Sequence analysis of the merozoite surface protein 7 (PvMSP-7) multigene family: vaccine candidates for *Plasmodium vivax*

Thesis submitted in accordance with the requirements of the University of Liverpool and Chulalongkorn University for the degree of Doctor in Philosophy by

Chew Weng Cheng

September 2018

Acknowledgements

For the completion of this thesis, I would like to express my heartfelt gratitude to my supervisors Andrew Jackson, Somchai Jongwutiwes, and Chaturong Putaporntip. Their kindness, motivation, and patient supports were enormous throughout my journey. They not only guided me in my research but also shaped my scientific reasoning skills and prepared me for the competitive world. This bulk of thesis would not be great without the scientific discussions from Steve Edward, Steve Paterson, and Britta Urban

This thesis would not have been possible without the sequencing facility at the Centre of Genomics Research (CGR) Liverpool to generate the sequence data for my research. Special thanks to Richard Gregory for all the bioinformatics support and solutions. My thanks to the medical officers and lab technologists from the malaria-endemic areas to collect patient samples on my behalf. I would like to particularly thank Urassaya Pattanawong for her expertly technical assistance and company during my research in Thailand. I would like to extend my sincere appreciation to Chulalongkorn University for providing the PhD scholarship under the 100th Anniversary Chulalongkorn University Fund for Doctoral Scholarship. I would also like to thank the 90th Anniversary Chulalongkorn University Fund (Ratchadaphiseksompotch Endowment Fund), Ratchadapiseksompotch Fund for Faculty of Medicine, and Department of Infection Biology for the research expenses. Two years research at the University of Liverpool were supported by Andrew Jackson and Overseas Research Fund of Chulalongkorn University.

Thanks to my fellow lab mates from the Department of Infection Biology, University of Liverpool and Department of Parasitology, Chulalongkorn University who shared their invaluable knowledge and helped me during my study. To everyone in the Department of Infection Biology, thank you all for the laughs during lunch hour and crazy social events. My lunch hour will never again be the same.

A million thanks to my parents whose are so supportive and always done so much for me through the endeavour. Lastly, to everyone who made this journey possible I am deeply grateful for your contributions.

Sequence analysis of the merozoite surface protein 7 (PvMSP-7) multigene family: vaccine candidates for *Plasmodium vivax*

Chew Weng Cheng

Abstract

P. vivax is found predominantly in Asia and the emergence of resistance to antimalarial drug and insecticides are major challenges to control of vivax malaria. Control is further complicated by the dormant liver stage of *P. vivax*, which produces an asymptomatic parasite reservoir. Malaria vaccine development is recognised as the most efficient control intervention globally. Currently, malaria vaccine development is concentrated in *P. falciparum* and RTS,S has been the most promising vaccine. This has encouraged similar initiatives to develop a *P. vivax* vaccine.

The subject of this thesis is the *Plasmodium vivax* merozoite surface protein 7 (PvMSP-7), which forms a multi-protein complex with other merozoite surface proteins and plays a principal role in erythrocyte invasion. Studies in *P. falciparum* have shown that targeting MSP-7 can impair erythrocyte invasion and regulate disease severity. For this reason, PvMSP-7 is a plausible vaccine candidate. However, several questions remain to be addressed before a vaccine can be developed; (i) the antigenic variation and expression pattern of the PvMSP-7 multigene family, (ii) which PvMSP-7 paralog is the most promising target, and (iii) which domain of the protein is most antigenically relevant. The main aim of this thesis is to characterise the structural and expression variation in PvMSP-7 paralogs in the Thai clinical setting, to pinpoint the optimal PvMSP-7 antigen for malaria vaccine development.

In Chapter 2, molecular diversity among 20 *P. vivax* clinical isolates from three malaria endemic areas in Thailand is assessed on a whole genome basis to establish parasite population structure in Thailand and its implications for the vaccine design. In Chapter 3, PvMSP-7 genetic diversity is examined among genomic data from Thai clinical isolates, showing that paralogs vary in their patterns of sequence variation. In Chapter 4, PvMSP-7E is evaluated as a specific *P. vivax* genetic marker, showing that genetic diversity is determined by protein secondary structure. In chapter 5, transcriptomic analysis of clinical RNA samples was used to determine the transcriptional profiles of all PvMSP-7 paralogs in diverse patients, showing that specific PvMSP-7 paralogs are expressed constitutively throughout the bloodstream infection cycle. In Chapter 6, serum from naturally infected patients was used to screen a high-density peptide microarray and identify immunogenic B-cell epitopes, showing that 14 universal epitopes belong to just six PvMSP-7 genes.

The conclusion of the thesis is that, of the 13 PvMSP-7 paralogs, PvMSP-7A is the most promising malaria vaccine candidate, being least polymorphic across parasite populations, expressed throughout the bloodstream infection cycle, and containing the greatest number of immunogenic B-cell epitopes. These immunogenic epitopes may confer a two-fold advantage in eliciting immunity and by impairing host cell invasion. This thesis provides a basis for the development of PvMSP-7A as an experimental vaccine, leading to the sustainable prevention of vivax malaria across the world.

List of publications

Paper I

Chew Weng Cheng, Chaturong Putaporntip, Somchai Jongwutiwes (2018) Polymorphism in merozoite surface protein-7E of *Plasmodium vivax* in Thailand: Natural selection related to protein secondary structure. PLoS ONE 13(5): e0196765. https://doi.org/10.1371/journal.pone.0196765

Paper II

Chew Weng Cheng, Somchai Jongwutiwes, Chaturong Putaporntip, Andrew P. Jackson (2018) Clinical expression and antigenic profiles of a *Plasmodium vivax* vaccine candidate: merozoite surface protein 7 (PvMSP-7). Malaria Journal. (*peer-review*)

Table of contents

Table of figures

Table of tables

Abbreviations

ACTs	Artemisinin-combination therapies
AMA-1	Apical membrane antigen 1
BAM	Binary Alignment Map
CPM	Count per millions
CSP	Circumsporozoite protein
CSV	comma separated values
CV	Cross-validation
DBPII	Duffy binding protein-II
DEG	Differentially expressed gene
EMP-1	Erythrocyte membrane protein 1
FDR	False discovery rate
FST	Fixation index
GATK	Genome Analysis Toolkit
GFP	Green fluorescent protein
GLURP	Glutamate-rich protein
IDC	Intraerythrocytic development cycle
IDR	intrinsically unstructured/disordered regions
IFNγ	Interferon-γ
IL4	Interleukin-4

LD	Linkage disequilibrium
LSA1	Liver-stage antigen 1
mdr1	Multidrug resistance 1
MSP	Merozoite surface protein
P. falciparum	*Plasmodium falciparum*
P. vivax	*Plasmodium vivax*
PCA	Principal component analysis
PCR	Polymerase chain reaction
PvMSP-7	*Plasmodium vivax* merozoite surface protein 7
r	Pearson's correlation coefficient
RAS	Radiated attenuated sporozoites
RBP	Reticulocyte-binding proteins
RH5	Reticulocyte binding-like protein 5
RIFIN	Repetitive interspersed genes
RNA-seq	RNA sequencing
RON	Rhoptry neck protein
SAM	Sequence Alignment Map
SERA	Serine repeat antigen
SNP	Single nucleotide polymorphism
SP	Sulfadoxine-pyrimethamine

SUB	Subtilisin-like protease
TRAG	Tryptophan-rich protein
TRAP	Thrombospondin-related anonymous protein
VCF	Variant Call Format
VIR	Variant interspersed repeat
WGS	Whole genome sequencing
WHO	World Health Organization
μL	Microliter
π	Nucleotide diversity

Chapter 1

General introduction

1.1 Malaria

Malaria is one of the most destructive tropical diseases caused by *Plasmodium* parasites. In 2016, it claimed 445 thousand lives and accounted for 216 million morbidities (WHO, 2017). There are four species of human malaria parasites, *P. falciparum*, *P. vivax*, *P. ovale*, and *P. malariae*. Ovale malaria can be further divided into two closely related but distinct species, *P. ovale curtisi* and *P. ovale wallikeri*. Recently, *P. knowlesi* has been identified as a zoonotic malaria parasite that impacts significantly on human populations (Jongwutiwes *et al*., 2011). The most prevalent *Plasmodium* species are *P. falciparum* and *P. vivax*, the latter being the most widespread species outside Africa. The *P. vivax* transmission in Africa is extremely low due to the lack of Duffy gene expression in erythroid cells among the population. The Duffy gene acts as a receptor for *P. vivax* merozoites to invade host erythrocytes (Miller *et al.*, 1976). Climatic factors are playing a pivotal role in malaria transmission in which the rainfall patterns, temperature, and humidity provide a nourishing breeding ground for *Anopheles* mosquitoes (Bi *et al.*, 2013; Lingala, 2017). For this reason, *P. vivax* has high transmission throughout South and Southeast Asia (Guerra *et al.*, 2006), Central and South America (Gething *et al.*, 2012). *P. vivax* infection was not a priority for malaria elimination in the past until a substantial risk of mortality was reported (Baird, 2013; Douglas *et al.*, 2014; Price *et al.*, 2009). *P. vivax* infection produces symptoms such as, shaking chills, headache, muscle aches, febrile paroxysms and higher proinflammatory cytokine levels (Hemmer *et al.*, 2006) which subsequently lead to severe outcomes. After infection with *P. vivax*, a proportion of sporozoites will develop into dormant forms in hepatocytes, known as hypnozoites (Hulden and Hulden, 2011). The hypnozoite stage in *P. vivax* causes multiple, unpredictable symptomatic episodes over many months, and possibly years (Robinson *et al.*, 2015). This biological characteristic has further complicated the case detection and eradication strategy. Emerging of drug-resistant strain in *P. vivax* has underscored the importance of vivax

malaria. Chloroquine was once the first-line treatment against *P. vivax* infection until 1989, where the first resistance case was noted (Rieckmann *et al.*, 1989). Subsequently, a number of antimalarial drug resistance have emerged from most malaria-endemic countries including, sulfadoxine-pyrimethamine (Imwong *et al.*, 2001) and mefloquine (Alecrim *et al.*, 1999). As a result of the widespread antimalarial drug resistance and increased global burden of *P. vivax*, a more effective approach is sought to control the transmission dynamics. Developing a malaria vaccine is considered an effective mode of the control strategy. The complexity of the parasite's life cycle has challenged the development of a universally effective vaccine. However, a subunit malaria vaccine is considered an essential part of the ideal control strategy, targeting the parasite circulating in endemic areas. Therefore, vaccine development is an avenue to reduce the morbidity and mortality.

1.2 History

Malaria is one of the world ancient diseases that was recorded more than 4,000 years ago. The first record was described in Chinese medical writings in 2,700 BC during the Huang Ti dynasty. Subsequently, a similar record was found by Greek, Roman, Assyrian, Indian, and Arabic writings. During the 4th century BCE, malaria occurred widely in Greece, where it affected many human populations. The first extensive reference to malaria was given by Hippocrates of Kos, who attributed malarial disease to the fumes originating from swamps. From that perspective, the "bad air" gave the disease its name: mal'aria in Italian.

The discovery of the malaria parasite begun with a French army surgeon (Charles Louis Alphonse Laveran) where he noticed the parasites in the blood of a patient who succumbed to malaria in the year 1880. He observed moving filaments under the light microscope, which today is believed to be exflagellation of a male gametocyte. The discovery led him to the Nobel Prize award in 1907 (Haas, 1999). By 1886, malaria was suggested to have multiple parasite species. Patients infected with malaria seemed to have varied symptoms, tertian malaria (48-hour periodicity) and quartan malaria (72-hour periodicity) (Cox, 2010). Camillo Golgi who discovered the malaria species was awarded the Nobel Prize in 1906. Malaria was generally accepted

to cause by a protozoan parasite in 1890. The first two human malaria parasites, *P. vivax* and *P. malariae* were introduced by two Italian scientists, Giovanni Batista Grassi and Raimando Felletti. *P. falciparum* was later named by an American scientist, William H. Welch in 1897 as the malignant tertian malaria parasite. Two malaria parasites, *P. ovale* and *P. knowlesi* were later introduced in 1922 and 1931, respectively. In 1897, Ronald Ross demonstrated that the malaria parasites were transmitted from infected host to mosquitoes. Ross carried out extensive investigations to identify mosquitoes as parasite vectors. For this discovery, he was awarded a Nobel Prize in 1902. During 1898 to 1899, sporogony life stages of three malaria parasites were described (*P. falciparum*, *P. vivax*, and *P. malariae*). By the late 19th century, malaria developmental stages were known to the community.

1.3 Global burden of malaria

As of 2016, 216 million cases of malaria were reported (WHO, 2017), largely in the African region (90%), followed by the Southeast Asian region (7%), and Eastern Mediterranean region (2%) (Figure 1.1), and an estimated 445 thousand deaths occurred due to the disease, 91% of these in Africa. Malaria remains endemic in 91 countries, however, the incidence rate is reported to have decreased by 18%, from 76 to 63 cases per 1000 populations (WHO, 2017). Out of 91 countries, 44 countries reported lower malaria incidence (< 10,000 cases) due to malaria elimination programs (WHO, 2017). Malaria transmission in the Southeast Asia region has seen the largest decline (48%). Two countries have been certified by WHO as malaria-free in 2016; Kyrgyzstan and Sri Lanka. In 2016, Figure 1.1 shows the status of malaria transmission in the respective country. The data presented in the world report could have underestimated the true magnitude of the disease due to conservative estimates by WHO. The discrepancy may contribute to the insufficient diagnostic facilities in the endemic areas.

Two most prevalent malaria species, *P. falciparum,* and *P. vivax* continue to account for most of the malaria incidence. In 2016, 99% of the malaria cases in Africa were contributed by *P. falciparum*. Outside Africa, *P. vivax* is the most prevalent species circulating in endemic areas. About 64% of cases were caused by *P. vivax* in the Americas region, followed by 40% and 30% in the Eastern Mediterranean and

Southeast Asia regions, respectively. *P. vivax* transmission is known to occur at a very low frequency in African populations due to Duffy-negativity. Contrary to this established scientific knowledge, recent studies have discovered the *P. vivax* infection in most parts of Africa (Gunalan *et al.*, 2018; Mendes *et al.*, 2011). This phenomenon raises the possibility of *P. vivax* adapted to infect Duffy-negative populations. If so, then *P. vivax* incidence would be expected to increase in the coming years. Therefore, studying *P. vivax* infection should now become a priority, as the frequency of *P. vivax* infection will hamper the malaria elimination strategy.

In Thailand, malaria cases recorded a declined between 2010 and 2016, from 32,480 cases to 11,522 cases (35% reduction). Malaria transmission occurs primarily along the international borders with Burma (Myanmar), Cambodia, and Malaysia. Factors affected malaria prevalence here are the forest fringe areas of these provinces, population movements, and the emergence of antimalarial drug resistance. Moreover, the expansion of rubber plantations in Thailand has led to sporadic cases over the past decade. Conflict in southern Thailand has challenged the control strategy, and a major malaria outbreak occurred in 2016. Figure 1.2 shows the provinces with malaria transmission in Thailand. Like other endemic areas, malaria burden in Thailand is due to the two most prevalent *Plasmodium* species, *P. falciparum,* and *P. vivax*. Contrary to other regions, *P. vivax* seems to dominate as a cause of malaria in Thailand since 2010 (>50% of cases contributed by *P. vivax*) (Figure 1.3).

An endemic area refers to a region in which malaria transmission is still active (Hay *et al.*, 2008). The definition can be extended to define the intensity of exposure. 'Holoendemic' describes an area with the perennial intense transmission, whereas 'hyperendemic' refers to an area with seasonally intense malaria transmission. 'Mesoendemic' is an area with the transmission that fluctuates with changes due to multiple local conditions, and 'hypoendemic' often refers to settings with low transmission and where effects are not significant. Endemicity is associated with host-vector interactions, population movements, antimalarial drug resistance, insecticides resistance, and the local demographic.

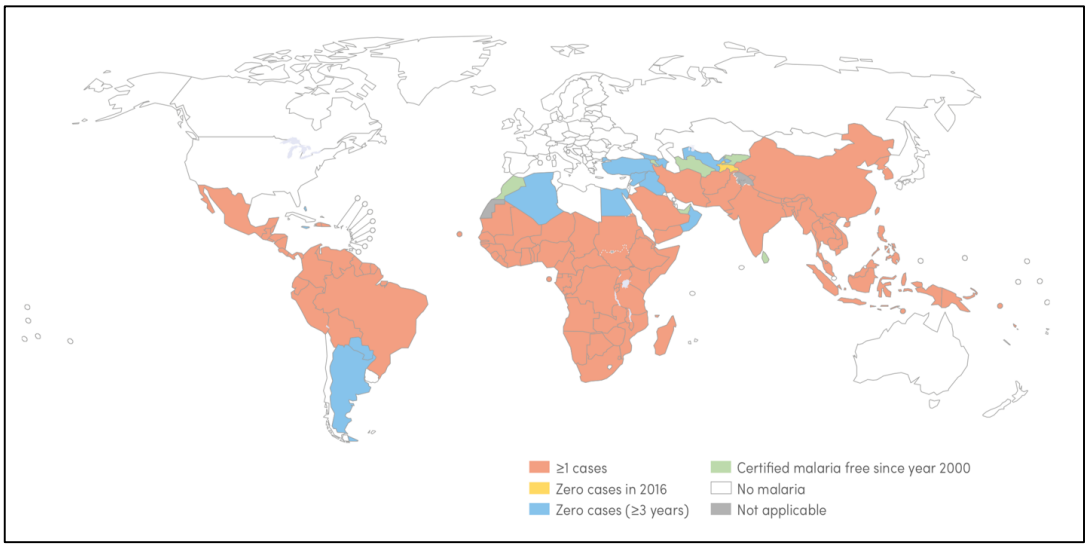


Figure 1.1 Countries with ongoing malaria transmission in 2016. Countries categorized as no malaria transmission (white), countries with zero malaria transmission over the past three years (sky blue), countries with zero malaria transmission in 2016 (yellow), countries certified with malaria free since year 2000 (light green), and countries with ongoing malaria transmission (blush). It is apparent that malaria transmission is still present in the central and south America, south and southeast Asia, and Africa. Source: World malaria report 2017 (WHO, 2017).

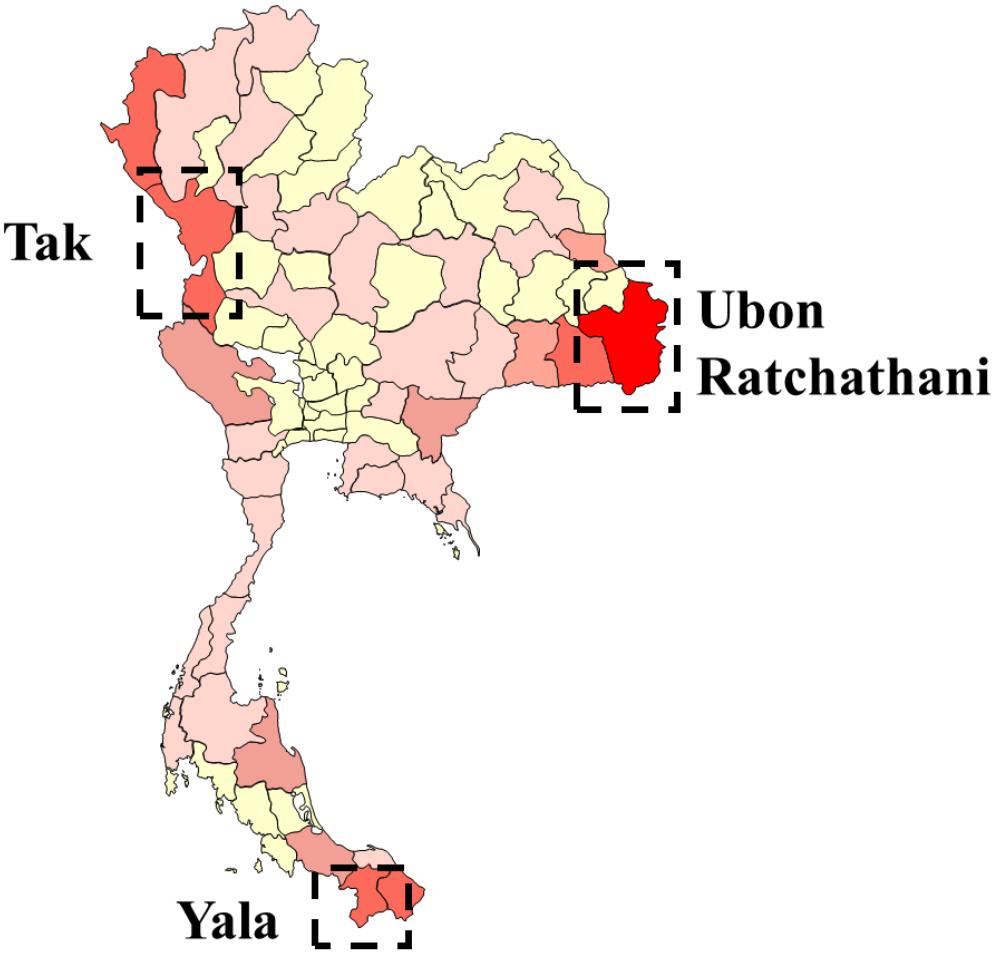


Figure 1.2 Provinces with ongoing malaria transmission in Thailand. Provinces with malaria transmission are coloured in pink. The provinces with darker shade indicates higher malaria transmission. From the map, most malaria endemic areas are along the international borders such as, Burma (Myanmar), Cambodia, and Malaysia. These areas are dense forest, reported antimalarial drug resistance, and populations movement between two countries. Three endemic areas (dashed boxes), Tak province, Ubon Ratchathani province, and Yala province are focused in this study.

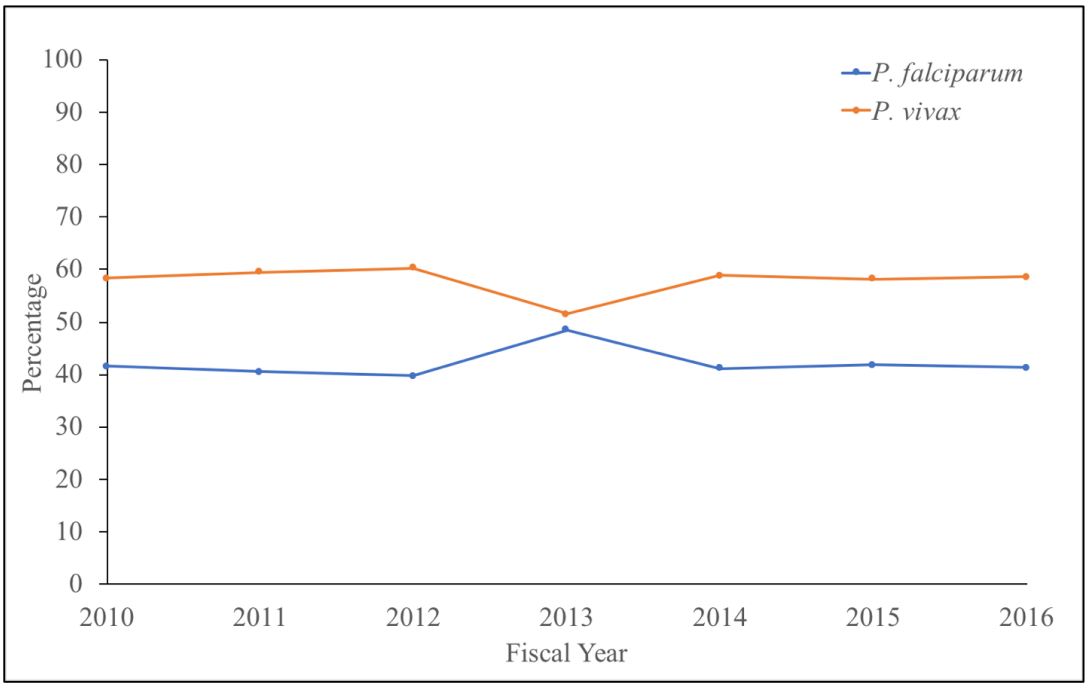


Figure 1.3 Proportion of malaria parasites species in Thailand from year 2010 to 2016. The number of cases between *P. falciparum* and *P. vivax* were expressed in percentage on the y-axis and fiscal year on the x-axis. The number of cases were adapted from the World malaria report 2017 (WHO, 2017). The malaria infections in Thailand were largely dominated by *P. vivax* since 2010.

1.4 Mosquito vectors

Anopheles mosquitoes are the main vector of malaria parasites. There are currently more than 3,500 species of mosquitoes in 41 genera known to transmit *Plasmodium* spp. (Ferraguti *et al.*, 2016). Of these, there are only about 70 *Anopheles* species responsible for human malaria transmission. Each of the species has its own breeding preferences, generally in water. The vector has four developmental stages, egg, larva, pupa, and adult. The first three developmental stages are aquatic and take approximately 5-14 days to complete. The cycle varies according to the *Anopheles* species and temperature. The female adult mosquitoes have a lifespan of around 1-2 weeks and require a blood meal for the development of eggs. There are four main factors in order to establish successful malaria transmission between a host and a mosquito including, abundance, longevity, capacity, and contact with humans. For the first criterion, the number of mosquitoes circulating in the area must be high enough to obtain a blood meal from an infected patient. Second, mosquitoes must survive a sufficiently long time after a blood meal. Longer survival allows the parasites to develop to the infective stages and travel to the salivary glands for transmission. Moreover, each mosquito should carry enough parasites in the salivary glands to ensure successful infection. Lastly, transmission is greatest when the mosquito breeding sites are closest to human homes.

Anopheles species differ from one malaria area to another, due to their biology and ecology, adaptation, and epidemiological patterns. *An. gambiae, An. fenestus*, *An. moucheti, An. nili* are four main species transmitting malaria in sub-Saharan Africa. In Thailand, *An. dirus*, *An. minimus*, and *An. maculatus* are the main vectors circulating in endemic areas of Thailand (Suwonkerd *et al.*, 2013). Environmental factors, human activities, and climate seasonality are three principle elements contributing to malaria transmission in Thailand. The malaria outbreak that occurred in Ubon Ratchathani province during 2016 was largely due to illegal logging by local people and migrants from other regions.

1.5 Clinical manifestation

The pathogenesis of malaria parasites is complex and the relapsing manifestation ranges from asymptomatic infection to acute disease and a chronic febrile disease. Periodic paroxysms are the most notable symptom of malaria infection. After the infection, the patient will undergo an incubation period of about a week with no symptoms. Prior to the first febrile attack, patients often experience symptoms like a headache, anorexia, myalgia, abdominal pain, cough, diarrhoea, restlessness, delirium, and anaemia (Malik *et al.*, 1998). The symptoms may last for 48 hours coinciding with the duration of the asexual developmental period. The first paroxysm can be divided into three stages, the cold stage, hot stage, and sweating stage. During the cold stage, patients will experience an intense cold coupled with vigorous shivering that lasts at least 15 minutes. The hot stage symptoms resemble extremely burning sensation up to six hours. The sweating stage is when patients sweat profusely and drop in body temperature for at least two hours. The parasites exert a profound effect upon the infected red cells and escape the vascular system. This characteristic explains how severe infections develop at significantly low levels of parasitaemia. In addition, accumulation of multiple clinical relapses also contributes to the severity of anaemia with low parasitaemia. Although vivax malaria is generally considered to be benign, severe complications have recently been recognized which include severe anaemia (<5 mg haemoglobin/dL), severe thrombocytopenia, acute pulmonary oedema, jaundice, splenic rupture, acute renal failure and rarely, cerebral malaria, and shock.

Moreover, the clinical presentations under low transmission settings of *P. vivax* malaria in children vary depending upon age and usually cannot be differentiated easily from other infectious complications (Anstey *et al*., 2012). Commonly, irrespective of age, symptoms presenting are fever, chills, and headache. However, in newborns, fever could be the only prominent symptom. Rupture of erythrocytes by mature schizonts stimulates the release of several inflammatory cytokines, which result in fever and myalgia. During initial infection, fever might be irregular. Upon synchronous red blood cells rupture, it leads to a typical cyclic fever (Stanley, 1997). If *P. vivax* malaria is left untreated during pregnancy, it causes severe anaemia in the mothers and can result in spontaneous abortion and intrauterine retardation of fetal growth.

1.6 Life cycle

Plasmodium parasites have complex developmental stages that involve switching between an asexual reproduction in a vertebrate host and a sexual reproduction in the mosquito (Figure 1.4). Through the developmental phase, the parasite transforms into multiple, distinctive morphological forms (rings, trophozoites, schizonts, merozoites, and gametocytes). *Plasmodium* parasites are haploid in both vertebrate host and mosquito, except a brief diploid period inside the mosquito midgut where gametocytes undergo sexual reproduction to a zygote. The asexual life cycle in vertebrate host begins with the entry of parasites from a blood meal, followed by replication inside hepatocytes, erythrocyte invasion, replication inside erythrocytes, and egress from erythrocytes either to transform into gametocytes or re-invade naïve erythrocytes. These transitions can be divided into three stages: pre-erythrocytic, intra-erythrocytic, and post-erythrocytic.

The developmental phase begins with a bite of an infected female *Anopheles* mosquito, where 100-125 infective sporozoites are inoculated into the subcutaneous tissue (Aly *et al*., 2009). Upon entry into the bloodstream, sporozoites take about one to three hours to leave the injected site. Sporozoites travel through the capillaries and invade hepatocytes. At this stage, patients infected with malaria parasites will not show any clinical manifestations. This phase is known as a productive invasion. Within the hepatocytes, the sporozoites will transform into the trophozoite stage and mature into a round schizont. The schizogony takes 47-52 hours to complete and releases 1,500-8,000 merozoites (Aly *et al*., 2009). After maturation, merozoites release from the disintegrated hepatocytes in membrane-bound vesicles called merosomes. The merosomes squeeze out of the liver and release merozoites into the bloodstream. These invasive merozoites are ready to infect naïve red blood cells within 30 minutes and correspond with parasitemia. However, in *P. vivax* and *P. ovale*, the parasites may enter a dormant stage in the liver (known as hypnozoite stage). These parasites are capable of inducing relapse infection after months or even years. This phenotype has made the study of *P. vivax* invasion a challenge, something exacerbated by the lack of a continuous culturing system.

Upon entry into erythrocytes, the parasite resides within the parasitophorous vacuole and feed on the haemoglobin. It will transform into a ring-stage form, then

progress to a larger, trophozoite form. The parasites feed on haemoglobin and modify red blood cell membrane to adhere to the uninfected red blood cells and the endothelium of blood vessels. Trophozoites are the most active feeding stage, they usually appear to be larger, rounded, and some of them assemble as membranous sacs (Maurer's clefts). Continuous feeding on haemoglobin generates a by-product known as hemozoin crystals. These distinctive pigments often define the trophozoite stage. The schizont stage is marked by nuclear divisions and the expression of proteins critical for erythrocyte invasion. At this developmental stage, nuclear divisions often form 16 nuclei. Multiple divisions simultaneously generate about 8-24 mature merozoites in *P. falciparum,* whilst 12-24 mature merozoites in *P. vivax*. Rupturing of the schizont releases merozoites into the circulation to invade new red blood cells. Merozoite invasion is central to the intra-erythrocytic stage. It sustains the parasite life cycle and malaria pathogenesis. In addition, the intra-erythrocytic stage is immunologically important due to the exposure of merozoites to the host immune system. This makes merozoite an attractive target for vaccine candidates.

After the release of merozoites into blood circulation, a small proportion of them will sexually differentiate into gametocytes (male: microgametocytes, female: macrogametocytes). These gametocyte cells are the precursor for male and female gametes. Mature gametocytes are transmitted into the anopheline mosquito during a blood meal. Male gametocytes differentiate to produce eight sperm-like gametocytes, whereas the female gametocytes generate a single and spherical macrogamete. Subsequently, a diploid zygote is formed in the fly midgut by fertilization of male and female gametes and developed into an ookinete. The motile ookinete penetrates the midgut epithelium and differentiates into an oocyst. After ten to thirteen days stretching the basal lamina overlying the oocyst, each rupture oocyst will release thousands of haploid sporozoites, which invade the salivary gland. The infective sporozoites are ready to infect vertebrate hosts during the next mosquito bite.

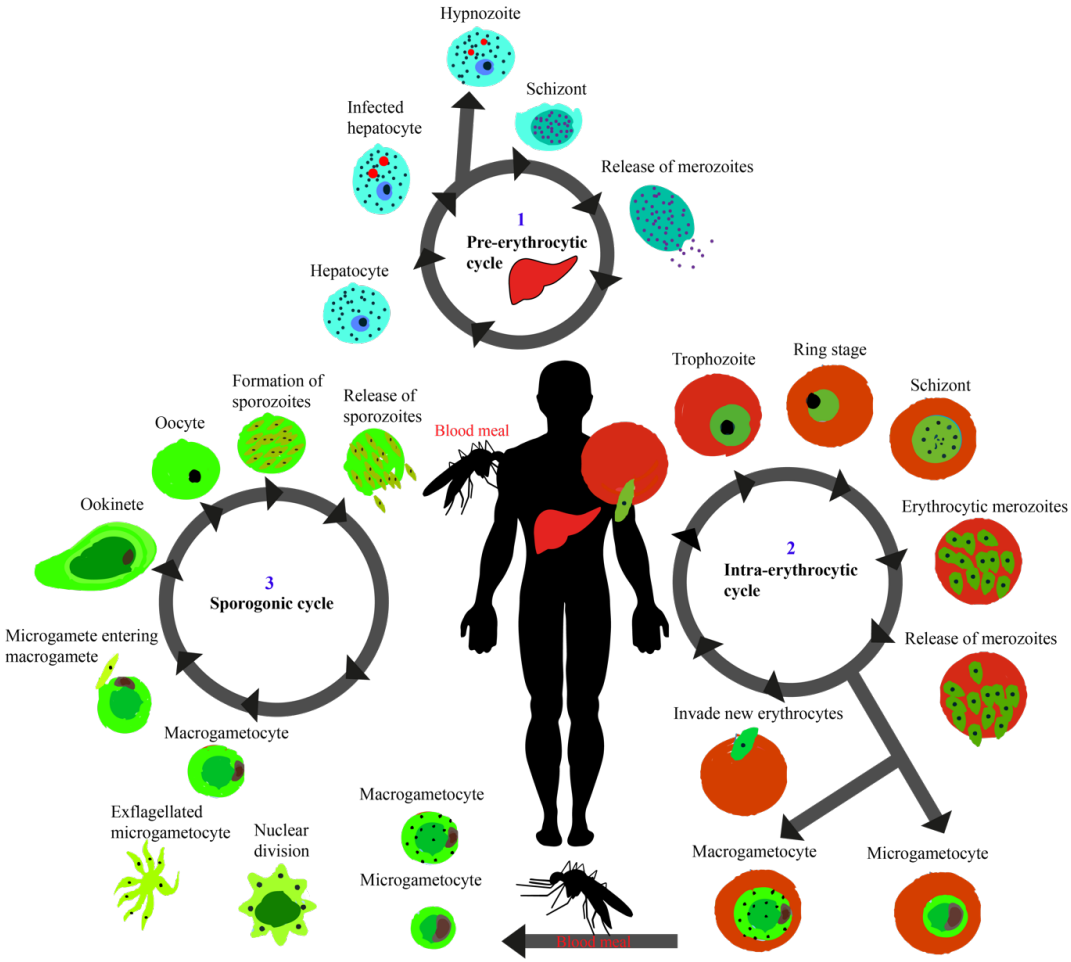


Figure 1.4 Life cycle of *Plasmodium* vivax. The life cycle can be divided into three cycles, pre-erythrocytic (blue), intra-erythrocytic cycle (red), and sporogony cycle (green). Two hosts are required to complete the life cycle, primary host (female Anopheles mosquito) and secondary host (human).

1.7 Prevention and control of malaria

Vector control strategies include the use of insecticide-treated bed nets (ITNs) with pyrethroids and indoor insecticide spray. This approach has been effective in reducing the global burden of malaria transmission. However, the widespread resistance of mosquitoes to insecticides is a major concern. It has been estimated that, should the mosquitoes become resistant to pyrethroid, more than 100,000 additional deaths would be expected (Mulamba *et al.*, 2014). Meanwhile, at the parasite intervention level, chemotherapeutic management is the most common approach against malaria. Vaccination is an attractive and sustainable malaria prevention strategy; however, it has yet to be achieved. There are tremendous research efforts being conducted to develop potential vaccine candidates, discussed in the next section.

Chloroquine, a group of 4-amino-quinoline, has been the first line treatment for *P. vivax* since 1946. However, there are reports about the emerging of chloroquine drug resistance in Papua New Guinea (Yung and Bennett, 1976), Indonesia (Baird *et al.*, 1991a) , Myanmar (Guthmann *et al.*, 2008), India (Shah *et al.*, 2011), and South America (Gonçalves *et al.*, 2014). Recently, reduced sensitivity of chloroquine against *P. vivax* was also reported in Ethiopia (Abreha *et al.*, 2017). Chloroquine is no longer in used for *P. falciparum* because resistance has developed in most endemic areas. In the case of mixed species infections, artemether-lumefantrine, which has the schizonticidal efficacy, is always the preferred antimalarial drug. However, this drug is not generally used outside Africa. As chloroquine and artemether-lumefantrine are not able to target liver stage infection, primaquine is the primary drug for *P. vivax*. Primaquine is a hypnozoiticide which targets the parasite in the liver to prevent the episode of relapse infection. However, the use of primaquine can have severe side effects. It causes severe hemolysis in glucose-6-phosphate-dehydrogenase deficient (G6PD) patients, pregnant women, and others hypersensitive to 4-aminoquinoline compounds.

Combination therapy is a regular practice in managing malaria. The drug combinations provide better efficacy and delay parasite replication which channel to a longer therapeutic life of monotherapy. Artemisinin-combination therapies (ACTs) is currently the frontline therapy against *P. falciparum*. Artemisinin contains a sesquiterpene lactone with an endoperoxide bridge can target malaria parasites within

minutes (Sun and Zhou, 2016). Quinine and quinidine are two active compounds that have a synergistic effect with artemisinin. The formulation is recommended against severe malaria. The use of artemisinin is not confined to *P. falciparum*; some studies reported efficacy in treating *P. vivax* infection (Karunajeewa *et al.*, 2008; Phan *et al.*, 2002). The study of artemisinin in Vietnam among vivax-infected patients showed a significant parasite clearance following the treatment. Other ACTs used to target *P. falciparum* include artesunate-amodiaquine, artesunate-sulfadoxine-pyrimethamine, artesunate-chlorgunanil-dapsone, and artesunate-pyronaridine.

Combination therapy has been the standard regimen for managing malaria in Thailand. Sulfadoxine-pyrimethamine (SP) was reported in deteriorating efficacy along the Thai-Cambodia border, and thereby no longer recommended for treatment (Thimasarn *et al.*, 1997). Mefloquine, a 4-aminoquinoline-methanol was used as a single therapy against *P. falciparum* in Thailand until the emergence of the resistant strain. Three days dihydroartemisinin-piperaquine is deployed across the country for uncomplicated *P. falciparum* malaria. Efficacy of chloroquine in Thailand was reported to have declined especially along the western Thailand border.

Studies revealed that drug resistance is associated with mutations of the multidrug resistance 1 (*mdr1*) gene. Y976F mutation in *mdr1* of *P. vivax* is linked to chloroquine resistance, where the multiple gene copies generated with susceptibility to chloroquine resistance (Golassa *et al.*, 2015). The *mdr1* of *P. vivax* is more prevalent along the western Thailand border, which leads to chloroquine resistance in Thailand (Imwong *et al.*, 2008). Besides that, there are other alternative antimalarial drugs that are sensitive to *P. vivax* such as, rifampicin, artesunate, sulfadoxine-pyrimethamine, artesunate-amodiaquine, and artesunate-pyronaridine (Chu and White, 2016). In conclusion, widespread drug-resistant strains may compromise the malaria control and research efforts should also focus on alternative approaches, such as the characterization of malaria vaccine candidates and the underlying biology to develop effective malaria control tools in the future.

1.8 Merozoite binding mechanisms

As discussed above, during the *Plasmodium* life cycle merozoites invade new erythrocytes once they emerge from infected erythrocytes. The interactions between merozoites and erythrocytes are therefore essential for malaria vaccine design. Merozoite recognition and invasion of erythrocytes involve multiple complex steps before a successful invasion is established. It involves attachment, reorientation, and invagination of the merozoite during the host-cell invasion (Figure 1.5). The invasion begins when the merozoites and erythrocytes establish primary contact at any point on the surface (Cowman and Crabb, 2006). This initial attachment is mediated by merozoite surface proteins that coat the outer surface of the parasite (Lin *et al.*, 2016). Subsequently, the parasite activates essential invasion organelles including, rhoptries, micronemes, and dense granules. These invasion organelles then release their contents periodically at the entry point. The rhoptries and micronemes act as a storage pocket for the merozoite proteins during schizogony and transported to the merozoite surface soon after merozoite egress from schizont. These two apical organelles have a higher binding capability to the receptors on erythrocyte surface which critical for invasion mechanism. The rhoptries are essential for host cell modification whereas micronemes are important for host cell adhesion and rupture (Kats *et al.*, 2008). Following attachment of merozoite to the erythrocyte, apical reorientation takes place to position its end adjacent to the erythrocyte membrane. Then, micronemes and rhoptries discharge their contents and an irreversible tight or moving junction forms between the apical end of the invading merozoite and target erythrocyte. The tight junction coordinates the connection between erythrocyte membrane, parasite, and actomyosin motor that drives invasion (Giovannini *et al.*, 2011). Upon contact with erythrocytes, thickening of the erythrocyte membrane is the first sight of junction formation. The entire invasion system is powered by an actin-myosin contractile system within the parasite itself.

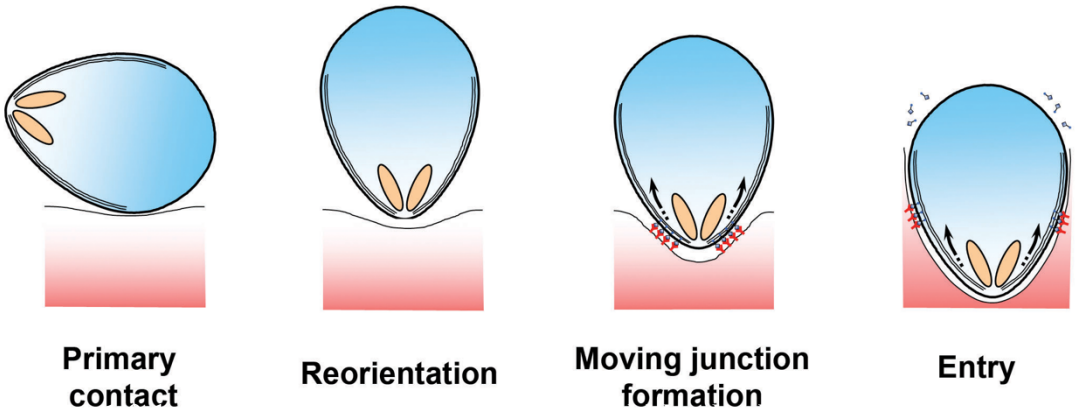


Figure 1.5 Merozoite invasion into host red blood cells. The merozoite invasion into the erythrocytes involve multiple steps including primary contact, reorientation, moving junction, and entry into host-cell. The diagram is available from (Wright and Rayner, 2014).

1.9 Vaccine development

Various antigens from different malaria life stages have been expressed as potential vaccine candidates. Although most vaccines targeting liver-stage indicates a positive direction against malaria distribution, immunity studies showed just partial efficacy in clinical trials. Multistage malaria vaccine candidates likely to offer an attractive perspective for malaria vaccine development targeting more than one life stage. Based on the malaria parasite life cycle, this approach is ideally to prevent initiation of infection, suppress clinical manifestations, and transmission of the disease (Boes *et al.*, 2015). To achieve this objective, key antigens from pre-erythrocytic, blood, and sexual antigens should be identified. Pre-erythrocytic stage vaccines target sporozoites and ultimately disturb parasite development before the symptomatic intra-erythrocytic stage ensues. Despite being asymptomatic in the infected patients, the parasite may be attacked to interrupt the initiation of infection. These vaccines aim to block the sporozoites invasion into hepatocytes. Blood stage vaccine development faces challenges owing to the antigenic variation on the merozoite surface proteins and infected erythrocyte surface proteins. The aim of blood-stage malaria vaccine

development is to inhibit the merozoite invasion into erythrocytes and suppress malaria clinical symptoms. Antigens discovery in blood stage should give priority to the functionally conserved regions to confer universally immune responses. Transmission-blocking vaccines aim at disrupting the transmission of the parasite from human to mosquito. The vaccines against the sexual stage of the parasites are aimed to prevent parasites from infecting the female *Anopheles* mosquito through a blood meal. Most of the vaccine developments are focusing on the *P. falciparum*. Identification of new antigens and evaluation *in vitro* remain elusive in *P. vivax*. In this section, antigen discovery in the three life stages of *Plasmodium* will be discussed, while focusing mainly on the intra-erythrocyte cycle and merozoite surface proteins of *P. vivax*.

1.9.1 Pre-erythrocytic stage vaccine

Currently, RTS,S/AS01 is a vaccine being developed for use against the deadliest species of *Plasmodium*, *P. falciparum*. The construction of RTS,S is based on the carboxy-terminal of the *P. falciparum* circumsporozoite protein (CSP) and formulated with an AS01 adjuvant system (Cohen *et al.*, 2010). The vaccine targets the parasite in its pre-erythrocytic stage, and that shows protective immunity among young children aged between 5-17 months (Rts, 2015). However, recent large-scale Phase III clinical trial introduced some doubt on its efficacy by suggesting that RTS,S/AS01 may not confer protection against severe malaria in infants aged between 6-12 weeks (Lell *et al.*, 2009). The vaccine efficacy declined from 27.0% to 18.3% after several months in follow-up studies (Gosling and Seidlein, 2016). Although RTS,S/AS01 did not perform as expected, it could still potentially be used to immunize other age groups. Besides RTS,S, there are also other novel vaccine candidates identified in the pre-erythrocytic stage including, Thrombospondin-related anonymous protein (TRAP) (Gantt *et al.*, 2000), liver-stage antigen 1 (LSA1) (Kurtis *et al.*, 2001), liver-stage antigen 3 (LSA3) (Brahimi *et al.*, 2001), and early transcribed membrane protein 5 (ETRAMP5) (Fontaine *et al.*, 2010).

Vaccines based on recombinant CSP have been tested in Phase III clinical trials in *P. falciparum*. The results from the trials were encouraging, in terms of protective immune responses and well tolerated in clinical patients (RTS, 2011, 2012). Moreover,

a portion of the CSP against *P. vivax* is currently in phase 1/2a vaccine trial (Bennett *et al.*, 2016). This vaccine candidate consists of the amino- and carboxy- fragments of the CSP and a truncated repeat region from two strains of parasites (VK210 and VK247) (Bennett *et al.*, 2016; Cheng *et al.*, 2013). Although the vaccine did not induce sterile protection, the parasite growth was significantly slower (Bennett *et al.*, 2016). TRAP is another promising vaccine candidate that plays a prime role for sporozoite gliding on the mosquito salivary gland (Duffy *et al.*, 2012; Kosuwin *et al.*, 2018; Long and Hoffman, 2002). Both of these proteins are essential for the sporozoite invasion (Steinbuechel and Matuschewski, 2009; Sultan *et al.*, 1997) and so vaccine design using these CSP and TRAP should essentially block the hepatocyte invasion by sporozoites.

1.9.2 Blood-stage vaccine

Although pre-erythrocytic stage vaccines have seen major progress in development, blood-stage vaccine candidates have made rather slower progress towards clinical testing. Vaccines that target blood-stages aim to prevent the development of clinical symptoms and impair parasite growth. Numerous blood-stage antigens have been studied, some with the potential to be a vaccine candidate (Miura, 2016). The merozoite is an attractive target for vaccine development because it is free in circulation for a brief period before entering another red blood cell and so vulnerable to destruction by antibodies. The initial release of merozoites into circulation can stimulate the immune memory of vaccinated individuals, leading to a high antibody production that offsets the further development of the blood infection at subsequent cycles (Carvalho *et al.*, 2002). Merozoite surface protein 1 (MSP-1), apical membrane antigen 1 (AMA1), and merozoite surface protein 3 (MSP-3) are the broadly studied surface antigens in blood-stage *Plasmodium*.

MSP-1 is involved in erythrocyte invasion and is a leading blood-stage vaccine candidate. MSP-1 undergoes secondary processing and immunization with a 42 kDa fragment is capable of reducing parasitemia (Singh *et al.*, 2006), although recent Phase 2b clinical trials of MSP-1 failed to protect children against malaria infection in Kenya despite having high antibodies titer (Ogutu *et al.*, 2009). Lyon and colleagues in 2008 showed that MSP-1 42 kDa fragment delivered in Freund's adjuvant derived strain-

specific immunity leading to reduce vaccine efficacy (Lyon *et al.*, 2008). It is likely that the strain-specific immunity has been influenced by the adjuvant system. Adjuvants have been used to enhance the level of immune responses to a vaccine where it guides the adaptive immune response (Coffman *et al.*, 2010). Lyon and colleagues suggested that a more appropriate adjuvant should be used to further assess the MSP-1 42 kDa fragment immunity in human (Lyon *et al.*, 2008). In addition, MSP-1 is a highly polymorphic antigen, multiple copies of the fragments are thought to be present in a recombinant vaccine (Draper *et al.*, 2009). An animal model constructed using MSP-1 Block 2 showed to induce immunogenic responses in all Block 2 serotypes (Cowan *et al.*, 2011). The N-terminal of the MSP-1 is known as Block 2 which encompasses most of the variants (Cowan *et al.*, 2011). The MSP-1 vaccine with a cocktail of polymorphic variants in Block 2 and conserved sequence in Block 1 was shown to elicit protective immune responses in African populations (Cowan *et al.*, 2011). Despite extensive polymorphism in the MSP-1 Block 2, the antigenic variants could still elicit protective immunity couples with the humoral responses confer by T-cell epitopes on the conserved Block 1 (Cowan *et al.*, 2011; Parra *et al.*, 2000).

AMA-1 is another promising surface antigen expressed during the intra-erythrocytic stage (Mitchell *et al.*, 2004). It plays an essential role in the invasion of the host cells through establishing the moving junction with the merozoite (Richard *et al.*, 2010). A phase I clinical trial of AMA-1 using AS02A adjuvant was conducted in Mali (Thera *et al.*, 2010) and North America (Thera *et al.*, 2008). The outcomes of both studies were encouraging, where it elicits high immunogenicity and well tolerated in natural infections (Thera *et al.*, 2008; Thera *et al.*, 2010). However, sequence diversity of AMA-1 is of similar magnitude as MSP-1 (Osier *et al.*, 2008), and therefore, protection may be similarly undermined by strain-specific polymorphism. Vaccine development based on this protein might require incorporating multi-copies of the AMA-1 in order to confer protection against all strains.

MSP-3 is a multigene family consisting of eight members in *P. falciparum* (Singh *et al.*, 2009) and 12 paralogs in *P. vivax* (Carlton *et al.*, 2008). The expansion of MSP-3 in *P. vivax* suggests a mechanism of immune evasion. Various studies have been conducted to investigate antibodies responses to MSP3, but protection efficacy remains to be examined in *P. vivax* at least. The immunogenicity of two MSP-3 paralogs in *P.*

vivax, MSP-3α and MSP-3β were tested in Brazilian populations. The recombinant proteins were highly immunogenic in natural infections administered via different adjuvants (Bitencourt *et al.*, 2013). This result is encouraging for vaccine development, even though the immune responses of other MSP-3 paralogs remain to be characterised. MSP-3 in *P. falciparum* has progressed further, a Phase 1b clinical trial was conducted in West African populations (Sirima *et al.*, 2011). The MSP-3 vaccine was substantially safe in the trial where the incidence rate was below two in 100 days. In addition, the trial highlights high antibody responses conferred by the MSP-3 vaccine and some indication of protection efficacy, despite only 45 children recruited for the study. The immune responses induced by MSP-3 in *P. falciparum* and *P. vivax* are consistent, therefore, the findings warrant further investigation as a subunit vaccine candidate in blood-stages.

Thus far, most vaccine candidates in blood-stages have shown encouraging immune responses, and protective efficacy in experimental models or initial clinical trials. The blood-stage vaccine components discussed above are not exhausted, there are further surface antigens that warrant further investigation. The MSP-7 multigene family, the subject of this thesis, are another possibility for blood-stage vaccine development, owing to their role in erythrocyte invasion and interaction with MSP-1 (Cheng *et al.*, 2018; Garzón-Ospina *et al.*, 2010; Garzón-Ospina *et al.*, 2016, 2014; Kadekoppala and Holder, 2010; Tewari *et al.*, 2005). A more comprehensive coverage of the MSP-7 multigene family will follow in the next section.

1.9.3 Sexual-stage vaccine

Blocking the malaria transmission at the mosquito stage is another approach to vaccination. This approach uses gametocyte or sexual-stage antigens to prevent transmission of parasites from host to mosquito. Some of the sexual-stage proteins exploited in transmission-blocking vaccine candidates include the gametocyte surface protein *Pfs* 230 (Eksi *et al.*, 2006), the ookinete protein *Pfs* 25/28 (Saxena *et al.*, 2007), and *Pfs* 48/45, a six-cysteine protein family found on the gamete cell surface and involved in gamete interaction within the mosquito gut (Dijk *et al*., 2011). *Pfs* 230 is an important transmission-blocking vaccine candidate in *P. falciparum*. It has shown to

inhibit the development of oocysts (Krause *et al.*, 2007). The inhibition of oocysts formation in the mosquito midgut will retard the development of thousands of sporozoites, so impairs the parasite transmission from mosquito to human (Hill, 2011). Six fragments of *Pfs* 230 were expressed as recombinant proteins in *E. coli.* Four out of six fragments elicited antisera that reduced *P. falciparum* infectivity to mosquitoes (Williamson *et al.*, 1995). *Pfs* 25/28 are two conserved vaccine candidates that showed to elicit immune responses in natural infection (Duffy and Kaslow, 1997). In humans, immunization with *Pfs* 25 demonstrated inhibition of sporozoites transmission from mosquito to human (Wu *et al.*, 2008). Immunization of rodents and primates with *Pfs* 48/45 inhibited oocyst formation by up to 95% after challenge with *P. falciparum* compared to controls (Chowdhury *et al.*, 2009).

From the multistage vaccine development perspective, transmission-blocking vaccines are capable of extending the life of other malaria vaccines by stopping the spread of the parasites. In any case, the antigen conformations and the vaccine efficacy need to be investigated further. The main hurdle to the further development of transmission-blocking vaccines has been the effectiveness of vaccine distribution. Every individual in malaria-endemic areas are likely to transmit the parasites, therefore, all residents within a community should be vaccinated. This translates into a mass vaccination investment, which has proven to be financially challenging. Since malaria transmission is a local and a focal feature of the landscape, deployment of transmission-blocking vaccine may be more sensible at a smaller local community in combination with other approaches.

1.10 Vaccine technology

Over the past few decades, many different vaccine formulations for malaria have been explored. Vaccine technologies can be divided into three categories, attenuated microbes, killed microbes or protein subunits (Hill, 2011). Vaccines that use attenuated microbes are the most successful form of a vaccine (Coelho *et al.*, 2017). These contain a weakened form of the microbe that protects against a cross-reactive pathogen. Such vaccines have been shown to elicit protective immune responses in infectious diseases such as human immunodeficiency virus (Blower *et al.*, 2001), influenza virus

(Bournazos and Ravetch, 2017), and smallpox (Minor, 2015). In malaria, three live attenuated malaria vaccines have been tested in clinical trials using RTS,S/AS01 (Keitany *et al.*, 2014). The parasites were attenuated by irradiation, drug coverage, and genetic attenuation. Attenuation by irradiation was the first approach against malaria parasites, in the form of radiation-attenuated sporozoites (RAS). Hoffmann and colleagues immunised 11 patients with RAS, of the 26 challenges, 24 were shown to induce protective immunity lasting up to 42 weeks (Hoffman *et al.*, 2002). In addition, immunization with a larger dose of cryopreserved sporozoites was shown to protect up to 60% of patients (Doll and Harty, 2014). Although the results seemed encouraging, this approach is laborious and expensive and so was superseded by other approaches.

Parasite attenuation by drug coverage using sporozoites and anti-malarial drug chloroquine inhibited the parasite intra-erythrocytic cycle (Keitany *et al.*, 2014). Immunological studies have shown that immune responses derived from this approach conferred long-lasting protection in four of six patients (Bijker *et al.*, 2013). To further verify the vaccine efficacy, more volunteers should be enrolled in the testing to determine the protection against sporozoite challenge. Genetic attenuation involves deletion of specific genes in *Plasmodium* sporozoites that precludes parasite development in the liver stage (Aly *et al.*, 2009; Tarun *et al.*, 2007). The deletion of genes including, UIS3, UIS4, p52, and sap1 were able to elicit long-lasting protection against sporozoite challenge in mice model. However, one out of six patients showed blood stage parasitemia after the second dose of injection, suggesting the mutant was not entirely attenuated (Aly *et al.*, 2009). Despite the disadvantages of each method, it would be interesting to compare the vaccine efficacy using the three strategies.

Lack of success using vaccines based on attenuated parasites has diverted attention to the development of killed-whole parasite formulations. The use of whole infectious pathogens was previously integrated into various vaccine designs, including tuberculosis, mumps, and rubella. Vaccine development using whole-killed parasite was seen to confer protection against blood-stage infection in malaria (Zepp, 2010). This approach covers a broad array of antigens exposed to the immune system and may overcome the limitation on vaccine efficacy caused by the antigenic polymorphism. A whole-killed parasite vaccine can be generated with both high temperature or chemicals. Challenge experiments in rodents and primates vaccinated with killed-whole

parasite can confer protective immunity (McCarthy and Good, 2010). In studies where monkeys were immunized with red cell lysate of three different *Plasmodium* species, *P. knowlesi* (Jiang *et al.*, 2009), *P. falciparum* (Butler *et al.*, 2012), and *P. yoelii* (Hagen *et al*., 1993), the monkeys were protected against parasite challenge. These results showed positive immunization of primates with killed-whole parasite during blood-stages. However, similar experiments are yet to be conducted in humans as a compatible adjuvant is still to be explored. Moreover, the cost and logistics to deliver the whole-killed parasite vaccine in endemic areas are concerns for this vaccine technology.

Subunit vaccine is another approach for vaccine technology. An example is the extensive studied RTS,S even though it provides sub-optimal protection (Kaslow and Biernaux, 2015). It is likely that the RTS,S together with other protein subunits can induce sterile protection. Synthetic peptides and chimeric protein vaccines are two examples of malaria subunit vaccines. Production of synthetic peptides in *P. falciparum* has been a challenge because it is difficult to produce epitopes on the protein surface in the correct conformation. A long synthetic peptide is likely to address the problem. However, the efficacy is unlikely to be superior to that of whole-killed parasite unless the conserved epitopes are recognized by the immune system. In addition, subunit malaria vaccines need to be highly reactive and required high titer to induce sterile protection (Cohen *et al.*, 2010). As for recombinant protein subunit vaccine, the type of adjuvant used in conjunction is crucial, as shown in RTS,S/AS01 (Didierlaurent *et al.*, 2017). The nature of adjuvant can enhance the immunogenicity of protein antigens. RTS,S/AS01 consists a number of virus-like particle delivery systems combined with repeat sequences and a C-terminal fragment of the circumsporozoite protein. These virus-like particle delivery systems have restricted size and only partial malaria sequences are used, which may compromise the vaccine efficacy (Draper *et al.*, 2015).

Another vaccine candidate that uses the similar approach is *Pfs*25. It is a transmission-blocking vaccine composed of protein-protein conjugates. Long-lasting protection was demonstrated using conjugates with aluminium hydroxide (Kubler-Kielb *et al.*, 2010). Additionally, similar protective immune responses were observed for dimeric *Pfs*25 conjugated to circumsporozoite protein repeat (Kubler-Kielb *et al.*, 2007). These observations show that the subunit vaccine that can inhibit both sporozoite replication in hepatocytes and transmission between host and vector. In parallel with

efforts to develop subunit vaccines against the pre-erythrocytic stage and sporogonic cycle, the intra-erythrocytic stage should be fully explored.

Subunit vaccines targeting the multiprotein complex formed during erythrocyte invasion is a new avenue of investigation, potentially able to elicit highly protective immune responses in humans. MSP-1 forms a multiprotein complex with MSP-6 and MSP-7 prior to erythrocyte invasion (Kauth *et al.*, 2006). Using antibodies against MSP-1/6/7 can prevent merozoite invasion by shedding of the multiprotein complex (Woehlbier *et al.*, 2010). Therefore, developing subunit vaccines that contain fragments of MSP-1/6/7 may provide an efficient vaccine. As multi-copy proteins, careful consideration of structural and functional diversity among paralogs, such as the antigen conformations, polymorphic regions, and immunogenicity, is necessary during vaccine design with these blood-stage antigens. Having discussed different vaccine technologies aimed at various life stages of malaria parasites, it is clear that each of the methods has its own advantages and disadvantages, and in fact, vaccine development should achieve a balance of immune effector roles to block replication of parasites at different life stages.

1.11 *Plasmodium vivax* merozoite surface protein 7 (PvMSP-7)

1.11.1 Molecular evolution

MSP-7 is a surface protein expressed by *Plasmodium* merozoites. The MSP-7 and MSP-7 related protein (MSRP) genes were first discovered in the *P. falciparum* merozoite surface as a 22-kDa ($MSP\text{-}7_{22}$) fragment (Pachebat *et al.*, 2001). The fragment binds non-covalently to the MSP-1 complex. Subsequently, the second fragment with 19-kDa ($MSP\text{-}7_{19}$) was reported to be derived from $MSP\text{-}7_{22}$ through a proteolytic event (Figure 1.6) MSP-7 is differentially expanded in *Plasmodium* genus, *P. vivax* has the highest copy number (Figure 1.7). The multigene family consists of 13 paralogs in *P. vivax,* nine paralogs in *P. falciparum*, five paralogs in *P. knowlesi*, four paralogs in *P. berghei*, four paralogs in *P. yoelii*, four paralogs in *P. chabaudi*, and seven paralogs in *P. reichenowi* (Figure 1.7) (Garzón-Ospina *et al.*, 2010). The

extensive copy number variation found across human and rodent malaria species suggests species-specific duplications or deletions. Evolutionary studies found several MSP-7 members undergo recombination events that led to the generation of new sequences (Castillo *et al.*, 2017; Garzón-Ospina *et al.*, 2010; Garzón-Ospina *et al.*, 2016). The evolution of the MSP-7 family appears to follow a birth-and-death model, where major events such as duplication, pseudogenizations, and gene loss occur frequently (Castillo *et al.*, 2017; Garzón-Ospina *et al.*, 2010).

As the MSP-7 family is conserved across the *Plasmodium* genus, this suggests that the proteins are playing multiple, essential roles. However, it is currently unknown about their exact function(s). Population genetic studies have shown that several *P. vivax* MSP-7 paralogs are under purifying selection, including PvMSP-7A, -7E, -7H, -7I, -7K, and -7L (Castillo *et al.*, 2017; Cheng *et al.*, 2018; Garzón-Ospina *et al.*, 2016, 2014; Garzón-Ospina *et al.*, 2012; Garzón-Ospina *et al.*, 2011). MSP-7 genes between *Plasmodium* genus have similar structure and organization, although they have a relatively low identity (Kadekoppala and Holder, 2010; Mongui *et al.*, 2006). Sequence similarity of MSP-7 genes between *P. falciparum* and *P. vivax* by Mongui *et al.* (2006) revealed that the proportion of sequence identity ranges between 9.9% to 41.8%. However, the further analysis focused on the C-terminal of MSP-7 in all *Plasmodium* species showed that this domain has remained highly conserved, suggesting this fragment could be especially important for protein function (Castillo *et al.*, 2017; Cheng *et al.*, 2018; Kadekoppala and Holder, 2010). Kadekoppala *et al.* (2008) showed that the C-terminal region of PfMSP-7 has high binding activity, and suggested that it might be implicated in host cell invasion. Furthermore, experimental knock-down of PfMSP-7's C-terminal produced a significant reduction in parasite invasion into erythrocytes (Kadekoppala *et al.*, 2008). The central region of most MSP-7 paralogs, especially in *P. vivax,* is highly polymorphic, suggesting that balancing selection is acting to maintain diversity. A closer look at MSP-7 proteins in *P. vivax*, shows that seven members are relatively conserved (PvMSP-7A, -7D, -7F, -7J, -7K, -7L, and -7M). Taken all these findings together, it is possible that MSP-7 paralogs are functionally differentiated and the family collectively performs multiple functions.

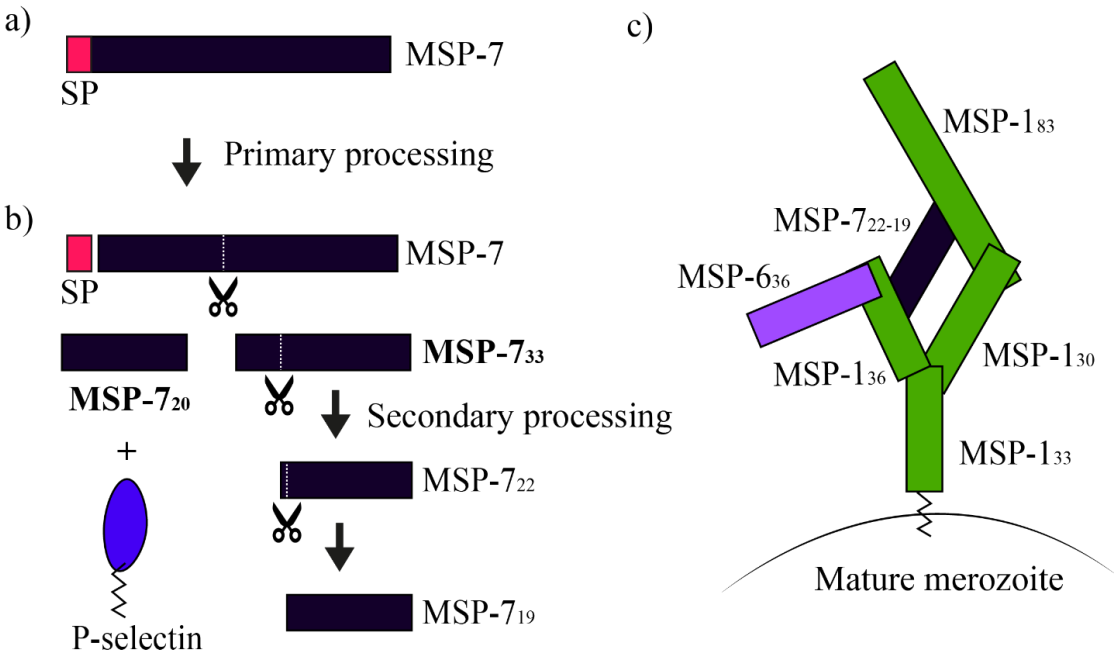


Figure 1.6 Schematic diagram of MSP-7 proteolytic events and a multiprotein complex. SP denotes signal peptide. **a)** MSP-7 was synthesized as a 48-kDa precursor in *P. falciparum*. **b)** MSP-7 undergoes two proteolytic events. The first proteolysis generates two protein fragments; 20-kDa fragment from the N-terminal and 33-kDa fragment from the C-terminal. The N-terminal fragment (MSP-7_{20}) interacts with P-selectin which modulates the disease severity (Perrin *et al.*, 2015). The C-terminal fragment (MSP-7_{33}) undergoes secondary proteolysis to generate a 22-kDa fragment and further cleave to generate 19-kDa fragment. These two fragments participate in the host-cell invasion with other merozoite surface antigens (Pachebat *et al.*, 2001). **c)** a multiprotein complex is formed between MSP-1 (83-, 36-, 33-, and 30-kDa), 36-kDa of MSP-6, and MSP-7 (22- and 19-kDa). This complex is thought to involve in the erythrocyte invasion.

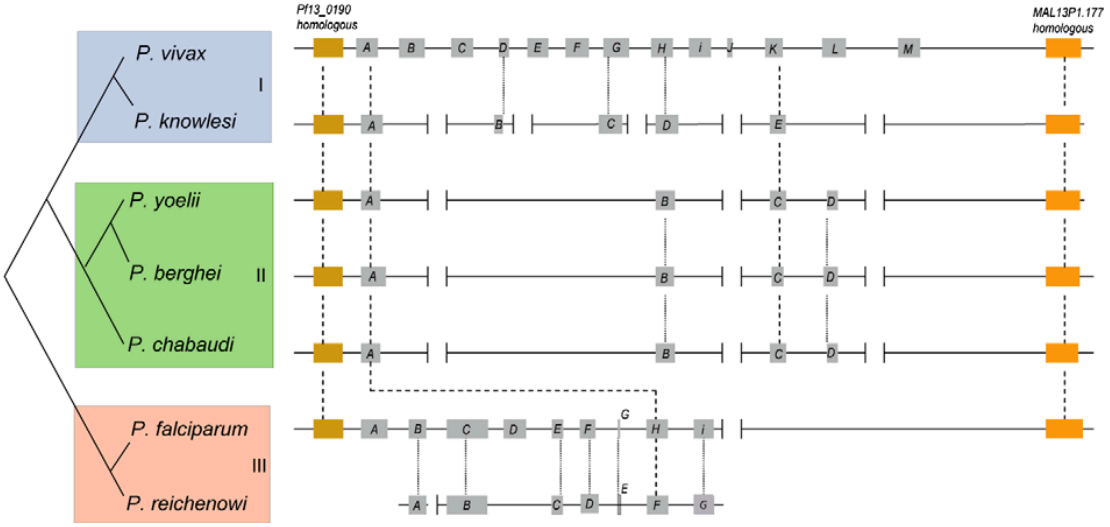


Figure 1.7 Schematic diagram of MSP-7 copy number in seven *Plasmodium* species. MSP-7 genes are arranged alphabetically from head to tail on a chromosome. Brown coloured boxes represent the regions flanking the MSP-7 multigene family array in each *Plasmodium* species. Each grey box depicts an MSP-7 gene, whilst the dotted lines connect genes with an orthologous relationship in phylogenies. The gap along *P. yoelii, P. berghei*, and *P. chaubadi* were introduced to allow gene positioning. The diagram is taken from (Garzón-Ospina *et al.*, 2010).

1.11.2 Proteolytic processing of MSP-7

MSP-7 is synthesized as a 48 kDa precursor undergoes two proteolytic cleavages, necessary for correct function (Figure 1.6). These proteolytic events are proposed to be essential for merozoite development and maturation (Pachebat *et al.*, 2007). In the primary proteolytic event, MSP-7 yields two protein fragments, 20-kDa (MSP-7_{20}, N-terminal) and 33-kDa (MSP-7_{33}, C-terminal) (Figure 1.6). Upon maturation of schizonts, MSP-7_{20} appears to be degraded on the merozoite surface. The 33-kDa fragment located at the C-terminal undergoes secondary proteolysis, generating the 22-kDa fragment and further cleaved to yield a 19-kDa fragment (MSP-7_{19}). The MSP-7_{22} fragment is tightly associated with MSP-1 multiprotein complex (Pachebat *et al.*, 2007). Other leading vaccine candidates including MSP-1 and AMA-1 undergo similar proteolytic processing like MSP-7, but only a small part of the C-terminal participates in the host-cell invasion (Blackman *et al.*, 1994; Urquiza *et al.*, 1996).

Upon erythrocyte invasion, MSP-7 found on the MSP-1 multiprotein complex as a 22- or 19-kDa fragment. This suggests MSP-7 interacts with the MSP-1 complex and the sequential proteolytic processing coincident with merozoite development and maturation. MSP-7 is non-covalently associated with the merozoite surface, forming complexes with other MSPs (Figure 1.6) (Cowman *et al.*, 2002). MSP-1 forms a protein-complex with MSP-6_{36}, MSP-7_{20}, and MSP-7_{19} which plays a role in the initial parasite-erythrocyte interaction (Kauth *et al.*, 2006). This large multiprotein complex interacts via the processed form of the C-terminal region. Based on the localization of MSP-7 on the merozoite surface, it is thought that the C-terminal interacts with host-cell through Band 3 (Garcia *et al.*, 2007). This complex modulates the erythrocyte invasion mechanism and is shed from the parasite surface following entry into the host cell. The fact that, not all parts of the MSP-7 are interacting with the MSP-1 multiprotein complex, vaccine design should, therefore, focus on the C-terminal of the protein.

1.11.3 The role of MSP-7

Antibodies targeting the multiprotein complex consisting of MSP-1, -6, and -7 were shown to interfere with the shedding of MSP-1 and prevent host-cell invasion in *P. falciparum* (Woehlbier *et al.*, 2010). Moreover, antibodies against different areas of this multiprotein complex revealed interference with the development of the parasite *in vitro* and shedding of the complex (Kauth *et al.*, 2006; Woehlbier *et al.*, 2006). However, PfMSP-7 knock-out parasites seem to survive and invade host-cell *in vivo* and *in vitro*, suggesting interruption of PfMSP-7 alone is not adequate to impede the invasion mechanism (Kadekoppala *et al.*, 2008; Woehlbier *et al.*, 2010). Meanwhile, deletion of a PfMSP-7 paralog (PF3D7_1335100) involved in the multiprotein complex led to a reduction in parasite's ability to invade host red blood cells by at least 20% *in vitro* (Kadekoppala *et al.*, 2008). Kadekoppala and colleagues also showed knocked-down of further five PfMSP-7 paralogs generated a null phenotype (Kadekoppala *et al.*, 2008). Likewise, deletion of MSP-7 in *P. berghei* revealed the impairment of parasite growth and predominantly reticulocytes preference (Tewari *et al.*, 2005). Furthermore, Kauth and colleagues (2006) reported the ability of MSP-1/6/7 in inducing strong humoral responses in falciparum-infected patients. Rabbit antibodies raised against PfMSP-6 and PfMSP-7 demonstrated the potential to inhibit parasite replication *in vitro*. Such polyclonal antibodies against the multiprotein complex imply the potential relevance of each component in vaccine development.

On the other hand, antibodies targeting the PfMSP-6_{36} and PfMSP-7_{22} showed the ability to interrupt secondary proteolytic processing of PfMSP-1 (Woehlbier *et al.*, 2010). The secondary proteolytic event is thought to be a precursor for forming the multiprotein complex and priming the invasion competent merozoites. Consequently, interrupting the secondary proteolytic event in MSP-1 could impede the primary step in the shedding of the multiprotein complex and inhibit the parasite's maturation pathway. Similarly, two MSP-7 paralogs in *P. yoelii* showed their ability to interact with the C-terminal of PyMSP-1 in a yeast two-hybrid system. One of those PyMSP-7 paralogs was the homologue of PfMSP-7 that previously isolated in the shed complex of MSP-1 (Mello *et al.*, 2004). All lines of evidence indicate that the MSP-7 paralogous genes have important roles in the invasion process.

Functional knock-out of certain MSP-7 paralogs in *P. berghei* and *P. falciparum* did not completely impair the parasite's invasion ability (Kadekoppala *et al.*, 2008; Tewari *et al.*, 2005). Likewise, only certain MSP-7 paralogs are found to interact with the MSP-1 multiprotein complex in *P. falciparum* and *P. yoelii*. This might suggest that not all MSP-7 paralogs participate in the red cell invasion mechanism (Mello *et al.*, 2004). Evidence from rodent infections indicated the immunomodulatory role of MSP-7. Infection with PbMSP-7 knock-out parasites resulted in a significantly lower death rate in mouse models (Mello *et al.*, 2004). Moreover, PbMSP-7 knock-out in laboratory-adapted parasite strain induced cerebral malaria in experimentally infected mice, but a wild-type strain did not (Spaccapelo *et al.*, 2011). MSP-7 paralogs in *P. falciparum* and *P. berghei* have also been shown to act as immunomodulators in regulating disease severity (Perrin *et al.*, 2015). P-selectin has been characterized as a host factor for mediating malaria-associated pathology (Combes *et al.*, 2004), and it interacts with PfMSP-7 the N-terminal region and the P-selectin C-type lectin and EGF-like domains (Figure 1.6) (Perrin *et al.*, 2015). The N-terminal region of PfMSP-7 was reported to be undetectable in the multiprotein complex, whilst only the C-terminal region participated in the erythrocyte invasion. This finding implies different fragments of MSP-7 have different biological functions. Interestingly, the same interaction was observed in PbMSP-7 (Tewari *et al.*, 2005). Therefore, these data suggesting MSP-7 paralogs have diverse and critical roles in addition to erythrocyte invasion.

1.11.4 Population genetics of MSP-7

The patterns of genetic diversity among MSP-7 family members vary between *P. falciparum* and *P. vivax* (Garzón-Ospina *et al*., 2012; Garzón-Ospina *et al*., 2011). The genetic diversity of PfMSP-7 has been reported to be rather conserved, possibly due to the evolutionary forces acting on the *P. falciparum* lineages (Castillo *et al.*, 2017; Garzón-Ospina *et al.*, 2010). However, sequence polymorphism among MSP-7 paralogs in *P. vivax* is more variable, some paralogs show extensive sequence variation, while others are uniform among strains. There are currently 13 MSP-7 genes arranged head-to-tail at chromosome 12 of *P. vivax* (Garzón-Ospina *et al.*, 2010). The PvMSP-7 paralogs are named alphabetically from A-Z (Table 1.1). The respective accession

number of PvMSP-7 in two reference strains are also detailed in the table below (Table 1.1); Salvador I (Carlton *et al.*, 2008) and PvP01 (Auburn *et al.*, 2016).

The population genetic diversity of most PvMSP-7 paralogs (PvMSP-7A, -7C, -7E, -7F, -7H, -7I, -7K, and -7L) has been evaluated in clinical isolates, largely within Colombian population (Cheng *et al.*, 2018; Garzón-Ospina *et al.*, 2014; Garzón-Ospina *et al.*, 2012; Garzón-Ospina *et al.*, 2011). PvMSP-7A, -7F, -7K, and -7L display low polymorphism compared to the others (Garzón-Ospina *et al.*, 2014; Garzón-Ospina *et al.*, 2012; Garzón-Ospina *et al.*, 2011). This shows that the PvMSP-7 family displays a heterogeneous pattern of genetic diversity: some members are highly conserved while the others are highly diverse, perhaps because they are exposed to different selective pressure or possess different biological constraints.

PvMSP-7C, -7H, and -7I were analysed by Garzón-Ospina *et al.* (2012) where 23 haplotypes were detected for PvMSP-7C, while 28 haplotypes were found for both PvMSP-7H and PvMSP-7I. PvMSP-7A and PvMSP-7K are highly conserved (Garzón-Ospina *et al.*, 2011). PvMSP-7A demonstrates little genetic diversity, with only four polymorphic sites, while PvMSP-7K has nine different haplotypes. PvMSP-7F and PvMSP-7L were also reported to have low genetic diversity (Garzón-Ospina *et al.*, 2014). These genes had only four and six segregating sites, respectively. Therefore, if PvMSP-7 paralogs were to use in vaccine development, selection of certain paralogs must consider the sequence polymorphism to avoid immune escape variants and allele-specific immune responses. The conserved proteins are the most attractive vaccine candidates because they have little capacity for vaccine escape.

Table 1.1 The nomenclature reflects each PvMSP-7 gene used in GenBank and PlasmoDB database. The table shows 13 PvMSP-7 genes named alphabetically and their respective accession numbers in Salvador I (Carlton *et al.*, 2008) and PvP01 (Auburn *et al.*, 2016) reference strain.

PvMSP-7	Salvador I	PvP01
A	PVX_082645	PvP01_1220400
B	PVX_082650	PvP01_1220300
C	PVX_082655	PvP01_1220200
D	PVX_082660	PvP01_1220100
E	PVX_082665	PvP01_1220000
F	PVX_082670	PvP01_1219900
G	PVX_082675	PvP01_1219800
H	PVX_082680	PvP01_1219700
I	PVX_082685	PvP01_1219600
J	PVX_082690	PvP01_1219500
K	PVX_082695	PvP01_1219400
L	PVX_082700	PvP01_1219300
M	PVX_082710	PvP01_1219200

1.11.5 MSP-7 transcript expression

While most research relating to MSP-7 has focused on the characterization of MSP-7 functions and antigenic variation, our understanding of MSP-7 gene expression is limited. MSP-7 transcripts were detected in the blood-stages of *P. falciparum* (Kadekoppala *et al.*, 2010; Otto *et al.*, 2010) *, P. yoelii* (Mello *et al.*, 2004; Otto *et al.*, 2014)*, P. berghei* (Otto *et al*., 2014*)*, and *P. vivax* (Bozdech *et al.*, 2008; Mello *et al.*, 2004). Consistently, all MSP-7 paralogs across four *Plasmodium* species showed an increase in transcript expression towards schizogony. The *P. vivax* transcriptome revealed the peak expression of 13 PvMSP-7 paralogs during late-schizont stage (Bozdech *et al.*, 2008). Although there is currently no evidence of PvMSP-7 present on the merozoite surface, the proteins have similarity to that of PfMSP-7 previously characterized on the merozoite surface. Similarly, MSP-7 paralogs in *P. berghei* and *P.*

yoelii were consistently expressed around the schizont stage using RNA sequencing approach (Mello *et al.*, 2004; Otto *et al.*, 2014). Moreover, the PbMSP-7 and PyMSP-7 paralogs were also evidenced by co-localize with the MSP-1 multiprotein complex on the merozoite (Mello *et al.*, 2004; Tewari *et al.*, 2005). Moreover, localization assay using immunofluorescence in *P. vivax* (Mello *et al.*, 2004), *P. yoelii* (Mello *et al.*, 2004), and *P. falciparum* (Kadekoppala *et al.*, 2010) revealed some MSP-7 paralos were found on the surface of merozoites. As previously discussed, only C-terminal of PfMSP-7 was shown to participate in the host-cell invasion and the N-terminal is likely to interact with P-selectin to modulate the disease severity (Perrin *et al.*, 2015). A 174 amino acid fragment derived from the N-terminal of PfMSP-7 was expressed and tagged with green fluorescent protein (GFP).

Current analysis of *P. vivax* transcripts is limited to two studies which used microarray technology (Bozdech *et al.*, 2008; Westenberger *et al.*, 2010). These two studies contributed significantly to the understanding of *P. vivax* transcription. Though, there are several limitations to these studies. Microarray data generated by Bozdech *et al.* (2008) are lacking the transcriptional profile of certain genes not present in Salvador I genome annotation (Carlton *et al.*, 2008). Furthermore, the data produced by Westenberger *et al.* (2010) did not cover the erythrocytic stage of the parasite which essential for understanding the invasion-related transcription. The study of *P. vivax* was hampered by the lack of effective continuous *in vitro* culture system that restricts scientists to dive right into the biology of this species (Noulin *et al.*, 2013). A study conducted by Zhu *et al.* (2016) shows a novel finding to study *P. vivax* transcriptome in clinical isolates. The group sequenced two clinical isolates from the field with asynchronous parasite composition. Intriguingly, the global transcriptome of two clinical isolates was correlated with the microarray-based results. However, the study was based on two clinical isolates which might not provide strong evidence for understanding the transcriptional changes of *P. vivax* in clinical isolates. Taken all these evidence together, RNA-seq was used to characterise the transcript abundance in ten clinical isolates and as a more definitive approach to characterise the transcriptional changes in the PvMSP-7 multigene family.

1.12 *Plasmodium vivax* in Thailand

The tremendous control efforts imposed a few decades back have successfully eliminated malaria from major cities in Thailand. Malaria transmission in Thailand has a unique feature where the transmission areas are separated by a malaria-free corridor in central Thailand. Currently, Thailand is regarded as a malaria-hypoendemic region (Cui *et al.*, 2003). Malaria infections are found along the border of two countries, such as Thailand-Myanmar, Thailand-Cambodia, and Thailand-Malaysia due to forest cover and movement of populations between two countries. Several studies have been conducted to define the spatial and temporal variation of malaria in Thailand. The incidence rate of malaria in Thailand was reported with high spatial heterogeneity, and most cases occurred along the Myanmar and Cambodia. Tak province (Thailand-Myanmar) recorded the highest malaria incidence possibly source from foreign workers, but precise information regarding migratory labourers and patterns of migration are inaccessible (Zhou *et al.*, 2005). *P. vivax* is highly diverse between different geographic regions (Chenet *et al.*, 2012; Hupalo *et al.*, 2016; Jennison *et al.*, 2015; Neafsey *et al.*, 2012) and mixed-clonal infections are common in Thailand. Malaria transmission in Thailand does not conform to a uniform pattern, therefore it is crucial to identify malaria risk areas. A major geographic division of *P. vivax* population structure was seen between western and eastern Thailand in a recent study of global diversity, which in line with the malaria-free region in the central region (Gupta *et al.*, 2016; Pearson *et al.*, 2016). However, this result was based on only four isolates (Pearson *et al.*, 2016).

Population genetic variability of *P. vivax* was also conducted at Mae Sod (Thailand-Myanmar) the level of variability was equally high compared to those from Papua New Guinea, which is a hyperendemic area (Cui *et al.*, 2003). *P. vivax* population bordering Cambodia (Chanthaburi) shows high haplotype and nucleotide diversities, the diversities are similar to those studies conducted at Thailand-Myanmar border (Kosuwin *et al.*, 2014). It is noteworthy that, malaria transmission is always low along the Thailand-Malaysia border (southern Thailand), unfortunately, it reappeared in several areas with sporadic outbreaks. Thus, the low level of polymorphisms in southern Thailand could be shaped by bottleneck effects (Cheng *et al.*, 2018; Jongwutiwes *et al.*, 2010; Kittichai *et al.*, 2017). Improved knowledge of genetic

polymorphisms will provide a clearer picture regarding the transmission dynamics of malaria in Thailand and further strengthen the infection control strategies.

1.13 Immunity to malaria

The mechanism underlying the immune responses in malaria parasite is not fully explained. Unlike viral infection which provides long-lasting immunity upon perhaps just a single infection, malaria patients only acquire immunity gradually. Immunity to malaria is known to i) be acquired gradually after exposure to parasite continuously (Baird, 1995; Cohen, 1979), ii) transfer passively to infants through maternal antimalarial antibodies (Diggs *et al.*, 1995; Dobbs and Dent, 2016), iii) develop partially in response to parasitemia (Schofield and Grau, 2005), and iv) provide levels of protection that correlate with clinical malaria (Kusi *et al.*, 2017). In malaria endemic areas, individuals are semi-immune as they are exposed to malaria parasite over time. This explains the malaria burden contributed mainly by the young children group (Arévalo-Herrera *et al.*, 2016). Furthermore, newborns within the age of three to six months are protected against clinical malaria in areas with intense transmission. The protection is likely to be derived from the transfer of IgG through the placenta in utero (Amaratunga *et al.*, 2011). Having said that, most of the infants in hyperendemic areas experience the first episode of malaria attack within the first few months of life.

The acquired immunity in adults seems to be non-sterile since they continue to present asymptomatic malaria with low-level parasitemia. This condition is known as premunition where it maintains parasite load below the threshold of pathogenicity and elicits chronic infection (Pérignon and Druilhe, 1994). Premunition is often seen in malaria hyperendemic areas. In addition, age and cumulative episodes of infection are two factors contribute to premunition. As suggested by Baird and colleagues in 1991, age plays an important role in the state of malaria immunity (Baird *et al.*, 1991b). Immigrants with naïve immunity to malaria developed asymptomatic malaria and lower parasitemia when they exposed to areas with high malaria transmission. The parasitemia was significantly lower than those children in hyperendemic areas after two years of exposure to malaria. Meanwhile, parasite exposure and clinical immunity showed a positive correlation with transmission dynamics (Snow *et al.*, 1997). Holo-

endemic communities in African showed children acquired immunity from complicated malaria earlier in life. From the evidence above, cumulative exposure and age are both playing an essential role in acquired immunity against malaria.

Immunoglobulins transferred from protected healthy individuals have been shown to confer immunity against malaria in children (McGregor, 1964). An experiment was conducted in Thai children, where they were treated with immunoglobulin from West-African adults (Sabchareon *et al.*, 1991). Interestingly, acquired immunity is acting independently of parasite life cycle, but not in passive transfer immunoglobulins. Immunoglobulins from the passive transfer have been shown to reduce parasitemia but do not confer sterile protection. It was suggested that the antibodies were targeting the blood stage of the malaria parasite (Marsh *et al.*, 1989). Findings based on a longitudinal study in Ghana revealed that the rate of re-infection after antimalarial treatment was identical to that of infants' cohort in the same endemic area, suggesting that the pre-erythrocytic stage is less likely to participate in the naturally acquired immunity (Owusu-Agyei *et al.*, 2001). In fact, antibodies appear to act against the erythrocytic stage, particularly targeting the merozoite. This is indicated by the lack of HLA-class molecules on the erythrocyte surface and parasite (Perrin and Dayal, 1982). Hence, immunity operating at the erythrocytic stage could be dominated by the humoral immunity. Despite enormous efforts to characterize the vaccine candidates, only certain antigens show some degree of immune protection, which warrants further understanding of malaria immunity.

1.13.1 Humoral immunity

Antibody-mediated immune responses confer primary immunity against blood-stages of *P. falciparum* infection, although the complete picture of this mechanism is still imperfect. Antimalarial antibodies have shown to inhibit invasion and replication of *P. falciparum* parasites, disruption binding to host receptors and mediate opsonisation of infected erythrocytes (Hill *et al.*, 2013). Immunisation of IgG antibodies in African children and Thai adults revealed the importance of humoral immunity in eliciting natural immunity against malaria (Bouharoun-Tayoun *et al.*, 1990; Cohen *et al.*, 1961). In addition, cytophilic subclasses IgG1 and IgG3 were shown to protect human against

infection (Weaver *et al.*, 2016). Several studies have characterised the role of antibodies against merozoite invasion. Intriguingly, the majority of the individuals in malaria-endemic areas displayed IgG3 subclass antibodies to the MSP-2 antigen (Stanisic *et al.*, 2009). The level of antibodies against MSP-2 was also correlated with protection efficacy in Gambian population (Taylor *et al.*, 1998).

Numerous studies have displayed the association between antibodies against *Plasmodium* antigens and the reduced risk of malaria episodes (Greenhouse *et al.*, 2011; Hill *et al.*, 2013; Stanisic *et al.*, 2015). Predominantly, most of the antigens identified are merozoite surface proteins and antigens lie within the invasion machinery (Beeson *et al.*, 2016; Woehlbier *et al.*, 2010). These components are thought to be promising vaccine candidates because they are readily exposed to antibodies (Beeson *et al.*, 2016). Given that polymorphic antigens are not considered promising vaccine candidates, the conserved regions within these antigens should be considered in subunit vaccine design because these parts are essential for the development of the humoral neutralizing antibody against the pathogen (López *et al.*, 2017). This approach was employed in vaccine design using *P. falciparum* erythrocyte membrane protein 1 (PfEMP-1) (Krause *et al.*, 2007), AMA-1 (Remarque *et al.*, 2012), and MSP-1 (Cavanagh *et al.*, 1998). In malaria holoendemic areas, stable immune responses were seen to confer by AMA-1 in all age groups of patients (Remarque *et al.*, 2012). Likewise, naturally acquired immunity was reported to induce by MSP-1 in *P. falciparum* and the correlation to protective immunity was subsequently reported (Cavanagh *et al.*, 1998; Moormann *et al.*, 2013).

1.13.2 Cellular immunity

Complete eradication of malaria parasites in the circulation is CD4+ T cell- and B cell-dependent (Langhorne *et al.*, 1998). CD4+ T cells confer protective immunity and also limit the parasite replication without B cells (Grun and Weidanz, 1983). The experiment conducted using CD4+ T cells from healthy individuals and exposed to *P. falciparum* antigens *in vitro* revealed secretion of cytokines and proliferation of T-cells (Rhee *et al.*, 2001). The CD4+ T cells response consists of two functionally distinct subsets, interferon-γ (IFNγ) and interleukin-4 (IL4). The secretion of IL-4 by T-cells especially is correlated with the antibody titre (Boström *et al.*, 2012). Populations in the malaria-

endemic areas often show no malaria symptoms and low T-cell responses to malaria antigen *in vitro* (Hviid *et al.*, 1996). Likewise, the similar observation was reported in Madagascar population where the individuals showed low T cell responses (Chougnet *et al.*, 1990). The lack of T cell responses could stem from the arresting antigen-specific T-cell outside the peripheral circulation (Hviid *et al.*, 1991) or from host genetic factors (Jepson *et al.*, 1997).

IFNγ plays a central role in the protective immunity (Inoue *et al.*, 2013). It participates in the activation of mononuclear and polymorphonuclear leukocytes. These components are essential in phagocytosis and lysis of infected red blood cells. On the other hand, as erythrocytes lack HLA class-I molecules, CD8+ T cells might be confined to the pre-erythrocytic stage with its cytotoxic role (Huang *et al.*, 2015; Tsuji, 2010). In contrast, the MHC-unrestricted γδ T cells might act during the erythrocytic stage where it showed an inhibitory effect in the *P. falciparum* cultures (Huang *et al.*, 2015).

1.14 Thesis aims and organisation

This thesis has two general objectives, i) to investigate the population structure of *P. vivax* in three malaria-endemic areas of Thailand, and ii) to examine the suitability of PvMSP-7 as vaccine candidate from the perspective of population genetics, gene expression, and immunogenicity.

The first aim of the thesis was to investigate the population structure of *P. vivax* from three malaria major endemic areas of Thailand using whole-genome approach. Patients were recruited from three different areas, Tak province (Northwest of Thailand), Ubon Ratchathani province (Northeast of Thailand), and Yala province (South of Thailand). Further to that, analysis was focused on the PvMSP-7 multigene family located at the chromosome 12 of *P. vivax*. This multigene family has been suggested to express during blood-stage infection and could affect the merozoite invasion of erythrocytes. For these reasons, PvMSP-7 paralogs are plausible vaccine candidates. The investigation of the population structure of *P. vivax*, antigenic variation of 13 PvMSP-7 paralogs, the transcriptional changes in natural infection, and novel

immunogenic epitopes will serve as ultimate starting points for further experimental work. The works present herein, will translate into the development of PvMSP-7 as a vaccine candidate.

The thesis is organized into five chapters explaining the main findings. In **Chapter 2**: population genomics of *Plasmodium vivax* in Thailand, describes the population structure of *P. vivax* parasite populations in Thailand. **Chapter 3**: sequence diversity of multigene family *Plasmodium vivax* merozoite surface protein 7 (PvMSP-7) genes in Thai clinical isolates, explains the antigenic variation of 13 PvMSP-7 paralogs in Thailand. The genetic diversity was explored across 13 PvMSP-7 paralogs which useful in the vaccine development perspective. **Chapter 4**: polymorphism in merozoite surface protein-7E of *Plasmodium vivax* in Thailand: Natural selection related to protein structures, presents a comprehensive analysis of the highly polymorphic locus in the PvMSP-7 multigene family which showed the potential of this locus as a genetic marker in Thailand. **Chapter 5**: clinical expression profiles of a *Plasmodium vivax* vaccine candidate: merozoite surface protein 7 (PvMSP-7), elucidates the transcriptional changes of this multigene family in natural infection during the IDC with co-expression analysis. **Chapter 6:** identification of antigenic epitopes within *Plasmodium vivax* merozoite surface protein 7 (PvMSP-7) in natural infection, uses the state-of-art high-density peptide array technology to screen novel immunogenic antibody epitopes across 13 PvMSP-7 paralogs.

Chapter 2

Population genomics of *Plasmodium vivax* in Thai clinical isolates

Abstract

The development of an effective malaria subunit vaccine has been hampered by the high magnitude of genetic diversity in *Plasmodium*. The global distribution of *P. vivax* is structured into distinct geographical regions and impact the design of a universally effective vaccine. Malaria transmission dynamics vary in Thailand where all the endemic regions are divided by a malaria free-corridor in the metropolitan city. Furthermore, understanding the population structure of *P. vivax* in Thailand is relevant to the efficacy of malaria vaccine development. To characterise the population structure of *P. vivax* in Thailand, 20 clinical samples were collected from three malaria endemic areas in Thailand. The 20 patients naturally experienced clinical malaria episodes were recruited from Tak province (Northwest of Thailand), Ubon Ratchathani (Northeast of Thailand), and Yala (South of Thailand). The whole-genome sequencing approach was used to sequence all clinical isolates. At the genome level, three distinct clusters were observed on the principal component analysis separating clinical isolates according to their geographical region. Pronounced genetic differentiation also revealed in the *P. vivax* populations from the three malaria-endemic areas (F_{ST}>0.1). This key finding highlights the extensive population structure of *P. vivax* in Thailand. Therefore, this discovery will guide a more effective vaccine development against *P. vivax*.

2.1 Introduction

Understanding the genetic structure of *Plasmodium* between locations provide a key insight into the genetics of the parasite which can translate into a more effective control and elimination strategy. The identification of genetic differences in parasite between geographical areas pinpoint the variants that present at different frequencies (Takala and Plowe, 2009). This knowledge will show the common allele and rare allele circulating between two populations. Having said that, should a malaria subunit vaccine include the rare allele it might elicit variant-specific immunity against the infection (Ferreira *et al.*, 2004). This is one of the obstacles in malaria vaccine design that contributes to lacking a universally effective vaccine. *P. vivax* often shows the high magnitude of genetic diversity and geographical variation (Chen *et al.*, 2017; Jennison *et al.*, 2015; Neafsey *et al.*, 2012). Therefore, a detailed characterisation of the *P. vivax* population structure between different endemic areas expands the strategy to design a more effective vaccine.

The genetic diversity of *P. vivax* is far higher than previously thought (Jennison *et al.*, 2015; Neafsey *et al.*, 2012). However, the variation patterns are not homogenous under different transmission settings due to host genetics and environmental factors. The population tends to cluster according to continental origins. *P. vivax* population in America found to have less variation relative to population in Asia or Oceania (Imwong *et al.*, 2007). Contrasting to the previous study, genetic diversity of *P. vivax* inferred from mitochondrial genomes in America was found comparable with those populations in other continental origins (Taylor *et al.*, 2013). This observation could be due to the limited sampling areas in the previous study and complex geography pattern drives the variation in *P. vivax* population (Taylor *et al.*, 2013). Two recent studies reported this pathogen is adapting to the selection pressure present in each local landscape which translates ongoing evolutionary interaction between the parasite and the environment (Hupalo *et al.*, 2016; Pearson *et al.*, 2016).

Microsatellite approach has been broadly used to infer the population structure of *P. vivax* in diverse endemic areas (Ferreira *et al.*, 2007; Kittichai *et al.*, 2017; Koepfli *et al.*, 2015; Liu *et al.*, 2014). The population structure of *P. vivax* across four continents was investigated using 11 microsatellites (Koepfli *et al.*, 2015). In total, 841 clinical isolates were collected between the year 1999 to 2008 from Central Asia, South

America, South-East Asia, and the South Pacific. Koepfli and colleagues reported the parasite diversity was greater in South-East Asia followed by South Pacific, South America, and Central Asia. The genetic differentiation was also pronounced in all parasite populations implying geographical factor influences the population structure (Koepfli *et al.*, 2015).

Likewise, a similar microsatellite strategy was used to examine the genetic structure of *P. vivax* in Amazonia (Ferreira *et al.*, 2007). Ferreira and colleagues also assessed the *P. vivax* transmission dynamics through the cross-sectional and longitudinal surveys. The analysis was conducted in 74 clinical samples using 14 markers (Ferreira *et al.*, 2007). Interestingly, a strong linkage disequilibrium and high frequency of haplotypes replacement were observed in the same pool *P. vivax* population over time. This is ultimately contributing to the increase antigenic variation within the population. Moreover, Kittichai and colleagues employed microsatellite approach to uncover the substantial population structure of *P. vivax* in Thailand (Kittichai *et al.*, 2017). Ten genetic markers were used to identify the population structure in two malaria endemic areas in Thailand. Based on the finding on 127 clinical samples, genetic differentiation was evidenced between endemic regions and no sharing haplotype was found (Kittichai *et al.*, 2017). Therefore, a more robust elimination strategy is required to tackle the malaria transmission in Thailand.

Two recent genomic analyses of *P. vivax* have revealed the global population structure (Hupalo *et al.*, 2016; Pearson *et al.*, 2016). The first study unravelled the global *P. vivax* population structure using 247 samples from Southeast Asia, Oceania, and a few clinical isolates from China, India, Sri Lanka, Brazil, and Madagascar (Pearson *et al.*, 2016). Closer looks into the sample collection in Thailand, 88 patients from Western Thailand and 4 patients from Eastern Thailand were involved in the analysis. Whole-genome sequencing approach was used to sequence the clinical isolates. In total, 726,077 high-quality single-nucleotide polymorphisms (SNPs) were derived. Interestingly, a phylogeny analysis revealed three distinct branches clustering samples from Western Southeast Asia, Eastern Southeast Asia, and the Pacific Island. The *P. vivax* populations in Thailand stratified into Western and Eastern groups that suggest malaria-free corridors have established in the metropolitan cities. The finding was in line with principal component analysis and ADMIXTURE analysis (Pearson *et*

al., 2016). Similarly, Hupalo and colleagues used the similar genomic approach to study the population stratification of *P. vivax* globally (Hupalo *et al.*, 2016). The study recruited 182 patients from 11 countries (Brazil, Cambodia, Colombia, India, Madagascar, Mexico, Myanmar, Korea, Papua New Guinea, Peru, Western Thailand, and Vietnam). The principal component analysis revealed distinct clusters that divide the isolates according to the geographical demography. The concordant result was obtained in phylogeny analysis and ADMIXTURE analysis (Hupalo *et al.*, 2016). Therefore, the global *P. vivax* population structure has been influenced by the geographical isolation. Understanding the local patterns of malaria transmission will improve future malaria vaccine development strategy.

In Thailand, *P. vivax* contributes significantly to the malaria incidence rate. The proportion of malaria incidence in *P. vivax* increase from 18.9% in the year 2011 to 63.2% in 2015 (Bureau of Vector Borne Disease, 2015). Malaria transmission in Thailand has a unique characteristic where all the endemic areas are separated by a malaria-free corridor in the central of Thailand (Parker *et al.*, 2015; Pearson *et al.*, 2016). The malaria endemic areas are clustered along the international borders such as Myanmar, Laos, Cambodia, and Malaysia (Parker *et al.*, 2015; Thimasarn *et al.*, 1995). This malaria transmission landscape is arising from the complex interactions between the ecological and socio-cultural factors (Thimasarn *et al.*, 1995). That said, careful evaluation is desired to implement malaria eradication strategy in these areas. The international border between Thailand and Myanmar contributes significantly to the malaria prevalence due to the political conflict and inefficient public health infrastructure (Parker *et al.*, 2015; Thimasarn *et al.*, 1995). Moreover, malaria transmission between the Thailand-Myanmar and Thailand-Southern Malaysia has a similarity where both areas are dominated by militants which retard the health service implementation (Thimasarn *et al.*, 1995). Population movement across the border further complicates the malaria transmission dynamics with transporting malaria parasites from one region to another region (Thimasarn *et al.*, 1995). On the other hand, Thailand-Cambodia border often records two seasonal peaks for malaria transmission, one in the dry season and one at the beginning of the rainy season (Thimasarn *et al.*, 1995). The malaria transmission in this region is contributed by the internal migration due to economic factors (Guyant *et al.*, 2015). Prosperous natural resources in the forest fringe areas drive the population to the areas where malaria transmission is prevalent.

Therefore, these factors contribute significantly to the malaria transmission dynamics in Thailand.

Therefore, mapping of global and local *P. vivax* population structure is playing a pivotal role in developing an antimalarial malaria vaccine. Previous studies determined the population structure of *P. vivax* in Western and Eastern of Thailand, although the size from the Eastern province was very low (n=4). The present study employed whole-genome sequencing approach to derive a more detailed picture of *P. vivax* population structure from three malaria endemic areas in Thailand (Tak province, Northwest of Thailand; Ubon Ratchathani, Northeast of Thailand; and Yala, South of Thailand). Using the finding from the genomic level, it will guide the development of a malaria-subunit vaccine in Thailand. Eventually, it will lead to a more effective strategy against malaria control and elimination in Thailand.

2.2 Methodology

2.2.1 Ethic Statement

Informed consent was obtained from all participants involved in the study. The subjects were informed regarding the purpose of the study and the potential risks involved. The research study was approved by the Institutional Review Board of the Faculty of Medicine, Chulalongkorn University (COA No. 322/2016 and IRB No. 104/59). All procedures performed in the study followed the international guidelines for human research protection as the Declaration of Helsinki, The Belmont Report, CIOMS Guideline and International Conference on Harmonization in Good Clinical Practice (ICH-GCP).

2.2.2 Study Population

This study was conducted across three rural areas along the international borders of Thailand. These study sites were hotspots of malaria transmission. The first study site

was located at Yala province, the border of Thailand and Malaysia (South of Thailand). Secondly, the analysis focused on Ubon Ratchathani province, the border of Thailand and Cambodia (Northeast of Thailand). Lastly, the third study site was Tak province, the border of Thailand and Myanmar (Northwest of Thailand). Yala occupies 4,521.1 km^2 with a total population of approximate 511,911 people. Ubon Ratchathani has approximate 1,000,000 of population occupies 15,744.850 km^2. Tak has more than 539,000 population occupies 16,406.6 km^2.

2.2.3 Sample Collection

Patients infected with malaria were allowed voluntary participation in this research study. Informed consents were requested from all patients in compliance with the Institutional Review Board. Samples collection was carried out between May to August 2016 in respective province hospitals. A trained medical officer and medical laboratory technician were responsible to confirm that the patients were infected with malaria through signs and symptoms of infection and microscopy diagnosis. Twenty patients (*n*=20) were recruited in the study, eight patients from Yala province (*n*=8), seven patients from Ubon Ratchathani province (*n*=7), and five patients from Tak province (*n*=5) (Table 2.1). The age of the 20 patients ranged from 16 to 50 years with the mean age of 31.15 years. Based on the clinical history and physical examination, all patients infected with only malaria parasite with no evidence of other concurrent infections. Approximate ten millilitres of venous blood sample was drawn from each subject and preserved in EDTA anticoagulant tubes. Blood samples preserved in EDTA anticoagulant tubes were transported on ice from the study sites to the laboratory at the Department of Parasitology, Faculty of Medicine, Chulalongkorn University. Upon arriving in the laboratory, clinical samples were processed immediately to avoid lysis of human leukocytes. To further characterise the population structure of *P. vivax* between the international borders of Thailand, 48 genome sequences were retrieved from the National Center for Biotechnology Information (NCBI) database. These data were used in the global study of *P. vivax* (Hupalo *et al.*, 2016; Pearson *et al.*, 2016). The isolates were sequenced from various continental origins, Brazil (*n*=3), Cambodia (*n*=10), Myanmar (*n*=8), Malaysia (*n*=6), East Thailand (*n*=3), and West Thailand (*n*=18) (Supplementary Table 1).

2.2.3.1 Inclusion Criteria

Febrile individuals with single *P. vivax* infection who expressed willingness to participate in the study. Microscopy examination and molecular testing were used to detect the presence of single vivax infection.

2.2.3.2 Exclusion Criteria

Children less than 5-year old, patients with severe malaria symptoms, and other known underlying immunodeficiency diseases were excluded in the study.

2.2.4 **Microscopy (**Species Identification and Parasitemia**)**

Microscopy examination was used to screen all the clinical isolates collected. Thin and thick blood smears were prepared to identify the presence of malaria parasites and possibly the species of the parasites. The blood smears were stained with Giemsa's stain which is the gold standard diagnosis in the laboratory. Parasitemia count was calculated from thin and thick blood smears per guidelines recommended by Centers for Disease and Prevention (CDC). One-hundred microscopic fields were examined under 100X objective for thin smears, while two-hundred leukocytes were counted for thick blood films. The equations for parasitemia calculation as follow:

Thin smear

$$\text{Infected erythrocytes in percent (\%)} = \frac{\text{Number of infected erythrocytes}}{\text{Total number of erythrocytes counted}} \times 100$$

Thick smear

$$\text{Number of parasites per microliter (μL) of blood} = \frac{8000}{\text{Number of leukocytes counted}} \times \text{number of parasites}$$

*Assumed 8,000 leukocytes per μL to quantify parasite density

Table 2.1 20 clinical isolates collected in the study. Patients were collected from three malaria endemic areas in Thailand (Yala, Ubon Ratchathani, and Tak province).

No.	Identifier	Location	Year of collection	Source	Age
1	YL002G16	Yala	2016	Clinical	50
2	YL003G16	Yala	2016	Clinical	48
3	YL004G16	Yala	2016	Clinical	24
4	YL005G16	Yala	2016	Clinical	43
5	YL007G16	Yala	2016	Clinical	14
6	YL008G16	Yala	2016	Clinical	47
7	YL009G16	Yala	2016	Clinical	35
8	YL010G16	Yala	2016	Clinical	24
9	UB001G16	Ubon Ratchathani	2016	Clinical	29
10	UB002G16	Ubon Ratchathani	2016	Clinical	25
11	UB003G16	Ubon Ratchathani	2016	Clinical	46
12	UB004G16	Ubon Ratchathani	2016	Clinical	23
13	UB005G16	Ubon Ratchathani	2016	Clinical	48
14	UB006G16	Ubon Ratchathani	2016	Clinical	24
15	UB007G16	Ubon Ratchathani	2016	Clinical	58
16	TAK001G16	Tak	2016	Clinical	20
17	TAK002G16	Tak	2016	Clinical	16
18	TAK003G16	Tak	2016	Clinical	29
19	TAK004G16	Tak	2016	Clinical	23
20	TAK005G16	Tak	2016	Clinical	17

2.2.5 Molecular identification

Nested PCR with species-specific primers were designed to identify the species of malaria. These primers were designed and optimized in Molecular Biology of Malaria and Opportunistic Parasites Research Unit, Department of Parasitology, Faculty of Medicine, Chulalongkorn University (Putaporntip et al., 2009a). The nested PCR primers were designed based on 18S ribosomal RNA with 100% specificity towards human malaria species (Table 2.2). The first reaction of the nested PCR contained 20 µL mixture from three µL DNA template, 0.13 µL genus-specific primers, 13.79 µL nuclease-free water, two microlitres 10X buffer, 0.6 µL magnesium chloride, 0.4 µL dNTP, and 0.08 µL Taq polymerase. The first amplification was set at 94ºC for one minute (denaturation), 40 cycles (denaturation) at 94 ºC for 40 seconds, first annealing at 50ºC for 30 seconds, second annealing at 72ºC for one minute, elongation at 72ºC for five minutes and finally hold at 20ºC. The secondary nested PCR reaction contained 20 µL mixture with three microlitre template from the first reaction. Species-specific primers (Table 2.2) were used for the secondary nest with amplification conditions the same as the first reaction, but the number of cycles reduced to 30. Two percent agarose gel was prepared and ran for 30 minutes. The agarose gel was visualized using ethidium bromide (EtBr). In each PCR reaction, a sample confirmed infected with *P. vivax* was used as a positive control while a blank contained only nuclease-free water was used as a negative control.

Table 2.2 Species-specific primers used to target different human malaria species.

Species	Primer	Primer Sequence (5' – 3')
Plasmodium spp.	F1	ATG CTT TAT TAT GGA TTG GAT GTC
	R1	CAG ACC GTA AGG TTA TAA TTA TGT
P. falciparum	PfF1	ATT ATT TAT TGT ATT ATT TTT TCT G
	PfR1	GTA TTG AGC GGA ACA AAT C
P. vivax	PvF1	AGT TAC CAC AAG ATA TTT TTG AAT TTT
	PvR1	TTG AGC AGA ACA ATA CAG
P. ovale	PoF1	ATA TCA TTT TTC TCC AGT GGG
	PoR1	ATG AGC AGA ACA ATA CAG
P. malariae	PmF1	ATA TCA TTC TTT TCT TAG TGG T
	PmR1	CTG TGC AGA ACA ATA CAG
P. knowlesi	PkF1	TAT TCT TCT TTT AGT GGA TTA TTT A
	PkR1	TAC ACT GAT TAG AAC AAT AC

2.2.6 Clonal detection

P. vivax is genetically diverse in Thailand owing to the great magnitude of malaria transmission along the borders. This transmission pattern channels to the multiple-strain or multiple clone infections (Gupta *et al.*, 2016; Havryliuk and Ferreira, 2009; Lin *et al.*, 2013). Isolates infected with multiple strain of parasites can lead to misidentification of the variants. Therefore, analysing isolates with a single parasite strain will improve the variant accuracy. Highly polymorphic markers are powerful and reliable tools to genotype *P. vivax* isolates. Several highly polymorphic genes are ideal for molecular genotyping, such as PvMSP-1 Belem strain (Putaporntip *et al.*, 2002), PvMSP-1 Salvador I strain, PvMSP-3α, PvMSP-3β (Putaporntip *et al.*,

2014), and PvMSP-3γ (Rice *et al.*, 2013; Rungsihirunrat *et al.*, 2011; Véron *et al.*, 2009) . PCR amplification using these molecular markers were performed on 20 samples collected from each endemic area. Each primer was designed to amplify the respective gene in *P. viva*x. Mixed infection was distinguished with more than one band detected on the gel (Figure 2.1). The PCR reaction was carried in a total volume of 20 µL. The reaction mixture contained three microlitres DNA template, 0.13 µL genus-specific primers, 13.79 µL nuclease-free water, two microlitres 10X buffer, 0.6 µL magnesium chloride, 0.4 µL dNTP, and 0.08 µL *Taq* polymerase. The amplified PCR products were analysed on one percent agarose gel stained with ethidium bromide.

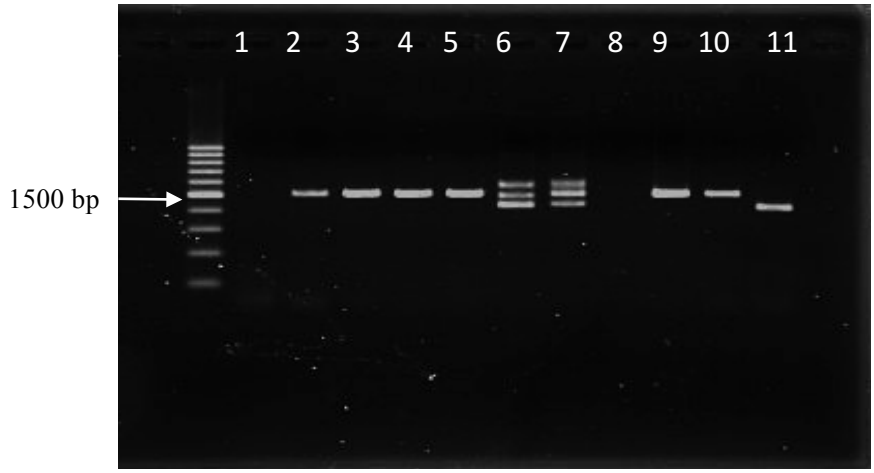


Figure 2.1 Clonal detection in isolates infected with *P. vivax*. PvMSP-3α was used as a genetic marker to screen 11 clinical isolates. Lane 1= negative control, lane 2 – 5= single infection, lane 6 – 7= mixed strain infections, lane 8 = negative sample, lane 9 – 11= single infection. Mixed clone infections are shown in lane 6 and 7 where three bands are observed.

2.2.7 Leukocytes Removal

The clinical samples were collected directly from the malaria patients which contained largely the human DNA. To minimize human DNA contamination in the downstream analyses, leukocytes were removed using C6288 cellulose (Sigma-Aldrich) column. The protocol was adapted from Sriprawat *et al.* (2009). Plasma from the clinical isolates

was removed via centrifugation at 5,000 rpm for 15 minutes. A 20 mL syringe with a centre outlet was tipped with two 2 cm^2 pieces of Grade 105 lens cleaning paper (Whatman®). Fifteen millilitres of loosely packed C6288 cellulose fibre was added to the syringe and packed down to ten millilitres mark on the syringe. Before the blood was filtered through the column, five millilitres of saline water was added to wet the column. Infected-blood samples were added to the syringe and allowed to pass through via gravity. Once the blood was no longer visible on top of the syringe, applied plunger onto the syringe to force last few drops of blood out of the column. The filtered blood was then centrifuged at 5,000 rpm for ten minutes and the supernatant was removed (Figure 2.2).

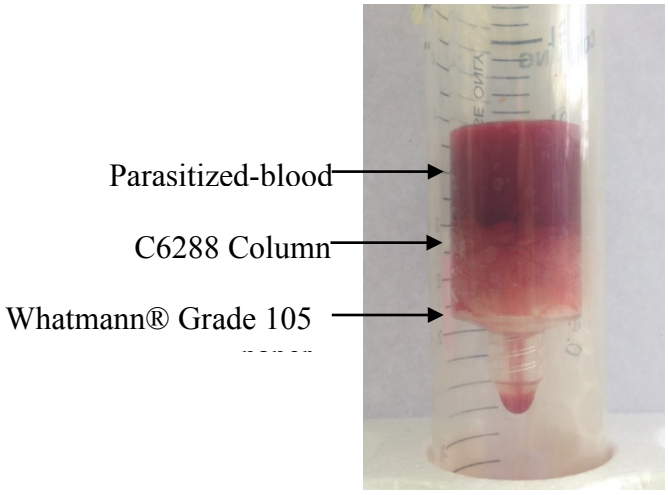


Figure 2.2 Packed C6288 cellulose column. The parasitized-blood was filtered through C6288 column into a collection tube.

2.2.8 DNA Extraction

DNA extractions were performed using the DNeasy Blood Mini and Midi Kit (Qiagen, Hilden, Germany) with slight modifications from the manufacturer's protocol. Plasma was removed from the clinical specimens and frozen at -80˚C for downstream applications. Two-hundred microliter of proteinase K enzyme was pipetted into the bottom of the 15 mL centrifuge tube. Two millilitres of blood was suspended in the 15 mL centrifuge tube, followed with 2.4 mL of lysis buffer. The suspension was mixed thoroughly and incubated at 70˚C for ten minutes. This step is important to ensure the cells are lysed and released of parasite DNA. Two millilitres of ethanol was added and

mixed vigorously to obtain the homogenous suspension. The mixture was passed through the Qiagen column which has a silica membrane to trap DNA and eliminate contaminants and proteins. The column was washed twice with washing buffer 1 and washing buffer 2. The column was spin dry to ensure the solutions from washing buffers are completely removed, to prevent possible interference with downstream analyses. The DNA was eluted with elution buffer provided in the kit.

2.2.9 DNA quantification

DNA concentration was quantified using three different approaches including *Qubit*, *TECAN*, and *NanoDrop*. The accuracy of DNA quantity is essential as it would influence the end results and library preparations of the whole-genome sequencing. A 0.5% agarose gel was prepared to assess the quality of purified DNA. The gel was left overnight at 30V and visualised with *SYBR Green* (Applied Biosystems, Carlsbad, USA). The single band should be observed on top of the gel which indicates DNA integrity.

2.2.9.1 *Qubit*

Total DNA amount was measured with Qubit® 3.0 fluorometer (Invitrogen). Qubit has higher sensitivity in quantifying DNA as it employs fluorometric principle. Fluorescent dye increases the accuracy of the measurement, as it only binds to the target of interest and concentrations measure through the intensity of fluorescence. DNA concentration was measured using Qubit® dsDNA BR Assay Kit (Invitrogen) designed specifically to quantify double-stranded DNA. Calibration of the fluorometer was performed according to the manufacturer's protocol. Two standards were prepared for the initial calibration, standard 1 and standard 2 (Table 2.3). Working solution was prepared by diluting Qubit® dsDNA BR reagent 1:200 in Qubit® dsDNA BR buffer. After the sample (1 μL) was added to the assay tube, it was vortexed for three seconds and incubated at room temperature for five minutes. Calibrations were performed according to the instructions displayed on the fluorometer. Sample volume and measurement units

were selected. Assay tube was inserted into the fluorometer and removed after the reading was recorded. Measurements were repeated triplicates. The Qubit® 3.0 fluorometer generates the readings automatically according to the equation as follows:

$$Concentration\ of\ sample = QF\ value \times \frac{200}{x}$$

QF = value given by the Qubit 3.0 Fluorometer

x = volume of sample added (µL)

Table 2.3 Calculation for DNA quantifications using Qubit® dsDNA BR Assay Kits and Qubit® RNA Broad-Range Assay Kits

	Sample volume (µL)	Working solution (µL)	Final volume (µL)
Standard 1	10	190	200
Standard 2	10	190	200
Sample	1	199	200

2.2.9.2 *TECAN*

Quant-iT™ Picogreen (TECAN) acts as a fluorescent dye that binds exclusively to DNA molecules. This dye is an ultrasensitive fluorescent nucleic acid stain that accurately quantifies double-stranded DNA. Five standards were prepared for initial calibration (Table 2.4). Samples were diluted to the ratio of 1:50, DNA to TE buffer. This concentration is preferred for samples with a low DNA quantity. Based on the quantification of Qubit, DNA concentration in all samples was not high. Therefore, dilution factor 1:50 was ideal to increase the accuracy of quantification. To achieve the 1:50 ratio, 4.2 µL of sample was mixed with 205.80 µL of TE buffer. The picogreen working solution was then added to the tubes and mixed homogeneously. The mixture was transferred to the respectively well on the 96-well microplate. After 5 minutes incubation under aluminium foil, the microplate was loaded to Infinite 200 PRO Microplate Reader (TECAN). Plate layout and final concentrations were set in the

Magellab™ Data Analysis Software. The preparation was then excited at 485 nm and emission measured at 520 nm. Data was output in an Excel spreadsheet and values were expressed in ng/µL. Sample concentration was then plotted against the DNA standard curve.

Table 2.4 Preparation of standards for Quant-iT™ Picogreen.

Standard	TE Buffer (µL)	DNA Standard (µL)	Picogreen (µL)	Concentration (ng/mL)
Standard 1	0	210	210	1000
Standard 2	189	21	210	100
Standard 3	207.9	2.1	210	10
Standard 4	209.79	0.21	210	1
Standard 5	210	0	210	0

2.2.9.3 *NanoDrop ND-2000*

NanoDrop (Thermo Scientific, Delaware, USA) is a microvolume sample retention technology that allows quantitation of low nucleic acid in a preparation. The principle based upon the fibre optic technology and surface tension properties that hold the sample. In the present study, as the parasitized blood was precious, one microlitre of each purified DNA was pipetted onto the measurement pedestal. The spectral measurement was initiated through a setup on the computer. Once the measurement was recorded, the sample was wiped off using lens cleaning paper. The A260/A230 ratio was recorded for each sample. This ratio indicates the purity of the nucleic acid. The ratio is preferably within the range 1.5 to 1.8, readings below or above the range could indicate possible contaminants present in the sample such as organic compounds carryover from the DNA purification procedure. To quantify the sample accurately, the upper and lower pedestals were cleaned with deionised water before each measurement was taken. This procedure would minimise the sample carryover and remove residues from the surface.

2.2.10 Whole-genome sequencing

Library preparation and sequencing was performed at the Centre for Genomic Research, University of Liverpool, United Kingdom. Sequencing was performed on HiSeq4000 platform. Modified TruSeq Nano DNA Library Prep Kit was prepared for genomic DNA samples. Paired-end sequencing with 350 bp insert libraries of 20 indexed samples was run on a single lane Illumina HiSeq4000 platform. The procedure began with the input of 50 ng for all samples to recover the material after fragmentation. Then, the half volume of reactions was used throughout the protocol. The whole-genome sequencing generated approximately 22 to 39 million paired-end reads for 20 clinical isolates. In addition, the genome-wide mean coverage of 20 clinical isolates was estimated using Qualimap 2 (Okonechnikov *et al.*, 2015). The depth of coverage for 20 isolates ranged from 2.11X to 147.02X (Table 2.5).

2.2.11 Bioinformatics Analysis

The raw sequences in the Fastq format were derived from the whole-genome sequencing. Fastq sequences contain quality score for each nucleotide and sequencing adapters. Before the sequences ready to use for downstream analyses, the raw Fastq reads were trimmed for the presence of Illumina adapter sequences using Cutadapt, version 1.2.1 (Martin, 2011). A specific option (-O 3) was set during the trimming procedure, any 3' end of the reads that match at least three base pairs of the adapter sequence were removed. Furthermore, reads that scored below the quality score of 20 and shorter than ten bases were removed using Sickle, version 1.200 (Joshi and Fass, 2011). The overall quality of the sequencing data was evaluated with FastQC (Andrews, 2010). The software assessed the GC content and proportion of duplicated reads. MultiQC (Ewels *et al.*, 2016) was used to merge all QC reports from FastQC into a single summary report which provides a more systemic visualisation of QC results. Based on the outputs, sequences passed the FastQC evaluation including the normal distribution of overall GC content, quality values of all bases were high (quality score within 20 to 34), and low overrepresented sequences. A high overrepresented sequence is usually an indication of library contamination.

2.2.12 Read mapping

P. vivax isolates were sequenced directly from the field and contained a high amount of host DNA. Therefore, the sequences were mapped to the human genome (GRCH37) using BWA-MEM (Li, 2013). The unmapped sequences were mapped against the *P. vivax* PvP01 reference genome using BWA-MEM (Auburn *et al.*, 2016). Due to a large number of reads, BWA-MEM aligner was preferred as it has been optimised for the fast alignment of illumina sequence reads (Li, 2013). In addition, BWA-MEM was also reported to generate high quality and improved accuracy of sequence alignment especially in a complex *Plasmodium* genome (Thankaswamy *et al.,* 2017). The mechanism behinds BWA-MEM is the sophisticated seeding algorithm with maximal exact matches (MEMs). The process then extends the seeds with the affine-gap Smith-Waterman algorithm (SW). For paired-end mapping, BWA-MEM works with a batch of reads independently. For each batch of the reads, mean and variance are calculated across the insert size distribution. These statistics will then transform into alignment scores and use to build final alignment (Li, 2013). BWA-MEM was used to align paired-end reads to the reference genomes. The alignments were stored in the Sequence Alignment Map (SAM) files. Default setting recommended by BWA-MEM was used throughout the alignment. Shorter split hits were marked using the command '-M'. This command was to improve the identification of duplicates. Mapping quality was assigned to each individual read. A read that aligned with no gap and no mismatch was assigned a high-quality score. However, its mates mapped within the complex regions were assigned a low-quality score because the point of origin cannot be determined. These quality scores indicate the probability of reported alignment is incorrect and useful in guiding the variant discovery. The quality score more than ten likely to indicate the alignment is unique. Once the reads aligned to the reference genome, they were sorted using SAMtools (Li *et al.*, 2009). SAMtools sorted SAM files to the BAM files for downstream analyses. BAM file and SAM file are comparable, however, BAM file is compressed to allow fast retrieval for indexed queries. BAM file was sorted by coordinate and indexed to achieve fast access to a specific aligned region. As the aim of the downstream analysis was to discover variants, more stringent criteria were applied in the GATK pipeline to avoid false positive variants.

In the present study, *P. vivax* PvP01 reference genome was preferred over the Salvador I reference genome because the quality of the assembly was improved greatly. The fragmentation was reduced to 226 scaffolds from 2500 scaffolds in Salvador I reference genome. The PvP01 reference genome also discovered additional 792 genes (Auburn *et al.*, 2016). The nuclear genome of PvP01 is 29 Mb in size, distributed across 14 chromosomes with 6,642 genes identified. The PvP01 reference genome was retrieved from PlasmoDB release 31 (http://plasmodb.org/plasmo/).

2.2.13 Variant calling

Genome Analysis Toolkit, version 3.7 (GATK) was used to discover variants (McKenna *et al.*, 2010). Procedures were performed according to the best practices pipeline in GATK. The best practices pipeline has been refined by the developers to discover high-quality variants using high-throughput technology. The GATK pipeline was extensively integrated in the genomic analysis of *P. falciparum* and *P. vivax* to derive a reliable dataset (Hostetler *et al.*, 2016; Hupalo *et al.*, 2016; Lukens *et al.*, 2014; Neafsey *et al.*, 2012; Park *et al.*, 2012; Pearson *et al.*, 2016). Picard, version 2.0.1 was used to pre-processing the BAM files before passing to GATK including AddorReplaceReadGroups, CleanSam, FixMateInformation, and MarkDuplicates. AddorReplaceReadGroups was used to assign a unique identifier to the BAM file. CleanSam was specified to perform soft-clipping extends beyond-end-of-reference alignment and MAPQ was set to 0 for unmapped reads. FixMateInformation was applied to verify all mate-pair information that matched between each read and its mate pair. MarkDuplicates function was used to identify multiple reads that match at a specific position in each BAM file. All duplicate reads were tagged, and a metrics file was generated for each BAM file contained a number of duplicates. The MarkDuplicates function is essential to remove PCR duplicates, failure of this procedure will cause over-representation of the overall sequence quality and depth of coverage. The resulting BAM files from Picard tool were used to perform local realignment in GATK. The local realignment was used to reduce the number of mismatching bases relative to the reference sequence. The presence of mismatching bases will introduce errors to the variant discovery process, as it might be mistaken as

a variant. HaplotypeCaller function was used to discover SNPs and indels in the dataset. The variant calling function was performed independently on each BAM file. The variants derived from each BAM file were joint using HaplotypeCaller. The variants were stored in the Variant Call Format (VCF) where it contained information about the position of the variant, quality score, and sample statistics. The joint genotyping procedure is effective to produce high-confidence genotype likelihood for every position (Pristo *et al*., 2011, Auwera *et al*., 2013) . Various artefacts in high-throughput technology were discussed including the incomplete reference genome and dependent errors (Li, 2014). In the present study, to reduce the artefacts in the analysis, variants from one sample were compared against another sample using HaplotyeCaller in GATK. The different consensus of variants was compared against each sample and the best match was chosen based on the variant quality score recalibration (VQSR). Variants with a VQSR score above 90.0 were considered as a true variant whereas variants below the threshold were omitted. The concordant variants were then used in the downstream analysis. As malaria parasite is haploid, *ploidy* -1 function was specified to consider only the heterozygous sites.

2.2.14 Variant filtering

Variant filtration was performed to achieve a highly reliable variant dataset. The sequence and variant calling parameters were assessed to eliminate low-quality variants, such as phred-like quality (QUAL), QualByDepth (QD), FisherStrand (FS), RMSMappingQuality (MQ), MappingQualityRankSumTest (MQRankSum), and ReadPosRankSumTest (ReadPosRankSum). The cut-off values for these parameters were used as recommended by GATK and other published data in *Plasmodium* (Hostetler *et al.*, 2016; Hupalo *et al.*, 2016; Lukens *et al.*, 2014; Neafsey *et al.*, 2012; Park *et al.*, 2012; Pearson *et al.*, 2016). QUAL refers to the phred-scaled quality at a variant site, high confidence calls would usually have high QUAL. QD is calculated from QUAL and unfiltered depth of samples which indicates variant confidence. FS is useful to detect any strand bias such that only one variant at a specific site observes either on the forward or reverse strand. The identification of strand bias is based on the phred-scaled *p*-value using Fisher's Exact Test. MQ indicates the mapping quality of all isolates. MQRankSum calculates the mapping qualities from the Mann-Whitney

Rank Sum Test. ReadPosRankSum is useful for heterozygous calls. The principle is similar to MQRankSum, but it is specific to determine the distance between alternate allele. In the initial analysis, the reads were mapped to the PvP01 reference genome, pre-processed using Picard tool, and indel realignment using GATK. This procedure derived a total of 761,249 consensus SNPs. After further filters applied to the dataset including VQSR, QUAL, QD, FS, MQ, MQRankSum, and ReadPosRankSum, the final haploid dataset contained 247,789 SNPs. This dataset was used for subsequent analysis. The following filters were used in the 20 clinical isolates:

- Variants with VQSR < 90.0 were excluded
- Variants with QUAL < 100.0 were excluded
- Variants with QD < 2.0 were excluded
- Variants with FS < 60.0 were excluded
- Variants with MQ < 40.0 were excluded
- Variants with MQRankSum < 12.5 were excluded
- Variants with ReadPosRankSum < -8.0 were excluded

2.2.15 Population structure

The population structure was used to infer the genetic ancestry differences in *P. vivax* population from different endemic areas in Thailand. Population stratification of *P. vivax* population in Thailand was determined using principal component analysis (PCA) implemented in SNPRelate (Zheng *et al.*, 2012) and ADMIXTURE analysis (Alexander *et al.*, 2013). The programs were freely available in R environment version 3.3.1 (R, 2016). The filtered VCF file contained 247,789 SNPs was filtered in PLINK version 1.9 (Purcell *et al.*, 2007) using the linkage disequilibrium (LD) pruning approach before the PCA was generated. The SNPs with strong LD can distort the PCA and ADMIXTURE analysis due to the tightly linked SNPs (Purcell *et al.*, 2007). The LD-pruning was, therefore, performed to reduce the variants number and uncover the true associations between samples (Sobota *et al.*, 2015). In the LD-pruning, a sliding window size of 50 was set across the genome, advanced with steps of five SNPs, and SNPs with threshold value above 0.5 were removed. The LD at 0.5 retained 24,524 independent SNPs used to infer the population structure of *P. vivax* in Thailand. The

PCA plot was constructed in SNPRelate and the population structure of *P. vivax* was revealed by the top two principal components (PCs). Genetic similarity or dissimilarity was identified through the distinct clusters. On the other hand, the population structure of *P. vivax* in Thailand was also achieved using ADMIXTURE analysis (Alexander *et al.*, 2013). The LD-pruned dataset was used in the analysis. ADMIXTURE program works on a maximum likelihood algorithm to predict the underlying admixture coefficients and ancestral allele frequencies. The number of ancestral populations (*K*) was estimated using five-fold cross-validation in haploid mode. Five *K* values between two to six were run to improve the likelihood of ancestral populations. The optimal *K* value was chosen based on the lowest cross-validation error (CV). The CV estimates the proportion of error in each *K* by partitions all the observed genotypes (Scheet and Stephens, 2006; Wold, 1978). Therefore, the optimal *K* value was used to construct the underlying population structure in bar plots.

2.2.16 Phylogeny analysis

The phylogeny tree was constructed to further validate the population structure observed in PCA. The phylogeny association has been used to explain the genetic relationships either on the population or individual level (Collins and Didelot, 2018). Phylogenetic trees were constructed using two models, maximum likelihood (Felsenstein, 1981) and the neighbour-joining method (Saitou and Nei, 1987). The maximum likelihood was performed in Randomized Axelerated Maximum Likelihood, version 8 (RAxML) (Stamatakis, 2014) using all 247,789 filtered SNPs. Maximum-likelihood estimates the evolutionary trees from nucleotide sequences and the evolutionary rate at each site is considered. RAxML works by constructing an initial tree, the algorithm will try to increase likelihood through improving each branch length and building local rearrangement. Bootstrap was set to 100-fold to increase the reliability. The best-fit model of nucleotide sequence was determined using jModelTest, version 2.0 (Posada, 2008). Usage of correct substitution is important as it will significantly affect the outcome of the phylogeny analysis. jModelTest calculates the probabilities of difference between DNA sequences along the branches of a phylogenetic tree. A sequence alignment contained all 20 sequences were passed into jModelTest, GTR substitution model with gamma rate variation was identified as the

best selection results. The neighbour-joining tree was constructed in MEGA 7.0 (Kumar *et al.*, 2016) using 500 bootstrap pseudoreplicates. The neighbour-joining tree differs from maximum likelihood where it constructs the phylogenetic tree from the genetic distance between sequences. It requires less computational power and ancestry is not considered in the analysis. That said, the neighbour-joining tree is suitable to infer the underlying population genetic structure.

2.2.17 Genetic differentiation

Fixation index (F_{ST}) was used to estimate the genetic differentiation among *P. vivax* population in Thailand. The F_{ST} was calculated in SNPRelate (Zheng *et al.*, 2012) using LD-pruned dataset. F_{ST} is influenced by the genetic polymorphism or allele frequencies, where 1 reflects the level of genetic differentiation is high and 0 indicates no population subdivision. The principle of the F_{ST} was based on the Weir & Cockerham (Weir and Cockerham, 1984). Weir and Cockerham (1984) estimate the F_{ST} between the population through the analysis of variance (ANOVA) methodology. This approach is unbiased especially in the study where the sample size is small (sample size < 6) and able to compensate for overestimating in the low magnitude of population differentiation (Willing *et al.*, 2012).

2.3 Results

2.3.1 Summary of sequencing data

Whole genome sequencing generated between 22 and 39 million paired-end reads for 20 samples (Table 2.5). Genomic studies of *P. vivax* have always been complicated by the presence of human DNA in the parasite-infected blood (Auburn *et al.*, 2013). Removal of human leukocytes from the clinical samples was performed using a modified cellulose column (Venkatesan *et al.*, 2012). However, the proportion of reads that mapped to the *P. vivax* PvP01 reference genome remained low for most of the isolates. The mean coverage between 2.11 to 147.02. The genome sequences were mapped to the human reference genome (GRCh37) before the unmapped reads were mapped to the *P. vivax* PvP01 reference genome. Most of the samples have a significant proportion of reads mapped to the human genome which suggests the clinical isolates were heavily contaminated with human DNA (average 74 million reads). Despite poor coverage in some of the isolates, the mapped reads would still able to provide insights about parasite population from these three malaria-endemic areas. Variant calling was callable at some sites even the samples had low coverage. Genotyping was suggested to be able to perform reliably at certain sites with known segregating variants (Winter *et al.*, 2015). Thus, all samples were included in the downstream analyses.

Table 2.5 Summary of sequencing data for 20 samples. The raw reads were mapped to the human genome (GRCh37) and the resulting unmapped reads were mapped to the *P. vivax* P01 reference genome using BWA-mem. The illumina adapter sequences were trimmed using Cutadapt version 1.2.1. Bases with window quality score of 20 were removed using Sickle version 1.2. Genome-wide mean coverage was estimated using *Qualimap 2*. [1]After adapter and quality trimming

Sample	Reads (million)[1]	Reads mapped to human	Proportion of reads mapped to human	Reads mapped to *P. vivax* P01	Proportion of reads mapped to *P. vivax* P01	Mean coverage
YL002G16	22,097,320	21,256,595	96.20	583,533	2.65	2.11
YL003G16	35,411,152	22,369,727	63.17	13,307,641	37.58	54.63
YL004G16	24,539,496	21,475,273	87.51	2,359,346	9.61	10.16
YL005G16	34.394,408	33,022,315	96.01	850,667	2.47	2.96
YL007G16	31,906,612	6,844,312	21.45	24,028,947	75.31	110.47
YL008G16	30,292,826	28,719,683	94.81	1,035,402	3.42	3.96
YL009G16	31,054,476	26,495,770	85.32	3,987,651	12.84	17.58
YL010G16	32,169,804	28,773,316	89.44	2,836,705	7.32	12.36
UB001G16	31,489,202	29,072,950	92.33	1,898,612	4.86	7.76
UB002G16	36,138,198	34,452,152	95.33	1,080,923	2.99	3.87
UB003G16	35,072,512	30,099,478	85.82	4,364,243	12.44	18.85
UB004G16	36,880,266	5,542,891	15.03	29,997,816	81.34	136.86
UB005G16	35,434,360	1,844,704	5.21	32,235,186	90.97	147.02
UB006G16	28,309,462	23,955,612	84.62	3,793,842	13.40	16.57
UB007G16	34,522,350	25,342,852	73.41	8,574,889	24.84	38.86
TAK001G16	35,599,990	32,406,132	91.03	2,781,068	7.81	11.96
TAK002G16	38,961,258	25,612,133	65.74	12,378,311	31.77	55.95
TAK003G16	30,069,032	29,139,898	96.91	481,104	1.60	3.41
TAK004G16	26,218,840	23,289,220	88.86	2,709,801	10.34	11.93
TAK005G16	35,144,328	31,791,202	90.46	2,829,512	8.05	12.11

2.3.2 Principal component analysis (PCA)

The main objective of the PCA was to analyse the geographical division of the *P. vivax* population in Thailand. The PCA plot was constructed using 24,524 LD-pruned SNPs (Figure 2.3). In total, 20 clinical isolates were used in the analysis, eight patients from Yala province, seven patients from Ubon Ratchathani province, and five patients from Tak province. Three distinct clusters were observed separating the 20 clinical isolates according to their geographical origin (Yala, Ubon Ratchathani, and Tak). A cluster contained clinical isolates from Ubon Ratchathani was located close proximity to Tak consistent with their geographical location (639 kilometres between two endemic areas). The present study analysed the population structure of *P. vivax* from Yala province (South of Thailand) for the first time. The first principal component and second principal component (PC) defined the greatest total variation of 16.5% and 7.8%, respectively. From Figure 2.3, these two PCs displayed population segregation of *P. vivax* according to their geographical location. One outlier from Yala province was identified on the top right of the PCA plot.

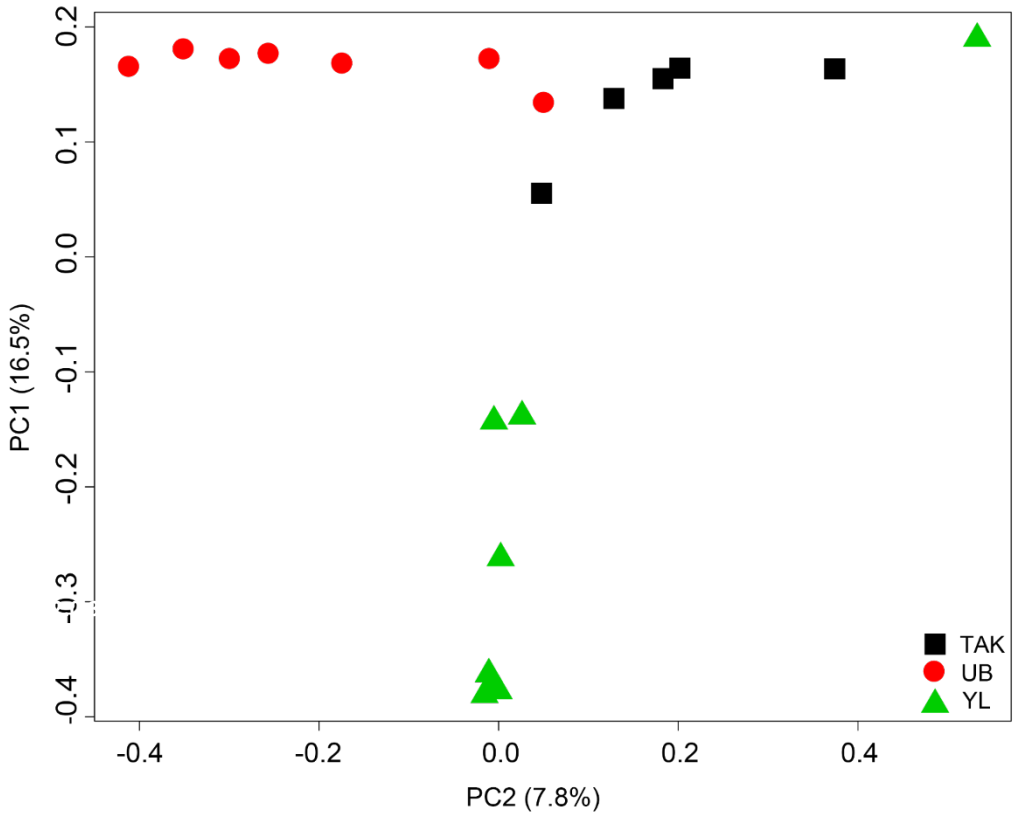


Figure 2.3 Principal component analysis of the 20 clinical isolates in Thailand. The analysis was based on the 24,524 LD-pruned SNPs using in PLINK version 1.9 (Purcell *et al.*, 2007). The PCA plot shows the geographical segregation of 20 clinical isolates according to their origins. Three distinct clusters were observed separating clinical isolates form Yala (YL), Ubon Ratchathani (UB), and Tak. The plot was generated with SNPRelate implemented in R (Zheng *et al.*, 2012). Each colour and shape of symbol denotes the respective endemic areas, black colour and square shape: Tak province, red colour and circle shape: Ubon Ratchathani province, and green colour and triangular shape: Yala province. Each clinical isolate is coloured according to the respective malaria-endemic area.

2.3.3 ADMIXTURE analysis

The ADMIXTURE analysis was conducted to assess *P. vivax* population structuring in Thailand. The most likely number of ancestral population (*K*) was estimated using a cross-validation method (CV). In the analysis, five *K* values were tested (*K* = 2 to 6). The optimal number of *P. vivax* population was achieved at the lowest cross-validation error (Figure 2.4). Based on the cross-validation error, *K* = 3 was identified as the least error separating the *P. vivax* in Thailand into three ancestral populations (Figure 2.5). At *K* = 3, the *P. vivax* population in Thailand separated into three subpopulations according to the geographical location: Yala province (South of Thailand), Ubon Ratchathani province (Northeast of Thailand), and Tak (Northwest of Thailand). Closer looks into the ADMIXTURE analysis in Figure 2.5, *P. vivax* in Yala province emerged as a distinct group from *K* = 2 to 6. However, one clinical sample from Yala province was seen to be admixed with Ubon Ratchathani and Tak province. This sample also appeared as an outlier in the PCA plot. On the other hand, *P. vivax* populations from Ubon Ratchathani and Tak province appeared to be more admixed suggesting the gene flow between two populations. This admixture pattern is consistent from *K* = 2 to 6. The geographical differentiation of *P. vivax* from ADMIXTURE analysis was in line with the PCA plot which further validates the finding.

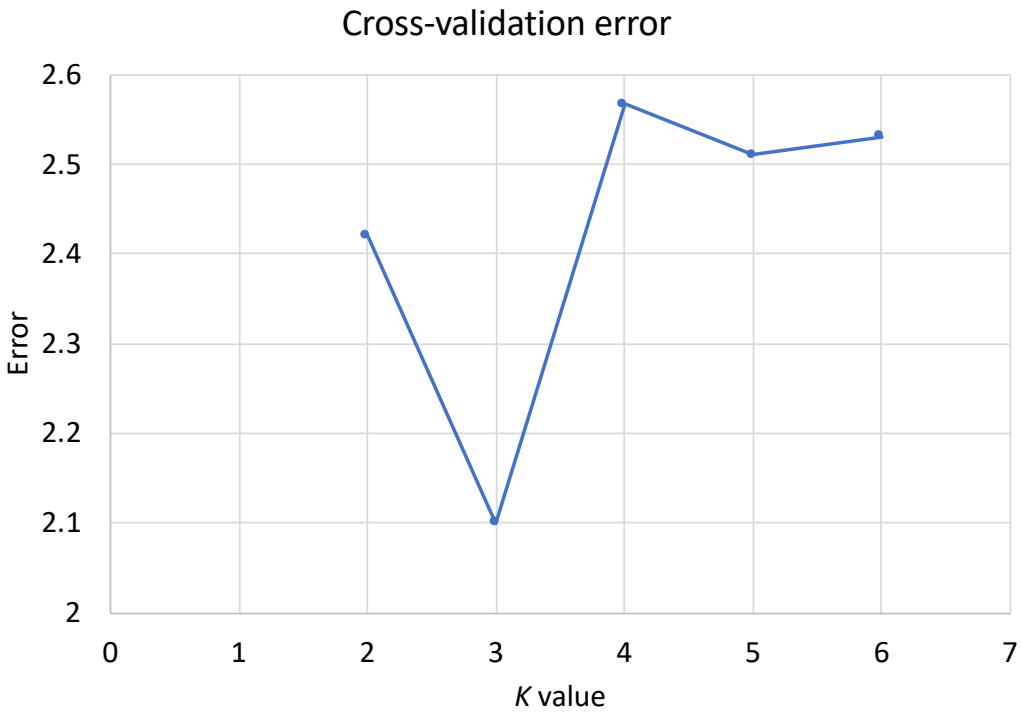


Figure 2.4 Cross-validation error (CV) to estimate the number of *P. vivax* population. The ADMIXTURE runs from $K = 2$ to 6 using the 247,789 SNPs from 20 clinical isolates. $K = 3$ revealed the lowest error and used to infer ancestral population of *P. vivax*.

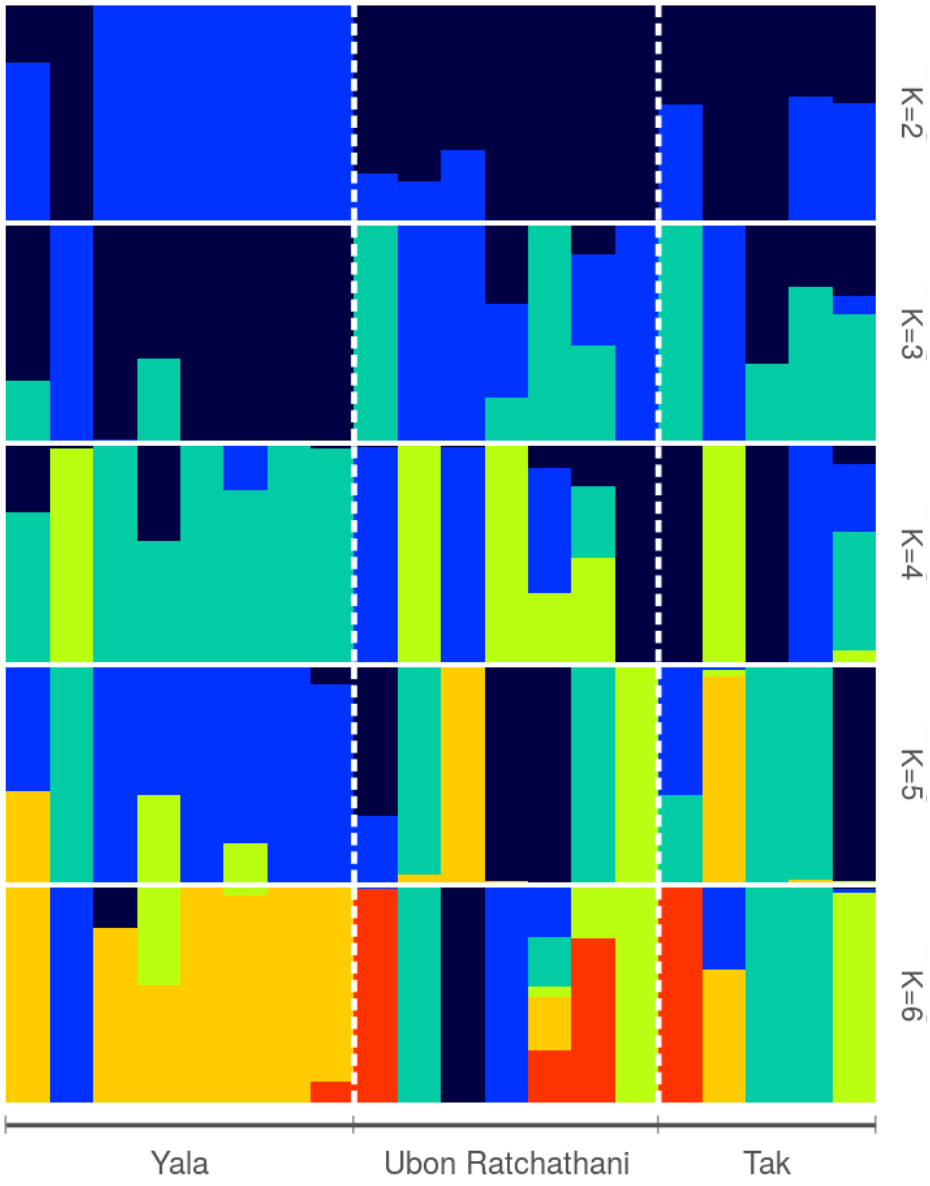


Figure 2.5 ADMIXTURE plots of *P. vivax* in three malaria endemic areas in Thailand. Five *K* values were plotted from *K* =2 to 6. *K* = 3 was used to infer the ancestral population as it achieved lowest CV error. A clear cluster is observed among the population from Yala province except an individual appeared to be admixed with population from the other two provinces. An admixed profile is observed between the populations from Ubon Ratchathani and Tak province implying the gene flow. In each plot, each population is represented by a different colour and each individual is represented by a vertical bar. The dotted white lines separate the individuals according to the geographical location. In total, 20 clinical isolates were included in the ADMIXTURE analysis.

2.3.4 Phylogeny analysis

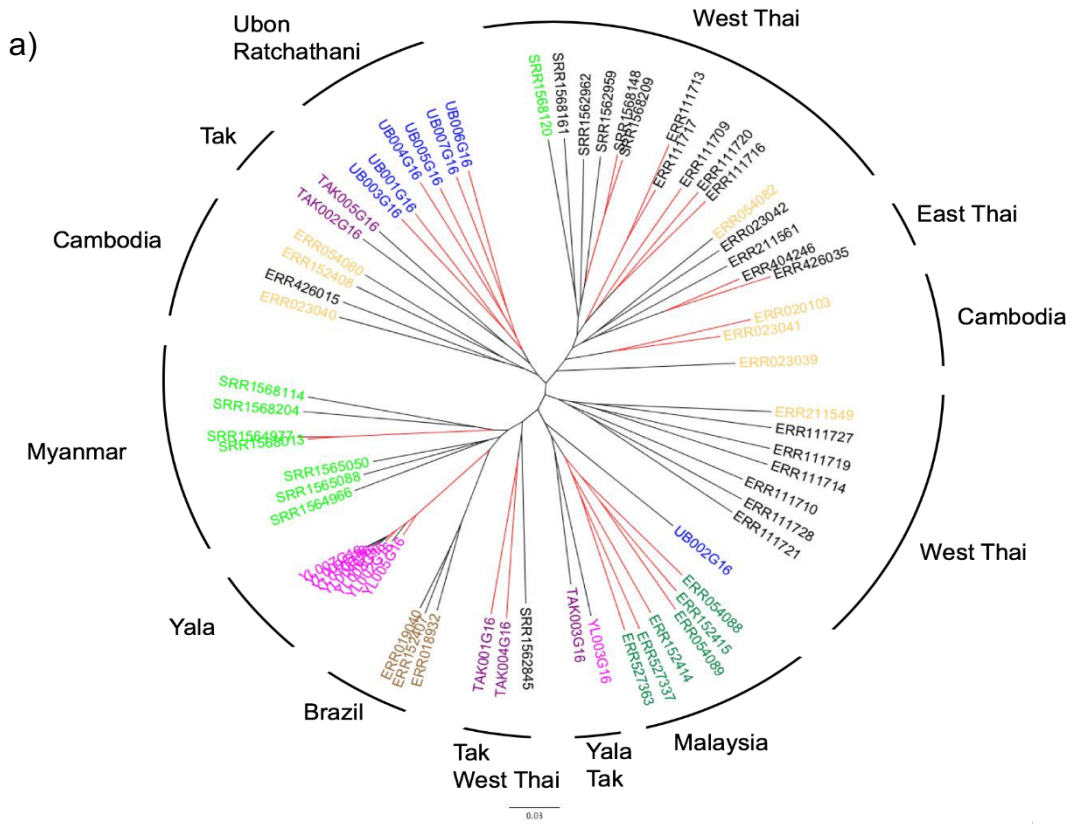
Two phylogeny models were constructed to validate the population structure of *P. vivax* in Thailand (Figure 2.6). A maximum likelihood tree was generated using RAxML version 8 (Stamatakis, 2014) (Figure 2.6a). The GTR substitution model with gamma rate variation was considered the best-fit model using jModelTest version 2.0 (Posada, 2008). In addition, the neighbour-joining tree was constructed in MEGA 7.0 (Kumar *et al.*, 2016) (Figure 2.6b). In total, 68 clinical isolates were used in the analysis where 20 genome sequences derived from the present study and 48 genome sequences retrieved from other studies (Hupalo *et al.*, 2016; Pearson *et al.*, 2016). Five isolates from Yala province were clustered together tightly in both maximum likelihood and neighbour-joining model. However, one isolate from Yala province (YL003G16) did not seem to reliably assign to a specific group, although a cluster was formed with an isolate from Tak province in maximum likelihood model. This observation is consistent with the PCA plot and the ADMIXTURE analysis where it appeared as an outlier. Furthermore, isolates from Yala province were too disparate from the isolates collected from Malaysia.

Furthermore, based on the maximum likelihood tree in Figure 2.6a, six isolates from Ubon Ratchathani province formed a high confidence branch (bootstrap support>80%). It also formed a cluster with the clinical isolates previously collected along the border (Cambodia). Two isolates from Tak province were also clustered with isolates from Cambodia and Ubon Ratchathani. One isolate from Ubon Ratchathani province failed to cluster with other isolates (UB002G16). One component on the maximum likelihood tree was seen to encompass clinical isolates from West Thailand, East Thailand, and Cambodia. Clinical isolates from Tak province did not form a distinct geographical cluster. Two isolates (TAK005G16 and TAK002G16) were clustered with Cambodia and Ubon Ratchathani whereas another two isolates (TAK001G16 and TAK004G16) were clustered with the sample previously collected in the same location (West Thailand). TAK003G16 located on the same branch as the isolate from Yala (YL003G16). Samples from Malaysia, Myanmar, and West Thailand were identified in several distinct branches. Meanwhile, in the neighbour-joining tree (Figure 2.6b) the clinical samples were clearly assigned to the respective branches according to geographical location. Clinical isolates from Ubon Ratchathani province,

East Thailand, and Cambodia were located on the distinctive cluster. Moreover, isolates from Tak province, West Thailand, and Myanmar can be observed on two clusters. Six isolates from Malaysia and three isolates from were assigned to a specific branch.

2.3.5 Genetic differentiation

Population differentiation between three endemic areas was estimated using the fixation index (F_{ST}). The F_{ST} was estimated by averaging in sliding windows, with a window size of 100 SNPs. The index ranges from 0 to 1 suggesting low genetic differentiation throughout the genome to complete genetic isolation between populations. The F_{ST} estimated across 24,524 LD-pruned SNPs revealed a low genetic differentiation between Tak and Ubon Ratchathani province (F_{ST} = 0.122, p-value > 0.05) implying gene flow between two endemic areas (Table 2.6). The F_{ST} values were relatively high between populations from Yala and Ubon Ratchathani F_{ST} = 0.297, p-value < 0.05), and between Yala and Tak (F_{ST} = 0.346, p-value < 0.05). These high F_{ST} values suggesting limited gene flow between these endemic areas. This finding is consistent with previous three analyses, the population of *P. vivax* is highly differentiated from the populations in the Northeast and Northwest province. In the PCA plot, admixture analysis, and phylogeny analysis, isolates from Yala province appeared to form a distinct cluster. The genetic divergence of Yala population is also consistent with the geographical location where it is more than 1000 km^2 away from Tak and Ubon Ratchathani province. The overall F_{ST} value between Tak and Ubon Ratchathani province was remarkably low (F_{ST} = 0.122, p-value > 0.05) suggesting modest geographical differentiation. The weak overall population structure was also revealed in the PCA plot, although two distinct clusters were observed, they were located next to each other. The admixture analysis and maximum likelihood tree also displayed an admixed relationship between two populations. Tak and Ubon Ratchathani province was located closer to each other compared to Yala province with approximately 639 km^2.


a)
Ubon Ratchathani
West Thai
Tak
East Thai
Cambodia
Cambodia
Myanmar
West Thai
Yala
Brazil
Malaysia
Tak
West Thai
Yala
Tak
0.03

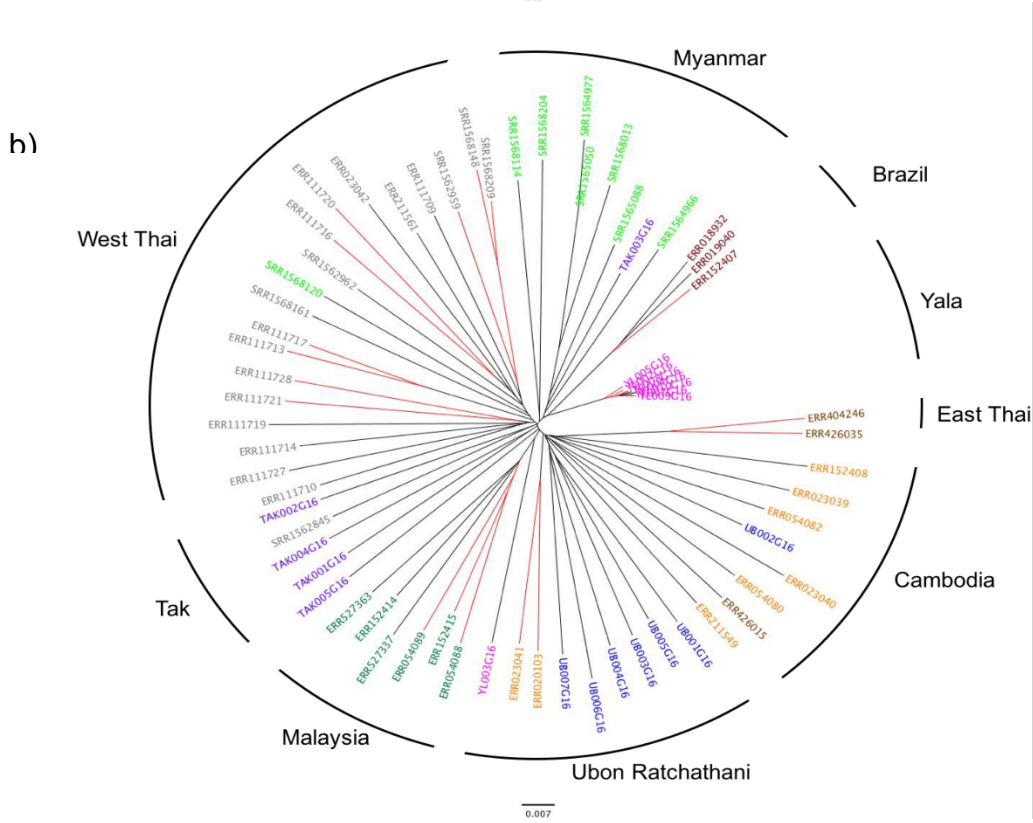

b)
Myanmar
Brazil
West Thai
Yala
East Thai
Tak
Cambodia
Malaysia
Ubon Ratchathani
0.007

Figure 2.6 Phylogeny trees of *P. vivax* in Thailand and other neighbouring countries. In total, 247,789 filtered SNPs were used to construct phylogenetic trees using two models, **a)** maximum likelihood model with 100 bootstraps and **b)** neighbour-joining tree with 500 bootstraps. The phylogeny trees show a distinct branch separating the *P. vivax* population in Yala from other two provinces (Ubon Ratchathani and Tak). Other previously studied isolates from the region and neighbouring countries were also added to the analysis. The branches clustering the endemic area and neighbouring country were less pronounced in the maximum-likelihood model. In the neighbour-joining tree, isolates are appeared to cluster with the isolates previously collected in the similar location and the bordering country such as Tak province with West Thai and Myanmar and Ubon Ratchathani province with East Thai and Cambodia. The maximum likelihood model was constructed using RAxML version 8 (Stamatakis, 2014) and the GTR substitution model with gamma rate variation was the best-fit model derived from the jModelTest version 2.0 (Posada, 2008). The neighbour-joining tree was constructed in MEGA 7.0 (Kumar *et al.*, 2016). Samples from each endemic area were coloured differently. 68 samples were included in the phylogeny analysis, 48 genome sequences were retrieved from the global study of *P. vivax* together with the 20 genome sequences derived in the present study. Branches with a bootstrap support of above 80% are highlighted in red.

Table 2.6 Genetic differentiation of *P. vivax* population in Thailand. The fixation index (F_{ST}) was estimated in SNPRelate (Zheng *et al.*, 2012) using 24,524 LD-pruned SNPs. Population differentiation is considered statistically significant when $p<0.05$.

Endemic area	Yala	Ubon Ratchathani	Tak
Yala	-		
Ubon Ratchathani	0.297*	-	
Tak	0.346*	0.122	-

2.4 Discussion

The population structure of *P. vivax* in Thailand was revealed using the whole-genome sequencing approach. A similar approach has been used to unravel the population structure of *P. vivax* globally (Hupalo *et al.*, 2016; Pearson *et al.*, 2016). In the present study, 20 clinical samples from diverse malaria endemic areas in Thailand have allowed me to identify the geographical divisions of vivax malaria in Thailand demographics. Patients infected with *P. vivax* were recruited from Yala province (South of Thailand), Ubon Ratchathani province (Northeast of Thailand), and Tak province (Northwest of Thailand). Using the SNP variants, the principal component analysis highlights a clear differentiation of *P. vivax* populations in Thailand according to their geographical location. Likewise, the ADMIXTURE analysis detected the population differentiation of *P. vivax* according to their geographical origin, although a certain degree of admixture population was observed in Ubon Ratchathani and Tak province. The distinct population differentiation of *P. vivax* will have an implication in the vaccine design strategy, highlighting the needs for careful consideration to develop an effective malaria vaccine.

The proportion of reads aligned to *P. vivax* PvP01 reference genome were inconsistent between 20 isolates, the percentage of mapped reads ranged from 2.65% to 81.34% (Table 2.5). This was stem from the high contamination of host DNA. *P. vivax* isolates always contain a high amount of host DNA due to its biology that infects only reticulocytes (Iyer *et al.*, 2007). Although the venous blood samples were filtered through C6288 cellulose columns to deplete host leukocytes (Venkatesan *et al.*, 2012), the approach has not been very effective in the study. Most of the reads mapped at least 80% to the human genome (GRCh37) (Table 2.5). However, two samples showed lower contamination of host DNA (YL007G17 and UB004G16). The coverage of the genome between each isolate was remarkable difference ranged from 2.11X to 147.02X (Table 2.5). To overcome this limitation, joint-variant calling implemented in GATK was used to identify the variants across samples (McKenna *et al.*, 2010). This approach has been described previously in the human genome with low sequence coverage (Jun *et al.*, 2015; Li *et al.*, 2010). The variant calling performed simultaneously across the samples showed to generate equally reliable dataset when compare with published variants (Jun

et al., 2015; Li *et al.*, 2010). This approach could increase the power to detect a concordant set of variants and reduce false-positive variant identification.

Geographical divisions of *P. vivax* populations in Thailand were identified by principal component analysis (Figure 2.3). Three distinct clusters distinguished the isolates from Southern Thailand (Yala province), Northeastern Thailand (Ubon Ratchathani province), and Northwestern Thailand (Tak province). This pattern is consistent with the geographical proximity of the endemic area and all of them are separated by a malaria-free corridor in the metropolitan cities (Suankratay *et al.*, 2001). Samples from Yala province were clustered distantly from other two populations and an outlier (YL003G16) was detected. This outlier could be due to the limited sample size or migration into Yala province with asymptomatic malaria. Samples from Ubon Ratchathani province and Tak province were loosely clustered together implying the presence of gene flow. Yala province is located further away from Ubon Ratchathani and Tak province, approximate 1681 km^2 and 1484 km^2, respectively. Ubon Ratchathani and Tak province are located closer to each other (approximate 639 km^2). The geographical distance is translated consistently with PCA plot produced.

The present study has a low sample size (n=20), to further describe the population structure of *P. vivax* around the bordering countries, 48 samples from two global studies were included in the PCA plot (Figure 2.7) (Hupalo *et al.*, 2016; Pearson *et al.*, 2016). As expected, most of the clinical isolates clustered according to the geographical region. Consistently, the clustering pattern is consistent with phylogeny trees. Isolates along the international borders were assigned to a specific group in PCA plot and phylogeny trees. Brazil isolates formed a distinct cluster in the PCA plot and phylogeny trees in line with the geographical location. On the other hand, isolates from Ubon Ratchathani province were clustered with samples from Eastern Thailand (EastThai) and Cambodia suggesting population movement along the border. Cross-border movement is not uncommon of between the Thai-Myanmar and Thai-Cambodia due to economics factor and political instability (Bhumiratana *et al.*, 2013; Guyant *et al.*, 2015). Therefore, the cross-border movement is a driver that homogenise the genetic structure of *P. vivax* along the border. Likewise, isolates from Tak province were clustered with samples from Western Thailand (WTH). However, isolates from Myanmar did not appear to cluster tightly with samples from Western Thailand and Tak

province. Myanmar isolates were collected from the Kachin State (Northmost state of Myanmar) which is distantly located from the Thai-Myanmar border (Hupalo *et al.*, 2016). Isolates from Yala province did not form a cluster with isolates from Malaysia. This pattern is stem from the sampling location of Malaysia isolates where they were collected from Peninsular Malaysia (approximate 4247 km^2 from Yala province) (Pearson *et al.*, 2016). As a whole, the pattern of population differentiation was in agreement with the global study, a major axis of differentiation was formed between western Southeast Asia, eastern Southeast Asia, and the Pacific Island (Hupalo *et al.*, 2016; Pearson *et al.*, 2016).

Admixture analysis identified lowest K value at $K = 3$ indicating the *P. vivax* populations from three malaria-endemic areas. Isolates from Yala province exhibits a high degree of homogeneity separating from the other two populations (Figure 2.5). The outlier identified in the PCA plot was admixed with population from Ubon Ratchathani province suggesting the patient could have migrated from the Northeast region. On the other hand, strong genetic differentiation was observed between Yala and Tak province (F_{ST}=0.346, p-value<0.05), and between Yala and Ubon Ratchathani province (F_{ST}=0.297, p-value<0.05). The phylogeny using two models (Figure 2.6) displayed a distinct branch with five Yala isolates. From the four analysis, *P. vivax* population from Yala province (Southern Thailand) was clearly distinguished from the populations in Ubon Ratchathani and Tak province. This pattern of differentiation was consistent with previously reported genes including TRAP and PvMSP-3 (Kittichai *et al.*, 2017; Kosuwin *et al.*, 2014). This may reflect the limited gene flow in the province. The malaria transmission in Yala province is mainly confined to the rubber plantation areas. Moreover, political unrest in the region since the year 2004 has hindered the migration to the area. Therefore, this factor has reduced the gene flow of the malaria parasite in Yala province.

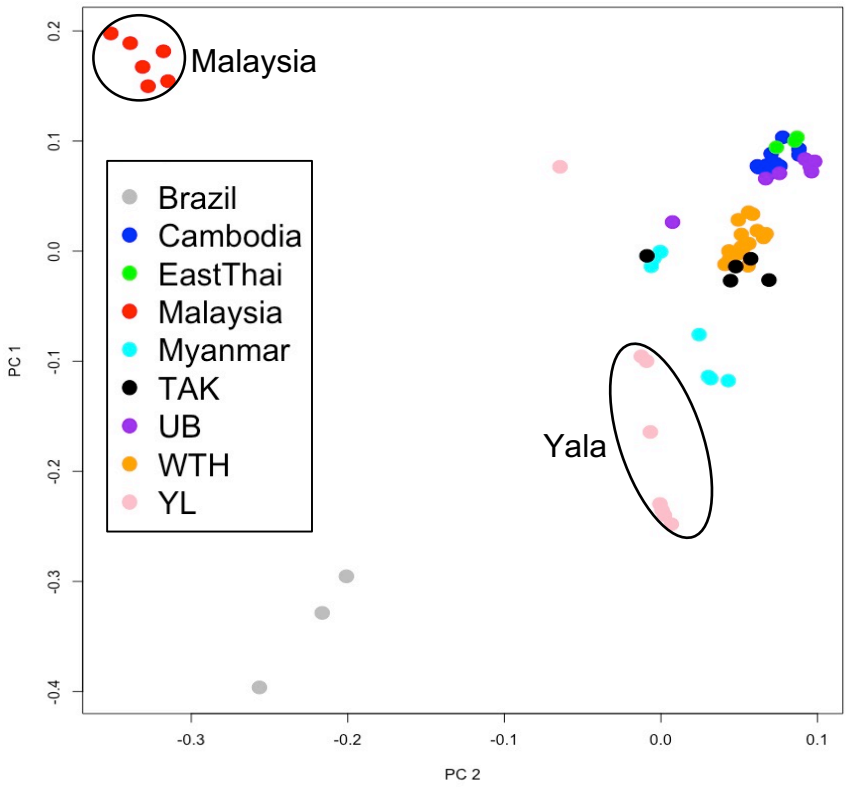


Figure 2.7 Principal component analysis of the 68 clinical isolates. The analysis was based on the 40,845 LD-pruned SNPs using in PLINK version 1.9 (Purcell *et al.*, 2007). The PCA plot shows the geographical segregation of 68 clinical isolates according to their origins. Each colour and shape of symbol denotes the respective endemic areas.

The admixed analysis between Ubon Ratchathani and Tak province revealed a degree of gene flow between two regions (Figure 2.5). This observation is in agreement with the PCA plot generated using clinical samples (Figure 2.3 and Figure 2.7). In addition, the low fixation index between two endemic areas supports the gene flow or genetic admixture. Population movement between Ubon Ratchathani and Tak province has been reported previously (Kosuwin *et al.*, 2014). Over the past few years, urbanisation in Ubon Ratchathani province has created employment opportunities which attract indigenous population from Tak province. Furthermore, political turmoil along the Thai-Myanmar border has displaced the residents from the province (Guyant *et al.*, 2015). Some of the indigenous population could harbour asymptomatic malaria and therefore, established gene flow between two areas

Given the population structure of *P. vivax* in Thailand, it will have a significant implication in vaccine design. Currently, a universally effective vaccine is yet to be achieved. The knowledge of the *P. vivax* population structure will guide the vaccine development to cover a higher diversity of haplotypes, relevant to the malaria vaccine efficacy either locally or globally (Barry and Arnott, 2014). The distribution of vaccine haplotypes is associated with the diversity of potential vaccine candidates. Genetic of malaria parasite is influenced by the local adaptation which translates into genetic isolation (Patz and Olson, 2006). In addition, malaria elicits allele-specific immunity which limits the protective effect in a large proportion of the population (Matuschewski and Mueller, 2007; Thera *et al.*, 2006). A cocktail vaccine contains multiple common haplotypes is likely to provide a higher protection against malaria in a large proportion of the world population. Therefore, investigating the population structure of *P. vivax* provides the basis for malaria vaccine development and predict the efficacy of a vaccine.

2.5 Conclusion

In summary, the current study presents the population structure of *P. vivax* in Thailand in parallel with the global studies. Three populations of *P. vivax* are identified in Thailand separating by their geographical location (Yala, South; Tak, Northeast; and Ubon Ratchathani, Northwest). The population structure is critical to assess the effectiveness of population in response to a malaria vaccine. Detailed consideration is required to deploy the vaccine development pipeline in Thailand as to achieve a universally effective vaccine. Now, this finding can be extended to the study of PvMSP-7 as a vaccine candidate which will be revealed in Chapter 3.

Chapter 3

Sequence diversity of *Plasmodium vivax* merozoite surface protein 7 (PvMSP-7) genes in Thai clinical isolates

Abstract

PvMSP-7 is a multigene family that expressed on the *Plasmodium* merozoite surface. Previous studies revealed the MSP-7 has evolved under a birth-and-death model and gene conversion was pronounced in some paralogous genes suggesting functional redundancy. Based on this finding, only certain PvMSP-7 paralogs should be considered in the vaccine development. Using the whole-genome sequences derived from the 20 clinical isolates in Chapter 2, the sequence diversity of the 13 PvMSP-7 paralogs in Thailand was uncovered. Furthermore, the structural variation in PvMSP-7 was examined. The results revealed not all PvMSP-7 paralogs showed the same extent of sequence variation owing to the functional differences. Some paralogous genes are conserved (PvMSP-7A, -7D, -7F, -7J, -7K, -7L, and -7M) while some are rather polymorphic (PvMSP-7B, -7C, -7E, -7G, -7H, and -7I). Structural variation was observed in all PvMSP-7 paralogs where the central region showed extensive sequence variation. Most sequence conservation was seen in the N- and C-terminal. In addition, evidence of intragenic recombination was more prevalence in the central region of PvMSP-7. The findings propose that the conserved PvMSP-7 genes and domains be used in the selection of malaria subunit vaccine development. Therefore, the PvMSP-7A, -7D, -7F, -7J, -7K, -7L, and -7M should be prioritised in the subunit vaccine design against malaria.

3.1 Introduction

Merozoite surface proteins have received great attention as vaccine candidates due to their role in erythrocyte invasion (Richards and Beeson, 2009). More recently, it has displayed inhibition effects on the blood-stage replication (Boyle *et al.*, 2013; Chandramohanadas *et al.*, 2014; Wilson *et al.*, 2015). The merozoite surface proteins are anchored to the membrane through glycosylphosphatidylinositol (GPI) membrane anchor and integral membrane anchor. The merozoite surface proteins (MSPs) localise to the merozoite surface prior to egress from the schizont. Once the MSPs ruptured from the schizont, the initial invasion is mediated by the GPI. Moreover, symptomatic malaria is caused by the blood-stage parasitaemia once merozoites are free from the schizont (Cowman *et al.*, 2012). Importantly, substantial malaria immunity targets the blood-stage antigens (Richards and Beeson, 2009). For this reason, MSPs are attractive malaria vaccine candidates. One hurdle in the development of the blood-stage vaccine is the antigenic diversity. Most of the important antigens displayed sequence variation as a mechanism to evade the host's immune system (Escalante *et al.*, 1998). However, the extent of polymorphism is varied in the antigen, some domains are conserved while some are rather diverse (Franks *et al.*, 2003). The conserved domains are likely to be a suitable malaria subunit vaccine. As such, understanding the sequence polymorphism is prime to develop MSPs in vaccine development. Merozoite surface protein 1 (MSP-1) and merozoite surface protein 3 (MSP-3) are two blood-stage vaccine candidates currently undergoing human vaccine trials (Bang *et al.*, 2011; Chitnis *et al.*, 2015).

The structural variation occurs in almost all the *Plasmodium* antigens. RTS,S is the most promising malaria vaccine candidate composed of the circumsporozoite protein (CSP) (Cohen *et al.*, 2010). The construction of RTS,S vaccine consists of polymorphic domains and conserved domains of CSP. The polymorphic regions are located in central and C-terminal of CSP while conserved domain located in N-terminal (Hughes, 1991; Jongwutiwes *et al.*, 1994; Putaporntip *et al.*, 2009c). Evidence of positive selection is found mainly in the C-terminal which potentially influencing the efficacy of the RTS,S vaccine. This polymorphism has shown to interrupt the T-cell reactivity to the specific epitope and influence the HLA binding (Takala and Plowe, 2009). This evidence demonstrated the antigenic variation will affect the outcome of

the vaccine design. Therefore, characterising the extent of sequence polymorphism of malaria antigens will resolve vaccine escape variants and improve vaccine efficacy.

Five multigene families in the whole-genome studies were reported to display high genetic diversity including MSP-3, variant interspersed repeat (VIR), MSP-7, serine repeat antigen (SERA), and reticulocyte-binding proteins (RBP) (Carlton *et al.*, 2004; Pearson *et al.*, 2016; Shen *et al.*, 2017). These families have a similarity where they are associated with host cell invasion and immune evasion. Intriguingly, MSP-7 was found under positive selection from the China-Myanmar border and the global *P. vivax* study involved 200 clinical isolates collected across the Asia-Pacific region (Pearson *et al.*, 2016; Shen *et al.*, 2017). The positive selection acting on the MSP-7 family is an indication that they are suitable vaccine candidates due to their functional importance. The functional antigens always under strong selection to evade the host immune system recognition (Chaurio *et al.*, 2016). The extensive genetic variation generates a pool of point mutations for selective pressure to act upon to avoid recognition by the host immunity (Goodswen *et al.*, 2018).

The PCR-based approach has been used to study several PvMSP-7 paralogs in Colombian isolates (Garzón-Ospina *et al.*, 2014; Garzón-Ospina *et al.*, 2012; Garzón-Ospina *et al.*, 2011). Heterogeneous sequence diversity was reported in the PvMSP-7 family, some paralogs demonstrated high level of sequence polymorphism (PvMSP-7C, -7E, -7H, and -7I) while some paralogs were rather conserved (PvMSP-7A, -7F, -7K, and -7L) (Garzón-Ospina *et al.*, 2014; Garzón-Ospina *et al.*, 2012; Garzón-Ospina *et al.*, 2011). The genetic diversity ranged from 0.0004 to 0.0039 in Colombian isolates. Likewise, a similar level of genetic diversity also reported in China-Myanmar border ranges between 0.0004 to 0.033 (Shen *et al.*, 2017). Closer looks into the sequence polymorphism along the protein, most of the polymorphic regions were concentrated in the central domain while N- and C-terminal were conserved (Garzón-Ospina *et al.*, 2014). The conserved regions are attractive to be incorporated in the malaria subunit vaccine design that likely to confer universal immunity. In view of the differences in genetic diversity between PvMSP-7 paralog, in-depth evaluation of sequence polymorphism is required to pinpoint the most promising paralog and the region into malaria vaccine development.

Using the phylogeny analysis, MSP-7 displays an uneven copy number across different *Plasmodium* lineages (Castillo *et al.*, 2017; Garzón-Ospina *et al.*, 2010). This difference has been proposed likely due to adaptations to human host (Castillo *et al.*, 2017). The evolution of MSP-7 also seen to consistent with a birth-and-death-model (Garzón-Ospina *et al.*, 2010). That said, duplication, pseudogenization, and gene loss events are common in MSP-7 evolutionary history. Moreover, MSP-7 paralogs in *P. vivax* have diverged from their orthologs in non-human primates' stem from the episodic positive selection pressure (Castillo *et al.*, 2017). Therefore, PvMSP-7 family might have undergone selection in parallel with the *P. vivax* lineage divergence from the Asian non-human primates (Castillo *et al.*, 2017; Garzón-Ospina *et al.*, 2010). These lines of evidence suggesting that the PvMSP-7 is functionally important.

Several MSP-7 paralogs in *P. falciparum* have shown to prime the erythrocyte invasion (Kauth *et al.*, 2006). They formed a multiprotein complex with MSP-1 prior to host cell invasion (Lin *et al.*, 2016). In addition, several pieces of evidence suggest that the MSP-7 could be playing redundant function (Kadekoppala and Holder, 2010). Based on the phylogeny analysis performed by Castillo and colleagues using Bayesian and maximum-likelihood approaches, MSP-7 paralogs were divided into three major groups from A to C (Castillo *et al.*, 2017). Group A consists of PvMSP-7A and PvMSP-7K while Group B comprises most of the PvMSP-7 paralogs (Castillo *et al.*, 2017). Interestingly, the PvMSP-7 paralogs in Group B displayed high sequence similarity suggesting gene duplication or gene conversion (Castillo *et al.*, 2017). On the other hand, Group C contains PvMSP-7L and PvMSP-7F (Castillo *et al.*, 2017). Consistently, the phylogenetic relationships are in line with the finding reported earlier by Garzón-Ospina and colleagues in 2010 (Garzón-Ospina *et al*., 2010). The complex evolutionary history in MSP-7 family indicates that careful evaluation is sought during vaccine development due to functional redundancy from MSP-7 proteins.

The efficacy of the vaccine is driven by the antigenic diversity; therefore, it is of utmost importance to characterise the level of antigenic variation of PvMSP-7 in different endemic areas and countries. To date, no study has addressed the antigenic diversity of PvMSP-7 in Thailand. In this chapter, sequence diversity of PvMSP-7 was revealed in three malaria major endemic areas in Thailand and pinpoint the specific paralogs to be included in the malaria subunit vaccine development.

3.2 Methodology

Experiment design, molecular diagnosis of the clinical samples, and bioinformatics processing were described in Chapter 2.

3.2.1 Multiple sequence alignment

The 13 PvMSP-7 sequences were retrieved from the 20 clinical isolates sequenced using the whole-genome approach described in Chapter 2. FASTA sequences were generated from 20 *P. vivax* infected individuals by using GATK FastaAlternateReferenceMaker (DePristo *et al.*, 2011). This command generates an alternative reference sequence replacing the reference bases at the variable sites with the bases derived from the variant calling. All the variant sites were assumed to be true variants as the dataset was filtered using stringent parameters. The FASTA files representing one PvMSP-7 paralog in each sample. DNA sequences were aligned using CLUSTAL W version 2.0 (Thompson *et al.*, 1994) against the PvP01 *P. vivax* reference genome (Auburn *et al.*, 2016). The accession number of the PvMSP-7 in PlasmoDB database is listed in Table 1.1 of Chapter 1. In the sequence alignment, all sites that postulated a gap were removed.

3.2.2 Genetic diversity

Population genetic metrics including genetic diversity and haplotype diversity were calculated using DnaSP v5 (Librado and Rozas, 2009). The genetic diversity was computed based on the principle illustrated in equations 10.5 or 10.6 (Nei, 1987). Nucleotide diversity (π) calculates the average number of nucleotide substitutions per site in each pairwise sequences. Sampling variance was calculated based on equation 10.7 and the square root of the indices to obtain standard deviation. Haplotype diversity was computed from aligned nucleotide sequences according to equations 8.4 and 8.12 (Nei, 1987). However, $2n$ was replaced with n with the assumption that the two populations experienced a similar evolutionary sampling procedure. Following

common practice, the standard error was calculated from the square root of the variance. Insertions and deletions in the sequence alignment were omitted from all estimates.

3.2.3 Tandem repeat detection

Tandem Repeats in each PvMSP-7 sequence was scanned iteratively using two different algorithms, mreps (Kolpakov *et al.*, 2003) and Tandem Repeat Finder (Benson, 1999). Both applications are available through public server interface (http://mreps.univ-mlv.fr and https://tandem.bu.edu/trf/trf.html). mreps is a sophisticated algorithm to identify repeat structures in the sequence alignments. The software finds the repeat using a permutation of combinatorial and heuristic approach. Repeat fragments are verified with mathematical paradigms and biologically relevant representations (Kolpakov *et al.*, 2003). Tandem Repeat Finder screened through the nucleotide alignments to find perfect and imperfect tandem repeats. The default setting was used in the procedure, tandem repeats were identified with the proportion of identity and pattern of repetition.

3.2.4 Natural selection

Selective pressures acting on individual sites of codon alignments were assessed using the Datamonkey web server (Pond and Frost, 2005). Five complementary methods were used to infer the selective pressures including single-likelihood ancestor counting (SLAC), fixed effects likelihood (FEL), internal branch FEL (iFEL), random effects likelihood (REL), and fast unconstrained Bayesian approximation (FUBAR). Statistical significance level (*p*-value) was based on the recommendation supplied by the programme (Pond and Frost, 2005). To minimise the false positive detection, significant codons were selected based on the consensus from at least two methods with *p*-value <0.1 for SLAC, FEL, and IFEL, FUBAR Posterior Probability >0.9, and/or REL Bayes Factor >50. SLAC has a more conservative approach in detecting selective pressure acting on each codon site. Hence, the false positive rate is relatively lower. It uses the maximum-likelihood and counting methods to calculate the rate of

synonymous substitutions per synonymous site (d_S) and the rate of nonsynonymous substitutions per nonsynonymous site (d_N) (Pond and Frost, 2005). FEL has a similar approach with SLAC, however, this approach assumes the selective force on each site is consistent. FEL uses an MG94xREV approach to infer d_N and d_S rate to each codon site (Kosakovsky Pond and Frost, 2005). Likewise, IFEL approach was used at the population level to assess if the sequences have been exposed to selective forces (Kosakovsky Pond and Frost, 2005). On the other hand, FUBAR identifies the d_N and d_S rate per codon site based on the Bayesian approach. The posterior probabilities above 0.9 indicate strong positive selective pressure (Murrell *et al.*, 2013). REL uses the enhanced Nielsen-Yang principle to detect codon sites under selection. This method has the flexibility to account for the nucleotide substitution biases, therefore, reduce the Type I error (Pond *et al*., 2011).

3.2.5 Recombination

Two approaches were used to uncover the intragenic recombination signals in 13 PvMSP-7 paralogs including calculation of the minimum recombination events (RM), and searching for recombination breakpoints through the genetic algorithm recombination detection (GARD) (Kosakovsky Pond *et al.*, 2006; Rozas and Rozas, 1997). The RM approach is implemented in the DnaSP version 5.0 (Librado and Rozas, 2009). The number of recombination event was identified through the Hudson 1987 approach (Hudson, 1987). It uses R=4Nr equation, where N refers to the population size and r is the recombination rate per sequence or between the contiguous sites. In addition, GARD was used to verify the recombination evidence present in the population sequence. It is implemented in the Datamonkey web server (Pond and Frost, 2005). GARD uses the phylogenetic inference on top of the appropriate model of nucleotide substitution to scan for recombination evidence. The goodness of fit between recombinant and nonrecombinant models was evaluated by AIC to support the recombination signals (Kosakovsky Pond *et al.*, 2006). Sequence segments were deemed statistically significant if *P*-value less than 0.05 using method of Shimodaira and Hasegawa (1999).

3.3 Results

3.3.1 Genetic diversity in PvMSP-7

In the present study, 20 clinical samples were used to assess the nucleotide diversity in the PvMSP-7 multigene family. Heterogenous genetic diversity was observed in PvMSP-7 paralogs (Table 3.1). The nucleotide diversity (π) ranged from 0.001 to 0.057. Four PvMSP-7 paralogs (PvMSP-7A, -7F, -7J, and -7L) have the lowest nucleotide diversity (π=0.001) whereas three PvMSP-7 paralogs have higher nucleotide diversity (PvMSP-7D π=0.002, PvMSP-7K π=0.003, and PvMSP-7M π=0.007). Furthermore, antigenic variation in six PvMSP-7 paralogs was more prevalent (PvMSP-7B π=0.021, PvMSP-7C π=0.021, PvMSP-7E π=0.057, PvMSP-7G π=0.056, PvMSP-7H π=0.045, and PvMSP-7I π=0.029). A boxplot represents the genetic diversity across 13 PvMSP-7 members is shown in Figure 3.1.

The structural variation in PvMSP-7 paralogs is illustrated in Figure 3.2. The nucleotide diversity was plotted on the per base basis to visualise the polymorphic region and conserved region spanned along the PvMSP-7 members. Notably, the central domain of the PvMSP-7 seems to harbour a high level of nucleotide diversity. Three peaks in the central region were observed correspond to PvMSP-7E, PvMSP-7H, and PvMSP-7G. On the other hand, the N- and C-terminal of the PvMSP-7 was relatively conserved. A small peak of nucleotide diversity was detected in the N-terminal correspond to PvMSP-7C, -7E, and -7G. Likewise, a peak was observed at the C-terminal of PvMSP-7B and PvMSP-7G (Figure 3.2).

In total, 20 haplotypes were found for PvMSP-7H while 19 haplotypes were found for PvMSP-7G (Table 3.1). In addition, 18, 17, and 16 haplotypes were identified from five PvMSP-7 paralogs (PvMSP-7B, -7E, -7I, -7C, and -7M). Also, 10, 8, 6, and 3 haplotypes were found for PvMSP-7A, -7F, -7K, -7L, -7D, and -7J (Table 3.1). Haplotype diversity (h) was higher in seven PvMSP-7 paralogs with h>0.900. Lower haplotype diversity was observed in five PvMSP-7 paralogs (PvMSP-7A h=0.853, PvMSP-7D h=0.674, PvMSP-7F h=0.711, PvMSP-7J h=0.195, PvMSP-7K h=0.884, and PvMSP-7L h=0.821).

Repeats in the PvMSP-7 paralogs were deduced using mreps (Kolpakov *et al*., 2003) and Tandem Repeat Finder (Benson, 1999). Repeat motifs were found in most of the PvMSP-7 paralogs. However, no repeats were detected in PvMSP-7D, -7E, -7H, -7J, -7L, and -7M. Two imperfect tandem repeats characterized by consensus repeats 'CAGCCGCAC' and 'GCAGCTGCA' were found in the central region of the PvMSP-7A. Meanwhile, in PvMSP-7B, one imperfect tandem repeat spanning C-terminal of the gene was detected with a sequence 'CCCCCTGCACCAGGACATCCACAAGCG'. One tandem repeat was found in the N-terminal of PvMSP-7C by a consensus sequence 'TTTCCTTTTCCTTTT'. In PvMSP-7F, a repeat characterised by a consensus sequence 'GAAGAAGCGGAGGAAGAAGCGGGGGA' spanned between nucleotide position 433-458. Two imperfect tandem repeats were detected in the N-terminal and central region of PvMSP-7G. These repeats were characterised by consensus sequences 'GAAGAGGAAGAGGAAGA' and 'GGAGGAGGAGGA'. Two tandem repeats were found in the central region of PvMSP-7I by consensus sequences 'AAGAAGAAGAAGA' and 'GAAGCAGAAGCAGAAGCAG'. Lastly, an imperfect tandem repeat was detected in the N-terminal of PvMSP-7K characterised by 'AGTGAGGGCCCCGCAAACATG'.

Table 3.1 DNA polymorphism measurement of PvMSP-7 sequences. The estimates were derived from 20 clinical isolates collected from three malaria endemic areas in Thailand.
n: number of isolates, Ss: number of segregating sites, S: number of singleton sites, Ps: number of parsimony-informative site, H: number of haplotypes, h: haplotype diversity, π: nucleotide diversity, (S.D): standard deviation

PvMSP-7	A	B	C	D	E	F	G	H	I	J	K	L	M
n	20	20	20	20	20	20	20	20	20	20	20	20	20
Sites	1134	1332	1188	531	1119	1236	1371	1197	1173	285	972	1263	1170
Ss	9	119	77	4	156	8	244	143	118	3	8	9	26
S	6	44	16	2	11	6	45	16	16	3	0	7	7
Ps	3	75	61	2	145	2	199	127	102	0	8	2	19
H	10	18	16	6	18	10	19	20	17	3	10	8	16
h (S.D)	0.853 (0.063)	0.989 (0.019)	0.974 (0.025)	0.674 (0.100)	0.984 (0.024)	0.711 (0.113)	0.995 (0.018)	1.000 (0.016)	0.979 (0.024)	0.195 (0.013)	0.884 (0.045)	0.821 (0.004)	0.974 (0.000)
π (S.D)	0.001 (0.000)	0.021 (0.003)	0.021 (0.003)	0.002 (0.000)	0.057 (0.003)	0.001 (0.000)	0.056 (0.007)	0.045 (0.003)	0.029 (0.003)	0.001 (0.000)	0.003 (0.000)	0.001 (0.000)	0.007 (0.001)

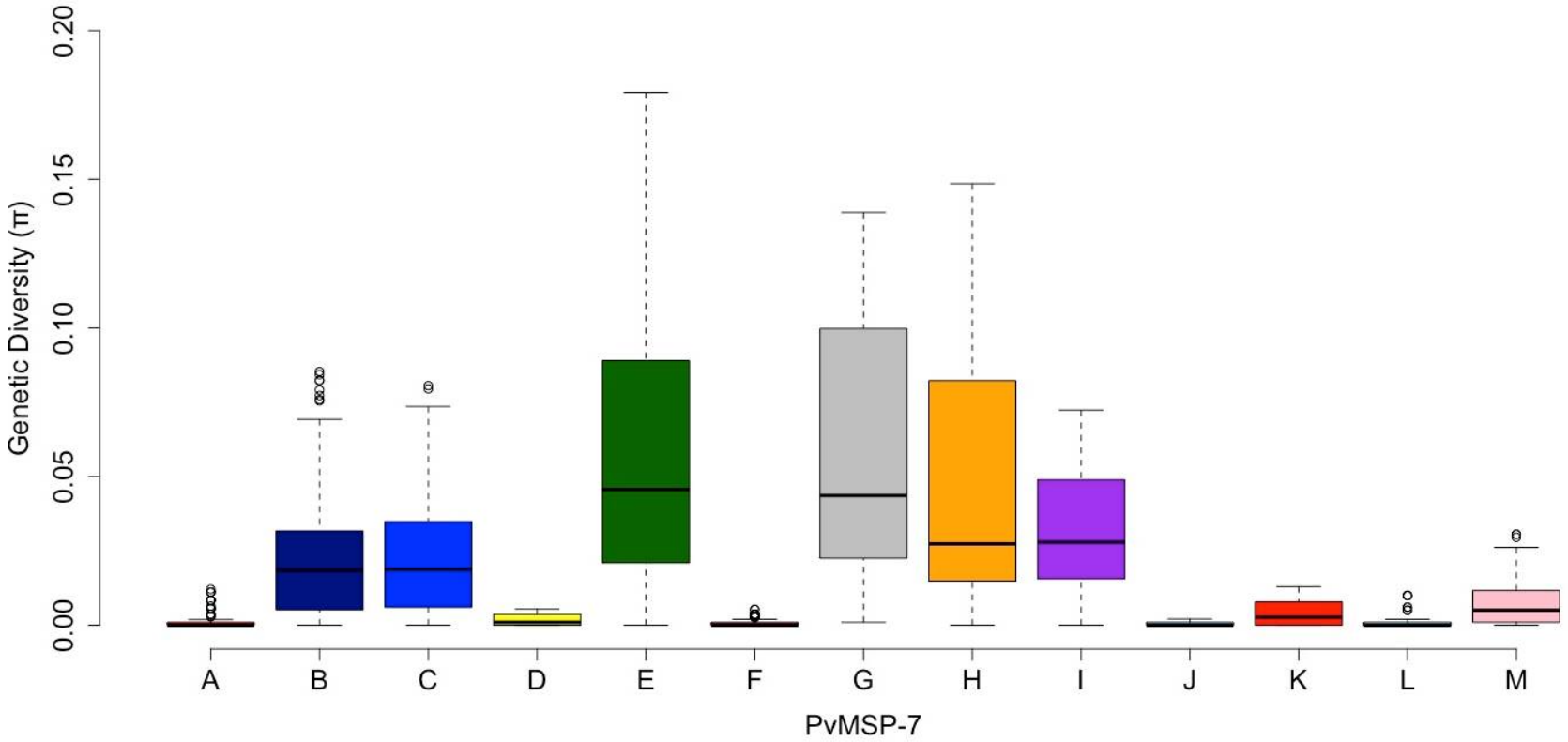


Figure 3.1 Boxplot of nucleotide diversity (π) for the 13 PvMSP-7 paralogs. The estimates were based on the average number of nucleotide substitutions per site in each pairwise sequences. All sites that postulated a gap were omitted in the analysis. The nucleotide diversity was derived from the DnaSP version 5.0. PvMSP-7 paralogs are represented by the different colour scheme.

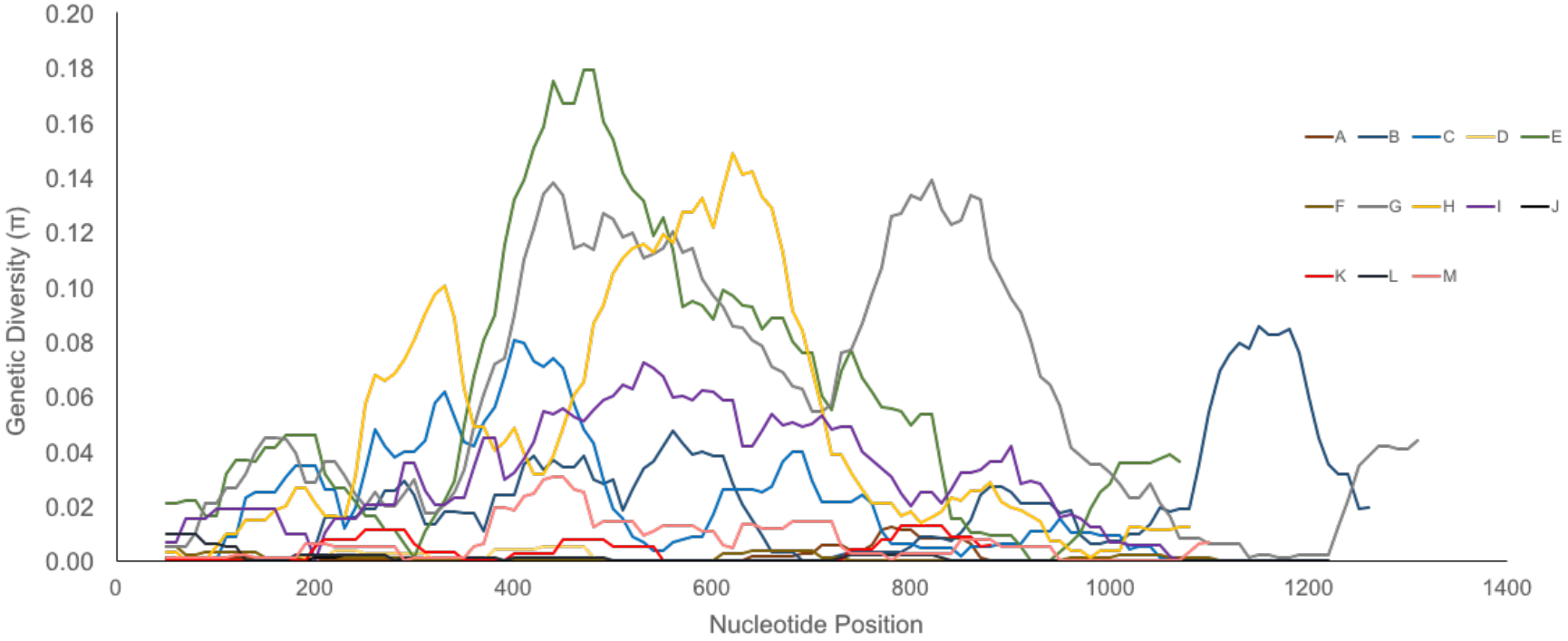


Figure 3.2 Structural variation of PvMSP-7 family. Nucleotide diversity (π) per site was estimated for each PvMSP-7 paralog. The π was derived from sequence alignment of 20 clinical isolates from Thailand. Each PvMSP-7 member is represented by a different colour in the figure.

3.3.2 Natural selection

Codon-based substitution models were used to measure the selective forces acting along the PvMSP-7 proteins. The positively selected residues might responsible for the host immunity evasion. Evidence of selective pressure on each codon of the PvMSP-7 multigene family member was tested using five approaches including SLAC, FEL, iFEL, REL, and FUBAR. Selective pressure was observed to act differently along the PvMSP-7 genes. Most of the positive selection signals were distributed in the central region of PvMSP-7 genes (Figure 3.3). PvMSP-7M was seen to have a higher frequency of positive selection signals compare to other members. In total, 15 codons were under strong positive selection signals spanning N-, central, and C-terminal. On the other hand, a lower number of positively selected residue was detected in PvMSP-7F and PvMSP-7L where only one selected site was found in codon 6 and codon 11, respectively. PvMSP-7A and PvMSP-7E had two codons under positive selection. In addition, positive selection signal was detected in the codons toward the C-terminal region in PvMSP-7B, -7C, -7D, -7G, and -7K. Meanwhile, no selective pressure was detected in PvMSP-7J.

Using the similar codon-based tests for departure from neutrality (SLAC, FEL, iFEL, REL, and FUBAR), residues under negative selection in the PvMSP-7 multigene family were identified (Figure 3.4). All the negatively selected sites were distributed outside the central domain of PvMSP-7 genes. It is noteworthy that these negatively selected residues were predominantly near the conserved N- and C-terminal. Overall, PvMSP-7H had 29 residues under negative selection while PvMSP-7G had 18 negatively selected sites (predicted by at least two approaches). Negatively selected codons were found for PvMSP-7B, -7C, -7E, and -7I ranged from 13 to 19 sites. Only one negative selected codon was identified in PvMSP-7D, -7K, -7L, and -7M. On the other hand, no negatively selected codons were detected in three PvMSP-7 paralogs (PvMSP-7A, -7F, and -7J).

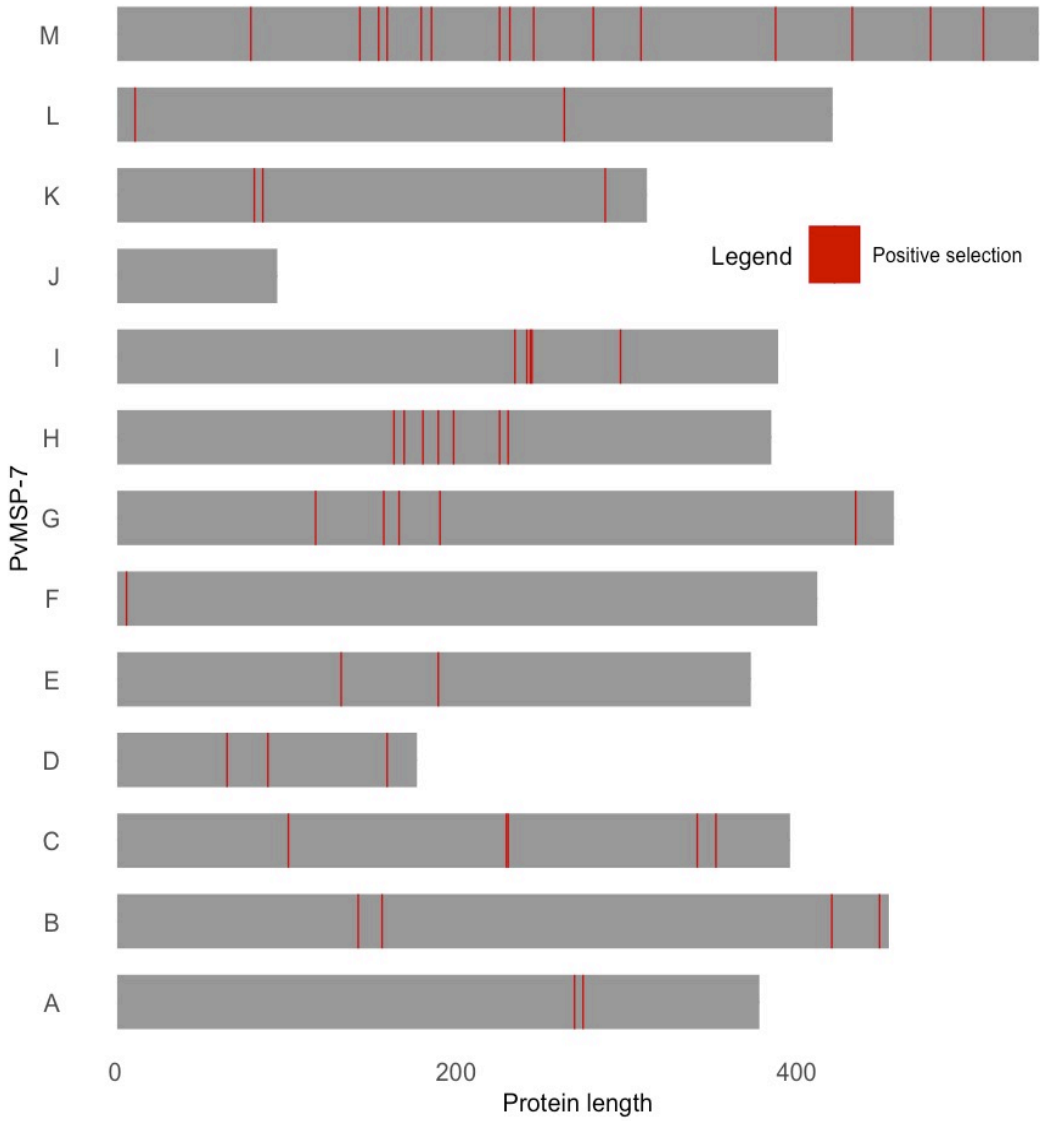


Figure 3.3 Positively selected codon sites in PvMSP-7 paralogs. The analysis was conducted with consideration of recombination. Five algorithms were used to infer the recombination position including Single-Likelihood Ancestor Counting (SLAC), Fixed effects likelihood (FEL), Random Effects Likelihood (REL), Internal Fixed effect likelihood (IFEL), and Fast Unconstrained Bayesian AppRoximation for inferring selection (FUBAR). Codons were selected based on the consensus derived from at least two methods with *p*-value <0.1 for SLAC, FEL, and IFEL, FUBAR Posterior Probability >0.9, and/or REL Bayes Factor >50. The red line represents the codon under positive selection and the PvMSP-7 proteins are drawn to the protein length.

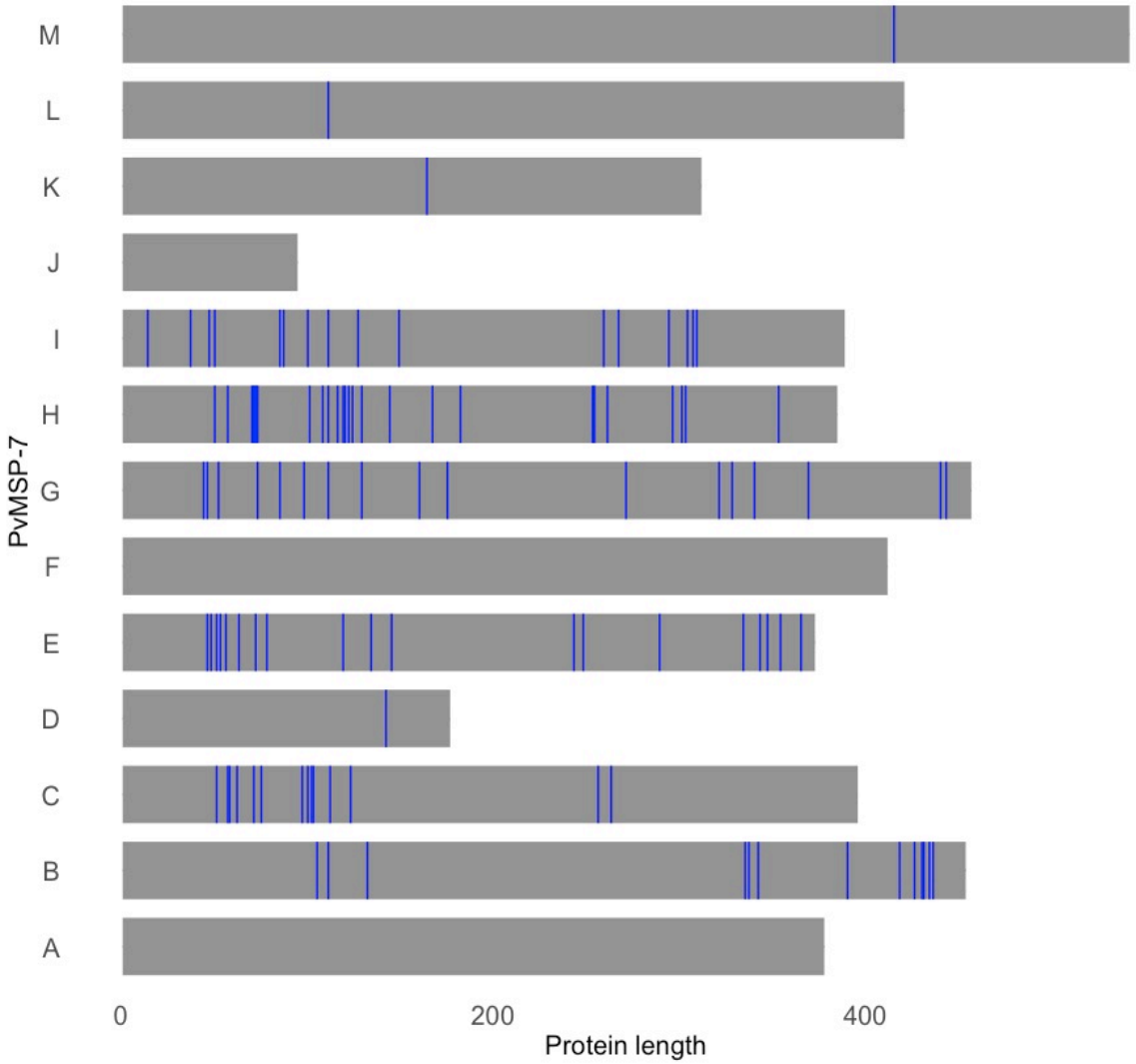


Figure 3.4 Negatively selected codon sites in PvMSP-7 paralogs. The codons under negative selection were detected using five approaches implemented in the Datamonkey web server (SLAC, FEL, REL, IFEL, and FUBAR). Negatively selected codons that achieved statistically significant from at least two methods were reported. The significant threshold (*p*-value) was used as recommended in the web server, *p*-value <0.1 for SLAC, FEL, and IFEL, FUBAR Posterior Probability >0.9, and/or REL Bayes Factor >50. The blue line indicates the codon under negative selection and the PvMSP-7 proteins are drawn according to the protein length.

3.3.3 Recombination

Genetic recombination is known to account for the majority of the variations observed in malaria antigens. Putative recombination positions were evaluated by two algorithms, such as minimum recombination events (RM) implemented in the DnaSP version 5.0 and genetic algorithm recombination detection (GARD) in Datamonkey web server. Consensus recombination positions from the two algorithms were presented herein (Table 3.2). Nine PvMSP-7 paralogs found to constitute of at least one recombination site. Notably, these recombination sites were distributed in the central domain of the genes. No recombination evidence was found in four PvMSP-7 paralogs (PvMSP-7A, -7F, -7J, and -7L).

Table 3.2 Significant recombination position detected in 13 PvMSP-7 genes. The recombination position was identified using the minimum recombination events (RM) and searching for recombination breakpoints through the genetic algorithm recombination detection (GARD). The RM is implemented in the DnaSP version 5.0 while the GARD is available from the Datamonkey web server. The reported recombination positions were detected in two algorithms. The recombination position was considered statistically significant with *p*-value<0.05.

PvMSP-7	Recombination position	*p*-value
A	No evidence	>0.05
B	632, 1130, 1197	<0.001
C	336, 599	<0.001
D	226	<0.001
E	356, 467, 600, 807	<0.001
F	No evidence	>0.05
G	349, 511, 691, 863	<0.001
H	321, 593, 692, 809	<0.001
I	334, 419, 523, 694, 801, 947	<0.001
J	No evidence	>0.05
K	540	<0.001
L	No evidence	>0.05
M	343, 574	<0.001

3.4 Discussion

An ideal vaccine candidate should show a high degree of sequence conservation to confer effective immune responses (Thera and Plowe, 2012). In this analysis, the antigenic diversity of PvMSP-7 and the potential vaccine candidates against malaria was described. Heterogeneous nucleotide diversity of PvMSP-7 in Thailand was identified from the 20 clinical samples. Some members demonstrated extensive sequence polymorphism (PvMSP-7B, -7C, -7E, -7G, -7H, and -7I) whilst some members are rather conserved (PvMSP-7A, -7D, -7F, -7J, -7K, -7L, and -7M). Structural variation was displayed in all PvMSP-7 paralogs. The central region of PvMSP-7 was shown to harbour high sequence variation while the N- and C-terminal was rather conserved. Codon-based substitution models identified most of the positively selected residues were distributed in the central region. On the other hand, negatively selected codons were predominantly located in the N- and C-terminal. Evidence of recombination was also more prevalent in the central region of the protein. Taken together all the findings, it is very encouraging that the conserved PvMSP-7 paralogs and the conserved region can be incorporated in the malaria blood-stage subunit vaccine development.

In the present study, sequence analysis of 13 PvMSP-7 paralogs has shown a marked difference in the nucleotide diversity. Six PvMSP-7 members (PvMSP-7B, -7C, -7E, -7G, -7H, and -7I) showed higher sequence variation. In contrast, seven PvMSP-7 paralogs (PvMSP-7A, -7D, -7F, -7J, -7K, -7L, and -7M) displayed limited sequence polymorphism. It is worth noting that this finding is in line with the studies conducted in Colombian isolates (Garzón-Ospina *et al.*, 2014; Garzón-Ospina *et al.*, 2012; Garzón-Ospina *et al.*, 2011) and other samples collected from multiple geographic regions (Castillo *et al.*, 2017). This is likely to stem from the different selective pressures or biological constraints acting upon PvMSP-7 members. Castillo and colleagues have indicated that several PvMSP-7 paralogs undergone episodic selection in their divergence from *P. cynomolgi* and might have promoted the genetic variability among the paralogous genes (Castillo *et al.*, 2017).

The data show that PvMSP-7E is the most diverse paralog among the PvMSP-7 multigene family. The level of nucleotide diversity (π=0.057) is identical with the study conducted in Colombia with 35 clinical sequences (Garzón-Ospina *et al.*, 2014).

Likewise, PvMSP-7B, -7C, -7G, -7H, and -7I displayed a high magnitude of sequence polymorphism similar to those of Colombian isolates and other malaria-endemic areas (Castillo *et al.*, 2017; Garzón-Ospina *et al.*, 2014; Garzón-Ospina *et al.*, 2012). The antigenic variation estimated was comparable to those malaria surface proteins including PvMSP-1 (Putaporntip *et al.*, 2002), PvMSP-3 (Rice *et al.*, 2014), PvMSP-5 (Putaporntip *et al.*, 2010), and AMA-1 (Escalante *et al.*, 2001). Despite the extensive level of nucleotide polymorphism observed in these malaria antigens, vaccine development could be focused on the conserved domain. In PvMSP-7, the vaccine development could be entailed the conserved region on the N- and C-terminal. PvMSP-7A, -7D, -7F, -7J, -7K, -7L, and -7M revealed a high degree of sequence conservation consistent with Colombian isolates and worldwide clinical sequences (Castillo *et al.*, 2017; Garzón-Ospina *et al.*, 2014; Garzón-Ospina *et al.*, 2012). Although these conserved PvMSP-7 paralogs are promising vaccine candidates, appropriate consideration is needed to avoid the probability of vaccine escape and allele-specific immunity because structural variation is pronounced in the PvMSP-7 multigene family.

All PvMSP-7 paralogous genes demonstrated a structural variation pattern. The central region of the genes was seen to harbour higher sequence polymorphism than the N- and C-terminal suggesting natural selection could have been acting differentially along the PvMSP-7 genes (Figure 3.2). Intriguingly, signals of recombination were more pronounced in the central region (Table 3.2). The high nucleotide diversity in the central region could possibly arise from the recombination events. Intragenic recombination occurs during meiosis is an essential event in generating new variants on most malaria antigens. The recombination events create allelic diversity where it removes deleterious variants or maintains beneficial traits (Hughes, 2008). Recombination and positive selection were also described previously in other organisms to drive antigenic variation (Andrews and Gojobori, 2004; Orsi *et al.*, 2007). In contrast, the nucleotide diversity in N- and C-terminal of the PvMSP-7 was lower. However, high level of sequence variation was seen at the N-terminal of PvMSP-7C, -7E, and -7G and two peaks of nucleotide diversity were also detected in the C-terminal correspond to PvMSP-7B and PvMSP-7G. Repeat motifs detected in the N-terminal of PvMSP-7C and PvMSP-7G could have resulted in the peak observed.

Differential selective pressure acting upon the genes could arise from the functional constraints. Codon-based substitution models detected the positive selected signals exclusively in the central of the PvMSP-7 genes suggesting the functional important. Consistently, these positive selection signals were located within the polymorphic and recombination region of the protein. The central region of the locus evolves rapidly probably attributed to the fact that the accumulation of amino acid substitutions for the parasite to evade the host's immune system recognition (Ferreira *et al.*, 2004). Moreover, the central domain has been implicated as a potential binding site to the primary proteolytic fragment of MSP-1, antigenic variation in this region could be responsible for immune evasion (Kadekoppala and Holder, 2010). Positive selection signal was also detected on a codon towards N-terminal in PvMSP-7F and PvMSP-7L implying the possible role in immune evasion. However, it still remains unknown the specific role of N-terminal during malaria infection.

Negative selection signal was detected mostly outside the central region of PvMSP-7 implying functional importance (Figure 3.4). This is consistent with the structural variation of PvMSP-7 where the conserved region located in the N- and C-terminal. Negative selection was shown to play a role in erythrocyte invasion in a malaria conserved antigen, rhoptry-associated protein 1 in *Plasmodium* (Pacheco *et al.*, 2010). Likewise, malaria antigens displaying low sequence polymorphism are likely to responsible for host cell invasion and often show negative selection signal (Garzón-Ospina *et al.*, 2018). Strong negative selection signals detected in N-terminal could arise from the involvement of the N-terminal in interacting with P-selectin. This interaction was shown in *P. berghei* where MSP-7 established a signal with P-selectin to modulate disease severity (Perrin *et al.*, 2015). However, this has not been shown in the *P. vivax*, future work should focus on the N-terminal of PvMSP-7 to validate the interaction of P-selectin. Furthermore, negative selection was also evidenced towards the codons in C-terminal. The C-terminal of MSP-7 has been suggested its role in erythrocyte invasion through MSP-1 multiprotein complex (Kadekoppala and Holder, 2010). Deletion of MSP-7 C-terminal in *P. falciparum* was found to inhibit parasite's ability to attach to the red blood cell (Kadekoppala *et al.*, 2008). Therefore, the N- and C-terminal of PvMSP-7 is an attractive target for malaria subunit vaccine development. However, these negatively selected codons should investigate carefully against the immune effectiveness in natural infection.

Closer looks into the codon-based tests of selection pressure across PvMSP-7 paralogs, no evidence of selection pressure was observed in PvMSP-7J. This paralog has a relatively short protein length (95 residues) which likely a pseudogene in the PvMSP-7 multigene family. PvMSP-7A and PvMSP-7L displayed evidence of positive selection in a few codons but no significant negative selection was detected. Likewise, the strong positive selection was detected along the PvMSP-7M locus, however, only one codon was under purifying selection. It supports the fact that these genes are under functional constraints where it tries to evade target by the host's immune system. Malaria surface antigens such as AMA-1, circumsporozoite protein, and MSP-1 displayed radical amino acid substitutions by strong balancing selection which likely to avoid detection by host's immune responses (Escalante *et al.*, 1998).

One of the major hurdles in the development of malaria subunit vaccine is the extensive sequence polymorphism in different malaria endemic areas around the world (Chenet *et al.*, 2012). To overcome this problem, the majority of the antimalarial vaccines have focused on the conserved candidates and domain to elicit universal immune responses. The PvMSP-7 paralogs with low genetic polymorphism are encouraged to be incorporated in the malaria vaccine development especially PvMSP-7A, -7D, -7F, -7J, -7K, -7L, and -7M. Although structural variation was demonstrated in most of the PvMSP-7 paralogs, selection of the conserved regions could enhance the immune efficacy against malaria. The high level of sequence conservation in the N- and C-terminal hold promise in the subunit vaccine design. The vaccine design encompasses the N- and C-terminal of PvMSP-7 could likely elicit superior effects in regulating disease severity and retard red-blood cell invasion. Therefore, antimalarial vaccines derived from PvMSP-7 conserved paralogs potentially impair parasite development.

3.5 Conclusion

Despite several indications that MSP-7 is important in erythrocyte invasion, caution should be taken during malaria subunit vaccine development as some of the proteins might have functional redundancy. The genetic diversity of PvMSP-7 multigene family in Thailand revealed heterogeneous sequence variation. Some paralogous genes revealed extensive sequence polymorphism while some members are rather conserved. From the vaccine design perspective, the conserved paralogs (PvMSP-7A, -7D, -7F, -7J, -7K, -7L, and -7M) and domains (N- and C-terminal) are ideally to be considered in vaccine development and likely to confer superior immune protection against malaria. Larger sampling size in Thailand is needed to validate the results. Furthermore, malaria immune response is known to be stage-specific, investigating the expression pattern of PvMSP-7 is therefore essential in vaccine design. In Chapter 5, the transcriptional changes of PvMSP-7 in natural infection will be revealed.

Chapter 4

Polymorphism in merozoite surface protein-7E of *Plasmodium vivax* in Thailand: Natural selection related to protein secondary structures

Abstract

PvMSP-7 multigene family shows heterogenous antigen variation among 13 paralogs. Among the 13 PvMSP-7 paralogs, PvMSP-7E is the most diverse paralog. To explore the potential of this protein as a genetic marker, 92 Thai isolates collected from three malaria endemic areas in Thailand (Tak province, Ubon Ratchathani province, and Yala and Narathiwat province) were further used to analyse the sequence polymorphism. The population genetic analysis estimated 52 unique haplotypes circulating in the endemic areas. Population structure based on this locus was observed between each endemic area. However, no evidence of genetic differentiation was found between populations collected from different periods in the same endemic area, indicating spatial but not temporal genetic variation. The sequence microheterogeneity is present within the N- and C- terminal regions with predicted four and six α-helical domains, respectively. In addition, evidence for purifying selection was found in the α-helices II-X, convincing structural or functional constraint in these domains. Conversely, the signal of positive selection was observed in α-helix I spanning the predicted signal peptide, in which amino acid substitutions may compromise the CD4+ T helper cell epitopes. The central region of PvMSP-7E encompassed the 5'-trimorphic and the 3'-dimorphic subregions. A signature positive selection was identified in the 3' dimorphic region within the central domain. A predicted intrinsically disorder protein has been shown to span across the central domain containing putative B cell epitopes and putative protein binding regions. Evidence of intragenic recombination was more distinguished in the central domain than the other domains of the gene. Therefore, the findings indicate that the antigen variation, occurrence of intragenic recombination, and evolutionary pressures in the PvMSP-7E locus appear to be differentially affected by protein secondary structures.

4.1 Introduction

In the previous chapter, the sequence variation of 13 PvMSP-7 members was described. From the result, PvMSP-7E was observed to display the highest nucleotide diversity than other paralogs within the multigene family. Although the conserved paralogs are ideally incorporated in the vaccine design, it is remaining unknown if only a specific paralog participates in the multiprotein complex prior to host cell invasion. Based on the previous study, several key malaria surface proteins seem to involve in erythrocyte invasion are under positive selection pressure and targeted for vaccine development (Garzón-Ospina *et al.*, 2018; Takala and Plowe, 2009). It has long recognized that the extensive sequence variation in malaria antigens arise from the balancing selection to allow the parasite to escape from host immune system by maintaining the sequence polymorphism. The evidence of natural selection has been characterized in three most studied vaccine candidates including, MSP-1 (Cheesman *et al.*, 2010), AMA-1 (Arnott *et al.*, 2014), and CSP (Neafsey *et al.*, 2015). As *Plasmodium* develops through different life stages, it expresses various stage-specific components, each of which will stimulate a specific immune response. Studies also reported that predominant alleles are varied between endemic areas, if a vaccine was to induce allele-specific immune responses, it might affect the vaccine efficacy between geographical locations (Takala and Plowe, 2009). Vaccine development using polymorphic antigens presented a major complication for vaccine design because vaccine candidates could elicit allele-specific immune responses. Thus, characterization of sequence polymorphism in malaria antigens from different geographic locations form the fundamental strategy for vaccine development.

MSP-1 is a prime surface protein that expresses during blood stages of the parasite. It undergoes several proteolytic cleavages and remains on the merozoite surface as a glycosylphosphatidylinositol-anchored complex. In the case of MSP-7, it undergoes similar proteolytic events as MSP-1 during schizogony (Baldwin *et al.*, 2015). Studies have shown that MSP-1 forms a non-covalent complex with MSP-6 and MSP-7 prior to host cell invasion. Disruption of MSP-1/6/7 in *P. falciparum* was evidenced to impair merozoite invasion into host cells (Woehlbier *et al.*, 2010). The MSP-1 sequence is divided into 17 blocks, where the last block is the main focus in vaccine development due to sequence conservation. Immunology studies have shown

cross-reactive antibody responses of MSP-1 block 17 in natural infections despite several SNPs in the domain. AMA-1 is another broadly studied vaccine candidate in blood stages of *Plasmodium*. The antigen is expressed on the apical surface of the merozoite and play a major role in host cell invasion by establishing the junction contact between merozoite and erythrocyte. No repeat region was found within the AMA-1 sequence, like other polymorphic antigens, conserved regions were incorporated in the vaccine development. Antibodies constructed using the conserved region was capable of stimulating protective immune responses in Papua New Guinea populations despite the small proportion of allele-specific antibodies (Cortés *et al.*, 2005). CSP is a pre-erythrocytic stage vaccine candidate which found on the surface of the sporozoite. The N- and C-terminal regions were found to responsible for parasite's hepatocyte-binding ligand (Swearingen *et al.*, 2016). Vaccine design based on this polymorphic antigen was focused on the repeat region. The vaccine efficacy induced by the repeat region has been in the pipeline to investigate the specificity of antibody responses. These studied identified the regions with greatest positive selection in relation to antigen-specific immune response. Based on the population genetics analyses applied on these polymorphic antigens, the domains of PvMSP-7E under positive selection were identified. This will translate into vaccine design which pinpoints the regions with most immunologically relevant based in this locus.

Much less is known about other merozoite proteins such as MSP-3 alpha (Zakeri *et al.*, 2006) and MSP-3 beta (Putaporntip *et al.*, 2014). Similar to MSP-7, MSP-3 is a multigene family consists of 12 paralogs encoded on chromosome 10 in *P. vivax*. These two paralogs have extensive sequence variation, consistently from the findings of two studies, the high genetic diversity indicating their potential as genetic markers in epidemiological studies other than being vaccine candidates. MSP-3 alpha of *P. vivax* in Iranian isolates is highly diverse between the northern and southern region using restriction fragment length polymorphism approach. Moreover, MSP-3 beta achieved a significant population differentiation between Thai and American parasite populations (*Fst*=0.28, *p*-value=<0.05) which in line with other molecular markers. From all these results, the high sequence variation of MSP-3 paralogs offers a powerful approach for genotyping *Plasmodium* isolates. This information provides an essential strategy for malarial drugs and vaccines implementation.

The extent of antigenic variation in PvMSP-7 has been characterized in natural infections from the previous chapter as well as isolates from Colombia. However, the sample size in these studies or sample populations might not well represented. PvMSP-7E appears to be the most polymorphic marker and evolved rapidly. Hence, this locus is potentially developed as a genetic marker for *P. vivax* populations further to involvement in host cell invasion. The rationale of this study is to evaluate the extent of sequence polymorphism among *P. vivax* populations from three major malaria endemic areas in Thailand.

4.2 Methodology

4.2.1 Human ethics statement

The study was approved by the Institutional Review Board in Human Research of Faculty of Medicine, Chulalongkorn University, Thailand (IRB No. 104/59). Blood samples were collected upon agreement from all participants or from their parents or guardians via written consent.

4.2.2 Study population

110 *P. vivax* malaria-positive blood samples were collected from patients with uncomplicated symptoms. The *P. vivax*-infected patients were diagnosed by microscopic examination of Giemsa stained blood films. Of these, 80 patients were recruited between the year 2008-2009 divided into 31 from Tak province, 16 from Narathiwat province, and nine from Yala province. To investigate if PvMSP-7E exhibits spatial or temporal variations, 24 blood samples collected during 1996 which preserved at -80°C were introduced into the analyses. Additional 30 samples from Ubon Ratchathani province collected during a malaria epidemic in the year 2014-2015 were used in the study. All blood samples were preserved in EDTA coagulant and stored at -30°C until used.

4.2.3 Amplification and sequencing of PvMSP-7E

The methods of DNA extraction, *Plasmodium* species identification and clonality detection have been described in Chapter 2. The complete coding region of PvMSP-7E (~1.1 kb) was amplified by nested PCR where two pairs of primers were designed targeting the outer and the inner region. Since only a small amount of blood volume was collected, nested PCR strategy is ideal in this scenario as the aim was to investigate other PvMSP-7 paralogs in the future. The PvMSP-7E was amplified by a pair of outer primers, PvMSP-7F (5'-CAT ACC TTC GAT ACG TGT ACT TC-3') and PvMSP-7R (5'-CAT TTC GCG TGT GCG TGT CTA TG-3') using the Salvador I reference strain (GenBank accession ID: XM_001614084, chromosome 12 from position 771164 to 772282). The inner primers were located before the start codon and after the stop codon of PvMSP-7E (PvMSP-7EF: 5'-AAT CGC CAC ACA TCG TCT GTG-3' and PvMSP-7ER: 5'-ATT TCA TCT TTA CTG TTG GGC AC-3). A schematic diagram of nested PCR primers was shown in Figure 4.1.

Primary PCR amplification was done in a total volume of 15 μL including PCR buffer, 200 μM dNTP, 0.2 μM of each primer, nuclease-free water, 2 μL of template DNA and 1.25 units of TaKaRa LA Taq (Takara, Seta, Japan). The thermal cycling profiles composed of a pre-amplification denaturation at 94ºC for 60 seconds, followed by 35 cycles of denaturation at 96ºC for 30 seconds, annealing at 50ºC for 30 seconds, polymerization at 72ºC for 7 minutes, and final elongation at 72ºC for 10 minutes. Following the primary PCR reaction, the secondary PCR amplification composed a total volume of 30 μL containing PCR buffer, 200 μM dNTP, 0.2 μM of each primer, nuclease-free water, 1 μL of template DNA from primary PCR and 1.25 units of ExTaq DNA polymerase (Takara, Seta, Japan). The amplification cycle for secondary PCR composed of denaturation at 94ºC for 60 seconds, followed by 30 cycles of denaturation at 96ºC for 30 seconds, annealing at 50ºC for 30 seconds, polymerization at 72ºC for 2 minutes, and final elongation at 72ºC for 5 minutes. The PCR amplification was performed in a GenAmp 9700 PCR thermal cycler (Applied Biosystems, Foster City, CA). The PCR products were analyzed on 1% agarose gel electrophoresis, stained with ethidium bromide and visualized under UV transillumination. Prior to DNA sequencing, PCR products were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany). DNA sequences were generated directly and bi-

directionally from the purified PCR products derived from the secondary PCR amplification using ABI PRISM BigDye Terminator v3.1 Ready Reaction Cycle Sequencing Kit (Applied Biosystems) and sequencing primers.

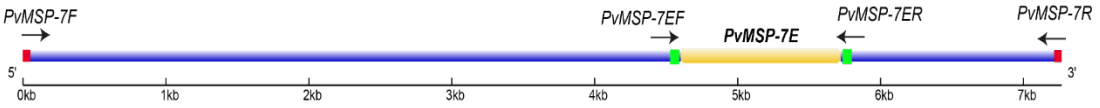


Figure 4.1. A schematic diagram of nested PCR primers was designed to amplify PvMSP-7E gene. PvMSP-7F and PvMSP-7R represent the outer primers whilst, PvMSP-7EF and PvMSP-7ER show the inner primers.

4.2.4 Data analysis and protein secondary structure prediction

DNA sequences were aligned using CLUSTAL W version 2.0 with the default setting (Larkin *et al.*, 2007). All sequences were aligned against the PvMSP-7E reference sequence derives from the Salvador I strain (GenBank and PlasmoDB accession no. PVX_082665). Published sequences from Colombian isolates were retrieved and added to the analysis (GenBank accession nos. KM212276-KM212294) (Garzón-Ospina *et al.*, 2014). In the alignment, all sites that postulated a gap were removed in pairwise comparisons of the analysis. DNA tandem repeat regions were determined by scanning the sequence with Tandem Repeats Finder version 4.09 program available from http://tandem.bu.edu/trf/trf.html. Protein secondary structure was reliably determined by Deep Convolutional Neural Filed Program (DeepCNF) (Wang *et al.*, 2016b). The advanced principle behind DeepCNF has accuracy up to 80% for determining the model complex sequence-structure relationship. The program is available freely via the RaptorX-Property web server (Wang *et al.*, 2016a). Intrinsically unstructured or protein disordered domains were identified by using the GeneSilico MetaDisorder service (Kozlowski and Bujnicki, 2012). The disordered domains are characterized by lack of stable tertiary structure and the intrinsic flexibility tolerates multiple interactions with other disordered proteins. Most of the major vaccine candidates are seen to present extensive intrinsically disordered regions along the gene for host-cell invasion (Guy *et al.*, 2015). Protein binding regions within the disordered protein were determined by using ANCHOR/IUPRED web server (Dosztányi *et al.*, 2009). The protein-protein

interaction regions within the disordered protein play a functional role in biological processes including regulation and signalling (Dyson and Wright, 2002).

4.2.5 Evolutionary genetic analysis

The molecular evolutionary analysis was performed using DnaSP version 5.10 (Librado and Rozas, 2009). The program calculates haplotype diversity, nucleotide diversity, and its variance. The nucleotide diversity (π) was estimated by the mean of pairwise sequence differences per site in the sample sequences. All sites were taken into the consideration based on that of Juke and Cantor model of nucleotide substitution (Jukes and Cantor, 1969). The rate of synonymous substitutions per synonymous site (d_S) and the rate of nonsynonymous substitutions per nonsynonymous site (d_N) were calculated using the method of Nei and Gojobori's model with Jukes-Cantor correction (Jukes and Cantor, 1969; Nei and Gojobori, 1986). Under positive selection, the ratio of d_N/d_S is expected to exceed 1, whilst d_N/d_S below 1 indicates purifying selection which selection pressure acts against protein changes. When the domain is not under selection pressure, $d_N/d_S = 0$ would be assumed. The standard errors with 1000 pseudo-sampling bootstraps were performed by comparing the nonsynonymous and synonymous substitutions implemented in the MEGA version 6.0 program (Tamura *et al.*, 2013). The statistical differences were calculated by using two-tailed Z-test and the statistics considered significant with $p < 0.05$.

Evolutionary pressures acting on each codon were estimated by using six complementary methods implemented in the Datamonkey web server (Pond and Frost, 2005). The six approaches used in the analyses were single-likelihood ancestor counting (SLAC), fixed effects likelihood (FEL), internal branch FEL (iFEL), random effects likelihood (REL), mixed effects model of evolution (MEME), and fast unconstrained Bayesian approximation (FUBAR). The significance level was based on the default setting as recommended by the web server. SLAC is a modified principle from the Suzuki-Gojobori counting approach (Suzuki and Gojobori, 1999). It has superior power to infer mutations at each site based on the maximum likelihood reconstruction of the ancestral sequences, provided the analysis does not void the assumption of neutral evolution. FEL is a less conservative model compares to SLAC,

where the ratio of nonsynonymous to synonymous is assessed independently at each codon position. iFEL has the similar principle as FEL, but the selection is estimated along the internal branch of the phylogeny. REL approach largely based on assumption where it allows synonymous rate variation (Yang and Nielsen, 2000). MEME has a sophisticated algorithm to detect individual sites under episodic and pervasive positive selection. MEME does not assume a *priori* distribution, instead, it allows each site to retain its previous selective history (Murrell *et al.*, 2012). FUBAR is a sophisticated algorithm to analyze large dataset based on Markov chain Monte Carlo. The approach hypothesizes the selection force is constant along the phylogeny (Murrell *et al.*, 2013). Selective influences based on amino acid properties were identified by using the TreeSAAP program (Woolley *et al.*, 2003). The actual probabilities changes for amino acid properties were deemed significant with their percent probability set at 99.9% based on categories 6-8.

4.2.6 Intragenic recombination

Evidence of recombination events was determined by using the Recombination Detection Program version 4 (RDP4) (Martin *et al.*, 2015). Recombination analyses were identified by a combination of methods such as GENCONV, Bootscanning, the maximum Chi-Square, CHIMAERA, Sister Scanning, and 3SEQ. Default settings were used to investigate the overall recombination signals. One sophisticated principle in RDP4 is the extensive exploration of recombination signals present within the nucleotide alignment. Recombination segments are fragmented into several different parts and scanned iteratively within the alignment until no recombination signals can be found. This mechanism is useful particularly in sequences with complex recombination patterns. The population genetic structure was expressed by using the fixation index (*Fst*) implemented in the Arlequin program version 3.11 (Excoffier *et al.*, 2005). The *Fst* was evaluated by the molecular variance (AMOVA) based on the principle by Weir and Cockerham (1984). AMOVA takes into consideration the total proportion of genetic variance including the number of mutations. Statistically significant of the fixation index was determined by a permutation test. Phylogenetic trees were constructed to infer the evolutionary relationship of the PvMSP-7E gene using the maximum likelihood method. The best substitution model was selected based

on the lowest Bayesian Information Criterion (Tamura *et al.*, 2013) score. The reliability of the phylogeny was evaluated by bootstrap method with 1000 pseudoreplicates.

4.2.7 B-cell and T-cell epitopes prediction

B-cell epitopes prediction is important in vaccine design, BCPRED web server was used to predict the linear B-cell epitopes with an epitope length of 20 amino acids. The estimation was set to 90% classifier specificity to achieve accurate epitope prediction (EL- Manzalawy *et al.*, 2008). CD4+ T cell epitopes confer protective immune responses in vaccination, PREDIVAC web server was used to predict MHC-II binding peptides (Oyarzún *et al.*, 2013). This server is the chosen for the analysis because it has more than 95% coverage of human HLA class II DR protein diversity. Five predominant HLA-DR alleles in Thai population were selected in the analysis including, DRB1*1202, DRB1*1502, DRB1*0701, DRB1*1501, and DRB5*1602 (Romphruk *et al.*, 1999).

4.3 Results

4.3.1 Genetic diversity in PvMSP-7E

The 110 samples infected with *P. vivax* were genotyped by PCR using block six of PvMSP-1. Of these, only 92 samples infected with a single strain of *P. vivax* were retained in the analyses. Electropherograms with non-superimposed signals were used to confirm the presence of single clone infections. In the analysis, Yala and Narathiwat provinces are defined as a single population because these areas are located next to each other and they have similar malaria transmission dynamics. Therefore, the PvMSP-7E sequences were characterised by geographical regions in Thailand as Tak (*n*=46), Ubon Ratchathani (*n*=22), and Yala-Narathiwat (*n*=24). However, of the 46 samples from Tak province, they were further subdivided into two populations based on the sampling period, the year 2008-2009 (*n*=28) and year 1996 (*n*=18) (Table 4.1). In total, 52 haplotypes were found for PvMSP-7E in Thai isolates, consisting of 194 nucleotide substitutions, 185 segregating sites, and 9 insertions/deletions. Haplotype #1 was predominantly found between populations from Ubon Ratchathani (*n*=2) and Yala-Narathiwat (*n*=14). Similarly, haplotypes #15-#17 were shared between Tak and Ubon Ratchathani populations. In contrast, the remaining 48 haplotypes were unique between the endemic areas (Figure 4.2). Nucleotide diversity varied from 0.0514±0.0048 (isolates from Yala-Narathiwat provinces) to 0.0620±0.0046 (isolates from Tak province). Although the level of nucleotide diversity was higher in Tak province than that of Ubon Ratchathani province, the differences were not statistically meaningful (Z-test, $p > 0.05$). The distribution of PvMSP-7E haplotypes in Yala-Narathiwat populations was skewed towards few haplotypes as shown by the haplotype diversity ($h = 0.540 \pm 0.062$) whereas, populations from Tak and Ubon Ratchathani province had a remarkably higher number of haplotypes and values of haplotype diversity, indicating an even distribution of haplotype frequencies in these endemic areas.

Table 4.1 Estimates of sequence diversity in the PvMSP-7E gene of *P. vivax* populations in Thailand. In total, 92 samples were used to infer the population genetic parameters implemented in DnaSP version 5.10. All sites that postulated a gap were excluded in the analysis.

	n	M	S	Indel	H	*h* ± S.D.	**π ± S.E.**
Tak	46	191	183	9	34	0.986 ± 0.007	0.0620 ± 0.0046
Tak 1996	18	182	174	9	16	0.987 ± 0.023	0.0496 ± 0.0039
Tak 2008-2009	28	189	181	9	22	0.984 ± 0.013	0.0677 ± 0.0050
Yala-Narathiwat	24	117	116	9	3	0.540 ± 0.062	0.0514 ± 0.0048
Ubon Ratchathani	22	161	153	9	19	0.987 ± 0.018	0.0586 ± 0.0047
Total	92	194	185	9	52	0.958 ± 0.013	0.0613 ± 0.0047

n: number of isolates, M: number of mutations, S: number of segregating sites, Indel: number of insertions or deletions, H: number of haplotypes, *h*: haplotype diversity, π: nucleotide diversity, S.E.: standard error.

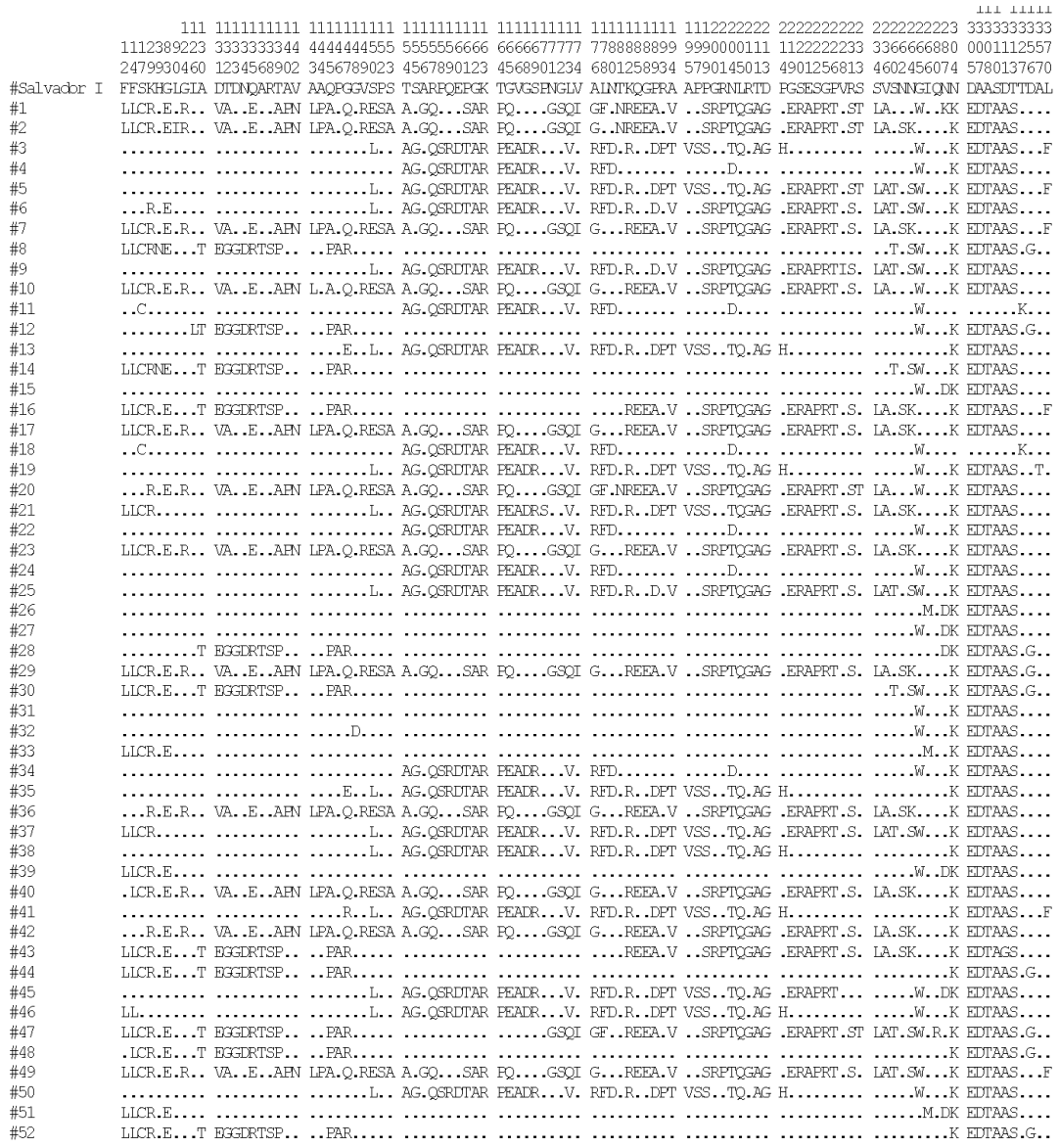


Figure 4.2 PvMSP-7E haplotypes among Thai isolates. In total, 52 haplotypes are distributed across the major malaria endemic areas in Thailand. The distribution of haplotypes according to endemic areas are as follows: haplotype #1 (Ubon Ratchathani *n*=2, Yala-Narathiwat *n*=14), haplotype #2 (Yala-Narathiwat *n*=9), haplotype #3-#4 (Tak *n*=3 each), haplotype #5-#12 (Tak *n*=2 each), haplotype #13-#14 (Ubon Ratchathani *n*=2 each), haplotype #15-#17 (Tak *n*=1, Ubon Ratchathani *n*=1), haplotype #18-38 (Tak *n*=1 each), haplotype #39-51 (Ubon Ratchathani *n*=1 each), and haplotype #52 (Yala-Narathiwat *n*=1).

4.3.2 Sequence variation in the 5’ and the 3’ regions of PvMSP-7E

The analysis of the complete sequence of the PvMSP-7E against the Salvador I reference sequence has shown two regions with relatively low nucleotide diversity (π=0.0224±0.0050 and 0.0273±0.0044). These conserved regions were located in the 5’ and 3’ regions of the gene, spanning 123 and 135-136 codons, respectively (Figure 4.3). Notably, the previous study from Colombian isolates did not analyse the 51 nucleotides at the 5’ end and 15 nucleotides at the 3’ end of the gene (Garzón-Ospina *et al.*, 2014). Closer looks into these regions, they contained four nonsynonymous codon changes at residues F12L, F14L, S17C and L368F. Among isolates examined herein, 20 and 51 nucleotide substitutions occurred in the 5’ and 3’ regions, respectively. Of the total 71 nucleotide substitutions, 69 were dimorphic substitutions and two sites at 786 and 926 at the 3’ region were trimorphic. Moreover, most of the Thai isolates had a deletion at codon 312 (94.6%) coding for proline in the 3’ region. In contrast, insertions between codons 150 and 151, and codons 215 and 216 account for 21 (22.8%) and 26 (28.3%) of the isolates with TTA (leucine) and GAA (glutamine).

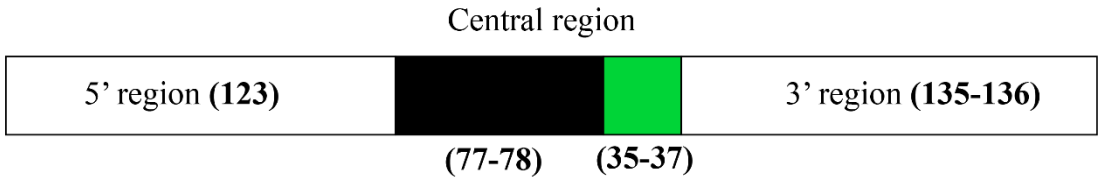


Figure 4.3 Schematic representation of PvMSP-7E depicting conserved (filled white boxes), variable trimorphic (filled black box) and dimorphic (filled green box) regions. The number of codons for each region is in parentheses.

4.3.3 Sequence variation in the central region of PvMSP-7E

The central region of PvMSP-7E spans between codons 124-234 of Salvador I reference strain (Figure 4.3). The nucleotide diversity along the central region displayed a higher magnitude than those at the 5' and 3' regions, the differences were statistically significant ($p < 0.0001$) (Table 4.2). In the analysis, 29 allelic types were observed in the central region. No repeat motifs were evidenced in the gene. Furthermore, the residues spanning 124-200 along the N-terminal of the central region displayed a mosaic organization of the sequences, suggesting it could have been derived from the genetic shuffling among three parental types. The genetic shuffling is represented in Figure 4.4(a) by Salvador I strain (type 1-5'), the APH5 isolate from Tak province (type II-5'), and an unknown strain (type III-5'). In contrast, the C-terminal within the central terminal could have been generated from two parental sequences (type I-3' and type II-3'). For this reason, nucleotide sequences spanning the N- and C- terminal of the central region are defined as 5'-trimorphic and 3'-dimorphic subregions, corresponding to 77-78 and 35-37 residues, respectively (Figure 4.3). The magnitude of nucleotide diversity for 5'-trimorphic region and 3'-trimorphic regions were significantly different ($p < 0.005$), the former exhibited higher diversity (Table 4.2). Additionally, two sites with insertion/deletions were located in the central region, each at the 5'trimorphic and the other at 3'-dimorphic subregions, respectively.

Table 4.2 Nucleotide diversity (π) and number of synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions per site in PvMSP-7E among Thai isolates

Region	n	M	*h*	$\pi \pm$ S.E.	$d_S \pm$ S.E.	$d_N \pm$ S.E.
5’	123	20	13	0.0224 ± 0.0050	0.0889 ± 0.0277**	0.0090 ± 0.0038
Central	112-113	122	20	0.1598 ± 0.0150§§§§	0.1125 ± 0.0261	0.1874 ± 0.0195*
5’-trimorphic	77	99	16	0.1939 ± 0.0198##	0.1483 ± 0.0394	0.2164 ± 0.0245
3’-dimorphic	36	23	10	0.0973 ± 0.0222	0.0349 ± 0.0253	0.1300 ± 0.0304*
3’	135-136	52	42	0.0273 ± 0.0044	0.0799 ± 0.0194***	0.0109 ± 0.0033
Total	369-371	194	52	0.0614 ± 0.0047	0.0919 ± 0.0123*	0.0549 ± 0.0056

N: number of mutations, *h*: number of haplotypes.

Tests of the hypothesis that d_S equals d_N; * $p < 0.05$; ** $p < 0.005$; *** $p < 0.001$.

Tests of the hypotheses that π in the central region equals the corresponding values in the 5’ and the 3’ regions; §§§§$p < 0.0001$.
Tests of the hypotheses that π in the 5’ trimorphic region equals that in the 3’ dimorphic region; ## $p < 0.005$.

a

```
Type      ←                          5′ central sequence                          →←          3′ central sequence          →
I         GKIKGQADTD NQAQRTAD-V AAQPGGVSPS TSARPQEPGK TGVTGSPNGL VEAGLVNTKT LQNVGPNGQR AADPQPGRAA NLPEGQRTND PQQ--GGSES TEGPAVTPRP SSTV
II        .......... ........-. .....D.... .......... .......... .......... .......... .......... .......... ...--..... .......... ....
III       .......... ........-. .......... AG.QSRDTAR PEA.DR...V ..R.F.D... .......... .......... D......... ...--..... .......... ....
IV        .......... ........-. ....E..L.. AG.QSRDTAR PEA.DR...V ..R.F.D.R. .....D...P TV.S.S.... TQ.....A.G H..-E..... .......... ....
V         .......... ........-. ....R..L.. AG.QSRDTAR PEA.DR...V ..R.F.D.R. .....D...P TV.S.S.... TQ.....A.G H..-E..... .......... ....
VI        .......... ........-. .......L.. AG.QSRDTAR PEA.DR...V ..R.F.D.R. .....D...P TV.S.S.... TQ.....A.G H..-E..... .......... ....
VII       .......... ........-. .......... AG.QSRDTAR PEA.DR...V ..R.F.D.R. .....D...P TV.S.S.... TQ.....A.G H..-E..... .......... ....
VIII      .......... ........-. .......L.. AG.QSRDTAR PEA.DR...V ..R.F.D.R. .....D...P TV.S.S.... TQ.....A.G H..EE..... .......... ....
IX        .......... ........-. .......... AG.QSRDTAR PEA.DR...V ..R.F.D.R. .....D...P TV.S.S.... TQ.....A.G H..EE..... .......... ....
X         .......... ........-. .......L.. AG.QSRDTAR PEA.DR...V ..R.F.D.R. .....D...P TV.S.S.... TQ.....A.G ...--.ERAP ..RT..... ....
XI        .......... ........-. .......L.. AG.QSRDTAR PEA.DR...V ..R.F.D.R. .....D...P TV.S.S.... TQ.....A.G ...--.ERAP ..RT....S. TL.A
XII       .......... ........-. .......L.. AG.QSRDTAR PEA.DR...V ..R.F.D.R. .....D...P TV.S.S.... TQ.....A.G ...--.ERAP ..RT....S. .L.A
XIII      .......... ........-. .......L.. AG.QSRDTAR PEA.DRS..V ..R.F.D.R. .....D...P TV.S.S.... TQ....GA.G ...--.ERAP ..RT....S. .L.A
XIV       .......... ........-. .......L.. AG.QSRDTAR PEA.DR...V ..R.F.D.R. .....D.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. .L.A
XV        .......... ........-. .......L.. AG.QSRDTAR PEA.DR...V ..R.F.D.R. .....D.... V....SRP.. TQ....GA.G ...--.ERAP ..RT.I..S. .L.A
XVI       .......... ........-. .......L.. AG.QSRDTAR PEA..... ........R. .E..EA.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. .L.A
XVII      ......TEGG DRT.SP..-. ..PAR..... .......... .......... .......... .......... .......... .......... ...--..... .......... ....
XVIII     ..L...TEGG DRT.SP..-. ..PAR..... .......... .......... .......... .......... .......... .......... ...--..... .......... ....
XIX       ......TEGG DRT.SP..-. ..PAR..... .......... .......... ........R. .E..EA.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. .L.A
XX        ......TEGG DRT.SP..-. ..PAR..... .......... .......... ....S...R. .E..EA.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. TL.A
XXI       ......TEGG DRT.SP..-. ..PAR..... .......... .......GSQ I.G.F...R. .E..EA.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. TL.A
XXII      ......TEGG DRT.SP..-. ..PQ.RELSA A.GQ...SAR PQ.....GSQ I.G.F...R. .E..EA.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. TL.A
XXIII     R......VA. .E...AP.NL P.PQ.RELSA A.GQ...SAR PQ.....GSQ I.G.F...R. .E..EA.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. TL.A
XXIV      R......VA. .E...AP.NL P.PQ.RELSA A.GQ...SAR PQ.....GSQ I.G.F...R. .E..EA.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. TL.A
XXV       R......VA. .E...AP.NL P.PQ.RELSA A.GQ...SAR .Q.....GRQ I.G.F...R. .E..EA.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. TL.A
XXVI      R......VA. .E...AP.NL P.PQ.RELSA A.GQ...SAR PQ.....GSQ I.G.F..NR. .E..EA.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. TL.A
XXVII     R......VA. .E...AP.NL P.PQ.RELSA A.GQ...SAR PQ.....GSQ I.G....NR. .E..EA.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. TL.A
XXVIII    R......VA. .E...AP.NL P.PQ.RELSA A.GQ...SAR PQ.....GSQ I.G.....R. .E..EA.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. .L.A
XXIX      R......VA. .E...AP.NL ..PQ.RELSA A.GQ...SAR PQ.....GSQ I.G.....R. .E..EA.... V....SRP.. TQ....GA.G ...--.ERAP ..RT....S. .L.A
```

b

5' Parental Sequences

```
Type I-5′     GKIKGQADTDNQAQRTAD-VAAQPGGVSPSTSARPQEPGKTGVTGSPNGLVEAGLVNTKTLQNVGPNGQRAADPQPGR
                                    ED
                                    R

Type II-5′    R......VA..E...AP.NLP.PQ.RELSAA.GQ...SARPQ.....GSQI.G.F..NR..E..EA....V....SRP
                                A                    T              S  T
                                                                    L

Type III-5′   ......TEGGDRT.SP..-...PAR..L..AG.QSRDTARPEA.DR...V..R.F.D.R......D...PTV.S.S..
                L                          S                 S
```

3' Parental Sequences

```
Type I-3′     AANLPEGQRTNDPQQ--GGSESTEGPAVTPRPSSTV
                D

Type II-3′    ..TQ....GA.G...--.ERAP..RT....S.TL.A
                      R   H  -E               S
                             EE
```

Figure 4.4 Sequence variation in the central region of PvMSP-7E. **(a)** Amino acid sequences. The boundaries of the 5' and 3' subregions are marked above the alignment. **(b)** Parental alleles of the 5' and 3' subregions. Dots are identical residues and dashes represent deletion/insertion. Amino acids shown under each parental sequence are variant residues.

4.3.4 Protein secondary structure prediction

Prediction of the protein secondary structure using DeepCNF method implemented in the RaptorX-Property web server showed that PvMSP-7E has ten α-helical domains. Of these, four α-helical domains were characterised in the N-terminal whilst, six α-helical domains were located in the C-terminal (Figure 4.5). Importantly, most of the non-helical regions appear to compose random coil structures. Three intrinsically unstructured or disordered regions were discovered by using MetaDisorder service implemented in the GeneSilico Metadisorder web server. These disordered regions span between codons 27-36, 46-102, and 121-241, the latter being the longest disordered structure and labelled as D1, D2, and D3. Furthermore, protein-protein binding regions were found spanning along the central fragment of PvMSP-7E (Figure 4.5).

4.3.5 Selective pressure on PvMSP-7E

Molecular evolution was used to determine whether the PvMSP-7E departure from neutrality. Analysis of patterns of substitution in the PvMSP-7E sequences by calculating d_S and d_N revealed that natural selection seems to operate differently along the gene. The d_S rate was significantly greater than d_N in both 5' and 3' regions ($p < 0.005$). This result suggested that the 5' and 3' domains were under natural purifying selection. In contrast, d_N had significantly exceeded d_S in the central region, indicating positive selection at certain residues in this gene ($p < 0.005$) (Table 4.2). Likewise, the results were consistent when analysis focused on each parasite population (Table 4.3). Closer looks into the central region of the sequences, d_N significantly greater than d_S in the 3'-dimorphic ($p < 0.05$). Meanwhile, 5'-trimorphic did not show evidence of departure from neutral expectations ($p > 0.05$) (Table 4.2). Moreover, to discover whether natural selective pressure operates specifically in a region of the sequence, the rate of synonymous and nonsynonymous substitutions per site were identified for each domain in relation to the predicted protein secondary structure. Due to the number of mutation sites in α-helical domains excluding α-helix-I, three adjacent helical regions were combined in further analysis. Results from the selection test showed that d_S significantly outnumbered d_N in α-helical domains II-IV, V-VII and VIII-X, implying purifying selection acting in these regions (Table 4.4). Meanwhile, purifying selection

was also seen in the predicted disordered domain 2 (D2). The random coiled regions in the gene and the predicted disordered domain 1 (D1) did not achieve significant differences between d_S and d_N. By contrast, the α-helix-I domain exhibited a higher magnitude of d_N than d_S and the difference was statistically meaning ($p < 0.05$), indicating positive selection in this domain. In domain D3, although d_N greater than d_S, the difference was not significant ($p > 0.05$). On the other hand, positive selection was evidenced in the 3' region of the D3 domain corresponding to the 3'-dimorphic subregion where a significantly higher rate of d_N than d_S was detected (Tables 4.2 and 4.4).

Tests for departure from neutrality were examined based on each codon as implemented in Datamonkey web server. Various principles of identification including SLAC, FEL, iFEL, REL, FUBAR and MEME revealed 2, 8, 11, 39, 12 and 20 positively selected sites, respectively (Table 4.5). Consistently, 18 positively selected sites were identified by using TreeSAAP program where various physicochemical properties of substituted amino acids were considered in the analysis. Besides that, SLAC, FEL, iFEL, REL and FUBAR methods found 23, 38, 31, 30 and 24 negatively selected sites, respectively (Table 4.6). A consensus of positively and negatively selected sites with at least two methods were used for further interpretations, thereby account for false positive and negative results. Closer looks into the findings, 80.77% (21 of 26) of the positively selected sites were mapped to the α-helix-I domains and the predicted disordered region. However, 85.29% (29 of 34) of the negatively selected sites were mapped outside these regions. The distribution of positively and negatively selected sites between α-helical domains and intrinsically unstructured regions was significant difference ($p < 0.0001$, Fischer exact probability test).

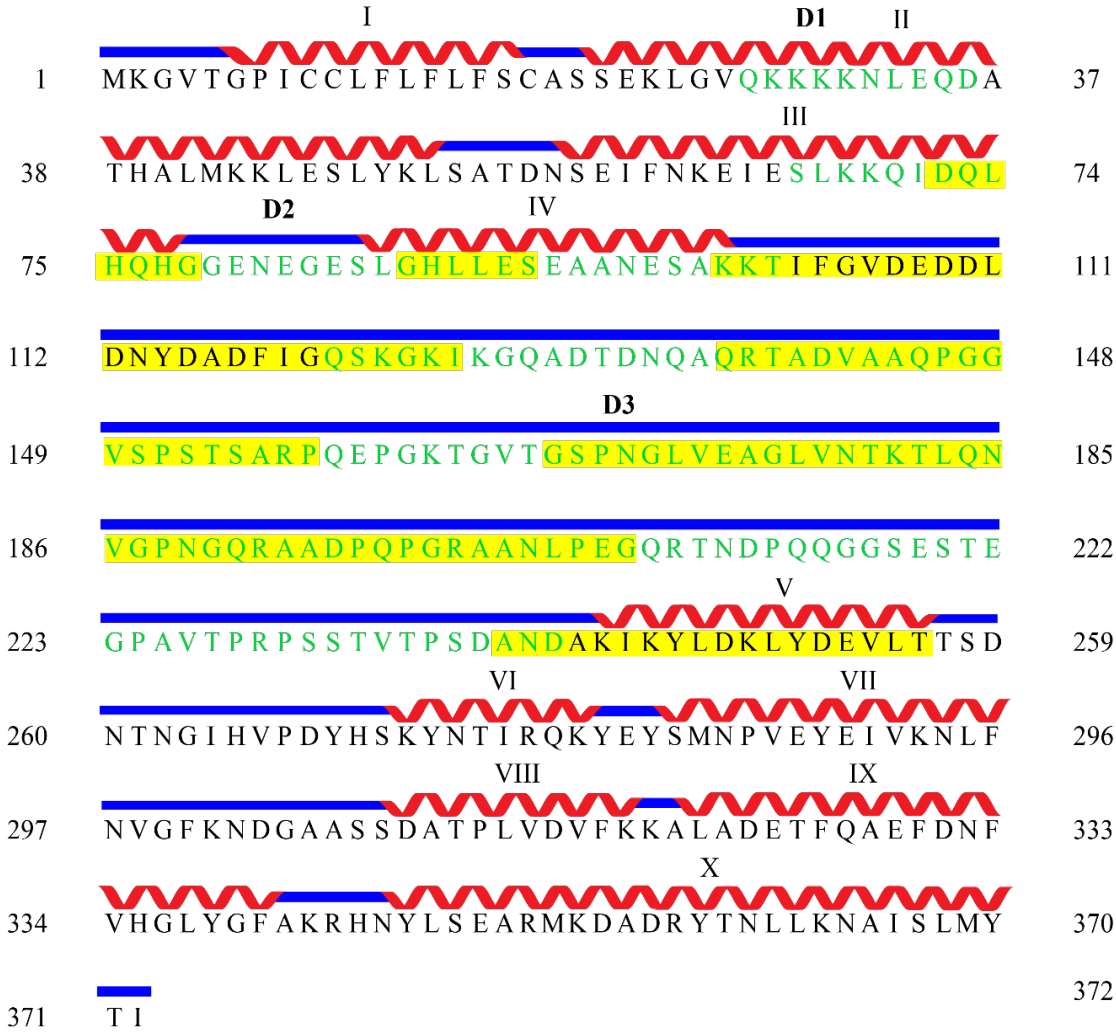


Figure 4.5 Predicted protein secondary structure of PvMSP-7E. The structure was generated by DeepCNF available from RaptorX-Property web server. Secondary structures are represented above the sequence by red helices (α-helices) and blue lines (coiled). Predicted intrinsically unstructured regions were determined using the GeneSilico Metadisorder service (residues in green). Protein-protein interaction regions within the disordered regions were analyzed by ANCHOR/IUPRED web server and highlighted in yellow. Ten α-helices domains were identified (I-X). Three unstructured disorder regions that span between codons 27-36, 46-102, and 121-241 were identified (D1-D3).

Table 4.3 Nucleotide diversity (π) and number of synonymous (d_S) and nonsynonymous (d_N) substitutions per site in *PvMSP-7E* among *P. vivax* populations in Thailand. Nei and Gojobori's model with Jukes-Cantor correction (Jukes and Cantor, 1969; Nei and Gojobori, 1986) was used to computed the rate of d_S and d_N. Standard error (S.E.) was derived from1000 pseudosamplings bootstrap method implemented in MEGA 6 suite (Tamura *et al.*, 2013).

Population	Region	π ± S.E.	d_S ± S.E.	d_N ± S.E.
Tak 1996	5'	0.0181 ± 0.0051	0.0690 ± 0.0262*	0.0083 ± 0.0036
	Central	0.1222 ± 0.0115	0.0809 ± 0.0206	0.1434 ± 0.0139*
	5'-trimorphic	0.1454 ± 0.0158	0.1051 ± 0.0295	0.1664 ± 0.0213
	3'-dimorphic	0.0800 ± 0.0184	0.0405 ± 0.0245	0.1008 ± 0.0263
	3'	0.0253 ± 0.0045	0.0935 ± 0.0194***	0.0138 ± 0.0033
	All	0.0496 ± 0.0041	0.0801 ± 0.0110**	0.0451 ± 0.0044
Tak 2016	5'	0.0248 ± 0.0053	0.0935 ± 0.0286**	0.0097 ± 0.0039
	Central	0.1680 ± 0.0151	0.1201 ± 0.0276	0.1942 ± 0.0213*
	5'-trimorphic	0.2007 ± 0.0205	0.1605 ± 0.0403	0.2276 ± 0.0253
	3'-dimorphic	0.1127 ± 0.0279	0.0556 ± 0.0304	0.1363 ± 0.0313
	3'	0.0352 ± 0.0052	0.1003 ± 0.0217***	0.0131 ± 0.0036
	All	0.0677 ± 0.0048	0.1029 ± 0.0135**	0.0580 ± 0.0061
Ubon Rachathani	5'	0.0231 ± 0.0052	0.0865 ± 0.0265**	0.0089 ± 0.0038
	Central	0.1588 ± 0.0148	0.1079 ± 0.0242	0.1874 ± 0.0202*
	5'-trimorphic	0.1963 ± 0.0208	0.1432 ± 0.0377	0.2255 ± 0.0254
	3'-dimorphic	0.0910 ± 0.0182	0.0485 ± 0.0286	0.1176 ± 0.0247
	3'	0.0189 ± 0.0038	0.0622 ± 0.0165**	0.0092 ± 0.0030
	All	0.0586 ± 0.0054	0.0838 ± 0.0132*	0.0546 ± 0.0055
Yala-Narathiwat	5'	0.0198 ± 0.0051	0.1239 ± 0.0386**	0.0114 ± 0.0049
	Central	0.1449 ± 0.0170	0.1554 ± 0.0376	0.2076 ± 0.0292
	5'-trimorphic	0.1794 ± 0.0263	0.2136 ± 0.0676	0.2594 ± 0.0460
	3'-dimorphic	0.0832 ± 0.0247	0.0708 ± 0.0430	0.1241 ± 0.0356
	3'	0.0148 ± 0.0038	0.0895 ± 0.0269**	0.0106 ± 0.0054
	All	0.0514 ± 0.0051	0.1204 ± 0.0175**	0.0611 ± 0.0079

Tests of the hypothesis that d_S equals d_N: # $p < 0.05$; ## $p < 0.005$; ### $p < 0.0001$.

Table 4.4 Number of synonymous (d_S) and nonsynonymous (d_N) substitutions per site in relation to protein secondary structure prediction of PvMSP-7E. The rate of d_S and d_N was estimated by Nei and Gojobori's model with Jukes-Cantor correction (Jukes and Cantor, 1969; Nei and Gojobori, 1986). Standard error (S.E.) was computed with 1000 pseudosamplings bootstrap method implemented in MEGA 6 suite (Tamura *et al.*, 2013).

Predicted domain	Nucleotides	d_S ± S.E.	d_N ± S.E.
α-helix I	36	0.0000 ± 0.0000	0.0608 ± 0.0310#
α-helices II-IV	192	0.0776 ± 0.0372#	0.0040 ± 0.0035
α-helices V-VII	108	0.0992 ± 0.0480#	0.0005 ± 0.0005
α-helices VIII-X	156	0.0480 ± 0.0215#	0.0051 ± 0.0029
α-helices II-X	456	0.0699 ± 0.0184###	0.0035 ± 0.0018
Remaining non-helical regions	630	0.1160 ± 0.0202	0.0967 ± 0.0101
Disorder I	30	0.0000 ± 0.0000	0.0205 ± 0.0208
Disorder II	171	0.2023 ± 0.0867#	0.0040 ± 0.0037
Disorder III	369	0.1116 ± 0.0250	0.1667 ± 0.0176

*Domains are demarcated as in Figure 4.5.

Tests of the hypothesis that d_S equals d_N: # $p < 0.05$; ## $p < 0.005$; ### $p < 0.0001$.

Table 4.5 Codon-based analysis of positive selection in PvMSP-7E. Codon-based test was conducted with five combination methods implemented in Datamonkey web server. The data was further evaluated with TreeSAAP (Woolley *et al.*, 2003). The table shows that the positively selected sites correspond to the predicted protein secondary structure. The tick marks show significant *p*-value as recommended.

Region	Codon	Amino acid		Test method						
		Sal-1	Variant	SLAC	FEL	IFEL	REL	FUBAR	MEME	TreeSAAP
Helix-I	12	F	L	✓	✓		✓	✓	✓	
Helix-I	14	F	L		✓	✓	✓	✓	✓	
Helix-I	17	S	C		✓		✓	✓	✓	✓
Helix-II/Disorder-I	29	K	R				✓	✓		
Disorder-II	83	G	E				✓	✓		
Disorder-III	131	D	V / E				✓			✓
Disorder-III	132	T	A / G		✓	✓		✓	✓	
Disorder-III	139	T	A / P			✓	✓			
Disorder-III	142	V	N				✓		✓	
Disorder-III	143	A	L				✓		✓	
Disorder-III	149	V	R						✓	✓
Disorder-III	150	S	E / L		✓	✓	✓	✓	✓	✓
Disorder-III	160	P	S / T		✓	✓	✓			
Disorder-III	164	G	Q / E			✓	✓			
Disorder-III	170	N	G					✓	✓	
Disorder-III	175	A	G / R		✓	✓	✓	✓	✓	✓
Disorder-III	188	P	A / D			✓	✓		✓	
Disorder-III	192	R	P			✓	✓			✓
Disorder-III	193	A	V / T			✓	✓		✓	✓
Disorder-III	203	N	T / D				✓		✓	
Disorder-III	209	R	G				✓			✓
Disorder-III	229	R	S			✓	✓		✓	
Non-helix	262	G	W					✓	✓	✓
Non-helix	263	I	M				✓		✓	
Helix-VII	285	N	K / D				✓	✓	✓	
Non-helix	368	L	F	✓	✓		✓	✓	✓	

Table 4.6 Codon-based analysis of negative selection in PvMSP-7E. Five methods of selection tests were used to compute the negatively selected sites as implemented in Datamonkey web server. The table shows that the negatively selected sites correspond to predicted protein secondary structure. Concordant results from 2 or more tests are shown. Tick marks indicate significant *p* values based on default option in the Datamonkey web server.

Region	Codon	Sal-1	Test method				
			SLAC	FEL	IFEL	REL	FUBAR
Helix-II	46	E	✓	✓	✓	✓	✓
Helix-II	48	L	✓	✓	✓	✓	✓
Non-helix	51	L	✓	✓	✓	✓	✓
Non-helix	52	S	✓	✓	✓	✓	✓
Non-helix	53	A	✓	✓	✓	✓	✓
Non-helix	56	N	✓	✓	✓	✓	✓
Helix-III	63	E	✓	✓	✓	✓	✓
Helix-III	64	I	✓	✓	✓	✓	✓
Helix-III	72	D	✓	✓	✓	✓	✓
Non-helix	78	G	✓	✓	✓	✓	✓
Disorder-III	127	K		✓	✓		
Disorder-III	134	N	✓			✓	
Disorder-III	148	G		✓	✓	✓	✓
Disorder-III	155	A		✓		✓	✓
Disorder-III	238	D		✓	✓		
Non-helix	242	A		✓	✓	✓	
Helix-V	247	L	✓	✓	✓	✓	✓
Helix-V	248	D	✓	✓	✓	✓	✓
Helix-V	249	K	✓	✓	✓	✓	✓
Helix-V	250	L		✓	✓	✓	
Helix-V	256	T	✓	✓	✓	✓	✓
Non-helix	257	T	✓	✓	✓	✓	✓
Non-helix	259	N	✓	✓	✓	✓	✓
Non-helix	261	T	✓	✓	✓	✓	✓
Non-helix	267	D	✓	✓	✓	✓	✓
Helix-VI	273	N		✓	✓		
Helix-VII	288	E	✓	✓	✓	✓	✓
Helix-VII	295	L	✓	✓	✓	✓	✓
Non-helix	307	S		✓	✓	✓	
Helix-VIII	313	L		✓		✓	
Helix-IX	331	D		✓	✓		
Helix-IX	335	H		✓	✓		
Helix-X	354	D	✓	✓	✓	✓	✓
Helix-X	365	A	✓	✓	✓	✓	✓

4.3.6 Recombination

Evidence of intragenic recombination in the PvMSP-7E locus was analysed by using various principles implemented in the RDP package. The methods of RDP, GENCONV, Bootscan, MaxChi, Chimera, Siscan, and 3Seq identified evidence of 11, 8, 8, and 1 recombination events in isolates from Tak (the year 2008-2009), Tak (the year 1996), Ubon Ratchathani, and Yala-Narathiwat, respectively. Importantly, at least half of the recombination sites (29 of 56) were spanned along the central region of the sequence (Table 4.7).

4.3.7 Population differentiation

Genetic differentiation between parasite populations was estimated from the fixation index (*Fst*). The fixation index lies between 0 indicates no genetic differentiation and 1 infers two populations are substantially distinct. The *Fst* values of PvMSP-7E locus between *P. vivax* populations in Thailand and Colombian were displayed in Table 4.8. The *Fst* values were significantly greatest ($p < 10^{-5}$) between the populations from Tak (regardless year of collection) and Yala-Narathiwat (21.75%), and between Ubon Ratchathani and Yala-Narathiwat (19.44%), suggesting genetic differentiation or restricted gene flow between these endemic areas. Although the *Fst* value between Ubon Ratchathani and Tak (all samples) was low (0.82%), the differences were statistically meaningful ($p = 0.045$). Furthermore, significant deviation from zero of the *Fst* values was evidenced when Tak parasite populations were analysed according to the collection period in relation to Yala-Narathiwat populations. Consistently, limited gene flow was observed between parasite populations from Ubon Ratchathani and Tak collected in 1996 ($p = 0.018$) despite a small *Fst* value. By contrast, no significant *Fst* value was observed between parasite populations from Ubon Ratchathani and Tak collected during 2008-2009, implying the gene flow had occurred. Meanwhile, genetic differentiation between Tak populations collected in 1996 and during 2008 and 2009 did not deviate from zero ($p = 0.108$), suggesting genetic stability of parasite population in Tak.

Table 4.7 Recombination breakpoints in PvMSP-7E of Thai isolates. Evidences of recombination events were determined by using the Recombination Detection Program version 4 (RDP4) (Martin *et al.*, 2015). Recombination analyses were identified by a combination of methods such as, GENCONV, Bootscanning, the maximum Chi Square, CHIMAERA, Sister Scanning, and 3SEQ. Default settings were used to investigate the overall recombination signals. The table shows the recombination breakpoints in each parasite population in Thailand and the respective position.

Population	Recombination breakpoints					Method (*p* value)						
	Total	Between positions		Between domains		RDP	GENECONV	Bootscan	Maxchi	Chimaera	SiScan	3SEQ
Tak 1996	8	9	855	5'	3'	NS	NS	NS	NS	$2.00x10^{-2}$	NS	$2.00x10^{-6}$
		10	561	5'	Central	$4.90x10^{-13}$	$4.80x10^{-10}$	$2.00x10^{-12}$	$5.00x10^{-12}$	$3.23x10^{-12}$	$7.75x10^{-24}$	$4.94x10^{-24}$
		164	561	5'	Central	$1.52x10^{-11}$	$3.81x10^{-10}$	$5.33x10^{-8}$	$2.35x10^{-15}$	$1.17x10^{-13}$	$5.01x10^{-18}$	$4.90x10^{-26}$
		302	728	5'	3'	NS	$3.77x10^{-7}$	$4.11x10^{-9}$	$3.41x10^{-7}$	$3.23x10^{-5}$	$7.03x10^{-4}$	$1.82x10^{-11}$
		302	855	Central	Central	$1.11x10^{-8}$	$5.50x10^{-10}$	$8.74x10^{-13}$	$3.34x10^{-9}$	$3.23x10^{-8}$	$9.31x10^{-4}$	$2.42x10^{-9}$
		416	957	Central	3'	$1.24x10^{-2}$	$1.65x10^{-3}$	NS	$3.04x10^{-4}$	$4.98x10^{-6}$	$7.13x10^{-16}$	$1.28x10^{-8}$
		447	558	Central	Central	NS	NS	NS	NS	NS	NS	$5.77x10^{-8}$
		562	612	Central	Central	NS	$3.89x10^{-3}$	NS	NS	NS	NS	$1.15x10^{-5}$
Tak 2011-2016	11	33	302	5'	5'	NS	NS	NS	NS	NS	NS	$6.85x10^{-3}$
		164	603	5'	Central	$1.24x10^{-13}$	$2.54x10^{-12}$	$2.39x10^{-16}$	$1.12x10^{-14}$	$7.38x10^{-15}$	$2.62x10^{-11}$	$1.48x10^{-25}$
		299	569	5'	Central	$1.59x10^{-8}$	$4.55x10^{-5}$	$1.61x10^{-8}$	$1.35x10^{-12}$	$1.08x10^{-10}$	$2.29x10^{-13}$	$2.94x10^{-20}$
		302	533	5'	Central	NS	$5.57x10^{-10}$	$1.09x10^{-12}$	$9.32x10^{-12}$	$9.32x10^{-12}$	$7.41x10^{-23}$	$3.77x10^{-26}$
		446	560	Central	Central	NS	NS	$7.29x10^{-3}$	NS	NS	NS	$1.02x10^{-7}$
		459	520	Central	Central	NS	NS	NS	NS	NS	$9.53x10^{-11}$	NS
		561	1099	Central	3'	$6.71x10^{-10}$	$4.47x10^{-9}$	$1.86x10^{-10}$	$8.69x10^{-14}$	$3.44x10^{-13}$	$6.34x10^{-12}$	$2.23x10^{-25}$
		645	726	Central	3'	$1.77x10^{-3}$	$7.61x10^{-3}$	$1.80x10^{-3}$	$4.09x10^{-3}$	$2.83x10^{-3}$	$6.91X10^{-7}$	$9.60x10^{-6}$
		687	834	Central	3'	NS	NS	NS	NS	NS	NS	$6.85x10^{-3}$
		727	884	3'	3'	NS	$2.19x10^{-3}$	$3.56x10^{-4}$	NS	NS	$9.07x10^{-3}$	$4.74x10^{-6}$
		888	1038	3'	3'	NS	$6.53x10^{-5}$	$1.38x10^{-3}$	$3.27x10^{-6}$	$1.15x10^{-5}$	$5.52x10^{-9}$	$6.91x10^{-11}$
Ubon Rachathani	8	14	536	5'	Central	NS	NS	NS	$1.74x10^{-2}$	$8.43x10^{-3}$	$5.18x10^{-10}$	$6.20x10^{-8}$
		68	784	5'	3'	NS	NS	NS	NS	NS	NS	$7.92x10^{-4}$
		302	536	5'	Central	$5.80x10^{-7}$	$4.98x10^{-7}$	$7.89x10^{-11}$	$1.90x10^{-11}$	$6.19x10^{-14}$	$1.86x10^{-25}$	$2.27x10^{-23}$
		302	963	5'	3'	NS	$1.48x10^{-3}$	$5.62x10^{-5}$	$4.12x10^{-4}$	$1.93x10^{-4}$	$5.37x10^{-3}$	$1.24x10^{-9}$
		446	644	Central	Central	NS	NS	NS	NS	NS	NS	$1.53x10^{-3}$
		446	703	Central	Central	$2.14x10^{-9}$	$1.91x10^{-6}$	$2.17x10^{-9}$	$2.09x10^{-10}$	$6.21x10^{-7}$	$1.60x10^{-13}$	$6.96x10^{-14}$
		462	538	Central	Central	NS	NS	NS	NS	NS	$3.26x10^{-3}$	NS
		502	1078	Central	3'	NS	$3.10x10^{-7}$	$1.03x10^{-9}$	$3.58x10^{-11}$	$6.09x10^{-11}$	$1.81x10^{-12}$	$1.21x10^{-24}$
Yala-Narathiwat	1	628	784	Central	3'	$1.59x10^{-7}$	$9.49x10^{-9}$	$1.38x10^{-7}$	$6.04x10^{-4}$	$2.14x10^{-4}$	NS	$3.62x10^{-9}$

Table 4.8 Interpopulation variance indices of *P. vivax* populations in Thailand inferred from PvMSP-7E. The fixation index (*Fst*) was estimated using Arlequin program version 3.11(Excoffier *et al.*, 2005). Pairwise *Fst* values (lower diagonal) between *P. vivax* populations and their *p* values by permutation test (upper diagonal). Dashes indicate no comparison was done.

Province	Tak 1996	Tak 2008-2009	Tak (all)	Ubon Rachathani	Yala-Narathiwat	Colombia
Tak 1996		0.396	-	0.018	$< 10^{-5}$	-
Tak 2008-2009	0.0263		-	0.108	$< 10^{-5}$	-
Tak (all)	-	-		0.045	$< 10^{-5}$	0.036
Ubon Rachathani	0.0108	0.0075	0.0082		$< 10^{-5}$	0.027
Yala-Narathiwat	0.2473	0.2328	0.2175	0.1944		$< 10^{-5}$
Colombia	-	-	0.0070	0.0131	0.2387	

4.3.8 Phylogeny analysis

A maximum likelihood phylogeny was constructed using the Hasegawa-Kishino-Yano model and gamma distribution given the evolutionary invariable yielded the lowest BIC score (Figure 4.6). As observed in the phylogenetic analysis in Figure 4.5, there was no geographical clustering pattern for the PvMSP-7E locus between the parasite populations from Thailand and Colombian. Tree topology showed two major clusters of sequences, however, the bootstrap value was relatively low. For this scenario, the mosaic organization in the central region of the gene is likely to arise from the recurrent interallelic recombination events. Hence, that could result in the phylogenetic homogenization of PvMSP-7E locus.

4.3.9 Predicted linear B-cell and helper T-cell epitopes

From the analysis, the majority of the B-cell epitopes predicted to lie in the central region of the PvMSP-7E gene (Figure 4.7). Moreover, five predominant HLA-DRB1 haplotypes in Thai population were considered in the analysis including DRB1*0701, DRB*1202, DRB1*1501, DRB1*1502 and DRB1*1602. The putative CD4+ T cell epitopes did not a yield specific pattern, where these epitopes scattered in all domains of the protein. Nevertheless, amino acid substitutions in these epitopes likely to affect the predicted HLA-binding scores (Table 4.9).

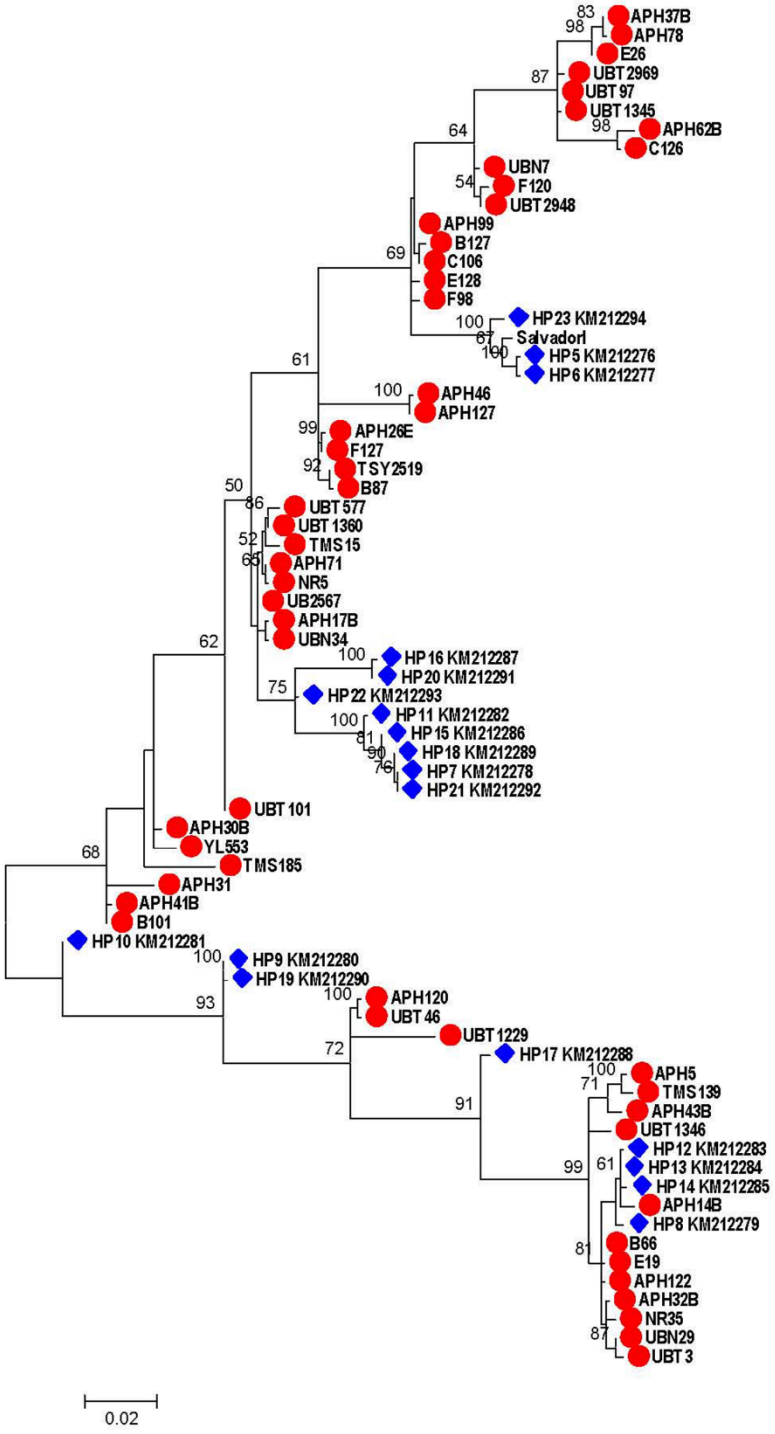


Figure 4.6 Maximum likelihood phylogenetic tree of *PvMSP-7E* based on Hasegawa-Kishino-Yano model and gamma distributed with invariant sites. The Tree was constructed using distinct sequences of Thai and Colombian isolates (closed triangle) comparing with the Salvador I strain (closed circle). Bootstrap values >50% are shown.

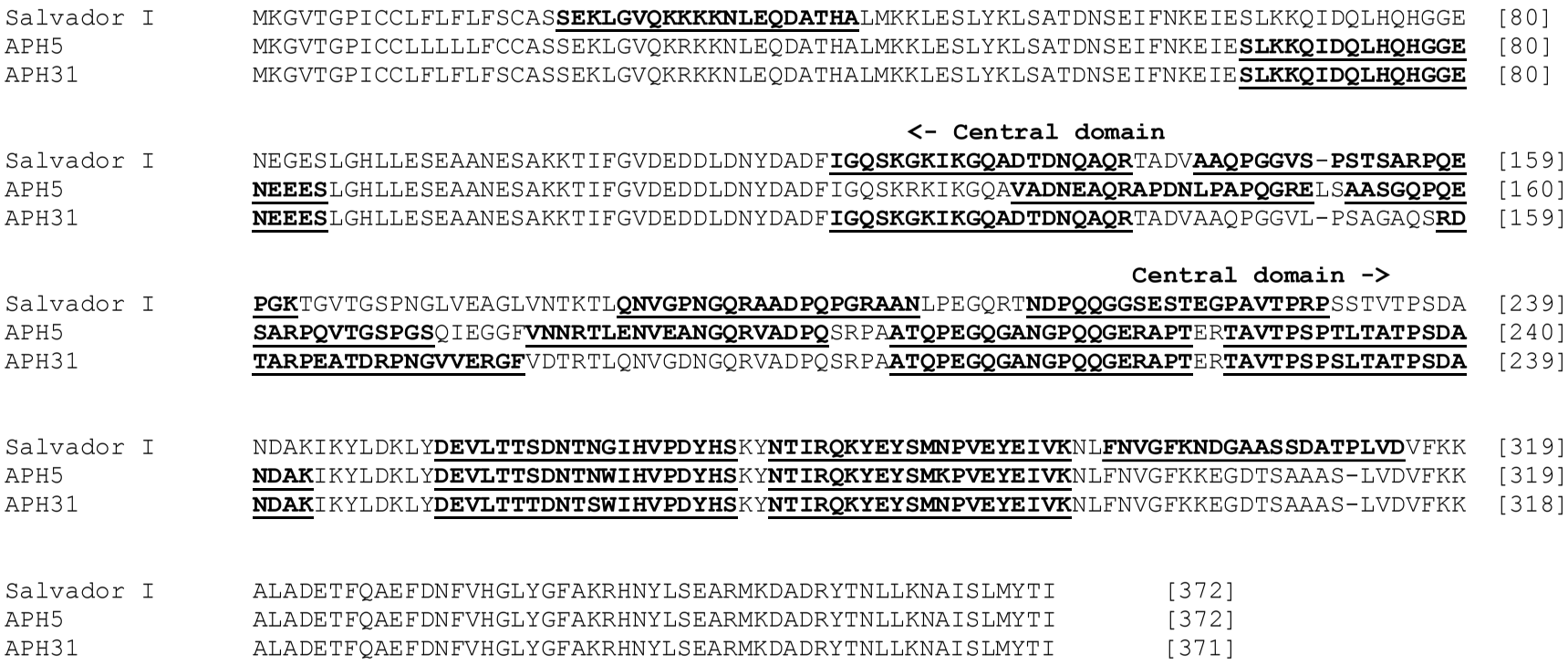


Figure 4.7 Predicted linear B-cell epitopes in PvMSP-7E of the Salvador I reference strain and two clinical isolates from Thailand (APH5 and APH31). The putative linear B-cell epitopes were predicted by using BCPred method implemented in BCPREDS B-cell epitope prediction web server (EL- Manzalawy *et al.*, 2008). The parameter was set to 90% classifier to improve the specificity.

Table 4.9 Putative CD4+ T cell epitopes in PvMSP-7E of the Salvador 1 reference sequence and two Thai isolates (APH5 and APH31) for predominant HLA-DRB1 haplotypes in Thai populations. The prediction was based on the PREDIVAC: CD4+ T cell epitopes prediction web-server with default settings (Oyarzún *et al.*, 2013). Five predominant HLA-DR alleles in Thai population were selected in the analysis including, DRB1*1202, DRB1*1502, DRB1*0701, DRB1*1501, and DRB5*1602 (Romphruk *et al.*, 1999).

HLA	Sequence	Residue	Domain	Score	Haplotype		
					Sal 1	APH 5	APH 31
DRB1* 1202	FIGQSK**R**KI	118-127	N-terminal	92.42		+	
	FIGQSK**G**KI	118-127	N-terminal	85.11	+		+
	VADN**E**AQR**A**	131-140	Central	85.42		+	
	DTDN**Q**AQR**T**	131-140	Central	<70	+		+
	V**GP**NGQR**A**A	186-195	Central	82.59	+		
	V**GD**NGQR**V**A	186-195	Central	82.59			+
	V**EA**NGQR**V**A	186-195	Central	82.59		+	
	YGFAKRHNY	338-347	C-terminal	88.22	+	+	+
	YTNLLKNAI	358-367	C-terminal	86.66	+	+	+
DRB1* 1502	**F**L**F**L**FS**CAS	12-21	N-terminal	86.32	+		+
	LLL**F**CCAS	12-21	N-terminal	<70		+	
	LDNYDADFI	111-120	N-terminal	80.92	+	+	+
	E**G**G**F**V**NNR**T	175-184	Central	80.62		+	
	E**A**G**L**V**NTK**T	175-184	Central	<70	+		
	E**R**G**F**V**DTR**T	175-184	Central	<70			+
	YEIVKNLFN	289-298	C-terminal	85.01	+	+	+
	FQAEFDNFV	326-335	C-terminal	84.40	+	+	+
DRB1* 0701	**F**L**F**LF**S**CAS	12-21	N-terminal	84.30	+		+
	LLL**F**CCAS	12-21	N-terminal	<70		+	
	F**S**CASSEKL	16-25	N-terminal	87.45	+		+
	F**C**CASSEKL	16-25	N-terminal	87.45		+	
	YKLSATDNS	49-58	N-terminal	81.44	+	+	+
	LSATDNSEI	51-60	N-terminal	84.99	+	+	+

HLA	Sequence	Residue	Domain	Score	Haplotype		
					Sal 1	APH 5	APH 31
	VT**GS**P**NGLV**	165-174	Central	81.40	+		
	VT**GS**P**GSQI**	165-174	Central	79.69		+	
	AT**DR**P**NGVV**	165-174	Central	<70			+
DRB1 *0701	LTATPSDAN	233-242	C-terminal	80.13		+	+
	STVTPSDAN	232-241	C-terminal	<70	+		
	YEYSM**N**PVE	280-289	C-terminal	82.32	+		+
	YEYSM**K**PVE	280-289	C-terminal	<70		+	
	FKKALADET	317-326	C-terminal	84.51	+	+	+
DRB1 *1501	MKGVTGPIC	1-10	N-terminal	81.77	+	+	+
	FL**F**LF**S**CAS	12-21	N-terminal	85.77	+		+
	LL**L**LF**C**CAS	12-21	N-terminal	<70		+	
	LFLFSCASS	13-22	N-terminal	82.31	+		+
	LLLFCCASS	13-22	N-terminal	82.31		+	
	LF**S**CASSEK	15-24	N-terminal	80.00	+		+
	LF**C**CASSEK	15-24	N-terminal	80.00		+	
	LESEAANES	90-99	N-terminal	80.85	+	+	+
	LDNYDADFI	111-120	N-terminal	86.10	+	+	+
	E**G**G**F**V**NNR**T	175-184	Central	84.19		+	
	E**A**G**L**V**NTK**T	175-184	Central	<70	+		
	E**R**G**F**V**DTR**T	175-184	Central	<70			+
	VDVFKKALA	314-323	C-terminal	81.79	+	+	+
	FQAEFDNFV	326-335	C-terminal	83.71	+	+	+
	LDKLYDEVL	247-256	C-terminal	81.62	+	+	+
	VKNLFNVGF	292-301	C-terminal	81.12	+	+	+
DRB5 *1602	E**G**G**F**V**NNR**T	175-184	Central	81.03		+	
	E**A**G**L**V**NTK**T	175-184	Central	<70	+		
	E**R**G**F**V**DTR**T	175-184	Central	<70			+

4.4 Discussion

MSP-7 in *P. falciparum* undergoes first proteolytic cleavage event to generate two protein fragments with 20-kDa ($MSP7_{20}$) and 33-kDa ($MSP\text{-}7_{30}$). Apparently, only the $MSP\text{-}7_{30}$ fragment in the C-terminal remains associated with the primary processing of MSP-1 (Pachebat *et al.*, 2007). Meanwhile, the secondary proteolytic event of PfMSP-7 derives a 19- or 22-kDa component which was detected in the MSP-1 complex. The cleavage sites were found to occur in the presence of glutamine residues, for instance between glutamine and glutamic acid, and glutamine and serine (Kadekoppala and Holder, 2010). Despite the association of PfMSP-7 and PfMSP-1 in the sequential proteolytic processing and involvement in invasion mechanism, nothing is known about the PvMSP-7 whether or not it undergoes the cleavage events parallel to PfMSP-7. Importantly, a consensus for *P. falciparum* subtilisin 1 (PfSUB1) cleavage site was discovered in the PvMSP-7E (Figure 4.8) (de Monerri *et al.*, 2011). For this reason, PvMSP-7E is likely to undergo a series of proteolytic cleavage events like PfMSP-7 and play a pivotal role in priming the merozoite prior to host cell invasion. PfSUB1 is known to mediate proteolytic events of several *P. falciparum* antigens including, MSP-1, MSP-6, MSP-7, and serine-repeat antigen protein (SERA) to regulate merozoite for its invasion process (Koussis *et al.*, 2009).

PfMSP7 (AF390150)	PLFQNLGLFGKNVLSKVKAQ▼SETDTQSKNEQEISTQGQEV	(176/177)
PvMSP7E (PVX_082665)	DLDNYDADFIGQSKGKIKGQ▼ADTDNQAQRTADVAAQPGGV	(129/130)
Variant 1	DLDNYDADFIGQSKRKIKGQ▼AVADNEAQRAPDNLPAPQGR	
Variant 2	DLDNYDADFIGQSKGKIKGQ▼TEGGDRTQSPADVAAPARGV	

PvMSP-7E. The cleavage site is represented by down-pointing triangles and amino acids residues between the cleavage sites are shown in parentheses after the sequences.

A study has been conducted on the nucleotide diversity of PvMSP-7E in Colombian population, albeit sample size was small (n=31) (Garzón-Ospina *et al.*, 2014). In the present study, PvMSP-7E in Thai isolates have shown comparable sequence diversity like those in Colombian isolates. The 5' and 3' regions of the gene were rather conserved. Although the 3' region contained more nucleotide substitutions than the 5' region, the extent of polymorphism was not statistically significant (Table

4.2). Purifying selection pressure was observed along the 5' and 3' region, suggesting the number of synonymous and nonsynonymous substitutions could stem from the structural or functional constraint of the protein. Furthermore, codon-based identification of deviation from neutrality also evidenced the majority of the negatively selected codons were found in these regions. Closer looks into the predicted protein secondary structure, a tight association between the protein secondary structure and the natural selection pressure was observed. Purifying selection seems to act in all helical fragments except α-helix-I, suggesting most of the α-helical structures were under the constraint of maintaining structure or function of the protein. Meanwhile, α-helix-I have been under positive selection as evidenced by the rate of d_N significantly outnumbered d_S. The putative signal peptide of PvMSP-7E encoded α-helix-I domain is likely to be shed from the precursor protein and remain independent with the MSP-1 complex (Kauth *et al.*, 2006). However, it is still to be explored whether the N-terminal signal peptide would confer immunogenicity during malaria infection. Positive selection spanning along the 5' signal peptide region suggesting its role in generating immune-evading mechanism. Amino acid substitutions at residues 12, 14, 16, and 17 of the α-helix-I domain likely to result in a change of CD4+ T-helper cell epitopes' predicted scores for peptide binding to the common HLA-DRB1 haplotypes among Thai isolates (Table 4.10) (Romphruk *et al.*, 1999).

Recombination clearly generates the PvMSP-7E nucleotide diversity in the central region. Despite the recombination signals distributed along the gene, most of the recombination breakpoints were located within the central region. Thus, a higher magnitude of nucleotide diversity in the central region may represent a mechanism of intragenic recombination between unique alleles. Closer looks into the 5' of the central region, the mosaic organisation of the sequences seem to arise from the interallelic recombination of three parental alleles. Meanwhile, it is noteworthy that the 3' fragment within the central region might stem from the recombination events by dimorphic parental alleles (Figure 4.4). This genetic organization is thought to be derived from distinct interallelic recombination during the sexual reproduction in anopheline vector and translate into sequence polymorphism at this locus. It has been proposed that the effective vector control may contribute to a reduction in nucleotide diversity among parasite population (Consortium, 2017). In addition, intragenic recombination may have a local impact on sequence polymorphism, where it preserves

adaptive traits or eliminates deleterious variants (Hughes, 2008). There is a consensus from the sequence analysis that the N-terminal and central polymorphic region of PvMSP-7E spans between predicted processing sites, similar to that of reported in subtilisin-like protease 1 in *P. falciparum* (Figure 4.7) (de Monerri *et al.*, 2011). Altogether these boundaries indicate the binding domains of MSP-7 to the first proteolytic event of MSP-1, where the C-terminal region of PfMSP-7 facilitates the protein-protein interaction (Kadekoppala and Holder, 2010).

A very high level of polar and charged amino acids such as glycine and proline residues was observed in the central region of the locus. In consensus, the entire central domain of PvMSP-7E displayed a predicted intrinsically unstructured or disordered protein. Despite intrinsically unstructured protein regions were also found in two clusters in the 5' domain (D1 and D2), they were relatively short spanning three α-helical domains (α-helix-II, III, and IV). Interestingly, these intrinsically disordered protein regions could provide a high degree of flexibility that enables the transition to structurally ordered regions upon functioning (Forman-Kay and Mittag, 2013; Guy *et al.*, 2015). Moreover, the N-terminal of PfMSP-7 has been identified to interact with host P-selectin receptors (Perrin *et al.*, 2015). P-selectin is a cell adhesion molecule that deposits on the host cell surface known to mediate disease severity during malaria infection (Combes *et al.*, 2004; Facer and Theodoridou, 1994). This feature on MSP-7 suggesting that the role in this family does not limit to erythrocyte invasion. Moreover, purifying selection operating in the N-terminal of the locus indicating the functional constraint present in the region.

The intrinsically disordered region spans along the central region of the PvMSP-7E, implying the structural plasticity is essential for modulating the molecular recognition or binding regions with other proteins (Guy *et al.*, 2015). Moreover, the protein binding regions were also predicted to lie within the central region of this protein (Figure 4.5). Meanwhile, departure from neutrality test had shown the rate of d_N higher than d_S supporting the positive selection pressure acting on the central region. Closer looks into the central region, the positive selection signal was exclusive to just the 3' domain which is the 3'-dimorphic subregion. Therefore, the 3'-dimorphic region not only predicted to be a binding region to the MSP-1 complex but also essential for the parasite to escape the host's immune system. Previous mice challenge model was

conducted using MSP-7 in *P. yoelii*, however, it did not confer protective immune responses against lethal infection (Mello *et al.*, 2004). In the present study, the amino acid substitutions in PvMSP-7E could potentially reduce the predicted scores for HLA-bindings for CD4+ T helper cell epitopes and predicted scores for linear B-cell epitopes, predominantly within the central region. However, these findings were predicted bioinformatically, the immunogenicity of PvMSP-7E in natural infection remains to be explored.

The malaria prevalence in Thailand has an overall decline pattern for the past three decades. However, the malaria cases have been fluctuating in several areas in Thailand, particularly those areas along the international borders including, Myanmar, Cambodia, and Malaysia. Analysis of sequence diversity in PvMSP-7E revealed that Yala-Narathiwat population has significantly lower haplotype diversity than those parasite populations in Ubon Ratchathani and Tak. Yala-Narathiwat, Ubon Ratchathani, and Tak showed 3, 19, and 34 number of haplotypes, respectively. Consistently, this pattern was reported in the previous population analysis in Thailand including, MSP-5, AMA-1, and PvTRAP (Kosuwin *et al.*, 2014; Putaporntip *et al.*, 2009b; Putaporntip *et al.*, 2010). Malaria transmission in Yala-Narathiwat used to keep at the minimum until a substantial increase in the past few years due to the lack of control strategy, resulting in bottleneck effects. It is noteworthy that, malaria transmission has been under control in most parts of Thailand, but transmigration of malaria cases is still persisting along the Thai-Myanmar and Thai-Cambodia borders. Thus, bottleneck effects can be envisaged among the parasite population in southern Thailand. Interallelic recombination is not uncommon of, as reported by previous malarial antigens in Thailand (Kosuwin *et al.*, 2014; Putaporntip *et al.*, 2009b; Putaporntip *et al.*, 2010). Analysis of recombination breakpoints in PvMSP-7E among Thai isolates showed between 1 (Yala-Narathiwat population) to 11 (Tak population collected in the year 2008-2009). Importantly, the number of recombination breakpoints is positively correlated with the level of haplotype diversity (r=0.941, p=0.059), implying intragenic recombination enhances the magnitude of haplotype diversity. Furthermore, the non-zero recombination breakpoints in parasite population from Yala-Narathiwat province further justified the bottleneck effects rather than the sudden clonal expansion in the area.

The phylogenetic relationship did not generate specific clusters belonging to the sequences from each malaria endemic areas in Thailand. Likewise, no unique clade was observed when Colombian isolates were introduced into the phylogeny analysis. Moreover, pairwise comparison of genetic differentiation between *P. vivax* populations in Thailand has shown significant strong population structure. Flight range of mosquito, cross-movement migration, and geographic distance are factors that could influence the genetic diversity. Meanwhile, *P. vivax* isolates collected during the year 1996 and year 2008-2009 did not yield significant population differentiation. This is consistent with the previous study using PvTRAP as a marker in Thailand (Kosuwin *et al.*, 2014). *P. vivax* populations in Thailand shows spatial but not temporal variation. Strikingly, the population differentiation between Ubon Ratchathani and Tak collected during the year 2008-2009 revealed a low fixation index and not statistically significant, implying gene flow between these two parasite populations. Malaria cases in Ubon Ratchathani were mainly indigenous before the illegal logging occurred at a large-scale. The illegal deforestation in Ubon Ratchathani during the sample collection period by locals and migrants from other provinces could have influenced the genetic diversity of parasite populations in Thailand.

4.5 Conclusion

Our results have shown that the extent of sequence polymorphism in PvMSP-7E locus among *P. vivax* populations in Thailand likely to be influenced by natural selection pressure and intra-allelic recombination. The levels of haplotype diversity are varied between endemic areas in Thailand, Yala-Narathiwat provinces revealed a low number of haplotypes suggesting bottleneck effects. Natural selection forces acting differently on the locus likely to associate with its predicted protein secondary structure. The α-helical domains are seen to be less tolerant to molecular adaptation than intrinsically unstructured domains. The insights gained in the present study could contribute to the rational design of the functional study and potential vaccine candidate.

Chapter 5

Clinical expression profiles of a *Plasmodium vivax* vaccine candidate: merozoite surface protein 7 (PvMSP-7)

Abstract

The previous chapter reveals the heterogeneous genetic diversity pattern of PvMSP-7 multigene family, suggesting not all paralogs are functionally equivalent. The precise roles of the PvMSP-7 paralogs have not been established, although certain of its orthologous genes in *P. falciparum* were shown to impair erythrocyte invasion. Using RNA-seq technology, it will channel to a better understanding of PvMSP-7 functional diversity by uncovering its expression profiles in natural infection. The transcriptional changes of PvMSP-7 paralogs through the intraerythrocytic development cycle (IDC) were shown using co-expression analysis. Ten field isolates present asynchronous parasite composition were sequenced by RNA-seq. Ten patients were divided into four clusters based on the principal component analysis using genome-wide expression profiles. Differentially expressed PvMSP-7 genes were identified through pairwise comparison of patients groups. The association of PvMSP-7 genes was assessed with cohorts of stage-regulated genes using co-expression analysis. Three PvMSP-7 paralogs, -7A, -7F, and -7M were shown to express constitutively in all clinical isolates. In contrast, PvMSP-7H and PvMSP-7I are significantly upregulated in two patients who experienced longer patency. These two genes demonstrated a signature co-expression with a schizont stage marker, while negatively correlated with liver stage and gametocyte stage markers. All lines of evidence support the developmental regulation of PvMSP-7 family during the IDC. The PvMSP-7A, -7F, and -7M were suggested to have additional functions besides host cell invasion. Therefore, the PvMSP-7 paralogs are not all functionally equivalent, comparatively brief expression of some PvMSP-7 paralogs should be a consideration in vaccine design.

5.1 Introduction

In Chapter 1 (section 1.11.3), the role of MSP-7 in *Plasmodium* was described. This role in the host-parasite interaction suggests that PvMSP-7 may have potential as a vaccine candidate. In this chapter, the functional distinctions among PvMSP-7 paralogs were examined using transcriptional profiling. To date, global transcriptional analyses of *P. vivax* have been conducted using microarray (Bozdech *et al.*, 2008) and RNA sequencing (Zhu *et al.*, 2016). Both approaches produced transcriptional profiles for synchronized *P. vivax* cell cultures across the 48-h intraerythrocytic developmental cycle (IDC). When the transcriptional profiles of IDC-specific genes were compared between *P. vivax* and *P. falciparum*, expression of syntenic genes in *P. vivax* was equally distributed across the IDC, but skewed slightly to the trophozoite-schizont transition, while non-syntenic genes were seen to express predominantly during schizont-ring stage transition (Bozdech *et al.*, 2008). Assuming that genes involved in host interaction are most dynamic and most likely to be non-syntenic, these observations suggest the initiation of host-parasite interaction during this transition.

Zhu and colleagues, in their RNA-seq study further refined the findings from microarray data. The study reveals that expression levels of highly expressed genes tend to peak during the late ring stage to mid schizont stage whereas, lower expressed genes seem to peak at the late schizont stage (Zhu *et al*. 2016). This pattern suggests that functional differences are clearly apparent in the comparison of different developmental stages.

MSP-7 transcripts were previously detected in the blood-stages of four *Plasmodium* species including *P. falciparum*, *P. vivax*, *P. berghei*, and *P. yoelii* (Bozdech *et al.*, 2008; Kadekoppala *et al.*, 2010; Mello *et al.*, 2004; Otto *et al.*, 2014; Otto *et al.*, 2010). It is currently assumed that all the MSP-7 paralogs have the similar transcriptional profile across the intraerythrocytic developmental cycle. The present study aims to investigate the MSP-7 expression pattern directly in the clinical isolates. This is essential in vaccine design because malaria confers stage-specific immunity (Cohen, 1979). The constitutive expression of a malaria antigen is likely to target different life stages of the parasite, therefore, provide a larger coverage to elicit malaria immune responses.

Recently, three clinical field isolates, composed of diverse parasite life stages, suggested that gene expression profiles are remarkably similar across infections (Kim *et al.*, 2017), which the researchers attributed to the dominant and homogenizing effect of asexual stage gene expression. One of the three isolates had a higher proportion of gametocytes, but although a few gametocyte genes were found to be significantly enriched such as *Pfs25*, its pattern of gene expression was not significantly different from the two other isolates (Kim *et al*., 2017). However, with only three isolates, the study has insufficient power to draw a conclusion, although it does demonstrate the feasibility of transcriptional profiling direct from the blood. To examine the gene expression dynamics of all MSP-7 genes in *P. vivax* clinical isolates, and so infer functional differences between paralogs based on distinct patterns of developmental regulation, ten vivax malaria patients were recruited from the field and generated *P. vivax* RNA-seq data from blood samples.

Co-expression analysis integrates the differential gene expression data between two conditions to identify the potential gene clusters. The clusters provide valuable indications on the genes with unknown biological functions (Dam *et al*., 2017) . Construction of gene expression clusters with the focus on PvMSP-7 genes could, therefore, provide some putative functions of this family. This approach has been widely applied to study disease-associated markers, to the identification of *P. falciparum* genes responsible for drug resistance mechanisms and parasite survival (Subudhi *et al.*, 2015). This study used co-expression analysis to elucidate the novel function of Cytochrome C heme-lyase in *P. falciparum*, based on its transcriptional profile, which clustered with a long chain fatty acid elongation enzyme, an aquaglyceroporin protein, and an acyl-CoA synythetase, which have known functions. Hence, co-expression analysis suggested that the Cytochrome C heme-lysae might play a significant role in osmotic protection during merozoite and ring stage, and therefore, might be a valid drug target (Subudhi *et al.*, 2015). Another study utilised 53 time-points of four intraerythrocytic development cycles in *P. falciparum* (Yu *et al.*, 2013). The analysis identified ten clusters with potential functions linked to DNA replication, adhesion to the host surface, and transcriptional regulation. For instance, the analysis found PFD0885c; a conserved *Plasmodium* protein clustered with three other genes SIP2, MAL8P1.153, and MSP-9 which potentially play a similar role in transcriptional regulation (Yu *et al.*, 2013).

Our understanding of gene structure and allelic diversity of MSP-7 multigene family in *P. vivax* has been described in population genetic studies. However, no studies have addressed the extent of transcription diversity of PvMSP-7 family in natural infection. In the study, RNA sequencing was used to profile the gene expression patterns of ten vivax-malaria patients present different proportion of parasite life stages. In the present study, gene expression patterns of PvMSP-7 and co-expression with developmentally regulated markers were revealed. Ten patients were divided into four groups based on the genome-wide expression profiles and characterised the differentially expressed genes (DEGs) between each group. The DEGs were subjected to co-expression analysis, and multiple co-expressed gene clusters were identified. Using the co-expression approach, the distinct role of PvMSP-7 genes was described, potentially undertake in different developmental stages. Critically, the work provides new insight into which PvMSP-7 paralogs are highly expressed and should be the primary focus of malaria subunit vaccine development.

5.2 Methodology

5.2.1 Study design and sample processing

Ten patients were recruited from two malaria endemic areas in Thailand. Of these, five samples were collected from Ubon Ratchathani province located Northeast of Thailand along the border between Thailand and Cambodia and five samples from Yala province located South of Thailand along the border between Thailand and Malaysia. They were asked to provide information on the days of fever they had (Table 5.1). The ten patients involved in the study were recruited voluntarily; blood specimens were collected after obtaining informed consent from the participants and all research procedures were performed in accordance with guidelines approved by the Institutional Review Board in Human Research of Faculty of Medicine, Chulalongkorn University, Thailand with registered number IRB No. 104/59. Patient blood was screened with a light microscope by an experienced laboratory technician, then around 600μL of venous blood samples were collected from *P. vivax*-infected patients. Blood specimens were preserved in RNAlater® solution (Ambion, Grand Island, NY, USA) with a 1:1 ratio. Additionally, 200μL of fresh blood samples without any preservative were spotted onto the Whatmann™ 3MM ChR 3 filter paper (Cat. No. 3030917, Maidstone, England). Molecular diagnosis of all samples collected was according to the methodology described in Chapter 2. All blood samples were stored at -20°C until processed. RNA was extracted from 500 ul of blood preserved in RNALater® using a QIAamp RNA blood mini kit (Qiagen, Hilden, Germany). All procedures were processed as per manufacturer's recommendations. First, the venous blood sample was mixed with buffer RLT. Then, the sample was vortexed with β-mercaptoethanol and centrifuged at 5000 rpm for an hour at 4°C. The supernatant was transferred into the RNeasy® spin column. The column was washed two times with Buffer RW1. Finally, the RNA was eluted with to 100 uL of elution buffer. RNA samples were quantified using a NanoDrop spectrophotometer (NanoDrop Technologies, Delaware, USA) and Qubit® 3.0 fluorometer (Thermo Fisher Scientific, Waltham, USA) and stored at -80°C.

Table 5.1 Ten patients diagnosed with *P. vivax* infection were recruited in the study from two malaria clinics in Thailand: Ubon Ratchathani and Yala. The age of the patient varied from 14 to 50 years old. Information regarding days of patency was noted when patients informed the medical officers during their visit.

Sample	Age (years)	Gender	Province	Days of patency
UBT3086	50	Male	Ubon Ratchathani	2
UBT3087	None	None	Ubon Ratchathani	3
UBT3089	41	Male	Ubon Ratchathani	2
UBT3090	14	Male	Ubon Ratchathani	2
UBT3091	37	Male	Ubon Ratchathani	3
YL3111	26	Female	Yala	2
YL3112	45	Male	Yala	2
YL3113	15	Male	Yala	7
YL3114	30	Female	Yala	2
YL3115	46	Female	Yala	7

5.2.2 RNA sequencing

Ten RNA samples were treated with DNase, followed by library preparation using Epicentre Globin-Zero Gold kit to deplete rRNA and globin transcript. The quality of the libraries was assessed by Agilent 2100 Bioanalyser. The RNAs appeared to have been degraded as smearing observed in the Bioanalyzer traces. Despite the degradation of RNAs, library preparation proceeded with Globin-Zero Gold kit, which is able to process degraded RNA effectively (Zhao *et al.*, 2018). RNA-seq libraries were prepared at the Centre for Genomics Research, the University of Liverpool using the NEBNext Ultra Directional protocol. The modified Globin-Zero protocol began with an initial input of 100ng RNA. The RNA-seq libraries were sequenced on an Illumina HiSeq4000 platform to generate approximately 30 million paired-end reads of 150 bp for each sample. The resulting RNA-seq data have been deposited in the ArrayExpress database at EMBL-EBI under accession number E-MTAB-6753 (www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6753).

5.2.3 Bioinformatics processing

The initial processing and quality assessment of the raw reads begun with base calling and de-multiplexing of indexed reads using *CASAVA* version 1.8.2 (Illumina) to produce ten samples from one lane of sequence data in fastq format. The Illumina adapter sequences were trimmed from the raw fastq files by using *Cutadapt* version 1.2.1 (Martin, 2011). The option "-O 3" was specified to trim off the 3' end of any reads that matched to the adapter sequence over at least three bp. The reads were further trimmed to remove low-quality bases, using *Sickle* version 1.200 with a minimum window quality score of 20. Before mapping the filtered reads to the reference genome, reads shorter than ten base pairs (bp) were removed. Subsequently, all the filtered reads were mapped to the human genome (GRCH37) using *TopHat2* (Kim *et al.*, 2013). The unmapped reads were then aligned to the PvP01 reference genome (Auburn *et al.*, 2016) using *TopHat2*. *TopHat2* is a sophisticated aligner designed specifically for RNA-seq data where it can detect splice junction and generate splice alignment. Short reads were split into smaller fragments and mapped independently. The segment alignments were then joined together in the final phase to produce end-to-end reads alignments. Paired-

end reads in the FASTQ file were used to generate the read alignment. In the alignment, mapped reads that have more than two mismatches were discarded and –read-realign-edit-dist 0 was activated to map every read against genome that is detected by the aligner. These parameters were able to address the problems of reads spanning multiple exons and improve the accuracy of spliced mapping in *TopHat2*. Following the alignment of reads to the *P. vivax* genome, the read counts of each gene were counted with *featureCounts* (Liao *et al.*, 2013). The read counts estimated by dividing the total number of mapped reads to the length of the gene then scaled the estimates to one million. The genome-wide coverage of each sample was determined using *Qualimap 2* (Okonechnikov *et al.*, 2015).

5.2.4 Differential genes expression

The read counts were estimated by *featureCounts* (Liao *et al.*, 2013) as input to *DESeq2* (Love *et al.*, 2014). *DESeq2* is a Bioconductor package implemented in R to identify the differentially expressed genes. An assumption was made about the distribution of read counts where the mean of read count was equal across each sample. The program identifies differentially expressed genes (DEGs) by normalising the read counts per million (CPM). Genes with a false discovery rate (FDR) threshold below 0.05 were considered significantly differentially expressed. Expression correlations were calculated using the normalised read counts. The correlation estimation was performed to assess the equality of the gene expression profile among ten samples. The high correlation coefficient indicates the reproducibility and reliability of the gene expression dataset. All expression values were added one, then log-transformed (log_2) to generate more equal variance with low read counts. Pearson correlation was estimated between each sample in R version 3.4.3 and corrplot (Taiyun and Viliam, 2017). Principal component analysis (PCA) was performed using the genome-wide CPM per gene derived from *DESeq2* in ten patients. The plot formed four clusters of individuals which subsequently used to estimate the DEGs in pairwise comparison manner (Figure 5.1). The analysis was also conducted using *EdgeR* (Robinson *et al.*, 2010) and *Cuffdiff* (Trapnell *et al.*, 2012). PCA plots were generated using the genome-wide expression values derived from *EdgeR* and *Cuffdiff* (Figure 1). Based on Figure 5.1, *EdgeR* algorithms clustered ten patients into four groups consistent with *DESeq2*.

However, in *Cuffdiff* the patients were scattered throughout the plot and did not form a specific pattern. This could stem from the different approaches used in each algorithm. *Cuffdiff* uses transcript-based method whilst, *DESeq2* and *EdgeR* use negative binomial model. A recent study has reported the poor performance of *Cuffdiff* in detecting DEGs owing to the uncertainty of read count using the conservative approach (Seyednasrollah *et al.*, 2013). For this reason, the results from *Cuffdiff* were not included in the study. Moreover, the DEGs identified using *DESeq2* and *EdgeR* did not deviate significantly. As the overall sensitivity of *DESeq2* and *EdgeR* seemed comparable, only the results from *DEseq2* were used for downstream analyses.

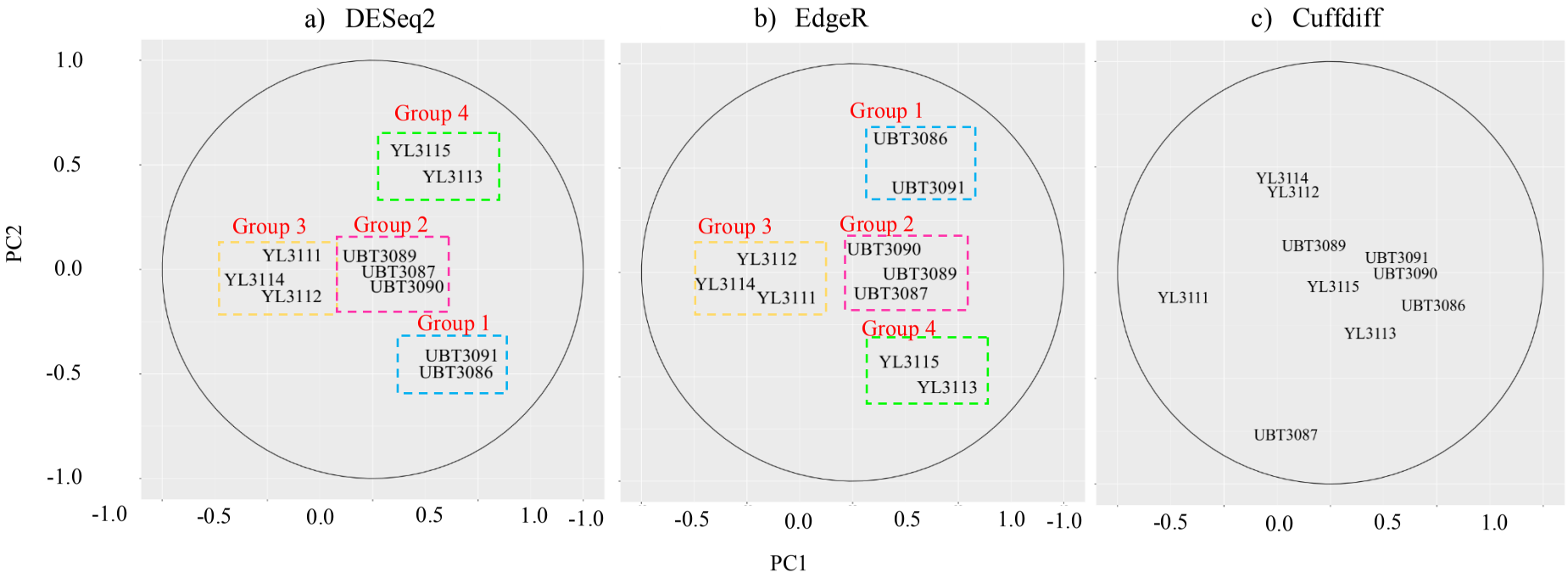


Figure 5.1 Principle component analysis (PCA) plots generated using three approaches. Genome-wide expression values were obtained using the respective approach in ten patients, a) DESeq2, b) EdgeR, and c) Cuffdiff. Patients were divided into groups based on the PCA plot. Each identifier represents each patient and the colour is assigned according to the groups (sky blue: group one, pink: group two, yellow: group three, and light green: group four).

5.2.5 Co-expression analysis

Ultimately, co-expression analysis was used to identify genes co-regulated with MSP-7 that with clear functions that could be used to infer the potential biological functions of PvMSP-7 genes during a specific developmental stage. *coseq*, an R-based package was used to estimate the clusters of co-expressed genes (Rau and Maugis-Rabusseau, 2017). The DEGs identified in *DESeq2* were imported into *coseq* for identifying clusters of co-expressed genes. The pipeline relies on the Gaussian mixture models with clustering all the genes based on the proportion of normalised counts in each expression profile. The data were fitted with a Gaussian mixture model on either arcsine- or logit-transformed normalised profiles. One hundred clusters were tested in each transformation with the following commands; arcsin_transformed <-coseq(counts, K=2:100, model="Normal", transformation="arcsin") and logit_transformed <-coseq(counts, K=2:100, model="Normal", transformation="logit"). To choose accurately between two transformation models, *coseq* calculated the corrected integrated completed likelihood (ICL) values from these two models, and the number of clusters and preferred model-transformation is selected via the highest corrected ICL value.

5.2.6 Enrichment analysis and pathway identification

Gene ontology analysis was used to investigate the enriched functions in each cohort of co-expressed genes. The identification was carried via PlasmoDB webserver release 35 (Aurrecoechea *et al.*, 2008). The PlasmoDB database has plugins Gene Ontology (GO) (Ashburner *et al.*, 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) which used to identify the biological processes and generate pathway maps for each cluster of genes. Heatmaps were generated based on the DEGs from each analysis. The heatmaps were divided into several sectors according to the clades generated in dendrograms and genes contained inside each sector were subject to gene ontology enrichment analysis. GO terms and pathways were considered significant such that the adjusted *p*-value below 0.05. Only unique and non-redundant biological and functional information was retained in the downstream interpretations. The GO terms derived from the same set of genes were removed from the analysis.

5.2.7 SNP discovery using RNA-seq samples

The expression profiles of PvMSP-7 were conducted extensively using differential gene expression and co-expression analysis. It is equally interesting to address whether genetic variants account for differences in gene expression changes across geographical locations. SNP calling was used to address the question concerning the association of geographical locations to gene expression profiles. Each RNA-seq sample was subjected to SNP calling using GATK version 3.7 (McKenna *et al.*, 2010). Reads in FASTQ files were processed bioinformatically and mapped to the PvP01 reference genome as explained in section 5.2.3. The resulting BAM files were used as the inputs for SNP calling based on the GATK's best practice pipeline. The best practice pipeline was verified by GATK developers using large-scale human dataset and made SNP calling with RNA-seq technically possible and reliable. There are certainly some limitations present in the pipeline as the sensitivity of SNP calling can be influenced by short-reads, read depth, and complex regions. However, some sophisticated aligner such as *TopHat2* (Kim *et al.,* 2013) is designed to address the problem of spliced reads alignment. In the SNP calling pipeline, Split"N"Trim was used in the first step of processing. RNA reads were gapped by intronic regions where "Ns" filled the regions. For this reason Split"N"Trim was used to split reads into exon segments and remove the "Ns" to prevent false positives. Local re-alignment was performed to correct mismatches contribute by indels. Base recalibration was activated to re-estimate base qualities to generate more accurate base quality scores for SNP calling. Lastly, SNP calling was executed using HaplotypeCaller in GATK with the capability to distinguish the intron-exon split regions (McKenna *et al.*, 2010). All variant sites with a Phred-scaled confidence threshold of at least 20 were output to an initial variant dataset in VCF format. In total, 68,258 SNPs were identified in the initial variant calling. Hard filters were used to filter the resulting dataset, to avoid false-positive variants. The filter clusters specific to RNA-seq analysis was performed with the default setting. Filter cluster of at least three SNPs within a window of 35 bases was applied, to improve sensitivity and specificity of detecting real variant. Variant sites with Fisher strand values of at least 30 were retained in the dataset. Fisher strand value is useful to detect any strand bias such that only one variant at a specific site observes either on the forward or reverse strand. Variant sites with a confidence value below two were discarded. The confidence values were determined by the Phred score and the depth of

the samples. However, as the criteria of hard filtering were very restrictive, some real variants could have been filtered in the process. The final dataset contained 42,988 high-quality SNPs that used in the downstream analyses.

5.2.8 Population analyses

The approaches used for the population analyses are similar to those described in section 2.2.15. The phylogenetic associations between ten samples were assessed using PCA and phylogenetic analyses based on 42,988 biallelic SNPs derived from SNP calling. PCA plot was generated using SNPRelate (Zheng *et al*., 2012) implemented in R environment version 3.3.1 (R, 2016). The top two principal components were used to plot the PCA. Phylogeny trees were further used to infer the population structure of ten samples. Variants of ten samples contained within the VCF file were concatenated into a single FASTA file using VCF-kit (Cook and Andersen, 2017) and used as an input for phylogenetic analyses. Maximum likelihood tree was generated using RAxML with the GTR substitution model (Stamatakis, 2014). The optimum substitution model was determined by jModelTest, version 2.0 (Posada, 2008). A Neighbour-joining tree was constructed using MEGA 7.0 (Kumar *et al*., 2016) based on the maximum composite likelihood method.

5.3 Results

5.3.1 Patient summary information

Across 10 RNA samples, RNA-seq generated an average of 30 million paired-end reads with 150 bp insert sizes (Table 5.2). As the samples were collected from clinical patients, the reads were aligned to the human genome (GRCh37) and subsequently used the unmapped reads to align to the *P. vivax* P01 reference genome (Auburn *et al.*, 2016). More than 70% of reads aligned to the human genome, while reads mapped to *P. vivax* ranged between 0.78 – 22.38%. The ten samples showed mean genome coverage of between 1.07X – 78.52X. UBT3089 has the highest mean coverage; 78.52X, while five samples had a level of coverage below 10X.

5.3.2 Sequencing metrics

Table 5.2 Summary statistics of mapping for the ten samples on to human genome GRCh37 and *P. vivax* P01 genome. The illumina adapter sequences were trimmed using *Cutadapt* version 1.2.1. The reads were further trimmed to remove low-quality bases, using *Sickle* version 1.2 with a minimum window quality score of 20. Genome-wide mean coverage was calculated using *Qualimap 2*.

Sample		Total read pair number[1]	Pair reads mapped to human	Percentage of reads mapped to human	Pair reads mapped to *P. vivax*	Percentage of reads mapped to *P. vivax*	Mean Coverage
1	UBT3086	29,157,609	24,317,087	83.40	346,899	1.19	1.82
2	UBT3087	36,549,438	31,324,232	85.70	977,753	2.68	17.67
3	UBT3089	30,083,802	24,588,757	81.73	3,035,144	10.09	78.52
4	UBT3090	30,397,809	23,639,952	77.77	3,267,128	10.75	17.59
5	UBT3091	28,550,673	19,658,903	68.86	6,388,456	22.38	42.07
6	YL3111	29,500,938	27,008,110	91.55	580,181	1.97	3.18
7	YL3112	32,578,915	28,978,835	88.95	913,421	2.80	4.95
8	YL3113	26,853,921	23,042,788	85.81	209,249	0.78	1.07
9	YL3114	31,860,822	28,820,223	90.46	350,337	1.10	1.88
10	YL3115	32,987,343	25,882,012	78.46	3,469,628	10.52	20.13

[1]After adapter and quality trimming

5.3.3 Estimation of transcript abundance values

The sequencing reads were mapped to the human reference genome (GRCH37). The unmapped reads were subsequently aligned to the *P. vivax* P01 reference genome using featureCounts (Liao *et al.*, 2013). In total, the mapped reads were counted for 6642 genes in ten samples to reveal the expression profiles. The read count of each gene was fitted to the negative binomial distribution to calculate the sample variance. This step is important to account for the read count bias in the differential expression analysis (Yoon and Nam, 2017). A dispersion estimates was generated by measuring the gene count of all samples. In Figure 5.2, a smooth red line was observed implying the average expression strength. Therefore, it is sensible to assume the read count corresponds to each gene is gene-specific variation.

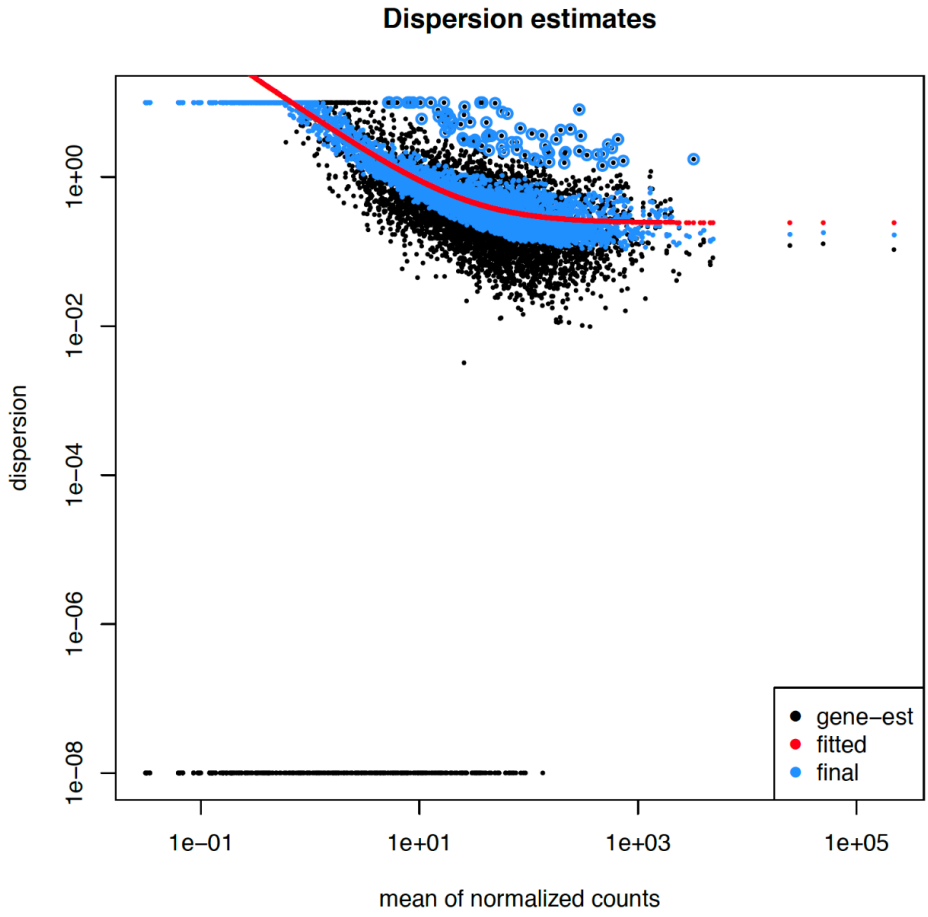


Figure 5.2 Dispersion estimates derived from DESeq2 and gene count measures between ten clinical samples. The black dot indicates the dispersion of each gene, the red line shows the dispersion of all samples, and the blue dot shows the corrected value for the gene.

5.3.4 Correlation of each RNA-seq sample

To evaluate variation in data quality among isolates, Pearson's correlation coefficient (*r*) was calculated between the corresponding transcript abundance values for each sample (Figure 5.3). The gene-level estimates after *DESeq2* normalisation showed a high degree of robustness as the variability evidenced by the Pearson's correlation between each isolate ranging from 0.59 to 0.91. Although a reduction in the correlation between sample UBT3090 and UBT3086 (r=0.59) was observed, in the downstream analyses this sample was grouped with other samples. Thus, individual variation among a group would not influence the results. The estimation was based on a set of 6,642 genes in ten samples.

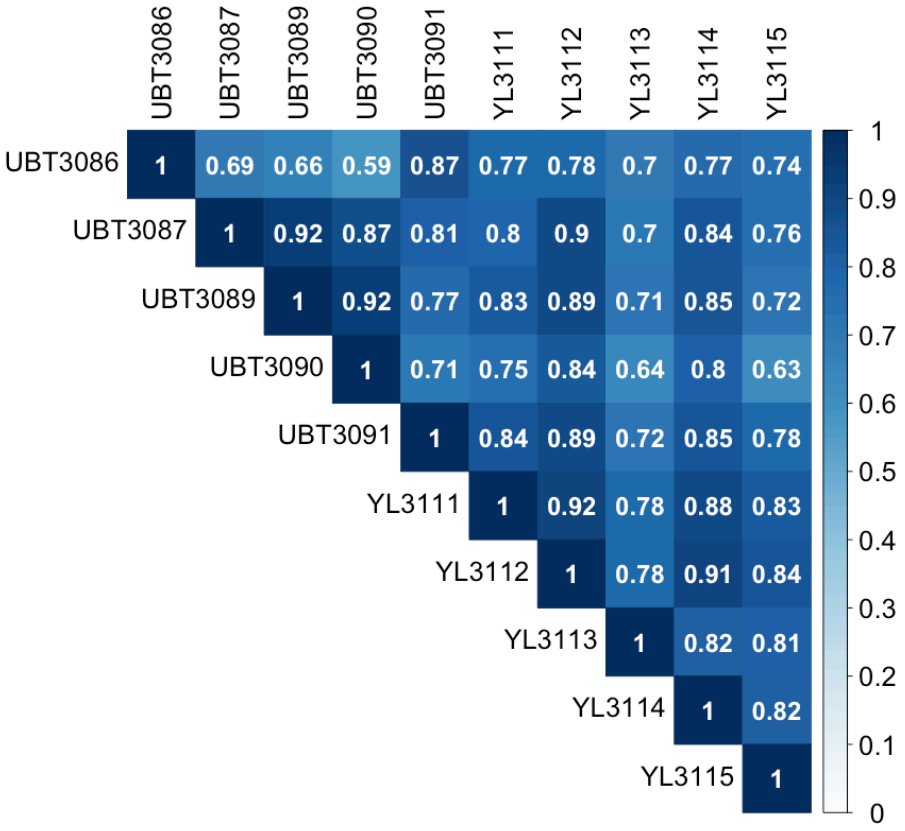


Figure 5.3 Correlation of gene expression patterns between each sample estimated based upon Pearson's correlation coefficient (*r*). The similarity was determined by the $\log_2$ transformation of normalised raw read counts derived from *DESeq2*. Highest correlation obtained was 0.91 whilst, the lowest was 0.59. Total read counts of 6,642 genes were used in the estimation.

5.3.5 Principal component analysis (PCA)

Clinical isolates infected with *P. vivax* often contain mixture a composition of parasite developmental stages. As samples were collected from the malaria clinic in the rural areas, microscopic slides were not preserved. Regrettably, the proportions of parasites developmental stages from microscopic slides were unable to be determined. Therefore, the relative proportions of parasite stages were characterised indirectly, using genome-wide expression profiles. Ten patients were separated into groups based on a principal component analysis of transcript abundance values. The principal component analysis was independent of age and days of fever. From the analysis, ten patients were clustered into four distinct groups using the expression profiles of 6,642 genes derived from the *DESeq2* package (Figure 5.4). The groups were labelled from one to four with the respective colours. Interestingly, two patients that both experienced longer patency (i.e. seven days) clustered in Group 4. Days of fever experienced by the patients solely relied on the conversation during their visit to the malaria clinic. Patients clustered in Group 1 to Group 3 had fever ranging from two to three days. As four distinct clusters separating ten patients were observed, it was of interest to know the genes that are significantly differentially expressed between each group, and if these could be used to infer the developmental state of the parasite population in a given sample, and if this had a specific relationship with expression of 13 PvMSP-7 paralogs. Differential gene expression between each group was performed in a pairwise manner. From the analyses, the highest number of DEGs were observed between Group 3 and Group 4, supporting the robustness of the data from asynchronous parasites indicating the expression changes between patients experienced shorter and longer days of patency.

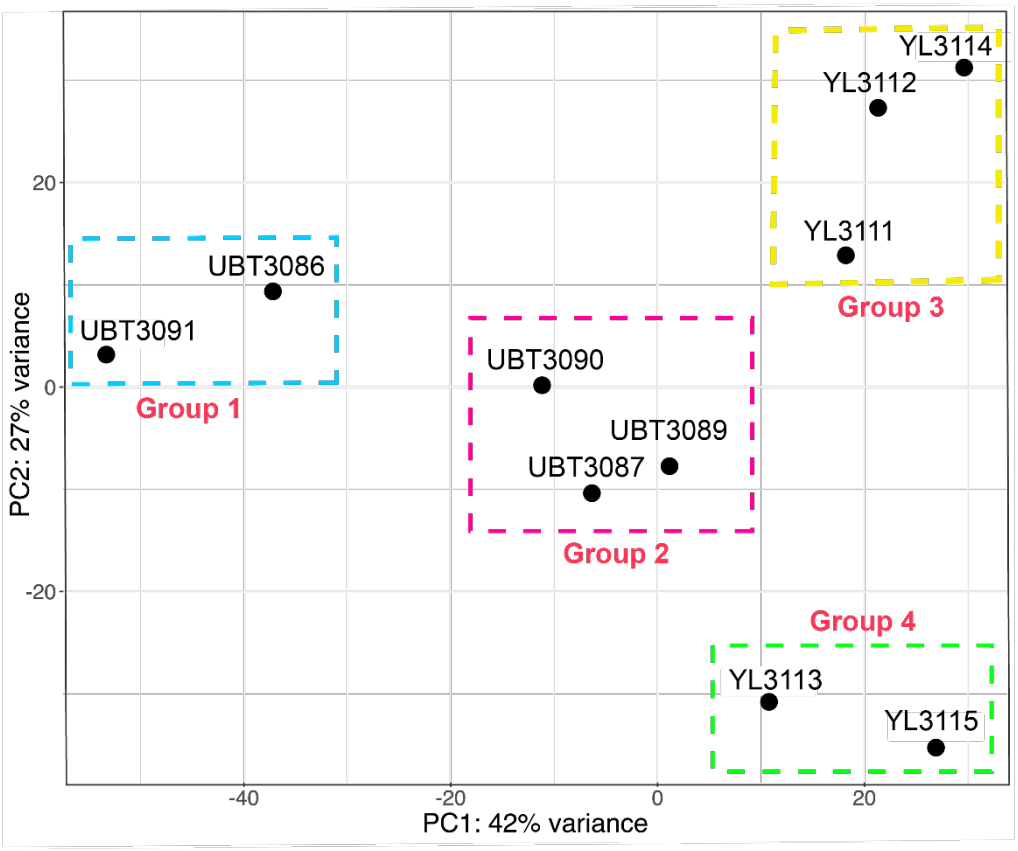


Figure 5.4 Principal component analysis (PCA) of ten patients based on genome-wide expression profile (6642 genes). Patients were divided into four groups based on the PCA. Each axis represents a principal component (PC1 and PC2) with 42% and 27% of total variance, respectively. The principal components were calculated from normalised read counts implemented in *DESeq2*. Each dot represents an individual, and the colour is assigned according to the groups (Group 1: sky blue, Group 2: pink, Group 3: yellow, and Group 4: light green).

5.3.6 PvMSP-7 expression profiles

To investigate the expression levels of 13 PvMSP-7 paralogs in ten patients, the normalised read counts from *DESeq2* were manually retrieved and plotted a heat map (Figure 5.5). From the heatmap, only group 4 patients showed a distinct PvMSP-7 expression profile, consistent with the PCA plot where group 4 patients formed a distinct cluster. The abundance of 13 PvMSP-7 paralogs was evaluated, they seemed to have varying expression patterns in ten patients. Interestingly, PvMSP-7A, -7F, and -7M were seen to express constitutively in all patients whereas, PvMSP-7H and -7I have higher expression profiles in the two patients experienced longer days of patency. A closer look at the log2 transformed expression values in ten patients, PvMSP-7A, -7F, and -7M showed higher abundance compared to other PvMSP-7 paralogs (mean expression value=7.76, 7.19, and 6.28). The mean expression values of PvMSP-7H and -7I were 4.03 and 3.58. On the other hand, other PvMSP-7 paralogs (-7B, -7C, -7D, -7E, -7G, -7J, -7K, -7L) were found to have relatively low expression levels with mean below 3.24.

To check whether PvMSP-7 paralogs are expressed significantly different between each group of patients, differential gene expression analysis was performed. PvMSP-7H and -7I was found to be significantly differentially expressed when Group 4 patients were involved in the pairwise analysis (S5.1). In addition, PvMSP-7L, -7K, and -7C were also found to be differentially expressed between Group 1 and Group 4. However, pairwise differential genes expression analysis between Group 1, Group 2, and Group 3 did not implicate PvMSP-7.

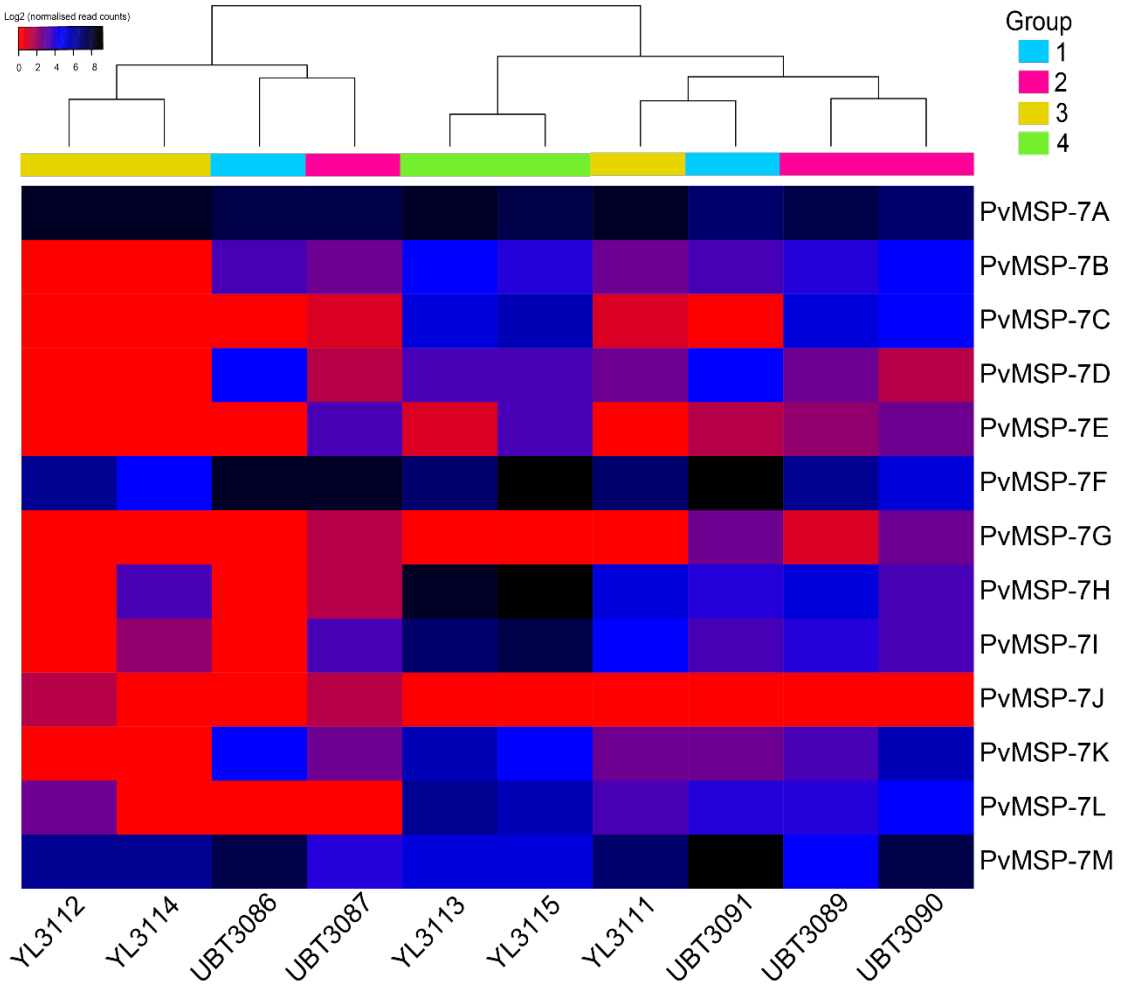


Figure 5.5 13 PvMSP-7 paralogs expression profiles in ten patients. The normalised reads from *DESeq2* were $\log_2$ transformed. The clustering on the X-axis of the dendrogram shows the similarity of PvMSP-7 expression profile between patients. The branching patterns in the dendrogram organised Group 4 patients into a group. For the heatmap, red indicates low expression level, and black indicates high expression level. Group of patients is labelled in different colours underneath the dendrogram (Group 1: sky blue, Group 2: pink, Group 3: yellow, and Group 4: light green).

5.3.7 Heat map of differentially expressed genes

Differential expression analysis was performe for the four patient groups identified by the PCA, to identify which transcripts uniquely defined these groups, and how this might relate to their MSP-7 and their developmental state. From the heat map in Figure 5.5, PvMSP-7H and -7I are highly expressed in Group 4 patients, suggesting PvMSP-7 paralogs could be specifically expressed at a particular developmental stage. Since the question relates to PvMSP-7 and only comparisons involving Group 4 showed a significant difference in PvMSP-7 expression, DEGs unique to Groups 1-3 were not evaluated in detail.

To illustrate the DEGs expression patterns between Group 3 and Group 4, Group 2 and Group 4, and Group 1 and Group 4, heatmaps were plotted in Figure 5.6 to 5.8. The heatmaps were divided into four or five sectors, grouping transcripts with similar expression profiles according to the dendrograms. Then, gene ontology enrichment analysis was performed on each sector. The genes in each sector were submitted to the PlasmoDB webserver release 35 (Aurrecoechea *et al.*, 2008) to predict biological processes, molecular functions, and metabolic pathways.

In total 1493 DEGs were identified between Group 3 and Group 4 (S5.1). Expression profiles for these genes are plotted in a heat map (Figure 5), which shows that the five samples separate into two distinct clades on the X-axis of the dendrogram (Figure 5.6). Individual transcripts arranged on the Y-axis dendrogram have been subdivided according to the cladistic structure; close inspection of the transcripts in sectors one (n=252) and two (n=456) shows that most of these transcripts are downregulated in Group 4. Gene ontology analysis was performed on each sector showed that GO terms associated with transcripts in sector one and two are enriched for functions relating to RNA binding, nucleic acid binding, ATP-dependent peptidase activity, helicase activity, tRNA processing, nitrogen compound metabolic process, and RNA metabolic process. Genes in sector four (n=192) were upregulated in Group 3 patients. Significantly enriched GO terms associated with these transcripts relate to macromolecular complex, eukaryotic translation initiation factor 3 complex, chaperonin-containing T-complex, single-organism catabolic process, small molecule metabolic process, protein folding, and translational initiation. Interestingly, genes in sector three (n=403) and five (n=190) were had higher expression values in Group 4

patients. The GO terms associated with these transcripts related to the rhoptry, cell surface, protein-DNA complex, nucleus, regulation of metabolic process, DNA binding, and protein binding. Five PvMSP-7 genes; -7K, -7I, -7H, -7C, and -7L were seen upregulated in Group 4. In summary, most of the genes expressed to a significantly greater extent in Group 4 relative to Group 3, are involved in erythrocyte invasion, which is consistent with the role of PvMSP-7 described previously.

351 DEGs were identified in the comparison of Group 2 and Group 4 (S5.1). Again, the patients of each group separate clearly a heat map (Figure 5.7). Subdivision of the DEGs comprising the heatmap into five sectors. Transcripts within sectors one (*n*=99) and two (*n*=49) were associated for GO terms for Maurer's cleft, origin recognition complex, host cell cytoplasm part, merozoite dense granule, and phosphopyruvate hydratase complex, but expression profiles in these sectors are not distinct in the two patient groups. By contrast, genes in sectors three (*n*=47) and four (*n*=142) are downregulated in Group 4 patients. These are associated with GO terms for purine metabolism pathway and host cell surface binding. No significant enrichment terms were detected in sector 5.

251 DEGs were identified in the comparison of Groups 1 and 4 (S5.1). Figure 5.8 shows that the two patient groups had distinct expression profiles. In sector one of the heat map, most of the genes have lower expression in sector four (*n*=125). Functional terms associated with these transcripts relate to the cell surface, Maurer's cleft, host cell surface binding, and uridylyltransferase activity. The precise genes that are responsible for this cell surface association are reticulocyte binding surface protein (PvP01_00004240), tryptophan-rich protein (PvP01_0504200), small heat shock protein HSP20 (PvP01_0518800). In sector two (*n*=71), expression levels are higher in Group 1 patients. The GO terms associated with these transcripts are crystalloid, cell surface, and host cell cytoplasm part. Sector three (*n*=22) and sector four (*n*=33) showed higher expression in group four patients. The predicted functional terms associated here are related to cell component, Maurer's cleft, rhoptry, cell surface, and proteolysis.

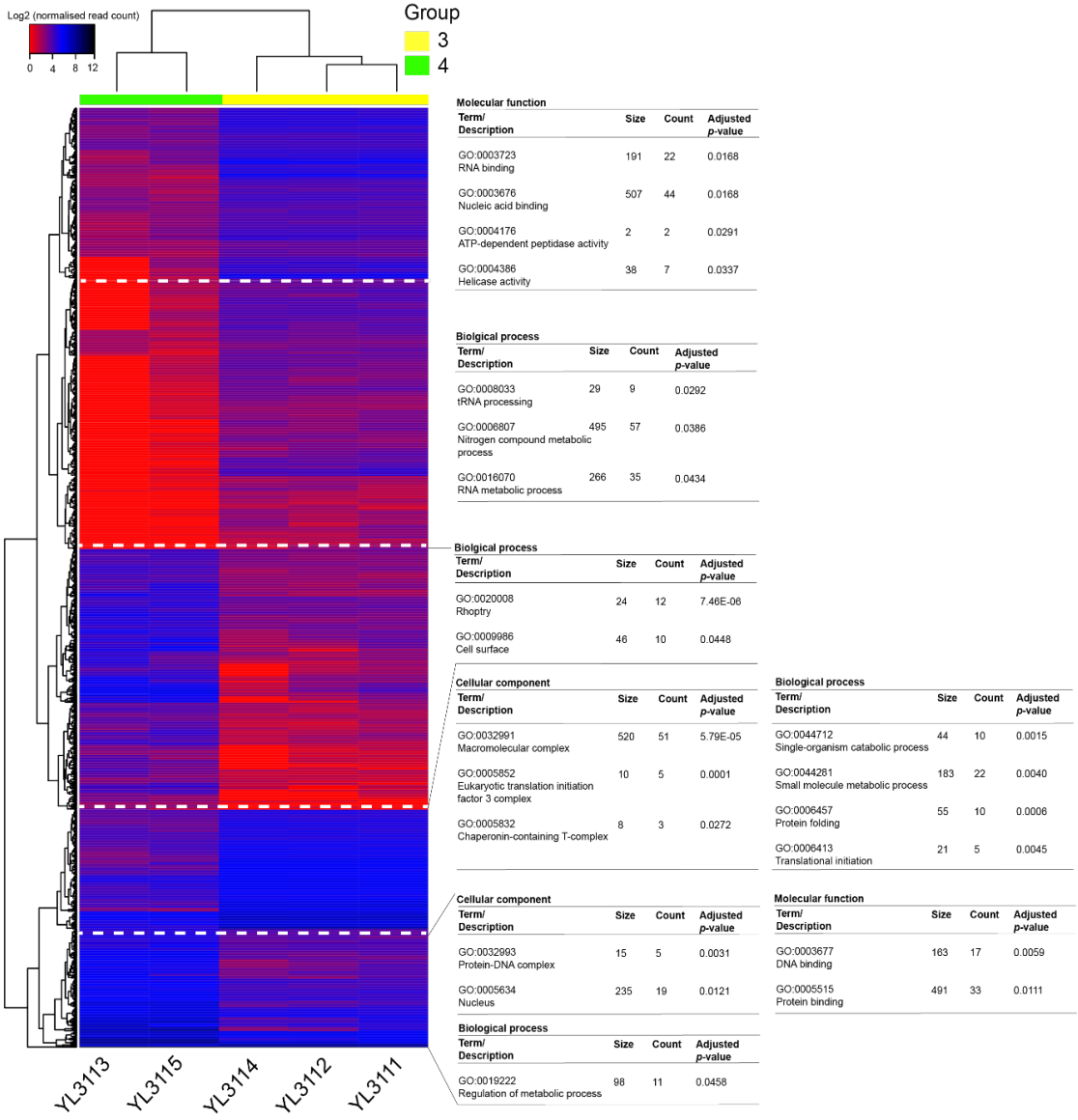


Figure 5.6 Genes differentially expressed (DEGs) between Group 3 and Group 4 (*n* = 1493). The heatmap was divided into five sectors. Significance GO terms and KEGG pathways in each sector are shown on the right. GO terms and KEGG pathways achieved adjusted *p*-value<0.05 were deemed significance from the *PlasmoDB* database. Data was generated from the $\log_2$ transformation of normalised reads in *DESeq2*. The colour scale represents the expression level of DEGs such that, red refers to lower expression while black refers to a higher expression. Hierarchical clusters of patients are represented by a vertical dendrogram on the X-axis. Group of patients is labelled in different colours underneath the dendrogram (Group 3: yellow and Group 4: light green).

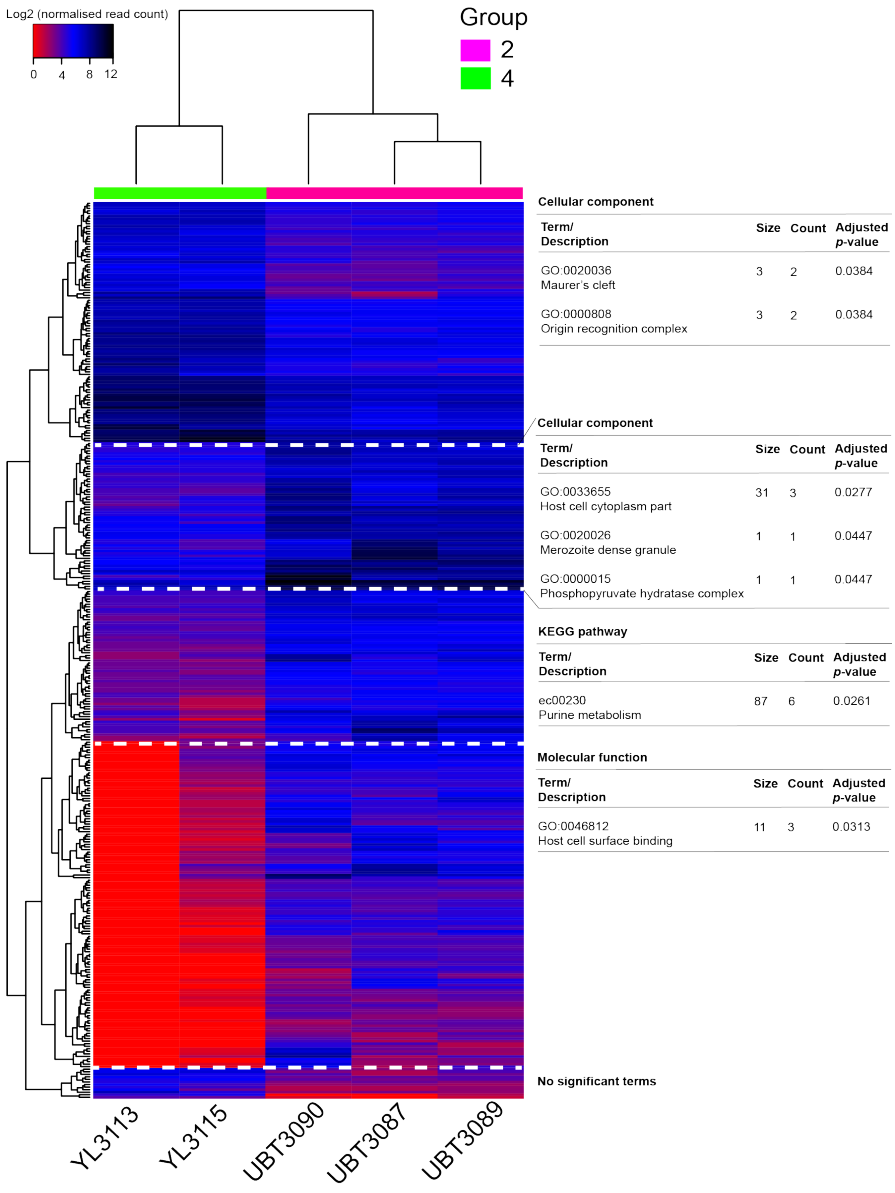


Figure 5.7 Genes differentially expressed between Group 2 and Group 4 (*n* = 351). The heatmap was divided into five sectors. Significance GO terms and KEGG pathways in each sector are shown on the right. GO terms and KEGG pathways achieved adjusted *p*-value<0.05 were deemed significance from the *PlasmoDB* database. Data was generated from the $\log_2$ transformation of normalised raw reads in *DESeq2*. The colour scale represents the expression level of DEGs such that, red refers to lower expression while black refers to a higher expression. Hierarchical clusters of patients are represented by a vertical dendrogram on the X-axis. Group of patients is labelled in different colours underneath the dendrogram (Group 2: pink and Group 4: light green).

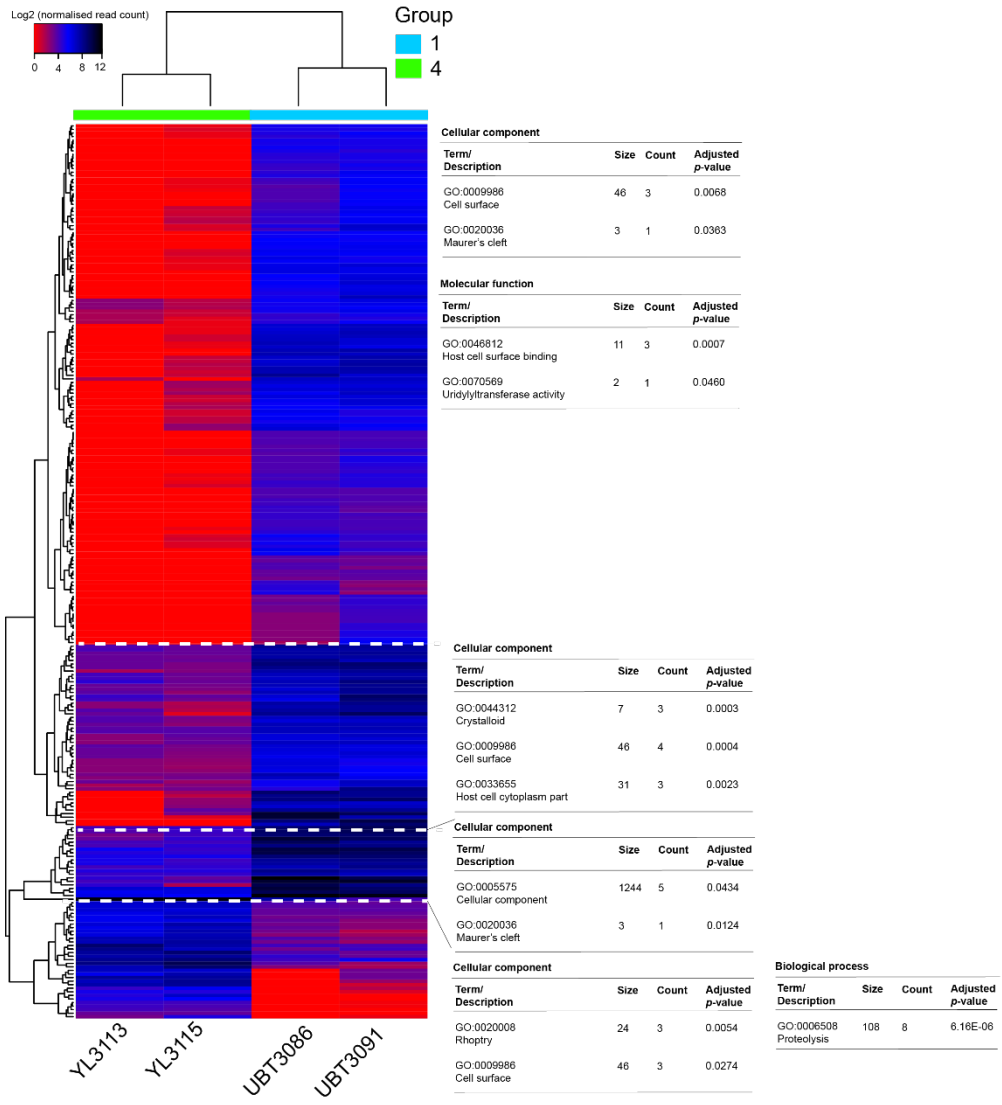


Figure 5.8 Genes differentially expressed between Group 1 and Group 4 (*n* = 251). The heatmap was divided into four sectors. Significance GO terms and KEGG pathways in each sector are shown on the right. GO terms and KEGG pathways achieved adjusted *p*-value<0.05 were deemed significance from the *PlasmoDB* database. Data was generated from the $\log_2$ transformation of normalised raw reads in *DESeq2*. The colour scale represents the expression level of DEGs such that, red refers to lower expression while black refers to a higher expression. Hierarchical clusters of patients are represented by a vertical dendrogram on the X-axis. Group of patients is labelled in different colours underneath the dendrogram (Group 1: sky blue and Group 4: light green).

5.3.8 Co-expression analysis

From the differential gene expression analysis, at least two PvMSP-7 paralogs (7H and 7I) were found significantly upregulated in Group 4 patients who experienced longer patency. This suggests that these PvMSP-7 paralogs could be developmentally regulated. Conversely, constitutive expression of three PvMSP-7 paralogs (-7A, -7F, and -7M) also indicates that functional distinctions exist between PvMSP-7 paralogs. To explore this further, and to identify other, stage-specific markers that were differentially expressed in the same manner as PvMSP-7H and PvMSP-7I, co-expression analysis was applied using the *coseq* package to cluster the expression differences observed between Group 3 and Group 4, Group 2 and Group 4, and Group 1 and Group 4 respectively. The analysis robustly identified DEG cohorts that were significantly co-expressed, that displayed the same transcriptional profile across the patient samples.

Differential expression analysis between Group 3 and Group 4 identified 1493 DEGs. These genes were used to construct the co-expression clusters (Figure 5.9). The analysis formed 14 clusters, of which cluster 1 contained PvMSP-7K, -7I, 7H, -7C, which were upregulated in Group 4 patients (Figure 5.9a, Supplementary Figure 1) while, PvMSP-7L increases found in cluster 13 (Figure 5.9b). The log-fold change of these five PvMSP-7 genes in Group 4 relative to Group 3 patients ranging from 3.64 - 6.45 (-7K; log-fold change = 4.54, -7I; log-fold change = 4.21, -7H; log-fold change = 4.60, -7C; log-fold change = 6.48, -7L; log-fold change = 3.64). Interestingly, they are co-expressed with several stage-specific markers with known function in erythrocyte invasion, such as early transcribed membrane protein (ETRAMP, PvP01_0618300), schizont egress antigen-1 (SEA1, PvP01_0607000), rhoptry neck protein 4 (RON4, PvP01_0916600) plasmepsin V (PMV, PvP01_1231100), merozoite organizing protein (MOP, PvP01_0715400), and rhoptry neck protein 5 (RON5, PvP01_0517600). Other stage markers were identified in other clusters including, gamete antigen 27/25 (PvP01_0422700), sporozoite and liver stage tryptophan-rich protein (TryThrA, PvP01_0532600), and liver-specific protein 3 (PvP01_0405000). Two high expression genes that co-expressed with these stage-specific markers are shown in the Figure 5.9c, 5.9d, and 5.9e, such as tryptophan-rich protein (TRAG36, PvP01_0119200), *Plasmodium* interspersed repeat (PIR, PvP01_0816000), heat shock protein 70 (HSP70,

PvP01_0515400), *Plasmodium* exported protein (EXP, PvP01_0300700), glyceraldehyde-3-phosphate dehydrogenase (GADPH, PvP01_1244000), and erythrocyte membrane-associated antigen (EMAA, PvP01_0103700). In Figure 5.9c, 12 tryptophan-rich proteins and 10 PIR proteins were co-expressed in the clusters. Pearson's correlation coefficient (*r*) was estimated for each stage-specific marker in relation to PvMSP-7H, demonstrating a positive correlation ($r>0.90$) with the expression of schizont-stage genes and a negative correlation ($r>-0.85$) with sporozoite, liver, and gametocyte stage markers.

Further comparison between Group 2 and Group 4 patients revealed 351 DEGs. Co-expression formed seven clusters using these DEGs (Supplementary Figure 2). Expression patterns for PvMSP-7 genes were similar to those observed in the previous analysis. PvMSP-7H and -7I were upregulated in Group 4 patients with log-fold change, 4.68 and 3.48, respectively (Figure 5.10a). In addition, SEA1 was also increased its expression level (log-fold change 2.12) within Group 4. Two stage-specific markers were detected in the analysis; liver-specific protein 3 (Figure 5.10c), and gamete antigen 27/25 (Figure 5.10d). The expression level of these two markers was reduced in Group 4 with log-fold change -3.31 and -4.64, respectively. Furthermore, five genes that are either schizont stage-specific markers or genes with known function in erythrocyte invasion, are positively correlated with PvMSP-7H ($r>0.75$). Parasite-infected erythrocyte surface protein (PIESP1, PvP01_0829800) was co-expressed with SEA1 in Figure 5.10b. As in the comparison with Group 3, PvMSP-7H is negatively correlated (*r* between -0.44 to -0.83) with gametocyte stage and liver stage markers. 19 *Plasmodium* exported proteins and 11 tryptophan-rich proteins were found to co-express with these two gametocyte and liver stage-specific markers.

The co-expression analysis between Group 1 and Group 4 patients was based on 251 DEGs. Six co-expression clusters were formed using *coseq* package (Supplementary Figure 3). PvMSP-7H and -7I genes were differentially expressed but they were placed into different clusters (Figure 5.11a and 5.11b). As with the previous two analyses, these two genes were upregulated in group four patients (7H; log-fold change = 5.31, 7C; log-fold change = 6.33). Two-stage markers; gamete release protein (PvP01_0115300) in Figure 5.11c and gamete antigen 27/25 in Figure 5.11d were downregulated in group four patients. The log-fold change was -6.25 and -7.22,

respectively. Two PvMSP-7 paralogs were also found to positively co-expressed (r>0.60) with PIESP1, PMV, rhoptry neck protein 3 (RON3, PvP01_1469200), serine-repeat antigen-1 (SERA, PvP01_0417100), subtilisin-like protease 3 (SUB3, PvP01_1026800), and high molecular weight rhoptry protein 3 (RhopH3, PvP01_0703800). On the other hand, these two PvMSP-7 paralogs are negatively correlated with two gametocyte stage-specific markers (r>-0.80). In line with the previous two analyses, these two gametocyte stage markers are co-expressed with 64 *Plasmodium* exported proteins and 12 tryptophan-rich proteins.

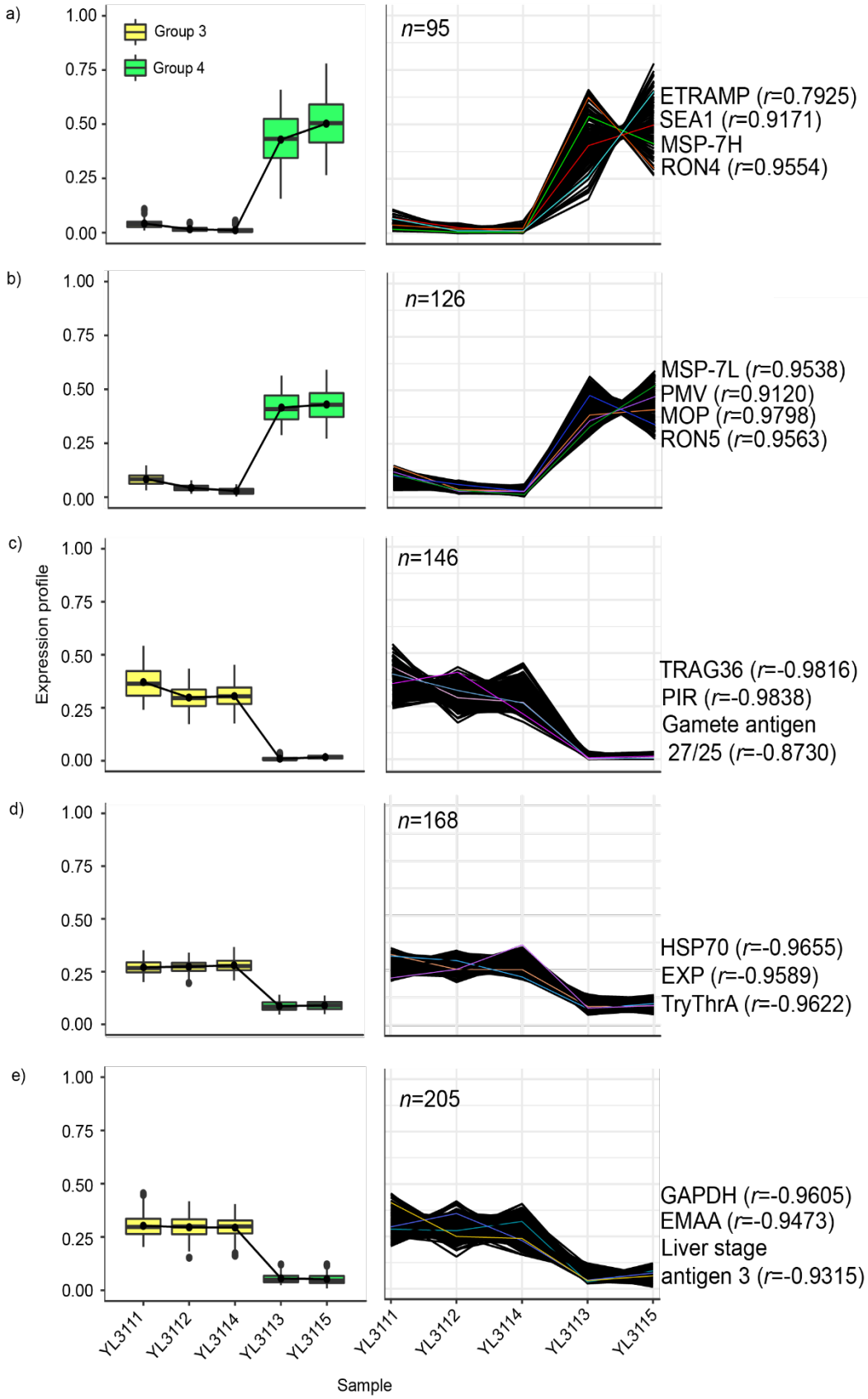

a)
b)
c)
d)
e)
Group 3
Group 4
Expression profile
Sample
YL3111
YL3112
YL3114
YL3113
YL3115
n=95
ETRAMP (r=0.7925)
SEA1 (r=0.9171)
MSP-7H
RON4 (r=0.9554)
n=126
MSP-7L (r=0.9538)
PMV (r=0.9120)
MOP (r=0.9798)
RON5 (r=0.9563)
n=146
TRAG36 (r=-0.9816)
PIR (r=-0.9838)
Gamete antigen
27/25 (r=-0.8730)
n=168
HSP70 (r=-0.9655)
EXP (r=-0.9589)
TryThrA (r=-0.9622)
n=205
GAPDH (r=-0.9605)
EMAA (r=-0.9473)
Liver stage
antigen 3 (r=-0.9315)

Figure 5.9 Co-expression analysis of five patients between Group 3 and Group 4. 1493 DEGs formed 14 clusters in co-expression analysis. The normalised reads derived from the *DESeq2* package were used to construct co-expression analysis implemented in *coseq*. Five MSP-7 genes were found co-expressed in two clusters, Figure 5.9a and 5.9b. In the Figure 5.9a, PvMSP-7K, -7I, -7H and -7C are co-expressed with 91 genes including the ETRAMP, SEA1, and RON4 whilst, PvMSP-7L is co-expressed with 125 genes including PMV, MOP, and RON5. Several stage-specific markers were also identified in Figure 5.9c, 5.9d, and 5.9e. Gamete antigen 27/25 is upregulated within condition three in Figure 5.9c. TryThrA is increased in expression within condition three with other 167 genes in Figure 5.9d. In Figure 5.9e, liver stage antigen 3 is upregulated in condition three. Two genes with high expression levels in group three patients were showed in Figure 5.9c, 5.9d, and 5.9e such as TRAG36, PIR, HSP70, EXP, GADPH, and EMAA. The boxplots on the left represent the patients, the colour depicts each condition, and the connected black lines on the boxplots indicate the mean expression of the genes. The line graph on the right shows the expression pattern of each gene in the individual sample, the lines correspond to the genes. The coloured lines on the line graph correspond to the specific genes labelled on the right. Correlations between PvMSP-7H and the selected genes showed on the right were assessed based on Pearson's correlation coefficient (r). The number of genes (n) in each cluster was shown on the top left of each line graph.

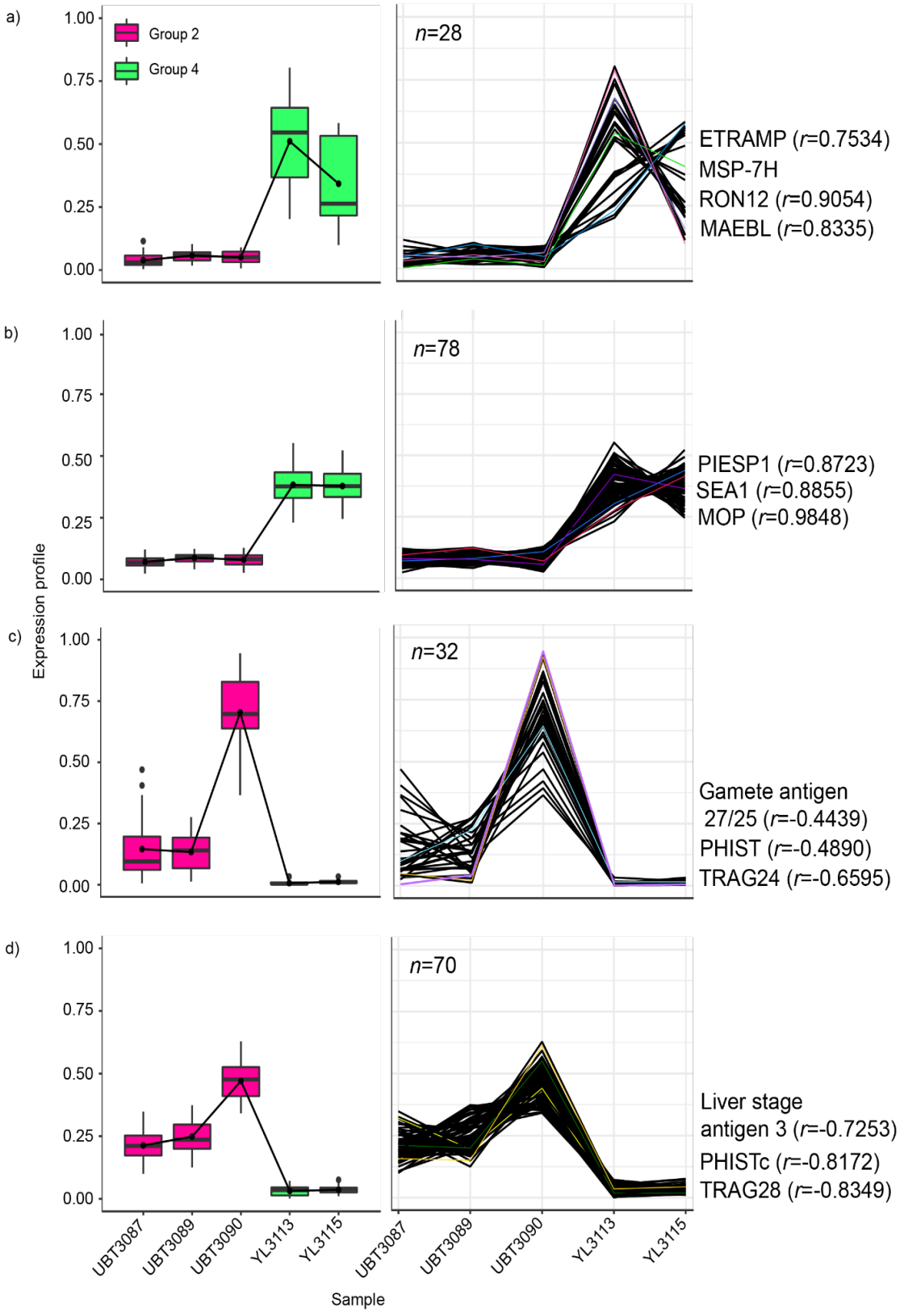

a)
b)
c)
d)
Group 2
Group 4
n=28
n=78
n=32
n=70
1.00
0.75
0.50
0.25
0.00
Expression profile
ETRAMP (r=0.7534)
MSP-7H
RON12 (r=0.9054)
MAEBL (r=0.8335)
PIESP1 (r=0.8723)
SEA1 (r=0.8855)
MOP (r=0.9848)
Gamete antigen 27/25 (r=-0.4439)
PHIST (r=-0.4890)
TRAG24 (r=-0.6595)
Liver stage antigen 3 (r=-0.7253)
PHISTc (r=-0.8172)
TRAG28 (r=-0.8349)
UBT3087
UBT3089
UBT3090
YL3113
YL3115
Sample

Figure 5.10 Co-expression analysis of five patients between Group 2 and Group 4. The normalised reads derived from the *DESeq2* package were used to construct co-expression analysis implemented in *coseq*. In Figure 5.10a, two MSP-7 genes; -7H and -7I are co-expressed with 26 genes including ETRAMP, RON12, and MAEBL. In Figure 5.10b, SEA1 is co-expressed with 77 genes such as PIESP1 and MOP. In Figure 5.10c, gamete antigen 27/25 clustered with 31 genes including PHIST and TRAG24, PvP01_1470100. In figure 5.10d, liver stage antigen 3 is clustered with 69 genes such as PHIST and TRAG28. Each boxplot depicts the individual patients, the colour represents each condition and the connected black lines on the boxplots represent the mean expression of each individual. The line graph on the right shows the expression pattern of each gene in an individual sample, the lines correspond to the genes. The coloured lines on the line graph correspond to the specific genes labelled on the right. Correlations between MSP-7H and the selected genes showed on the right were assessed based on Pearson's correlation coefficient (r). The number of genes (n) in each cluster was shown on the top left of each line graph.

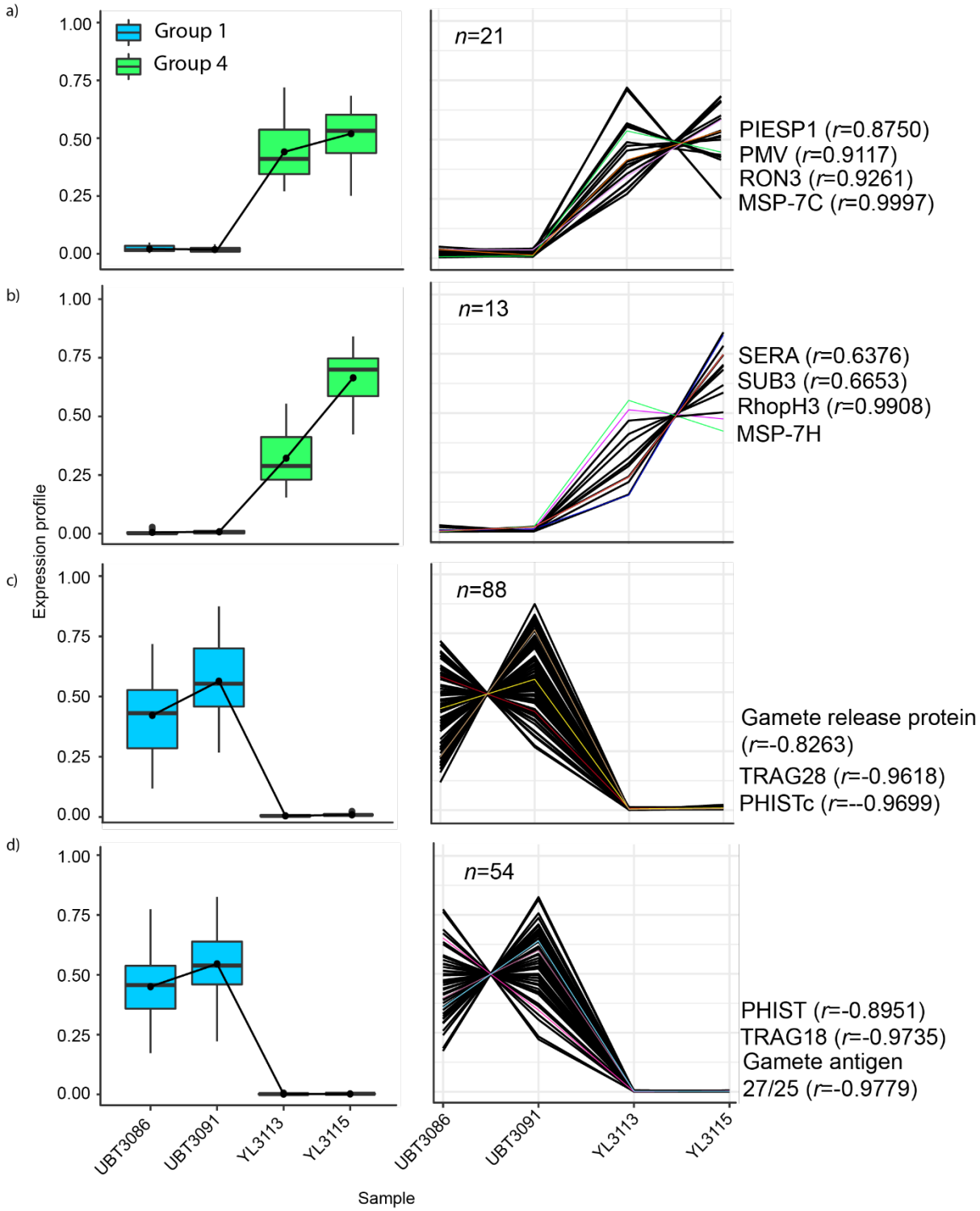

a)
b)
c)
d)
Group 1
Group 4
Expression profile
1.00
0.75
0.50
0.25
0.00
n=21
n=13
n=88
n=54
PIESP1 (r=0.8750)
PMV (r=0.9117)
RON3 (r=0.9261)
MSP-7C (r=0.9997)
SERA (r=0.6376)
SUB3 (r=0.6653)
RhopH3 (r=0.9908)
MSP-7H
Gamete release protein (r=-0.8263)
TRAG28 (r=-0.9618)
PHISTc (r=--0.9699)
PHIST (r=-0.8951)
TRAG18 (r=-0.9735)
Gamete antigen 27/25 (r=-0.9779)
UBT3086
UBT3091
YL3113
YL3115
Sample

Figure 5.11 Co-expression analysis of four patients between Group 1 and Group 4. The normalised reads derived from the *DESeq2* package were used to construct co-expression analysis implemented in *coseq*. Two PvMSP-7 genes; -7C and -7H were found in this analysis as shown in Figure 5.11a and 5.11b, respectively. PvMSP-7C is co-expressed with 20 genes such as PIESP1, PMV, and RON3 whilst PvMSP-7H is co-expressed with 12 other genes including SERA, SUB3, and RhopH3. Two gametocyte markers; gamete release protein and gamete antigen 27/25 were each found in the clusters as shown in Figure 5.11c and 5.11d. The gamete release protein is co-expressed with TRAG28 and PHISTc whilst, gamete antigen 27/25 is co-expressed with PHIST and tryptophan-rich protein 18 (TRAG18). The boxplots on the left represent the patients, the colour depicts each condition, and the connected black lines on the boxplots indicate the mean expression of the genes. The line graph on the right shows the expression pattern of each gene in the individual patient the lines correspond to the genes. The coloured lines on the line graph correspond to the specific genes labelled on the right. Correlations between MSP-7H and the selected genes showed on the right were assessed based on Pearson's correlation coefficient (r). The number of genes (n) in each cluster was shown on the top left of each line graph.

5.3.9 SNP discovery

PCA plot was used to investigate if the genetic variants influencing the gene expression of ten RNA samples. The PCA constructed using 42,988 genome-wide SNPs revealed population structure of ten RNA samples. In Figure 5.12, the plot of the two highest principal components highlight the distinct clusters of samples from Yala and Ubon Ratchathani province. Five samples from Yala province clustered closely with one another. However, samples from Ubon Ratchathani province did not cluster tightly together. Three individuals UBT3090, UBT3089, and UBT3091 are located away from the main cluster.

The phylogeny analysis was performed to validate the result derived from PCA. The maximum likelihood tree was generated in Randomized Axelerated Maximum Likelihood, version 8 (RAxML) (Stamatakis, 2014) using all 42,988 SNPs. Bootstrap was set to 100-fold to increase the reliability. The best-fit model of nucleotide sequence was determined using jModelTest, version 2.0 (Posada, 2008). A sequence alignment file contained all SNPs in ten samples were passed into jModelTest, GTR substitution model with gamma rate variation was identified as the best selection results. The neighbour-joining tree was constructed in MEGA 7.0 (Kumar *et al.*, 2016) using 1000 bootstrap pseudoreplicates. Maximum composite likelihood method based on the Tamura 3-parameter model was used to construct the neighbour joining tree. The Maximum likelihood and neighbour-joining tree in Figure 5.13 and Figure 5.14, respectively corroborated finding from PCA. Two distinct clusters were observed from the trees with high bootstrap support values. Isolates from Yala province were separated from Ubon Ratchathani province. However, one sample from Yala province (YL3113) did not cluster with other samples from the same location in the maximum likelihood tree (Figure 5.13).

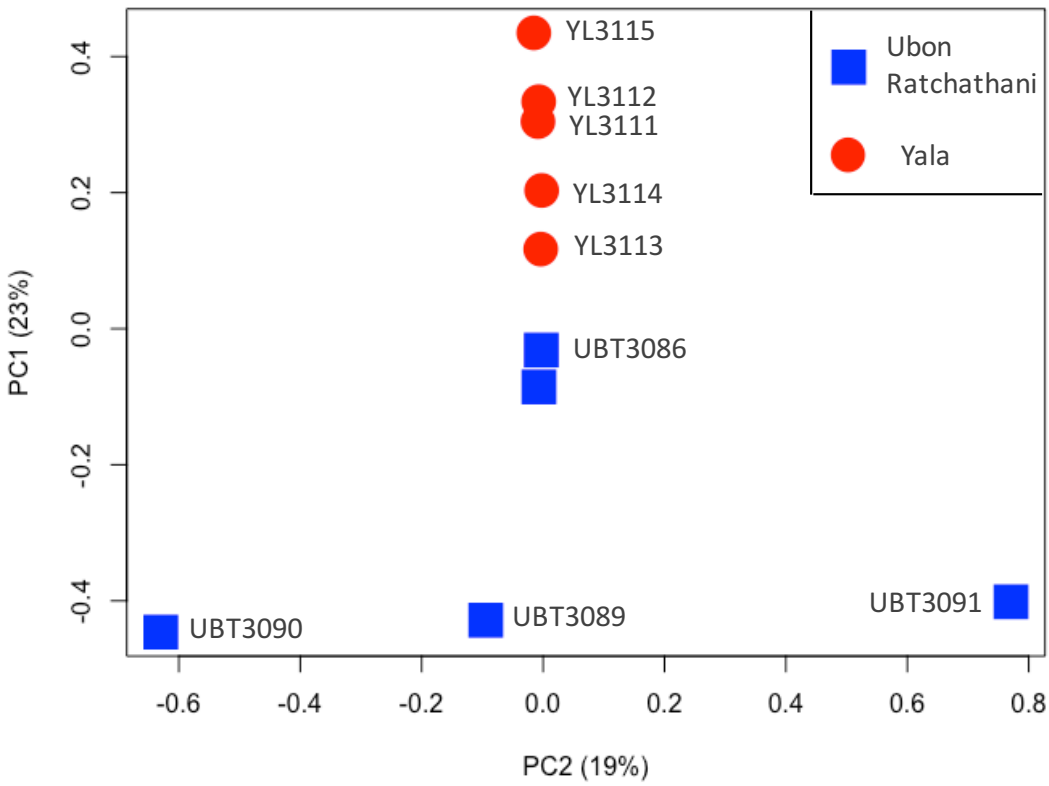


Figure 5.12 Principal component analysis of ten clinical isolates in Thailand. The analysis was based on the 42,988 genome-wide SNPs. The PCA plot shows geographical segregation of 10 clinical isolates according to their origins. Isolates from Ubon Ratchathani are loosely clustered. The plot was generated using SNPRelate in R package (Zheng *et al*., 2012). The colour and shape of symbol represents the respective malaria endemic area, blue colour and square shape: Ubon Ratchathani province, and red colour and circle shape: Yala province.

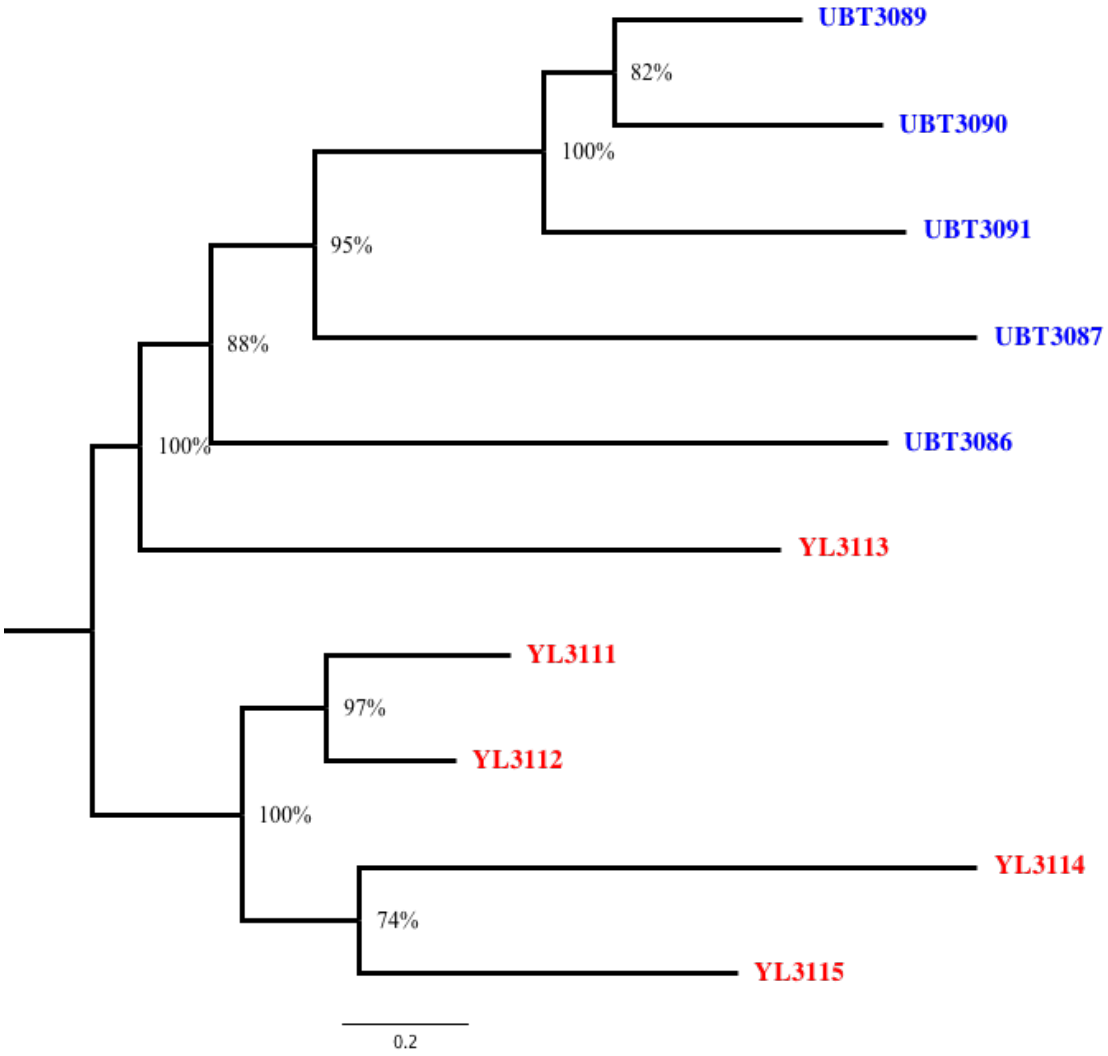


Figure 5.13 Maximum likelihood tree (midpoint rooted) of ten clinical isolates from Thailand. The tree was estimated by RAxML using 42,988 biallelic SNPs. GTR substitution model was used. Ten samples are separated into two major groups based on the geographical location except YL3113. Boostrap values at the nodes were generated from 100 replicates. Labels in blue are isolates from Ubon Ratchathani Province and labels in red are isolates from Yala province.

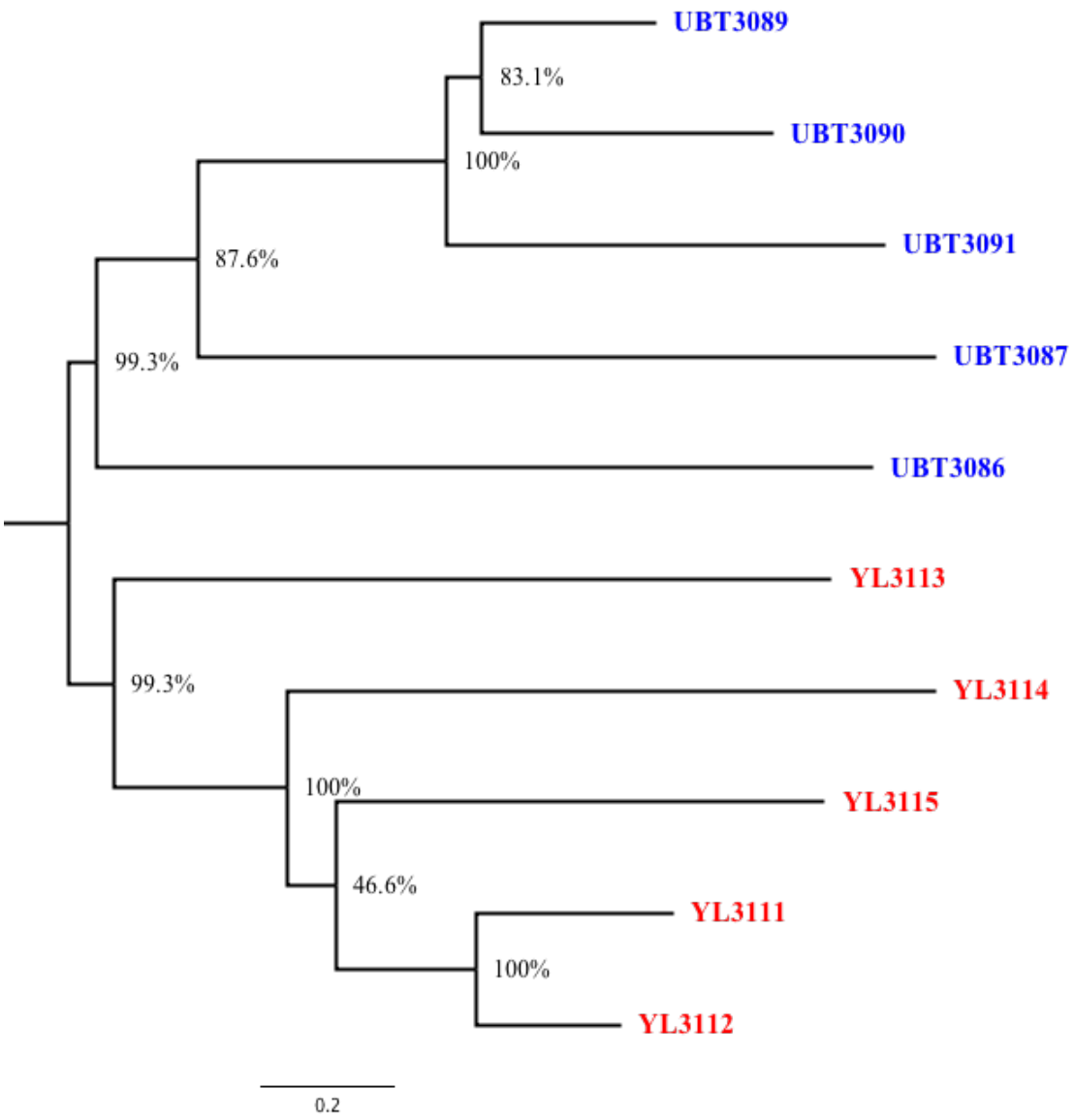


Figure 5.14 Neighbour-joining tree (midpoint rooted) of ten clinical isolates from Thailand. The tree was estimated by MEGA7 using 42,988 biallelic SNPs. Maximum composite likelihood method based on the Tamura 3-parameter model was used. Ten samples are separated into two distinct groups based on the geographical location. Boostrap values at the nodes were generated from 1000 replicates. Labels in blue are isolates from Ubon Ratchathani Province and labels in red are isolates from Yala province.

5.4 Discussion

In this chapter, *P. vivax* transcriptomes were generated for 10 clinical bloodstream infections from naturally infected patients. These infections varied in the combination of parasite developmental stages that composed the parasite population, which is typical of clinical samples. A recent study of similar material containing asynchronous parasite populations suggested that parasite expression patterns remain consistent, regardless of the differences in parasite developmental status (Kim *et al.*, 2017). My analysis of PvMSP-7expression patterns of in these clinical isolates reach a different conclusion, showing that developmental differences are clearly discernible among patients, apparently coinciding with the patent period, specifically, that there are substantial differences between the expression patterns of the PvMSP-7 paralogs through the IDC. Three PvMSP-7 members; -7A, -7F, and -7M were constitutively expressed in all patients, while PvMSP-7H and -7I increased in expression level in patients that had experienced longer patency. Co-expression analysis showed that PvMSP-7H and –7I co-express with a schizont stage marker but are inversely associated with sporozoite stage, liver stage, and gametocyte stage markers. Therefore, these findings suggest that, while all MSP-7 paralogs function during the bloodstream infection, some paralogs may be developmentally regulated within the context of the IDC.

It is sensible, therefore, to inspect transcriptomes from synchronized parasite cultures for consistent differences in paralog expression profiles, although this was not reported in the original descriptions of *P. vivax* transcriptomes. The results were compared with previous synchronous parasite cultures in *P. vivax* (Bozdech *et al.*, 2008)*, P. falciparum* (López-Barragán *et al.*, 2011)*,* and *P. berghei* (Otto *et al.*, 2014). In Figure 5.12, the expression profile of MSP-7 in three *Plasmodium* species is illustrated using a heat map. The first *P. vivax* transcriptome, described in 2008, used microarray analysis to describe the expression abundance of the genes to capture transcriptional changes over 48 hours IDC of three distinct isolates from the early ring stage to schizont stage (Bozdech *et al.*, 2008). A subsequent study by López-Barragán and colleagues applied RNA-seq to cultured parasites from ring stage to ookinete stage, profiling the gene expression of sexual and asexual stages in *P. falciparum* (López-Barragán *et al.*, 2011). In 2014, an extensive transcriptome of rodent malaria was published (Otto *et al.*, 2014), again based on RNA-seq but applied to rodent malaria.

These studies focused on the global transcriptional profiles in the *Plasmodium* species but, did not highlight specifically on the expression patterns of MSP-7 genes. The MSP-7 expression levels were retrieved from these data and made comparisons between the transcriptional profiles (Figure 5.15). Note that the size of the multigene family differs between species: 13 genes in *P. vivax*, nine genes in *P. falciparum*, and four genes in *P. berghei*. Of the 13 genes in *P. vivax*, eleven paralogs had increased expression from the early schizont stage to late schizont stage. PvMSP-7A, -7F and -7M were expressed constitutively over the whole IDC (Bozdech *et al.*, 2008). In contrast, PvMSP-7G and -7J were minimally expressed. In *P. falciparum*, all paralogs were upregulated within late trophozoite to schizont transition (López-Barragán *et al.*, 2011). Two paralogs, PfMSP-7A and -7I maintained their expression through the IDC. In addition, three paralogs in *P. berghei* were seen to have their expression peaked at schizont stage, except MSP-7D which increased its transcription level at gametocyte stage (Otto *et al.*, 2014). As in the other species, the expression of a subset of paralogs (i.e. PbMSP-7A, -7C, and -7D) was maintained throughout the bloodstream life cycle, from the ring stage to ookinete.

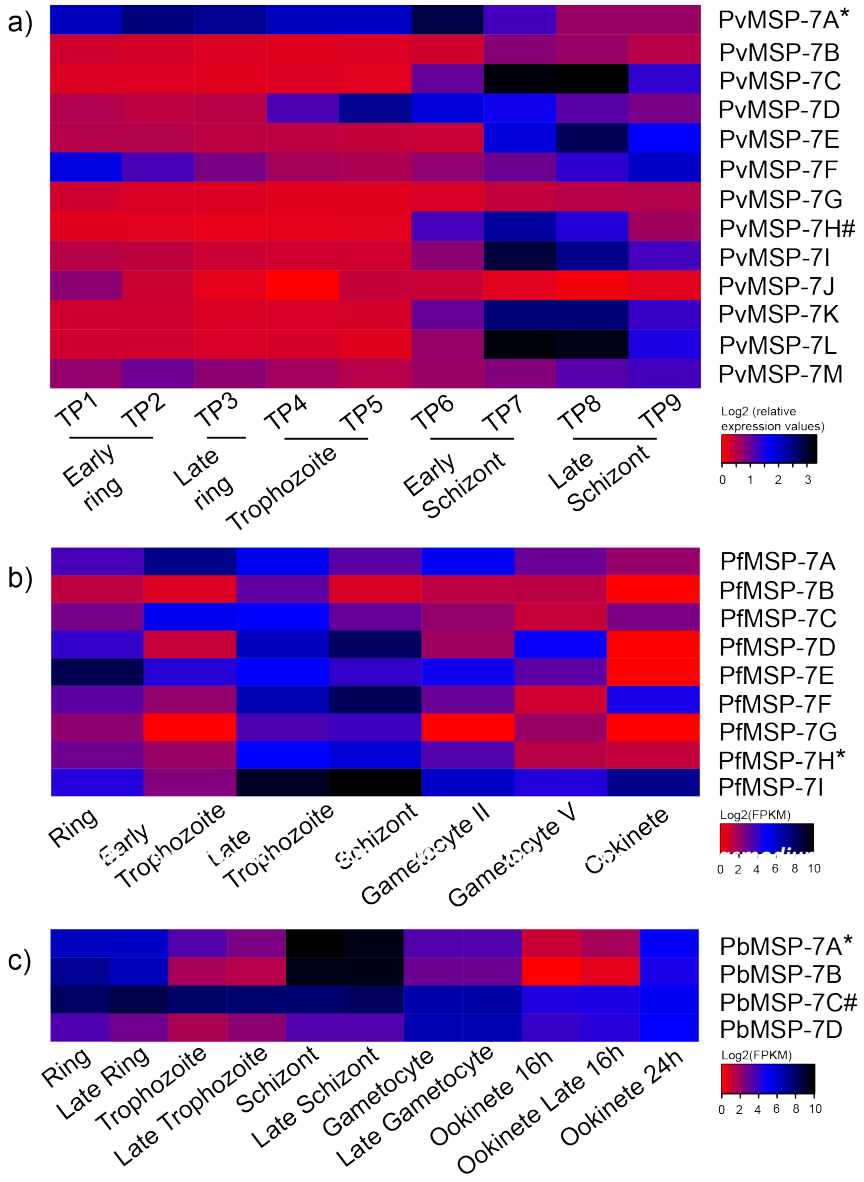


Figure 5.15 Intra-erythrocytic cycle (IDC) of MSP-7 expression profiles in *Plasmodium* spp. a) 13 MSP-7 expression profiles in *P. vivax*. The transcriptional profiles were processed using microarray (Bozdech *et al*., 2008), **b)** Nine MSP-7 expression profiles in *P. falciparum* across seven time-points in the life cycle (Lopez-Barragan *et al.* 2011), and **c)** four MSP-7 expression levels in *P. berghei* across five life stages (Otto *et al.*, 2014). All values are $\log_2$ transformed. The colour scale represents the expression pattern of each gene from red to black. Symbols (* and #) indicate the orthologous genes between three species.

Having observed consistent patterns across the three species, with some paralogs expressed throughout the IDC and others restricted to schizont or late trophozoite stages, the hypothesis was tested relating to that these might be stable phenotypes, maintained across *Plasmodium* species, by asking if phenotypically similar genes in different species were orthologs. By consulting the published gene phylogeny (Castillo *et al.*, 2017; Garzón-Ospina *et al.*, 2016), it is clear that genes orthologous with MSP-7A in *P. vivax*, (i.e. MSP-7H in *P. falciparum*, and MSP-7A in *P. berghei*) do not show similar expression patterns across the IDC. MSP-7A in *P. vivax* is expressed through the early ring stage to early schizont stage and peaks in abundance at early schizont stage, whereas in *P. falciparum* MSP-7H is expressed from late trophozoite to gametocyte II and silent after gametocyte stage V. In *P. berghei*, MSP-7A expression is maintained throughout the IDC, peaking in the schizont but continuing at a low level in the ookinete at 16h. Similarly, the ortholog of PvMSP-7K (MSP-7C in *P. berghei*) does not have a similar expression profile; MSP-7K in *P. vivax* was abundant exclusively during early schizont stage to late schizont stage, while MSP-7C in *P. berghei* was highly expressed throughout the IDC. This evidence suggests that the transcriptional profiles of individual MSP-7 gene lineages are not conserved between *Plasmodium* species. Rather, the profile of individual paralogs is flexible over evolutionary time, indeed, differences in gene number show that paralogs may be gained and lost easily. Thus, while differences in expression profile among MSP-7 paralogs is a conserved feature of *Plasmodium* life cycles, it is not derived from conservation of specific gene lineages, which perhaps points to common physiological demands across species, which have been met by diverse MSP-7 lineages through time.

Closer looks into the expression patterns were seen to have elevated expression levels. Specific up-regulation of PvMSP-7H and -7I was observed in two patients (Group 4) who experienced longer patency. Co-expression analysis revealed that these two paralogs are co-expressed with a schizont stage marker, while negatively correlated with liver-stage and gametocyte-stage markers. This finding is consistent with the *in vitro* culture of *P. vivax*, where PvMSP-7H and -7I were upregulated during the schizont stage. Invasion-related genes in *P. falciparum* have been reported to be abundantly transcribed during schizont stage. The association of PvMSP-7H and -7I with known host cell invasion proteins such as the early transcribed membrane protein (MacKellar *et al.*, 2011), schizont egress antigen-1 (Raj *et al.*, 2014), rhoptry neck

protein 4 (Lebrun *et al.*, 2005), plasmepsin V (Sedwick, 2014), merozoite organizing protein (Absalon *et al.*, 2016), and rhoptry neck protein 5 (Curtidor *et al.*, 2014), suggests that these two paralogs may play a specific role in erythrocyte invasion.

MSP-7 paralogs are thought to be involved in host cell invasion (Beeson *et al.*, 2016; Cowman and Crabb, 2006; Garzón-Ospina *et al.*, 2010; Garzón-Ospina *et al.*, 2016; Gomez *et al.*, 2011; Kadekoppala *et al.*, 2008; Kauth *et al.*, 2006; Spaccapelo *et al.*, 2011; Tewari *et al.*, 2005), based mainly on the surge in its expression during the schizont stage. Previously, MSP-7 paralogs have been treated as functionally identical, or redundant at least. Garzón-Ospina and colleagues (2016) suggested that MSP-7 paralogs were functionally divergent, based on differences in their evolutionary rates. Population studies also indicate heterogeneous allelic diversity among MSP-7 paralogs (Garzón-Ospina *et al.*, 2014; Garzón-Ospina *et al.*, 2012; Garzón-Ospina *et al.*, 2011). My observation that three PvMSP-7 paralogs, -7A, -7F, and -7M, are most highly expressed in all patients, consistent with constitutive expression throughout the IDC, is the first evidence for stable phenotypic variation among within the MSP-7 family. Coupled with the fact that three paralogs have higher sequence conservation compared to other members (Castillo *et al.*, 2017), suggestive of stronger purifying selection, these lines of evidence indicating that these three paralogs are not simply merozoite proteins, but have additional functions in addition to their role in invasion while expressed on the merozoite surface. Resolving these functions, whether or not they involve host-parasite interaction or recognition of host factors, such as P-selectin (Perrin *et al.*, 2015) will require further experimental work. Recently, MSP-7 in *P. falciparum* showed to interact with P-selectin which known for its role in host immunomodulation, helping to maintain parasite survival in the longer term (Perrin *et al.*, 2015). However, whether or not MSP-7 in *P. vivax* establishes the similar interaction remains to be addressed. Avidity-based extracellular protein interaction screening (AVEXIS) can be used to assess the host-parasite interaction, further support the additional role of PvMSP-7.

The RNA-seq SNPs analysis inferred from the PCA plot and phylogeny trees suggesting that the geographical location is a confounding factor for population stratification as described in Chapter 2. Ten RNA samples were seen to cluster according to their geographical origin. The samples from Ubon Ratchathani province

were clustered into one clade (Figure 5.13 and 5.14). The other samples from Yala province were clustered into another clade. The clustering pattern suggesting that the genetic variation for each individual originated from the same geographical location was similar. Strikingly, this piece of information did not coincide with patients' length of patency derived in differential gene expression analysis. PCA plots generated from the transcriptome and RNA-seq SNPs analysis did not yield similar clustering pattern. Two patients experienced longer patency clustered as a group in the transcriptome PCA (Figure 5.4), however, no obvious separation was seen between these two patients and other patients from the same location in RNA-seq SNPs PCA (Figure 5.12). A study conducted in Kenya revealed that the stages of malaria infection significantly impact the gene expression pattern (Griffiths *et al.*, 2005). Acutely ill patients infected with malaria had a distinct gene expression likely to derive from the immune-related responses and cell activity (Griffiths *et al.*, 2005). This further suggests PvMSP-7 gene expression is developmentally regulated and genetic variation is independent from the length of patency.

In addition, the gene expression has been suggested to influence by host genetic polymorphisms (De Mendonça *et al.*, 2012; Driss *et al.*, 2011). Multiple variants and mutations in human hosts might contribute to disease susceptibility. Multiple SNPs found in the superoxide dismutase-1 (SOD-1) enzyme was reported to impact the expression pattern in patients infected with malaria (De Mendonça *et al.*, 2012). Moreover, expression of southeast Asian ovalocytosis has been linked with malaria infection (Kidson *et al.*, 1981). Southeast Asian ovalocytosis is an inherited dominant trait that involves 27-pair deletion of band 3 protein in the erythrocyte membrane (Williams, 2006). Therefore, further investigation of patients' medical history and host gene expression profile would enable a better understanding of malaria susceptibility. The major limitation of the study was the low read depth and low sample size which may have introduced biases in the analysis. Overcoming these limitations may increase the strength of the current study.

The number of vaccine candidates in *P. vivax* lags far behind than that of *P. falciparum*, therefore, antigen discovery is crucial. MSP-7 paralogs are promising vaccine candidates due to their assumed presence on the merozoite surface, although the previous paragraph shows how this assumption might not be valid. If PvMSP-7

paralogs have distinct functions, not all members would be equally valuable for vaccine design. Potential vaccine candidates from this multigene family should have limited sequence polymorphism while eliciting protective immune responses against natural infections. Such candidates are often cell surface proteins, encoded by multigene families. Few studies have explicitly addressed how variation within a gene family could affect vaccine efficacy, but the Reticulocyte-binding protein (RBP) family, a promising vaccine candidate either for *P. vivax* (Han *et al.*, 2016) or *P. falciparum* (Baum *et al.*, 2009; Campeotto *et al.*, 2017), is one case. *P. falciparum* reticulocyte-binding protein homolog 5 (PfRH5) is currently under clinical trials as a vaccine target. It is found to be highly conserved in field isolates, essential in erythrocyte invasion, and shown to stimulate immunogenicity in an animal model (Baum *et al.*, 2009; Campeotto *et al.*, 2017). Of the eleven RBP gene members in *P. vivax*, two paralogs PvRBP1a and PvRBP1b are both as vaccine candidates due to their localisation at the microneme during schizont stage and the role in host-interaction (Galinski *et al.*, 2000). Han and colleagues (2016) also reported the choice of these two paralogs as vaccine candidates over other gene members due to the conservation in antigenicity and proven protective properties in mouse models. A closer look at their expression patterns during the IDC shows that these genes are highly expressed in the schizont stage (Bozdech *et al.*, 2008).

5.5 Conclusion

The analyses have addressed the PvMSP-7 gene expression in ten *P. vivax* patients with asynchronous parasite populations. The PvMSP-7 paralogs revealed different expression profiles within the IDC. Three members (-7A, -7F, and -7M) have a stable expression over the whole IDC, which suggests additional roles besides merozoite invasion. Two further paralogs are restricted in expression to the late schizont (i.e. merozoite surface) only. The presence of differential expression seems to be a consistent property of *Plasmodium* species. Therefore, from the perspective of vaccine design, careful evaluation of regulatory differences within multicopy gene families is necessary; constitutively expressed MSP-7 paralogs may make better vaccine candidates as they are exposed to host factors for greater periods.

Chapter 6

Identification of antigenic B-cell epitopes within *Plasmodium vivax* merozoite surface protein 7 (PvMSP-7)

Abstract

PvMSP-7 family members are promising vaccine candidates for blood stage infection. Previous chapters have shown that heterogeneity in sequence variation and gene expression among PvMSP-7 paralogs are important considerations in vaccine development. In this chapter, the host immune response to the paralogous genes of PvMSP-7 proteins was considered. Naturally acquired antibodies to PvMSP-7 are yet to be characterized, so a high-density peptide microarray was used to identify immunodominant epitopes in PvMSP-7 proteins. 1173 different amino acid peptides covering the entire sequences of all 13 PvMSP-7 paralogs were spotted on to a microarray chip. Each peptide was 15-mer in length with an overlap of 11 amino acids printed in duplicate. Five pools of sera from naturally infected human patients were divided into age groups and used to screen for IgG reactivity against PvMSP-7. A higher number of differentially detected peptides was found in three younger age groups (0-14, 15-29, and 30-44) compared to the two older groups (45-59 and 60-74). 14 immunogenic linear B-cell epitopes were observed that exhibited cross-reactivity in all age groups. Most of these epitopes are located within the intrinsically unstructured/disordered region and random coiled-coil structures of PvMSP-7 proteins, which are promising targets for antibodies. Among all PvMSP-7 paralogs, the greatest number of naturally acquired, cross-reactive IgG immune responses was made against PvMSP-7A and PvMSP-7L, indicating that differences in antigenicity among paralogs must be an important consideration in developing a PvMSP-7 vaccine. As a result of this work, the immunodominant linear B-cell epitopes found within the conserved domain of PvMSP-7A represent the best candidates for vaccine development.

6.1 Introduction

In previous chapters, the PvMSP-7 paralogs demonstrated variation in their sequence polymorphism and in stage-specific expression patterns during infections. For vaccine development, it is important to learn about relevant epitopes of candidate antigens. Poor immunogenicity has been the main obstacle for malaria vaccine development (Matuschewski and Mueller, 2007). To identify the immunogenic epitopes across 13 PvMSP-7 paralogs and pinpoint promising paralogs to be incorporated in vaccine development, serum from naturally infected hosts was screened for all linear immunogenic epitopes encoded by the 13 protein sequences, using a high-density peptide microarray. This approach is different from a conventional protein microarray which requires expression of recombinant protein in its soluble form and is highly laborious. The peptide microarray offers a highly cost-effective approach to translate potential peptides into vaccine development. To date, peptide microarray has been used to identify the immunogenic epitopes in *P. falciparum* schizont egress antigen 1 (PfSEA-1A) (Nixon *et al.*, 2017), *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) (Quintana *et al.*, 2018), repetitive interspersed genes (RIFIN) (Quintana *et al.*, 2018), *P. falciparum* surface-associated interspersed genes (SURFIN) (Quintana *et al.*, 2018), and *P. vivax* Duffy Binding Protein (PvDBP) (Chootong *et al.*, 2010).

Using other studies as a guide, the immunogenic epitopes were determined and they may confer protective humoral immunity. Nixon *et al.* (2017) attempted to identify immunoreactive epitopes of PfSEA-1A using peptide microarray technology. The array was designed with 15-mers amino acid covering position 810 to 1023. Five peptides of 273 overlapping peptides were found highly cross-reactive with infected sera from *P. falciparum* patients in Kenya. The authors further developed these five linear B-cell epitopes into vaccines with an adjuvant and performed challenge experiment in an animal model. All challenged animals responded to one or more of the immunogenic peptides discovered in peptide microarray. Combination of multiple epitopes showed a synergistic effect with a significant reduction of parasitaemia in experimental animals (Nixon *et al.*, 2017). The magnitude of protection conferred by these epitopes showed that a peptide microarray can be highly effective in screening possible B cell epitopes for immunogenicity. Likewise, a similar approach was employed by Quintana *et al.* (2018) to investigate the immune responses of three gene families (PfEMP1, RIFIN,

and SURFIN) responsible for severe malaria in *P. falciparum* (Quintana *et al.*, 2018). The investigation was carried out in a cohort of children from a malaria endemic area. A peptide microarray was designed encompassing whole surface protein families, which identified multiple reactive peptides. at several different locations of the antigen sequences. Overall, 19 epitopes were identified in PfEMP1, three epitopes in RIFIN-A, and eleven epitopes in $SURFIN_{4.2}$, and the authors intend to characterise the role of these epitopes in rosette formation (Quintana *et al*., 2018).

In *P. vivax*, a peptide microarray was designed to screen the linear B-cell epitopes encompassed the entire *P. vivax* Duffy Binding Protein (Chootong *et al.*, 2010). A total of 178 peptides with a 12-mer overlap were arranged into a 96-well format. Ten B-cell epitopes, mostly located in the central domain, displayed strong immune responses. This result is consistent with previous observations, in where the central region was suggested to be essential for receptor recognition (Hans *et al.*, 2005). Antibody affinity tests using total IgG were performed on the ten immunogenic epitopes, showing that the ten epitopes displayed different magnitudes of inhibitory binding effect to erythrocyte. Strong immune responses were also observed from all PvDBP conserved epitopes using ELISA (Ntumngia *et al.*, 2012). Clearly, further work on the protective immunity induced by these epitopes is required to design a broadly effective malaria vaccine, but these two studies advanced our understanding of PvDBP conserved epitopes and immunogenicity. Likewise, the initial identification of linear B-cell epitopes across 13 PvMSP-7 paralogs is an essential stage and constitutes a starting point for vaccine development.

Immune responses to malaria have been described in Chapter 1, section 1.14. Antibody-mediated immune responses prime the malaria immunity. A Gambian population displayed a stable total IgG antibody level to PfMSP-2, which contributed to sterile protective immunity in the same community (Taylor *et al.*, 1998). No studies have yet investigated the PvMSP-7 epitopes involved in natural immune responses. Immunogenic epitopes should be readily identified in different age groups from naturally infected blood serum. Naturally acquired immunity is prevalent in human populations where malaria transmission is high. However, differences in serological response frequently occur with respect to age (Aponte *et al.*, 2007; Doolan *et al.*, 2009; Pinkevych *et al.*, 2012). Most complicated malaria cases occur in young children in

holoendemic areas. In contrast, immune responses to malaria increase with age after a series of exposures (Pinkevych *et al.*, 2012). Slower replication of malaria parasite and reduction of parasitaemia have both been demonstrated to correlate with age (Pinkevych *et al.*, 2012). Therefore, using the high-density peptide microarray technology, the highly reactive PvMSP-7 peptides present in each age-group, and the cross-reactive peptides in all age groups will be identified.

6.2 Methodology

6.2.1 Human ethics statement

The study was approved by the Institutional Review Board in Human Research of Faculty of Medicine, Chulalongkorn University, Thailand (IRB No. 104/59). Written consent was obtained from all participants or from their parents or guardians' prior admission into the study.

6.2.2 Human sera

P. vivax infections were confirmed using microscopy and the same molecular approach as described in Chapter 2. Patients who showed complicated malaria symptoms and underlying immunodeficiency disorders were excluded from the study. 64 patients infected with vivax-malaria were recruited from two major malaria endemic areas in Thailand, Ubon Ratchathani province and Tak province. Of these, 15 serum samples from Tak province were collected in 2013, while 49 serum samples from Ubon Ratchathani province were collected during 2014-2016. Negative controls, i.e. malaria naïve donors (n=20), were recruited from non-malaria endemic areas at Chulalongkorn hospital. Approximately two millilitres of venous blood was collected from patients with single-strain *P. vivax* infection. Five experimental groups comprised serum from vivax-infected patients based on age-group (Table 6.1). A master pool of serum for each age –group was prepared by pooling five microliters of serum from each vivax-infected

patient. A similar pool of serum was prepared from 22 malaria naïve donors. The antibody responses from the five groups of vivax-infected patients were compared against the negative controls in the analysis. Pooling of serum has been demonstrated as a feasible approach in seroprevalence testing without affecting the sensitivity and specificity in human immunodeficiency virus (Cahoon-Young *et al.*, 1989). The similar serum pooling approach was used in multiple antigen peptide vaccines against *P. falciparum* (Mahajan *et al*., 2010). All samples were preserved at -80°C until used.

Table 6.1 The number of patients in five different age groups.

Group	Age	Number of samples (*n*)
1	0-14	11
2	15-29	22
3	30-44	19
4	45-59	10
5	60-74	2
6	Negative controls	22

6.2.3 Microarray screening

A custom peptide microarray was constructed using 13 PvMSP-7 paralogous sequences. The peptide microarray involved primary and secondary antibodies. The primary antibodies (infected serum) was first bound to the microarray surface and the specific secondary antibody (anti-human) with fluorescent dye was added to bind specifically to the antigen. The fluorescent signal emits from each peptide represents the antibody response. The microarray contains 15-mer peptides of each PvMSP-7 gene with an overlap of 11 amino acids, printed in duplicate. In total, 13 PvMSP-7 paralogs translated into 1173 different amino acid peptides (2,346 peptides in duplicate). Each peptide slide consisted of three identical array copies to ensure the reliability of the results. The peptide arrays were framed by flag anti-HA (YPYCVPDYAG, 52 spots) as a quality control measurement. The signal intensities of these control peptides

indicate the spot uniformity and binding specificities. The peptide microarrays were produced by PepperPrint (Heidelberg, Germany). The peptide microarrays were coated with poly(ethylene glycol)-based graft copolymer with a thickness of 13.5 nm and an additional three amino acid linker (β-alanine, aspartic acid, and β-alanine). The addition of three amino acids is to ensure optimal epitope orthogonal attachment and presentation. All microarray slides were used within a month after delivered by PepperPrint. The microarray slides stored at -20°C were stable for at least six months without losing its reactivity.

Before the microarray testing was conducted for each group of patients, optimisation of primary and secondary antibody was performed using a single subarray. The optimization of secondary antibody began with the lowest dilution at 1:5000 then 1:2500 with standard buffer. Goat anti-human IgG (H+L) DyLight680 antibody was used as a secondary antibody. Pooled sera of vivax-infected patients were tested with the dilution of 1:1000, 1:500, and 1:50. Dilution of polyclonal serum at 1:50 and secondary antibody at 1:2500 was seen to be the optimum conditions for testing, as the fluorescent intensity was clearly observed on each spot (fluorescence intensity>2000). Subsequently, the peptide microarray was probed with vivax-infected patients' polyclonal serum with goat anti-human IgG (H+L) DyLight680 antibody (Rockland Immunochemical, Gilbertsville, USA).

The patients' polyclonal serum was incubated at a dilution of 1:50 in the presence of goat anti-human IgG (H+L) DyLight680 antibody at a dilution of 1:2500. Each spot intensities were quantified on an Agilent G2565CA microarray scanner equipped with Agilent SureScan Technology to ensure precision microarray scanning. The peptides were scanned with the AgilentHD red colour single channel at 10 μm resolution and output to a 16-bit greyscale tiff files. The microarray analysis was completed with a PepSlide analyser and saved as TIF files. The PepSlide analyser performed global normalisation of all signal intensities between arrays to remove the noise in each TIF file. The mean and median of local background and foreground intensity values for all peptide spots were determined by the PepSlide analyser. The fluorescence intensity was quantified and exported into the comma separated values (CSV) files for further analysis. Each of the CSV files contains information about the location, peptide, and fluorescent intensity on each peptide spot.

6.2.4 Microarray incubation

The experimental procedure of peptide microarray was performed as described by the manufacturer's manual (PePperPrint, 2017). Four reagents were prepared as follows:

i) Standard buffer: Phosphate-buffered saline (PBS) with 0.05% Tween20, pH 7.4
ii) Blocking buffer: Standard buffer with 1% BSA
iii) Staining buffer: PBS with 0.05% Tween20 and 10% blocking buffer
iv) Dipping buffer: 1mM Tris, pH 7.4

Firstly, the peptide microarrays were treated with blocking buffer for 30 minutes at room temperature. The peptide microarrays were incubated with goat anti-human IgG (H+L) DyLight680 antibody that was diluted 1:2500 in staining buffer. Next, peptide microarrays were incubated with patients' sera overnight at 4°C. The pool of serum was diluted 1:50 with staining buffer. The peptide microarray was washed three times with standard buffer using an orbital shaker at 140 rpm. The peptide microarrays were dipped in the dipping buffer until all contaminants were washed off from the peptide microarray surface. Finally, the peptide microarrays were dried by tapping the edge of the slide against a pad of tissue paper. The steps were repeated with Cy3 conjugated anti-HA control antibody supplied by PepperPrint (Heidelberg, Germany).

6.2.5 Pre-processing methods

Peptide array data, pre-processed by the PepSlide analyser and saved in CSV files, were used to identify the peptides producing significant responses, i.e. greater than background responses in controls. Plots were generated to assess the quality of the arrays. Background correction and intensity normalisation were performed using the LIMMA package (Smyth, 2005) implemented in R version 3.4.3 (R, 2017) . Background correction was performed on each array using the subtraction method (Figure 6.1). This step was performed to eliminate the effects of non-specific binding across the arrays. The subtraction method subtracts the local background estimates them from the foreground intensity and is commonly applied in microarray analysis

(Ritchie *et al.*, 2007). Figure 6.1 shows the intensity level on the X-axis and the total number of measurement values with a given log intensity on the Y-axis. The total area under the pile of each array corresponds to the total number of observations. By comparing the measurement values before and after the background correction, is can be seen that the density of each experimental group became more homogenous after the correction. Following background correction, normalisation of intensities for between-array variation (Figure 6.2) was conducted, before the identification of significantly responsive peptides. Normalisation was accomplished between-arrays using the quantile principle in LIMMA (Smyth, 2005). The quantile function generates a more uniform intensity distribution, compensating for systematic measurement errors between-array, leaving only the true biological differences in the dataset. Figure 6.2, each box shows the distribution of expression values within one array as boxplots. Fluorescent intensity before normalisation showed some deviation (Figure 6.2a); however, the median of boxplots within each experimental group was more uniform after normalisation. After normalisation, the data were analysed statistically using LIMMA.

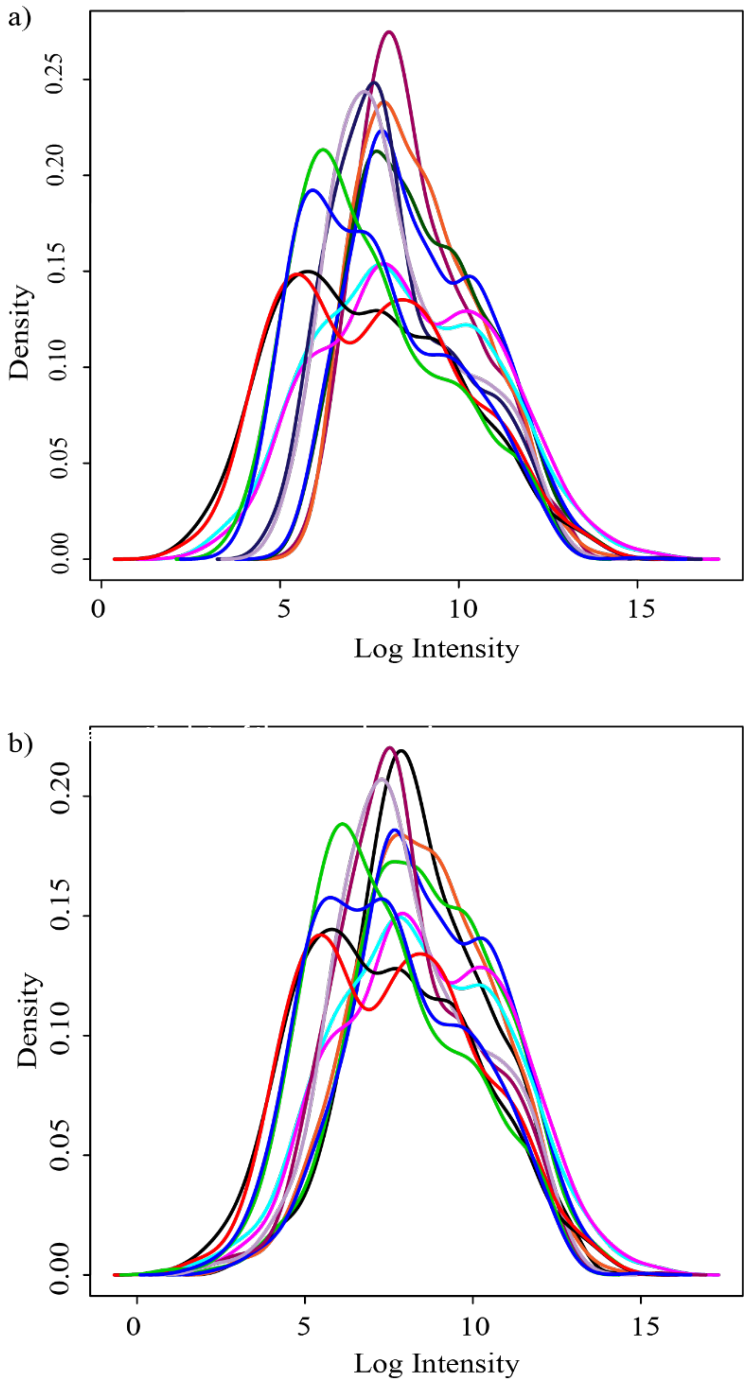


Figure 6.1 Diagnostic plots of the one-colour microarray. Figure **a)** shows the intensities before background correction. Figure **b)** displays the intensities after background correction using the subtraction method (Ritchie *et al.*, 2015) implemented in LIMMA (Smyth, 2005). On the X-axis of each plot, it shows the log intensity of each peptide microarray, whilst the Y-axis implies the number of measured values with a given log intensity. The area under the pile of each array indicates the total number of observations. Each colour on the plot corresponds to the individual array.

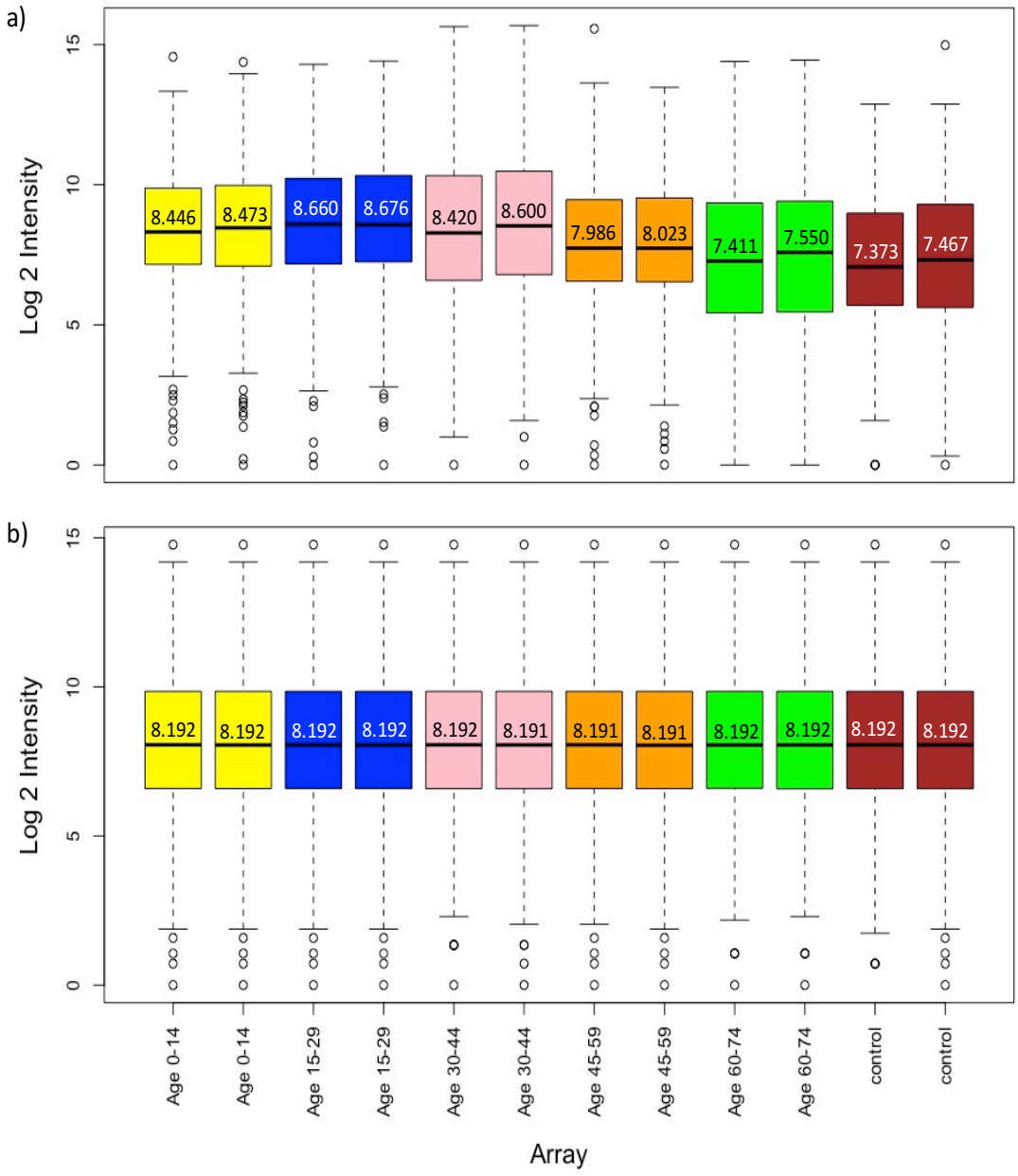


Figure 6.2 Boxplots display the intensities before and after normalisation. Figure **a)** shows the microarray log intensity without normalisation. Some deviation was observed before intensity was normalised within each experimental group. Figure **b)** illustrates the log-transformed normalised intensities using the quantile method. The mean and the range of intensity between each experimental group was more uniform after the normalisation. Each experimental group has a replicate which represents the same colour on the boxplot. The mean $\log_2$ intensity for each peptide array is shown on the boxplot.

6.2.6 Statistical analysis

Statistical analysis was performed using the LIMMA package (Smyth, 2005) implemented in R version 3.4.3 (R, 2017) . Following the normalisation procedure, the data were fitted to a linear model implemented in the LIMMA package. The package processed the data with *t*-statistics. The aim of the statistical analysis was to identify the significantly detected peptides in all experimental groups by comparing the log transformation of expression intensity. After fitting the data to the linear model, a simple Bayesian model was used to estimate significant peptides between the experimental group and negative control (Smyth, 2004). This model is similar to *t*-statistics except the standard errors have been moderated across peptides. The summary statistics were computed by the eBayes() function where a list of significantly responsive peptides were displayed. Benjamini and Hochberg's method or false discovery rate (FDR) was used to correct the observed *P*-value. FDR < 0.05 were considered statistically significant. The outputs from the statistical analysis include log fold change, t-statistics, *P*-value, and FDR.

6.2.7 Protein secondary structures, protein disordered region

The conformation structure of the protein has been suggested to associate with immunogenicity (Scheiblhofer *et al.*, 2017). The protein structure determines the presentation of the immunogenic epitopes on major histocompatibility complex for triggering immune responses (Scheiblhofer *et al.*, 2017). Therefore, immunogenic epitopes of PvMSP-7 were mapped to the predicted protein secondary structure to evaluate the immunogenicity and enhance the vaccine development strategy. The protein secondary structure for 13 PvMSP-7 paralogs was predicted using JPred4 (Drozdetskiy *et al.*, 2015) implemented in Jalview version 2.10.1 (Waterhouse *et al.*, 2009). JPred4 accurately predicts the secondary structure based on the JNet algorithm. Overall, the accuracy of the secondary structure prediction is reported to be 82% (Drozdetskiy *et al.*, 2015). It provides three states of secondary structure predictions, alpha-helix, beta-strand, and coiled-coil. The predicted protein secondary structures were saved as an SVG image. Several leading malarial vaccines with roles in host-cell invasion contain an extensive intrinsically unstructured domain, and protein disordered

domains have been shown to affect adaptive immunity against *P. falciparum* (Guy *et al.*, 2015). Moreover, linear B-cell epitopes are also known to be enriched along protein disorder domains (Guy *et al.*, 2015), which have enhanced antibodies affinity (Sormanni *et al.*, 2015). In the present study, the intrinsically unstructured domains spanning along the PvMSP-7 paralogs were identified. This prediction was used to investigate the association of PvMSP-7 immunogenic epitopes and the intrinsically unstructured protein domains which likely to implicate for vaccine design. The intrinsically unstructured or protein disordered domains were identified by using the GeneSilico MetaDisorder service (Kozlowski and Bujnicki, 2012). It reliably predicts the protein disordered regions using 13 arbitrarily disorder predictors, which were found to be outperformed with other primary methods.

6.2.8 *In silico* B-cell epitopes predictions

To evaluate the prediction accuracy of *in silico* B-cell epitopes prediction, the proportion of theoretical epitopes was compared with the experimental immunogenic epitopes. Bepipred linear epitope prediction (Larsen *et al.*, 2006) implemented in the IEDB database (http://tools.iedb.org/bcell/) was used to predict B-cell epitopes in 13 PvMSP-7 paralogs (Vita *et al.*, 2014). The algorithm uses Hidden Markov models (HMMs), which are known to yield reliable B-cell epitopes. Protein sequences in FASTA format for all PvMSP-7 paralogs were compared to the database, which predicted a score for each amino acid. The window size was set to 15 amino acids and 50% specificity in all sequences in order to obtain reliable linear B-cell epitopes.

6.3 Result

6.3.1 *In silico* analysis of putative linear B-cell epitopes in PvMSP-7 proteins

The potential linear B-cell epitopes in 13 PvMSP-7 proteins were predicted using Bepipred linear epitope prediction (Larsen *et al.*, 2006) implemented in the IEDB database (http://tools.iedb.org/bcell/). As shown in Figure 6.3, most of the putative B-cell epitopes were distributed in the central region of the PvMSP-7 protein structure. All the high score linear epitopes located within the central domain with prediction scores of at least 2.0. Moreover, these epitopes appeared to be a long stretch of amino acid residues. PvMSP-7D was an exception to this trends, with the location of predicted epitopes in this paralog biased toward the C-terminal. The number of predicted B-cell epitopes was found to vary among different 13 PvMSP-7 paralogs. Certain PvMSP-7 proteins were found to have fewer antigenic epitopes, for example, PvMSP-7D and -7J (Figure 6.3). Using a threshold score of 0.5, 11 PvMSP-7 showed a minimum of ten predicted immunogenic epitopes. The threshold score was estimated based on the Karplus and Schulz flexibility on each of the residue in the epitopes (Karplus and Schulz, 1985). The higher the score, defines the probability to be an immunogenic epitope. In the analysis, epitopes with a threshold score above 0.5 were considered immunogenic epitopes. Threshold score was set at 0.5 to achieve optimal sensitivity and specificity for detecting linear B-cell epitopes. Previous studies showed a low threshold score tends to yield high sensitivity but low specificity in detecting immunogenic epitopes or vice versa, leading to false positive or negative outputs (Guy *et al.*, 2015).

Protein secondary structures of PvMSP-7 were examined using JPred4 (Drozdetskiy *et al.*, 2015). The predicted protein secondary structures in PvMSP-7 revealed that most of the alpha helices were located in the N- and C-terminal, while the central region had a predominantly random coiled-coil structure. From the analysis, PvMSP-7F contains only one alpha helical structure located in the N-terminal, whereas the remaining structure is composed of random coiled-coil. PvMSP-7M consists of the highest number of alpha-helices, four and five alpha-helices in the N- and C-terminal, respectively. Beta strands were also detected in five PvMSP-7 paralogs encoded in the N- and C-terminal (Figure 6.3). Overall, all predicted linear B-cell epitopes displayed

a specific feature, one where they span the random coiled-coil central region of PvMSP-7, but were routinely distant from the alpha-helices and beta strands of protein secondary structures.

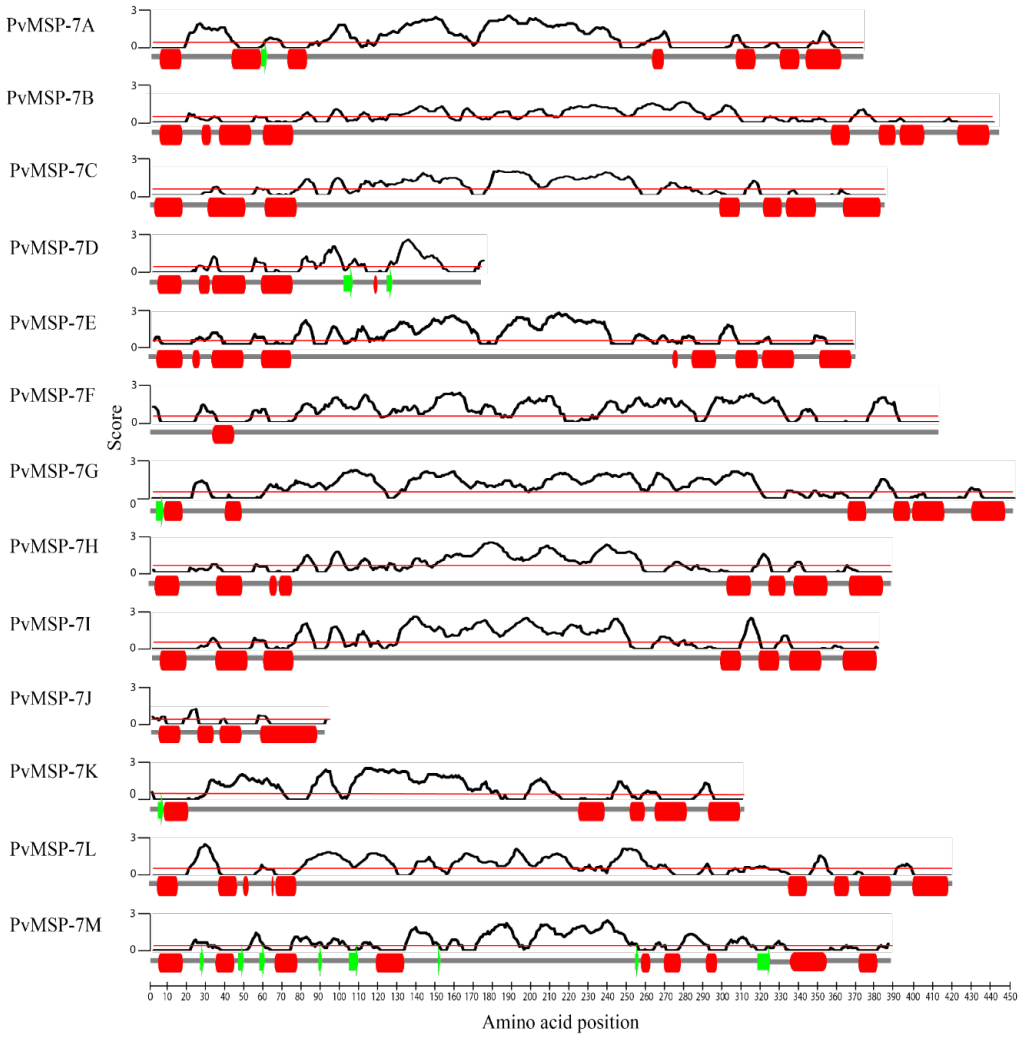


Figure 6.3 ***In silico*** **predicted linear B-cell epitopes in the context of predicted protein secondary structure for the PvMSP-7 multigene family.** Linear B-cell epitopes were predicted using Bepipred linear epitope prediction (Larsen *et al.*, 2006) implemented in the IEDB database. The predicted score for each amino acid residue is shown on the line graph, defines the probability to be an immunogenic epitope. The predicted values above the threshold (red line) are considered significant and likely to have B-cell epitopes. The threshold was set at 50% to increase the reliability of the B-cell epitopes prediction. Predicted protein secondary structures for 13 PvMSP-7 proteins were produced using JPred4, and are shown below each plot. Red boxes: alpha-helix, green arrows: beta-strand, and grey lines: coiled-coil.

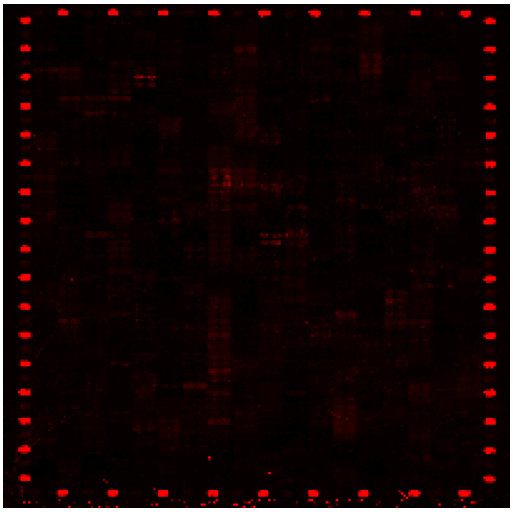
6.3.2 Mapping of PvMSP-7 linear B-cell epitopes by peptide microarray

To date, the immunogenicity of 13 PvMSP-7 paralogs in natural infection remains unknown. To identify immunoreactive linear B-cell epitopes in the PvMSP-7 family, the peptide microarray was incubated with sera from patients with single *P. vivax* infection, pooled according to five age groups (Figure 6.4). The high-density peptide array consists of 1173 peptides of 15-mers spanning the complete coding sequence of 13 PvMSP-7 paralogs. The serum peptide-reactivity profile in *P. vivax* patients was detected by goat anti-human IgG attached with a fluorophore (DyLight 680). Based on the fluorescent signals captured on the peptide microarray, pool of serum in the 30-44 age group displays the strongest reactivity to IgG antibody (Figure 6.4c). The red spot intensities were greater compared to other experimental groups, indicating a clear interaction pattern. The weakest immunoreactivity was observed from the 45-59 age group. Fluorescence intensity in the 0-14 and 15-29 age groups showed comparable fluorescent intensity (Figure 6.4a and 6.4b). A pool of serum from malaria-naïve individuals was served as a negative control in the analysis (Figure 6.4f). Fluorescence signals on the negative control were observed on the array, although they appeared to be very weak. Each of the arrays was framed by the HA control peptides, the spot intensities were homogenous as seen in Figure 6.4. However, the fluorescence intensity of HA control peptides in Figure 6.4e was not as sharp compared to other arrays.

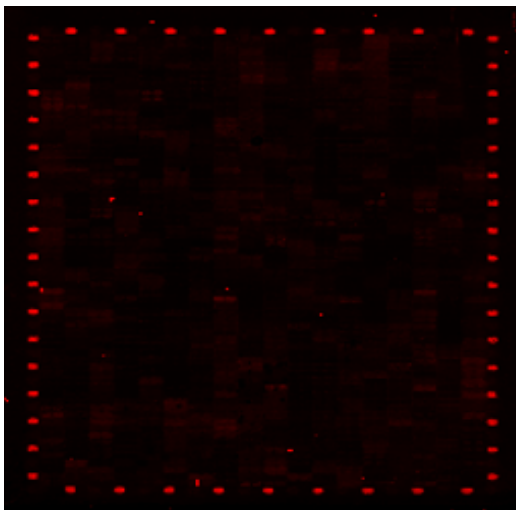
The number of naturally immunogenic peptides distributed among 13 PvMSP-7 proteins varied. Figure 6.5 shows the distribution of significantly responsive peptides according to each experiment group. Five PvMSP-7 paralogs (PvMSP-7A, -7B, -7K, -7L, and -7M) revealed a high number of antigenic peptides spanning the N-, central, and C-terminal. The immunogenic peptides in PvMSP-7I were biased toward the central and C-terminal. Six other PvMSP-7 proteins (PvMSP-7C, -7E, -7F, -7G, -7H, and -7J) have lower number of immunogenic peptides. Meanwhile, no antigenic peptides were predicted for PvMSP-7D. The relationship of immunogenic epitopes and predicted protein secondary structures were consistent with the *in silico* prediction, in that most of the immunoreactive epitopes were contained within the random coiled-coil motifs.

Comparing the results of *in silico* and experimental B -cell epitope prediction (Figure 6.3, 6.5), there was consensus in the predicted epitope sequences for PvMP-7A,

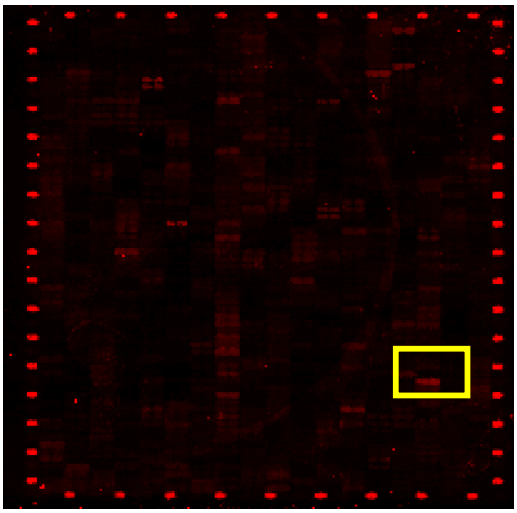
-7B, -7F, -7G, -7H, -7I, -7K, -7L, and -7M, although the number of epitopes varied. Moreover, the length of the peptide derived from *in silico* prediction and experimental was different as peptide microarray was designed based on 15-mer amino acid residue. On average, the *in silico* prediction derived approximate 30 amino acid residues in an epitope. Nevertheless, the naturally immunogenic epitopes were contained within the *in silico* predicted sequences. For instance, two predicted peptides spanning PvMSP-7A in amino acid residue 19 – 44 and 120 – 169 were found to be significant responders in five experimental groups (peptide position; 10 – 25 and 158 – 173). There were some exceptions to the consensus between methods. While PvMSP-7A, -7B, -7I, -7K, -7L, and -7M displayed most of the consensus epitopes in the central region, PvMSP-7H did not, despite *in silico* analysis indicating a long epitope spanning amino acid residues 136 – 247. The similar pattern is observed in PvMSP-7C and PvMSP-7E. Likewise, no immunogenic epitopes were detected in PvMSP-7D despite *in silico* prediction revealed seven antigenic peptides.



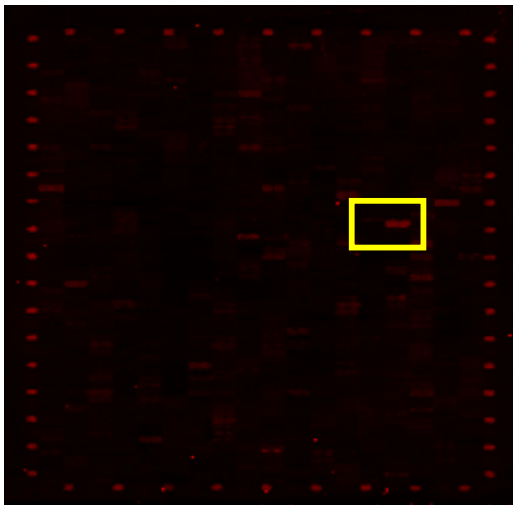
a) 0-14



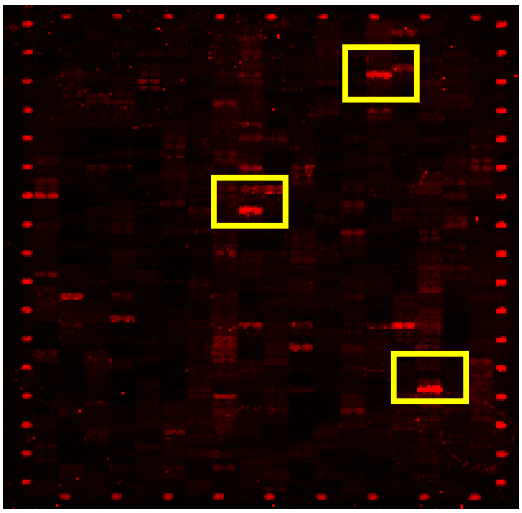
d) 45-59



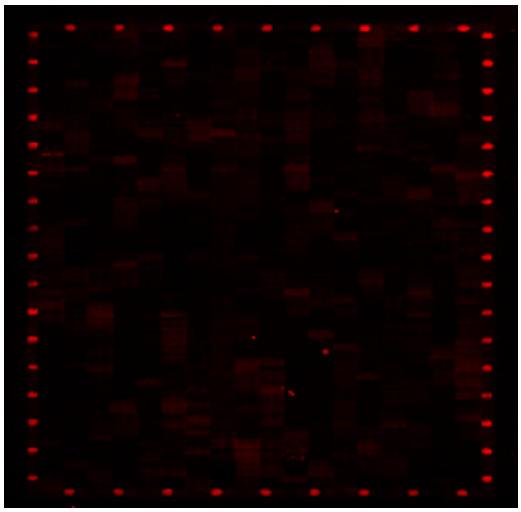
b) 15-29



e) 60-74



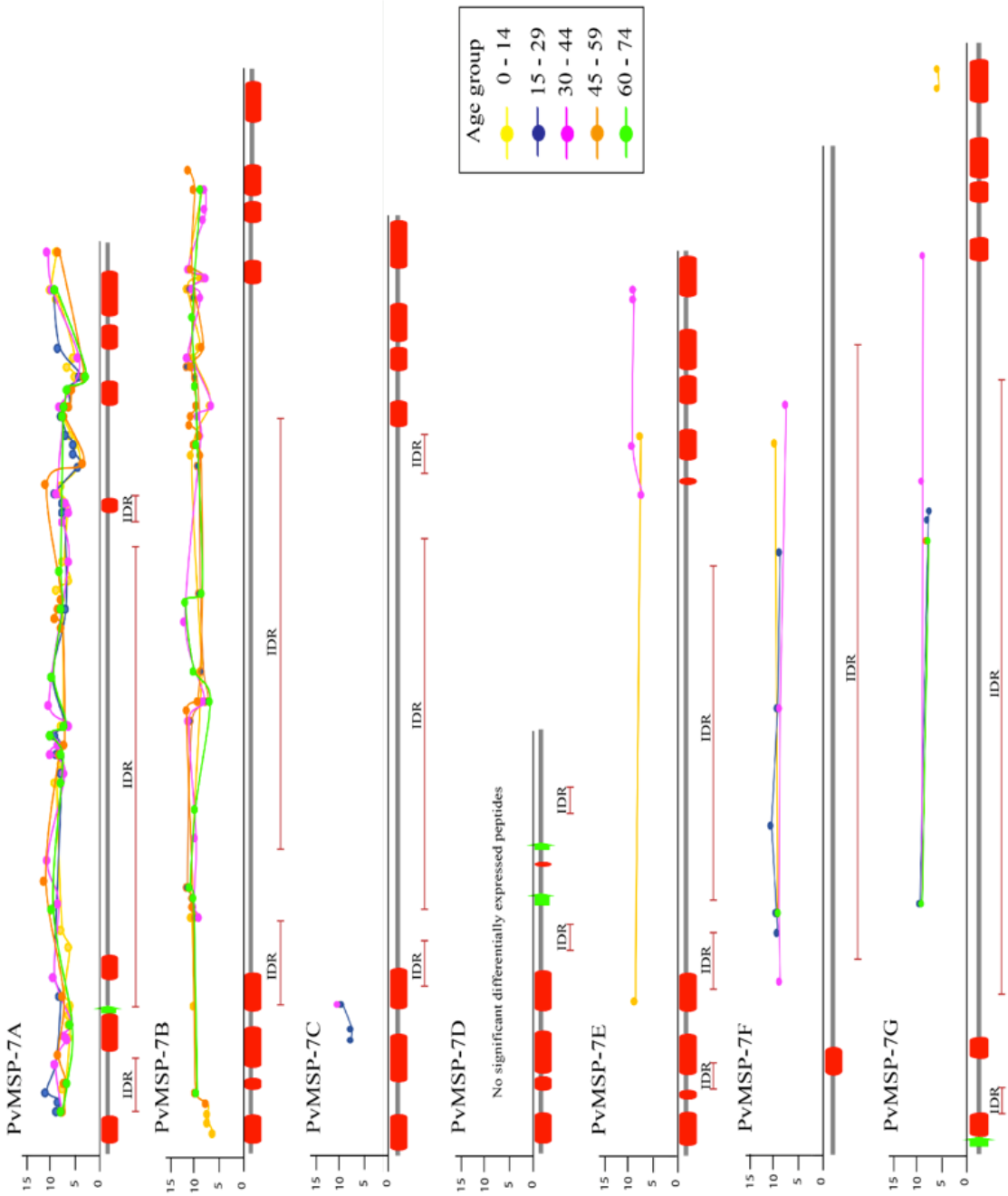
c) 30-44



f) Negative control

Figure 6.4 Mapping of 13 PvMSP-7 epitopes by peptide microarray. The peptide microarray was designed as spot duplicates. Six identical sub-arrays, each consists of 1173 amino acid peptides (2,346 peptides in duplicate). The analysis was conducted in five different age groups, **a)** 0 -14, **b)** 15 - 29, **c)** 30 - 44, **d)** 45 – 59, **e)** 60 – 74, and **f)** negative controls. Incubation of the peptide microarray with the pool of polyclonal antibodies from vivax-infected patients at a dilution of 1:50. The procedure followed by staining peptide microarray with the secondary goat anti-human IgG antibody at a dilution of 1:2500. Control peptides (HA) were located around the border of each peptide array and stained with Cy3 (red) conjugated anti-HA antibody. The fluorescent intensity was estimated by PepSlide Analyser. Each spot on the peptide microarray corresponds to one peptide. The strongest responders as indicated by the bright red fluorescent signal are highlighted in the yellow box.

6.3.3 Naturally immunogenic linear B-cell epitopes within PvMSP-7 paralogs



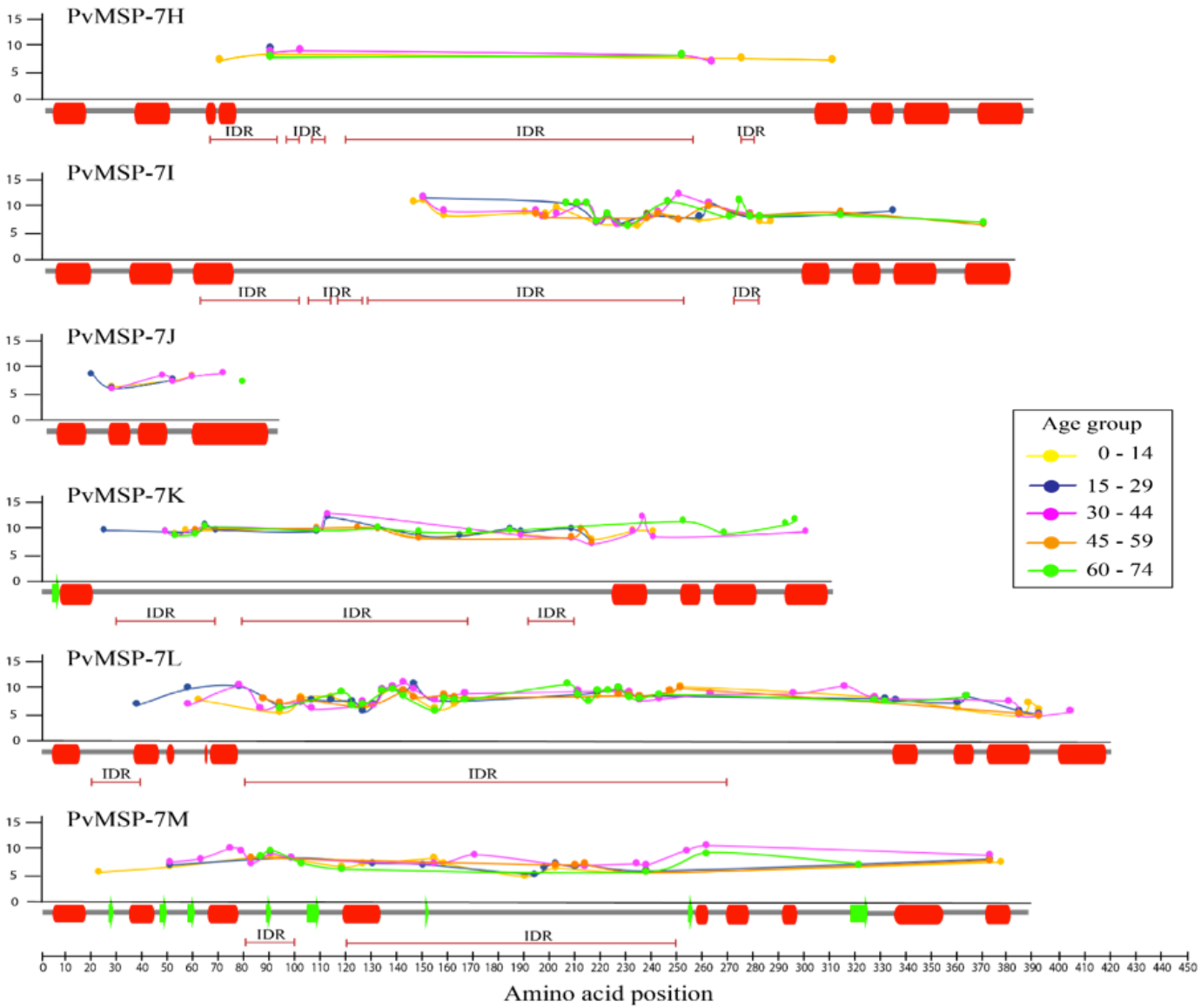

PvMSP-7H
PvMSP-7I
PvMSP-7J
PvMSP-7K
PvMSP-7L
PvMSP-7M
IDR
Age group
0 - 14
15 - 29
30 - 44
45 - 59
60 - 74
Amino acid position

Figure 6.5 Schematic diagram of naturally immunogenic linear B-cell epitopes within 13 PvMSP-7 paralos, and intrinsically unstructured/disordered regions. Linear B-cell epitopes within each PvMSP-7 protein were identified using a custom peptide array screened with serum from naturally infected patients. The intensity of each epitope is shown on the y-axis and was derived from the fluorescence signal followed by background subtraction, quantile normalisation, and log transformation. Pools of serum divided into five age groups; 0-14; 15-29; 30-44; 45-59; and 60-74. Peptides found to respond significantly for each age-group are related to protein structure, with each epitope represented by a coloured dot placed at its middle amino acid position. The colour of epitopes corresponds to age group (see key). Predicted protein secondary structures for each PvMSP-7 paralog derived from JPred4 (Drozdetskiy *et al.*, 2015) are shown below each plot. Red boxes: alpha-helix, green arrows: beta-strand, and grey lines: coiled-coil. The intrinsically unstructured/disordered regions (IDR) predicted using GeneSilico MetaDisorder service (Kozlowski and Bujnicki, 2012) are shown below the predicted protein secondary structure.

6.3.4 Differentially detected peptides

The significant differentially responsive peptides in each experimental group were identified in the analysis. Subsequently, the analysis focused only on the novel immunogenic peptides that present in all age groups. The antigenic peptides contain within the PvMSP-7 repertoire were analysed using the LIMMA package (Smyth, 2005). Different proportion of significant differentially responsive peptides were found between age groups (Table 6.2). Age group of 30-44 obtained the highest number of the significant differentially responsive peptides (number of peptides= 141). Followed by age group 0-14, 15-29, 60-74, and 45-59. Older age group (age 45-59) has the lowest number of differentially responsive peptides (number of significant peptides= 77). The number of differentially responsive peptides was in line with the fluorescence signals captured in the peptide microarray (Figure 6.4). From Figure 6.4, the number of spots and fluorescent intensity in age group 30-44 were clearer compared to other experimental groups. All differentially detected peptides are listed in S6.2.

A Venn diagram (Figure 6.6) was used to illustrate the consensus peptides present in each experiment group. A total of 120 unique differentially responsive peptides were obtained in five experimental groups. The number of unique peptides with respect to age group ranged from 8 to 38 peptides. Age group 30-44 has a higher number of differentially responsive peptides, followed by 0-14, 60-74, 15-29, and 45-59.

Of the 1173 peptides evaluated, 14 novel antigenic peptides in all age groups were differentially responsive in response to IgG antibody (Table 6.3). These consensus antigenic peptides were found in six PvMSP-7 proteins (PvMSP-7A, -7B, -7H, -7I, -7L, and -7M). Closer looks into the position of these novel peptides, they were predominantly distributed in the central domain of the gene (Figure 6.7). An exception was seen in PvMSP-7A, antigenic peptides encoded in three domains of the protein. Of the 14 novel peptides present in all five experimental groups, the majority of these antigenic peptides were derived from PvMSP-7A (42.86%). Six significantly responsive peptides were found in PvMSP-7A, one peptide located in the N-terminal, one peptide encoded in the central terminal, and four peptides span the C-terminal. Among the consensus immunogenic epitopes in PvMSP-7A, the average expression

value ranged from 5.521 to 10.937, which reflected the magnitude of the naturally acquired IgG response.

Meanwhile, four differentially responsive peptides span along PvMSP-7L (Figure 6.7). Three antigenic peptides located in the N-terminal whilst one peptide located in the central domain. Like PvMSP-7A, the IgG response at different peptides was varied (average expression values ranged from 8.860 to 11.886). A higher IgG response was observed in 139-EAVDEEAEKEDTAVI-154. The average expression value across five experimental group was 11.886 (log-fold change= 5.863 ± 1.743). In contrast, only one consensus differentially responsive peptide present in PvMSP-7B, -7H, -7I, and -7M. The IgG response infers from the average expression values ranged from 8.860 to 9.866. Closer looks into the position of these 14 novel epitopes on the predicted protein secondary structure (Figure 6.7). The majority located in the random coiled-coil motifs. Two peptides in PvMSP-7A, 10- CLLLLCAGPVLGDDD-25 and 314-KNTLIKTFKKALYDK-329 spanned between the alpha-helices and random coiled-coil regions.

Table 6.2 Significantly responsive **peptides in five groups of patients.** Significantly responsive peptides in each experimental group were identified using the LIMMA package (Smyth, 2005). Peptides with false discovery rate (FDR) below 0.05 are considered statistically significant.

Group	Age	Significantly responsive peptides (FDR<0.05)
1	0-14	135
2	15-29	119
3	30-44	141
4	45-59	77
5	60-74	90

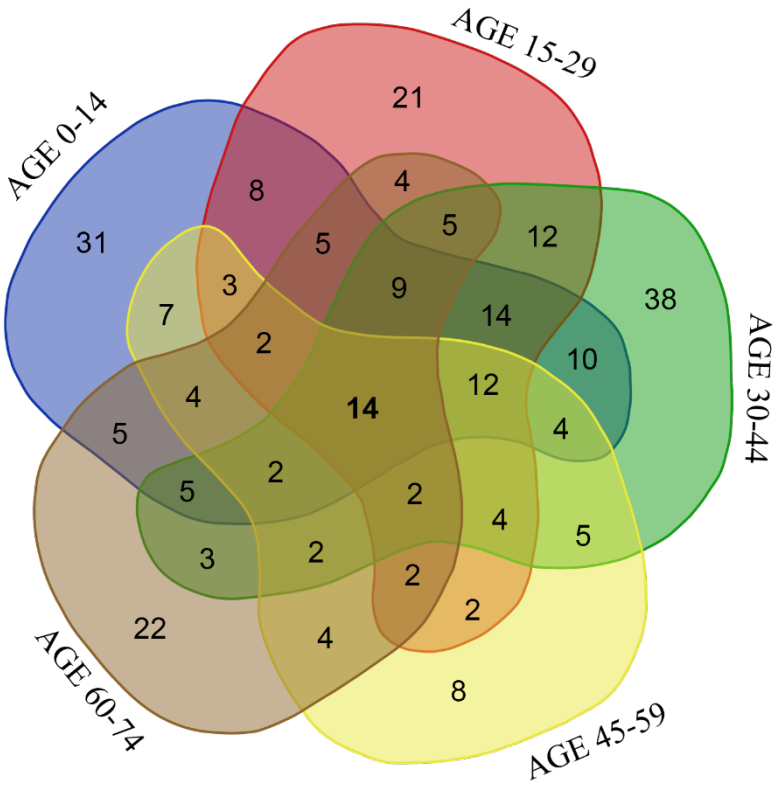


Figure 6.6 Venn diagram showing the overlap in of linear B-cell epitopes predicted by peptide microarray for five patient age groups. 14 peptides that gave significant responses in all experimental groups were observed to be present in all age groups. The significantly detected peptides were identified using *t*-statistics, pairwise comparison between each age group against the control. Peptides with FDR<0.05 were considered to be differentially detected.

Table 6.3 Sequence, parent gene and structural position of 14 PvMSP-7 peptides that gave significant responses on the peptide microarray in all age groups, compared to the negative control.

Peptide	PvMSP-7	Domain	Position
CLLLLCAGPVLGDDD	A	N	10 - 25
EAVQWGPATEEVVAE	A	Central	158 - 173
KLLDTMLTNGQVERE	A	C	298 - 313
TMLTNGQVEREKKNT	A	C	302 - 317
EREKKNTLIKTFKKA	A	C	310 - 325
KNTLIKTFKKALYDK	A	C	314 - 329
YESIHGEDEPQVVPS	B	Central	178 - 193
EEESLGHLLESEDAD	H	N	83 - 98
DEIHVPPFHSKYNDF	I	C	272 - 287
EDTTPKEQQEDQNVS	L	N	91 - 106
QEENTQVKNVIFTEK	L	N	123 - 138
EAVDEEAEKEDTAVI	L	N	139 - 154
SSAESAPNEPDVNTT	L	Central	207 - 222
SVKSGDDGEEEDGAT	M	Central	232 - 247

N: N-terminal, Central: Central-terminal, and C: C-terminal

Consensus immunogenic epitopes in all age groups

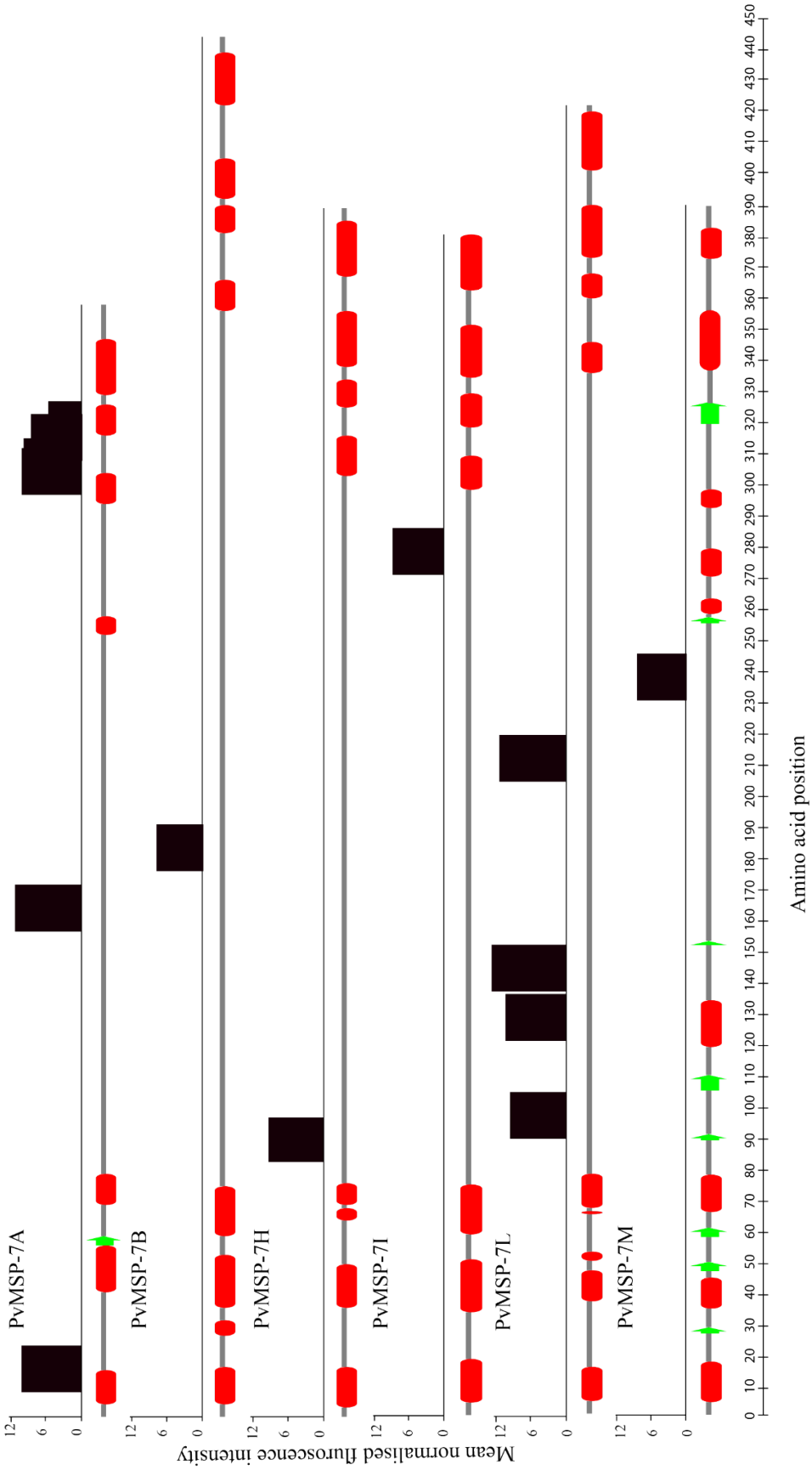


Figure 6.7 Schematic diagram of 14 naturally immunogenic, linear B-cell epitopes present in all age groups, in relation to predicted protein structure. 14 peptides with significant responses on the peptide microarray (FDR<0.05) in all age groups were identified using the LIMMA package (Smyth, 2005). These are arranged in relation to the predicted protein secondary structures of each PvMSP-7 paralog derived from JPred4. Red boxes: alpha-helix, green arrows: beta-strand, and grey lines: coiled-coil. The y-axis shows the mean fluorescent intensity of the observed epitope across the age groups after background subtraction, quantile normalisation, and log transformation.

6.4 Discussion

Using a high-density peptide microarray approach has allowed me to efficiently identify linear B-cell epitopes. A similar approach has been used to characterise antigenic epitopes in *P. falciparum* (Lu *et al.*, 2015; Quintana *et al.*, 2018). Individuals naturally exposed to *P. vivax* in malaria endemic areas of Thailand were used to identify the immunogenic epitopes. Five PvMSP-7 proteins (PvMSP-7A, -7B, -7K, -7L, and -7M) contained the most immunogenic B-cell epitopes (Figure 6.5), and14 peptide sequences in PvMSP-7A, -7B, -7H, -7I, -7L, and -7M were confirmed epitopes in all patient age-groups (Figure 6.7). The high prevalence of cross-reactive IgG responses to PvMSP-7 under conditions of natural exposure supports vaccine development, and the precise identification of epitopes will facilitate the further development of MSP-7 as a subunit vaccine against *P. vivax*.

In silico B-cell epitope mapping has been used to identify potential epitopes in 13 PvMSP-7 proteins. The prediction was performed using Bepipred linear epitope prediction (Larsen *et al.*, 2006). This algorithm was used to map major vaccine candidates including, PvMSP-1 (Soares *et al.*, 2014), PvMSP-9 (Rodrigues *et al.*, 2016) , and PvAMA-1 (Bueno *et al.*, 2011). Consistently, most of the predicted epitopes from these major vaccine candidates showed a remarkable immune response in experimental studies (Bueno *et al.*, 2011; Rodrigues *et al.*, 2016; Soares *et al.*, 2014). It was determined in this study that the predicted B-cell epitopes are scattered along the genes in PvMSP-7 (Figure 6.3). A closer look into the B-cell epitope distribution shows that there is a long stretch of B-cell epitopes located within the central regions of the proteins. The central region of the PvMSP-7 proteins is known to display extensive sequence polymorphism, which strongly implies functional constraints (Cheng *et al.*, 2018; Garzón-Ospina *et al.*, 2010; Garzón-Ospina *et al.*, 2014; Garzón-Ospina *et al.*, 2012). Similarly, the polymorphic central domain of the circumsporozoite protein (CSP) was shown to encode immunodominant B-cell epitopes in *P. falciparum* (Zavala *et al.*, 1983) and *P. vivax* (Arévalo-Herrera *et al.*, 1998), and this domain has subsequently been used as the basis for a vaccine that elicits effective humoral and cellular immune responses against malaria infection (Nardin *et al.*, 2000; Wang *et al.*, 1998). Based on the observations in this chapter, the central domain of PvMSP-7 appears to be the main target of antibodies and the best region of MSP-7 for use in a

vaccine. Only in two cases, PvMSP-7D, and PvMSP-7E, where epitopes not found in the central region, however, these two proteins are relatively short in length, 176 amino acids and 94 amino acids, respectively. It is possible that PvMSP-7D and PvMSP-7E are truncated pseudogenes, which results in fewer epitopes.

In the present study, the B-cell epitopes derived from *in silico* prediction were compared against the naturally immunogenic epitopes in PvMSP-7 proteins. Several consensus epitopes were found particularly in PvMP-7A, -7B, -7F, -7G, -7I, -7K, -7L, and -7M (Figure 6.3 and Figure 6.5). The position of naturally immunogenic B-cell epitopes was not located at the precise predicted position because the peptide array was designed with 15-mers. Nonetheless, the immunogenic epitopes in natural infection still seen to overlap with the *in silico* linear B-cell epitopes. Essentially, the *in silico* epitope prediction is able to distinguish comparable naturally immunogenic regions. Therefore, care should be exercised when the analysis is dependent on *in silico* screen alone.

Intrinsically unstructured/disordered regions are known to associate with immune responses (Guy *et al.*, 2015). Several major malaria vaccine candidates that have been developed for use against *P. falciparum* were reported to contain a long, intrinsically unstructured/disordered region. For example, MSP-2 (Adda *et al.*, 2009), MSP-3 (Van *et al*., 2014) , Glutamate-rich protein (GLURP) (Feng *et al.*, 2006), AMA-1 (Guy *et al.*, 2015), serine repeat antigen 5 (SERA) (Yagi *et al.*, 2014), and CSP (Foquet *et al.*, 2014) are composed of partially or completely intrinsically disordered regions. Moreover, the intrinsically unstructured/disordered regions of SERA5 and CSP have been shown to contain epitopes that induce protective immune responses (Foquet *et al.*, 2014; Yagi *et al.*, 2014). Intriguingly, increased antibody recognition of linear B-cell epitopes was also demonstrated along the intrinsically disordered region (Guy *et al.*, 2015). This feature may be due to greater intrinsic plasticity relative to structural domains, which perhaps facilitates molecular recognition and interactions with other invasive protein targets (Uversky and Dunker, 2013; Wright and Dyson, 2015). Consistently, a majority of the antigenic linear B-cell epitopes in PvMSP-7 (Figure 6.5) are present along the long disordered region in the central domain. Protein-protein interactions validated using ANCHOR have shown potential binding sites within all predicted disordered regions in PvMSP-7 (Rajamani *et al.*, 2004). Although malaria research has not yet reported the association between protein secondary

structure and immune responses, dominant epitopes of *Echinococcus multilocularis* Emy162 revealed random coiled-coil regions with strong antigenicity (Li *et al.*, 2013). Therefore, multiple lines of evidence suggest that linear B-cell epitopes within the PvMSP-7 central domain could play a role in antibody binding and mediate immune responses.

Most of the PvMSP-7 paralogs (PvMSP-7A, -7B, -7C, -7E, -7H, -7I, -7K, -7H, and -7M) also exhibited a short segment of intrinsically unstructured/disordered region in the N-terminal. These disordered segments are flanked by structured regions where they undergo a disordered to structured transition upon functioning (Forman-Kay and Mittag, 2013). MSP-7 undergoes two proteolytic events where the N-terminal is cleaved after primary proteolysis (Pachebat *et al.*, 2007). The fate of the N-terminal after proteolysis was unclear until recently; Perrin *et al.* (2015) demonstrated that the cleaved N-terminal domain as a ligand for the host's P-selectin. This interaction seems to regulate disease severity because P-selectin plays a primary role in recruiting leukocytes to the injury sites during an inflammatory response (Klintman *et al.*, 2004). Therefore, the N-terminal of PvMSP-7 could change its structure upon contact with the target components and regulate pathogenicity.

In the present study, 14 novel epitopes that were immunogenic in all experimental groups under conditions of natural exposure were identified (Figure 6.7). These 14 epitopes encoded in six PvMSP-7 paralogs have shown to elicit IgG cross-reactivity (fold-changes= 3.307 – 6.311). It is noteworthy that, of the 14 novel epitopes, nine immunogenic epitopes were derived from PvMSP-7A and PvMSP-7L. The high log fold-change in two peptides EREKKNTLIKTFKKA (PvMSP-7A$_{310\text{-}315}$) and EAVDEEAEKEDTAVI (PvMSP-7L$_{139\text{-}154}$) are located in the conserved fragment of the antigens. Deletion of full-length PvMSP-7A orthologue in *P. falciparum* reported retarding the merozoites invasion of erythrocytes (Kadekoppala *et al.*, 2008). This finding highlights the functional importance of PvMSP-7A, and a subunit vaccine incorporating the PvMSP-7A could potentially lead to immune responses that reduce pathogenicity.

Sera from malaria-naïve patients served as a negative control in the study. Cross-reactivity was observed in peptide microarray of negative control to a lesser extent (Figure 6.4f). Consistently, this phenomenon was discussed in several studies

where malaria-naïve volunteers displayed cross-reactivity towards *Plasmodium* antigens (H. FELL *et al.*, 1994; Wipasa *et al.*, 2011; Zevering *et al.*, 1992). This could arise from the proliferation of immune memory cells in non-exposed malaria individuals in response to malaria peptides (Good and Saul, 1993). Likely that the differentiation of memory cells are stimulated by microorganisms or other vaccine antigens which present the similar epitopes to malaria antigens (Wipasa *et al.*, 2011). However, the present study unable to draw a conclusion about the fluorescence signals captured on the negative control (Figure 6.4f) as the clinical history of the malaria naïve individuals recruited for the study was inaccessible.

6.5 Conclusion

This study presents the first linear B-cell epitopes of PvMSP-7 in natural infection based on the high-density peptide microarray. This novel approach has identified immunodominant linear B-cell epitopes within the PvMSP-7 repertoire. Based on the evidence presented in this chapter, 14 novel PvMSP-7 peptides are universally targeted by naturally acquired IgG antibodies. Two highly immunogenic PvMSP-7 paralogs, PvMSP-7A and PvMSP-7L are promising vaccine candidates. The naturally immunogenic linear epitopes of PvMSP-7A and PvMSP-7L span along the conserved motif of PvMSP-7 should be prioritised in subunit vaccine development. Although the magnitude of IgG responses conferred by each epitope was demonstrated, the protective efficacy remains to be explored. A larger cohort of patients is required to validate the findings presented herein.

Chapter 7

General discussion

This thesis contributes significantly to the understanding of *Plasmodium vivax* merozoite surface protein 7 as a vaccine candidate. Various reverse vaccinology approaches were used to progress the use of PvMSP-7 in future vaccine design including, antigenic variation, gene expression, and identification of immunogenic epitopes spanning the entire PvMSP-7 repertoire.

In **Chapter 2**, the extensive population structure of *P. vivax* was observed between three main malaria major endemic areas in Thailand (Tak province, Ubon Ratchathani province, and Yala province). This finding is consistent with the geographic distribution of malaria parasite in Thailand and implicates in the vaccine development strategy. In **Chapter 3**, PvMSP-7 multigene family displays heterogeneous sequence variation, certain paralogs are highly polymorphic, and others are rather conserved. Most of the positive selection signals were identified in the central region of the gene owing to the functional constraint. Vaccine development should prioritise the conserved paralogs and terminal to elicit immune responses in a larger proportion of the world population. In **Chapter 4**, PvMSP-7E exhibits comparable genetic diversity to other genetic markers for *P. vivax*. Evolutionary pressures act differentially along the locus and seem to be differentially affected by predicted protein secondary structure. In **Chapter 5**, three PvMSP-7 paralogs (PvMSP-7A, -7F, and -7M) have shown to express constitutively across the developmental stages. Interestingly, certain PvMSP-7 paralogs (PvMSP-7H and -7I) demonstrated stage-specific expression in patients experienced longer infection suggesting PvMSP-7 family is developmentally regulated. Lastly, **Chapter 6** shows total IgG antibodies level to PvMSP-7 paralogs quantified using peptide microarray. In total, 14 highly immunogenic peptides were identified and belong to six PvMSP-7 paralogs. PvMSP-7A contains the highest number of immunogenic epitopes. Taking all evidence, PvMSP-7A is a plausible vaccine candidate in *P. vivax* and established a foundation for pre-clinical vaccine development.

The major barrier in the study is the genomic coverage of *P. vivax*. The accuracy of the finding is influenced by the number of reads mapped to the genome. Patients recruited in the study demonstrated low parasitaemia and only a limited volume of

venous blood collected. Moreover, the genomic DNAs collected in the study also suffered from high-level human DNA contamination. The sequencing results showed 2X to 147X genome coverage, the inconsistency in the coverage might introduce bias in the analysis. More patients should be recruited in the future study to validate the population of *P. vivax* in three malaria endemic areas in Thailand. The latest technology using selective whole-genome amplification is an alternative to achieve higher genomic coverage, thereby improve the data quality (Cowell *et al.*, 2017). Furthermore, a larger sample size could pinpoint chromosome or genes under selective pressures for drug resistance surveillance. All these approaches allow the fine scale of *P. vivax* analysis in Thailand.

As *P. vivax* is lacking an effective culture system, the transcriptional changes of PvMSP-7 were analysed based on the clinical isolates present multiple parasite life-stages. Two genes (PvMSP-7H and -7I) were seen upregulated in the patients who experienced longer patency while three others (PvMSP-7A, -7F, and -7M) were constitutively expressed across the IDC. This suggests PvMSP-7 is developmentally regulated across the IDC. A larger sample size is needed to validate this novel finding. The coverage of the RNA-seq data also demonstrated some extent of inconsistency where the coverage ranged from 1X to 79X. In general, transcripts with high expression have a higher probability to be sequenced than those lowly expressed ones. Therefore, to characterise the lowly expressed transcript, higher sequencing depth is required. In conclusion, a larger sample size and higher genome coverage are required to fine scale the findings in this thesis.

Mapping of short sequences fragments to a genome is a challenge. It is the main question to the research community aiming to uncover the novel mutations underlying diseases (Trapnell and Salzberg, 2009). The reads should be aligned without allowing large gaps in the alignment. A large gap can introduce alignment error. A more challenging problem occurs in RNA-seq where alignments can have huge gaps due to introns. Furthermore, the multigene families in *Plasmodium* species further complicate the read mapping. It is difficult to map reads to multiple repeated regions where the aligner must decide which is the true location. Currently, short-read aligners are rapidly growing and trained to map highly divergent regions. It has been suggested that using

different aligners and inspect the alignment manually represent a more sensible approach, although this step is often time-consuming (Tian *et al.*, 2016).

Variant calling is the next analysis after the alignment against the reference genome. Short reads may be mismapped to the reference genome and contribute to inaccurate SNPs especially the reads mapped to complex locations. It has a higher chance to introduce genotyping errors to paralogues or tandem repeat regions (Torkamaneh *et al.*, 2016). As such, to minimise the impact of mismatches on variant calling, the SNP catalogue can be introduced using different pipelines to achieve concordant SNP results. The mismatches of multiple reads mapped to the positions were reported to be detected by some sophisticated tools such as GATK realigner where it can filter out false SNPs (Tian *et al.*, 2016; Van der Auwera *et al.*, 2013). Therefore, it is critical for the researchers to change multiple parameters in the SNP calling pipeline to enhance the outcome.

RNA-seq was employed in the present study to identify differentially expressed genes between groups of patients. However, biases could have been introduced during the genome mapping using 150 bp paired-end reads. Recently, a group was comparing the effect arising from the length of reads in RNA-seq between 50 bp and 100 bp paired-end reads (Chhangawala *et al.*, 2015). Strikingly, the differential gene expression analysis did not yield substantially contrasting results. However, the length of reads had a significant effect on the splice junction detection (Chhangawala *et al.*, 2015). Sequencing depth and sample size are two major factors affecting the differential expression analysis. The number of differentially expressed genes are positively correlated with the sequence depth (Zhao *et al.*, 2016). As discussed previously, a higher sequencing depth and larger sample size will refine the study. All the solutions discussed above have its limitations and introduce bias in the analysis. The emerging of long sequence reads and more advanced alignment strategy would improve the precision of the findings in the study.

This thesis has presented evidence that PvMSP-7 should be considered in malaria subunit vaccine development. Much of what have observed in the study compares well with past and present *Plasmodium* vaccine candidates, many of which were comparable multigene families, for example, the widely studied *Plasmodium* interspersed repeat (pir) gene family (Cunningham *et al*., 2010) . Pir is the largest multigene family in

Plasmodium, consisting of 68 – 838 genes (Carlton *et al.*, 2008; Gardner *et al.*, 2002; Pain *et al.*, 2008). The pir gene family is found towards the end of all *P. vivax* chromosomes (Portillo *et al*., 2001). The paralogs in this family have been located on the erythrocyte surface implying a role in the invasion mechanism. The paralogs are also transcribed differentially through the IDC suggesting that distinct functions exist among paralogs. Some of the genes were expressed continuously through the erythrocytic stage (Carlton *et al.*, 2008). Similar expression pattern of pir genes was observed in *P. yoelii* where some genes only transcribed during the erythrocytic cycle (Cunningham *et al*., 2005). This transcriptional pattern is believed to reflect different functions among the paralogs. Several studies have highlighted pir gene functions other than host-cell invasions, such as signalling, trafficking, and binding to host cells (Rénia and Goh, 2016; Yam *et al.*, 2016). The transcription level pir genes in *P. yoelii* was also reported to be regulated by host immunity as shown in the mice model (Cunningham *et al*., 2005) . Consistently, PvMSP-7 paralogs have demonstrated similar transcriptional changes through the IDC as pir genes. However, the relationship between functional importance and the transcriptional changes of PvMSP-7 through the IDC still requires further investigation. The stage-specific expression of pir family in *P. vivax* found to have no distinct clustering in the phylogeny (Cunningham *et al*., 2010) . This agrees somewhat with the observations made for PvMSP-7 in Chapter 5, where the paralogs expressed constitutively (PvMSP-7A, -7F, and -7M) do not form a clade in phylogenies, but are instead paraphyletic. In contrast, PvMSP-7H and -7I expressed exclusively in patients experienced longer patency formed a cluster within the PvMSP-7 tree. This implies that the expression profiles of PvMSP-7H and -7I may be associated with a species-specific gene duplication.

Multigene families in *Plasmodium* have evolved to encode variant surface antigens (Kyes *et al.*, 2007). Each of the variant antigens typically retains a conserved, functional region, while the variation of non-conserved regions allow the parasite to evade the host immune responses (Rénia and Goh, 2016). Allelic variation is known to prime the immune responses in malaria infection (Marsh, 1992). A malaria subunit vaccine should ideally focus on the conserved domain, to confer cross-reactive immune responses (Cao *et al.*, 2016). P48/45 is a broadly studied transmission-blocking vaccine candidate (Dijk *et al*., 2001). Several studies have reported that the conserved domain of P48/45 is functional, and the disruption of the gene in *P. falciparum* and *P. berghei*

hampered the development of male gametes (Outchkourov *et al*., 2008, Dijk *et al*., 2001). Recently, conserved epitopes located towards the C-terminal of the antigen were shown to confer cross-reactive immune responses between *P. falciparum* and *P. vivax* (Cao *et al.*, 2016). A challenge experiment was conducted in a mouse model and the immune responses were quantified by ELISA (Cao *et al.*, 2016). However, protective immunity is yet to be demonstrated in populations naturally exposed to *P. falciparum* and *P. vivax*.

Likewise, merozoite surface protein 2 (MSP-2) has been investigated for its immune response in blood-stages (Beeson *et al.*, 2016). MSP-2 is expressed on the merozoite surface similar to MSP-7. Intrinsically disordered regions of MSP-2 were observed along the antigens (Morales *et al.*, 2015). The central domain of MSP-2 displays extensive antigenic variation and is flanked by conserved domains at the N- and C-termini (MacRaild *et al.*, 2015). Immune responses were evaluated in the polymorphic (Flück *et al.*, 2004) and conserved regions (Seow *et al.*, 2017). The C-terminal of MSP-2 encoded the main epitopes shown to recognised by the mouse monoclonal antibodies (Seow *et al.*, 2017). A subunit vaccine based on the polymorphic and conserved regions was shown to reduce parasitaemia by 62% (Flück *et al.*, 2004). Drawing on the experience of these past and present vaccine candidates, the conserved domains of PvMSP-7 are the starting point for its vaccine development. The N- and C-terminal of PvMSP-7 are highly conserved, and in the case of the C-terminal, this is because it interacts with MSP-1 prior to host-cell invasion (Castillo *et al.*, 2017; Kadekoppala and Holder, 2010; Kadekoppala *et al.*, 2008). Whether the C-terminal of PvMSP-7 would induce selective immune responses or protective immunity still requires further exploration. Polymorphism in the central region could still provide a benefit to an experimental vaccine, as it seems to for MSP-2, but careful evaluation is necessary. Therefore, in line with P48/45 and MSP-2 vaccine candidates, the C-terminal of PvMSP-7 should be prioritised in vaccine design.

Another important aspect of vaccine design is deciding which PvMSP-7 paralogs should be included, if not all. In **Chapters 3** and **5,** the heterogeneous patterns of sequence diversity and transcriptional profiles were observed among PvMSP-7 paralogs. Therefore, it seems that PvMSP-7 paralogs do not perform identical functions, neither are they exposed to host immunity to equal extents. The reticulocyte-binding

protein family has been considered as vaccine candidates in *P. vivax* (Han *et al.*, 2016) and *P. falciparum* (Baum *et al.*, 2009; Campeotto *et al.*, 2017). Of the eleven RBP paralogs in *P. vivax*, PvRBP1a and PvRBP1b were selected in vaccine design owing to the expression at the microneme in schizont stage (Han *et al.*, 2016). Moreover, sequence analysis of PvRBP1a and PvRBP1b displayed conservation in antigenicity and induced protective immune responses in the animal model (Han *et al.*, 2016). Based on this evidence, PvMSP-7 paralogs that show high sequence conservation should be considered in vaccine development. The precise location of PvMSP-7 during blood-stages is yet to be characterised in *P. vivax*. However, the orthologous gene of PvMSP-7A in *P. berghei* was present on the merozoite surface (Kadekoppala *et al.*, 2010). The disruption of PbMSP-7A did exhibit a significant reduction in erythrocyte invasion, suggesting the utility of this paralog in vaccine design (Kadekoppala *et al.*, 2008). Future work should focus on the localisation of all PvMSP-7 paralogs to underpin the vaccine design.

Rapid identification of novel vaccine candidates has been described using chemical peptide synthesis and serological screening (Valencia *et al.*, 2011). A total of 50 *P. vivax* orthologous genes in *P. falciparum* were chemically synthesised and evaluated for their immunogenicity using ELISA (Villard *et al.*, 2007). In the present study, a high-density peptide microarray was used to screen all immunogenic epitopes across PvMSP-7 paralogs. This state-of-art technology efficiently identified the immune responses against PvMSP-7 in natural infection. The technology has been used to identify the B-cell epitopes in *P. falciparum* Schizont Egress Antigen 1 (Nixon *et al.*, 2017). Of the five immunogenic epitopes tested, three epitopes displayed protective immune responses with a significant reduction in parasitaemia (Nixon *et al.*, 2017). The immune response was positively correlated with the protective efficacy (Nixon *et al.*, 2017). Immunogenic peptides were identified in PvMSP-7, however uneven distribution of immunogenic peptides was observed in the family. This likely stem from the functional difference of each paralog. These immunogenic peptides should be prioritised in the immunology testing, to which whether they elicit sterile protection malaria.

Although a high-density peptide microarray provides high-resolution identification of immunogenic epitopes, the conformational structure of the epitope should not be

neglected. The native state of epitope influences the antibodies binding specificity (Forsström *et al.*, 2015). A recent study conducted using polyclonal sera from patients has revealed the antibody recognised differently with the conformational structure of epitopes (Forsström *et al.*, 2015). Of the eight linear epitopes studied, antibodies recognised three conformational epitopes (Forsström *et al.*, 2015). Having said that, despite antibodies binding to the linear B-cell epitopes, the specificity of immune responses is likely to influence by the protein native state. The linear B-cell epitopes spanning across PvMSP-7 were characterised, but the nature of antibody response against the conformational structure of the epitopes warrants further investigation. This investigation can be accomplished using recombination protein strategy where the immunogenic B-cell epitopes are synthesised as recombination protein fragments. The binding intensity of antibody to protein fragments in relative to peptides will channel to the understanding of conformation-specific antibody.

The malaria-specific antibody IgG plays a pivotal role in clearing parasites and reducing the risk of malaria (Dobbs and Dent, 2016). The total IgG demonstrated a peak in uncomplicated malaria infection, where the mean concentrations of IgG antibodies were significantly higher than that of IgE antibody in uncomplicated malaria conducted in Thai patients (Perlmann *et al.*, 2000). A study has reported a positive correlation between IgG responses to the C-terminal domain of MSP-1 and subsequent reductions in parasite density (Branch *et al.*, 1998; Riley *et al.*, 1992). This evidence supports a role for IgG antibodies in mediating disease severity. In Chapter 6, the total IgG immune response was determined on a peptide microarray and found that epitopes were not evenly distributed between paralogs. Six PvMSP-7 paralogs have a higher number of epitopes and, within these, epitopes associated with the C-terminal, in a similar fashion to MSP-1 (Baldwin *et al.*, 2015). For this reason, as PvMSP-7A contains a higher number of immunogenic epitopes in the C-terminal, it should, therefore, prioritised in the vaccine development.

In Chapter 6, the total IgG antibodies response to PvMSP-7 was quantified, which could potentially reduce the parasite density during blood-stage infection, but without specifying the IgG isotypes responsible. Specific IgG subtypes do mediate protective immune responses in *Plasmodium* (Stanisic *et al*., 2009). The erythrocytic developmental stage is predominantly affected by IgG1 and IgG3 subclasses, while the

IgG2 and IgG4 act as antagonists (Aucan *et al.*, 2000; Taylor *et al.*, 1998; Weaver *et al.*, 2016). In a field study, sterile protection induced by 19 kDa C-terminal of MSP-1 was recognised by IgG1 and IgG3 antibodies (Diallo *et al.*, 2001). Furthermore, a cross-sectional study was carried out in West Africa where 178 individuals were assessed for their IgG responses to MSP-2 (Taylor *et al.*, 1998), and the antibody response was predominantly conferred by IgG1 and IgG3. IgG1 antibodies level was higher in children below the age of ten while IgG3 antibodies primarily found in adolescents and adults (Taylor *et al.*, 1998). These results suggest the subclass-specific antibodies influence the protective immune responses in malaria. For this reason, it will be necessary to characterize the cytophilic antibodies responses against PvMSP-7 to provide a complete understanding of protective efficacy.

Age-dependent immune responses were demonstrated in blood-stage antigens such as MSP-1, MSP-2, AMA-1, and 175-kDa erythrocyte binding antigen (EBA-175) (Dobaño *et al.*, 2011; Taylor *et al.*, 1998). These studies revealed a consistent pattern in IgG responses where the level seems to decrease with age in early infancy and increases again by age 2 years. Naturally acquired immunity to malaria requires uninterrupted exposure to the parasite and the responses are species- and stage-specific (Doolan *et al.*, 2009; Nhabomba *et al.*, 2014; Schüffner, 1938). In the present study, total IgG responses to PvMSP-7 were investigated in five different age groups. Totally, 14 cross-reactive immunogenic epitopes were detected in PvMSP-7, supporting the notion that these epitopes are immune targets. A difference in the number of antigenic epitopes was observed between each age group. The number of immunogenic epitopes was higher in the three groups of patients (age group 0-14, 15-29, and 30-44) while a drop in antigenic epitopes was seen in two older age groups (age group 45-59 and 60-74). The increase of immunogenic epitopes in three younger age groups is consistent with other studies because naturally acquired immunity to malaria establishes over time. To our knowledge, no study so far has investigated the total IgG responses using a large age pooled (0-74 years old). The discrepancy in the number of immunogenic epitopes between age groups is likely due to the patients' exposure to the parasite. Therefore, information about the exposure to the malaria parasite and immune responses to PvMSP-7 in each clinical patient would refine the finding.

Having discussed the population genetics, transcriptional patterns, and immunogenicity of PvMSP-7, the next section discusses possible steps to transform PvMSP-7 in vaccine development, with comparison to the broadly studied vaccine candidate in *P. falciparum*, RTS,S (Neafsey *et al.*, 2015; Olotu *et al.*, 2013; Olotu *et al.*, 2016). RTS,S is a pre-erythrocytic stage vaccine candidate that has demonstrated partial and complete protection against infection in an experimental model (Moorthy and Ballou, 2009). RTS,S is made up of 19 NANP repeats and the C-terminal of the circumsporozoite protein (CSP) fused to the hepatitis B surface antigen (Cohen *et al.*, 2010). It is hypothesized that the repeat regions will enhance its recognition of the host's immune system. It has been tested up to Phase III in the clinical setting where up to 50% of the children protected against infection upon vaccination (RTS,S, 2012) . Multiple Phase II trials have supported the safety of this vaccine in all age groups including infants and young children in sub-Saharan Africa (Gosling and von Seidlein, 2016). Moreover, the results from Phase II clinical trials also displayed protective immune responses against *P. falciparum* (Gosling and von Seidlein, 2016). The milestones achieved during the development of the RTS,S vaccine could guide the development of a vaccine PvMSP-7. In similar fashion to the CSP C-terminal domain, the conserved regions of PvMSP-7 should be prioritised in the vaccine design as suggested they are likely to elicit the protective immune response in natural infection.

Several constructs of CSP based in the central region were evaluated for their immune efficiency (Espinosa *et al.*, 2013). Some candidates were well tolerated and highly immunogenic (Espinosa *et al.*, 2013). Furthermore, regions of CSP that failed to elicit protective immune responses in Kenyan and Thai volunteers during Phase IIb trial were abandoned (Cohen *et al.*, 2010). In the present study, the cross-reactive epitopes were detected using peptide microarray. In total 14 highly immunogenic B-cell epitopes should proceed further to evaluate their ability to induce protective immunity using a mouse model similar to that used in the CSP vaccine approach (Espinosa *et al.*, 2013). The epitopes that demonstrate poor protective immunity will then be excluded. An expression system for PvMSP-7 would greatly benefit our ability to evaluate its protective effects. The appropriate expression system will produce the recombinant protein vaccines of PvMSP-7 to gain insights into its immune protection efficacy.

It will also be necessary to express the specific domains of the preferred PvMSP-7 paralogs in recombinant form. Generating recombinant proteins through bacterial-based systems is a challenge due to the formation of inclusion bodies and bacterial toxins that reduce biologically active components. However, several vaccine candidates have been successfully produced through bacterial-based systems using *Escherichia coli*, *Saccharomyces cerevisiae,* and *Pichia pastoris* (Gurkan and Ellar, 2005). They are the most effective expression hosts owning to the alcohol oxidase promoter that allows expression of foreign genes at ease. Recombinant proteins can be produced in several forms including soluble proteins, fusions antigens, lengthy artificial peptides, and self-antigen arrays on virus-like particles (Powles *et al.*, 2015). The RTS,S vaccine is based on the virus-like particle platform using the hepatitis B surface antigen (Oyarzún and Kobe, 2016). The C-terminal region of CSP is part of the RTS,S vaccine where it found to involve in the attachment to the parasite hepatocytes and demonstrated a conserved epitope (Wang *et al.*, 2009). This construction was shown to elicit significant protection in human (Kazmin *et al.*, 2017).

A universal influenza virus vaccine generated using recombinant DNA protein was approved in 2013 (Soema *et al.*, 2015). This highly effective vaccine was based on the conserved protein regions containing B-cell epitopes. Interestingly, this vaccine was shown to induce protective immunity and increased cross-reactivity against various influenza strains (Soema *et al.*, 2015). The Multimeric-001 influenza vaccine contains nine conserved epitopes which translate into a single 50-kDa synthetic protein (Atsmon *et al.*, 2012). The construction was based on the *E. coli* standard fermentation and purification approaches. The recombinant vaccine has proven to be safe and stimulate humoral and cellular immunity in patients and currently in stage III of the clinical trial (Atsmon *et al.*, 2012). Thus, it may be possible to pursue a similar strategy for PvMSP-7, combining the immunogenic and conserved regions of PvMSP-7A in a recombinant protein.

Another critical consideration will be selecting the appropriate adjuvants. All subunit vaccines in their native form must be coupled with an adjuvant. Subunit vaccines include only the protective regions of the protein. The choice of adjuvant is important to deliver the vaccine and induce durable immunity by presenting the vaccine antigens to the host immune system (Pasquale *et al.*, 2015). The RTS,S vaccine has

been formulated with the AS01 and AS02 adjuvant systems (Leroux-Roels *et al.*, 2014). The AS01 adjuvant system contains a liposome-based adjuvant, a Toll-like receptor 4 ligand, and QS-21 (Leroux-Roels *et al.*, 2014). The QS-21 acts as an immune response enhancer (Kensil *et al.*, 2006). RTS,S/AS01 induced a protective effect in young infants and children over a three to four year period (Gosling and Seidlein, 2016) . Administration of a booster dose prolonged the protection against infection (RTS, 2015) . The AS02 adjuvant system contains 3-deacylated monophosphoryl lipid and QS-21 oil-in-water emulsion (Ballou, 2009). In Phase I clinical trial, approximately 32% of the population protected against the sporozoite challenge (Ballou, 2009). Both AS01 and AS02 adjuvant systems demonstrated satisfying safety and immunogenicity, however, the AS01 has protection over 50% upon sporozoite challenge (Ockenhouse *et al.*, 2015). Having said that, a suitable adjuvant system will enhance the immune responses and protective effect. Furthermore, the Phase II clinical trial of RTS,S/AS01 in children from Tanzania and Kenya displayed fewer adverse complications (Bejon *et al.*, 2008). After observing the vaccine efficacy from the Phase II clinical trial, RTS,S/AS01 was chosen to proceed further in the Phase III clinical trial. It has been demonstrated the importance of choosing a suitable adjuvant system in the subunit vaccine. Specific adjuvant systems are likely to have high purity and reduce adverse effects, however, it might reduce the immunogenicity (Christensen, 2016).

This thesis represents a comprehensive study of PvMSP-7 using high-throughput technologies. In anticipation of further development of a PvMSP-7 subunit vaccine, there is a need to address the appropriate protein expression system, adjuvants, and epitopes that confer sterile protective responses. Based on the findings in this study, PvMSP-7A should be prioritised in subunit vaccine development because it shows greatest sequence conservation among the family members and constitutive expression during the IDC. In addition, six immunodominant epitopes along the PvMSP-7A displayed cross-reactivity in all clinical isolates. Five cross-reactive epitopes were identified in the N- and C-termini, which, based on previous research, is likely to modulate disease severity and impair parasite invasion. Therefore, PvMSP-7A is a promising vaccine candidate and should be developed further as a malaria vaccine candidate.

References

1 Abreha, T., Hwang, J., Thriemer, K., Tadesse, Y., Girma, S., Melaku, Z., Assef, A., Kassa, M., Chatfield, M.D., and Landman, K.Z. (2017). Comparison of artemether-lumefantrine and chloroquine with and without primaquine for the treatment of Plasmodium vivax infection in Ethiopia: A randomized controlled trial. PLoS medicine *14*, e1002299.

2 Absalon, S., Robbins, J.A., and Dvorin, J.D. (2016). An essential malaria protein defines the architecture of blood-stage and transmission-stage parasites. Nature communications *7*, 11449.

3 Adda, C.G., Murphy, V.J., Sunde, M., Waddington, L.J., Schloegel, J., Talbo, G.H., Vingas, K., Kienzle, V., Masciantonio, R., and Howlett, G.J. (2009). Plasmodium falciparum merozoite surface protein 2 is unstructured and forms amyloid-like fibrils. Molecular and biochemical parasitology *166*, 159-171.

4 Alecrim, M.d.G.C., Alecrim, W., and Macêdo, V. (1999). Plasmodium vivax resistance to chloroquine (R2) and mefloquine (R3) in Brazilian Amazon region. Revista da Sociedade Brasileira de Medicina Tropical *32*, 67-68.

5 Alexander, D.H., Novembre, J., and Lange, K. (2013). Admixture 1.23 Software Manual.

6 Aly, A.S., Vaughan, A.M., and Kappe, S.H. (2009). Malaria parasite development in the mosquito and infection of the mammalian host. Annual review of microbiology *63*, 195-221.

7 Amaratunga, C., Lopera-Mesa, T.M., Brittain, N.J., Cholera, R., Arie, T., Fujioka, H., Keefer, J.R., and Fairhurst, R.M. (2011). A role for fetal hemoglobin and maternal immune IgG in infant resistance to Plasmodium falciparum malaria. PloS one *6*, e14798.

8 Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

9 Andrews, T.D., and Gojobori, T. (2004). Strong positive selection and recombination drive the antigenic variation of the PilE protein of the human pathogen Neisseria meningitidis. Genetics *166*, 25-32.

10 Anstey, N.M., Douglas, N.M., Poespoprodjo, J.R., and Price, R.N. (2012). *Plasmodium vivax:* clinical spectrum, risk factors and pathogenesis. Adv Parasitol *80*, 151-201.

11 Aponte, J.J., Menendez, C., Schellenberg, D., Kahigwa, E., Mshinda, H., Vountasou, P., Tanner, M., and Alonso, P.L. (2007). Age interactions in the development of naturally acquired immunity to Plasmodium falciparum and its clinical presentation. PLoS medicine *4*, e242.

12 Arévalo-Herrera, M., Lopez-Perez, M., Dotsey, E., Jain, A., Rubiano, K., Felgner, P.L., Davies, D.H., and Herrera, S. (2016). Antibody profiling in naïve and semi-immune individuals experimentally challenged with Plasmodium vivax sporozoites. PLoS neglected tropical diseases *10*, e0004563.

13 Arévalo-Herrera, M., Roggero, M., Gonzalez, J., Vergara, J., Corradin, G., Lopez, J., and Herrera, S. (1998). Mapping and comparison of the B-cell epitopes recognized on the Plasmodium vivax circumsporozoite protein by immune Colombians and immunized Aotus monkeys. Annals of Tropical Medicine & Parasitology *92*, 539-551.

14 Arnott, A., Wapling, J., Mueller, I., Ramsland, P.A., Siba, P.M., Reeder, J.C., and Barry, A.E. (2014). Distinct patterns of diversity, population structure and evolution in the AMA1 genes of sympatric Plasmodium falciparum and Plasmodium vivax populations of Papua New Guinea from an area of similarly high transmission. Malaria journal *13*, 233.

15 Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., and Eppig, J.T. (2000). Gene Ontology: tool for the unification of biology. Nature genetics *25*, 25.

16 Atsmon, J., Kate-Ilovitz, E., Shaikevich, D., Singer, Y., Volokhov, I., Haim, K.Y., and Ben-Yedidia, T. (2012). Safety and immunogenicity of multimeric-001—a novel universal influenza vaccine. Journal of clinical immunology *32*, 595-603.

17 Auburn, S., Böhme, U., Steinbiss, S., Trimarsanto, H., Hostetler, J., Sanders, M., Gao, Q., Nosten, F., Newbold, C.I., and Berriman, M. (2016). A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. Wellcome open research *1*.

18 Auburn, S., Marfurt, J., Maslen, G., Campino, S., Rubio, V.R., Manske, M., MacHunter, B., Kenangalem, E., Noviyanti, R., and Trianty, L. (2013). Effective preparation of Plasmodium vivax field isolates for high-throughput whole genome sequencing. PLoS One *8*, e53160.

19 Aucan, C., Traoré, Y., Tall, F., Nacro, B., Traoré-Leroux, T., Fumoux, F., and Rihet, P. (2000). High immunoglobulin G2 (IgG2) and low IgG4 levels are associated with human resistance to Plasmodium falciparum malaria. Infection and immunity *68*, 1252-1258.

20 Aurrecoechea, C., Brestelli, J., Brunk, B.P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., and Harb, O.S. (2008). PlasmoDB: a functional genomic database for malaria parasites. Nucleic acids research *37*, D539-D543.

21 Baird, J.K. (2013). Evidence and implications of mortality associated with acute Plasmodium vivax malaria. Clinical microbiology reviews *26*, 36-57.

22 Baird, J.K. (1995). Host age as a determinant of naturally acquired immunity to Plasmodium falciparum.

23 Baird, J.K., Basri, H., Bangs, M.J., Subianto, B., Patchen, L.C., and Hoffman, S.L. (1991a). Resistance to chloroquine by Plasmodium vivax in Irian Jaya, Indonesia. The American journal of tropical medicine and hygiene *44*, 547-552.

24 Baird, J.K., Jones, T.R., Danudirgo, E.W., Annis, B.A., Bangs, M.J., Basri, P.H., and Masbar, S. (1991b). Age-dependent acquired protection against Plasmodium falciparum in people having two years exposure to hyperendemic malaria. The American journal of tropical medicine and hygiene *45*, 65-76.

25 Baldwin, M.R., Li, X., Hanada, T., Liu, S.-C., and Chishti, A.H. (2015). Merozoite surface protein 1 recognition of host glycophorin A mediates malaria parasite invasion of red blood cells. Blood, blood-2014-2011-611707.

26 Ballou, W. (2009). The development of the RTS, S malaria vaccine candidate: challenges and lessons. Parasite immunology *31*, 492-500.

27 Bang, G., Prieur, E., Roussilhon, C., and Druilhe, P. (2011). Pre-clinical assessment of novel multivalent MSP3 malaria vaccine constructs. PLoS One *6*, e28165.

28 Barry, A.E., and Arnott, A. (2014). Strategies for designing and monitoring malaria vaccines targeting diverse antigens. Frontiers in immunology *5*, 359.

29 Baum, J., Chen, L., Healer, J., Lopaticki, S., Boyle, M., Triglia, T., Ehlgen, F., Ralph, S.A., Beeson, J.G., and Cowman, A.F. (2009). Reticulocyte-binding protein homologue 5–an essential adhesin involved in invasion of human erythrocytes by Plasmodium falciparum. International journal for parasitology *39*, 371-380.

30 Beeson, J.G., Drew, D.R., Boyle, M.J., Feng, G., Fowkes, F.J., and Richards, J.S. (2016). Merozoite surface proteins in red blood cell invasion, immunity and vaccines against malaria. FEMS microbiology reviews *40*, 343-372.

31 Bejon, P., Lusingu, J., Olotu, A., Leach, A., Lievens, M., Vekemans, J., Mshamu, S., Lang, T., Gould, J., and Dubois, M.-C. (2008). Efficacy of RTS, S/AS01E vaccine against malaria in children 5 to 17 months of age. New England Journal of Medicine *359*, 2521-2532.

32 Bennett, J.W., Yadava, A., Tosh, D., Sattabongkot, J., Komisar, J., Ware, L.A., McCarthy, W.F., Cowden, J.J., Regules, J., and Spring, M.D. (2016). Phase 1/2a trial of Plasmodium vivax malaria vaccine candidate VMP001/AS01B in malaria-naive adults: safety, immunogenicity, and efficacy. PLoS neglected tropical diseases *10*, e0004423.

33 Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research *27*, 573-580.

34 Bhumiratana, A., Intarapuk, A., Sorosjinda-Nunthawarasilp, P., Maneekan, P., and Koyadun, S. (2013). Border malaria associated with multidrug resistance

on Thailand-Myanmar and Thailand-Cambodia borders: transmission dynamic, vulnerability, and surveillance. BioMed research international *2013*.

35 Bi, Y., Yu, W., Hu, W., Lin, H., Guo, Y., Zhou, X.-N., and Tong, S. (2013). Impact of climate variability on Plasmodium vivax and Plasmodium falciparum malaria in Yunnan Province, China. Parasites & vectors *6*, 357.

36 Bijker, E.M., Bastiaens, G.J., Teirlinck, A.C., van Gemert, G.-J., Graumans, W., van de Vegte-Bolmer, M., Siebelink-Stoter, R., Arens, T., Teelen, K., and Nahrendorf, W. (2013). Protection against malaria after immunization by chloroquine prophylaxis and sporozoites is mediated by preerythrocytic immunity. Proceedings of the National Academy of Sciences, 201220360.

37 Bitencourt, A.R., Vicentin, E.C., Jimenez, M.C., Ricci, R., Leite, J.A., Costa, F.T., Ferreira, L.C., Russell, B., Nosten, F., and Rénia, L. (2013). Antigenicity and immunogenicity of Plasmodium vivax merozoite surface protein-3. PLoS One *8*, e56061.

38 Blackman, M.J., Scott-Finnigan, T.J., Shai, S., and Holder, A.A. (1994). Antibodies inhibit the protease-mediated processing of a malaria merozoite surface protein. Journal of Experimental Medicine *180*, 389-393.

39 Blower, S., Koelle, K., Kirschner, D.E., and Mills, J. (2001). Live attenuated HIV vaccines: predicting the tradeoff between efficacy and safety. Proceedings of the National Academy of Sciences *98*, 3618-3623.

40 Boes, A., Spiegel, H., Voepel, N., Edgue, G., Beiss, V., Kapelski, S., Fendel, R., Scheuermayer, M., Pradel, G., and Bolscher, J.M. (2015). Analysis of a multi-component multi-stage malaria vaccine candidate—tackling the cocktail challenge. PloS one *10*, e0131456.

41 Boström, S., Giusti, P., Arama, C., Persson, J.-O., Dara, V., Traore, B., Dolo, A., Doumbo, O., and Troye-Blomberg, M. (2012). Changes in the levels of cytokines, chemokines and malaria-specific antibodies in response to Plasmodium falciparum infection in children living in sympatry in Mali. Malaria Journal *11*, 109.

42 Bouharoun-Tayoun, H., Attanath, P., Sabchareon, A., Chongsuphajaisiddhi, T., and Druilhe, P. (1990). Antibodies that protect humans against Plasmodium falciparum blood stages do not on their own inhibit parasite growth and invasion in vitro, but act in cooperation with monocytes. Journal of Experimental Medicine *172*, 1633-1641.

43 Bournazos, S., and Ravetch, J.V. (2017). Attenuated vaccines for augmented immunity. Cell host & microbe *21*, 314-315.

44 Boyle, M.J., Wilson, D.W., and Beeson, J.G. (2013). New approaches to studying Plasmodium falciparum merozoite invasion and insights into invasion biology. International journal for parasitology *43*, 1-10.

45 Bozdech, Z., Mok, S., Hu, G., Imwong, M., Jaidee, A., Russell, B., Ginsburg, H., Nosten, F., Day, N.P., and White, N.J. (2008). The transcriptome of Plasmodium vivax reveals divergence and diversity of transcriptional regulation in malaria parasites. Proceedings of the National Academy of Sciences *105*, 16290-16295.

46 Brahimi, K., Badell, E., Sauzet, J.-P., BenMohamed, L., Daubersies, P., Guérin-Marchand, C., Snounou, G., and Druilhe, P. (2001). Human Antibodies against Plasmodium falciparumLiver-Stage Antigen 3 Cross-React with Plasmodium yoelii Preerythrocytic-Stage Epitopes and Inhibit Sporozoite Invasion In Vitro and In Vivo. Infection and immunity *69*, 3845-3852.

47 Branch, O.H., Udhayakumar, V., Hightower, A.W., Oloo, A.J., Hawley, W.A., Nahlen, B.L., Bloland, P.B., Kaslow, D.C., and Lal, A.A. (1998). A longitudinal investigation of IgG and IgM antibody responses to the merozoite surface protein-1 19-kiloDalton domain of Plasmodium falciparum in pregnant women and infants: associations with febrile illness, parasitemia, and anemia. The American journal of tropical medicine and hygiene *58*, 211-219.

48 Bueno, L.L., Lobo, F.P., Morais, C.G., Mourão, L.C., de Ávila, R.A.M., Soares, I.S., Fontes, C.J., Lacerda, M.V., Olórtegui, C.C., and Bartholomeu, D.C. (2011). Identification of a highly antigenic linear B cell epitope within Plasmodium vivax apical membrane antigen 1 (AMA-1). PloS one *6*, e21289.

49 Butler, N.S., Vaughan, A.M., Harty, J.T., and Kappe, S.H. (2012). Whole parasite vaccination approaches for prevention of malaria infection. Trends in immunology *33*, 247-254.

50 Cahoon-Young, B., Chandler, A., Livermore, T., Gaudino, J., and Benjamin, R. (1989). Sensitivity and specificity of pooled versus individual sera in a human immunodeficiency virus antibody prevalence study. Journal of Clinical Microbiology *27*, 1893-1895.

51 Campeotto, I., Goldenzweig, A., Davey, J., Barfod, L., Marshall, J.M., Silk, S.E., Wright, K.E., Draper, S.J., Higgins, M.K., and Fleishman, S.J. (2017). One-step design of a stable variant of the malaria invasion protein RH5 for use as a vaccine immunogen. Proceedings of the National Academy of Sciences *114*, 998-1002.

52 Cao, Y., Bansal, G.P., Merino, K., and Kumar, N. (2016). Immunological Cross-Reactivity between Malaria Vaccine Target Antigen P48/45 in Plasmodium vivax and P. falciparum and Cross–Boosting of Immune Responses. PloS one *11*, e0158212.

53 Carlton, J., Silva, J., and Hall, N. (2004). The genome of model malaria parasites, and comparative genomics. Malaria parasites: genome and molecular biology, 33-63.

54 Carlton, J.M., Adams, J.H., Silva, J.C., Bidwell, S.L., Lorenzi, H., Caler, E., Crabtree, J., Angiuoli, S.V., Merino, E.F., and Amedeo, P. (2008). Comparative

genomics of the neglected human malaria parasite Plasmodium vivax. Nature *455*, 757.

55 Carvalho, L.J., Daniel-Ribeiro, C.T., and Goto, H. (2002). Malaria vaccine: candidate antigens, mechanisms, constraints and prospects. Scand J Immunol *56*, 327-343.

56 Castillo, A.I., Pacheco, M.A., and Escalante, A.A. (2017). Evolution of the merozoite surface protein 7 (msp7) family in Plasmodium vivax and P. falciparum: A comparative approach. Infection, Genetics and Evolution *50*, 7-19.

57 Cavanagh, D.R., Elhassan, I.M., Roper, C., Robinson, V.J., Giha, H., Holder, A.A., Hviid, L., Theander, T.G., Arnot, D.E., and McBride, J.S. (1998). A longitudinal study of type-specific antibody responses to Plasmodium falciparum merozoite surface protein-1 in an area of unstable malaria in Sudan. The Journal of Immunology *161*, 347-359.

58 Chandramohanadas, R., Russell, B., Liew, K., Yau, Y.H., Chong, A., Liu, M., Gunalan, K., Raman, R., Renia, L., and Nosten, F. (2014). Small molecule targeting malaria merozoite surface protein-1 (MSP-1) prevents host invasion of divergent plasmodial species. The Journal of infectious diseases *210*, 1616-1626.

59 Chaurio, R.A., Pacheco, M.A., Cornejo, O.E., Durrego, E., Stanley Jr, C.E., Castillo, A.I., Herrera, S., and Escalante, A.A. (2016). Evolution of the transmission-blocking vaccine candidates Pvs28 and Pvs25 in Plasmodium vivax: geographic differentiation and evidence of positive selection. PLoS neglected tropical diseases *10*, e0004786.

60 Cheesman, S., O'Mahony, E., Pattaradilokrat, S., Degnan, K., Knott, S., and Carter, R. (2010). A single parasite gene determines strain-specific protective immunity against malaria: the role of the merozoite surface protein I. International journal for parasitology *40*, 951-961.

61 Chen, S.-B., Wang, Y., Kassegne, K., Xu, B., Shen, H.-M., and Chen, J.-H. (2017). Whole-genome sequencing of a Plasmodium vivax clinical isolate exhibits geographical characteristics and high genetic variation in China-Myanmar border area. BMC genomics *18*, 131.

62 Chenet, S.M., Tapia, L.L., Escalante, A.A., Durand, S., Lucas, C., and Bacon, D.J. (2012). Genetic diversity and population structure of genes encoding vaccine candidate antigens of Plasmodium vivax. Malaria journal *11*, 68.

63 Cheng, C.W., Putaporntip, C., and Jongwutiwes, S. (2018). Polymorphism in merozoite surface protein-7E of Plasmodium vivax in Thailand: Natural selection related to protein secondary structure. PloS one *13*, e0196765.

64 Cheng, Y., Ito, D., Sattabongkot, J., Lim, C.S., Kong, D.-H., Ha, K.-S., Wang, B., Tsuboi, T., and Han, E.-T. (2013). Serological responses to a soluble

recombinant chimeric Plasmodium vivax circumsporozoite protein in VK210 and VK247 population. Malaria journal *12*, 323.

65 Chhangawala, S., Rudy, G., Mason, C.E., and Rosenfeld, J.A. (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome biology *16*, 131.

66 Chitnis, C.E., Mukherjee, P., Mehta, S., Yazdani, S.S., Dhawan, S., Shakri, A.R., Bharadwaj, R., Gupta, P.K., Hans, D., and Mazumdar, S. (2015). Phase I clinical trial of a recombinant blood stage vaccine candidate for Plasmodium falciparum malaria based on MSP1 and EBA175. PloS one *10*, e0117820.

67 Chootong, P., Ntumngia, F.B., VanBuskirk, K.M., Xainli, J., Cole-Tobian, J.L., Campbell, C.O., Fraser, T.S., King, C.L., and Adams, J.H. (2010). Mapping epitopes of the Plasmodium vivax Duffy binding protein with naturally acquired inhibitory antibodies. Infection and immunity *78*, 1089-1095.

68 Chougnet, C., Deloron, P., Lepers, J.P., Rason, M.D., Savel, J., and Coulanges, P. (1990). Longitudinal study of the cellular response to Pf155/RESA and circumsporozoite protein in Madagascar. Immunology letters *25*, 231-235.

69 Chowdhury, D.R., Angov, E., Kariuki, T., and Kumar, N. (2009). A potent malaria transmission blocking vaccine based on codon harmonized full length Pfs48/45 expressed in Escherichia coli. PloS one *4*, e6352.

70 Christensen, D. (2016). Vaccine adjuvants: Why and how. Human vaccines & immunotherapeutics *12*, 2709-2711.

71 Chu, C.S., and White, N.J. (2016). Management of relapsing Plasmodium vivax malaria. Expert review of anti-infective therapy *14*, 885-900.

72 Coelho, C.H., Doritchamou, J.Y.A., Zaidi, I., and Duffy, P.E. (2017). Advances in malaria vaccine development: report from the 2017 malaria vaccine symposium (Nature Publishing Group).

73 Coffman, R.L., Sher, A., and Seder, R.A. (2010). Vaccine adjuvants: putting innate immunity to work. Immunity *33*, 492-503.

74 Cohen, J., Nussenzweig, V., Vekemans, J., and Leach, A. (2010). From the circumsporozoite protein to the RTS, S/AS candidate vaccine. Human vaccines *6*, 90-96.

75 Cohen, S. (1979). Review lecture-Immunity to malaria. Proc R Soc Lond B *203*, 323-345.

76 Cohen, S., McGregor, I., and Carrington, S. (1961). Gamma-globulin and acquired immunity to human malaria. Nature *192*, 733-737.

77 Collins, C., and Didelot, X. (2018). A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS computational biology *14*, e1005958.

78 Combes, V., Rosenkranz, A.R., Redard, M., Pizzolato, G., Lepidi, H., Vestweber, D., Mayadas, T.N., and Grau, G.E. (2004). Pathogenic role of P-selectin in experimental cerebral malaria: importance of the endothelial compartment. The American journal of pathology *164*, 781-786.

79 Consortium, A.g.G. (2017). Genetic diversity of the African malaria vector Anopheles gambiae. Nature *552*, 96.

80 Cook, D.E., and Andersen, E.C. (2017). VCF-kit: assorted utilities for the variant call format. Bioinformatics *33*, 1581-1582.

81 Cortés, A., Mellombo, M., Masciantonio, R., Murphy, V.J., Reeder, J.C., and Anders, R.F. (2005). Allele specificity of naturally acquired antibody responses against Plasmodium falciparum apical membrane antigen 1. Infection and immunity *73*, 422-430.

82 Cowan, G.J., Creasey, A.M., Dhanasarnsombut, K., Thomas, A.W., Remarque, E.J., and Cavanagh, D.R. (2011). A malaria vaccine based on the polymorphic block 2 region of MSP-1 that elicits a broad serotype-spanning immune response. PLoS One *6*, e26616.

83 Cowell, A.N., Loy, D.E., Sundararaman, S.A., Valdivia, H., Fisch, K., Lescano, A.G., Baldeviano, G.C., Durand, S., Gerbasi, V., and Sutherland, C.J. (2017). Selective whole-genome amplification is a robust method that enables scalable whole-genome sequencing of Plasmodium vivax from unprocessed clinical samples. MBio *8*, e02257-02216.

84 Cowman, A.F., Baldi, D.L., Duraisingh, M., Healer, J., Mills, K.E., O'Donnell, R.A., Thompson, J., Triglia, T., Wickham, M.E., and Crabb, B.S. (2002). Functional analysis of *Plasmodium falciparum* merozoite antigens: implications for erythrocyte invasion and vaccine development. Philos Trans R Soc Lond B Biol Sci *357*, 25-33.

85 Cowman, A.F., Berry, D., and Baum, J. (2012). The cellular and molecular basis for malaria parasite invasion of the human red blood cell. J Cell Biol *198*, 961-971.

86 Cowman, A.F., and Crabb, B.S. (2006). Invasion of red blood cells by malaria parasites. Cell *124*, 755-766.

87 Cox, F.E. (2010). History of the discovery of the malaria parasites and their vectors. Parasites & vectors *3*, 5.

88 Cui, L., Mascorro, C.N., Fan, Q., Rzomp, K.A., Khuntirat, B., Zhou, G., Chen, H., Yan, G., and Sattabongkot, J. (2003). Genetic diversity and multiple

infections of *Plasmodium vivax* malaria in Western Thailand. Am J Trop Med Hyg *68*, 613-619.

89 Cunningham, D., Lawton, J., Jarra, W., Preiser, P., and Langhorne, J. (2010). The pir multigene family of Plasmodium: antigenic variation and beyond. Molecular and biochemical parasitology *170*, 65-73.

90 Cunningham, D.A., Jarra, W., Koernig, S., Fonager, J., Fernandez-Reyes, D., Blythe, J.E., Waller, C., Preiser, P.R., and Langhorne, J. (2005). Host immunity modulates transcriptional changes in a multigene family (yir) of rodent malaria. Molecular microbiology *58*, 636-647.

91 Curtidor, H., Patiño, L.C., Arévalo-Pinzón, G., Vanegas, M., Patarroyo, M.E., and Patarroyo, M.A. (2014). Plasmodium falciparum rhoptry neck protein 5 peptides bind to human red blood cells and inhibit parasite invasion. Peptides *53*, 210-217.

92 De Mendonça, V.R., Goncalves, M.S., and Barral-Netto, M. (2012). The host genetic diversity in malaria infection. Journal of tropical medicine *2012*.

93 de Monerri, N.C.S., Flynn, H.R., Campos, M.G., Hackett, F., Koussis, K., Withers-Martinez, C., Skehel, J.M., and Blackman, M.J. (2011). Global identification of multiple substrates for PfSUB1, an essential malarial processing protease. Infection and immunity.

94 del Portillo, H.A., Fernandez-Becerra, C., Bowman, S., Oliver, K., Preuss, M., Sanchez, C.P., Schneider, N.K., Villalobos, J.M., Rajandream, M.-A., and Harris, D. (2001). A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax. Nature *410*, 839.

95 DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., and Hanna, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics *43*, 491.

96 Diallo, T.O., Spiegel, A., Diouf, A., Perraut, R., Kaslow, D.C., and Garraud, O. (2001). IgG1/IgG3 antibody responses to various analogs of recombinant ypfmsp119--a study in immune adults living in areas of Plasmodium falciparum transmission. The American journal of tropical medicine and hygiene *64*, 204-206.

97 Didierlaurent, A.M., Laupèze, B., Di Pasquale, A., Hergli, N., Collignon, C., and Garçon, N. (2017). Adjuvant system AS01: helping to overcome the challenges of modern vaccines. Expert review of vaccines *16*, 55-63.

98 Diggs, C., Hines, F., and Wellde, B. (1995). Plasmodium falciparum: passive immunization of Aotus lemurinus griseimembra with immune serum. Experimental parasitology *80*, 291-296.

99 Disease, B.o.V.B. (2015). Malaria Situation. In Annual Report, Chinanonwes, ed. (Bangkok, Thailand: Aksorn Graphic and Design), pp. 14-18.

100 Dobaño, C., Quelhas, D., Quintó, L., Puyol, L., Serra-Casas, E., Mayor, A., Nhampossa, T., Macete, E., Aide, P., and Mandomando, I. (2011). Age-dependent IgG subclass responses to Plasmodium falciparum EBA-175 are differentially associated with incidence of malaria in Mozambican children. Clinical and Vaccine Immunology, CVI. 05523-05511.

101 Dobbs, K.R., and Dent, A.E. (2016). Plasmodium malaria and antimalarial antibodies in the first year of life. Parasitology *143*, 129-138.

102 Doll, K.L., and Harty, J.T. (2014). Correlates of protective immunity following whole sporozoite vaccination against malaria. Immunologic research *59*, 166-176.

103 Doolan, D.L., Dobaño, C., and Baird, J.K. (2009). Acquired immunity to malaria. Clinical microbiology reviews *22*, 13-36.

104 Dosztányi, Z., Mészáros, B., and Simon, I. (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics *25*, 2745-2746.

105 Douglas, N.M., Pontororing, G.J., Lampah, D.A., Yeo, T.W., Kenangalem, E., Poespoprodjo, J.R., Ralph, A.P., Bangs, M.J., Sugiarto, P., and Anstey, N.M. (2014). Mortality attributable to Plasmodium vivax malaria: a clinical audit from Papua, Indonesia. BMC medicine *12*, 217.

106 Draper, S.J., Angov, E., Horii, T., Miller, L.H., Srinivasan, P., Theisen, M., and Biswas, S. (2015). Recent advances in recombinant protein-based malaria vaccines. Vaccine *33*, 7433-7443.

107 Draper, S.J., Goodman, A.L., Biswas, S., Forbes, E.K., Moore, A.C., Gilbert, S.C., and Hill, A.V. (2009). Recombinant viral vaccines expressing merozoite surface protein-1 induce antibody-and T cell-mediated multistage protection against malaria. Cell host & microbe *5*, 95-105.

108 Driss, A., Hibbert, J.M., Wilson, N.O., Iqbal, S.A., Adamkiewicz, T.V., and Stiles, J.K. (2011). Genetic polymorphisms linked to susceptibility to malaria. Malaria journal *10*, 271.

109 Drozdetskiy, A., Cole, C., Procter, J., and Barton, G.J. (2015). JPred4: a protein secondary structure prediction server. Nucleic acids research *43*, W389-W394.

110 Duffy, P.E., and Kaslow, D.C. (1997). A novel malaria protein, Pfs28, and Pfs25 are genetically linked and synergistic as falciparum malaria transmission-blocking vaccines. Infection and immunity *65*, 1109-1113.

111 Duffy, P.E., Sahu, T., Akue, A., Milman, N., and Anderson, C. (2012). Pre-erythrocytic malaria vaccines: identifying the targets. Expert review of vaccines *11*, 1261-1280.

112 Dyson, H.J., and Wright, P.E. (2002). Coupling of folding and binding for unstructured proteins. Current opinion in structural biology *12*, 54-60.

113 Eksi, S., Czesny, B., Van Gemert, G.J., Sauerwein, R.W., Eling, W., and Williamson, K.C. (2006). Malaria transmission-blocking antigen, Pfs230, mediates human red blood cell binding to exflagellating male parasites and oocyst production. Molecular microbiology *61*, 991-998.

114 EL-Manzalawy, Y., Dobbs, D., and Honavar, V. (2008). Predicting linear B-cell epitopes using string kernels. Journal of Molecular Recognition: An Interdisciplinary Journal *21*, 243-255.

115 Escalante, A.A., Grebert, H.M., Chaiyaroj, S.C., Magris, M., Biswas, S., Nahlen, B.L., and Lal, A.A. (2001). Polymorphism in the gene encoding the apical membrane antigen-1 (AMA-1) of Plasmodium falciparum. X. Asembo Bay Cohort Project. Molecular and biochemical parasitology *113*, 279-287.

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

117 Espinosa, D.A., Yadava, A., Angov, E., Maurizio, P.L., Ockenhouse, C.F., and Zavala, F. (2013). Development of a chimeric Plasmodium berghei strain expressing the repeat region of the P. vivax circumsporozoite protein for in vivo evaluation of vaccine efficacy. Infection and immunity *81*, 2882-2887.

118 Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047-3048.

119 Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary bioinformatics *1*, 117693430500100003.

120 Facer, C.A., and Theodoridou, A. (1994). Elevated plasma levels of P-selectin (GMP-140/CD62P) in patients with Plasmodium falciparum malaria. Microbiology and immunology *38*, 727-731.

121 Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution *17*, 368-376.

122 Feng, Z.-P., Zhang, X., Han, P., Arora, N., Anders, R.F., and Norton, R.S. (2006). Abundance of intrinsically unstructured proteins in P. falciparum and other apicomplexan parasite proteomes. Molecular and biochemical parasitology *150*, 256-267.

123 Ferraguti, M., Martínez-de La Puente, J., Roiz, D., Ruiz, S., Soriguer, R., and Figuerola, J. (2016). Effects of landscape anthropization on mosquito community composition and abundance. Scientific reports *6*, 29002.

124 Ferreira, M.U., da Silva Nunes, M., and Wunderlich, G. (2004). Antigenic diversity and immune evasion by malaria parasites. Clinical and diagnostic laboratory immunology *11*, 987-995.

125 Ferreira, M.U., Karunaweera, N.D., da Silva-Nunes, M., Da Silva, N.S., Wirth, D.F., and Hartl, D.L. (2007). Population structure and transmission dynamics of Plasmodium vivax in rural Amazonia. The Journal of infectious diseases *195*, 1218-1226.

126 Flück, C., Smith, T., Beck, H.-P., Irion, A., Betuela, I., Alpers, M.P., Anders, R., Saul, A., Genton, B., and Felger, I. (2004). Strain-specific humoral response to a polymorphic malaria vaccine. Infection and Immunity *72*, 6300-6305.

127 Fontaine, A., Pophillat, M., Bourdon, S., Villard, C., Belghazi, M., Fourquet, P., Durand, C., Lefranc, D., Rogier, C., and Fusai, T. (2010). Specific antibody responses against membrane proteins of erythrocytes infected by Plasmodium falciparum of individuals briefly exposed to malaria. Malaria journal *9*, 276.

128 Foquet, L., Hermsen, C.C., van Gemert, G.-J., Van Braeckel, E., Weening, K.E., Sauerwein, R., Meuleman, P., and Leroux-Roels, G. (2014). Vaccine-induced monoclonal antibodies targeting circumsporozoite protein prevent Plasmodium falciparum infection. The Journal of clinical investigation *124*, 140-144.

129 Forman-Kay, J.D., and Mittag, T. (2013). From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. Structure *21*, 1492-1499.

130 Forsström, B., Axnäs, B.B., Rockberg, J., Danielsson, H., Bohlin, A., and Uhlen, M. (2015). Dissecting antibodies with regards to linear and conformational epitopes. PloS one *10*, e0121673.

131 Franks, S., Baton, L., Tetteh, K., Tongren, E., Dewin, D., Akanmori, B.D., Koram, K.A., Ranford-Cartwright, L., and Riley, E.M. (2003). Genetic diversity and antigenic polymorphism in Plasmodium falciparum: extensive serological cross-reactivity between allelic variants of merozoite surface protein 2. Infection and immunity *71*, 3485-3495.

132 Galinski, M.R., Xu, M., and Barnwell, J.W. (2000). Plasmodium vivax reticulocyte binding protein-2 (PvRBP-2) shares structural features with PvRBP-1 and the Plasmodium yoelii 235 kDa rhoptry protein family. Molecular and biochemical parasitology *108*, 257-262.

133 Gantt, S., Persson, C., Rose, K., Birkett, A.J., Abagyan, R., and Nussenzweig, V. (2000). Antibodies against thrombospondin-related anonymous protein do not inhibit Plasmodium sporozoite infectivity in vivo. Infection and immunity *68*, 3667-3673.

134 Garcia, Y., Puentes, A., Curtidor, H., Cifuentes, G., Reyes, C., Barreto, J., Moreno, A., and Patarroyo, M.E. (2007). Identifying merozoite surface protein 4 and merozoite surface protein 7 Plasmodium falciparum protein family members specifically binding to human erythrocytes suggests a new malarial parasite-redundant survival mechanism. Journal of medicinal chemistry *50*, 5665-5675.

135 Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., and Bowman, S. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. Nature *419*, 498.

136 Garzón-Ospina, D., Buitrago, S.P., Ramos, A.E., and Patarroyo, M.A. (2018). Identifying potential Plasmodium vivax sporozoite stage vaccine candidates: An analysis of genetic diversity and natural selection. Frontiers in genetics *9*, 10.

137 Garzón-Ospina, D., Cadavid, L.F., and Patarroyo, M.A. (2010). Differential expansion of the merozoite surface protein (msp)-7 gene family in Plasmodium species under a birth-and-death model of evolution. Molecular phylogenetics and evolution *55*, 399-408.

138 Garzón-Ospina, D., Forero-Rodríguez, J., and Patarroyo, M.A. (2016). Evidence of functional divergence in MSP7 paralogous proteins: a molecular-evolutionary and phylogenetic analysis. BMC evolutionary biology *16*, 256.

139 Garzón-Ospina, D., Forero-Rodríguez, J., and Patarroyo, M.A. (2014). Heterogeneous genetic diversity pattern in Plasmodium vivax genes encoding merozoite surface proteins (MSP)-7E,− 7F and-7L. Malaria journal *13*, 495.

140 Garzón-Ospina, D., López, C., Forero-Rodríguez, J., and Patarroyo, M.A. (2012). Genetic diversity and selection in three Plasmodium vivax merozoite surface protein 7 (Pvmsp-7) genes in a Colombian population. PloS one *7*, e45962.

141 Garzón-Ospina, D., Romero-Murillo, L., Tobón, L.F., and Patarroyo, M.A. (2011). Low genetic polymorphism of merozoite surface proteins 7 and 10 in Colombian Plasmodium vivax isolates. Infection, Genetics and Evolution *11*, 528-531.

142 Gething, P.W., Elyazar, I.R., Moyes, C.L., Smith, D.L., Battle, K.E., Guerra, C.A., Patil, A.P., Tatem, A.J., Howes, R.E., and Myers, M.F. (2012). A long neglected world malaria map: Plasmodium vivax endemicity in 2010. PLoS neglected tropical diseases *6*, e1814.

143 Giovannini, D., Späth, S., Lacroix, C., Perazzi, A., Bargieri, D., Lagal, V., Lebugle, C., Combe, A., Thiberge, S., and Baldacci, P. (2011). Independent roles of apical membrane antigen 1 and rhoptry neck proteins during host cell invasion by apicomplexa. Cell host & microbe *10*, 591-602.

144 Golassa, L., Erko, B., Baliraine, F.N., Aseffa, A., and Swedberg, G. (2015). Polymorphisms in chloroquine resistance-associated genes in Plasmodium vivax in Ethiopia. Malaria journal *14*, 164.

145 Gomez, N.D., Safeukui, I., Adelani, A.A., Tewari, R., Reddy, J.K., Rao, S., Holder, A., Buffet, P., Mohandas, N., and Haldar, K. (2011). Deletion of a malaria invasion gene reduces death and anemia, in model hosts. PLoS One *6*, e25477.

146 Gonçalves, L.A., Cravo, P., and Ferreira, M.U. (2014). Emerging Plasmodium vivax resistance to chloroquine in South America: an overview. Memórias do Instituto Oswaldo Cruz *109*, 534-539.

147 Good, M.F., and Saul, A.J. (1993). Molecular Immunological Considerations in Malaria Vaccine Development (CRC Press).

148 Goodswen, S.J., Kennedy, P.J., and Ellis, J.T. (2018). A gene-based positive selection detection approach to identify vaccine candidates using Toxoplasma gondii as a test case protozoan pathogen. Frontiers in Genetics *9*, 332.

149 Gosling, R., and von Seidlein, L. (2016). The future of the RTS, S/AS01 malaria vaccine: an alternative development plan. PLoS medicine *13*, e1001994.

150 Greenhouse, B., Ho, B., Hubbard, A., Njama-Meya, D., Narum, D.L., Lanar, D.E., Dutta, S., Rosenthal, P.J., Dorsey, G., and John, C.C. (2011). Antibodies to Plasmodium falciparum antigens predict a higher risk of malaria but protection from symptoms once parasitemic. Journal of Infectious Diseases *204*, 19-26.

151 Griffiths, M.J., Shafi, M.J., Popper, S.J., Hemingway, C.A., Kortok, M.M., Wathen, A., Rockett, K.A., Mott, R., Levin, M., and Newton, C.R. (2005). Genomewide analysis of the host response to malaria in Kenyan children. Journal of Infectious Diseases *191*, 1599-1611.

152 Grun, J., and Weidanz, W. (1983). Antibody-independent immunity to reinfection malaria in B-cell-deficient mice. Infection and immunity *41*, 1197-1204.

153 Guerra, C.A., Snow, R.W., and Hay, S.I. (2006). Mapping the global extent of malaria in 2005. Trends in parasitology *22*, 353-358.

154 Gunalan, K., Niangaly, A., Thera, M.A., Doumbo, O.K., and Miller, L.H. (2018). Plasmodium vivax Infections of Duffy-Negative Erythrocytes: Historically Undetected or a Recent Adaptation? Trends in parasitology.

155 Gupta, B., Parker, D.M., Fan, Q., Reddy, B.N., Yan, G., Sattabongkot, J., and Cui, L. (2016). Microgeographically diverse Plasmodium vivax populations at the Thai-Myanmar border. Infection, Genetics and Evolution *45*, 341-346.

156 Gurkan, C., and Ellar, D.J. (2005). Recombinant production of bacterial toxins and their derivatives in the methylotrophic yeast Pichia pastoris. Microbial Cell Factories *4*, 33.

157 Guthmann, J.P., Pittet, A., Lesage, A., Imwong, M., Lindegardh, N., Min Lwin, M., Zaw, T., Annerberg, A., De Radiguès, X., and Nosten, F. (2008). Plasmodium vivax resistance to chloroquine in Dawei, southern Myanmar. Tropical Medicine & International Health *13*, 91-98.

158 Guy, A.J., Irani, V., MacRaild, C.A., Anders, R.F., Norton, R.S., Beeson, J.G., Richards, J.S., and Ramsland, P.A. (2015). Insights into the immunological properties of intrinsically disordered malaria proteins using proteome scale predictions. PLoS One *10*, e0141729.

159 Guyant, P., Canavati, S.E., Chea, N., Ly, P., Whittaker, M.A., Roca-Feltrer, A., and Yeung, S. (2015). Malaria and the mobile and migrant population in Cambodia: a population movement framework to inform strategies for malaria control and elimination. Malaria journal *14*, 252.

160 H. FELL, A., CURRIER, J., and F. GOOD, M. (1994). Inhibition of Plasmodium falciparum growth in vitro by CD4+ and CD8+ T cells from non-exposed donors. Parasite immunology *16*, 579-586.

161 Haas, L. (1999). Charles Louis Alphonse Laveran (1845-1922). Journal of Neurology, Neurosurgery & Psychiatry *67*, 520-520.

162 Han, J.-H., Lee, S.-K., Wang, B., Muh, F., Nyunt, M.H., Na, S., Ha, K.-S., Hong, S.-H., Park, W.S., and Sattabongkot, J. (2016). Identification of a reticulocyte-specific binding domain of Plasmodium vivax reticulocyte-binding protein 1 that is homologous to the PfRh4 erythrocyte-binding domain. Scientific reports *6*, 26993.

163 Hans, D., Pattnaik, P., Bhattacharyya, A., Shakri, A.R., Yazdani, S.S., Sharma, M., Choe, H., Farzan, M., and Chitnis, C.E. (2005). Mapping binding residues in the Plasmodium vivax domain that binds Duffy antigen during red cell invasion. Molecular microbiology *55*, 1423-1434.

164 Havryliuk, T., and Ferreira, M.U. (2009). A closer look at multiple-clone Plasmodium vivax infections: detection methods, prevalence and consequences. Memorias do Instituto Oswaldo Cruz *104*, 67-73.

165 Hay, S.I., Smith, D.L., and Snow, R.W. (2008). Measuring malaria endemicity from intense to interrupted transmission. The Lancet infectious diseases *8*, 369-378.

166 Hemmer, C.J., Holst, F.G.E., Kern, P., Chiwakata, C.B., Dietrich, M., and Reisinger, E.C. (2006). Stronger host response per parasitized erythrocyte in Plasmodium vivax or ovale than in Plasmodium falciparum malaria. Tropical Medicine & International Health *11*, 817-823.

167 Hill, A.V. (2011). Vaccines against malaria. Philosophical Transactions of the Royal Society B: Biological Sciences *366*, 2806-2814.

168 Hill, D.L., Eriksson, E.M., Suen, C.S.L.W., Chiu, C.Y., Ryg-Cornejo, V., Robinson, L.J., Siba, P.M., Mueller, I., Hansen, D.S., and Schofield, L. (2013). Opsonising antibodies to P. falciparum merozoites associated with immunity to clinical malaria. PLoS One *8*, e74627.

169 Hoffman, S.L., Goh, L.M., Luke, T.C., Schneider, I., Le, T.P., Doolan, D.L., Sacci, J., de la Vega, P., Dowler, M., and Paul, C. (2002). Protection of humans against malaria by immunization with radiation-attenuated Plasmodium falciparum sporozoites. The Journal of infectious diseases *185*, 1155-1164.

170 Hostetler, J.B., Lo, E., Kanjee, U., Amaratunga, C., Suon, S., Sreng, S., Mao, S., Yewhalaw, D., Mascarenhas, A., and Kwiatkowski, D.P. (2016). Independent origin and global distribution of distinct Plasmodium vivax Duffy binding protein gene duplications. PLoS neglected tropical diseases *10*, e0005091.

171 Huang, J., Tsao, T., Zhang, M., Rai, U., Tsuji, M., and Li, X. (2015). A sufficient role of MHC class I molecules on hepatocytes in anti-plasmodial activity of CD8+ T cells in vivo. Frontiers in microbiology *6*, 69.

172 Hudson, R.R. (1987). Estimating the recombination parameter of a finite population model without selection. Genetics Research *50*, 245-250.

173 Hughes, A.L. (1991). Circumsporozoite protein genes of malaria parasites (Plasmodium spp.): evidence for positive selection on immunogenic regions. Genetics *127*, 345-353.

174 Hughes, A.L. (2008). Near neutrality. Annals of the New York Academy of Sciences *1133*, 162-179.

175 Hulden, L., and Hulden, L. (2011). Activation of the hypnozoite: a part of Plasmodium vivax life cycle and survival. Malaria journal *10*, 90.

176 Hupalo, D.N., Luo, Z., Melnikov, A., Sutton, P.L., Rogov, P., Escalante, A., Vallejo, A.F., Herrera, S., Arévalo-Herrera, M., and Fan, Q. (2016). Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax. Nature genetics *48*, 953.

177 Hviid, L., Kurtzhals, J., Dodoo, D., Rodrigues, O., Rønn, A., Commey, J., Nkrumah, F.K., and Theander, T.G. (1996). The gamma/delta T-cell response to Plasmodium falciparum malaria in a population in which malaria is endemic. Infection and immunity *64*, 4359-4362.

178 Hviid, L., Theander, T.G., Abdulhadi, N.H., Abu-Zeid, Y.A., Bayoumi, R.A., and Jensen, J.B. (1991). Transient depletion of T cells with high LFA-1 expression from peripheral circulation during acute Plasmodium falciparum malaria. European journal of immunology *21*, 1249-1253.

179 Imwong, M., Nair, S., Pukrittayakamee, S., Sudimack, D., Williams, J.T., Mayxay, M., Newton, P.N., Kim, J.R., Nandy, A., and Osorio, L. (2007). Contrasting genetic structure in Plasmodium vivax populations from Asia and South America. International journal for parasitology *37*, 1013-1022.

180 Imwong, M., Pukrittakayamee, S., Looareesuwan, S., Pasvol, G., Poirreiz, J., White, N.J., and Snounou, G. (2001). Association of Genetic Mutations inPlasmodium vivax dhfr with Resistance to Sulfadoxine-Pyrimethamine: Geographical and Clinical Correlates. Antimicrobial agents and chemotherapy *45*, 3122-3127.

181 Imwong, M., Pukrittayakamee, S., Pongtavornpinyo, W., Nakeesathit, S., Nair, S., Newton, P., Nosten, F., Anderson, T.J., Dondorp, A., and Day, N.P. (2008). Gene amplification of the multidrug resistance 1 gene of Plasmodium vivax isolates from Thailand, Laos, and Myanmar. Antimicrobial agents and chemotherapy *52*, 2657-2659.

182 Inoue, S.-I., Niikura, M., Mineo, S., and Kobayashi, F. (2013). Roles of IFN-γ and γδ T cells in protective immunity against blood-stage malaria. Frontiers in immunology *4*, 258.

183 Iyer, J., Grüner, A.C., Rénia, L., Snounou, G., and Preiser, P.R. (2007). Invasion of host cells by malaria parasites: a tale of two protein families. Molecular microbiology *65*, 231-249.

184 Jennison, C., Arnott, A., Tessier, N., Tavul, L., Koepfli, C., Felger, I., Siba, P.M., Reeder, J.C., Bahlo, M., and Mueller, I. (2015). Plasmodium vivax populations are more genetically diverse and less structured than sympatric Plasmodium falciparum populations. PLoS neglected tropical diseases *9*, e0003634.

185 Jepson, A., Banya, W., Sisay-Joof, F., Hassan-King, M., Nunes, C., Bennett, S., and Whittle, H. (1997). Quantification of the relative contribution of major histocompatibility complex (MHC) and non-MHC genes to human immune responses to foreign antigens. Infection and Immunity *65*, 872-876.

186 Jiang, G., Shi, M., Conteh, S., Richie, N., Banania, G., Geneshan, H., Valencia, A., Singh, P., Aguiar, J., and Limbach, K. (2009). Sterile protection against Plasmodium knowlesi in rhesus monkeys from a malaria vaccine: comparison of heterologous prime boost strategies. PloS one *4*, e6559.

187 Jongwutiwes, S., Buppan, P., Kosuvin, R., Seethamchai, S., Pattanawong, U., Sirichaisinthop, J., and Putaporntip, C. (2011). Plasmodium knowlesi malaria in humans and macaques, Thailand. Emerging infectious diseases *17*, 1799.

188 Jongwutiwes, S., Putaporntip, C., and Hughes, A.L. (2010). Bottleneck effects on vaccine-candidate antigen diversity of malaria parasites in Thailand. Vaccine *28*, 3112-3117.

189 Jongwutiwes, S., Tanabe, K., Hughes, M.K., Kanbara, H., and Hughes, A.L. (1994). Allelic variation in the circumsporozoite protein of Plasmodium falciparum from Thai field isolates. The American journal of tropical medicine and hygiene *51*, 659-668.

190 Joshi, N., and Fass, J. (2011). sickle - A windowed adaptive trimming tool for FASTQ files using quality.

191 Jukes, T.H., and Cantor, C.R. (1969). Evolution of protein molecules. Mammalian protein metabolism *3*, 132.

192 Jun, G., Wing, M.K., Abecasis, G.R., and Kang, H.M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. Genome research, gr. 176552.176114.

193 Kadekoppala, M., and Holder, A.A. (2010). Merozoite surface proteins of the malaria parasite: the MSP1 complex and the MSP7 family. International journal for parasitology *40*, 1155-1161.

194 Kadekoppala, M., O'Donnell, R.A., Grainger, M., Crabb, B.S., and Holder, A.A. (2008). Deletion of the Plasmodium falciparum merozoite surface protein 7 gene impairs parasite invasion of erythrocytes. Eukaryotic cell *7*, 2123-2132.

195 Kadekoppala, M., Ogun, S.A., Howell, S., Gunaratne, R.S., and Holder, A.A. (2010). Systematic genetic analysis of the Plasmodium falciparum MSP7-like family reveals differences in protein expression, location, and importance in asexual growth of the blood-stage parasite. Eukaryotic cell *9*, 1064-1074.

196 Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research *28*, 27-30.

197 Karplus, P., and Schulz, G. (1985). Prediction of chain flexibility in proteins. Naturwissenschaften *72*, 212-213.

198 Karunajeewa, H.A., Mueller, I., Senn, M., Lin, E., Law, I., Gomorrai, P.S., Oa, O., Griffin, S., Kotab, K., and Suano, P. (2008). A trial of combination antimalarial therapies in children from Papua New Guinea. New England Journal of Medicine *359*, 2545-2557.

199 Kaslow, D.C., and Biernaux, S. (2015). RTS, S: toward a first landmark on the Malaria Vaccine Technology Roadmap. Vaccine *33*, 7425-7432.

200 Kats, L.M., Cooke, B.M., Coppel, R.L., and Black, C.G. (2008). Protein trafficking to apical organelles of malaria parasites–building an invasion machine. Traffic *9*, 176-186.

201 Kauth, C.W., Woehlbier, U., Kern, M., Mekonnen, Z., Lutz, R., Mücke, N., Langowski, J., and Bujard, H. (2006). Interactions between merozoite surface proteins 1, 6, and 7 of the malaria parasite Plasmodium falciparum. Journal of Biological Chemistry *281*, 31517-31527.

202 Kazmin, D., Nakaya, H.I., Lee, E.K., Johnson, M.J., Van Der Most, R., Van Den Berg, R.A., Ballou, W.R., Jongert, E., Wille-Reece, U., and Ockenhouse, C. (2017). Systems analysis of protective immune responses to RTS, S malaria vaccination in humans. Proceedings of the National Academy of Sciences *114*, 2425-2430.

203 Keitany, G.J., Vignali, M., and Wang, R. (2014). Live attenuated pre-erythrocytic malaria vaccines. Human vaccines & immunotherapeutics *10*, 2903-2909.

204 Kensil, C.R., Liu, G., Anderson, C., and Storey, J. (2006). Effects of QS-21 on innate and adaptive immune responses. In Vaccine Adjuvants (Springer), pp. 221-234.

205 Kidson, C., Lamont, G., Saul, A., and Nurse, G.T. (1981). Ovalocytic erythrocytes from Melanesians are resistant to invasion by malaria parasites in culture. Proceedings of the National Academy of Sciences *78*, 5829-5832.

206 Kim, A., Popovici, J., Vantaux, A., Samreth, R., Bin, S., Kim, S., Roesch, C., Liang, L., Davies, H., and Felgner, P. (2017). Characterization of P. vivax blood stage transcriptomes from field isolates reveals similarities among infections and complex gene isoforms. Scientific reports *7*, 7761.

207 Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology *14*, R36.

208 Kittichai, V., Koepfli, C., Nguitragool, W., Sattabongkot, J., and Cui, L. (2017). Substantial population structure of Plasmodium vivax in Thailand facilitates identification of the sources of residual transmission. PLoS neglected tropical diseases *11*, e0005930.

209 Klintman, D., Li, X., and Thorlacius, H. (2004). Important role of P-selectin for leukocyte recruitment, hepatocellular injury, and apoptosis in endotoxemic mice. Clinical and diagnostic laboratory immunology *11*, 56-62.

210 Koepfli, C., Rodrigues, P.T., Antao, T., Orjuela-Sánchez, P., Van den Eede, P., Gamboa, D., Van Hong, N., Bendezu, J., Erhart, A., and Barnadas, C. (2015). Plasmodium vivax diversity and population structure across four continents. PLoS neglected tropical diseases *9*, e0003872.

211 Kolpakov, R., Bana, G., and Kucherov, G. (2003). mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic acids research *31*, 3672-3678.

212 Kosakovsky Pond, S.L., and Frost, S.D. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. Molecular biology and evolution *22*, 1208-1222.

213 Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., and Frost, S.D. (2006). GARD: a genetic algorithm for recombination detection. Bioinformatics *22*, 3096-3098.

214 Kosuwin, R., Feng, M., Makiuchi, T., Putaporntip, C., Tachibana, H., and Jongwutiwes, S. (2018). Naturally acquired IgG antibodies to thrombospondin-related anonymous protein of Plasmodium vivax (Pv TRAP) in Thailand predominantly elicit immunological cross-reactivity. Tropical Medicine & International Health.

215 Kosuwin, R., Putaporntip, C., Tachibana, H., and Jongwutiwes, S. (2014). Spatial variation in genetic diversity and natural selection on the thrombospondin-related adhesive protein locus of Plasmodium vivax (PvTRAP). PloS one *9*, e110463.

216 Koussis, K., Withers-Martinez, C., Yeoh, S., Child, M., Hackett, F., Knuepfer, E., Juliano, L., Woehlbier, U., Bujard, H., and Blackman, M.J. (2009). A multifunctional serine protease primes the malaria parasite for red blood cell invasion. The EMBO journal *28*, 725-735.

217 Kozlowski, L.P., and Bujnicki, J.M. (2012). MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC bioinformatics *13*, 111.

218 Krause, D.R., Gatton, M.L., Frankland, S., Eisen, D.P., Good, M.F., Tilley, L., and Cheng, Q. (2007). Characterization of the antibody response against Plasmodium falciparum erythrocyte membrane protein 1 in human volunteers. Infection and immunity *75*, 5967-5973.

219 Kubler-Kielb, J., Majadly, F., Biesova, Z., Mocca, C.P., Guo, C., Nussenzweig, R., Nussenzweig, V., Mishra, S., Wu, Y., and Miller, L.H. (2010). A bicomponent Plasmodium falciparum investigational vaccine composed of protein-peptide conjugates. Proceedings of the National Academy of Sciences *107*, 1172-1177.

220 Kubler-Kielb, J., Majadly, F., Wu, Y., Narum, D.L., Guo, C., Miller, L.H., Shiloach, J., Robbins, J.B., and Schneerson, R. (2007). Long-lasting and transmission-blocking activity of antibodies to Plasmodium falciparum elicited in mice by protein conjugates of Pfs25. Proceedings of the National Academy of Sciences *104*, 293-298.

221 Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular biology and evolution *33*, 1870-1874.

222 Kurtis, J.D., Hollingdale, M.R., Luty, A.J., Lanar, D.E., Krzych, U., and Duffy, P.E. (2001). Pre-erythrocytic immunity to Plasmodium falciparum: the case for an LSA-1 vaccine. Trends in parasitology *17*, 219-223.

223 Kusi, K.A., Manu, E.A., Gwira, T.M., Kyei-Baafour, E., Dickson, E.K., Amponsah, J.A., Remarque, E.J., Faber, B.W., Kocken, C.H., and Dodoo, D.

(2017). Variations in the quality of malaria-specific antibodies with transmission intensity in a seasonal malaria transmission area of Northern Ghana. PloS one *12*, e0185303.

224 Kyes, S.A., Kraemer, S.M., and Smith, J.D. (2007). Antigenic variation in Plasmodium falciparum: gene organization and regulation of the var multigene family. Eukaryotic cell *6*, 1511-1520.

225 Langhorne, J., Cross, C., Seixas, E., Li, C., and Von Der Weid, T. (1998). A role for B cells in the development of T cell helper function in a malaria infection in mice. Proceedings of the National Academy of Sciences *95*, 1730-1734.

226 Larkin, M.A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., and Lopez, R. (2007). Clustal W and Clustal X version 2.0. bioinformatics *23*, 2947-2948.

227 Larsen, J.E.P., Lund, O., and Nielsen, M. (2006). Improved method for predicting linear B-cell epitopes. Immunome research *2*, 2.

228 Lebrun, M., Michelin, A., El Hajj, H., Poncet, J., Bradley, P.J., Vial, H., and Dubremetz, J.F. (2005). The rhoptry neck protein RON4 relocalizes at the moving junction during Toxoplasma gondii invasion. Cellular microbiology *7*, 1823-1833.

229 Lell, B., Agnandji, S., Von Glasenapp, I., Haertle, S., Oyakhiromen, S., Issifou, S., Vekemans, J., Leach, A., Lievens, M., and Dubois, M.-C. (2009). A randomized trial assessing the safety and immunogenicity of AS01 and AS02 adjuvanted RTS, S malaria vaccine candidates in children in Gabon. PLoS one *4*, e7611.

230 Leroux-Roels, G., Leroux-Roels, I., Clement, F., Ofori-Anyinam, O., Lievens, M., Jongert, E., Moris, P., Ballou, W.R., and Cohen, J. (2014). Evaluation of the immune response to RTS, S/AS01 and RTS, S/AS02 adjuvanted vaccines: randomized, double-blind study in malaria-naive adults. Human vaccines & immunotherapeutics *10*, 2211-2219.

231 Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997.

232 Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics *30*, 2843-2851.

233 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078-2079.

234 Li, Y., Liu, X., Zhu, Y., Zhou, X., Cao, C., Hu, X., Ma, H., Wen, H., Ma, X., and Ding, J.-B. (2013). Bioinformatic prediction of epitopes in the Emy162

antigen of Echinococcus multilocularis. Experimental and therapeutic medicine *6*, 335-340.

235 Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology *34*, 816-834.

236 Liao, Y., Smyth, G.K., and Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.

237 Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics *25*, 1451-1452.

238 Lin, C.S., Uboldi, A.D., Epp, C., Bujard, H., Tsuboi, T., Czabotar, P.E., and Cowman, A.F. (2016). Multiple Plasmodium falciparum merozoite surface protein 1 complexes mediate merozoite binding to human erythrocytes. Journal of Biological Chemistry *291*, 7703-7715.

239 Lin, J.T., Patel, J.C., Kharabora, O., Sattabongkot, J., Muth, S., Ubalee, R., Schuster, A.L., Rogers, W.O., Wongsrichanalai, C., and Juliano, J.J. (2013). Plasmodium vivax isolates from Cambodia and Thailand show high genetic complexity and distinct patterns of P. vivax multidrug resistance gene 1 (pvmdr1) polymorphisms. The American journal of tropical medicine and hygiene *88*, 1116-1123.

240 Lingala, M.A. (2017). Effect of meteorological variables on Plasmodium vivax and Plasmodium falciparum malaria in outbreak prone districts of Rajasthan, India. Journal of infection and public health *10*, 875-880.

241 Liu, Y., Auburn, S., Cao, J., Trimarsanto, H., Zhou, H., Gray, K.-A., Clark, T.G., Price, R.N., Cheng, Q., and Huang, R. (2014). Genetic diversity and population structure of Plasmodium vivax in Central China. Malaria journal *13*, 262.

242 Long, C.A., and Hoffman, S.L. (2002). Malaria--from Infants to Genomics to Vaccines. Science *297*, 345-347.

243 López, C., Yepes-Pérez, Y., Hincapié-Escobar, N., Díaz-Arévalo, D., and Patarroyo, M.A. (2017). what is Known about the immune Response induced by Plasmodium vivax Malaria vaccine Candidates? Frontiers in immunology *8*, 126.

244 López-Barragán, M.J., Lemieux, J., Quiñones, M., Williamson, K.C., Molina-Cruz, A., Cui, K., Barillas-Mury, C., Zhao, K., and Su, X.-z. (2011). Directional gene expression and antisense transcripts in sexual and asexual stages of Plasmodium falciparum. BMC genomics *12*, 587.

245 Love, M., Anders, S., and Huber, W. (2014). Differential analysis of count data–the DESeq2 package. Genome Biol *15*, 550.

246 Lu, Y., Li, Z., Teng, H., Xu, H., Qi, S., He, J.a., Gu, D., Chen, Q., and Ma, H. (2015). Chimeric peptide constructs comprising linear B-cell epitopes: application to the serodiagnosis of infectious diseases. Scientific reports *5*, 13364.

247 Lukens, A.K., Ross, L.S., Heidebrecht, R., Gamo, F.J., Lafuente-Monasterio, M.J., Booker, M.L., Hartl, D.L., Wiegand, R.C., and Wirth, D.F. (2014). Harnessing evolutionary fitness in Plasmodium falciparum for drug discovery and suppressing resistance. Proceedings of the National Academy of Sciences *111*, 799-804.

248 Lyon, J.A., Angov, E., Fay, M.P., Sullivan, J.S., Girourd, A.S., Robinson, S.J., Bergmann-Leitner, E.S., Duncan, E.H., Darko, C.A., and Collins, W.E. (2008). Protection induced by Plasmodium falciparum MSP142 is strain-specific, antigen and adjuvant dependent, and correlates with antibody responses. PloS one *3*, e2830.

249 MacKellar, D.C., Vaughan, A.M., Aly, A.S., DeLeon, S., and Kappe, S.H. (2011). A systematic analysis of the early transcribed membrane protein family throughout the life cycle of Plasmodium yoelii. Cellular microbiology *13*, 1755-1767.

250 MacRaild, C.A., Zachrdla, M., Andrew, D., Krishnarjuna, B., Nováček, J., Žídek, L., Sklenář, V., Richards, J.S., Beeson, J.G., and Anders, R.F. (2015). Conformational dynamics and antigenicity in the disordered malaria antigen merozoite surface protein 2. PLoS One *10*, e0119899.

251 Mahajan, B., Berzofsky, J.A., Boykins, R.A., Majam, V., Zheng, H., Chattopadhyay, R., de la Vega, P., Moch, J.K., Haynes, J.D., and Belyakov, I.M. (2010). Multiple antigen peptide vaccines against Plasmodium falciparum malaria. Infection and immunity *78*, 4613-4624.

252 Malik, G.M., Seidi, O., El-Taher, A., and Mohammed, A.S. (1998). Clinical aspects of malaria in the Asir Region, Saudi Arabia. Ann Saudi Med *18*, 15-17.

253 Marsh, K. (1992). Malaria-a neglected disease? Parasitology *104*, S53-S69.

254 Marsh, K., Otoo, L., Hayes, R., Carson, D., and Greenwood, B. (1989). Antibodies to blood stage antigens of Plasmodium falciparum in rural Gambians and their relation to protection against infection. Transactions of the Royal Society of Tropical Medicine and Hygiene *83*, 293-303.

255 Martin, D.P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. Virus evolution *1*.

256 Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal *17*, pp. 10-12.

257 Matuschewski, K., and Mueller, A.K. (2007). Vaccines against malaria–an update. The FEBS journal *274*, 4680-4687.

258 McCarthy, J.S., and Good, M.F. (2010). Whole parasite blood stage malaria vaccines: a convergence of evidence. Human vaccines *6*, 114-123.

259 McGregor, I. (1964). Studies in the acquisition of immunity to Plasmodium falciparum infections in Africa. Transactions of the Royal Society of Tropical Medicine and Hygiene *58*, 80-92.

260 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., and Daly, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research *20*, 1297-1303.

261 Mello, K., Daly, T.M., Long, C.A., Burns, J.M., and Bergman, L.W. (2004). Members of the merozoite surface protein 7 family with similar expression patterns differ in ability to protect against Plasmodium yoelii malaria. Infection and immunity *72*, 1010-1018.

262 Mendes, C., Dias, F., Figueiredo, J., Mora, V.G., Cano, J., de Sousa, B., Do Rosário, V.E., Benito, A., Berzosa, P., and Arez, A.P. (2011). Duffy negative antigen is no longer a barrier to Plasmodium vivax–molecular evidences from the African West Coast (Angola and Equatorial Guinea). PLoS neglected tropical diseases *5*, e1192.

263 Miller, L.H., Mason, S.J., Clyde, D.F., and McGinniss, M.H. (1976). The resistance factor to Plasmodium vivax in blacks: the Duffy-blood-group genotype, FyFy. New England Journal of Medicine *295*, 302-304.

264 Minor, P.D. (2015). Live attenuated vaccines: historical successes and current challenges. Virology *479*, 379-392.

265 Mitchell, G., Thomas, A., Margos, G., Dluzewski, A., and Bannister, L. (2004). Apical membrane antigen 1, a major malaria vaccine candidate, mediates the close attachment of invasive merozoites to host red blood cells. Infection and immunity *72*, 154-158.

266 Miura, K. (2016). Progress and prospects for blood-stage malaria vaccines. Expert review of vaccines *15*, 765-781.

267 Mongui, A., Perez-Leal, O., Soto, S.C., Cortes, J., and Patarroyo, M.A. (2006). Cloning, expression, and characterisation of a Plasmodium vivax MSP7 family merozoite surface protein. Biochemical and biophysical research communications *351*, 639-644.

268 Moormann, A.M., Sumba, P.O., Chelimo, K., Fang, H., Tisch, D.J., Dent, A.E., John, C.C., Long, C.A., Vulule, J., and Kazura, J.W. (2013). Humoral and cellular immunity to Plasmodium falciparum merozoite surface protein 1 and

protection from infection with blood-stage parasites. The Journal of infectious diseases *208*, 149-158.

269 Moorthy, V.S., and Ballou, W.R. (2009). Immunological mechanisms underlying protection mediated by RTS, S: a review of the available data. Malaria journal *8*, 312.

270 Morales, R.A., MacRaild, C.A., Seow, J., Krishnarjuna, B., Drinkwater, N., Rouet, R., Anders, R.F., Christ, D., McGowan, S., and Norton, R.S. (2015). Structural basis for epitope masking and strain specificity of a conserved epitope in an intrinsically disordered malaria vaccine candidate. Scientific reports *5*, 10103.

271 Mulamba, C., Riveron, J.M., Ibrahim, S.S., Irving, H., Barnes, K.G., Mukwaya, L.G., Birungi, J., and Wondji, C.S. (2014). Widespread pyrethroid and DDT resistance in the major malaria vector Anopheles funestus in East Africa is driven by metabolic resistance mechanisms. PloS one *9*, e110058.

272 Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S.L., and Scheffler, K. (2013). FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Molecular biology and evolution *30*, 1196-1205.

273 Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., and Pond, S.L.K. (2012). Detecting individual sites subject to episodic diversifying selection. PLoS Genet *8*, e1002764.

274 Nardin, E.H., Oliveira, G.A., Calvo-Calle, J.M., Castro, Z.R., Nussenzweig, R.S., Schmeckpeper, B., Hall, B.F., Diggs, C., Bodison, S., and Edelman, R. (2000). Synthetic malaria peptide vaccine elicits high levels of antibodies in vaccinees of defined HLA genotypes. The Journal of infectious diseases *182*, 1486-1496.

275 Neafsey, D.E., Galinsky, K., Jiang, R.H., Young, L., Sykes, S.M., Saif, S., Gujja, S., Goldberg, J.M., Young, S., and Zeng, Q. (2012). The malaria parasite Plasmodium vivax exhibits greater genetic diversity than Plasmodium falciparum. Nature genetics *44*, 1046.

276 Neafsey, D.E., Juraska, M., Bedford, T., Benkeser, D., Valim, C., Griggs, A., Lievens, M., Abdulla, S., Adjei, S., and Agbenyega, T. (2015). Genetic diversity and protective efficacy of the RTS, S/AS01 malaria vaccine. New England Journal of Medicine *373*, 2025-2037.

277 Nei, M. (1987). Molecular evolutionary genetics (Columbia university press).

278 Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular biology and evolution *3*, 418-426.

279 Nhabomba, A.J., Guinovart, C., Jiménez, A., Manaca, M.N., Quintó, L., Cisteró, P., Aguilar, R., Barbosa, A., Rodríguez, M.H., and Bassat, Q. (2014). Impact of age of first exposure to Plasmodium falciparum on antibody responses to malaria in children: a randomized, controlled trial in Mozambique. Malaria journal *13*, 121.

280 Nixon, C.E., Park, S., Pond-Tor, S., Raj, D., Lambert, L.E., Orr-Gonzalez, S., Barnafo, E.K., Rausch, K.M., Friedman, J.F., and Fried, M. (2017). Identification of protective B-cell epitopes within the novel malaria vaccine candidate Plasmodium falciparum schizont egress antigen 1. Clinical and Vaccine Immunology *24*, e00068-00017.

281 Noulin, F., Borlon, C., Van Den Abbeele, J., D'Alessandro, U., and Erhart, A. (2013). 1912-2012: a century of research on *Plasmodium vivax* in vitro culture. Trends Parasitol *29*, 286-294.

282 Ntumngia, F.B., Schloegel, J., Barnes, S.J., McHenry, A.M., Singh, S., King, C.L., and Adams, J.H. (2012). Conserved and variant epitopes of Plasmodium vivax Duffy binding protein as targets of inhibitory monoclonal antibodies. Infection and immunity, IAI. 05924-05911.

283 Ockenhouse, C.F., Regules, J., Tosh, D., Cowden, J., Kathcart, A., Cummings, J., Paolino, K., Moon, J., Komisar, J., and Kamau, E. (2015). Ad35. CS. 01-RTS, S/AS01 heterologous prime boost vaccine efficacy against sporozoite challenge in healthy malaria-naive adults. PloS one *10*, e0131571.

284 Ogutu, B.R., Apollo, O.J., McKinney, D., Okoth, W., Siangla, J., Dubovsky, F., Tucker, K., Waitumbi, J.N., Diggs, C., and Wittes, J. (2009). Blood stage malaria vaccine eliciting high antigen-specific antibody concentrations confers no protection to young children in Western Kenya. PloS one *4*, e4708.

285 Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2015). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics *32*, 292-294.

286 Olotu, A., Fegan, G., Wambua, J., Nyangweso, G., Awuondo, K.O., Leach, A., Lievens, M., Leboulleux, D., Njuguna, P., and Peshu, N. (2013). Four-year efficacy of RTS, S/AS01E and its interaction with malaria exposure. New England Journal of Medicine *368*, 1111-1120.

287 Olotu, A., Fegan, G., Wambua, J., Nyangweso, G., Leach, A., Lievens, M., Kaslow, D.C., Njuguna, P., Marsh, K., and Bejon, P. (2016). Seven-year efficacy of RTS, S/AS01 malaria vaccine among young African children. New England Journal of Medicine *374*, 2519-2529.

288 Orsi, R., Ripoll, D., Yeung, M., Nightingale, K., and Wiedmann, M. (2007). Recombination and positive selection contribute to evolution of Listeria monocytogenes inlA. Microbiology *153*, 2666-2678.

289 Osier, F.H., Fegan, G., Polley, S.D., Murungi, L., Verra, F., Tetteh, K.K., Lowe, B., Mwangi, T., Bull, P.C., and Thomas, A.W. (2008). Breadth and magnitude of antibody responses to multiple Plasmodium falciparum merozoite antigens are associated with protection from clinical malaria. Infection and immunity *76*, 2240-2248.

290 Otto, T.D., Böhme, U., Jackson, A.P., Hunt, M., Franke-Fayard, B., Hoeijmakers, W.A., Religa, A.A., Robertson, L., Sanders, M., and Ogun, S.A. (2014). A comprehensive evaluation of rodent malaria parasite genomes and gene expression. BMC biology *12*, 86.

291 Otto, T.D., Wilinski, D., Assefa, S., Keane, T.M., Sarry, L.R., Böhme, U., Lemieux, J., Barrell, B., Pain, A., and Berriman, M. (2010). New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. Molecular microbiology *76*, 12-24.

292 Outchkourov, N.S., Roeffen, W., Kaan, A., Jansen, J., Luty, A., Schuiffel, D., van Gemert, G.J., van de Vegte-Bolmer, M., Sauerwein, R.W., and Stunnenberg, H.G. (2008). Correctly folded Pfs48/45 protein of Plasmodium falciparum elicits malaria transmission-blocking immunity in mice. Proceedings of the National Academy of Sciences *105*, 4301-4305.

293 Owusu-Agyei, S., Koram, K., Baird, J.K., Utz, G., Binka, F.N., Nkrumah, F., Fryauff, D., and Hoffman, S. (2001). Incidence of symptomatic and asymptomatic Plasmodium falciparum infection following curative therapy in adult residents of northern Ghana. The American journal of tropical medicine and hygiene *65*, 197-203.

294 Oyarzún, P., Ellis, J.J., Bodén, M., and Kobe, B. (2013). PREDIVAC: CD4+ T-cell epitope prediction for vaccine design that covers 95% of HLA class II DR protein diversity. BMC bioinformatics *14*, 52.

295 Oyarzún, P., and Kobe, B. (2016). Recombinant and epitope-based vaccines on the road to the market and implications for vaccine design and production. Human vaccines & immunotherapeutics *12*, 763-767.

296 Pachebat, J.A., Kadekoppala, M., Grainger, M., Dluzewski, A.R., Gunaratne, R.S., Scott-Finnigan, T.J., Ogun, S.A., Ling, I.T., Bannister, L.H., and Taylor, H.M. (2007). Extensive proteolytic processing of the malaria parasite merozoite surface protein 7 during biosynthesis and parasite release from erythrocytes. Molecular and biochemical parasitology *151*, 59-69.

297 Pachebat, J.A., Ling, I.T., Grainger, M., Trucco, C., Howell, S., Fernandez-Reyes, D., Gunaratne, R., and Holder, A.A. (2001). The 22 kDa component of the protein complex on the surface of Plasmodium falciparum merozoites is derived from a larger precursor, merozoite surface protein 7. Molecular and biochemical parasitology *117*, 83-89.

298 Pacheco, M.A., Ryan, E.M., Poe, A.C., Basco, L., Udhayakumar, V., Collins, W.E., and Escalante, A.A. (2010). Evidence for negative selection on the gene

encoding rhoptry-associated protein 1 (RAP-1) in Plasmodium spp. Infection, Genetics and Evolution *10*, 655-661.

299 Pain, A., Böhme, U., Berry, A., Mungall, K., Finn, R., Jackson, A.P., Mourier, T., Mistry, J., Pasini, E., and Aslett, M. (2008). The genome of the simian and human malaria parasite Plasmodium knowlesi. Nature *455*, 799.

300 Park, D.J., Lukens, A.K., Neafsey, D.E., Schaffner, S.F., Chang, H.-H., Valim, C., Ribacke, U., Van Tyne, D., Galinsky, K., and Galligan, M. (2012). Sequence-based association and selection scans identify drug resistance loci in the Plasmodium falciparum malaria parasite. Proceedings of the National Academy of Sciences *109*, 13052-13057.

301 Parker, D.M., Carrara, V.I., Pukrittayakamee, S., McGready, R., and Nosten, F.H. (2015). Malaria ecology along the Thailand–Myanmar border. Malaria journal *14*, 388.

302 Parra, M., Hui, G., Johnson, A.H., Berzofsky, J.A., Roberts, T., Quakyi, I.A., and Taylor, D.W. (2000). Characterization of conserved T-and B-cell epitopes in Plasmodium falciparum major merozoite surface protein 1. Infection and immunity *68*, 2685-2691.

303 Pasquale, A.D., Preiss, S., Silva, F.T.D., and Garçon, N. (2015). Vaccine adjuvants: from 1920 to 2015 and beyond. Vaccines *3*, 320-343.

304 Patz, J.A., and Olson, S.H. (2006). Malaria risk and temperature: influences from global climate change and local land use practices. Proceedings of the National Academy of Sciences *103*, 5635-5636.

305 Pearson, R.D., Amato, R., Auburn, S., Miotto, O., Almagro-Garcia, J., Amaratunga, C., Suon, S., Mao, S., Noviyanti, R., and Trimarsanto, H. (2016). Genomic analysis of local variation and recent evolution in Plasmodium vivax. Nature genetics *48*, 959.

306 PePperPrint (2017). PEPperCHIP Immunoassay Protocol (PePperPrint).

307 Pérignon, J.L., and Druilhe, P. (1994). Immune mechanisms underlying the premunition against Plasmodium falciparum malaria. Memórias do Instituto Oswaldo Cruz *89*, 51-53.

308 Perlmann, P., Perlmann, H., Looareesuwan, S., Krudsood, S., Kano, S., Matsumoto, Y., Brittenham, G., Troye-Blomberg, M., and Aikawa, M. (2000). Contrasting functions of IgG and IgE antimalarial antibodies in uncomplicated and severe Plasmodium falciparum malaria. The American journal of tropical medicine and hygiene *62*, 373-377.

309 Perrin, A.J., Bartholdson, S.J., and Wright, G.J. (2015). P-selectin is a host receptor for Plasmodium MSP7 ligands. Malaria journal *14*, 238.

310 Perrin, L.H., and Dayal, R. (1982). Immunity to Asexual Erythrocytic Stages of Plasmodium Falciparum: Role of Defined Antigens in the Humoral Response 1. Immunological reviews *61*, 245-269.

311 Phan, G.T., De Vries, P.J., Tran, B.Q., Le, H.Q., Nguyen, N.V., Nguyen, T.V., Heisterkamp, S.H., and Kager, P.A. (2002). Artemisinin or chloroquine for blood stage Plasmodium vivax malaria in Vietnam. Tropical Medicine & International Health *7*, 858-864.

312 Pinkevych, M., Petravic, J., Chelimo, K., Kazura, J.W., Moormann, A.M., and Davenport, M.P. (2012). The dynamics of naturally acquired immunity to Plasmodium falciparum infection. PLoS computational biology *8*, e1002729.

313 Pond, S.L.K., and Frost, S.D. (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics *21*, 2531-2533.

314 Posada, D. (2008). jModelTest: phylogenetic model averaging. Molecular biology and evolution *25*, 1253-1256.

315 Powles, L., Xiang, S.D., Selomulya, C., and Plebanski, M. (2015). The use of synthetic carriers in malaria vaccine design. Vaccines *3*, 894-929.

316 Price, R.N., Douglas, N.M., and Anstey, N.M. (2009). New developments in Plasmodium vivax malaria: severe disease and the rise of chloroquine resistance. Current opinion in infectious diseases *22*, 430-435.

317 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., and Daly, M.J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics *81*, 559-575.

318 Putaporntip, C., Hongsrimuang, T., Seethamchai, S., Kobasa, T., Limkittikul, K., Cui, L., and Jongwutiwes, S. (2009a). Differential prevalence of Plasmodium infections and cryptic Plasmodium knowlesi malaria in humans in Thailand. The Journal of infectious diseases *199*, 1143-1150.

319 Putaporntip, C., Jongwutiwes, S., Grynberg, P., Cui, L., and Hughes, A.L. (2009b). Nucleotide sequence polymorphism at the apical membrane antigen-1 locus reveals population history of Plasmodium vivax in Thailand. Infection, Genetics and Evolution *9*, 1295-1300.

320 Putaporntip, C., Jongwutiwes, S., and Hughes, A.L. (2009c). Natural selection maintains a stable polymorphism at the circumsporozoite protein locus of Plasmodium falciparum in a low endemic area. Infection, genetics and evolution *9*, 567-573.

321 Putaporntip, C., Jongwutiwes, S., Sakihama, N., Ferreira, M.U., Kho, W.-G., Kaneko, A., Kanbara, H., Hattori, T., and Tanabe, K. (2002). Mosaic organization and heterogeneity in frequency of allelic recombination of the

Plasmodium vivax merozoite surface protein-1 locus. Proceedings of the National Academy of Sciences *99*, 16348-16353.

322 Putaporntip, C., Miao, J., Kuamsab, N., Sattabongkot, J., Sirichaisinthop, J., Jongwutiwes, S., and Cui, L. (2014). The Plasmodium vivax merozoite surface protein 3β sequence reveals contrasting parasite populations in southern and northwestern Thailand. PLoS neglected tropical diseases *8*, e3336.

323 Putaporntip, C., Udomsangpetch, R., Pattanawong, U., Cui, L., and Jongwutiwes, S. (2010). Genetic diversity of the Plasmodium vivax merozoite surface protein-5 locus from diverse geographic origins. Gene *456*, 24-35.

324 Quintana, M.d.P., Ch'ng, J.-H., Moll, K., Zandian, A., Nilsson, P., Idris, Z.M., Saiwaew, S., Qundos, U., and Wahlgren, M. (2018). Antibodies in children with malaria to PfEMP1, RIFIN and SURFIN expressed at the Plasmodium falciparum parasitized red blood cell surface. Scientific reports *8*, 3262.

325 R Core Team (2017). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).

326 Raj, D.K., Nixon, C.P., Nixon, C.E., Dvorin, J.D., DiPetrillo, C.G., Pond-Tor, S., Wu, H.-W., Jolly, G., Pischel, L., and Lu, A. (2014). Antibodies to PfSEA-1 block parasite egress from RBCs and protect against malaria infection. Science *344*, 871-877.

327 Rajamani, D., Thiel, S., Vajda, S., and Camacho, C.J. (2004). Anchor residues in protein–protein interactions. Proceedings of the National Academy of Sciences *101*, 11287-11292.

328 Rau, A., and Maugis-Rabusseau, C. (2017). Transformation and model choice for RNA-seq co-expression analysis. Briefings in bioinformatics, bbw128.

329 Remarque, E.J., Roestenberg, M., Younis, S., Walraven, V., van der Werff, N., Faber, B.W., Leroy, O., Sauerwein, R., Kocken, C.H., and Thomas, A.W. (2012). Humoral immune responses to a single allele PfAMA1 vaccine in healthy malaria-naive adults. PLoS One *7*, e38898.

330 Rénia, L., and Goh, Y.S. (2016). Malaria parasites: the great escape. Frontiers in immunology *7*, 463.

331 Rhee, M., Akanmori, B., Waterfall, M., and Riley, E. (2001). Changes in cytokine production associated with acquired immunity to Plasmodium falciparum malaria. Clinical & Experimental Immunology *126*, 503-510.

332 Rice, B.L., Acosta, M.M., Pacheco, M.A., Carlton, J.M., Barnwell, J.W., and Escalante, A.A. (2014). The origin and diversification of the merozoite surface protein 3 (msp3) multi-gene family in Plasmodium vivax and related parasites. Molecular phylogenetics and evolution *78*, 172-184.

333 Rice, B.L., Acosta, M.M., Pacheco, M.A., and Escalante, A.A. (2013). Merozoite surface protein-3 alpha as a genetic marker for epidemiologic studies in Plasmodium vivax: a cautionary note. Malaria journal *12*, 288.

334 Richard, D., MacRaild, C.A., Riglar, D.T., Chan, J.-A., Foley, M., Baum, J., Ralph, S.A., Norton, R.S., and Cowman, A.F. (2010). Interaction between Plasmodium falciparum apical membrane antigen 1 and the rhoptry neck protein complex defines a key step in the erythrocyte invasion process of malaria parasites. Journal of Biological Chemistry, jbc. M109. 080770.

335 Richards, J.S., and Beeson, J.G. (2009). The future for blood-stage vaccines against malaria. Immunology and cell biology *87*, 377-390.

336 Rieckmann, K., Davis, D., and Hutton, D. (1989). Plasmodium vivax resistance to chloroquine? The Lancet *334*, 1183-1184.

337 Riley, E., Allen, S., Wheeler, J., Blackman, M., Bennett, S., Takacs, B., SCHONFELD, H.J., Holder, A., and Greenwood, B. (1992). Naturally acquired cellular and humoral immune responses to the major merozoite surface antigen (Pf MSP1) of Plasmodium falciparum are associated with reduced malaria morbidity. Parasite immunology *14*, 321-337.

338 Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research *43*, e47-e47.

339 Ritchie, M.E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G.K. (2007). A comparison of background correction methods for two-colour microarrays. Bioinformatics *23*, 2700-2707.

340 Robinson, L.J., Wampfler, R., Betuela, I., Karl, S., White, M.T., Suen, C.S.L.W., Hofmann, N.E., Kinboro, B., Waltmann, A., and Brewster, J. (2015). Strategies for understanding and reducing the Plasmodium vivax and Plasmodium ovale hypnozoite reservoir in Papua New Guinean children: a randomised placebo-controlled trial and mathematical model. PLoS medicine *12*, e1001891.

341 Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

342 Rodrigues-da-Silva, R.N., da Silva, J.H.M., Singh, B., Jiang, J., Meyer, E.V., Santos, F., Banic, D.M., Moreno, A., Galinski, M.R., and Oliveira-Ferreira, J. (2016). In silico identification and validation of a linear and naturally immunogenic B-cell epitope of the Plasmodium vivax malaria vaccine candidate merozoite surface protein-9. PloS one *11*, e0146951.

343 Romphruk, A., Puapairoj, C., Romphruk, A., Barasrux, S., and Leelayuwat, Y.U. (1999). Distributions of HLA-DRB1/DQB1 alleles and haplotypes in the North-eastern Thai population: indicative of a distinct Thai population with

Chinese admixtures in the Central Thais. European journal of immunogenetics *26*, 129-133.

344 Rozas, J., and Rozas, R. (1997). DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Computer Applications in the Biosciences *13*, 307-311.

345 Rts, S. (2015). Efficacy and safety of RTS, S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. The Lancet *386*, 31-45.

346 RTS, S.C.T.P. (2011). First results of phase 3 trial of RTS, S/AS01 malaria vaccine in African children. New England Journal of Medicine *365*, 1863-1875.

347 RTS, S.C.T.P. (2012). A phase 3 trial of RTS, S/AS01 malaria vaccine in African infants. New England Journal of Medicine *367*, 2284-2295.

348 Rungsihirunrat, K., Chaijaroenkul, W., Siripoon, N., Seugorn, A., and Na-Bangchang, K. (2011). Genotyping of polymorphic marker (MSP3α and MSP3β) genes of Plasmodium vivax field isolates from malaria endemic of Thailand. Tropical Medicine & International Health *16*, 794-801.

349 Sabchareon, A., Burnouf, T., Ouattara, D., Attanath, P., Bouharoun-Tayoun, H., Chantavanich, P., Foucault, C., Chongsuphajaisiddhi, T., and Druilhe, P. (1991). Parasitologic and clinical human response to immunoglobulin administration in falciparum malaria. The American journal of tropical medicine and hygiene *45*, 297-308.

350 Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution *4*, 406-425.

351 Saxena, A.K., Wu, Y., and Garboczi, D.N. (2007). Plasmodium p25 and p28 surface proteins: potential transmission-blocking vaccines. Eukaryotic cell *6*, 1260-1265.

352 Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics *78*, 629-644.

353 Scheiblhofer, S., Laimer, J., Machado, Y., Weiss, R., and Thalhamer, J. (2017). Influence of protein fold stability on immunogenicity and its implications for vaccine design. Expert review of vaccines *16*, 479-489.

354 Schofield, L., and Grau, G.E. (2005). Immunological processes in malaria pathogenesis. Nature Reviews Immunology *5*, 722.

355 Schüffner, W. (1938). Two subjects relating to the epidemiology of malaria. Journal of the Malaria Institute of India *1*.

356 Sedwick, C. (2014). Plasmepsin V, a secret weapon against malaria. PLoS biology *12*, e1001898.

357 Seow, J., Morales, R.A., MacRaild, C.A., Krishnarjuna, B., McGowan, S., Dingjan, T., Jaipuria, G., Rouet, R., Wilde, K.L., and Atreya, H.S. (2017). Structure and characterisation of a key epitope in the conserved C-terminal domain of the malaria vaccine candidate MSP2. Journal of molecular biology *429*, 836-846.

358 Seyednasrollah, F., Laiho, A., and Elo, L.L. (2013). Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in bioinformatics *16*, 59-70.

359 Shah, N.K., Dhillon, G.P., Dash, A.P., Arora, U., Meshnick, S.R., and Valecha, N. (2011). Antimalarial drug resistance of Plasmodium falciparum in India: changes over time and space. The Lancet infectious diseases *11*, 57-64.

360 Shen, H.-M., Chen, S.-B., Wang, Y., Xu, B., Abe, E.M., and Chen, J.-H. (2017). Genome-wide scans for the identification of Plasmodium vivax genes under positive selection. Malaria journal *16*, 238.

361 Shimodaira, H., and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular biology and evolution *16*, 1114-1114.

362 Singh, S., Miura, K., Zhou, H., Muratova, O., Keegan, B., Miles, A., Martin, L.B., Saul, A.J., Miller, L.H., and Long, C.A. (2006). Immunity to recombinant Plasmodium falciparum merozoite surface protein 1 (MSP1): protection in Aotus nancymai monkeys strongly correlates with anti-MSP1 antibody titer and in vitro parasite-inhibitory activity. Infection and immunity *74*, 4573-4580.

363 Singh, S., Soe, S., Weisman, S., Barnwell, J.W., Pérignon, J.L., and Druilhe, P. (2009). A conserved multi-gene family induces cross-reactive antibodies effective in defense against Plasmodium falciparum. PloS one *4*, e5410.

364 Sirima, S.B., Cousens, S., and Druilhe, P. (2011). Protection against malaria by MSP3 candidate vaccine. New England Journal of Medicine *365*, 1062-1064.

365 Smyth, G.K. (2005). Limma: linear models for microarray data. In Bioinformatics and computational biology solutions using R and Bioconductor (Springer), pp. 397-420.

366 Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology *3*, 1-25.

367 Snow, R.W., Omumbo, J.A., Lowe, B., Molyneux, C.S., Obiero, J.-O., Palmer, A., Weber, M.W., Pinder, M., Nahlen, B., and Obonyo, C. (1997). Relation between severe malaria morbidity in children and level of Plasmodium falciparum transmission in Africa. The Lancet *349*, 1650-1654.

368 Soares, L.A., Evangelista, J., Orlandi, P.P., Almeida, M.E., Sousa, L.P.d., Chaves, Y., Barbosa-Filho, R., Lacerda, M.V., Mariuba, L.A., and Nogueira, P.A. (2014). Genetic diversity of MSP1 Block 2 of Plasmodium vivax isolates from Manaus (central Brazilian Amazon). Journal of immunology research *2014*.

369 Sobota, R.S., Shriner, D., Kodaman, N., Goodloe, R., Zheng, W., Gao, Y.T., Edwards, T.L., Amos, C.I., and Williams, S.M. (2015). Addressing population-specific multiple testing burdens in genetic association studies. Annals of human genetics *79*, 136-147.

370 Soema, P.C., van Riet, E., Kersten, G., and Amorij, J.-P. (2015). Development of cross-protective influenza A vaccines based on cellular responses. Frontiers in immunology *6*, 237.

371 Sormanni, P., Aprile, F.A., and Vendruscolo, M. (2015). Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins. Proceedings of the National Academy of Sciences *112*, 9902-9907.

372 Spaccapelo, R., Aime, E., Caterbi, S., Arcidiacono, P., Capuccini, B., Di Cristina, M., Dottorini, T., Rende, M., Bistoni, F., and Crisanti, A. (2011). Disruption of plasmepsin-4 and merozoites surface protein-7 genes in Plasmodium berghei induces combined virulence-attenuated phenotype. Scientific reports *1*, 39.

373 Sriprawat, K., Kaewpongsri, S., Suwanarusk, R., Leimanis, M.L., Phyo, A.P., Snounou, G., Russell, B., Renia, L., and Nosten, F. (2009). Effective and cheap removal of leukocytes and platelets from Plasmodium vivax infected blood. Malaria journal *8*, 1.

374 Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312-1313.

375 Stanisic, D.I., Fowkes, F.J., Koinari, M., Javati, S., Lin, E., Kiniboro, B., Richards, J.S., Robinson, L.J., Schofield, L., and Kazura, J.W. (2015). Acquisition of antibodies against Plasmodium falciparum merozoites and malaria immunity in young children and the influence of age, force of infection, and magnitude of response. Infection and immunity *83*, 646-660.

376 Stanisic, D.I., Richards, J.S., McCallum, F.J., Michon, P., King, C.L., Schoepflin, S., Gilson, P.R., Murphy, V.J., Anders, R.F., and Mueller, I. (2009). Immunoglobulin G subclass-specific responses against Plasmodium falciparum merozoite antigens are associated with control of parasitemia and protection from symptomatic illness. Infection and immunity *77*, 1165-1174.

377 Stanley, J. (1997). Malaria. Emerg Med Clin North Am *15*, 113-155.

378 Steinbuechel, M., and Matuschewski, K. (2009). Role for the Plasmodium sporozoite-specific transmembrane protein S6 in parasite motility and efficient malaria transmission. Cellular microbiology *11*, 279-288.

379 Suankratay, C., Wilde, H., and Berger, S. (2001). Thailand: country survey of infectious diseases. Journal of travel medicine *8*, 192-203.

380 Subudhi, A.K., Boopathi, P.A., Pandey, I., Kaur, R., Middha, S., Acharya, J., Kochar, S.K., Kochar, D.K., and Das, A. (2015). Disease specific modules and hub genes for intervention strategies: A co-expression network based approach for Plasmodium falciparum clinical isolates. Infection, Genetics and Evolution *35*, 96-108.

381 Sultan, A.A., Thathy, V., Frevert, U., Robson, K.J., Crisanti, A., Nussenzweig, V., Nussenzweig, R.S., and Ménard, R. (1997). TRAP is necessary for gliding motility and infectivity of Plasmodium sporozoites. Cell *90*, 511-522.

382 Sun, C., and Zhou, B. (2016). The molecular and cellular action properties of artemisinins: what has yeast told us? Microbial Cell *3*, 196.

383 Suwonkerd, W., Ritthison, W., Ngo, C.T., Tainchum, K., Bangs, M.J., and Chareonviriyaphap, T. (2013). Vector biology and malaria transmission in Southeast Asia. In Anopheles mosquitoes-New insights into malaria vectors (InTech).

384 Suzuki, Y., and Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. Molecular biology and evolution *16*, 1315-1328.

385 Swearingen, K.E., Lindner, S.E., Shi, L., Shears, M.J., Harupa, A., Hopp, C.S., Vaughan, A.M., Springer, T.A., Moritz, R.L., and Kappe, S.H. (2016). Interrogating the Plasmodium sporozoite surface: identification of surface-exposed proteins and demonstration of glycosylation on CSP and TRAP by mass spectrometry-based proteomics. PLoS pathogens *12*, e1005606.

386 Taiyun, W., and Viliam, S. (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84).

387 Takala, S.L., and Plowe, C.V. (2009). Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'. Parasite immunology *31*, 560-573.

388 Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular biology and evolution *30*, 2725-2729.

389 Tarun, A.S., Dumpit, R.F., Camargo, N., Labaied, M., Liu, P., Takagi, A., Wang, R., and Kappe, S.H. (2007). Protracted sterile protection with Plasmodium yoelii pre-erythrocytic genetically attenuated parasite malaria vaccines is independent of significant liver-stage persistence and is mediated by CD8+ T cells. The Journal of infectious diseases *196*, 608-616.

390 Taylor, J.E., Pacheco, M.A., Bacon, D.J., Beg, M.A., Machado, R.L.D., Fairhurst, R.M., Herrera, S., Kim, J.-Y., Menard, D., and Póvoa, M.M. (2013). The evolutionary history of Plasmodium vivax as inferred from mitochondrial

genomes: parasite genetic diversity in the Americas. Molecular biology and evolution, mst104.

391 Taylor, R.R., Allen, S.J., Greenwood, B.M., and Riley, E.M. (1998). IgG3 antibodies to Plasmodium falciparum merozoite surface protein 2 (MSP2): increasing prevalence with age and association with clinical immunity to malaria. The American journal of tropical medicine and hygiene *58*, 406-413.

392 Ten Hagen, T., Sulzer, A.J., Kidd, M.R., Lal, A.A., and Hunter, R.L. (1993). Role of adjuvants in the modulation of antibody isotype, specificity, and induction of protection by whole blood-stage Plasmodium yoelii vaccines. The Journal of Immunology *151*, 7077-7085.

393 Tewari, R., Ogun, S.A., Gunaratne, R.S., Crisanti, A., and Holder, A.A. (2005). Disruption of Plasmodium berghei merozoite surface protein 7 gene modulates parasite growth in vivo. Blood *105*, 394-396.

394 Thankaswamy-Kosalai, S., Sen, P., and Nookaew, I. (2017). Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. Genomics *109*, 186-191.

395 Thera, M.A., Doumbo, O.K., Coulibaly, D., Diallo, D.A., Kone, A.K., Guindo, A.B., Traore, K., Dicko, A., Sagara, I., and Sissoko, M.S. (2008). Safety and immunogenicity of an AMA-1 malaria vaccine in Malian adults: results of a phase 1 randomized controlled trial. PloS one *3*, e1465.

396 Thera, M.A., Doumbo, O.K., Coulibaly, D., Diallo, D.A., Sagara, I., Dicko, A., Diemert, D.J., Heppner Jr, D.G., Stewart, V.A., and Angov, E. (2006). Safety and allele-specific immunogenicity of a malaria vaccine in Malian adults: results of a phase I randomized trial. PLoS Clinical Trials *1*, e34.

397 Thera, M.A., Doumbo, O.K., Coulibaly, D., Laurens, M.B., Kone, A.K., Guindo, A.B., Traore, K., Sissoko, M., Diallo, D.A., and Diarra, I. (2010). Safety and immunogenicity of an AMA1 malaria vaccine in Malian children: results of a phase 1 randomized controlled trial. PloS one *5*, e9041.

398 Thera, M.A., and Plowe, C.V. (2012). Vaccines for malaria: how close are we? Annual review of medicine *63*, 345-357.

399 Thimasarn, K., Jatapadma, S., Vijaykadga, S., Sirichaisinthop, J., and Wongsrichanalai, C. (1995). Epidemiology of malaria in Thailand. Journal of Travel Medicine *2*, 59-65.

400 Thimasarn, K., Sirichaisinthop, J., Chanyakhun, P., Palananth, C., and Rooney, W. (1997). A comparative study of artesunate and artemether in combination with mefloquine on multidrug resistant falciparum malaria in eastern Thailand. The Southeast Asian journal of tropical medicine and public health *28*, 465-471.

401 Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through

sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research *22*, 4673-4680.

402 Tian, S., Yan, H., Neuhauser, C., and Slager, S.L. (2016). An analytical workflow for accurate variant discovery in highly divergent regions. BMC genomics *17*, 703.

403 Torkamaneh, D., Laroche, J., and Belzile, F. (2016). Genome-wide SNP calling from genotyping by sequencing (GBS) data: a comparison of seven pipelines and two sequencing technologies. PLoS One *11*, e0161333.

404 Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols *7*, 562.

405 Trapnell, C., and Salzberg, S.L. (2009). How to map billions of short reads onto genomes. Nature biotechnology *27*, 455.

406 Tsuji, M. (2010). A retrospective evaluation of the role of T cells in the development of malaria vaccine. Experimental parasitology *126*, 421-425.

407 Urquiza, M., RODRIGUEZ, L.E., SUAREZ, J.E., GUZMÁN, F., OCAMPO, M., CURTIDOR, H., SEGURA, C., TRUJILLO, E., and PATARROYO, M.E. (1996). Identification of Plasmodium falciparum MSP-1 peptides able to bind to human red blood cells. Parasite immunology *18*, 515-526.

408 Uversky, V.N., and Dunker, A.K. (2013). The case for intrinsically disordered proteins playing contributory roles in molecular recognition without a stable 3D structure. F1000 biology reports *5*.

409 Valencia, S.H., Rodríguez, D.C., Acero, D.L., Ocampo, V., and Arévalo-Herrera, M. (2011). Platform for Plasmodium vivax vaccine discovery and development. Memorias do Instituto Oswaldo Cruz *106*, 179-192.

410 van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J.P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. Briefings in bioinformatics, bbw139.

411 Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., and Thibault, J. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics *43*, 11.10. 11-11.10. 33.

412 Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., and Jones, D.T. (2014). Classification of intrinsically disordered regions and proteins. Chemical reviews *114*, 6589-6631.

413 van Dijk, M.R., Janse, C.J., Thompson, J., Waters, A.P., Braks, J.A., Dodemont, H.J., Stunnenberg, H.G., van Gemert, G.-J., Sauerwein, R.W., and Eling, W. (2001). A central role for P48/45 in malaria parasite male gamete fertility. Cell *104*, 153-164.

414 Venkatesan, M., Amaratunga, C., Campino, S., Auburn, S., Koch, O., Lim, P., Uk, S., Socheat, D., Kwiatkowski, D.P., and Fairhurst, R.M. (2012). Using CF11 cellulose columns to inexpensively and effectively remove human DNA from Plasmodium falciparum-infected whole blood samples. Malaria journal *11*, 41.

415 Véron, V., Legrand, E., Yrinesi, J., Volney, B., Simon, S., and Carme, B. (2009). Genetic diversity of msp3α and msp1 _b5 markers of Plasmodium vivax in French Guiana. Malaria journal *8*, 40.

416 Villard, V., Agak, G.W., Frank, G., Jafarshad, A., Servis, C., Nébié, I., Sirima, S.B., Felger, I., Arevalo-Herrera, M., and Herrera, S. (2007). Rapid identification of malaria vaccine candidates based on α-helical coiled coil protein motif. PloS one *2*, e645.

417 Vita, R., Overton, J.A., Greenbaum, J.A., Ponomarenko, J., Clark, J.D., Cantrell, J.R., Wheeler, D.K., Gabbard, J.L., Hix, D., and Sette, A. (2014). The immune epitope database (IEDB) 3.0. Nucleic acids research *43*, D405-D412.

418 Wang, R., Doolan, D.L., Le, T.P., Hedstrom, R.C., Coonan, K.M., Charoenvit, Y., Jones, T.R., Hobart, P., Margalith, M., and Ng, J. (1998). Induction of antigen-specific cytotoxic T lymphocytes in humans by a malaria DNA vaccine. Science *282*, 476-480.

419 Wang, R., Smith, J.D., and Kappe, S.H. (2009). Advances and challenges in malaria vaccine development. Expert reviews in molecular medicine *11*.

420 Wang, S., Li, W., Liu, S., and Xu, J. (2016a). RaptorX-Property: a web server for protein structure property prediction. Nucleic acids research *44*, W430-W435.

421 Wang, S., Peng, J., Ma, J., and Xu, J. (2016b). Protein secondary structure prediction using deep convolutional neural fields. Scientific reports *6*, 18962.

422 Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.

423 Weaver, R., Reiling, L., Feng, G., Drew, D.R., Mueller, I., Siba, P.M., Tsuboi, T., Richards, J.S., Fowkes, F.J., and Beeson, J.G. (2016). The association between naturally acquired IgG subclass specific antibodies to the PfRH5 invasion complex and protection from Plasmodium falciparum malaria. Scientific reports *6*, 33094.

424 Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. evolution *38*, 1358-1370.

425 Westenberger, S.J., McClean, C.M., Chattopadhyay, R., Dharia, N.V., Carlton, J.M., Barnwell, J.W., Collins, W.E., Hoffman, S.L., Zhou, Y., and Vinetz, J.M. (2010). A systems-based analysis of Plasmodium vivax lifecycle transcription from human to mosquito. PLoS neglected tropical diseases *4*, e653.

426 WHO (2017). World malaria report 2017 (Luxembourg: World Health Organization).

427 Williams, T.N. (2006). Human red blood cell polymorphisms and malaria. Current opinion in microbiology *9*, 388-394.

428 Williamson, K.C., Keister, D.B., Muratova, O., and Kaslow, D.C. (1995). Recombinant Pfs230, a Plasmodium falciparum gametocyte protein, induces antisera that reduce the infectivity of Plasmodium falciparum to mosquitoes. Molecular and biochemical parasitology *75*, 33-42.

429 Willing, E.-M., Dreyer, C., and Van Oosterhout, C. (2012). Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. PloS one *7*, e42649.

430 Wilson, D.W., Goodman, C.D., Sleebs, B.E., Weiss, G.E., de Jong, N.W., Angrisano, F., Langer, C., Baum, J., Crabb, B.S., and Gilson, P.R. (2015). Macrolides rapidly inhibit red blood cell invasion by the human malaria parasite, Plasmodium falciparum. BMC biology *13*, 52.

431 Winter, D.J., Pacheco, M.A., Vallejo, A.F., Schwartz, R.S., Arevalo-Herrera, M., Herrera, S., Cartwright, R.A., and Escalante, A.A. (2015). Whole genome sequencing of field isolates reveals extensive genetic diversity in Plasmodium vivax from Colombia. PLoS neglected tropical diseases *9*, e0004252.

432 Wipasa, J., Okell, L., Sakkhachornphop, S., Suphavilai, C., Chawansuntati, K., Liewsaree, W., Hafalla, J.C., and Riley, E.M. (2011). Short-lived IFN-γ effector responses, but long-lived IL-10 memory responses, to malaria in an area of low malaria endemicity. PLoS pathogens *7*, e1001281.

433 Woehlbier, U., Epp, C., Hackett, F., Blackman, M.J., and Bujard, H. (2010). Antibodies against multiple merozoite surface antigens of the human malaria parasite Plasmodium falciparum inhibit parasite maturation and red blood cell invasion. Malaria journal *9*, 77.

434 Woehlbier, U., Epp, C., Kauth, C.W., Lutz, R., Long, C.A., Coulibaly, B., Kouyaté, B., Arevalo-Herrera, M., Herrera, S., and Bujard, H. (2006). Analysis of antibodies directed against merozoite surface protein 1 of the human malaria parasite Plasmodium falciparum. Infection and immunity *74*, 1313-1322.

435 Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. Technometrics *20*, 397-405.

436 Woolley, S., Johnson, J., Smith, M.J., Crandall, K.A., and McClellan, D.A. (2003). TreeSAAP: selection on amino acid properties using phylogenetic trees. Bioinformatics *19*, 671-672.

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

438 Wright, P.E., and Dyson, H.J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. Nature reviews Molecular cell biology *16*, 18.

439 Wu, Y., Ellis, R.D., Shaffer, D., Fontes, E., Malkin, E.M., Mahanty, S., Fay, M.P., Narum, D., Rausch, K., and Miles, A.P. (2008). Phase 1 trial of malaria transmission blocking vaccine candidates Pfs25 and Pvs25 formulated with montanide ISA 51. PloS one *3*, e2636.

440 Yagi, M., Bang, G., Tougan, T., Palacpac, N.M., Arisue, N., Aoshi, T., Matsumoto, Y., Ishii, K.J., Egwang, T.G., and Druilhe, P. (2014). Protective epitopes of the Plasmodium falciparum SERA5 malaria vaccine reside in intrinsically unstructured N-terminal repetitive sequences. PLoS One *9*, e98460.

441 Yam, X.Y., Brugat, T., Siau, A., Lawton, J., Wong, D.S., Farah, A., Twang, J.S., Gao, X., Langhorne, J., and Preiser, P.R. (2016). Characterization of the Plasmodium Interspersed Repeats (PIR) proteins of Plasmodium chabaudi indicates functional diversity. Scientific reports *6*, 23449.

442 Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular biology and evolution *17*, 32-43.

443 Yoon, S., and Nam, D. (2017). Gene dispersion is the key determinant of the read count bias in differential expression analysis of RNA-seq data. BMC genomics *18*, 408.

444 Yu, F.-D., Yang, S.-Y., Li, Y.-Y., and Hu, W. (2013). Co-expression network with protein–protein interaction and transcription regulation in malaria parasite Plasmodium falciparum. Gene *518*, 7-16.

445 Yung, A., and Bennett, N.M. (1976). Chloroquine-resistant falciparum malaria in Papua New Guinea. Medical Journal of Australia, 320-321.

446 Zakeri, S., Barjesteh, H., and Djadid, N.D. (2006). Merozoite surface protein-3α is a reliable marker for population genetic analysis of Plasmodium vivax. Malaria journal *5*, 53.

447 Zavala, F., Cochrane, A.H., Nardin, E.H., Nussenzweig, R.S., and Nussenzweig, V. (1983). Circumsporozoite proteins of malaria parasites contain a single immunodominant region with two or more identical epitopes. Journal of experimental medicine *157*, 1947-1957.

448 Zepp, F. (2010). Principles of vaccine design—lessons from nature. Vaccine *28*, C14-C24.

449 Zevering, Y., Amante, F., Smillie, A., Currier, J., Smith, G., Houghten, R.A., and Good, M.F. (1992). High frequency of malaria-specific T cells in non-exposed humans. European journal of immunology *22*, 689-696.

450 Zhao, Q.-Y., Gratten, J., Restuadi, R., and Li, X. (2016). Mapping and differential expression analysis from short-read RNA-Seq data in model organisms. Quantitative Biology *4*, 22-35.

451 Zhao, S., Zhang, Y., Gamini, R., Zhang, B., and Schack, D. (2018). Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. Scientific reports *8*, 4781.

452 Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326-3328.

453 Zhou, G., Sirichaisinthop, J., Sattabongkot, J., Jones, J., Bjornstad, O.N., Yan, G., and Cui, L. (2005). Spatio-temporal distribution of *Plasmodium falciparum* and *Plasmodium vivax* malaria in Thailand. Am J Trop Med Hyg *72*, 256-262.

454 Zhu, L., Mok, S., Imwong, M., Jaidee, A., Russell, B., Nosten, F., Day, N.P., White, N.J., Preiser, P.R., and Bozdech, Z. (2016). New insights into the Plasmodium vivax transcriptome using RNA-Seq. Scientific reports *6*, 20498.

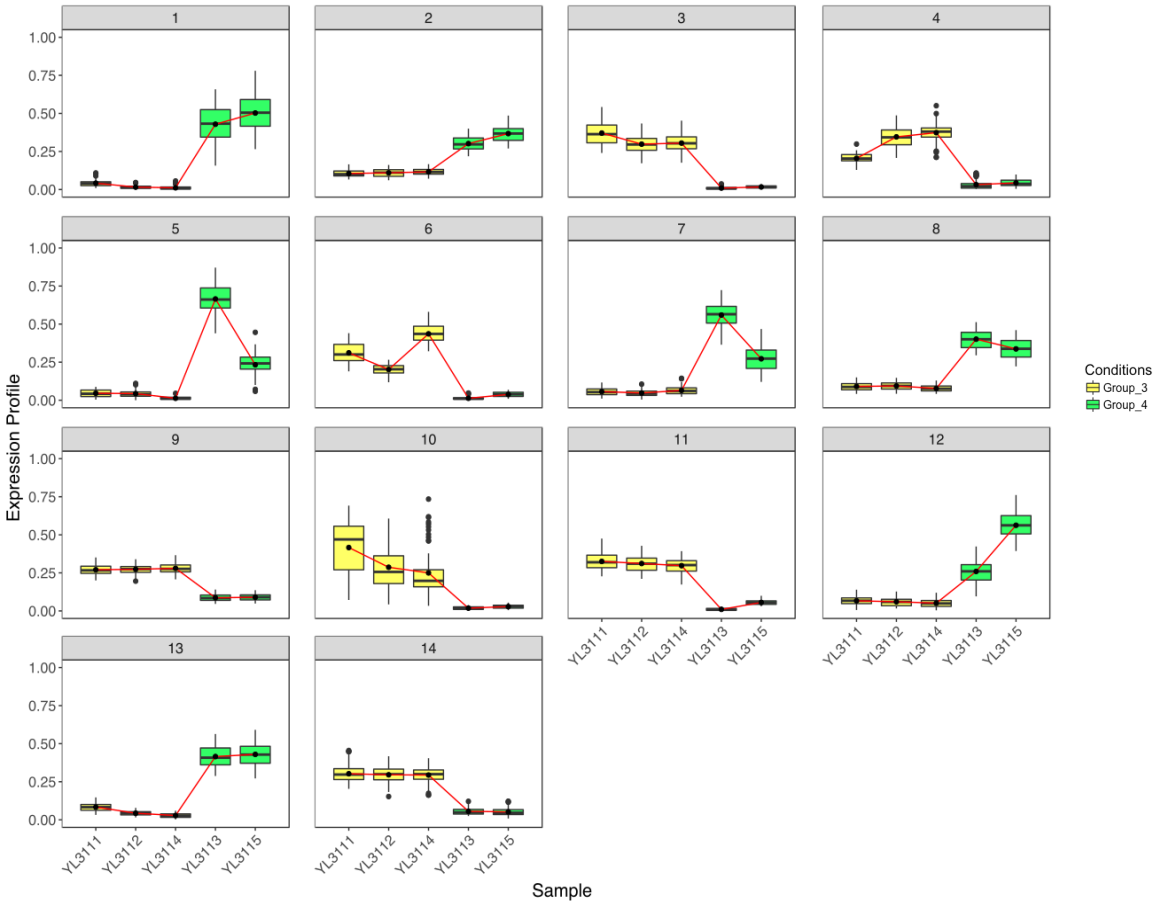
Appendix

Supplementary Table 1. 48 sequences retrieved from National Center for Biotechnology Information (NCBI) database. NA: not available.

No.	Short Read Archive Identifier	Country	Location	Year of collection	Sequencing approach
1	ERR018032	Brazil	NA	2008	Illumina
2	ERR019040	Brazil	NA	2008	Illumina
3	ERR152407	Brazil	NA	2008	Illumina
4	ERR020103	Cambodia	NA	2010	Illumina
5	ERR023039	Cambodia	NA	2010	Illumina
6	ERR023040	Cambodia	NA	2010	Illumina
7	ERR023041	Cambodia	NA	2010	Illumina
8	ERR054080	Cambodia	NA	2010	Illumina
9	ERR054082	Cambodia	NA	2010	Illumina
10	ERR152408	Cambodia	NA	2010	Illumina
11	ERR211549	Cambodia	NA	2010	Illumina
12	ERR211561	Cambodia	NA	2010	Illumina
13	ERR23042	Cambodia	NA	2010	Illumina
14	ERR054088	Malaysia	NA	2011	Illumina
15	ERR054089	Malaysia	NA	2011	Illumina
16	ERR152414	Malaysia	NA	2011	Illumina
17	ERR152415	Malaysia	NA	2011	Illumina
18	ERR527337	Malaysia	NA	2013	Illumina
19	ERR527363	Malaysia	NA	2013	Illumina
20	SRR1564966	Myanmar	Kachin State	2012/2013	Illumina
21	SRR1564977	Myanmar	Kachin State	2012/2013	Illumina
22	SRR1565050	Myanmar	Kachin State	2012/2013	Illumina
23	SRR1565088	Myanmar	Kachin State	2012/2013	Illumina
24	SRR1568013	Myanmar	Kachin State	2012/2013	Illumina

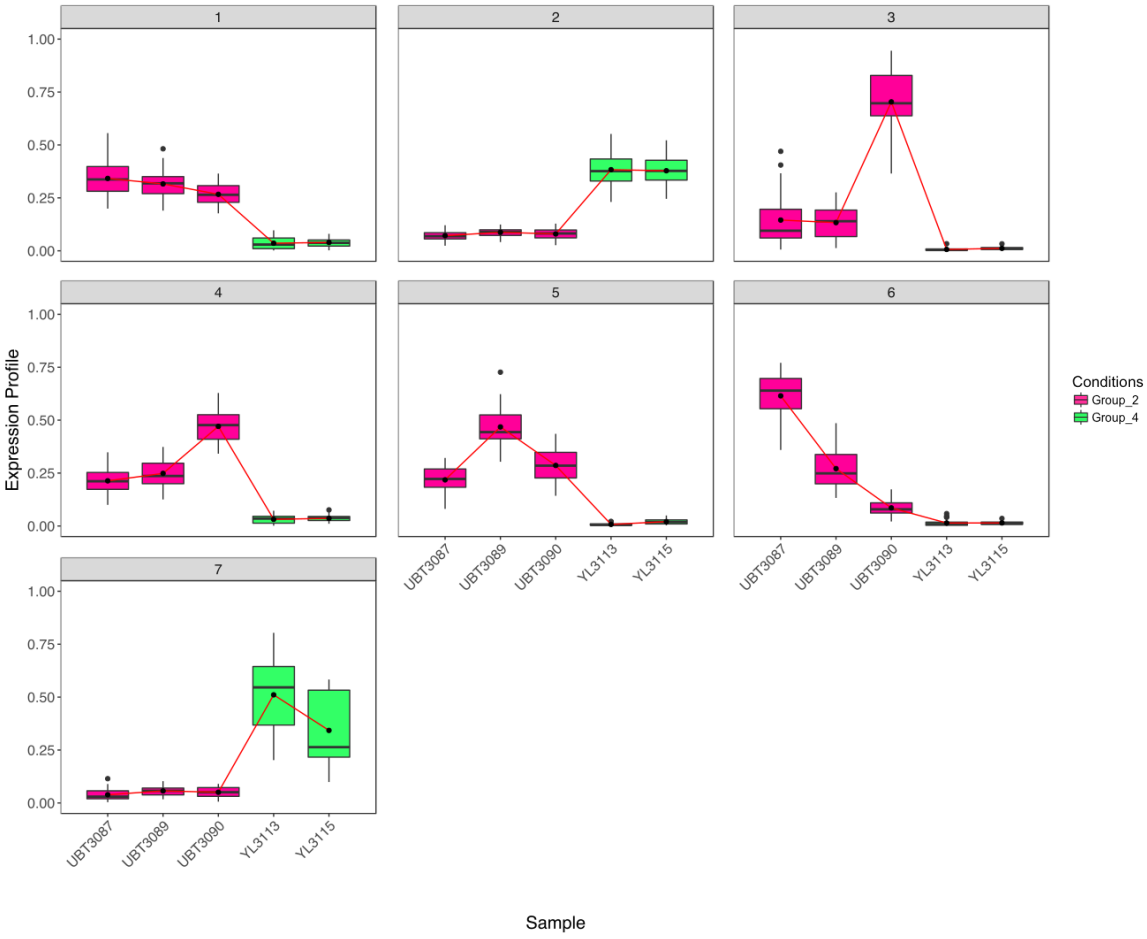
25	SRR1568114	Myanmar	Kachin State	2012/2013	Illumina
26	SRR1568120	Myanmar	Kachin State	2012/2013	Illumina
27	SRR1568204	Myanmar	Kachin State	2012/2013	Illumina
28	ERR404246	Thailand	East	2013	Illumina
29	ERR426015	Thailand	East	2012	Illumina
30	ERR426033	Thailand	East	2013	Illumina
31	ERR428035	Thailand	East	2013	Illumina
32	ERR 111728	Thailand	West	2007	Illumina
33	ERR111709	Thailand	West	2011	Illumina
34	ERR111710	Thailand	West	2007	Illumina
35	ERR111713	Thailand	West	2013	Illumina
36	ERR111714	Thailand	West	2011	Illumina
37	ERR111716	Thailand	West	2011	Illumina
38	ERR111717	Thailand	West	2006	Illumina
39	ERR111719	Thailand	West	2007	Illumina
40	ERR111720	Thailand	West	2012	Illumina
41	ERR111721	Thailand	West	2012	Illumina
42	ERR111727	Thailand	West	2011	Illumina
43	SRR1562845	Thailand	West	2012	Illumina
44	SRR1562959	Thailand	West	2013	Illumina
45	SRR1562962	Thailand	West	2012	Illumina
46	SRR1568148	Thailand	West	2013	Illumina
47	SRR1568161	Thailand	West	2012	Illumina
48	SRR1568209	Thailand	West	2013	Illumina

Supplementary Figure 1



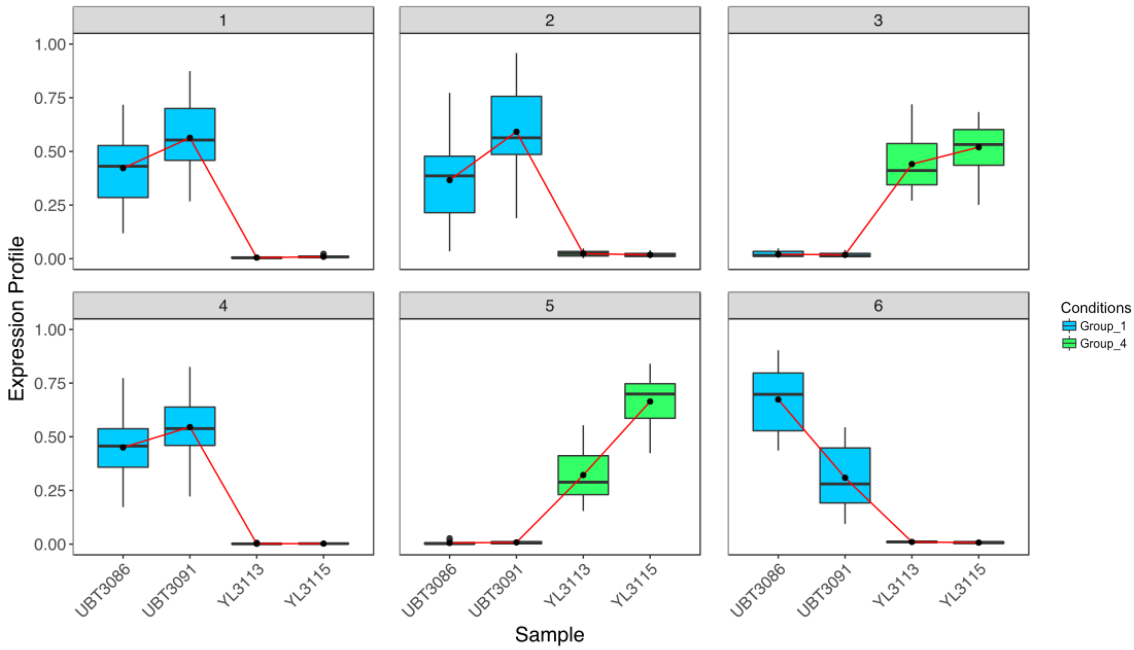
Supplementary Figure 1. Co-expression analysis between Group 3 and Group 4. Genes are clustered into 14 groups from the *coseq* package implemented in R. The input normalised reads were derived from the *DESeq2* package. The boxplot depicts each individual, and the colour represents each group of patients. The genes in each cluster are listed in S5.2.

Supplementary Figure 2



Supplementary Figure 2. Co-expression analysis between Group 2 and Group 4. Genes are clustered into seven groups from the *coseq* package implemented in R. The input normalised reads were derived from the *DESeq2* package. The boxplot depicts each individual, and the colour represents each group of patients. The genes in each cluster are listed in S5.2.

Supplementary Figure 3



Supplementary Figure 3. Co-expression analysis between Group 1 and Group 4. Genes are clustered into six groups from the *coseq* package implemented in R. The input normalised reads were derived from the *DESeq2* package. The boxplot depicts each individual, and the colour represents each group of patients. The genes in each cluster are listed in S5.2.

Supplementary legends

S5.1 Full table of differentially expressed genes between each group of patients. The spreadsheet contains DEGs of pairwise comparison between Group 1 and 2, Group 1 and 3, Group 1 and 4, Group 2 and 3, Group 2 and 4, and Group 3 and Group 4. DEGs with FDR<0.05 were considered significantly differentially expressed. The DEGs were identified using DESeq2.

S5.2 Full table of co-expressed genes in each cluster derived from a pairwise comparison of DEGs between Group 3 and 4, Group 2 and 4, and Group1 and 4. The analysis was conducted using *coseq* implemented in R. The table contains the identifier, gene name, and product description.

S6.1 Full table of differentially detected peptides in five groups of patients. The differentially detected peptides were identified from the pairwise comparison of each group of patients to negative control. The analysis was performed using *t*-statistics implemented in LIMMA package. The table contains differentially detected peptides in each group of patients, log fold-change, average expression, and the adjusted *p*-value. The peptides with adjusted *p*-value < 0.05 were considered statistically significant.