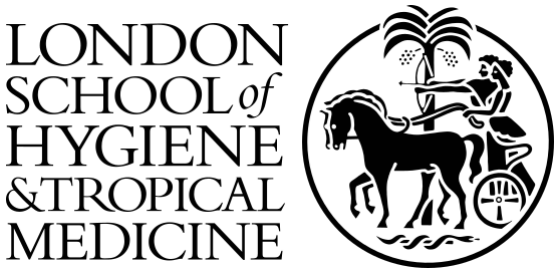

Applying next generation sequencing of genomes and transcriptomes to investigate the population structure and biology of *Plasmodium* species

Suzanne Elizabeth Hocking

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy

University of London
September 2019

Department of Infection Biology

Faculty of Infectious Tropical Diseases

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by the BBSRC

Research supervisor: Professor David Conway

Declaration of Work

I, Suzanne Elizabeth Hocking, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

I would firstly like to thank my supervisor, Prof. David Conway, for his continued support, encouragement and enthusiasm over the last three and a half years, and for helping to train me as an independent and curious scientist. Also to Prof. Damer Blake, my second supervisor, for his insights and guidance throughout my PhD. I am extremely grateful to Dr. Sarah Tarr for her support and interest in my research, for helping me develop my lab skills and for the occasional swapping of gardening tips. I am also extremely grateful to Ms. Lindsay Stewart, who has provided unending guidance and help, not just with skills like malaria culturing (which while I never learnt to love, I was at least capable of) and interpretation of results, but also for general support and encouragement, particularly when research was hard and uncooperative.

Thank you also to the other various members of the Conway group who have come and gone over the years, in particular to Dr. Ofelia Diaz, who was often a sympathetic ear when things went wrong and a lot of fun to have around in the lab. Dr. Paul Divis has been extremely helpful, providing me with sequencing and bioinformatics help at the beginning of my PhD, and on that note, thank you also to Ozan Gundogdu for maintaining the sequencing facility and Eloise Thompson for maintaining malaria culture lab and to Dr. Sam Alsford who has provided advice and support throughout my PhD as the Department Research Degrees Coordinator.

Thank you to the team at Singer Instruments for hosting me for my internship, but mostly for introducing me to Carl, who also gets a thank you for putting up with me over the last couple of years and for generally being a pretty great human being. Also, thanks to all the people at LIDo, Nadine Mogford who has cheered us all on from the sidelines and made sure everything has run smoothly over the last four years. My long-suffering housemate, Fiona, along with Caroline and Michael, we are forever bonded by our PhDs. Lindsay and Becky, we have supported each other throughout two degrees now, and their advice, jokes, and memes have been a welcome distraction over the last few years. I must also thank Emma, whose friendship over the last 20 years has helped me get to the point I'm at now and who has always been there when I've needed her. Finally, a thank you to my parents for everything they've done for me, for diligently helping me move house every year, and pretending to be interested when I explain the results of my latest experiment (that goes for Carl as well).

Abstract

Plasmodium parasites, the causative agents of malaria, are responsible for a significant burden of disease worldwide. Understanding populations and parasite biology is critical to developing effective control mechanisms and questions relating to these have been addressed in this thesis using whole genome and transcriptome sequencing. In Southeast Asia, the zoonotic species *Plasmodium knowlesi* now causes a significant amount of malarial disease, particularly in Malaysia. Clinical isolates of *P. knowlesi* from peninsular Malaysia were whole genome sequenced and single nucleotide polymorphisms were used to inform population genomic analyses. This work confirmed the existence of a third, divergent population of *P. knowlesi* in peninsular Malaysia and uncovered evidence of selection acting on these parasites. In sub-Saharan Africa, *P. falciparum* is responsible for virtually all malarial disease. Clinical isolates sampled from West Africa were investigated using whole transcriptome sequencing of *ex vivo* parasites, focusing on a gene known as *mspdbl2*, which is a possible marker of gametocytogenesis. Schizonts from clinical isolates were assessed for MSPDBL2 expression by immunofluorescence assays and were whole transcriptome sequenced. Analysis of gene expression was carried out correlating to expression of MSPDBL2, revealing enrichment of genes with known or suspected involvement in gametocytogenesis. A limiting factor of this investigation was the very low amount of material available from *ex vivo* culture. An alternative avenue for investigating limiting material is through single-cell sequencing. One method for single-cell transcriptomics was trialled in this thesis and extensive optimisation resulted in the sequencing of transcriptomes from a small number of *P. falciparum* schizonts, offering a future methodology for investigating gene expression using very few parasites.

Figures

Tables

Appendices

1. Introduction

1.1 Malaria remains a global disease

Despite extensive control attempts around the world, and elimination strategies launched in some selected locations, *Plasmodium* species continue to cause extensive malarial disease and death. In 2017 there were an estimated 219 million new malaria cases, reported from 87 countries, with sub-Saharan Africa holding a disproportionate burden of disease, accounting for over 90% of all malaria cases and deaths (World Health Organisation 2018). Whilst malaria remains a serious, global disease, its effect has been lessened over the years, with the number of malaria deaths being reduced from over 800,000 in 2000 to 435,000 in 2017 (Figure 1.1), although the rate of decrease of malarial disease in at risk populations has slowed in recent years. The use of insecticide-treated bed nets has been the cause of the greatest decline in malaria cases since 2000, although evidence suggests that other control strategies (residual indoor spraying and rapid diagnosis and treatment by artemisinin-combination therapy) also have an important impact on disease when they are implemented intensively (Bhatt et al. 2015). An additional and ongoing problem is the lack of an available and effective vaccine. Only one vaccine candidate has been through phase 3 trials, and has been shown to provide only moderate and short term protection, estimated to be effective at preventing ~40% of malaria cases for up to a year (RTS 2015).

Out of 87 malaria endemic countries, 43 reported an increase in the incidence of malarial disease between 2010 and 2017, most of which are in Africa (World Health Organisation 2018). This increase is mainly thought to be as a result of improved surveillance, reporting, and diagnostic techniques, but could also be the result of poor vector control and the limited use of effective diagnosis and treatment in many populations (World Health Organisation 2018). The development and spread of

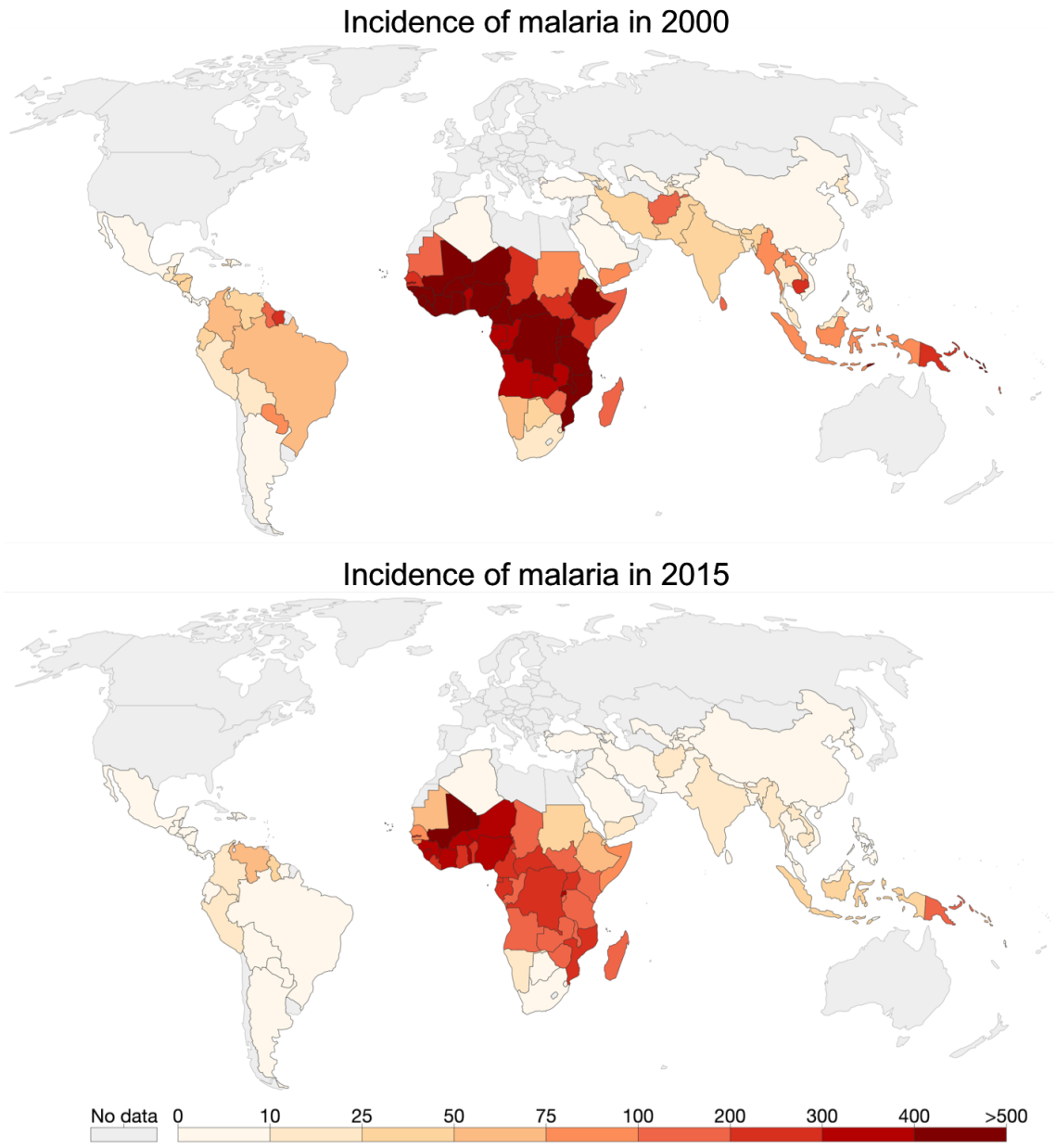

Figure 1.1. The incidence of malaria measured as the number of new cases of malaria per year per 1,000 people at risk between 2000 and 2015. Malaria incidence has steadily declined since 2000 but remains a global disease of substantial impact (ourworldindata.org/malaria).

antimalarial drug resistance in *Plasmodium* species, along with insecticide resistance in mosquitos pose increasing concerns for the re-emergence of disease (World Health Organization 2016). *Plasmodium falciparum* is the main cause of malaria in almost all endemic regions, with *P. vivax* being responsible for the majority of the remaining cases. Despite a 40% decrease in the incidence of *P. vivax* malaria since 2000, it still represents an important cause of disease (Battle et al. 2019). In South America, 75% of disease in endemic countries is caused by *P. vivax*. The other three malaria parasite species known to cause human disease, *P. malariae, P. ovale* (which comprises two cryptic species), and *P. knowlesi*, do so at a much lower incidence rate.

1.2 Life cycle of *Plasmodium* parasites

Plasmodium parasites have a complex life cycle requiring a vertebrate intermediate host and a mosquito definitive host which acts as the parasite vector (illustrated in Figure 1.2). Parasites undergo many rounds of asexual replication within the circulatory system of the vertebrate host, the portion of the life cycle that is responsible for symptomatic disease. Infection of the vertebrate host begins when an infected mosquito introduces *Plasmodium* sporozoites into the skin and they circulate through the lymphatic system until they reach the liver. Sporozoites enter hepatocytes in the liver and develop intracellularly over several days to form hepatocytic schizonts, within which asexual replication produces thousands of merozoites. These are polarised cells containing specialised organelles and structures at their apical ends to facilitate erythrocyte invasion. The merozoites are eventually released into the blood where they are able to invade host erythrocytes, initiating the asexual intraerythrocytic development cycle. Once invasion has occurred, the parasite exports hundreds of proteins that must traverse the parasitophorous vacuole – a membrane-bound compartment created during invasion that surrounds the parasite within the erythrocyte – to enter the host cell and many of

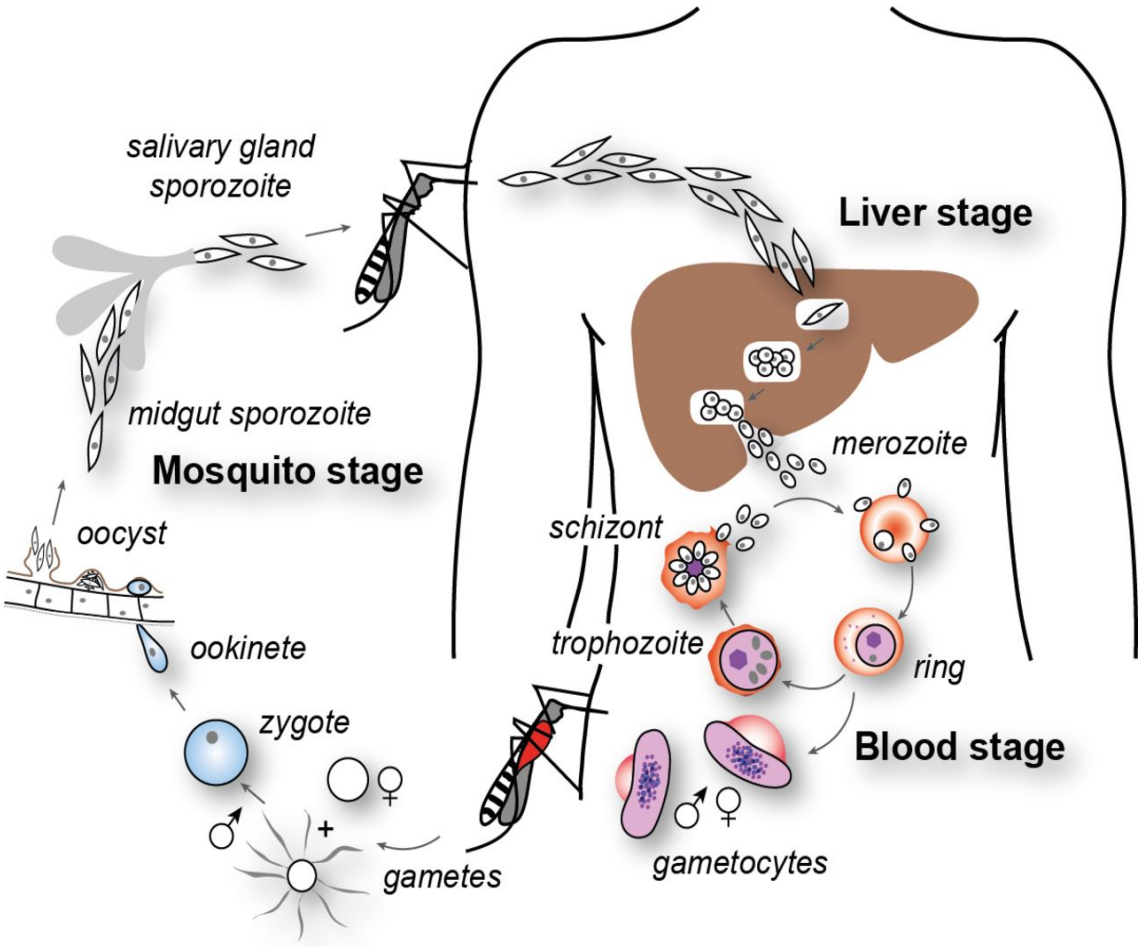

Figure 1.2. The life cycle of *Plasmodium* species. *Plasmodium* parasites infect a vertebrate intermediate host and mosquito vector definitive host. Injection of sporozoites into the vertebrate skin and vascular system begins the infection. Sporozoites move through the lymphatic system to the liver, where they enter hepatocytes and undergo round of nuclear division to produce thousands of merozoites ('Liver stage'). These are released back into the blood, where they circulate and undergo repeated rounds of asexual replication ('Blood stage'). Within each cycle, up to 32 daughter merozoites develop from a single mature schizont. A small proportion of parasites undergo gametocytogenesis and develop into male or female gametocytes. These are taken up by a mosquito taking a blood meal, where they undergo genetic recombination, resulting in the production of sporozoites, ready to infect another host. (Source: Cowman, Berry, & Baum, 2012).

these proteins are responsible for issuing extensive alterations to the cells (Lingelbach and Joiner 1998; Marti et al. 2004). The intraerythrocytic parasite now undergoes morphological development, first into a 'ring' stage early trophozoite before growing rapidly to develop into a haemozoin-sequestering late trophozoite. Trophozoites continue to grow, and undergo nuclear division to form multi-nucleate schizonts – the mature stage of the parasite's asexual life cycle. Nuclear division allows the formation of up to 32 merozoites within a single schizont, and these are released upon rupture of the host erythrocyte. Free merozoites enter the blood stream and will typically reinvade erythrocytes in less than five minutes (Gilson and Crabb 2009; Boyle et al. 2010). In many cases it is synchronised merozoite egress that is responsible for the periodic and cyclical nature of malarial disease symptoms. In addition, *P. falciparum* trophozoites and schizonts adhere to and sequester themselves within organ microvasculature in a process mediated by the polymorphic PfEMP1 proteins, encoded by the multi-gene family, *var*, which can lead to severe disease complications, such as cerebral malaria (MacKintosh et al. 2004). Cytoadhesion has also been reported for *P. vivax* and *P. berghei* parasites and proposed for *P. knowlesi*, although it may not be required for these species and the molecular mechanisms involved are still under investigation (Carvalho et al. 2010; Fatih et al. 2012; El-Assaad et al. 2013). The length of a single erythrocyte invasion cycle varies between species, and lasts for approximately 48 hours in *P. falciparum*, *P. vivax,* and *P. ovale*, 72 hours in *P. malariae*, and only 24 hours in *P. knowlesi*, whose quotidian life cycle can lead to rapid development of complications and severe malaria (Cox-Singh et al. 2008).

A small proportion of parasites within the asexual intraerythrocytic cycle will develop into male or female gametocytes rather than progressing through further rounds of asexual replication and it is these sexual parasites that are transmitted to mosquitos

during a blood meal, where they fertilise to form zygotes which undergo genetic recombination in the mosquito midgut during the only diploid portion of the parasite's life cycle.

1.3 *Plasmodium falciparum* is a major global cause of disease

Plasmodium falciparum is most prevalent malaria parasite species in most endemic regions except South America and parts of Asia, and is the most likely species to cause severe disease (Figure 1.3A). It disproportionately affects children under 5 years old, and in the WHO African region (Figure 1.3B) was responsible for 99.7% of malaria cases in 2017 (Figure 1.3A) (World Health Organisation 2018). In other WHO regions (Figure 1.3B), it was responsible for up to two thirds of all malarial disease. Since the year 2000, control mechanisms and rapid treatment with artemisinin combination therapy (ACT) have reduced the incidence of *P. falciparum* malarial disease by ~40% in Africa (Bhatt et al. 2015), where the number of deaths due to *P. falciparum* malaria are estimated to have dropped by 57% since 2000 (Gething et al. 2016). In the two decades prior to that, Africa saw a probable increase in mortality rates as resistance to the then front line anti-malarial drug (chloroquine) became rampant and widespread, and it was only after the introduction of new drugs that mortality decreased substantially after 2005 (Murray et al. 2012). Currently there are still regions of Africa which carry a high burden of malaria mortality, particularly these that have poor coverage with anti-malarial treatment and insecticide-treated bed nets (Gething et al. 2016).

Resistance to anti-malarial drugs has been reported in *P. falciparum* since the 1950s and is now widespread throughout *P. falciparum* endemic regions, presenting a growing concern regarding malaria treatment and elimination (Mita and Tanabe 2012). Chloroquine was a first line anti-malarial treatment for many years (and is still in some

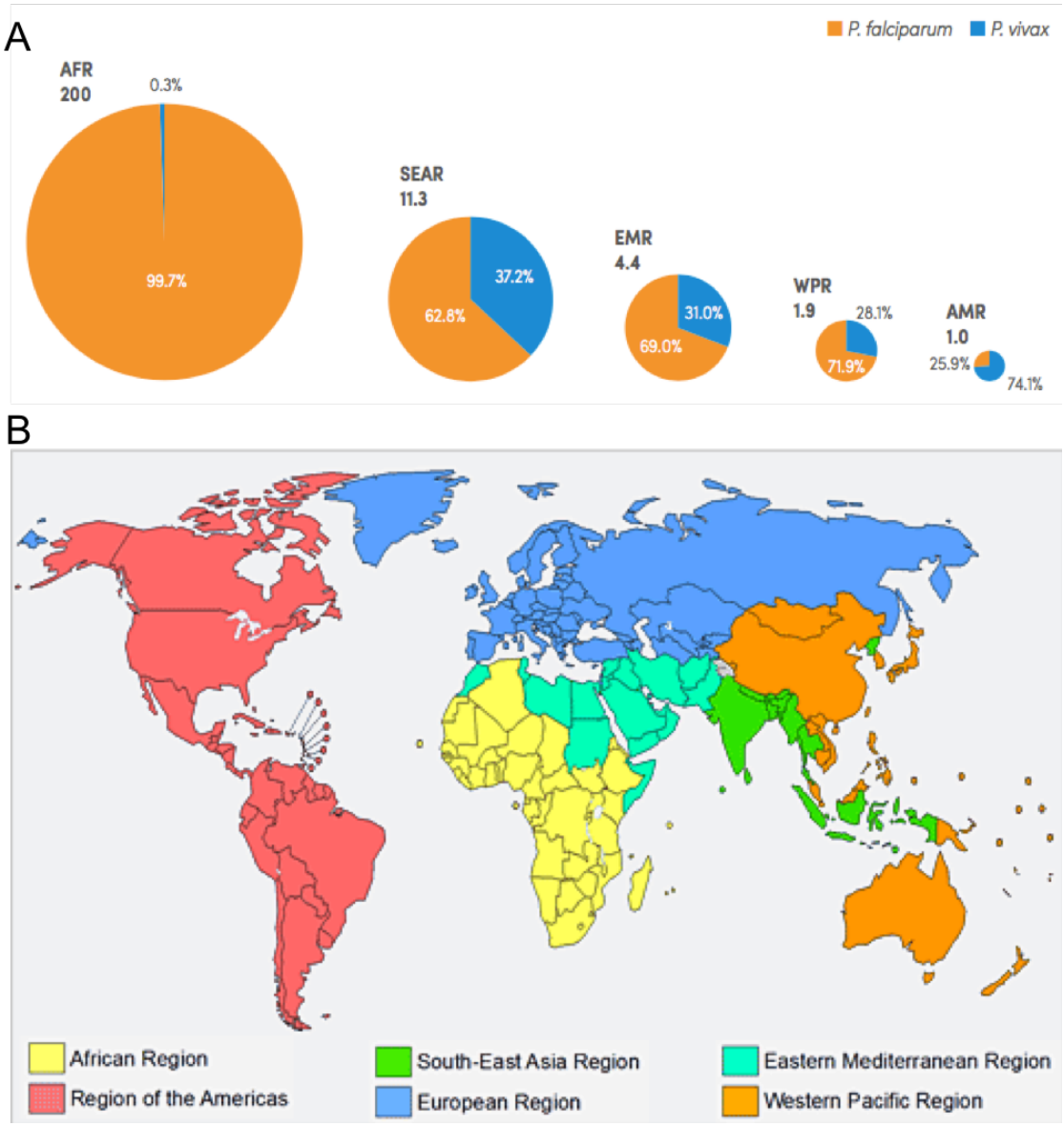

Figure 1.3. Proportion of malaria caused by *P. falciparum* and *P. vivax* in WHO geographical regions A. Proportion of malaria cases caused by the two most prevalent malarial species, *P. falciparum* (orange) and *P. vivax* (blue), in World Health Organisation geological regions (shown in **B**). ~200 million cases reported in the African region (AFR), 99.7% caused by *P. falciparum*, ~11 million cases in the Southeast Asia region (SEAR), 63% caused by *P. falciparum*, ~4.4 million cases in the Eastern Mediterranean region (EMR), 69% caused by *P. falciparum*, and ~1.9 million cases in the Western Pacific region (WPR), 72% caused by *P. falciparum*. The region of the Americas (AMR) is the only malaria endemic region where *P. vivax* is the predominant cause (74%) of malaria. **B.** Map of the world divided into the World Health Organisation geological regions corresponding to regions in **A.**

regions in which *P. falciparum* remains chloroquine-susceptible), before the spread of resistance resulted in need for other drugs. Resistance to chloroquine was reported independently in the 1950s in Southeast Asia (along the Thailand-Cambodia border) and South America (along the Panama-Colombia border) (Wellems and Plowe 2001) and is mediated by the chloroquine-resistance transporter (*pfcrt*) gene (Fidock et al. 2000). Genetic analysis of the *pfcrt* gene revealed that resistance originating from Southeast Asia had spread throughout Africa (Ariey et al. 2006) and into India by the 1980s (Shah et al. 2011), and resistance originating from the Panama-Colombia border had spread to all South American endemic regions by the early 1980s (Mita and Tanabe 2012). Similar geographic patterns of drug resistance spread can be seen for parasites with resistance against other anti-malarials (Mita and Tanabe 2012).

In 2005 Artemisinin-combination therapies (ACT) were recommended by the WHO as the first line treatment for *P. falciparum* malaria. These treatments combine artemisinin with another anti-malarial drug, and result in rapid and substantial clearance of parasites. Increased usage and availability of ACTs have had a profound impact on malaria control (Bhatt et al. 2015). However, as with older anti-malarial drugs, resistance to artemisinin has now developed in *P. falciparum*. Resistance was first reported in Cambodia, but is now found throughout Southeast Asia and is associated with slow clearance of parasites caused by multiple point mutations in the gene *kelch13* that affect the "propeller" region of the protein (Ashley et al. 2014). Evidence has also suggested that patients carrying resistant parasites had higher than expected amounts of gametocytes in peripheral blood before and after ACT treatment, suggesting that these parasites may also have a transmission advantage over artemisinin-susceptible parasites (Ashley et al. 2014). Artemisinin resistance has not yet been clearly reported in Africa, with a single isolated case identified in 2013 in Equatorial Guinea, located on the West

coast of Central Africa (Lu et al. 2017) and an analysis of *kelch13* point mutations in isolates from Africa showed them to have no effect on parasite clearance (Ménard et al. 2016). However, recent evidence from *P. falciparum* isolates collected from both Kenya (East Africa) and Ghana (West Africa) has uncovered single nucleotide polymorphisms present within genes that have been linked to delayed parasite clearance upon treatment with ACT (Henriques et al. 2014; Adams et al. 2018). Single nucleotide polymorphisms and gene expansions have also been found throughout Southeast Asia in genes linked to resistance to both drugs in an ACT (Hamilton et al. 2019).

Aside from drug resistance, an ongoing concern in the context of anti-malarial drugs is that the vast majority are ineffective at killing gametocytes, resulting in the risk of transmission from treated patients. Development of "transmission-blocking" drugs is ongoing, with several that are active against sexual stages in various pre-clinical and clinical phases of the development pipeline (Ashley and Phyo 2018).

1.4 *Plasmodium knowlesi* malaria in Southeast Asia

In much of Southeast Asia, *P. falciparum* remains the leading cause of malaria, with *P. vivax* being the second most common causal species (Figure 1.3A) (WHO 2017). However, the species *P. knowlesi* is now recognised as an emerging and underestimated cause of disease, and is particularly important in Malaysia. Unlike the common malaria parasite species that infect humans, *P. knowlesi* is a zoonotic parasite which primarily infects two species of macaque monkey which occur throughout Southeast Asia (Coatney et al. 1972). Although *P. knowlesi* parasites are potentially capable of human-mosquito-human transmission, human infections are thought to be acquired sporadically from mosquitos that have previously bitten monkeys (Chin et al. 1968). Human infections are usually uncomplicated, but do have the capacity to result in severe and sometimes fatal disease (Singh et al. 2004; Cox-Singh et al. 2008; Daneshvar et al.

2009). In a study of patients from Sarawak in Malaysian Borneo confirmed to be infected with *P. knowlesi*, a small proportion presented with or went on to develop severe disease (Daneshvar et al. 2009). In Sabah, Malaysian Borneo, another study identified a higher proportion of patients with *P. knowlesi* malaria that developed severe disease (William et al. 2011), and the most common severe complication reported is respiratory distress (Cox-Singh et al. 2008; Daneshvar et al. 2009; William et al. 2011; Rajahram et al. 2012). Severe disease leading to death has since been identified retrospectively in cases of malaria identified incorrectly with *P. malariae* (Cox-Singh et al. 2008; Singh and Daneshvar 2013). Research has indicated that a high fatality rate may be caused by a delay in administering appropriate treatment because of *P. malariae* mis-diagnosis (a species that typically causes mild disease), and also due to possible misconception that severe disease is limited to *P. falciparum* infection (Rajahram et al. 2012).

Human infections with *P. knowlesi* have been mostly reported from Malaysia (Figure 1.4). The first report of human *P. knowlesi* infection came in 1965 from an infected traveller returning from peninsular Malaysia where he had spent several nights in the forest (Chin et al. 1965; Coatney 1971). However it was not until the early 2000s that the first report of widespread infection appeared, located in Kapit district of Sarawak state in Malaysian Borneo (Singh et al. 2004). In this study, blood samples were taken from over 200 malaria patients, two thirds of which had been diagnosed with *P. malariae* by microscopy. PCR analysis of these samples reveal that none were infected with *P. malariae,* and in fact over 50% were infected with *P. knowlesi*, with mis-diagnosis with *P. malariae* being a result of the morphological similarities between the two species, making microscopic diagnosis challenging (Singh et al. 2004). While the traveller in 1965 who contracted *P. knowlesi* malaria was also briefly mis-diagnosed

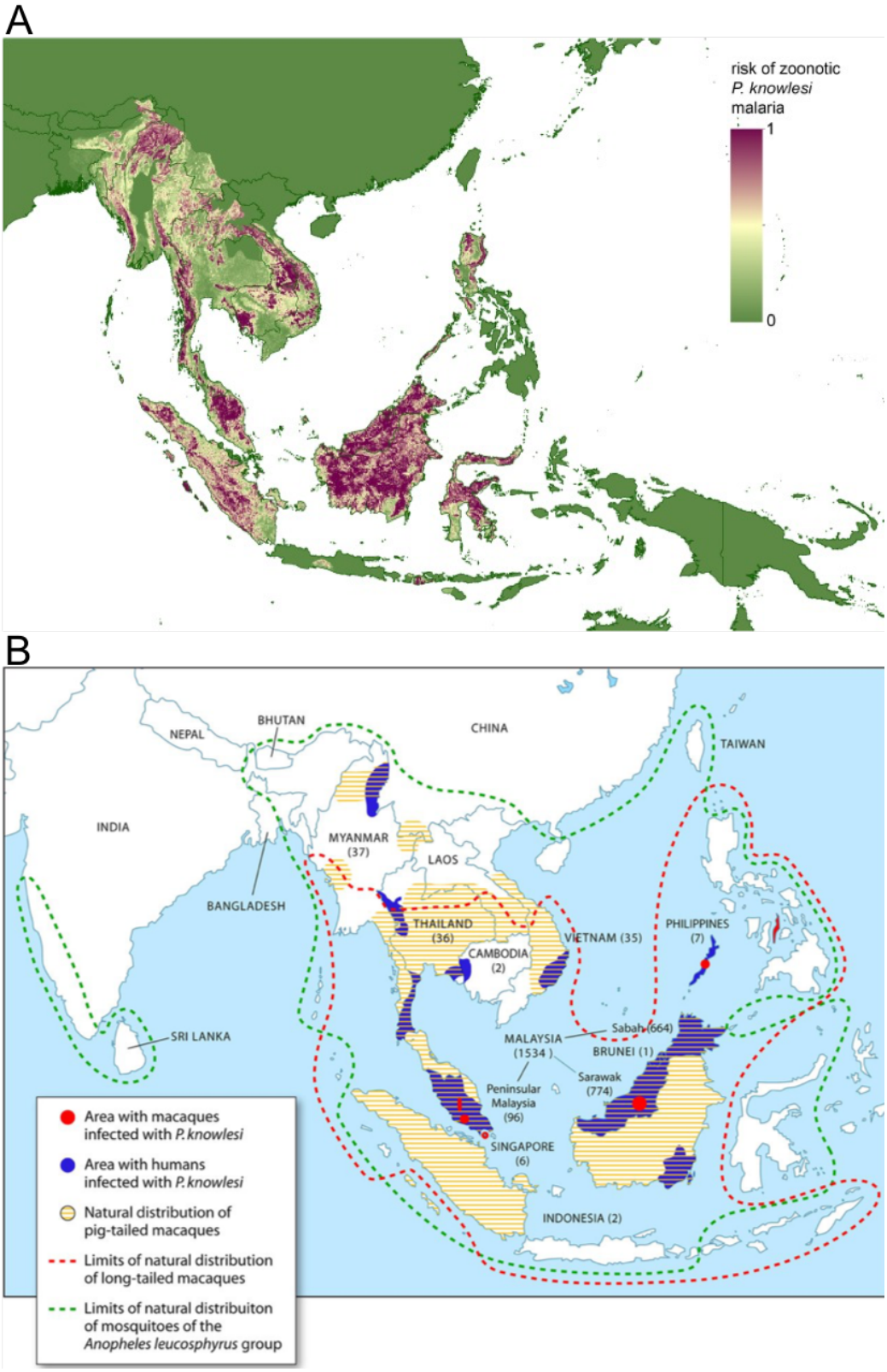

Figure 1.4. *P. knowlesi* at-risk regions of Southeast Asia and distribution of reservoir host and vector species. A. Regions of Southeast Asia considered to be at risk of human *P. knowlesi* malaria, where "1" (purple) is considered high risk and "0" (green) zero risk. Cases have already been reported in all Southeast Asian countries except Laos. Adapted from (Shearer et al. 2016). **B.** Distribution of the mosquito vectors (yellow) and macaque reservoir hosts (red and green outlines) of *P. knowlesi* along with regions in which human (blue shading) and macaque (red shading) *P. knowlesi* infection has been reported. Source: (Singh and Daneshvar 2013).

with *P. falciparum* and then *P. malariae* prior to his correct diagnosis with *P. knowlesi*, the report from Kapit was the first evidence of the chronic and widespread mis-diagnosis that has unquestionably had an impact on the epidemiological status of *P. knowlesi* in Malaysia and probably elsewhere in Southeast Asia. Retrospective PCR analysis of blood films originally classified by microscopy as *P. malariae* from 1996 onwards in Malaysia revealed that nearly 100% of samples contained *P. knowlesi* (Lee et al. 2009).

Since this first report of widespread *P. knowlesi* malaria in Sarawak, further investigation has revealed it to be prevalent in other part of Malaysia as well. In the Sabah state of Malaysian Borneo, much like in Sarawak, mis-diagnosis of patients with *P. malariae* has led to underreporting of the true extent of *P. knowlesi* malaria, with adults bearing the brunt of this zoonotic disease (Barber et al. 2011). Over 4500 cases of *P. knowlesi* malaria have now been formally identified in Malaysian Borneo (World Health Organisation 2017), although actual numbers of cases will be higher due to previous mis-reporting. Species-specific PCR analysis has been carried out on human blood samples collected throughout peninsular Malaysia as well, and *P. knowlesi* infection was confirmed for between 50% and 70% of malaria cases, across several states, showing that the parasite is not limited to Malaysian Borneo (Vythilingam et al. 2008; Yusof et al. 2014). Subsequent collection of blood samples from over 100 wild long-tailed and pig-tailed macaques (*Macaca fascicularis* and *M. nemestrina*, respectively) throughout the Kapit region in Sarawak and PCR analysis revealed that multi-species and multi-genotype infections were common, and that *P. knowlesi* was present in almost all of the long-tailed macaque samples, and in half of the pig-tailed macaque samples (Lee et al. 2011). Similar sampling from wild macaques (predominantly long-tailed macaques) in peninsular Malaysia revealed that half were

positive for *P. knowlesi*, and only a few macaques were positive for non-*P. knowlesi* malaria (Vythilingam et al. 2008). In Kuala Lipis in the state of Pahang in peninsular Malaysia, a smaller proportion of macaques were positive for *P. knowlesi* parasites, but in contrast no parasite-positive macaques were found in the state of Selangor, indicating that infection is not homogenous throughout all of Malaysia (Vythilingam et al. 2008). Sequence analysis of the small subunit ribosomal RNA (SSUrRNA) gene locus, as well as the circumsporozoite surface protein (*csp*) gene locus from a panel of *P. knowlesi* isolates sampled from wild macaques and human infections in Malaysian Borneo revealed no clustering based on host species, indicating that transmission is not host-restricted, and these results implicate the long-tailed and pig-tailed macaques as being the primary reservoir hosts for human infection (Lee et al. 2011). Consequently, malaria caused by *P. knowlesi* is now known to be widespread and is recognised as the leading cause of malaria in Malaysia (Singh et al. 2004; Cox-Singh et al. 2008; Vythilingam et al. 2008; Lee et al. 2011; Yusof et al. 2014). In Sabah, the incidence of *P. knowlesi* appears to be rising alongside decreasing rates of *P. falciparum* and *P. vivax* (William et al. 2013; Cooper et al. 2019). This is likely a combination of much improved diagnostics and acknowledgment of *P. knowlesi* malaria but also due to human behaviour, with deforestation and changing land use altering the behaviour of macaques and resulting in closer proximity between them and humans (William et al. 2013; Stark et al. 2019). With Malaysia on the World Health Organisation "E-2020" list of countries aiming to eliminate human transmitted malaria by 2020, understanding and control of *P. knowlesi* disease is critical.

Outside of Malaysia, *P. knowlesi* malaria is more rarely reported, although human cases have been identified in nine countries in Southeast Asia, with others considered to be at risk (Figure 1.4A) (Shearer et al. 2016). With several countries in Southeast Asia on

track to achieve malaria elimination, *P. knowlesi* presents a unique challenge. Developing long-term strategies for this infection will be critical as it cannot be removed as part of general malaria elimination. Indonesia is one such country that is aiming for malaria elimination by 2030 (Sitohang et al. 2018), and which has recently reported a significant number of local cases of *P. knowlesi* malaria (Herdiana et al. 2016; Lubis et al. 2017; Coutrier et al. 2018; Herdiana et al. 2018). A study carried out in North Sumatra (on the island of Sumatra, geographically isolated from mainland Southeast Asia and Borneo) has indicated that an unexpectedly high proportion of individuals were PCR-positive for *P. knowlesi* (Lubis et al. 2017). On another isolated island of Indonesia, Sabang, which is predominantly forested and home to long-tailed macaques, no case of malaria had been reported since 2011 until active case detection uncovered two clusters of human *P. knowlesi* infection. One cluster was associated with construction workers known to spend nights in the forest and were therefore in close proximity to macaques, while the other occurred within one family at a residential location (Herdiana et al. 2018).

In Thailand, the first naturally acquired *P. knowlesi* infection in a human was reported in 2004 from a forested region near the Thailand-Myanmar border (Jongwutiwes et al. 2004), and more endemic cases have since been identified throughout Thailand (Putaporntip et al. 2009; Jongwutiwes et al. 2011). Natural infections have also been reported in Singapore (Ng et al. 2008; Jeslyn et al. 2011), Vietnam (Eede et al. 2009), Cambodia (Khim et al. 2011), and Myanmar (Ghinai et al. 2017). In the Philippines, cases of *P. knowlesi* infection in travellers have been reported on the island of Palawan (De Canale et al., 2017; Ennis et al., 2009), and imported cases have also been reported from Indonesian Borneo (Figtree et al. 2010) and Thailand (Müller and Schlagenhauf 2014). The geographical area reporting human *P. knowlesi* infection is therefore

substantial, and the wide distributions of mosquito vectors and macaque reservoir species make control of this parasite challenging (Figure 1.4B).

Understanding the population structure of *P. knowlesi* may be useful in devising effective control strategies against the parasite. In Malaysian Borneo multi-locus microsatellite analysis of *P. knowlesi* has shown that it forms two, genetically distinct sympatric populations, associated with different macaque reservoir hosts (Divis et al. 2015; Divis et al. 2017). These divergent populations have been confirmed by whole-genome sequencing (Assefa et al. 2015; Pinheiro et al. 2015; Divis et al. 2018). In peninsular Malaysia, a third population, genetically distinct from those of Malaysian Borneo has been identified, originally comprising a small number of old laboratory isolates of *P. knowlesi*, but microsatellite analysis of recent clinical samples has since confirmed this third population as accounting for all cases analysed in peninsular Malaysia (Assefa et al. 2015; Divis et al. 2017).

Outside of Malaysia however, the genetic structure of *P. knowlesi* is almost entirely unknown. Limited research has been carried out using clinical isolates collected from Thailand, based on the amplification and sequencing of a single genetic loci, and it seems that these parasites cluster most closely with those from peninsular Malaysia indicating some relatedness of parasites on mainland Southeast Asia that have diverged from *P. knowlesi* on Malaysian Borneo (Ahmed et al. 2018), although further sampling and deeper sequencing methods will need to be used to verify these results. Increased effort for sampling and in-depth genomic analysis will be necessary to understand *P. knowlesi* genomics more broadly in Southeast Asia. Further detail regarding the genomic population structure of *P. knowlesi* populations in Malaysia will be given in Chapter 3.

1.5 Development of genome sequencing techniques and their use in *Plasmodium* research

The ability to sequence an organism's genome has revolutionised genetics. The first genome to be fully sequenced was that of bacteriophage Φ174 in 1977, using "plus and minus" DNA sequencing, a predecessor of Frederick Sanger's hugely successful di-deoxy chain termination method ("Sanger sequencing") (Sanger and Coulson 1975; Sanger et al. 1977; Sanger et al. 1978). The rapid development, ease of use, and reliability of Sanger sequencing eventually over took the plus and minus method and other sequencing techniques available at the time, and has been used extensively since. The advent and commercialisation of "next-generation" sequencing in the late 1990s and early 2000s has since provided an option for very high-throughput DNA sequencing (Heather and Chain 2016).

The first "next-generation" sequencing technique was pyrosequencing developed in the late 1990s, and this was the first high-throughput, scalable next-generation sequencing platform to be commercialised in 2005. Pyrosequencing is a sequencing-by-synthesis reaction, where an isolated DNA fragment is used as a template for DNA polymerase in the presence of several other compounds, one of which is luciferase. The addition of a base by DNA polymerase is controlled by washing the DNA template with one of the four nucleotides (dATP, dTTP, dCTP, or dGTP) and incorporation of a nucleotide is accompanied by the release of pyrophosphate, which is converted to ATP and acts as a substrate for luciferase. This is detected as light, and allows the DNA sequence to be elucidated (Nyrén et al. 1993; Ronaghi et al. 1998). Pyrosequencing has fallen out of use in recent years in favour of other high-throughput shotgun sequencing techniques, such as Illumina's sequence-by-synthesis chemistry, and so-called "third-generation" sequencing which encompasses "single molecule real-time" (SMRT) sequencing from a

single intact DNA molecule, with the commercially available PacBio system released in 2010, followed in 2015 by the first portable SMRT sequencer (the MinION) from Oxford Nanopore. One of the advantages of SMRT sequencing over short read sequencing, such as what is provided by Illumina, is that as DNA is sequenced as a single molecule ("long-read" sequencing), the data can be used to uncover structural polymorphisms (for example, copy number variants) that are not easily resolved using short-read data.

Illumina sequencing is currently the most widely used technology, and sequencing begins with the capture and immobilisation of DNA fragments onto a glass slide (the "flow cell"). DNA fragments are then clonally amplified by "bridge amplification", during which the forward DNA fragments bend over to bind an oligo sequence attached to the flow cell. DNA polymerase then builds the reverse complementary strand of DNA and the two strands release, with the original forward strand being washed away and the newly synthesised reverse strand remaining attached to the flow cell. The newly synthesised strand of DNA then bends over to bind another oligo, forming a bridge and DNA polymerase builds a complementary strand. These two strands are then denatured, leaving two single stranded DNA fragments attached to the flow cell, which can go on to form further bridges. This process is repeated many times, forming clusters of identical DNA strands which will then be sequenced. These clonally amplified clusters are an important quality control step and are used to differentiate between mutations in the genome and errors introduced during sequencing. After amplification, sequencing-by-synthesis is carried out. DNA polymerases are primed on the DNA fragments and synthesis occurs one base at a time. Each nucleotide fluoresces at a different wavelength upon addition to the DNA chain, which is interpreted computationally to build the DNA sequence.

Plasmodium falciparum was the first malaria parasite to be whole-genome sequenced, in a project spanning 7 years from 1995-2002, using Sanger sequencing, carried out on individually fragmented and clones chromosomes (Foster et al. 1995; Gardner et al. 2002). The *P. vivax* and *P. knowlesi* genomes were also sequenced using the Sanger method with dye-terminator chemistry and published several years later in 2008 (Carlton et al. 2008; Pain et al. 2008). The publication of these later genomes coincided with the first commercial launch of Illumina's sequencing-by-synthesis next-generation sequencing chemistry. The continued decreasing cost of Illumina sequencing, combined with robust and reliable sequencing capabilities make it the most used sequencing technology (Heather and Chain 2016). Large scale genomic sequencing of *Plasmodium* parasites now allows analysis of population structure (Assefa et al. 2015), characterisation and spread of drug resistance loci (Miotto et al. 2013; Ariey et al. 2014), and insight into parasite biology and evolution (Volkman et al. 2007; Manske et al. 2012). Nearly 3000 genomes of *P. falciparum* have been sequenced in an effort to gain a comprehensive understanding of variation within the species around the world as part of the MalariaGEN Pf3k project (https://www.malariagen.net/projects/pf3k), just one example of how high-throughput sequencing techniques are being implemented for large-scale projects.

1.5.1 *Plasmodium* genomics

All of the parasite species that cause human malaria now have high quality whole-genome reference sequences available (Gardner et al. 2002; Carlton et al. 2008; Pain et al. 2008; Rutledge et al. 2017). Sequencing of these species has allowed substantial analysis of their particular genomics and revealed a surprising number of differences between them. The genome of the *P. falciparum* laboratory line 3D7 was completed first in 2002 (Gardner et al. 2002). It is 23Mb in length, comprises 14 chromosomes and

contains ~5300 protein-coding genes for a gene density of approximately one gene per every ~4300bp. In the initial analysis, only 30% of protein-coding genes had high enough similarity to genes in other organisms to assign putative functions. The genome has an extremely high content of A + T bases, averaging ~80% genome-wide, rising to ~90% in introns and intergenic regions. This extreme AT bias is also seen in other members of the *Plasmodium* genus (Figure 1.5), but in particular is shared with other species of the *Laverania* "sub-genus" of *Plasmodium*, which includes other primate malarias, *P. reichenowi* and *P. gaboni* alongside *P. falciparum*. Evidence has suggested that there is significant bias towards G:C to A:T transition mutations that could be the cause of the extreme A + T content seen in *P. falciparum* and the other *Laverania* species, leading to possible mutational advantages such as a propensity for gene expansions and accumulation of insertion/deletion events affecting gene expression (Hamilton et al. 2017).

Within the *P. falciparum* genome, each chromosome has highly conserved sub-telomeric regions comprising of large and complex repeat structures, which form five sub-telomeric "blocks". Three antigenic, low complexity and highly polymorphic gene families of *P. falciparum* are located in the sub-telomeric regions, the *var, rif*, and *stevor* families (Gardner et al. 2002). Members of these three families are formed by two exons and their arrangement within the sub-telomeres is highly structured. The *var* gene family encodes highly diverse PfEMP1 proteins that are exposed on the infected erythrocyte cell surface and mediate cytoadhesion, which is an integral part of *P. falciparum's* pathogenicity (Su et al. 1995). The function of *rif* proteins (rifins) and *stevor* proteins is unknown, but they are also exposed on the erythrocyte cell surface and show antigenic variation (Petter et al. 2007). The development of these large gene families within the sub-telomeres may be a reflection of the nature of these genomic

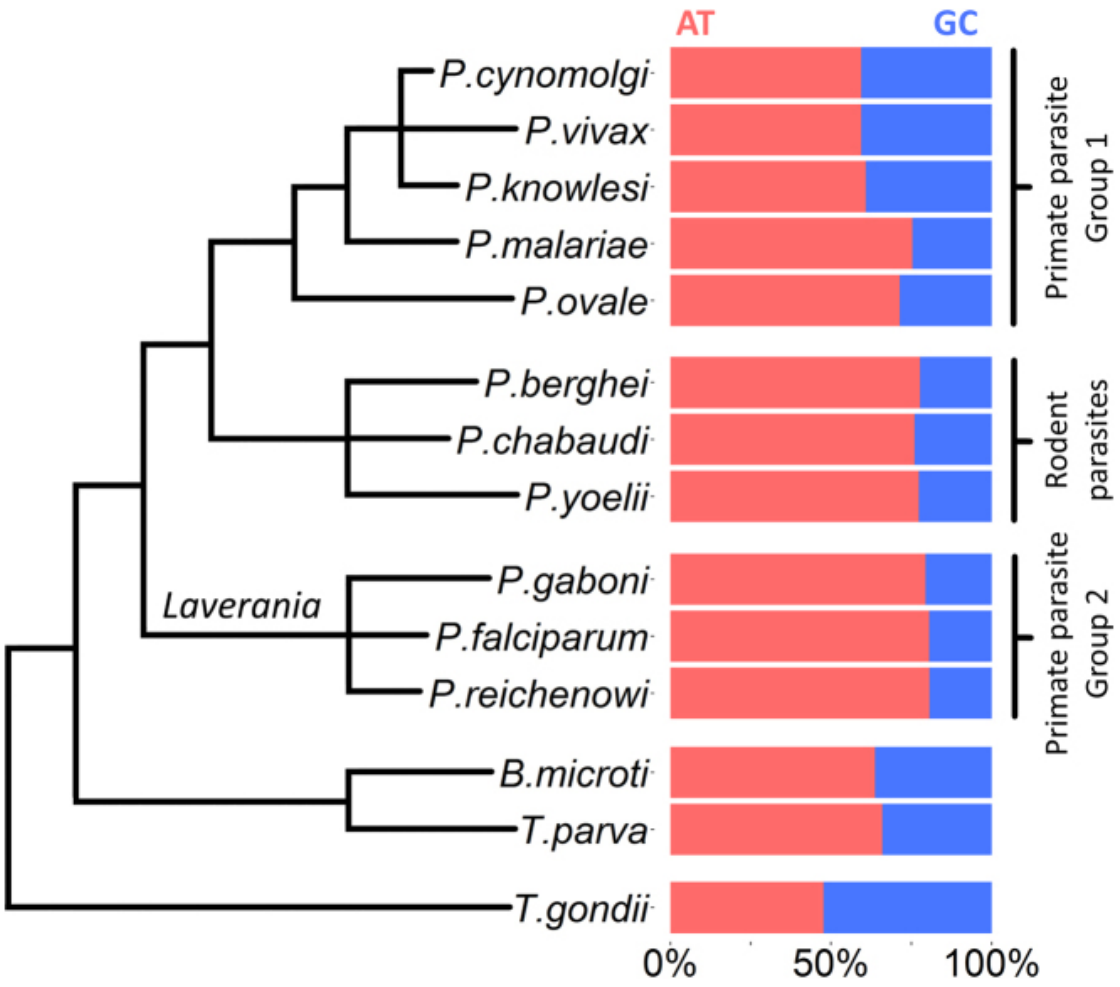

Figure 1.5. Phylogenetic tree representing relationships between *Plasmodium* clades alongside A+T genome content. Also shown are other non-*Plasmodium* apicomplexan parasites, *Babesia microti, Theileria parva*, and *Toxoplasma gondii*. Branches do not accurately represent evolutionary distances. Source (Hamilton et al. 2017).

regions, which are highly polymorphic and interspersed with repeat sequences, undergoing frequent recombination and gene duplication events (Fischer et al. 1997).

There has been substantial genomic variation reported in *P. falciparum*, with the most rapidly evolving genes being transmembrane proteins, and those involved in antigenic variation, cytoadhesion, and cell invasion, whereas essential genes such as those encoding metabolic enzymes show complete conservation (Jeffares et al. 2007; Volkman et al. 2007). Some of these polymorphic, multi-allelic genes are expressed in the erythrocytic stage and as such interact with the host environment, representing targets of natural immunity. These have evolved under diversifying, balancing selection, which produces highly polymorphic loci, for example the merozoite surface proteins such as *msp3*, *mspdbl1,* and *mspdbl2*, and antigenic genes such as apical membrane antigen-1 (*ama1*) and erythrocyte-binding antigens, for example *eba175* (Tetteh et al. 2009; Ochola et al. 2010; Amambua-Ngwa et al. 2012). Other polymorphic genes have evolved under directional selection and these are often drug resistance loci, such as the chloroquine resistance genes *pfcrt* and *mdr1*. The selection acting on these genes is often causes a selective sweep which increases the frequency of an advantageous allele and the flanking chromosomal sequencing which tends to be population specific, and causes of selection are driven different treatment programmes in different locations (Wootton et al. 2002; Mobegi et al. 2014).

The first reference genome of *P. knowlesi* to be sequenced was reported as 23.5Mb in length distributed across 14 chromosomes (Pain et al. 2008). Unlike *P. falciparum*, *P. knowlesi* has a less skewed A + T content of 63%. The genome encodes ~5200 protein-coding genes, 80% of which have orthologues in *P. falciparum*. In contrast to *P. falciparum*, in which large multi-copy antigenic gene families are mostly in the sub-

telomeres, genes from the *P. knowlesi*-specific polymorphic *kir* and *SICAvar* families are found throughout the chromosomes, often surrounded by intrachromosomal telomeric-like sequences. In addition to the *SICAvar* and *kir* gene families, five more *P. knowlesi*-unique families were identified (*Pk-fam-a* to *Pk-fam-e*) (Pain et al. 2008). The genome of *P. knowlesi* is largely similar to those of other malarias that fall in the simian malaria clade, *P. vivax* (which exclusively infects humans), and *P. cynomolgi* (which while capable of human infection, exclusively infects monkeys naturally), with comparable A + T content (Figure 1.5) although it is slightly smaller than the 26 Mb genomes of *P. vivax* and *P. cynomolgi*. *P. cynomolgi* also contains the intrachromosomal telomeric sequences seen in *P. knowlesi*, although these are absent in *P. vivax*. Almost all of the genes identified in *P. knowlesi* have orthologues in *P. vivax* and *P. cynomolgi* (Tachibana et al. 2012).

The *SICAvar* genes are multi-exon and encode antigens exposed on the infected erythrocyte cell surface and are associated with virulence. This family is unrelated to the *P. falciparum var* gene family but is also thought to undergo expression switching (Al-Khedery et al. 1999). *SICAvar* genes fall into two types: type I containing 7-14 exons and type II containing 3-4 (Pain et al. 2008). Both SICAvar and KIR proteins have structured domain organisation (Figure 1.6). SICAvar proteins contain a number of highly diverse cysteine-rich domains, followed by a transmembrane domain and a SICAvar conserved domain (Figure 1.6A). Their intronic regions are low complexity and highly repetitive. The *kir* gene family is part of the *pir* superfamily and like *SICAvar* genes they are thought to encode antigenic proteins expressed at the erythrocyte cell surface (although their precise function remains unknown) and are divided into four types based on exon number. KIR proteins consist of 1-3 domains, followed by a transmembrane domain (Figure 1.6B) and several were found to contain amino-acid sequences with 100% identity to *Macaca* CD99 protein (an immune protein

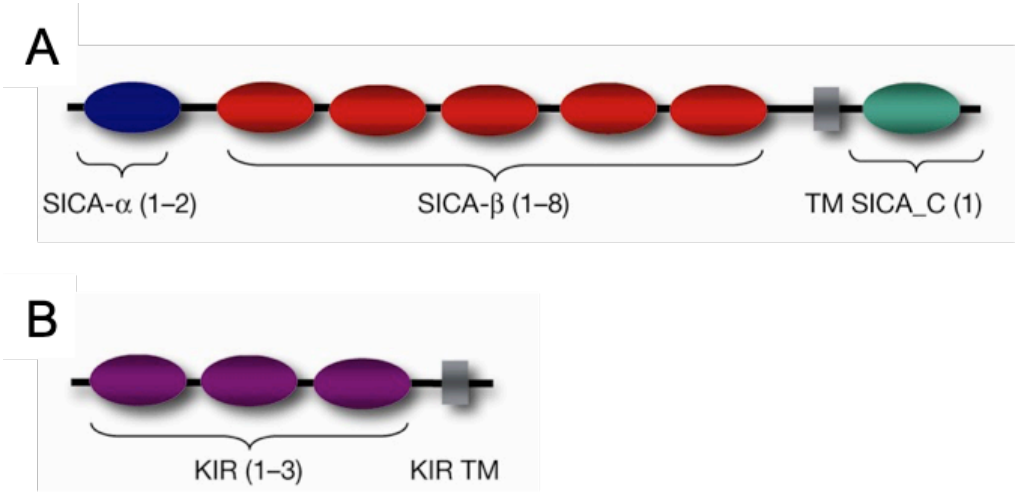

Figure 1.6. Domain organisation of SICAvar proteins (A) and KIR proteins (B) gene families in *P. knowlesi*. A. SICAvar proteins are formed by highly variable cysteine-rich domains (1-2 SICA-α domains and 1-8 SICA-β domains), a transmembrane domain (TM), and a cytoplastic domain (SICA_C). **B.** KIR proteins fall into four types (based on the number of exons), but all have the same predicted domain organisation: 1-3 KIR domains, and a transmembrane domain (KIR TM).

with a critical role in T-cell function) and significant identity to human CD99 (Pain et al. 2008). Interestingly, different KIR proteins showed identity to different regions of CD99 such that when combined they covered over 50% of CD99's extracellular domain (Pain et al. 2008).

1.5.2 Using next-generation genome sequencing to investigate population structure in *Plasmodium*

The advent of genetic sequencing drastically changed the way in which biology and evolution could be investigated and has allowed the identification and characterisation of populations, their structure, and the individuals and selection pressures that affect them at the DNA level. Adaptation and selection occur as a result of mutations such as

insertion/deletions (INDELs), copy number variants (CNVs), microsatellites, and single nucleotide polymorphisms (SNPs). In *P. falciparum*, INDELs outnumber SNPs and occur most frequently in non-coding regions and CNVs covering drug resistance loci have been identified (Nair et al. 2007; Miles et al. 2016). Single nucleotide polymorphisms are a very common form of genetic variation, and are the most straight forward to interpret, and so they have informed vast amounts of research. SNPs can be used to infer genomic diversity within and between populations, identify loci under specific selection pressures that are driving population structure and differentiation, and identify mutations associated with phenotypic changes.

Prior to the advent of high-throughput next-generation sequencing, analysis relied upon pre-existing knowledge of SNPs, which required painstaking identification of variants which were used to inform the design of DNA microarrays that could then be used for calling variants in other samples. Next-generation sequencing has since revolutionised the field of SNP discovery and analysis. As more and more high-quality reference genomes are sequenced and released, SNP discovery and analysis has become easier, but is still a challenge when working with non-model organisms for which high-quality reference genomes are not available (Kumar et al. 2012). The nature of short-read sequencing means that for every base of the genome, multiple independently sequenced DNA fragments will exist, providing consolidatory data and allowing reliable identification of novel polymorphisms. In addition, analysis of populations is not just limited to pre-identified SNPs, but instead can be based on an entire genome worth of data. The accumulation of SNPs is a natural mutational process, however external selection pressures that can ultimately lead to adaptation influence SNP patterns, and can be used to understand the diversity and evolutions of populations. Whole-genome sequencing and SNP discovery has been used to identify loci under selection between

different populations of *Plasmodium* parasites. For example, the whole-genome sequencing of parasites taken from two populations in West Africa, one in a high transmission region and the other from a low transmission region found detectable differences between the populations including divergent directional selection on drug resistance loci (correlating to variable drug use in each region), and high differentiation surrounding the *gdv-1* locus, a gene known to be critical for gametocytogenesis (Mobegi et al. 2014). Long-read sequencing from *P. falciparum* clinical isolates from West Africa has since revealed a significant structural dimorphism, consisting of allele-specific deletions in the 3'-intergenic region of *gdv-1* that are under differential local selection and found at different frequencies in different populations (Duffy et al. 2018).

The breadth of information available from high-quality SNP discovery from next-generation data is impressive. Malaysia is considered to be in the pre-elimination stage for *P. vivax* malaria. High-throughput investigations using whole-genome sequencing and SNP discovery has allowed comparative analysis of *P. vivax* parasites from Malaysia with populations of parasites elsewhere in Southeast Asia where transmission remains high (Auburn et al. 2018). From over 200 isolates, whole genome sequencing allowed the identification of over 500,000 high-quality SNPs, which informed analysis regarding population structure and evolution of drug resistance loci, and in response to the selection pressures induced by near-elimination. Results revealed a high level of relatedness among Malaysian *P. vivax* samples, with one extremely large cluster indicating a rapid clonal expansion of a particular strain. SNP analysis also revealed differential selection acting on multiple loci in Malaysian samples, many of which are known to be involved in drug resistance (Auburn et al. 2018). Analyses such as these provide insights into how a changing environment impacts parasites at a molecular level, and will be crucial in achieving and maintaining malaria elimination in the future.

1.5.3 Next-generation transcriptomic sequencing

Alongside development of next-generation sequencing for DNA, methods for capturing and analysing the RNA from samples were also being developed, allowing high-throughput analysis of gene expression profiles (transcriptomes). RNA-seq was first demonstrated over a decade ago (Bainbridge et al. 2006), and prior to next-generation sequencing, cDNA microarrays were most commonly used to assess gene expression. However microarrays have several limitations, primarily the need for existing gene information (Wang et al. 2009). Next-generation sequencing on the other hand is capable of measuring gene expression *de novo*, and at a much higher scale. Methods for RNA-seq are on the whole the same as those used for DNA-seq, however RNA must first go through some additional steps before sequencing libraries can be prepared. Due to the instability of RNA it is first reverse transcribed into cDNA, this is typically achieved using oligo(dT) primers, designed to enrich for mRNA molecules and remove the rRNA molecules that make up ~90% of a cell's total RNA. Transcriptomic data for *Plasmodium* species is invaluable and (among other applications) RNA-seq has been used to interrogate gene expression changes throughout the life cycle of *P. falciparum* and *P. vivax* (Otto et al. 2010; Zhu et al. 2016), identify genes involved in the parasite gametocytogenesis pathway (Bancells et al. 2019), and understand the mechanisms underlying anti-malarial drug resistance (Antony et al. 2016; Kim et al. 2019).

1.6 The advent of single-cell RNA sequencing

Traditional RNA-seq does come with some limitations. These can be technical in nature, such as for researchers investigating early stages of development where there are simply too few cells available to carry out reliable RNA-seq. There can also be biological limitations, as gene expression is plastic and changes may occur in response to environmental and cellular cues that we are not always privy to, resulting in

individual cells from an identical genetic background having highly variable transcriptomic profiles (Tang et al. 2009). As such, when RNA-seq is carried out using a surplus of cells/RNA ("bulk" RNA-seq), the acquired data presents an 'average' transcriptome that is contributed to by the unique transcriptomic profiles of each cell (Figure 1.7).

Efforts were first pursued for amplifying and sequencing RNA from single cells in the 1990s. The first report of RNA successfully amplified from single cells was in 1990, analysed by *in situ* hybridisation (Brady et al. 1990), and the very first transcriptome to be sequenced from a single cell was reported in 1992 (Eberwine et al. 1992). With the increased availability and affordability of next-generation sequencing, the first transcriptome from a single cell to be sequenced using high-throughput methods was a mouse blastomere as reported in 2009 (Tang et al. 2009), and significant progress has been made since then. The pilot experiment by Tang et al. (2009) detected expression from over 5,000 more genes than an equivalent microarray approach that used hundreds of cells, and almost all of the genes shown to be expressed by microarray showed expression by single-cell RNA-seq (Tang et al. 2009). Applications for single cell RNA-seq have been extensive since then. Over the last decade, it has been used to investigate early stages of development (Tang et al. 2010; Reinius et al. 2016), profile transcriptomic heterogeneity in cancerous tissues (Ramsköld et al. 2012; Patel et al. 2014; Chung et al. 2017), and recent developments are allowing its' use with bacterial cells, aimed at understanding transcriptional changes in cells in response to environmental pressures (Wang et al. 2015). All single-cell RNA-seq methods follow the same basic protocol: isolation of single cells, lysis of cell membranes, reverse transcription of RNA, amplification of cDNA, followed by library preparation and next-generation sequencing.

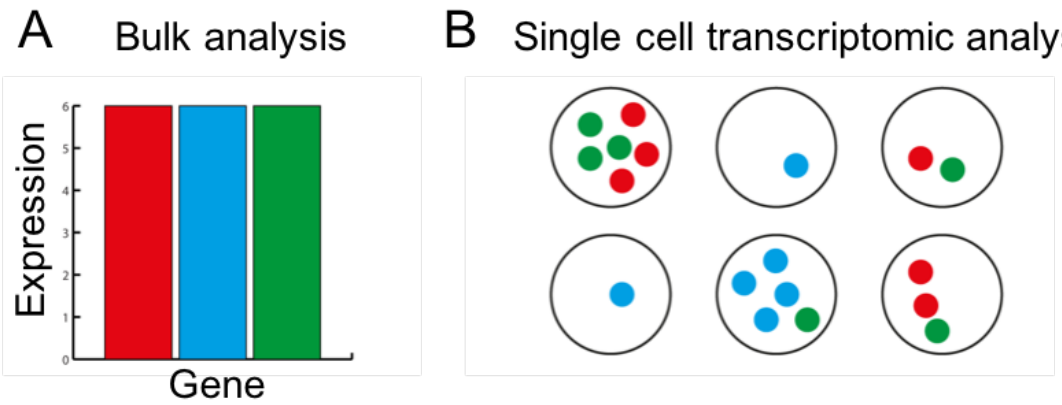

Figure 1.7. Individual cells contain unique transcriptomic profiles. RNA-seq from bulk material produces an 'average' transcriptional profile (A) that is made up from the individually varying profiles (B) (Adapted from Macaulay & Voet, 2014).

Isolation of single cells is generally carried out by fluorescence-activated cell sorting (FACS) (Jaitin et al. 2014), microfluidics (Streets et al. 2014; Wu et al. 2014), or droplet-based techniques (Klein et al. 2015; Macosko et al. 2015). Lower throughput techniques such as cell picking and laser capture microdissection can also be used (Islam et al. 2011; Ramsköld et al. 2012; Boone et al. 2013; Picelli et al. 2014). Reverse transcription is typically carried out using oligo(dT) primers in order to select for mRNA molecules, which make up a small proportion of a cell's total RNA.

Multiple Displacement Amplification (MDA) is one method that was first adapted for use with limiting amounts of human genomic DNA (Dean et al. 2002), before being tested on DNA from single cells (Spits et al. 2006a; Lasken 2007) and has since been used for amplification of cDNA in RNA-seq experiments (Pan et al. 2013). Multiple displacement amplification (MDA) of RNA begins with oligo(dT) primed reverse

transcription, followed by ligation of the individual cDNA molecules into long, chimeric cDNA strands (Figure 1.8). These long lengths of cDNA facilitate the MDA reaction. Amplification is carried out by the bacteriophage Φ29 DNA polymerase, which is used for several reasons. It is a high-fidelity DNA polymerase, with intrinsic proof-reading capability, it is highly processive, capable of generating DNA strands >70kb in length, and, critically important for MDA, it has DNA strand displacement activity (Paez et al. 2004; Pinard et al. 2006). After cDNA molecules have been ligated together, random hexamer primers bind throughout the chimeric molecules, and are used by Φ29 DNA polymerase to build the second strand of cDNA. When a Φ29 DNA polymerase reaches the start of the next, newly synthesised DNA strand, its strand displacement activity displaces the new strand and Φ29 DNA polymerase continues along the cDNA molecule. The newly displaced cDNA strands are themselves now bound by the random hexamer primers and Φ29 DNA polymerase begins second strand synthesis. This results in massively branched chimeric cDNA molecules, which can then be prepared for sequencing.

Alternatively, PCR-based amplification can be used, a popular method of which is Smart-seq2, a protocol utilising the "SMART™" template-switching chemistry developed by Clontech (Zhu et al. 2001). Reverse transcription is initiated with specialised oligo(dT) priming and carried out by the Moloney Murine Leukaemia Virus (MMLV), which facilitates the addition of 2-5 non-template nucleotides at the 3' end of the newly synthesised cDNA strand (first-strand synthesis). PCR amplification then takes place using a limited number of cycles (18 recommended in the protocol), with second-strand synthesis occurring by "template-switching" of the MMLV, primed by a specialised oligo that binds to the newly added untemplated nucleotides (Figure 1.9). Samples can now be prepared directly for next-generation sequencing. Compared to

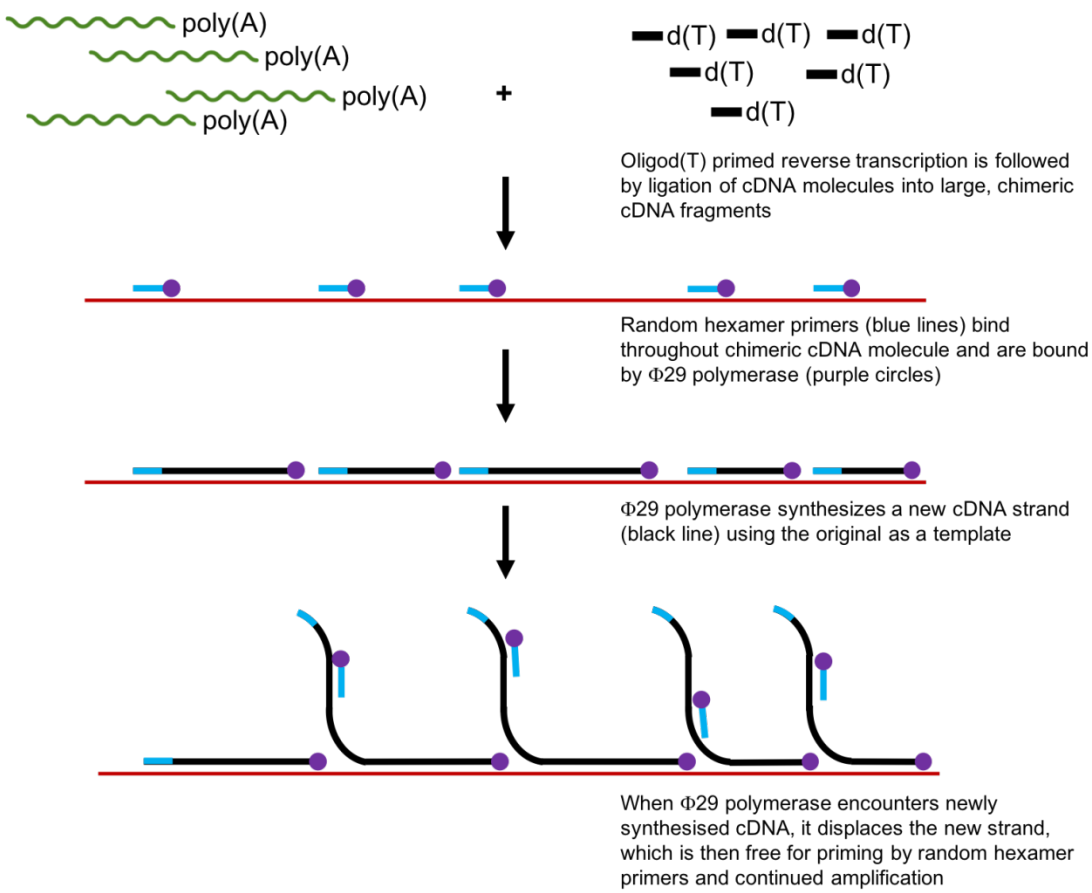

Figure 1.8. Representation of how multiple displacement amplification (MDA) amplifies RNA from single cells. After oligo(dT) primed reverse transcription, individual cDNA molecules are ligated together into large, chimeric cDNA strands which are bound by random hexamer primers. Bacteriophage Φ29 DNA polymerase synthesises the second strand of cDNA using the original chimera as a template. When Φ29 polymerase encounters newly synthesised DNA, it displaces the strand and continues to synthesise DNA. Displaced strands can now be bound by random hexamer primers and used as templates for Φ29 polymerase resulting in complex, branched cDNA structures formed by mass amplification. Author rendition, inspired by (Spits et al. 2006b).

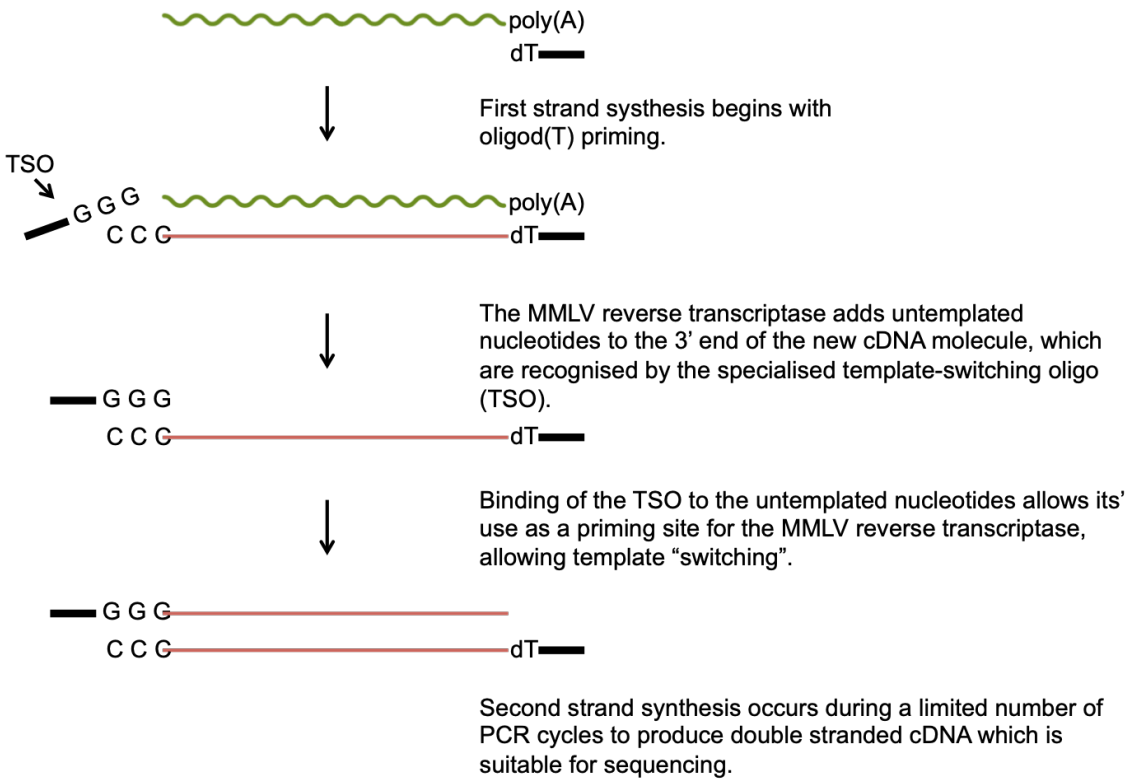

Figure 1.9. Representation of the Smart-seq2 protocol for amplifying RNA from single cells. Smart-seq2 is a protocol based on the use of a "template-switching" reverse transcriptase. RNA is first reverse transcribed using oligo(dT) primers and the Moloney Murine Leukaemia Virus (MMLV) reverse transcriptase. The reverse transcriptase adds 2-5 untemplated nucleotides to the end of the cDNA strand (represented in this figure by "C C C"). These untemplated nucleotides are bound by a specialist template-switching oligo (TSO) which facilitates the "switching" of the MMLV reverse transcriptase, which is now able to synthesis the second strand of cDNA after the original RNA molecule has been degraded. Author rendition, inspired by (Picelli et al. 2014).

PCR based amplification, MDA is thought to result in less bias, more uniform amplification and lower introduction of error due to the high fidelity of Φ29 DNA polymerase (Esteban et al. 1993; Dean et al. 2002; Pan et al. 2013). However, the amount of amplification needed to obtain sequenceable amounts of material may inevitably lead to some bias.

One method of accounting for technical noise and bias is by using RNA spike-ins, molecules of known sequence and added at a known concentration, allowing the indirect estimation of and correction for the effect of technical noise. Quantified using an RNA dilution series, the amount of starting RNA has been found to be highly predictive of the amount of technical noise that can be expected in the dataset, using 5000pg of starting material resulted in technical noise comparable to that expected from bulk RNA-seq analysis (Brennecke et al. 2013). As the amount of starting RNA is decreased, the amount of technical noise increases, and genes with a low-read count showed highly variable 'expression' between replicates, an effect mitigated in genes with a high-read count (Brennecke et al. 2013). This was confirmed by spiking single *Arabidopsis thaliana* cells with HeLa total RNA, which found high technical variation for HeLa genes with a read count of < 100. *Arabidopsis thaliana* cells were also spiked with ERCC RNA spike-ins to allow correction for technical noise, but these were only effective when large cell numbers were used (Brennecke et al. 2013).

An alternative method has been developed which allows for the direct counting of RNA molecules. These 'unique molecular identifiers' (UMIs) are 5bp barcodes, ligated to cDNA molecules during reverse transcription. After sequencing, the number of distinct UMIs mapping to each position can be directly counted. Fragments originating from PCR artefacts will be represented by the same UMI and can be subsequently removed

from the dataset. The use of UMIs is particularly powerful when investigating genes with lower read counts, but the nature of UMI ligation during reverse transcription means that just the 5' end of mRNA molecules can be sequenced (Islam et al. 2014).

In light of the development of numerous single-cell sequencing methods such as Drop-seq, Smart-seq2, MDA techniques, CEL-seq, MARS-seq, to name a few (Dean et al. 2002; Jaitin et al. 2014; Picelli et al. 2014; Macosko et al. 2015; Hashimshony et al. 2016), the commercial sector is rapidly developing its own repertoire of products. Qiagen now produces a whole-genome and a whole-transcriptome single-cell amplification kit using MDA chemistry. A protocol similar to Smart-seq2, utilising a template-switching oligo, is available through Clontech, and is also available integrated with the high-throughput microfluidic platform (the C1™) produced by Fluidigm, which isolates cells on a microfluidic chip prior to reverse transcription and amplification. Another microfluidic platform is available through 10X Genomics which allows even higher throughput analysis of thousands of cells in parallel.

1.7 *Plasmodium* parasites have highly plastic transcriptomes

The transcriptomic profiles of *Plasmodium* parasites are not fixed. Like other pathogens, continual interaction with a host's immune system, along with selection pressure introduced by anti-malarial drugs and changing transmission dynamics require *Plasmodium* parasites to have flexible and dynamic gene expression.

Microarray analysis of ~2700 genes and RNA-seq of ~4000 genes carried out throughout the *P. falciparum* intraerythrocytic life cycle have shown global variation in gene expression, with transcriptional profiles changing in a cyclical manner, correlating to developmental stage (Figure 1.10) (Bozdech et al. 2003a; Otto et al. 2010). Antigenic variation in *Plasmodium* parasites is well-documented and is a critical requirement of

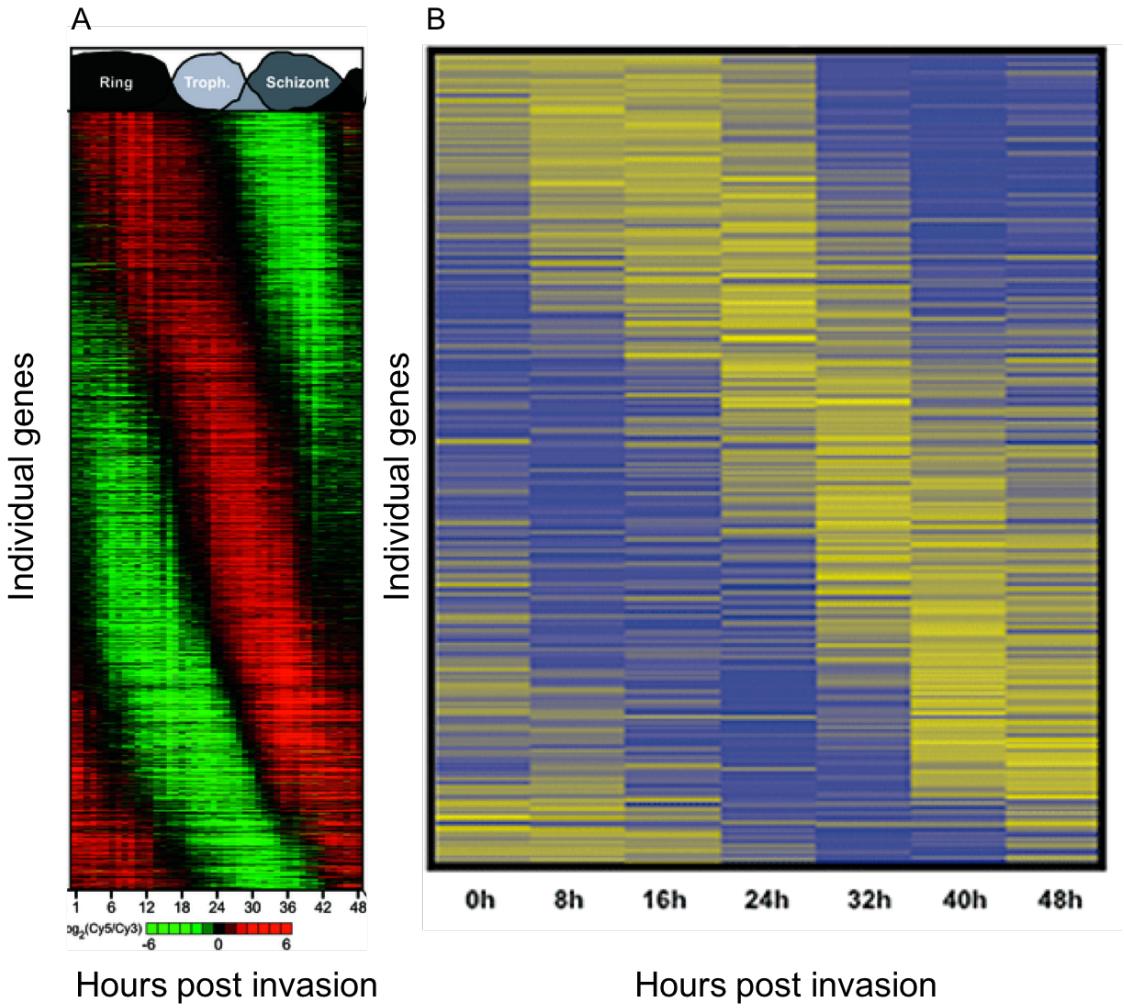

Figure 1.10. Heat maps showing the variability of gene expression throughout the *P. falciparum* intraerythrocytic life cycle analysed by microarray and RNA-seq. Each horizontal bar represents a gene. **A.** Expression of 2712 genes detected by microarray over the intraerythrocytic life cycle of *P. falciparum* HB3. Red indicates higher expression of that gene, and green indicates lower expression. The cyclical cascade of gene expression correlating to different life stages can clearly be seen **B.** RNA-seq analysis of tightly synchronized *P. falciparum* 3D7 parasites, taken at eight-hour increments along the intraerythrocytic life cycle. Each bar represents one of 3975 genes. Yellow indicates higher expression and blue indicates lower expression. The periodic pattern of gene expression is maintained at the greater resolution provided by RNA-seq analysis. (Adapted from Bozdech et al. 2003; Otto et al. 2010).

the parasite to evade specific immune responses. In *P. falciparum*, there are 50-100 *var* genes (encoding the cytoadhesive PfEMP-1 proteins) present in the genome (Su et al. 1995), with infected erythrocytes expressing only a single, distinct and clonally variant PfEMP-1 protein at any one time, but parasites are capable of 'switching' their expression, a key immune evasion strategy for maintaining infection that is utilised by bacteria and other protozoan parasites (Scherf et al. 1998). Expression of *var* genes is controlled at the transcriptional level and there is evidence to suggest this is through long non-coding RNA mediated silencing (Amit-Avraham et al. 2015). Other malaria parasites also have clonally variant antigenic gene families, including the *vir* family in *P. vivax*, and the *kir* and *SICAvar* families in *P. knowlesi*. The *vir*, *kir*, and *rif*/*stevor* families together belonging to the "*pir*" superfamily of genes, also contains orthologues in the rodent malarias (Janssen et al. 2002). While less is known about these families compared to the *var* genes, they do show clonally variant expression, and antigenic switching has been documented to occur in the *SICAvar* gene family in *P. knowlesi* (Al-Khedery et al. 1999) and throughout the *pir* superfamily (Cunningham et al. 2010).

Transcriptional variability is not limited to highly polymorphic, antigenic gene families. By taking sub-clones of several *P. falciparum* laboratory lines and carrying to limited culture, Rovira-Graells et al. (2009) used a *P. falciparum* gene expression microarray to identify genes showing clonally variant gene expression. It was found that between genetically distinct *P. falciparum* laboratory lines, there was extensive overlap in the genes showing transcriptional variability. A large proportion of these belonged to the highly polymorphic antigenic gene families, however a considerable amount of transcriptional variation was also found in to be members of other gene families. The Apicomplexan *Apetala-2* (*ap2*) family of transcription factors (with well-established roles in gametocytogenesis), along with three gene families encoding exported proteins

were represented by a number of variably expressed genes (Rovira-Graells et al. 2012). Experimental testing of the parasite clones' reaction to heat shock found that transcriptionally diverse parasite lines adapted much quicker and more successfully than newly sub-cloned parasites (with low transcriptional diversity). The authors theorised that an adaptation advantage of the transcriptionally diverse line was due to selection of a pre-existing sub-population of parasites which had a more suitable transcriptional profile (Rovira-Graells et al. 2012).

Transcriptional variability can also be observed through a process known as invasion switching. In *P. falciparum*, invasion ligands display considerable redundancy, and some laboratory lines are capable of switching the ligands used for invasion not only in response to an external stimuli but also spontaneously in *in vitro* culture (Dolan et al. 1990; Awandare et al. 2018). A key residue used for invasion is sialic acid (SA) which is found in glycophorins on the erythrocyte cell surface, and parasite invasion can be broadly characterised as being either SA-dependent or SA-independent (Dolan et al. 1990). Research has shown that *P. falciparum* parasites are capable of switching between invasion phenotypes, either by culturing SA-dependent parasites with erythrocytes treated with neuraminidase, a compound which cleaves and removes sialic acid residues, or by genetic ablation of the invasion gene erythrocyte-binding antigen-175 (*eba175*), the protein for which binds erythrocytes in a SA-dependent manner (Duraisingh et al. 2003a). Using microarray analysis of gene expression, the gene reticulocyte binding-like protein 4 (*Rh4*) was identified as being considerably upregulated in SA-independent parasites compared to SA-dependent parasites of the same genetic background and genetic ablation of this gene resulted in parasites that were unable to switch to SA-independent invasion (Stubbs et al. 2005). Laboratory line parasites with both functional *eba175* and *Rh4* show reversible pathway switching and

this is thought to occur through silencing of *Rh4* (Stubbs et al. 2005) and evidence based on gene expression profiling suggests invasion pathway switching of this nature occurs naturally in clinical isolates of *P. falciparum* for as of yet unknown reasons, but possibly in association with virulence or as a method of immune evasion (Nery et al. 2006).

In addition, there is evidence showing that *Plasmodium* parasites that result in severe disease (encompassing complications such as cerebral malaria, respiratory distress, and pregnancy-associated malaria) have different transcriptomic profiles to those that cause uncomplicated malaria. Genome-wide analysis of cDNA transcripts from *P. falciparum* collected from the placentas of women with pregnancy-associated malaria revealed a panel of genes with differential expression when compared to *in vitro* cultured *P. falciparum* 3D7 parasites, with enrichment of genes located in sub-telomeric regions (Ndam et al. 2008). Transcriptional profiling of parasites collected from children with cerebral malaria uncovered two distinct and discriminatory transcriptomic profiles of parasites significantly associated with high and low parasitemia, with each cluster showing differential gene set enrichment (Milner et al. 2012). In addition, comparison of profiles between patients with and without retinopathy (damage to the retina resulting from cerebral malaria) revealed that parasites from children who developed retinopathy had an enrichment of genes involved in the induction of invasion, adhesion, and DNA replication (Milner et al. 2012).

1.8 Gametocytogenesis is represented by a rare population of cells in the blood

The transmission potential of a *Plasmodium* infection is directly related to the number of gametocytes produced in that infection, resulting in a cost-benefit trade-off regarding the rate of conversion between increasing transmission or improving within-host

survival potential, and it is thought that this is why gametocytogenesis typically occurs at a low rate in *Plasmodium* parasites (Taylor and Read 1997; Carter et al. 2013). However, this rate is flexible and varies in response to environmental changes and cues, and is known to increase when parasites are under stress (Carter et al. 2013). Application of anti-malarial drugs appears to increase the rate of gametocytogenesis, a phenomenon that has been demonstrated *in vivo* in response to several anti-malarial drugs in *P. falciparum* (Buckling et al. 1999b; Peatey et al. 2009) and in the rodent malaria *P. chabaudi* (Buckling et al. 1999a).

The vast majority of anti-malarial drugs are ineffective on gametocytes, leaving the window open for further transmission from a treated patient and also leading to relapse of disease. Just two anti-malarial drugs (primaquine and tafenoquine, both belonging to the 8-aminoquinoline family) have shown effective killing of gametocytes. Both, however, can have serious side effects when administered to glucose-6-phosphate dehydrogenase deficient patients (Tarlov et al. 1962; Rueangweerayut et al. 2017) and also cannot be administered during pregnancy, in breast-feeding mothers, or to babies less than 6 months of age, and this has limited their widespread use (Abdul-Ghani et al. 2015). In addition, further investigations are needed to produce the high quality evidence needed for their use as gametocytocidal drugs (Eziefula et al. 2012). Understanding the process of gametocytogenesis may help towards developing new control tools.

Gametocytogenesis occurs in at least two stages, commitment to sexual development by the asexual parasite, and subsequent morphological conversion into a gametocyte. The physiological triggers that initiate commitment to gametocytogenesis are not well understood, however it is widely thought that it occurs as a stress response due to

environmental deterioration, such as when parasitaemia is high or due to drug therapy (Bruce et al. 1990). These environmental processes trigger the transcription of *ap2-g* (which is usually marked by H3K9me3 repression), which signals the start of the commitment process, and expression of *ap2-g* is subsequently maintained in a positive feedback loop (Kafsack et al. 2014). Transcriptional changes can then be observed throughout each stage of gametocyte development (Young et al. 2005).

After commitment, the parasite undergoes five morphological transitions (stage I – V gametocytes) to develop into mature male or female gametocytes, a process that takes 9-12 days in *P. falciparum* and less than 60 hours in other *Plasmodium* species (Gautret and Motard 1999). Each of these shifts is preceded and accompanied by a drastic and widespread change in gene expression, and there is evidence that the gametocyte transcriptome is associated with particular epigenetic motifs (Van Biljon et al., 2019). Male and female gametocytes are distinguishable by microscopy in stage IV and V gametocytes, and it has been reported that two thirds of the gametocyte's transcriptome is differentially transcribed in male and female gametocytes in *P. falciparum* (Lasonder et al. 2016). It is now accepted that post-transcriptional regulation has an essential role to play in gene expression changes in *Plasmodium* parasites, and has a profound impact on the rate of transcript degradation at different points in the asexual life cycle (Shock et al., 2007). There is good evidence for the transcription of a large number gametocyte-specific genes early in gametocytogenesis, where they remain untranslated until the gamete and ookinete stages where they act as "maternal" mRNAs that initiate developmental cascades, and this may be a common regulation method in *Plasmodium* (Kaslow et al., 1988, Liu et al., 2011). However, despite extensive interest and research, the genetic and transcriptomic networks controlling gametocytogenesis are still

relatively unknown, and as such this represent an important phase of the parasite's life cycle to research and target through drug and vaccine interventions.

Identification of new players in the gametocytogenesis pathway is vital to encouraging the development of new therapies specifically targeting the transmission portion of the *Plasmodium* lifestyle. Several interesting genes have been identified in recent years, from the more well-known gametocytogenesis activators, such as *gametocyte development 1* (*gdv-1*) and *ap2-g*, to lesser known genes which have more subtle roles in the control of gametocyte development. An example of one of these latter genes would be *ap2-g2*, which is involved in the repression of the asexual development gene network, and as such assists in gametocytogenesis in a manner unlike the more straight forward transcriptional activators (Yuda et al., 2015). Several other genes have been implicated in the gametocytogenesis pathway, many of them downstream of *gdv-1* or *ap2-g*, but their functions are even more elusive. One gene of interest that falls in this category is *mspdbl2*, a gene encoding a merozoite surface protein. Multiple lines of evidence have suggested involvement of *mspdbl2* in gametocyte commitment, and the protein has a striking "on/off" expression profile in mature schizonts, the basis of which remains unknown (Amambua-Ngwa et al., 2012; Tarr et al., 2018; Filarsky et al., 2018). The nature of the involvement of *mspdbl2* is not yet known, and it may be a marker of gametocytogenesis, rather than an active player. Chapter 4 explores this gene in further detail and aims to uncover whether or not it could be important in gametocyte commitment.

1.9 Aims and Objectives

Deep sequencing of genomes and transcriptomes was applied to *P. falciparum* and *P. knowlesi* parasites in order to gain deeper understanding of population genomics and parasite biology. The first aim of this thesis was to profile the population genomics of *P. knowlesi* clinical isolates collected from peninsular Malaysia using whole-genome sequencing and investigate divergence and selection based on the analysis of single nucleotide polymorphisms. The second aim was to use low input techniques to obtain high quality transcriptomic sequence data from clinical malaria isolates for which only a limited amount of parasite material is available in order to investigate the expression profile of *P. falciparum* parasites that express the merozoite surface protein MSPDBL2, considered as a possible marker of gametocytogenesis. Finally, in an extension of these low input techniques, transcriptomics was attempted from single *P. falciparum* infected erythrocytes.

1. Population genomics of *P. knowlesi* in peninsular Malaysia

P. knowlesi is a major cause of malaria in Malaysia. Whole genome sequencing has revealed dramatic diversity between populations of *P. knowlesi* parasites throughout Malaysia. The proposed population Cluster for peninsular Malaysia (Cluster 3) is not understood on a genome-wide level, and profiling these parasites is essential for understanding their evolution and adaptation as human infections increase.

To do this, genomic DNA extracted from *P. knowlesi* clinical isolates from peninsular Malaysia underwent sequencing using Illumina chemistry to obtain genome-wide sequence data. High quality genome-wide sequence data were compared to the *P. knowlesi* reference genome to identify single nucleotide polymorphisms, which were used to inform population genomics and investigate selection and diversity within these

parasites and between *P. knowlesi* populations in peninsular Malaysia and Malaysian Borneo.

2. Transcriptomic profiling of MSPDBL2 in *P. falciparum* clinical isolates

Gametocytogenesis is an essential stage of the *Plasmodium* life cycle, however is poorly understood from a genetic and mechanistic perspective. The gene *mspbl2* is suspected to be involved in, or be a marker for gametocyte commitment, but is not yet characterised. In order to further understand where this gene fits into the gametocytogenesis pathway, its expression was investigated in *P. falciparum* clinical isolates. *Plasmodium falciparum* clinical isolates from West Africa were cultured *ex vivo*, and schizonts were harvested after approximately 48 hours prior to re-invasion. Immunofluorescence assays were used to profile the expression of MSPDBL2 protein in schizonts. Clinical isolates underwent whole transcriptome sequencing using Illumina chemistry and the transcriptomic data was analysed to identify genes with expression correlating to MSPDBL2 protein expression.

3. Applying single-cell transcriptomics to *P. falciparum* schizonts

Plasmodium parasites display striking transcriptional plasticity, and this can be difficult to investigate using bulk methods. Some genes (*mspdbl2*, for example) show strikingly different expression between clonal parasites and can be investigated with single cell sequencing. In order to apply these methods to *Plasmodium* parasites, two were tested and optimised for use on *P. falciparum* laboratory lines.

A commercially available kit was tested and used to sequence transcriptomes from flow-sorted single schizont-infected erythrocytes. A microfluidic platform was tested with schizont-infected erythrocytes as a potential alternative to the method based on cell

sorting. The quality and reliability of the transcriptomes sequenced from the single cells was assessed by bioinformatic analysis.

2 Materials and Methods

2.1 *Plasmodium* culturing methods

2.1.1 Thawing of *Plasmodium* parasites

Cryopreserved *Plasmodium* isolates were removed from liquid nitrogen and allowed to thaw at room temperature. Half the sample volume of room temperature 12% NaCl was added dropwise while mixing gently by hand and the sample left to incubate at room temperature for 5 minutes. Following this, 10x times the sample volume of 1.6% NaCl was added slowly in a 50ml falcon tube to facilitate consistent mixing. Samples were then pelleted by centrifugation at 500g for 5 minutes. The supernatant was removed, and the pellet was washed twice by resuspending in RPMI-1640 (Sigma-Aldrich) media supplemented with 0.2% AlbuMAX™ II (Thermo Fisher Scientific, MA, USA), 2mM L-glutamine and 20µg/ml of Gentamycin (complete media) and pelleting by centrifugation at 500g for 5 minutes, before being resuspended in complete media and freshly washed red blood cells at approximately 2.5% haematocrit.

2.1.2 Preparation of red blood cells

Whole blood of varying blood types was collected from anonymous donors at the London School of Hygiene and Tropical Medicine and stored at 4˚C. Working in a sterile environment, the buffy coat layer was first removed from the blood, before the erythrocytes were washed three times with complete media by centrifugation at 2500g for 5 minutes. Washed blood was then diluted to 50% haematocrit with complete media and stored at 4˚C. A+ research red cells (NHS Blood and Transplant Service) were taken from 4˚C storage and washed once in complete media prior to use, before being resuspended at 50% haematocrit in complete media and stored at 4˚C. Washed blood was used for no longer than 7 days or until lysis could be seen in Giemsa-stained parasite smears.

2.1.3 *Plasmodium falciparum* laboratory line culture conditions

Parasites from the *Plasmodium falciparum* 3D7 laboratory line were maintained in culture at < 5% haematocrit in complete media. Parasites were grown with either washed whole blood or A+ research red cells prepared according to Section 2.1.2. Cultures were incubated shaking at 37°C in an atmosphere of 95% N_2 and 5% CO_2, and media was changed no less than every two days.

2.1.4 *Plasmodium falciparum* clinical isolate culture conditions

Clinical isolates of *P. falciparum* were maintained in culture at 2.5% haematocrit in complete media. Parasites were cultured with washed whole blood (various blood types) incubated with 5% O_2, 5% CO_2, and 90% N_2, at 37˚C while shaking at 60 R.P.M. *P. falciparum ex vivo* cultures were grown to maturity within original host erythrocytes and harvested prior to reinvasion.

2.1.5 Synchronisation of *Plasmodium falciparum* parasites using Percoll®

Long-term laboratory lines of *P. falciparum* were synchronised using a Percoll® (GE Healthcare, IL, USA) gradient. Firstly, cultures were spun down for 2 minutes at 500g and resuspended at approximately 25% haematocrit. In a separate tube, 70% Percoll® with 2.9% Sorbitol in 1X PBS was overlaid with 35% Percoll® (1.4% Sorbitol). The 25% haematocrit culture was then carefully pipetted to lay on top of the 35% Percoll®. This was then spun at 2500g for 10 minutes at 24˚C with zero break to prevent disturbing the Percoll® gradient. Cellular debris remained suspended on the 35% gradient and was discarded. Schizonts suspended on the 70% Percoll® layer were transferred into a fresh falcon tube and washed with complete media by centrifugation at 500g for 2 minutes. The late-stage enriched pellet was then either put back into culture (Section 2.1.3) or parasites were harvested for downstream use.

2.1.6 Enrichment *of Plasmodium falciparum* schizonts from clinical isolates by magnetic separation

As clinical isolates do not reliably allow purification of schizonts using Percoll®, these were enriched by magnetic activated cell sorting (MACS®) using a manual separator with magnetic LD separation columns (Miltenyi Biotec, Germany). As parasites mature, they digest iron obtained from red blood cells, the product of which is deposited within the parasite as haemozoin. The haemozoin concentration increases as a parasite matures, and young parasites (rings and early trophozoites) do not contain any haemozoin. Magnetic-activated separation uses ferromagnetic beads held in a column, magnetised by an external magnetic stand. As parasites are passed through the tube, the haemozoin-containing later parasite stages bind, and the remaining parasites are washed away. These late parasites can then be collected by removing the tube from the magnetic field and washing the beads (Ribaut *et al.,* 2008).

For the separation, parasite cultures were first pelleted by centrifugation at 500g for 2 minutes and resuspended in no less than 1ml of fresh complete media. LD columns were placed on the magnetic stand and charged by running through with 12ml of complete media. Resuspended parasite cultures were then applied to the column and allowed to run through. The column was then washed with 3ml of complete media, or until the run-through was clear. The column was then removed from the magnetic stand and placed in a fresh 50ml falcon tube. 1ml of media was applied to the column and forced through to dislodge bound parasites. Parasites were then pelleted by centrifugation at 500g for five minutes. 0.5µl of pelleted cells was used for microscopy to assess staging of the parasites using Giemsa staining. Remaining pelleted cells were resuspended in <500µl of complete media with 10µM of E64 to prevent schizont egress in a round-bottom 96-well plate and incubated for 4.5 hours in an atmosphere of 5% O_2,

5% CO_2,, and 90% N_2 at 37˚C. 0.5µl of these parasites was used to make a smear to assess staging at the end of the 4.5 hours using Giemsa staining.

2.2 Immunofluorescence Assays

For immunofluorescence assays, previously produced and published polyclonal mouse serum raised against MSPDBL2 was used. This was generated from a construct designed against the conserved N-terminal of MSPDBL2. The recombinant protein was expressed in *E. coli* and antibodies were raised by inoculation into laboratory mice, with the final serum collection occurring seven days after the final inoculation (Amambua-Ngwa et al., 2012).

Schizonts from *P. falciparum* clinical isolates in Chapter 4 were isolated by magnetic separation, matured in the presence of E64 and mixed with uninfected erythrocytes to give a parasitemia of 2-3% before being washed three times in cold 1% bovine serum albumin (BSA) in 1X PBS. On the final wash, the supernatant was removed, and the cell pellet resuspended in 1% BSA to give a final haematocrit of 2%. 15µl of this was spotted into 12-well slides (Hendley-Essex, U.K.) and left to dry overnight. In the morning, slides were carefully wrapped in tissue and stored with silica pouches in sealed bags at -40˚C. For long-term storage, slides were moved into purpose-built boxes packed with silica gel pouches and, kept at -40˚C.

For antibody staining, slides were removed from -40˚C and thawed in a “desiccation” chamber (an air-tight plastic box lined with silica pouches). Slides were then fixed in 4% paraformaldehyde, 0.0075% glutaraldehyde in 1X PBS for 30 minutes. The fixative was exchanged for 0.01% Triton™ X-100 in 1X PBS and left to incubate for 10 minutes at room temperature. Finally, slides were blocked overnight at room temperature in 3% BSA/0.01% sodium azide in 1X PBS. The next day, slides were

washed in excess sterile-filtered 1X PBS for five minutes with shaking, and then dried on a heat block set at 50˚C. On each slide one well was designated as a "secondary antibody-only control" well and 15µl of 3% BSA in PBS was added to this well. The remaining wells were incubated with 15µl of a 1:200 dilution of αMSPDBL2 polyclonal mouse serum (diluted into 3% BSA with 0.01% sodium azide in 1X PBS) for 3 hours at room temperature in a sealed box to minimise evaporation. Slides were then rinsed three times in excess 1X PBS as done previously and dried on a heat block. Next, the slides were incubated in a dark sealed box with a 1:1000 dilution of the secondary antibody goat anti-mouse IgG Alexa Fluor™ 555 (diluted into 3% BSA in PBS) for one hour (Thermo Fisher Scientific, MA, USA). Slides were then washed three times and dried as described previously. Vectashield® mounting fluid with DAPI (Sigma Aldrich, MO, USA) was added to each well and the slide sealed with a coverslip and nail varnish. Slides were stored in the dark at 4˚C prior to fluorescence microscopy. Microscopy was carried out on a manual Leica microscope with Retiga colour camera.

2.3 Extraction of RNA from parasites using TRIzol® reagent

RNA from non-limiting parasites samples was extracted using the phenol-chloroform protocol. Parasite cultures were spun down, and the supernatant removed. The packed cells were then resuspended in 5X the pellet volume off TRIzol® and immediately stored at -80˚C. For phenol-chloroform extraction, first 200µl of chloroform was added per 1ml of TRIzol® preserved sample, which were mixed prior to incubating at room temperature for 10 minutes. Samples were then centrifuged at 13,000rpm for 15 minutes. After centrifugation the upper phase containing the RNA was transferred to a new tube. To clean up the extracted RNA, the RNeasy® Micro kit (Qiagen, Germany) was used following the manufacturer's instructions, with the inclusion of the additional DNase step and using an elution volume of 30µl. Eluted RNA was stored at -80˚C.

2.4 RNA extraction from limited starting material

In Chapter 4 *ex vivo* clinical isolates of *P. falciparum* were harvested for mature schizonts from limited amounts (~100-1000µl) of infected blood. Parasite material destined for RNA extraction was stored in either TRIzol® reagent or RNA*later*™ (Thermo Fisher Scientific, MA, USA). Material preserved in RNA*later*™ was initially stored overnight at 4˚C to allow full penetration of the cells before being moved to -20˚C for long-term storage. Samples in TRIzol® were kept at -80 ˚C. RNA was extracted from samples stored in TRIzol® as described in Section 2.3.

RNA was extracted from the material stored in RNA*later*™ by phenol-chloroform extraction using TRIzol® LS reagent at 10X the volume of RNA*later*™ and RNA was cleaned up using the NucleoSpin® RNA XS extraction kit (Macherey-Nagel, Germany), following manufacturer's instructions. RNA quality was assessed on the Bioanalyzer using the RNA 6000 Pico kit (Agilent Technologies, CA, USA).

2.5 Reverse transcription and amplification of low input RNA from clinical isolates

Samples showing successful RNA extraction were reverse transcribed and amplified using the SmartSeq® v4 Ultra® Low Input RNA Kit for Sequencing (Takara Bio. Inc., Shiga Prefecture, Japan). SmartSeq® v4 reactions were carried out following manufacturer's instructions but using ½ volumes of reagents as this has been shown to give equivalent quality results (Dr. Sarah Tarr, personal communications, 2018). 12 PCR cycles were used during the amplification step. Amplified cDNA was run on the Bioanalyzer using the High-Sensitivity DNA analysis chemistry (Agilent Technologies, CA, USA) to check for successful amplification and fluorometrically quantified using a Qubit® 2.0 with dsDNA high-sensitivity reagents (Thermo Fisher Scientific, MA, USA).

2.6 Whole transcriptome amplification from single cells using REPLI-g®

In Chapter 5, the REPLI-g® whole transcriptome amplification kit (Qiagen, Germany) was used to obtain single cell transcriptomes. All single-cell work was completed in a PCR-free environment and within a Laminar Flow Hood cleaned with RNaseZap® and 70% ethanol and sterilised under a UV light for 15 minutes prior to use. The Manufacturer's protocol was followed initially before several permutations were tested in order to improve output. In the final protocol parasites were isolated (either by limiting dilution or FACS) directly into the REPLI-g® lysis buffer, supplemented with recombinant RNasin® (rRNasin®) at a final concentration of 1U/µl (Promega, WI, USA), and diluted to the appropriate volume using sterile, molecular grade, and RNase/DNase free 1X PBS. After cell lysis, further rRNasin® was added to the reaction at 1U/µl, before the genomic DNA was disrupted using the 'gDNA Wipeout buffer'. RNA then underwent reverse transcription using an oligo(dT) primer. The cDNA fragments are then ligated together into long strands up to 12kb in length. The REPLI-g® 'SensiPhi DNA Polymerase' amplifies the cDNA strands by MDA in an isothermal reaction. Alongside each experiment, 50pg of previously purified, DNase-treated *P. falciparum* 3D7 RNA was included as a control.

2.7 Fixation of parasites and whole transcriptome amplification using the Fluidigm C1™ platform

The Fluidigm C1™ system was used to isolate single parasites fixed with dithobis (succinimidyl propionate) (DSP) and carry out whole transcriptome amplification in Chapter 5. The cell fixation protocol using DSP was modified from Attar et al. (2018), as follows. After incubating parasites with E64, the cultures were spun down and the schizont pellet washed and then aliquoted out at the appropriate concentration according to manufacturer's instructions in 1X PBS. 200µl of DSP was added to the cells which

were then incubated at room temperature for 30 minutes. DSP was subsequently quenched using 1M Tris-HCl (pH 7.5), for a final concentration of 2mM. Fixed cells were stored at 4˚C. Prior to running on a C1™, cells were supplemented with 2% BSA to reduce cell stickiness and approximately 1700 cells were loaded onto the mRNA Integrated Fluidic Circuit (IFC) microfluidic chip (Fluidigm, CA, USA). The standard C1™ SmartSeq® v4 protocol was followed with the inclusion of 50mM DTT in the C1™ lysis mixture in order to reverse the DSP crosslinking. Capture sites into which cells are isolated by microfluidics were photographed and number of cells per site was recorded along with presence of any debris in each site's microfluidics.

2.8 Library preparation for Illumina sequencing

2.8.1 TruSeq Nano LT library preparation for DNA

Genomic DNA (gDNA) sequencing libraries were prepared using the TruSeq Nano LT DNA kit (Illumina, CA, USA), following the manufacturer's instructions. Genomic DNA was first fragmented into 350bp or 550bp fragments using the M220 Focused-ultrasonicator™ (Covaris, MA, USA) using 300ng of input gDNA. The fragmentation process results in DNA fragments with overhanging ends. The next step in library preparation converts these into blunt ends. Following this, the DNA fragments of desired size (550bp for reads 350bp in length) are selected for, to remove large and small fragments. Next, the blunt 3' ends of the DNA fragments are adenylated with a single "A", firstly to prevent DNA-DNA ligation during the adaptor ligation process, and also to facilitate the ligation of the adaptors, which contains a single overhanging "T" at their 3' ends. DNA fragments were then enriched by PCR, using primers that anneal to the ends of the adaptor sequences.

Quantification of completed libraries was carried out by quantitative PCR (qPCR) using the KAPA Library Quantification kit for Illumina® (Kapa Biosystems, MA, USA) on a

7500 Fast Real-Time PCR System (Thermo Fisher Scientific, MA, USA). Library quality and size distribution of fragments was assessed with the Agilent High Sensitivity DNA kit on the Agilent Bioanalyzer (Agilent Technologies, CA, USA) allowing accurate calculation of gDNA concentration and appropriate dilution factors. Samples showing significant primer dimer on the Bioanalyzer or containing less than 4nM of DNA by qPCR were not carried forward for sequencing. Samples passing quality criteria were normalised to 4nM and pooled equimolar in batches of no more than twelve. Library pools were denatured using freshly prepared 0.2N NaOH and diluted to a final concentration of 15pM and spiked with 1% PhiX. Paired-end sequencing was carried out on the Illumina MiSeq using the 600-cycle v3 MiSeq reagents using the 'generate FASTQ' programme and a read length of 300bp.

2.8.2 Nextera XT library preparation

For cDNA originating from low input, limiting samples the Nextera XT DNA Library Preparation Kit was used following manufacturer's instructions (Illumina, CA, USA), however reactions were carried out in ¼ volumes of those recommended in the reference guide. In brief, samples containing sufficient cDNA were first fragmented and tagged with DNA adaptor sequences using the 'tagmentation' chemistry. Tagged cDNA fragments were then amplified by PCR over 12 cycles, which also ligates the Index 1 and Index 2 adaptor sequences which are required for cluster generation and sequencing. Amplified libraries were cleaned by bead purification using Agencourt AMPure XP beads (Beckman Coulter, CA, USA). Libraries were run on a Bioanalyzer using the High-Sensitivity DNA analysis reagents to assess quality. Libraries were normalised by fluorometric quantification on a Qubit® 2.0 device using dsDNA high-sensitivity reagents and diluted to 4nM in RSB. Normalised libraries were pooled equimolar and stored at -20˚C prior to sequencing. Library pools were denatured in

0.1M NaOH and diluted to 15pM, and spiked with 1% PhiX. Paired-end sequencing was carried out using the 150-cycle MiSeq v3 reagents on an Illumina MiSeq using the 'generate FASTQ' programme.

2.9 Assembly of Illumina short reads

2.9.1 Genome Assembly

Prior to assembling the short reads to the reference genomes, the quality of each sample was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). In FastQC the GC% of each library was investigated, along with visual assessment of the average base quality along the reads. Short reads were then trimmed of low quality bases with Trimmomatic v0.3.2 using default parameters, which will drop reads if they are less than 36 bases in length, remove leading and trailing bases with low quality (less than 3), and move along the read in a 4-base sliding window, cutting the read when the average quality of the window drops below 15 (Bolger Lohse 2014). Trimmed reads were aligned to the *P. knowlesi* H-strain version 2.0 reference genome using Burrows-Wheeler Aligner version 0.7.15 using the bwa-mem algorithm and default parameters (Li and Durbin 2010), using paired reads outputted by Trimmomatic. Aligned reads were output in "sam" format and these were converted to "bam" format, sorted and indexed using samtools version 1.3.1 (Li et al. 2009). Reads originating from PCR duplication were removed using Picard version 2.9.1 "MarkDuplicates" function (https://broadinstitute.github.io/picard/). Average read depth for each isolate was estimated using samtools "depth" function, this calculation does not include sites that have zero coverage and calculated read depth for each region before averaging across the genome.

2.9.2 Transcriptome Assembly

Whole transcriptome short read sequence data were assembled by alignment to the *P. falciparum* 3D7 v3 reference genome using HISAT2 (Kim et al. 2015) with default parameters to generate 'sam' alignment files, which were converted into 'bam' format using samtools as was carried out in Section 2.9.1. Reads with MAPQ scores of <60 were removed using a custom, in-house python script.

2.10 Analysis of differential gene expression

Raw gene counts for each sample were made using the summarizeOverlaps function of the GenomicAlignments package (Lawrence, Huber et al., 2013) in the R statistical framework (R Development Core Team 2013). A custom made GFF reference file which had been prepared previously in which *P. falciparum* sub-telomeric, polymorphic gene families (*var, rifin, stevor*) and regions of the genome which showed allelic diversity between the *P. falciparum* 3D7 version 3 reference genome and ten schizont stage (40 h.p.i.) clinical isolate transcriptomes was used for generating read counts per gene (Tarr et al., 2018). Staging of samples was carried out by comparison of gene transcript levels in the clinical isolates with RNA-seq data obtained from a developmental time course of *P. falciparum* 3D7 (Otto et al. 2010). In this time course, RNA was sequenced from parasites harvested from seven timepoints (0 – 48 hours post invasion in 8 hour increments) (Otto et al. 2010). Normalised transcript levels for the clinical isolates and time course samples were calculated using Fragments Per Kilobase of transcript per Million mapped reads (FPKM) for all the genes using the DESeq2 "fpkm" function (Love et al. 2014). FPKM values for genes in each of the clinical isolates were correlated against the time course data using Spearman's Rank correlation. Differential expression analysis was carried out in R using the package DESeq2 (Love et al. 2014).

DESeq2 estimates differential expression using negative binomial generalised linear models takes a table of raw read counts per gene with an associated table holding phenotype data for each sample. Firstly, the data is normalised internally to correct for RNA composition bias and library size. It also estimates dispersion, calculated within a group, which describes how much the observed value is expected to be away from the mean value (Love et al. 2014). For visualisation of volcano plots, the R package EnhancedVolcano was used (Blighe 2019).

Genes with known or suspected involvement in gametocytogenesis were identified firstly by carrying out gene searches on PlasmoDB (https://plasmodb.org), each gene was investigated to identify those with known gametocytogenesis involvement for which literature was available. To identify genes with a suspected role in gametocytogenesis, which is not always immediately obvious from PlasmoDB, a master gene list of 121 genes (Appendix 1) was created using data gathered from publications relevant to the current research that covers a spectrum of gametocyte-related gene expression (Silvestrini et al. 2010; Filarsky et al. 2018; Josling et al. 2019). All genes identified in differential expression analysis were cross-referenced with this list to find overlap.

2.11 Identification of single nucleotide polymorphisms

Single nucleotide polymorphisms were identified for the *P. knowlesi* populations in Chapter 3. The first round of SNP calling was carried out on each sample independently. SNPs were initially marked using samtools mpileup run with the following flags: -B -I -Q 23 -d 2000 -C 50 and marked SNPs were then called using bcftools call -m -v and filtered using vcfutils.pl varFilter -d 10 -D 2000. The resulting per-sample SNP lists were then filtered and positions with a read depth of less than 30x were discarded. A read depth of 30x was used based on recommendation from Illumina,

and in line with previous research (Assefa et al., 2015; Divis et al., 2018). The SNPs for each sample were then compiled into a single file and filtered to exclude redundant SNPs, resulting in a 'unique SNP list' representing the *P. knowlesi* population, containing SNP positions and alternate bases.

It is difficult to assemble short reads to some parts of the genome such as low complexity, repetitive regions, and these were therefore masked from analysis. This includes the sub-telomeres and the multi-copy *kir* and *SICAvar* gene families, and low complexity intergenic regions between these. Boundaries of sub-telomeric regions were defined as in previous analysis (Divis et al. 2018), while *kir* and *SICAvar* genes were identified using the gene search function of PlasmoDB (www.plasmodb.org) and verified by inspecting the *P. knowlesi* annotation EMBL files in Artemis (Rutherford et al. 2000). In genomic regions where several polymorphic genes belonging to *kir* and *SICAvar* families occurred sequentially, the masked region was extended to include all genes and intergenic sequence, and extended on each side until the first SNP called after the final masked gene in that region. SNPs falling into these regions were filtered out of the unique SNP list using bedtools intersect v2.27.0 (Quinlan and Hall 2010). A second round of SNP calling and filtering was then implemented using custom made Perl scripts (Assefa et al. 2015) in order to make the major genotype call at each SNP position for each sample. SNP positions that had a 50:50 split in mixed base calls for a particular sample were deemed uncalled and replaced with 'N'. SNPs that were missing in more than 10% of the infection samples were not analysed.

2.12 Using single nucleotide polymorphisms to inform population genomic analysis

The packages adegenet (Jombart 2008) and pegas (Paradis 2010) in the R statistical framework were used to carry out principal component analysis and generate a

Neighbour-Joining tree of the samples using a genetic distance matrix based on SNP data (R Development Core Team 2013). For population structure analysis, SNP data was loaded into R and the package PopGenome (Pfeifer et al. 2014) was used to calculate nucleotide diversity, the within-population Tajima's D index, and the between-population fixation index (F_{ST}). Nucleotide diversity was calculated genome-wide in non-overlapping sliding windows of 50kb. To scan for genes that may be under balancing selection, Tajima's D was calculated on a gene-by-gene basis, with genes only being considered if they contained three or more SNPs. The fixation index (F_{ST}) was calculated for all individual SNPs across the genome with a minor allele frequency of at least 10%, and mean values were calculated in all sliding windows of 500 consecutive SNPs, overlapping by 250 SNPs, across the genome.

A scan for loci under recent positive selection was performed by identifying SNPs with an allele associated with extended haplotype homozygosity, using the R package rehh (Gautier et al. 2012). The standardised integrated haplotype score (|iHS|) is a statistic that is used to look for evidence that recent positive selection has acted at a particular locus. The score is based on the level of Extended Haplotype Homozygosity (EHH) surrounding a given ("core") SNP on the ancestral allele relative to the derived allele, where the EHH measures the decay of a haplotype around the core SNP. At each SNP, the iHS gives a measure for the strength of evidence for positive selection having acted on that SNP (Voight et al., 2006).

The |iHS| was calculated for biallelic SNPs with no missing calls and with a minor allele frequency of at least 10%. SNPs with |iHS| values in the top 0.01% were set as "core" SNPs, around which putative windows of selection were identified using the extended haplotype homozygosity (EHH) score, plotted until the EHH signal declined to less than 0.05 on each side. Overlapping windows of EHH resulting from more than

one core SNP were merged to produce one overall putative selection window for that region, and any gaps of over 20kb between SNPs with elevated |iHS| values were considered to break a putative window of selection. The Rsb statistic was also calculated in rehh and this is also derived from the EHH and comes from contrasting the haplotype lengths associated with the same SNP in different populations to identify regions under different selection in different populations (Tang et al., 2007).

Within-isolate mixedness was assessed for each *P. knowlesi* cluster independently using the F_{WS} metric, carried out as previously described (Manske et al. 2012) using custom R scripts kindly provided by Prof. Antoine Claessens (Montpellier University). For each SNP, the within-isolate (H_W) and within-population (H_S) heterozygosity were calculated and plotted in R to show the relationship between H_W and H_S for each isolate. The F_{WS} score is calculated from the gradient of this line and indicates the genome-wide estimate of heterozygosity. Isolates with scores ≥0.95 were considered to contain a predominant single genotype.

2.13 *msp-4* PCR for discriminating genomic DNA and cDNA

To quickly and reliably differentiate between cDNA and genomic DNA (gDNA) in Chapter 5, an intron-spanning PCR was developed within the gene *msp-4*. PCR was carried out with either HiFi (Clontech) or Phusion (Thermo Fisher Scientific, MA, USA) reagents and polymerases. Primers were designed within exons 1 and 2 of *msp-4* using the NCBI Primer-BLAST tool to encompass the *msp-4* intron (Figure 2.1). Forward primer sequence 5'-3' (within exon 1): CCCCCAATTTATCTGACGCA. Reverse primer sequence 5'-3' (within exon 2): TCATCTCCACAACCCCCATT. Cycling conditions were based on manufacturer guidelines with an extension time of

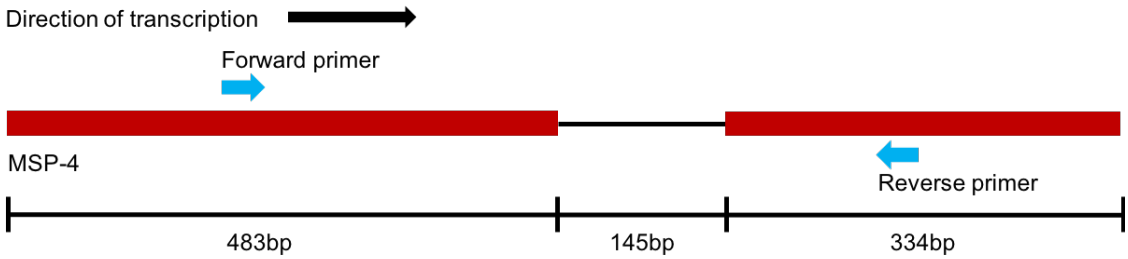

Figure 2.1 Primer strategy for the intron-spanning *msp-4* PCR. In order to assess whether single cell transcriptomic techniques had amplified cDNA originating from RNA or genomic DNA, the *msp-4* intron-spanning PCR was designed. Amplification of *msp-4* gDNA results in a 589bp product due to inclusion of the 145bp intron. Amplification of *msp-4* cDNA results in a 440bp product due to intron splicing.

one minute and an annealing temperature of 60˚C. The expected amplicon sizes are 589bp for gDNA and 440bp for cDNA.

2.15 Analysis of transcriptomes obtained from single *P. falciparum* schizonts

Analysis of single-cell and low cell number transcriptomes was carried out in R statistical software using the package Scater (McCarthy et al. 2016) within the Bioconductor framework. Gene counts for the transcriptomes obtained from single-cells were generated using HTSeq-count (Anders et al. 2015) with the "intersection strict" option using the *P. falciparum* GFF reference file described in Section 2.10. For the low input *P. falciparum* clinical isolates, the summarizeOverlaps function of the R GenomicAlignments package was used, so as to integrate with the DESeq2 recommended procedures (Lawrence, Huber et al., 2013). Contamination of non-*Plasmodium* microbial origin was identified using the taxonomic assignment software, Kraken, using the MiniKraken database (Wood and Salzberg 2014).

3. Genome-wide analysis of population structure and adaptation of *Plasmodium knowlesi* in peninsular Malaysia

3.1 Introduction

All human malaria parasite species originated as zoonotic cross-infections from non-human primates (Liu et al. 2010; Rutledge et al. 2017; Sundararaman et al. 2018), and the most prevalent species, *P. falciparum and P. vivax* now exclusively infect humans. *Plasmodium knowlesi* is one of the recognised human malarias that has remained primarily a zoonosis. Human infection from zoonotic malaria had been considered to be extremely rare, but findings in Malaysia (Singh et al. 2004; Cox-Singh et al. 2008), and subsequent surveys elsewhere in Southeast Asia have revealed that many human malaria cases are due to *Plasmodium knowlesi*, a malaria species associated with macaque hosts (Shearer et al. 2016). In Malaysia, overlapping distributions of the natural macaque hosts, mosquito vector species, and human populations have resulted in *P. knowlesi* being the cause of almost all human malaria cases (Singh and Daneshvar 2013). Transmission is thought to still be maintained exclusively in macaques, with human infection remaining opportunistic in nature. As several countries in Southeast Asia are working towards eliminating malaria, *P. knowlesi* represents a unique public health challenge. Due to the presence of wild reservoir hosts, elimination is unlikely, and the problem would worsen if the parasite adapts, or environments change to enable effective transmission between humans (William et al. 2013).

Investigations into parasite population genomics and dynamics are essential to understand whether there has been recent adaptation. In *P. knowlesi*, this understanding has been achieved by microsatellite genotyping (Divis et al. 2015; Divis et al. 2017) and whole-genome sequencing (Assefa et al. 2015; Divis et al. 2018). However, gaining a

full picture of the *P. knowlesi* population is complicated by its reservoir hosts. Whole-genome sequencing from human samples is usually straight-forward, as a human patient typically carries just a single *Plasmodium* infection. However, macaques are commonly host to several *Plasmodium* species, and often these are multi-genotype infections, which makes whole-genome sequencing using standard, short-read methods very challenging.

In Malaysian Borneo, *P. knowlesi* forms two genetically divergent sympatric populations, (termed Clusters 1 and 2), that are respectively associated with different reservoir hosts; long-tailed macaques (*Macaca fascicularis*) and pig-tailed macaques (*M. nemestrina*) (Divis et al. 2015). In peninsular Malaysia, on the Asian mainland, it is clear that there is a different genetic subpopulation of *P. knowlesi* (termed Cluster 3). This was initially identified by genome sequencing of a few old laboratory isolates (Assefa et al. 2015), however subsequent microsatellite analysis of many clinical cases from peninsular Malaysia has confirmed these to all be Cluster 3 type (Divis et al. 2017). Whole-genome sequencing of clinical isolates from Clusters 2 found that these parasites have a heterogenous pattern of nucleotide diversity throughout their genomes, unlike the fairly uniform pattern seen in Cluster 1 genomes, and analysis of inter-cluster divergence between Cluster 1 and 2 revealed large genomic regions of high and low divergence, indicative of recent introgression between the populations, with the genetic isolation between the clusters being maintained within the regions of high divergence (Divis et al. 2018). Independent, secondary analysis of whole-genome sequence data from Cluster 1 and 2 isolates has also indicated that these parasites have recently undergone introgression events (Diez Benavente et al. 2017).

With regards to future adaptation and epidemiology of *P. knowlesi*, any one of the three independently occurring Clusters of *P. knowlesi* might emerge further, particularly as human and macaque populations continue to encroach and overlap and it is an increasing cause for concern in Southeast Asia, particularly in Malaysia. In this Chapter, I aimed to increase our understanding of the population structure of *P. knowlesi* in peninsular Malaysia, following on from research that has shown that a population of parasites exists here that is geographically isolated and divergent from the populations profiled in Malaysian Borneo. The Tajima's D statistic will be used to uncover any signatures of balancing selection that may indicate whether there are any loci under increasing selection pressures from both the human and macaque host. Finally, signatures of positive selection will be identified and analysed to understand if there is evidence of any selective sweeps acting on the population from peninsular Malaysia and to see if these relate to possible adaptation to the human host. The degree of divergence between the three population clusters is impressive, and by analysing the Cluster 3 population structure, I hope to uncover possible genomic locations driving some of this divergence.

3.2 Materials and Methods

3.2.1 Sample collection and DNA extraction

The study was approved by the Medical Research and Ethics Committee of the Malaysian Ministry of Health, and by the Ethics Committee of the London School of Hygiene and Tropical Medicine. Heparinised venous blood samples of up to 10 ml were collected from 56 patients presenting with *P. knowlesi* malaria at five hospitals in Peninsular Malaysia between February 2016 and January 2018, after written informed consent from each patient was obtained Leukocytes were depleted by passing each blood sample through a CF11 cellulose column, to thereby increase the proportion of

parasite compared to host DNA. Genomic DNA (gDNA) was extracted using QIAamp DNA Mini kits (Qiagen, Germany), and all infections were confirmed to contain only *P. knowlesi* by nested PCR assays testing for all locally known malaria parasite species (Lee et al. 2011). Sample collection and gDNA extraction was carried out by colleagues at the University of Malaysia, Sarawak. Genomic DNA was lyophilised prior to transport to the UK, following which it was dissolved in 30µl of nuclease-free water and quantified on a spectrophotometer using the Quant-IT™ PicoGreen® ds kit (Thermo Fisher Scientific, MA, U.S.A). Samples containing at least 300ng of DNA were processed for sequencing.

3.2.2 Whole-genome sequencing of *P. knowlesi* isolates

Genomic DNA (gDNA) sequencing libraries were prepared using the TruSeq Nano LT DNA kit (Illumina), following the manufacturer's instructions (Section 2.8.1). Concentration and quality of the prepared libraries was assessed by qPCR and on the Bioanalyzer (Agilent Technologies, CA, USA) (Section 2.8.1), and those showing significant primer dimer or containing less than 4nM of DNA by qPCR were not carried forward for sequencing. Library pools were denatured diluted to a final concentration of 15pM and spiked with 1% PhiX. Paired-end sequencing using a read length of 350bp was carried out on the Illumina MiSeq (Section 2.8.1).

Quality of short reads was assessed, and reads were appropriately trimmed before assembly to the *P. knowlesi* H-strain version 2.0 reference genome (Section 2.9.1). Samples with an average read depth of less than 20x were excluded from further analysis. Samples with a low mapping rate and low coverage were presumed to be contaminated; with low mapping rates correlating to a GC content differing from 38%. After assembly and quality filtering data from 28 samples were available for

downstream analysis, representing samples from five hospitals in peninsular Malaysia (Figure 3.1). For comparison with samples from elsewhere, Illumina short reads were retrieved from previous studies (Assefa et al. 2015; Pinheiro et al. 2015; Divis et al. 2018) and assembled using the pipeline described in Section 2.9.1.

3.2.3 Analysis of whole-genome sequence data

Single nucleotide polymorphisms were called using a full repertoire of *P. knowlesi* genomes, including 28 new sequenced infection samples from peninsular Malaysia obtained in this study as well as 74 previous samples from Malaysian Borneo (40 from Cluster 1, 34 from Cluster 2), and 5 laboratory isolates (107 in total) (Assefa et al. 2015; Pinheiro et al. 2015; Divis et al. 2018). SNPs were called in two rounds, the first to obtain per-sample SNP lists, which were then filtered to remove those that fell into polymorphic regions before the second round of SNP calling was implemented to make the major genotype call at each SNP position (Section 2.11)

Analysis of SNP data was carried out in the R statistical framework using the packages adegenet (Jombart 2008) and pegas (Paradis 2010), used to elucidate population structure (Section 2.12). The R package PopGenome (Pfeifer et al. 2014) was used to calculate nucleotide diversity, the within-population Tajima's D index, and the between-population fixation indices (F_{ST}) using parameters and methods outlined in Section 2.12. A scan for loci under recent positive selection was performed by identifying SNPs with an allele associated with extended haplotype homozygosity using the integrated haplotype score (|iHS|) and Rsb metrics, calculated in the R package rehh (Gautier et al. 2012) (Section 2.12). Within-isolate mixedness was estimated with the F_{WS} score, calculated as described in Section 2.12.

3.3 Results

3.3.1 Population structure of *P. knowlesi* Cluster 3 in peninsular Malaysia

After processing and quality filtering, genome sequences were successfully obtained from 28 clinical samples of *P. knowlesi* from cases recruited at five locations in peninsular Malaysia (Figure 3.1, Table 3.1), none of the samples from Kuala Krai hospital met the criteria for sequencing. From the new infection samples, 994,761 SNPs were called initially, of which 40,934 were removed as they were in polymorphic regions which show generally unreliable short read mapping (*kir* and *SICAvar* genes, and sub-telomeres), leaving 953,827 SNPs present in the rest of the genome. SNPs in isolates that had mixed base calls were replaced with "N", and SNPs that had missing data in >10% of individual infection samples were removed (affecting less than 1% of SNPs), leaving a total of 474,109 SNPs were included in subsequent analysis.

A Neighbour-Joining (NJ) tree was generated using pairwise genetic distances among individual samples, which showed that all the newly sequenced samples belonged to a population (Cluster 3) genetically divergent from those in Malaysian Borneo (Clusters 1 and 2) (Figure 3.2A). These clinical samples clustered with the old, previously sequenced laboratory isolates, which had initially shown the existence of a third major genetic population within this species (Assefa et al. 2015). The overall genome-wide nucleotide diversity (π) for Cluster 3 as a whole was 4.13×10^{-3}. This is slightly higher than the nucleotide diversity for Cluster 1 ($\pi = 3.89 \times 10^{-3}$) and much higher than for Cluster 2 ($\pi = 2.13 \times 10^{-3}$) (Figure 3.2B).

Analysis of SNP allele frequencies genome-wide confirmed that the Cluster 3 population in peninsular Malaysia is highly divergent from each of the separate Clusters 1 and 2 in Malaysian Borneo, showing respective genome-wide mean F_{ST} values of 0.32

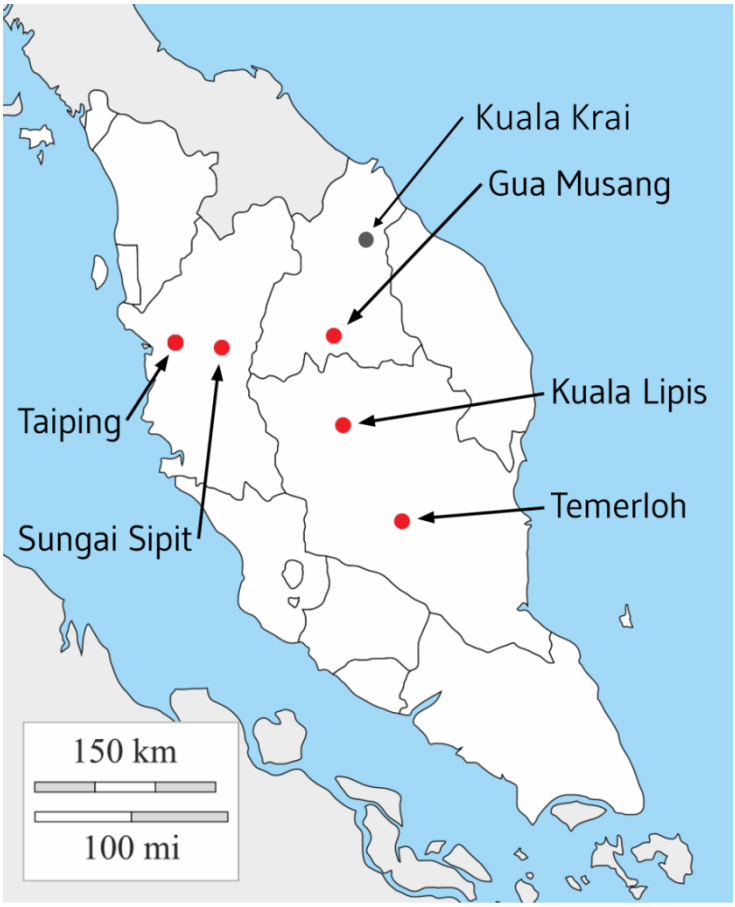

Figure 3.1. Locations of *P. knowlesi* clinical sampling in peninsular Malaysia. Sampling took place in the states of Perak (Taiping and Sungai Sipit), Kelantan (Gua Musang and Kuala Krai), and Pahang (Kuala Lipis and Temerloh). Out of 56 infection samples, 32 yielded *P. knowlesi* DNA of sufficient quantity and purity for Illumina Sequencing. Of these, 28 yielded high coverage genome-wide sequences for population genomic analysis (Table 3.1). No samples from Kuala Krai were of sufficient quality to be included in genomic analysis, and this location is therefore not represented by the final dataset.

Table 3.1. Details of the 28 Clinical isolates of *Plasmodium knowlesi* that were sequenced successfully and included in downstream analysis.

Sample	Actual Parasitemia (p/ul)	Year	gDNA concentration (ng/ul)	Library concentration (nM)	total reads ($x10^6$)	GC%	Mapped reads ($x10^6$)	Alignment rate (%)	Average coverage
GMK02	88164	2017	214	48.3	5.8	38%	5.7	98.9	47x
GMK03	16480	2017	26.8	68.1	5.6	40%	4.9	87.6	40x
GMK06	98800	2018	68.8	3.50	8.7	38%	8.6	98.5	74x
GMK07	12900	2018	12.5	5.99	7.9	38%	7.7	97.2	66x
KLK02	1932	2017	28.5	209	3.2	39%	3.1	97.1	40x
KLK04	50938	2017	147	298	4.0	38%	3.9	98.6	32x
KLK05	27251	2017	122	81.9	5.0	38%	4.5	98.2	36x
KLK06	661875	2017	283	45.2	5.8	38%	5.4	98.8	43x
KLK08	84400	2017	80.8	314	3.6	38%	3.6	98.7	91x
KLK12	7236	2017	22.8	89.5	3.6	38%	3.3	96.3	27x
KLK14	33060	2017	62.9	84.7	5.8	38%	5.7	98.1	46x
KLK15	136408	2017	304	80.4	5.4	38%	4.7	95.3	37x
KLK16	13756	2017	33.8	57.3	5.4	39%	5.1	85.4	37x
KLK17	22554	2017	13.3	15.6	5.8	38%	5.4	95.5	44x
KLK19	46035	2018	10.0	11.9	4.4	38%	4.2	94.3	37x
KLK21	60950	2018	59.6	9.87	4.5	38%	4.5	98.4	39x
KLK23	23288	2018	19.7	6.13	8.3	38%	8.3	98.9	71x
KLK24	141069	2018	93.9	7.15	5.7	38%	5.6	98.4	47x
KLK25	54352	2018	26.6	12.5	3.7	38%	3.7	98.1	32x
KLK27	69160	2018	22.7	4.71	7.8	38%	7.6	97.5	64x
SSK01	6027	2017	13.1	7.30	8.8	47%	5.1	61.2	40x
SSK02	2763	2017	13.5	15.9	7.6	38%	7.4	94.6	62x
SSK04	6750	2017	16.6	74.6	3.6	40%	2.9	83.4	22x
SSK05	6750	2017	90.2	316	3.8	38%	3.5	98.4	30x
TK01	3806	2017	12.1	7.60	4.4	47%	2.6	64.3	21x
TK03	10323	2017	51.0	80.0	6.0	39%	5.8	95.7	49x
TPK03	4000	2018	6.88	6.06	9.8	46%	6.3	64.3	55x
TPK06	12494	2018	12.8	7.36	8.8	38%	8.5	96.0	74x

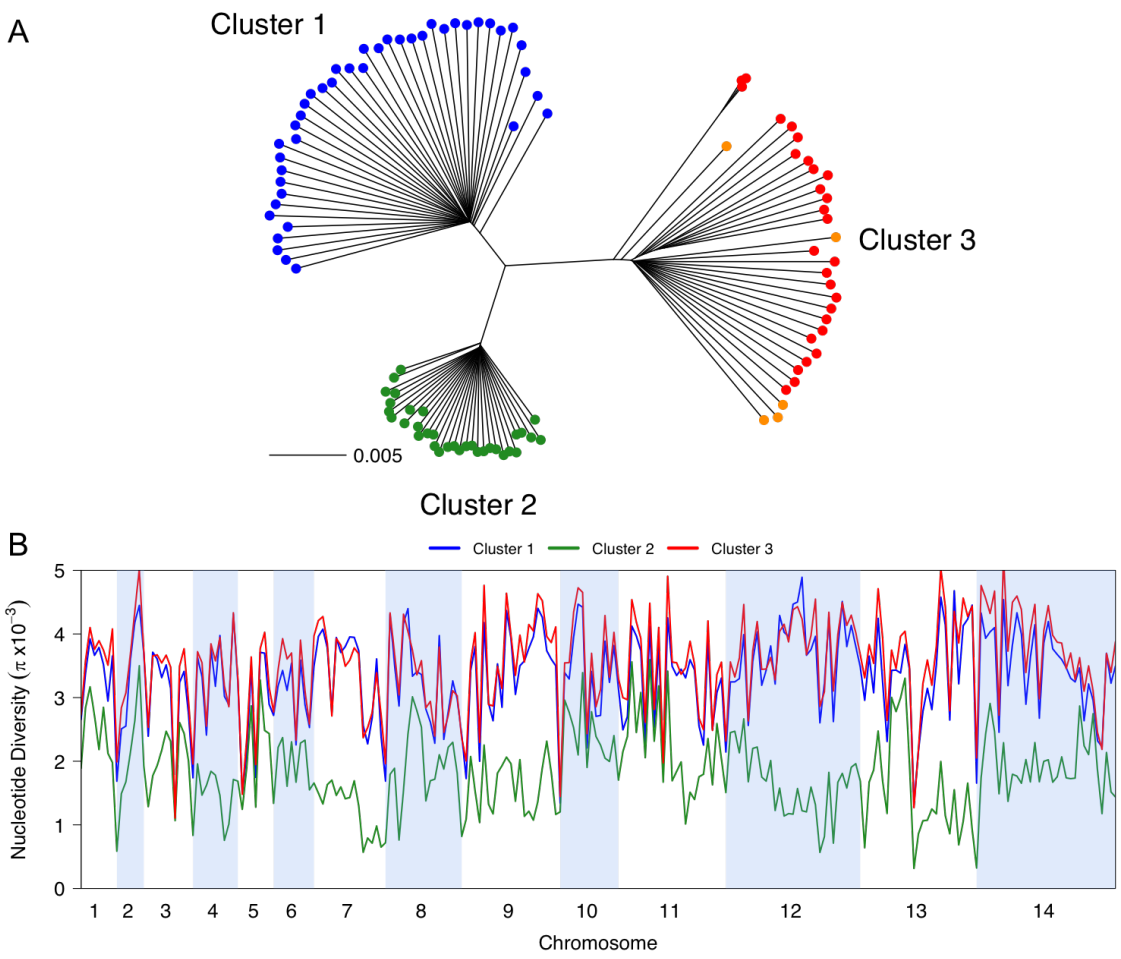

Figure 3.2. Population structure of Cluster 3 in the context of Clusters 1 and 2. A. Neighbour-Joining tree based on a pairwise genetic distance matrix between individual *P. knowlesi* infection samples, for the 28 new clinical samples from peninsular Malaysia (shown in red), five previously sequenced laboratory isolates (orange, most were originally isolated from peninsular Malaysia many years ago) (Assefa et al. 2015), and 74 samples from Malaysian Borneo that grouped into separate clusters (Cluster 1 shown in blue, Cluster 2 in green) (Assefa et al. 2015; Pinheiro et al. 2015; Divis et al. 2018). All the clinical isolates from peninsular Malaysia grouped into Cluster 3 together with the laboratory isolates. The distance matrix is based on the proportion of all SNPs with differences between each infection sample (scale bar shows branch length for 0.5% of SNPs differing), with the majority of reads within each infection sample determining the allele scored for each SNP. **B.** Genome-wide scan of nucleotide diversity (π) for *P. knowlesi* among the clinical isolates in peninsular Malaysia (Cluster 3, red), compared with diversity seen in the sub-populations in Malaysian Borneo (Clusters 1 and 2). The sliding window plot shows values of nucleotide diversity for non-overlapping windows of 50 kb in each of the 14 chromosomes.

and 0.42 (Figure 3.3), with 3713 and 6738 SNPs being at complete fixation in the respective comparisons (Figure 3.4 and 3.5, respectively). There is little variation across the genome in the level of divergence when comparing Cluster 3 with Cluster 1 (Figure 3.3A), but more variation in the comparison with Cluster 2 (Figure 3.3B). This is due to a previously described mosaic pattern of diversity across the genome of Cluster 2 (Divis et al. 2018), which contributes to greater genome-wide heterogeneity in divergence between Clusters 1 and 2 in Malaysian Borneo (Figure 3.3C) than between either of these and Cluster 3.

The Neighbour-Joining tree also indicated branching among the Cluster 3 clinical samples from peninsular Malaysia into three different sub-clusters; containing fifteen, ten, and three of the infection samples (Figure 3.2A). To examine this further, Principal Component Analysis (PCA) based on the SNP data was carried out, and the individual samples separated into three groups as predicted (Figure 3.6A). Three samples clustered tightly, away from the rest along Principal Component 1 (which accounted for 10.5% of the overall variation), while the others clustered into two less tightly separated groups along Principal Component 2 (which accounted for 6.5% of the overall variation). These three groups existing within Cluster 3 in peninsular Malaysia are considered to be sub-populations, and have been labelled here as sub-cluster A (15 infections), B (10 infections), and C (three infections). They are not separated geographically within peninsular Malaysia, as samples from each group have been detected at multiple sites (Figure 3.6B), and one hospital (in Kuala Lipis) had infections of all three sub-clusters.

The most divergent sub-cluster (C) consisted of infections that were highly related to each other, being virtually identical in large parts of the genome (Figure 3.6C); despite the fact that each of these infections came from different hospitals, located in three

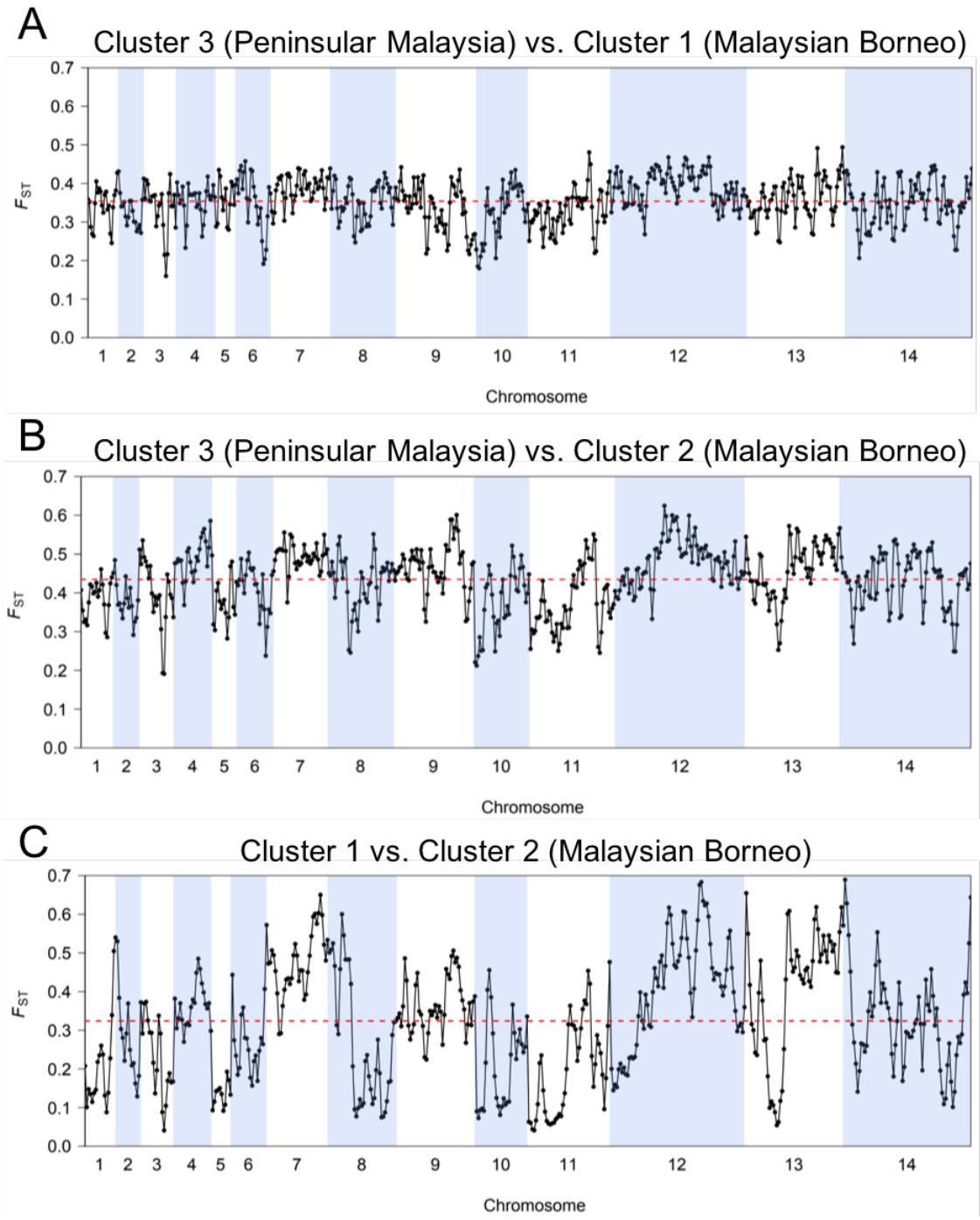

Figure 3.3. Genome-wide F_{ST} scan of divergence between *P. knowlesi* Clusters 1, 2, and 3. All SNPs with overall allele frequencies of at least 10% were included, and analyses show values for windows containing 500 consecutive SNPs, centred by the midpoint of each sequential window and overlapping by 250 SNPs. Red dotted lines show the genome-wide average F_{ST}. **A.** The level of divergence between Cluster 3 in peninsular Malaysia and Cluster 1 in Malaysian Borneo does not differ greatly throughout the genome (mean F_{ST} = 0.32). **B.** Divergence between Cluster 3 in peninsular Malaysia and Cluster 2 in Malaysian Borneo is slightly higher (mean F_{ST} = 0.42), and shows more heterogeneity between genomic regions due to mosaic structure of diversity in Cluster 2 (as explained by the bottom panel). **C.** Divergence between Clusters 1 and 2 in Malaysian Borneo, showing marked heterogeneity across the genome, correlating with the more modest variation in the panel above due to mosaic structure of diversity in Cluster 2 as previously reported (Divis et al. 2018).

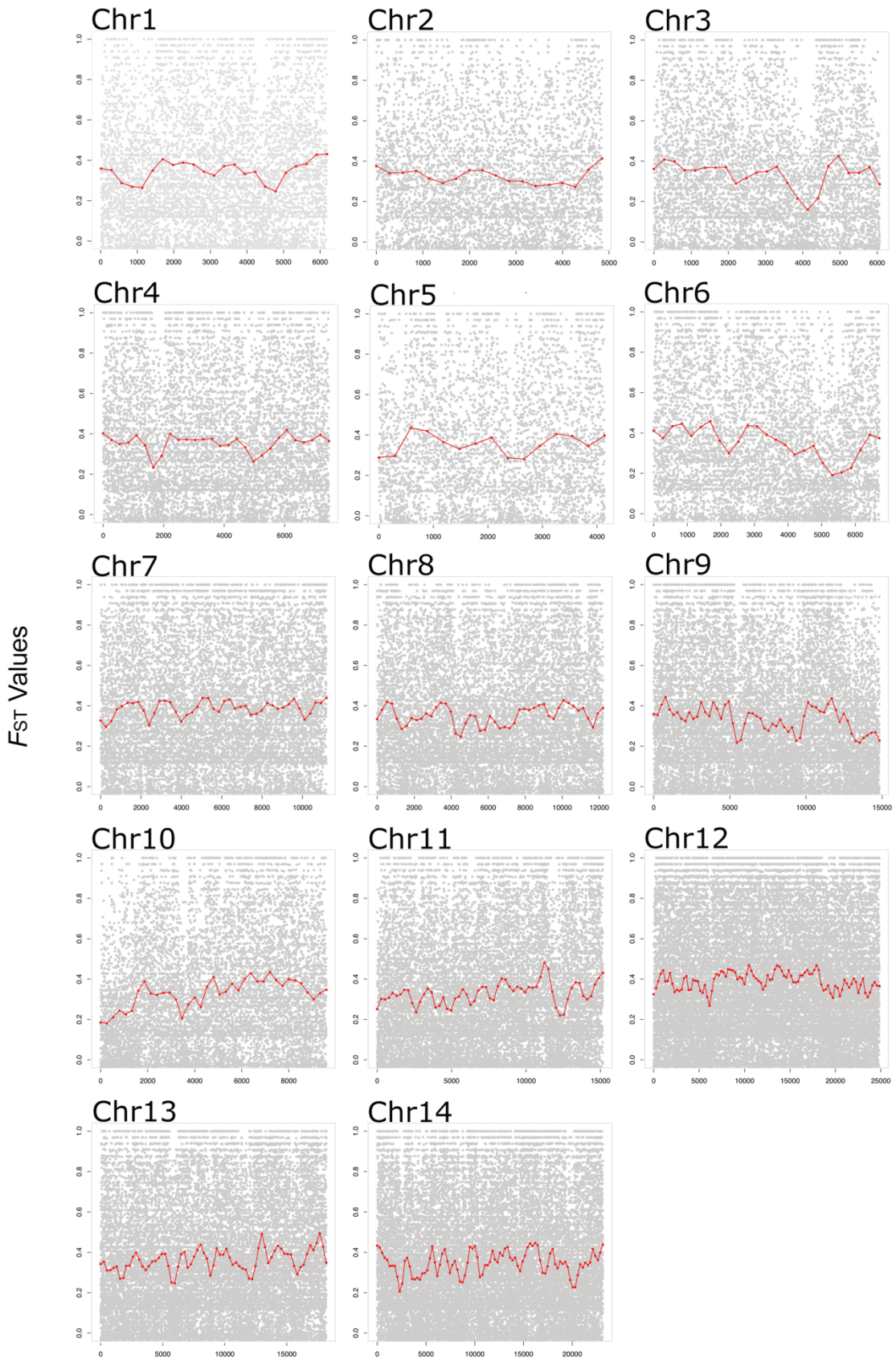

Figure 3.4. F_{ST} values shown for all individual SNPs comparing between *P. knowlesi* Cluster 3 in peninsular Malaysia and Cluster 1 in Malaysian Borneo. SNPs with a minor allele frequency of < 10% were excluded. In red are shown the means for overlapping windows of 500 SNPs (step size of 250 SNPs along each chromosome).

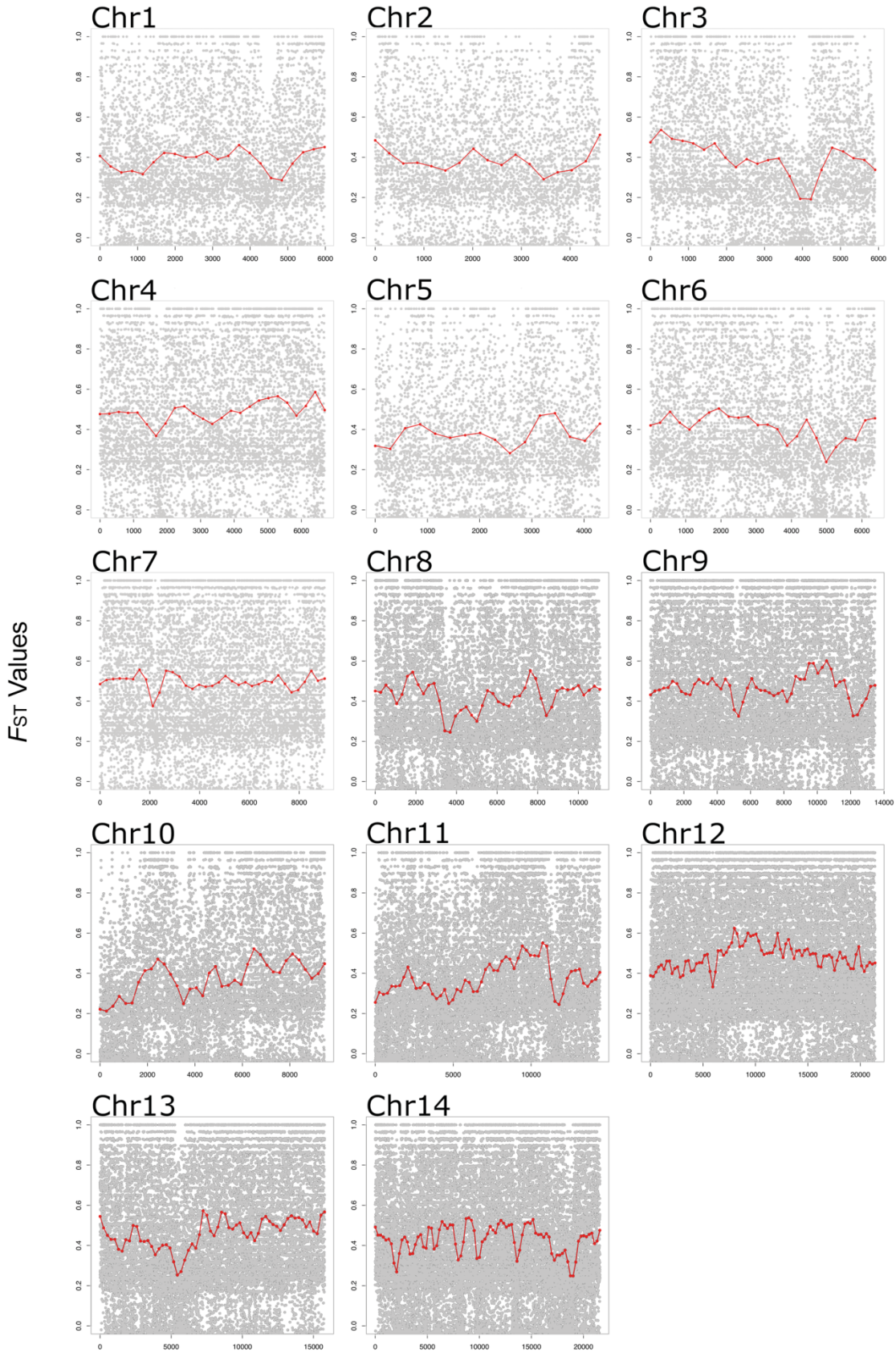

Figure 3.5. F_{ST} values shown for all individual SNPs comparing between *P. knowlesi* Cluster 3 in peninsular Malaysia and Cluster 2 in Malaysian Borneo. SNPs with a minor allele frequency of < 10% were excluded. In red are shown the means for overlapping windows of 500 SNPs (step size of 250 SNPs along each chromosome).

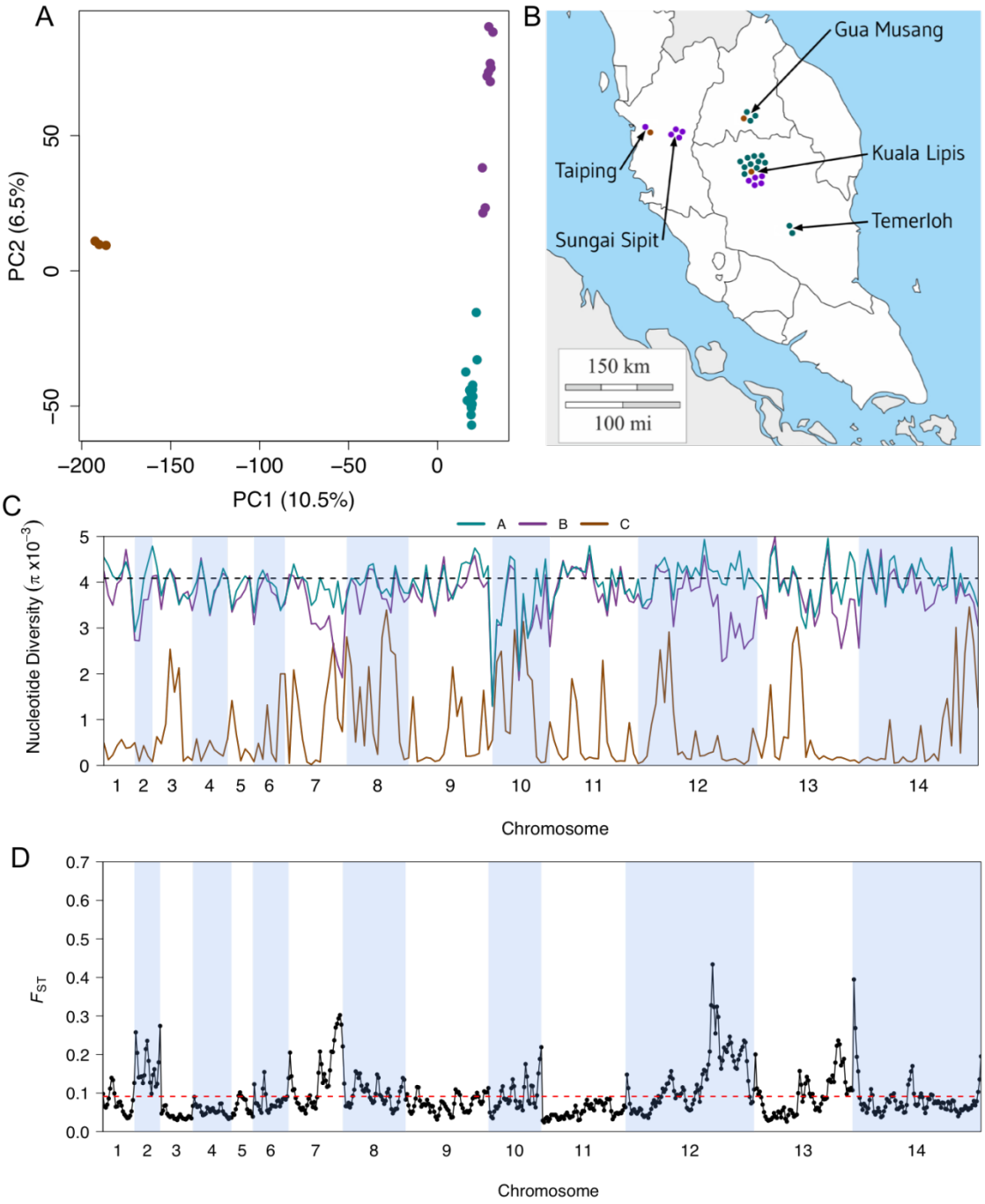

Figure 3.6. Population substructure *of P. knowlesi* Cluster 3 in peninsular Malaysia. A. Principal Component Analysis (PCA) of the 28 Cluster 3 *P. knowlesi* clinical isolates from peninsular Malaysia shows that these cluster into three groups, termed sub-clusters A (15 isolates), B (10 isolates) and C (three isolates). The assignment of all samples to these three sub-clusters is completely consistent with their placement in the within-Cluster 3 branching of the Neighbour-Joining tree based on the pairwise distance matrix (Figure 3.2A). **A.** The first Principal Component accounts for 10.5% of overall variation and separated sub-cluster 3 from the others, while the second Principal Component accounts for 6.5% of overall variation and separated sub-clusters A and B. **B.** Each of the Cluster 3 *P. knowlesi* sub-clusters was detected at multiple sites within peninsular Malaysia (points shown at each of the five sampling sites show individual infections with colours for the different sub-clusters as in the previous panel). The site with most samples had all three sub-clusters co-occurring locally. **C.** Genome-wide scan of diversity shows that the sub-cluster C samples are virtually

identical in large parts of the genome, whereas sub-clusters A and B are both highly diverse throughout the genome, with only a few genomic regions showing lower diversity in sub-cluster B compared to A (in chromosomes 2, 7, 12 and 13). **D.** Genome-wide scan of differentiation between sub-clusters A and B by sliding window F_{ST} analysis shows peaks of differentiation corresponding to regions with differences in diversity. Most notable is a large region of chromosome 12 having many windows with F_{ST} values exceeding 0.2, and containing some individual SNPs with fixed differences (Figure 3.8).

different states in peninsular Malaysia (Figure 3.6B). The validity of the low nucleotide diversity seen in sub-cluster C was assessed by calculating nucleotide diversity of three random samples from sub-cluster A and sub-cluster B (Figure 3.7). It was found that the nucleotide diversities from the three sub-cluster A samples and three sub-cluster B samples were broadly comparable to the estimates for the entire sub-clusters. This indicates that the regions of extremely low nucleotide diversity seen in sub-cluster C are biologically meaningful, and represent isolates that are highly related, perhaps where a more clonal population may have developed due to selection pressures unique from sub-clusters A and B.

Although sub-clusters A and B had similar levels of nucleotide diversity to each other genome wide (overall $\pi = 4.02 \times 10^{-3}$ and 3.78×10^{-3} respectively), sliding window analysis indicated a few genomic regions in which sub-cluster B shows lower diversity than sub-cluster A, the most prominent of these regions covering half of chromosome 12 (Figure 3.6C).

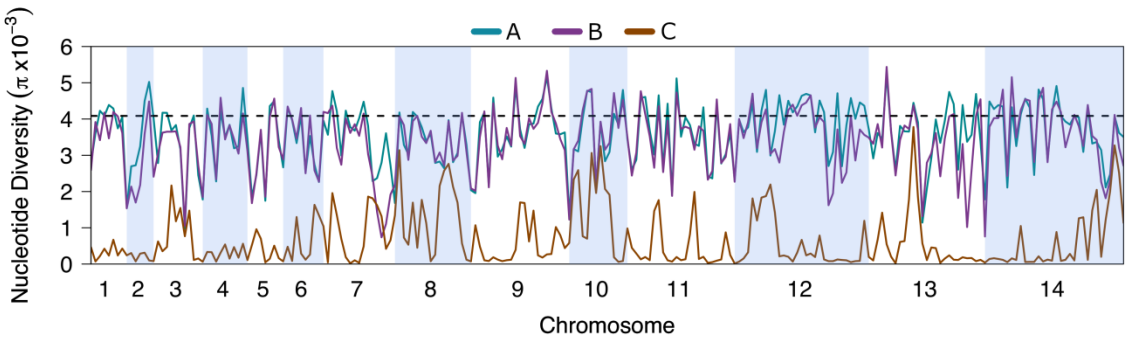

Figure 3.7. Comparison of nucleotide diversity between sub-cluster C and three samples from each sub-cluster A and sub-cluster B. In order to validate the low diversity seen in sub-cluster C and to explore whether this was due to the limited number of samples, or is biological meaningful, three samples were randomly selected from sub-clusters A and B and genome-wide nucleotide diversity was calculated for these. As can be seen, the sub-cluster A and B diversities from limited samples were comparable to each other and to Figure 3.6C. This indicates that the low diversity seen in sub-cluster C is biological in nature, and not an artefact of limited sample numbers.

A comparison of differentiation by F_{ST} sliding window analysis within Cluster 3 between sub-clusters A and B revealed a background of low differentiation, interspersed with peaks of high differentiation. The genome-wide average F_{ST} was 0.073. The regions that showed differences in levels of nucleotide diversity are also the most differentiated between the sub-clusters, most notably a large region of chromosome 12, in which many windows had F_{ST} values exceeding 0.2 (Figure 3.6D), and also containing some individual SNPs with fixed differences (Figure 3.8). The other regions of differing nucleotide diversity between sub-clusters A and B on chromosomes 7 and 13 also show elevated F_{ST} values.

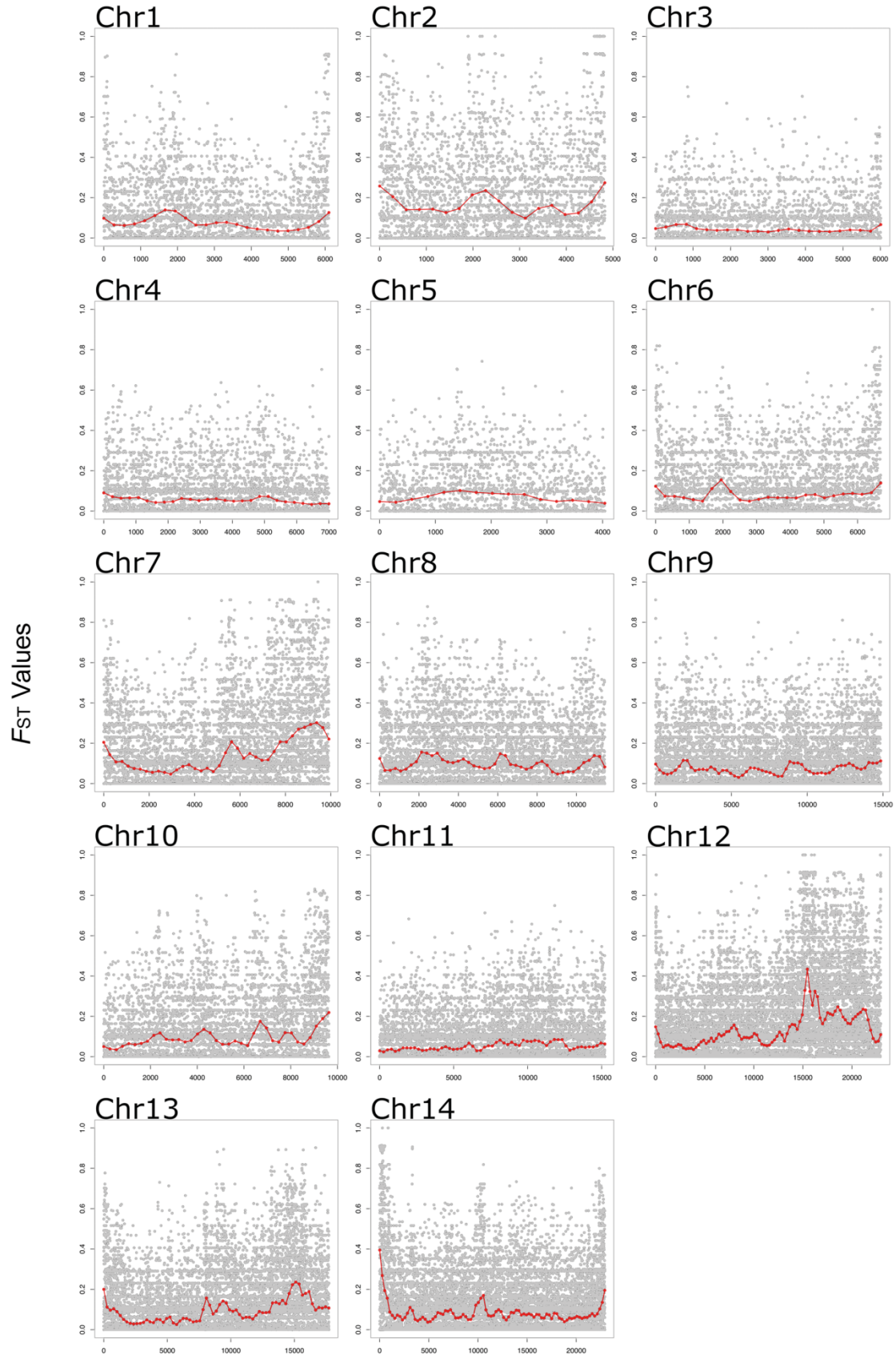

Figure 3.8. F_{ST} values shown for all individual SNPs comparing between *P. knowlesi* Cluster 3 sub-cluster A and sub-cluster B. SNPs with a minor allele frequency of < 10% were excluded. In red are shown the means for overlapping windows of 500 SNPs (step size of 250 SNPs along each chromosome).

3.3.2 Potential signatures of balancing selection in Cluster 3

In order to scan for genomic loci that may be under different selection pressures in *P. knowlesi* in peninsular Malaysia, nucleotide site allele frequencies were first summarised by calculating the Tajima's D index for all 4742 genes containing three or more SNPs. Overall, values were negatively skewed, with a mean of -0.86 (Figure 3.9A), and only 215 genes had values above zero, or which just eight had values above 1 (Figure 3.9B). The genome-wide pattern is consistent with expectations if there had been long-term population size expansion. Individual genes with the highest Tajima's D values may be under balancing selection, and should be considered separately (Appendix 2). The gene with the highest Tajima's D (D) value of 1.81 (locus PKNH_0309800) encodes a protein of unknown function. Some genes with high values have orthologues in other malaria parasite species that are likely targets of immunity, including a tryptophan-rich protein (PKNH_1472400, D = 0.98), a 6-cysteine protein (PKNH_1254400, D = 0.61), an exported protein PHIST (PKNH_0808500, D = 0.57), and an MSP7-like protein (PKNH_1265900, D = 1.15). However, some genes with orthologues considered to be targets of immunity in other malaria parasite populations have negative Tajima's D values here, including the circumsporozoite protein (*csp*) gene which had the highest Tajima's D value genome-wide in Cluster 1 *P. knowlesi* in Malaysian Borneo (Assefa et al. 2015), as well as the apical membrane antigen 1 gene (*ama1*, D = -1.35), the duffy binding protein alpha (*DBPα*, D = -0.89) and normocyte-binding protein gene (*NBPXa*, D = -0.42).

Tajima's D was also estimated for the larger two Cluster 3 sub-clusters (A and B). It is important to note that these results exist against a background of limited sample numbers (15 and 10), which will artificially inflate Tajima's D values. As such, precise Tajima's D values are less useful than investigating trends between the different sub-

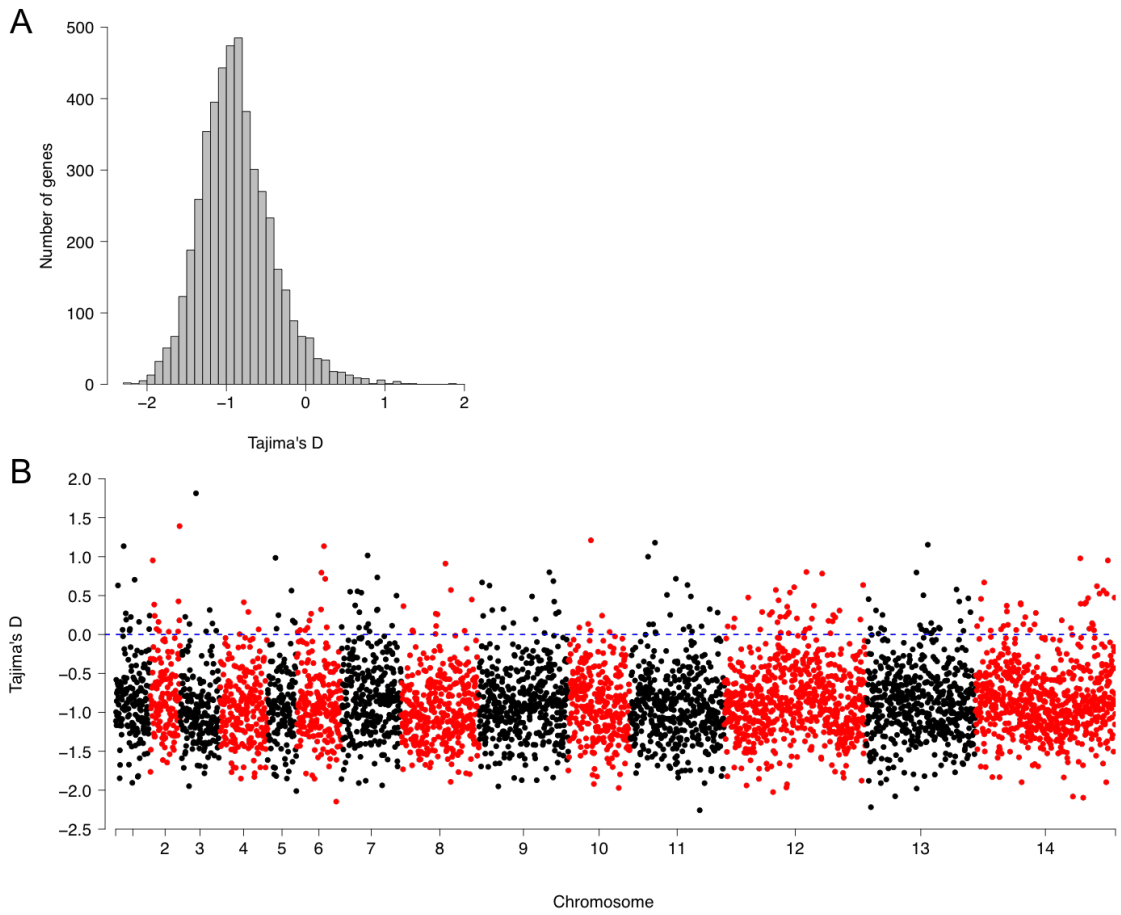

Figure 3.9. Variation in nucleotide site allele frequency spectra (summarised by Tajima's D index) across 4742 *P. knowlesi* genes. Only genes with three or more SNPs were included in the analysis. **A.** Overall values were negatively skewed with a mean Tajima's D of -0.86, consistent with a pattern that would be caused by long-term population size expansion. **B.** The minority of genes with high values are distributed throughout the genome, some of which may be under balancing selection (individual values for all genes are shown in Appendix 2).

clusters. Both sub-clusters had negative genome-wide Tajima's D, and despite the limitations imposed by small sample numbers, the region of divergent nucleotide diversity, and high F_{ST} differentiation between sub-clusters A and B on chromosome 12 is also shown to have divergent Tajima's D values.

3.3.3 Selection acting upon chromosome 12 in Cluster 3

Chromosome 12 was consistently presented as a region of divergence between Cluster 3 sub-clusters A and B, showing divergent nucleotide diversity, high fixation, and elevated Tajima's D values between the sub-clusters. To more accurately identify the region of divergence, the moving average of Tajima's D values for both sub-clusters was calculated and plotted in windows of 50 genes (with a step size of 1 gene per window) (Figure 3.10). This method identified a region of nearly one million base pairs (1,833,887bp-2,823,486bp), which corresponded to 193 genes, that were contributing to the divergent Tajima's D values in chromosome 12, with elevated values seen in sub-cluster B. A second, smaller region with a positive moving average occurred subsequently, and was just under 200,000 base pairs in length (2,855,684bp-3,051,598bp), and corresponded to 63 genes. The gene with the highest Tajima's D (1.96) in the larger or the two divergent regions was locus PKNH_1252400, encoding a TBC domain protein. The overall mean Tajima's D for the genes in this region for sub-cluster B was 0.07, which is much higher than for the same region for sub-cluster A (-0.74) and higher than the average Tajima's D across the whole of chromosome 12 for sub-cluster B (-0.29). Other genes in these regions, with Tajima's D values of greater than one include *kelch13* (PKNH_1257700, TD = 1.38), an early transcribed membrane protein gene orthologous to *etramp14.2* in *P. falciparum* (PKNH_1246400, TD = 1.31), and the gene encoding an MSP7-like protein (PKNH_1265900, TD = 1.15), along with several uncharacterised genes of unknown function.

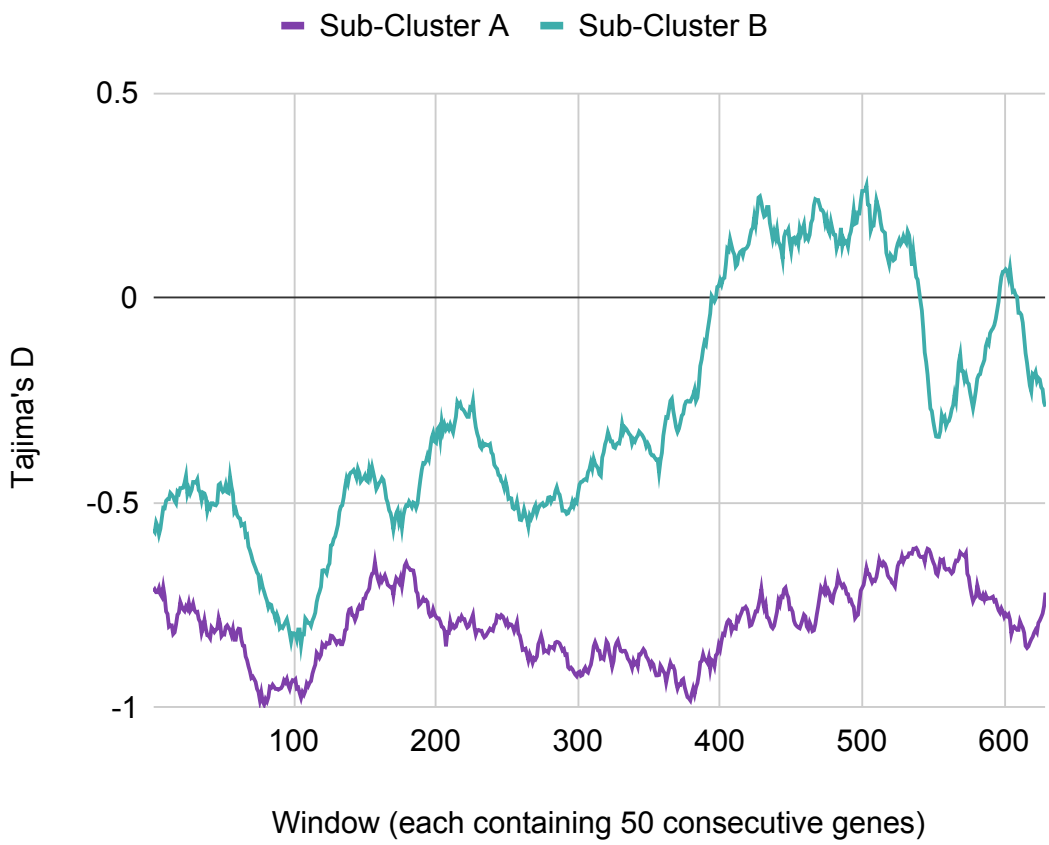

Figure 3.10. Variation in nucleotide site allele frequency spectra (summarised by Tajima's D index), plotted as a moving gene average across chromosome 12. Moving average has been plotted for Cluster 3 sub-cluster A (blue) and B (red). Sub-cluster A shows a negative Tajima's D distribution across chromosome 12, whereas sub-cluster B has two regions of positive Tajima's D distribution. The positive regions correlate with the region in chromosome 12 showing lower diversity in sub-cluster B.

3.3.4 Signals of positive selection in Cluster 3

The standardised integrated haplotype score (|iHS|) index was used to scan for evidence of genomic regions affected by recent positive directional selection. When analysing the full population complement of clinical isolates from peninsular Malaysia, SNPs with standardised |iHS| values in the top 0.01% had the extent of their influence evaluated by examination of the ranges of their extended haplotype homozygosity. This identified four distinct genomic windows of extended haplotypes (Figure 3.11 and Appendix 3). Two of these windows (in chromosomes 1 and 9) spanned across regions which included *SICAvar* and *kir* gene clusters that had been masked from SNP calling and analysis. This was an interesting observation, indicating unlikely positive selection of the regions around these polymorphic genes. However, these genes are not well mapped and do not have well characterised gene boundaries, and it is also likely that the extreme selection upon these regions is affecting the surrounding genomic regions. As it is not known if this phenomenon is biological or technical in nature and is likely due to poor mapping in these regions, these EHH windows were not further analysed. The other two windows that did not span *SICAvar* and *kir* genes covered ~28 kb (11 genes) on chromosome 9 and ~315 kb (81 genes) on chromosome 12. The large region on chromosome 12 coincides with the region that has high genomic divergence and differentiation between Cluster 3 population sub-cluster A and B (Figure 3.4D). Although analysis of |iHS| in sub-clusters A and B separately is statistically weak due to low sample sizes, examination of the data shows that the chromosome 12 region has differing |iHS| values sub-cluster A and sub-cluster B (Figure 3.12). Overall, evidence suggests that recent selection has operated strongly on a locus on chromosome 12, and that this has particularly affected part of the *P. knowlesi* population in peninsular Malaysia.

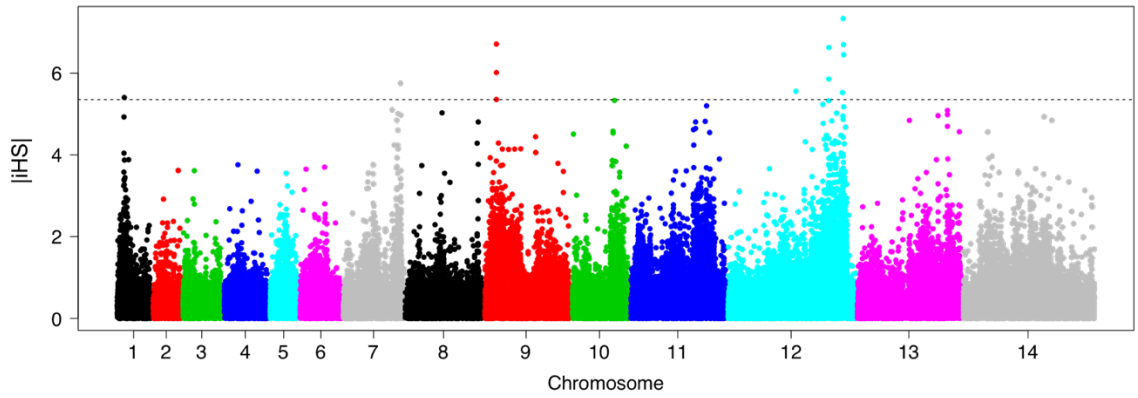

Figure 3.11. Scan for evidence of genomic regions affected by recent positive directional selection in *P. knowlesi* Cluster 3 using the standardised integrated haplotype score |iHS| index. Examination of the ranges of extended haplotype homozygosity for individual SNPs with high |iHS| values identified four distinct genomic windows of extended haplotypes (specified in detail in Table S3). Two of these (in chromosomes 1 and 9) spanned across *SICAvar* and *kir* genes which were masked from SNP calling, while the other two did not include *SICAvar* or *kir* genes but covered ~28 kb on chromosome 7 and ~ 315 kb on chromosome 12. The large region on chromosome 12 is a merged window, comprised of five high |iHS| core SNPs having overlapping windows of extended haplotype homozygosity, and coincides with the region of chromosome 12 that has the highest genomic divergence between Cluster 3 population sub-clusters A and B (Figure 3.4D).

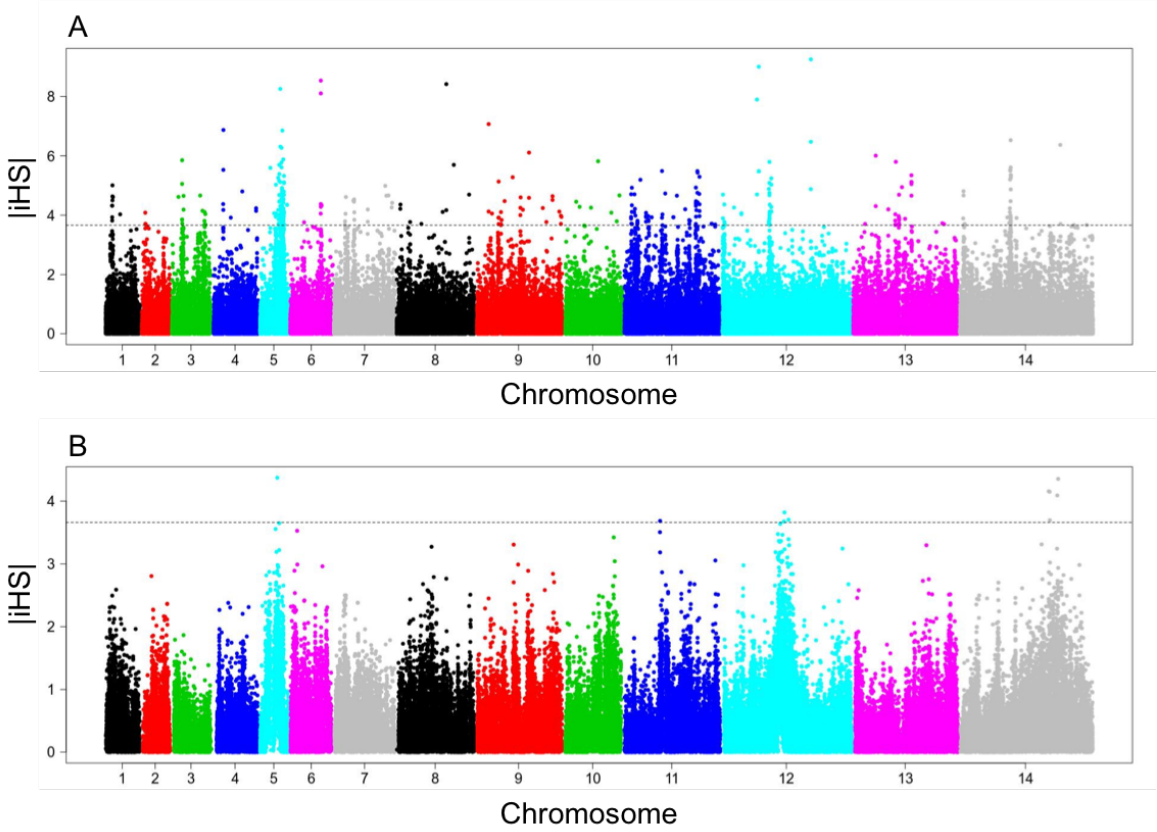

Figure 3.12. Scan for evidence of genomic regions affected by recent positive directional selection in *P. knowlesi* Cluster 3 sub-clusters using the standardised integrated haplotype score |iHS| index. A. |iHS| index for SNPs in sub-cluster **A.** B. |iHS| index for SNPs in sub-cluster B, which has lower values genome-wide, with particularly low values for SNPs in the region of divergence on chromosome 12.

To scan for differences in recent selection between populations of *P. knowlesi*, the Rsb metric was calculated for pairwise populations of *P. knowlesi*. When comparing Cluster 3 to the Malaysian Borneo Cluster 1 and 2 (Figure 3.13A and 3.13B), two regions were identified in which identical SNPs showed elevated Rsb values (Appendix 4A and 3B). Peaks in these regions were not seen in the Cluster 3 sub-cluster A and sub-cluster B pairwise comparison (Figure 3.12C), and possibly represent a geographic difference in selection. SNPs in these shared regions occur clustered together over small genomic areas, and occur both within genes and in intergenic DNA. The first of these regions is on chromosome 10. SNPs here did not fall in protein-coding DNA, and instead are restricted to the region between a gene encoding a V-type proton ATPase 16 kDA proteolipid subunit (PKNH_1013800) and one encoding protoheme IX farnesyltransferase (PKNH_1013900). The second region is on chromosome 14, with SNPs occurring within and in the region downstream of DNA gyrase subunit B (PKNH_1459600). Other SNPs with high Rsb values occurred mostly in isolation or alongside a single other SNP. The cross-population Rsb was also calculated for the Cluster 3 sub-clusters A and B (Appendix 4C). Four regions were identified with more than one SNP with elevated Rsb in close proximity, on chromosomes 6, 8, 11, and 14. None of these regions overlap with those identified in the between Cluster 3 and Clusters 1 and 2. The region with the most number of SNPs with elevated Rsb values was on Chromosome 11, occurring both within and upstream to the gene encoding RAP protein PKNH_1120900.

3.3.5 Estimation of *P. knowlesi* mixed infections

The within-infection F_{WS} fixation index was used to calculate the genetic diversity within infections (Auburn et al. 2012; Manske et al. 2012). The F_{WS} was calculated independently for each of the *P. knowlesi* population clusters (Appendix 5). In the

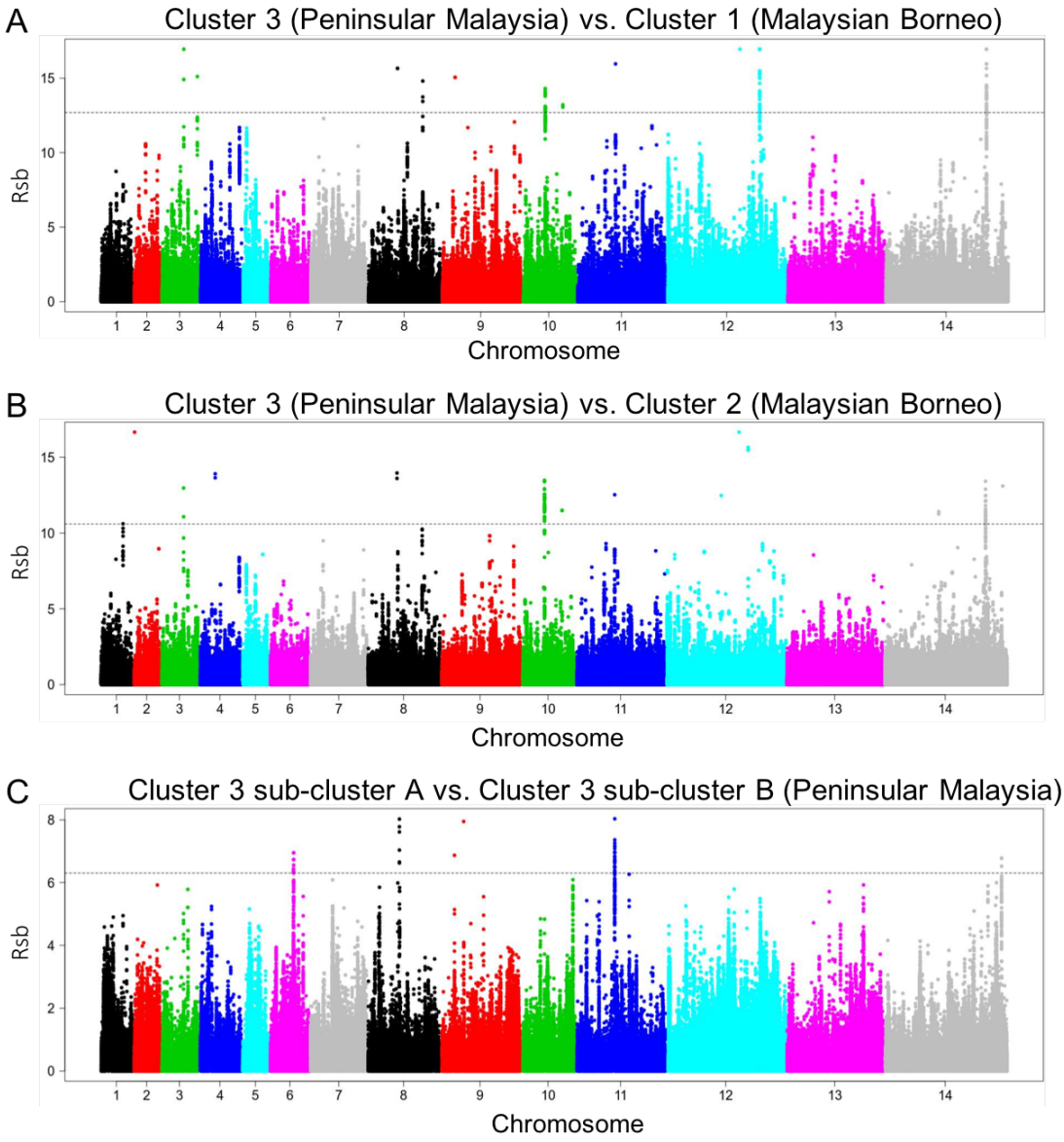

Figure 3.13. Scan for evidence of positive directional selection acting on a population level using the Rsb metric, across all SNPs identified for *P. knowlesi*. A. Pairwise population comparison using Rsb between Cluster 3 and Cluster 1. This identified five regions containing more than one elevated SNP. **B.** Pairwise population comparison between Cluster 3 and Cluster 2, identifying seven regions containing more than one elevated SNP. Two regions (on chromosome 10 and chromosome 14) had high Rsb values for identical SNPs between Clusters 3 and 1 and Clusters 2 and 1, indicating a possible geographical separation in these regions. **C.** Rsb metric for Cluster 3 sub-cluster A and sub-cluster B. No SNP in this comparison reaches the Rsb values seen for inter-Cluster tests. Four regions contained more than one SNP with elevated Rsb values.

process, allele frequencies for each SNP were calculated within each individual infection (H_W) and for the local population as a whole (H_S), and the gradient of H_W/H_S was used to calculate the F_{WS} score for each infection. An F_{WS} score of ≥0.95 (close to the complete fixation value of 1.0) is considered to represent a predominantly single genotype infection, whereas samples with an F_{WS} of <0.95 are very clearly composed of more than one genotype (Manske et al. 2012). Cluster 1, comprised of 41 clinical isolates had an F_{WS} values ranging from 0.82 to 0.99 (mean = 0.98), and four of these samples (9.8%) had scores of <0.95 (Figure 3.14 A, D). Cluster 2, comprised of 33 samples had an F_{WS} values ranging from 0.90 to 0.99 (mean = 0.99), and five of these samples (15.2%) had scores of <0.95 (Figure 3.13 B, D). Cluster 3, comprised of 28 clinical isolates, had an F_{WS} values ranging from 0.76 – 0.99 (mean = 0.98), with four out of the 28 isolates (14.3%) having F_{WS} scores of <0.95 (Figure 3.14 C, D). There was no significant difference found between the F_{ws} estimates for each Cluster (Cluster 1 vs. Cluster 2 t=-0.52, P=0.60; Cluster 1 vs. Cluster 3 t=0.04, P=0.96; Cluster 2 vs. Cluster 3 t=-0.41, P=0.69). The high F_{WS} scores for the *P. knowlesi* isolates indicate that the genotyping and SNP calling methods used in this chapter were appropriate for the data.

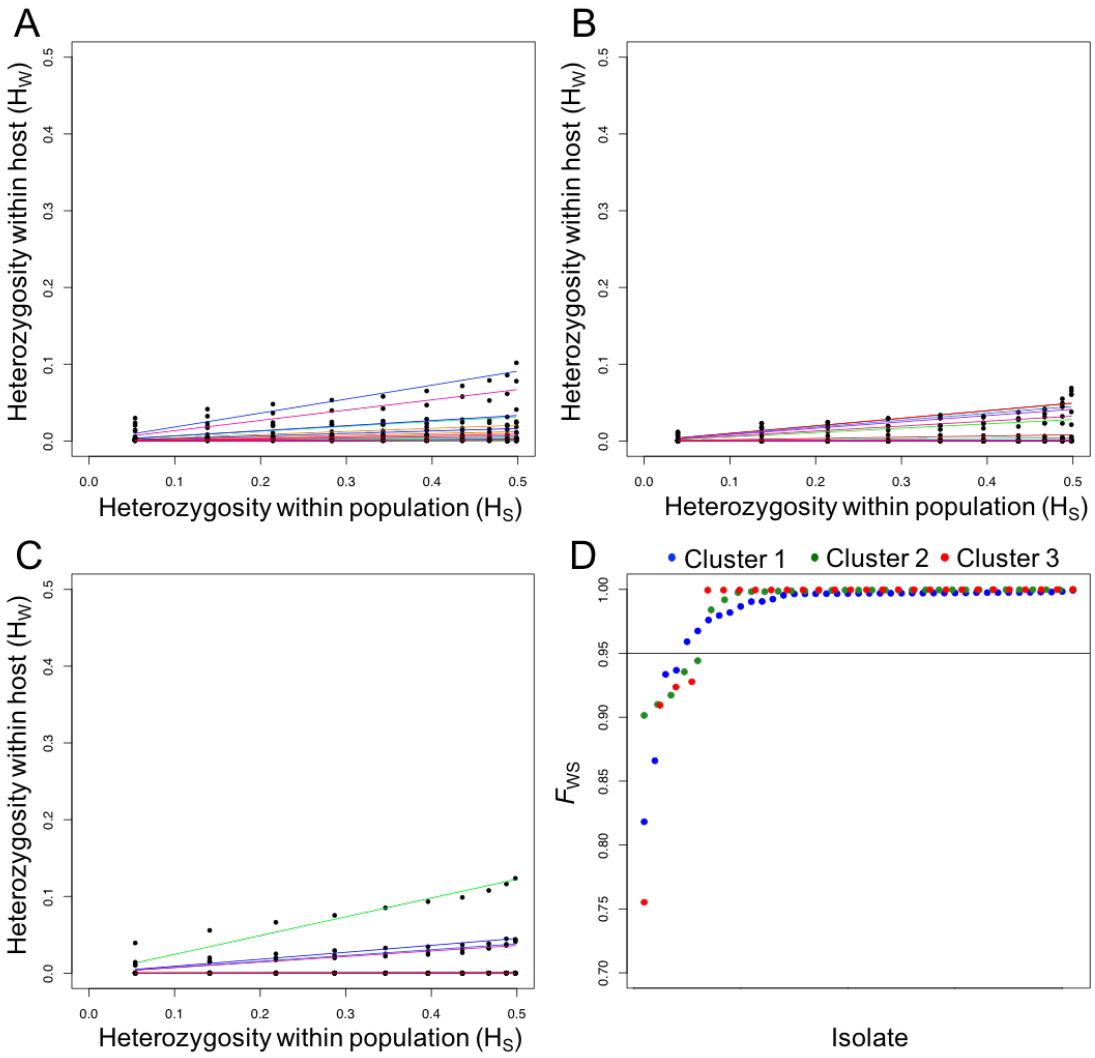

Figure 3.14 Within-host and within-population diversity measured using the F_{WS} metric for *P. knowlesi* population clusters. A-C. Within-host heterozygosity (H_W) and within-population heterozygosity (H_S) based on SNP allele frequencies show a linear relationship. Each line represents this relationship for individual isolates of Cluster 1 (**A**), Cluster 2 (**B**), and Cluster 3 (**C**). **D**. The F_{WS} metric, calculated from the gradient of the H_W/H_S relationship, for each isolate of Cluster 1 (blue dots), Cluster 2 (green dots), and Cluster 3 (red dots). The horizontal line at F_{WS} 0.95 shows the cut-off above which isolates are considered to be predominantly single genotype infections.

3.4 Discussion

Genomic analysis has revealed unexpected parasite population structure and evidence of recent selection in *P. knowlesi* in peninsular Malaysia. Following previous multi-locus microsatellite analysis, it was expected that whole genome sequencing of samples from peninsular Malaysia would reveal a population distinct from those in Malaysian Borneo but grouping together with several old laboratory isolates. This was clearly shown, all samples from peninsular Malaysia belonged to a genetic population termed Cluster 3 that is highly divergent from both of the Clusters 1 and 2 in Malaysian Borneo (genome-wide mean F_{ST} values of 0.32 and 0.42). In addition, and as has been indicated by previous whole genome sequencing (Assefa et al., 2015), Cluster 3 has a high genome-wide nucleotide diversity, being comparable to the diversity seen in Cluster 1. Cluster 2 has a much lower nucleotide diversity due to the mosaic structure of parasites in this population, as has been described previously (Divis et al., 2018).

As mentioned, Cluster 3 is highly diverse, and the clinical samples from peninsular Malaysia diverged into three distinct sub-clusters. The cause of the genetic population structure within peninsular Malaysia needs to be determined. However, as the old laboratory isolates that had been previously sequenced are dispersed throughout the sub-structure of the peninsular Malaysia clinical isolates, it is likely that this separation is well-established. There does not appear to be any geographic separation, as each of the three sub-clusters was found in overlapping locations, and all were detected from the hospital with most samples analysed. The population structure may reflect more than one zoonotic transmission cycle, or could be a sign of recent selection and potentially emergence of a sub-population of *P. knowlesi* transmitted more effectively between humans.

In Malaysian Borneo, long-tailed and pig-tailed macaques are respective reservoir hosts for the Cluster 1 and 2 populations of *P. knowlesi* transmitted to humans (Divis et al. 2015; Divis et al. 2017), but it is unknown whether different reservoir hosts contribute to the parasite population structure within peninsular Malaysia. Microsatellite analysis of *P. knowlesi* in long-tailed macaques from peninsular Malaysia has indicated that most of them belong to Cluster 3, although some samples from long-tailed macaques from two locations had indeterminate cluster assignments (Divis et al. 2017). It will be important to sample and genotype parasites from pig-tailed macaques in peninsular Malaysia, as well as to analyse more samples from long-tailed macaques, to investigate whether the parasite sub-clusters have different reservoir host species locally.

Genetic sub-clusters of *P. knowlesi* in peninsular Malaysia may also be transmitted by different mosquito species. *Plasmodium knowlesi* parasites are transmitted by the *Anopheles* leucosphyrus group of mosquitoes, containing a diverse array of species found throughout Southeast Asia (Sallum et al. 2005), including *An. latens, An. cracens, An. Introlatus,* and *An. hackeri* in which *P. knowlesi* has been detected in peninsular Malaysia, and other species that have been shown to be infected elsewhere (Vythilingam et al. 2018). *Anopheles* leucosphyrus group mosquitoes predominantly inhabit forested areas (Sinka et al. 2011; Moyes et al. 2016), so changes to forest areas and ongoing deforestation will affect human exposure. The potential vector species vary in relative abundance among different sampling sites in peninsular Malaysia (Vythilingam et al. 2008; Jiram et al. 2012; Vythilingam et al. 2014), but more surveys are required to determine the relative extent to which they transmit *P. knowlesi*, and whether they transmit different populations of the parasite corresponding to those described here.

Genome-wide scans revealed a few chromosomal regions of divergence between Cluster 3 sub-clusters A and B, the most prominent of which is a large region on Chromosome 12. This region showed lower nucleotide diversity in sub-cluster B compared to A, and was also the genomic region with the strongest evidence of recent direction selection, shown by the extended haplotype homozygosity analysis. It would seem that selection is operating on a locus in this chromosomal region of sub-cluster B, and might be driving this signature of local population structure. Local population genetic sub-structure has been seen in the endemic malaria parasite *P. falciparum* in Cambodia under strong antimalarial drug selection (Miotto et al. 2015), although antimalarial treatment is unlikely to be a major cause of selection on *P. knowlesi* as long as it remains primarily a zoonotic infection. It is notable that the Chromosome 12 region did not show a signature indicating recent selection in Malaysian Borneo (Assefa et al. 2015), indicating that this signature of selection is specific to sub-cluster B of Cluster 3.

More unexpected is the observation of another subpopulation of parasites, here termed sub-cluster C, which are highly related and virtually identical to each other throughout most of the genome. Although less common than infections belonging to sub-cluster A and B, clinical cases with this parasite type presented in different hospitals in three different states in peninsular Malaysia. Local population genetic sub-structure has been previously seen in the human malaria parasites *P. falciparum* (Anthony et al. 2005) and *P. vivax* (Auburn et al. 2018) in Malaysia, although that has been interpreted as indicating fragmented populations that are close to being eliminated. It is likely that zoonotic *P. knowlesi* populations are sub-structured for other reasons, as seen in Malaysian Borneo where the two divergent parasite genetic populations seen in human cases are associated with different reservoir hosts (Divis et al. 2015; Divis et al. 2017).

Cluster 3 had a skewed allele frequency spectrum with an overall negative Tajima's D, reflecting an excess of low frequency alleles and suggesting long term population expansion, as seen in most malaria parasite populations including the separate *P. knowlesi* Clusters 1 and 2 in Malaysian Borneo (Assefa et al. 2015; Divis et al. 2018). This is notable, as the overall trend remains apparent in Cluster 3 despite population sub-structure that can push Tajima's D values in the other direction. Scanning for genes that have unusually positive values and that might be under balancing selection showed different results from those seen in other parasite populations. For example, in the Cluster 1 population in Malaysian Borneo, the highest Tajima's D value was seen for the circumsporozoite surface protein (*csp*) gene, which was also within a window of extended haplotype homozygosity indicating recent selection, whereas in peninsular Malaysia *csp* did not have an elevated Tajima's D value, nor did it fall into an extended haplotype homozygosity window. As previously noted, these indices are less suited for analysis of Cluster 2 which shows more profound genome-wide mosaicism in diversity, likely affected by occasional introgression with sympatric Cluster 1 in Borneo (Diez Benavente et al. 2017; Divis et al. 2018).

Multi-genotype *P. knowlesi* infections have been shown using microsatellite analysis to be common in macaques but rarer in humans, perhaps due to the differences in transmission intensity in the two host species (Divis et al. 2015). Previous microsatellite-based analysis did not find any significant difference in the frequency of multi-genotype infections based on location of Cluster 1 and 2 infections in Malaysian Borneo and this was reflected by F_{WS} calculation of Clusters 1 and 2 carried out in a subsequent study (Assefa et al. 2015; Divis et al. 2015). Likewise, the F_{WS} re-calculations for Clusters 1 and 2 carried out here, which included all the previously analysed samples along with additional Cluster 2 isolates which had been sequenced

subsequent to the previous calculation, showed similar levels of within-host diversity, with no significant difference found between the Cluster estimates. The F_{WS} calculation for Cluster 3 also did not show any significant difference in mixedness compared to Clusters 1 and 2. While multi-genotype infections can be an indicator for the potential of recombination events occurring after parasites are taken up by mosquitoes, in the case of *P. knowlesi* the high proportion of predominantly single genotype infections in a population is most likely due to the low transmission intensity affecting this species, particularly among human hosts, and it could also be influenced by geographic isolation between human infections (Manske et al. 2012). F_{WS} scores for *P. falciparum* in West Africa have been shown to reflect differing transmission intensities, with higher transmission regions showing lower mean F_{WS} scores, but in general these parasites have much lower F_{WS} scores than any of the *P. knowlesi* clusters (Auburn et al. 2012; Mobegi et al. 2014; Duffy et al. 2015). In the case of *P. knowlesi*, which as far as we know remains exclusively a zoonotic disease, the low rate of infection mixedness is likely a function of intermittent, opportunistic transmission.

Experimental studies on *P. knowlesi* have been conducted only on a few strains, of the Cluster 3 type (Assefa et al. 2015; Divis et al. 2017), isolated many years ago and used to infect laboratory monkeys (Coatney et al. 1972). Diverse observations have suggested that parasite virulence increased as a result of serial passage of blood stage parasites from one monkey to another, or from one human to another as part of induced malaria fever therapy for neurosyphilis. One of these laboratory strains has been adapted twice to *in vitro* culture in human erythrocytes, as a result of independent experimental approaches involving culture with mixtures of macaque and human erythrocytes prior to growth in human erythrocytes alone (Lim et al. 2013; Moon et al. 2013). The short-term adaptability of this single parasite strain is further shown by

selection for culture in long-tailed macaque erythrocytes, enhanced growth being associated with the loss of an invasion pathway needed for invading human erythrocytes (Moon et al. 2016). These limited examples illustrate it is very likely that the highly diverse natural parasite populations are adapting to changing conditions in Malaysia.

Population genomic analysis of *P. knowlesi* so far has focused on parasites from Malaysia, where most reported cases of *P. knowlesi* malaria have been described. However, cases of *P. knowlesi* malaria in humans have now been reported from all Southeast Asian countries, and the actual numbers may be much higher in many areas, as there has been very limited molecular identification of parasite species in most parts of the region. It is not yet known if there are other local zoonotic sub-populations throughout the region, or whether all parasites belong to the major sub-populations seen in Malaysia. It will be particularly interesting to conduct population genetic analysis to determine the situation at sites where *P. knowlesi* has only very recently been realized to occur in humans (Herdiana et al. 2018; Imwong et al. 2019).

4. Transcriptomic profiles of *P. falciparum* clinical isolates expressing variable levels of MSPDBL2 – a possible marker of gametocytogenesis

4.1 Introduction

Transmission of malaria parasites from humans to mosquitoes is vital for continuation of the parasites' life cycle, and requires the development of male and female gametocytes from asexual precursors. This process of gametocytogenesis has been the focus of much research and represents an attractive target for drug development. However, although we now have some understanding of the processes involved, we still do not know the precise mechanisms. Gametocytogenesis occurs in two main stages, the first is commitment, whereby an asexual parasite commits irreversibly to developing into a gametocyte, a process which is indicated and initiated by a transcriptional 'switch' from an asexual programme of gene expression to a sexual programme of gene expression (Kafsack et al. 2014). The second stage is conversion, where the committed parasite physically develops from a trophozoite into a gametocyte by undergoing massive morphological and transcriptional changes. The gene identified to be at the centre of the switch to sexual commitment is a particular member of the apicomplexan *Apetala-2* (*ap2*) gene family, termed *ap2-g* (Kafsack et al. 2014). This gene is a transcription factor that has been recognised as a 'master regulator' of sexual commitment, which controls many of the downstream gene expression cascades needed for commitment to be established. Evidence for this comes from whole genome sequencing of gametocyte non-producing lines of *P. berghei* which revealed that *ap2-g* was the only gene to carry independently derived mutations, and likewise whole genome sequencing of gametocyte non-producing *P. falciparum* laboratory lines revealed again that the only gene containing loss of function mutations was *ap2-g* (Kafsack et al. 2014; Sinha et al. 2014).

The commitment stage of gametocytogenesis was originally described as occurring in the intraerythrocytic cycle previous to the one in which parasites underwent the morphological conversion (Bruce et al. 1990a). It was originally found that the progeny from a single schizont tended to all either be gametocytes or asexual parasites, suggesting that the fate of a parasite's progeny is fixed prior to the subsequent invasion cycle in which gametocyte conversion occurs (Bruce et al. 1990b). However, recent evidence has now suggested that (provided certain parameters are met) commitment and conversion can occur within the same invasion cycle (Bancells et al. 2019). Under this newly suggested model, if *ap2-g* transcription begins early enough in the ring stage parasite for sufficient AP2-G protein to be achieved prior to schizogony then parasites are able to convert within the same invasion cycle (Bancells et al. 2019). If the threshold level of AP2-G is not reached during the same invasion cycle, the parasite will continue along the asexual pathway, re-invade, and will then convert into a gametocyte in the subsequent invasion cycle (Bancells et al. 2019). In this way, AP2-G expression strictly controls the rate at which gametocytogenesis occurs in asexual parasites.

The upstream region of *ap2-g* contains H3K9me3 methylation marks associated with Heterochromatin protein-1 (HP1), which binds the H3K9me3 and recruits methyltransferases to facilitate heterochromatin formation and heritable silencing of the targeted locus (Flueck et al. 2009; Lopez-Rubio et al. 2009). Experimental conditional knock-out of the HP1 locus in *P. falciparum* resulted in a 25-fold increase in the number of parasites expressing the early gametocyte marker *Pfs16*, and a high proportion of gametocytes were seen in the cultures. Restoring HP1 expression in HP1-knockout parasites prior to schizogony halted the inflated conversion to gametocytes (Brancucci et al. 2014). These experimental data strongly indicate that expression of *ap2-g* is controlled by HP1-dependent gene silencing. A method proposed for the

control of AP2-G expression is through gametocyte development protein-1 (GDV-1), which was one of the first genes identified as being critical to gametocytogenesis (Eksi et al. 2012). The mechanism of GDV-1 mediated activation of genes is suggested to be through destabilisation of heterochromatin through interaction of GDV-1 with HP1. Co-immunoprecipitation assays have shown that GDV-1 and HP1 co-purify and it is thought that the two proteins form a complex, co-localising at the nuclear periphery (Filarsky et al. 2018). In a parasite line containing a conditional knock-out of *gdv-1*, induction of the gene significantly induced transcription of *ap2-g* along with eight other genes compared to *gdv-1* knockout parasites, all of which contain the upstream H3K9me3 marks of HP1-mediated silencing. One of these genes encode the merozoite surface protein MSPDBL2 (Filarsky et al. 2018).

MSPDBL2 is a merozoite surface protein and is one of two MSP3-like proteins that contains a duffy-binding like (DBL) domain (Singh et al. 2009). MSPDBL2 has been found to co-localise with MSP-1, in a classic pattern of expression consistent with that of a protein expressed on the merozoite cell surface (Singh et al. 2009; Hodder et al. 2012). Genome-wide analysis of single nucleotide polymorphism frequencies in *P. falciparum* populations has revealed the *mspdbl2* gene to be under strong balancing selection (Ochola et al. 2010; Amambua-Ngwa et al. 2012). Further analysis of *mspdbl2* transcription in *ex vivo* clinical isolates and long-term adapted *P. falciparum* laboratory lines revealed generally very low transcript levels. Consistent with this, MSPDBL2 protein expression is restricted to a small proportion of schizonts representing each isolate (Amambua-Ngwa et al. 2012). The number of schizonts expressing MSPDBL2 varied between isolates, but was expressed by less than 1% of schizonts in all culture-adapted lines except HB3, which expressed the protein in ~10% of schizonts (Amambua-Ngwa et al. 2012). A limited amount of data have suggested that MSPDBL2

could be involved in drug resistance (Van Tyne et al. 2011; Van Tyne et al. 2013), or in the development of protective immunity against repeated *P. falciparum* infection (Tetteh et al. 2013; Chiu et al. 2015). More recently, it has been suggested that MSPDBL2 could have a role in gametocytogenesis. Induction of *gdv-1* has been shown to result in upregulation of *mspdbl2* alongside other genes known to be involved in gametocytogenesis. An independent experiment in which *gdv-1* antisense RNA was disrupted in the F12 *P. falciparum* laboratory line, which is unable to produce gametocytes due to a nonsense mutation within *ap2-g* (Kafsack et al. 2014), also identified *mspdbl2* as being upregulated compared to wild-type F12 parasites in which *gdv-1* antisense RNA is active (Filarsky et al. 2018). Strand-specific RNA-seq has shown that *gdv-1* is likely regulated by its own antisense RNA and disruption of the antisense RNA results in a marked induction of GDV-1 expression (Broadbent et al. 2015; Filarsky et al. 2018). Variation in *mspdbl2* transcription has also been noted to be differential between *P. falciparum* laboratory lines and clinical isolates (Tarr et al. 2018). Utilising multiple biological replicates of four laboratory lines and six clinical isolates from Ghana, it was found that among ten highly expressed genes showing significant differences in gene expression between the groups of samples only two had increased expression in the clinical isolates, one of these was gamete antigen 27/25 (PF3D7_1371600) and the other was *mspdbl2* (PF3D7_1036300) (Tarr et al. 2018). While these lines of evidence suggest some association between the expression of *mspdbl2* and the onset of gametocytogenesis, it is not known whether the gene is marker of the process or whether it plays a more active role.

Gametocytogenesis is a key part of the *Plasmodium* life cycle, and recent efforts have increased our understanding of the genetics behind this process, and what environmental cues may trigger it. In this chapter, I am aiming to determine whether the

gene *mspdbl2*, proposed as either being involved in, or a marker for, gametocytogenesis is associated with the transcription of other genes in the gametocytogenesis pathway. I will undertake analysis of MSPDBL2 protein in order to profile its expression in a panel of clinical isolates collected from West Africa to determine the frequency of isolates with high *mspdbl2* expression. These isolates will undergo RNA-Seq and genes that are differentially expressed between MSPDBL2-high and MSPDBL2-low expressors will be identified to determine if they are involved in gametocytogenesis. The association of *mspdbl2* expression with gametocytogenesis genes will provide strong grounds for continuing investigations into this gene in this context.

4.2 Materials and Methods

Blood samples from clinical malaria cases collected from patients attending local health facilities in Ghana (Kintampo), Mali (Nioro), and Senegal (Pikine) between 2012 and 2013 were used in these studies. Patients aged between 2 and 14 years were eligible if they had uncomplicated clinical malaria, had not taken antimalarial drugs in the 72 hours preceding sample collection and tested positive for *P. falciparum* malaria by lateral flow rapid diagnostic test and slide microscopy. Up to 5ml of venous blood was collected in heparinised, anti-coagulation BD Vacutainer® tubes (BD Biosciences). Blood samples were leukocyte depleted by centrifugation, separation of plasma, and removal of leukocyte buffy coat layer. Erythrocytes were cryopreserved in glycerolyte and stored at -80˚C or in liquid nitrogen before shipment on dry ice to the London School of Hygiene and Tropical Medicine. Ethical approval for the collection and analysis of clinical samples was granted by the Ethics Committee of the Ministry of Health in Senegal, the Ethics Committee of the Ministry of Health in Mali, the Ethics Committee of the Ghana Health Service, the Noguchi Memorial Institute for Medical Research, University of Ghana, the Kintampo Health Research Centre, and the London

School of Hygiene and Tropical Medicine. Written informed consent was obtained from parents or legal guardians of participating children and additional assent was received from participating children.

At the London School of Hygiene and Tropical Medicine, the clinical isolates were thawed in batches of eight (Section 2.1.1) and maintained in culture *ex vivo* (Section 2.1.4). Microscope slides were made for each isolate upon thawing, and then again periodically to assess their progression into schizogony. Isolates containing schizonts on the second day after thawing were purified by magnetic separation and parasites were allowed to mature in the presence of E64 to prevent schizont rupture, matured parasites were centrifuged to leave the erythrocyte pellet (Section 2.1.6). Clinical isolates with sufficiently large erythrocyte pellets were used for both immunofluorescence assays as well as RNA-seq and are referred to as "matched" isolates.

Erythrocyte pellets containing matured schizonts were prepared for immunofluorescence assays by washing and resuspending parasites in 1% BSA and spotting into individual wells of 12-well slides (Hendley-Essex), dried slides were stored at -40˚C before use (Section 2.2). Antibody staining was carried out on thawed slides by incubation with αMSPDBL2 polyclonal mouse serum and subsequently with goat anti-mouse IgG Alexa Fluor® 555 secondary antibody. Vectashield® mounting fluid containing DAPI was used to visualise nuclei (Section 2.2). Mature schizonts (containing >8 nuclei) were counted using DAPI and scored for MSPDBL2 expression on a manual Leica microscope with a 100x objective, a sample size of approximately 1000 mature schizonts were counted per isolate.

Parasite material for RNA extraction was stored in either TRIzol® or RNA*later*™ (Thermo Fisher Scientific, MA, USA) and RNA was extracted from these samples by

phenol-chloroform extraction and cleaned up using the NucleoSpin® RNA XS extraction kit (Macherey-Nagel, Germany), following manufacturer's instruction (Section 2.4). Samples showing successful RNA extraction on the Bioanalyzer were reverse transcribed and amplified using the SmartSeq® v4 Ultra® Low Input RNA Kit for Sequencing (Takara Bio. Inc., Shiga Prefecture, Japan) (Section 2.5). Successfully amplified isolates were then prepared for sequencing on the Illumina MiSeq using the Nextera XT Library Prep kit using 75bp paired-end reads (Illumina, California, USA) (Section 2.8.2). Samples were pooled equimolar at 4nM with no more than 12 samples per pool and were run on an Illumina MiSeq using the 150-cycle MiSeq reagent kit v3. Due to the amplification process used for the isolates, it was not possible to use strand-specific RNA sequencing.

Whole transcriptome short read sequence data were assembled by alignment to the *P. falciparum* 3D7 reference genome using HISAT2 (Kim et al. 2015) (Section 2.9.2), and read counts per gene were calculated. Differential gene expression analysis was carried out in DESeq2, where MSPDBL2 protein expression in schizonts in each isolate (based on IFA counting) was used to categorise or rank the isolates for differential gene expression analysis (Section 2.10). For the differential gene expression analysis using discrete groups with a 1% cut-off, 12 samples fell into the "low expressor" group, and 5 samples fell into the "high expressor" group. For the analysis using discrete groups with a 3% cut-off, 14 samples fell into the "low expressor" group and 3 samples fell into the "high expressor" group. For the analysis using discrete groups comparing >3% expression with <1% expression, 3 samples had expression >3% and 12 had expression <1%. For the continuous analysis, samples were binned at 0% expression and then in increments of 0.5%. As there were few samples with high expression of MSPDBL2, there were five bins that only contained a single replicate.

4.3 Results

92 clinical isolates from three West African countries were thawed for *ex vivo* culture. Six from Senegal, 36 from Ghana, and 50 from Mali. Out of these, 48 contained sufficient schizont material to yield both IFA slides and RNA ("matched" samples). 43 of the 48 samples that contained parasite material were harvested for schizonts at the first round before re-invasion (approximately 48 hours after thawing); three of the 48 samples were cultured for 10 days prior to collection and two of the 48 samples had been grown in culture for 73 days. 24 isolates did not contain enough material for IFA slides and instead 100% of the material was stored in RNA*later*™ for potential separate RNA-seq.

4.3.1 MSPDBL2 protein expression varies among clinical isolates from West Africa

Multiwell slides containing schizonts prepared from the 48 "matched" isolates (those from which IFA slides were made as well as RNA-seq transcriptome data), along with an additional four isolates whose slides had been previously prepared, were stained with polyclonal sera against MSPDBL2 and sealed with mounting medium containing DAPI; a nuclear stain used to identify and count merozoites. A schizont was considered to be positive for MSPDBL2 if it contained eight or more merozoites and was clearly fluorescent above the background. As MSPDBL2 is usually expressed in only a small proportion of schizonts, in order to get accurate counts, approximately 1000 schizonts with 8 or more nuclei were counted per parasite isolate. Of the 48 matched samples, ten did not contain sufficient schizonts to count 1000, this was due to either high levels of contamination with younger parasites (mid-trophozoite to early schizont stages), or a parasite density either too low or too high to allow for accurate merozoite counting and scoring of MSPDBL2. Of the 38 samples with sufficient schizonts for counting, 24 had an MSPDBL2 expression of less than 1% (four of these had no MSPDBL2-positive

schizonts scored). Fourteen of the 38 isolates expressed MSPDBL2 in more than 1% of their schizonts and five isolates had MSPDBL2 expression in over 3% of schizonts. The range of expression was 0% to 6.6% in these samples (Table 4.1). MSPDBL2 protein was localised to the merozoites, as expected for a surface protein. In addition to MSPDBL2-positive 8+ nuclei schizonts, any "ambiguous" positives were also scored separately. These included early schizonts and late trophozoites showing fluorescence, the presence of MSPDBL2 in these younger parasites was not unexpected, as it is expressed from approximately 32 hours post invasion (h.p.i) (Otto et al. 2010).

These new data from the 38 isolates were combined with MSPDBL2-expression count data from previously studied clinical isolates for which IFA expression data are available (prepared and analysed by Dr. Sarah Tarr, and Ms. Lindsay Stewart at LSHTM), taking the total number of isolates assessed for MSPDBL2 protein expression by IFA up to 77, collected from five West African countries (Ghana n=43, Senegal n=10, The Gambia n=4, Guinea n=1, Mali n=19). Including these previous results, MSPDBL2 expression ranged from 0% to 73% (Figure 4.1). An exceptional isolate with most schizonts expressing MSPDBL2 (73%) was INV236 from Senegal. INV236 was cultured *ex vivo* and RNA was extracted by Ms. Lindsay Stewart (LSHTM). RNA was cleaned and amplified using the SmartSeq® v4 Ultra® Low Input RNA Kit for Sequencing and sequenced by Dr. Sarah Tarr (LSHTM) using the methods outlined in Section 4.2.

4.3.2 Obtaining samples with matched IFA and RNA-seq data

RNA was extracted from the schizont preparations of 38 samples which were analysed by IFA. RNA was run on the Bioanalyzer to obtain the RNA Integrity Number (RIN) for that isolate and to qualitatively assess the presence or absence of the 18s and 28s

Table 4.1. MSPDBL2 protein expression and RNA-seq details of clinical isolates

Sample ID	MSPDBL2 IFA expression	Harvested RNA?	RIN	Library concentration (nM)	Total reads ($x10^6$)	Alignment rate (%)
INV236*	73.0%	Yes	N/A	N/A	0.8	82.0
INV374*	0.4%	Yes	9.8	13.1	1.1	84.9
INV375	0.3%	No	-	-	-	-
INV378	0.2%	No	-	-	-	-
INV382	0.4%	No	-	-	-	-
INV384	0.2%	No	-	-	-	-
INV394	0.2%	No	-	-	-	-
INV395*	6.5%	Yes	7.6	7.80	3.0	84.9
INV396*	2.0%	Yes	8.2	8.60	3.8	88.2
INV273/2	0.6%	No	-	-	-	-
INV284/2	0.9%	No	-	-	-	-
INV401*	1.1%	Yes	7.9	8.10	2.0	84.6
INV406*	0.6%	Yes	8.8	12.8	1.6	87.5
INV408	0.7%	No	-	-	-	-
INV409	1.4%	No	-	-	-	-
INV410*	1.0%	Yes	8.1	10.8	2.6	83.9
INV419*	0.0%	Yes	7.2	15.3	2.4	80.3
INV420	4.5%	No	-	-	-	-
INV425*	3.7%	Yes	8.2	20.1	1.4	84.8
INV427	4.4%	No	-	-	-	-
INV428	1.5%	No	-	-	-	-
INV429	1.5%	No	-	-	-	-
INV430	1.0%	No	-	-	-	-
INV431	0.3%	No	-	-	-	-
INV432	1.1%	No	-	-	-	-
INV438	3.3%	No	-	-	-	-
INV440*	0.7%	Yes	7.4	11.7	2.1	84.6
INV441	1.1%	No	-	-	-	-
INV443	0.0%	No	-	-	-	-
INV444	0.5%	No	-	-	-	-
INV445*	0.2%	Yes	6.4	14.0	1.8	83.3
INV446*	0.3%	Yes	6.5	11.6	2.0	85.4
INV447	0.6%	No	-	-	-	-
INV448*	0.0%	Yes	9.0	15.5	2.9	77.6
INV449*	0.8%	Yes	8.7	12.5	3.2	83.6
INV450*	0.1%	Yes	6.7	18.2	1.7	82.8
INV455*	0.2%	Yes	6.0	13.7	1.4	86.9
INV458	0.4%	No	-	-	-	-
INV459*	0.0%	Yes	8.2	10.1	2.3	82.5

39 *ex vivo* clinical isolates were counted for their number of MSPDBL2-positive schizonts. Of these, 17 were of sufficient quality to undergo RNA-seq from low input RNA material. Only samples with a RIN of six or higher were considered high enough quality for downstream analysis and these have matched MSPDBL2 expression data analysed by IFA. RNA-seq yielded between $7.7x10^5$ to $3.8x10^6$ reads per sample. MSPDBL2 expression was assessed by IFA in approximately 1000 schizonts with eight or more nuclei per isolate. 17 samples gave high-quality RNA-seq data which informed differential gene expression analysis and are indicated with *.

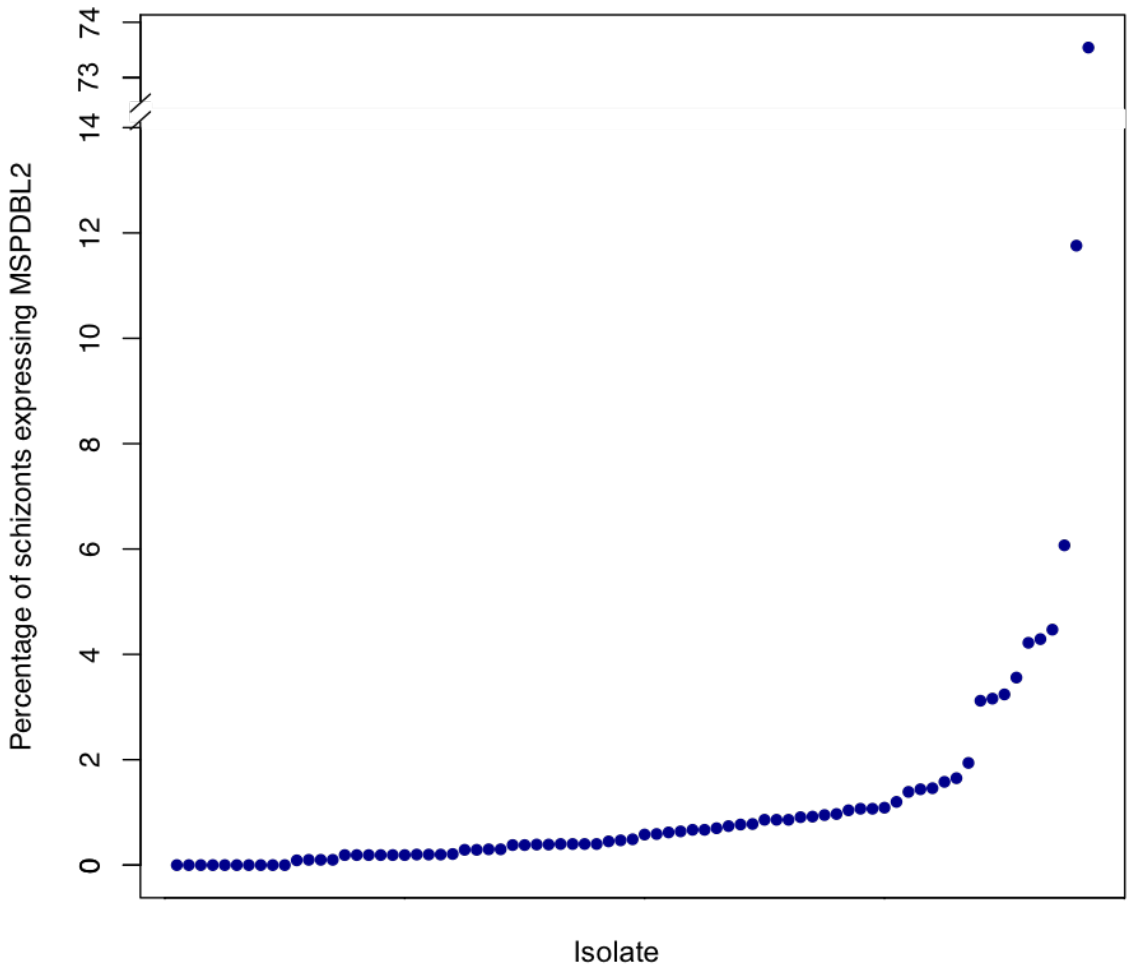

Figure 4.1. Expression of MSPDBL2 in *P. falciparum* schizonts in 77 clinical isolates, collected from five countries in West Africa (Ghana n=43, Senegal n=10, The Gambia n=4, Guinea n=1, Mali n=19). Approximately 1000 mature schizonts containing eight or more nuclei were scored in immunofluorescence assays, using anti-MSPDBL2 polyclonal mouse serum. Ten of the samples had 0% expression, and 21 had expression over 1%. Included are 39 clinical samples previously analysed by IFA for MSPDBL2 expression (prepared and analysed by Dr. Sarah Tarr, and Ms. Lindsay Stewart, LSHTM), and the new samples with matched IFA and RNA-seq data.

ribosomal RNA (rRNA) peaks (Figure 4.2A). 30 of the 38 samples showed distinct 18s and 28s rRNA and it was decided that these would all be carried forward for reverse transcription and amplification, and subsequently sequencing, despite the fact that several had low RIN values (9 isolates had RIN values of < 6). After reverse transcription and amplification five samples did not show the characteristic 'hump' of cDNA that was expected by Bioanalyzer analysis, which results from many amplified cDNA fragments of varying lengths, indicating that either the reverse transcription or the amplification step had failed (Figure 4.2B). Library preparation for sequencing using the Nextera XT kit (Illumina) was carried out for the remaining 25 samples and was successful for all but one sample. Successfully prepared libraries showed an expected distribution of read lengths (Figure 4.2C). The 24 isolate libraries were pooled equimolar into two batches of 12 and successfully sequenced. Eight of these transcriptomes were generated from RNA with a RIN of less than 6 and it was decided at this point to remove these samples from the dataset in order to minimise the introduction of errors or bias. INV236 (73% MSPDBL2-positive schizonts), which was prepared previously, was added to the transcriptome dataset. This left a final sample size of 17 isolates with matched IFA and RNA-seq data of sufficiently high quality. Of these, 5 had MSPDBL2 expression by IFA of greater than 1%.

4.3.3 Quality of Transcriptomic data from low input samples

Due to the amplification carried out on the isolates prior to sequencing, it was not possible to carry out strand-specific RNAseq. It is therefore not possible to assess the presence and abundance of non-coding antisense RNA molecules, which could affect the measurement of gene expression. However, during the reverse transcription stage of the amplification process, mRNA molecules are enriched for by the use of oligod(T) primers. Polyadenylation of antisense RNA is not commonplace, and in *Plasmodium* it

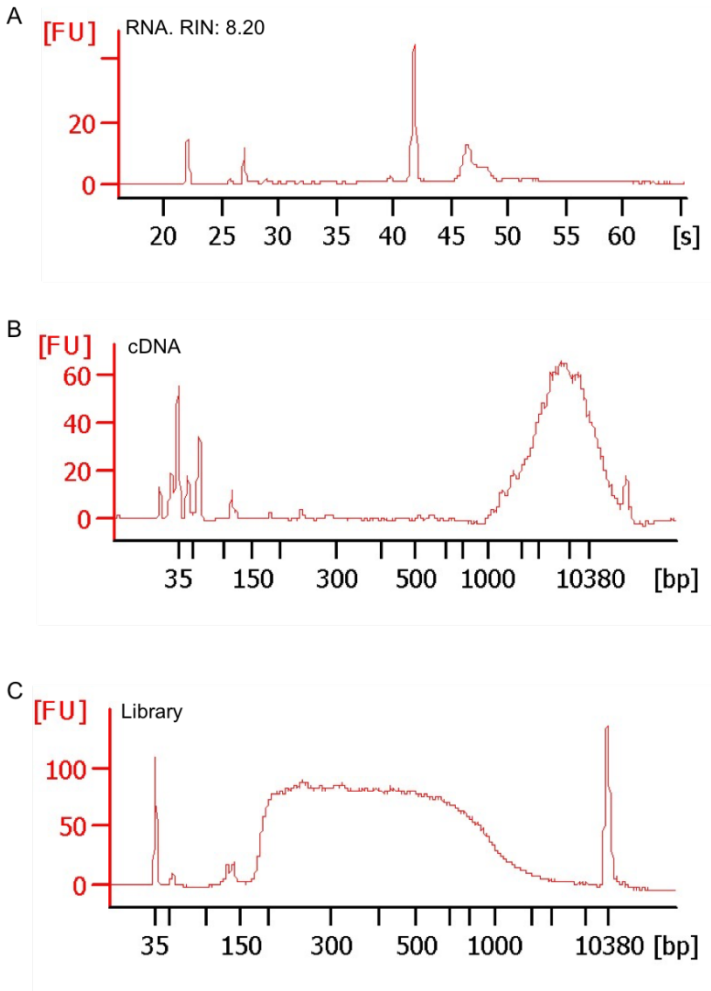

Figure 4.2. Illustrative representation of Bioanalyzer traces at each successful step of the RNA extraction, reverse transcription, amplification and library preparation process. A. Quality of RNA was assessed using the RNA 6000 Pico kit (Agilent Technologies, CA, USA). The two peaks between 40s and 50s represent the 18s and 28s ribosomal RNA peaks and presence of these peaks, alongside the RNA Integrity Number (RIN) indicate the quality of the RNA. **B.** After reverse transcription and amplification of RNA using the SmartSeq® v4 chemistry, the expected "hump" of cDNA can be seen, resulting from a range of cDNA fragment lengths. This indicates successful generation of amplified cDNA material. The very small peaks around the 35bp lower marker likely originate from degraded RNA. **C.** Libraries were prepared using the Nextera XT library preparation kit (Illumina, CA, USA). The distribution of fragment sizes allows for accurate calculation of library concentration and normalisation.

is thought to affect a very small number of transcripts (<160) (Siegel et al., 2014). For these reasons, it is not expected that the presence of antisense RNA will have a significant impact on the expression levels for genes in this chapter.

Whole transcriptomes were obtained for 17 novel clinical isolates (16 prepared in the current study plus INV236). Short reads from the 17 samples were assembled against the *P. falciparum* v3 3D7 reference genome. Trimming the paired-end 75bp raw short reads was not necessary as the quality of the reads did not diminish significantly towards the ends. A visual assessment of multi-exon genes was undertaken in Artemis to ensure that reads did not extend into introns; this indicated by proxy that cDNA rather than genomic DNA had been captured (Figure 4.3). Percentages of reads aligning to the *P. falciparum* 3D7 reference genome ranged from 74% to 88%, indicating that there were not high levels of contamination within the preparations, with total number of reads per sample ranging from 862,000 – 3.7 million (Table 4.1). Transcript levels of *mspdbl2* were assessed using the "Fragments Per Kilobase of transcript per Million mapped reads" (FPKM) metric, which measures the number of short DNA fragments mapping to a particular gene, normalised for the size of the sequencing library and for the length of the gene. Isolate INV236 (with 73% MSPDBL2-positive schizonts by IFA) had an extremely high FPKM value for *mspdbl2* of 22,000 relative to that of the other isolates whose FPKM values ranged from 0 – 340.

In order to test for bias and any batch effects that might be acting on the dataset, transcriptome data from the 17 samples were loaded into the R package DESeq2, using a masked GFF annotation file that had the *var*, *rifin*, and *stevor* gene families removed (Tarr et al. 2018). In addition to removal of these subtelomeric genes, portions of other protein-coding genes that show high allelic diversity were masked to ensure mapping

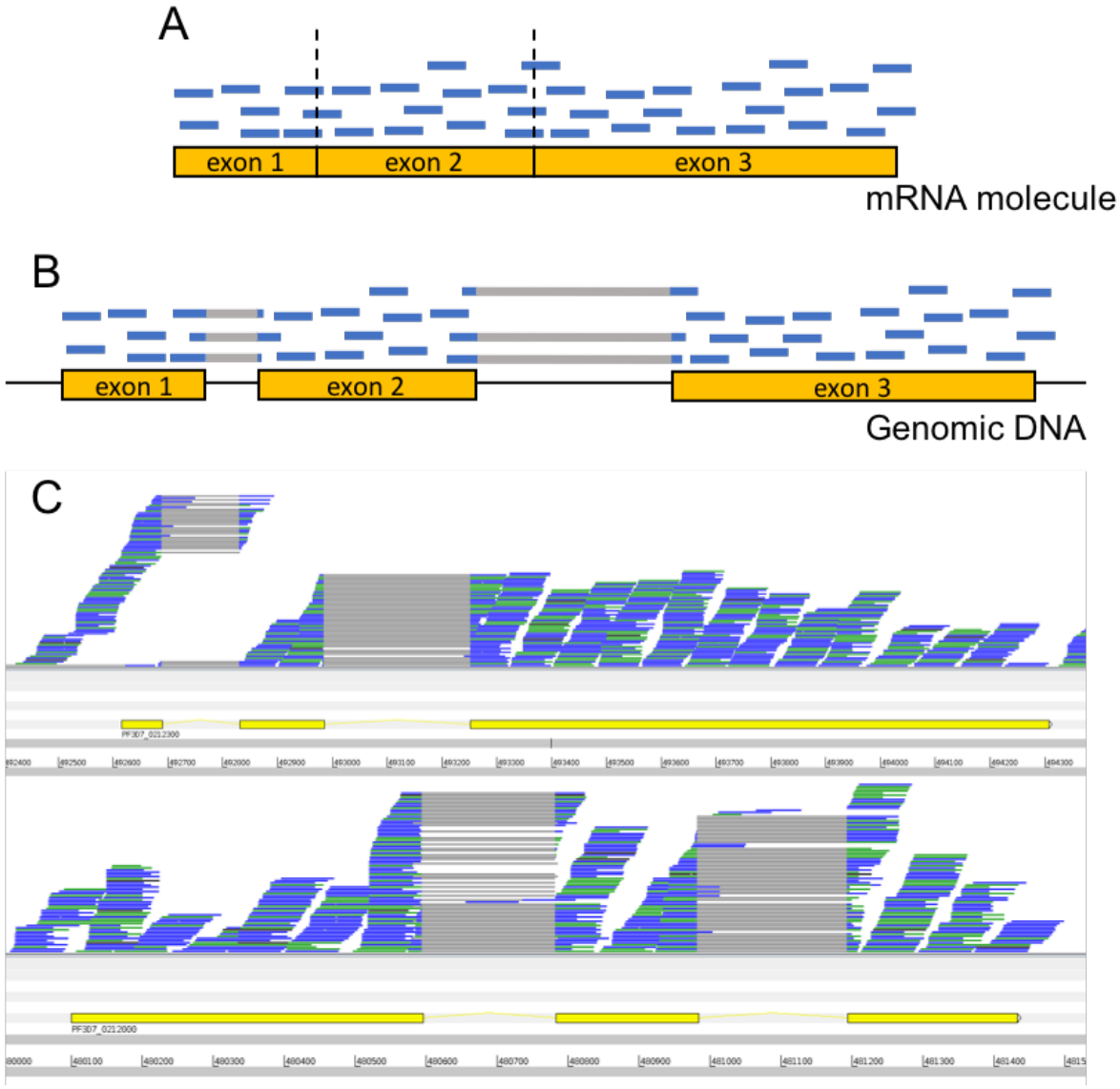

Figure 4.3. Representation of the expected distribution of Illumina short reads throughout multi-exon genes for RNA-seq data. **A.** When short reads (blue bars) are created from mRNA molecules, they are distributed throughout the spliced gene, with some reads covering exon-exon boundaries. **B.** When these short reads are mapped back to the genome which contains the exon and intron sequences, the short reads are 'split' across exon-exon boundaries, resulting in an absence of sequence data covering the introns (shown by the grey bars). **C.** Representation of short read distribution (blue, green, and black bars) viewed in Artemis from INV374 for two multi-exon genes mapped to the *P. falciparum* 3D7 reference genome, clearly showing the absence of read data in introns, indicating that the sequencing library represented cDNA and not genomic DNA.

occurred only in conserved regions of those genes to give a more accurate calculation of expression and minimise allelic bias among samples (Tarr et al. 2018).

Principal Component Analysis (PCA) was carried out in DESeq2 to identify clustering of samples, which would identify potential technical bias. Principal components (PC) 1-4 were examined (Figure 4.4). The data was normalised using the DESeq2 rlog normalisation prior to PCA. When based on genes which contained a minimum of 10 reads in any one sample, the majority of the samples broadly clustered together along principal component 1 (PC1), which covered 94% of the variance. Two samples (INV236 and INV374) were separated along PC2, which accounted for 2% of the total variation. The only technical difference shared between these two samples and unique from the other samples was that schizont material from them had been stored in TRIzol® prior to extraction, rather than RNA*later*™. When an alternative approach is taken whereby PCA is carried out on the top 500 genes with the highest variance across isolates, INV236 and INV374 are separated from the main cluster along PC1, accounting for 45% of the total variation, indicating that the RNA preservation method may be influencing read counts, but that there was little bias introduced by the library preparation and sequencing processes. There was not any obvious clustering along PC1-PC4 related to MSPDBL2 protein expression (Figure 4.4).

In addition to PCA, other methods can be used to investigate the quality and bias of a dataset. A distance matrix between samples was constructed from regularised log (rlog) normalised data. This allowed us to look at sample similarity and the heatmap shows that the samples disperse into three groups (Figure 4.5). As seen in the PCA, the TRIzol®-preserved samples grouped separately from the others, while the remaining

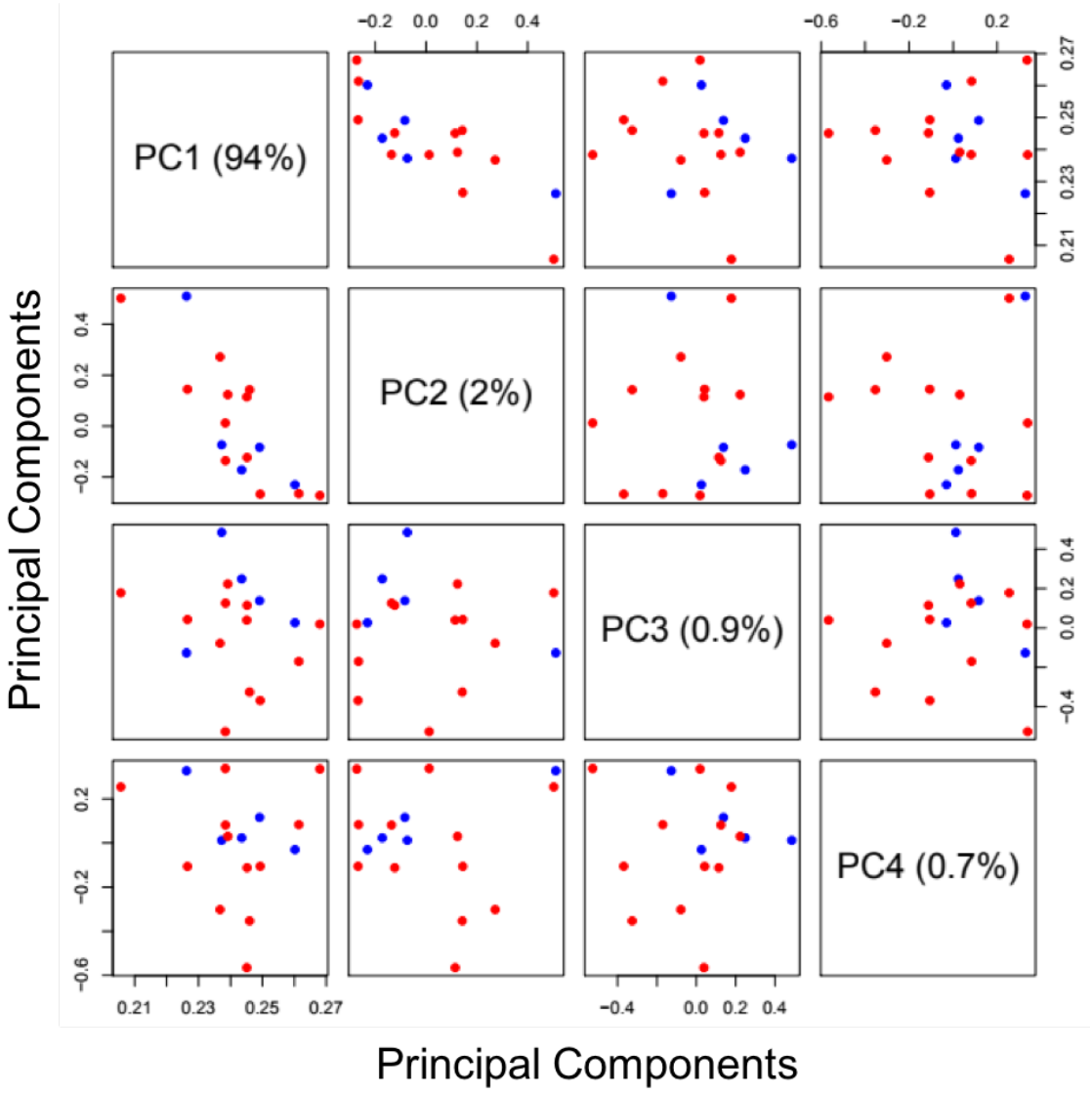

Figure 4.4. Principal component analysis of gene expression from 17 clinical isolates
RNA-seq data for the 17 clinical isolates with matched IFA MSPDBL2 expression data were assessed by Principal Component Analysis to investigate any clustering that may be the result of technical bias, or from biological differences between isolates. Principal Components 1-4 have been calculated based on all genes which have ten of more mapped reads (to reduce background noise). The isolates are coloured red were represented by less than 1% of schizonts expressing MSPDBL2 by IFA, and the blue by greater than 1% of schizonts expressing MSPDBL2 by IFA. PC1 accounts for almost all (98%) of the variance, and does not appear to be driven by significant clustering. INV236 is the only sample placed slightly away from the majority. Along PC2 (2% of the variance), two samples are placed slightly away from the remainder. These are INV236 and INV374; two samples whose RNA was preserved in TRIzol® rather than RNA*later*™. Along PC3 and PC4 no obvious clustering of samples can be seen, and none of the principal components are associated with the proportions of schizonts showing MSPDBL2 protein expression. This indicated that a comparison of transcriptomes to test for associations with MSPDBL2 will not contain any systematic bias.

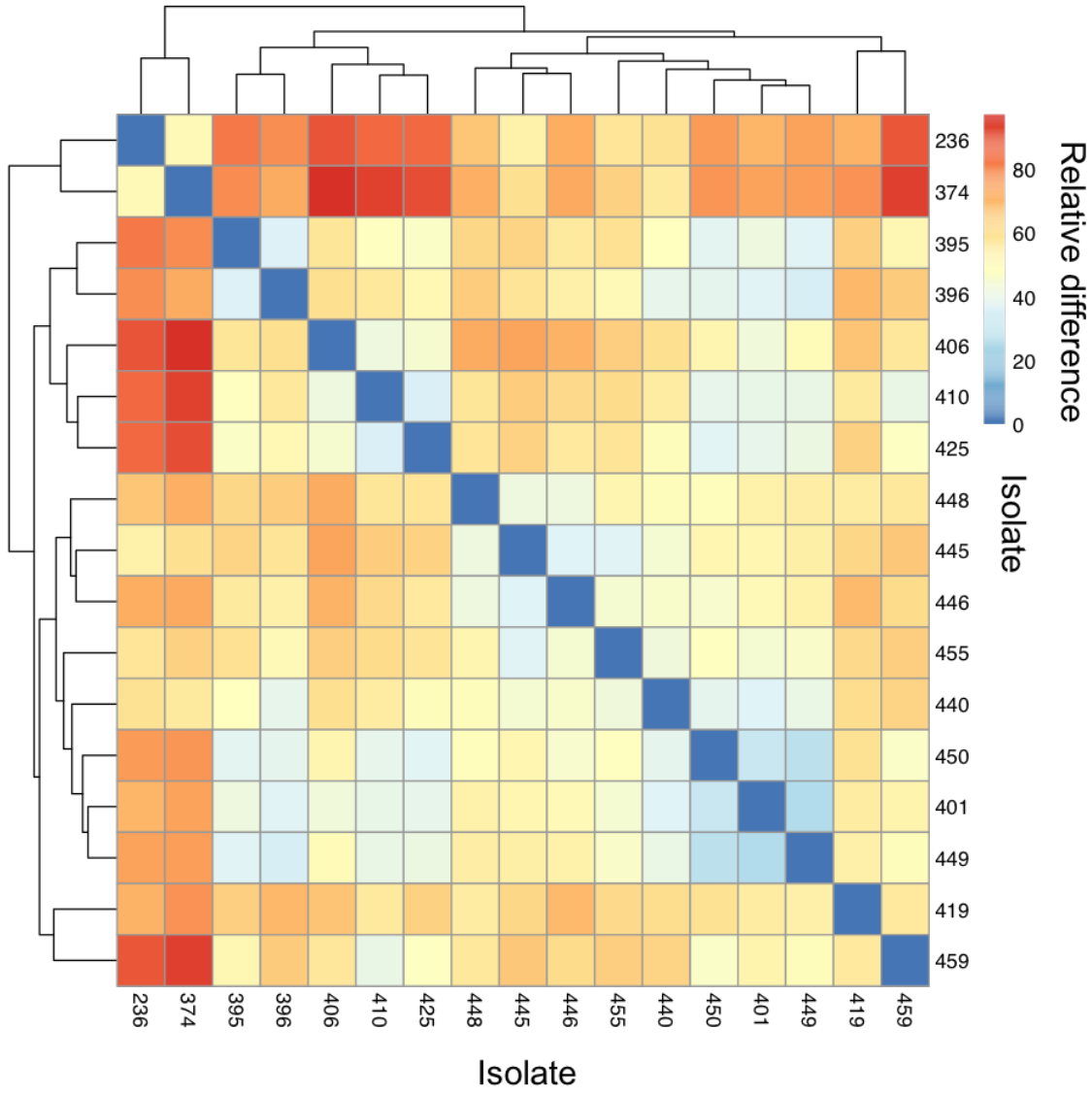

Figure 4.5. Heat map showing sample similarity of transcriptomes between the clinical isolates. Sample similarity was assessed by constructing a distance matrix of rlog normalised read counts of each sample. The samples are shown as pairwise comparisons. The lighter blue indicates increasing sample difference. The samples cluster into three groups; one of these is formed by INV236 and INV374, which also cluster together in principal component analysis (Figure 4.4). The remaining samples fall broadly into two groups.

samples separate broadly into two groups. To investigate the technical variability further, the number of genes per sample with zero mapped reads was plotted (Figure 4.6A) alongside the average number of reads per gene, per sample (Figure 4.6B). The distribution of untransformed read data shown in Figure 4.6C shows sample variation, however this was normalised by the rlog transformation, as expected (Figure 4.6D). These methods of interrogating samples for technical bias show that, despite INV236 and INV374 clustering separately by PCA, there was no evidence of poorer amplification or sequencing of these samples.

4.3.4 Obtaining first-round schizonts from *ex vivo* culture

Significant effort was put in at the *ex vivo* culturing stage of this project to ensure that the maximum possible number of parasites had reached schizogony prior to harvesting. However, clinical isolates have lifecycle lengths varying from each other and from those of laboratory lines, and as these were mostly *ex vivo* samples they were not tightly synchronised. MACS® purification filtered out the majority of younger parasites, and the four-hour incubation with E64 allowed parasites to mature without the most mature schizonts rupturing. To assess the success of methods used to promote maturity in this project, the FPKM values for each isolate were compared to an RNA-seq time course generated from transcriptome data derived from tightly synchronised *P. falciparum* 3D7 parasites, which were harvested at seven timepoints in eight hour increments of hours post invasion, h.p.i. (0 h.p.i, 8 h.p.i, 16 h.p.i, 24 h.p.i, 32 h.p.i, 40 h.p.i, and 48 h.p.i) and whole transcriptome sequenced (Otto et al. 2010). Spearman's rank correlation was calculated for each isolate against each timepoint in the time course. This approach has been used previously to assess the overall correlations of life stages of clinical isolates and laboratory lines (Tarr et al. 2018). Results are shown in Figure 4.7 and the Spearman's rank correlations shown in Table 4.2. Of the 17 isolates, 13 were found to

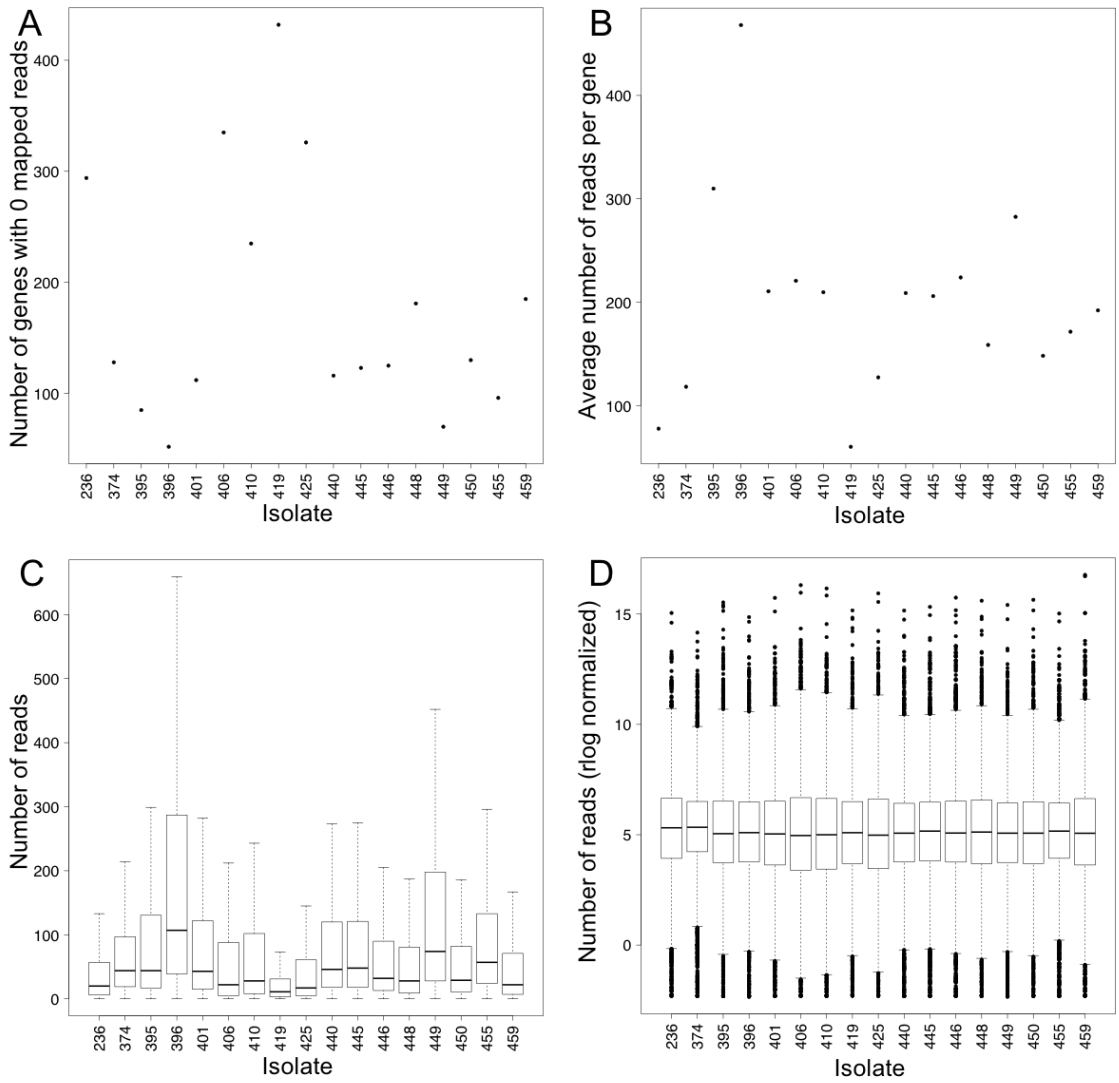

Figure 4.6. Graphs assessing technical differences between clinical isolate sequencing libraries A. The number of genes with zero mapped reads per isolate. **B.** The average number of reads per gene. **C.** Distribution of the numbers of reads per gene based on raw read data. **D.** Variation in number of reads representing genes is normalised by rlog transformation.

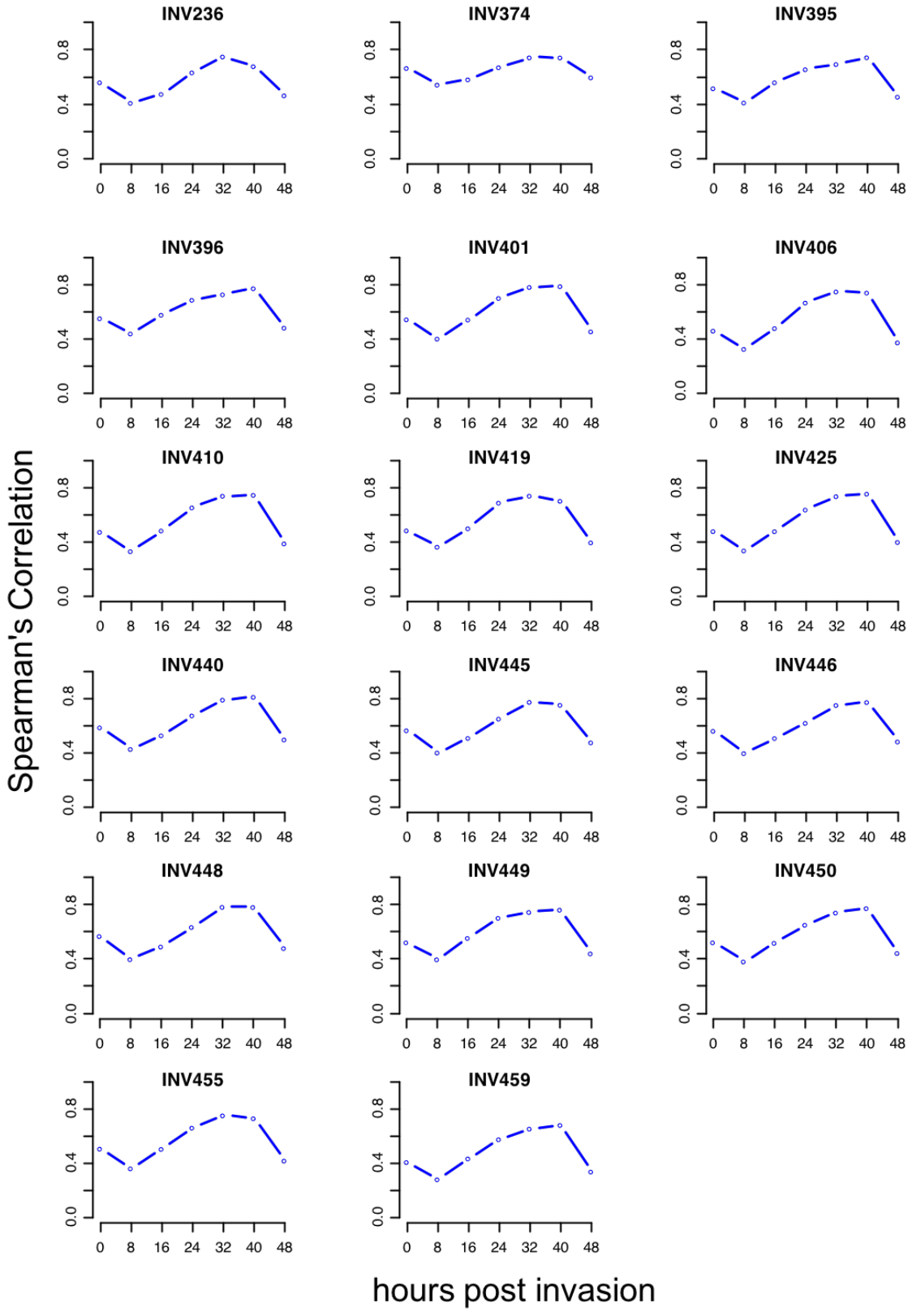

Figure 4.7. Estimated development of the 17 clinical isolates. FPKM values for the 17 clinical isolates, correlated with previous data collected across seven timepoints in the *P. falciparum* life cycle (Otto et al. 2010). *Ex vivo* preparation of these samples involved MACS® purification to obtain later stage parasites, followed by 4.5 hours with the addition of E64 to block schizont egress with the aim of capturing late stage schizonts. The majority of samples were captured at 32 h.p.i. or 40 h.p.i.

Table 4.2. Spearman's rank correlation between clinical isolates and RNA-seq timecourse data

	Spearman's rank correlation with reference data at:						
Isolate	**0 h.p.i.**	**8 h.p.i.**	**16 h.p.i.**	**24 h.p.i.**	**32 h.p.i.**	**40 h.p.i.**	**48 h.p.i.**
INV236	0.57	0.41	0.48	0.64	0.75	0.68	0.47
INV374	0.67	0.55	0.59	0.68	0.75	0.75	0.60
INV395	0.52	0.42	0.57	0.66	0.70	0.75	0.46
INV396	0.56	0.44	0.58	0.69	0.73	0.78	0.49
INV401	0.55	0.41	0.55	0.70	0.79	0.79	0.46
INV406	0.46	0.33	0.48	0.67	0.75	0.74	0.38
INV410	0.48	0.34	0.49	0.66	0.74	0.75	0.39
INV419	0.49	0.37	0.51	0.69	0.75	0.71	0.40
INV425	0.48	0.34	0.48	0.64	0.74	0.76	0.40
INV440	0.59	0.43	0.53	0.68	0.80	0.82	0.50
INV445	0.57	0.41	0.51	0.66	0.78	0.76	0.48
INV446	0.57	0.40	0.51	0.63	0.76	0.78	0.49
INV448	0.57	0.40	0.50	0.64	0.79	0.79	0.48
INV449	0.53	0.40	0.56	0.71	0.75	0.77	0.45
INV450	0.53	0.39	0.52	0.66	0.75	0.78	0.45
INV455	0.51	0.37	0.51	0.67	0.76	0.74	0.42
INV459	0.41	0.29	0.44	0.58	0.66	0.69	0.34

FPKM values for genes in the 17 isolates were correlated using Spearman's Rank with previous RNA-seq data harvested from tightly synchronised *P. falciparum* 3D7 parasites at seven timepoints in eight hour increments across the parasite life cycle (0 h.p.i. – 48 h.p.i.) (Otto et al. 2010).

correlate most strongly with 40 h.p.i 3D7 parasites and four isolates had the strongest correlation with 32 h.p.i. This analysis indicates that the majority of parasites in most samples were either approaching or in schizogony. Based on time course data available at PlasmoDB.org from *P. falciparum* 3D7, *mspdbl2* gene expression is lowest during the trophozoite stage (at around 16 h.p.i) and can be seen to increase during early schizogony (around 32 h.p.i) before peak expression is reached at 40 h.p.i. Expression remains at 48 h.p.i and lower expression can still be seen in merozoites and re-invaded rings at around 8 h.p.i. (Otto et al. 2010). As such, no samples were excluded based on their staging, as expression of MSPDBL2 can be seen throughout schizogony.

4.3.5 Analysing differential gene expression between discrete MSPDBL2 phenotype groups

Differential gene expression analysis was carried out in DESeq2 in R. In each comparison, genes for which expression differed between groups at a significance level of $P < 0.01$ were considered. Samples were split into discrete groups based on MSPDBL2 protein expression. Firstly, a 1% threshold for MSPDBL2 expression was used: with "low-expressers" having <1% MSPDBL2-positive schizonts and "high-expressers" having >1% MSPDBL2-positive schizonts. This identified 42 genes with increased expression and 18 with decreased expression in the MSPDBL2 "high-expressers" (Figure 4.8A, Appendix 6A). Of the 42 genes with increased expression, four are known or suspected to have a role in gametocytogenesis: exported protein factor 1 *(epf1,* PF3D7_0114000) (Silvestrini et al. 2010), nucleoporin *nup116/nsp116* (PF3D7_1473700) (Brancucci et al. 2014; Josling et al. 2019), a zinc finger protein (PF3D7_0315600) (Josling et al. 2019), and an unknown conserved *Plasmodium* protein PF3D7_1467600 (Josling et al. 2019). Secondly, the threshold for low-expressers and high-expressers was set to 3%. Three isolates had MSPDBL2 expression

over 3%. This analysis identified 38 genes with increased expression and 4 with decreased expression (Figure 4.8B, Appendix 6B). Five of the genes with increased expression are known or suspected to have a role in gametocytogenesis: *nup116/nsp116* (PF3D7_1473700), lysine-specific histone demethylase *lsd2* (PF3D7_0801900) (Josling et al. 2019), a zinc finger protein (PF3D7_1134600) (Josling et al. 2019), early gametocyte enriched phosphoprotein *egxp* (PF3D7_1466200) (Nixon et al. 2018), and a gene encoding a conserved *Plasmodium* protein of unknown function (PF3D7_1467600) (Filarsky et al. 2018; Josling et al. 2019).

A third comparison based on discrete grouping was made using a "buffer" zone within which intermediate level expression isolates were excluded. The low-expressers were those with <1% MSPDBL2 expression and the high-expressers were the isolates in the >3% category. This comparison was carried out in order to reduce noise between phenotype groups. This approach identified 35 genes with increased expression and 9 with decreased expression (Figure 4.8C, Appendix 6C). Four of the genes with increased expression are known or suspected to have a role in gametocytogenesis: *nup116/nsp116* (PF3D7_1473700), *lsd2* (PF3D7_0801900), a zinc finger protein (PF3D7_1134600) and *egxp* (PF3D7_1466200).

The gene lists from the three discrete group comparisons were cross-referenced to identify genes that were consistently higher in the high MSPDBL2 groups. Approximately half of the genes identified in one comparison were also present in one or both of the others (35 genes out of 71), and 26 genes were present in two out of the three comparisons, with nine present in all three comparisons (Table 4.3). Of the 26 genes present in more than one analysis, three genes (*nup116/nsp116, lsd2*, the zincfinger protein PF3D7_1134600, the protein of unknown function PF3D7_1467600, and *egxp*) have known or suspected involvement in gametocytogenesis (Appendix 1).

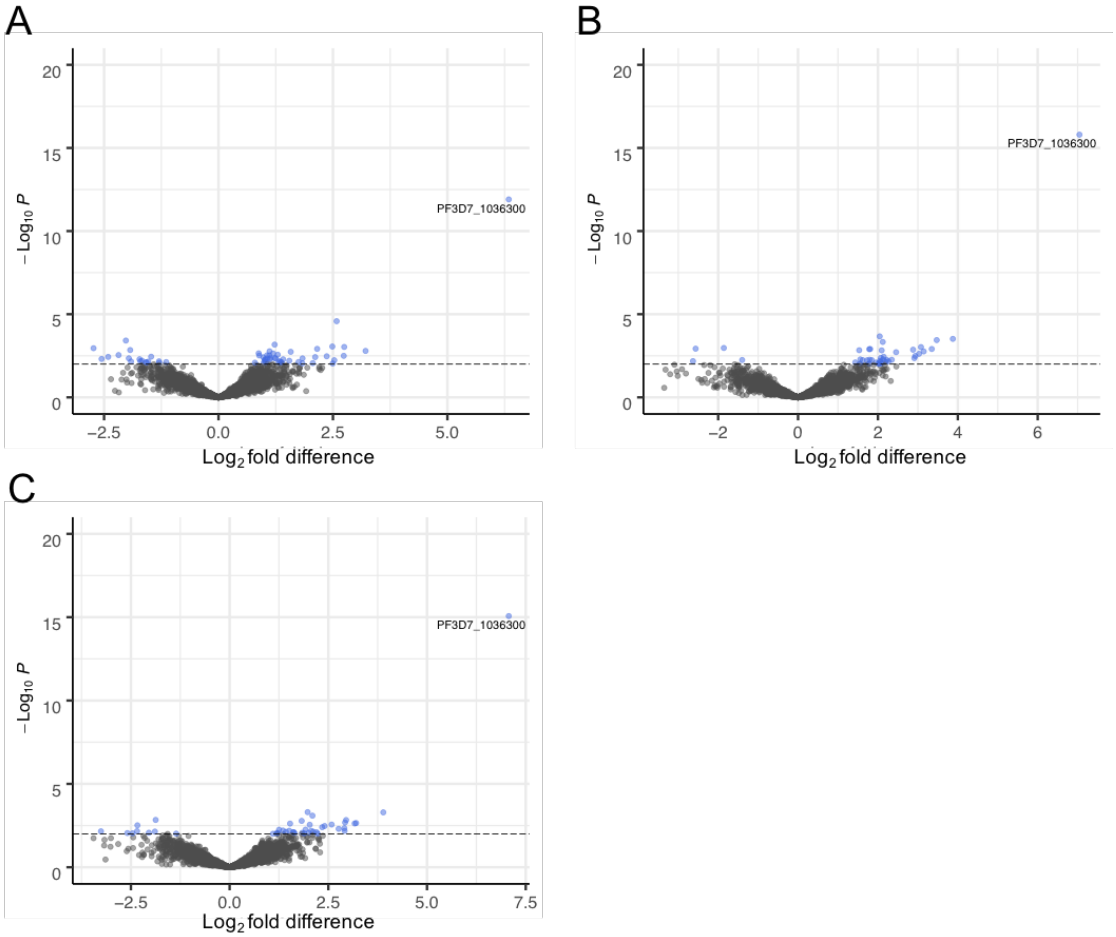

Figure 4.8. Volcano plots showing the log2 fold difference in gene expression between samples placed in discrete groups of MSPDBL2 expression. Samples were placed into high or low MSPDBL2 expressing groups based on IFA expression. The Y-axis shows the P-values, with the threshold of 0.01 indicated by the dotted line; dots shaded blue had significant P-values below 0.01. MSPDBL2 is labelled on each plot. **A.** Samples were split into groups of those with <1% of schizonts expressing MSPDBL2 and >1% expressing MSPDBL2. This identified 42 genes with higher expression in the >1% group. **B.** Samples were split into groups with <3% of schizonts expressing MSPDBL2 and >3% expression MSPDBL2. This identified 38 genes with higher expression in the high group. **C.** Samples were grouped into those with <1% expression of MSPDBL2 in schizonts (n=12) and >3% expression (n=3), this left a "buffer" zone (1-3%) in which no samples were included, designed to reduce noise and identified only the most significant hits. This approach identified 35 genes with higher expression in the >3% group. Of all the genes identified in the three analyses (72), 35 of these were present in two comparisons, and nine were present in all three.

Table 4.3. Analysis of gene expression correlated with expression of MSPDBL2 was assessed in three comparisons of discrete groupings

Gene ID	1% cut-off	3% cut-off	<1% and >3%	Description
PF3D7_0110800	1.03	NS	1.25	transcription initiation factor TFIIB
PF3D7_0112100	2.11	2.35	2.58	conserved *Plasmodium* protein, UF
PF3D7_0202000*	NS	2.02	2.12	knob-associated histidine-rich protein
PF3D7_0214300*	NS	3.03	2.91	conserved *Plasmodium* protein, UF
PF3D7_0309200	NS	2.14	2.07	asparagine synthetase, putative
PF3D7_0416100	1.35	1.8	1.82	glutamyl-tRNA(Gln) aminotransferase subunit A
PF3D7_0416300	NS	1.53	1.53	DNA helicase MCM9, putative
PF3D7_0416900	NS	2.04	1.97	conserved *Plasmodium* protein, UF
PF3D7_0516500	NS	1.85	1.92	major facilitator superfamily domain-containing protein, putative
PF3D7_0516900	1.11	NS	1.22	60S ribosomal protein L2
PF3D7_0722400	1	NS	1.2	Obg-like ATPase 1, putative
PF3D7_0723400	NS	2.22	2.2	conserved *Plasmodium* protein, UF
PF3D7_0801900	NS	2.45	2.41	lysine-specific histone demethylase
PF3D7_0803400	NS	1.56	1.42	DNA repair/recombination RAD54
PF3D7_0829400*	NS	2.13	2	prolyl 4-hydroxylase subunit alpha
PF3D7_0831800	2.15	NS	2.34	histidine-rich protein II
PF3D7_1023100	NS	1.79	1.64	dynein heavy chain, putative
PF3D7_1027300*	2.75	2.93	2.76	Peroxiredoxin
PF3D7_1028900	NS	1.43	1.51	inner membrane complex protein
PF3D7_1036300*	6.34	7.05	7.07	MSPDBL2
PF3D7_1134600	NS	3.08	2.92	zinc finger protein
PF3D7_1232900	1.31	NS	1.6	nucleotidyltransferase
PF3D7_1319600*	NS	2.91	2.91	ACDC domain-containing protein
PF3D7_1327200	NS	1.62	1.63	ribonuclease P protein subunit RPR2
PF3D7_1361200*	NS	2.09	2.03	conserved *Plasmodium* protein, UF
PF3D7_1362700*	2.49	3.35	3.2	conserved *Plasmodium* protein, UF
PF3D7_1408000	1.55	2.12	2.1	plasmepsin II
PF3D7_1408600	0.98	NS	1.1	40S ribosomal protein S8e, putative
PF3D7_1413800	1.19	NS	1.35	diphthamide biosynthesis protein 1, putative
PF3D7_1431400*	NS	2.05	1.89	surface-related antigen SRA
PF3D7_1461800*	2.53	3.15	3.17	conserved *Plasmodium* protein, UF
PF3D7_1466200*	NS	2.26	2.23	early gametocyte enriched phosphoprotein EGXP
PF3D7_1467600*	2.74	3.47	NS	conserved *Plasmodium* protein, UF
PF3D7_1473700*	3.21	3.88	3.89	nucleoporin NUP116/NSP116
PF3D7_1474000*	2.36	2.88	2.95	conserved *Plasmodium* protein, UF

Differential gene expression analysis carried out using a 1% MSPDBL2 expression cut-off, a 3% MSPDBL2 expression cut-off, and a third using a "buffer" zone, comparing samples with <1% MSPDBL2 expression to samples with >3% MSPDBL2 expression. Genes listed are those that were identified in two or three of the comparisons. Values shown in the table indicate the log2 fold difference between the groups. NS indicates the gene was not considered to have differential expression in that comparison (P>0.01). Genes highlighted in bold have known or suspected roles in gametocytogenesis. Genes with * were also present in the gene expression analysis associated with isolates continuously ranked by their MSPDBL2 protein expression with a significance of <0.001. UF: protein has unknown function.

A small number of genes in each discrete phenotype group analysis were identified as having lower expression in the group with higher MSPDBL2 protein expression (Appendix 7A-C). None of these genes from any of the three analyses are known to be involved in gametocytogenesis, based on investigation using PlasmoDB and cross-referencing with gene lists suspected to be involved in gametocytogenesis (Silvestrini et al. 2010; Filarsky et al. 2018; Josling et al. 2019) (Appendix 1)

4.3.6 Analysing differential gene expression through a continuum of MSPDBL2 protein expression

An additional approach was taken whereby instead of splitting isolates into discrete groups of "low expressers" and "high expressers", MSPDBL2 protein expression was treated as a continuous variable, and differential gene expression was carried out on samples based on their MSPDBL2 IFA expression. Samples were divided into multiple bins based on their expression of MSPDBL2 (Section 4.2), using an interval of 0.5% between bins. However, due to the very few samples expression MSPDBL2 at a high level, these bins only contain one replicate, which is a limitation of this analysis. Differential gene expression is then calculated between each of the bins. Due to the stochastic nature of gene expression, no gene will have "zero" change when samples are placed into multiple bins and as such every gene will have an "increased" or "decreased" skew of expression, therefore a higher P-value stringency helps to reduce background noise. At a P-value of < 0.01, 194 genes with increased expression and 304 genes with decreased expression were identified (Appendix 8 and 9). Of the 194 genes that increased alongside MSPDBL2, 17 are known to be, or to potentially be involved in early gametocytogenesis, based on previous studies. For example, exported protein family-1 *(epf1*, PF3D7_0114000) (Silvestrini et al. 2010), *Pfs16* (PF3D7_0406200) (Bruce et al. 1994), *gexp-13, -15, -02*, and *-04* (PF3D7_0831300, PF3D7_1031600,

PF3D7_1036300, PF3D7_1102500, and PF3D7_1372100**,** respectively) (Silvestrini et al. 2010), Two members of the AP2 family of transcription factors, *ap2* (PF3D7_1342900) and *ap2-g2* (PF3D7_1408200) (Sinha et al. 2014; Yuda et al. 2015), and a 6-cysteine protein (*P47*, PF3D7_1346800) (van Schaijk et al. 2006), among others.

The continuous variable analysis identified many more genes that were differentially expressed than the discrete analysis. In order to identify the most significant hits and reduce the effect of 'noise', the P-value cut-off was tightened to 0.005, which left 142 genes with higher expression correlated with MSPDBL2 expression (Appendix 8); the stringency was then tightened even further to 0.001, which reduced the list to a subset of 52 genes (Table 4.4). This list contained 13 significant hits (including *mspdbl2*) with known or suspected roles in gametocytogenesis (25% of the gene list), including a 6-cysteine protein, encoding the female-specific marker *P47* (PF3D7_1346800), the gene encoding the early gametocyte marker *Pfs16* (Pf3D7_0406200), and the gene encoding *epf1* (Pf3D7_0114000). The 52 genes with increased expression with P-values <0.001 had their rlog normalised expression plotted as a heatmap (Figure 4.9) to look at how the genes clustered based on their expression. The genes formed two main clusters (shown by the distance tree along the y-axis of Figure 4.9). Nine of the 12 genes with gametocytogenesis involvement were found in one cluster, with MSPDBL2 and the remaining two genes in the second cluster. *mspdbl2* was the most highly significant gene, showing the greatest change in expression when correlated to its own protein expression.

For comparison, the 130 genes with lower expression ($P < 0.001$) compared to increasing MSPDBL2 protein expression were investigated (Appendix 9). Using

Table 4.4. List of genes with increased expression across isolates ranked by their MSPDBL2 protein expression

Gene ID	Increase between ranked samples*	P-value	Gene description
PF3D7_1036300	8.24	5.35E-25	duffy binding-like merozoite surface protein 2
PF3D7_1476600	5.95	6.67E-08	*Plasmodium* exported protein UF
PF3D7_1474000	5.29	7.59E-08	conserved *Plasmodium* protein UF
PF3D7_1102500	5.38	1.32E-07	*Plasmodium* exported protein (PHISTb) UF
PF3D7_1461800	5.94	2.96E-06	conserved *Plasmodium* protein UF
PF3D7_1445700	3.99	1.38E-05	conserved *Plasmodium* protein UF
PF3D7_0814200	3.82	1.74E-05	DNA/RNA-binding protein Alba 1
PF3D7_1466200	4.02	2.04E-05	early gametocyte enriched phosphoprotein EGXP
PF3D7_0114000	4.23	2.07E-05	exported protein family 1
PF3D7_1372100	4.86	2.07E-05	*Plasmodium* exported protein (PHISTb) UF
PF3D7_0215000	3.61	2.20E-05	acyl-CoA synthetase
PF3D7_1362700	4.97	2.56E-05	conserved *Plasmodium* protein UF
PF3D7_1473700	6.46	2.83E-05	nucleoporin NUP116/NSP116, put
PF3D7_0829400	3.67	2.84E-05	prolyl 4-hydroxylase subunit alpha, put
PF3D7_1027300	4.27	5.18E-05	peroxiredoxin
PF3D7_0515000	3.87	6.17E-05	pre-mRNA-splicing factor CWC2, put
PF3D7_1346800	5.19	6.23E-05	6-cysteine protein
PF3D7_1132600	3.58	6.36E-05	pre-mRNA-splicing factor 38A, put
PF3D7_1477700	4.51	9.14E-05	*Plasmodium* exported protein (PHISTa) UF
PF3D7_1431400	3.13	9.94E-05	surface-related antigen SRA
PF3D7_1140200	3.67	1.06E-04	conserved *Plasmodium* protein UF
PF3D7_1467600	4.81	1.16E-04	conserved *Plasmodium* protein UF
PF3D7_1130200	2.83	1.23E-04	60S ribosomal protein P0
PF3D7_0202000	3.64	1.26E-04	knob-associated histidine-rich protein
PF3D7_1429100	4.15	1.43E-04	apicoplast ribosomal protein L15 precursor, put
PF3D7_0315600	4.54	1.66E-04	zinc finger protein, put
PF3D7_0611600	3.54	1.81E-04	basal complex transmembrane protein 1
PF3D7_0927300	3.23	1.9E-04	fumarate hydratase
PF3D7_0815500	3.15	1.92E-04	conserved *Plasmodium* protein UF
PF3D7_1236200	3.48	2.31E-04	conserved *Plasmodium* protein UF
PF3D7_0403600	3.04	2.32E-04	conserved *Plasmodium* protein UF
PF3D7_1317000	3.67	2.65E-04	U4/U6.U5 tri-snRNP-associated protein 2 put
PF3D7_1233200	3.84	2.91E-04	conserved *Plasmodium* protein UF
PF3D7_1437900	3.06	3.0E-04	HSP40, subfamily A
PF3D7_0419600	2.89	3.53E-04	ran-specific GTPase-activating protein 1, put
PF3D7_0912900	2.68	4.88E-04	26S proteasome regulatory subunit RPN8, put
PF3D7_1432800	2.92	4.94E-04	HP12 protein homolog, put
PF3D7_0208000	2.99	5.1E-04	serine repeat antigen 1
PF3D7_1361200	3.69	5.27E-04	conserved *Plasmodium* protein UF
PF3D7_1136300	3.02	5.41E-04	tudor staphylococcal nuclease
PF3D7_1312100	3.51	5.72E-04	GYF domain-containing protein, put
PF3D7_0318900	3.25	5.93E-04	conserved protein UF
PF3D7_0527500	2.68	6.15E-04	Hsc70-interacting protein
PF3D7_1428300	2.88	6.25E-04	proliferation-associated protein 2g4, put
PF3D7_1002400	2.95	6.41E-04	transformer-2 protein homolog beta, put
PF3D7_0420100	3.32	7.59E-04	serine/threonine protein kinase RIO2
PF3D7_0406200	2.98	8.6E-04	sexual stage-specific protein precursor *pfs16*
PF3D7_1134000	3.27	8.89E-04	heat shock protein 70
PF3D7_0106900	3.68	8.98E-04	cytidylyltransferase, put
PF3D7_0408500	3.27	9.15E-04	NYN domain-containing protein, put
PF3D7_1145400	2.61	9.31E-04	dynamin-like protein
PF3D7_0214300	4.66	9.94E-04	conserved *Plasmodium* protein UF

Genes identified with expression that increased in samples alongside MSPDBL2, treated as a continuous variable. Using the high stringent p-value cut-off of <0.001, 52 genes were identified to be highly significant and unlikely to be a result of stochastic noise. Of these, 12 alongside *mspdbl2* (25%) are known or are suspected to be

involved in gametocytogenesis (gene ID highlighted in bold), compared to only 10 (7%) of genes among the 130 showing lower expression with increasing MSPDBL2 protein expression. *mspdbl2* is the most highly significant, and most variable gene identified, as expected. 13 of these highly significant genes (Table 4.5) are also present in two or more of the differential expression analyses utilising discrete groups of MSPDBL2 expression (Table 4.3). UF: protein has unknown function. Put: putative. * The log2 of the difference in transcription of the gene between the ranked sampled.

PlasmoDB, along with cross-referencing of gene lists with suspected involvement in gametocytogenesis (Silvestrini et al. 2010; Filarsky et al. 2018; Josling et al. 2019), out these 130 genes, only 10 genes (7%) have known or suspected involvement in gametocytogenesis. When cross-referencing the continuous analysis with the three analyses using discrete sample groups of MSPDBL2 expression, 13 of the highly significant (P <0.001) genes with increased expression in the MSPDBL2 continuous analysis were also present in two or all of the discrete group comparisons (Table 4.5), this included *mspdbl2*, early gametocyte enriched phosphoprotein *egxp*, and nucleoporin *nup116/nsp116*. At the lower stringency P-value threshold of <0.01, 18 of the 35 genes identified in more than one discrete analysis were also present in the continuous gene list (Table 4.3, genes indicated with asterixis).

In all four comparisons (the three discrete comparisons with different cut-off of MSPDBL2 expression alongside the continuous one), *mspdbl2* (PF3D7_1036300) was identified as the gene showing the highest difference in gene expression. This was an informative “positive control” showing that MSPDBL2 gene expression was indeed associated with the protein expression seen by IFA.

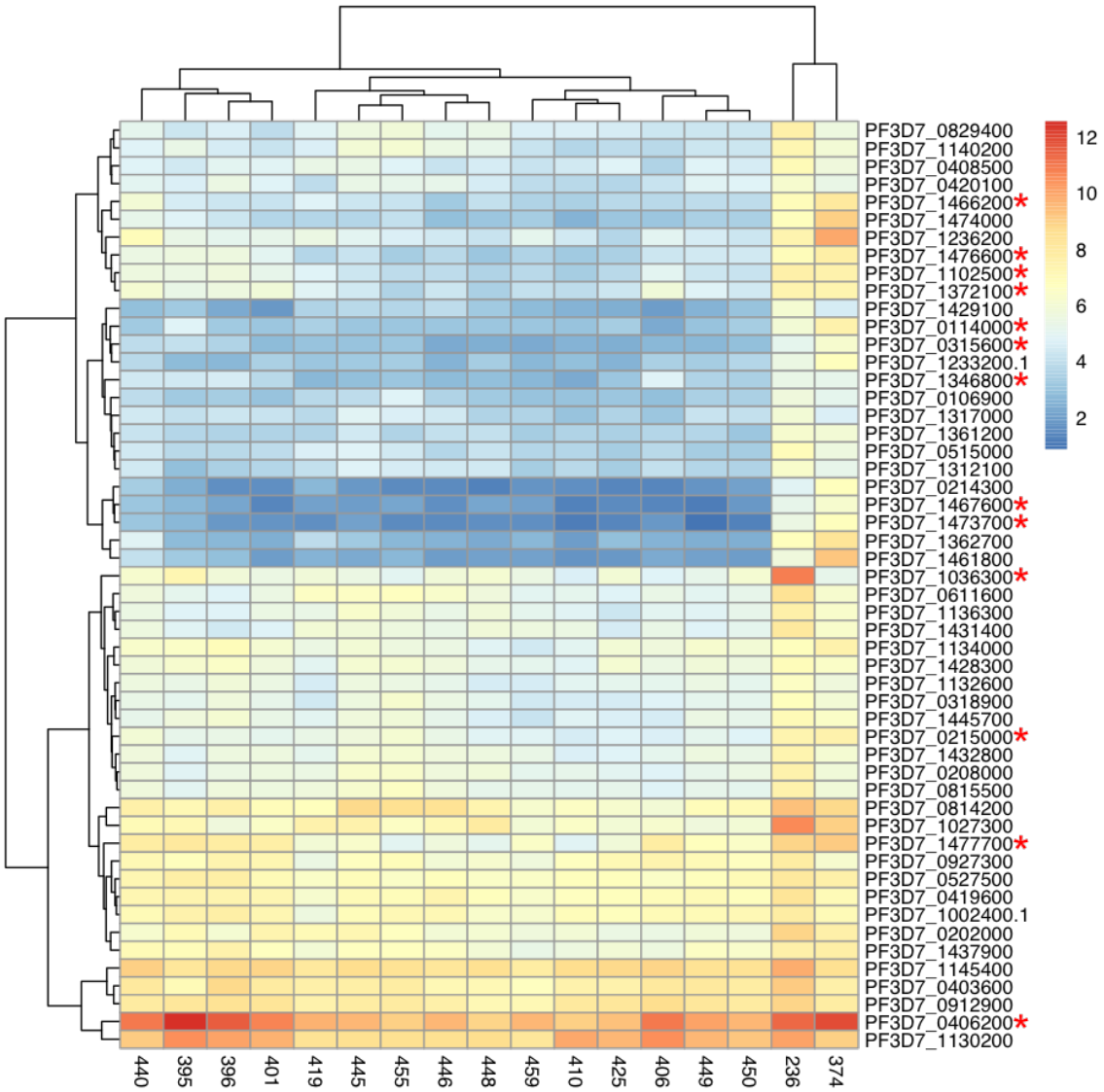

Figure 4.9. Clustering of genes showing higher expression in isolates with higher MSPDBL2 expression by IFA. Isolates were ranked based on their MSPDBL2 protein expression and highly significant (P-value < 0.001) genes with increased expression correlating with increasing MSPDBL2 protein expression were identified. The normalised read counts for these genes were extracted and plotted as a heatmap to investigate clustering of genes based on expression. The genes fall into two broad clusters. Twelve of the 52 genes (alongside *mspdbl2*) have known or suspected involvement in gametocytogenesis and these are indicated by red asterixis. Nine of the 12 genes fall into one cluster of genes. MSPDBL2 and the three remaining genes fall into the second cluster.

Table 4.5. Genes with increased expression across isolates ranked by MSPDBL2 protein expression that also had increased expression in one or more of the discrete phenotype group analyses

Gene ID	Increase in Expression	P-value	Description
PF3D7_1036300	8.24	5.35E-25	duffy binding-like merozoite surface protein 2
PF3D7_1474000	5.29	7.59E-08	conserved *Plasmodium* protein, UF
PF3D7_1461800	5.94	2.96E-06	conserved *Plasmodium* protein, UF
PF3D7_1466200	4.02	2.04E-05	early gametocyte enriched phosphoprotein EGXP
PF3D7_1362700	4.97	2.56E-05	conserved *Plasmodium* protein, UF
PF3D7_1473700	6.46	2.83E-05	nucleoporin NUP116/NSP116, putative
PF3D7_0829400	3.67	2.84E-05	prolyl 4-hydroxylase subunit alpha, putative
PF3D7_1027300	4.27	5.18E-05	peroxiredoxin
PF3D7_1431400	3.13	9.94E-05	surface-related antigen SRA
PF3D7_1467600	4.81	0.000116	conserved *Plasmodium* protein, UF
PF3D7_0202000	3.64	0.000126	knob-associated histidine-rich protein
PF3D7_1361200	3.69	0.000527	conserved *Plasmodium* protein, UF
PF3D7_0214300	4.66	0.000994	conserved *Plasmodium* protein, UF

13 genes identified as having increased expression associated with increase in MSPDBL2 protein expression in ranked samples were also identified in the previously described analyses using discrete phenotype groups. Gene IDs highlighted in bold are known or suspected to have a role in gametocytogenesis. UF: protein has unknown function.

To understand the extent to which the DGE analysis was being driven by isolate INV236 with extreme MSPDBL2 expression, the analysis was repeated after removal of INV236 from the dataset. The results from this analysis revealed that three genes from the original analysis that included INV236 were also identified when INV236 is removed at a p-value of < 0.001. One of these was the sexual stage precursor Pfs16 (PF3D7_0406200), a ribosomal gene (PF3D7_1130200), and an unknown protein PF3D7_1130200, which was the second most highly differentially expressed gene in the original analysis that included INV236. At a relaxed p-value of < 0.01, nine genes were shared between the two analyses. Interestingly, although MSPDBL2 was shown to be differentially expressed in the analysis excluding INV236 (log2FoldChange of 2.2), this was at a less significant p-value of 0.013.

4.3.7 Genes differentially expressed in correlation to *mspdbl2* FPKM values

To extend the analysis of samples ranked by MSPDBL2 protein expression and to consolidate the results from this analysis, the FPKM values for *mspdbl2* were extracted for each sample using DESeq2. The IFA MSPDBL2 positivity and FPKM values showed a weak, non-significant correlation (Spearman's rho = 0.21, P-value = 0.43). The correlation between IFA and FPKM expression is more clear at higher levels, particularly at expression levels of >1% by IFA (Figure 4.10), but is noticeably weak at very low IFA and FPKM values (Figure 4.10B), probably due to random noise. Therefore, it was decided to use the FPKM values for *mspdbl2* to identify genes with higher expression in samples with higher transcript levels. At a P-value of <0.001, this approach identified 42 genes with higher expression and of these, 17 (40%) were also identified when using ranked MSPDBL2 protein level (Table 4.6) and 16 of the 42 genes (38%) were also present in two or more of the gene lists produced using the discrete MSPDBL2 expression phenotype groups (Table 4.3). Of these 42 genes, 12

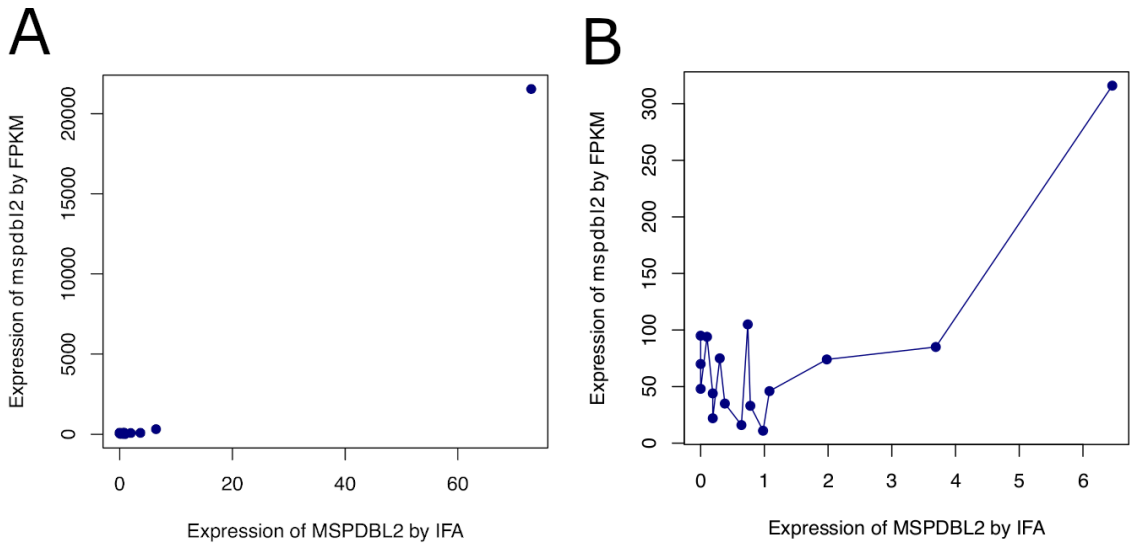

Figure 4.10. Plot showing the correlation between MSPDBL2 protein expression measured by IFA and *mspdbl2* transcript expression measured by FPKM. A. Isolate INV236, which had unprecedentedly high expression of MSPDBL2 by IFA, also had high expression by FPKM. **B.** A correlation can also be seen at IFA expression levels > 1%, however at lower expression there is no obvious correlation, possibly due to the effect of stochastic noise.

(28%) have known or suspected involvement in gametocytogenesis (Table 4.6), which were identified as described previously (Section 2.10). Seven of the 12 genes were also identified in the MSPDBL2 ranked protein expression analysis (Section 4.3.6) and six were identified in the analyses using discrete MSPDBL2 phenotype groups (Section 4.3.5). There were three genes with known or suspected involvement in gametocytogenesis which did not appear on any of the gene lists produced by previous methods, PF3D7_1148700, encoding the exported protein GEXP12 (Silvestrini et al. 2010), PF3D7_1408200, encoding an AP2 transcription factor, AP2-G2 (Yuda et al. 2015), and PF3D7_1472200, encoding histone deacetylase-1 (Filarsky et al. 2018). Finally, the list of genes with increased expression in isolates with higher *mspdbl2* FPKM values was intersected with the list of genes showing increased expression in

Table 4.6. Genes with increased expression correlating to *mspdbl2* transcript levels measured by FPKM

Gene ID	Increase between ranked samples[+]	P-value	Product Description
PF3D7_0114000*	4.53	2.63E-09	exported protein family 1
PF3D7_0207800	2.92	5.33E-05	serine repeat antigen 3
PF3D7_0214300*,**	5.45	1.11E-06	conserved *Plasmodium* protein UF
PF3D7_0309200**	3.81	3.31E-04	serine/threonine protein kinase, put
PF3D7_0315600*	3.74	1.81E-04	zinc finger protein, put
PF3D7_0402200	4.36	9.99E-04	surface-associated interspersed protein 4.1 pseudo
PF3D7_0501400	3.69	1.49E-05	interspersed repeat antigen
PF3D7_0519500	3.18	7.43E-05	CCR4 domain-containing protein 1, put
PF3D7_0723400**	3.85	4.46E-04	conserved *Plasmodium* protein, UF
PF3D7_0724100	3.50	3.94E-04	conserved *Plasmodium* protein, UF
PF3D7_0801900**	4.29	2.01E-05	lysine-specific histone demethylase, put
PF3D7_0829400*,**	3.54	5.28E-04	prolyl 4-hydroxylase subunit alpha, put
PF3D7_0930300	3.82	9.74E-04	merozoite surface protein 1
PF3D7_1014300	3.09	7.43E-04	SPRY domain-containing protein, put
PF3D7_1027300*,**	5.57	2.30E-06	peroxiredoxin
PF3D7_1036300*,**	9.37	2.59E-45	duffy binding-like merozoite surface protein 2
PF3D7_1102500*	4.56	2.40E-04	*Plasmodium* exported protein (PHISTb), UF
PF3D7_1132400	2.08	5.86E-04	conserved *Plasmodium* membrane protein, UF
PF3D7_1133700	3.24	1.27E-04	FHA domain-containing protein, put
PF3D7_1133800	2.11	9.79E-04	RNA (uracil-5-)methyltransferase, put
PF3D7_1134600**	4.73	8.65E-05	zinc finger protein, putative
PF3D7_1138800	2.34	6.24E-04	WD repeat-containing protein, put
PF3D7_1142100	3.87	7.22E-04	conserved *Plasmodium* protein, UF
PF3D7_1148700	2.98	9.80E-04	*Plasmodium* exported protein (PHISTc), UF
PF3D7_1212700	3.41	1.49E-04	eukaryotic translation initiation factor 3.A put
PF3D7_1228300	2.72	7.52E-05	NIMA related kinase 1
PF3D7_1233200*	3.45	1.57E-04	conserved *Plasmodium* protein, UF
PF3D7_1235300	3.28	7.10E-05	CCR4-NOT transcription complex s4, put
PF3D7_1236200*	3.65	1.33E-04	conserved *Plasmodium* protein, UF
PF3D7_1327300	4.04	2.35E-04	conserved *Plasmodium* protein, UF
PF3D7_1361200*,**	3.61	8.35E-06	conserved *Plasmodium* protein, UF
PF3D7_1362700*,**	5.96	2.60E-08	conserved *Plasmodium* protein, UF
PF3D7_1408200	4.00	4.85E-05	AP2 domain transcription factor AP2-G2
PF3D7_1431400*,**	3.50	3.94E-04	surface-related antigen SRA
PF3D7_1437200	1.89	8.05E-04	ribonucleoside-diphosphate reductase subunit,
PF3D7_1461800*,**	5.41	2.52E-06	conserved *Plasmodium* protein, UF
PF3D7_1466200*,**	4.38	2.17E-07	early gametocyte enriched phosphoprotein EGXP
PF3D7_1467600*,**	5.71	6.00E-07	conserved *Plasmodium* protein, UF
PF3D7_1469600	3.29	2.51E-04	acetyl-CoA carboxylase
PF3D7_1472200	3.29	3.09E-07	histone deacetylase, put
PF3D7_1473700*,**	5.88	2.89E-06	nucleoporin NUP116/NSP116, put
PF3D7_1474000*,**	4.77	1.01E-05	conserved *Plasmodium* protein, UF

Genes identified as having higher expression in samples with higher *mspdbl2* FPKM transcript levels with P-values < 0.001. Genes highlighted bold have known or suspected roles in gametocytogenesis. * indicate that these genes were also identified as having higher expression correlating to MSPDBL2 IFA expression (Table 4.4). ** indicates that these genes were also present in two or more of the discrete phenotype group analyses (Table 4.3). UF: protein has unknown function. Put: putative. [+] The log2 of the difference in transcription of the gene between the ranked sampled.

isolates with higher MSPDBL2 IFA expression with p-values < 0.01, and this identified 31 genes out of the 42 that also had higher expression in the MSPDBL2 IFA gene list.

31 genes were identified as having lower expression in samples with higher *mspdbl2* FPKM transcript levels (Appendix 10), and only 3 of these have known or suspected roles in gametocytogenesis (9%), showing that there is enrichment for gametocyte genes in samples with high *mspdbl2* FPKM values.

4.4 Discussion

Through various methods, combining protein and gene-level analysis, and using the DESeq2 framework in R, this project has identified a series of genes whose expression is associated with MSPDBL2 protein expression in schizonts. Several methods of identifying genes whose transcript levels covary with that of MSPDBL2 were tested in the DESeq2 framework. In each of these, *mspdbl2* was identified as the most differentially expressed gene, indicating that MSPDBL2 protein expression, as measured by IFA, can be recapitulated on the RNA level using low input RNA methods.

Attention was drawn to *mspdbl2* when it was identified as being under strong balancing selection in a panel of *P. falciparum* clinical isolates, and subsequent profiling of MSPDBL2 protein expression by IFA showed it to be expressed in an "on/off" fashion in schizonts (Ochola et al. 2010; Amambua-Ngwa et al. 2012). Investigation using RT-qPCR revealed variation in the rate at which *mspdbl2* was transcribed among both clinical isolates and laboratory lines, and IFA with specific antibodies showed the MSPDBL2 protein in less than 1% of schizonts (Amambua-Ngwa et al. 2012). These results have been recapitulated by the IFA data from the new clinical isolates from West Africa presented here. While most of the newly investigated isolates had MSPDBL2

expression in less than 1% of schizonts, expression ranged from 0%-6.6%, except for a single isolate (INV236) showing an exceptionally high number of MSPDBL2-positive schizonts (73%). This was matched by a very high transcript level, with an FPKM of over 21,000 for INV236, indicating that protein expression of MSPDBL2 could be reflected by the level of *mspdbl2* transcription. It should be noted that INV236 was observed in culture to yield mostly gametocytes at the subsequent intraerythrocytic cycle and there proved to be insufficient asexual parasites present to allow continuous *in vitro* culture (unpublished data from Ms. Lindsay Stewart, LSHTM).

Multiple approaches were used in this study to identify genes with altered transcription correlating to expression of MSPDBL2. The approach in which MSPDLB2 expression by IFA was treated as a continuous variable identified 52 highly significant (P-value < 0.001) genes that had higher transcription associated with higher MSPDBL2 protein expression. Of these, 12 (alongside *mspdbl2*) were previously known or suspected to be involved in gametocytogenesis. The equivalent list of genes which showed decreased expression in this analysis (130) contained only 10 known or suspected gametocytogenesis genes, indicating that the genes with increased transcript levels correlated with MSPDBL2 expression are enriched for gametocyte genes (25% of all genes with increased expression vs. 7% of all genes with decreased expression at a significance of $P < 0.001$). In every differential expression analysis carried out, genes with increased expression correlating to higher MSPDBL2 expression were enriched for gametocyte genes compared to genes with decreased expression.

In the analysis was carried out using *mspdbl2* transcript level, measured by FPKM, 40% of the genes identified were also shown to have higher transcription correlated with MSPDBL2 protein expression. The discrepancy between gene lists could be the result

of a temporal difference between transcription and expression of a protein, or due to the correlation of MSPDBL2 protein expression and transcript levels being poor at low expression and transcript levels.

When differential gene expression analysis based on IFA expression was carried out excluding INV236, there was some overlap in the genes identified, however this was not extensive, and MSPDBL2 was shown to be differentially expressed at a less significant p-value than in the other analyses. This indicates that the extremely high expression of MSPDBL2 in INV236 is driving some of the signal seen in these samples, the correlation seen between the protein (by IFA) and transcript (by FPKM) also showed significant noise at lower expression levels. In the future this will have to be addressed, and alternative, experimental methodologies should be investigated for obtaining isolates with higher *mspdbl2* expression.

One key regulator of gametocytogenesis commitment is *ap2-g*, a target of *gdv-1* and considered to be a critical element required for switching onto the sexual development pathway (Kafsack et al. 2014; Filarsky et al. 2018). Using an *ap2-g* conditional knock-out line recently led to the identification of 42 genes that had higher transcription in parasites with stabilised and transcribed *ap2-g* compared to *ap2-g* knock-out parasites (Josling et al. 2019). Five of these genes were found to have increased expression associated with increasing MSPDBL2 expression (5/52); PF3D7_0315600, *egxp* (PF3D7_1466200), PF3D7_1467600, *nup116/nsp116* (PF3D7_1473700), and PF3D7_1476600. The gene *nup116* has also been identified being de-repressed in HP1 knockout parasites, presumably as a result of *gdv-1* mediated *ap2-g* de-repression (Brancucci et al. 2014), as HP1 is a known repressor of *ap2-g* transcription (Filarsky et al. 2018). In the set of genes showing decreased expression associated with increasing

MSPDBL2 expression here, just one gene (1/130) (PF3D7_2467700) was identified as a target of *ap2-g* (Josling et al. 2019).

Additionally, genes with differential transcription were identified using isolates placed into discrete phenotype groups based on MSPDBL2 expression (using a 1%, 3% and <1% >3% cut-off). Cross-referencing the gene lists from these three analyses identified 35 genes that had higher transcription in the high MSPDBL2 expression groups in more than one analysis (Table 4.3). Five of these genes have been identified as targets of *ap2-g* using the conditional knock out line (Josling et al. 2019). It is interesting that a number of potential *ap2-g* target genes are showing increasing expression associated with MSPDBL2, as *ap2-g* itself was not identified in any of the analyses carried out. Likewise, *mspdbl2* was not identified as a target of AP2-G (Josling et al. 2019).

Targets of GDV-1, which is the upstream activator of *ap2-g*, have been identified using a parasite line in which ectopic *gdv-1* is tagged to a destabilisation domain, in a similar manner to the line created to investigate *ap2-g* targets (Filarsky et al. 2018). Targets of GDV-1 included *ap2-g*, *mspdbl2*, a PHISTa gene *pfg14_748* (PF3D7_1477700), and PF3D7_1477400. In the analysis of gene transcription using MSPDBL2 protein expression as a continuous variable, *pfg14_748* was shown to have correlated increased expression (Table 4.4), and although this gene was identified as a target of GDV-1, it does not seem to be downstream of AP2-G (Josling et al. 2019). Additional research has suggested that *pfs14_748* is expressed in parasites prior to their morphological conversion into gametocytes, with levels of the protein increasing as parasites develop along the gametocytogenesis pathway, and GFP-tagged Pfg14_748 was detectable alongside the known early gametocytogenesis marker Pfs16 (PF3D7_0406200) (Bruce et al. 1994) in asexual parasite cultures before gametocytes were observed to develop (Eksi et al. 2005). A number of AP2-G targets have been identified using chromatin

immunoprecipitation combined with whole transcriptome sequencing to identify *ap2-g* binding motifs throughout the genome. While several genes (including *nup116, gexp02*, and *pfs16*) containing the *ap2-g* binding motif were identified as also being upregulated in sexually-committed schizonts, sexually-committed rings, and stage I gametocytes, *mspdbl2* was not one of them which suggests that it is not activated by *ap2-g* (Josling et al. 2019).

The early gametocyte marker *pfs16* also had increased expression associated with MSPDBL2 protein expression, and was also found to be differentially expressed between the destabilisation domain-tagged GDV-1^{ON} and GDV-1^{OFF} parasites (Filarsky et al. 2018). Temporally, the increase in MSPDBL2 expression in parasites with stabilised *gdv-1* occurred within the same invasion cycle, after *gdv-1* induction and parallel to that of *ap2-g*, whereas genes thought to be targets of AP2-G did not show sharp increases in expression until the subsequent invasion cycle (Filarsky et al. 2018).
It might be expected that *ap2-g* and *gdv-1* would be identified as differentially expressed in isolates with higher *mspdbl2* expression, if *mspdbl2* is involved in gametocytogenesis, particularly when many of the targets of *gdv-1* and *ap2-g* have been identified. As the isolates used in this chapter were synchronised and allowed to develop into mature schizonts to obtain peak *mspdbl2* expression, it is likely that the expression profiles of *gdv-1* and *ap2-g* have already peaked and decreased by the time *mspdbl2* expression has increased, as indicated in Filarsky et al. (2018).

Two of the gamete-exported protein (GEXP) family genes had increased expression associated with increasing MSPDBL2 protein expression (*gexp02* PF3D7_1102500 and *gexp04* PF3D7_1372100), and *gexp02* was also shown to have increase transcription at higher *mspdbl2* FPKM levels. Members of the GEXP family are involved in protein

export occurring during gametocytogenesis (Silvestrini et al. 2010). Although the GEXP proteins were identified by mass spectrometry and gene enrichment analysis from the early gametocyte proteasome, *gexp02* has also been independently identified as part of the downstream cascade resulting from induced *gdv-1* expression, alongside *mspdbl2* and other early gametocyte markers (Filarsky et al. 2018). It has also been shown that *gexp*02 is de-repressed in HP1 conditional knock out parasites when HP1 is disrupted, presumably due to the activation of *gdv-1* (Brancucci et al. 2014).

Together, this information suggests *mspdbl2* could be expressed very early during the commitment process, possibly downstream of *gdv-1* and together with other early gametocytogenesis genes, but perhaps not in the *ap2-g* downstream cascade. This is supported by the absence of AP2-G binding sites upstream of MSPDBL2, but the presence of HP1 binding motifs, implicating it as either a target of GDV-1 or activated alongside *gdv-1* (Filarsky et al. 2018; Josling et al. 2019).

This exploratory work aimed at understanding *mspdbl2* was carried out on a panel of clinical isolates from West Africa, however more extensive research is required to confirm if and how *mspdl2* is directly involved in gametocytogenesis, whether that be directly or indirectly. If the gene is involved, then the timeline of events within individual cells needs to be explored and validated, along with identification of the upstream regulators and downstream targets of *mspdbl2*, as it does not appear to be a target of *ap2-g*. Due to the level of noise seen between isolates with low MSPDBL2 expression, experimental approaches should be investigated as a priority.

One experimental approach could be to create a *P. falciparum* laboratory line in which *mspdbl2* is tagged with GFP. This would allow cells expressing MSPDBL2 to be isolated by fluorescence activated cell sorting. These viable, sorted parasites would then be maintained in culture to determine whether they proceed along a sexual or asexual

pathway. In parallel, I would suggest carrying out RNA-seq on MSPDBL2-positive sorted parasites; the equivalent MSDPBL2-negative parasites from the same culture, and a collection taken from the mixed population before sorting. This could be done with long-term adapted laboratory lines (for example HB3, which has an MSPDBL2 expression rate in schizonts of ~10%), and also selected clinical isolates with relatively high proportions of MSPDBL2-positive parasites. This approach would allow more direct and experimentally comparable analysis to be undertaken between MSPDBL2-positive sorted cells and MSPDBL2-negative sorted cells. As these populations would originate from the same bulk culture, differentially expressed genes would be identified from an otherwise comparable background, increasing the reliability of the comparison.

This study has utilised RNA-seq generated from limited sample material, and another avenue that could be used to investigate *mspdbl2* is by utilising single-cell transcriptomics. Transcriptomes have been obtained from single malaria parasites (Poran et al. 2017; Ngara et al. 2018; Reid et al. 2018), although the nature of sequencing from RNA in a single cell results in significant challenges. Low input RNA-seq from a population of "pure" MSPDBL2-positive sorted cells provides an alternative to single-cell approaches and will be vital for validation when single-cell transcriptomics is undertaken.

A complementary approach to pure transcriptomic sequencing would involve inducing gametocytogenesis in laboratory lines and/or clinical isolates and periodically taking samples for IFA and RNA-seq across a time-course during gametocytogenesis induction. Immunofluorescence assays would reveal whether MSPDBL2 is induced during gametocyte induction, and RNA-seq would identify associated genes. This would complement previous studies in which RNA has been harvested along the timeline of gametocyte induction (Filarsky et al. 2018; Josling et al. 2019).

5. Application of single-cell transcriptomics to *P. falciparum* parasites

5.1 Introduction

Single-cell RNA-seq provides an avenue to investigate aspects of cell biology unattainable using traditional techniques. Originally implemented for interrogating cell populations in multi-cellular organisms with a particular focus on mammalian early development (Tang et al. 2009; Tang et al. 2010) and tumour biology (Ramsköld et al. 2012), it is increasingly being applied to unicellular organisms, both eukaryotic and prokaryotic in origin (Wang et al. 2015; Liu et al. 2017). Transcriptomics has been used broadly in malaria research (Otto et al. 2010; Rovira-Graells et al. 2012; Tarr et al. 2018), however RNA-seq from bulk preparations overlooks transcriptomic differences between parasites and presents an "average" transcriptome contributed to by millions of cells. *Plasmodium* infections are frequently clonal within a host, but within these parasites exist a rich variety of transcriptomic profiles designed to provide malaria parasites with redundancies and strategies for a number of processes, such as evasion from a host's immune system, erythrocyte invasion, and production of gametocytes (Duraisingh et al. 2003b; Rovira-Graells et al. 2012; Painter et al. 2017). Differing transcriptomic profiles can also be found between parasites causing uncomplicated malaria and parasites causing severe disease (Milner et al. 2012; Tonkin-Hill et al. 2018). Some, but not all, of the transcriptomic clonal variability is contributed to by genes that are part of highly polymorphic, antigenic families, such as the *var* gene family in *P. falciparum* which encode the cytoadhesive PfEMP1 proteins, implicated in virulence, pathogenicity, and development of severe malaria (MacKintosh et al. 2004).

Research utilising single cell transcriptomes has greatly expanded in recent years. Techniques originally developed for use with larger, mammalian cells (Tang et al. 2009; Picelli et al. 2014) have now been tested on unicellular organisms which are typically

much smaller than mammalian cells and contain far less RNA. (Kolisko et al. 2014; Liu et al. 2017). Transcript recovery rate has been observed to decrease and technical noise to increase with decreasing cell size and RNA amount, representing a major limitation for obtaining reliable transcriptomic profiles for unicellular organisms (Brennecke et al. 2013; Kolisko et al. 2014; Liu et al. 2017; Ngara et al. 2018). Despite these limitations, investigations into using single-cell transcriptomics on *Plasmodium* parasites have proceeded, and in recent years a number of studies have been published which use these techniques to interrogate single parasites. These studies have required a degree of adaptation and testing of standard protocols to increase suitability for *Plasmodium* parasites and maximise mRNA capture, but have none-the-less generated robust single-cell data, correlating to tens of thousands of cells (Poran et al. 2017; Ngara et al. 2018; Reid et al. 2018). For a gene such as *mspdbl2*, which displays a striking "on/off" expression pattern, single-cell analysis would give a superior view of the transcriptomic differences between MSPDBL2-expressing and non-expressing parasites, particularly from parasites isolated from the same genetic background.

A notable and extremely important example of within-clone variability is the production of gametocytes. Gametocytes and sexually committed asexual parasites represent a rare cell population within a background of non-committed asexual parasites. Conversion rate from asexual to sexual development occurs in only a fraction of parasites, typically at a rate of less than 10% (Smalley et al. 1981), although this varies between parasite isolates and is also influenced by the parasites' environment (Carter et al. 2013; Schneider et al. 2018). As only small proportions of parasites within a clonal population will be committed to the sexual development pathway at any one time, single-cell RNA-seq is a valuable tool for investigating the network of gene expression that results in gametocytogenesis. A key regulator of commitment to gametocytogenesis is the

transcriptional regulator *ap2-g*, without which parasites are unable to produce gametocytes (Kafsack et al. 2014). In a pioneering study, single-cell RNA-seq was used to profile *Plasmodium* parasites entering the gametocyte pathway and identify downstream targets of AP2-G with minimal interference from the transcriptomic profiles of non-committed parasites (Poran et al. 2017). The study used a *P. falciparum* parasite line in which *ap2-g* was tagged to a 'destabilisation domain' and used for single cell RNA-seq. Parasites of this line are only able to express AP2-G when cultured in the presence of the ligand, Shld-1, which allowed direct comparison of the transcriptomic profiles of AP2-G^{ON} and AP2-G^{OFF} parasites from an otherwise identical genetic background. Over 18,000 single *Plasmodium* infected erythrocytes were isolated throughout the invasion cycle and sequenced, and the resulting transcriptomic profiles were sufficiently robust to differentiate parasites based on their position in the life cycle (Poran et al. 2017). Analysis of differential expression identified a panel of genes expressed exclusively in the AP2-G^{ON} parasites that were are not known to be involved in gametocytogenesis, but are likely activated by AP2-G, indicating the effectiveness of single-cell RNA-seq techniques (Poran et al. 2017).

Gametocytogenesis has also been a focus of other *Plasmodium* single-cell transcriptomic studies, perhaps owing to the drastic differences seen between asexual and sexually committed parasites. Using the Smart-seq2 protocol (described in Section 1.6), single *Plasmodium* infected erythrocytes were collected at six timepoints during the intraerythrocytic life cycle to investigate temporal changes to transcriptomic profiles (Ngara et al. 2018). While the sequenced parasites broadly grouped with their respective timepoints, they showed substantial heterogeneity in their gene expression which prevented tight clustering of parasites from the same isolation time. However, a small number of sexually-committed parasites were identified by detection of an early

gametocytogenesis genetic marker, *etramp3.0*, and subsequently a small panel of genes induced in these cells was identified, members of which may have involvement in gametocytogenesis (Ngara et al. 2018).

An ongoing challenge is being able to measure the reliability of transcriptomes sequenced from single cells. To test the specificity of the techniques with *Plasmodium*, erythrocytes infected with *P. falciparum* (tagged with GFP) or *P. berghei* (tagged with mCherry) were mixed and isolated by FACS. Single cell transcriptomic profiles of these parasites were found to be >98% specific to the correct species (Reid et al. 2018). By comparing single cell transcriptomes to bulk RNA-seq time course dataset collected throughout the intraerythrocytic life cycle (Otto et al. 2010), Reid et al. (2018) were able to differentiate single cells as asexual parasites, male gametocytes or female gametocytes and predicted abrupt transcriptomic changes associated with progression through the life cycle, in contrast to the smooth, overlapping transitions previously described (described in Section 1.7 and Figure 1.10).

Another popular technique for both genomic and transcriptomic amplification is multiple displacement amplification (MDA) (described in detail in Section 1.6). First developed from a loop-based amplification technique for use with bacteria and bacteriophages, it was first adapted for use with genomic DNA from limited material (tested from 300ng – 0.3ng) (Dean et al. 2002). Compared to PCR based methods, MDA was shown to produce much more uniform amplification of genomic DNA which resulted in 3-fold amplification bias, compared to between 10^2 and 10^6-fold for the PCR techniques tested (Dean et al. 2002), making it ideal for transcriptomics where sampling error affecting amplification of gene transcripts can drastically affect the interpretation of the data.

Plasmodium parasites exhibit remarkably plastic gene expression, all the way down to the single-parasite level. Some aspects of *Plasmodium* biology, such as gametocytogenesis, are difficult to investigate due to the background 'noise' introduced by the vast numbers of parasites of different stages. Single-cell RNA-seq is a relatively novel technique that has only been optimised and carried out on *Plasmodium* parasites over the last few years. The ability to sort parasites based on a fluorescently-tagged gene of interest followed by single-cell RNA-seq is a method of obtaining transcriptomes from parasites without the interference of other parasites. In this chapter, my aim was to trial single-cell RNA-seq techniques with *P. falciparum* parasites in order to optimise a protocol for these unique cells and obtain reliable and representational single cell transcriptomes. These techniques could then be applied to investigate rare cell populations, such as those parasites expressing *mspdbl2*.

5.2 Materials and Methods

5.2.1 Parasite culturing conditions and schizont isolation

The *P. falciparum* laboratory line 3D7 was cultured according to conditions specified in Section 2.1.3 using washed research red blood cells (Section 2.1.2). Parasites were synchronised on a Percoll® gradient (Section 2.1.5) approximately 48 hours prior to purification and Percoll® was then used to isolate schizonts from synchronised parasites. Low number and single parasites were obtained by either serial dilution or by Fluorescence-Activated Cell Sorting (FACS). The concentration of schizonts obtained by serial dilution was assessed using a Neubauer improved C-Chip™ disposable haemocytometer (DHC-N01, NanoEnTek Inc., South Korea). Schizonts which were isolated by FACS were first stained with Vybrant® DyeCycle™ Green DNA stain (Life Technologies) and were sorted by Dr. Sarah Tarr (LSHTM) in a sterile environment on

a BD FACSAria™ Fusion cell sorter using an 85µm nozzle and purity mask, which prevents a cell from being sorted if it is in close proximity to another cell.

5.2.2 Whole transcriptome amplification from limiting cell numbers

Schizonts were aliquoted into either single-cells or low-number multiple cells (1 cell, 3 cells, 10 cells, 20 cells, 100 cells or 1000 cells). Reverse transcription and amplification of cDNA was carried out using the REPLI-g® whole transcriptome amplification (WTA) kit (Section 2.6). The manufacturer's protocol was followed initially before several permutations were tested in order to improve output. The final protocol is outlined in Section 2.6 and incorporated the use of the RNase inhibitor recombinant RNasin® (Promega, WI, USA). Alongside each experiment, 50pg of previously purified, DNase-treated *P. falciparum* 3D7 RNA was included as a control (extracted using methods outlined in Section 2.3).

5.2.3 Whole transcriptome amplification using the Fluidigm C1™ system

The Fluidigm C1™ system which integrates with the SmartSeq® v4 chemistry (Clontech), based on the template-switching PCR-based amplification methods (Section 1.6) was used to isolate single parasites fixed with dithobis (succinimidyl propionate) (DSP) and carry out whole transcriptome amplification (Section 2.7). Schizonts were isolated by Percoll® purification (Section 2.1.5) and fixed with DSP (Section 2.7). Cells were supplemented with BSA to reduce stickiness and loaded onto the mRNA Integrated Fluidic Circuit (IFC) microfluidic chip. Reagents needed for the C1™ SmartSeq® v4 reaction were also loaded onto the IFC along with DTT to reverse the DSP cross linking (Section 2.7).

5.2.4 Whole transcriptome amplification quality control

Amplified low input material was quantified fluorometrically either on a Qubit® 2.0 using the High-Sensitivity dsDNA reagents (Thermo Fisher Scientific, MA, USA), or on a Spectramax M3 using the Quant-IT™ PicoGreen® dsDNA reagents (Thermo Fisher Scientific, MA, USA). Quality of amplified material was checked on a Bioanalyzer using high-sensitivity dsDNA reagents (Agilent Technologies, CA, USA). To ensure the cDNA amplified originated from RNA, an intron-spanning PCR was developed to amplify part of the sequence within the gene *msp-4* (Section 2.13) and used to discriminate between cDNA and gDNA produced in the single cell experiments.

5.2.5 Sequencing and transcriptomic analysis of cDNA from low cell number and single cell material

Samples were prepared for sequencing using the TruSeq Nano LT DNA Library preparation kit (Illumina) using a 350bp fragment size and were sequenced paired end over 150 cycles according to Section 2.8.1. Short reads were assessed for quality and assembled to the *P. falciparum* 3D7 (Pf3D7) reference genome, (Section 2.9.2). Analysis of single-cell and low cell number transcriptomes was carried out in R statistical software using gene counts generated using HTSeq-count (Anders et al. 2015) (Section 2.14), using the custom made *P. falciparum* GFF reference file with polymorphic regions masked described in Section 2.14. Non-*Plasmodium* microbial contamination was assessed using Kraken (Section 2.14).

5.3 Results

Initially, the REPLI-g® protocol was carried out according to Qiagen manufacturer instructions, with no additions or changes. Matured and purified *P. falciparum* 3D7 schizonts were isolated into single cell aliquots or 100 cell aliquots. Two types of negative control were included, following the REPLI-g® protocol without the addition of any biological material; the first were wells which remained open during FACS, and the second were sealed during FACS. All samples were treated identically throughout the REPLI-g® protocol (FACS and REPLI-g® protocol carried out by Dr. Sarah Tarr, LSHTM). From this experiment, four single cells and one of the 100 cell aliquots, along with two negative control wells (one open during sorting, the other sealed) were selected for whole transcriptome sequencing on an Illumina MiSeq.

Alignment rates of short sequence reads to the *P. falciparum* 3D7 genome ranged between 37% - 89% for the samples containing schizonts (Table 5.1). For the single cells, all except one had an overall alignment rate of over 70%. The 100-cell sample had an alignment rate of 89%, the percentage of non-aligned reads is a reflection of the degree of contamination within a sample. Non-*Plasmodium* microbial contamination was assessed for each sample using Kraken (Wood and Salzberg 2014), a taxonomic assignment programme for assessing diversity of short sequence reads, using the MiniKraken database, representative of bacterial, archaeal, and viral species (Table 5.2). For each of the negative reactions (open and sealed during sorting), approximately 50% of short reads could be assigned to microbial species. The remaining 50% was likely either unclassifiable or belonged to non-microbial species, possibly resulting from human contamination. Approximately 40% of reads sequenced from the negative well which was sealed during FACS belonged to Enterobacteriaceae (Phylum: Proteobacteria), with 3% assigned to *Escherichia coli*. The negative well kept open

Table 5.1. Sequencing statistics of the *P. falciparum* 3D7 schizont samples containing limited cell numbers.

Sample	Number of cells	Number of reads (millions)	Alignment rate to 3D7 (%)
1G	100	2.6	89.2
2G	1	3.3	80.8
2H	1	2.3	70.0
3D	1	2.7	37.2
3E	1	1.8	77.2

Samples containing 100 cells or one cell were isolated by FACS and whole transcriptome amplified by Dr. Sarah Tarr (LSHTM). Alignment of short sequencing reads to the *P. falciparum* 3D7 reference genome ranged from 37% to 89%.

during FACS contained 7% of reads mapping to *E. coli* and 4% of reads mapping to within the Burkholderiales (phylum: Proteobacteria). Reactions which contained *Plasmodium* parasites had considerably fewer reads mapping to contaminants with between 8% and 33% mapping to non-*Plasmodium* microbial species, and mostly these were assigned to bacterial species. A few phyla of bacteria were represented in multiple samples (Actinobacteria, Firmicutes, Spirochaetes, Proteobacteria). The species *E. coli* and *Borrelia duttonii* were present in all samples containing *Plasmodium* cells.

To assess the *Plasmodium* transcriptome data qualitatively, the bam files which contain the mapped short sequence reads were viewed in the Artemis genome browsing software against the *P. falciparum* v3 3D7 reference genome. It was found that, although coverage across the transcriptome was high, sequencing reads were not restricted to exonic regions. Instead sequence data were observed extending into and throughout intronic and intergenic regions, showing no differentiation between exons

Table 5.2 Contamination present in sequence data from whole transcriptome amplified samples.

Taxonomic name and rank		Percentage of short reads mapping to the named taxa for each sample						
		1B (-ve closed)	1C (-ve open)	1G (100 cells)	2G (1 cell)	2H (1 cell)	3D (1 cell)	3E (1 cell)
Bacteria	Domain	51.0	51.0	4.50	6.50	8.90	31.7	9.35
Proteobacteria	Phylum	43.5	45.0	1.10	3.29	6.39	24.6	6.76
…Gammaproteobacteria	Clade	43.1	40.3	0.66	2.83	3.36	24.0	6.32
…Betaproteobacteria	Clade	0.30	4.30	0.04	0.15	2.79	0.21	0.05
…Alphaproteobacteria	Clade	0.06	0.04	0.24	0.22	0.18	0.33	0.34
Actinobacteria	Phylum	1.25	1.37	0.21	0.57	0.23	0.37	0.25
Firmicutes	Phylum	0.09	0.22	0.64	0.58	0.48	0.35	0.73
…Bacilli	Clade	0.09	0.22	0.27	0.22	0.19	0.25	0.40
…Clostridia	Clade	0.00	0.00	0.25	0.27	0.24	0.06	0.20
Tenericutes	Phylum	0.00	0.00	0.26	0.18	0.16	0.09	0.12
…Mollicutes	Clade	0.00	0.00	0.26	0.18	0.16	0.09	0.12
Spirochaetes	Phylum	0.00	0.00	1.07	0.70	0.65	0.32	0.49
Viruses	Domain	0.61	0.14	0.92	0.77	0.68	0.38	0.57
Archaea	Domain	0.00	0.00	0.57	0.55	0.40	0.10	0.37
Euryarchaeotra	Phylum	0.00	0.00	0.57	0.55	0.40	0.10	0.37
…Methanococci	Clade	0.00	0.00	0.3	0.26	0.23	0.06	0.17
…Methanomicrobia	Clade	0.00	0.00	0.21	0.25	0.14	0.03	0.18

The software Kraken was used to assign taxonomic labels to the short sequence reads that did not map to *P. falciparum* 3D7. A number of Phyla and clades were represented at above 0.1% of short sequence reads per sample, which are listed above. The vast majority of contamination originated from bacteria. Other taxa were represented at lower levels. The negative wells contained far greater amounts of contamination due to no competition for reagents with *Plasmodium* DNA.

and introns, and therefore showing no evidence of splicing (Figure 5.1). This lack of spliced material and the presence of substantial data in intergenic regions indicated that the material sequenced could not have been cDNA originating from RNA, but instead was likely to include genomic DNA contamination originating from the same cells, despite the DNase "gDNA WipeOut" step included in the REPLI-g® protocol.

5.3.1 Design and testing of a cDNA and gDNA discriminatory PCR on single cell samples

As the first single-cell experiment resulted in the amplification and sequencing of material originating from *P. falciparum* genomic DNA, a PCR assay was developed to rapidly and effectively discriminate between gDNA and cDNA before sequencing was undertaken. Three genes known to be expressed during schizogony were selected for primer design, *cyrpa*, *msp-4*, and *clag9* (Bozdech et al. 2003b). Two discriminatory strategies were tested. The first was to design primer sets with one primer spanning an exon-exon boundary, aimed at preventing primer binding and therefore amplification from gDNA in which introns are present. The second was an intron-spanning strategy designed to have forward and reverse primers in separate exons to use amplicon size difference to distinguish between gDNA and cDNA products. After testing on bulk prepared *P. falciparum* 3D7 gDNA and cDNA, the *msp-4* intron-spanning strategy was selected (PCR strategy shown in Figure 2.1) due to the amplicon size difference with cDNA and gDNA products being clearly identifiable. This PCR was used to assess all material prepared by whole transcriptome amplification and ensure samples showing only the *msp-4* cDNA product were sequenced.

The *msp-4* intron-spanning PCR was tested on the WTA samples that were sequenced, along with other single cells which were amplified but not sequenced. A gDNA positive

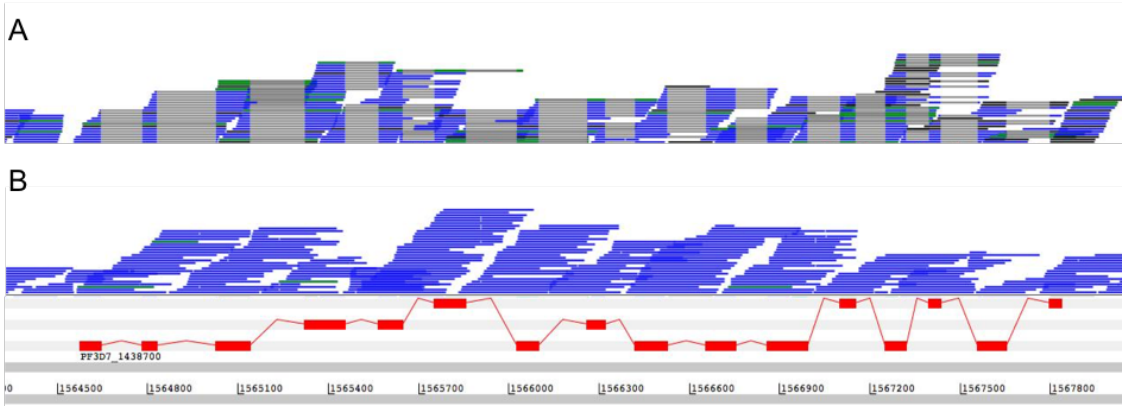

Figure 5.1. Sequence data from bulk RNA-seq and single-cell RNA-seq mapping to multi-exon genes shown in the Artemis genome browser. Schizonts from *P. falciparum* 3D7 were isolated by FACS and dispensed into single or 100 cell aliquots. A selection of these samples were whole transcriptome amplified by the REPLI-g® MDA chemistry and sequenced. Artemis was used to view transcriptomes to assess quality. PF3D7_1438700 is a multi-exon gene on chromosome 14, and is shown here as a representation of multi-exon genes across the genome. **A.** The expected distribution of sequencing short reads from bulk RNA-seq of *P. falciparum* 3D7 parasites (Dr. Sarah Tarr, unpublished data), showing sharp boundaries between exonic sequence data (blue and green bars) and the absence of intronic sequence data (grey), indicative and expected of data originating from RNA due to the action of splicing. **B.** Sequence reads pooled from the single cells and the 100 cell sample that were amplified using the REPLI-g® method and sequenced. Reads extend throughout intronic regions, showing no evidence of splicing. This indicates that the sequence data originated at least in part from *P. falciparum* 3D7 genomic DNA contamination rather than from RNA.

control and a cDNA positive control, both originating from bulk 3D7 parasite culture were also included (Figure 5.2). The gDNA positive control showed a strong band approximately 600bp in size (expected product 589bp), while the cDNA control showed a strong band approximately 400bp in size (expected product 440bp), along with a weak band the same size as the gDNA product; thought to originate from unspliced *msp-4* product. A third much larger band seen in the PCR product after amplifying cDNA reaction is likely to be non-specific. The results from the PCR from samples of 100 and 1000 parasite cells show the REPLI-g® kit had exclusively amplified gDNA and not cDNA, as only *msp-4* gDNA product was visible (Figure 5.2).

5.3.2 Optimisations of the REPLI-g® protocol to minimise gDNA contamination

Several modifications to the standard REPLI-g® protocol were trialled in order to eliminate contamination from *Plasmodium* gDNA. Two steps of the REPLI-g® protocol were suspected to be performing with low efficiency to allow gDNA amplification and at the same time preclude effective generation of cDNA. One of these is the 'gDNA WipeOut' DNase treatment step, which must be ineffective for gDNA to be present during the amplification. In addition, as no cDNA was seen (rather than a mix of gDNA and cDNA) it was possible that the reverse transcription (RT) step was failing or performing inefficiently. Reverse transcription failure could be due to inefficiencies in the reaction itself or could be because of an underlying lack of high enough quality RNA for the reaction to proceed due to RNA degradation.

To directly address the gDNA contamination, aliquots of 1000 and 100 schizont-infected erythrocytes were prepared by serial dilution. Alongside each iteration of the protocol tested, a positive control of previously purified and DNase-treated bulk *P. falciparum* 3D7 RNA was included. A subset of reactions was set up to include an

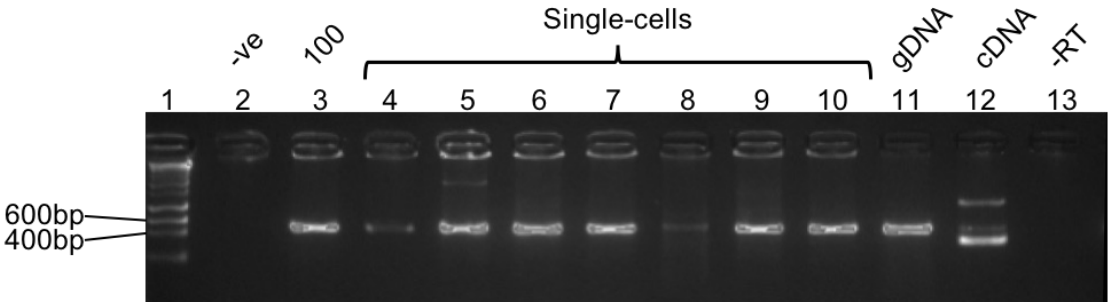

Figure 5.2. *msp-4* intron-spanning PCR confirms the absence of cDNA in single cells. Next-generation sequencing revealed that the REPLI-g® protocol had failed to sufficiently capture and amplify the RNA from single cell and 100 cell samples. A PCR assay was designed to test for gDNA contamination in amplified samples prior to whole transcriptome sequencing. The gene *msp-4* is a multi-exon gene known to be reliably expressed in *P. falciparum* schizonts. Primers were designed within exon 1 and exon 2 of this gene to produce different sized products from cDNA and gDNA based on exclusion/inclusion of the intron. Expected product sizes are 589bp for gDNA and 440bp for cDNA. The *msp-4* assay was tested on the four single cells (lanes 4-7) and 100 cell sample (lane 3) that underwent sequencing, along with three additional single cells that did not undergo sequencing (lanes 8-10). Genomic DNA and cDNA independently extracted from *P. falciparum* 3D7 bulk culture were run as positive controls (lanes 11 and 12, respectively) and show that the gDNA shows a larger product than cDNA, as expected. The REPLI-g® negative control (lane 2) and the reverse transcription negative control (lane 13) were both negative. As suspected from the sequencing data, all the cells prepared with the REPLI-g® protocol contained only gDNA *msp-4* product.

additional treatment with DNase 1 (Qiagen) alongside the REPLI-g® gDNA WipeOut step. Another subset of reactions was set up to address possible inefficiencies in the RT reaction resulting from the unusual AT richness of the *P. falciparum* genome (~80% genome wide) and these samples were run with a lower RT incubation temperature of 37°C rather than the recommended 42°C. A third subset of reactions was set up to combine these alterations, which underwent an additional DNase treatment step and were incubated at 37˚C during the RT reaction. The success of each alteration was assessed using the *msp-4* intron-spanning PCR (Figure 5.3). In this experiment, all but one of the reactions failed outright, based on the *msp-4* PCR, and the one sample that had successful amplification showed *msp-4* gDNA product. This suggested that there was a more fundamental issue that was affecting reaction success, particularly as the purified RNA control showed a product whose size was consistent with cDNA, in addition to some residual gDNA (Figure 5.3, lane 2). This indicates not only that the protocol can be successful given ideal starting conditions, but also that something present specifically in the whole cell reactions was impacting the reaction success. One avenue that seemed likely was the introduction of RNases into the reactions, and this was explored further.

5.3.3 RNA degradation is a probable cause of REPLI-g® failure

One possible entry route for RNases was thought to be through the filtered 1X PBS used to isolate, wash, and dilute parasites. Therefore, sterile, molecular-grade 1X PBS, free from RNases and DNases, was tested alongside the originally used PBS for cell isolation and dilution. Reactions were set up using 1000 cells and the purified RNA positive control that was used previously. The first purified RNA reaction was set up using the same 1X PBS that was used for the whole cell dilutions and the second reaction was set up with molecular-grade 1X PBS, using 50pg of purified RNA. Both of

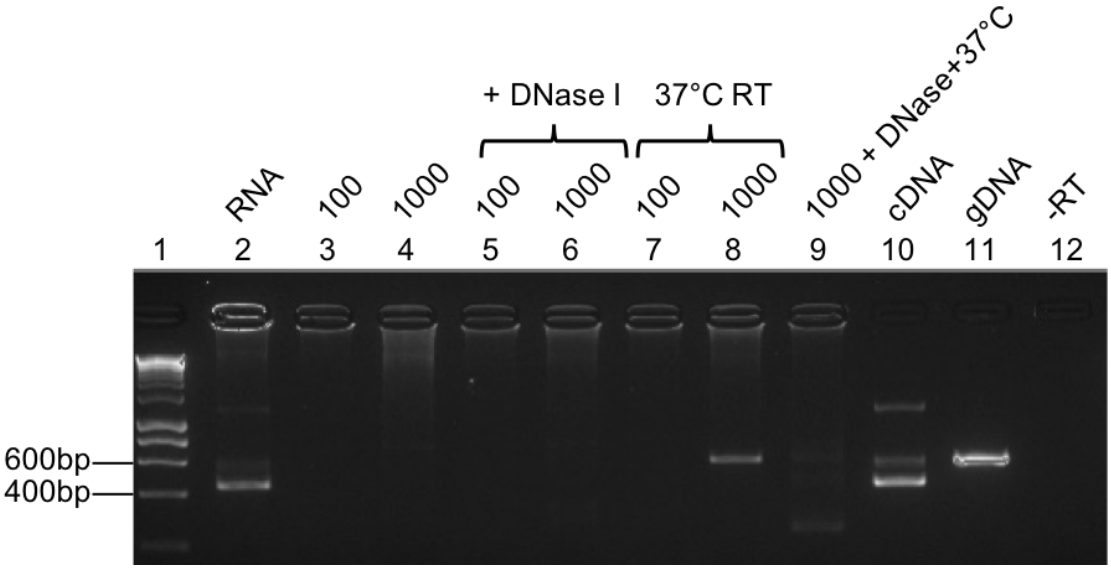

Figure 5.3. Alteration of reverse transcription incubation and addition of DNase I treatment does not improve REPLI-g® success. The REPLI-g® protocol was carried out on 100 and 1000 cell aliquots of *P. falciparum* schizonts to reduce genomic DNA contamination. The *msp-4* intron-spanning PCR was used to determine the success of each alteration. Lane 2 contains preciously purified 3D7 RNA (1ng), which underwent the REPLI-g® protocol following the manufacturer's instructions; the *msp-4* PCR showed that cDNA had been produced, mixed with a small amount of gDNA, seen as a faint band (lane 2). The standard REPLI-g® protocol was also carried out on 100 cell (lane 3) and 1000 cell (lane 4) aliquots, and these did not produce gDNA or cDNA. The first alteration tested was the addition of 1µl DNase 1 alongside the standard gDNA WipeOut step, this was tested on 100 cell (lane 5) and 1000 cell (lane 6) aliquots and these reactions failed to produce cDNA or gDNA *msp-4* amplicons. The second alteration was to lower the temperature of the reverse transcription reaction from 42°C to 37°C, again tested on 100 cells (lane 7) and 1000 cells (lane 8), which generated *msp-4* gDNA product. The final alteration was to combine the addition of DNase 1 with a 37°C RT temperature in a reaction of 1000 cells (lane 9). Lanes 10-12: cDNA positive control, gDNA positive control, and negative control, respectively.

these reactions showed *msp-4* product consistent with cDNA. Two 1000 cell reactions were set up, one with the original 1X PBS and the second with molecular-grade PBS; these both only produced *msp-4* gDNA. The final two reactions contained 1000 cells spiked with 0.5ng of purified RNA, one with the old PBS and one with the molecular-grade PBS to test whether the REPLI-g® kit could generate cDNA from the purified RNA in the presence of cellular material., and these also only showed *msp-4* gDNA (Figure 5.4). These results indicate that it is unlikely to be an external factor such as PBS resulting in the destruction of the RNA in a reaction. The failure of the whole cell reactions spiked with purified RNA to produce cDNA indicated that there were RNases being introduced into the reactions by association with the cells themselves resulting in RNA degradation as soon as the cells are lysed by the REPLI-g® lysis buffer.

5.3.4 Testing a modified protocol with the addition of an RNase inhibitor

Introduction of RNases into the reaction would have to occur during the initial steps of the REPLI-g® protocol, likely prior to cell lysis and resulting in degradation of the RNA before the reverse transcription reaction is carried out. This would explain why the whole cell reactions spiked with purified RNA only produced gDNA product. While the REPLI-g® kit contains an RNase inhibitor, details beyond this are proprietary information and as such the inhibitor is of unknown type and concentration. Other protocols and chemistries for single-cell transcriptomics use an RNase inhibitor at the cell lysis step, and again immediately afterwards (Poran et al. 2017). The REPLI-g® protocol was adjusted to include the RNase inhibitor, recombinant RNasin® (rRNasin®, Promega) in the initial lysis mix into which whole cells were directly diluted into. As rRNasin® is denatured at temperature higher than 50°C, additional rRNasin® was added immediately following the 95°C cell lysis step and these

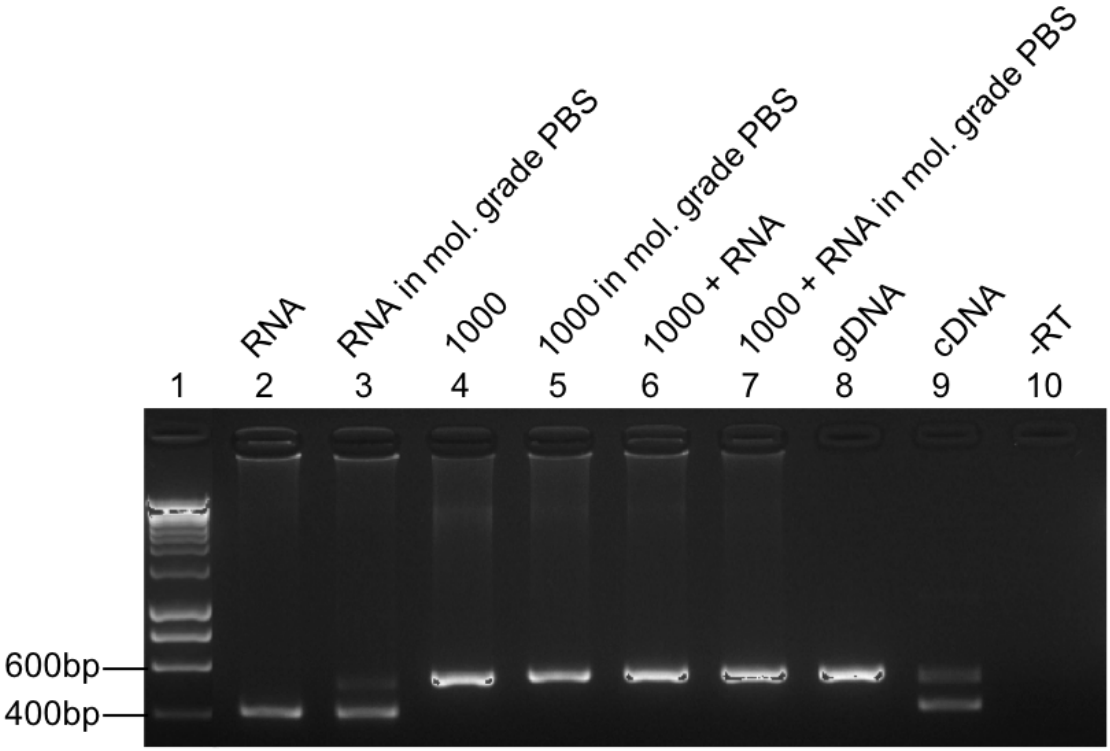

Figure 5.4. *msp-4* PCR testing the effect of using DNase/RNase free PBS on gDNA contamination. A possible cause for the failed reactions and gDNA contamination was through introduction of RNases through the PBS used to isolate whole cells. As such, molecular-grade, RNase/DNase free 1X PBS was tested against the sterile-filtered PBS used previously. The *msp-4* intron-spanning PCR was carried out on the samples. Reactions containing 50pg of purified RNA were used as controls; one diluted into the original PBS (lane 2) and one diluted into the molecular-grade PBS (lane 3) all showed *msp-4* cDNA product. When using 1000 whole cells in a reaction with the old (lane 4) and molecular grade (lane 5) PBS, only *msp-4* gDNA product was seen. Two reactions containing 1000 cells were spiked with 0.5ng purified RNA, to test if this was able to be reverse transcribed in the presence of whole cells; one of these reactions was diluted with the old PBS (lane 6) and the second with molecular-grade PBS (lane 7); these both showed only *msp-4* gDNA product. Lanes 8-10: gDNA and cDNA positive controls, and negative control, respectively.

alterations were tested on purified RNA and on reactions containing 1000 cells (Figure 5.5). Two purified RNA (50pg) positive controls were run without the addition of rRNasin®, both of which produced cDNA in the *msp-4* PCR. A third experimental reaction contained 50pg of purified RNA, set up with the addition of 1U/µl (the manufacturer recommended concentration) of rRNasin® in the lysis mix and an additional 1U/µl added immediately after cell lysis, to ensure that the rRNasin® was not having a detrimental effect on the WTA reaction, this reaction also produced *msp-4* cDNA product. A reaction containing 1000 whole cells diluted directly into lysis mix, without any rRNasin® also produced cDNA, however at an approximately 50:50 mix with gDNA. Another reaction (a repeat of one carried out in the previous experiment) contained 1000 whole cells, spiked with 0.5ng purified RNA, and unlike in the previous experiment this reaction produced a mix of cDNA and gDNA. Two reactions contained 1000 cells each, with rRNasin® added at a final concentration of 1U/µl and 3U/µl and these reactions produced the greatest amount of cDNA. Finally, 1000 whole cells were spiked with 0.5ng purified RNA and had rRNasin® added at 1U/µl, resulting in a reaction that produced very faint bands of mixed *msp-4* cDNA and gDNA (Figure 5.5). Over all of the REPLI-g® experiments, the success or failure of reactions has not been consistent, however a possible contributory factor is that in this experiment the whole cells were diluted directly into the REPLI-g® lysis buffer and used immediately, whereas previously these cells had been frozen in 1000 cell aliquots at -80C and stored overnight before whole transcriptome amplification possibly resulting in RNA degradation. The 1000 whole cell reactions with added rRNasin® contained approximately double the concentration of DNA compared to those without rRNasin®. The higher concentration of rRNasin® did not seem to improve cDNA production, and together these results suggest that the adapted protocol with additional RNase inhibitor to be a promising improvement.

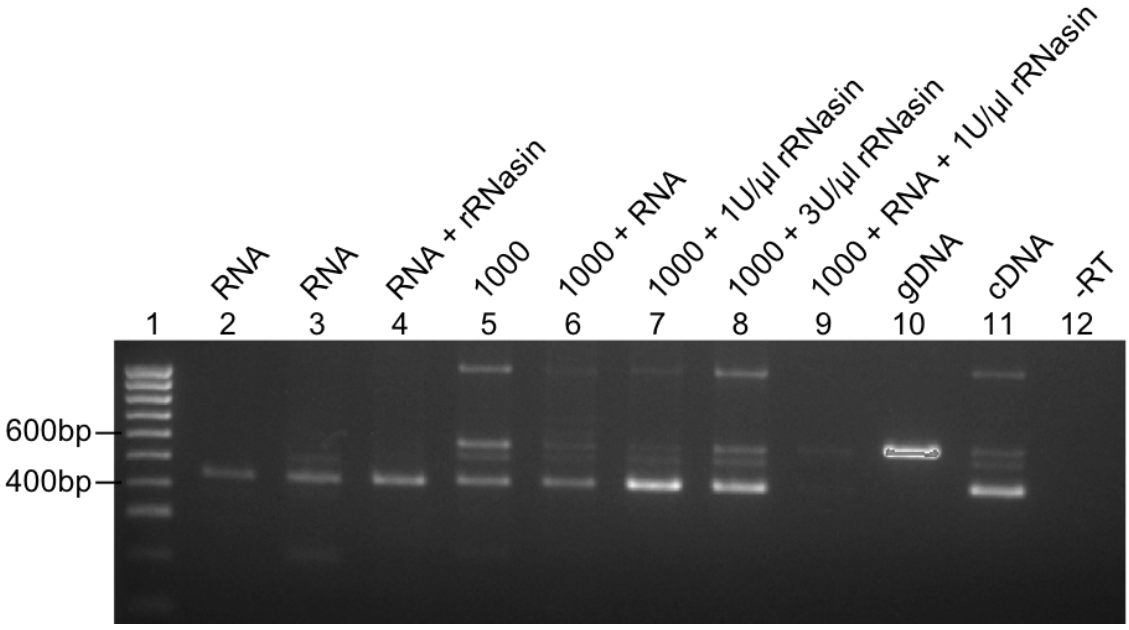

Figure 5.5. *msp-4* PCR showing the effect of adding an RNase inhibitor on the REPLI-g® protocol. Reactions were set up to test whether the introduction of rRNasin® into the reactions before and after cell lysis would prevent high degradation of RNA. The *msp-4* intron-spanning PCR was carried out on the reactions. Three purified RNA controls, two following the standard REPLI-g® protocol (lanes 2 and 3) and one with the rRNasin® addition (lane 4) all produced *msp-4* cDNA. The remaining reactions were set up using "fresh" aliquots of 1000 parasites (not frozen between collection and WTA as had been done previously). The 1000 cell reaction following the REPLI-g® protocol produced a mix of cDNA and gDNA (lane 5). A repeat of a previous reaction contained 1000 cells spiked with 0.5ng purified RNA also produced a cDNA and gDNA mix, with stronger cDNA product (lane 6). When 1000 cell reactions included rRNasin® at two concentrations, cDNA products were notably stronger than gDNA product (rRNasin® included at 1U/µl and 3U/µl in lanes 7 and 8, respectively). 1000 cells spiked with 0.5ng purified RNA plus 1U/µl rRNasin® produced very weak mixed cDNA and gDNA product. Lanes 10-12 show the gDNA, cDNA and negative controls.

The sensitivity of the RNase inhibitor adapted protocol was tested to assess whether the success of the 1000 whole cell reactions could be recreated when using far fewer cells per reaction; with the target being a single cell per reaction. The REPLI-g® protocol was followed, utilising molecular-grade PBS, freshly isolated and diluted parasites, and with the addition of rRNasin® at a final concentration of 1U/µl with the lysis buffer into which cells were diluted and again immediately following the 95°C cell lysis step. Parasites were diluted to the appropriate concentration to obtain 1000, 100, 10, 3, or 1 cell per reaction. The *msp-4* PCR was carried out to determine success (Figure 5.6). The reaction containing 1000 whole cells was unsuccessful. Two reactions were set up to contain 100 parasites; one of these produced predominantly gDNA, while the second produced a mix of gDNA and cDNA. Two reactions contained 10 whole cells, and these both showed strong, clean *msp-4* cDNA product. Three reactions contained three cells each. One of these failed, the second produced a mix of cDNA and gDNA, and the final reaction showed cDNA product. Three reactions contained single cells. Two of these failed, while the third produced a PCR product consistent with cDNA (Figure 5.6).

With the success of the low cell number experiment, particularly when using between one and ten cells, the next experiment was set up to include reactions set up by limiting dilution and reactions containing cells isolated by FACS (FACS was carried out by Dr. Sarah Tarr, LSHTM). Parasites were prepared and enriched for schizonts as a single culture, before being split into 'FACS isolated' and 'serially diluted' aliquots, in order to minimise technical bias. FACS cells were sorted directly into the lysis buffer with 1U/µl rRNasin®. For both cell isolation methods, reactions were set up with 100, 10, 3, and single cells. Based on the *msp-4* PCR, both of the 100 cell reactions (FACS and limiting dilution isolated) contained gDNA. Of the remaining reactions containing serially diluted cells, one 3 cell reaction produced a strong *msp-4* cDNA band, and one

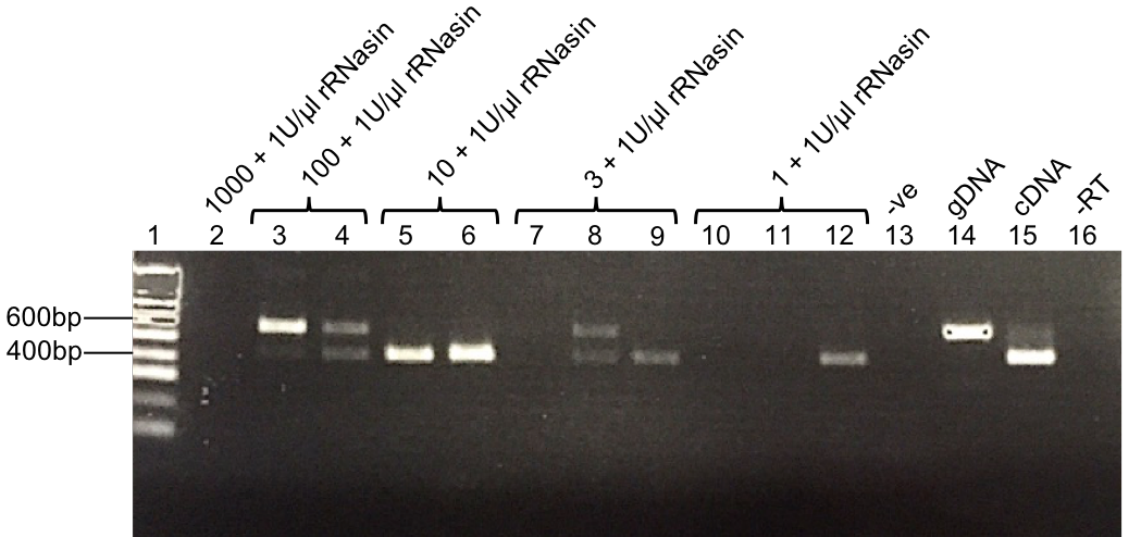

Figure 5.6. *msp-4* PCR carried out on samples containing 1 – 100 schizonts amplified using the RNase inhibitor-supplemented REPLI-g® protocol. The REPLI-g® protocol was adapted to include the addition of rRNasin® in the cell lysis mix into which cells were deposited and again just after the cell lysis 95˚C incubation step. The protocol was tested on several reactions containing 1000 cells down to single cells, obtained by limiting dilution. The *msp-4* PCR was used to assess gDNA and cDNA product. The 1000 cell reaction did not produce any DNA (lane 2). The reactions containing 100 cells produced mixes of cDNA and gDNA PCR product; with one (lane 3) producing far greater gDNA than cDNA, and the other (lane 4) producing them in roughly equal amounts. Both of the reactions containing 10 cells produced strong bands consistent with cDNA (lanes 5 and 6). One out of three reactions containing three cells failed (lane 7), another produced a mix of cDNA and gDNA (lane 8), and the final one had a weak but clean cDNA product. Of three single cell reactions, two failed (lanes 10 and 11), and the final one produced a clean cDNA *msp-4* band (lane 12). Lanes 14-16 show to gDNA, cDNA and negative controls.

single cell produced a weak *msp-4* cDNA band. The reactions containing cells isolated by FACS were more successful. Both 10 cell reactions, one 3 cell reaction, and two single cell reactions produced *msp-4* cDNA, although the product from the single cells was very faint (Figure 5.7). The single cell reactions showing *msp-4* cDNA product (one reaction contained a cell isolated by limiting dilution, two by FACS), the FACS-isolated 3 cell reaction showing *msp-4* cDNA product, and the FACS-isolated 10 cell reaction showing *msp-4* cDNA product underwent whole transcriptome sequencing.

5.3.5 Whole transcriptome sequencing of material from low input and single parasite samples

Five of the reactions from the REPLI-g® protocol supplemented with rRNasin® were whole transcriptome sequenced. These consisted of one single cell reaction isolated by limiting dilution (1A, Figure 5.7A Lane 10), two single cells isolated by FACS (1D and 1E, Figure 5.7B Lanes 8 and 10), one reaction containing three cells isolated by FACS (1C, Figure 5.B Lane 5) and one reaction containing 10 cells isolated by FACS (1B, Figure 5.B Lane 4). Sequencing reads were mapped to the *P. falciparum* 3D7 genome. Alignment rates were generally good (Table 5.3), except for one of the FACS-isolated single cells (1D) which had an alignment rate of 2.6%, indicating that nearly 100% of reads did not belong to *P. falciparum* 3D7. MiniKraken was used to assess contamination in the newly sequenced transcriptomes (Table 5.4). As with the previously sequenced gDNA contaminated samples, bacterial species accounted for the bulk of microbial contamination, with Enterobacteria being the most highly represented. The samples had varying degrees of contamination with species that were found in the originally sequenced single cells. Single cell 1A had 17% of reads mapping to the Firmicutes phylum, which is represented in all of the *Plasmodium* samples but at much lower amounts. The single cell sample (1D) which had a very low *Plasmodium*

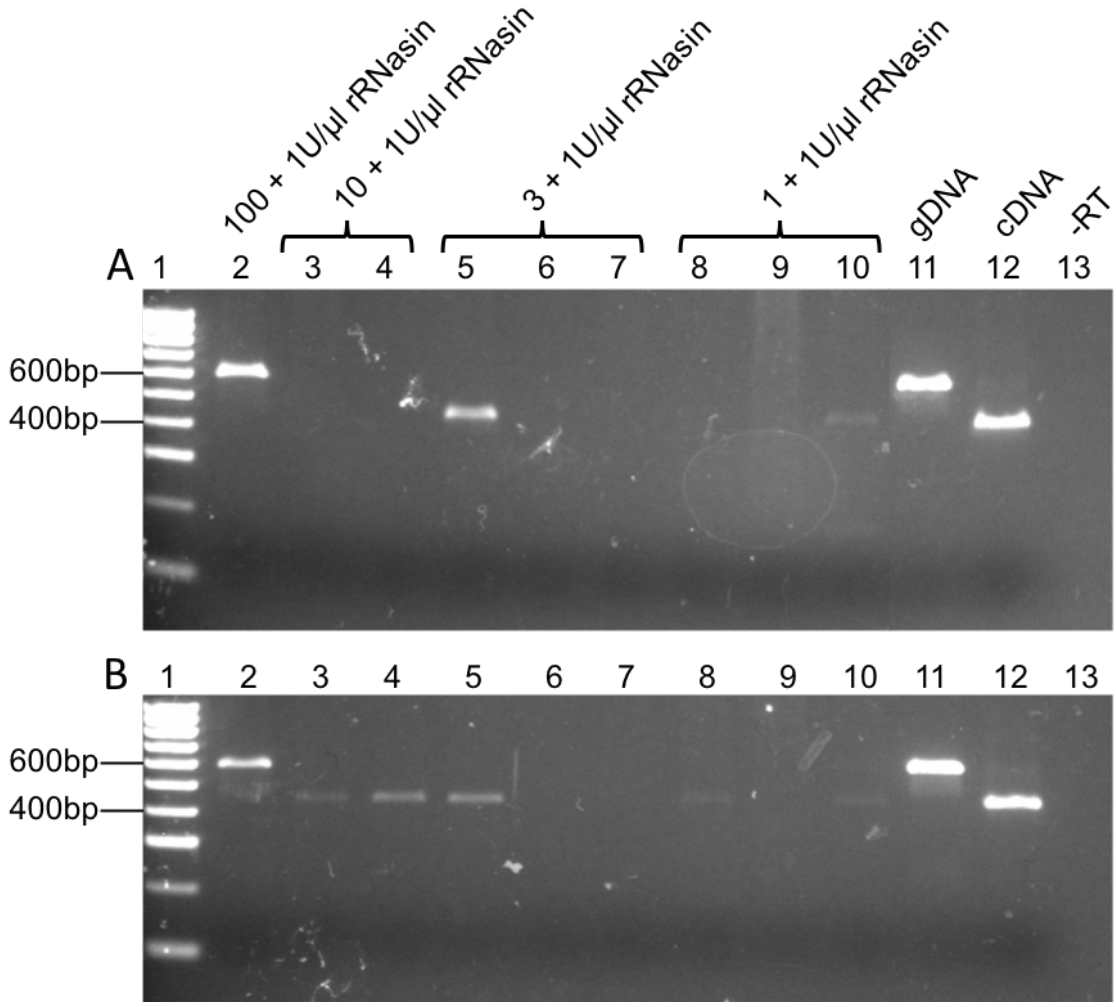

Figure 5.7. *msp-4* PCR carried out on samples containing 1 – 100 schizonts isolated by FACS and limiting dilution using the amended REPLI-g® protocol. So far, the REPLI-g® protocol alterations had been tested on large pools of 1000 parasites and on purified RNA. The altered protocol was carried out on reactions containing 100, 10, 3, or single parasites, and material was used in the *msp-4* PCR **A.** Parasite aliquots of 100, 10, 3, and single cells were obtained by serial dilution. Reactions underwent WTA with the addition of rRNasin®. One reaction was set up with 100 cells, and the *msp-4* PCR showed gDNA product (lane 2). Two reactions containing 10 cells failed (lanes 3 and 4). One out of three reactions containing 3 cells showed *msp-4* cDNA product (lane 5). Of three reactions containing a single cell, one produced a very weak *msp-4* cDNA product (lane 10). **B.** Parallel reactions set up using FACS to isolate cells. The 100 cell reaction showed *msp-4* gDNA (lane 2). The two 10 cell reactions and one 3 cell reaction showed cDNA products (lanes 3-5). Two single cells showed very weak *msp-4* cDNA product (lanes 8 and 10). Lanes 11-12 show the gDNA and cDNA controls. Lane 13 shows the negative control.

Table 5.3. Sequencing statistic from the *P. falciparum* 3D7 schizonts samples amplified using the amended REPLI-g® protocol, supplemented with an RNase inhibitor.

Sample	No. of cells	No. of reads (millions)	Alignment rate to 3D7 (%)	No. of genes with >0 reads	No. of genes with >10 reads	Isolation method
1A (Fig.5.7A, L10*)	1	10.4	37	1736	652	Dilution
1B (Fig.5.7B, L4*)	10	9.0	69.2	4422	2514	FACS
1C (Fig.5.7B, L5*)	3	6.9	64.7	2045	867	FACS
1D (Fig.5.7B, L8*)	1	5.3	2.6	807	117	FACS
1E (Fig.5.7B, L10*)	1	14.85	45.6	2076	889	FACS

One sample contained a schizont-infected erythrocyte isolated by limiting dilution, four samples (one with 10 infected erythrocytes, one with 3 and two with 1) were isolated by FACS. Alignment to the *P. falciparum* 3D7 reference genome ranged from 2% to 69%. * L indicates the electrophoresis lane showing the MSP-4 amplicon for each sample in Figure 5.7.

alignment rate (2.5%), had 5% of reads mapping to the microbial database, with human contamination being a possibility for the remainder of the reads.

To ensure that cDNA has indeed been captured and sequenced, the sequence data were viewed in Artemis. Several multi-exon genes were viewed, where intron-exon boundaries were clearly visible, confirming that the transcriptomes had indeed been sequenced (Figure 5.8). Sequence data obtained from a limiting number of cells are likely to show "patchy" coverage across the transcriptome due to the effect of random sampling error, particularly for low transcript number genes (Brennecke et al. 2013). To investigate the extent to which this impacted the single cell data, the WTA samples were compared to five schizont-specific *P. falciparum* 3D7 transcriptomes generated from bulk culture (prepared and sequenced by Dr. Sarah Tarr). Most genes (>4500) sequenced from the bulk cultures are detectable (with one or more mapped reads)

Table 5.4 Contamination present in sequence data from samples whole transcriptome amplified using the amended REPLI-g® protocol.

Taxonomic name and rank		Percentage of short reads mapping to the named taxa for each sample				
		1A (1 cell)	1B (100 cells)	1C (3 cells)	1D (1 cell)	1E (1 cell)
Bacteria	Domain	24.7	16.0	11.9	4.88	13.0
Proteobacteria	Phylum	4.79	14.8	9.22	4.05	9.10
…Gammaproteobacteria	Clade	4.62	13.6	8.73	3.81	9.03
…Betaproteobacteria	Clade	0.14	0.08	0.16	0.03	0.05
…Alphaproteobacteria	Clade	0.01	0.08	0.23	0.03	0.01
Actinobacteria	Phylum	0.40	0.01	0.26	0.00	0.50
Firmicutes	Phylum	18.1	0.66	0.52	0.03	0.19
…Bacilli	Clade	18.1	0.05	0.52	0.03	0.18
Viruses	Domain	0.15	0.22	0.35	0.20	0.14

The software Kraken was used to assign taxonomic labels to the short sequence reads that did not map to *P. falciparum* 3D7. A number of Phyla and Clades were represented at above 0.1% of short sequence reads per sample, which are listed above. The vast majority of contamination originated from bacteria. Other taxa were represented at lower levels.

(Figure 5.9A), and these whole transcriptomes are represented mostly by moderately expressed genes, with a small number having zero counts, and fewer being expressed at very high levels (Figure 5.9B).

The single and three cell samples had a far greater number of genes with no detectable expression (Figure 5.9A) and these represented the bulk of genes (Figure 5.9D and E). The single cell (1D) with very low alignment rate had the greatest number of genes with no mapped reads, and only had 117 genes with more than 10 mapped reads (Table 5.3), indicating that the number of detected genes is more likely to be a function of poor data sampling and sparseness, rather than true biological variation in which genes are expressed in a single cell or small subset of cells. The three cell and the other single cell transcriptomes had roughly equal numbers of genes with zero mapped reads and had 867 and 889 genes with more than 10 mapped reads, respectively (Table 5.3). The

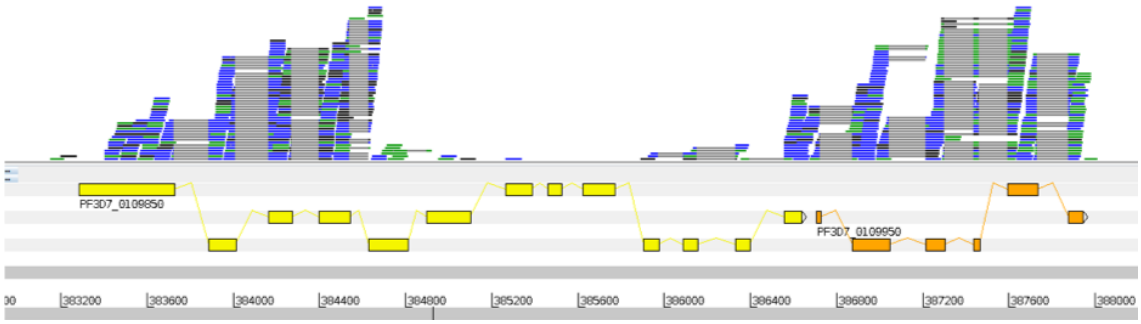

Figure 5.8. Sequence data from the samples whole transcriptome amplified using the amended REPLI-g® protocol mapping to multi-exon genes shown in the Artemis genome browser. The distribution of short sequencing reads pooled from reactions containing 10, 3, and single cells across two multi-exon genes show sharp boundaries at intron-exon boundaries, represented by the grey bars that signify when a short sequence read (blue and green bars) has been split across an intron. This indicates that sequence data has originated from RNA, rather than from genomic DNA, in contrast to the read distribution seen in Figure 5.1B from the first single cell experiment.

transcriptome from the ten pooled cells (1B) recapitulated more closely the number of detected genes seen in the five bulk transcriptomes and had 2514 genes with more than 10 mapped reads (Table 5.3), and perhaps indicates that this number of cells provided sufficient RNA to reduce the transcriptome-wide sparseness. The read distributions for the single, 3 cell and 10 cell samples are heavily skewed towards zero and low count read mapping (Figure 5.9C-E). While the 10 cell sample is shown to have slightly more genes with moderate numbers of read counts (Figure 5.9C), few genes have moderate of high read counts in the single cell transcriptomic sequence data (Figure 5.9D).

Further investigation into the distribution skew of read mapping revealed that while the library sizes for the single and low number samples were reasonably large (approximately $2x10^6$ reads or more), the vast majority of the reads mapped to

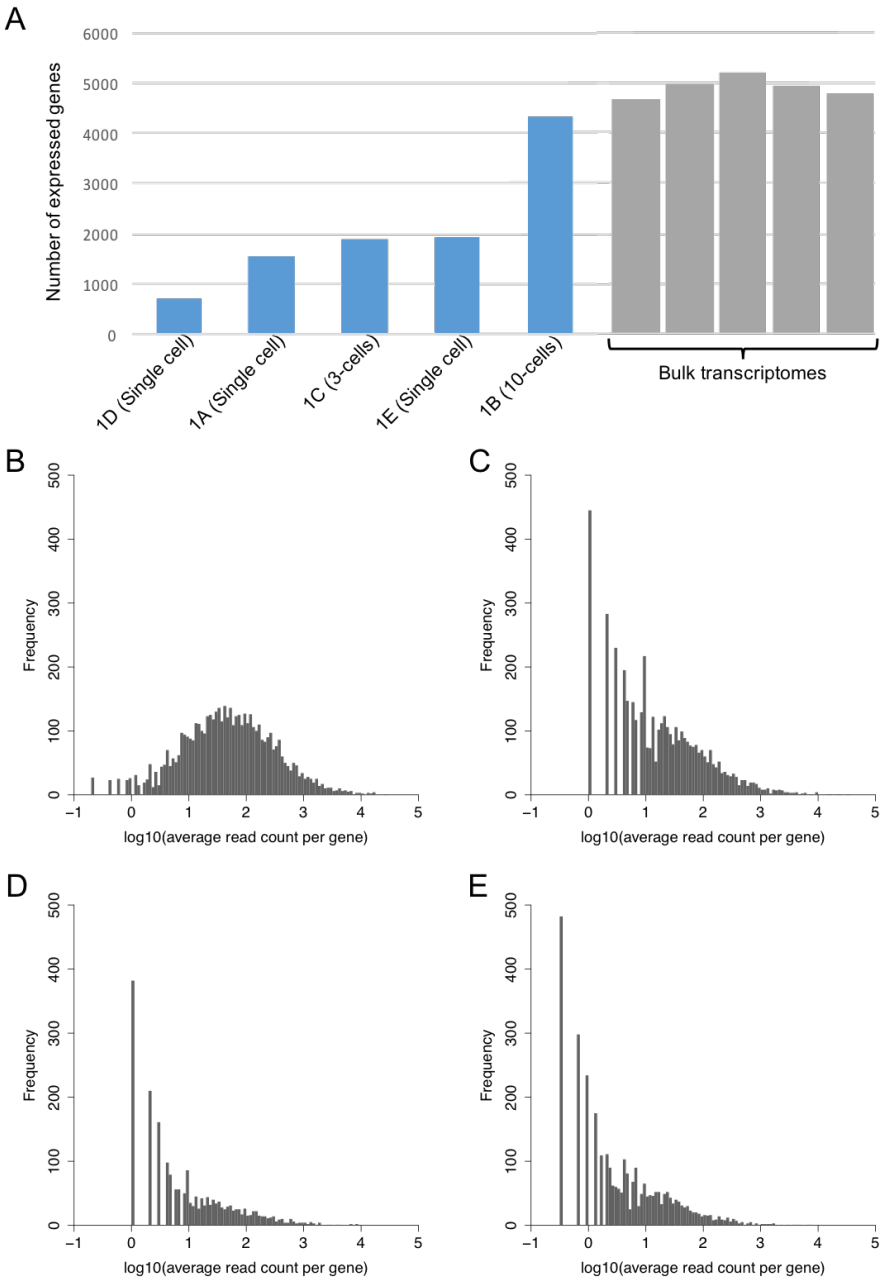

Figure 5.9. Technical graphs showing the number of detectable genes and distribution of read counts per gene in each of the whole transcriptome sequenced samples. Using the REPLI-g® protocol, whole transcriptomes were sequenced from three single *P. falciparum* schizonts, along with a transcriptome from three pooled cells, and one from ten pooled cells. **A.** The number of detectable genes (those with 1+ mapped reads) is substantially lower at lower cell numbers (blue bars) when compared to transcriptomes generated from bulk culture material (grey bars). The reaction containing ten cells contained numbers of detectable genes comparable to that of the bulk transcriptomes from schizonts previously prepared. **B-E.** The frequency distributions of number of reads per gene. Bulk transcriptomes (**B**) show an even distribution of gene expression levels, with the majority of genes being moderately expressed. The sample containing ten cells (**C**) showed some moderately expressed genes, however the majority of genes were either not detectable, or contained less than 5 mapped reads. The reactions containing three cells (**D**) and single cells (**E**) showed an even greater proportion of genes with zero mapped reads, and few genes with moderate expression.

ribosomal RNA genes rather than to protein-coding genes (Figure 5.10). 83% of reads from the ten pooled cells mapped to rRNA genes, 88% of reads from the three pooled cells samples mapped to rRNA genes, and between 94% and 96% of the reads from the three single cells mapped to rRNA genes, resulting in the number of reads mapping to protein coding genes ranging between 6,056 (single cell 1D) to 710,326 (ten pooled cells). By contrast, transcriptomes sequenced from bulk *P. falciparum* 3D7 schizonts had an average of only 14% of reads mapping to rRNA genes, ranging from 0.2%-35%.

5.3.6 The Fluidigm C1™ system for high-throughput acquisition of single cells

The commercially available microfluidic system, the C1™ (Fluidigm) was tested as an alternative method to flow sorting and use of the REPLI-g® kit. The C1™ handles cell isolation, reverse transcription and amplification on a microfluidic chip ("Integrated fluidic circuit") with minimal user input after the chip has been set up and utilises the SmartSeq® v4 template-switching PCR-based commercially available single cell kit (Clontech). Amplified material from the C1™ can then be prepared for whole transcriptome sequencing. The microfluidic chip used contained 96 "capture sites" through which the cell suspension flows and into which cells are isolated, ideally one per capture site. To test the C1™ system, *P. falciparum* 3D7 schizonts were purified and fixed with DSP. Schizonts were stained with Vybrant® DyeCycle™ Green to fluoresce DNA, and assessed microscopically after fixation to ensure that erythrocytes were intact and undamaged (Figure 5.11). The schizont suspension was subsequently loaded onto the C1™ microfluidic chip and after cell capture had taken place each of the capture sites was viewed microscopically to assess the state of each capture site. Of the 96 sites, 33 contained a single schizont, six contained a single cell plus unidentifiable "debris", 19 contained more than one cell (one site contained five cells), and 37 sites were empty.

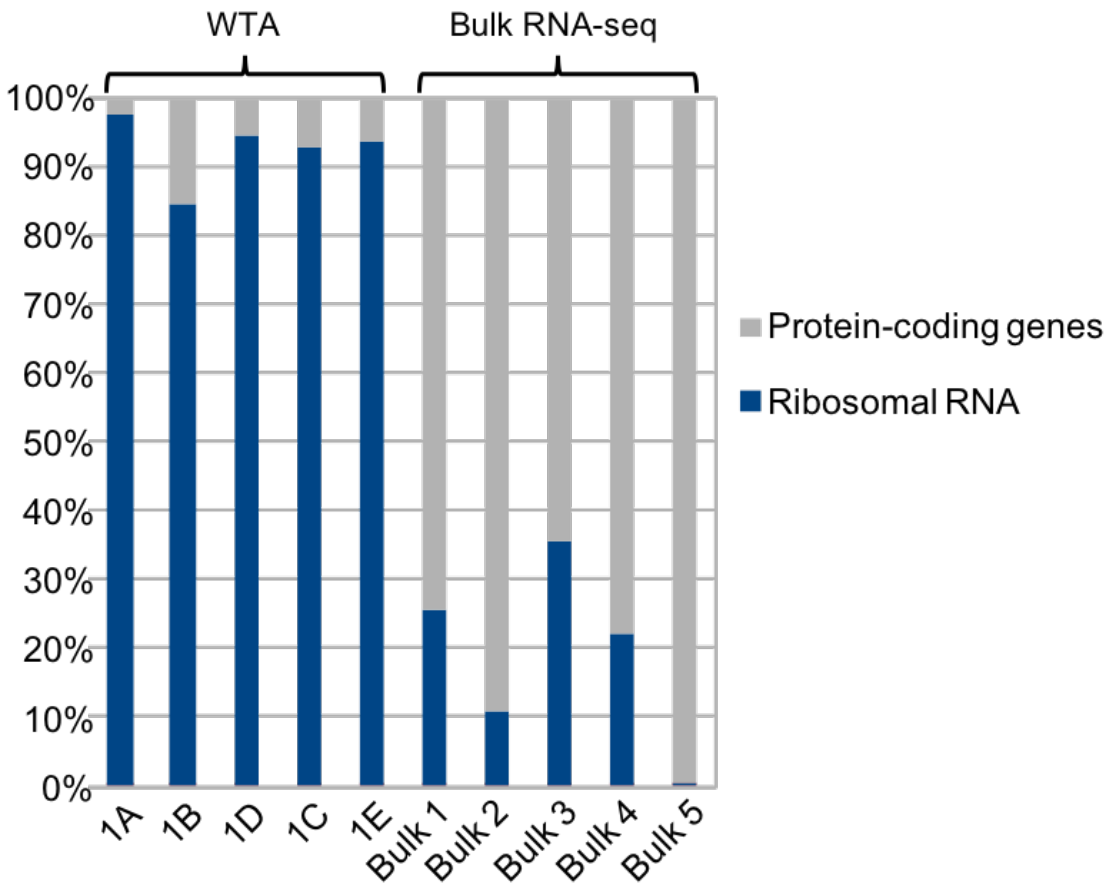

Figure 5.10. The proportion of reads mapping to *P. falciparum* genes encoding rRNA and protein-coding genes for the whole transcriptome sequence samples, compared to bulk RNA-seq data. RNA-seq data from five transcriptomes obtained from *P. falciparum* 3D7 bulk preparations (Bulk 1 – 5) showed that an average of 14% of reads mapped to genes encoding rRNAs. The three single cells (1A, 1D and 1E), the three pooled cells (1C) and the ten pooled cells (1B) that were prepared by whole transcriptome amplification (WTA) using the amended REPLI-g® protocol by comparison contained between 83% and 96% of reads mapping to rRNA genes.

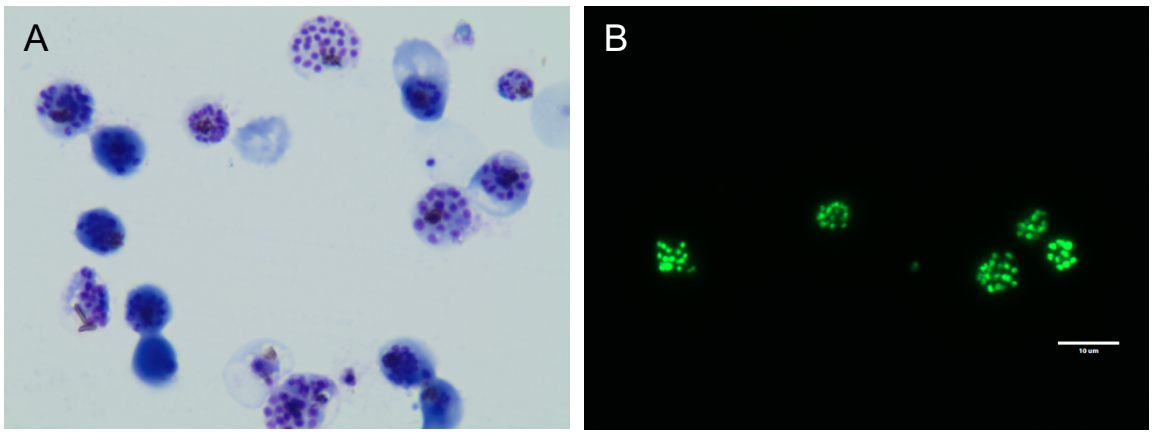

Figure 5.11. DSP fixation of *P. falciparum* 3D7 schizonts. The Fluidigm C1™ platform was tested for single-cell RNA-seq. *P. falciparum* schizonts were isolated and fixed with DSP, following tested protocols (Attar et al. 2018), stained with Vybrant® DyeCyle™ Green DNA dye, before being loaded onto the C1 microfluidic chip. **A.** After DSP fixation, schizont-infected erythrocytes remained intact and retained their expected morphology. **B.** The Vybrant® DyeCyle™ Green DNA dye was used to stain the schizonts to allow viewing of cells after isolation on the microfluidic chip, showing individual merozoites within the fixed schizonts.

The average concentration of DNA per single cell was 0.8 ng/ul. Quality of DNA from the C1™ samples was assessed on the Bioanalyzer, to evaluate suitability of samples for whole transcriptome sequencing. Two distinct features were apparent among the Bioanalyzer trace. The first is a small (~40bp) peak of low intensity which can be seen just adjacent to the lower marker (Figure 5.12A), either in isolation, or along with a large peak, between ~7kb and ~10kb in size, often overlapping with the upper marker (Figure 5.12B). The expected distribution of fragment sizes for material generate by the C1™ is ~ 600bp - ~2000bp (Figure 5.12C) and the SmartSeq® v4 chemistry is not expected to result in very long, ligated cDNA fragments (unlike the previously used MDA REPLI-g® method) and so the *msp-4* intron-spanning PCR was used to determine whether the large DNA fragments represent gDNA. The PCR revealed that out of the 33 single cells, 15 were positive for *msp-4* gDNA, the remaining 18 did not

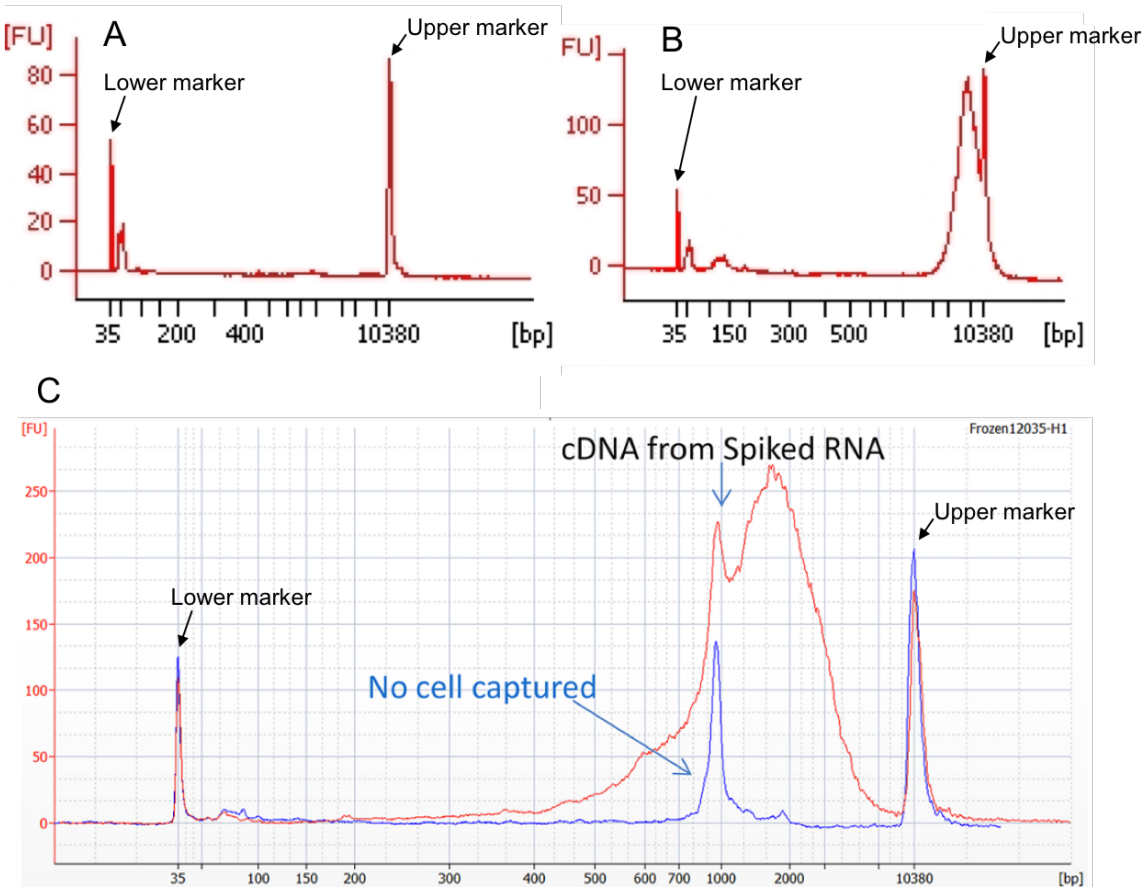

Figure 5.12. Bioanalyzer traces showing DNA fragments obtained using the C1™ platform. Samples containing cells isolated using the Fluidigm C1™ had cDNA run on a Bioanalyzer to determine their suitability for whole transcriptome sequencing. Several samples did not show any peaks on the BioAnalyzer. Those that did have peaks showed one of two patterns. **A.** Some samples contained a single, low intensity peak of low molecular weight (~40bp), believed to be cDNA originating from heavily degraded RNA. **B.** Some samples showed the ~40bp peak, along with a high intensity, high molecular weight peak representing fragments approximately 7kb-10kb in size. It is thought that this peak originated from *P. falciparum* genomic DNA. **C.** The expected Bioanalyzer trace from cDNA successfully amplified from single cells (spiked with purified RNA) using the Fluidigm C1™, showing the expected peak size distribution of ~600 - ~2000bp (Source: "Using C1 to generate Single-Cell cDNA libraries for mRNA sequencing" manual by Fluidigm).

have any product. Without confirmation, it is likely that the small peak on the Bioanalyzer at ~40bp (Figure 5.12A) represents very short cDNA fragments originating from highly degraded RNA, and that the large peak represents amplified gDNA.

5.4 Discussion

In this chapter, methods for whole transcriptome amplification and sequencing from *P. falciparum* single schizonts were tested, resulting in the sequencing of three single cell transcriptomes after numerous method optimisations. Multiple displacement amplification was used to amplify cDNA obtained from RNA in single cell and low number cells isolated by FACS or by serial dilution. Initial experiments revealed that the amplification had failed to capture any cDNA, and instead the sequence data originated from *P. falciparum* genomic DNA. Following a series of optimisation steps, the addition of an RNase inhibitor allowed the amplification of cDNA from three single cells, one sample containing three cells, and one sample containing 10 cells. These samples underwent whole transcriptome sequencing. Comparison to transcriptomes obtained from bulk parasite culture exposed the level of sparseness seen in data belonging to very low cell number reactions. Among the genes with non-zero read counts, the vast majority had less than 10 mapped reads. These data illustrate the technical challenges that need to be overcome before single cell transcriptomes can reliably represent true biological differences.

Contamination of reactions with non-target DNA was minimised by working within PCR-free environments, with pre-amplification steps being carried out in a laminar flow hood, cleaned and UV-sterilised before use. However, of the short reads sequenced, alignment to the *P. falciparum* 3D7 genome varied from 2% to 70% of

reads; with the single cell reactions aligning on the lower end of this scale (2% to 45%), indicating that despite the precautions taken, contamination still occurred to a significant level. This level of contamination within single-cell work is not unprecedented (Reid et al. 2018). Bacterial species made up the vast majority of contamination in all samples. The exact composition of the contamination varied between samples, indicating that it occurred stochastically, rather than representing a systematic contamination problem. However there were some common contaminants, and these belonged to the Actinobacteria and Firmicutes phyla and the Enterobacteriacae family, species of which have previously been identified as common sequencing contaminants; the presence and volume of which was noted to be compounded by low concentration of the intended input material (Salter et al. 2014).

Ribosomal RNA makes up around 90% of all total cellular RNA, which is problematic for transcriptomics. Commercial next-generation sequencing kits typically contain an rRNA depletion step prior to reverse transcription, designed to enrich for mRNA and ensure that it is these that make up the bulk of sequencing data. However, these protocols are not suited for such low concentrations of RNA found in single cells without significant depletion of mRNA molecules (Fang and Akinci-Tolun 2016). High levels of rRNA contamination in single cell sequencing has been reported previously both in *Plasmodium* and also in other single cell organisms, comprising up to 90% of sequencing reads (Kolisko et al. 2014; Reid et al. 2018). These studies utilised FACS followed by application of a single cell transcriptomic sequencing method known as Smart-seq2 (Picelli et al. 2014). A separate study used a different technique called Drop-seq; a droplet based method of single cell RNA-seq, which incorporated a tRNA and rRNA depletion step, resulting in less than 5% of reads on average mapping to these RNAs in the sequence data (Poran et al. 2017). This method resulted in the

detection of far fewer protein-coding genes (~650) compared to experiments using an optimised Smart-seq2 protocol (~1900), although another study utilising the Smart-seq2 protocol detected only around 300-800 genes per parasite (Ngara et al. 2018; Reid et al. 2018).

Another issue needing to be addressed was the amplification of material from genomic DNA, rather than from RNA. It is thought that this resulted from a combination of RNA degradation prior to reverse transcription, and an insufficient DNase step, with the genomic composition of *P. falciparum* almost certainly being a compounding factor. Genomic G + C content is typically around 30% - 50% for prokaryotes and eukaryotes (Li and Du 2014). *P. falciparum* has an extremely rich AT genome, comprising around 80% of DNA (up to 90% in intronic and intergenic regions), which is unusually high even for *Plasmodium* species generally (Figure 1.5) (Nikbakht et al. 2015). The reverse transcription step of the REPLI-g® protocol utilises an oligo(dT) primer, designed to bind to the poly(A) tails of mRNA molecules; precluding amplification from gDNA (or from tRNA and rRNA molecules). However, the extreme AT-richness of the *P. falciparum* genome is probably sufficient to allow for some (however inefficient) binding of the oligo(dT) primers in genomic DNA. The combination of all of these factors could add up to allow the massive amplification of gDNA to the exclusion of cDNA seen in the REPLI-g® optimisations, despite the relatively low level of gDNA in a schizont compared to mRNA molecules.

The remaining sequencing reads that did map to protein-coding genes allowed detection of a variable number of genes per sample; within the expectations of previous studies (Poran et al. 2017; Ngara et al. 2018; Reid et al. 2018). Expectedly, as the number of cells per sample was increased to 10, a higher proportion of the genome's complement

of genes were detectable. The samples containing single or three cells had far fewer detectable genes, which appears to be a result or poor representation due to the technical bias experienced, with the highly contaminated single cell having less than 1000 genes with one or more reads. However, detectable does not mean reliably expressed. In all the samples, almost all of the detected genes were represented by less than 10 mapped read, below the technical threshold for reliable "expression". In previous studies, these low detectable gene numbers have none-the-less been used to accurately assign cells to specific life stages, indicating that biological information is retained, even when sequence data is incomplete (Poran et al. 2017; Ngara et al. 2018; Reid et al. 2018).

The REPLI-g® protocol uses MDA chemistry and a high-fidelity Φ29 polymerase with an extremely low error rate, making it ideal for repeated amplification of cDNA from reverse transcribed RNA. An alternative, and arguably more popular, method for single cell transcriptomics is the aforementioned Smart-seq2 chemistry, which uses PCR amplification and a template-switching reverse transcriptase (Picelli et al. 2014). The high-throughput Fluidigm C1™ utilises microfluidics and the commercially available SmartSeq® v4 reagents (Clontech) to prepare cells for single-cell transcriptomics. This is the same reagent kit used in Chapter 4 to capture transcriptomes from low input clinical isolates. The C1™ is capable of capturing 96 single cells on one microfluidic chip, however only 1/3 of these "capture sites" contained a single parasitised erythrocyte, and none of these generated cDNA, indicating that the standard protocol would need adjustment akin to that of the Smart-seq2 protocol optimised by for use with *P. falciparum* cells (Reid et al. 2018) before being used successfully. Parasites used on the C1™ were preserved with DSP, use of which has been tested on single cell destined for transcriptomics on the C1 platform (Attar et al. 2018). As found with the *Plasmodium* schizonts, the DSP fixation was found not to impact cell morphology and

these were found to be captured by the C1™ microfluidics similarly to fresh cells. Most importantly, the cDNA profiles from fixed and fresh cells were found to be comparable, although the yield of cDNA was less for the fixed cells. This did not seem to impact the overall transcriptomic data obtained, with library complexity being maintained from fresh to fixed cells, indicating the DSP is most likely a safe fixation protocol to use to preserve the RNA in single cells. The *Plasmodium* parasites captured on the C1™ all failed to produce cDNA and roughly half produced gDNA and it seems unlikely that this is due to the DSP fixation process, and instead indicates that some modifications to the SmartSeq® v4 protocol are needed for this system.

Despite efforts to optimise the REPLI-g® protocol, because the SmartSeq® v4 protocol was successful in Chapter 4, future work would be best directed using this approach, although utilising the non-commercial Smart-seq2 alternative would be preferable due to increased flexibility of integrating and optimising suggested protocols. Isolating cells by FACS would allow targeted separation of cells with a specific gene tagged, and presents an advantage over the use of the Fluidigm C1™ microfluidic system. It is likely that some optimisations would still be necessary when using the Smart-seq2 protocol, and these can be guided by knowledge gathered through the REPLI-g® optimisation, along with those made in other *Plasmodium* single-cell transcriptomic studies (Poran et al. 2017; Ngara et al. 2018; Reid et al. 2018).

In Chapter 4, the commercial Smart-Seq protocol was used to successfully obtain transcriptomes originating from limiting material in *P. falciparum* clinical isolates. Transcriptomic differences relating to the MSPDBL2 schizont expression phenotype were detectable in these samples. As MSPDBL2 is expressed in a small subset of schizonts, single cell sequencing would be a good avenue to pursue. If a *P. falciparum*

parasite line was developed in which *mspdbl2* is tagged with GFP, single MSDBL2-expressing parasites could be sorted by FACS and undergo single-cell transcriptomic analysis. This would provide a population of single parasites whose transcriptomes can be compared to non-MSPDBL2 expressing parasites of the same background to identify differentially expressed genes between the two populations of cells.

6 General discussion and concluding remarks

Whole genome and transcriptome sequencing have been used throughout this thesis to address questions relating to *Plasmodium* populations and parasite biology. In addition, research undertaken has demonstrated both the power and limitations of working with clinical isolates of *Plasmodium*, for which there is often a very limited amount of available material. Chapters 4 and 5 in particular have addressed the very real technical limitations that are an integral part of working with very limited, low input material. In Chapter 3, sampling was carried out in peninsular Malaysia for clinical isolates of *P. knowlesi*, and 53% of samples collected contained sufficient genomic DNA to proceed with whole genome sequencing using the previously defined protocol. In Chapter 4, *P. falciparum* clinical isolates were thawed and cultured ex vivo, and only 17% of the samples yielded sufficient material for immunofluorescence assays and RNA-seq. Chapter 5 went further to directly test some of the techniques currently available for obtaining reliable sequence data from very low input samples containing just one or very few cells and to address the challenges faced when working with material of this nature, the methods tested produced single cell transcriptomes, however further optimisation or testing of alternative techniques is needed to produce high-quality data. The research presented in this thesis should go some way to contributing to our biological understanding of malaria parasites, while also highlighting the continuing need for development and innovation of the processing and sequencing of low input and clinical samples.

Recognition of *P. knowlesi* as a major cause of human malaria in Malaysia has encouraged efforts towards understanding the genomics of this species with respect to the human host, and importantly how to develop strategies to limit its impact (Cox-

Singh et al. 2008). Population genomics has played a role in this, firstly through the use of multi-locus microsatellite analysis, and more recently by whole genome sequencing (Divis et al. 2015; Divis et al. 2018). Whole genome sequencing provides a far greater amount of data than previous approaches and has been used to further investigate and deep diversity of the sympatric populations of *P. knowlesi* that co-exist in Malaysian Borneo (Divis et al. 2018). Thus far, the population genomics of *P. knowlesi* in peninsular Malaysia (and also throughout the rest of Southeast Asia) has remained limited, mainly based on the whole genome sequences of a few old laboratory lines from Southeast Asia which indicted the presence of a third divergence cluster, separate to those of Malaysian Borneo. Microsatellite analysis of *P. knowlesi* from humans and macaques from peninsular Malaysia has previously confirmed the existence of this third cluster (Assefa et al. 2015; Divis et al. 2017) and single loci genomic analyses from *P. knowlesi* in other countries in Southeast Asia have provided evidence that Cluster 3 is found beyond peninsular Malaysia (Ahmed et al. 2018).

The research carried out in Chapter 3 has been the first deep genomic sequencing of *P. knowlesi* parasites collected throughout peninsular Malaysia from hospital patients presenting with malaria and identified as *P. knowlesi*. As was expected from previous research, these samples showed clear divergence from the sympatric Clusters 1 and 2 in Malaysian Borneo, evidencing ancient population subdivision. This division possibly stems from the geographic isolation of Malaysian Borneo from mainland Southeast Asia when the last ice age ended around 13,000 years ago, although the divergence could predate this (Lee et al. 2011). What was of particular interest was evidence of some subdivision among the Cluster 3 samples, which indicates that there is possible selection driving divergence between the parasites, the nature of which remains unknown, and warrant further investigation. Reports of *P. knowlesi* malaria are now

commonplace throughout Southeast Asia, where its reservoir hosts are distributed and human *P. knowlesi* infections have now been reported in all countries in Southeast Asia. Sampling for clinical isolates should be undertaken not only in peninsular Malaysia, but in other countries as well, and obtaining sufficient material for whole genome sequencing should be matter of course. With divergence seen within Cluster 3 parasites from peninsular Malaysia, the genomic analyses of parasites from further into the Southeast Asian mainland and from other island countries such as Indonesia and the Philippines should prove extremely interesting. A wider repertoire of genomes would be very useful for determining the drivers of divergence for these zoonotic parasites, particularly with the ongoing risk of adaptation to human infection and the development of human-to-human transmission.

An on-going 'blind spot' with *P. knowlesi* research is the relative lack of available material from the wild macaque reservoir hosts, and also from the wild vectors. Sampling from the blood of wild macaques in Malaysian Borneo followed by microsatellite analysis led to the discovery that the divergence between the sympatric Clusters 1 and 2 had been driven by the preferential infection of different macaque species by each cluster (Divis et al. 2015), although the biology underpinning this is still not understood. Sampling in peninsular Malaysia confirmed that Cluster 3 was found in both human and wild macaques, and that there were a subset of macaque infections which appeared to have intermediate assignments between Clusters 2 and 3 (Divis et al. 2017). At this time, no whole genomes from wild macaque-sampled *P. knowlesi* have been sequenced. Obtaining usable samples from wild macaques that are suitable for whole genome sequencing is difficult as they often have low parasitaemia, multi-genotype, multi-species infections which would make direct bulk sequencing of *P. knowlesi* parasites highly inefficient. A solution for tackling multi-species infections

would be to utilise single-cell techniques using FACS to isolate *P. knowlesi*-infected erythrocytes through the use of antibodies specific to a *P. knowlesi* protein and sequencing from single parasites. This would be technically challenging, though single-cell genomics is generally more straight forward than the single-cell transcriptomics attempted in Chapter 5 as single-cell genomics precludes the need for preserving and accurately amplifying highly unstable RNA molecules. Nonetheless, it is hoped that the methods trialled in Chapter 5 will not just be useful for investigating *P. falciparum*, but will also be applicable for *P. knowlesi*, in order to streamline single cell investigations in this species.

Study of clinical isolates of *P. falciparum* here focused on schizont transcriptomes correlated with expression of *mspdbl2*, which may be an early marker of commitment to gametocytogenesis. These samples contained very small amounts of infected red blood cells and parasites were developed *ex vivo* until most parasites had gained maturity and were harvested prior to reinvasion. The process of culture adaptation presents significant selection pressures on clinical isolates and there are detectable changes in transcriptomics and phenotypes of parasites adapted to culture (LeRoux et al. 2009; Lapp et al. 2015). In addition, long-term culture has been shown to result in the accumulation of genomic loss-of-function mutations, often affecting the gametocytogenesis pathway (Claessens et al. 2017). The genetics and transcriptomics underlying gametocytogenesis are complex. The genes *gdv-1* and *ap2-g* have been identified as being required for gametocytogenesis, and mutations in these genes are responsible for loss of gametocytogenesis in laboratory lines and for the spontaneous loss in culture adapted clinical isolates (Eksi et al. 2012; Kafsack et al. 2014; Claessens et al. 2017). In addition, allelic divergence and structural dimorphism has been identified at the *gdv-1* locus, and it is not yet known how this affects expression of *gdv-*

1 and other genes (Duffy et al. 2018). Studies utilising over-expression and knockout of *gdv-1* and *ap2-g* have begun to unravel the gene expression pathways necessary for gametocyte commitment to occur. However, further research – not just focused on laboratory lines, but also using clinical isolates – is needed to gain a clear picture of the gametocytogenesis pathway from a genomic and transcriptomic perspective (Filarsky et al. 2018; Josling et al. 2019; Usui et al. 2019). Due to the accumulation of genetic mutations, specifically affecting gametocytogenesis, and the preceding changes in environmental selection pressure caused by long-term *in vitro* culture, short-term *ex vivo* culture of clinical isolates is preferable to obtain transcriptomic profiles from parasites which have been minimally affected by the *in vitro* culturing process, even if these approaches do present technical difficulties and limitations.

Analysis of gene expression across the *P. falciparum* transcriptome based on MSPDBL2 protein expression in every case identified *mspdbl2* to be the gene with the most different expression, indicating that the approaches used were reliably identifying transcriptional changes despite the limited amount of starting material. The technical limits beyond using *ex vivo* clinical isolates, in which there may be millions of parasites present in a sample, are still being explored. Single-cell transcriptomics have been applied to malaria parasites over the past few years, leading to biological inference for processes such as virulence and gametocytogenesis, however these approaches still rely on the sequencing of hundreds or thousands of individual cells to generate enough transcriptome information from these cells (Poran et al. 2017; Ngara et al. 2018; Reid et al. 2018).

As seen in Chapter 5, sequencing of the mRNA molecules from single cells is still a factor of random sampling error and the stochastic nature of reverse transcription and

amplification, seen by the low number of genes with more than ten mapped reads. Previous evidence has shown how the recovery of mRNA molecules in single-cell transcriptomics is a factor of the amount of starting transcript levels. This makes it difficult to work with a small unicellular organism within which exists only a small number of mRNA molecules when compared to mammalian cells, such that only the genes with the highest numbers of transcripts can be very accurately quantified in individual cells (Brennecke et al. 2013). This in itself is a limitation, but not an absolute barrier and single-cell transcriptomics presents one of the best avenues of continued investigation into the work carried out in Chapter 4. For the clinical isolates investigated in Chapter 4, expression of the MSPDBL2 protein was usually only in a small proportion of parasites, and the transcriptomic profiles of these parasites therefore are mostly contributed to by *mspdbl2* non-expressing parasites. A targeted single-cell approach guided by developments tested in Chapter 5 could allow further investigation into *mspdbl2,* in particular to test its role in parasite differentiation.

Concluding remarks

Diverse approaches have been carried out utilising whole genome and transcriptome sequencing with an emphasis here on the study of clinical isolates, many of which contain limiting numbers of parasites and amounts of nucleic acid. Whole genomes of *P. knowlesi* in clinical isolates have been sequenced and used to investigate diversity and selection. Identifying drivers of selection within these parasites will be crucial to understanding the transmission dynamics affecting *P. knowlesi*, particularly as the risk of human infection may be increasing.

In addition, low input techniques have been implemented to investigate how *mspdbl2* may be involved in gametocytogenesis in *P. falciparum*. With the current lack of safe and effective drugs that target gametocytes, increasing our understanding of the genetic

networks involved in this process should be central to driving development of new therapies that could also block transmission. Continued development and innovation in these low input approaches will prove crucial to future research into clinical material.

7 Bibliography

Abdul-Ghani, R., Basco, L.K., Beier, J.C. and Mahdy, M.A.K. 2015. Inclusion of gametocyte parameters in anti-malarial drug efficacy studies: filling a neglected gap needed for malaria elimination. *Malaria Journal* 14, 413. doi: 10.1186/s12936-015-0936-4.

Adams, T., Ennuson, N.A.A., Quashie, N.B., Futagbi, G., Matrevi, S., Hagan, O.C.K., Abuaku, B., Koram, K.A., et al. 2018. Prevalence of *Plasmodium falciparum* delayed clearance associated polymorphisms in adaptor protein complex 2 mu subunit (*pfap2mu*) and ubiquitin specific protease 1 (*pfubp1*) genes in Ghanaian isolates. *Parasites and Vectors* 11, 175. doi: 10.1186/s13071-018-2762-3.

Ahmed, M.A., Chu, K.B., Vythilingam, I. and Quan, F.S. 2018. Within-population genetic diversity and population structure of *Plasmodium knowlesi* merozoite surface protein 1 gene from geographically distinct regions of Malaysia and Thailand. *Malaria Journal* 17, 442. doi: 10.1186/s12936-018-2583-z.

Al-Khedery, B., Barnwell, J.W. and Galinski, M.R. 1999. Antigenic variation in malaria: A 3' genomic alteration associated with the expression of a *P. knowlesi* variant antigen. *Molecular Cell* 3, pp. 131–141. doi: 10.1016/S1097-2765(00)80304-4.

Amambua-Ngwa, A., Tetteh, K.K.A., Manske, M., Gomez-Escobar, N., Stewart, L.B., Deerhake, M.E., Cheeseman, I.H., Newbold, C.I., et al. 2012. Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genetics* 8. doi: 10.1371/journal.pgen.1002992.

Amit-Avraham, I., Pozner, G., Eshar, S., Fastman, Y., Kolevzon, N., Yavin, E. and Dzikowski, R. 2015. Antisense long noncoding RNAs regulate var gene activation in the malaria parasite *Plasmodium falciparum*. *Proceedings of the National Academy of Sciences* 112, pp. 982–991. doi: 10.1073/pnas.1420855112.

Anders, S., Pyl, P.T. and Huber, W. 2015. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, pp. 166–169. doi: 10.1093/bioinformatics/btu638.

Anthony, T.G., Conway, D.J., Cox-Singh, J., Matusop, A., Ratnam, S., Shamsul, S. and Singh, B. 2005. Fragmented Population Structure of *Plasmodium falciparum* in a Region of Declining Endemicity. *The Journal of Infectious Diseases* 191, pp. 1558–1564. doi: 10.1086/429338.

Antony, H.A., Pathak, V., Parija, S.C., Ghosh, K. and Bhattacherjee, A. 2016. Whole transcriptome expression analysis and comparison of two different strains of *Plasmodium falciparum* using RNA-Seq. *Genomics Data* 8, pp. 110–112. doi: 10.1016/j.gdata.2016.04.004.

Ariey, F., Fandeur, T., Durand, R., Randrianarivelojosia, M., Jambou, R., Legrand, E., Ekala, M.T., Bouchier, C., et al. 2006. Invasion of Africa by a single *pfcrt* allele of South East Asian type. *Malaria Journal* 5, 34. doi: 10.1186/1475-2875-5-34.

Ariey, F., Witkowski, B., Amaratunga, C., Beghain, J., Langlois, A.C., Khim, N., Kim, S., Duru, V., et al. 2014. A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature* 505, 50. doi: 10.1038/nature12876.

Ashley, E.A., Dhorda, M., Fairhurst, R.M., Amaratunga, C., Lim, P., Suon, S., Sreng,

S., Anderson, J.M., et al. 2014. Spread of Artemisinin Resistance in *Plasmodium falciparum* Malaria. *New England Journal of Medicine* 371, pp. 441–423. doi: 10.1056/NEJMoa1314981.

Ashley, E.A. and Phyo, A.P. 2018. Drugs in Development for Malaria. *Drugs* 9, pp. 861–879. doi: 10.1007/s40265-018-0911-9.

Assefa, S., Lim, C., Preston, M.D., Duffy, C.W., Nair, M.B., Adroub, S. a, Kadir, K. a, Goldberg, J.M., et al. 2015. Population genomic structure and adaptation in the zoonotic malaria parasite *Plasmodium knowlesi*. *Proceedings of the National Academy of Sciences of the United States of America* 112, pp. 13027–13032. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26438871.

Attar, M., Sharma, E., Li, S., Bryer, C., Cubitt, L., Broxholme, J., Lockstone, H., Kinchen, J., et al. 2018. A practical solution for preserving single cells for RNA sequencing. *Scientific Reports* 8, 2151. doi: 10.1038/s41598-018-20372-7.

Auburn, S., Benavente, E.D., Miotto, O., Pearson, R.D., Amato, R., Grigg, M.J., Barber, B.E., William, T., et al. 2018. Genomic analysis of a pre-elimination Malaysian *Plasmodium vivax* population reveals selective pressures and changing transmission dynamics. *Nature Communications* 9, 2585. doi: 10.1038/s41467-018-04965-4.

Auburn, S., Campino, S., Miotto, O., Djimde, A.A., Zongo, I., Manske, M., Maslen, G., Mangano, V., et al. 2012. Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. *PLoS ONE* 7. doi: 10.1371/journal.pone.0032891.

Awandare, G.A., Nyarko, P.B., Aniweh, Y., Ayivor-Djanie, R. and Stoute, J.A. 2018. *Plasmodium falciparum* strains spontaneously switch invasion phenotype in suspension culture. *Scientific Reports* 8, 5782. doi: 10.1038/s41598-018-24218-0.

Bainbridge, M.N., Warren, R.L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., et al. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* 7, 246. doi: 10.1186/1471-2164-7-246.

Bancells, C., Llorà-Batlle, O., Poran, A., Nötzel, C., Rovira-Graells, N., Elemento, O., Kafsack, B.F.C. and Cortés, A. 2019. Revisiting the initial steps of sexual development in the malaria parasite *Plasmodium falciparum*. *Nature Microbiology* 4, pp. 144–154. doi: 10.1038/s41564-018-0291-7.

Barber, B.E., William, T., Jikal, M., Jilip, J., Dhararaj, P., Menon, J., Yeo, T.W. and Anstey, N.M. 2011. *Plasmodium knowlesi* malaria in children. *Emerging infectious diseases* 17, pp. 814–820. doi: 10.3201/eid1705.101489.

Battle, K.E., Lucas, T.C.D., Nguyen, M., Howes, R.E., Nandi, A.K., Twohig, K.A., Pfeffer, D.A., Cameron, E., et al. 2019. Mapping the global endemicity and clinical burden of *Plasmodium vivax*, 2000–17: a spatial and temporal modelling study. *The Lancet* 394, pp. 332–343. doi: 10.1016/s0140-6736(19)31096-7.

Bhatt, S., Weiss, D.J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C.L., et al. 2015. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature* 526, 207. doi: 10.1038/nature15535.

Boone, D.R., Sell, S.L. and Hellmich, H.L. 2013. Laser capture microdissection of enriched populations of neurons or single neurons for gene expression analysis after traumatic brain injury. *Journal of Visualized Experiments* 74, e50308. doi:

10.3791/50308.

Boyle, M.J., Wilson, D.W., Richards, J.S., Riglar, D.T., Tetteh, K.K.A., Conway, D.J., Ralph, S.A., Baum, J., et al. 2010. Isolation of viable *Plasmodium falciparum* merozoites to define erythrocyte invasion events and advance vaccine and drug development. *Proceedings of the National Academy of Sciences* 107, pp. 14378–14383. doi: 10.1073/pnas.1009198107.

Bozdech, Z., Llinás, M., Pulliam, B.L., Wong, E.D., Zhu, J. and DeRisi, J.L. 2003a. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biology* 1, e5. doi: 10.1371/journal.pbio.0000005.

Bozdech, Z., Zhu, J., Joachimiak, M., Cohen, F., Pulliam, B. and DeRisi, J. 2003b. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biology* 4, R9. Available at: http://genomebiology.com/2003/4/2/R9.

Brady, G., Barbara, M. and Iscove, N.N. 1990. Representative in Vitro cDNA Amplification From Individual Hemopoietic Cells and Colonies. *Methods in Molecular and Cell Biology* 2, pp. 17–25.

Brancucci, N.M.B., Bertschi, N.L., Zhu, L., Niederwieser, I., Chin, W.H., Wampfler, R., Freymond, C., Rottmann, M., et al. 2014. Heterochromatin protein 1 secures survival and transmission of malaria parasites. *Cell Host and Microbe* 16, pp. 165–176. doi: 10.1016/j.chom.2014.07.004.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., et al. 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* 8, pp. 207–211. doi: 10.1038/nmeth.2645.

Broadbent, K.M., Broadbent, J.C., Ribacke, U., Wirth, D., Rinn, J.L. and Sabeti, P.C. 2015. Strand-specific RNA sequencing in *Plasmodium falciparum* malaria identifies developmentally regulated long non-coding RNA and circular RNA. *BMC Genomics* 16, 454. doi: 10.1186/s12864-015-1603-4.

Bruce, M.C., Alano, P., Duthie, S. and Carter, R. 1990a. Commitment of the malaria parasite *Plasmodium falciparum* to sexual and asexual development. *Parasitology* 100, pp. 191–200. doi: 10.1017/S0031182000061199.

Bruce, M.C., Baker, D.A., Alano, P., Rogers, N.C., Graves, P.M., Targett, G.A.T. and Carter, R. 1990b. Sequence coding for a sexual stage specific protein of *Plasmodium falciparum*. *Nucleic Acids Research* 18, pp. 3171–3174. doi: 10.1093/nar/18.16.4991-b.

Bruce, M.C., Carter, R.N., Nakamura, K. ichiro, Aikawa, M. and Carter, R. 1994. Cellular location and temporal expression of the *Plasmodium falciparum* sexual stage antigen Pfs16. *Molecular and Biochemical Parasitology* 65, pp. 11–22. doi: 10.1016/0166-6851(94)90111-2.

Buckling, A., Crooks, L. and Read, A. 1999a. *Plasmodium* chabaudi: Effect of antimalarial drugs on gametocytogenesis. *Experimental Parasitology* 93, pp. 45–54. doi: 10.1006/expr.1999.4429.

Buckling, A., Ranford-Cartwright, L.C., Miles, A. and Read, A.F. 1999b. Chloroquine increases *Plasmodium falciparum* gametocytogenesis *in vitro*. *Parasitology* 118, pp. 339–346. doi: 10.1017/s0031182099003960.

De Canale, E., Sgarabotto, D., Marini, G., Menegotto, N., Masiero, S., Akkouche, W., Biasolo, M.A., Barzon, L., et al. 2017. *Plasmodium knowlesi* malaria in a traveller returning from the Philippines to Italy, 2016. *New Microbiologica* 40, pp. 291–

294.

Carlton, J.M., Adams, J.H., Silva, J.C., Bidwell, S.L., Lorenzi, H., Caler, E., Crabtree, J., Angiuoli, S. V., et al. 2008. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 445, pp. 757–763. doi: 10.1038/nature07327.

Carter, L.M., Kafsack, B.F.C., Llinás, M., Mideo, N., Pollitt, L.C. and Reece, S.E. 2013. Stress and sex in malaria parasites Why does commitment vary? *Evolution, Medicine, and Public Health* 1, pp. 135–147. doi: 10.1093/emph/eot011.

Carvalho, B.O., Lopes, S.C.P., Nogueira, P.A., Orlandi, P.P., Bargieri, D.Y., Blanco, Y.C., Mamoni, R., Leite, J.A., et al. 2010. On the cytoadhesion of *Plasmodium vivax* –infected erythrocytes. *The Journal of Infectious Diseases* 202, pp. 638–647. doi: 10.1086/654815.

Chin, W., Contacos, P.G., Coatney, G.R. and Kimball, H.R. 1965. A naturally acquired quotidian-type malaria in man transferable to monkeys. *Science* 149, 865. doi: 10.1126/science.149.3686.865.

Chin, W., Contacos, P.G., Collins, W.E., Jeter, M.H. and Alpert, E. 1968. Experimental mosquito-transmission of *Plasmodium knowlesi* to man and monkey. *The American journal of tropical medicine and hygiene* 17, pp. 155–358. doi: 10.4269/ajtmh.1968.17.355.

Chiu, C.Y.H., Hodder, A.N., Lin, C.S., Hill, D.L., Li Wai Suen, C.S.N., Schofield, L., Siba, P.M., Mueller, I., et al. 2015. Antibodies to the *Plasmodium falciparum* proteins MSPDBL1 and MSPDBL2 opsonize merozoites, inhibit parasite growth, and predict protection from clinical malaria. *Journal of Infectious Diseases* 212, pp. 406–415. doi: 10.1093/infdis/jiv057.

Chung, W., Eum, H.H., Lee, H.O., Lee, K.M., Lee, H.B., Kim, K.T., Ryu, H.S., Kim, S., et al. 2017. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications* 8, 15081. doi: 10.1038/ncomms15081.

Claessens, A., Affara, M., Assefa, S.A., Kwiatkowski, D.P. and Conway, D.J. 2017. Culture adaptation of malaria parasites selects for convergent loss-of-function mutants. *Scientific Reports* 7, 41303. doi: 10.1038/srep41303.

Coatney, G., Collins, W., Mcwilson, W. and Contacos, P. 1972. The primate malarias: US Department of Health, Education and Welfare.

Coatney, G.R. 1971. The simian malarias: zoonoses, anthroponoses, or both? *The American journal of tropical medicine and hygiene* 20, pp. 795–803. doi: 10.4269/ajtmh.1971.20.795.

Cooper, D.J., Rajahram, G.S., William, T., Jelip, J., Mohammad, R., Benedict, J., Alaza, D.A., Malacova, E., et al. 2019. *Plasmodium knowlesi* malaria in Sabah, Malaysia, 2015-2017: ongoing increase in incidence despite near-elimination of the human-only *Plasmodium* species. *Clinical Infectious Diseases* 20. doi: 10.1093/cid/ciz237.

Coutrier, F.N., Tirta, Y.K., Cotter, C., Zarlinda, I., González, I.J., Schwartz, A., Maneh, C., Marfurt, J., et al. 2018. Laboratory challenges of *Plasmodium* species identification in Aceh Province, Indonesia, a malaria elimination setting with newly discovered *P. knowlesi*. *PLoS Neglected Tropical Diseases* 12, e0006924. doi: 10.1371/journal.pntd.0006924.

Cowman, A.F., Berry, D. and Baum, J. 2012. The cellular and molecular basis for

malaria parasite invasion of the human red blood cell. *Journal of Cell Biology* 198, pp. 961–971. doi: 10.1083/jcb.201206112.

Cox-Singh, J., Davis, T.M., Lee, K.S., Shamsul, S.S., Matusop, A., Ratnam, S., Rahman, H.A., Conway, D.J., et al. 2008. *Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening. *Clin.Infect.Dis.* 46, pp. 165–171. doi: 10.1086/524888.

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, pp. 65–73. doi: 10.1016/j.molbiopara.2009.12.010.

Daneshvar, C., Davis, T.M.E., Cox-Singh, J., Rafa'ee, M.Z., Zakaria, S.K., Divis, P.C.S. and Singh, B. 2009. Clinical and laboratory features of human *Plasmodium knowlesi* infection. *Clinical Infectious Diseases* 49, pp. 852–860. doi: 10.1086/605439.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., et al. 2002. Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences* 99, pp. 5261–5266. doi: 10.1073/pnas.082089499.

Diez Benavente, E., Florez de Sessions, P., Moon, R.W., Holder, A.A., Blackman, M.J., Roper, C., Drakeley, C.J., Pain, A., et al. 2017. Analysis of nuclear and organellar genomes of *Plasmodium knowlesi* in humans reveals ancient population structure and recent recombination among host-specific subpopulations. *PLoS Genetics* 13, e1007008. doi: 10.1371/journal.pgen.1007008.

Divis, P.C.S., Duffy, C.W., Kadir, K.A., Singh, B. and Conway, D.J. 2018. Genome-wide mosaicism in divergence between zoonotic malaria parasite subpopulations with separate sympatric transmission cycles. *Molecular Ecology* 27, pp. 860–870. doi: 10.1111/mec.14477.

Divis, P.C.S., Lin, L.C., Rovie-ryan, J.J., Kadir, K.A., Anderios, F., Hisam, S., Sharma, R.S.K., Singh, B., et al. 2017. Three divergent subpopulations of the malaria Parasite *Plasmodium knowlesi*. *Emerging Infectious Diseases* 23, pp. 616–624.

Divis, P.C.S., Singh, B., Anderios, F., Hisam, S., Matusop, A., Kocken, C.H., Assefa, S.A., Duffy, C.W., et al. 2015. Admixture in humans of two divergent *Plasmodium knowlesi* populations associated with different macaque host species. *PLoS Pathogens* 11, e1004888. doi: 10.1371/journal.ppat.1004888.

Dolan, S.A., Miller, L.H. and Wellems, T.E. 1990. Evidence for a switching mechanism in the invasion of erythrocytes by *Plasmodium falciparum*. *Journal of Clinical Investigation* 86, pp. 618–624. doi: 10.1172/JCI114753.

Duffy, C.W., Amambua-Ngwa, A., Ahouidi, A.D., Diakite, M., Awandare, G.A., Ba, H., Tarr, S.J., Murray, L., et al. 2018. Multi-population genomic analysis of malaria parasites indicates local selection and differentiation at the *gdv1* locus regulating sexual development. *Scientific Reports* 8, 15763. doi: 10.1038/s41598-018-34078-3.

Duffy, C.W., Assefa, S.A., Abugri, J., Amoako, N., Owusu-Agyei, S., Anyorigiya, T., MacInnis, B., Kwiatkowski, D.P., et al. 2015. Comparison of genomic signatures of selection on *Plasmodium falciparum* between different regions of a country with high malaria endemicity. *BMC Genomics* 16, 527. doi: 10.1186/s12864-015-1746-3.

Duraisingh, M.T., Maier, A.G., Triglia, T. and Cowman, A.F. 2003a. Erythrocyte-

binding antigen 175 mediates invasion in *Plasmodium falciparum* utilizing sialic acid-dependent and -independent pathways. *Proceedings of the National Academy of Sciences* 10, pp. 4796–4801. doi: 10.1073/pnas.0730883100.

Duraisingh, M.T., Triglia, T., Ralph, S.A., Rayner, J.C., Barnwell, J.W., McFadden, G.I. and Cowman, A.F. 2003b. Phenotypic variation of *Plasmodium falciparum* merozoite proteins directs receptor targeting for invasion of human erythrocytes. *EMBO Journal* 22, pp. 1047–1057. doi: 10.1093/emboj/cdg096.

Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M. and Coleman, P. 1992. Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences* 89, pp. 3010–3014. doi: 10.1073/pnas.89.7.3010.

Eede, P. Van Den, Van, H.N., Van Overmeir, C., Vythilingam, I., Duc, T.N., Hung, L.X., Manh, H.N., Anné, J., et al. 2009. Human *Plasmodium knowlesi* infections in young children in central Vietnam. *Malaria Journal* 8, 249. doi: 10.1186/1475-2875-8-249.

Eksi, S., Haile, Y., Furuya, T., Ma, L., Su, X. and Williamson, K.C. 2005. Identification of a subtelomeric gene family expressed during the asexual-sexual stage transition in *Plasmodium falciparum*. *Molecular and Biochemical Parasitology* 143, pp. 90–99. doi: 10.1016/j.molbiopara.2005.05.010.

Eksi, S., Morahan, B.J., Haile, Y., Furuya, T., Jiang, H., Ali, O., Xu, H., Kiattibutr, K., et al. 2012. *Plasmodium falciparum* gametocyte development 1 (*Pfgdv1*) and gametocytogenesis early gene identification and commitment to sexual development. *PLoS Pathogens* 8, e1002964. doi: 10.1371/journal.ppat.1002964.

El-Assaad, F., Wheway, J., Mitchell, A.J., Lou, J., Hunt, N.H., Combes, V. and Grau, G.E.R. 2013. Cytoadherence of *Plasmodium berghei*-infected red blood cells to murine brain and lung microvascular endothelial cells *in vitro*. *Infection and Immunity* 8, pp. 3984–3991. doi: 10.1128/iai.00428-13.

Ennis JG, Teal AE, Habura A, Madison-Antenucci S, Keithly JS, Arguin PM, Barnwell JW, Collins WE, Mali S, Slutsker L, Dasilva A, H. 2009. Simian malaria in a U.S. traveler-New York, 2008. *Morbidity and Mortality Weekly Report* 58, pp. 229–232.

Esteban, J.A., Salas, M. and Blanco, L. 1993. Fidelity of φ29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *Journal of Biological Chemistry* 268, pp. 2719–2726.

Eziefula, A.C., Gosling, R., Hwang, J., Hsiang, M.S., Bousema, T., Von Seidlein, L. and Drakeley, C. 2012. Rationale for short course primaquine in Africa to interrupt malaria transmission. *Malaria Journal* 11, 360. doi: 10.1186/1475-2875-11-360.

Fang, N. and Akinci-Tolun, R. 2016. Depletion of ribosomal RNA Sequences from single-cell RNA-sequencing library. *Current Protocols in Molecular Biology* 115, pp. 7–27. doi: 10.1002/cpmb.11.

Fatih, F.A., Siner, A., Ahmed, A., Woon, L.C., Craig, A.G., Singh, B., Krishna, S. and Cox-Singh, J. 2012. Cytoadherence and virulence - The case of *Plasmodium knowlesi* malaria. *Malaria Journal* 11, 33. doi: 10.1186/1475-2875-11-33.

Fidock, D.A., Nomura, T., Talley, A.K., Cooper, R.A., Dzekunov, S.M., Ferdig, M.T., Ursos, L.M.B., Bir Singh Sidhu, A., et al. 2000. Mutations in the *P. falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Molecular Cell* 6, pp. 861–871. doi: 10.1016/S1097-

2765(05)00077-8.

Figtree, M., Lee, R., Bain, L., Kennedy, T., Mackertich, S., Urban, M., Cheng, Q. and Hudson, B.J. 2010. *Plasmodium knowlesi* in human, Indonesian Borneo. *Emerging Infectious Diseases* 14, pp. 672–674. doi: 10.3201/eid1604.091624.

Filarsky, M., Fraschka, S.A., Niederwieser, I., Brancucci, N.M.B., Carrington, E., Carrió, E., Moes, S., Jenoe, P., et al. 2018. GDV1 induces sexual commitment of malaria parasites by antagonizing HP1-dependent gene silencing. *Science* 359, pp. 1259–1263. doi: 10.1126/science.aan6042.

Fischer, K., Horrocks, P., Preuss, M., Wiesner, J., Wünsch, S., Camargo, A.A. and Lanzer, M. 1997. Expression of *var* genes located within polymorphic subtelomeric domains of *Plasmodium falciparum* chromosomes. *Molecular and Cellular Biology* 17, pp. 3679–3686. doi: 10.1128/mcb.17.7.3679.

Flueck, C., Bartfai, R., Volz, J., Niederwieser, I., Salcedo-Amaya, A.M., Alako, B.T.F., Ehlgen, F., Ralph, S.A., et al. 2009. *Plasmodium falciparum* heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. *PLoS Pathogens* 5, e1000569. doi: 10.1371/journal.ppat.1000569.

Foster, J., Thompson, J. and Wellcome Trust Malaria Genome Collaboration 1995. The *Plasmodium falciparum* genome project: A resource for researchers., pp. 1–4. doi: 10.1016/0169-4758(95)80092-1.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419, pp. 498–511. doi: 10.1038/nature01097.

Gautier, M., Vitalis, R., Baillarguet, D. and Cedex, F.-M. 2012. rehh : an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* 28, pp. 1176–1177. doi: 10.1093/bioinformatics/bts115.

Gautret, P. and Motard, A. 1999. Periodic infectivity of *Plasmodium* gametocytes to the vector. A review. *Parasite* 6, pp. 103–111. doi: 10.1051/parasite/1999062103.

Gething, P.W., Casey, D.C., Weiss, D.J., Bisanzio, D., Bhatt, S., Cameron, E., Battle, K.E., Dalrymple, U., et al. 2016. Mapping *Plasmodium falciparum* Mortality in Africa between 1990 and 2015. *New England Journal of Medicine* 375, pp. 2435–2445. doi: 10.1056/nejmoa1606701.

Ghinai, I., Cook, J., Hla, T.T.W., Htet, H.M.T., Hall, T., Lubis, I.N., Ghinai, R., Hesketh, T., et al. 2017. Malaria epidemiology in central Myanmar: identification of a multi-species asymptomatic reservoir of infection. *Malaria Journal* 16, 16. doi: 10.1186/s12936-016-1651-5.

Gilson, P.R. and Crabb, B.S. 2009. Morphology and kinetics of the three distinct phases of red blood cell invasion by *Plasmodium falciparum* merozoites. *International Journal for Parasitology* 39, pp. 91–96. doi: 10.1016/j.ijpara.2008.09.007.

Hamilton, W.L., Amato, R., van der Pluijm, R.W., Jacob, C.G., Quang, H.H., Thuy-Nhien, N.T., Hien, T.T., Hongvanthong, B., et al. 2019. Evolution and expansion of multidrug-resistant malaria in southeast Asia: a genomic epidemiology study. *The Lancet Infectious Diseases* 3099, pp. 10–15. doi: 10.1016/S1473-3099(19)30392-5

Hamilton, W.L., Claessens, A., Otto, T.D., Kekre, M., Fairhurst, R.M., Rayner, J.C. and Kwiatkowski, D. 2017. Extreme mutation bias and high AT content in *Plasmodium falciparum*. *Nucleic acids research* 45, pp. 1889–1901. doi: 10.1093/nar/gkw1259.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., et al. 2016. CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology* 17, 77. doi: 10.1186/s13059-016-0938-8.

Heather, J.M. and Chain, B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics* 107, pp. 1–8. doi: 10.1016/j.ygeno.2015.11.003.

Henriques, G., Hallett, R.L., Beshir, K.B., Gadalla, N.B., Johnson, R.E., Burrow, R., Van Schalkwyk, D.A., Sawa, P., et al. 2014. Directional selection at the *pfmdr1*, *pfcrt*, *pfubp1*, and *pfap2mu* loci of *Plasmodium falciparum* in Kenyan children treated with ACT. *Journal of Infectious Diseases* 210, pp. 2001–2008. doi: 10.1093/infdis/jiu358.

Herdiana, H., Cotter, C., Coutrier, F.N., Zarlinda, I., Zelman, B.W., Tirta, Y.K., Greenhouse, B., Gosling, R.D., et al. 2016. Malaria risk factor assessment using active and passive surveillance data from Aceh Besar, Indonesia, a low endemic, malaria elimination setting with *Plasmodium knowlesi*, *Plasmodium vivax*, and *Plasmodium falciparum*. *Malaria Journal* 15, 468. doi: 10.1186/s12936-016-1523-z.

Herdiana, H., Irnawati, I., Coutrier, F.N., Munthe, A., Mardiati, M., Yuniarti, T., Sariwati, E., Sumiwi, M.E., et al. 2018. Two clusters of *Plasmodium knowlesi* cases in a malaria elimination area, Sabang Municipality, Aceh, Indonesia. *Malaria Journal* 17, 186. doi: 10.1186/s12936-018-2334-1.

Hodder, A.N., Czabotar, P.E., Uboldi, A.D., Clarke, O.B., Lin, C.S., Healer, J., Smith, B.J. and Cowman, A.F. 2012. Insights into duffy binding-like domains through the crystal structure and function of the merozoite surface protein MSPDBL2 from *Plasmodium falciparum*. *Journal of Biological Chemistry* 287, pp. 32922–32939. doi: 10.1074/jbc.M112.350504.

Imwong, M., Madmanee, W., Suwannasin, K., Kunasol, C., Peto, T.J., Tripura, R., Von Seidlein, L., Nguon, C., et al. 2019. Asymptomatic natural human infections with the simian malaria parasites *Plasmodium cynomolgi* and *Plasmodium knowlesi*. *Journal of Infectious Diseases* 219, pp. 695–702. doi: 10.1093/infdis/jiy519.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P. and Linnarsson, S. 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research* 21, pp. 1160–1167. doi: 10.1101/gr.110882.110.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P. and Linnarsson, S. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods* 11, pp. 163–166. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24363023.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., et al. 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343, pp. 776–779. doi: 10.1126/science.1247651.

Janssen, C.S., Barrett, M.P., Turner, C.M.R. and Stephen Phillips, R. 2002. A large gene family for putative variant antigens shared by human and rodent malaria parasites. *Proceedings of the Royal Society B: Biological Sciences* 269, pp. 431–436. doi: 10.1098/rspb.2001.1903.

Jeffares, D.C., Pain, A., Berry, A., Cox, A. V., Stalker, J., Ingle, C.E., Thomas, A., Quail, M.A., et al. 2007. Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. *Nature Genetics* 39, pp. 120–125. doi: 10.1038/ng1931.

Jeslyn, W.P.S., Huat, T.C., Vernon, L., Irene, L.M.Z., Sung, L.K., Jarrod, L.P., Singh, B. and Ching, N.L. 2011. Molecular epidemiological investigation of *Plasmodium knowlesi* in humans and macaques in Singapore. *Vector-Borne and Zoonotic Diseases* 11, pp. 131–135. doi: 10.1089/vbz.2010.0024.

Jiram, A.I., Vythilingam, I., Noorazian, Y.M., Yusof, Y.M., Azahari, A.H. and Fong, M.Y. 2012. Entomologic investigation of *Plasmodium knowlesi* vectors in Kuala Lipis, Pahang, Malaysia. *Malaria Journal* 11, 213. doi: 10.1186/1475-2875-11-213.

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

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, pp. 1799–1806. doi: 10.3201/eid1710.110349.

Jongwutiwes, S., Putaporntip, C., Iwasaki, T., Sata, T. and Kanbara, H. 2004. Naturally acquired *Plasmodium knowlesi* malaria in human, Thailand. *Emerging Infectious Diseases* 10, pp. 2211–2213. doi: 10.3201/eid1012.040293.

Josling, G.A., Venezia, J., Orchard, L., Russell, T.J., Painter, H.J. and Llinas, M. 2019. Regulation of sexual differentiation is linked to invasion in malaria parasites. *bioRxiv* . doi: 10.1101/533877.

Kafsack, B.F.C., Rovira-Graells, N., Clark, T.G., Bancells, C., Crowley, V.M., Campino, S.G., Williams, A.E., Drought, L.G., et al. 2014. A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature* 507, pp. 248–252. doi: 10.1038/nature12920.

Kaslow, D.C., Quakyi, I.A., Syin, C., Raum, M.G., Keister, D.B., Coligan, J.E., McCutchan, T.F. and Miller, L.H., 1988. A vaccine candidate from the sexual stage of human malaria that contains EGF-like domains. *Nature*, 333, pp.74-76.

Khim, N., Siv, S., Kim, S., Mueller, T., Fleischmann, E., Singh, B., Divis, P.C.S., Steenkeste, N., et al. 2011. *Plasmodium knowlesi* infection in humans, Cambodia, 2007-2010. *Emerging Infectious Diseases* 17, pp. 1900–1902. doi: 10.3201/eid1710.110355.

Kim, A., Popovici, J., Menard, D. and Serre, D. 2019. *Plasmodium vivax* transcriptomes reveal stage-specific chloroquine response and differential regulation of male and female gametocytes. *Nature Communications* 10, 371. doi: 10.1038/s41467-019-08312-z.

Kim, D., Langmead, B. and Salzberg, S.L. 2015. Hisat: a fast spliced aligner with low memory requirements. *Nature methods* 12, pp. 357–360. doi: 10.1038/nmeth.3317.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., et al. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, pp. 1187–1201. doi: 10.1016/j.cell.2015.04.044.

Kolisko, M., Boscaro, V., Burki, F., Lynn, D.H. and Keeling, P.J. 2014. Single-cell transcriptomics for microbial eukaryotes. *Current Biology* 24, pp. 1081–1082. doi: 10.1016/j.cub.2014.10.026.

Kumar, S., Banks, T.W. and Cloutier, S. 2012. SNP discovery through next-generation sequencing and its applications. *International Journal of Plant Genomics* 2012. doi: 10.1155/2012/831460.

Lapp, S.A., Mok, S., Zhu, L., Wu, H., Preiser, P.R., Bozdech, Z. and Galinski, M.R. 2015. *Plasmodium knowlesi* gene expression differs in *ex vivo* compared to *in vitro* blood-stage cultures. *Malaria Journal* 14, 110. doi: 10.1186/s12936-015-0612-8.

Lasken, R.S. 2007. Single-cell genomic sequencing using Multiple Displacement Amplification. *Current Opinion in Microbiology* 10, pp. 510–516. doi: 10.1016/j.mib.2007.08.005.

Lasonder, E., Rijpma, S.R., Van Schaijk, B.C.L., Hoeijmakers, W.A.M., Kensche, P.R., Gresnigt, M.S., Italiaander, A., Vos, M.W., et al. 2016. Integrated transcriptomic and proteomic analyses of *P. falciparum* gametocytes: Molecular insight into sex-specific processes and translational repression. *Nucleic Acids Research* 44, pp. 6087–6101. doi: 10.1093/nar/gkw536.

Lee, K.S., Cox-Singh, J., Brooke, G., Matusop, A. and Singh, B. 2009. *Plasmodium knowlesi* from archival blood films: Further evidence that human infections are widely distributed and not newly emergent in Malaysian Borneo. *International Journal for Parasitology* 39, pp. 1125–1128. doi: 10.1016/j.ijpara.2009.03.003.

Lee, K.S., Divis, P.C.S., Zakaria, S.K., Matusop, A., Julin, R.A., Conway, D.J., Cox-Singh, J. and Singh, B. 2011. *Plasmodium knowlesi*: Reservoir hosts and tracking the emergence in humans and macaques. *PLoS Pathogens* 7, e1002015. doi: 10.1371/journal.ppat.1002015.

LeRoux, M., Lakshmanan, V. and Daily, J.P. 2009. *Plasmodium falciparum* biology: analysis of *in vitro* versus *in vivo* growth conditions. *Trends in Parasitology* 25, pp. 474–481. doi: 10.1016/j.pt.2009.07.005.

Li, H. and Durbin, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, pp. 589–595. doi: 10.1093/bioinformatics/btp698.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25, pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, X.Q. and Du, D. 2014. Variation, evolution, and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. *PLoS ONE* 9, e88339. doi: 10.1371/journal.pone.0088339.

Lim, C., Hansen, E., DeSimone, T.M., Moreno, Y., Junker, K., Bei, A., Brugnara, C., Buckee, C.O., et al. 2013. Expansion of host cellular niche can drive adaptation of a zoonotic malaria parasite to humans. *Nature communications* 4, 1638. doi: 10.1038/ncomms2612.

Lingelbach, K. and Joiner, K.A. 1998. The parasitophorous vacuole membrane surrounding *Plasmodium* and *Toxoplasma*: an unusual compartment in infected cells. *Journal of cell science* 111, pp. 1467–1475.

Liu, W., Li, Y., Learn, G.H., Rudicell, R.S., Robertson, J.D., Keele, B.F., Ndjango, J.B.N., Sanz, C.M., et al. 2010. Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. *Nature* 467, pp. 420–425. doi: 10.1038/nature09442.

Liu, Z., Hu, S.K., Campbell, V., Tatters, A.O., Heidelberg, K.B. and Caron, D.A. 2017. Single-cell transcriptomics of small microbial eukaryotes : limitations and potential. *The ISME Journal* 11, pp. 1–4. doi: 10.1038/ismej.2016.190.

Liu, Z., Miao, J. and Cui, L., 2011. Gametocytogenesis in malaria parasite: commitment, development and regulation. *Future microbiology*, 6, pp.1351-1369.

Lopez-Rubio, J.J., Mancio-Silva, L. and Scherf, A. 2009. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell Host and Microbe* 5, pp. 179–190. doi: 10.1016/j.chom.2008.12.012.

Love, M.I., Huber, W. and Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15, 550. doi: 10.1186/s13059-014-0550-8.

Lu, F., Culleton, R., Zhang, M., Ramaprasad, A., von Seidlein, L., Zhou, H., Zhu, G., Tang, J., et al. 2017. Emergence of indigenous artemisinin-resistant *Plasmodium falciparum* in Africa. *New England Journal of Medicine* 376, pp. 991–993. doi: 10.1056/nejmc1612765.

Lubis, I.N.D., Wijaya, H., Lubis, M., Lubis, C.P., Divis, P.C.S., Beshir, K.B. and Sutherland, C.J. 2017. Contribution of *Plasmodium knowlesi* to multispecies human Malaria infections in North Sumatera, Indonesia. *Journal of Infectious Diseases* 215, pp. 1148–1155. doi: 10.1093/infdis/jix091.

Macaulay, I.C. and Voet, T. 2014. Single cell genomics: advances and future perspectives. *PLoS Genetics* 10, e1004126. doi: 10.1371/journal.pgen.1004126.

MacKintosh, C.L., Beeson, J.G. and Marsh, K. 2004. Clinical features and pathogenesis of severe malaria. *Trends in Parasitology* 20, pp. 597–603. doi: 10.1016/j.pt.2004.09.006.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., et al. 2015. Highly parallel genome-wide expression profiling of individual cells esing nanoliter droplets. *Cell* 161, pp. 1202–1214. doi: 10.1016/j.cell.2015.05.002.

Manske, M., Miotto, O., Campino, S., Auburn, S., Almagro-Garcia, J., Maslen, G., O'Brien, J., Djimde, A., et al. 2012. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* 487, pp. 375–379. doi: 10.1038/nature11174.

Marti, M., Good, R.T., Rug, M., Knuepfer, E. and Cowman, A.F. 2004. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* 306, pp. 1930–1933. doi: 10.1126/science.1102452.

McCarthy, D.J., Campbell, K.R., Lun, A.T.L. and Wills, Q.F. 2016. scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R. *Bioinformatics* 33, pp. 1179–1186. Available at: http://biorxiv.org/lookup/doi/10.1101/069633.

Ménard, D., Khim, N., Beghain, J., Adegnika, A.A., Shafiul-Alam, M., Amodu, O., Rahim-Awab, G., Barnadas, C., et al. 2016. A worldwide map of *Plasmodium falciparum* K13-propeller polymorphisms. *New England Journal of Medicine* 374, pp. 2453–2464. doi: 10.1056/nejmoa1513137.

Miles, A., Iqbal, Z., Vauterin, P., Pearson, R., Campino, S., Theron, M., Gould, K., Mead, D., et al. 2016. Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. *Genome Research* 26, pp. 1288–1299. doi: 10.1101/gr.203711.115.

Milner, D.A., Pochet, N., Krupka, M., Williams, C., Seydel, K., Taylor, T.E., van de Peer, Y., Regev, A., et al. 2012. Transcriptional profiling of *Plasmodium falciparum* parasites from patients with severe malaria identifies distinct low vs. high parasitemic clusters. *PLoS ONE* 7, e40739. doi:

10.1371/journal.pone.0040739.

Miotto, O., Almagro-Garcia, J., Manske, M., MacInnis, B., Campino, S., Rockett, K.A., Amaratunga, C., Lim, P., et al. 2013. Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia. *Nature Genetics* 45, pp. 648–655. doi: 10.1038/ng.2624.

Miotto, O., Amato, R., Ashley, E.A., Macinnis, B., Almagro-Garcia, J., Amaratunga, C., Lim, P., Mead, D., et al. 2015. Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. *Nature Genetics* 47, pp. 226–234. doi: 10.1038/ng.3189.

Mita, T. and Tanabe, K. 2012. Evolution of *Plasmodium falciparum* drug resistance: Implications for the development and containment of artemisinin resistance. *Japanese Journal of Infectious Diseases* 65, pp. 465–475. doi: 10.7883/yoken.65.465.

Mobegi, V.A., Duffy, C.W., Amambua-Ngwa, A., Loua, K.M., Laman, E., Nwakanma, D.C., MacInnis, B., Aspeling-Jones, H., et al. 2014. Genome-wide analysis of selection on the malaria parasite *Plasmodium falciparum* in West African populations of differing infection endemicity. *Molecular Biology and Evolution* 31, pp. 1490–9. doi: 10.1093/molbev/msu106.

Moon, R.W., Hall, J., Rangkuti, F., Ho, Y.S., Almond, N., Mitchell, G.H., Pain, A., Holder, A. a, et al. 2013. Adaptation of the genetically tractable malaria pathogen *Plasmodium knowlesi* to continuous culture in human erythrocytes. *Proceedings of the National Academy of Sciences of the United States of America* 110, pp. 531–6. doi: 10.1073/pnas.1216457110.

Moon, R.W., Sharaf, H., Hastings, C.H., Ho, Y.S., Nair, M.B., Rchiad, Z., Knuepfer, E., Ramaprasad, A., et al. 2016. Normocyte-binding protein required for human erythrocyte invasion by the zoonotic malaria parasite *Plasmodium knowlesi*. *Proceedings of the National Academy of Sciences of the United States of America* 113, pp. 7231–7236. doi: 10.1073/pnas.1522469113.

Moyes, C.L., Shearer, F.M., Huang, Z., Wiebe, A., Gibson, H.S., Nijman, V., Mohd-Azlan, J., Brodie, J.F., et al. 2016. Predicting the geographical distributions of the macaque hosts and mosquito vectors of *Plasmodium knowlesi* malaria in forested and non-forested areas. *Parasites and Vectors* 9, 242. doi: 10.1186/s13071-016-1527-0.

Müller, M. and Schlagenhauf, P. 2014. *Plasmodium knowlesi* in travellers, update 2014. *International Journal of Infectious Diseases* 22, pp. 55–64. doi: 10.1016/j.ijid.2013.12.016.

Murray, C.J.L., Rosenfeld, L.C., Lim, S.S., Andrews, K.G., Foreman, K.J., Haring, D., Fullman, N., Naghavi, M., et al. 2012. Global malaria mortality between 1980 and 2010: A systematic analysis. *The Lancet* 379, pp. 413–431. doi: 10.1016/S0140-6736(12)60034-8.

Nair, S., Nash, D., Sudimack, D., Jaidee, A., Barends, M., Uhlemann, A.C., Krishna, S., Nosten, F., et al. 2007. Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Molecular Biology and Evolution* 24, pp. 562–573. doi: 10.1093/molbev/msl185.

Ndam, N.T., Bischoff, E., Proux, C., Lavstsen, T., Salanti, A., Guitard, J., Nielsen, M.A., Coppée, J.Y., et al. 2008. *Plasmodium falciparum* transcriptome analysis reveals pregnancy malaria associated gene expression. *PLoS ONE* 3, e1855. doi: 10.1371/journal.pone.0001855.

Nery, S., Deans, A.M., Mosobo, M., Marsh, K., Rowe, J.A. and Conway, D.J. 2006. Expression of *Plasmodium falciparum* genes involved in erythrocyte invasion varies among isolates cultured directly from patients. *Molecular and Biochemical Parasitology* 149, pp. 208–215. doi: 10.1016/j.molbiopara.2006.05.014.

Ng, O.T., Eng, E.O., Cheng, C.L., Piao, J.L., Lee, C.N., Pei, S.W., Tian, M.T., Jin, P.L., et al. 2008. Naturally acquired human *Plasmodium knowlesi* infection, Singapore. *Emerging Infectious Diseases* 14, pp. 814–816. doi: 10.3201/eid1405.070863.

Ngara, M., Palmkvist, M., Sagasser, S., Hjelmqvist, D., Björklund, Å.K., Wahlgren, M., Ankarklev, J. and Sandberg, R. 2018. Exploring parasite heterogeneity using single-cell RNA-seq reveals a gene signature among sexual stage *Plasmodium falciparum* parasites. *Experimental Cell Research* 371, pp. 130–138. doi: 10.1016/j.yexcr.2018.08.003.

Nikbakht, H., Xia, X. and Hickey, D.A. 2015. The evolution of genomic GC content undergoes a rapid reversal within the genus *Plasmodium*. *Genome* 57, pp. 507–511. doi: 10.1139/gen-2014-0158.

Nixon, C.P., Nixon, C.E., Michelow, I.C., Silva-Viera, R.A., Colantuono, B., Obeidallah, A.S., Jha, A., Dockery, D., et al. 2018. Antibodies to PfsEGXP, an early gametocyte-enriched phosphoprotein, predict decreased *Plasmodium falciparum* gametocyte density in humans. *Journal of Infectious Diseases* 218, pp. 1792–1801. doi: 10.1093/infdis/jiy416.

Nyrén, P., Pettersson, B. and Uhlén, M. 1993. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical Biochemistry* 208, pp. 171–175. doi: 10.1006/abio.1993.1024.

Ochola, L.I., Tetteh, K.K.A., Stewart, L.B., Riitho, V., Marsh, K. and Conway, D.J. 2010. Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in *Plasmodium falciparum*. *Molecular Biology and Evolution* 27, pp. 2344–2351. doi: 10.1093/molbev/msq119.

Otto, T.D., Wilinski, D., Assefa, S., Keane, T.M., Sarry, L.R., Böhme, U., Lemieux, J., Barrell, B., et al. 2010. New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Molecular Microbiology* 76, pp. 12–24. doi: 10.1111/j.1365-2958.2009.07026.x.

Paez, J.G., Lin, M., Beroukhim, R., Lee, J.C., Zhao, X., Richter, D.J., Gabriel, S., Herman, P., et al. 2004. Genome coverage and sequence fidelity of φ29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic acids research* 32, e71. doi: 10.1093/nar/gnh069.

Pain, A., Bohme, U., Berry, A.E., Mungall, K., Finn, R.D., Jackson, A.P., Mourier, T., Mistry, J., et al. 2008. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* 455, pp. 799–803. doi: 10.1038/nature07306.

Painter, H.J., Carrasquilla, M. and Llinás, M. 2017. Capturing *in vivo* RNA transcriptional dynamics from the malaria parasite *Plasmodium falciparum*. *Genome Research* 27, pp. 1074–1086. doi: 10.1101/gr.217356.116.

Pan, X., Durrett, R.E., Zhu, H., Tanaka, Y., Li, Y., Zi, X., Marjani, S.L., Euskirchen, G., et al. 2013. Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proceedings of the National Academy of Sciences* 110, pp. 594–599. doi: 10.1073/pnas.1217322109.

Paradis, E. 2010. Pegas: An R package for population genetics with an integrated-

modular approach. *Bioinformatics* 26, pp. 419–420. doi: 10.1093/bioinformatics/btp696.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B. V., et al. 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, pp. 1396–1401. doi: 10.1126/science.1254257.

Peatey, C.L., Skinner-Adams, T.S., Dixon, M.W.A., McCarthy, J.S., Gardiner, D.L. and Trenholme, K.R. 2009. Effect of antimalarial drugs on *Plasmodium falciparum* gametocytes. *The Journal of Infectious Diseases* 200, pp. 1518–1521. doi: 10.1086/644645.

Petter, M., Haeggström, M., Khattab, A., Fernandez, V., Klinkert, M.Q. and Wahlgren, M. 2007. Variant proteins of the *Plasmodium falciparum* RIFIN family show distinct subcellular localization and developmental expression patterns. *Molecular and Biochemical Parasitology* 156, pp. 51–61. doi: 10.1016/j.molbiopara.2007.07.011.

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

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S. and Sandberg, R. 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols* 9, pp. 171–81. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24385147.

Pinard, R., de Winter, A., Sarkis, G.J., Gerstein, M.B., Tartaro, K.R., Plant, R.N., Egholm, M., Rothberg, J.M., et al. 2006. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC genomics* 7, 216. doi: 10.1186/1471-2164-7-216.

Pinheiro, M.M., Ahmed, M.A., Millar, S.B., Sanderson, T., Otto, T.D., Lu, W.C., Krishna, S., Rayner, J.C., et al. 2015. *Plasmodium knowlesi* genome sequences from clinical isolates reveal extensive genomic dimorphism. *PLoS ONE* 10, e01211303. doi: 10.1371/journal.pone.0121303.

Poran, A., Nötzel, C., Aly, O., Mencia-Trinchant, N., Harris, C.T., Guzman, M.L., Hassane, D.C., Elemento, O., et al. 2017. Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. *Nature* 551, pp. 95–99. doi: 10.1038/nature24280.

Putaporntip, C., Hongsrimuang, T., Seethamchai, S., Kobasa, T., Limkittikul, K., Cui, L. and Jongwutiwes, S. 2009. Differential prevalence of *Plasmodium* infections and cryptic *Plasmodium knowlesi* malaria in humans in Thailand. *The Journal of Infectious Diseases* 199, pp. 1143–1150. doi: 10.1086/597414.

Quinlan, A.R. and Hall, I.M. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, pp. 841–842. doi: 10.1093/bioinformatics/btq033.

R Development Core Team 2013. R: A language and environment for statistical computing. doi: 10.1007/978-3-540-74686-7.

Rajahram, G.S., Barber, B.E., William, T., Menon, J., Anstey, N.M. and Yeo, T.W. 2012. Deaths due to *Plasmodium knowlesi* malaria in Sabah, Malaysia: Association with reporting as *Plasmodium malariae* and delayed parenteral artesunate. *Malaria Journal* 11, 284. doi: 10.1186/1475-2875-11-284.

Ramsköld, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., et al. 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology* 30, pp. 777–782. doi: 10.1038/nbt.2282.

Reid, A.J., Talman, A.M., Bennett, H.M., Gomes, A.R., Sanders, M.J., Illingworth, C.J.R., Billker, O., Berriman, M., et al. 2018. Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *eLife* 7, e33105. doi: 10.7554/elife.33105.

Reinius, B., Plaza Reyes, A., Edsgärd, D., Codeluppi, S., Petropoulos, S., Deng, Q., Lanner, F., Sandberg, R., et al. 2016. Single-cell RNA-Seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* 165, pp. 1012–1026. doi: 10.1016/j.cell.2016.03.023.

Ribaut, C., Berry, A., Chevalley, S., Reybier, K., Morlais, I., Parzy, D., Nepveu, F., Benoit-Vical, F. and Valentin, A., 2008. Concentration and purification by magnetic separation of the erythrocytic stages of all human Plasmodium species. *Malaria journal*, 7, p.45.

Ronaghi, M., Uhlén, M. and Nyrén, P. 1998. A sequencing method based on real-time pyrophosphate. *Science* 281, pp. 363–365. doi: 10.1126/science.281.5375.363.

Rovira-Graells, N., Gupta, A.P., Planet, E., Crowley, V.M., Mok, S., De Pouplana, L.R., Preiser, P.R., Bozdech, Z., et al. 2012. Transcriptional variation in the malaria parasite *Plasmodium falciparum*. *Genome Research* 22, pp. 925–938. doi: 10.1101/gr.129692.111.

RTS, S.C.T.P. 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, pp. 31–45. doi: 10.1016/S0140-6736(15)60721-8.

Rueangweerayut, R., Bancone, G., Harrell, E.J., Beelen, A.P., Kongpatanakul, S., Möhrle, J.J., Rousell, V., Mohamed, K., et al. 2017. Hemolytic potential of tafenoquine in female volunteers heterozygous for Glucose-6-Phosphate Dehydrogenase (G6PD) Deficiency (G6PD Mahidol Variant) versus G6PD-Normal volunteers. *American Journal of Tropical Medicine and Hygiene* 97, pp. 702–711. doi: 10.4269/ajtmh.16-0779.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. and Barrell, B. 2000. Artemis: Sequence visualization and annotation. *Bioinformatics* 16, pp. 944–945. doi: 10.1093/bioinformatics/16.10.944.

Rutledge, G.G., Böhme, U., Sanders, M., Reid, A.J., Cotton, J.A., Maiga-Ascofare, O., Djimdé, A.A., Apinjoh, T.O., et al. 2017. *Plasmodium malariae* and *P. ovale* genomes provide insights into malaria parasite evolution. *Nature* 542, pp. 101–104. doi: 10.1038/nature21038.

Sallum, M.A.M., Peyton, E.L. and Wilkerson, R.C. 2005. Six new species of the *Anopheles leucosphyrus* group, reinterpretation of *An. elegans* and vector implications. *Medical and Veterinary Entomology* 19, pp. 158–199. doi: 10.1111/j.0269-283X.2005.00551.x.

Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., et al. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology* 12, 87. doi: 10.1186/s12915-014-0087-z.

Sanger, F. and Coulson, A.R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* 94, pp. 441–448. doi: 10.1016/0022-2836(75)90213-2.

Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchison, C.A., et al. 1978. The nucleotide sequence of bacteriophage φX174. *Journal of Molecular Biology* 125, pp. 225–246. doi: 10.1016/0022-2836(78)90346-7.

Sanger, F., Nicklen, S. and Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74, pp. 5463–5467.

van Schaijk, B.C.L., van Dijk, M.R., van de Vegte-Bolmer, M., van Gemert, G.J., van Dooren, M.W., Eksi, S., Roeffen, W.F.G., Janse, C.J., et al. 2006. Pfs47, paralog of the male fertility factor Pfs48/45, is a female specific surface protein in *Plasmodium falciparum*. *Molecular and Biochemical Parasitology* 149, pp. 216–222. doi: 10.1016/j.molbiopara.2006.05.015.

Scherf, A., Hernandez-Rivas, R., Buffet, P., Bottius, E., Benatar, C., Pouvelle, B., Gysin, J. and Lanzer, M. 1998. Antigenic variation in malaria: *In situ* switching, relaxed and mutually exclusive transcription of *var* genes during intra-erythrocytic development in *Plasmodium falciparum*. *EMBO Journal* 17, pp. 5418–5426. doi: 10.1093/emboj/17.18.5418.

Schneider, P., Greischar, M.A., Birget, P.L.G., Repton, C., Mideo, N. and Reece, S.E. 2018. Adaptive plasticity in the gametocyte conversion rate of malaria parasites. *PLoS Pathogens* 14, e1007371. doi: 10.1371/journal.ppat.1007371.

Shah, N.K., Dhillon, G.P.S., 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, pp. 57–64. doi: 10.1016/S1473-3099(10)70214-0.

Shearer, F.M., Huang, Z., Weiss, D.J., Wiebe, A., Gibson, H.S., Battle, K.E., Pigott, D.M., Brady, O.J., et al. 2016. Estimating geographical variation in the risk of zoonotic *Plasmodium knowlesi* infection in countries eliminating malaria. *PLoS Neglected Tropical Diseases* 10, e0004915. doi: 10.1371/journal.pntd.0004915.

Shock, J.L., Fischer, K.F. and DeRisi, J.L., 2007. Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome biology*, 8, p.R134.

Silvestrini, F., Lasonder, E., Olivieri, A., Camarda, G., van Schaijk, B., Sanchez, M., Younis Younis, S., Sauerwein, R., et al. 2010. Protein export marks the early Phase of gametocytogenesis of the human malaria parasite *Plasmodium falciparum*. *Molecular & Cellular Proteomics* 9, pp. 1437–1448. doi: 10.1074/mcp.m900479-mcp200.

Siegel, T.N., Hon, C.C., Zhang, Q., Lopez-Rubio, J.J., Scheidig-Benatar, C., Martins, R.M., Sismeiro, O., Coppée, J.Y. and Scherf, A., 2014. Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum. *BMC genomics*, *15*(1), p.150.

Singh, B. and Daneshvar, C. 2013. Human infections and detection of *Plasmodium knowlesi*. *Clinical Microbiology Reviews* 26, pp. 165–184. doi: 10.1128/CMR.00079-12.

Singh, B., Sung, L.K., Matusop, A., Radhakrishnan, A., Shamsul, S.S., Cox-Singh, J.,

Thomas, A. and Conway, D.J. 2004. A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *The Lancet* 363, pp. 1017–1024. doi: 10.1016/S0140-6736(04)15836-4.

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. doi: 10.1371/journal.pone.0005410.

Sinha, A., Hughes, K.R., Modrzynska, K.K., Otto, T.D., Pfander, C., Dickens, N.J., Religa, A.A., Bushell, E., et al. 2014. A cascade of DNA-binding proteins for sexual commitment and development in *Plasmodium*. *Nature* 507, pp. 253–257. doi: 10.1038/nature12970.

Sinka, M.E., Bangs, M.J., Manguin, S., Chareonviriyaphap, T., Patil, A.P., Temperley, W.H., Gething, P.W., Elyazar, I.R., et al. 2011. The dominant *Anopheles* vectors of human malaria in the Asia-Pacific region: Occurrence data, distribution maps and bionomic précis. *Parasites and Vectors* 4, 89. doi: 10.1186/1756-3305-4-89.

Sitohang, V., Sariwati, E., Fajariyani, S.B., Hwang, D., Kurnia, B., Hapsari, R.K., Laihad, F.J., Sumiwi, M.E., et al. 2018. Malaria elimination in Indonesia: halfway there. *The Lancet Global Health* 6, pp. 604–606. doi: 10.1016/s2214-109x(18)30198-0.

Smalley, M.E., Brown, J. and Bassett, N.M. 1981. The rate of production of *Plasmodium falciparum* gametocytes during natural infections. *Transactions of the Royal Society of Tropical Medicine and Hygiene* . doi: 10.1016/0035-9203(81)90349-7.

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I. and Sermon, K. 2006a. Optimization and evaluation of single-cell whole-genome multiple displacement amplification. *Human Mutation* 27, pp. 496–503. doi: 10.1002/humu.20324.

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I. and Sermon, K. 2006b. Whole-genome multiple displacement amplification from single cells. *Nature Protocols* 1, pp. 1965–1970. doi: 10.1038/nprot.2006.326.

Stark, D.J., Fornace, K.M., Brock, P.M., Abidin, T.R., Gilhooly, L., Jalius, C., Goossens, B., Drakeley, C.J., et al. 2019. Long-tailed macaque response to deforestation in a *Plasmodium knowlesi*-endemic area. *EcoHealth* , pp. 1–9. doi: 10.1007/s10393-019-01403-9.

Streets, A.M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., Yang, L., Fu, Y., et al. 2014. Microfluidic single-cell whole-transcriptome sequencing. *Proceedings of the National Academy of Sciences* 111, pp. 7048–7053. doi: 10.1073/pnas.1402030111.

Stubbs, J., Simpson, K.M., Triglia, T., Plouffe, D., Tonkin, C.J., Duraisingh, M.T., Maier, A.G., Winzeler, E.A., et al. 2005. Molecular mechanism for switching of *P. falciparum* invasion pathways into human erythrocytes. *Science* 309, pp. 1384–1387. doi: 10.1126/science.1115257.

Su, X. zhuan, Heatwole, V.M., Wertheimer, S.P., Guinet, F., Herrfeldt, J.A., Peterson, D.S., Ravetch, J.A. and Wellems, T.E. 1995. The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell* 82, pp. 89–100. doi: 10.1016/0092-8674(95)90055-1.

Sundararaman, S.A., Hahn, B.H., Avitto, A.N., Sharp, P.M., Liu, W., MacLean, O.A., Giles, J., Trimboli, S., et al. 2018. Evolutionary history of human *Plasmodium vivax* revealed by genome-wide analyses of related ape parasites. *Proceedings of the National Academy of Sciences* 115, pp. 8450–8459. doi: 10.1073/pnas.1810053115.

Tachibana, S.I., Sullivan, S.A., Kawai, S., Nakamura, S., Kim, H.R., Goto, N., Arisue, N., Palacpac, N.M.Q., et al. 2012. *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. *Nature Genetics* 44, pp. 1051–1055. doi: 10.1038/ng.2375.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K. and Surani, M.A. 2010. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-seq analysis. *Cell Stem Cell* 6, pp. 468–478. doi: 10.1016/j.stem.2010.03.015.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* 6, pp. 377–382. doi: 10.1038/nmeth.1315

Tarlov, A.R., Brewer, G.J., Carson, P.E. and Alving, A.S. 1962. Primaquine sensitivity: glucose-6-phosphate dehydrogenase deficiency: an inborn error of metabolism of medical and biological significance. *Archives of Internal Medicine* 109, pp. 209–234. doi: 10.1001/archinte.1962.03620140081013.

Tarr, S.J., Díaz-Ingelmo, O., Stewart, L.B., Hocking, S.E., Murray, L., Duffy, C.W., Otto, T.D., Chappell, L., et al. 2018. *Plasmodium falciparum* mature schizont transcriptome variation among clinical isolates and laboratory-adapted clones. *BMC Genomics* 19, 894. doi: 10.1101/329532.

Taylor, L.H. and Read, A.F. 1997. Why so few transmission stages? Reproductive restraint by malaria parasites. *Parasitology Today* 13, pp. 135–140. doi: 10.1016/S0169-4758(97)89810-9.

Tetteh, K.K.A., Osier, F.H.A., Salanti, A., Kamuyu, G., Drought, L., Failly, M., Martin, C., Marsh, K., et al. 2013. Analysis of antibodies to newly described *Plasmodium falciparum* merozoite antigens supports MSPDBL2 as a predicted target of naturally acquired immunity. *Infection and Immunity* 81, pp. 3835–3842. doi: 10.1128/iai.00301-13.

Tetteh, K.K.A., Stewart, L.B., Ochola, L.I., Amambua-Ngwa, A., Thomas, A.W., Marsh, K., Weedall, G.D. and Conway, D.J. 2009. Prospective identification of malaria parasite genes under balancing selection. *PLoS ONE* 4, e5568. doi: 10.1371/journal.pone.0005568.

Tonkin-Hill, G.Q., Trianty, L., Noviyanti, R., Nguyen, H.H.T., Sebayang, B.F., Lampah, D.A., Marfurt, J., Cobbold, S.A., et al. 2018. The *Plasmodium falciparum* transcriptome in severe malaria reveals altered expression of genes involved in important processes including surface antigen–encoding *var* genes. *PLoS Biology* 16, e2004328. doi: 10.1371/journal.pbio.2004328.

Trape, J.F. 2001. The public health impact of chloroquine resistance in Africa. *American Journal of Tropical Medicine and Hygiene* 64, pp. 12–17. doi: https://doi.org/10.4269/ajtmh.2001.64.12.

Van Biljon, R., Van Wyk, R., Painter, H.J., Orchard, L., Reader, J., Niemand, J., Llinás, M. and Birkholtz, L.M., 2019. Hierarchical transcriptional control regulates Plasmodium falciparum sexual differentiation. *BMC genomics*, 20, pp.1-16.

Van Tyne, D., Park, D.J., Schaffner, S.F., Neafsey, D.E., Angelino, E., Cortese, J.F., Barnes, K.G., Rosen, D.M., et al. 2011. Identification and functional validation of the novel antimalarial resistance locus PF10_0355 in *Plasmodium falciparum*. *PLoS Genetics* 7, e1001383. doi: 10.1371/journal.pgen.1001383.

Van Tyne, D., Uboldi, A.D., Healer, J., Cowman, A.F. and Wirth, D.F. 2013. Modulation of PF10-0355 (MSPDBL2) alters *Plasmodium falciparum* response to antimalarial drugs. *Antimicrobial Agents and Chemotherapy* 57, pp. 2937–2941. doi: 10.1128/AAC.02574-12.

Usui, M., Prajapati, S.K., Ayanful-Torgby, R., Acquah, F.K., Cudjoe, E., Kakaney, C., Amponsah, J.A., Obboh, E.K., et al. 2019. *Plasmodium falciparum* sexual differentiation in malaria patients is associated with host factors and GDV1-dependent genes. *Nature Communications* 10, 2140. doi: 10.1038/s41467-019-10172-6.

Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K., 2006. A map of recent positive selection in the human genome. *PLoS biology*, *4*.

Volkman, S.K., Sabeti, P.C., Decaprio, D., Neafsey, D.E., Schaffner, S.F., Milner, D.A., Daily, J.P., Sarr, O., et al. 2007. A genome-wide map of diversity in *Plasmodium falciparum*. *Nature Genetics* 39, pp. 113–117. doi: 10.1038/ng1930.

Vythilingam, I., Lim, Y.A.L., Venugopalan, B., Ngui, R., Leong, C.S., Wong, M.L., Khaw, L., Goh, X., et al. 2014. *Plasmodium knowlesi* malaria an emerging public health problem in Hulu Selangor, Selangor, Malaysia (2009-2013): epidemiologic and entomologic analysis. *Parasites & Vectors* 7, 436. Available at: https://doi.org/10.1186/1756-3305-7-436.

Vythilingam, I., Noorazian, Y.M., Huat, T.C., Jiram, A.I., Yusri, Y.M., Azahari, A.H., Norparina, I., Noorrain, A., et al. 2008. *Plasmodium knowlesi* in humans, macaques and mosquitoes in peninsular Malaysia. *Parasites and Vectors* 1, 26. doi: 10.1186/1756-3305-1-26.

Vythilingam, I., Wong, M.L. and Wan-Yussof, W.S. 2018. Current status of *Plasmodium knowlesi* vectors: a public health concern? *Parasitology* 145, pp. 32–40. doi: 10.1017/s0031182016000901.

Wang, J., Chen, L., Chen, Z. and Zhang, W. 2015. RNA-seq based transcriptomic analysis of single bacterial cells. *Integrative Biology* 7, pp. 1466–1476. doi: 10.1039/c5ib00191a.

Wang, Z., Gerstein, M. and Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* 10, pp. 57–63. doi: 10.1038/nrg2484.

Wellems, T.E. and Plowe, C. V 2001. Chloroquine-resistant malaria. *The Journal of Infectious Diseases* 184, pp. 770–776.

William, T., Menon, J., Rajahram, G., Chan, L., Ma, G., Donaldson, S., Khoo, S., Fredrick, C., et al. 2011. Severe *Plasmodium knowlesi* malaria in a tertiary care hospital, Sabah, Malaysia. *Emerging Infectious Diseases* 17, pp. 1248–1255. doi: 10.3201/eid.1707.101017.

William, T., Rahman, H.A., Jelip, J., Ibrahim, M.Y., Menon, J., Grigg, M.J., Yeo, T.W., Anstey, N.M., et al. 2013. Increasing incidence of *Plasmodium knowlesi* malaria following control of *P. falciparum* and *P. vivax* malaria in Sabah, Malaysia. *PLoS Neglected Tropical Diseases* 7, e2026. doi: 10.1371/journal.pntd.0002026.

Wood, D.E. and Salzberg, S.L. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* 15, R46. doi: 10.1186/gb-

2014-15-3-r46.

Wootton, J.C., Feng, X., Ferdig, M.T., Cooper, R.A., Mu, J., Baruch, D.I., Magill, A.J. and Su, X.Z. 2002. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* 418, pp. 320–323. doi: 10.1038/nature00813.

World Health Organisation 2017. *Outcomes from the Evidence Review Group on Plasmodium knowlesi*. Available at: https://www.who.int/malaria/mpac/mpac-mar2017-*Plasmodium*-knowlesi-presentation.pdf.

World Health Organisation 2018. *World Malaria Report*. doi: ISBN 978 92 4 156483 0.

World Health Organization 2016. *Artemisinin and artemisinin-based combination therapy resistance: status report (No. WHO/HTM/GMP/2016.11)*.

Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., et al. 2014. Quantitative assessment of single-cell RNA-sequencing methods. *Nature methods* 11, pp. 41–46. doi: 10.1038/nmeth.2694.

Young, J.A., Fivelman, Q.L., Blair, P.L., De La Vega, P., Le Roch, K.G., Zhou, Y., Carucci, D.J., Baker, D.A., et al. 2005. The *Plasmodium falciparum* sexual development transcriptome: A microarray analysis using ontology-based pattern identification. *Molecular and Biochemical Parasitology* 143, pp. 67–79. doi: 10.1016/j.molbiopara.2005.05.007.

Yuda, M., Iwanaga, S., Kaneko, I. and Kato, T. 2015. Global transcriptional repression: An initial and essential step for *Plasmodium* sexual development. *Proceedings of the National Academy of Sciences* 12, pp. 12824–12829. doi: 10.1073/pnas.1504389112.

Yusof, R., Lau, Y.L., Mahmud, R., Fong, M.Y., Jelip, J., Ngian, H.U., Mustakim, S., Hussin, H.M., et al. 2014. High proportion of *knowlesi* malaria in recent malaria cases in Malaysia. *Malaria journal* 13, 168. doi: 10.1186/1475-2875-13-168.

Zhu, L., Mok, S., Imwong, M., Jaidee, A., Russell, B., Nosten, F., Day, N.P., White, N.J., et al. 2016. New insights into the *Plasmodium vivax* transcriptome using RNA-Seq. *Scientific Reports* 6, 20498. doi: 10.1038/srep20498.

Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. and Siebert, P.D. 2001. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques* 30, pp. 892–897. doi: 10.2144/01304pf02.

8 Appendices

Appendix 1. Genes with suspected roles in gametocytogenesis based on relevant publications (separate Excel spreadsheet on CD-ROM)

Appendix 2. Tajima's D values for all *P. knowlesi* Cluster 3 genes (separate Excel spreadsheet on CD-ROM).

Appendix 3. Genomic coordinates and genes within four regions of extended haplotype homozygosity in the *P. knowlesi* population in peninsular Malaysia (separate Excel spreadsheet on CD-ROM).

Appendix 4. Genomic co-ordinates of SNPs with elevated Rsb values for Cluster 3 vs. Cluster 1 (**A**), Cluster 3 vs. Cluster 2 (**B**), and Cluster 3 sub-cluster A vs. sub-cluster B (**C**) (separate Excel spreadsheet on CD-ROM).

Appendix 5. F_{WS} values for all *P. knowlesi* clinical isolates (separate Excel spreadsheet on CD-ROM).

Appendix 6A. Genes with increased expression in clinical isolates with >1% MSPDBL2-positive parasites compared with isolates with <1% MSPDBL2-positive parasites (separate Excel spreadsheet on CD-ROM).

Appendix 6B. Genes with increased expression in clinical isolates with >3% MSPDBL2-positive parasites compared with isolates with <3% MSPDBL2-positive parasites (separate Excel spreadsheet on CD-ROM).

Appendix 6C. Genes with increased expression in clinical isolates with >3% MSPDBL2-positive parasites compared with isolates with <1% MSPDBL2-positive parasites (separate Excel spreadsheet on CD-ROM).

Appendix 7A. Genes with decreased transcription in clinical isolates with >1% MSPDBL2-positive parasites compared with isolates with <1% MSPDBL2-positive parasites.

Gene ID	log2FoldChange	P-value	Product Description
PF3D7_0211600	-1.67	6.20E-03	UDP-N-acetylglucosamine transferase subunit
PF3D7_0308000	-1.30	8.38E-03	DNA polymerase delta small subunit, putative
PF3D7_0318700	-2.73	1.10E-03	conserved *Plasmodium* protein, unknown function
PF3D7_0408000	-2.41	3.67E-03	conserved *Plasmodium* protein, unknown function
PF3D7_0608200	-1.47	3.58E-03	conserved *Plasmodium* protein, unknown function
PF3D7_0611700	-2.18	2.86E-03	60S ribosomal protein L39
PF3D7_0623800	-1.71	6.82E-03	tyrosine kinase-like protein, putative
PF3D7_0721900	-1.72	5.07E-03	V-type ATPase V0 subunit e, putative
PF3D7_0819800	-2.02	3.82E-04	conserved *Plasmodium* protein, unknown function
PF3D7_0825200	-1.57	7.88E-03	translation initiation factor IF-3
PF3D7_1024100	-1.91	6.97E-03	conserved *Plasmodium* protein, unknown function
PF3D7_1110000	-1.59	7.89E-03	conserved *Plasmodium* protein, unknown function
PF3D7_1325500	-1.14	7.52E-03	conserved *Plasmodium* protein, unknown function
PF3D7_1349700	-1.96	4.49E-03	peptidase, putative
PF3D7_1426900	-1.53	7.09E-03	cytochrome b-c1 complex subunit 6, putative
PF3D7_1428500	-1.30	6.82E-03	protein kinase, putative
PF3D7_1463600	-1.93	1.43E-03	conserved *Plasmodium* protein, unknown function

Appendix 7B. Genes with decreased transcription in clinical isolates with >3% MSPDBL2-positive parasites compared with isolates with <3% MSPDBL2-positive parasites.

Gene ID	log2FoldChange	P-value	Product Description
PF3D7_0906600	-2.56	1.18E-03	Zinc finger protein, putative
PF3D7_1111300	-2.63	6.62E-03	Protein transport protein BOS1, putative
PF3D7_1409600	-1.86	1.07E-03	Conserved *Plasmodium* protein, unknown function
PF3D7_1439500	-1.40	5.65E-03	oocyst rupture protein 2, putative

Appendix 7C. Genes with decreased transcription in clinical isolates with >3% MSPDBL2-positive parasites compared with isolates with <1% MSPDBL2-positive parasites.

Gene ID	log2FoldChange	P-value	Product description
PF3D7_0814500	-2.46	9.23E-03	conserved protein, unknown function
PF3D7_0819800	-1.89	7.22E-03	conserved *Plasmodium* protein, unknown function
PF3D7_0906600	-2.34	3.04E-03	zinc finger protein, putative
PF3D7_1111300	-2.59	8.89E-03	protein transport protein BOS1, putative
PF3D7_1349700	-2.35	6.86E-03	peptidase, putative
PF3D7_1409600	-1.87	1.45E-03	conserved *Plasmodium* protein, unknown function
PF3D7_1439500	-1.36	9.59E-03	oocyst rupture protein 2, putative
PF3D7_1463600	-2.05	8.49E-03	conserved *Plasmodium* protein, unknown function

Appendix 8. Genes with higher expression correlated to ranked MSPDBL2 protein expression at $P < 0.01$ (separate Excel spreadsheet on CD-ROM).

Appendix 9. Genes with lower expression correlated to ranked MSPDBL2 protein expression at $P < 0.01$ (separate Excel spreadsheet on CD-ROM).

Appendix 10. Genes with decreased transcription correlated to increased *mspdbl2* transcript levels at a P-value of <0.001 (separate Excel spreadsheet on CD-ROM).