs used as the starting point for barcode building for country classification.

Selection of highly informative barcoding SNPs using a tagging and machine learning approach

In order to further reduce number of markers used for barcoding construction, we applied the software *TAGster*[30] to identify tagging SNPs that summarise blocks of high linkage disequilibrium (LD), as estimated by the correlation metric r^2 across windows of size 500 kbp. The genetic variability can be captured by tagging SNPs due to the strong LD structure found in *P. vivax* populations [27], and we identified 1,173 SNPs, which summarised variation in 4,896 neighbouring SNPs, almost 40% of the total dataset. The 1,173 highly informative SNPs were then used as the pool for a further selection using a random forest modelling approach. Prior to implementation, missing alleles (4.4%) were imputed with high accuracy (<1% error, see **Online Methods**), and a neighbour joining tree was constructed using the resulting data, thereby confirming that no bias was introduced to the clustering patterns **(Supplementary figure 2, top).** We sought to show that the 1,173 SNPs predict not only large geographic differences within isolates, but also similarities within intra-border isolates. In particular, we studied the correlation of genome distance based on 1,173 SNPs against those representing intra regions (genomic distance < 20,000 SNPs) **(Supplementary figure 2, bottom)**. This analysis revealed that the 1,173 high frequency tagging SNPs can not only detect strong inter-border differentiation but also closely related samples within the same country based on genomic distance (r^2 = 0.96).

Application of the random forest classification approach to the classification of country involved constructing 500 trees, partitioning the dataset randomly into 80% training (n=357) and 20% for validation (n=89), and using 34 variables at each nodal split. Classification error rates became stable when 100 trees were averaged **(Supplementary figure 3, left)**. The final model performed with an overall out-of-bag error rate of 17.6%, where the main classification errors were found across

Southeast Asian populations (Thailand, Vietnam, Cambodia, and Myanmar), which were identified as highly related populations using the genomic distance (**Figure 2, right**), as well as by others [11]. The random forest model was then used to identify the 60 SNPs with the highest predictive importance across the trees **(Supplementary figure 3, right)**. A further 11 SNPs were chosen to summarise strong inter-countries based on the fixation index (F_{ST}> 0.7), leading to a final barcoding set of 71 SNPs (**Table 1**). Within this SNP set, there are differences in allele frequency across the two main regions (Southeast Asia and South America), but no markers were fixed across the populations. The 71 markers are in low LD **(Supplementary figure 4)** (r^2 mean: 0.151, Inter-quartile range: 0.019 - 0.235), but some blocks of correlation are observed, which can be explained by an uneven geographic distribution of the isolates.

In order to assess the potential of the barcode for geographic classification, a PCA was generated using only the 71 SNPs **(Supplementary figure 5, top),** and similar clustering patterns were observed to those using the genome-wide (741k including a "42-SNP barcode") SNPs (**Figure 1**). When comparing the genomic distance obtained using the 71 barcoding to genome-wide SNPs (n=741,074) in intra-borders pair-wise comparisons, we observed 88.3% correlation proving the potential of this barcode to not only identify geographical origin but also provide insights into the relatedness of intra-border isolates **(Supplementary Figure 5, bottom).** The potential to infer intra-border relatedness and therefore, provide insights into transmission dynamics, was also supported by the fact that 87.9% (392/446) of the haplotypes obtained were unique in the dataset. Furthermore, the 71-SNPs were used to blindly predict the geography of the 20% of the isolates (n=89) not used to develop the random forest, and were complemented by a further 16 Brazilian in-house sequenced isolates. The model using only the 71-SNPs yielded an overall accuracy of 93.4% (6.6% out-of-bag error) and predicted correctly the origin of all the provided isolates. It outperformed a published 42-SNP barcode [26], which under the same random forest model conditions (80%/20% training/prediction and 500 trees) obtained a 79.3% accuracy. The low accuracy of the 42-SNP barcode can also be observed in the ambiguous clustering found in the PCA (**Supplementary Figure 6, top**) and neighbour-joining tree (**Supplementary Figure 6, bottom**). Furthermore, this SNP barcode showed a low correlation with the genome-wide (741k) SNPs distance (r^2= 0.60).

In-silico **testing of the barcode in a near-elimination setting in Malaysia.**

A field-ready SNP barcode with the potential for being deployed in low-transmission settings has to be proven efficacious in settings where the identification of foci of infection and imported cases are of key relevance. We used an independent dataset comprising 60 *P. vivax* isolates, sourced from Sabah Malaysia, where the genomic population dynamics have been extensively studied using microsatellite markers and whole genome sequencing methodologies [24]. *In silico* application of the 71-SNP barcode identified two main populations (referred to as K1 and K2), where one (K2) comprised of 26 almost genetically identical isolates in a known transmission outbreak, previously supported using a set of 9 microsatellites [24]. Using the 71 SNPs the estimated the genetic distance across isolates was strongly correlated with genome-wide SNP-based estimates (r^2 = 0.83). A PCA plot based on 71-SNPs revealed the two distinct populations **(Figure 3A;** K1 grey; K2 yellow**),** and the outbreak isolates from K2 tightly clustered and shared the same available haplotype across the 72 SNPs. There was one notable exception (ERR1475456) which presented a larger genetic distance (**Figure 3A**, green label). This isolate shared the same microsatellite haplotype as the outbreak K2, but the distribution of genomic distance when compared to the rest of the outbreak samples was greater, thus being less related to the outbreak isolates and potentially being an error in the genotyping (**Figure 3B**). Furthermore, the generation of a neighbour-joining tree using the 71 SNPs revealed clustering by administrative division in Sabah Malaysia, which could be used for tracking of cases from different health authorities (**Figure 3C**). Altogether, these results show the potential of the 71-SNP barcode (or future versions with data-driven machine learning updates) to be used in field settings, not only to predict geographical origin but also to infer transmission events. The Malaysian analysis highlights its utility in low transmission and near-elimination settings.

DISCUSSION

Plasmodium vivax accounts for a high global malaria burden, with marked incidence outside of sub-Saharan Africa [2], and growing levels of drug resistance and severe human infections [31–33]. The resilience of the parasite, especially its re-appearance in regions where malaria transmission had previously been halted [8], poses strong challenges to malaria elimination. Microsatellite genotyping has been used to study *P. vivax* genomic and population dynamics, but it underestimates variability in natural populations[34], as shown in our analysis. Whole genome sequencing is the gold standard approach, but is currently logistic- and cost-inefficient for large genomic epidemiological studies, especially in under-resourced endemic areas. To bridge the gap, we have a developed a 71 SNP barcode, informed by whole genome sequencing data from 446 isolates across 16 different countries worldwide. To date, barcodes in *Plasmodium* species – particularly *P. falciparum* - have focused on determination of the geographical origin of the samples using nuclear or organellar genomic markers [22,25], or to infer transmission intensity using the frequency of unique haplotypes [21,26] or by modelling the complexity of the infections [23], but not together as adopted here. Further, *P. vivax* barcodes have been based on small datasets and they have been demonstrated to target specific geographic populations [27], rather than our global approach that is more robust to the potentially high mobility of individuals through access to air travel. In fact, application of a 42-SNP barcode[26] to our dataset led to a lower accuracy (79.3%) when predicting origin of the isolates at a country level. The barcode also performed sub-optimally when estimating SNP-based relatedness of isolates, and therefore may not be suited to the inference of intra-border transmission.

Our barcode was constructed by recognising that common SNPs are more robust markers and harbour greater explanatory power for both geographical and transmission inference. By triaging SNPs using established linkage disequilibrium tagging methods, it was possible to apply random forest methodology to select 60 markers, augmented by 11 inter-continental SNPs, to accurately predict country (93.4%), with an 6.6% out-of-bag error. Whilst there are alternatives to the random forest algorithm, it has become an established data analysis tool in bioinformatics, with an excellent performance in settings where the number of variables is much larger than the number of observations. Further, the methodology has an ability to explore complex interaction structures and highly correlated variables, and return useful measures of variable importance [35,36]. The set of "important" SNP predictors of country determined by the random forest approach was robust to initial model parameterisation, and the final model and set of SNPs were validated perfectly on a 20% subset of the original samples augmented by prospectively collected ones from Brazil. The 71-SNP barcode and intermediate SNP sets informing its construction were used to reconstruct principal components plots. These plots were consistent with the overall population structure based on 741k markers, and therefore confirm there was no major loss of geographical specificity.

Overall, our 71-SNP barcode is the first that shows such strong levels of accuracy in geographic prediction in *P. vivax* for a highly diverse dataset, making it a highly valuable tool for the detection of imported cases of malaria. Further, as whole genome data becomes available, especially from sites with currently poor coverage such as central America, Africa and south Asia, the machine learning approach can rapidly update the SNPs in the barcode. The barcode was also designed to be able to provide information about potential transmission trends and therefore be useful in field settings, where large genomic differences are less likely, with the exception of imported cases. The proportion of unique haplotypes identified across the dataset was high (88%; 392/446 haplotypes observed), which allows greater scope for informing on intra-border haplotype diversity, including low diversity such as in an outbreak setting. This utility was demonstrated by the use of intra-border highly related isolates, where we confirmed that the SNP distances based on the 71-SNP barcode and genome-wide 741k markers were highly correlated, and represent an improvement on comparative reported values for microsatellites (r^2 = 0.70)[34]. Further, the use of the 71 SNP barcode was validated *in silico* using data from 60 *P. vivax* sourced from a low endemic and near-elimination setting in Malaysia, where it was possible to partition the population into highly structured subclades and confirm the presence of an outbreak cluster, previously identified using both microsatellite genotyping and sequencing. Also, by using the 71 SNP barcode we identified a misclassified "outbreak" isolate with an identical microsatellite haplotype, and evidence of regional clustering by the different administrative divisions, further supporting the utility of the tool and superior performance to microsatellite genotyping.

In summary, we have presented a new molecular barcode for *P. vivax* that can provide information on both geographical origin and on closely-related isolates within country borders to help infer transmission events and identify foci of infection. The 71-SNP barcode out-competes previous genotyping methods and is a powerful and affordable solution that could potentially enable the execution of large genomic epidemiological studies, with high throughput assessment of large numbers of parasites. By leveraging off growing and large datasets of whole genome sequencing data, and the power of machine learning algorithms, it will be possible to update the barcode, augment it with drug resistance markers, and implement it rapidly in field-based settings using portable technologies. Ultimately, insights into genetic diversity will assist the much-needed understanding of the dynamics of *P. vivax* populations and inform clinical and disease control decision making.

MATERIALS AND METHODS

Genomic data generation

Illumina sequenced data from previous studies [11–13] was downloaded from ENA repository to form a total dataset of 834 isolates. Each isolate data was mapped against the PvP01_v1 reference (obtained from http://genedb.org) using *bwa-mem* [37]. SNPs (n=1,522,046) were called from the resulting alignments using *samtools* software suite [38], as previously described [27]. Isolates and SNPs were excluded if presenting with >20% missing or heterozygous genotype calls, and further SNPs were removed if from hypervariable gene regions (e.g. *vir* genes). The final dataset consisted of 741,074 high-quality SNPs and 446 isolates.

Population Structure and tag SNP selection

The 741,074 high-quality SNPs and its subsets were used calculate matrices of Manhattan distance between samples. These matrices were used to generate principal component analysis plots and neighbour-Joining trees (R *ape* package [39]). The Pearson's r^2 metric, calculated using the R base function *cor,* was used to estimate the correlation between distance matrices. The software *TAGster*[30] was used to identify SNPs which summarise blocks of high linkage disequilibrium, as estimated using the genetic r^2 metric. We specified an r^2 threshold of at least 0.7 for inclusion in a block and a window size of 500 kbp, leading to 16,075 SNPs with minor allele frequency > 0.3 being included for analysis. The resulting 1,173 tag SNPs identified were then further characterized for downstream analysis. Fixation index (F_{ST}) was calculated using in house scripts in R.

Barcode SNP selection using Random Forests (RF)

The selected 1,173 highly informative SNPs obtained using a combination of minor allele frequency and tag SNP approaches were extracted from the original dataset and imputation was performed on the missing data points (4.4%) in the dataset using the R package *missForest*[40]. This yielded an estimated out-of-bag error in the imputation of 16.2% which left a total of 0.7% potentially erroneous calls. After imputation, a random selection of 80% of the dataset was assigned as training set and the remaining 20% was assigned as test dataset and country was tested as the predicted variable. Then, 500 trees were calculated using the *RandomForest*[41] package in R in order to determine the SNPs in the dataset with highest importance for classification of samples into countries. We then selected the

60 SNPs with highest importance and used the R package *LDcorSV*[42] to calculate the correlation between the markers. Further PCA plots and neighbour-joining trees were calculated for this subset of SNPs.

REFERENCES

1. Howes, R. E. *et al.* Global Epidemiology of Plasmodium vivax. *Am. J. Trop. Med. Hyg.* **95,** 15–34 (2016).
2. WHO. *World Malaria Report 2017*. (2017).
3. Tjitra, E. *et al.* Multidrug-Resistant Plasmodium vivax Associated with Severe and Fatal Malaria: A Prospective Study in Papua, Indonesia. *PLOS Med.* **5,** e128 (2008).
4. Poespoprodjo, J. R. *et al.* Adverse Pregnancy Outcomes in an Area Where Multidrug-Resistant Plasmodium vivax and Plasmodium falciparum Infections Are Endemic. *Clin. Infect. Dis.* **46,** 1374–1381 (2008).
5. Poespoprodjo, J. R. *et al.* Vivax Malaria: A Major Cause of Morbidity in Early Infancy. *Clin. Infect. Dis.* **48,** 1704–1712 (2009).
6. Price, R. N. *et al.* Global extent of chloroquine-resistant Plasmodium vivax: a systematic review and meta-analysis. *Lancet. Infect. Dis.* **14,** 982–991 (2014).
7. Cotter, C. *et al.* The changing epidemiology of malaria elimination: new strategies for new challenges. *Lancet (London, England)* **382,** 900–911 (2013).
8. Sattabongkot, J., Tsuboi, T., Zollner, G. E., Sirichaisinthop, J. & Cui, L. Plasmodium vivax transmission: chances for control? *Trends Parasitol.* **20,** 192–198 (2004).
9. Winter, D. J. *et al.* Whole Genome Sequencing of Field Isolates Reveals Extensive Genetic Diversity in Plasmodium vivax from Colombia. *PLoS Negl. Trop. Dis.* **9,** e0004252 (2016).
10. Parobek, C. M. *et al.* Selective sweep suggests transcriptional regulation may underlie Plasmodium vivax resilience to malaria control measures in Cambodia. *Proc. Natl. Acad. Sci.*

U. S. A. **113,** E8096–E8105 (2016).

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

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

13. de Oliveira, T. C. *et al.* Genome-wide diversity and differentiation in New World populations of the human malaria parasite Plasmodium vivax. *PLoS Negl. Trop. Dis.* **11,** e0005824 (2017).

14. Abdullah, N. R. *et al.* Plasmodium vivax population structure and transmission dynamics in Sabah Malaysia. *PLoS One* **8,** e82553 (2013).

15. Getachew, S. *et al.* Variation in Complexity of Infection and Transmission Stability between Neighbouring Populations of Plasmodium vivax in Southern Ethiopia. *PLoS One* **10,** e0140780 (2015).

16. Trimarsanto, H. *et al.* VivaxGEN: An open access platform for comparative analysis of short tandem repeat genotyping data in Plasmodium vivax Populations. *PLoS Negl. Trop. Dis.* **11,** (2017).

17. Pava, Z. *et al.* Genetic micro-epidemiology of malaria in Papua Indonesia: Extensive P. vivax diversity and a distinct subpopulation of asymptomatic P. falciparum infections. *PLoS One* **12,** e0177445 (2017).

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

19. Diez Benavente, E. *et al.* Analysis of nuclear and organellar genomes of Plasmodium knowlesi in humans reveals ancient population structure and recent recombination among host-specific subpopulations. *PLOS Genet.* **13,** e1007008 (2017).

20. Daniels, R. *et al.* A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking. *Malar. J.* **7,** 223 (2008).

21. Daniels, R. F. *et al.* Modeling malaria genomics reveals transmission decline and rebound in

Senegal. *Proc. Natl. Acad. Sci.* **112,** 7067–7072 (2015).

22. Preston, M. D. *et al.* A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. *Nat. Commun.* **5,** 4052 (2014).

23. Chang, H.-H. *et al.* THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. *PLOS Comput. Biol.* **13,** e1005348 (2017).

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

25. Rodrigues, P. T. *et al.* Using mitochondrial genome sequences to track the origin of imported plasmodium vivax infections diagnosed in the United States. *Am. J. Trop. Med. Hyg.* **90,** 1102–1108 (2014).

26. Baniecki, M. L. *et al.* Development of a Single Nucleotide Polymorphism Barcode to Genotype Plasmodium vivax Infections. *PLoS Negl. Trop. Dis.* **9,** e0003539 (2015).

27. Diez Benavente, E. *et al.* Genomic variation in Plasmodium vivax malaria reveals regions under selective pressure. *PLoS One* **12,** e0177134 (2017).

28. Nag, S. *et al.* High throughput resistance profiling of Plasmodium falciparum infections based on custom dual indexing and Illumina next generation sequencing-technology. *Sci. Rep.* **7,** (2017).

29. Benavente, E. D. *et al.* Genomic variation in Plasmodium vivax malaria reveals regions under selective pressure. *PLoS One* **12,** (2017).

30. Xu, Z., Kaplan, N. L. & Taylor, J. A. TAGster: efficient selection of LD tag SNPs in single or multiple populations. *Bioinformatics* **23,** 3254–3255 (2007).

31. Poespoprodjo, J. R. *et al.* Adverse Pregnancy Outcomes in an Area Where Multidrug-Resistant Plasmodium vivax and Plasmodium falciparum Infections Are Endemic. *Clin. Infect. Dis.* **46,** 1374–1381 (2008).

32. Poespoprodjo, J. R. *et al.* Vivax Malaria: A Major Cause of Morbidity in Early Infancy. *Clin. Infect. Dis.* **48,** 1704–1712 (2009).

33. Price, R. N. *et al.* Global extent of chloroquine-resistant Plasmodium vivax: a systematic review and meta-analysis. *Lancet. Infect. Dis.* **14,** 982–991 (2014).

34. Vali, U., Einarsson, A., Waits, L. & Ellegren, H. To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Mol. Ecol.* **17,** 3808–3817 (2008).

35. Díaz-Uriarte, R. & de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7,** 3 (2006).

36. Chen, X. & Ishwaran, H. Random Forests for Genomic Data Analysis. *Genomics* **99,** 323–329 (2012).

37. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

38. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

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

40. Stekhoven, D. J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28,** 112–118 (2012).

41. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).

42. Mangin, B. *et al.* Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb).* **108,** 285–291 (2012).

FIGURES

Figure 1. Principal component (PC) analysis plot generated using 741,074 high quality SNPs across 446 *P. vivax* isolates reveals geographic clustering. Isolates were coloured according to country of origin. Clustering by region can be found, with Southeast Asian isolates appearing cluster at the bottom right of the plot, Oceania at the top right, and South American isolates clustering on the centre left side of the plot. A relative degree of clustering by country can be observed, especially for isolates from Oceania and to a lesser extent Southeast Asia.

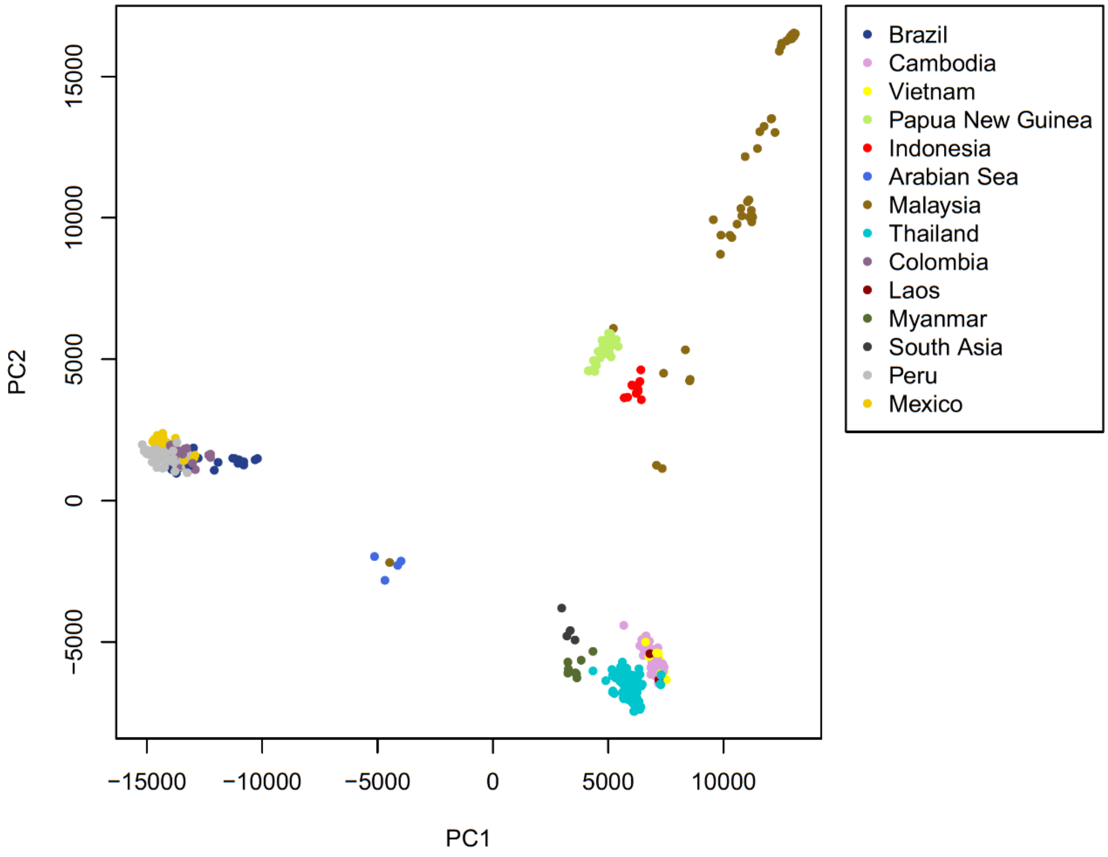


Figure 2. Subsetting of SNPs by minimum allele frequency (MAF) reveals the strong explanatory power of high frequency SNPs in *P. vivax*. Three equally sized divisions (i.e. tertiles) of the SNP dataset were used based on MAF, left [0 - 0.2 %], centre [0.4 - 1.3 %] and right [1.3 – 50 %]. Each of these subsets were used to construct a neighbour-Joining tree **(top)** revealing only clear geographic clustering when the high frequency SNPs are used **(right)**. Furthermore, the correlation of the genome distance calculation using all SNPs and each subset separately, reveals the poor correlation for the low frequency SNPs (0.14, left and 0.18, centre) and a strong correlation for the high frequency subset (0.93, right).

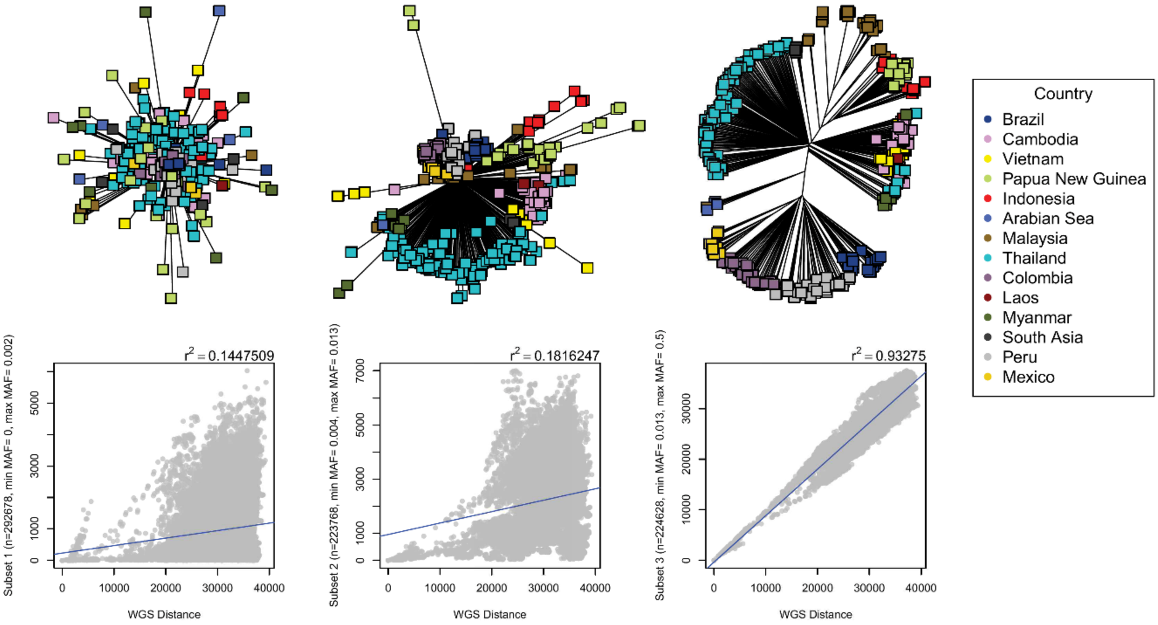
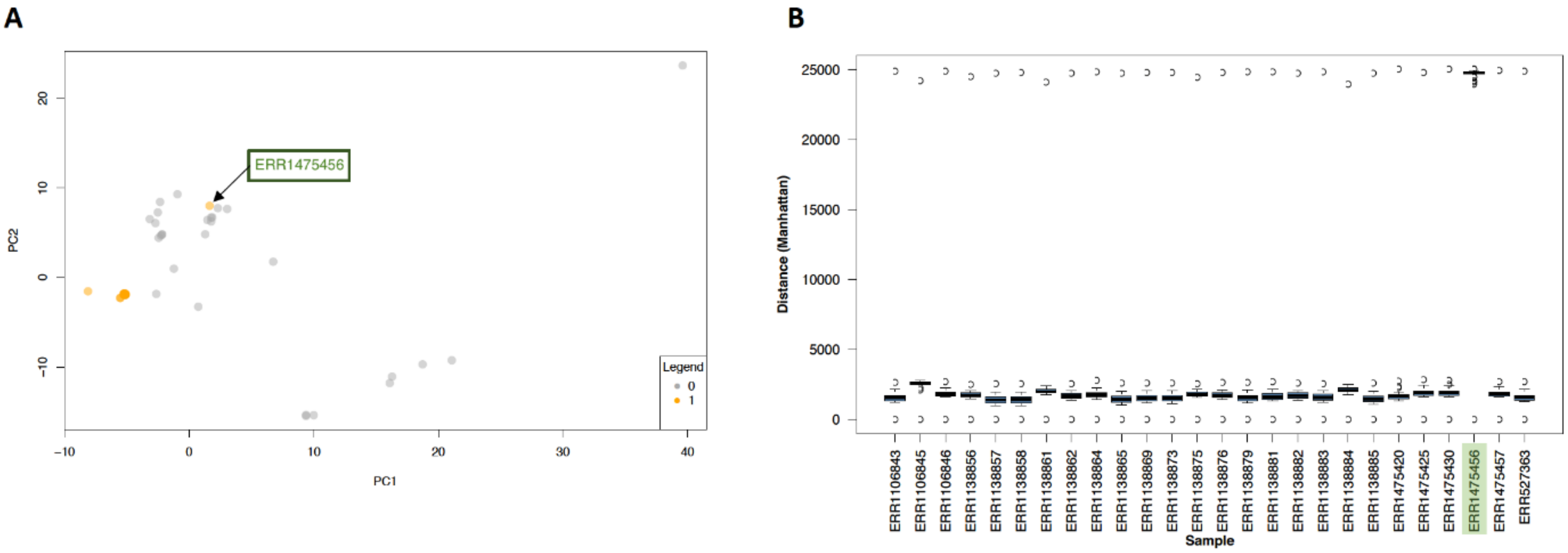
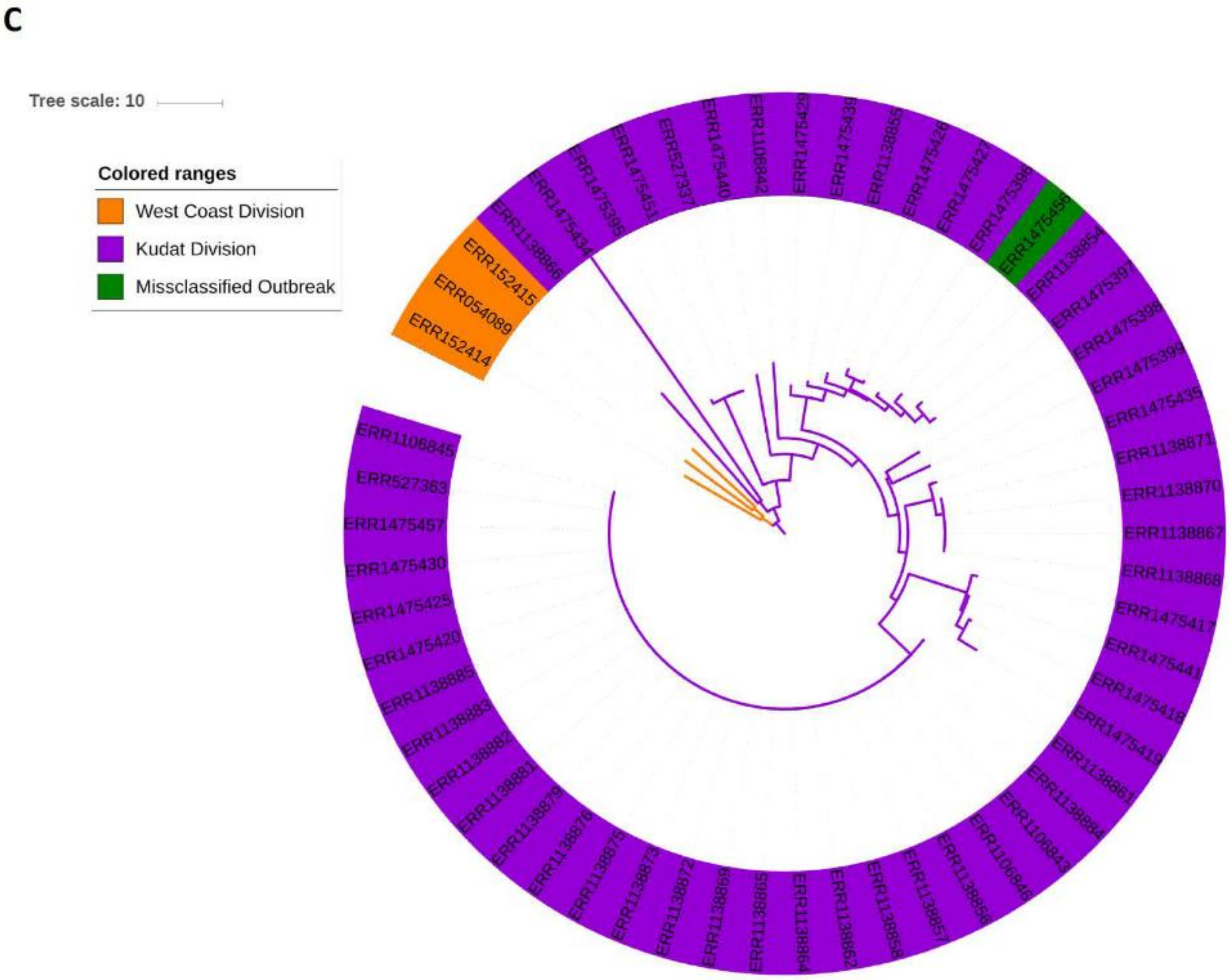


Figure 3. Use of 71 SNPs barcode in P. vivax isolates from Sabah, Malaysia reveals patterns of transmission. A dataset of 60 isolates from a near-elimination setting that has been exhaustively characterised by whole genome sequencing in [24] was characterized by means of a PCA using the 71 SNP barcode from this study (A). The PCA revealed a highly fidelity to the results obtained in the study using whole genome sequencing, with the outbreak population K2 (yellow in A and 1 in legend) clustering together. The only isolate showing distant clustering (ERR1475456) was further characterized using the genome distance by pair-wise comparison to other outbreak isolates (B) showing it not to be as closely related to the outbreak as indicated by microsatellite genotyping in [24]. Furthermore, a neighbour-joining tree was generated and coloured according to their district in Sabah, and revealed clustering patterns of the samples from the West Coast Division in Sabah (C).

(Figure on pages 226 [A and B] and 227 [C])



A
ERR1475456
Legend
0
1
PC1
PC2
−10
0
10
20
30
40
B
Distance (Manhattan)
0
5000
10000
15000
20000
25000
Sample
ERR1106843
ERR1106845
ERR1106846
ERR1138856
ERR1138857
ERR1138858
ERR1138861
ERR1138862
ERR1138864
ERR1138865
ERR1138869
ERR1138873
ERR1138875
ERR1138876
ERR1138879
ERR1138881
ERR1138882
ERR1138883
ERR1138884
ERR1138885
ERR1475420
ERR1475425
ERR1475430
ERR1475456
ERR1475457
ERR527363


C
Tree scale: 10
Colored ranges
🟧 West Coast Division
🟪 Kudat Division
🟩 Missclassified Outbreak
ERR152414
ERR054089
ERR152415
ERR1138866
ERR1475434
ERR1475395
ERR1475451
ERR527337
ERR1475440
ERR1106842
ERR1475429
ERR1475439
ERR1138855
ERR1475426
ERR1475427
ERR1475396
ERR1475456
ERR1138854
ERR1475397
ERR1475398
ERR1475399
ERR1475435
ERR1138871
ERR1138870
ERR1138867
ERR1138868
ERR1475417
ERR1475441
ERR1475418
ERR1475419
ERR1138861
ERR1138884
ERR1106843
ERR1106846
ERR1138856
ERR1138857
ERR1138858
ERR1138862
ERR1138864
ERR1138865
ERR1138869
ERR1138872
ERR1138873
ERR1138875
ERR1138876
ERR1138879
ERR1138881
ERR1138882
ERR1138883
ERR1138885
ERR1475420
ERR1475425
ERR1475430
ERR1475457
ERR527363
ERR1106845

TABLES

Table 1. Selection of 71 barcode SNPs for *Plasmodium vivax*. The barcode has both geographic origin prediction power and transmission inference ability. A certain degree of clustering across populations is found in the SNPs without complete fixation, allowing for increased number of haplotypes to be traced and transmission events to be investigated, the colours represent the gradient from 0 to 1 of the frequency of each SNP in each population, blue for highly frequent SNPs and red for low frequent SNPs.

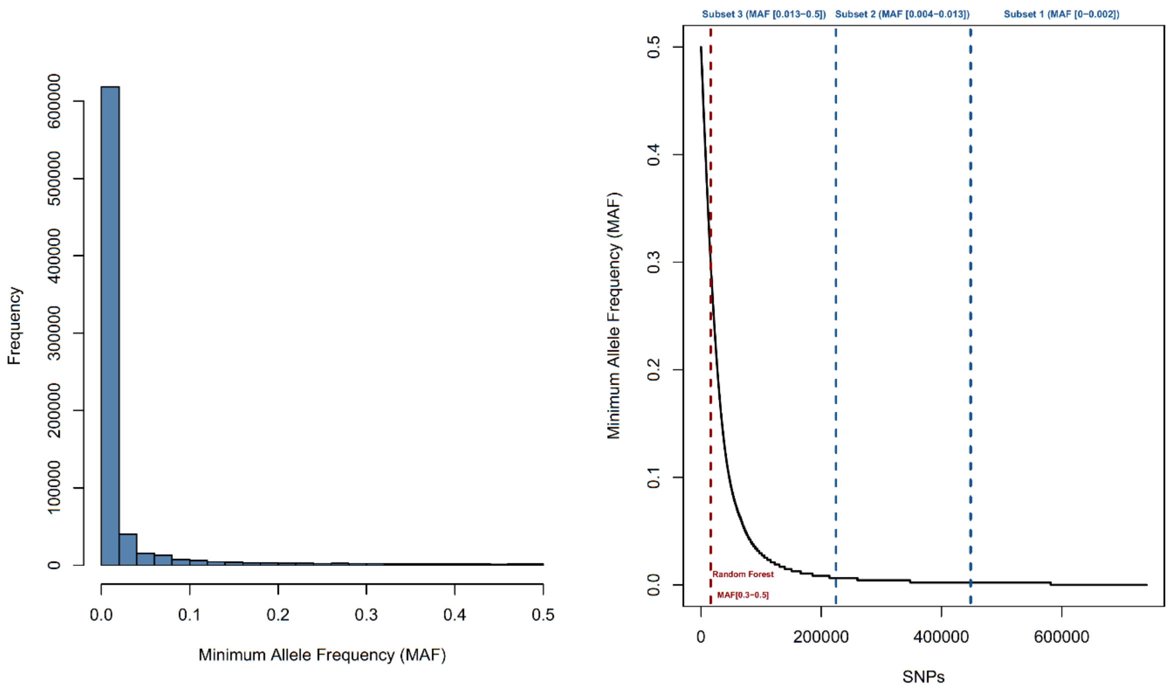
(Table in next page)

Table 1. Selection of 71 barcode SNPs for *Plasmodium vivax*. The barcode has both geographic origin prediction power and transmission inference ability. A certain degree of clustering across populations is found in the SNPs without complete fixation, allowing for increased number of haplotypes to be traced and transmission events to be investigated.

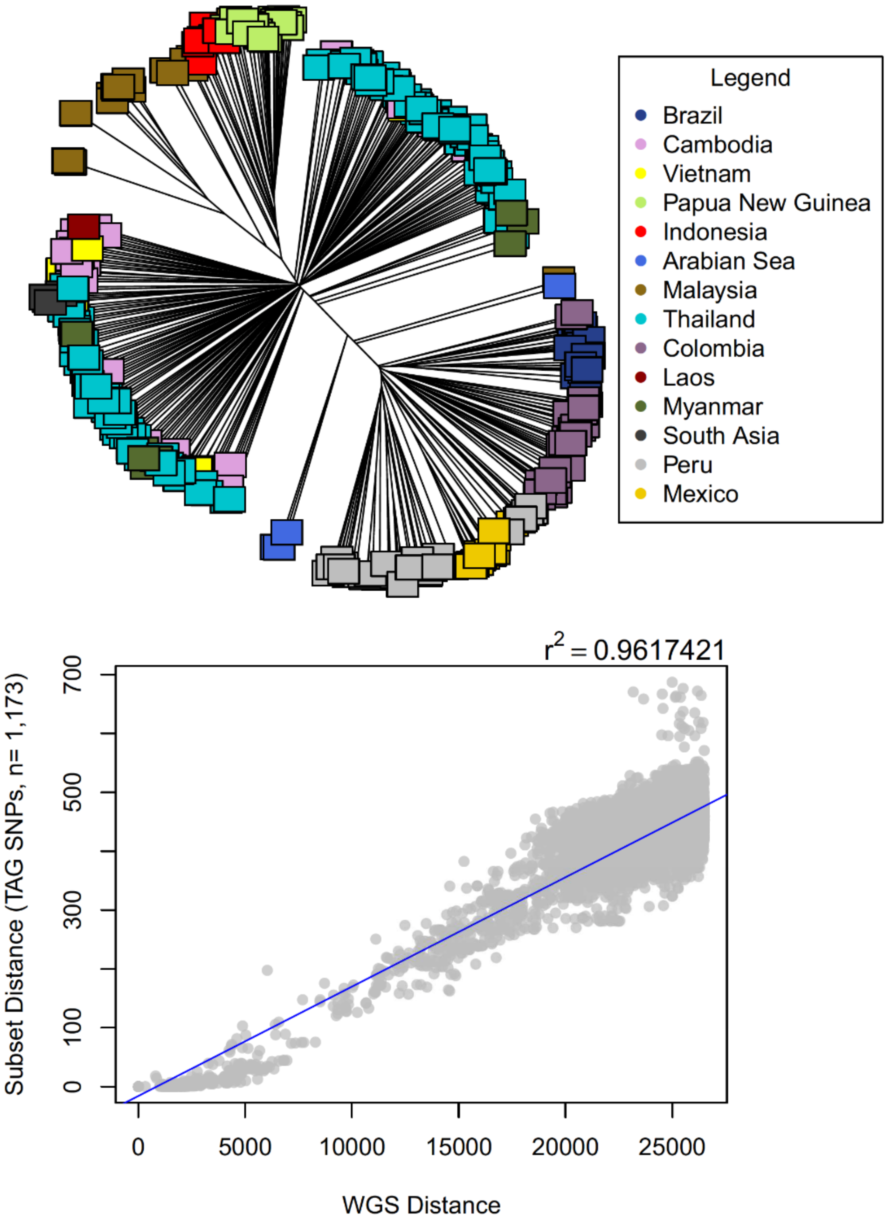
					South East Asia					South East Asia/Oceania			South America					
Chromosome	Position	Refe	Alteri	Gene	Thailand	Cambodia	Laos	Myanmar	Vietnam	Papua New Guinea	Indonesia	Malaysia	Colombia	Brazil	Peru	Mexico	Arabian Sea	South Asia
PvP01_09_v1	253550	C	A	PVP01_0903800	1.00	1.00	1.00	0.89	1.00	0.88	0.73	0.97	0.19	0.36	0.00	0.00	0.00	0.64
PvP01_13_v1	340505	G	A	PVP01_1307300	0.99	0.94	1.00	1.00	1.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.75	1.00
PvP01_14_v1	1071214	C	T	PVP01_1424900	0.98	0.94	1.00	0.67	0.86	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.45
PvP01_13_v1	924465	T	A	.	0.97	0.88	1.00	1.00	0.86	0.04	0.45	0.40	0.23	0.00	0.07	0.00	1.00	1.00
PvP01_03_v1	536764	C	T	PVP01_0312200	0.95	0.81	1.00	0.89	0.93	0.12	0.55	0.92	0.00	0.00	0.00	0.00	0.25	0.45
PvP01_06_v1	639238	G	A	PVP01_0615100	0.95	0.72	1.00	0.56	0.86	0.65	0.45	0.89	0.67	0.64	0.02	0.10	0.50	0.82
PvP01_14_v1	1270401	G	C	PVP01_1429500	0.95	0.00	0.00	0.67	0.00	0.00	0.09	0.89	0.00	0.00	0.00	0.00	0.25	0.09
PvP01_04_v1	692311	T	C	PVP01_0416900	0.95	0.94	1.00	0.89	0.86	0.42	0.55	0.13	0.70	0.61	0.55	0.15	0.50	1.00
PvP01_07_v1	497459	A	G	PVP01_0709800	0.94	1.00	1.00	1.00	0.93	0.04	0.36	0.98	0.00	0.61	0.00	0.00	0.50	0.36
PvP01_13_v1	769284	C	T	.	0.93	0.72	0.00	0.67	0.71	0.31	0.18	0.08	1.00	0.93	0.91	1.00	0.75	0.82
PvP01_02_v1	155305	G	T	PVP01_0203000	0.92	0.97	1.00	0.44	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09
PvP01_04_v1	649508	T	A	PVP01_0415900	0.92	0.69	1.00	0.78	1.00	0.04	0.18	0.38	0.00	0.00	0.00	0.00	0.00	0.27
PvP01_08_v1	59546	C	T	PVP01_0801100	0.91	0.69	0.50	0.89	0.71	0.73	0.27	0.03	0.16	0.29	0.14	0.05	0.50	0.36
PvP01_10_v1	1161740	G	C	PVP01_1026700	0.91	0.66	1.00	0.67	0.64	0.00	0.27	0.92	0.00	0.00	0.00	0.00	0.00	0.55
PvP01_11_v1	1372724	C	G	PVP01_1132000	0.91	0.50	1.00	0.67	0.64	0.35	0.64	0.86	0.26	0.18	0.17	0.45	1.00	0.64
PvP01_13_v1	815186	A	G	PVP01_1317300	0.91	0.63	0.00	0.78	0.86	0.04	0.18	0.02	0.00	0.00	0.09	0.10	0.50	0.27
PvP01_12_v1	2751196	C	T	.	0.91	0.94	0.50	0.78	0.86	0.00	0.09	0.05	0.81	0.50	0.78	0.75	1.00	1.00
PvP01_03_v1	799618	A	G	PVP01_0319300	0.90	0.56	1.00	0.78	0.57	0.04	0.09	0.10	0.16	0.25	0.28	0.05	0.75	0.82
PvP01_05_v1	1206169	G	T	.	0.89	0.75	1.00	0.78	0.86	0.08	0.09	0.90	0.00	0.00	0.00	0.00	0.25	0.36
PvP01_07_v1	503559	A	G	PVP01_0709800	0.88	0.69	0.50	0.89	0.50	0.00	0.27	0.49	0.00	0.68	0.00	0.00	1.00	0.45
PvP01_13_v1	1231631	T	C	PVP01_1329100	0.88	0.94	1.00	0.56	0.86	0.96	0.00	0.94	0.12	0.04	0.09	1.00	0.50	0.64
PvP01_01_v1	157446	T	A	PVP01_0103400	0.88	0.50	1.00	0.78	0.50	0.27	0.45	0.11	0.79	0.89	1.00	0.10	0.75	0.55
PvP01_05_v1	318005	C	T	PVP01_0507200	0.88	0.47	1.00	0.78	0.64	0.77	0.45	0.08	0.88	0.54	0.84	0.60	1.00	0.45
PvP01_10_v1	274067	T	C	.	0.88	0.84	1.00	0.78	1.00	0.19	0.00	0.03	0.93	0.39	0.84	0.10	1.00	0.82
PvP01_12_v1	2741883	T	C	PVP01_1265900	0.87	0.84	0.50	1.00	0.79	0.04	0.09	0.10	0.35	0.21	0.26	0.95	1.00	0.91
PvP01_05_v1	1087255	T	G	PVP01_0526800	0.86	0.28	0.50	0.89	0.36	0.15	0.55	0.92	0.02	0.36	0.76	0.00	0.75	0.82
PvP01_03_v1	812961	C	G	PVP01_0319600	0.83	0.78	1.00	0.89	0.93	0.00	0.09	0.92	0.00	0.00	0.00	0.00	0.00	0.73
PvP01_14_v1	1338047	T	C	PVP01_1430700	0.80	0.63	0.00	0.67	0.29	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00
PvP01_12_v1	2568422	C	T	.	0.79	0.84	1.00	0.89	0.79	0.00	0.18	0.97	0.00	0.00	0.00	0.00	0.50	0.36
PvP01_05_v1	1066232	G	A	PVP01_0526300	0.75	0.09	0.00	0.67	0.00	0.00	0.00	0.24	0.02	0.04	0.10	0.00	1.00	0.45
PvP01_09_v1	1728276	G	T	.	0.74	0.78	0.50	0.67	0.57	0.12	0.45	0.02	0.00	0.00	0.02	0.00	1.00	0.27
PvP01_03_v1	101653	A	G	.	0.74	0.78	1.00	0.56	0.64	0.65	0.45	0.62	0.16	0.14	0.69	0.15	0.75	0.55
PvP01_03_v1	101866	G	T	.	0.74	0.69	1.00	0.56	0.71	0.85	0.64	0.60	0.28	0.14	0.83	0.15	0.75	0.64
PvP01_03_v1	101610	G	A	.	0.71	0.78	1.00	0.56	0.57	0.65	0.45	0.11	0.26	0.14	0.71	0.15	0.50	0.45
PvP01_13_v1	1232713	G	A	PVP01_1329100	0.70	0.50	0.50	0.89	0.79	1.00	0.00	0.08	0.95	1.00	0.98	1.00	1.00	1.00
PvP01_14_v1	1451245	C	T	.	0.66	0.09	0.00	0.33	0.21	0.00	0.00	0.02	1.00	0.89	0.97	1.00	1.00	0.64
PvP01_05_v1	1181738	G	A	.	0.63	0.19	0.00	0.78	0.14	0.00	0.09	0.02	0.81	0.36	0.90	0.25	0.25	0.36
PvP01_08_v1	987176	C	A	PVP01_0822400	0.61	0.16	0.00	0.11	0.07	0.00	0.27	0.13	0.93	0.86	0.83	1.00	0.75	0.45
PvP01_05_v1	1252615	C	A	.	0.53	0.03	0.00	0.89	0.00	0.00	0.00	0.02	1.00	0.68	0.93	1.00	1.00	0.45
PvP01_04_v1	401576	A	G	PVP01_0409900	0.47	0.47	0.50	0.67	0.29	0.00	0.00	0.02	1.00	1.00	0.97	0.95	1.00	0.55
PvP01_08_v1	1508826	A	G	PVP01_0835600	0.47	0.75	0.50	0.56	0.64	0.35	0.00	0.16	0.98	0.00	0.16	0.55	0.00	0.36
PvP01_03_v1	101401	T	C	.	0.43	0.19	0.00	0.56	0.14	0.19	0.36	0.37	0.72	0.50	0.09	0.85	0.75	0.45
PvP01_13_v1	354018	T	C	PVP01_1307600	0.36	0.03	0.00	0.56	0.79	1.00	0.82	0.78	0.56	0.61	0.59	0.30	0.00	0.55
PvP01_14_v1	2799980	C	T	.	0.34	0.19	0.50	0.56	0.36	0.88	0.73	0.06	1.00	1.00	0.98	1.00	1.00	0.45
PvP01_12_v1	287842	A	C	PVP01_1207100	0.27	0.03	0.00	0.56	0.00	0.96	0.82	0.03	0.98	0.96	0.95	1.00	1.00	0.36
PvP01_14_v1	585620	C	T	PVP01_1413400	0.26	0.16	0.00	0.22	0.21	0.88	0.64	0.94	0.67	0.21	0.59	0.15	1.00	0.55
PvP01_04_v1	667734	T	G	PVP01_0416400	0.24	0.09	0.00	0.22	0.07	0.00	0.18	0.03	0.93	0.82	1.00	1.00	1.00	0.45
PvP01_09_v1	1217464	C	T	PVP01_0927600	0.18	0.00	0.00	0.11	0.00	0.54	0.18	0.02	0.93	0.82	0.97	0.50	0.25	0.18
PvP01_12_v1	2330891	T	C	.	0.18	0.06	0.00	0.00	0.07	0.46	0.09	0.89	0.91	0.93	0.93	0.05	0.50	0.36
PvP01_02_v1	697366	T	G	PVP01_0216200	0.16	0.00	0.00	0.00	0.07	0.00	0.00	0.02	0.77	0.43	0.84	0.30	1.00	0.00
PvP01_03_v1	435501	T	A	PVP01_0309300	0.16	0.03	0.00	0.11	0.07	0.00	0.00	0.00	0.95	0.79	0.93	1.00	0.00	0.00
PvP01_03_v1	382736	T	C	PVP01_0307900	0.15	0.03	0.00	0.22	0.00	0.00	0.00	0.03	0.12	0.07	0.00	0.00	1.00	0.00
PvP01_08_v1	1364769	A	G	.	0.15	0.03	0.00	0.44	0.00	0.00	0.00	0.70	1.00	1.00	0.95	1.00	0.75	0.18
PvP01_09_v1	641801	T	C	PVP01_0913800	0.14	0.06	0.00	0.44	0.00	0.04	0.00	0.02	1.00	0.82	0.93	1.00	1.00	0.82
PvP01_12_v1	629287	T	C	PVP01_1214900	0.14	0.03	0.00	0.56	0.00	0.00	0.00	0.00	0.81	0.89	0.98	1.00	1.00	0.27
PvP01_13_v1	1696786	C	G	.	0.13	0.00	0.00	0.44	0.00	0.08	0.00	0.02	1.00	1.00	0.97	1.00	1.00	0.45
PvP01_05_v1	606507	A	G	PVP01_0514500	0.12	0.03	0.00	0.00	0.29	0.46	0.64	0.94	0.72	0.68	0.59	0.35	0.25	0.18
PvP01_14_v1	2096518	A	G	PVP01_1448200	0.10	0.00	0.00	0.00	0.00	0.85	0.45	0.87	0.86	0.46	0.26	0.65	0.25	0.36
PvP01_08_v1	115537	A	C	.	0.08	0.03	0.00	0.11	0.00	0.00	0.00	0.02	1.00	0.93	1.00	1.00	0.50	0.27
PvP01_12_v1	1534427	T	C	.	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.93	1.00	0.95	1.00	0.18
PvP01_05_v1	213306	G	A	.	0.05	0.00	0.00	0.00	0.00	0.85	0.00	0.10	0.30	0.00	0.14	0.00	0.25	0.27
PvP01_11_v1	1790451	G	A	PVP01_1142200	0.05	0.16	0.50	0.11	0.00	0.96	0.73	0.89	0.44	0.32	0.33	0.75	0.00	0.09
PvP01_14_v1	801261	A	T	PVP01_1418100	0.02	0.09	0.00	0.00	0.00	0.96	0.27	0.03	1.00	0.89	0.79	1.00	0.75	0.45
PvP01_06_v1	531114	A	G	PVP01_0612100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.75	0.18
PvP01_08_v1	1510168	C	T	.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.61	0.72	0.40	0.00	0.00
PvP01_08_v1	1546424	C	G	PVP01_0836700	0.00	0.00	0.00	0.00	0.00	0.92	0.27	0.02	1.00	1.00	0.98	1.00	0.00	0.00
PvP01_11_v1	1548740	G	T	.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.82	1.00	0.00	0.00	0.00
PvP01_12_v1	323603	C	T	PVP01_1208000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.96	0.98	1.00	0.00	0.00
PvP01_12_v1	492536	T	C	.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.79	0.95	0.00	0.00	0.00
PvP01_13_v1	593388	T	C	PVP01_1313200	0.00	0.09	0.00	0.00	0.00	0.77	0.55	0.08	1.00	0.82	0.97	1.00	0.00	0.36
PvP01_14_v1	2151758	G	A	.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.37	0.82	0.03	0.00	0.00	0.00

SUPPLEMENTARY INFORMATION

Supplementary Figure 1. (Left) Distribution of SNPs according to minor allele frequency (MAF); **(Right)** SNPs partitioned into three equally sized divisions based on MAF (blue dashed lines), and a cut-off of MAF >0.3 (red dashed line) was used to pre-select SNPs for downstream analysis.

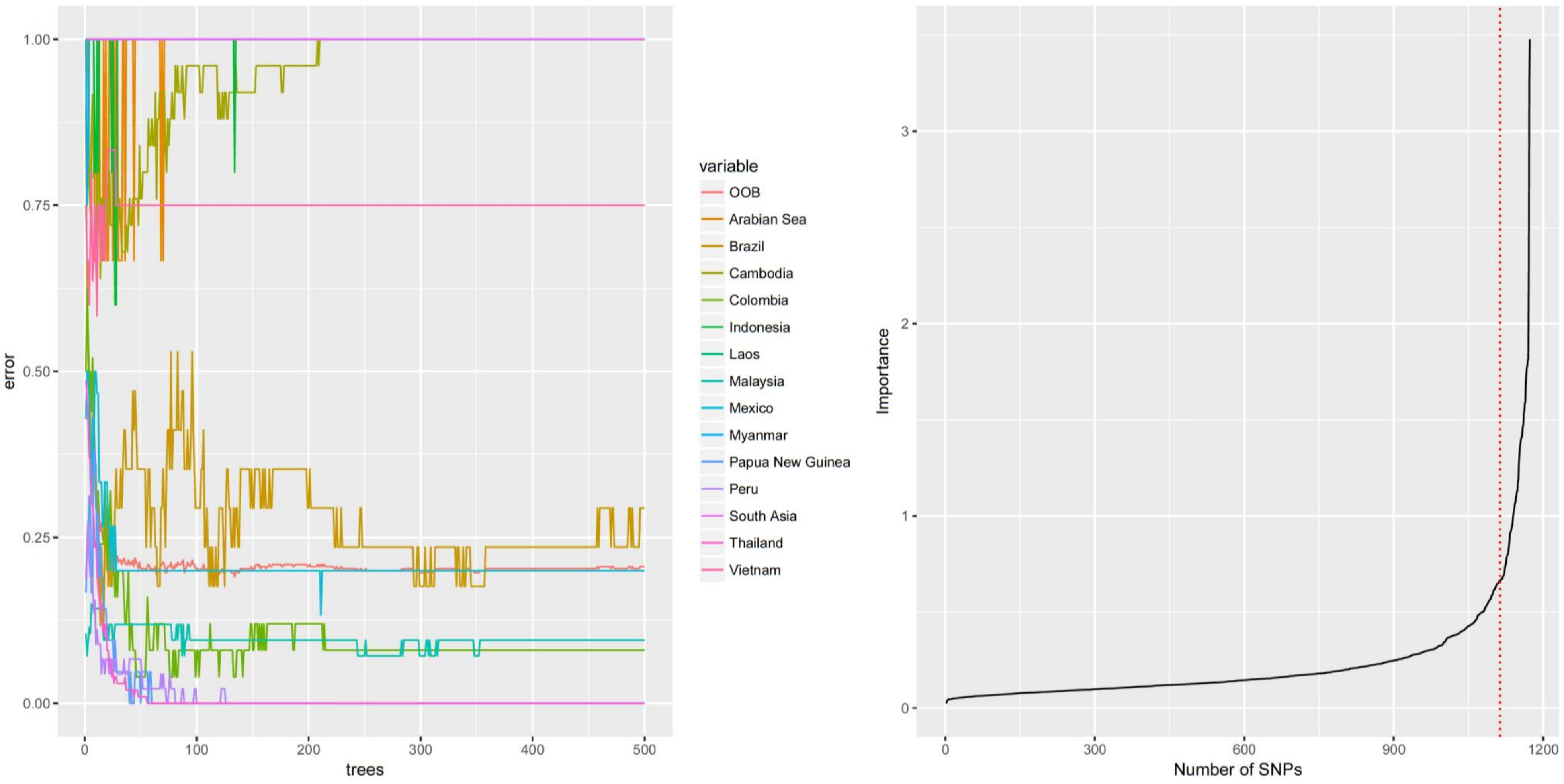


Supplementary figure 2. (Top) Neighbour-joining tree based on 1,173 tagging SNPs in *P. vivax* selected using the *TAGster* [30] software shows strong similarity with the tree from **Figure 2 (right**), observing a strong geographical signal; **(Bottom)** The correlation of genome distance based on whole genome sequencing (WGS; 741k SNPs) with the subset of tagging (1,173) SNPs is high (r^2 = 0.96).

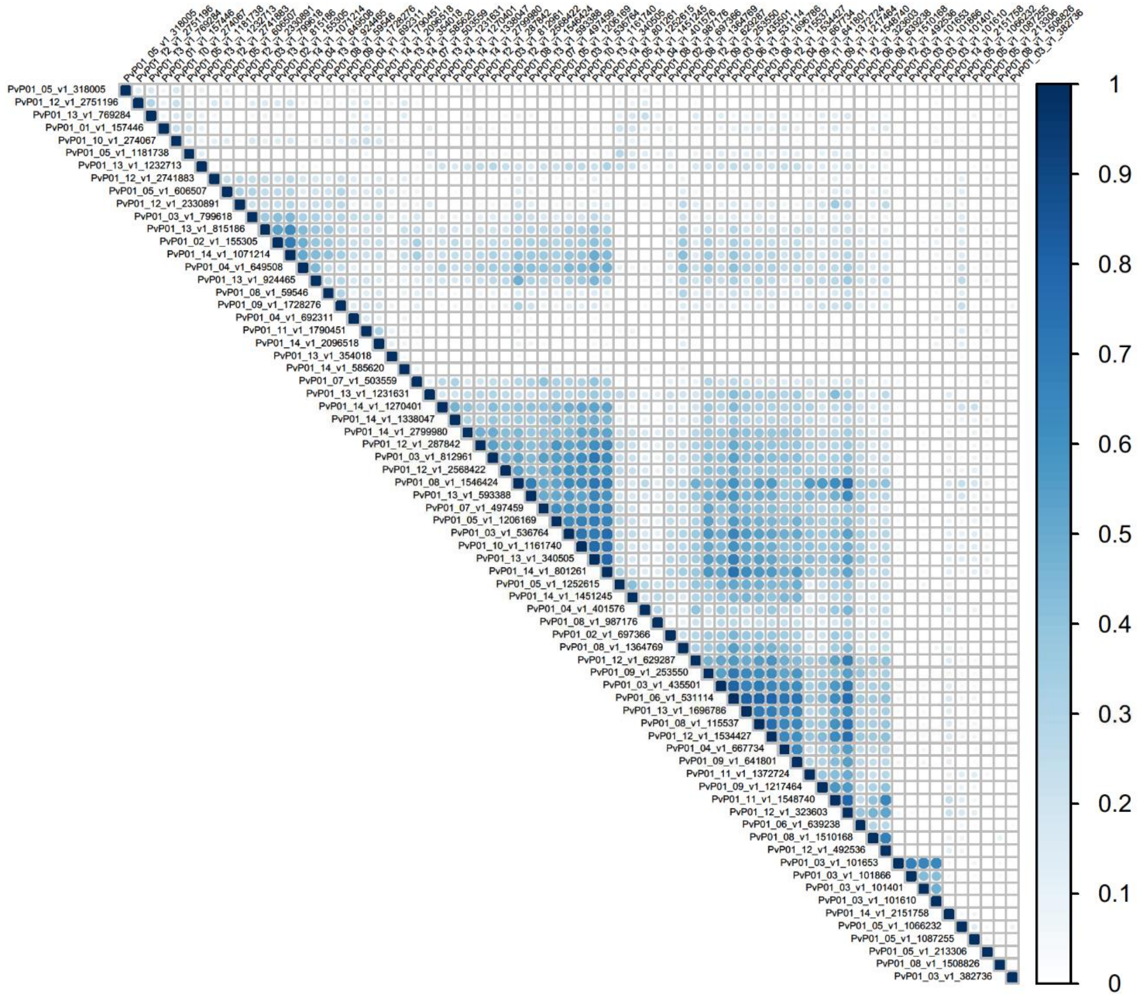


Supplementary figure 3. **(Left)** Classification error for the different geographic categories across the 500 trees in the random forest model reaches stability when 100 trees are averaged (OOB). **(Right)** Variable importance estimated from the random forest model for the number of 1,173 tag SNPs. The red dashed line is the cut-off based on importance, which is the threshold used to determine SNP inclusion in the barcode based on the SNPs with the highest importance.

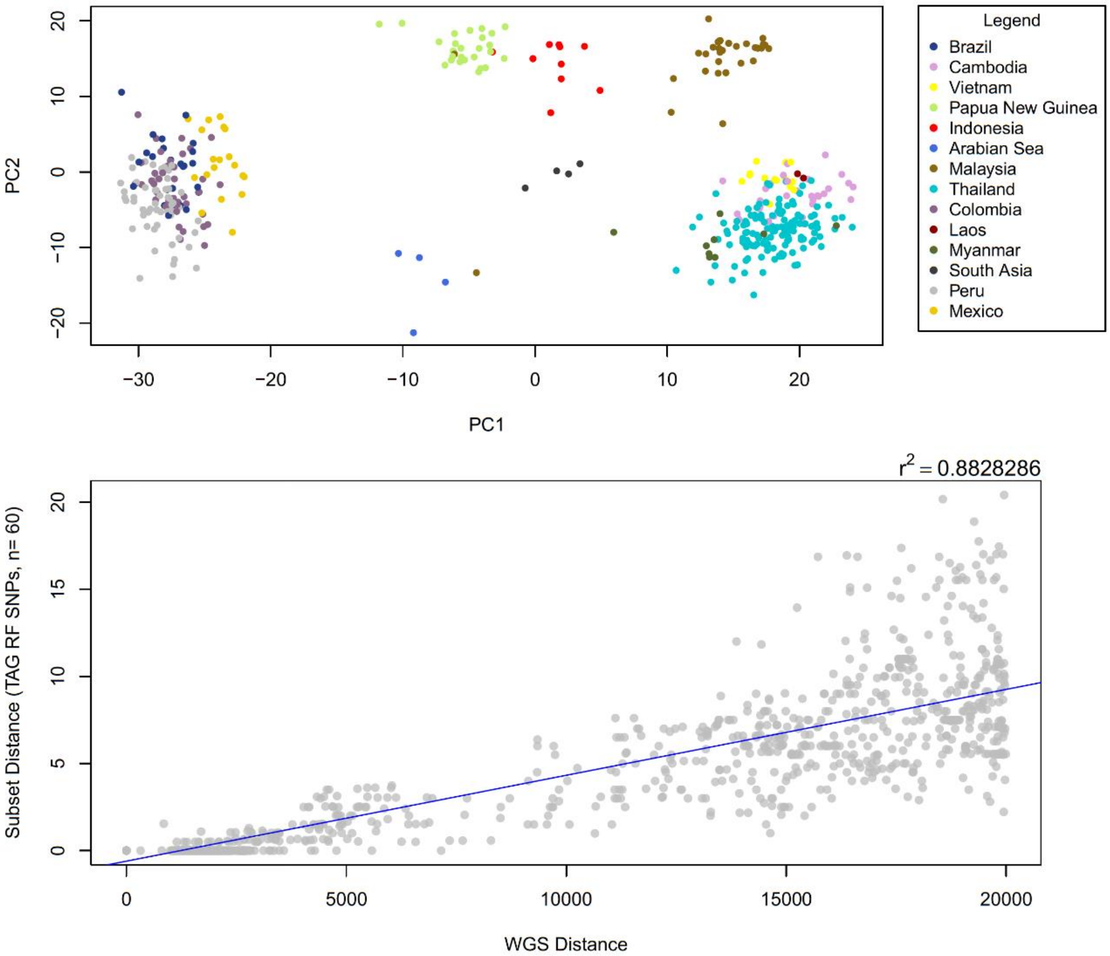
(Figure in next page)


error
1.00
0.75
0.50
0.25
0.00
0
100
200
300
400
500
trees
variable
OOB
Arabian Sea
Brazil
Cambodia
Colombia
Indonesia
Laos
Malaysia
Mexico
Myanmar
Papua New Guinea
Peru
South Asia
Thailand
Vietnam
Importance
3
2
1
0
0
300
600
900
1200
Number of SNPs

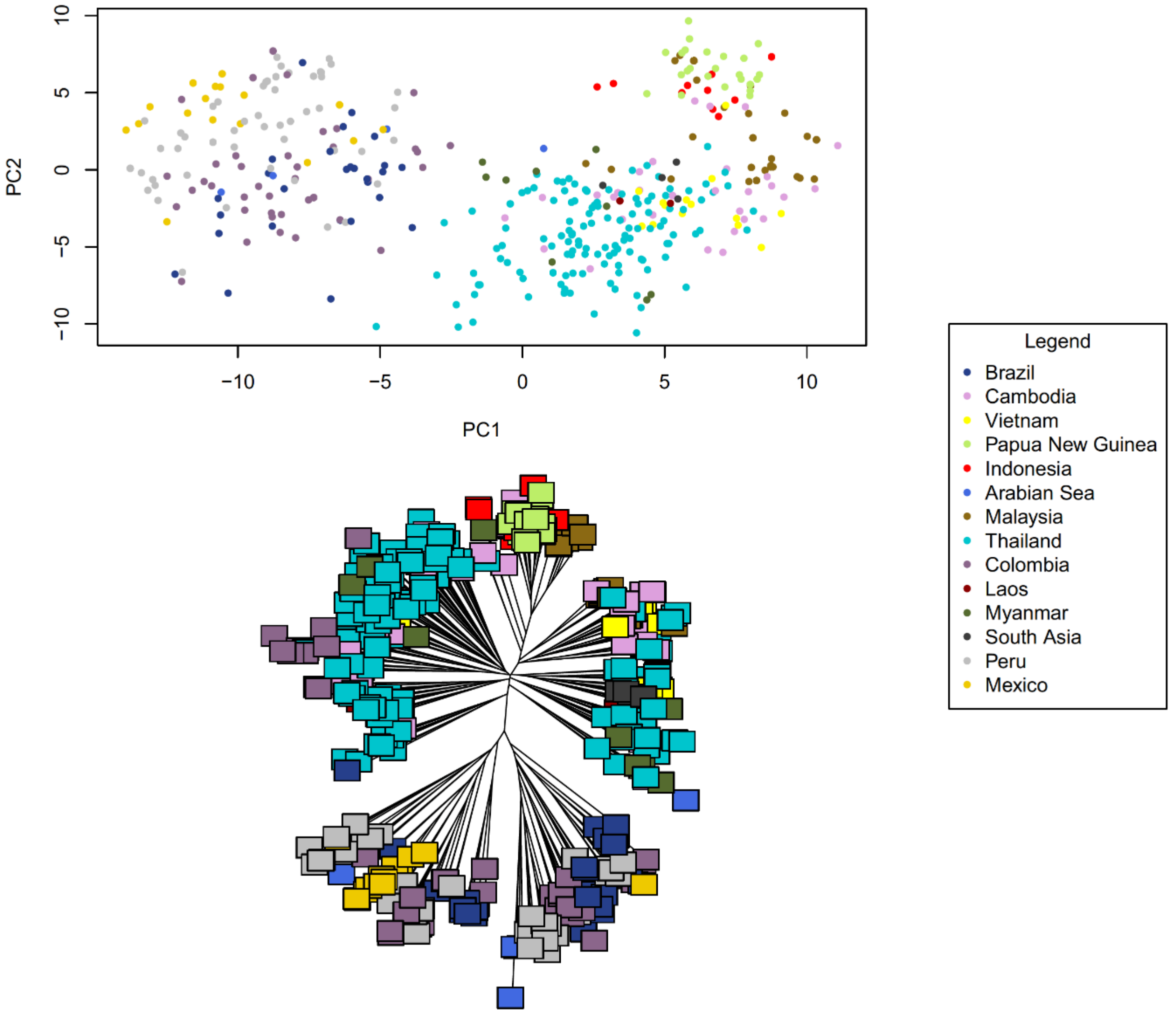
Supplementary Figure 4. Correlation of the 71 SNP barcode for *Plasmodium vivax*. Low overall correlation was found between the 71 SNPs selected (mean r^2= 0.15). Blocks of correlation were observed which correspond with SNPs with geographic signal (i.e. Southeast Asian high frequency SNPs).



Supplementary Figure 5. PCA plot reveals geographic clustering of *Plasmodium vivax* isolates using 71 SNPs barcode. Clustering by region and by country was observed when using the 71 SNP barcode as observed in the top panel. Furthermore, a strong correlation of 0.88 was observed between the distances based on whole genome (n=741k) and 71 barcoding SNPs, revealing the potential for the barcode to identify closer related intra-border isolates.



Supplementary Figure 6. The PCA plot and neighbour-joining tree using a previously published 42-SNP barcode [26] shows ambiguous geographic clustering of *Plasmodium vivax* isolates. Geographical clustering by region was apparent, although a degree of overlap was observed and separation by country was not clear. This result is suggested by the low accuracy (79.3 %) obtained when predicting geographical origin using a random forest model formed with the set of 42 SNPs.



Chapter 8
Discussion and Conclusion

DISCUSSION

This thesis covers the population genetic analysis of whole genome sequencing data from isolates of *Plasmodium* malaria infections. In **Chapter 2**, the PacBio long-read sequencing technology was used to generate a new highly robust reference genome for the zoonotic *P. knowlesi* human-adapted strain A1-H.1 [1]. Advances in the use of this parasite strain in the laboratory have allowed the *in vitro* culture of *P. knowlesi* using human blood, thereby avoiding the use of difficult to access macaque blood. Concerns have been raised about the quality of the assembly of the available *P. knowlesi* H strain reference [2], therefore the generation of this new reference can be a useful alternative or comparative resource. Use of the A1-H.1 reference led to superior mapping coverage across the available Illumina isolate collections, out-performing the H strain. The use of PacBio sequencing can also provide insights into methylation events in the genome. The coverage levels obtained in the sequencing did not allow for the detection of m5C methylation events, which have shown to be of relevance in *P. falciparum*[3]. However, I reported here for the first time the presence of m6A and m4C methylation events in the *P. knowlesi* genome, which have been observed in other eukaryotic organisms [4] where they are thought to play a role in transcription regulation [4]. Furthermore, several regions within the genome with a higher density of methylation events were identified. These regions included several methyltransferases and the *NBPXa* gene, which is an essential gene for human red-blood cell invasion [5]. The robust reference genome and the insights described here, in conjunction with the ability to culture *P. knowlesi in-vitro,* create the perfect framework to allow for follow-up studies of parasite biology. Potential studies could involve the use of RNAseq sequencing to investigate the possible impact of the epigenetic events on transcription levels and regulation, similar to investigations in other organisms [6]. Furthermore, the use of new molecular technologies like genome editing will make it possible to understand the biology of the parasite. CRISPR editing has been successfully applied to other *Plasmodium* parasites [7]. A robust, strain-specific reference genome is a highly primordial tool for the design of genetic and molecular experiments, for example, producing mutants that would allow the study of the role played by different genes in the parasite life cycle. This approach has been previously performed with invasion-related genes in *P. knowlesi* [5].

In **Chapter 3,** the mapping of all the available short-read Illumina sequenced isolates against the A1-H.1 reference was performed. Studies have shown that application of erroneous bioinformatic analytical approaches to Plasmodium isolates with complex infections (multiplicity of infection [MOI] > 1) could bias the results [8]. By applying a stringent filtering of isolates presenting high MOI, a more complex population structure is revealed that is complementary to the previously described genomic clusters associated with either Peninsular Malaysia, or one of the two macaque species that are natural hosts for *P. knowlesi* [9]. The genomic comparison of the two new regional clusters in Betong and Kapit belonging to the long-tailed macaque cluster (*Mf-Pk*) revealed the presence of mosaicism in their haplotypes (determined initially by allele frequency differences using F_{ST}). These mosaicisms were determined to be introgression events from the pig-tailed genomic cluster (*Mn-Pk*). This is the first report that the two host-specific subpopulations of *P. knowlesi* are not genetically isolated, as it was previously thought [10]. An in-depth study of the genomic composition of the individual isolates revealed the introgressed fragments to be present across multiple isolates and spanning all chromosomes in the genome. Despite this, the only strong signal of selection for such fragments was found in a chromosome 8 region, harbouring genes related with mosquito life stages of the parasite, suggesting a possible mosquito-driven effect in the parasite genome. Analysis of organellar genomes also supported introgression and suggested the possibility of such events occurring in non-recent times. This is the first time that such strong and recent events are described in any *Plasmodium* parasite, and the relevance of such events in the context of the ecological change that Malaysian Borneo is experiencing is still to be determined. In the future, studies that focus on understanding the geographical breadth as well as the population depth of these events would be of high importance. Prospective studies with representative sampling from each of the different regions in Borneo would be required to understand the extent of introgression. Compartmentation of the geographical areas based on ecological characteristics such as land use (forest, plantation, urban, transition between them), and mosquito vector distribution could help elucidate what factors might be driving such exchanges. Further, the study of retrospective samples collections could provide insights into the historical relevance of such events and whether they played a role in the establishment of the genetic subpopulations observed currently in Malaysian *P. knowlesi* parasites.

Building on the introgression insights, and in order to overcome the difficulty to sequence low-parasitiemia infections of *P. knowlesi* infections, a new methodology for selective whole genome amplification (SWGA) was implemented and tested in several newly collected isolates from both Peninsula and Borneo Malaysia **(Chapter 4)**. The new methodology had been previously used in other *Plasmodium* species [11,12] and primers specific for *P. knowlesi* were designed. The new method yielded a significant improvement in overall parasite DNA concentration compared to human DNA for almost all the samples tested. This approach led to ~40% of the genome covered with at least 5 reads, thereby allowing the calling of SNPs. Using laboratory data generated by our group, it was possible to demonstrate that the method is reliable to amplify sample isolates across all three subpopulations of *P. knowlesi*, including some complex infections (MOI > 1). This approach demonstrated that although the previously sequenced laboratory strains had been collected long ago from Peninsular Malaysia, they still are representative of this subpopulation when compared to the recently collected isolates. Further investigation of the Peninsular isolates revealed signatures of introgression events, which seemed to originate from the *Mn-Pk* subpopulation. However, in this case the genes affected did not seem to be associated with mosquito-related genes, but with invasion related loci such as *DBP*β and *NBPXa,* as well as exported proteins from the subtelomeric regions. This observation, although represented only by single case events, might suggest that these introgression events also can affect invasion related genes, where differences in the haplotypes of the *DBP*α gene have shown to affect binding affinity of such proteins to human receptors [13]. It is worth noting that given that the haplotypes observed originate from human infections, they are all thought to be able to invade human red blood cells, and therefore might not be a good representation of the natural pool of invasion haplotypes in the macaque reservoir host. Despite this, the dynamics of these human-selected haplotypes are of relevance as they might provide different invasion efficiencies and could be of relevance in order to understand the spread of this newly emerging form of human malaria. A comprehensive study of the introgression patterns in the 5 *P. knowlesi* RBP and DBP genes revealed that extensive exchange events have occurred within the *NBPXb* gene for the Borneo subpopulations *MF-Pk* and *Mn-Pk,* which suggests more complex patterns related to the invasion dynamics of the different clusters. Follow-up work could involve additional sampling of isolates, with linked GPS coordinates. Microsatellite data has shown that the distribution of the clusters across regions of Peninsular Malaysia is uneven [9], and exploring how this distribution correlates with introgressions would assist with understanding how these events arose. Furthermore, data collected in a longitudinal manner would provide insight into whether these events are sporadic, self-contained or are currently spreading across the population. The biology underlying the events could be explored using CRISPR systems [7]. For example, the different haplotypes found to be introgressed could be inserted into a laboratory strains (e.g. A1-H.1), thereby assessing their efficiency on invading human red blood cells, and providing insights into how these exchanges respond to natural adaptations to new, more efficient invasion routes. This new finding suggests that other biologically relevant pathways can be targeted and studying such event will help understand how these deeply differentiated populations originated, helping to understand whether this phenomenon is due to adaptation to the host macaques, the anopheline vectors or both.

Like *P. knowlesi*, the population structure of *P. falciparum* is broadly geographically distributed and differentiated. However, it is unclear if highly variable gene families can be used to barcode the geographical source of infections. In **Chapter 5** a study of the genetic and structural diversity of a vaccine candidate and a key functional gene related with malaria in pregnancy, the *var2csa,* is shown. Given the hypervariability of this gene, mapping strategies do not succeed at unravelling the genetic diversity. Therefore, I used the long-read sequencing technologies to generate robust sequences spanning the whole gene for several laboratory strains and clinical isolates, and to confirm the presence of isolates with extra genic copies and the degree of similarity of the copies found. The short-read sequencing was then used on the same isolates to establish a pipeline to identify samples with extra gene copies and to map distribution of such isolates geographically, identifying an increased prevalence of isolates with extra different copies in West African populations. Using these same isolates, I established a *de-novo* assembly pipeline and calibrated it using the long-read assembled genes and generated a set of more than 1,200 sequences spanning almost 80% of the gene (>7 kbp). These sequences allowed the study of the diversity of the amino-acid sequence of the ID1-DBL2Xb region, which includes the minimal binding domain of the CSA receptor [14] where the protein binds and is the target of both vaccines currently in Phase 2 clinical trials[15,16]. I identified 4 clades of sequences, two of which had been previously described [17] and Cluster 1 had been associated with low birthweight in pregnancy in West Africa [17]. Furthermore, mapping the clades geographically showed Clade 1 to be present across all the populations, although at different frequencies. Moreover, higher diversity was found in African populations than in Southeast Asian populations and two of the Clades seemed predominantly African. From a within gene diversity perspective it was found that the ID1-DBL2Xb region, which is the main target of the two main vaccine candidates, harbours a surprisingly high number of small, low frequency insertions and deletions. This high level of polymorphism prevents the correct calculation of the nucleotide diversity in this region and underestimates the overall diversity. The impact of such small structural variants is yet undetermined, especially with respect to both binding affinity to the CSA receptor during pregnancy and vaccine antibody recognition. One approach could involve grouping *var2csa* gene variants based on the different clades, the generation of consensus sequences for each clade, and then an assessment of the impact of inter-clade differences in binding with the CSA receptor. Currently, new protein structure modelling algorithms are being developed in order to estimate the effect that combined mutations might have on the binding affinity and functionality of other proteins in *P. falciparum* [18] and other organisms [19]. The application of similar approaches could streamline the study of the impact of such diversity patterns in the functionality of the VAR2CSA protein. Further, a study of the diversity of the CSA receptor in the human host could also be performed. This approach could reveal whether diversity patterns respond to host-associated adaptation, and would require the collection of host and parasite samples from clinical cases.

The study of the *P. vivax* malaria parasite has historically been neglected in favour of the deadlier *P. falciparum* species [20]. In order to design barcodes that can tackle the study of such species in an epidemiological manner, firstly genomic diversity and population structure of this parasite has to be investigated. In **Chapter 6** I collated the largest collection of *P. vivax* whole genome sequenced clinical isolates publicly available at the time, and performed a genomic diversity analysis. Using the SNPs found across the dataset it was possible to confirm that *P. vivax* parasites harboured more genetic diversity than *P. falciparum,* and that the two main populations in Southeast Asia and South America present a greater genetic distance compared to Africa and South East Asia for *P. falciparum*. It is worth mentioning that due to the high level of MOI observed (present in 47% of isolates) and the approach followed, using the major calls; therefore studying the most abundant strain in each sample, the results obtained would underestimate the level of genomic complexity in these isolates. Construction of a neighbour-joining tree confirmed the presence of clustering by geography, in particular, a South American collection of isolates from different countries and a South East Asian population, mainly consisting of isolates from Thailand. Using these populations, genetic regions with allele frequency differences revealed using the F_{ST} metric and selection pressure methods were detected, and loci belonging to mosquito life-stages related genes and drug resistance (e.g. *MRP1* and *DHPS*) were identified. Furthermore, novel deletions in the *MDR1* gene promoter region, and gene duplications in the *PvDBP* invasion-related gene were found. SNPs with regional classification power where found in the apicoplast, and the analysis of mixed infections, revealed the power of the mitochondrial genome to be used as a barcoding tool for different *Plasmodium spp.*

This work was limited in sample size, but two large *P. vivax* genomic studies with new isolates have been published subsequently [21,22]. In **Chapter 7** I gathered the new data together with previously analysed and in-house sequenced isolates in order to design a new barcode that could tackle both the prediction of geographic origin and the inference of transmission dynamics within the parasite populations. The set consisted of >741k SNPs across 446 isolates from 16 countries across the globe. Using a neighbour-joining tree method the geographical clustering of the isolates observed in **Chapter 6** was confirmed and it was established that intra-border isolates are closer related based on genomic distance than inter-border isolates therefore making a SNP barcode for country classification possible. In order to reduce the number of SNPs the analysis was stratified based on partitioning the minor allele frequency (MAF) into tertiles. It was found that the SNPs that harbour the highest MAF across all the populations drive the geographical clustering observed in the data. The 16,075 pre-selected SNPs by MAF > 0.3 were then screened using a SNP tagging software, which identified 1,173 SNPs that capture the variability harboured by 40% of the pre-selected SNPs. These SNPs were used to construct a random forest model to classify samples according to country of origin. The 60 SNPs with the highest classification power were then selected, and complemented by 11 high F_{ST} SNPs in order to improve classification power. The final set of 71 SNPs were used to train a set comprised of 80% of the isolates, accomplishing a 93.4% accuracy in classification by country. Furthermore, the same SNP barcode was used to replicate the results of a genomic study in a low-endemic setting in Malaysia [23], and confirmed the transmission dynamic trends observed using whole genome sequencing. This work shows that the 71-SNP barcode outperforms previous results using microsatellites and previously published barcodes [24]. This barcode was constructed, and can be updated, using a straight-forward pipeline and can accommodate large sets of isolate data. Ideally, new collections and datasets should be prioritised from geographical regions with less sequencing coverage, such as Central America, East Africa and the Middle East. One potential study would be the prospective sampling and genotyping using the SNP barcode of isolates from malaria endemic regions. By combining mathematical modelling approaches and barcode-informed isolate relatedness, it would be possible to establish the migration patterns of the parasite populations across different regions. The reduction in the costs of sequencing and genotyping, including of barcode SNPs using PCR-based amplification (\$5-10 per sample [25]), will facilitate large-scale genomic epidemiological studies. Further extensions of the presented work, include the creation of barcodes for other *Plasmodium spp*., such as *P. falciparum* for which there is a collection of more than 3,000 isolates currently publicly available. These barcodes could be complemented with genetic regions that can reveal drug resistance profile of parasites and *Plasmodium spp.* in the same assay.

CONCLUSION

Whole genome sequencing can be used for the study of *Plasmodium* parasite populations and its genomic diversity. Different sequencing-based technologies can tackle different problems, such as long-read platforms for the generation of genome references and resolving complex gene morphologies, or short-read sequencing for lower-cost genomic studies of genetic diversity in candidate genes. The use of different methodologies such as mapping, *de-novo assembly*, statistical methods, population genetics and phylogenetics, combined with a robust quality control in pre-analysis steps are the key to enable the correct interpretation of the information harboured by the sequencing data. In this thesis, I have demonstrated the potential of such methodologies to address issues such as: (i) the generation of high quality robust new reference genomes for *Plasmodium spp.*, (ii) to provide insights into the epigenetics of *Plasmodium* parasites, (iii) determination of genomic dynamics of malaria in complex ecological niches, (iv) inform vaccine deployment and provide insights into the potential impact of genetic diversity in vaccine development efforts against malaria, (v) determining the genomic population structure within a parasite *Plasmodium* population, and (vi) the generation of genotyping tools for the study of *Plasmodium* populations and its transmission dynamics.

THE FUTURE OF PLASMODIUM GENOMIC STUDIES

It is an exciting time for the study of *Plasmodium* genomics, with technological improvements in high-throughput sequencing and genotyping technologies, at lower costs. These innovations make the typing of thousands of samples logistically feasible and economically affordable will allow the study of large genomic epidemiological studies. The recent WHO 2017 malaria report shows a worrying trend regarding funding to fight for malaria elimination, with a large funding gap [26]. This fact combined with the heterogeneity of malaria transmission [27–29] has forced the key funders and control programmes working in the endemic regions to reassess their efforts. This assessment has led to a drive to create and deploy tools that can determine foci of infection or hot-spots of transmission, thereby assisting the more efficient distribution and application of available resources.

There is still scope for the improvement in the collection of isolates from specific regions of the world, to fill in the gaps, such as the Middle East and Africa for *P. vivax*, and *P. falciparum* populations in Africa. This improvement in geographical resolution would allow for new genomic studies and characterization of unknown diversity. Despite this, the tools designed in this thesis could help to further understand the dynamics of the parasite populations, especially regarding micro-epidemiology within country boundaries. Prospective studies would need to be implemented in order to determine the resolution that can be reached in tracking parasite transmission using these tools. These tools could also complement those already deployed such as the serological surveillance assays and the use of rapid diagnostic tests, both in high transmission and low transmission or near-elimination settings.

The combination of genomic tools together with phenotypic information, such as drug resistance, disease severity, vaccine efficacy, or even geographically located data could allow researchers to tackle a large number of the still unanswered questions in malaria epidemiology. In the Greater Mekong Sub-region (GMS), the WHO has established the target of eliminating artemisinin-resistant malaria [30] as the only realistic way to avoid the spread of resistant parasites to other regions of the world. In order to achieve this goal, there is an urgent need for tools that can identify both the drug resistant parasites, as well as methods that can predict and survey the dynamics of the parasite populations. As shown in **Chapter 7,** using the relatedness between the parasites to establish transmission patterns and events in epidemiological studies is possible. The use of new drug resistance molecular markers such as the *kelch-13* mutations [31] or the *plasmepsin II–III* copy number [32] could be combined with a *P. falciparum* barcode, developed in a similar way to the one characterised in **Chapter 7**. If implemented in a portable genotyping device, such a barcode could track drug-resistant *P. falciparum* populations across the region, and inform targeted interventions by malaria control programmes.

In **Chapters 2, 3 and 4,** for *P. knowlesi* I described for the first time in the *Plasmodium* genera a marked mosaicism across highly differentiated subpopulations like the one observed in other parasites such as trypanosomes [33]. In order to characterize the extent and relevance of such events, further isolate collections from different regions are needed, and the phenotypic characterization of the different haplotypes presented here need to be established. Prospective longitudinal studies, which would track these events across populations, could assist with determining the forces behind such events. The sequencing of previously collected samples using the approach described in **Chapter 4** could help elucidate if such events have played a role in past evolutionary trends, or are effectively a result of the current ongoing ecological changes in Malaysia. It is assumed that infections of *P. knowlesi* are in the majority of cases zoonotic events, but these exchanges and introgression events generate an immense pool of genetic diversity for the parasite. This diversity combined with the disruption of its natural niche, could drive the parasite populations to draw from this genetic pool in order to transition to more stable transmission routes, such as human-human infection.

Finally, in **Chapter 5** I described the global structural and genetic diversity of the *var2csa* gene in *P. falciparum*, associated with malaria in pregnancy and a vaccine candidate. For the first time I reported a surprisingly high number of insertions and deletions in the target region ID1-DBL2Xb of the vaccines being developed. The impact of such variability should be investigated both from the point of view of the disease phenotype (binding to placental CSA receptor) as well as from the point of view of vaccine efficacy. Such experiments would need to be approached from a protein modelling point of view, given the high abundancy of insertions and deletions found across the isolates. Despite this, using the sequences generated in **Chapter 5** consensus protein sequences could be created and purified for each of the Clades, and their binding affinity studied using more traditional binding affinity assays. There is a need to characterize the impact of the parasites with extra copies of the *var2csa* gene, which in this study are shown to be present in all populations; although more prevalent in West Africa. A follow-up study could enrol pregnant woman infected with malaria, and collect birth data, any adverse outcomes, and a placenta biopsy sample. Parasite material could be obtained from the mothers and the placental biopsies, their *var2csa* sequences compared, and copy number associations with clinical phenotype investigated.

The *var2csa* analysis revealed four "ID1-DBL2Xb" Clades that had uneven distribution across geographical regions. The geographic distribution of the clades suggests that there is a need for multi-clonal vaccine strategies. Further, the appearance of 2 African driven clades needs follow-up investigation, particularly to ascertain their prevalence and any association with disease severity during pregnancy and adverse outcomes at birth. Furthermore, it is unclear if the association observed between some of these clades with adverse outcomes in West Africa can be extrapolated to other populations where the 3D7-like clade has been found, such as Southeast Asia. This data could also be used to inform geographical locations where the first trials for vaccine deployment should be tested. In particular, deployment of a vaccine is expected to be more effective in regions harbouring parasite populations with lower diversity. Finally, the methodology described here and the data generated could assist the implementation of a baseline survey in regions selected for vaccine trials to compare pre- and post-vaccination deployment parasite populations in order to track selection for clade specific variants.

In conclusion, current and emerging technologies have the potential to expand and improve on the findings of the work presented in this thesis. The use of sequence-based barcoding of parasites will inform on transmission trends, the geographical origin of the isolates, and drug resistance, and thereby assist malaria control and elimination efforts. The study of genetic diversity in *Plasmodium* populations can provide insights into their evolutionary and adaptation history, as well as guide the development of vaccine design and implementation. More widely, the application of 'omics approaches in epidemiological studies with integration across malaria parasites, vectors and hosts has the potential to lead to biological insights to assist the development of new tools for disease control.

REFERENCES

1. Moon, R. W. *et al.* Adaptation of the genetically tractable malaria pathogen Plasmodium knowlesi to continuous culture in human erythrocytes. *Proc. Natl. Acad. Sci.* **110,** 531–536 (2013).
2. Pain, A. *et al.* The genome of the simian and human malaria parasite Plasmodium knowlesi. *Nature* **455,** 799–803 (2008).
3. Ponts, N. *et al.* Genome-wide Mapping of DNA Methylation in the Human Malaria Parasite Plasmodium falciparum. *Cell Host Microbe* **14,** 696–706 (2013).
4. Greer, E. L. *et al.* DNA Methylation on N^6-Adenine in <em>C.&#xa0;elegans</em>. *Cell* **161,** 868–878 (2016).
5. Moon, R. W. *et al.* Normocyte-binding protein required for human erythrocyte invasion by the zoonotic malaria parasite Plasmodium knowlesi. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 7231–7236 (2016).
6. Maurano, M. T. *et al.* Role of DNA Methylation in Modulating Transcription Factor

Occupancy. *Cell Rep.* **12,** 1184–1195 (2015).

7. Ghorbal, M. *et al.* Genome editing in the human malaria parasite Plasmodium falciparum using the CRISPR-Cas9 system. *Nat. Biotechnol.* **32,** 819 (2014).

8. Zhong, D., Koepfli, C., Cui, L. & Yan, G. Molecular approaches to determine the multiplicity of Plasmodium infections. *Malar. J.* **17,** 172 (2018).

9. Divis, P. C. S. *et al.* Three Divergent Subpopulations of the Malaria Parasite Plasmodium knowlesi. *Emerg. Infect. Dis.* **23,** 616–624 (2017).

10. Assefa, S. *et al.* Population genomic structure and adaptation in the zoonotic malaria parasite Plasmodium knowlesi. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 13027–13032 (2015).

11. Oyola, S. O. *et al.* Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. *bioRxiv* (2016). doi:10.1101/067546

12. Cowell, A. N. *et al.* Selective Whole-Genome Amplification Is a Robust Method That Enables Scalable Whole-Genome Sequencing of Plasmodium vivax from Unprocessed Clinical Samples. *MBio* **8,** (2017).

13. Lim, K. L., Amir, A., Lau, Y. L. & Fong, M. Y. The Duffy binding protein (PkDBPαII) of Plasmodium knowlesi from Peninsular Malaysia and Malaysian Borneo show different binding activity level to human erythrocytes. *Malar. J.* **16,** 331 (2017).

14. Salanti, A. *et al.* Evidence for the Involvement of VAR2CSA in Pregnancy-associated Malaria. *J. Exp. Med.* **200,** 1197–1203 (2004).

15. European Vaccine Initiative. http://www.euvaccine.eu/portfolio/project-index/placmalvac.

16. European Vaccine Initiative. http://www.euvaccine.eu/portfolio/project-index/primalvac.

17. Patel, J. C. *et al.* Increased risk of low birth weight in women with placental malaria associated with P. falciparum VAR2CSA clade. *Sci. Rep.* **7,** 7768 (2017).

18. Mwangi, H. N., Wagacha, P., Mathenge, P., Sijenyi, F. & Mulaa, F. Structure of the 40S ribosomal subunit of Plasmodium falciparum by homology and de novo modeling. *Acta Pharm. Sin. B* **7,** 97–105 (2017).

19. Phelan, J. *et al.* Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.* **14,** 31 (2016).

20. Howes, R. E. *et al.* Global Epidemiology of Plasmodium vivax. *Am. J. Trop. Med. Hyg.* **95,** 15–34 (2016).

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

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

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

24. Baniecki, M. L. *et al.* Development of a Single Nucleotide Polymorphism Barcode to Genotype Plasmodium vivax Infections. *PLoS Negl. Trop. Dis.* **9,** e0003539 (2015).

25. Nag, S. *et al.* High throughput resistance profiling of Plasmodium falciparum infections based on custom dual indexing and Illumina next generation sequencing-technology. *Sci. Rep.* **7,** (2017).

26. WHO. *World Malaria Report 2017*. (2017).

27. Bousema, T., Kreuels, B. & Gosling, R. Adjusting for Heterogeneity of Malaria Transmission in Longitudinal Studies. *J. Infect. Dis.* **204,** 1–3 (2011).

28. Bousema, T. *et al.* The Impact of Hotspot-Targeted Interventions on Malaria Transmission in Rachuonyo South District in the Western Kenyan Highlands: A Cluster-Randomized Controlled Trial. *PLoS Med.* **13,** e1001993 (2016).

29. Tusting, L. S., Bousema, T., Smith, D. L. & Drakeley, C. Measuring changes in Plasmodium falciparum transmission: Precision, accuracy and costs of metrics. *Adv. Parasitol.* **84,** 151–208 (2014).

30. World Health Organization. *Eliminating malaria from the Greater Mekong subregion*. (2015).

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

32. Bopp, S. *et al.* Plasmepsin II–III copy number accounts for bimodal piperaquine resistance among Cambodian Plasmodium falciparum. *Nat. Commun.* **9,** 1769 (2018).

33. Messenger, L. A. & Miles, M. A. Evidence and importance of genetic exchange among field populations of Trypanosoma cruzi. *Acta Trop.* **151,** 150–155 (2015).