Methods for analysis of deep sequencing data from mixtures of *Plasmodium falciparum* clones or stage-specific transcriptomes

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Anita Lerch

aus

Wynigen BE

Basel, 2018

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Ingrid Felger und Prof. Dr. Mark D. Robinson

Basel, den 27. März 2018

Prof. Dr. Martin Spiess

Dekan

To those persons
who encouraged me
to do this thesis

TABLE OF CONTENT

SUMMARY

Malaria is a life-threatening infectious disease caused by *Plasmodium* parasites transmitted to humans through bites of infected *Anopheles* mosquitos. An estimated 445,000 people die every year by an infection with *Plasmodium* parasites, most of them children living in sub-Saharan Africa. As a result of increased malaria control, the mortality was greatly reduced in the last decades. To develop new tools for elimination and to evaluate the impact of control, a good understanding of the epidemiology and biology of malaria parasites is required.

Studies of infection and transmission dynamics of *Plasmodium* parasites were greatly improved by distinguishing individual parasite clones and monitoring their infection dynamics over time. In regions with high transmission of *Plasmodium* parasites, individuals are often infected with several clones concurrently. Individual parasites clones can be identified by genotyping. The current standard method used for genotyping is amplification of highly length-polymorphic merozoite surface protein 2 (*msp2*) or other antigen genes followed by sizing of the amplicon by capillary electrophoresis (CE). The sensitivity to detect low-abundant clones (minority clones) of *msp2*-CE genotyping is however limited, resulting in an underestimation of multiplicity of infection (MOI). A shortfall of this genotyping method is that frequency of individual clones within a sample cannot be determined. This urges the search for new genotyping methods that rely on sequencing of genomic fragments with extensive single nucleotide polymorphism (SNP).

Improvement in next generation sequencing (NGS) technologies permitted the use of amplicon sequencing (Amp-Seq) in epidemiological studies. Genotyping by amplicon sequencing has a higher sensitivity to detect minority clones, can quantify the frequency of each clone within a sample, and allows the use of SNP polymorphic markers. In the frame of this thesis, a new Amp-Seq genotyping assay was developed, including known SNP polymorphic markers, and novel marker '*cpmp*', as well as a bioinformatic analysis workflow. This genotyping assay was applied on field samples from a longitudinal study conducted in Papua New Guinea. A comparison to *msp2*-CE genotyping confirmed the higher sensitivity to detect minority clones by Amp-Seq genotyping method and showed a significant underestimation of MOI by classical size polymorphic marker. However, no significant increase in molecular force of infection ($_{\mathrm{mol}}$FOI), i.e. number of new infections per individual per year, was observed.

Quantification of the frequency of individual clones in longitudinal samples permitted to infer multi-locus haplotypes. Multi-locus haplotypes increased discriminatory power of genotyping and robustly distinguished new infections from those detected in an individual earlier. For calculating the density of clones from multi-clone infections the within-host clone frequency is multiplied by parasitaemia of this infection determined by quantitative PCR. Density of individual parasites clones in multi-clone infections over time is a new parameter for epidemiological studies. It will permit to study the dynamics, and thus fitness, of parasite clones exposed to within-host competition or to acquired natural immunity.

NGS also gained great importance in gene expression studies of *Plasmodium* parasites in patient samples. Transcriptome studies are complicated by the mixture of different developmental stages present concurrently in samples collected from patients. Even in *in vitro* cultured samples after tight synchronisation or enrichment of a specific developmental stage, small fractions of other development stages are still found. This problem is of particular relevance for *P. vivax*, as the absence of continuous *in vitro* culture so far has hampered the study of isolated parasite stages. For example, the transcriptome of *P. vivax* gametocytes, one of the stages found in peripheral blood and infective to mosquitos, has not yet been described.

A solution for differentiating mixed transcription may come from deconvolution methods, which either infer the stage proportion in samples or stage-specific transcriptome signatures. A large selection of different deconvolution methods has been developed for the analysis of heterogeneous tissues, e.g. cancer tissues or hematopoietic cell, but these methods have rarely been applied to mixed stages of malaria parasites. The best suited combination of normalisation and deconvolution methods for analysis of RNA sequencing (RNA-Seq) data from mixed-stage samples of *Plasmodium* parasites was evaluated based on experimentally mixed highly synchronised blood stages. Normalisation by count per million and deconvolution with a negative binomial regression model followed by selection of genes with significant fold change resulted in the best agreement with transcriptomes as observed in single stages. This strategy can easily be transferred to *Plasmodium* field samples with known stage proportions. This analysis performed in cultured parasites of defined mixed stages served as proof-of-concept and confirmed that identification of stage-specific genes is feasible also in field samples, notably in species that cannot be cultivated, such as *P. vivax*.

NGS permits fundamentally new approaches to study *Plasmodium* parasites. This thesis presents a novel marker and data analysis platform for highly sensitive *P. falciparum* genotyping. Furthermore, a best practice workflow was identified to infer stage-specific gene expression from parasite infections consisting of mixed developmental stages. This provides a crucial tool for the analysis of gene expression data generated from *Plasmodium* field samples.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support of many people. First and foremost, I would like to thank my supervisor Ingrid Felger for giving me the opportunity to do this PHD with her at the Swiss and Public Health Institute. In particular, I would like to thank her for her flexibility that allowed me to do large part of this PhD at WEHI, and to spend the final months in California. Furthermore, I am grateful for her openness to accept me with a background of bioinformatics into her group of molecular epidemiologists. Ingrid has been a great mentor and she always put me back on track whenever I got lost in details.

I also would like to thank Ivo Mueller who suggested to do this PhD jointly supervised with Ingrid Felger and himself. He has been a tremendous help in getting everything organized and started once we had agreed to do this PhD.

Further I would thank Liam O'Connor who offered to supervise the systems biological and computational part of this thesis. Liam also provided office space in his group and computational infrastructure. His guidance greatly helped to make better use of my knowledge as a computational scientist by pointing out the right methods do be used for the analysis. I also appreciate his good sense of humour, which included making me aware of my very Swiss attitudes.

Special thanks go to my partner Cristian Köpfli, who encouraged me to do this thesis. In addition to his own work, he spent a lot of time to discuss my projects and raised essential questions that helped to improve my work. He was incredible patient and tried to cheer me up whenever I was frustrated when things did not progress.

Special thanks goes do Sebastian Rusch, Natalie Hoffman, Armin Passecker and Camilla Messerli, who conducted laboratory work for the experimental mixed transcriptome, and to Stephen Wilcox who supervised sequencing and helped with the complex multiplexing of the amplicon sequencing libraries.

Further, I would like to thank Melanie Bahlo, Gordon Smyth, Tony Papenfuss, Peter Pemberton, Amanda Ross, Till Voss, Nicolas Brancucci, Connie Li Wai Suen, Saskia Freytag and Peter Beck, they all took time to discuss specific issues of the data analysis with me.

A would thank to Connie, Sofonias, Camila, Abby and all other friends at WEHI for having a lot of good times. Their friendship helped to make Melbourne our home away from home. I thank all members of the SwissTPH especially the group MolDia, MolPar and GeneReg, who welcomed me every time I was back in Switzerland.

I also thank our new friends in California, in particular our neighbour Irene Vorontsova for help with proofreading of the thesis. I also thank her and her husband Haylesh to apply their barista skills to supply us with excellent coffee when an extra boost of energy was needed.

With the upcoming defense in mind I would thank my PhD committee: Mark Robinson as external co-referee Pascale Mäser as adviser and Ivo Mueller as expert. Further, I would like to thank Maria Grünberg for printing my thesis and taking it to the university while I was in the USA.

Many people have helped to make it possible to successfully complete this PhD while moving across several continents and raising a small child. In particular, I would like to thank to my mother Hanni Lerch and Cristian's parents Verena and Georges for looked after my son Fabrice. A big thank to Herbert and Ursula, who gave us a home for the whole family when we were in Switzerland and helped with many small, but essential organisational issues. A very special thanks goes to our nanny Julieta. Having someone so dedicated and reliable looking after Fabrice while I was writing my thesis was an invaluable support.

And last but not least I thank the Swiss National Science Foundation, Novartis Foundation for Medical-Biological Research and Stiftung Emilia-Guggenheim-Schnurr for financial support, as well as the Swiss Society of Tropical Medicine and Parasitology and International PhD Program in Infection Biology for their financial support to attend conference and courses.

ABBREVIATIONS

ACT	artemisinin-based combination therapy
ama1	apical membrane protein 1
Amp-Seq	amplicon sequencing
ART	artesunate
bp	base pair
CE	capillary electrophoresis
cpmp	PF3D7_0104100, conserved Plasmodium membrane protein
CQ	chloroquine
csp	circumsporozoite protein
DE	Differential expression (DE)
DGE	Differential gene expression
DNA	deoxyribonucleic acid
DTT	dichloro-diphenyl-trichloroethane EIR entomological inoculation rate
ESTs	expressed sequence tags
FACS	Fluorescence activated cell sorting
G6PD	glucose-6-phosphatase dehydrogenase glurp glutamate-rich protein
glurp	glutamine rich protein
H_e	expected heterozygosity
HRP	histidine-rich protein
IRS	indoor residual spraying
ITN	insecticide-treated bednet
LD	linkage disequilibrium
LLIN	long-lasting insecticide-treated bednet
LM	light microscopy
MalariaGEN	Malaria Genomic Epidemiology Network
MIPs	Molecular inversion probes
MOI	multiplicity of infection
$_{mol}$FOI	molecular force of infection
msp	merozoite surface protein
ORF	open reading frame
PCR	polymerase chain reaction
pPCR	primary PCR
nPCR	nested PCR

qPCR	quantitative PCR
qRT-PCR	quantitative reverse-transcription PCR
PfEMP1	Plasmodium falciparum erythrocyte membrane protein 1
pfs25/pvs25	Plasmodium falciparum/ Plasmodium vivax 25 kDa ookinete surface antigen
pLDH	Plasmodium lactate dehydrogenase
PNG	Papua New Guinea
PNG IMR	Papua New Guinea Institute of Medial Reasearch
PQ	Primaquine
RBC	red blood cell
RDT	rapid diagnostic test
RFLP	restriction fragment length polymorphism
(m)RNA	(messenger) ribonucleic acid
RNA-Seq	RNA sequencing
scRNA-Seq	Single cell RNA sequencing
SNP	single nucleotide polymorphism
Swiss TPH	Swiss Tropical and Public Health Institute
TARE-2	telomere-associated repeat element 2
UID	Unique identifier
varATS	var gene acidic terminal sequence
WGS	Whole genome sequencing

CHAPTER 1: INTRODUCTION

1.1 MALARIA

Malaria is a life-threatening infectious disease caused by *Plasmodium* parasites. *Plasmodia* are transmitted to humans through bites of infected *Anopheles* mosquitoes. Today, no human should die from an infection with *Plasmodium,* as an infection with *Plasmodium* parasites can be prevented and treated [1]. However, in 2016, about 216 million cases of malaria resulting in an estimated 445,000 deaths were reported worldwide by the World Health Organization (WHO) (Figure 1) [2]. Most deaths occur among children living in Africa [3]. Even though mortality was reduced by half in the last 10 years (881,000 deaths in 2006 [4]), more has to be done to reach zero mortality. To achieve this goal, further research into the epidemiology and biology of malaria parasites is required.

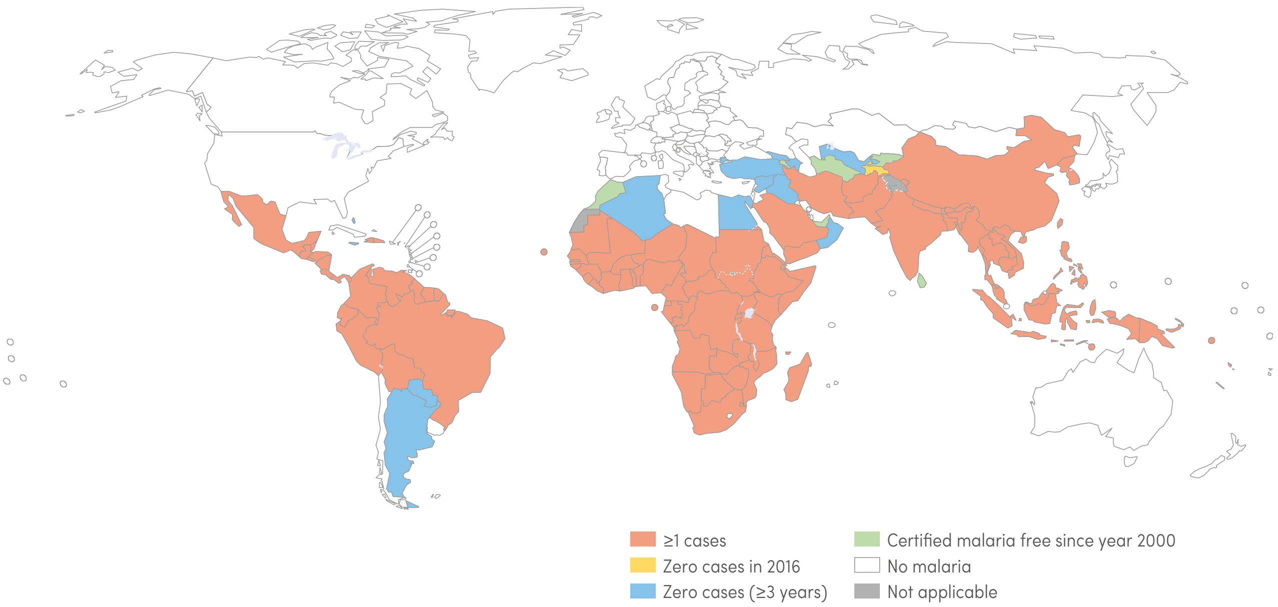


Figure 1: World map of indigenous cases of *Plasmodium* infection (source World Malaria Report 2017)

Human malaria exists since pre-historical times and the associated fever was already described in ancient times in China, the Middle East, India and the Mediterranean area [5]. But the parasites causing this fever were only discovered in the 19th century [5]. At the beginning of the 20th century first attempts to control malaria were undertaken by minimising mosquito to human contacts to prevent transmission [5]. At the same time, efforts were undertaken to decrease the mosquito breeding sites by draining marshes [6]. In the 1940s, with the development of residual insecticides and synthetic anti-malarials further achievements in malaria control were made [7,8]. Encouraged by this success, in 1955 the WHO formulated a plan for worldwide malaria eradication (Global Malaria Eradication Programme, GMEP) [6,9], resulting in elimination of *Plasmodium* in Europa and USA [6]. The GMEP was stopped in 1969 [6]. The emergence of mosquito resistance to insecticides and parasite resistance to anti-malarials was one reason of the eradication campaign failure [7,8,10]. Another reason was that little effect was achieved in some continental tropical countries of Asia, South America and Africa [9]. The operational logistic was often too complex for countries with weak infrastructure [9]. The subsequent weakening of the control efforts resulted in a resurge of malaria [6]. Since then, several programs were launched and organisations founded to coordinate the control of malaria globally, e.g. the Roll Back Malaria (RBM) partnership [9]. Global eradication of malaria was put back on the global agenda in 2007 when Bill and Melinda Gates announced not just to treat malaria or to control it, but to plan a long-term course to eradication [11].

Today, the main strategies to control malaria are to prevent transmission by indoor residual spraying (IRS), insecticide-treated bed nets (ITNs), and rapid treatment of the patient. However, gametocytes, that represent the human infective reservoir to mosquitoes, are only partially cleared by artemisinin combination therapy (ACT) [12]. Only treatment with low-dose Primaquine (PQ) clears *P. falciparum* gametocytes [12,13]. In addition, adults in endemic countries infected with *Plasmodium* are often asymptomatic and therefore undiagnosed [14]. Currently it is unclear how much they contribute to the transmission from human to mosquitoes, but infection of mosquitos feeding on asymptomatic individuals has been reported [15]. A vaccine might be a key tool for malaria elimination. After decades of research, the first vaccine against *P. falciparum* sporozoites RTS,S completed clinical trial phase III and was approved by the European Medicines Agency in 2015. However, the efficacy of RTS,S for children between 5-17 months is only ~50% [16].

To reach malaria elimination, a better understanding of the epidemiology and molecular biology of the parasite is needed. In the last decade, next-generation sequencing technology (NGS) has permitted fundamentally new approaches to study biology, and it also has great potential to study infectious diseases. NGS approaches applied to malaria parasites, however, yield unique challenges to data analysis. In this thesis, novel approaches are presented to analyse NGS data from isolates containing mixed clones or mixed parasite life stages.

1.1.1 *Plasmodium* species

Plasmodium parasites are Protozoa belonging to the phylum apicomplexa. They evolved over thousands of years and are widespread in reptiles, birds and mammals [5]. All *Plasmodium* parasites need two hosts in their life cycle, a dipteran insect host and a vertebrate host [5]. The sexual reproduction occurs always in the insect host. Over 250 *Plasmodium* species are known to infect vertebrates [17]. More than one hundred of these are transmitted by mosquitoes

Five *Plasmodium* species are known to cause malaria in humans: *P. falciparum, P. vivax*, *P. malariae*, *P. knowlesi*, *P. ovale* (with subspecies *P. ovale wallikeri* and *P. ovale curtisi*) (Figure 2). Of these 6 species, *P. falciparum* and *P. vivax* are by far the most prevalent ones. *P. falciparum* occurs worldwide and is the predominant species in Africa. *P. falciparum* is almost exclusively responsible for malaria mortality (99% of deaths) [2]. *P. vivax* predominates in Latin America, India and South East Asia, and threatens almost 40% of the world's population [18]. All human *Plasmodium* species are transmitted by the Anopheles mosquito [19].

The high mortality and morbidity of Malaria had selective pressure on the human genome. Several genetic modification evolved that give a certain degree of protection against infection or severe malaria, like sickle cell disease, thalassaemia, glucose-6-phosphate dehydrogenase (G6PD) deficiency and the absence of Duffy antigens on red blood cells [20,21].

1.1.2 Life cycle of the human malaria parasites

Plasmodia have a complex life cycle (Figure 3) [19,22]. The first step of a human infection, the exo-erythrocytic cycle (duration ~8 days), starts with the bite of an infected mosquito vector [23,24]. The mosquito injects sporozoites into the dermis of the skin, where they transmigrate the dermal tissues to reach small blood vessels and via circulating blood migrate to the liver. In the liver, sporozoites invade liver cells and develop into liver trophozoites. The trophozoite develops further into a schizont, which consist of thousands of merozoites. Upon infection of the liver with sporozoites, *P. vivax, P. cynomolgi and P. ovale* form additional dormant stages called 'hypnozoites' [19,25,26]. The hypnozoites cause clinical relapses weeks to months after the first infection.

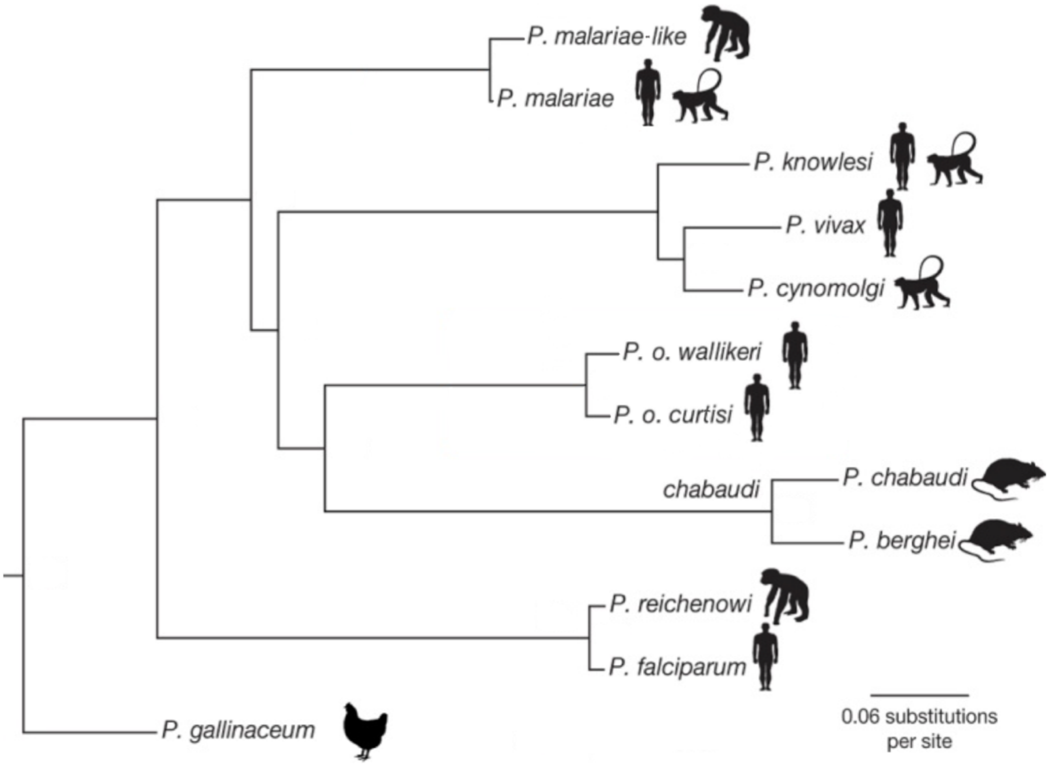


Figure 2: Maximum likelihood phylogenetic tree of *Plasmodium* genus. Silhouettes show host of the different species. (image source Rutledge et al. 2017).

Erythrocytic cycle (duration ~48h), begin with release of the liver merozoites into the blood stream. In the blood stream, the merozoites invade red blood cells (RBC), where they develop in ~32h into trophozoites [19,27]. The trophozoites develop further into schizonts, which contain 12-32 merozoites [19]. These merozoites are then again released into the blood stream, and invade new RBCs.

Some of the merozoites develop in the RBC into male or female gametocytes. With the ingestion of gametocytes during the blood meal by a female mosquito (2-5µl of blood [28]), the sexual cycle begins, called sporogonic cycle. In the midgut of the mosquito the ingested female and male gametocytes develop into macrogametes and 8 microgametes formed from the microgametocytes by exflagellation. After fertilization a diploid zygote is formed which further develops in to an ookinete. The ookinete transmigrates the peritrophic membrane and midgut epithelium. For about 2 weeks the parasite remains underneath the basal membrane of the midgut and develops into an oocyst which contains thousands of sporozoites finally released into the haemocoel of the mosquito [29]. Sporozoites migrate to the salivary glands and dozens or up to a few hundred are injected into the dermis of the human skin when the mosquito takes a next blood meal.

Hypnozoites present a particular challenge to the control and elimination of *P. vivax*, as drugs against blood stages do not target them, which results in frequent relapses. Hypnozoites can only be cleared with the drug Primaquine (PQ). Treatment with PQ lasts for 14 days and can cause haemolysis in patients with G6PD deficiency [30], which is prevalent across most of the malaria-endemic countries [31]. Development of new drugs targeting the hypnozoite stage is therefore urgently needed.

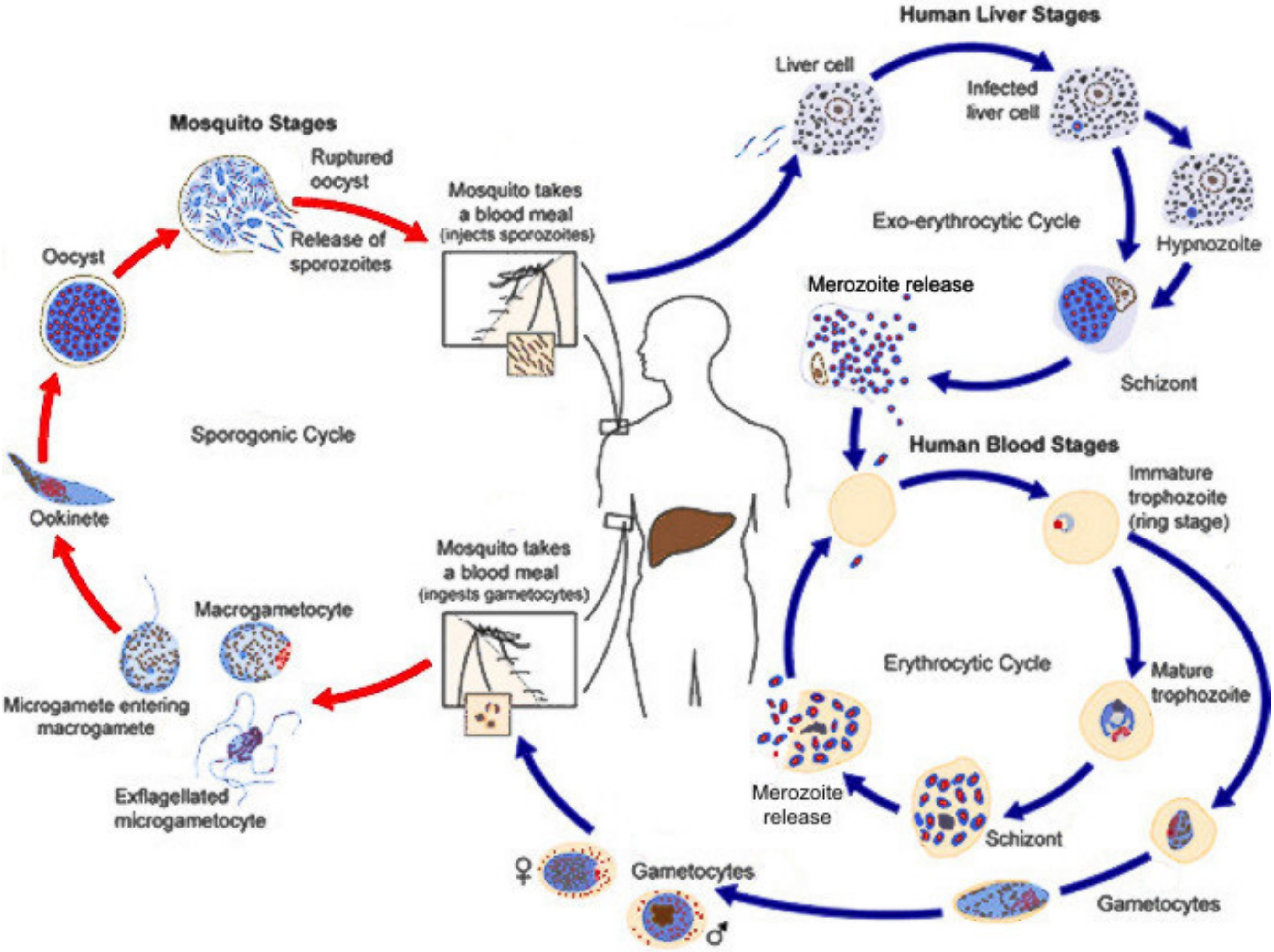


Figure 3: Life cycle of *Plasmodium vivax* (image modified from CDC)

1.2 Molecular Epidemiology of Malaria

1.2.1 Molecular Epidemiological Parameters

Identification of individual clones and monitoring them in the course of an infection is an important aspect of epidemiological studies on parasite infection and transmission dynamics. Several parameters are used to describe the dynamics of malaria ~~i~~nfectious and to measure the outcome of interventions. Important molecular epidemiological parameters for *Plasmodium* infections are multiplicity of infection (MOI), duration of infection, and molecular force of infection ($_{\text{mol}}$FOI) [32].

Multiplicity of infection (MOI) is defined as the number of co-infecting parasite clones. Individuals in countries with high transmission of *Plasmodia* are often infected with several clones concurrently [33,34] This superinfection can be caused by multiple infective mosquito bites or by a single mosquito bite injecting multiple genetically distinct parasite clones.

Molecular force of infection ($_{\text{mol}}$FOI) is defined as the number of genetically distinct new infections acquired over time. It is a measure of exposure. It provides a robust measure of transmission as it differentiates between persistent and new infections. Longitudinal studies are needed to determine $_{\text{mol}}$FOI.

Duration of infection for untreated *Plasmodium* infections is defined as the time from the first observation of a parasite clone in the blood until clearance of this clone by the human immune system. The duration of infection depends on the age, it was found to be longest in children of 5-9 years with a duration of ~180 days [35].

Parasitaemia is defined as the parasite load respective density in the blood. Parasite density is either determined by light microscopy (LM) of Giemsa-stained blood smears (limit of detection ~100 parasites per µl of blood), or by qPCR of single- or multi-copy genes (limit of detection ~3 parasite per µl of blood or lower)[36]. The parasite density in the blood of an infected individual is influenced by several factors [37]. For example, the parasite load depends on: (i) the acquired immunity of the host with children often showing higher parasite densities; (ii) the duration of an infection, with longer persisting infection showing lower parasite densities; and (iii), for *P. falciparum*, the stage of the parasite within its 48 h cycle, as the mature blood stages are sequestered in inner organs and therefore apparently absent in peripheral blood.

Duration of infection, $_{\text{mol}}$FOI, and MOI are all determined by genotyping individual clones and therefore depend on the limit of detection of the assays to diagnose and genotype infections [38–40]. The duration of a clonal infection and $_{\text{mol}}$FOI are difficult to determine if the density of individual parasite clones is around the limit of detection, and they frequently escape detection. This imperfect detection must be distinguished from parasite clearance and reinfection with a genetically indistinguishable clone as it biases the estimates of $_{\text{mol}}$FOI and duration of infection. Modelling approaches are therefore used to estimate $_{\text{mol}}$FOI and duration of infection [35,41,42].

1.2.2 Genotyping of *Plasmodium* parasites

Individual parasite clones are identified by genotyping. Genotyping is not only used to determine MOI, $_{\text{mol}}$FOI or duration of infection, but also to study population structure or phenotypes like drug resistance. Depending on the genotyping application, different marker sets are selected [43,44]. A single marker of high resolution is often sufficient for epidemiological studies where individual clones need to be identified. For studying phenotypes like drug resistance, markers covering all mutations (e.g. several SNPs within a gene, or several genes) associated with resistance must be typed. In population genetics studies, multiple genome-wide markers are required that are unlinked from each other and not under selection pressure. For recrudescence typing in anti-malarial drug efficacy trials the use of three unlinked markers with high resolution are recommended by the WHO [45].

The first methods to genotype *P. falciparum* used amplification of the highly length-polymorphic merozoite surface protein 2 (*msp2*) and subsequent sizing either by full length fragment or by restriction fragment length polymorphism (RFLP) [46–49]. In 2006, PCR-RFLP was modified to capillary electrophoresis (CE). This change increased resolution by using different fluorescent-labels for the FC27 and 3D7 allelic families [50,51]. CE simplified analysis by omitting the interpretation of the RFLP size pattern which was difficult to analyse, especially when RFLP size patterns of multiple concurrent clones were superimposed. Currently, the recommended marker and method for genotyping in drug trials is merozoite surface protein 1 (*msp1*), *msp2* and glutamine rich protein (*glurp*) by CE [45].

Another genotyping method is typing of 24-42 SNPs (SNP barcode) that are distributed over the whole genome. This multi-locus SNP-typing can determine genome-wide diversity and is suited for population studies, as selected SNPs are unlinked to each other. Mutations of SNPs are determined by either High Resolution Melting, Oligonucleotide Ligation or TaqMan [52–54]. However, SNP-typing requires a lot of DNA template, as each SNP is typed as an independent assay. Another difficulty is the haplotype inference in case of multi-clone infections. The haplotypes of sample with mixed infection is difficult to resolve if the genotypes are unlinked to each other (see Section 1.4.1).

Improvement in next generation sequencing technologies (Illumina, 454/Roche or Ion Torrent) towards longer sequence reads and lower sequencing cost per sample by multiplexing of samples permitted the use of amplicon sequencing in epidemiological studies. Amplicon sequencing (Amp-Seq) genotyping has a higher sensitivity, quantifies proportion of different variants and can detect low-abundant clones (minority clones) in samples with multiple concurrent infections. However, the higher sensitivity of Amp-Seq comes at the cost of calling false alleles caused by sequencing error or PCR artefacts. First Amp-Seq genotyping of *P. falciparum* used the length polymorphic markers *msp1* and *msp2*, as well as the SNP polymorphic region of circumsporozoite protein (*csp*) [16,55,56].

In the past few years, whole genome sequencing (WGS) of single clone infections also became an option to determine genotypes. However, the cost per sample is high and the sequence library preparation is too labour intensive for large studies. For mixed clone infections, WGS is not feasible as the minority clone can only be detected at very high sequence costs. For example, to detect a minority clone in a sample at a within-host frequency of 1:500, at least 120Gb (25Mb genome size multiplied by 500-fold coverage) needs to be sequenced. This corresponds to one Illumina NextSeq run with a sequence cost of approximately USD4000.

A recent study showed a bias in size polymorphic genotyping towards the shorter fragments in samples with multiple concurrent infections [57]. The resulting underestimation of multiplicity of infection (MOI) urges the search for new SNP polymorphic marker genes. Amp-Seq of SNP polymorphic markers might represent the best alternative to genotype with size polymorphic markers. An earlier study claimed that Amp-Seq has a higher sensitivity to detect minority clones compared to *msp2*-CE genotyping [55], but nothing is known about the specificity of the method and how the higher sensitivity to detect minority clones impacts the molecular epidemiological parameter MOI, $_{\text{mol}}$FOI and duration of infection. A comprehensive comparison of *msp2*-CE genotyping and Amp-Seq genotyping with new markers was the topic of this thesis and can be found in more detail in Chapters 2 and 3.

1.3 Genomics and Transcriptomics of *Plasmodium* Parasites

1.3.1 Genomics

The genome of the human *Plasmodium* species encodes ~5000 genes on 14 chromosomes in ~25Gb nucleotides. *Plasmodia* also carry a mitochondrion and apicoplast genome [58]. First approaches to sequence *P. falciparum* and *P. vivax* was made by Sanger sequencing of expressed sequence tags (ESTs) from cloned cDNA fragments, leading to the discovery of more than 600 genes [59–62]. Later, whole chromosome shotgun Sanger sequencing method was used to sequence the genome of *P. falciparum* [63]. In short, individual chromosomes were separated, isolated and shared. The shared fragments were then cloned into yeast artificial chromosomes (YAC) and Sanger sequenced. The sequences were first assembled by YAC and then by chromosomes. The publication of *P. falciparum* genome enabled systematic analysis of the proteome and showed that a large proportion of genes were devoted to immune evasion and host-parasite interactions. Since then, the whole genome of all human malaria parasites were sequenced and published: *P. vivax* and *P. knowlesi* in 2008 [64,65] and *P. malariae* and *P. ovale* in 2017 [26]. Also the closest related *Plasmodium* species of *P. falciparum* and *P. vivax* were sequenced: the chimpanzee malaria parasites *P. reichenowi* and the monkey malaria parasite *P. cynomolgi* [25,66] (Figure 1).

Comparative genomics between the different *Plasmodium* species gave insight into the evolutionary history and showed that ~77% of the genes are orthologous and in conserved gene synteny [64]. Genes in synteny indicate a conserved metabolome, as they belong to the metabolic pathways, housekeeping and membrane transporter genes. Species-specific genes are located at syntenic break points and have mostly a host-parasite interaction function. Of the human *Plasmodium* species, only *P. falciparum* is routinely cultured for gene function studies. Comparative analysis of the genomes of other species can be used to identify group-specific genes associated with traits like development of hypnozoite e.g. *P. vivax*, *P. ovale* and *P. cynomolgi* or the ability to infect human and monkeys e.g. *P. knowlesi*, P. *malariae* and *P. cynomolgi* [67].

Efforts to describe the genetic variation *P. falciparum* and *P. vivax* were undertaken by the Malaria Genomic Epidemiology Network (MalariaGEN, https://www.malariagen.net) in 2005. Today >3,000 genomes of *P. falciparum* and >480 genomes of *P. vivax* are available from multiple publications [24,68–74], describing >900,000 SNPs of *P. falciparum* and >300,000 SNPs of *P. vivax*. In addition to SNPs, microsatellite-length polymorphisms, intragenic repeats and copy number variation add to the genetic diversity of *Plasmodia*. The high genetic variation of *Plasmodium* parasites is required for the immune evasion mechanism, but also represents adaptation to the human and mosquito host, or resulting from drug pressure [70,75].

1.3.2 Transcriptomics

The availability of the annotated whole genome sequences enables to study the whole transcriptome of *P. falciparum* and *P. vivax*. The annotated genes were discovered by scanning the whole genome for open reading frames (ORF) or by using EST sequences [63,64]. The first transcriptomes of the erythrocytic cycle of *P. falciparum* and *P. vivax* using a DNA microarray platform were both published shortly after the whole genome sequence [76–78]. Advances in high-throughput RNA sequencing (RNA-Seq) permitted the study of the transcriptome without knowledge of the underlying genomic sequence [79]. RNA-Seq of the *P. falciparum* and *P. vivax* transcriptomes improved the existing annotation for both species by identifying new genes, splice sites and splice variants [78–80].

The complex life cycle of *Plasmodia* with two different hosts and three different cycles (exo-erythrocytic, erythrocytic and sporogonic cycle) is transcriptionally and post-transcriptionally regulated [81,82]. Each stage of the life cycle has a characteristic gene expression pattern [76,77,83,84]. The transcriptome of *P. falciparum* shows a highly ordered cascade of gene expression over the parasite's life cycle produced by transcriptional regulation [76]. The highly ordered expression permits functional annotation of genes with so far unknown function by co-expression analysis, as genes with similar function are often co-expressed [77,85]. Furthermore, *Plasmodia* also have stage specific copies of ribosomal RNA [63,86–88].

Comparative analysis between the transcriptomes of *P. falciparum* and *P. vivax* explained differences in the biology of the two species [78]: For example, the genes for *P. vivax* immune evasion or red blood cell invasion mechanism differed substantially from those in *P. falciparum,* because most of those genes are not in syntenic order. This helps to explain why mature erythrocytic stages of *P. vivax* circulate in the peripheral blood, whereas they are sequestered in *P. falciparum*, or why *P. vivax* infects only reticulocytes. Furthermore, RNA-Seq of *P. vivax* also revealed unusually long 5′ untranslated regions and multiple transcription start sites [80].

Currently, gene expression data exist for every developmental stage of *Plasmodia* except for the oocyst stage in the mosquito. However, for none of the *Plasmodium* species the whole life cycle is covered (Table xy3). Most of the available transcriptomics data in PlasmoDB (http://plasmodb.org [89]) are used to study the gene regulation mechanism of *P. falciparum* erythrocytic cycle or specific phenotypes [90,91]. Basic research on *P. vivax* is greatly hampered by a lack of continuous *in vitro* parasite culture. The available *P. vivax* transcriptome data of the erythrocytic and sporogonic cycle (except sporozoite stage) originated from enriched and short-term cultured field samples [78,92]. In view of the difficulties in culturing *P. vivax,* the published transcriptome data may likely not fully represent the gene expression in the human host. For example, stress-related genes might be overexpressed, while genes required to escape the human immune system or clearance in the spleen might not be expressed. Moreover, the transcriptome data of *P. vivax* gametocytes and liver stages (developing liver schizonts and hypnozoites) are still not available.

P. vivax hypnozoites are a major problem for elimination. Hypnozoites cause relapses weeks to months after the initial infection and sustain transmission [93]. As a model organism to study *P. vivax* liver stages, the monkey malaria parasite *P. cynomolgi* is studied. Recently, transcriptomes of the *P. cynomolgi* liver schizont and hypnozoite were published [94,95]. Yet, the commitment to form hypnozoites is not understood, and may already be determined in the sporozoite. During the course of this thesis contributions to the study of *P. vivax* sporozoites transcriptome were made, which might yield novel drug targets. The manuscript of this additional project is presented in Chapter 6.

A better understanding of the *P. vivax* gametocyte transcriptome is highly relevant, as its development differs to the one of *P. falciparum*. *P. vivax* gametocytes develop much faster than *P. falciparum* gametocytes and appear in the peripheral blood before clinical symptoms occur [19,22,96]. In contrast, gametocytes of *P. falciparum* develop in 10-12 days sequestered in the bone marrow and start to circulate in the peripheral blood as mature gametocytes only after clinical symptoms occur [22,37].

A challenge for the study of developmental stage-specific gene expression is the mixture of different stages present in samples collected from patients. This is the case for clinical isolates of all species, e.g. when gametocytes and asexual blood stages are present. The problem applies particularly to *P. vivax*, as the absence of continuous *in vitro* culture prevents the study of isolated parasite stages. Methods to de-convolute transcriptomes from mixes stages will be of great help to understand *P. vivax* gametocyte development.

During the course of this thesis, methods to infer stage-specific gene expression were assessed using RNA-Seq data from experimentally mixed stages of highly synchronized *P. falciparum* culture (Chapter 4). In the future, these methods will be applied to infer the *P. vivax* gametocyte transcriptome from field samples containing a mixture of stages, which has been the far aim of this thesis.

Table xy3: Overview of published microarray or RNAseq transcriptome data for the two most important human malaria species, *P. falciparum* and *P. vivax*, as well as *P. cynomolgi*, which is closely related to *P. vivax*. NA, no transcriptome available.

Development stage	***P. falciparum*[1]**	***P. vivax***	***P. cynomolgi***
Exo-erythrocytic cycle:			[97][2]
Trophozoite	NA	NA	NA
Schizont	NA	NA	[94,95][3]
Hypnozoite	-	NA	[94,95] [3]
Erythrocytic cycle:	[98][4]		[97][2]
Merozoite	[77][2]	NA	NA
Ring	[76,77][2]	[78][2] [80][3]	
Trophozoite	[76,77][2]	[78][2] [80][3]	
Schizont	[76,77][2]	[78][2] [80][3]	
Gametocyte	[77][2] [99][2]	NA	
Female & male gametocyte	[100][3]	NA	
Sporogonic cycle:			
Macro & Microgamete		[92][2]	
Zygote		[92][2]	
Ookinete		[92][2]	
Oocyst			
Sporozoites (mosquito saliva)	[77][2] [101][3]	[92][2] [101][3]	[97][2]

[1] Only a selection of available transcriptomes is shown. Selection criteria were initial publication or quality of transcriptome data.

[2] Microarray

[3] RNA sequencing

[4] Single cell RNA sequencing

1.4 Overview of Bioinformatics Methods

In the last decade, next generation sequencing (NGS), also called high throughput sequencing or deep sequencing, became widely applicable to field samples from molecular epidemiology studies. Performing NGS on field samples is much more challenging than on samples from laboratory cultivated parasites and requires more robust analysis methods. In case of *Plasmodium* field samples collected from patients the main challenges for the laboratory work are that the amount of input material is limited and contaminated with host DNA or RNA. For data analysis, the large biological variation between field samples is a challenge. Field samples can contain complex mixtures of infecting clones or development stages. As a result, often no biological replicates are feasible, because each patient harbours a unique parasite strain and a unique mixture of stages. Most NGS analysis methods are not developed for complex field isolates and therefore need adaptions to be applicable on such samples.

1.4.1 Haplotype inference and MOI estimation

SNP-based haplotype inference of a sample containing a single-clone infection is done by calling the predominant SNPs in the sequence reads, thus identifying the haploid genome. SNPs of low frequency are regarded as amplification or sequencing errors. Several software are available for SNP calling, e.g. samtools or GATK framework [102,103]. However, SNP calling is much more complex in samples containing multi-clone infections with unknown multiplicity. The situation resembles SNP calling in polyploid genomes where the ploidy is unknown. In addition, the frequency of each clone in a multi-clone infection is unknown and can be even less than 1%.

Most of the software for SNP calling and haplotype inference were developed for diploid genomes or require prior knowledge of the ploidy, e.g. ReadBackedPhasing, HapCUT, HaplotypeCaller, HapCompass, BEAGLE, IMPUTE2, SHAPEIT [103–110]. Such software cannot be used for multi-clonal infections with unknown ploidy.

The approach chosen to infer haplotypes in multi-clone infections depends on whether SNPs are linked or unlinked by sequence reads. In Amp-Seq, multiple SNPs are linked usually by a single sequence read. Haplotype inference in such data can be done by clustering of those sequence reads, e.g. SeekDeep, Swarm [111–113]. The clustering combines similar sequence reads together that differ because of amplification or sequencing errors. However, also sequence reads from closely related clones cluster together, if they differ in only one SNP.

For data from WGS or SNP barcodes, where SNPs are unlinked or only partly overlapping by sequence reads, the number of co-infecting clones needs to be estimated before haplotype inference can be performed. The MOI estimation software use either (i) a sliding window to cluster reads that partly overlap locally, e.g. estMOI [114] or (ii) estimate MOI directly from SNP frequencies without using any information about SNP linkage, e.g. COIL, pfmix, THE REAL McCOIL [115–117].

Haplotype inference on partly linked SNPs of small genomes is performed by a sliding window or by extending a smaller section of the genome where sequence reads have significant overlap and clustering of reads can be applied, e.g. shorah, PredictHaplo, QuRe, ViSpA, HaploClique, HapCompass-Tumor [118–123]. On unlinked SNPs, haplotype inference is performed by assembling SNPs sharing a similar proportion of reads by using a Markov chain Monte Carlo (MCMC) approach, e.g. DEploid [124].

SeekDeep is currently the most commonly used method to analyse Amp-Seq genotyping data of *Plasmodia*. However, SeekDeep can only be used on a cluster with a large working memory capacity. Swarm in contrast, runs very efficiently on a standard personal computer. Both methods called false haplotypes when samples with controlled mixture were analysed. In this thesis, an in-depth analysis of false haplotype calls was made and a new workflow put together for simple analysis of Amp-Seq genotyping data (Chapter 2). Furthermore, the potential of using longitudinal Amp-Seq genotyping for multi-locus haplotype inference in complex infections was explored (Chapter 3).

1.4.2 Differential expression and deconvolution of mixed transcriptomes

Differential expression (DE) analysis is used to study the difference in gene expression between phenotypes, groups or cell stages. The most commonly used software for gene expression analysis by microarray is limma, while for RNA-Seq data edgeR, DESeq and Cufflinks is often used [125–129]. The workflow of these software is similar. They first normalise the gene expression data and then fit a linear model to get an estimate of the variation in the data and the fold change between the different groups. The main difference between microarray and RNA-Seq gene expression analysis is that microarray data are normally distributed, whereas RNA-Seq data follow a negative binomial distribution. The software for RNA-Seq differ in their methods used to normalise and estimate the variation in the data. RNA-Seq data also provides the possibility to study alternative splice forms. Some isoforms might have different functions and are often expressed in different cell types. Following software amongst others are used for gene expression analyses at the exon level: DEXSeq, edgeR or MISO [126,130,131].

A single cell-type or developmental stage cannot always be isolated from biological samples, e.g. the hematopoietic subsets in the human blood. In this case, the observed transcriptome represents a mixture of cell-type specific transcriptomes. Several deconvolution methods have been developed either to infer the relative cell-type fraction in the sample or to infer the different cell-type specific transcriptome signatures, e.g. csSam, PERT, CIBERSORT, DeconRNASeq, DSection, xCell [132–141]. A comprehensive review of deconvolution methods can be found in Mohammedi et al. 2017. In general, deconvolution methods make the following assumptions[142]: (i) linearity, meaning that the observed mixed transcriptomes correspond to the sum of individual transcriptome signatures weighted by the relative cell-type fractions; (ii) non-negativity, meaning that neither the transcriptome signatures nor the relative cell-type fractions are negative; (iii) sum up to one, meaning that the relative fractions of cell-types sum up to one; and (iv) similar cell quantity, that the signature profiles and corresponding mixture must be normalised to ensure to represent gene expression level of the same number of cells.

So far, deconvolution of *P. falciparum* blood stages has been performed only on microarray transcriptome data [143]. Applying the same approach to RNA-Seq data does not give satisfactory results. One explanation for this is that the similar cell quantity assumption is not valid for transcriptome data from *Plasmodium* blood stages, as the parasite genome replicates during the erythrocytic cycle. Normalisation methods are used to ensure that expression levels represent similar cell quantity. Evaluating normalization and deconvolution methods for RNA-Seq data to infer stage specific transcriptomes from experimentally mixed *Plasmodium* blood stages is one of the topics of this thesis and is presented in Chapter 4.

1.5 Aim and Objectives of this Thesis

The overall aim of this thesis is two-fold: Firstly, to develop a novel protocol and analysis pipeline to infer haplotypes of multi-clone infections from deep sequencing data and comparing these haplotypes to genotyping data based on size polymorphism. Secondly, to evaluate normalisation and de-convolution methods to infer stage-specific transcriptome signatures from experimental mixed stage samples of *P. falciparum* with known stage composition as proof of concept for inferring the transcriptome of *P. vivax* gametocytes from field samples.

Specific objectives include:

Objective 1: Development of a new Amp-Seq genotyping assay for multi-clone *P. falciparum* infections.

a) Screening *P. falciparum* genomes for highly polymorphic loci.

b) Identifying a novel marker suited for Amp-Seq.

c) Developing a highly multiplexed Amp-Seq genotyping assay, suited for large epidemiological studies.

Objective 2: Development of an analysis pipeline for Amp-Seq genotyping data of multi-clone *P. falciparum* infections.

a) Developing a bioinformatics pipeline to analyse Amp-Seq genotyping data.

b) Evaluating the impact of amplification and sequence errors on genotype calling in experimental mixtures.

c) Defining a detection limit and filtering criteria for genotype calling.

Objective 3: Comparative analysis of SNP-based and length-polymorphic-based genotyping method in longitudinal samples from a cohort study in PNG.

a) Applying the developed Amp-Seq assay and analysis pipeline to archived field samples from a longitudinal study comprising samples with multi-clone infections.

b) Comparison of the resolution, sensitivity, and specificity of Amp-Seq genotyping markers with that of the length-polymorphic genotyping marker *msp2*.

c) Comparison of molecular parameters (MOI, $_{\mathrm{mol}}$FOI) describing *P. falciparum infection* dynamics obtained from Amp-Seq genotyping markers with those from length-polymorphic markers.

d) Exploring the limitation of multi-locus haplotype inference from Amp-Seq genotyping data.

e) Exploring the suitability of Amp-Seq to study clone dynamics and density of each clone in longitudinal samples comprising multi-clone infections.

Objective 4: De-convolution of mixed stage transcriptome data.

a) Assessing stage purity of highly synchronised *P. falciparum* culture-derived parasites.

b) Differential expression analysis of highly synchronised *P. falciparum* samples.

c) Evaluating normalisation methods for RNA-Seq data from samples with varying total RNA levels.

d) Evaluating de-convolution methods to infer a stage-specific transcriptome from mixed stage transcriptomes of known stage composition.

e) Evaluating de-convolution methods to estimate stage composition in mixed stage transcriptome data from stage-specific transcriptome signatures.

f) Assessing the feasibility of inferring the transcriptome of *P. vivax* gametocytes from field samples containing enriched gametocytes mixed with late blood stages.

Additional project: Preliminary analysis of the transcriptome and epigenome of *P. vivax* sporozoites.

a) Processing of RNA and chromatin immunoprecipitation (ChIP) sequencing data.

b) Exploring correlation between transcriptional activity and histone modifications.

c) Identifying transcriptionally silenced regions by histone modifications containing genes of the multi-gene family *Pv-fam*.

REFERENCES

1. WHO. World Malaria Report 2012. 2012.

2. WHO. World Malaria Report 2017. 2017.

3. Shetty P. The numbers game. Malaria, Nat Outlook. **2012**; :4–5.

4. World Health Organization. World Malaria Report 2008. World. **2008**; :6–14.

5. Coatney G, Collins W, Warren M, Contacos P. The Primate Malarias. Atlanta: Center for disease control and prefention; 1971.

6. Nájera JA, González-Silva M, Alonso PL. Some lessons for the future from the global malaria eradication programme (1955-1969). PLoS Med. **2011**; 8(1).

7. Livadas GA, Georgopoulos G. Development of resistance to DDT by Anopheles sacharovi in Greece. Bull World Health Organ. **1953**; 8(4):497–511.

8. Antony H, Parija S. Antimalarial drug resistance: An overview. Trop Parasitol. **2016**; 6(1):30.

9. Nelson KE, Williams CM. Infectious Disease Epidemiology. 2nd ed. Jones & Bartlett Learning; 2014.

10. Harinasuta T, Suntharasamai P, Viravan C. Chloroquine-resistant falciparum malaria in Thailand. Lancet (London, England). **1965**; 2(7414):657–60.

11. Roberts L, Enserink M. Malaria. Did they really say ... eradication? Science. **2007**; 318(5856):1544–5.

12. Bousema T, Okell L, Shekalaghe S, et al. Revisiting the circulation time of Plasmodium falciparum gametocytes: molecular detection methods to estimate the duration of gametocyte carriage and the effect of gametocytocidal drugs. Malar J. **2010**; 9:136.

13. Betuela I, Rosanas-Urgell A, Kiniboro B, et al. Relapses contribute significantly to the risk of Plasmodium vivax infection and disease in Papua New Guinean children 1-5 years of age. J Infect Dis. **2012**; 206(11):1771–80.

14. Bousema T, Drakeley C. Determinants of Malaria Transmission at the Population Level. Cold Spring Harb Perspect Med. **2017**; .

15. Kiattibutr K, Roobsoong W, Sriwichai P, et al. Infectivity of symptomatic and asymptomatic Plasmodium vivax infections to a Southeast Asian vector, Anopheles dirus. Int J Parasitol. **2017**; 47(2–3):163–170.

16. Neafsey DE, Juraska M, Bedford T, et al. Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine. N Engl J Med. **2015**; 373(21):2025–37.

17. Faust C, Dobson AP. Primate malarias: Diversity, distribution and insights for zoonotic Plasmodium. One Heal. The Authors; **2015**; 1:66–75.

18. Price RN, Tjitra E, Guerra C a, Yeung S, White NJ, Anstey NM. Vivax malaria: neglected and not benign. Am J Trop Med Hyg. **2007**; 77(6 Suppl):79–87.

19. Galinski MR, Meyer EVS, Barnwell JW. Plasmodium vivax: modern strategies to study a persistent parasite's life cycle. Adv. Parasitol. Elsevier; 2013.

20. Kwiatkowski DP. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. Am J Hum Genet. **2005**; 77(2):171–192.

21. Hedrick PW. Population genetics of malaria resistance in humans. Heredity (Edinb). Nature Publishing Group; **2011**; 107(4):283–304.

22. Bousema T, Drakeley C. Epidemiology and infectivity of Plasmodium falciparum and Plasmodium vivax gametocytes in relation to malaria control and elimination. Clin Microbiol Rev. **2011**; 24(2):377–410.

23. Hermsen CC, Vlas SJ De, Gemert GJA Van, Telgt DSC, Verhage DF, Sauerwein RW. Testing vaccines in human experimental malaria: Statistical analysis of parasitemia measured by a quantitative real-time polymerase chain reaction. Am J Trop Med Hyg. **2004**; 71(2):196–201.

24. McCarthy JS, Griffin PM, Sekuloski S, et al. Experimentally induced blood-stage Plasmodium vivax infection in healthy volunteers. J Infect Dis. **2013**; 208(10):1688–94.

25. Tachibana S-I, Sullivan S a, Kawai S, et al. Plasmodium cynomolgi genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. Nat Genet. Nature Publishing Group; **2012**; 44(9):1051–5.

26. Rutledge GG, Böhme U, Sanders M, et al. Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. Nature. Nature Publishing Group; **2017**; 542(7639):101–104.

27. Brancucci NMB, Bertschi NL, Zhu L, et al. Heterochromatin Protein 1 Secures Survival and Transmission of Malaria Parasites. Cell Host Microbe. **2014**; 16(2):165–176.

28. Clements AN. The Biology of Mosquitoes: Development, Nutrition and Reproduction. Chapman and Hall; 1992.

29. Oaks SC, Mitchell VS, Pearson GW. Malaria: Obstacles and Opportunities. J. Carpenter, Ed. Comm. Study Malar. Prev. Control Div. Int. Heal. 1991.

30. Howes RE, Battle KE, Satyagraha AW, Baird JK, Hay SI. G6PD Deficiency. Global Distribution, Genetic Variants and Primaquine Therapy. Adv Parasitol. Elsevier; **2013**; 81:135–201.

31. Howes RE, Piel FB, Patil AP, et al. G6PD Deficiency Prevalence and Estimates of Affected Populations in Malaria Endemic Countries: A Geostatistical Model-Based Map. PLoS Med. **2012**; 9(11).

32. Felger I, Maire M, Bretscher MT, et al. The Dynamics of Natural Plasmodium falciparum Infections. Gosling RD, editor. PLoS One. Public Library of Science; **2012**; 7(9):e45542.

33. Miller RH, Hathaway NJ, Kharabora O, et al. A deep sequencing approach to estimate Plasmodium falciparum complexity of infection (COI) and explore apical membrane antigen 1 diversity. Malar J. BioMed Central; **2017**; 16(1):490.

34. Fola AA, Harrison GLA, Hazairin MH, et al. Higher Complexity of Infection and Genetic Diversity of Plasmodium vivax Than Plasmodium falciparum Across All Malaria Transmission Zones of Papua New Guinea. Am J Trop Med Hyg. **2017**; 96(3):630–641.

35. Bretscher MT, Maire N, Felger I, Owusu-Agyei S, Smith T. Asymptomatic Plasmodium falciparum infections may not be shortened by acquired immunity. Malar J. BioMed Central; **2015**; 14(1):294.

36. Hofmann N, Mwingira F, Shekalaghe S, Robinson LJ, Mueller I, Felger I. Ultra-Sensitive Detection of

Plasmodium falciparum by Amplification of Multi-Copy Subtelomeric Targets. Seidlein L von, editor. PLOS Med. **2015**; 12(3):e1001788.

37. Bousema T, Okell L, Felger I, Drakeley C. Asymptomatic malaria infections: detectability, transmissibility and public health relevance. Nat Rev Microbiol. Nature Publishing Group; **2014**; (October):1–8.

38. Koepfli C, Schoepflin S, Bretscher M, et al. How much remains undetected? Probability of molecular detection of human Plasmodia in the field. PLoS One. **2011**; 6(4):e19010.

39. Mueller I, Schoepflin S, Smith T a., et al. Force of infection is key to understanding the epidemiology of Plasmodium falciparum malaria in Papua New Guinean children. Proc Natl Acad Sci. **2012**; 109(25):10030–10035.

40. Bretscher MT, Valsangiacomo F, Owusu-Agyei S, Penny M a, Felger I, Smith T. Detectability of Plasmodium falciparum clones. Malar J. **2010**; 9:234.

41. Smith T, Felger I, Fraser-Hurt N, Beck HP. Effect of insecticide-treated bed nets on the dynamics of multiple Plasmodium falciparum infections. Trans R Soc Trop Med Hyg. **1999**; 93 Suppl 1:53–7.

42. Sama W, Owusu-Agyei S, Felger I, Vounatsou P, Smith T. An immigration-death model to estimate the duration of malaria infection when detectability of the parasite is imperfect. Stat Med. **2005**; 24(21):3269–88.

43. Koepfli C, Mueller I. Malaria Epidemiology at the Clone Level. Trends Parasitol. Elsevier Ltd; **2017**; xx:1–12.

44. Volkman SK, Neafsey DE, Schaffner SF, Park DJ, Wirth DF. Harnessing genomics and genome biology to understand malaria biology. Nat Rev Genet. **2012**; 13(5):315–28.

45. World Health Organization. Methods and techniques for clinical trials on antimalarial drug efficacy: genotyping to identify parasite populations. 2008.

46. Felger I, Beck H-P. Genotyping of Plasmodium falciparum. PCR-RFLP analysis. Methods Mol Med. **2002**; 72:117–29.

47. Felger I, Tavul L, Kabintik S, et al. Plasmodium falciparum: extensive polymorphism in merozoite surface antigen 2 alleles in an area with endemic malaria in Papua New Guinea. Exp. Parasitol. 1994. p. 106–116.

48. Mercereau-Puijalon O, Fandeur T, Bonnefoy S, Jacquemot C, Sarthou JL. A study of the genomic diversity of Plasmodium falciparum in Senegal 2. Typing by the use of the polymerase chain reaction. Acta Trop. **1991**; 49(4):293–304.

49. Kain KC, Lanar DE. Determination of genetic variation within Plasmodium falciparum by using enzymatically amplified DNA from filter paper disks impregnated with whole blood. J Clin Microbiol. **1991**; 29(6):1171–1174.

50. Falk N, Maire N, Sama W, et al. Comparison of PCR-RFLP and Genescan-based genotyping for analyzing infection dynamics of Plasmodium falciparum. Am J Trop Med Hyg. **2006**; 74(6):944–50.

51. Jafari S, Bras J Le, Bouchaud O, Durand R. Plasmodium falciparum clonal population dynamics during

malaria treatment. J Infect Dis. **2004**; 189:195–203.

52. Baniecki ML, Faust AL, Schaffner SF, et al. Development of a single nucleotide polymorphism barcode to genotype Plasmodium vivax infections. PLoS Negl Trop Dis. **2015**; 9(3):e0003539.

53. Gan LSH, Loh JP. Rapid identification of chloroquine and atovaquone drug resistance in Plasmodium falciparum using high-resolution melt polymerase chain reaction. Malar J. **2010**; 9(1):134.

54. Daniels R, Volkman SK, Milner DA, et al. A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking. Malar J. **2008**; 7(1):223.

55. Juliano JJ, Porter K, Mwapasa V, et al. Exposing malaria in-host diversity and estimating population diversity by capture-recapture using massively parallel pyrosequencing. Proc Natl Acad Sci U S A. **2010**; 107(46):20138–43.

56. Bailey J a, Mvalo T, Aragam N, et al. Use of massively parallel pyrosequencing to evaluate the diversity of and selection on Plasmodium falciparum csp T-cell epitopes in Lilongwe, Malawi. J Infect Dis. **2012**; 206(4):580–7.

57. Messerli C, Hofmann NE, Beck H-P, Felger I. Critical Evaluation of Molecular Monitoring in Malaria Drug Efficacy Trials and Pitfalls of Length-Polymorphic Markers. Antimicrob Agents Chemother. **2017**; 61(1):AAC.01500-16.

58. Saxena V, Garg S, Tripathi J, et al. Plasmodium vivax apicoplast genome: a comparative analysis of major genes from Indian field isolates. Acta Trop. Elsevier B.V.; **2012**; 122(1):138–49.

59. Reddy GR, Chakrabarti D, Schuster SM, Ferl RJ, Almira EC, Dame JB. Gene sequence tags from Plasmodium falciparum genomic DNA fragments prepared by the “genease” activity of mung bean nuclease. Proc Natl Acad Sci U S A. **1993**; 90(21):9867–71.

60. Chakrabarti D, Reddy GR, Dame JB, et al. Analysis of expressed sequence tags from Plasmodium falciparum. Mol Biochem Parasitol. **1994**; 66(1):97–104.

61. Carlton JMR, Muller R, Yowell CA, et al. Profiling the malaria genome: A gene survey of three species of malaria parasite with comparison to other apicomplexan species. Mol Biochem Parasitol. **2001**; 118(2):201–210.

62. Cui L, Fan Q, Hu Y, et al. Gene discovery in Plasmodium vivax through sequencing of ESTs from mixed blood stages. Mol Biochem Parasitol. **2005**; 144(1):1–9.

63. Gardner MJ, Hall N, Fung E, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature. **2002**; 419(6906):498–511.

64. Carlton JM, Adams JH, Silva JC, et al. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature. Macmillan Publishers Limited. All rights reserved; **2008**; 455(7214):757–63.

65. Pain A, Böhme U, Berry a E, et al. The genome of the simian and human malaria parasite Plasmodium knowlesi. Nature. **2008**; 455(7214):799–803.

66. Otto TD, Rayner JC, Böhme U, et al. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. Nat Commun. **2014**; 5:4754.

67. Hall N. Genomic insights into the other malaria. Nat Genet. Nature Publishing Group; **2012**; 44(9):962–3.

68. Auburn S, Böhme U, Steinbiss S, et al. A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. Wellcome Open Res. **2016**; 1(0):4.

69. Manske M, Miotto O, Campino S, et al. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature. Nature Publishing Group; **2012**; 487(7407):375–379.

70. MalariaGEN Plasmodium falciparum Community Project. Genomic epidemiology of artemisinin resistant malaria. Elife. **2016**; 5:1–29.

71. Neafsey DE, Galinsky K, Jiang RHY, et al. The malaria parasite Plasmodium vivax exhibits greater genetic diversity than Plasmodium falciparum. Nat Genet. Nature Publishing Group; **2012**; 44(9):1046–50.

72. Pearson RD, Amato R, Auburn S, et al. Genomic analysis of local variation and recent evolution in Plasmodium vivax. Nat Genet. **2016**; 48(8):959–964.

73. Parobek CM, Lin JT, Saunders DL, et al. Selective sweep suggests transcriptional regulation may underlie *Plasmodium vivax* resilience to malaria control measures in Cambodia. Proc Natl Acad Sci. **2016**; 113(50):E8096–E8105.

74. Hupalo DN, Luo Z, Melnikov A, et al. Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax. Nat Genet. **2016**; 48(8):953–958.

75. Scherf A, Lopez-Rubio JJ, Riviere L. Antigenic Variation in Plasmodium falciparum. Annu Rev Microbiol. **2008**; 62(1):445–470.

76. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. PLoS Biol. **2003**; 1(1):E5.

77. Roch KG Le, Zhou Y, Blair PL, et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science. **2003**; 301(5639):1503–8.

78. Bozdech Z, Mok S, Hu G, et al. The transcriptome of Plasmodium vivax reveals divergence and diversity of transcriptional regulation in malaria parasites. Proc Natl Acad Sci U S A. **2008**; 105(42):16290–5.

79. Otto TD, Wilinski D, Assefa S, et al. New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. Mol Microbiol. **2010**; 76(1):12–24.

80. Zhu L, Mok S, Imwong M, et al. New insights into the Plasmodium vivax transcriptome using RNA-Seq. Sci Rep. Nature Publishing Group; **2016**; 6(February):20498.

81. Llinás M, Deitsch KW, Voss TS. Plasmodium gene regulation: far more to factor in. Trends Parasitol. **2008**; 24(12):551–6.

82. Vembar SS, Droll D, Scherf A. Translational regulation in blood stages of the malaria parasite Plasmodium spp.: systems-wide studies pave the way. Wiley Interdiscip Rev RNA. **2016**; 7(6):772–792.

83. Mamoun C Ben, Gluzman IY, Hott C, et al. Co-ordinated programme of gene expression during asexual

intraerythrocytic development of the human malaria parasite Plasmodium falciparum revealed by microarray analysis. Mol Microbiol. **2001**; 39(1):26–36.

84. Horrocks P, Dechering K, Lanzer M. Control of gene expression in Plasmodium falciparum. Mol Biochem Parasitol. **1998**; 95(2):171–81.

85. Pelle KG, Oh K, Buchholz K, et al. Transcriptional profiling defines dynamics of parasite tissue sequestration during malaria infection. Genome Med. **2015**; 7(1):19.

86. Wampfler R, Mwingira F, Javati S, et al. Strategies for Detection of Plasmodium species Gametocytes. Paul RE, editor. PLoS One. **2013**; 8(9):e76316.

87. Nishimoto Y, Arisue N, Kawai S, et al. Evolution and phylogeny of the heterogeneous cytosolic SSU rRNA genes in the genus Plasmodium. Mol Phylogenet Evol. **2008**; 47(1):45–53.

88. Li J, Gutell RR, Damberger SH, et al. Regulation and trafficking of three distinct 18 S ribosomal RNAs during development of the malaria parasite. J Mol Biol. **1997**; 269(2):203–213.

89. Bahl A, Brunk B, Crabtree J, et al. PlasmoDB: The Plasmodium genome resource. A database integrating experimental and computational data. Nucleic Acids Res. 2003. p. 212–215.

90. Kensche PR, Hoeijmakers WAM, Toenhake CG, et al. The nucleosome landscape of Plasmodium falciparum reveals chromatin architecture and dynamics of regulatory sequences. Nucleic Acids Res. **2016**; 44(5):2110–2124.

91. Mok S, Ashley EA, Ferreira PE, et al. Drug resistance. Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance. Science. **2015**; 347(6220):431–5.

92. Westenberger SJ, McClean CM, Chattopadhyay R, et al. A systems-based analysis of Plasmodium vivax lifecycle transcription from human to mosquito. PLoS Negl Trop Dis. **2010**; 4(4):e653.

93. Wampfler R, Hofmann NE, Karl S, et al. Effects of liver-stage clearance by Primaquine on gametocyte carriage of Plasmodium vivax and P. falciparum. PLoS Negl Trop Dis. **2017**; 11(7):1–15.

94. Cubi R, Vembar SS, Biton A, et al. Laser capture microdissection enables transcriptomic analysis of dividing and quiescent liver stages of Plasmodium relapsing species. Cell Microbiol. **2017**; :e12735.

95. Voorberg-van der Wel A, Roma G, Gupta DK, et al. A comparative transcriptomic analysis of replicating and dormant liver stages of the relapsing malaria parasite Plasmodium Cynomolgi. Elife. **2017**; 6(1).

96. Mueller I, Galinski MR, Baird JK, et al. Key gaps in the knowledge of Plasmodium vivax, a neglected human malaria parasite. Lancet Infect Dis. Elsevier Ltd; **2009**; 9(9):555–66.

97. Ylostalo J, Randall AC, Myers T a, Metzger M, Krogstad DJ, Cogswell FB. Transcriptome profiles of host gene expression in a monkey model of human malaria. J Infect Dis. **2005**; 191(3):400–9.

98. Poran A, Nötzel C, Aly O, et al. Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. Nature. Nature Publishing Group; **2017**; 551(7678):95–99.

99. Young J a, Fivelman QL, Blair PL, et al. The Plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification. Mol Biochem Parasitol. **2005**; 143(1):67–79.

100. Lasonder E, Rijpma SR, Schaijk BCL Van, et al. Integrated transcriptomic and proteomic analyses of P. Falciparum gametocytes: Molecular insight into sex-specific processes and translational repression. Nucleic Acids Res. **2016**; 44(13):6087–6101.

101. Sporozoite V, Ivo C, Jex AR, Kappe SHI. Integrated transcriptomic, proteomic and epigenomic analysis of Plasmodium vivax salivary-gland sporozoites. bioRxiv. **2017**; .

102. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. **2011**; 27(21):2987–2993.

103. Depristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. **2011**; 43(5):491–501.

104. Bansal V, Bafna V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics. **2008**; 24(16):i153-9.

105. Aguiar D, Istrail S. HapCompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data. J Comput Biol. **2012**; 19(6):577–90.

106. Aguiar D, Istrail S. Haplotype assembly in polyploid genomes and identical by descent shared tracts. Bioinformatics. **2013**; 29(13):352–360.

107. Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum Genet. **2007**; 81(5):1084–1097.

108. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. **2009**; 5(6).

109. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. **2012**; 9(2):179–181.

110. O'Connell J, Gurdasani D, Delaneau O, et al. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genet. **2014**; 10(4).

111. Mahé F, Rognes T, Quince C, Vargas C de, Dunthorn M. Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ. **2015**; 3:e1420.

112. Mahé F, Rognes T, Quince C, Vargas C de, Dunthorn M. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ. **2014**; 2:e593.

113. Hathaway NJ, Parobek CM, Juliano JJ, Bailey JA. SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. Nucleic Acids Res. Oxford University Press; **2017**; (December):1–13.

114. Assefa SA, Preston MD, Campino S, Ocholla H, Sutherland CJ, Clark TG. estMOI: estimating multiplicity of infection using parasite deep sequencing data. Bioinformatics. **2014**; 30(9):1292–4.

115. O'Brien JD, Iqbal Z, Wendler J, Amenga-Etego L. Inferring Strain Mixture within Clinical Plasmodium falciparum Isolates from Genomic Sequence Data. PLoS Comput Biol. **2016**; 12(6):1–20.

116. Chang H-H, Worby CJ, Yeka A, et al. THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. Pascual M, editor. PLoS Comput Biol. **2017**; 13(1):e1005348.

117. Galinsky K, Valim C, Salmier A, et al. COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. Malar J. **2015**; 14(1):4.

118. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics. BioMed Central Ltd; **2011**; 12(1):119.

119. Töpfer A, Marschall T, Bull R a, Luciani F, Schönhuth A, Beerenwinkel N. Viral quasispecies assembly via maximal clique enumeration. PLoS Comput Biol. **2014**; 10(3):e1003515.

120. Aguiar D, Wong WSW, Istrail S. Tumor haplotype assembly algorithms for cancer genomics. Pac Symp Biocomput. **2014**; :3–14.

121. Prabhakaran S, Rey M, Zagordi O, Beerenwinkel N, Roth V. HIV haplotype inference using a propagating dirichlet process mixture model. IEEE/ACM Trans Comput Biol Bioinforma. **2014**; 11(1):182–191.

122. Astrovskaya I, Tork B, Mangul S, et al. Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC Bioinformatics. **2011**; 12(Suppl 6):S1.

123. Prosperi MCF, Salemi M. QuRe: Software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics. **2012**; 28(1):132–133.

124. Zhu SJ, Almagro-Garcia J, McVean G. Deconvolution of multiple infections in Plasmodium falciparum from high throughput sequencing data. Bioinformatics. **2017**; .

125. Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Stat Appl Genet Mol Biol. **2004**; 3(1):1–25.

126. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. **2010**; 26(1):139–40.

127. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **2014**; 15(12):550.

128. Trapnell C, Williams B a, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. Nature Publishing Group; **2010**; 28(5):511–5.

129. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. BioMed Central Ltd; **2010**; 11(10):R106.

130. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. **2010**; 7(12):1009–15.

131. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Res. **2012**; 22(10):2008–17.

132. Qiao W, Quon G, Csaszar E, Yu M, Morris Q, Zandstra PW. PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. Bonneau R, editor. PLoS Comput Biol. **2012**; 8(12):e1002838.

133. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous

tissue samples based on mRNA-Seq data. Bioinformatics. **2013**; 29(8):1083–5.

134. Gaujoux R, Seoighe C. CellMix: a comprehensive toolbox for gene expression deconvolution. Bioinformatics. **2013**; 29(17):2211–2.

135. Gaujoux R, Seoighe C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. Infect Genet Evol. Elsevier B.V.; **2012**; 12(5):913–21.

136. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One. **2009**; 4(7):e6098.

137. Zhong Y, Wan Y-W, Pang K, Chow LML, Liu Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinformatics. **2013**; 14:89.

138. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. **2017**; 18(1):220.

139. Shen-Orr SS, Tibshirani R, Khatri P, et al. Cell type–specific gene expression differences in complex tissues. Nat Methods. Nature Publishing Group; **2010**; 7(4):287–289.

140. Erkkilä T, Lehmusvaara S, Ruusuvuori P, Visakorpi T, Shmulevich I, Lähdesmäki H. Probabilistic analysis of gene expression measurements from heterogeneous tissues. Bioinformatics. **2010**; 26(20):2571–7.

141. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. Nature Publishing Group; **2015**; (MAY 2014):1–10.

142. Mohammadi S, Zuckerman N, Goldsmith A, Grama A. A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues. Proc IEEE. **2017**; 105(2):340–366.

143. Joice R, Narasimhan V, Montgomery J, et al. Inferring Developmental Stage Composition from Gene Expression in Human Malaria. Przytycka TM, editor. PLoS Comput Biol. **2013**; 9(12):e1003392.

CHAPTER 2: AMP-SEQ GENOTYPING: MARKER, ASSAY AND ANALYSIS PIPELINE

Lerch *et al. BMC Genomics* (2017) 18:864
DOI 10.1186/s12864-017-4260-y

BMC Genomics

METHODOLOGY ARTICLE Open Access

Development of amplicon deep sequencing markers and data analysis pipeline for genotyping multi-clonal malaria infections

Anita Lerch[1,2,3], Cristian Koepfli[3,4], Natalie E. Hofmann[1,2], Camilla Messerli[1,2], Stephen Wilcox[3,4], Johanna H. Kattenberg[5,6], Inoni Betuela[5], Liam O'Connor[3,4], Ivo Mueller[3,4,7] and Ingrid Felger[1,2*]

Abstract

Background: Amplicon deep sequencing permits sensitive detection of minority clones and improves discriminatory power for genotyping multi-clone *Plasmodium falciparum* infections. New amplicon sequencing and data analysis protocols are needed for genotyping in epidemiological studies and drug efficacy trials of *P. falciparum*.

Methods: Targeted sequencing of molecular marker *csp* and novel marker *cpmp* was conducted in duplicate on mixtures of parasite culture strains and 37 field samples. A protocol allowing to multiplex up to 384 samples in a single sequencing run was applied. Software "HaplotypR" was developed for data analysis.

Results: *Cpmp* was highly diverse ($H_e = 0.96$) in contrast to *csp* ($H_e = 0.57$). Minority clones were robustly detected if their frequency was >1%. False haplotype calls owing to sequencing errors were observed below that threshold.

Conclusions: To reliably detect haplotypes at very low frequencies, experiments are best performed in duplicate and should aim for coverage of >10'000 reads/amplicon. When compared to length polymorphic marker *msp2*, highly multiplexed amplicon sequencing displayed greater sensitivity in detecting minority clones.

Keywords: Plasmodium falciparum, Malaria, Amplicon sequencing, SNP, Haplotype clustering, Multi-clone infections, msp2, csp, cpmp, HaplotypR software

Background

In infection biology of malaria as well as of many other pathogens, detection of minority clones is a crucial task. In areas of high malaria transmission, most infected hosts harbour multiple clones of the same *Plasmodium* species. To better understand the epidemiology and infection dynamics of malaria, individual parasite clones are tracked over time to measure the acquisition, elimination and persistence of individual clones in a human host. The incidence of new clones per host serves as surrogate measure for the exposure of an individual and for the transmission intensity in a population [1]. Identification of new infections is also crucial in clinical trials of antimalarial drugs, where persisting clones need to be distinguished from new clones in post-treatment samples from patients with recurrent parasitaemia [2, 3]. For such diverse applications, genotyping methods based on length polymorphic markers had been applied for decades, particularly by targeting microsatellite markers or genes encoding parasite surface antigens such as merozoite surface proteins 1 and 2 (*msp1*, *msp2*) [4, 5].

Despite their wide use in many malaria research laboratories, length polymorphic markers have important limitations. For example, microsatellite typing suffers from frequent occurrence of stutter peaks, possibly resulting from polymerase slippage on stretches of simple tandem repeats. A cut-off requirement for a minimal peak height (e.g. 33% of the predominant peak [6]) is required to

* Correspondence: Ingrid.Felger@unibas.ch
[1]Swiss Tropical and Public Health Institute, Basel, Switzerland
[2]University of Basel, Basel, Switzerland
Full list of author information is available at the end of the article

prevent scoring of artefact fragments. However, this cut-off makes it impossible to detect minority clones falling below the selected threshold. Another limitation of length polymorphic marker, particularly the highly polymorphic parasite surface antigens, consists in the usually large size differences between alleles. Major size differences lead to bias in amplification, preferring the shorter PCR fragments in samples that concurrently contain multiple *P. falciparum* infections [7].

Deep sequencing of short amplicons has the potential to overcome some of the shortfalls of length polymorphic genotyping markers, in particular the influence of fragment length of a marker on the detectability of minority clones. Earlier studies used two different approaches for genotyping of *P. falciparum* and *P. vivax* by amplicon deep sequencing: (i) Sequencing of the classical length polymorphic genotyping markers, such as *P. falciparum msp1* and *msp2* [8]. Alternatively, sequencing targeted non-repetitive regions that harbour extensive single nucleotide polymorphism (SNP), such as the *P. falciparum* circumsporozoite protein (*csp*) or *P. vivax msp1* [9–11]. The strength of these approaches is that all SNPs within an amplicon are linked by a single sequence read, leading directly to haplotype identification. (ii) Sequencing of multiple loci of genome-wide distribution, whereby each locus comprises one SNP [12]. This latter approach is particularly suited for population genetic investigations, as these loci are not linked. The downside is that the haplotype of each infecting clone has to be reconstructed, which is difficult or even impossible for samples with a high number of co-infecting clones per host [13]. Thus, genotyping of samples containing multi-clone infections remains an unresolved challenge when multiple genome-wide loci are targeted.

In previous studies, amplicon deep sequencing was performed on two platforms, 454/Roche or Ion Torrent. In the past these technologies have produced longer sequences than the 37 bp reads obtained by the Illumina sequencing platform. Now Illumina MiSeq generates reads of up to 600 bp length (Illumina, MiSeq Reagent Kit v3). Sequencing error rates of 454/Roche and Ion Torrent technologies were high, owing to insertion and deletion (indel) errors occurring predominantly in homopolymeric regions [14–16]. Illumina sequencing is less susceptible to indel errors and has an overall smaller error rate [16].

The present report outlines a strategy and protocols for identifying highly diverse markers for SNP-based genotyping of *P. falciparum* by amplicon sequencing. The primary aim was to thoroughly assess the analytical sensitivity and specificity of amplicon sequencing in detecting minority clones. In epidemiological studies involving hundreds of samples sequencing costs per sample are crucial. Therefore we designed a highly multiplexed protocol, allowing sequencing of up to 384 barcoded *P. falciparum* amplicons in a single Illumina MiSeq run. Because multiple concurrent *P. falciparum* clones may differ greatly in density, sequencing analysis strategies need to identify alleles of very low abundance. To distinguish true minority clones from sequencing errors, quality checks were designed based on replicates of samples and integrated into the sequence analysis pipeline. The newly created data analysis software package was validated using experimental mixtures of *P. falciparum* in vitro culture strains, and tested on field samples.

Results

Marker selection

A protocol for deep sequencing and data analysis was developed for two molecular markers, namely the *P. falciparum csp* gene (PF3D7_0304600) and gene PF3D7_0104100, annotated in the malaria sequence database PlasmoDB as "conserved *Plasmodium* membrane protein" (*cpmp).* Results from these two markers were compared with classical length polymorphic genotyping using the highly diverse marker *msp2.* Sizes of *msp2* fragments amplified for genotyping range from 180 to 515 bp in PNG using published primers (Additional file 1: Table S1). Marker *csp* has been used for deep sequencing in the past [9] and the previously published primers (Additional file 1: Table S1) were used. The *csp* amplicon spans the T-cell epitope of the circumsporozoite protein from nucleotide position 858 to 1186 of the 3D7 reference sequence.

The newly validated marker *cpmp* was identified by calculating heterozygosity in 200 bp windows of 3′411 genomic *P. falciparum* sequences from 23 countries available from the MalariaGEN dataset [17]. Genes from multi-gene families or regions of poor sequence alignments, often caused by length polymorphism of intragenic tandem repeats, were excluded from the list of potential markers. A 430 bp fragment of *cpmp* spanning nucleotide positions 1895 to 2324 scored highest in expected heterozygosity (H_e) and was prioritized as candidate for a highly diverse amplicon sequencing marker. H_e in the worldwide dataset was 0.93 for *cpmp* compared to 0.86 for *csp* (Table 1, Additional file 1: Figure S1 and S2). Genomes originating from Papua New Guinea (PNG) revealed 9 haplotypes in 22 genomes for *cpmp* and 3 haplotypes in 30 genomes for *csp.*

Assessment of sequence quality

Csp and *cpmp* amplicons from 37 field samples and 13 mixtures of *P. falciparum* culture strains HB3 and 3D7 were sequenced on Illumina MiSeq in paired-end mode. A total of 5′810′566 paired raw sequences were retrieved. Of all reads, 326′302 mapped to the *phiX* reference sequence. 4′989′271 paired sequence reads were successfully de-multiplexed to yield a set of amplicon sequences per individual sample. 4′411′214 reads could be

Table 1 Diversity of markers *cpmp* and *csp* based on 3'411 genomes of the MalariaGen dataset

Marker	H_e^a	No. of SNPs	Fragment size[b]	No. of Haplotypes
cpmp	0.930[c]	20[c]	383[c]	82 of 980[c,d]
csp	0.857	40	287	77 of 1323[d]

[a] Expected heterozygosity
[b] Fragment size without primer sequence
[c] Trimming of reads in the here presented experiments led to a reduction of variation (Characteristics for a shorter *cpmp* fragment size of 310 bp: He = 0.913, SNPs = 14 and number of haplotypes = 47)
[d] From 3411 genomes only genomes with non-ambiguous SNP calls in selected region were used

assigned to individual amplicons. Median sequence coverage over all sequenced samples was 1′490 for *cpmp* (1st and 3rd quartiles: [537, 2183]) and 731 for *csp* (1st and 3rd quartiles: [524, 1092]). The discrepancy in median sequence coverage was deliberate and resulted from our pooling strategy to underrepresent *csp* amplicons to prevent their predominance in the sequencing library due to this amplicon's shorter length (Additional file 1).

The quality of the sequence run was assessed by investigating the sequencing error rate in sequence reads of the spiked-in *phiX* control. The mean mismatch rate per nucleotide of *phiX* control reads with respect to the *phiX174* genome was 5.2% (median 0.34%). The mismatch rate increased towards the end of sequence reads, up to 11% for forward reads and 54% for reverse reads (Additional file 1: Figure S3). To censor regions of high mismatch rates, forward and reverse sequence reads were trimmed before any further analyses to a length of 240 and 170 nucleotides, respectively. After trimming, the mean mismatch rate per nucleotide of *phiX* control reads was 0.50%.

As further quality check, the sequencing error rate was assessed in sequences of Linkers F and R (Additional file 1: Figure S4). These linker sequences never get amplified but are joined to the product in PCR, therefore any mismatch detected in these stretches will derive from either sequencing or initial primer synthesis. The average number of sequence mismatches in this part was 0.12% per sample per nucleotide (Additional file 1: Table S2 and Figure S5). The sequencing error rate also was assessed in regions corresponding to the primers of each marker (Additional file 1: Figure S4). Mismatches with respect to the known sequences of the PCR primers may derive from amplification errors or from errors in sequencing or primer synthesis during preparation of the sequencing library. The average number of mismatches in the primer regions was 0.28% for *cpmp* and 0.71% for *csp* per nucleotide per sample (Additional file 1: Figure S6 and Table S2).

Finally, the sequencing error rate was assessed in amplicons obtained from various mixtures of *P. falciparum* culture strains HB3 and 3D7. Potential sources of mismatches with respect to the reference sequence of strains 3D7 and HB3 include amplification error, sequencing error and errors due to de-multiplexing of samples [18]. The average number of sequence mismatches after trimming to lengths of 240 and 170 nucleotides respectively for forward and reverse reads was 0.38% for *cpmp* and 0.46% for *csp* (Fig. 1, Additional file 1: Table S2). This equates to 1–2 mismatches per read of 310 nucleotides. On average 87.5% of reads for *cpmp* and 85.5% for *csp* from mixtures of strains HB3 and 3D7 contained ≤2 mismatches per read with respect to the strains' reference sequences. Together the analyses of *phiX* and HB3/3D7 sequences indicated an intrinsic sequence error rate of 0.4–0.5%. The error rate of the linker sequence suggested that one third of these errors were sequencing errors, while two thirds were amplification errors.

Limit of detection assessed in serial dilutions of parasite culture

To test the feasibility to also genotype blood samples of low parasite density, serial dilutions of *P. falciparum* strain 3D7 over 5 orders of magnitude (5–50′000 parasites/μl) were sequenced (Additional file 1: Table S3). The 3D7 haplotype was detected in all dilutions. However, sequence coverage for dilutions harbouring 5 and 50 parasites/μl was below 550 reads. This indicated that the desired equimolar representation of amplicons was not achieved by our pooling strategy (Additional file 1 Pooling of samples - Pool for PCR without visible product on agarose gel). Our approach did not fully counterbalance lower amounts of amplicon.

Assessment of minority clone detectability

Defined mixtures of *P. falciparum* strains HB3 and 3D7 were sequenced to assess the detectability of minority clones under controlled conditions. The minority clone was detected in all tested dilution ratios up to 1:3000 (Table 2, Additional file 1: Table S4 and S5). Reads comprising obvious PCR artefacts (indels and chimeras) were detected in these mixtures up to a frequency of 0.48% for marker *cpmp* and 6.2% for *csp*. Up to 8.4% of reads for *cpmp* and 10.8% for *csp* were singletons or failed to cluster with 3D7 or HB3 haplotypes. This proportion of reads is therefore most likely an estimate of the cumulative background noise of the methodology. These reads fell below the default cut-off criteria (details below) and were thus excluded.

Simulations by bootstrap resampling were applied to estimate the probability to detect a minority clone at increasing sequencing coverage and decreasing ratios of the minority clone in a mixture of two strains. Resampling was repeated 1000 times and included only sequence data from mixtures of strains that were sequenced at a coverage of >3000 reads. At a coverage of 10′000 sampled reads the minority clone

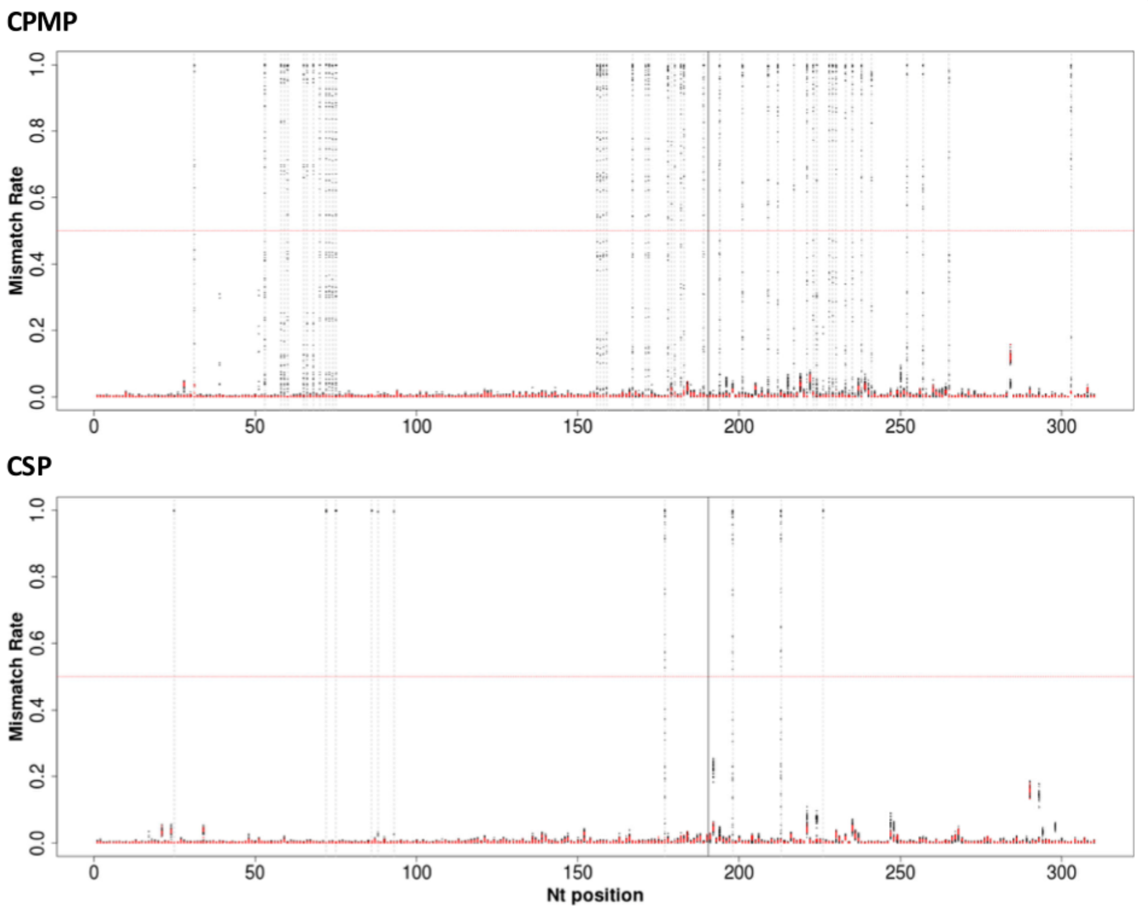


Fig. 1 Mismatch rate per nucleotide position derived from all samples sequenced for markers *cpmp* and *csp*. Each data point represents the mean observed mismatch rate observed in all reads of one sample at the respective nucleotide position. Red data points: control samples (*P. falciparum* culture strains); black data points: field samples; X-axis: nucleotide position in sequenced fragment; Y-axis: mismatch rate with respect to the reference sequence (for control samples: sequences of strains 3D7 and HB3, for field samples, 3D7 sequence); dashed grey lines represent SNPs with a mismatch rate of >0.5 in >1 sample; red dotted horizontal line indicates a mismatch rate of 0.5; solid black vertical line: position of concatenation of forward and reverse reads

was robustly detected at ratios 1:1 to 1:1000 for *cpmp* and up to 1:500 for *csp* (Fig. 2, Additional file 1: Figure S7 and S8). The cut-off set for haplotype positivity required that a haplotype was detected ≥3 times and represented ≥0.1% of all reads from the respective blood sample. More stringent criteria to call a haplotype (i.e. a higher minimum number of reads) would require a higher coverage for the detection of minority clones. Thus, more stringency in haplotype definition on the one hand reduces sensitivity, but increases specificity by eliminating false haplotypes attributable to background noise (Additional file 1: Figure S7 and S8).

Specification of default cut-off settings in software HaplotypR

Cut-off values for the analysis of sequencing data were defined to support removal of background noise caused by sequencing and amplification errors. The following values represent minimal stringency and can be adjusted to higher stringency to increase specificity in the HaplotypR pipeline:

(i) Cut-off settings for SNP calling were defined by a population-based approach. A SNP was required to be dominant (>50% of all reads) in ≥2 samples. A single dominant occurrence of a SNP is likely caused by amplification or sequencing error.
(ii) Cut-off settings for haplotype calling required a haplotype to be supported by ≥3 reads in ≥2 samples (including independent replicates of the same sample). Per haplotype a minimum of 3 reads are needed to distinguish SNPs from sequencing errors,

Table 2 Detectability of the minority clone in defined ratios of *P. falciparum* strains HB3 and 3D7

Ratios in mixtures	*cpmp*					*csp*					Minimum Coverage HaplotypR[c]
HB3:3D7	3D7[a] %	HB3[a] %	PCR artefacts %	Back-ground[b] %	Coverage	3D7[a] %	HB3[a] %	PCR artefacts %	Back-ground[b] %	Coverage	
1:1	34.6	57.4	0.48	7.53	40,768	34.7	50.5	5.79	9.01	9009	6
1:10	75.6	16.4	0.40	7.59	13,037	76.1	10.1	5.63	8.08	3341	30
1:50	88.8	3.15	0.06	7.95	4953	82.7	2.88	6.23	8.16	14,711	150
1:100	90.9	1.53	0.36	7.26	13,311	83.5	2.25	5.41	8.88	11,975	300
1:500	90.8	0.48	0.27	8.44	5649	84.0	0.46	4.76	10.8	3508	1500
1:1000	91.5	0.23	0.03	8.26	3039	85.7	0.22	5.09	9.02	1807	3000
1:1500	92.5	0.11	0.48	6.94	55,887	86.3	0.08[d]	5.71	7.91	23,619	4500
1:3000	92.5	0.09[4]	0.38	7.00	7417	85.0	0.04[d]	5.87	9.10	2318	9000

[a] Percent of reads that cluster with 3D7 and HB3 reference sequences

[b] Singleton reads and reads that failed to cluster with 3D7 or HB3 haplotypes

[c] Theoretical minimum required coverage needed to detect minority clone by software HaplotypR with default cut-off values

[d] Haplotypes considered as noise by software HaplotypR (default cut-off: ≥3 reads per haplotype and a minority clone detection limit of 1:1000)

because a consensus sequence cannot be determined from 2 disparate reads alone. Random sequencing and amplification errors would unlikely lead repeatedly to a particular haplotype.

(iii)Cut-off settings for calling minority clones required that a minority clone would represent at least 0.1% of all reads of a sample, which corresponds to a detection limit for minority clones of 1:1000. For the current project, the cut-off was justified by the results obtained from artificial mixtures of culture strains, which defined the technical limit of detection for a minority clone. This parameter may be set to more stringent values.

Application of these three default cut-off values to mixtures of culture strains had the effect that HaplotypR missed the minority clone for both markers in the greatest dilution ratio of the two strains tested (1:3000). For marker *csp* the minority clone fell below the cut-off even in the 1:1500 ratio (Table 2). No false-positive haplotypes were called after applying default cut-off criteria, even in samples with a very high coverage in the controlled mixtures (up to 55′000 reads) and in simulations by bootstrapping (up to 100′000 sampled reads) (Additional file 1: Table S4 and S5).

Validation of SNP calling

The above criteria were validated on reads from culture strains and primer sequences. The background sequencing error rate at each individual nucleotide position was measured to distinguish sequencing and amplification errors from true SNPs. Mismatch rates of up to 22% was measured in primer sequences (Additional file 1: Figure S6), and 18% in amplicons from culture strains (Fig. 1, Additional file 1: Table S2). None of these mismatches

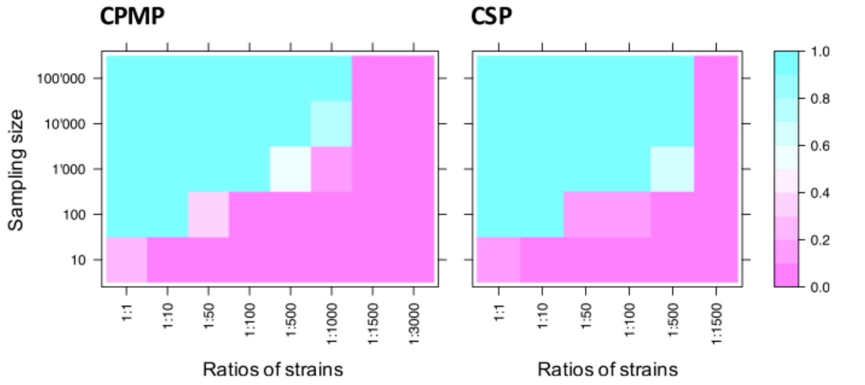


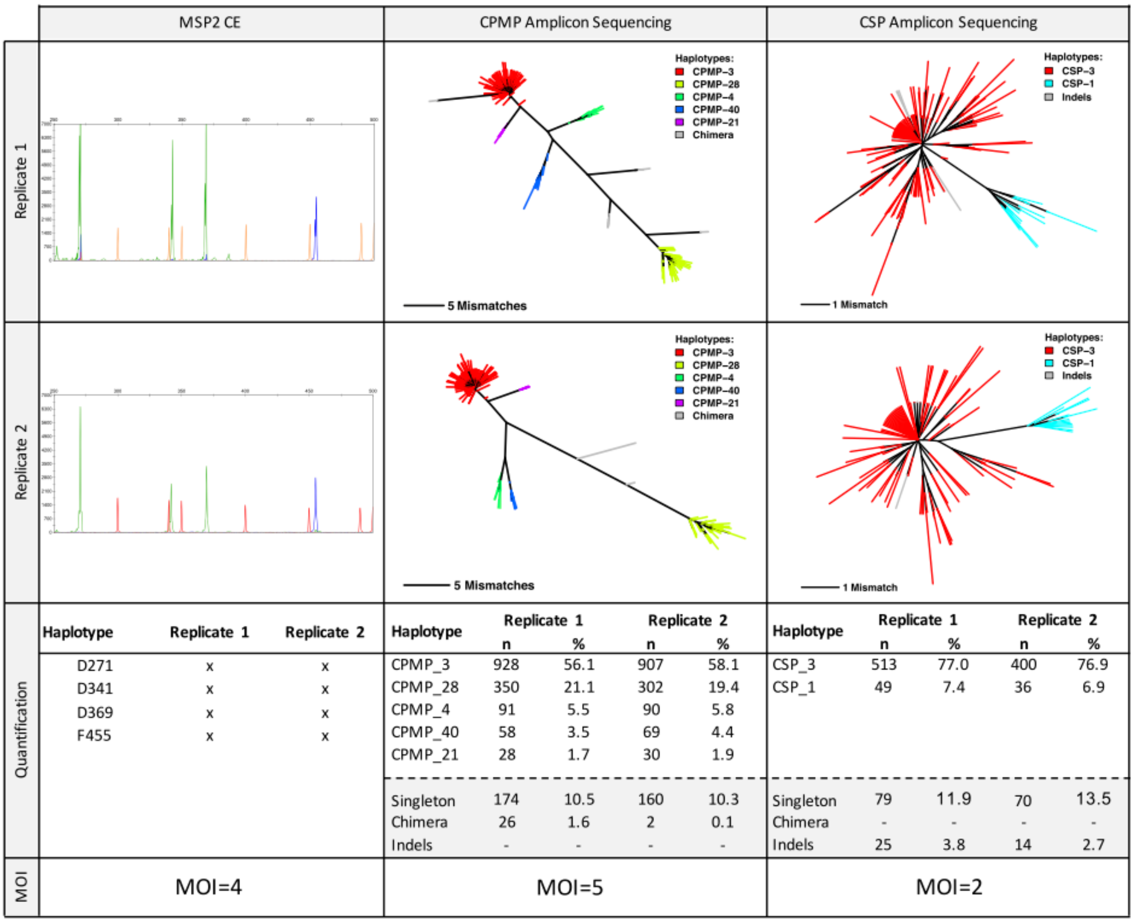
Fig. 2 Simulation of minority clone detectability by bootstrapping for marker *cpmp* and *csp*. Cut-off for acceptance of a haplotype was a minimum coverage per haplotype of 3 reads and a minority clone detection limit of 1:1000. Samples were drawn from reads of defined mixtures of *P. falciparum* strain 3D7 and HB3. X-axis shows dilution ratios of strains 3D7 and HB3; Y-axis indicates the sampling size (number of draws from sequence reads) for each mixture of strains. Sampling was repeated 1000 times to estimate mean minority clone detectability

led to calling of a SNP after the above cut-off was applied (i.e. >50% of reads in ≥2 samples).

Validation of amplicon sequencing in field samples

37 *P. falciparum* samples from PNG were genotyped by amplicon sequencing. Dendrograms were produced for each marker from raw sequencing reads (Fig. 3, Additional file 1: Figure S9). Branch lengths in these dendrograms represent the number of SNPs that differ between any sequences compared. Branches with sequences belonging to the same haplotype (defined as "clusters") are labelled in the same colour. Haplotype frequencies within each individual sample were determined from the reads of the sample before applying cut-offs (Fig. 3, panel "Quantification"). When analysing the genetic diversity in field sample, haplotypes were only counted as true haplotypes if both replicates pass the haplotype calling cut-off. This more stringent criterion was introduced to prevent erroneous over-estimation of multiplicity due to false haplotypes.

All samples were genotyped for length polymorphic marker *msp2* using capillary electrophoresis (CE) for fragment sizing. *Msp2* genotyping was reproducible and consistent between different laboratories (Fig. 3, Additional file 1: Figure S9: left column). A mean multiplicity of infection (MOI) of 2.2 was observed in 37 field samples analysed by *msp2* genotyping and 25/37 (67.5%) of samples harboured multiple clones (Fig. 4, Additional file 1: Table S6). Mean MOI and H_e were compared between the



Haplotype	Replicate 1	Replicate 2
D271	x	x
D341	x	x
D369	x	x
F455	x	x
MOI=4		

Haplotype	Replicate 1		Replicate 2	
	n	%	n	%
CPMP_3	928	56.1	907	58.1
CPMP_28	350	21.1	302	19.4
CPMP_4	91	5.5	90	5.8
CPMP_40	58	3.5	69	4.4
CPMP_21	28	1.7	30	1.9
Singleton	174	10.5	160	10.3
Chimera	26	1.6	2	0.1
Indels	-	-	-	-
MOI=5				

Haplotype	Replicate 1		Replicate 2	
	n	%	n	%
CSP_3	513	77.0	400	76.9
CSP_1	49	7.4	36	6.9
Singleton	79	11.9	70	13.5
Chimera	-	-	-	-
Indels	25	3.8	14	2.7
MOI=2				

Fig. 3 Comparison of genotyping by length-polymorphic marker *msp2* and amplicon sequencing of *cpmp* and *csp*. Raw data from length-polymorphism- and SNP-based genotyping for one *P. falciparum*-positive field sample. Left panel: Capillary electropherograms (CE) for *msp2* nested PCR products (duplicate experiments); X-axis: fragment length, Y-axis: peak heights (arbitrary intensity units); size standards: red/orange peaks; 3D7-type *msp2* genotypes: green peaks; FC27-type *msp2* genotypes: blue peaks. Middle and right panel: Dendrograms derived from sequence reads of marker *cpmp* (middle) and *csp* (right); coloured lines represent membership to a specific, colour-coded haplotype; Grey lines: sequence reads of PCR artefacts (later excluded by cut-off settings); line length: number of mismatches according to bar insert. Bottom panels: Read counts (n) and percentage of reads (%) per haplotype and final multiplicity call

genotyping methods (Table 3, Fig. 4, Additional file 1: Table S6). The resolution of marker *cpmp* was slightly higher than that of *msp2* with 27 *cpmp* haplotypes versus 25 *msp2* alleles, H_e of 0.96 versus 0.95 and a higher mean MOI of 2.41 versus 2.19, respectively. Overall the two methods agreed well, with good concordance of MOI (Cohen's Kappa 0.71, equal weights, z = 6.64, *p*-value = 3.04e-11). Compared to *msp2* the discriminatory power of *csp* was substantially lower with only 4 *csp* haplotypes found in 37 samples, H_e of 0.57 and mean MOI of 1.54. Concordance between *csp* and *msp2* MOI was poor (Cohen's Kappa 0.38, equal weights, z = 4.48, p-value = 7.61e-6).

Reproducibility of amplicon sequencing in field samples

Csp and *cpmp* haplotypes obtained from 37 field samples were compared between replicates to investigate reproducibility of the molecular and bioinformatic analyses. For both replicates of the field samples the default cut-off criteria for haplotype calling (≥3 reads and minority clone detection limit of 1:1000) were applied. Concordance between replicates was very good with Cohen's Kappa 0.84 (equal weights, z = 7.769, p-value = 7.99e-15) for *cpmp* and 0.91 (equal weights, z = 6.466, p-value = 1.01e-10) for *csp*. Comparison of replicates permitted to investigate the amount of false haplotype calls. True haplotypes should be detected in both replicates, unless the sequence depth

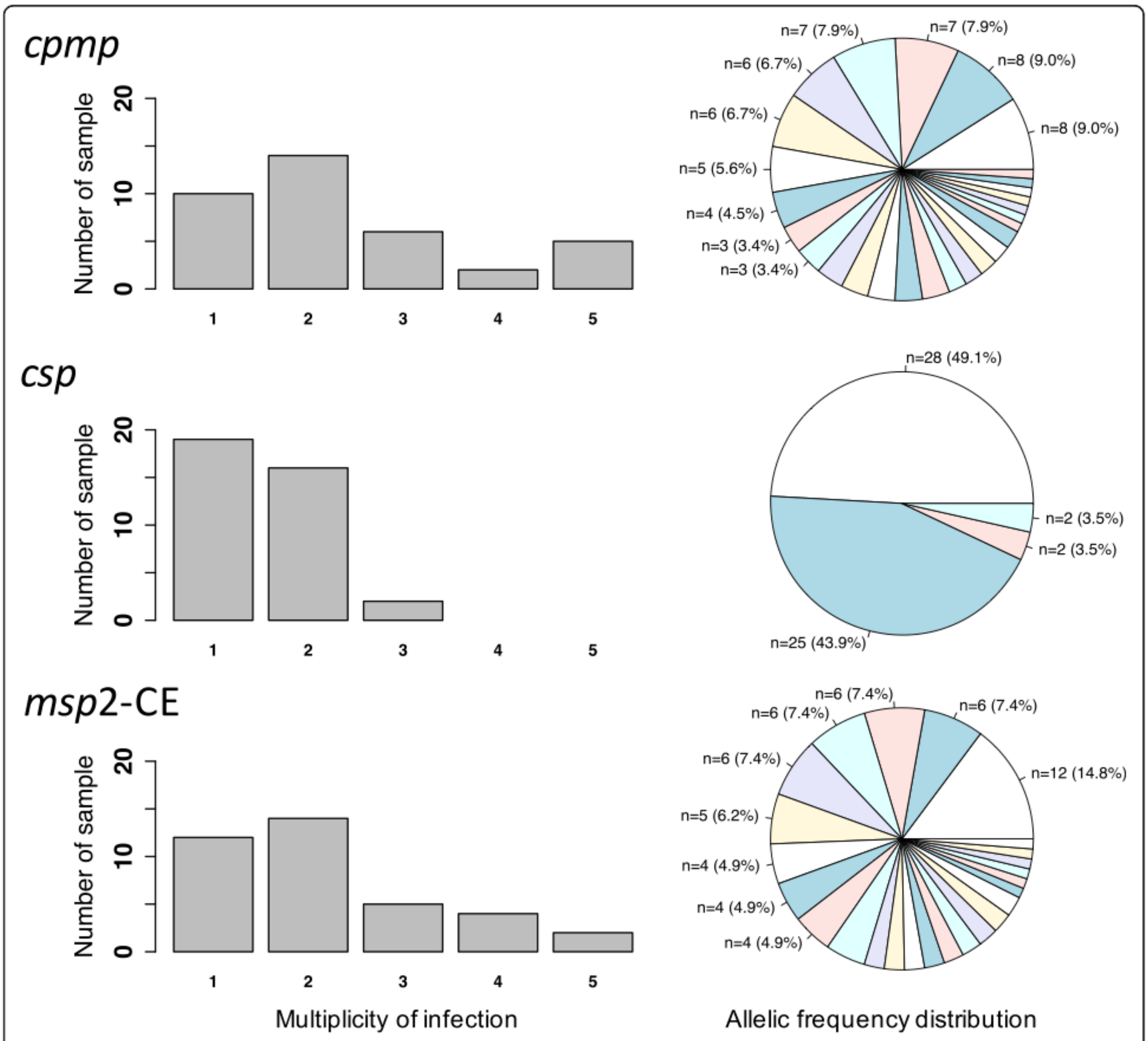


Fig. 4 Frequency distribution of multiplicity of infection and allelic frequencies of *cpmp*, *csp* and *msp2*-CE. 37 *P. falciparum* positive samples from PNG were analysed for the 3 markers *cpmp* (27 haplotypes), *csp* (4 haplotypes) and *msp2*-CE (25 genotypes). Pie charts represent allelic frequency distribution for each marker in 37 samples

Table 3 Summary of genotyping results from three molecular markers analysed in 37 field samples

Marker	H_e	Mean MOI	Number SNPs[a]	Number Haplotypes	Concordance of MOI K
msp2 CE	0.948	2.19	NA	25	Reference
cpmp	0.957	2.41	45	27	0.71[b] (good)
csp	0.574	1.54	10[c]	4	0.38[d] (poor)

[a] With respect to the reference sequence of *P. falciparum* strain 3D7
[b] Cohen's Kappa (2 raters, weights = equal): z = 6.64, p-value = 3.04e-11
[c] 4/10 SNPs are fixed within these 37 field samples
[d] Cohen's Kappa (2 raters, weights = equal): z = 4.48, *p*-value = 7.61e-6

is not sufficient for detecting a minority clone in one of the replicates. *Cpmp* minority clones that had a frequency > 1.0% of all reads were consistently detected with ≥3 reads in both replicates (Table 4, Additional file 1: Figure S10). For *csp* this was achieved for minority clones with a frequency of >0.70%. 18 *cpmp* haplotypes were detected with ≥3 reads in only one of the replicates. In three instances one of the replicate did not pass the cut-off criteria due to low coverage. For marker *csp*, 2 haplotypes with ≥3 reads were detected in one replicate only. In summary, a comparison of replicates indicated 15 potentially false haplotype calls for *cpmp* and 2 for *csp*. These calls stem from reads with a frequency < 1%, Therefore, performing replicates are essential to prevent erroneous overestimation of multiplicity due to false haplotypes.

An attempt was made to investigate the influence of the number of PCR cycles performed during amplicon library preparation on the generation of artefacts. This was possible by using 25 and 15 cycles in the nested PCR for replicate 1 and 2, respectively. Cycle number had no influence on the proportion of singleton and indel reads. However, the proportion of chimeric *cpmp* reads was higher in replicate 1 using 25 cycles than in replicate 2 using 15 cycles (0.63% versus 0.13%, Student's t-Test *P*-value = 0.0221). No chimeric *csp* reads were detected in the field samples (Table 5).

Discussion

This report presents the development of a new genotyping methodology for *P. falciparum* based on amplicon deep sequencing. The search for new markers was prompted by severe limitations of length polymorphic markers, which represent the currently used standard for genotyping malaria parasites. A strong bias towards preferential amplification of shorter fragments in multi-clone infections was observed, so that larger fragments were lost even if only 5-fold underrepresented compared to shorter fragments from the same sample [7]. This called for an alternative approach that relies on haplotypes created from several SNPs rather than length polymorphism. With respect to minority clone detectability, amplicon sequencing overcame this pitfall of length polymorphism methods and also performed very well in field samples.

Amplicon sequencing showed an excellent resolution when using the novel genotyping marker *cpmp* (PF3D7_0104100). The strategy applied for down-selecting highly diverse regions in the genome suggested *cpmp* as the top candidate. *Cpmp* is most abundantly expressed in sporozoite stages [19], but the function of the encoded protein is unknown. The gene is under balancing selection with a Tajima's D of 1.16 in Guinea and 1.05 in Gambia [20]. In this study, *cpmp* revealed a genetic diversity similar to the length polymorphic region of the widely-used marker *msp2*. 45 SNPs were observed in the 37 field samples of this study, leading to the designation of 27 haplotypes for marker *cpmp*. With increasing number of field samples processed, additional rare SNPs and even more haplotypes are likely to be found. The diversity of *cpmp* was high also in the global MalariaGEN

Table 4 Concordance of haplotype calls in replicates of 37 field samples

	cpmp	*csp*	Passed cut-off[a]
Haplotype frequency within sample ≥ 1%			
present in both replicates	**87**	**57**	**yes**
present in single replicate only	0	0	no
Haplotype frequency within sample < 1%			
present in both replicates at ≥ 3 reads[b]	**2**	**0**	**yes**
present in both replicates one ≥ 3 reads[b] and one < 3 reads[b]	1[c]	0	yes/no[d]
present in single replicate at ≥ 3 reads[b]	17[e]	2	yes/no[d]
present in both replicates at < 3 reads[b]	1	0	no
present in single replicate at < 3 reads[b]	10	5	no

Bold rows indicate haplotypes that did pass cut-off criteria in both replicates
[a] Default cut-off criteria to accepted haplotype ≥3 reads and a minority clone detection limit of 1:1000
[b] Owing to default cut-off for haplotype call
[c] Second replicate had too low coverage to detect ≥3 reads
[d] Potential false haplotype calls as only one replicate passed cut-off criteria
[e] In 2 instances second replicate had too low coverage to detect minority clone

Table 5 Mean proportion of singleton or chimeric reads and indels detected in both field sample replicates

Marker	Replicate 1			Replicate 2		
	Singletons %	Indels %	Chimera %	Singletons %	Indels %	Chimera %
csp	11.55	3.78[a]	0.00	11.47	4.05[a]	0.00
cpmp	9.76	0.073[b]	0.631[c]	9.74	0.034[b]	0.130[c]

[a] Marker *csp*: Indels Replicate 1 versus 2; Student's t-Test: $t = -1.336$, df = 71.052, p-value = 0.1858

[b] Marker *cpmp*: Indels Replicate 1 versus 2; Student's t-Test: $t = 1.3304$, df = 71.94, p-value = 0.1876

[c] Marker *cpmp*: Chimera Replicate 1 versus 2; Student's t-Test: $t = 2.3552$, df = 55.4, p-value = 0.02208

dataset (H_e = 0.93); its resolution as genotyping marker in other geographic regions remains to be shown. In contrast, marker *csp*, analysed in parallel to *cpmp* and also used in earlier studies, showed a limited diversity with only 4 haplotypes detected in 37 field samples. Earlier studies reported similar low diversity for *csp* in regions of Asia Pacific [21]. Thus, *csp* is not suited to serve as a single genotyping marker in PNG. However, the global diversity of *csp* according to the MalariaGEN dataset seems to be high (H_e = 0.86), and high diversity has also been observed in African isolates [21].

Implementing amplicon sequencing required parallel development of a bioinformatics pipeline. A known problem in sequence analysis is the robust detection of minority clones from a background of experimentally induced artefacts. We addressed this problem with the design of HaplotypR, a software package dedicated to stepwise analyse of sequence reads for samples containing multiple clones. The HaplotypR pipeline can be divided into three steps: In the first step, this pipeline de-multiplexes and clusters raw sequence reads to clusters of related sequences, so called "representative haplotypes". This step employs Swarm2 software, which expands pools of amplicons (identical sequence reads) by iteratively joining other pools of amplicons that are separated by a defined number of mismatches (e.g. one substitution, insertion or deletion) [22, 23]. This strategy permits unbiased clustering of sequence reads without the need to define a list of SNPs. This enables capturing of previously unknown SNPs without any adjustments to the pipeline. In the next step HaplotypR checks all representative haplotypes for presence of PCR artefacts (indels and chimeras), and labels and censors these. In the final step HaplotypR removes background noise by applying defined cut-offs and reports a list of final haplotypes calls.

Validation of HaplotypR was made possible by reads from serial dilutions of *P. falciparum* culture strain 3D7 and from controlled mixtures of strains HB3 and 3D7. On those control samples the impact of amplification and sequencing errors could be assessed. An increased frequency of sequence mismatches relative to the 3D7 reference sequence of up to 22% was observed at a few specific genomic locations including the sequences of amplified primers. To differentiate these sequencing errors from true genotypes of rare minority clones, we defined a SNP calling cut-off where a genotype was required to be dominant (>50% of all reads) in at least 2 samples. This cut-off is critical to distinguish true positive genotypes that are rare in the population from sequencing errors.

To prevent reporting of false haplotypes, HaplotypR pipeline applies two types of cut-offs: firstly, a cut-off for singleton exclusion, whereby a SNP or haplotype needed to be supported by more than one sample. It is unlikely that these cut-offs would remove true haplotypes, except if the sample size was very small. In this case, it is recommended to amplify and sequence samples in duplicate, as in this study. A true haplotype is expected to be present in both replicates and thus will not get excluded. Secondly, a cut-off for haplotype coverage was defined requiring that a haplotype is supported by a user-defined number of sequence reads. This flexible cut-off can be selected for each marker. The coverage cut-off removes false or weakly supported haplotypes and thus improves specificity. On the other hand, the ability to detect minority clones (i.e. sensitivity) will be limited by a cut-off based on coverage. Sequence reads from a minority clone were detected in all ratios up to 1:3000 in the mixtures of strains HB3 and 3D7. However, due to high background noise, false haplotypes with a frequency of up to 0.01% were also detected, making the definition of a cut-off to remove background noise obligatory. Applying these default cut-offs in HaplotypR decreased minority clone detectability from 1:3000 to 1:1000.

In an other publication a parasite density specific cut-off was applied in addition to a default cut-off [24].

The potential of amplicon sequencing for genotyping samples of very low parasitaemia was assessed in serial dilutions of strain 3D7. Sequence reads were retrieved from samples of a parasitaemia as low as 5 parasites/μl, however coverage was below 100 reads for the lowest level of parasitaemia. To reliably genotype samples spanning a wide range of parasitaemias, similar sequence coverage (and thus unbiased normalization of input material) for all samples is needed. The inexpensive strategy used to adjust amplicon concentrations of individual samples to equal levels prior to pooling for highly multiplexed sequencing still resulted in fluctuation in the sequence coverage, but a commercial DNA normalisation kit may improve equimolar pooling of samples [25, 26].

All samples in this study were sequenced in 2 replicates. This was done to assess the reproducibility of amplicon sequencing method of genotyping very low abundant minority clones, and to investigate the effect of nested PCR cycle number on artefacts. Analysing replicates of field samples revealed that haplotypes with a frequency of >1% were consistently detected in both replicates. In contrast, haplotypes with a frequency of <1% were frequently detected only in a single replicate. If minority clones of <1% frequency are to be reliably detected, amplifying and sequencing two or more replicates for each sample would be essential to call true haplotypes.

To detect minority clones with high sensitivity and specificity, samples need to be sequenced at high coverage and in replicates. As sensitivity may be adjusted by sequence coverage, choices have to be made in a trade-off between sequencing costs and sensitivity. The specific genotyping application can guide this choice. For example, in large scale field studies with many samples, a high degree of multiplexing of samples at moderate sequence coverage may be chosen to keep sequencing cost low. Furthermore, a less sensitive approach without performing replicates may be sufficient when detection of very rare minority clones is less of an issue. Another important application of genotyping of malaria parasites is the example of "recrudescence typing" during in vivo drug efficacy trials. To distinguish a new infection from one present as a minority clone prior to drug administration requires highest sensitivity and every clone must be reliably detected. In such cases a sequencing approach with less multiplexing is desired to achieve high coverage and maximal detection of minority clones.

The power of high sequencing coverage was shown for example in a study assessing the subclonal diversity in carcinomas [27]. Minority variants with a frequency of 1:10′000 were detected with a sequence depth of 100′000 reads per sample. Our results reported from malaria field samples does not have sufficient sequence depth to achieve such sensitivity, as median sequencing depth per sample was 1′490 reads for *cpmp* and 731 reads for *csp* owing to a high number of samples and of markers sequenced in parallel. A total of 352 samples were multiplexed in a single sequence run. Samples simultaneously processed but not included in the present analysis served for an unrelated research question. According to our protocol for PCR-based sequencing library preparation, costs per sample for Amp-Seq were twice that of msp2-CE genotyping [5]. Thus, the approach applied by us is cost effective as it permits parallel processing of several hundred samples, a range typically encountered in population-wide studies.

Targeted amplicon sequencing is not only used for investigating genetic diversity of *Plasmodium* parasites, but also widely applied in other fields, e.g. to study diversity of other pathogens, diversity in eco-systems or sequence alteration caused by CRISPER/Cas9 [8–11, 24, 28–31]. For pooling of multiple samples in one sequencing run, individual samples are generally either labelled by ligating a sequencing adapter that carries an index sequence [28] or by amplification with the sequencing adapter carrying an index and linker sequence [29, 30]. Data analysis follows two main strategies to retrieve haplotypes either by clustering of the full sequence read [10, 22, 24] or by SNP calling and optional haplotype inference [13, 28].

Conclusions

Short amplicon sequencing has the advantage that no multi-locus haplotype reconstruction is needed, as all SNPs are linked by a single paired-end read. This allows the reliable analysis of samples of very high MOI, a prerequisite for genotyping in areas of high malaria endemicity. An additional strength of this method is that previously undescribed or newly evolving haplotypes can be captured without any adjustment of the typing methodology or the HaplotypeR pipeline. The main limiting factor for the detection of minority clones was the sequence depth per sample. The sequence coverage in the present study was in the order of 1000 reads (median number of reads for *cpmp* was 1490 and for *csp* 731). This permitted detection of minority clones at a frequency of >0.3% of the total parasite load. To robustly detect minority clones at 0.1% frequency, a coverage of 10′000 reads is recommended. In addition, experiments should be performed in duplicate. The need to detect such low-abundance clones depends on the specific research question, which should guide experimental decision on number of samples and multiplexed amplicons as well as on the desired sequence depth.

The specification of amplification and sequencing errors presented here as well as the developed bioinformatic tools to handle such complex analytical tasks are relevant to all amplification-based genotyping methods of multiple clones or quasi-species within a sample. The newly developed pipeline can be used to analyse any amplicon sequencing based genotyping data irrespective of marker or organism.

Methods

Parasite genomic DNA

P. falciparum in vitro culture strains HB3 and 3D7 were mixed in 8 different proportions to generate well defined control samples with known MOI and well defined ratios of genomes. The ratios in these HB3-3D7 mixtures ranged from 1:1 to 1:3′000. Five additional control samples represented a dilution series of strain 3D7 with parasite densities ranging from 50′000 to 5 parasite/μl. Dilutions were prepared in human gDNA to reconstitute the nucleic acid concentration of a human blood sample. Details of parasite quantification were published previously [7]. Thirty-seven archived field samples from a cohort study conducted in East Sepik Province, Papua New Guinea (PNG) in 2008 were used to validate the performance of protocols for genotyping and data analysis in natural *P. falciparum* infections [32].

Genotyping using length polymorphic marker *msp2*

For determination of mean MOI field samples were genotyped using the classical *P. falciparum* genotyping marker *msp2*. Fluorescently labelled nested PCR products were sized by CE on an automated sequencer and analysed using GeneMapper software according to previously published protocols [5]. Each DNA sample was genotyped twice in independent laboratories to assess reproducibility of clone multiplicity (Fig. 3 and Additional file 1: Figure S9).

Amplicon deep sequencing marker selection and assay development

3′411 genomes from 23 countries, published by the *Plasmodium falciparum* Community Project (MalariaGEN), were screened to identify highly diverse markers for SNP-based genotyping [17]. The *P. falciparum* genomes were divided in 200 bp windows and H_e was calculated for each window as follows: $H_e = \frac{n}{n-1}\left[1-\sum p_i^2\right]$ where n is the number of clones and p_i the frequency of allele i. Annotated genes (PlasmoDB v11.0) that overlapped with windows of high heterozygosity were selected for further evaluation. Genes belonging to gene families, such as *var, rifin, stevor* and surf families, were excluded from the list, as well as genes with high heterozygosity that is usually caused by length polymorphism (Additional file 2).

Primers for marker *cpmp* were designed manually. Location of primers was selected to flank a region of maximum diversity (Additional file 1: Figure S1 and S2). Amplicon sizes were limited to a maximum of 500 bp to conform to possible read lengths of the Illumina MiSeq platform. Quality control of primers was assessed with online tools for secondary structure and primer dimer interaction (https://www.thermofisher.com/us/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html) [33]. Primer sequences are listed in Additional file 1: Table S1 and H_e values for amplicons are shown in Additional file 1: Figure S1 and S2.

Sequencing library preparation

The sequencing library was generated by 3 rounds of PCR with KAPA HiFi HotStart ReadyMix PCR Kit as described earlier [30]. A first round of 25 cycles amplified the gene of interest. A second marker-specific nested PCR amplified the primary product with primers that carried a 5′ linker sequence. We compared different cycle numbers for this second round: 25 cycles for replicate 1 and 15 cycles for replicate 2. This comparison was done to test for effects of cycle number on sequence diversity caused by imperfect polymerase fidelity [34]. To allow pooling and later de-multiplexing of amplicons, a third and final amplification was performed using primers binding to the F and R Linker sequence at the 3′ end, that introduced a sample-specific molecular barcode sequence plus the Illumina sequence adapter at the 5′ end. The relative positions of all these elements are depicted in the schematic in Additional file 1: Figure S4. A detailed PCR protocol containing primer sequences, cycle conditions und pooling steps are described in Additional file 1.

PCR products were purified with NucleoMag beads. The expected fragment size of the sequencing library was confirmed by Agilent 2200 Tapestation System. DNA concentration of the sequencing library was quantified by Qubit Fluorometer (Thermo Fisher Scientific). Sequencing was performed on an Illumina MiSeq platform in paired-end mode using Illumina MiSeq reagent kit v2 (500-cycles) together with a Enterobacteria phage PhiX control (Illumina, PhiXControl v3).

Bioinformatic analysis pipeline "HaplotypR"

Sequence reads were mapped with bowtie2 (parameter: end-to-end and sensitive) [35] to the *phiX174* genome (Accession: J02482.1) for assessing the quality of the sequencing run and calculating sequencing error per nucleotide position. Reads were then de-multiplexed to separate individual samples and different genotyping markers (Fig. 5). Primer sequences were truncated, the sequence was trimmed according to the quality of the *phiX* control sequence reads and paired reads were fused together.

For analysis of control samples, fused reads were mapped to the corresponding primers and *P. falciparum* reference sequences of strains 3D7 and HB3 (Accession: AL844502.2, AL844501.2, AB121018.1, AANS01000117.1). Rates of mismatches to primer and reference sequences were calculated for each individual sample at each nucleotide position. A SNP was defined as a nucleotide position with a > 50% mismatch rate in the sequence reads from at least two independent samples.

For prediction of haplotypes, fused reads were clustered individually per sample with Swarm2 software (parameters: boundary = 3 and fastidious mode) [22, 23]. The centre of each cluster represents the most abundant sequence of the cluster and thus constitutes a predicted haplotype. The cluster size represents the within-sample clone frequency in the tested sample. Haplotypes with a cluster size of 1 were classified as singletons and considered background noise. Haplotypes were checked for PCR artefacts such as indels and chimeric reads. Indels are caused by polymerase slippage which occurred primarily at stretches of homopolymers. Chimeric reads, caused by incomplete primer extension and inhomologous re-annealing, were identified

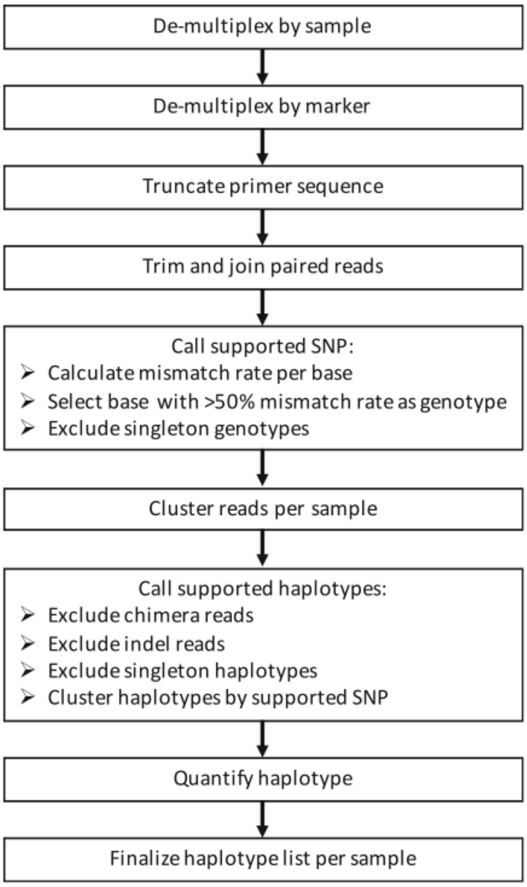

De-multiplex by sample
De-multiplex by marker
Truncate primer sequence
Trim and join paired reads
Call supported SNP:
➢ Calculate mismatch rate per base
➢ Select base with >50% mismatch rate as genotype
➢ Exclude singleton genotypes
Cluster reads per sample
Call supported haplotypes:
➢ Exclude chimera reads
➢ Exclude indel reads
➢ Exclude singleton haplotypes
➢ Cluster haplotypes by supported SNP
Quantify haplotype
Finalize haplotype list per sample

Fig. 5 Bioinformatic analysis pipeline applied on highly multiplexed deep sequencing data

with vsearch software (parameters: uchime_denovo mode, mindiffs = 3, minh = 0.2) [36]. To distinguish chimera haplotypes resulting from PCR artefacts from true recombined haplotypes, a population-wide approach (combining all samples of the entire study) is implemented in HaplotypR. A chimera was classified as such if a haplotype was identified as chimera by vsearch at all instances it occurred. On the other hand, if a chimera was detected in only some of the samples, it was not classified as chimera, but as a true haplotype. However, in such instances this haplotype was always flagged and the outcome "true chimera" or "true haplotype" was resolved by using replicates. This approach is justified, as it is expected that a true recombinant haplotype would be transmitted without its parent haplotypes.

The full analysis pipeline, named HaplotypR, was implemented as R package and is illustrated in Figure 5 (https://github.com/lerch-a/HaplotypR.git).

Estimated detectability of minority clones by sampling

Detectability of minority clones was estimated by bootstrapping from the reads of the control samples with defined HB3-3D7 strain ratios. Reads were randomly sampled with replacement until the required coverage was reached. These resampled set of reads were processed in the same manner as the original samples using HaplotypR. For resampling only sequence files from HB3-3D7 mixtures were used that had a coverage of >3000 reads.

Additional files

Additional file 1: Supporting information. Supplementary text, figures and tables. (DOCX 2111 kb)

Additional file 2: List of 200 bp H_e windows of whole *P. falciparum* genome. (XLSX 96 kb)

Additional file 3: List of haplotype calls. (XLSX 67 kb)

Abbreviations

CE: Capillary electrophoresis; CPMP: PF3D7_0104100, conserved *Plasmodium* membrane protein; CSP: Circumsporozoite protein; H_e: Expected heterozygosity; indel: Insertion and deletion; MOI: Multiplicity of infection; MSP: Merozoite surface proteins; PCR: Polymerase chain reaction; PNG: Papua New Guinea; SNP: Single nucleotide polymorphism; SRA: Sequence read archive

Acknowledgements

We are grateful to the study participants and their guardians and to the field and laboratory team of the PNG Institute of Medical Research, in particular Anna Rosanas-Urgell and Alice Ura. This publication uses data from the MalariaGEN *Plasmodium falciparum* Community Project as described in Genomic epidemiology of artemisinin resistant malaria, eLife, 2016 (DOI: https://doi.org/10.7554/eLife.08714).

Funding

This work was supported by the Swiss National Science Foundation [310030_159580] and the International Centers of Excellence in Malaria Research [U19 AI089686]. AL was partly funded by Novartis Foundation for Medical-Biological Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

Sequencing data have been deposited at the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRR5412316. A list with the haplotype calls can be found in Additional file 3. HaplotypR can be downloaded from https://github.com/lerch-a/HaplotypR.git.

Authors' contributions

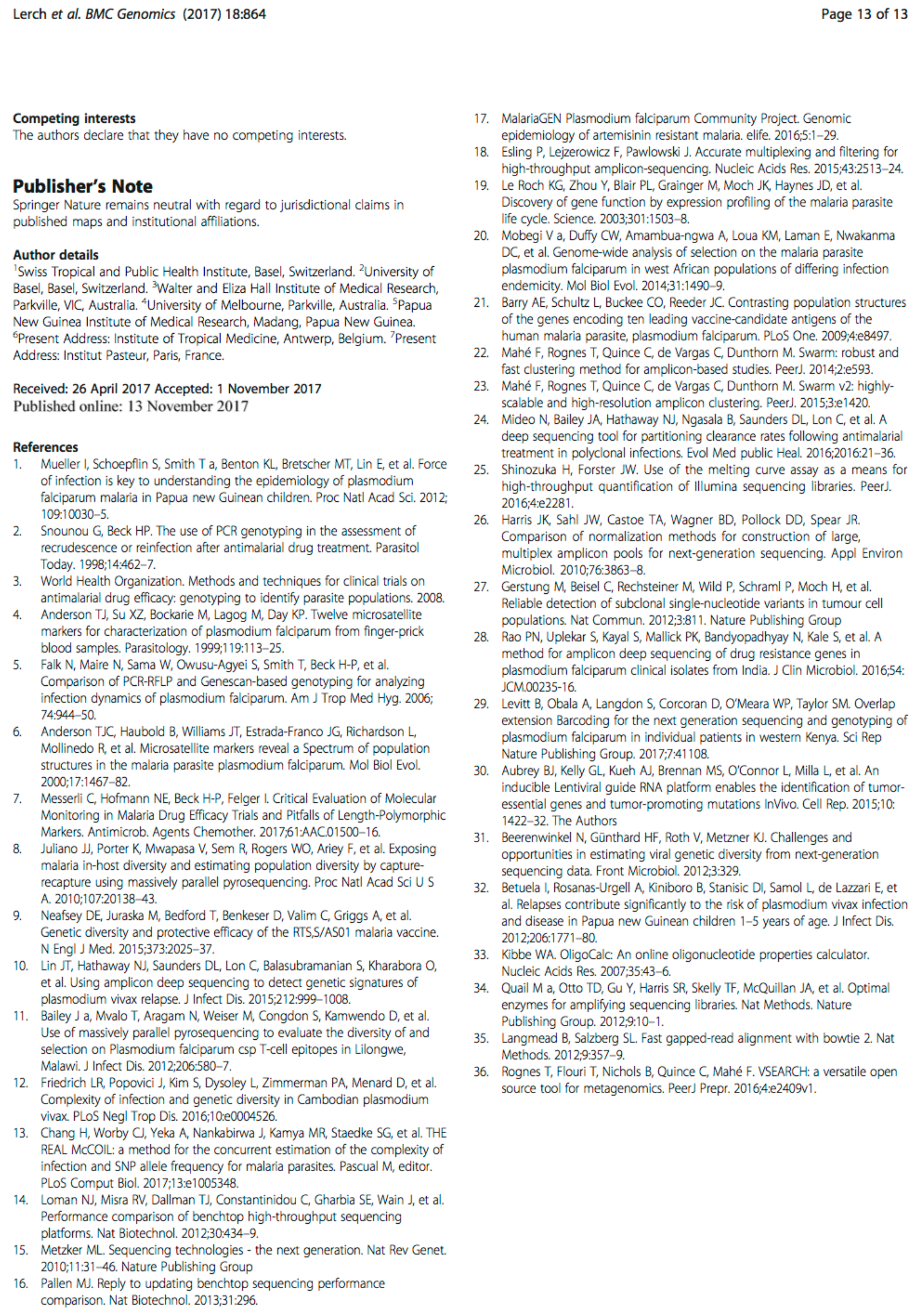
Conceived and designed the experiments: IF, IM, AL, CK, SW. Performed the experiments: AL, CK, JHK, NH, CM, SW. Supervised field work and responsible for acquisition of samples: IB. Analysed the data: AL. Supervision: IF, IM, LOC. Writing - draft: AL, IF. Writing - review & editing: CK, NH, JHK, IM, LOC. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Ethical clearance was obtained from PNG Institute of Medical Research Institutional Review Board (IRB 07.20) and PNG Medical Advisory Committee (07.34). Informed written consent was obtained from all parents or guardians prior to recruitment of each child. No medical records were used for this study.

Consent for publication

Not Applicable.



Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Swiss Tropical and Public Health Institute, Basel, Switzerland. [2]University of Basel, Basel, Switzerland. [3]Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia. [4]University of Melbourne, Parkville, Australia. [5]Papua New Guinea Institute of Medical Research, Madang, Papua New Guinea. [6]Present Address: Institute of Tropical Medicine, Antwerp, Belgium. [7]Present Address: Institut Pasteur, Paris, France.

Received: 26 April 2017 Accepted: 1 November 2017
Published online: 13 November 2017

References

1. Mueller I, Schoepflin S, Smith T a, Benton KL, Bretscher MT, Lin E, et al. Force of infection is key to understanding the epidemiology of plasmodium falciparum malaria in Papua new Guinean children. Proc Natl Acad Sci. 2012; 109:10030–5.
2. Snounou G, Beck HP. The use of PCR genotyping in the assessment of recrudescence or reinfection after antimalarial drug treatment. Parasitol Today. 1998;14:462–7.
3. World Health Organization. Methods and techniques for clinical trials on antimalarial drug efficacy: genotyping to identify parasite populations. 2008.
4. Anderson TJ, Su XZ, Bockarie M, Lagog M, Day KP. Twelve microsatellite markers for characterization of plasmodium falciparum from finger-prick blood samples. Parasitology. 1999;119:113–25.
5. Falk N, Maire N, Sama W, Owusu-Agyei S, Smith T, Beck H-P, et al. Comparison of PCR-RFLP and Genescan-based genotyping for analyzing infection dynamics of plasmodium falciparum. Am J Trop Med Hyg. 2006; 74:944–50.
6. Anderson TJC, Haubold B, Williams JT, Estrada-Franco JG, Richardson L, Mollinedo R, et al. Microsatellite markers reveal a Spectrum of population structures in the malaria parasite plasmodium falciparum. Mol Biol Evol. 2000;17:1467–82.
7. Messerli C, Hofmann NE, Beck H-P, Felger I. Critical Evaluation of Molecular Monitoring in Malaria Drug Efficacy Trials and Pitfalls of Length-Polymorphic Markers. Antimicrob. Agents Chemother. 2017;61:AAC.01500–16.
8. Juliano JJ, Porter K, Mwapasa V, Sem R, Rogers WO, Ariey F, et al. Exposing malaria in-host diversity and estimating population diversity by capture-recapture using massively parallel pyrosequencing. Proc Natl Acad Sci U S A. 2010;107:20138–43.
9. Neafsey DE, Juraska M, Bedford T, Benkeser D, Valim C, Griggs A, et al. Genetic diversity and protective efficacy of the RTS,S/AS01 malaria vaccine. N Engl J Med. 2015;373:2025–37.
10. Lin JT, Hathaway NJ, Saunders DL, Lon C, Balasubramanian S, Kharabora O, et al. Using amplicon deep sequencing to detect genetic signatures of plasmodium vivax relapse. J Infect Dis. 2015;212:999–1008.
11. Bailey J a, Mvalo T, Aragam N, Weiser M, Congdon S, Kamwendo D, et al. Use of massively parallel pyrosequencing to evaluate the diversity of and selection on Plasmodium falciparum csp T-cell epitopes in Lilongwe, Malawi. J Infect Dis. 2012;206:580–7.
12. Friedrich LR, Popovici J, Kim S, Dysoley L, Zimmerman PA, Menard D, et al. Complexity of infection and genetic diversity in Cambodian plasmodium vivax. PLoS Negl Trop Dis. 2016;10:e0004526.
13. Chang H, Worby CJ, Yeka A, Nankabirwa J, Kamya MR, Staedke SG, et al. THE REAL McCOIL: a method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. Pascual M, editor. PLoS Comput Biol. 2017;13:e1005348.
14. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012;30:434–9.
15. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11:31–46. Nature Publishing Group
16. Pallen MJ. Reply to updating benchtop sequencing performance comparison. Nat Biotechnol. 2013;31:296.
17. MalariaGEN Plasmodium falciparum Community Project. Genomic epidemiology of artemisinin resistant malaria. elife. 2016;5:1–29.
18. Esling P, Lejzerowicz F, Pawlowski J. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. Nucleic Acids Res. 2015;43:2513–24.
19. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science. 2003;301:1503–8.
20. Mobegi V a, Duffy CW, Amambua-ngwa A, Loua KM, Laman E, Nwakanma DC, et al. Genome-wide analysis of selection on the malaria parasite plasmodium falciparum in west African populations of differing infection endemicity. Mol Biol Evol. 2014;31:1490–9.
21. Barry AE, Schultz L, Buckee CO, Reeder JC. Contrasting population structures of the genes encoding ten leading vaccine-candidate antigens of the human malaria parasite, plasmodium falciparum. PLoS One. 2009;4:e8497.
22. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ. 2014;2:e593.
23. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ. 2015;3:e1420.
24. Mideo N, Bailey JA, Hathaway NJ, Ngasala B, Saunders DL, Lon C, et al. A deep sequencing tool for partitioning clearance rates following antimalarial treatment in polyclonal infections. Evol Med public Heal. 2016;2016:21–36.
25. Shinozuka H, Forster JW. Use of the melting curve assay as a means for high-throughput quantification of Illumina sequencing libraries. PeerJ. 2016;4:e2281.
26. Harris JK, Sahl JW, Castoe TA, Wagner BD, Pollock DD, Spear JR. Comparison of normalization methods for construction of large, multiplex amplicon pools for next-generation sequencing. Appl Environ Microbiol. 2010;76:3863–8.
27. Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, Moch H, et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nat Commun. 2012;3:811. Nature Publishing Group
28. Rao PN, Uplekar S, Kayal S, Mallick PK, Bandyopadhyay N, Kale S, et al. A method for amplicon deep sequencing of drug resistance genes in plasmodium falciparum clinical isolates from India. J Clin Microbiol. 2016;54: JCM.00235-16.
29. Levitt B, Obala A, Langdon S, Corcoran D, O'Meara WP, Taylor SM. Overlap extension Barcoding for the next generation sequencing and genotyping of plasmodium falciparum in individual patients in western Kenya. Sci Rep Nature Publishing Group. 2017;7:41108.
30. Aubrey BJ, Kelly GL, Kueh AJ, Brennan MS, O'Connor L, Milla L, et al. An inducible Lentiviral guide RNA platform enables the identification of tumor-essential genes and tumor-promoting mutations InVivo. Cell Rep. 2015;10: 1422–32. The Authors
31. Beerenwinkel N, Günthard HF, Roth V, Metzner KJ. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbiol. 2012;3:329.
32. Betuela I, Rosanas-Urgell A, Kiniboro B, Stanisic DI, Samol L, de Lazzari E, et al. Relapses contribute significantly to the risk of plasmodium vivax infection and disease in Papua new Guinean children 1–5 years of age. J Infect Dis. 2012;206:1771–80.
33. Kibbe WA. OligoCalc: An online oligonucleotide properties calculator. Nucleic Acids Res. 2007;35:43–6.
34. Quail M a, Otto TD, Gu Y, Harris SR, Skelly TF, McQuillan JA, et al. Optimal enzymes for amplifying sequencing libraries. Nat Methods. Nature Publishing Group. 2012;9:10–1.
35. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9:357–9.
36. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ Prepr. 2016;4:e2409v1.

Additional file 1

Protocol: Sequencing library preparation

Primary PCR

Multiplexed primary PCR was performed in a total volume of 15µl including 2µl template DNA (1:2 diluted), 250nM of each primary primer (GeneWorks Pty Ltd, Australia) and 7.5µl 2xKAPA HiFi HotStart Ready Mix. Cycling conditions were as follows: initial denaturation 95°C for 3 minutes followed by 25 cycles of 20 seconds denaturation at 98°C, 15 seconds annealing at 52°C and 45 seconds elongation at 72°C plus a final elongation of 2 minutes at 72°C.

Nested PCR

Marker-specific nested PCRs were performed in a total volume of 15µl including 1µl primary PCR product diluted 1:10 in dH_2O, 250nM of the respective nested primer pair (GeneWorks) and 7.5µl 2x KAPA HiFi HotStart Ready Mix (KAPA Biosystems). Cycling conditions were as follows for replicate 1 or 2: initial denaturation 95°C for 3 minutes followed by 15 or 10 cycles of 20 seconds denaturation at 98°C, 15 seconds annealing at 55°C for marker *cpmp* or 56°C for marker *csp* and 45 seconds elongation at 72°C. After 15 or 10 cycles the annealing temperature was increased to 62°C for further 10 or 5 cycles, respectively. Eventually a final elongation of 2 minutes at 72°C was performed. In total, 25 cycles were performed for replicate 1 and 15 cycles for replicate 2.

Pooling of amplicons per sample

Nested PCR products were run on a 1.5% agarose gel for visual inspection of fragment size and quantity. DNA concentration of nested products was estimated in relation to size standard fragments (Solis BioDyne 100bp DNA Ladder). *Cpmp* and *csp* nested PCR products of each sample were pooled in equimolar concentrations. Visual estimation of DNA concentration was difficult as amplicons of marker *csp* and *cpmp* differed in length. To prevent predominance of csp amplicons in the sequencing library due its shorter length, *csp* amplicons were undervalued. This lead to a lower median read coverage for marker *csp* compared to *cpmp*. In case the amplification product was not visible in the agarose gel, equal volumes of both the nested cpmp and csp PCR products nevertheless were pooled.

Sequencing library preparation PCR

PCRs for constructing the sequencing library were carried out in a total volume of 15µl and included 1µl pooled nested products diluted 1:20, 250nM of each sequencing adapter primer and 7.5µl 2xKAPA HiFi HotStart Ready Mix. Cycling conditions were as follows: initial denaturation 95°C for 3 minutes followed by 10 cycles of 20 seconds denaturation at 98°C, 15 seconds annealing at 65°C and 45 seconds elongation at 72°C plus a final elongation of 2 minutes at 72°C.

Pooling of samples

DNA concentrations after these sequencing library PCRs were estimated on a 1.5% agarose gel. All sequencing library PCR products were pooled in equimolar concentrations. This was achieved by pooling equal volumes of all products showing similar band intensity complemented by a pool for PCRs without visible products on agarose gel. These 5 pools were purified with 0.6 volumes of NucleoMag beads (size selection > 300bp) and quantified by Qubit Fluorometer (Thermo Fisher Scientific). Eventually all 5 pools were combined to a final sequencing library by adjusting the volume used from each pool according to its DNA concentration and number of samples combined in a pool.

Sequence library cleanup and sequencing

The expected fragment sizes of the sequencing library were confirmed by Agilent 2200 Tapestation System. The DNA concentration of the final sequencing library pool was quantified by Qubit Fluorometer (Thermo Fisher Scientific). Sequencing was performed on an Illumina MiSeq platform in paired-end mode using MiSeq reagent kit v3 (500-cycles) together with a Enterobacteria phage PhiX Control v3 (Illumina).

Table S1: PCR Primer sequence for *msp2* CE genotyping and sequence library preparation.

Primer for primary PCR	
cpmp_prim_F	CGATACAGGACATATAGA
cpmp_prim_R	TTCAATAACATTTACTAGG
csp_prim_F	ATCAAGGTAATGGACAAG
csp_prim_R	ACTCAAACTAAGATGTGTTC
Primer for nested PCR	
csp_F_Linker	GTGACCTATGAACTCAGGAGTCAAATGACCCAAACCGAAATGT
csp_R_Linker	CTGAGACTTGCACATCGCAGCGGAACAAGAAGGATAATACCA
cpmp_F_Linker	GTGACCTATGAACTCAGGAGTCCATAAGTCATTAAAATTTATGGAT
cpmp_R_Linker	CTGAGACTTGCACATCGCAGCCGTTACTATCAAGATCGTTAATATC
Primer for msp2 CE genotyping	
msp2_S2_fw	GAAGGTAATTAAAACATTGTC
msp2_S3_rev	GAGGGATGTTGCTGCTCCACAG
msp2_S1-fw	GCTTATAATATGAGTATAAGGAGAA
msp2_FC27-rev	GCATTGCCAGAACTTGAA
msp2_3D7-rev	CTGAAGAGGTACTGGTAGA
Primer for sequence library PCR (XXXXXX=barcode)	
Forward	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXXXXGTGACCTATGAACTCAGGAGTC
Reverse	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTXXXXXXXXCTGAGACTTGCACATCGCAGC

Forward barcode		**Reverse barcode**	
Fwd_1	TAGATCGC	Rev_1	TAAGGCGA
Fwd_2	CTCTCTAT	Rev_2	CGTACTAG
Fwd_3	TATCCTCT	Rev_3	AGGCAGAA
Fwd_4	AGAGTAGA	Rev_4	TCCTGAGC
Fwd_5	GTAAGGAG	Rev_5	GGACTCCT
Fwd_6	ACTGCATA	Rev_6	TAGGCATG
Fwd_7	AAGGAGTA	Rev_7	CTCTCTAC
Fwd_8	CTAAGCCT	Rev_8	CAGAGAGG
Fwd_13	TGGTGGTA	Rev_9	GCTACGCT
Fwd_14	TTCACGCA	Rev_10	CGAGGCTG
Fwd_15	AGCACCTC	Rev_11	AAGAGGCA
Fwd_16	CAAGGAGC	Rev_12	GTAGAGGA
Fwd_17	ATTGGCTC	Rev_13	ATGCCTAA
Fwd_18	CACCTTAC	Rev_14	ACGCTCGA
Fwd_19	CTAAGGTC	Rev_15	AGTCACTA
Fwd_20	GAACAGGC	Rev_16	ATCCTGTA
		Rev_17	CGCATACA
		Rev_18	CTGGCATA
		Rev_19	GATAGACA
		Rev_20	GCTAACGA
		Rev_21	GTGTTCTA
		Rev_22	TCCGTCTA
		Rev_23	CCTAATCC
		Rev_24	GACAGTGC

Table S2: Summary of mismatch rates for linker sequences, marker primers used in primary and nested amplification and for amplicons[1] of markers *cpmp* and *csp* generated from controlled mixtures of two *P. falciparum* strains 3D7 and HB3

	Linkers	Primers		Amplicons	
	%	cpmp %	csp %	cpmp %	csp %
MIN	0.00	0.00	0.00	0.00	0.00
1ST QU.	0.00	0.00	0.03	0.06	0.07
MEDIAN	0.08	0.09	0.21	0.15	0.18
MEAN	0.12	0.28	0.71	0.38	0.46
3RD QU.	0.19	0.20	0.42	0.35	0.43
MAX	1.93	10.92	22.01	15.76	18.13

[1] Mismatch rate was calculated relative to 3D7 and HB3 reference sequence.

Table S3: Percent of sequence reads clustering to 3D7 reference sequence, percent of PCR artefacts and percent of singleton reads in serial dilution of culture strain 3D7.

Parasitaemia per µl	*cpmp*				*csp*			
	3D7[1] %	PCR artefact %	Singletons %	Coverage	3D7[1] %	PCR artefact %	Singletons %	Coverage
50,000	93.3	0.3	6.4	6,758	86.4	5.2	8.4	1,623
5,000	92.2	0.2	7.6	2,382	85.3	4.9	9.8	1,374
500	91.9	0.2	7.9	3,751	85.4	5.0	9.6	3,725
50	93.3	0.0	6.7	165	83.0	6.1	10.9	540
5	66.7	0.0	33.3	6	87.1	0.0	12.9	70

[1] Percent of reads that cluster with 3D7 reference sequence.

Table S4: Detectability of the minority clone in defined ratios of *P. falciparum* strains HB3 and 3D7 for marker *cpmp*. Read counts clustering correctly with 3D7 and HB3 haplotypes, as well as read failed to cluster with 3D7 and HB3 haplotypes: false haplotype CPMP-15 (below cut-off criteria), singleton and obvious PCR artefacts (indels and chimeras).

Ratios in mixtures	3D7		HB3		Singleton		Indels		Chimera		CPMP-15		Coverage
HB3:3D7	n	%	n	%	n	%	n	%	n	%	n	%	n
1:1	14,103	34.6	23,399	57.4	3,067	7.52	107	0.26	92	0.23	0	0.00	40,768
1:10	9,854	75.6	2,141	16.4	987	7.57	39	0.30	15	0.12	1	0.01	13,037
1:50	4,400	88.8	156	3.15	394	7.95	3	0.06	0	0.00	0	0.00	4,953
1:100	12,093	90.9	204	1.53	966	7.26	48	0.36	0	0.00	0	0.00	13,311
1:500	5,130	90.8	27	0.48	476	8.43	16	0.28	0	0.00	0	0.00	5,649
1:1000	2,780	91.5	7	0.23	251	8.26	1	0.03	0	0.00	0	0.00	3,039
1:1500	51,680	92.5	60	0.11	3,876	6.94	268	0.48	3	0.01	0	0.00	55,887
1:3000	6,863	92.5	7	0.09	518	6.98	29	0.39	0	0.00	0	0.00	7,417

Table S5: Detectability of the minority clone in defined ratios of *P. falciparum* strains HB3 and 3D7 for marker *csp*. Read counts clustering correctly with 3D7 and HB3 haplotypes, as well as read failed to cluster with 3D7 and HB3 haplotypes: false haplotype CSP-9 (below cut-off criteria), singleton and obvious PCR artefacts (indels and chimeras).

Ratios in mixtures	3D7		HB3		Singleton		Indels		Chimera		CSP-9		Coverage
HB3:3D7	n	%	n	%	n	%	n	%	n	%	n	%	n
1:1	3,126	34.7	4,550	50.5	812	9.01	497	5.52	24	0.26	0	0.00	9,009
1:10	2,544	76.1	339	10.2	270	8.08	186	5.57	2	0.06	0	0.00	3,341
1:50	12,170	82.7	424	2.88	1,201	8.16	890	6.05	26	0.18	0	0.00	14,711
1:100	9,996	83.5	269	2.25	1,063	8.88	644	5.38	3	0.03	0	0.00	11,975
1:500	2,948	84.0	16	0.46	377	10.75	166	4.73	1	0.03	0	0.00	3,508
1:1000	1,548	85.7	4	0.22	163	9.02	92	5.09	0	0.00	0	0.00	1,807
1:1500	20,381	86.3	19	0.08	1,866	7.90	1,349	5.71	1	0.00	3	0.01	23,619
1:3000	1,970	85.0	1	0.04	211	9.10	136	5.87	0	0.00	0	0.00	2,318

Table S6: Multiplicity of infection of 37 field sample measured by length polymorphic marker *msp2* and SNP polymorphic markers *cpmp* and *csp*.

MOI	***msp2*** **n**	***cpmp*** ***n***	***csp*** ***n***
1	12	10	19
2	14	14	16
3	5	6	2
4	4	2	
5	2	5	
Mean	2.2	2.5	1.5

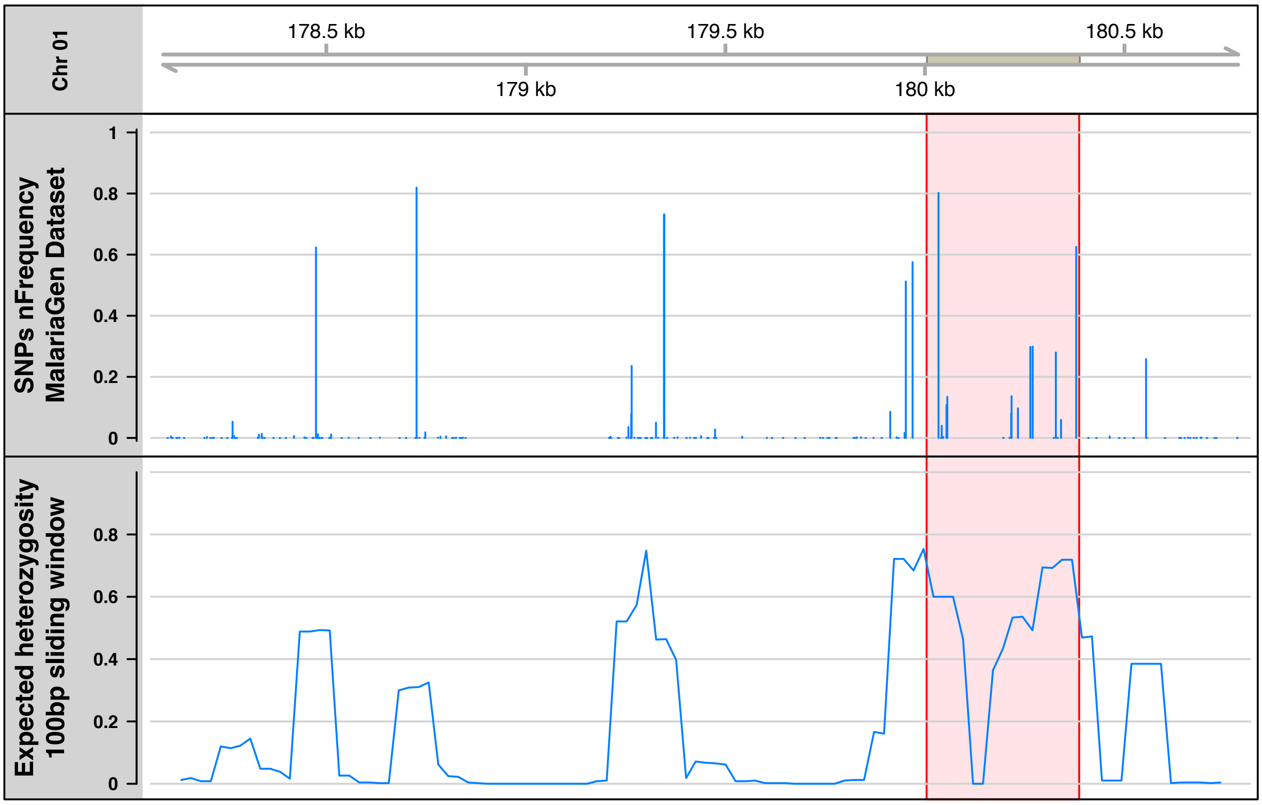


Figure S1: Genomic distribution of single nucleotide polymorphism in sequenced alleles of *P. falciparum* gene PF3D7_0104100 (*cpmp* marker). The top panel represents alleles of global origin (MalariaGEN *P. falciparum* Community Project, 2016). The bottom panel shows expected heterozygosity values for sliding windows of 100bp across the entire gene. Red box highlights region selected for amplification.

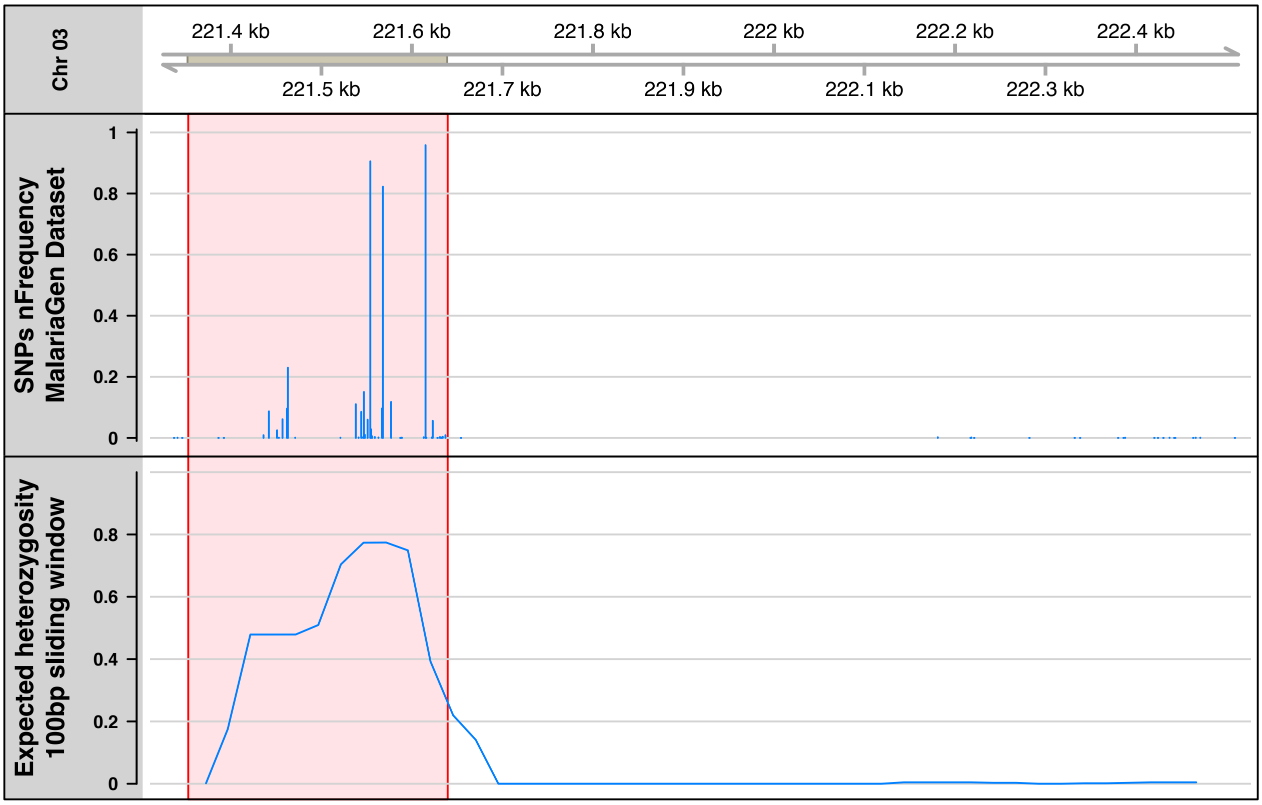


Figure S2: Genomic distribution of single nucleotide polymorphism in sequenced alleles of the *P. falciparum* circumsporozoite protein *(csp)*. The top panel represents alleles of global origin (MalariaGEN *P. falciparum* Community Project, 2016). The bottom panel shows expected heterozygosity values for sliding windows of 100bp across the entire gene. Red box highlights region selected for amplification.

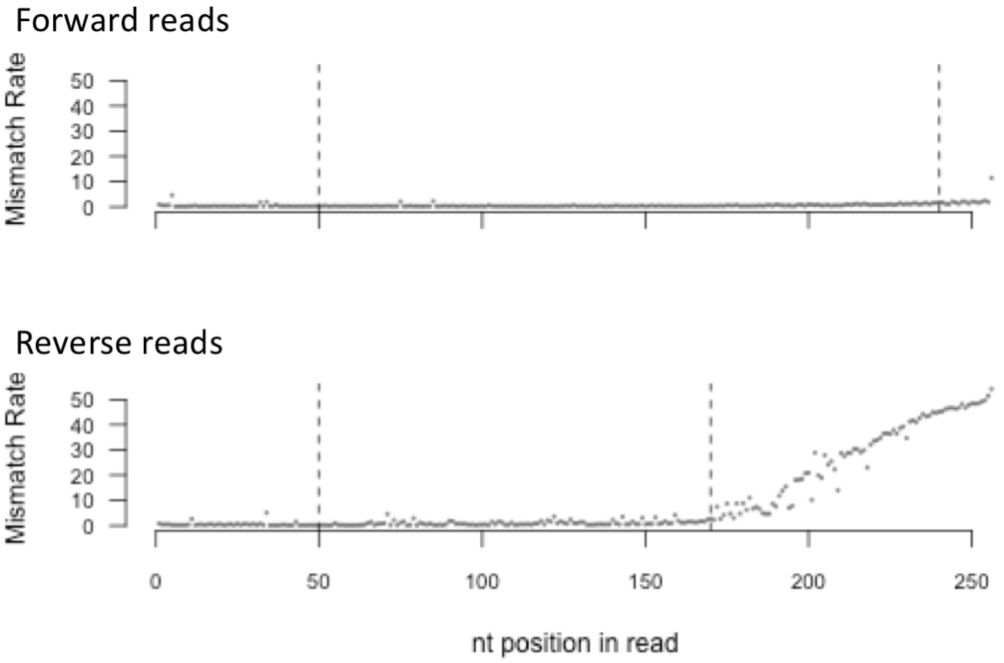


Figure S3: **Average mismatch rate per nucleotide position in reads of spiked in control DNA *phiX*.** X-axis: nucleotide position in *phiX* read. Y-axis: mismatch rate with respect to *phiX* reference sequence. Each data point represents the average mismatch rate of all *phiX* reads at a given nucleotide position.

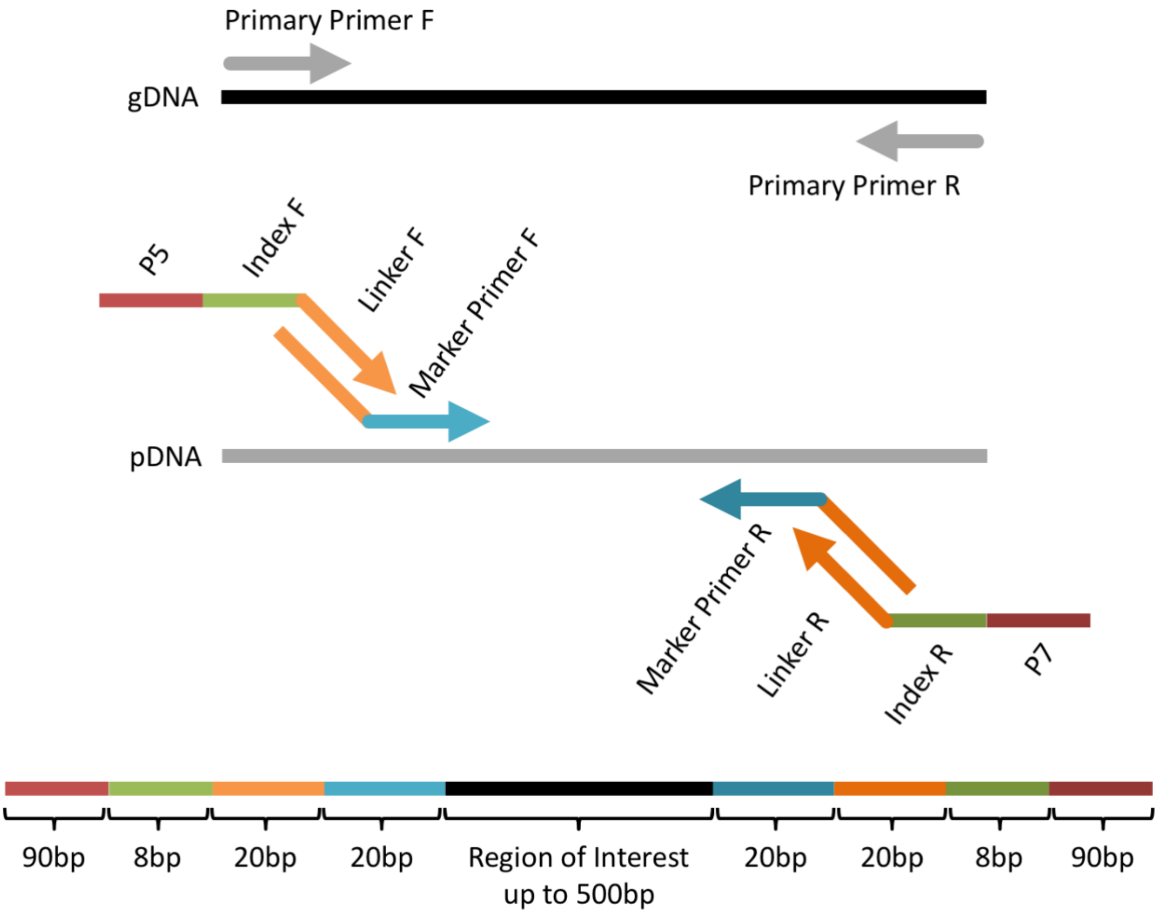


Figure S4: Design of the amplicon sequencing library. Primary primers target the gene of interest. Primary PCR is followed by nested PCR using marker-specific primers that carry F and R linker sequences at their 5' ends. The primers for the final round of amplification target the F and R linker sequences. These primers carry sample-specific indices (barcodes) plus Illumina sequencing adapter P5 and P7 at their 5' ends. The line at the bottom indicates the sizes of the various elements.

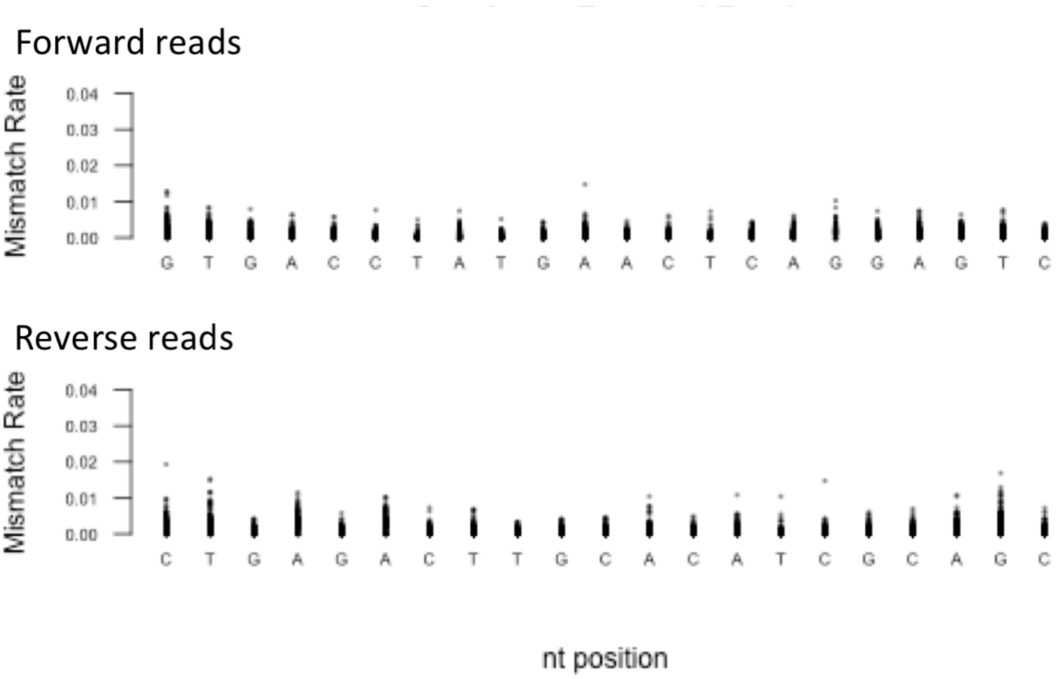


Figure S5: Observed mismatch rate at each nucleotide position in forward and reverse reads of linker sequence. Data derived from all samples analysed. Each data point represents the mean mismatch rate of all reads from an individual sample. X-axis: nucleotides of forward and reverse linker (5' to 3'). Y-axis: mismatch rate with respect to known linker sequence.

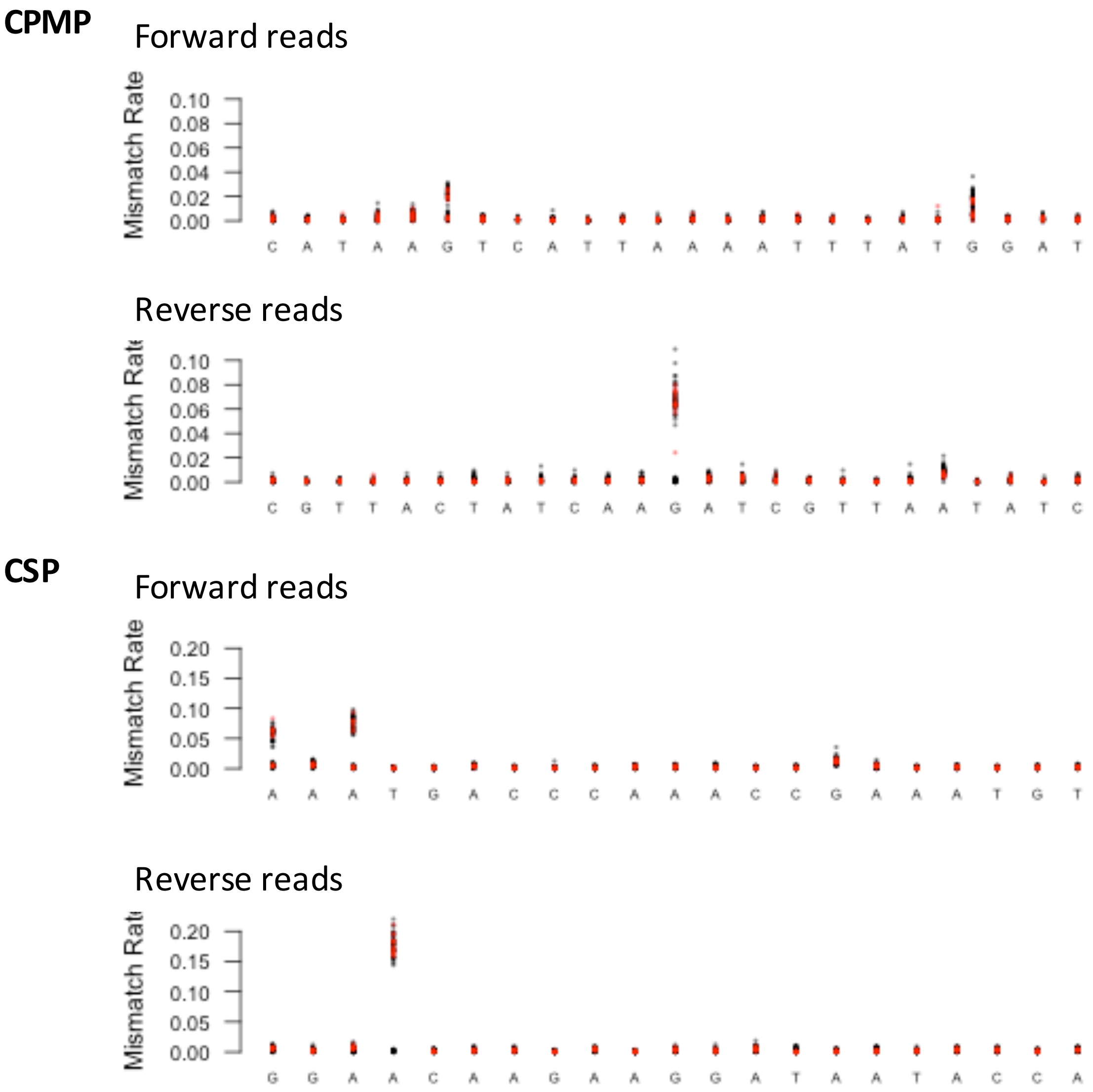


Figure S6: Mismatch rate per nucleotide position in forward and reverse primers of markers cpmp and csp. Data derived from all samples analysed. Each data point represents the mean observed mismatch rate of all reads from an individual sample. Red data points: control samples (P. falciparum culture strains); black data points: field samples; X-axis: nucleotides of forward and reverse primers (5' to 3'); Y-axis: mismatch rate with respect to the known primer sequences.

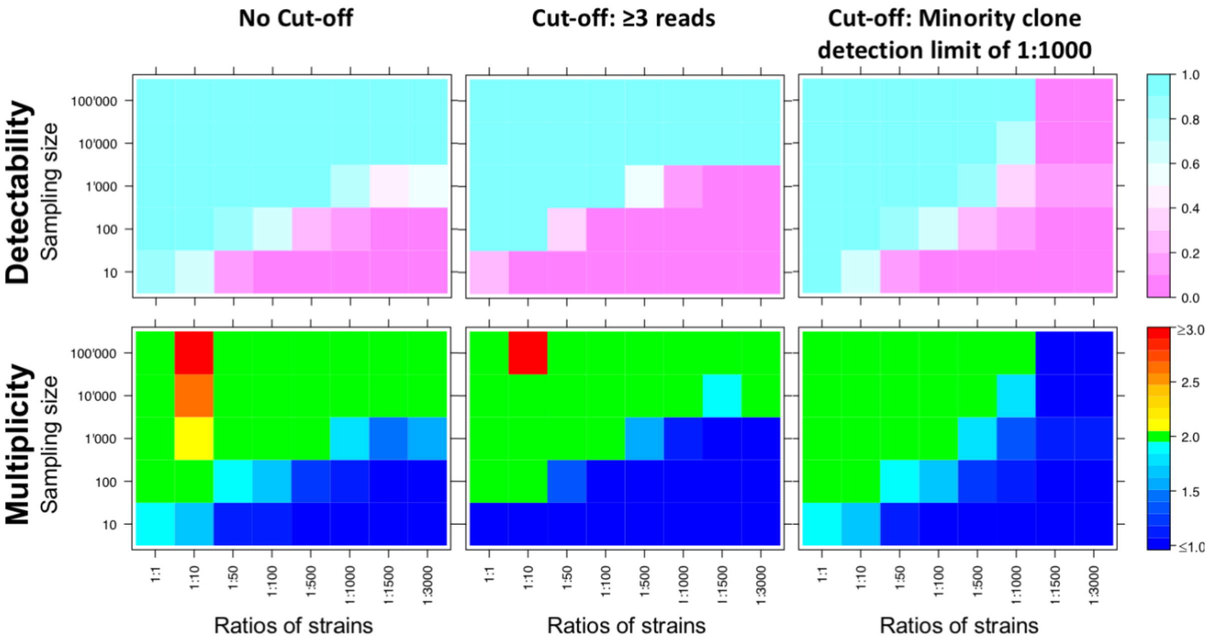


Figure S7: Simulation of the detectability of a minority clone (top panel) and of measured multiplicity of infection (bottom panel) by bootstrapping for marker *cpmp*. Cut-off settings: no cut-off (left panel); ≥3 read per haplotype (middle panel); minority clone detection limit of 1:1000 (right panel). Samples were drawn from reads of defined mixtures of *P. falciparum* strains 3D7 and HB3. X-axis represents ratios of strains 3D7 and HB3. Y-axis indicates the sampling size (number of draws from the sequence reads (coverage >3000) for each mixture of strains. Sampling was repeated 1000 times to estimate the mean detectability of a minority clone.

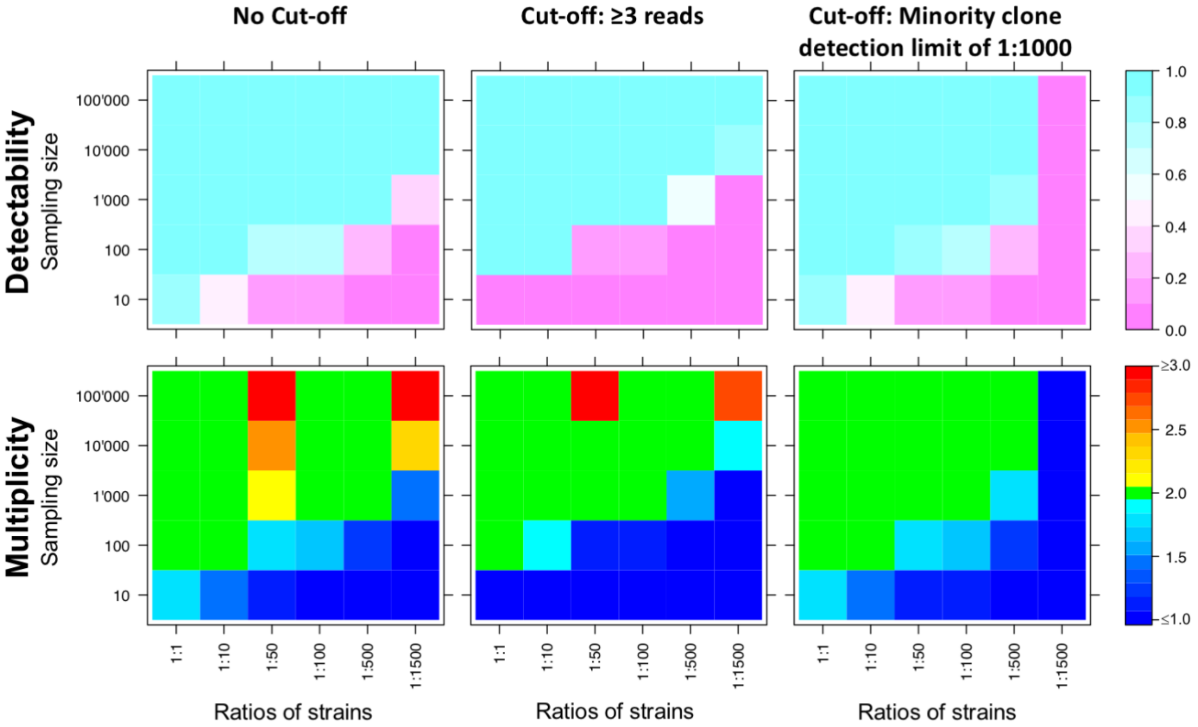
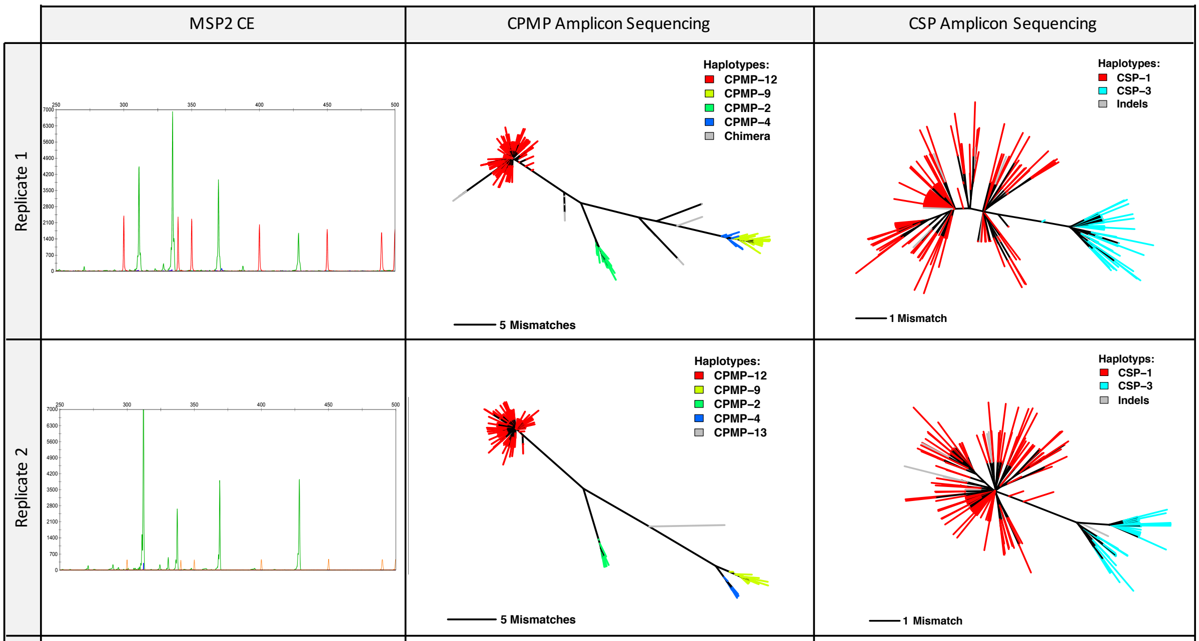


Figure S8: Simulation of the detectability of a minority clone (top panel) and of measured multiplicity of infection (bottom panel) by bootstrapping for marker *csp.* Cut-off settings: no cut-off (left panel); ≥3 read per haplotype (middle panel); 0.1% minority clone detection limit of 1:1000 (right panel). Samples were drawn from reads of defined mixtures of *P. falciparum* strains 3D7 and HB3. X-axis represents dilution ratios of strains 3D7 and HB3. Y-axis indicates the sampling size (number of draws from the sequence reads (coverage >3000) for each mixture of strains. Sampling was repeated 1000 times to estimate the mean detectability of a minority clone.



Quantification

Haplotype	Replicate 1	Replicate 2
D311	x	x
D334	x	x
D369	x	x
D428	x	x

Haplotype	Replicate 1		Replicate 2	
	n	%	n	%
CPMP_12	1513	68.4	754	70.1
CPMP_9	187	8.5	107	9.9
CPMP_2	180	8.1	88	8.2
CPMP_4	48	2.2	31	2.9
CPMP_13	-	-	3	0.3
Singleton	235	10.6	93	8.6
Chimera	42	1.8	-	-
Indels	6	0.3	-	-

Haplotype	Replicate 1		Replicate 2	
	n	%	n	%
CSP_1	563	68.7	630	65.7
CSP_3	136	16.6	156	16.3
Singleton	100	12.1	130	13.6
Chimera	-	-	-	-
Indels	21	2.7	43	4.5

	MSP2 CE	CPMP Amplicon Sequencing	CSP Amplicon Sequencing
MOI	MOI=4	MOI=4	MOI=2

Figure S9: Comparison of genotyping by length-polymorphic marker *msp2* and amplicon sequencing of markers *cpmp* and *csp* exemplified in 1 field sample. Capillary electropherograms (CE) and dendrograms represent the raw data of markers *msp2*-CE, *cpmp* and *csp* (two top panels). Quantification of haplotypes and final multiplicity call (two bottom panels). Grey shading indicates haplotypes and reads filtered out by cut-off settings (example discussed in detail in results section, paragraph “Validation of amplicon sequencing in field samples”).

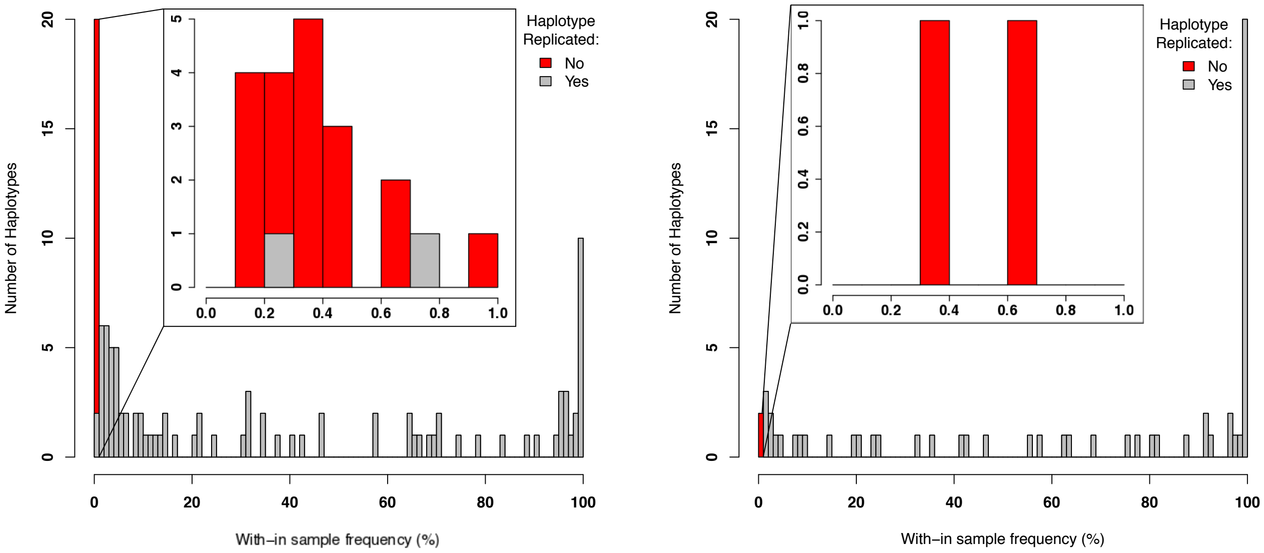


Figure S10: Reproducibility of amplicon sequencing in field samples. Haplotype calls that passed default cut-off criteria were compared between replicates to investigate reproducibility. In grey: number of haplotypes detected in both replicates, in red: number of haplotypes detected only in a single replicate. Inserts present frequency distributions below 1% at a higher resolution.

CHAPTER 3: AMP-SEQ GENOTYPING: LONGITUDINAL TRACKING OF COMPLEX INFECTIONS

Longitudinal tracking of Plasmodium falciparum clones in complex infections by amplicon deep sequencing

Anita Lerch[a,b,c], Cristian Koepfli[c,d,*], Natalie E. Hofmann[a,b], Johanna H. Kattenberg[e,*], Anna Rosanas-Urgell[e,*], Inoni Betuela[e], Ivo Mueller[c,d*], Ingrid Felger[a,b,#]

[a] Swiss Tropical and Public Health Institute, Basel, Switzerland
[b] University of Basel, Basel, Switzerland
[c] Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia
[d] University of Melbourne, Parkville, Australia
[e] Papua New Guinea Institute of Medical Research, Madang, Papua New Guinea
*current affiliations: CK: University of California Irvine, Irvine, USA. JHK, ARU: Institute of Tropical Medicine, Antwerp, Belgium. IM: Institut Pasteur, Paris, France.
[#] Corresponding author

– Manuscript is submitted to Journal of Open Forum Infectious Diseases –
– Published in bioRxiv April 30, 2018, DOI: https://doi.org/10.1101/306860 –

Abstract

Background
Longitudinal tracking of individual *Plasmodium falciparum* strains in multi-clonal infections is essential for investigating infection dynamics of malaria. The traditional genotyping techniques did not permit tracking changes in individual clone density during persistent natural infections. Amplicon deep sequencing (Amp-Seq) offers a tool to address this knowledge gap.

Methods
The sensitivity of Amp-Seq for relative quantification of clones was investigated using three molecular markers, *ama1*-D2*, ama1*-D3*,* and *cpmp*. Amp-Seq and length-polymorphism based genotyping were compared for their performance in following minority clones in longitudinal samples from Papua New Guinea.

Results
Amp-Seq markers were superior to length-polymorphic marker *msp2* in detecting minority clones (sensitivity Amp-Seq: 95%, *msp2:* 85%). Multiplicity of infection (MOI) by Amp-Seq was 2.32 versus 1.73 for *msp2*. The higher sensitivity had no effect on estimates of force of infection because missed minority clones were detected in preceding or succeeding bleeds. Individual clone densities were tracked longitudinally by Amp-Seq despite MOI>1, thus providing an additional parameter for investigating malaria infection dynamics.

Conclusion
Amp-Seq based genotyping of longitudinal samples improves detection of minority clones and estimates of MOI. Amp-Seq permits tracking of clone density over time to study clone competition or the dynamics of specific, i.e. resistance-associated genotypes.

INTRODUCTION

Molecular-epidemiological parameters used to describe the infection dynamics of *Plasmodium falciparum* include the number of co-infecting parasite clones (multiplicity of infection, MOI), the rate at which different genotypes are acquired over time (molecular force of infection, $_{mol}$FOI) and duration of infection [1]. These measures are based on monitoring the presence or absence of clones in cross-sectional or longitudinal samples collected in regular intervals. In earlier studies individual parasite clones in multi-clonal field samples were distinguished and tracked over time by genotyping the length-polymorphic marker merozoite surface protein 2 (*msp2*) by capillary electrophoresis-based fragment sizing (CE) [2–4]. Yet, *msp2*-CE genotyping has limited sensitivity for minority clone detection [3,5]. Alternative typing methods instead could perform better in detecting minority clones, but might impact measures of MOI and $_{mol}$FOI [6]. So far quantification of individual clones within multi-clonal infections was not feasible, as this would have required highly complex allele-specific quantitative PCR (qPCR).

SNP-based genotyping by deep amplicon sequencing (Amp-Seq) can detect low-abundant *P. falciparum* clones at ratios of 1:1000 in mixed infections [7]. Most importantly, genotyping by Amp-Seq also quantifies precisely the relative abundance of clones, as shown with artificial mixtures of clones [7–9]. From these ratios the absolute density of each clone (i.e. a certain haplotype) within a multi-clone infection can be deduced if the total parasitaemia of the sample was established by qPCR [9]. When analysing consecutive samples from a given study participant, presence and fluctuations in density of clones can be tracked. We explore how longitudinal information can be used to improve identification of minority clones with low densities around the detection limit.

A previous study has estimated clonal density with Amp-Seq in multi-clone infections to estimate clearance rates after antimalarial treatment [9]. We apply the same approach to track parasite clones longitudinally in untreated natural infections. In addition, we increase the resolution of genotyping by combining sequence information from several markers into multi-locus haplotypes.

METHODS

Study design

A subset of 153 archived *P. falciparum* genomic DNA samples from 33 children (mean 4.3 samples [min: 2, max: 11]) aged 1-5 years were available from an cohort study with blood sampling over 40 weeks (first 12 weeks every fortnightly, then monthly) in Papua New Guinea (PNG) [10]. The two conditions for selection of children were: ≥2/14 bleeds PCR positive, and MOI>1 in at least one of the samples of each child. Ethical clearance was obtained from PNG Institute of Medical Research Institutional Review Board (IRB 07.20) and PNG Medical Advisory Committee (07.34). Informed written consent was obtained from all parents or guardians prior to recruitment of each child.

Genotyping using length polymorphic marker *msp2*

Samples were genotyped using the classical *P. falciparum* marker *msp2* according to published protocols [11]. Fluorescently labelled nested PCR products were sized by CE on an automated sequencer and analysed using GeneMarker software. Fragments were accepted if the following cut-off criteria were met: peak height >500 intensity units and >10% of the height of the majority peak. Electropherograms were inspected visually to exclude obvious stutter peaks. All DNA samples were genotyped in 2 independent laboratories to assess reproducibility of clone detection and measures of MOI.

Marker selection for Amplicon deep sequencing

Amp-Seq was performed on three amplicons located in two different *P. falciparum* marker genes, namely PF3D7_0104100, "conserved *Plasmodium* membrane protein" (*cpmp*), and PF3D7_1133400, "apical membrane antigen 1" (*ama1*) whose genetic diversity has been studied in great detail [12–14]. Previously published primers were used for marker *cpmp* [7]. For *ama1* two amplicons of 479 and 516 bp were selected that span regions of maximum diversity, i.e. subdomains 2 and 3 of the ectodomain [15]. Primer sequences and exact amplicon positions are listed in Tables S1 and S2.

Sequencing library preparation

Sequencing libraries were generated by three rounds of PCR, according to previously published protocols [7]. After primary PCR, a 5' linker sequence was added during nested PCR. Nested PCR products were subject to another PCR round with primers binding to the linker sequences and carrying Illumina sequence adapters plus an eight nucleotide long sample-specific molecular index to permit pooling of amplicons for sequencing and later de-multiplexing. The final sequence library was purified with NucleoMag beads prior to sequencing on an Illumina MiSeq platform in paired-end mode using Illumina MiSeq reagent kit v2 (500-cycles) together with Enterobacteria phage PhiX control (Illumina, PhiXControl v3).

Sequence read analysis and haplotype calling

Samples yielding a sequence coverage of <25 reads were excluded from the analysis. An overview of sequence read coverage for all Amp-Seq markers is given in Table S3. Sequence reads were analysed using software HaplotypR [7], (https://github.com/lerch-a/HaplotypR.git). To remove low quality sequences, reads were trimmed to 240bp for forward and 170bp for reverse reads. As reference sequence *P. falciparum* strain 3D7 was used (PlasmoDB release 34, [16]). The term genotype refers to a single nucleotide polymorphism (SNP). Calling a SNP required a >50% mismatch rate in the sequence reads of this nucleotide position in at least two independent samples. A haplotype was defined as sequence variant of an entire amplicon. Haplotypes containing inserts or deletions (indels) were filtered out, as well as haplotypes resulting from chimeric reads or singleton reads. The number of reads of a given haplotype over all remaining reads of the same marker within a sample is denoted by the term "within-host haplotype frequency". Cut-off criteria for haplotype calling were as follows: a minimum of 3 reads coverage per sample, a within-host haplotype frequency ≥0.1% and an occurrence of this haplotype in at least 2 samples.

Multi-locus haplotype inference in longitudinal samples

Amp-Seq quantifies the frequency of each haplotype within a sample, which permits to infer multi-locus haplotypes. A multi-locus haplotype was deduced in multiple rounds. In the first round, the multi-locus haplotype of the dominant clone of a sample was inferred by selecting each marker's dominant haplotype (>54% within-host haplotype frequency, i.e. 50%+3.8% standard deviation in within-host haplotype frequency between replicates). After each round the identified dominant haplotype was ignored and in the following round the dominant haplotype was identified among the remaining reads. If several haplotypes occurred in a sample at similar frequencies, it may be impossible to identify the dominant haplotype. This was resolved by analysing the change in within-host haplotype frequency between the observed and preceding or succeeding sample of the same host. An example of our approach to multi-locus haplotype inference is shown in detail the Supplemental Text S1.

The final step of multi-locus haplotype inference addressed the problem of clones of a multiple infection that share by chance the same allele of one of the markers. As a consequence, the within-host frequency of a shared haplotype amounts to the sum of two or more independent clones carrying the same allele. In such cases multi-locus haplotypes were inferred by assigning the shared alleles to those haplotypes that summed up to the same proportion in the other two markers. Samples for which the multi-locus haplotype could not be established by this approach were considered unresolvable (Table S4).

Reproducibility, sensitivity and false discovery rate

Samples were analysed in duplicates with Amp-Seq markers and *msp2*-CE. Performing duplicates permitted to identify and exclude false-positive haplotypes and thus prevented erroneous over-estimation of MOI. Each haplotype was classified into one of four groups (example see FIG S1): (1) True-positive (TP) haplotype, i.e. it passed the haplotype calling cut-off in both replicates or in one replicate plus in the preceding or succeeding bleed; (2) False-positive (FP) haplotype, i.e. it passed the haplotype calling cut-off in only one replicate and was not detected in any of the preceding or succeeding samples of that individual; (3) False-negative (FN_i) haplotype, i.e. it was detected in one or both replicates but did not pass the cut-off criteria at that occasion, whereas it was detected in the preceding or succeeding bleed as TP (at least once) or FN haplotype; (4) Background noise (all other cases).

Additionally, false-negative (FN_{ii}) haplotypes were imputed for samples in which no sequence read was detected. These false-negative haplotypes were imputed only when (a) the haplotype was detected in the preceding as well as the succeeding bleed as a true-positive. Presence in only one of preceding or succeeding sample was not considered sufficient evidence for assuming a case of missed detection. For the Amp-Seq markers but not msp2-CE, false-negative haplotypes were also imputed when (b) data for the other two markers was present and the corresponding multi-locus haplotype was established in the preceding or succeeding sample.

The sensitivity to detect parasite clones was estimated based on selected individuals who had not received antimalarial treatment during the timespan analysed and harboured at least one haplotype that was detected at 3 consecutive bleeds. Sensitivity was defined as the true positive rate of a genotyping method and was calculated as TP/(TP+FN). The risk to falsely assign a haplotype not present in the sample was measured as the “false discovery rate” (FDR), calculated as FP/(TP+FP). This rate represents the extent of false haplotype calls of a genotyping method.

The reproducibility of clone detection in technical replicates (comprising all experiential procedures from PCR to sequence run) was calculated as $\frac{2n_2}{n_1+2n_2}$, where n_1 is the number of haplotypes detected in a single replicate and n_2 the number of haplotypes detected in both replicates [17]. Only TP haplotypes were used to estimate reproducibility.

Epidemiological parameters: clone density, diversity, MOI and FOI

The density of a parasite clone was calculated by multiplying within-host haplotype frequency by parasitaemia (measured by qPCR). Clone density is expressed as copies of target gene per microliter, quantified by qPCR targeting the 18S rRNA gene of *P. falciparum* [18]. The technical detection limit of qPCR was 0.4 copies/μl whole blood.

Based on true positive haplotypes, the expected heterozygosity (H_e) and mean MOI were determined from baseline (or first bleed available) samples for each marker as described [7]. H_e was also estimated for combined markers in samples that had a resolvable multi-locus haplotype and that were separated by a treatment plus ≥2 consecutive *P. falciparum* negative samples from the same child.

$_{mol}$FOI was estimated on longitudinal sets of sample that had a complete set of replicates. Haplotypes were counted as new infection if a haplotype was (i) not present in the baseline sample but in a subsequent sample, (ii) not detected at ≥2 consecutive preceding bleeds or (iii) not detected after antimalarial treatment plus after at least one negative sample. Time at risk was calculated as the timespan between baseline and last sampling, minus 14 days for each antimalarial treatment (to account for the prophylactic effect of treatment).

An overview of sample selection criteria applied for different types of analyses is listed in Table S5.

RESULTS

Genetic diversity of markers

The discriminatory power of Amp-Seq markers *cpmp*, *ama1*-D2 and *ama1*-D3, as well as length-polymorphic marker *msp2*-CE was estimated in 33 baseline samples. The resolution was highest for amplicon marker *cpmp* (H_e=0.961) that distinguished 30 haplotypes and gave a mean MOI=2.45 (Table 1, MOI distribution by marker in FIG S2). The second-best resolution was obtained by marker *msp2*-CE (H_e=0.940) that distinguished 20 haplotypes and measured a mean MOI=1.73. Haplotype and SNP frequencies of Amp-Seq markers are shown in FIG 1 and S2.

Discriminatory power can be increased by combining multiple markers. Inference of multi-locus haplotypes was possible for 66 clones in 46 selected samples. Combining marker *cpmp* with either of the two *ama1* fragments yielded very high diversity (53 and 55 haplotypes, H_e=0.992 and 0.994 for *cpmp*/*ama1*-D2 and *cpmp*/*ama1*-D3) (Table 2 and FIG S3). Combining all 3 markers did not increase discriminatory power any further.

Using longitudinal genotyping data to increase detectability of clones

Imperfect detectability of parasite clones has been described previously in longitudinal genotyping studies [1,19–21]. Data from replicates and longitudinal samples can be used to make assumptions on missed clones. This permits imputing of missed haplotypes and thus improves the tracking of clonal infections within an individual over time. Two types of missed haplotypes respective false-negative haplotypes were distinguished: (FN_i) haplotypes that were detected below the cut-off and (FN_{ii}) haplotypes that were not detected but imputed (Table 3). FIG 2 shows an example of different type of missed haplotypes for all Amp-Seq markers.

The sensitivity to detect parasite clones was estimated for each genotyping marker by enumerating false-negative haplotypes. Sensitivity was higher for the Amp-Seq markers than for *msp2*-CE (in decreasing order

96.5%, 95.0%, 93.9% and 85.1% for *ama1*-D2, *cpmp*, *ama1*-D3 and *msp2*-CE) (Table 4). For ≥57% of the identified false-negative haplotypes, reads were detected but fell below cut-off criteria (category (i) above). If such haplotypes were counted as positives by relaxing the cut-off criteria, sensitivity would increase to 99.1%, 97.5% and 97.4% for Amp-Seq markers *ama1*-D2, *cpmp* and *ama1*-D3 (Table 4).

The false discovery rate of haplotypes for Amp-Seq markers was in the range of 0.9-4.2% (Table 4). Reproducibility to detect parasite clones in technical replicates was greater for Amp-Seq markers than for marker *msp2*-CE (in decreasing order 0.95, 0.95, 0.94 and 0.91 for *ama1*-D3, *cpmp*, *ama1*-D2 and *msp2*-CE) (Table S6 and FIG S4).

Determination of $_{\text{mol}}$FOI by different molecular markers and methods

A higher sensitivity of the genotyping method does not necessary impact molFOI, i.e. new clones/year, because a missed minority clone could be detected at one of the successive bleeds. We investigated the number of new infections acquired during 40 weeks follow-up in 27 children from whom a complete data set was available (on average 4.3 samples per child [min: 2, max: 7]). Mean molFOI was 2.7, 2.7, 2.3 and 2.2 new infections per year for markers ama1-D3, cpmp, msp2-CE and ama1-D2 (negative binomial regression p-value for comparison of *msp2*-CE to *ama1*-D3, *cpmp* and *ama1*-D2: 0.596, 0.649 and 0.877) (FIG S5). Thus, no substantial difference in mean $_{\text{mol}}$FOI was found for the different molecular markers and different genotyping methods.

Quantitative dynamics of multiple infecting *P. falciparum* clones

Densities of individual clones was calculated from the total parasitaemia by qPCR and the within-host haplotype frequency. Examples of individual clone density dynamics in children with multi-clone infections are shown for three Amp-Seq markers (FIG 3). The density of some clones remained constant over time, whereas other clones showed fluctuations in density over 3 orders of magnitude (FIG 3A and B). In some children the dominant clone remains dominant over the observation period (FIG 3A), whereas in others switch-over between minority clone and dominant clone was observed (FIG 3B). In highly complex field samples some clones might share the same haplotype of a given marker (FIG 3C). Such clones can only be differentiated and quantified if multiple markers are typed and at least one of the markers is not shared between concurrent clones.

After artemisinin combination therapy, some of the parasite clones from multi-clone infections were cleared 14 days after antimalarial treatment, whereas others were still detectable (FIG 3A, B and C). These persisting clones had decreased clone densities (<21 copies/μl) and likely represent remaining late gametocyte stages of cleared asexual infections [22]. Some new infections following antimalarial treatment (artesunate-primaquine) showed a rapid increase in clone density within the first 14 days after re-infection of a host, followed by a slow decrease in clone density until clearance (FIG 3D), whereas in other infections clone density remained constant (FIG 3C).

Discussion

While MOI and $_{\text{mol}}$FOI have been extensively described as epidemiological parameters, the ratio and density of individual clones within complex infections has not yet been investigated. This gap in knowledge was due to shortfalls of traditional length-polymorphic markers, where the length of a fragment greatly influences the amplification efficiency in multi-clone infections with fragments competing in PCR and a strong bias favouring smaller fragments [5]. As a result, multi-locus haplotypes could not be inferred from traditional genotyping data in a reliable way. Such inference is required, for example, for phylogenetic or population genetic studies. In such studies, multiple-clone infections were usually excluded or only the predominant haplotype included [23,24]. With the possibility to establish multi-locus haplotypes from complex infections the discriminatory power will be greatly improved in future.

Single Amp-Seq markers *cpmp, ama1*-D2, *ama1*-D3, and *msp2*-CE yielded similar resolution. Combining *cpmp* with either of the *ama1* fragments increased further discriminatory power. The excellent performance of Amp-Seq marker *cpmp* had been demonstrated earlier [7]. Such increased resolution is of great practical value for PCR-correction in clinical drug efficacy trials, where new infections need to be reliably distinguished from those present in an individual earlier. Robust methods for this application are urgently needed.

For infections with high multiplicity (MOI≥3), inference of multi-locus haplotypes remains challenging (example in FIG S6). Inference is straightforward if haplotypes occur at distinctive abundance in any of the longitudinal samples. If haplotypes are equally abundant in a sample and remain so over time, the multi-locus haplotype cannot be inferred. The same is true for complex patterns of shared haplotypes. In the present study, multi-locus haplotypes up to MOI=3 were inferred. For higher multiplicity, sophisticated statistical methods like Markov chain Monte Carlo on longitudinal samples could be applied [25].

Genotyping longitudinal samples in duplicates enabled an evidence-based approach to identify false-negative haplotypes. This permitted to estimate each marker's sensitivity to detect minority clones. Amp-Seq genotyping with markers *ama1*-D2*, ama1*-D3 and *cpmp* missed less clones compared to *msp2*-CE genotyping (Amp-Seq in average 5.4% versus 14.9% *msp2*-CE)*.* This difference is likely due to less stringent cut-off criteria for Amp-Seq compared to *msp2* genotyping. Minority clone detection by *msp2*-CE is limited by peak calling cut-off criteria, which are usually a fixed minimal signal intensity plus a minimum peak height of 10% (used in our study) or more of the dominant peak. Minority clones with an abundance of <10% of all amplified fragments will not pass these criteria. An increase of *msp2*-CE sensitivity would require a lower cut-off, which would lead to more false positive signals from either stutter peaks or background noise. In contrast, Amp-Seq allows to remove PCR artefacts before haplotype calling and thus can support a much lower cut-off of <1% [7].

In cohort studies where Amp-Seq genotyping is performed in successive follow up samples of the same patient, an even more relaxed definition of Amp-Seq cut-off criteria would be justifiable. In this scenario, the same evidence-based strategy of using successive samples can be used to recover minority haplotypes that were detected with read counts below the haplotype calling cut-offs. If recovery would be performed in this study, ≥57% of all false-negative haplotypes would be identified. Such recovery would increase detectability of parasite clones by Amp-Seq to >97%. In addition, multi-locus haplotypes could provide additional evidence for accurate recovery.

The higher sensitivity of Amp-Seq to detect minority clones compared to *msp2*-CE substantially increased MOI, but did not affect mean $_{\mathrm{mol}}$FOI. Any estimation of $_{\mathrm{mol}}$FOI needs to account for temporary absence of clones from the peripheral blood caused by sequestration [1,19–21]. A clone that is temporarily undetectable owing to density fluctuations is likely observed at either the preceding or succeeding bleed. Therefore, a clone is usually only counted as new infection if it was not detected in ≥2 consecutive blood samples. As a consequence, a clone missed at a single bleed will not necessarily lead to a decrease of $_{\mathrm{mol}}$FOI.

A clone that was intermittently missed at one bleed by *msp2*-CE was always detected by Amp-Seq. This observation supports the practice in earlier papers where intermittently missed clones were imputed [21]. Counting a recurrent haplotype as new infection after a single negative bleed would lead to an overestimation of $_{\mathrm{mol}}$FOI [1,19–21]. The statistical power of this study was limited and a larger study is needed to fully explore the effect of the typing method used on estimates of MOI, $_{\mathrm{mol}}$FOI, or even prevalence rates.

A major advantage of Amp-Seq over *msp2*-CE is that the density of an individual clone in multi-clone infections can be calculated. Quantifying the density of individual parasites clones over time permits to study dynamics, and thus fitness, of parasite clones exposed to within-host competition [26]. For example, the relative densities of new infections can be compared to clones already persisting in a host, and their densities in respect to extrinsic factors or clinical symptoms can be investigated.

CONCLUSION

Amplicon sequencing improves clone detectability compared to *msp2*-CE owing to its greater sensitivity for detection of minority clones. Our results confirm earlier assumptions on clone persistence with intermittent missed observation. This validates the imputation of false negatives to correct for imperfect detection of clones, a strategy also used in previous studies on clone dynamics. Using multi-locus haplotypes for genotyping permitted to identify robustly individual clones and improved differentiation between new and recurring clones. Construction of multi-locus haplotypes are of great value to compensate the effects of highly abundant haplotypes in the population. The option to quantify individual clones enables new approaches to investigate effects of parasite fitness or superinfection in multi-clone infections.

Acknowledgement

We are grateful to the study participants and their guardians and to the field and laboratory team of the PNG Institute of Medical Research, in particular Alice Ura. We would also like to thank Stephen Wilcox for supervising sequence library preparation and sequencing.

Authors Contributions

Conceived and designed the experiments: IF, IM, AL, CK. Performed the experiments: AL, CK, JHK, NH, ARU. Supervised field work: IB, ARU. Analysed the data: AL. Supervision: IF. Writing - draft: AL, IF. All Co-authors have read the manuscript and agreed with the final version.

Financial support

This work was supported by the Swiss National Science Foundation [grant number 310030_159580] and the International Centers of Excellence for Malaria Research [grant number U19 AI089686]. AL was partly funded by Novartis Foundation for Medical-Biological Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of interest

All authors: No reported conflicts.

Reference

1. Felger I, Maire M, Bretscher MT, et al. The Dynamics of Natural Plasmodium falciparum Infections. Gosling RD, editor. PLoS One. Public Library of Science; **2012**; 7(9):e45542.

2. Hofmann NE, Karl S, Wampfler R, et al. The complex relationship of exposure to new Plasmodium infections and incidence of clinical malaria in Papua New Guinea. Elife. **2017**; 6:1–23.

3. Koepfli C, Schoepflin S, Bretscher M, et al. How much remains undetected? Probability of molecular detection of human Plasmodia in the field. PLoS One. **2011**; 6(4):e19010.

4. Sonden K, Doumbo S, Hammar U, et al. Asymptomatic Multiclonal Plasmodium falciparum Infections Carried Through the Dry Season Predict Protection Against Subsequent Clinical Malaria. J Infect Dis. **2015**; 212(4):608–616.

5. Messerli C, Hofmann NE, Beck H-P, Felger I. Critical Evaluation of Molecular Monitoring in Malaria Drug Efficacy Trials and Pitfalls of Length-Polymorphic Markers. Antimicrob Agents Chemother. **2017**; 61(1):AAC.01500-16.

6. Miller RH, Hathaway NJ, Kharabora O, et al. A deep sequencing approach to estimate Plasmodium falciparum complexity of infection (COI) and explore apical membrane antigen 1 diversity. Malar J. BioMed Central; **2017**; 16(1):490.

7. Lerch A, Koepfli C, Hofmann NE, et al. Development of amplicon deep sequencing markers and data analysis pipeline for genotyping multi-clonal malaria infections. BMC Genomics. BMC Genomics; **2017**; 18(1):864.

8. Levitt B, Obala A, Langdon S, Corcoran D, O'Meara WP, Taylor SM. Overlap Extension Barcoding for the Next Generation Sequencing and Genotyping of Plasmodium falciparum in Individual Patients in Western Kenya. Sci Rep. Nature Publishing Group; **2017**; 7(September 2016):41108.

9. Mideo N, Bailey JA, Hathaway NJ, et al. A deep sequencing tool for partitioning clearance rates following antimalarial treatment in polyclonal infections. Evol Med public Heal. **2016**; 2016(1):21–36.

10. Betuela I, Rosanas-Urgell A, Kiniboro B, et al. Relapses contribute significantly to the risk of Plasmodium vivax infection and disease in Papua New Guinean children 1-5 years of age. J Infect Dis. **2012**; 206(11):1771–80.

11. Falk N, Maire N, Sama W, et al. Comparison of PCR-RFLP and Genescan-based genotyping for analyzing infection dynamics of Plasmodium falciparum. Am J Trop Med Hyg. **2006**; 74(6):944–50.

12. Arnott A, Wapling J, Mueller I, et al. Distinct patterns of diversity, population structure and evolution in the AMA1 genes of sympatric Plasmodium falciparum and Plasmodium vivax populations of Papua New Guinea from an area of similarly high transmission. Malar J. **2014**; 13(1):233.

13. Cortés A, Mellombo M, Masciantonio R, Murphy VJ, Reeder JC, Anders RF. Allele specificity of naturally acquired antibody responses against Plasmodium falciparum apical membrane antigen 1. Infect Immun. **2005**; 73(1):422–30.

14. Cortés A, Mellombo M, Mueller I, Benet A, Reeder JC, Anders RF. Geographical structure of diversity and differences between symptomatic and asymptomatic infections for Plasmodium falciparum vaccine candidate AMA1. Infect Immun. **2003**; 71(3):1416–26.

15. Hodder AN, Crewther PE, Matthew ML, et al. The disulfide bond structure of *Plasmodium* apical membrane antigen-1. J Biol Chem. **1996**; 271(46):29446–52.

16. Bahl A, Brunk B, Crabtree J, et al. PlasmoDB: The Plasmodium genome resource. A database integrating experimental and computational data. Nucleic Acids Res. 2003. p. 212–215.

17. Bretscher MT, Valsangiacomo F, Owusu-Agyei S, Penny M a, Felger I, Smith T. Detectability of Plasmodium falciparum clones. Malar J. **2010**; 9:234.

18. Rosanas-Urgell A, Mueller D, Betuela I, et al. Comparison of diagnostic methods for the detection and quantification of the four sympatric Plasmodium species in field samples from Papua New Guinea. Malar J. **2010**; 9(1):361.

19. Sama W, Owusu-Agyei S, Felger I, Dietz K, Smith T. Age and seasonal variation in the transition rates and detectability of Plasmodium falciparum malaria. Parasitology. **2006**; 132(Pt 1):13–21.

20. Sama W, Owusu-Agyei S, Felger I, Vounatsou P, Smith T. An immigration-death model to estimate the duration of malaria infection when detectability of the parasite is imperfect. Stat Med. **2005**; 24(21):3269–88.

21. Smith T, Felger I, Fraser-Hurt N, Beck HP. Effect of insecticide-treated bed nets on the dynamics of multiple Plasmodium falciparum infections. Trans R Soc Trop Med Hyg. **1999**; 93 Suppl 1:53–7.

22. Bousema T, Okell L, Shekalaghe S, et al. Revisiting the circulation time of Plasmodium falciparum gametocytes: molecular detection methods to estimate the duration of gametocyte carriage and the effect of gametocytocidal drugs. Malar J. **2010**; 9:136.

23. MalariaGEN Plasmodium falciparum Community Project. Genomic epidemiology of artemisinin resistant malaria. Elife. **2016**; 5:1–29.

24. Barry AE, Schultz L, Buckee CO, Reeder JC. Contrasting population structures of the genes encoding ten leading vaccine-candidate antigens of the human malaria parasite, Plasmodium falciparum. PLoS One. **2009**; 4(12):e8497.

25. Zhu SJ, Almagro-Garcia J, McVean G. Deconvolution of multiple infections in Plasmodium falciparum from high throughput sequencing data. Bioinformatics. **2017**; .

26. Roode JC de, Culleton R, Cheesman SJ, Carter R, Read AF. Host heterogeneity is a determinant of competitive exclusion or coexistence in genetically diverse malaria infections. Proc R Soc B Biol Sci. **2004**; 271(1543):1073–1080.

FIGURES

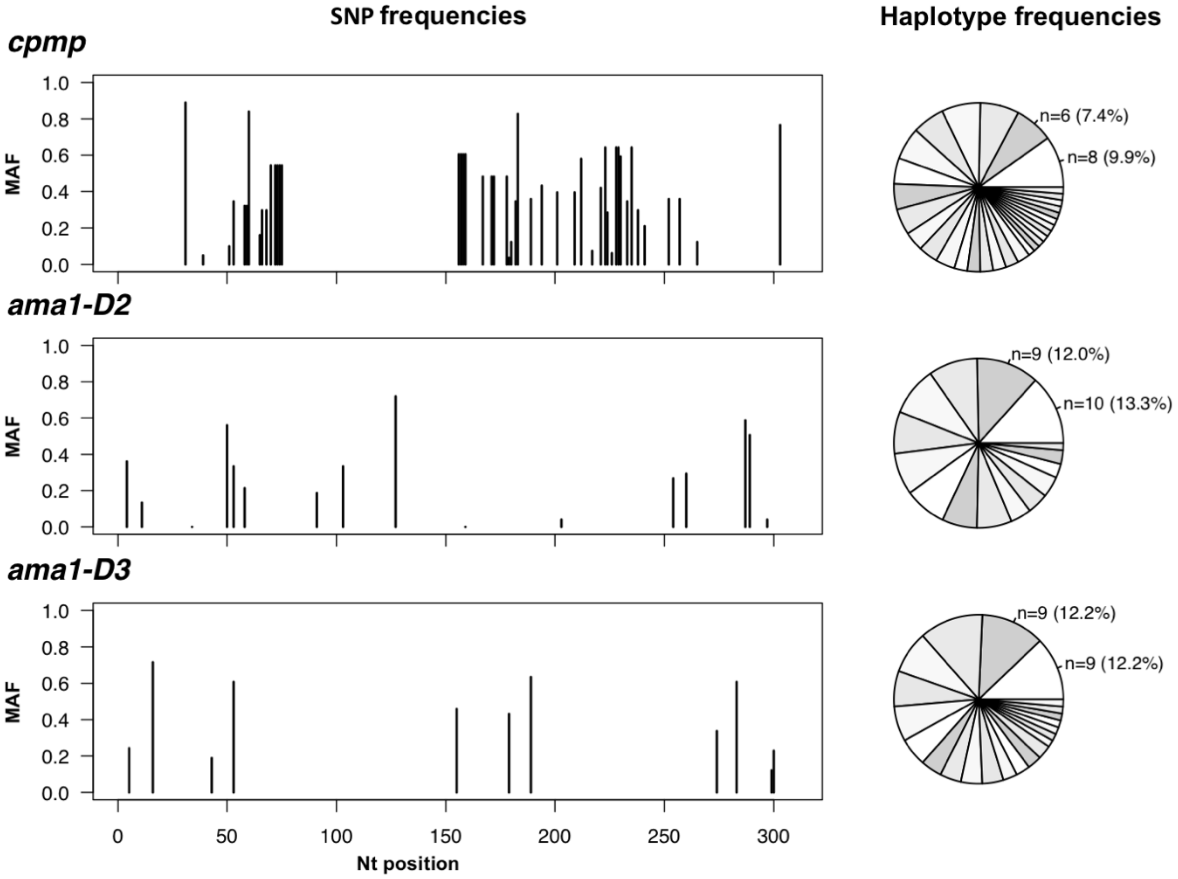
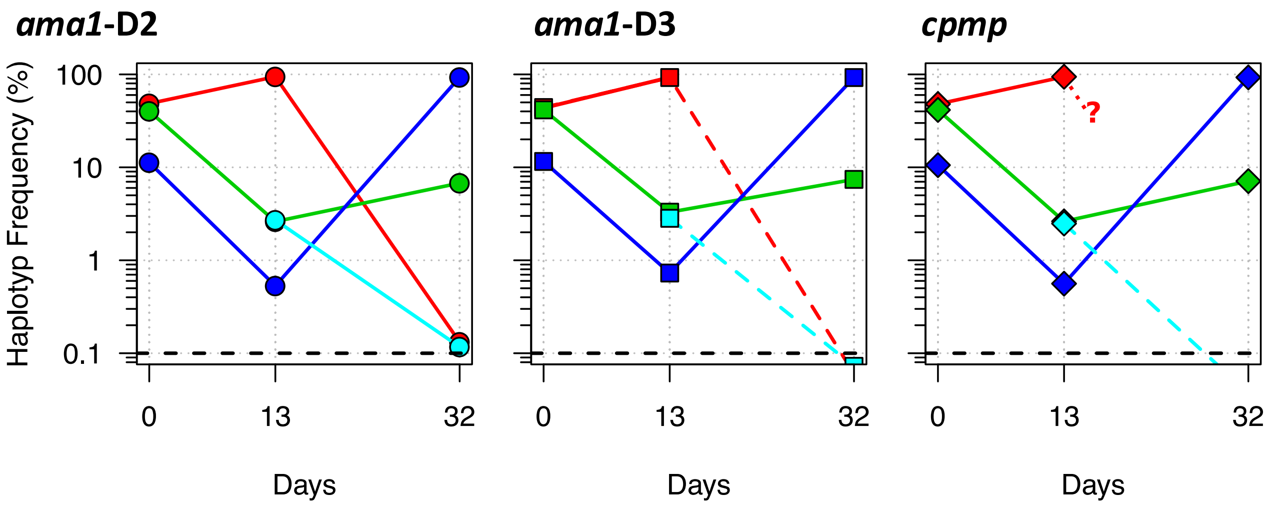


FIG 1: Frequency of individual SNPs and haplotypes of three markers in 33 baseline samples from PNG. Minor allelic frequency (MAF) of each SNP (left) and frequency of haplotypes in these baseline samples (right). n, number of observations per haplotype shown for 2 most prevalent haplotypes. Total number of different haplotypes: 30 for *cpmp*, 15 for *ama1*-D2 and 22 for *ama1*-D3. (Frequency of haplotypes for markers *msp2*-CE given in FIG S2).



Days	Clone 1			Clone 2			Clone 3			Clone 4		
										○	□	◇
0	48.5	44.0	48.0	40.2	41.5	41.4	11.3	11.6	10.6	-	-	-
13	94.4	92.6	94.3	2.61	3.29	2.62	0.53	0.73	0.56	2.71	2.83	2.49
32	0.13	0.06	n.d.	6.77	7.42	7.10	93.1	92.5	92.9	0.12	0.07	0.04

FIG 2: Within-host haplotype frequencies of Amp-Seq markers in longitudinal samples from one child. Inserted table lists within-host multi-locus haplotype frequencies in percent. Multi-locus haplotypes have the same colour-code in figures and table. Solid line represents persisting haplotypes above cut-off criteria (true-positive haplotypes). Dashed line represents persisting haplotypes falling below cut-off criteria (false-negative haplotypes detected below cut-off criteria). Dotted line and question mark indicate a false-negative haplotype that was not detected (n.d.) but could be imputed based on the established multi-locus haplotypes from the preceding sample. Black dashed line represents cut-off criteria of the Amp-Seq genotyping method.

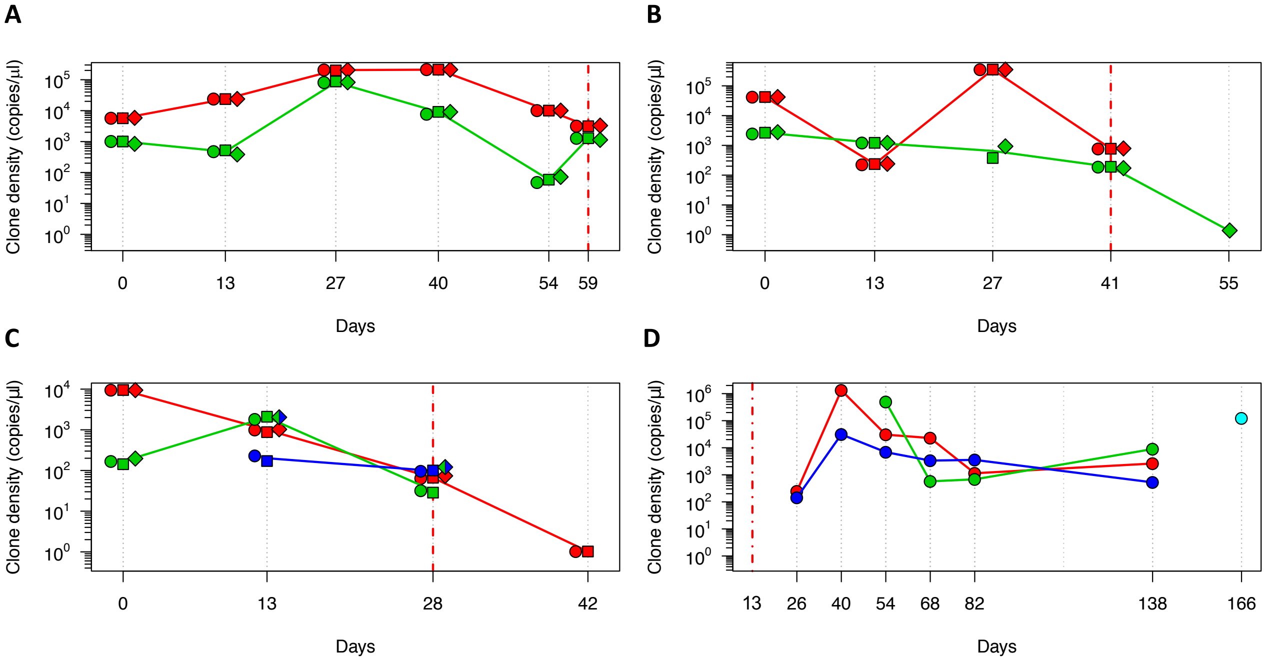


FIG 3: Dynamics of multi-clone infections in 4 children. Multi-marker haplotypes could be generated in panels A, B and C. Inference of multi-locus haplotypes was not possible for the child in panel D; here the dynamics of individual clones tracked by marker ama1-D2 are shown. Each colour represents a clone. individual markers represented by different shapes: *cpmp* (diamonds), *ama1*-D2 (circles) and *ama1*-D3 (squares). Solid line connecting multi-locus haplotypes represents their median frequency. Grey dotted vertical lines represent sampling dates. Red dashed lines represent day of artemisinin combination therapy. Red dash-dotted line represents end of radical cure (artesunate-primaquine) at baseline.

TABLES

Table 1: Genotyping results of 4 molecular markers analysed in 33 baseline field samples.

Marker	**H_e**	**Mean MOI**	**Number of clones[1]**	**Number of haplotypes**	**Number of SNPs[2]**
msp2 CE	0.940	1.73 [3]	57	20	n/a
cpmp	0.961	2.45 [3]	81	30	48
ama1-D2	0.928	2.27 [3]	75	15	17
ama1-D3	0.939	2.24 [3]	74	22	11

H_e, expected heterozygosity.

MOI, multiplicity of infection.

[1] Sum of all haplotypes in all samples.

[2] With respect to the reference sequence of *P. falciparum* strain 3D7.

[3] Pairwise comparison using two-sided paired t-test with adjusted p-value by Holm: p-value=0.008 for *ama1*-D2 vs *msp2*-CE, p-value=0.036 for *ama1*-D3 vs *msp2*-CE, and p-value=0.005 for *cpmp* vs *msp2*-CE.

Table 2: Genotyping results of 3 molecular markers analysed in 47 independent field samples with 66 different clones. H_e, expected heterozygosity.

Marker	**H_e**	**Number of Haplotypes**
cpmp	0.948	25
ama1-D2	0.926	16
ama1-D3	0.938	21
cpmp + ama1-D2	0.992	53
cpmp + ama1-D3	0.994	55
cpmp + ama1-D2 + ama1-D3	0.994	55

Table 3: Numbers of missed haplotypes due to imperfect detection either at baseline, in any intermediate sample, or prior to haplotype clearance. Haplotypes from 48 longitudinal samples from 12 children were classified into true-positive (TP) and false-negative haplotypes. Two types of false-negative haplotypes (missed clones) can be differentiated: (FN_i) False-negative haplotypes detected but below cut-off criteria and (FN_{ii}) false-negative haplotypes not detected but imputed.

Marker	Baseline sample	Any intermediate sample	Sample prior to clearance
	n	n	n
TP haplotypes			
msp2-CE	29	34	23
cpmp	39	45	31
ama1-D2	36	44	29
ama1-D3	36	43	29
FN_i haplotypes			
msp2-CE	2	6	2
cpmp	1	0	3
ama1-D2	0	1	2
ama1-D3	1	0	3
FN_{ii} haplotypes			
msp2-CE	n/a [1]	5	n/a [1]
cpmp	0	2	1
ama1-D2	1	0	0
ama1-D3	1	2	0

[1] FN_{ii} haplotypes cannot be imputed at the beginning of infection or prior to clearance for marker *msp2*-CE.

Table 4: Sensitivity and false discovery rate (FDR) of the genotyping method. Sensitivity and FDR was estimated based on persistent clones in 48 longitudinal samples from 12 individuals. Detectability of minority clone can be increased by including missed persistent haplotypes detected below the cut-off criteria. TP, true-positive haplotypes. FN$_i$, false-negative haplotypes detected, but below cut-off criteria. FN$_{iiab}$, false-negative haplotypes with no read detected.

Marker	**TP**	**FN**			**FP**	**Sensitivity**	**FDR**	**Detected Haplotypes**[1]
	n	n$_i$	n$_{iia}$	n$_{iib}$	n	TP/(TP+FN$_{i+iiab}$)	FP/(TP+FP)	(TP+FN$_i$)/(TP+ FN$_{i+iiab}$)
msp2-CE	86	10	5	n/a[2]	n/a[3]	0.851[4]	n/a[3]	0.950
cpmp	115	4	2	1	5	0.943	0.042	0.975
ama1-D2	109	3	0	1	1	0.965	0.009	0.991
ama1-D3	108	4	2	1	3	0.939	0.027	0.974

[1] Detected true-positive and false-negative haplotypes.

[2] Not imputed for *msp2*-CE as multi-locus haplotypes cannot be established.

[3] Length-polymorphic data generated in different laboratories do not provide replicates suited for determination of false-positive haplotype calls and estimation of FDR.

[4] Without haplotypes, that were imputed based on multi-locus haplotypes at the beginning or end of an infection.

Supplemental Figures

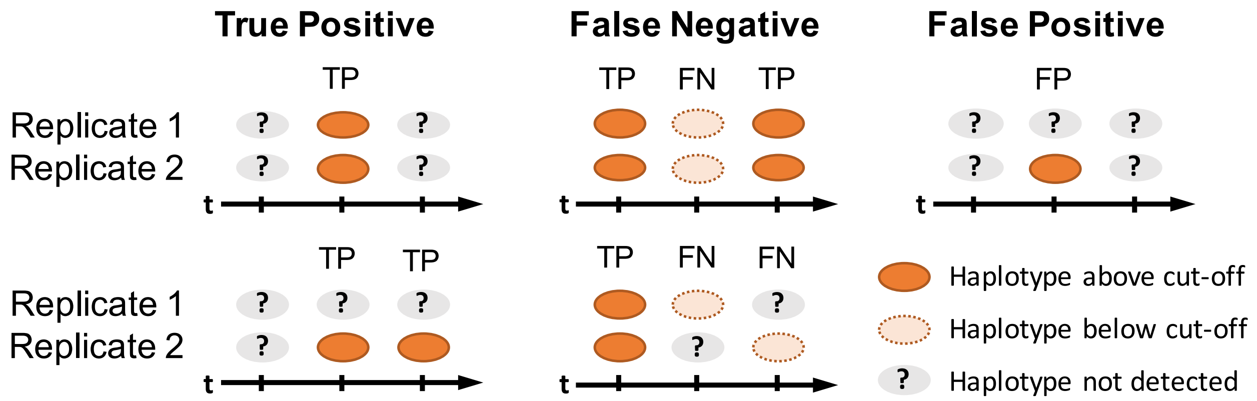


FIG S1: Schematic of haplotype classification. Examples show the classification of haplotypes in true-positive (TP), false-negative (FN) and false-positive (FP), based on their detection either in duplicates or in the preceding or succeeding bleeds.

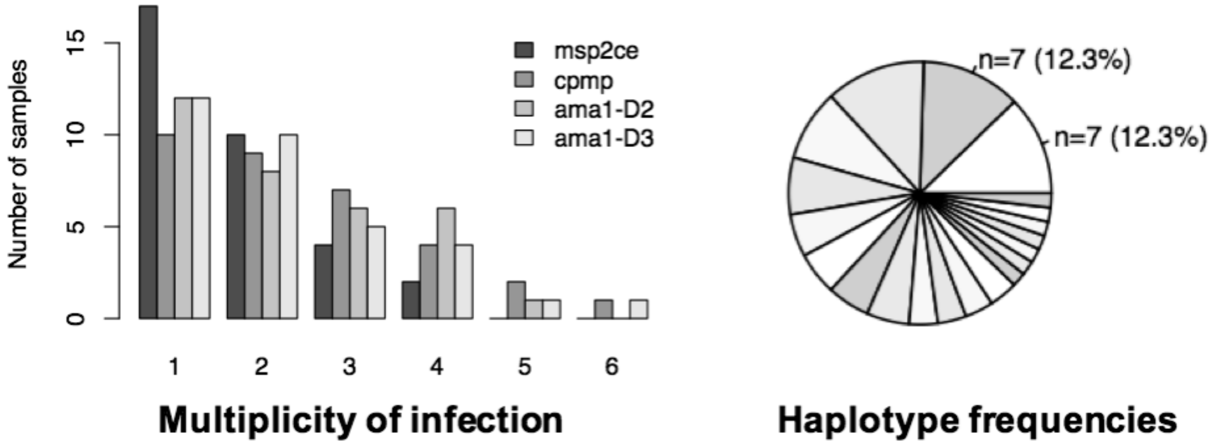


FIG S2: Frequency distribution of multiplicity of infection by marker (left) and frequency of *msp2*-CE haplotypes (right) in 33 baseline samples. Marker *msp2*-CE identified 20 different haplotypes. (Frequency distribution of haplotypes of Amp-Seq markers given in FIG 1).

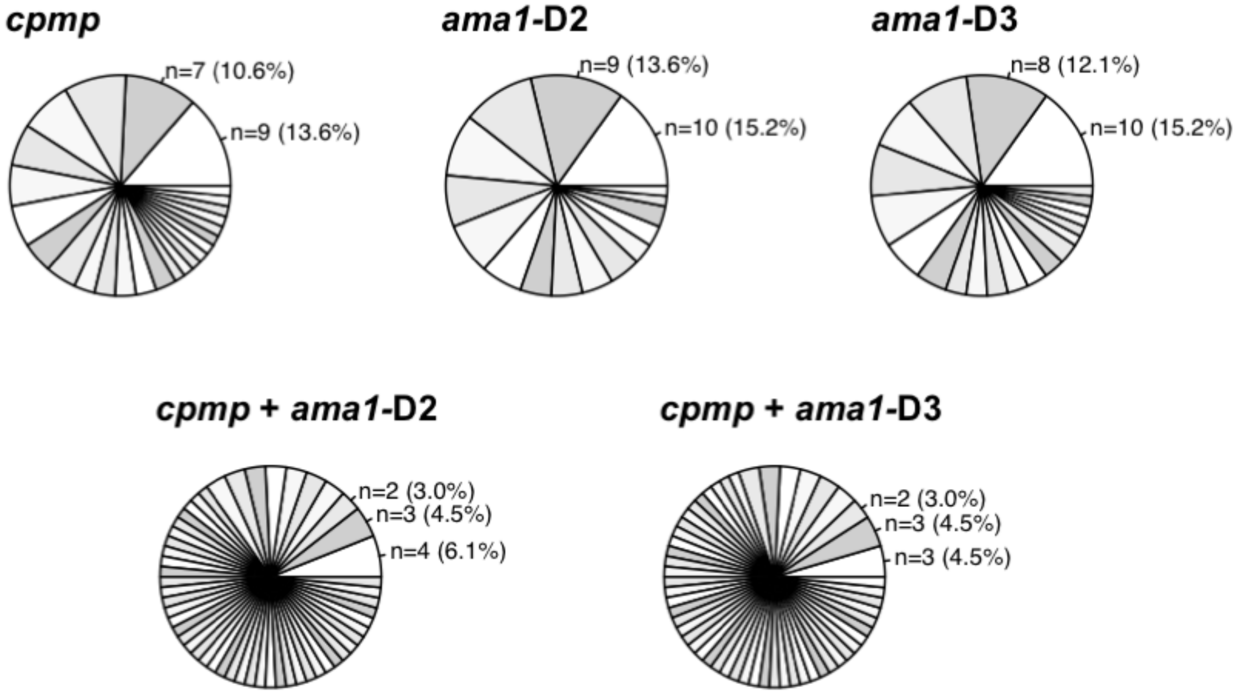


FIG S3: Haplotype frequencies by marker in 46 independent samples comprising 66 clones. For marker *cpmp* 25 different alleles were identified, for *ama1*-D2 16 haplotypes and for *ama1*-D3 21 haplotypes. Top panel: haplotypes base on single markers; bottom panel: two-marker haplotypes.

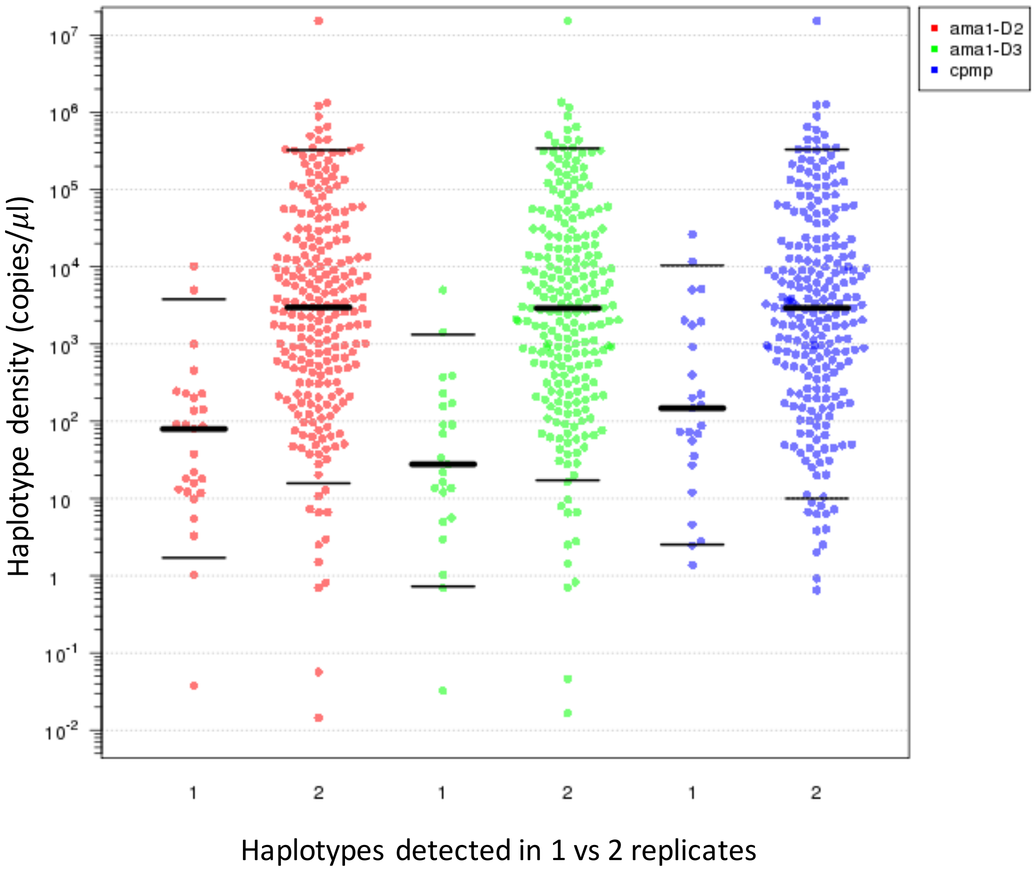


FIG S4: Density of true-positive haplotypes detected in only one or both replicates. X-axis, haplotypes detected in 1 versus 2 replicates by Amp-Seq marker. Y-axis, haplotype density by qPCR measured as 18S rRNA gene copies per µl whole blood. Points represent individual haplotypes; colours represent individual markers. Black horizontal bar represents 5, 50 and 95-percentile. Wilcoxon rank sum test with continuity correction: W=1000 and p-value=2x10^{-9} for *ama1*-D2, W=700 and p-value=5x10^{-9} for *ama1*-D3, W=1000 and p-value=5x10^{-5} for *cpmp*.

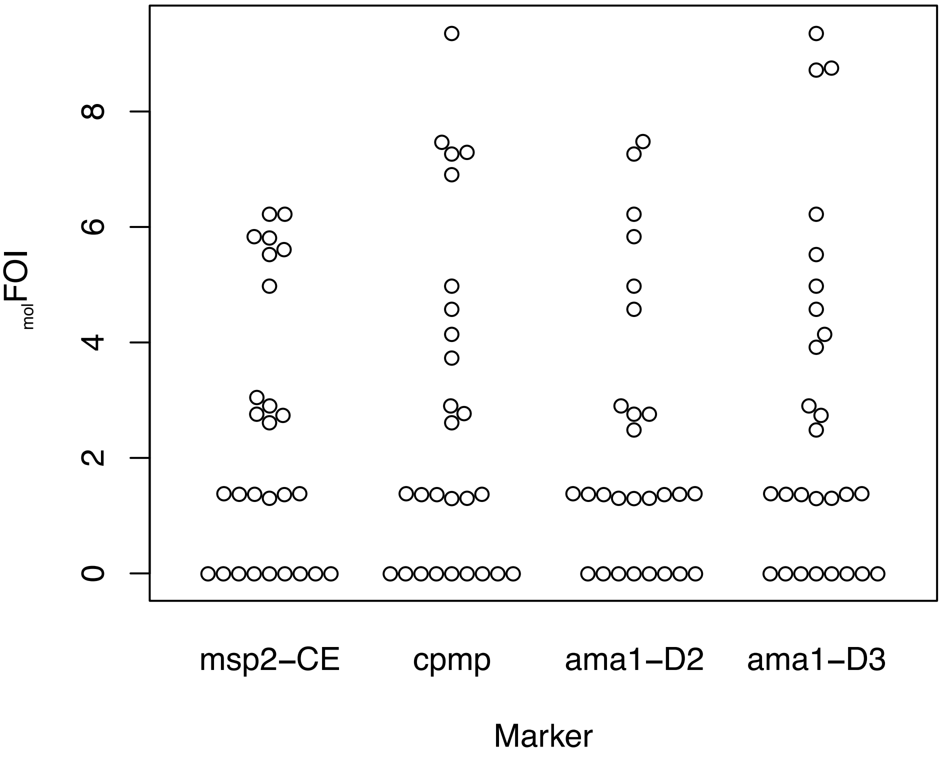
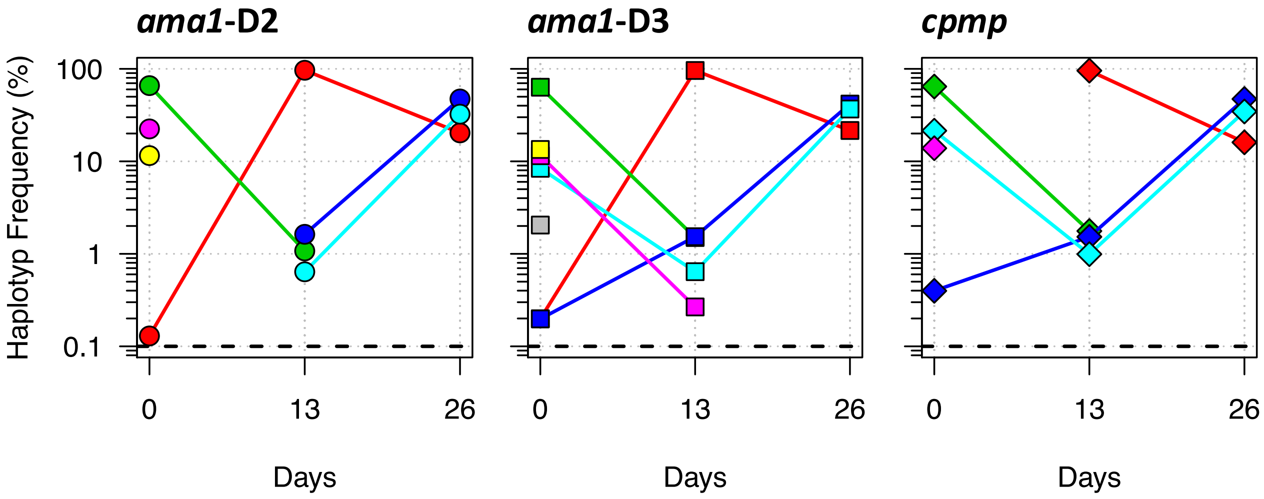


FIG S5: Frequency distribution of molecular force of infection ($_{mol}$FOI) by marker. A total of 117 samples from 27 individuals (on average 4.3 samples per individual [min: 2, max: 7]) were used to estimate force of infection (FOI).



Multi-locus haplotype names	Day 0			Day 13			Day 32		
	ama1-D2 %	*ama1*-D3 %	*cpmp* %	*ama1*-D2 %	*ama1*-D3 %	*cpmp* %	*ama1*-D2 %	*ama1*-D3 %	*cpmp* %
Haplotype 1	65.9	63.3	64.4	1.1	1.5	1.8			
Haplotype 2	0.1	0.2		96.6	96	95.6	20.3	21.6	16.1
Haplotype 3		0.2	0.4	1.6	1.5	1.5	47.3	41.6	47.2
Haplotype 4		8.4	21.4	0.6	0.6	1	32.4	36.8	34.6
Haplotype 5	22.4								
Haplotype 6	11.5								
Haplotype 7		11.5			0.3				
Haplotype 8		13.4							
Haplotype 9		2.0							
Haplotype 10			13.9						

FIG S6: Within-host haplotype frequency of Amp-Seq markers in longitudinal samples from 1 child representing an unresolvable multi-locus haplotype. Inserted table lists within-host haplotype frequencies for all markers with a possible solution of partly established multi-locus haplotypes for the major haplotypes. Multi-locus haplotypes 1-3 match well in frequencies of individual haplotypes at day 0, 13 and 32. In contrast, multi-locus haplotype 4 does not match in frequencies of individual haplotypes at day 0. This could be explained by a complex shared haplotype situation with one or several clones detected only at day 0 and 13, e.g. haplotypes 5-10. Solid lines represent persisting haplotypes.

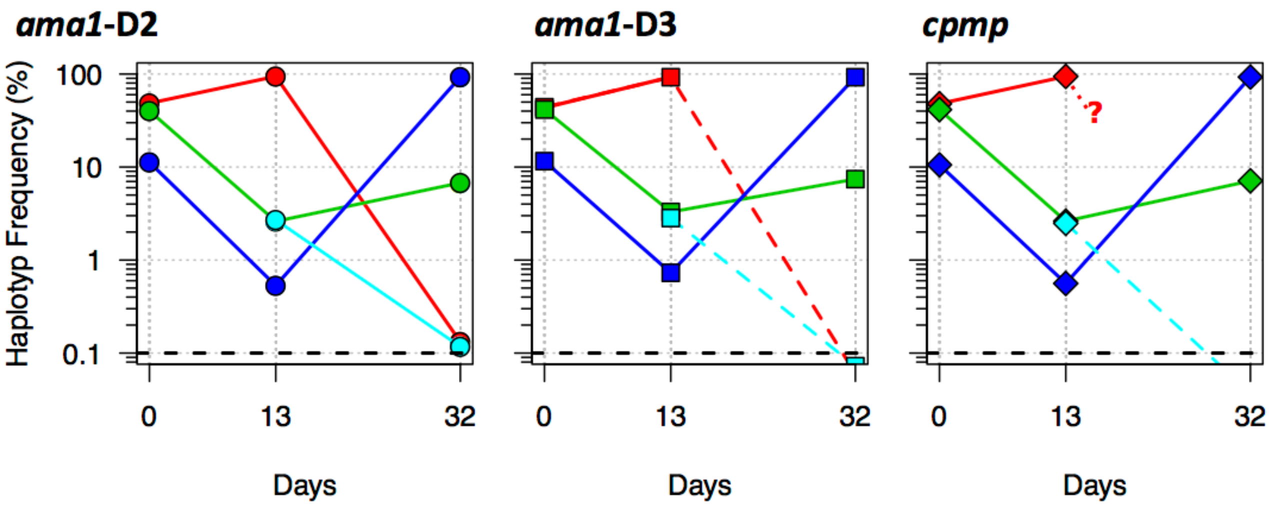


FIG S7: Within-host haplotype frequencies of Amp-Seq markers in longitudinal samples from one child. Multi-locus haplotypes have the same colour-code in figures. Solid line represents persisting haplotypes above cut-off criteria (true-positive haplotypes). Dashed line represents persisting haplotypes falling below cut-off criteria (false-negative haplotypes detected below cut-off criteria). Dotted line and question mark indicate a false-negative haplotype that was not detected but could be imputed based on the established multi-locus haplotypes from the preceding sample. Black dashed line represents cut-off criteria of the Amp-Seq genotyping method.

SUPPLEMENTAL TABLES

Table S1: PCR Primer sequence for Amp-Seq and *msp2*-CE genotyping and sequence library preparation.

Primer for primary PCR	
cpmp_prim_F	CGATACAGGACATATAGA
cpmp_prim_R	TTCAATAACATTTACTAGG
Pfama1_F5	TGCGTATTATTATTGAGC
Pfama1_R613	GTGTTGTATGTGATGCTC
Primer for nested PCR	
ama1_D2_F_Linker	GTGACCTATGAACTCAGGAGTC**GGTCCTAGATATTGTAATAAAG**
ama1_D2_R_Linker	CTGAGACTTGCACATCGCAGC**CATGTTGGTTTGACATTAAA**
ama1_D3_F_Linker	GTGACCTATGAACTCAGGAGTC**TACTACTGCTTTGTCCCATC**
ama1_D3_R_Linker	CTGAGACTTGCACATCGCAGC**TCAGGATCTAACATTTCATC**
cpmp_F_Linker	GTGACCTATGAACTCAGGAGTC**CATAAGTCATTAAAATTTATGGAT**
cpmp_R_Linker	CTGAGACTTGCACATCGCAGC**CGTTACTATCAAGATCGTTAATATC**
Primer for msp2 CE genotyping	
msp2_S2_fw	GAAGGTAATTAAAACATTGTC
msp2_S3_rev	GAGGGATGTTGCTGCTCCACAG
msp2_S1-fw	GCTTATAATATGAGTATAAGGAGAA
msp2_FC27-rev	GCATTGCCAGAACTTGAA
msp2_3D7-rev	CTGAAGAGGTACTGGTAGA
Primer for sequence library PCR (XXXXXX=barcode)	
Forward	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXXXXGTGACCTATGAACTCAGGAGTC
Reverse	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTXXXXXXXXCTGAGACTTGCACATCGCAGC

Forward barcode		**Reverse barcode**	
Fwd_1	TAGATCGC	Rev_1	TAAGGCGA
Fwd_2	CTCTCTAT	Rev_2	CGTACTAG
Fwd_3	TATCCTCT	Rev_3	AGGCAGAA
Fwd_4	AGAGTAGA	Rev_4	TCCTGAGC
Fwd_5	GTAAGGAG	Rev_5	GGACTCCT
Fwd_6	ACTGCATA	Rev_6	TAGGCATG
Fwd_7	AAGGAGTA	Rev_7	CTCTCTAC
Fwd_8	CTAAGCCT	Rev_8	CAGAGAGG
Fwd_13	TGGTGGTA	Rev_9	GCTACGCT
Fwd_14	TTCACGCA	Rev_10	CGAGGCTG
Fwd_15	AGCACCTC	Rev_11	AAGAGGCA
Fwd_16	CAAGGAGC	Rev_12	GTAGAGGA
Fwd_17	ATTGGCTC	Rev_13	ATGCCTAA
Fwd_18	CACCTTAC	Rev_14	ACGCTCGA
Fwd_19	CTAAGGTC	Rev_15	AGTCACTA
Fwd_20	GAACAGGC	Rev_16	ATCCTGTA
		Rev_17	CGCATACA
		Rev_18	CTGGCATA
		Rev_19	GATAGACA
		Rev_20	GCTAACGA
		Rev_21	GTGTTCTA
		Rev_22	TCCGTCTA
		Rev_23	CCTAATCC
		Rev_24	GACAGTGC

Table S2: Location and size of the amplicons.

	cpmp	***ama1*-D2**	***ama1*-D3**
From	1895	775	1281
To	2324	1253	1796
Size	430	479	516

Table S3: Summery of sequence coverage (total read numbers) by Amp-Seq marker.

	cpmp	*ama1*-D2	*ama1*-D3
1st Qu.	247	2292	2997
Median	794	3386	4716
Mean	1117	3682	5189
3rd Qu.	1632	5143	6906
Max	6376	11570	34240

Table S4: Summary of multi-locus haplotype (MLH) inference based on longitudinal samples from 33 children.

Status of MLH inference	**Samples**	**Multi-locus haplotypes**	**Single-locus haplotypes**		
			cpmp	***ama1*-D2**	***ama1*-D3**
	n	**n**	**n**	**n**	**n**
Full established MLH	78 [1]	116 [1]	116	103	103
Partly established MLH [2]	49	64	135	130	126
Unresolvable MLH [3]	8	0	20	18	18
Incomplete datasets [4]	13	0	7	11	11
Total	140	180	258 [5]	244 [5]	240 [5]

n number of samples or haplotypes.

[1] 45 out of 78 samples with fully established multi-locus haplotypes were single clone infections.

[2] Samples were multi-locus haplotypes could be established for some but not for all clones of a sample.

[3] Samples were no multi-locus haplotype could be established.

[4] Samples with missing genotyping results for any of the markers.

[5] Total number of parasite clones detected in 140 samples was 277.

Table S5: Overview of sample selection criteria applied for different types of analyses.

Analysis Type	**Samples n**	**Children n**	**Selection Criteria**
Baseline H_e and MOI	33	33	Baseline (or first bleed available) sample.
Multi-locus H_e	46	33	Samples with a resolvable multi-locus haplotype that were separated by a treatment plus ≥2 consecutive *P. falciparum* negative samples from the same child.
$_{mol}$FOI	117	27	Children with a complete set of replicates.
Sensitivity and false discovery rate	48	12	Children that did not received antimalarial treatment during the timespan analysed and harboured at least one haplotype that was detected at 3 consecutive bleeds.
Reproducibility	139	33	True-positive haplotypes.

Table S6: Reproducibility of true-positive haplotypes in technical replicates. Reproducibility only decreased when clone densities fell below 1000 copies 18S rRNA gene per µl whole blood and/or within-host frequency below 1% (FIG S5).

	cpmp			***ama1*-D2**			***ama1*-D3**		
	n_1	**n_2**	**q**	**n_1**	**n_2**	**q**	**n_1**	**n_2**	**q**
	25	235	0.949	28	228	0.942	23	226	0.952
Haplotype density (copies/µl)									
>1000	7	148	0.977	2	146	0.993	2	142	0.993
100-1000	6	52	0.945	8	50	0.926	5	51	0.953
10-100	8	23	0.852	13	22	0.772	10	22	0.815
<10	4	12	0.857	5	10	0.800	6	11	0.786
Haplotype proportion within a sample (%)									
>10	13	172	0.964	16	165	0.954	11	167	0.968
1-10	4	55	0.965	4	47	0.959	5	46	0.948
<1	8	8	0.667	8	16	0.800	7	13	0.788

n_1 number of clones detected only with one of the replicates.

n_2 number of clones detected with both replicates.

q detectability as descripted in Bretscher et al. 2010.

SUPPLEMENTAL TEXT

Example of multi-locus haplotype inference

Below an example of *P. falciparum* infection dynamics is shown for one child in great detail to illustrate our strategy for inferring a multi-locus haplotype that combines SNP data from three molecular markers *ama1*-D2, *ama1*-D3, and *cpmp.* Within-host haplotype frequency data of the example is shown in Table S8 and corresponding graphic illustration in FIG S7.

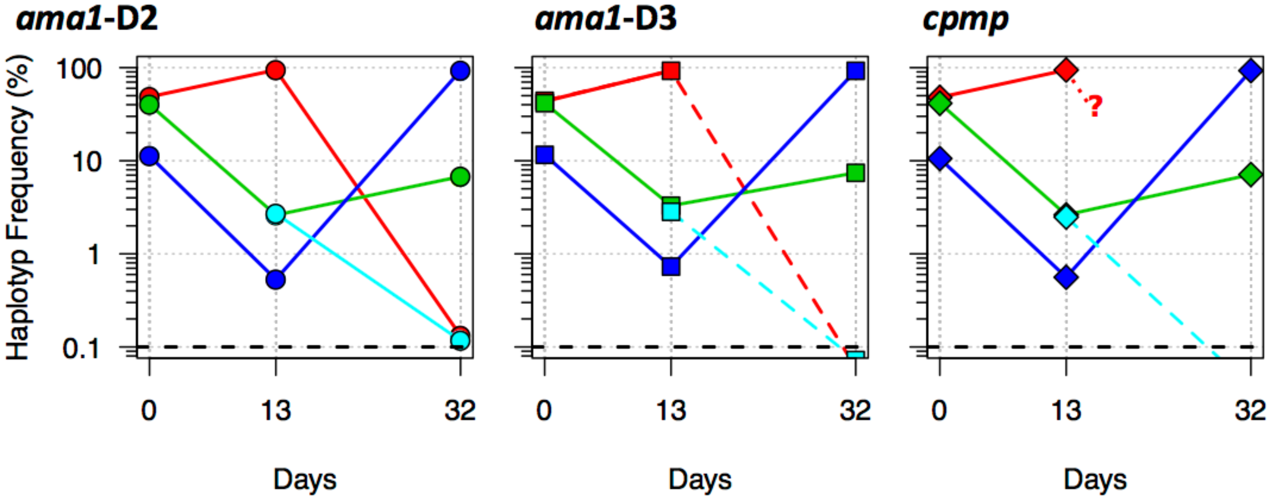


FIG S7: Within-host haplotype frequencies of Amp-Seq markers in longitudinal samples from one child. Multi-locus haplotypes have the same colour-code in figures. Solid line represents persisting haplotypes above cut-off criteria (true-positive haplotypes). Dashed line represents persisting haplotypes falling below cut-off criteria (false-negative haplotypes detected below cut-off criteria). Dotted line and question mark indicate a false-negative haplotype that was not detected but could be imputed based on the established multi-locus haplotypes from the preceding sample. Black dashed line represents cut-off criteria of the Amp-Seq genotyping method.

Table S8: Within-host haplotype frequencies (WHHF) in percent of individual Amp-Seq markers observed in longitudinal samples from one child. Haplotypes of individual markers (termed alleles) are sorted by WHHF of day 0. **Haplotypes 1-4** represent multi-loci haplotypes composed of one allele of each of the 3 markers.

Multi-locus haplotype names	**Day 0** ***ama1*-D2 %**	 ***ama1*-D3 %**	 ***cpmp* %**	**Day 13** ***ama1*-D2 %**	 ***ama1*-D3 %**	 ***cpmp* %**	**Day 32** ***ama1*-D2 %**	 ***ama1*-D3 %**	 ***cpmp* %**
Haplotype 1	48.5	44.0	48.0	94.4	92.6	94.3	0.13	0.06	
Haplotype 2	40.2	41.5	41.4	2.61	3.29	2.62	6.77	7.42	7.10
Haplotype 3	11.3	11.6	10.6	0.53	0.73	0.56	93.1	92.5	92.9
Haplotype 4	-	-	-	2.71	2.83	2.49	0.12	0.07	0.04

The inference of multi-marker haplotypes started with identification of alleles that belong to the dominant parasite clone. A dominant Haplotype was defined by a within-host haplotype frequencies (WHHF) >54%.

Inference of multi-marker Haplotypes at Day 0

At Day 0 of this example, 2 different alleles per marker occurred at similar WHHF (listed by marker in Supplemental Table S8. At Day 0 no dominant Haplotype was evident, therefore any increase or decrease of in WHHF of these alleles at Day13 was interrogated: one allele of each of the 3 markers showed an increase of approx. +46%, while the remaining 3 alleles of similar frequency revealed a decrease by approx. -38%. Based on these recoded frequency changes we combined those alleles from each marker, which all increased by approx. +46%, into multi-locus **Haplotype 1** (FIG S7, Day 0 in red).

Alleles that constituted **Haplotype 1** were not considered in next steps of inference. Additional multi-locus haplotypes of Day 0 were inferred by combining the alleles of similar frequency which showed a decrease in WHHF for all 3 markers of approx. -38%, thus defining multi-locus **Haplotype 2** (FIG S7, Day 0 in green). For the next steps of inference, all alleles associated with multi-locus **Haplotypes** 1 and **2** were no more considered. The remaining alleles constituted multi-locus **Haplotype 3** with ~11% WHHF for all markers (FIG S7, Day 0 in blue).

Multi-marker Haplotypes at Day 13

The dominant alleles in all 3 markers of the Day 13 sample were consistent with multi-locus **Haplotype 1** characterized by ~93% WHHF for all 3 markers (FIG S7, Day 13 in red). Again this multi-locus haplotype was no more considered in the next steps of Day 13 haplotype inference. Next two multi-locus haplotypes with similar WHHF were observed. In agreement with allele combinations found at Day 0, multi-locus **Haplotype 2** was identified by an increase of these alleles at Day 32 of approx. +4% (FIG S7, Day 13 in green). After excluding alleles constituent multi-locus **Haplotypes 1** and **2** an additional new multi-locus **Haplotype 4** with similar WHHF as **Haplotype 2** was found (FIG S7, Day 13 in light blue). The remaining alleles, all with frequencies below 1%, corresponded to multi-locus **Haplotype 3** (FIG S7, Day 13 in blue).

Multi-marker Haplotypes at Day 32

The dominant clone in the Day 32 sample corresponds to multi-locus **Haplotype 3**, characterized in this sample by a steep increase of, ~93% WHHF for all markers (FIG S7, Day 32 in blue). Alleles of this dominant clones are no more considered in the next step of inference. The dominant clone in this step corresponds to **Haplotype 2** with ~7% WHHF for all markers (FIG S7, Day 32 in green). In the next step all alleles of **Haplotypes 3** and **2** were no more considered. But no further multi-locus haplotypes could be established, as WHHF of the remaining alleles were below the 0.1% WHHF cut-off criteria for some of the markers. However, as the inferred multi-locus haplotypes of Day 0, 13 and 32 match for all samples and marker ama1-D2 showed a WHHF above the cut-off criteria, the multi-locus Haplotypes **1** and **4** could be imputed (FIG S7, Day 32 in light blue and red).

CHAPTER 4: DECONVOLUTION OF MIXED-STAGE TRANSCRIPTOMES

NORMALISATION AND DECONVOLUTION OF RNA-SEQ DATA FROM MIXED TRANSCRIPTOMES OF *PLASMODIUM FALCIPARUM* BLOOD STAGES

Anita Lerch[1,2], Sebastian Rusch[1,2], Natalie E. Hofmann[1,2], Armin Passecker[1,2], Camilla Messerli[1,2], Cristian Koepfli[3,4], Hans-Peter Beck[1,2], Ingrid Felger[1,2,#]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia
[4] University of Melbourne, Parkville, Australia
[#] Corresponding author

ABSTRACT

Background
Study gene expression of *Plasmodium* parasites in field samples is of great importance, e.g. to understand mechanism of drug resistance or in absence of a continuous *in vitro* culture system. Transcriptome studies of field samples are complicated by the mixture of different developmental stages present concurrently in the samples. Deconvolution methods permit to infer stage specific gene expression from mixed stage samples with known stage proportions. However, fold increase of total RNA during intra-erythrocytic development cycle complicates deconvolution of mixed stage samples.

Methods
Several deconvolution and normalisation methods were evaluated with experimental mixtures of highly synchronised *P. falciparum* stages. Permutation testing was used to sub-select those genes which had fold change large enough to still be identified as differentially expressed after deconvolution. Inferred significant fold changes (p-value<0.05) were compared to fold changes as observed in stage-specific transcriptomes from highly synchronised *P. falciparum* samples.

Results
Negative binomial regression together with normalisation by the total number of sequence reads showed best agreement in up or down regulation: 96.8% of 239 genes with significant fold changes between ring and trophozoite stage, 99.5% of 1318 genes between ring and schizont stage, and 99.5% of 3627 genes between trophozoite and schizont stage. Significant fold-changes of gene expression identified by permutation testing provided a robust selection criterion for genes which could be successfully deconvoluted.

Conclusion
The identified strategy for deconvolution of mixed-stage transcriptomes and identification significant fold change after deconvolution can be transferred to field samples of any *Plasmodium* species with known stage proportions.

INTRODUCTION

The life cycle of *Plasmodium falciparum* is highly regulated and shows a cascading expression profile (Bozdech et al. 2003; Le Roch et al. 2003). The study of stage-specific gene expression provides important basic knowledge for malaria research, e.g. to understand mechanism of artemisinin drug resistance which shows a decelerated development at young ring stage (Mok et al. 2015). Various time-course studies of the intra-erythrocytic development cycle (IDC) exist for cultured *P. falciparum* strains, giving insights into the stage specific transcriptomes (Bozdech et al. 2003; Le Roch et al. 2003; Otto et al. 2010; Bártfai et al. 2010; Kensche et al. 2016). All time-course transcriptomes can be accessed by PlasmoDB (http://plasmodb.org) (Bahl et al. 2003). In contrast, transcriptome data for *P. vivax* IDC are very limited. Only one study of synchronised short-term cultured blood stages samples exists from infected patients (Bozdech et al. 2008; Zhu et al. 2016). The study of *P. vivax* stage specific gene expression is greatly hampered by a lack of continuous *in vitro* parasite culture. Gene expression has to be studied from *P. vivax* positive blood samples collected in the field that consist of a mixture of different developmental stages. Despite enrichment of a specific developmental stage or after tight synchronisation, small fractions of other stages are found. The transcriptome of *P. vivax* gametocytes, one of the stages found in peripheral blood, has not yet been described. Because *P. vivax* transcriptome analysis must rely on deconvolution of mixed stages, robust approaches to tackle RNA sequencing (RNA-Seq) data from mixed life stage are urgently needed.

Several approaches have been presented in the past to deconvolute observed mixed cell-type transcriptomes measured by microarray or RNA-Seq (Table 1). These deconvolution methods infer either cell-type specific transcriptomes based on known proportions of cell-types in the mixture, or cell-type proportions in the mixture based on known cell-type specific transcriptomes, called signatures (Abbas et al. 2009; Erkkilä et al. 2010; Shen-Orr et al. 2010; Qiao et al. 2012; Gong et al. 2011; Gaujoux & Seoighe 2012; Zhong & Liu 2012; Gong & Szustakowski 2013; Gaujoux & Seoighe 2013; Newman et al. 2015; Joice et al. 2013). All deconvolution approaches assume similar cell quantity, meaning that the transcriptome of each individual stage within a mixture originate from the similar number of cells. This assumption is not valid for gene expression data gained from *Plasmodium species*, as the parasite genome replicates during the IDC. By completion of the IDC, the parasite has reached the schizont stage with up to 32 merozoites. During the IDC the parasite undergoes substantial increase in the amount of total RNA (Bártfai et al. 2010; Sims et al. 2009; Kensche et al. 2016). Therefore, deconvolution methods cannot be applied without taking into account this increase in the total amount of RNA.

Normalisation methods are used to adjust gene expression data for biological differences in RNA composition between samples. Different methods are required for microarray or RNA-Seq data. Most common used methods for RNA-Seq are based on Reads Per Kilobase per Million mapped reads (RPKM), Trimmed Mean of M values (TMM), relative log expression (RLE), or Remove Unwanted Variation (RUV) (Dillies et al. 2013; Mortazavi et al. 2008; Robinson & Oshlack 2010; Anders & Huber 2010; Risso et al. 2014). RPKM normalisation only adjusts for difference in total read counts and gene length, whereas TMM, RLE, and RUV normalisation also adjust for difference in RNA composition between the samples. TMM and RLE normalisation make use of the assumption that the majority of genes are expressed at a constant level. Those normalisation methods might not work for transcriptome data of *Plasmodium* species, as most genes have a periodically fluctuating gene expression. Another possibility for normalisation is to use Biological Scaling Normalization (BSN), which normalises by an experimentally measured parameter, the so called 'scale' (Aanes et al. 2014).

An approach to deconvolute mixed transcriptomes of *Plasmodium* parasites field samples measured by affymetrix microarray (ThermoFisher Scientific) was presented earlier (Joice et al. 2013). Applying this approach to RNA-Seq data from experimentally mixed stage transcriptomes of *P. falciparum* did not provide satisfying results. To determine if another deconvolution method is better suited for RNA-Seq data of *Plasmodium* parasites field samples, several existing normalisation and deconvolution methods were evaluated with experimentally mixed-stage transcriptomes of *P. falciparum*. The specific aim of this work was to infer stage-specific transcriptomes from samples composed of mixed-stages, by using stage counts determined by light microscopy (LM) for all samples analysed. Solving this problem by using experimentally mixed, highly synchronized stages of *P. falciparum* as a proof-of-concept forms the basis for the ultimate goal of estimating the gametocyte transcriptome from enriched *P. vivax* field samples. As an additional step, knowledge of the expression signature of all *P. vivax* parasite stages might permit to infer the composition of *P. vivax* parasite stages from field samples with unknown stage composition based either on qRT-PCR or RNA-Seq data.

MATERIALS AND METHODS

Extraction of viral/HIV RNA

Extracted viral RNA from a supernant of non-infectious HIV-1 virus (4×10^9 RNA copies/ml) was used as spike-in control and was kindly provided by Department of Biomedicine, University of Basel. The non-infectious HIV-1 virus originating from parental HIV-1 NL4-3 strain carries a large deletion in the *env* gene Viral RNA was extracted using the QIAamp Viral RNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol.

Cultivation and synchronisation of P. falciparum HB3 cultures

Plasmodium falciparum strain HB3 was cultured in RPMI 1640 (Gibco life technologies), and 0.5% Albumax (Invitrogen), 50mg Hypoxanthin, 25mM HEPES (Sigma) and 5% haematocrit according to standard procedures (Trager & Jensen 1976).

Mixed stage HB3 Parasites (5-10%) were pre-synchronised 2 cycles before sample collection using two sorbitol synchronisations 10h apart in cycle -2. Sorbitol synchronisations were performed as follows (Lambros & Vanderberg 1979): The parasite culture was centrifuged for 5 min at 1900 rpm; The pellet (with a concentration of 10^7 RBC/μl) was resuspended in 6 volumes of 5% sorbitol (Sigma) and incubated at 37°C for 5 min. After 5 min centrifugation at 1900 rpm the supernatant was removed. The pellet was resuspended in culture medium. Parasites were expanded in cycle -1 to yield a parasitaemia of about 5%. At the end of cycle (-1) late stage parasites (44h-48h) were percoll separated. (Radfar et al. 2009): Separated schizonts were pooled and seeded in fresh dishes containing 1.5ml RBCs (10^7 RBC/μl) and 30ml fresh culture medium. These plates were incubated for 4 h to allow re-invasion. To reduce double infections, the plates were placed on an orbital shaker (200rpm) at 37°C. After 4h the cultures were pooled and synchronised using sorbitol (synchronisation window 4h). The ring stage culture was equally split into 14 30ml dishes (5% haematocrit) and incubated at 37°C. Per time-point two dishes of parasites culture were harvest 8h, 15h, 24h, 33h and 48h after percoll separation. After centrifugation 1.2 ml of infected RBCs (10^7 RBC/μl) and 20μl of viral HIV RNA were combined with 8ml Ribozol and frozen at -80°C.

Additionally, highly pure ring and schizont stage samples were produced to reduce the fraction of other stages to a minimum. To remove mature stage parasites that could contaminate the sample of "pure ring stage", two dishes of highly synchronized 8h parasites culture were purified using a MACS CS magnetic column (Miltennyi Biotec) with a flow resistor 22G (flow rate 3.5 ml/min) attached. The flow-thru (containing the non-magnetic ring stages) was collected and centrifuged. 1.2 ml of the cell pellet was lysed in 8 ml Ribozol and frozen at -80°C. For removal of potential ring stages (non-magnetic) that could contaminate the sample of "pure schizont stages", two dishes of the highly synchronized 48h parasites culture were additionally purified using a MACS CS column with a flow resistor 21G (flow rate 4 ml/min) attached. The flow through was discarded and the column washed with 30ml culture medium. The column was removed from the magnet and the magnetic schizonts were eluted with 20ml culture medium. The collected infected RBCs were centrifuged and supplemented with 3ml packed uninfected RBCs to generate the same conditions during RNA extraction as for other time-points. 1.2 ml of packed cells were added to 8ml Ribozol and frozen at -80°C.

RNA extraction from in vitro cultured P. falciparum strain HB3 and RNA quantification experiments

Total RNA was isolated and purified using Ribozol (Amresco) and RNeasy Kit (Qiagen) according to the manufacturer's protocol. Genomic DNA (gDNA) was removed by DNase digestion with Ambion DNase I Kit (ThermoFisher). RNA samples were tested for gDNA contamination by two *P. falciparum* specific qPCR assays that target 18S rRNA or varATS genes using a StepOne Plus Real-Time PCR System (Applied Biosystems) (Hofmann et al. 2015). Total RNA concentration was measured on Nanodrop (Thermo Fisher). Concentration of spiked-in HIV RNA, human gene *β-globin*, gene *pfs25* and gene *pfpk4* in the extracted RNA was quantified in triplicate by qRT-PCR as described previously (Labhardt et al. 2016; Brancucci et al. 2014; Wampfler et al. 2013; Irenge et al. 2005). Composition of reaction mixes and thermocycler conditions in Table S1.

Experimental mixtures of synchronized developmental stages of P. falciparum strain HB3

To adjust for loss of RNA during the extraction process (due to different amounts of total RNA used in extractions and possible saturation of extraction columns), the same amount of viral RNA was added into each sample before RNA extraction, permitting to restore the original total RNA concentration for a comparison of changes in RNA concentration between time-points. Sample 8h and 33h were diluted to restore the same concentration of spiked-in HIV RNA as in sample 48h. Experimental mixtures were prepared according to Table S2. Sample 8h represents ring (R) stages, sample 33h trophozoite (T) stages and sample 48h schizont (S) stages.

Counts of P. falciparum development stages

Parasites of each developmental stage of *P. falciparum* were quantified by light microscopy on Giemsa stained slides and FACS counting. Giemsa stained slides were prepared from a smear of 3-5µl of each *P. falciparum* culture. The smear was air dried and fixed for 2 min in 100% methanol. After fixation slide was transferred to a 10-15% Giemsa staining solution for 15-20 min. Slides were scanned with a Zeiss Axio Scan.Z1 Slide Scanner (Carl Zeiss GmbH, Jena, Germany), with a 20x objective for Giemsa stained slides.

For FACS analysis, 50 *µ*l of *P. falciparum* culture were spun and re-suspended in 100 *µ*l SYBR Green I nucleic acid gel stain (Sigma-Aldrich) diluted 1:5000 in parasite culture medium (PCM), incubated for 20-30 min at 37°C (in the dark) and washed 3 times in 1 ml PCM. 1.5 *µ*l of stained cells were transferred to 1ml FACS flow, vortexed and analysed by flow cytometer BD FACS Calibur (BD Biosciences). Cell Quest Pro Software was used to determine parasitaemia and stage counts (Figure S1).

High throughput sequencing

RNA-seq libraries were prepared using the Illumina TrueSeq Stranded mRNA Library Preparation Kit. Libraries were sequenced on a HiSeq 2500 125 cycle single read with added Illumina PhiX Control. Sequence reads were mapped with tophat (Version 2, parameters: read-mismatches=4, read-edit-dist=4 min-intron-length=10, max-intron-length=10000, max-multihits=1) to *P. falciparum* 3D7 reference sequence (PlasmoDB, release 11) (Trapnell et al. 2009; Bahl et al. 2003). Raw read counts were extracted with htseq-count (parameters: stranded=no type=gene idattr=gene_id mode=intersection-nonempty) using gene annotation (PlasmoDB, release 11) (Anders et al. 2015; Bahl et al. 2003).

Normalisation of RNA-Seq and qRT-PCR data

Raw read counts were normalised with different methods. Normalisation by counts per million (CPM) was used to represent the RPKM method. TMM and RLE normalisation was performed by using the R package edgeR (Robinson et al. 2010). In short, TMM and RLE normalisation factors were calculated with the 'calcNormFactor' function and multiplied by total sequence library size. Using the function 'cpm' normalised read counts were obtained from the adjusted library sizes. Normalisation with housekeeping gene *pk4* (PF3D7_0628200) annotated as "eukaryotic translation initiation factor 2-alpha kinase" was performed by dividing raw read count by *pk4* read count and multiplied by the mean *pk4* read count of all samples. Finally, BSN normalisation was performed as previously described (Aanes et al. 2014). In short, CPM normalised counts were multiplied by a biological scaling factor and the mean library size. Three different biological scaling factor were used for BSN normalisation in this study: the relative gene expression of housekeeping gene *pk4* to spiked-in viral RNA, both measured by qRT-PCR for $\mathrm{BSN}_{\mathrm{Bio}}$, and TMM or RLE normalisation factors for $\mathrm{BSN}_{\mathrm{TMM}}$ or $\mathrm{BSN}_{\mathrm{RLE}}$ respectively.

Gene expression of gene *pfs25*, *pk4*, and *β-globin* by qRT-PCR was normalised by calculating the gene expression relative to the spiked-in viral RNA. To compare normalised gene expression by RNA-Seq and qRT-PCR, fold change of genes *pfs25*, *pk4*, and *β-globin* was calculated by dividing gene expression of each individual time-point by mean gene expression over all time-point samples of that gene. Fold changes were compared by Spearman correlation.

Differential gene expression (DGE) analysis

The R package edgeR (Robinson et al. 2010) was used to identify differentially expressed genes in R, T and S stages represented by the 8h, 33h and 48h time-course samples. Normalisation was performed with CPM, TMM, PK4 and $\mathrm{BSN}_{\mathrm{RLE}}$ methods. PK4 and $\mathrm{BSN}_{\mathrm{RLE}}$ normalisation was not implemented in edgeR. Therefore, PK4 and $\mathrm{BSN}_{\mathrm{RLE}}$ normalised read counts were used as input for DGE analysis and no further normalisation was performed during DGE analysis. Dispersion of each gene (tagwise dispersion) was estimation based on time-course and highly pure stage samples to account for missing replicates. R stage was represented by the two samples, "highly pure ring stage" and "8h" samples, T stage by "24h" and "33h" samples, and S stage by "highly pure schizonts stage" and "48h" samples. The model was fit with function glmFit and a design matrix including an intercept, followed by a likelihood ratio test using function glmLRT. Genes with a significant differential gene expression were identified with function decideTestsDGE.

Deconvolution of mixed-stage samples

Mixed-stage transcriptomes were deconvoluted into stage-specific signatures based on known stage compositions of the samples. R Package CellMix (http://web.cbio.uct.ac.za/~renaud/CRAN/web/CellMix) was used for csSam, csLsfit and csQprog deconvolution and function lm from R package stats for deconvolution by linear regression. Deconvolution by negative binomial regression was performed with R package edgeR with a design matrix containing the stage proportions of samples and estimated tagwise dispersion. Impact of fold increase of RNA in mixed stage samples was visually inspected by multidimensional scaling analysis with function plotMDS from R package edgeR (Figure S2). Significance of gene expression fold changes was tested by 400 permutations of the stage proportions of mixed-stage samples. P-values were estimated with permp function of R package statmod (Phipson & Smyth 2010) and adjusted for multiple testing by calculating the false discovery rate (FDR) with function ‘p.adjust’ of R package stats (Benajmini & Hochberg 1995). Significant fold changes of deconvoluted R_{Est}, T_{Est} and S_{Est} stage transcriptom with fold change >1 were compared to fold changes of R, T and S stage samples by calculating the Pearson correlation factor R^2 and percentage of up or down regulation agreement. Comparison was performed on $\log_2$ transformed inferred stage-specific transcriptomes with an added prior of 0.5 to avoid taking log of zero.

Results

Fold increase in total RNA and parasite genomes during intra-erythrocytic development cycle

The amount of total RNA in *P. falciparum* parasites was reported to increase substantially during the intra-erythrocytic development cycle (IDC) (Bártfai et al. 2010; Sims et al. 2009; Kensche et al. 2016). To estimate fold increase in total RNA of a highly synchronised *P. falciparum* culture, equal amounts of viral RNA was added to each sample before RNA extraction. After RNA extraction and purification, concentrations of the extracted RNAs were adjusted based on the measured viral RNA concentration to restore the true RNA concentration prior to losses of RNA during the extraction procedure. Total RNA concentration increased by 18-fold during IDC, respectively between 8h and 48h samples (Table 2).

Normalisation and differential gene expression (DGE) of time-course samples

Time course samples were used to evaluate different normalisation methods that best fitted the fold change profile of genes *pk4*, *pfs25*, and human *beta-globin* measured by qRT-PCR and adjusted according to spiked-in viral RNA. BSN_{Bio} normalisation correlated best to the fold change profile measured by qRT-PCR (spearman correlation R^2 in decreasing order 0.96, 0.81, 0.65, 0.52 and 0.46 for BSN_{Bio}, PK4, CPM, TMM and RLE) (Table 3, Figure 1 left panel). Normalisation by the empirically selected gene *pk4* (PF3D7_0628200) achieved second best correlation. *Pk4* is known to be relative constantly expressed and had been used in an earlier publication as a reference/housekeeping gene (Brancucci et al. 2014). CPM, TMM and RLE normalisation were not able to adjust for fold increase of total RNA. The biological scaling factor (relative gene expression of *pk4* to spiked-in viral RNA, both measured by qRT-PCR) used for the BSN_{Bio} normalisation cannot be used for mixed stage field samples. As an alternative, TMM or RLE scaling factor can be used for the BSN normalisation (Aanes et al. 2014). BSN_{RLE} and BSN_{TMM} correlated better to the fold change profile than TMM and RLE normalisation, but less good than BSN_{Bio} (spearman correlation R^2 in decreasing order 0.83 and 0.81 for BSN_{RLE} and BSN_{TMM}) (Table 3, Figure 1 right panel).

Differentially expressed genes were identified by DGE analysis of the 8h, 33h and 48h samples representing ring (R), trophozoite (T) and schizont (S) stages. The number of significant differential expressed genes (FDR<0.05) depended on the normalisation method used and was higher for BSN_{RLE} and PK4 normalisation than for CPM and TMM normalisation (Table 4).

Evaluation of various deconvolution methods based on experimental mixtures

Experimental mixtures of highly synchronous cultures were used to determine the best suited normalisation and deconvolution method for estimation of stage-specific signatures of gene expression from mixed-stage samples with known proportion. Only genes showing a significant inferred fold change (p-value <0.05) and with fold change >1 between ring and trophozoite stage (R-T), ring and schizont stage (R-S) or trophozoite and schizont stage (T-S) were used for comparison of the normalisation and deconvolution methods (Table 5).

Deconvolution by csSam, csLsfit and csQprog resulted in almost identical inferred signatures (Figure S3). Therefore, method csQprog was chosen as representative for those methods for further analysis, csQprog includes a non-negative constraint of estimated gene expression signatures. Normalisation methods CPM, PK4 and BSN_{RLE} gave identical results when applied in combination with the deconvolution method csQProg (Table 5, 6 and 7).

All normalisation and deconvolution methods yielded Pearson correlation of inferred and measured signatures for genes with a significant fold change in the range of 0.78 and 0.95 for the ring stage signature, 0.89 and 0.99 for the trophozoite stage, and 0.72 and 0.95 for the schizont stage (Table 6). The choice of normalisation method most affected the inferred ring and schizont signatures. The highest correlation coefficient for inferred ring and schizont signatures was observed by the combination of methods PK4 and edgeR (Table 6). The combination TMM-csQprog showed the lowest correlation coefficient for inferred schizont signatures. For inferred ring signatures the combination BSN_{RLE}-glm seems least suitable.

Irrespective of deconvolution and normalization method, the comparison of inferred and measured fold changes showed that Pearson correlation was highest for the T-S fold change (R^2 between 0.91 and 0.93). For inferring the R-T and R-S changes, no good correlation was obtained with deconvolution methods csQprog and lm (R^2 between 0.34 and 0.53 for R-T and between 0.10 and 0.62 for R-S) (Table 6). Deconvolution method edgeR achieved the best agreement in up or down regulation for genes with a significant fold change between R-T and R-S (R^2 between 0.85 and 0.87 for R-T and between 0.91 and 0.94 for R-S) (Table 6 and 7, Figure 2). When deconvolution method edgeR was used, the choice of normalisation method had little impact on results. However, normalisation with CPM and PK4 resulted in slightly higher agreements for all possible fold changes.

Discussion

Most of the methods available for deconvolution of samples of heterogeneous composition were developed for gene expression data from microarray platforms and apply a linear model for deconvolution. Microarray data measures gene expression as fluorescence intensity, whereas RNA-Seq data counts the reads observed (Robinson et al. 2010). Therefore, an overdispersed Poisson model, e.g. negative binomial regression, is more appropriate for RNA-Seq deconvolution. However, the linearity of the data is no longer maintained after log-transformation conducted during the negative binomial regression (Zhong & Liu 2012). This dilemma could not be circumvented and was not resolved in this study.

Deconvolution of the transcriptomes of mixed *Plasmodium* developmental stages is particularly challenging as total RNA increases 18-fold during the 48 h blood stage cycle. An increase of total RNA is expected because the intracellular parasite undergoes several rounds of mitosis and by the end of the cycle has replicated into 16 to 32 merozoites. Unfortunately, no additional material remained from this study that could be used for DNA extraction to precisely measure increase of genomes by qPCR. Therefore, it remains unclear whether the increase of total RNA and was proportional to the growing number of genomes during the time course.

A change in total RNA, as the observed in this experiment, caused by very high expression of a number of genes, may cause a bias in the measurement of gene expression. During RNA-Seq, samples are pooled at equal molarity to achieve similar amount of total sequence reads for each sample. A large amount of upregulated genes at one of the time-points could have consumed a substantial proportion of the total number of sequenced reads and caused under-sampling of the remaining genes (Robinson & Oshlack 2010). In this case the median gene expression might appear higher for time-course samples owing to their increased total RNA, whereas low expressed genes might be missed. If this sampling artefact is not accounted for by normalisation, some genes falsely might appear to be downregulated. On the other side, if the wrong normalisation method is chosen, incorrect expression patterns can be generated.

Our evaluation of different RNA-Seq normalisation methods showed that the commonly used methods, e.g. TMM, RLE and CPM, were unable to adjust for the total RNA increase. These normalisation methods assume that most genes are not differentially expressed. However, during the IDC of *P. falciparum* most genes are differentially expressed. BSN_{Bio} normalisation uses a biological scaling factor that represent the differences in RNA concentration between samples. BSN_{Bio} offered the best adjustment for total RNA increase. However, BSN_{Bio} normalisation cannot be used for field samples as spike-in of viral RNA is not possible. The second best option was normalisation with housekeeping gene PK4. PK4 normalisation can always be applied to RNA-Seq data. PK4 is a highly expressed housekeeping gene and has been used before to normalise qRT-PCR gene expression data (Brancucci et al. 2014). Normalisation by only a single housekeeping gene carries the risk of an unstable normalization. Multiple empirically selected housekeeping genes expressed at different levels, followed by RUV-Seq normalisation, might be a better choice for normalisation (Risso et al. 2014), but this approach depends on the availability of well validated empirically selected housekeeping genes.

Comparison of DGE analysis of R, T and S stages with different normalisation methods indicated that BSN_{RLE} or PK4 normalisation identified approximately 1.6-fold more genes showing significant differential gene expression compared to normalisation by CPM or TMM (1814 or 1642, versus 1030 or 1019 significant genes). In our study the number of genes with significant differential expression might have been underestimated for all normalisation methods, because no biological replicates were available to estimate dispersion. Instead, dispersion was estimated by combining the samples of pure stages with samples from time course experiments. This approach to estimate dispersion likely overestimates dispersion and as a consequence would reduce the number of genes showing significant differential expression.

One schizont parasite carries on average as much RNA as 18 ring stage parasites (Figure S2). As a consequence, in mixed-stage samples, the 18-fold increase in total RNA resulted in a predominance of late stage parasites in the observed mixed transcriptome. Additionally, increase in total RNA can lead to a non-linear mixing of transcripts in the sample.

The comparison of normalisation and deconvolution with several methods showed that the combination of edgeR and CPM are best suited to infer stage-specific transcriptomes. This combination showed highest agreement between significant fold changes of inferred and measured stage-specific genes. The different normalisation methods only weakly influenced the outcome of the deconvolution with edgeR. This result was unexpected. The comparison of time-course samples to gene expression measured by qRT-PCR indicated that the BSN_{Bio} normalisation method best reflected gene expression of RNA-Seq data. An explanation for this results could be that the log transformation of the observed transcriptomes (during negative binomial regression of edgeR) reduces the effect of an 18-fold increase in total RNA. This would also explain why in our analyses the normalisation method mattered when deconvolution was performed with method csQprog or lm. Another explanation why edgeR performs better than csQprog or lm is that log transformation of RNA-Seq data leads to better homogeneity of variance, and consequently to a more accurate estimation of variance during permutation testing.

Permutation testing allowed robust identification of successfully deconvoluted genes i.e. genes that were indeed differentially expressed among stages. However, the 400 permutations performed in this study likely were insufficient for the comparison of >5000 genes. This shortcoming became apparent after adjusting p-values for multiple testing by calculating false discovery rate (FDR). Much fewer significant genes with FDR<0.05 could be identified (Table S3). More permutations were not possible for the complex comparison of five deconvolution methods and four normalisation methods. The number of permutation should be greatly increased for future analysis when only one normalisation and deconvolution method combination is performed. And consequently, only genes with a FDR<0.05 should be selected for identification of stage specific genes.

Conclusion

By comparing observed stage specific transcriptomes of highly synchronized *P. falciparum* cultures to inferred stage-specific transcriptomes of mixed-stages of the same samples, this study showed that deconvolution of stage-specific transcriptomes is feasible. The 18-fold increase in total RNA between rings and schizonts proofed to be a particular challenge for accurate deconvolution. Best approach for deconvolution of mixed developmental stages of malaria parasites was deconvolution with method edgeR and CPM normalisation. Genes with a fold-change large enough to be successfully deconvoluted could be identified by permutation testing

This deconvolution approach can be transferred to field samples of any *Plasmodium* species with known stage proportions. This proof-of-concept study paves the way for inferring gene expression of *P. vivax* gametocytes from field samples, under the condition that the proportions of developmental stages of the parasite in the sample is known.

Table 1: Overview of deconvolution methods.

Method name	**Estimates**		**Description/ Comments**	**Reference**
	Signature	**Proportions**		
lsfit	yes	yes	Least-squares	Abbas 2009
qprog	yes	yes	Quadratic programming	Gong 2010
DeconRNASeq	no	yes	Quadratic programming	Gong 2013
csSAM	yes	yes		Shen-Orr2010
lm and qprog	yes	yes	Method 'lm' for signature and 'qprog' for proportions	Joice 2013
DSection	yes	(no)	MCMC	Erkkila2010
PERT	no	yes	NNLS with LDA	Qiao 2012
DSA	yes[1]	yes[1]	Requires marker genes	Zhong 20013
ssKL	yes[1]	yes[1]	Requires marker genes	Gaujoux 2011
ssFrobenius	yes[1]	yes[1]	Requires marker genes	Gaujoux 2011
deconf	yes[1]	yes[1]	Alternating NNLS No longer running in R	Repsilber 2010
TEMT	yes	no	probabilistic model-based Supports mixture of two cell types	Li 2013
DeMix	yes	yes	SW no longer accessible	Ahn 2013
ISOpure	no	yes	Use signature to estimate cancer profile and cell-type composition	Quon 2013
-	no	yes	SW no longer accessible	Clarker 2010
CIBERSORT	no	yes	Robust linear regression and ν-SVR	Newman 2015
DCQ	no	yes	Elastic net Digital Cell quantification	Altboum 2014
PSEA			No implementation available. Uses 'lm' method needs marker gene	Kuhn 2011
ISOLATE	no	yes	LDA Estimates cancer signature and cell-type composition	Quon2009
xCell	no	yes	Application developed specifically for human gene expression can be easily transferred to Plasmodium species	Aran 2017

NNLS Non-negative least squares

NNML Non-negative maximum likelihood model

LDA Latent Dirichlet Allocation

[1] Complete deconvolution

Table 2: Total RNA concentration and fold increase over time course of extracted RNA from highly synchronized *P. falciparum* cultures spiked with equal quantities of viral RNA. Ring (R), trophozoite (T), and schizont (S) stage sample were used for experimental mixed stage transcriptomes.

Sample	Total RNA ng/μl	Spiked-in viral RNA[1] copies/μl	Adjusted total RNA[2] ng/μl	Total RNA fold increase
R / 8h	138	8.0×10^5	42.9	1
15h	110.9	4.5×10^5	60.9	1.4
24h	445.7	8.2×10^5	135.9	3.2
T / 33h	645.1	4.1×10^5	388.8	9.1
S / 48h	801.4	2.5×10^5	801.4	18.7

[1] Concentration of spiked-in viral RNA by qRT-PCR after RNA extraction.

[2] Adjusted total RNA concentration to restore equal concentration spiked-in viral RNA in each sample.

Table 3: Comparison of normalisation methods between fold change of gene expression measured by qRT-PCR and RNA-Seq for 3 genes at 5 time-points.

Normalisation Method	Correlation Spearman *rho*	Concordance Cohn's *Kappa* (p-value)
CPM	0.65	0.42 (8.1×10^{-3})
TMM	0.52	0.36 (0.024)
RLE	0.46	0.32 (0.044)
BSN_{Bio}	0.96	0.78 (9.7×10^{-7})
PK4	0.81	0.55 (3.6×10^{-3})
BSN_{TMM}	0.81	0.57 (2.7×10^{-4})
BSN_{RLE}	0.83	0.62 (8.9×10^{-5})

Table 4: Comparison of different normalization methods for differential gene expression analysis of RNA-Seq data.

Normalisation Method	Differential expressed genes [1] n_{sig}	Fold Change					
		R-T [2]		R-S [3]		T-S [4]	
		$n_{fc>1}$	$n_{fc<-1}$	$n_{fc>1}$	$n_{fc<-1}$	$n_{fc>1}$	$n_{fc<-1}$
CPM	1019	287	87	250	132	103	250
TMM	1030	211	121	101	337	151	229
PK4	1642	381	106	762	308	228	478
BSN_{RLE}	1814	356	121	822	353	314	621

[1] 5308 genes were included in this analysis.

[2] Fold change between ring and trophozoite stage.

[3] Fold change between ring and schizont stage.

[4] Fold change between trophozoite and schizont stage.

n_{sig} Number of genes with a significant fold change (FDR<0.05).

$n_{fc>1}$ Number of genes with significant fold change and with fold change > 1 (up regulated).

$n_{fc<-1}$ Number of genes with significant fold change and with fold change < -1 (down regulated).

Table 5: Number of genes with a significant fold change (p-value<0.05) after deconvolution and with fold change >1 for 3 different deconvolution method methods.

Deconvolution Method	Normalisation Method	Significant genes [1] n_{sig}	R-T [2]		R-S [3]		T-S [4]	
			$n_{fc>1}$	n_{sig}	$n_{fc>1}$	n_{sig}	$n_{fc>1}$	n_{sig}
csQprog	CPM	3959	1110	1148	1036	1153	1336	3728
	TMM	4197	778	873	1422	1697	1842	3924
	PK4	3959	1110	1148	1036	1153	1336	3728
	BSN_{RLE}	3959	1110	1148	1036	1153	1336	3728
lm	CPM	3720	1023	1044	1073	1138	1440	3566
	TMM	3720	1023	1044	1073	1138	1440	3566
	PK4	4079	511	769	943	1555	1608	3559
	BSN_{RLE}	3529	1107	1115	592	626	922	3322
edgeR	CPM	4059	179	275	538	1397	1314	3769
	TMM	4259	128	308	594	1869	1792	3932
	PK4	4023	195	262	545	1109	1304	3733
	BSN_{RLE}	4071	180	288	534	1388	1319	3784

[1] Total of 5308 genes were included in the analysis.

[2] Fold change between ring and trophozoite stage.

[3] Fold change between ring and schizont stage.

[4] Fold change between trophozoite and schizont stage.

n_{sig} Number of genes with a significant fold change (p-value<0.05) by permutation test.

$n_{fc>1}$ Number of genes with significant fold change and with fold change>1.

Table 6: Pearson correlation of genes with significant fold change (p-value <0.05) after deconvolution and with a fold change >1.

Deconvolution	Normalisation	Signature			Fold change		
Method	Method	R	T	S	R-T [1]	R-S [2]	T-S [3]
csQprog	CPM	0.80	0.99	0.84	0.47	0.28	0.93
	TMM	0.83	0.99	0.72	0.44	0.36	0.92
	PK4	0.80	0.99	0.84	0.47	0.28	0.93
	BSN_{RLE}	0.80	0.99	0.84	0.47	0.28	0.93
lm	CPM	0.82	0.99	0.76	0.46	0.26	0.92
	TMM	0.82	0.99	0.76	0.46	0.26	0.92
	PK4	0.82	0.99	0.80	0.53	0.62	0.93
	BSN_{RLE}	0.78	0.98	0.92	0.34	0.10	0.91
edgeR	CPM	0.94	0.90	0.94	0.87	0.91	0.91
	TMM	0.93	0.90	0.94	0.86	0.91	0.92
	PK4	0.95	0.89	0.95	0.85	0.94	0.92
	BSN_{RLE}	0.94	0.90	0.94	0.87	0.92	0.91

[1] Fold change between ring and trophozoite stage.

[2] Fold change between ring and schizont stage.

[3] Fold change between trophozoite and schizont stage.

Table 7: Agreement among 3 analysis methods in up or down regulation of genes with significant fold change (p-value <0.05) and with a fold change >1.

Deconvolution	Normalisation	R-T [1]		R-S [2]		T-S [3]	
Method	Method	n	%	n	%	n	%
csQprog	CPM	686	61.4%	877	82.7%	3485	99.4%
	TMM	664	84.3%	1234	82.9%	3538	99.7%
	PK4	686	61.4%	877	82.7%	3485	99.4%
	BSN_{RLE}	686	61.4%	877	82.7%	3485	99.4%
lm	CPM	644	62.8%	890	81.4%	3436	99.7%
	TMM	738	71.9%	940	85.9%	3429	99.5%
	PK4	470	84.2%	1120	88.6%	3348	99.6%
	BSN_{RLE}	559	50.3%	532	88.4%	3223	98.2%
edgeR	CPM	239	96.8%	1318	99.5%	3527	99.5%
	TMM	203	95.8%	1624	98.9%	3540	99.5%
	PK4	226	96.6%	1032	99.5%	3493	99.5%
	BSN_{RLE}	251	96.5%	1307	99.5%	3535	99.4%

[1] Fold change between ring and trophozoite stage.

[2] Fold change between ring and schizont stage.

[3] Fold change between trophozoite and schizont stage.

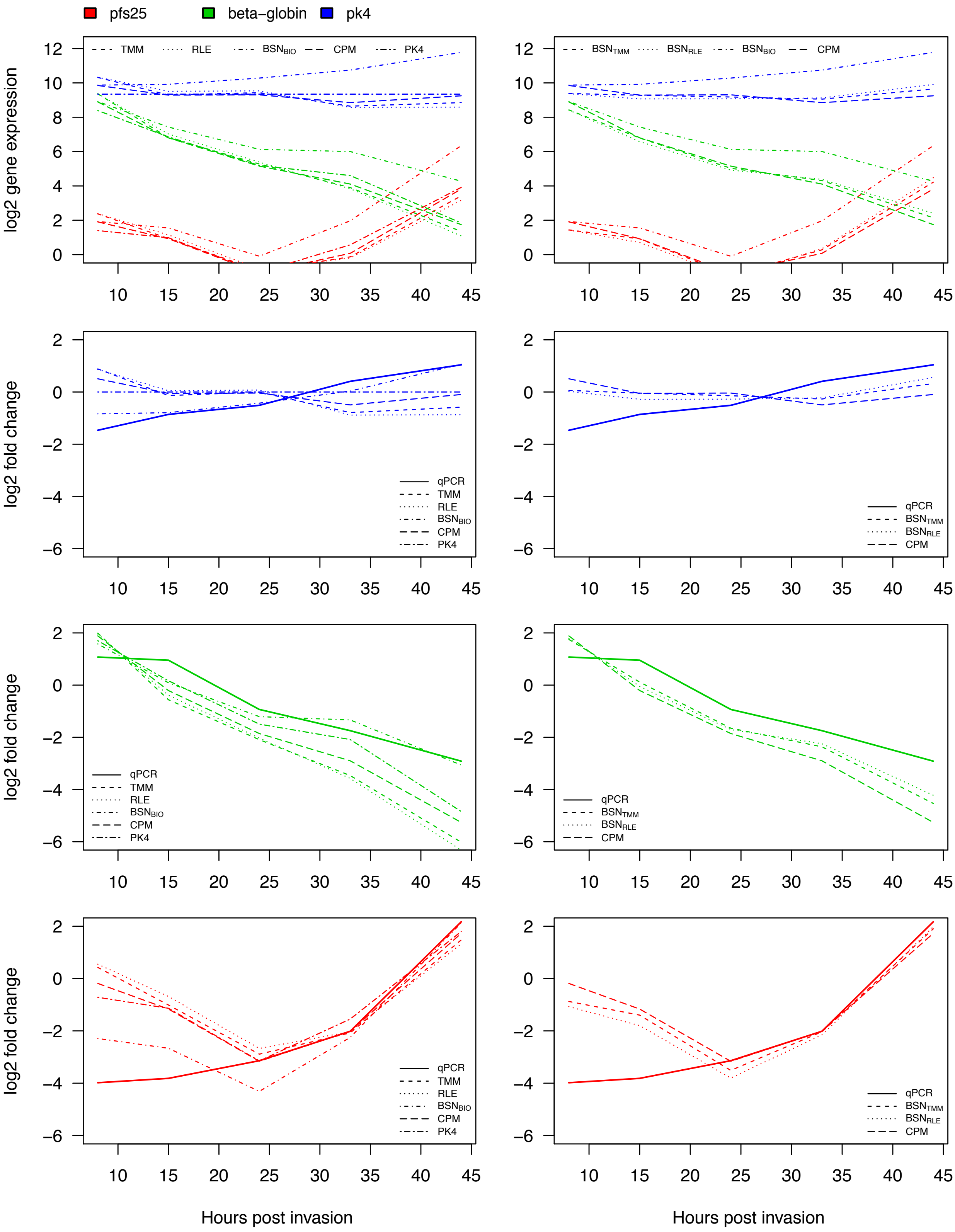


Figure 1: Effects of different normalisation methods on gene expression of *pfs25*, *pk4* and beta-globin measured by RNA-Seq in comparison to expression levels by qRT-PCR. For better overview, different normalisation methods were split on left and right panels. CPM and BSN$_{Bio}$ Normalisation are found on both sides for better comparison. Top panels: gene expression in log$_2$ count per million. Other panels: fold change relative to average gene expression.

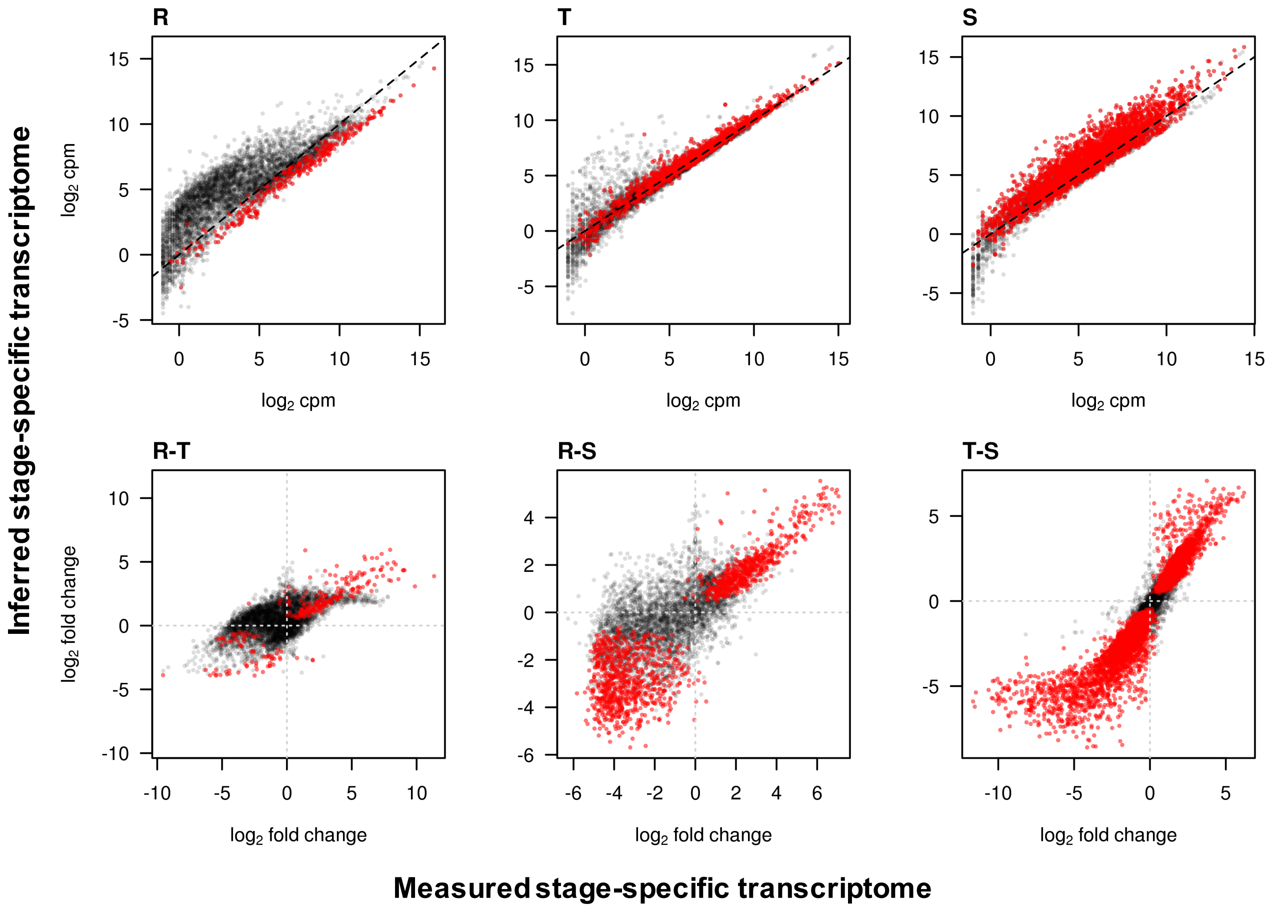


Figure 2: Inferred transcriptomes and fold changes by deconvolution method edgeR. Top panel, ring (R), trophozoite (T) and schizont (S) stage gene expression in counts per million (cpm). Bottom panel, fold changes between R and T (R-T), between R and S (R-S) and between T and S (R-S). X-axis, observed stage-specific transcriptomes or fold changes of stage. Y-axis, inferred stage-specific transcriptomes or fold changes. Red points represent genes with significant (p-value<0.05) signature or fold changes.

Supplemental Material

Table S1: qPCR reaction mix and cycle conditions.

	HIV	Pfs25	PK4	PfS18S	varATS	β-globin
Reaction Mix:						
RNAse free H_2O	6 µl	6 µl	6 µl	2.5 µl	2.5 µl	2.5 µl
Taqman RT buffer (2x)	10 µl	10 µl	-	-	-	-
Taqman RT Enzym (40x)	0.5 µl	0.5 µl	-	-	-	-
Power SYBR® Green mix	-	-	10 µl	-	-	-
GeneEx buffer (2x)	-	-	-	6 µl	6 µl	6 µl
Primer mix (10uM)	1 µl	1 µl	2 µl	1 µl	1 µl	1 µl
Probe (10uM)	0.5 µl	0.5 µl	-	0.5 µl	0.5 µl	0.5 µl
RNA Template (1:10)	2 µl	2 µl	-	2 µl	2 µl	-
cDNA Template (1:10)	-	-	2 µl [1]	-	-	2 µl [1]
Cycle condition:						
Pre-incubation	48 °C, 15 min	48 °C, 15 min	50 °C, 2 min	50 °C, 2 min	50 °C, 2 min	50 °C, 2min
Initial denaturation	95 °C, 10 min	95 °C, 10 min	95 °C, 10 min	95 °C, 10 min	95 °C, 10 min	95 °C, 10min
Denaturation	95 °C, 15 s	95 °C, 15 s	95 °C, 15 s	95 °C, 15 s	95 °C, 15 s	95 °C, 15s
Annealing & Elongation	60 °C, 1 min	58 °C, 1min	58 °C, 1 min	58 °C, 1 min	55 °C, 1 min	55 °C, 1min
Number of cycles	45x	45x	40x	45x	45x	45x
Melt Curve	-	-	95 °C, 15 s 58 °C, 1 min	-	-	-
Standard curve:						
Template	HIV cDNA	Plasmid	3D7 cDNA	Plasmid	3D7 cDNA	HB3 cDNA [1]
Unit	copies/ul	copies/ul	parasites/ul	copies/ul	parasites/ul	arbitrary
From	10^5	10^6	10^5	10^6	10^5	10^5
To	1	10	1	10	1	10

[1] 10µl *P. falciparum* strain HB3 culture was reverse transcribed to 40µl cDNA and adjusted to 100µl.

Table S2: List of proportions used for experimental mixtures of ring, trophozoite, and schizont stage sample.

Sample Name	Ring stage sample	Trophozoite stage sample	Schizont stage sample
Ratio1	0.75	0.25	0.00
Ratio2	0.50	0.50	0.00
Ratio3	0.25	0.75	0.00
Ratio4	0.00	0.75	0.25
Ratio5	0.00	0.50	0.50
Ratio6	0.00	0.25	0.75
Ratio7	0.25	0.00	0.75
Ratio8	0.50	0.00	0.50
Ratio9	0.75	0.00	0.25
Ratio10	0.10	0.80	0.10
Ratio11	0.10	0.10	0.80
Ratio12	0.80	0.10	0.10

Table S3: Number of genes with a significant fold change (FDR[1] <0.05) and with fold change >1.

Deconvolution Method	Normalisation Method	Significant genes [2] n_{sig}	R-T [3]		R-S [4]		T-S [5]	
			$n_{fc>1}$	n_{sig}	$n_{fc>1}$	n_{sig}	$n_{fc>1}$	n_{sig}
csQprog	CPM	3564	0	0	259	268	1304	3503
	TMM	3872	0	0	853	911	1797	3752
	PK4	3564	0	0	259	268	1304	3503
	BSN_{RLE}	3564	0	0	259	268	1304	3503
lm	CPM	3260	0	0	0	0	1388	3260
	TMM	3260	0	0	0	0	1388	3260
	PK4	3235	0	0	0	0	1543	3235
	BSN_{RLE}	2159	0	0	0	0	765	2159
edgeR	CPM	3604	0	0	0	0	1272	3604
	TMM	3791	0	0	0	0	1744	3791
	PK4	3557	0	0	0	0	1261	3557
	BSN_{RLE}	3613	0	0	0	0	1272	3613

[1] False discovery ratio (FDR) calculated to adjust p-values for multiple testing.

[2] Total of 5308 genes were included in the analysis.

[3] Fold change between ring and trophozoite stage.

[4] Fold change between ring and schizont stage.

[5] Fold change between trophozoite and schizont stage.

n_{sig} Number of genes with a significant fold change (p-value<0.05) by permutation test.

$n_{fc>1}$ Number of genes with significant fold change and with fold change>1.

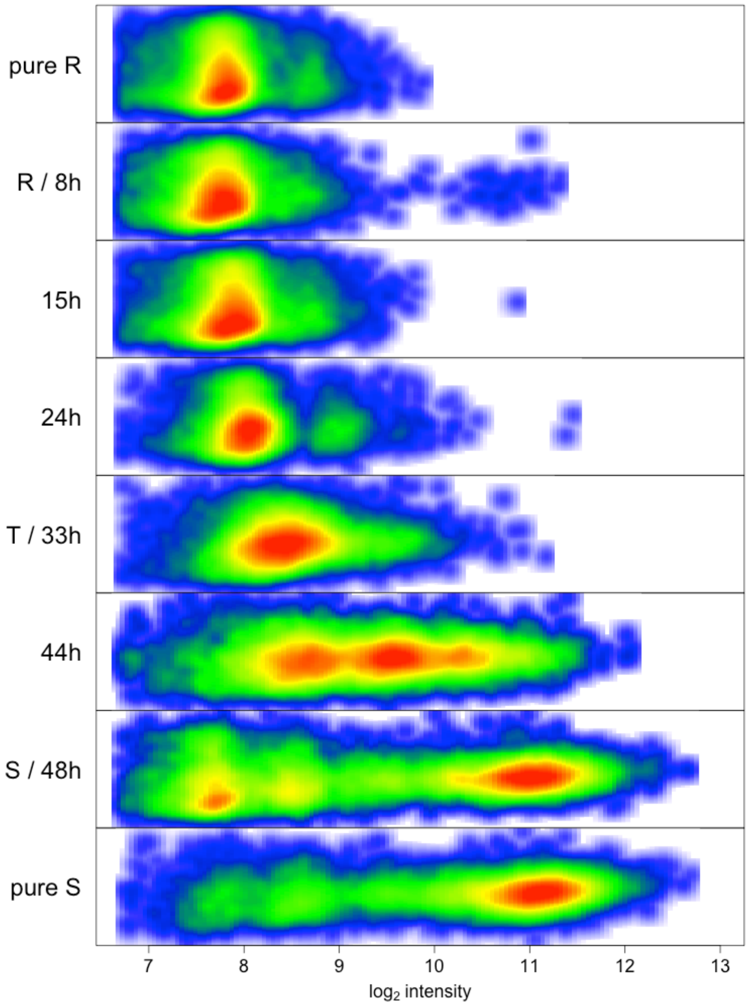


Figure S1: FACS counts of SYBR green stained time course and highly pure stage samples. X-axis, $\log_2$ fluorescence intensity. Y-axis, side scatter. Colours shows density distribution with red as high and blue low density.

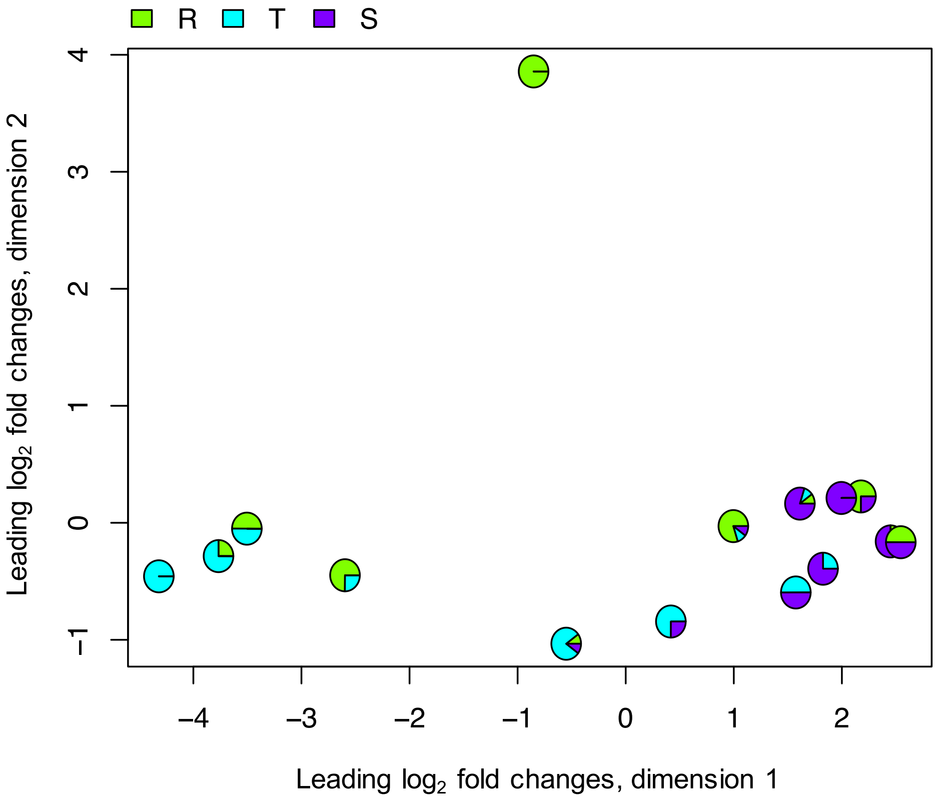


Figure S2: Multidimensional scaling plot showing predominance of late stage parasites and non-linear mixing of stage-specific transcriptomes. X and Y-axis, distance in $\log_2$ fold changes between gene expression profiles. Pies showing stage proportions of a sample.

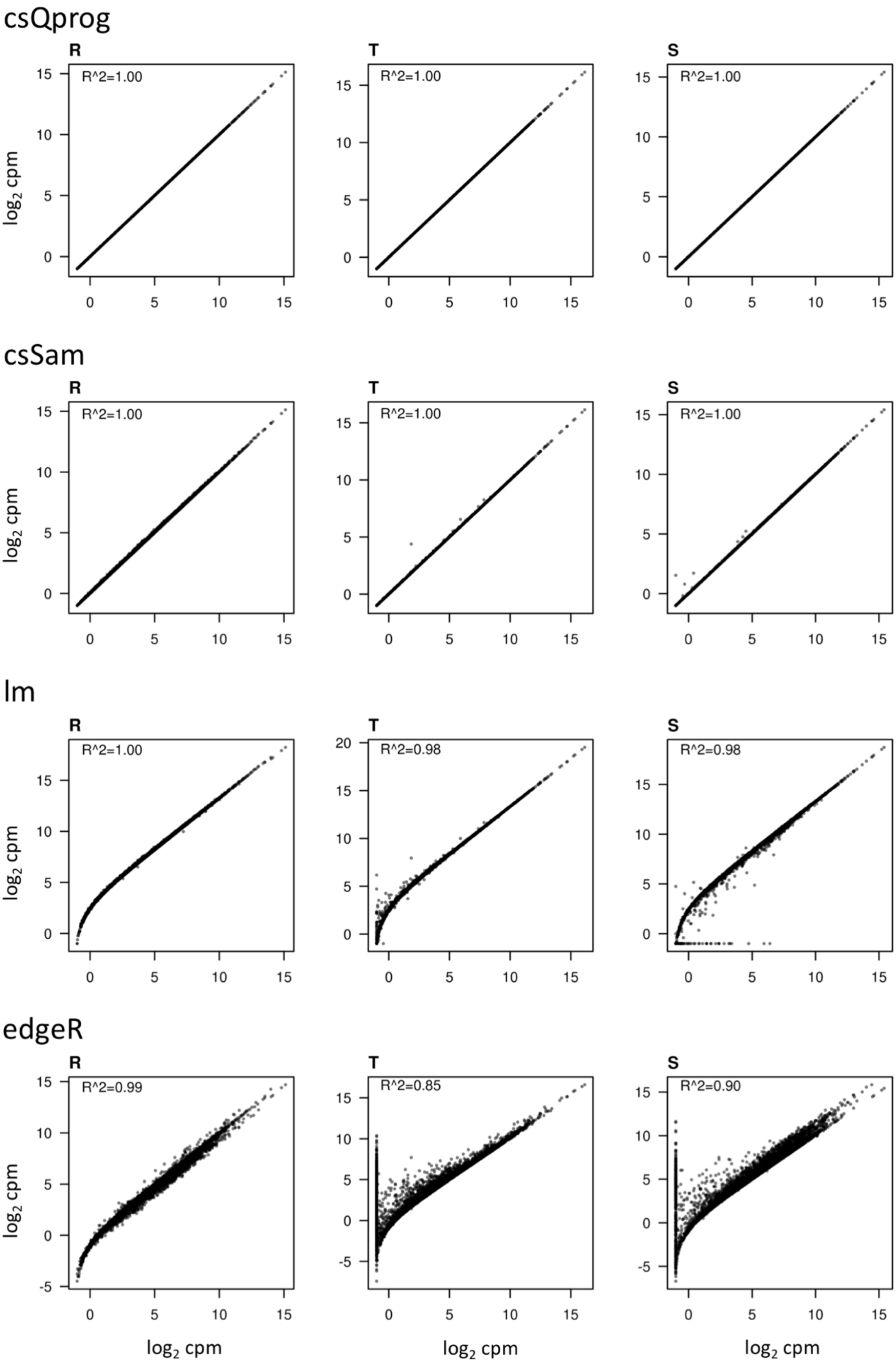


Figure S3: Inferred transcriptomes by different deconvolution methods and CPM normalisation method. X-axis, estimated coefficient of csLsfit deconvolution. Y-axis, estimated coefficient of method named in panel header. R, estimated ring stage signature (left panel). T, estimated trophozoite stage signature (middle panel). S, estimated schizont stage signature (right panel).

CHAPTER 5: GENERAL DISCUSSION

The overall aim of this thesis was to develop tools for analysis of deep sequencing data from mixtures of *P. falciparum* clones or stage-specific transcriptomes. In the first case, samples of individuals concurrently infected with several different parasite clones were analysed. Individual parasite clones in these samples were identified by Amp-Seq and subsequent clustering of sequence reads. The resulting within-host clone frequency was then used to infer multi-locus haplotypes and to estimate density of the clones. Identification of individual parasite clones are required for determination of molecular epidemiological parameters e.g. MOI, $_{\mathrm{mol}}$FOI, and duration of infection, which were then used to study the epidemiology of *Plasmodium* parasites. In the second case, samples of experimentally mixed-stage transcriptomes were analysed to infer stage-specific transcriptomes based on their known stage composition. Existing deconvolution and normalisation methods were evaluated to find the best approach to analyse such data. Comparison to stage-specific transcriptomes showed that genes with a significant fold change can be used to identify stage-specific genes in field samples. This thesis provides proof-of-concept analysis for inferring *P. vivax* gametocyte-specific-genes in samples enriched for mature stage parasites, but still containing different stages of asexual parasites.

This chapter discusses the overall findings and limitations of the analysis conducted, and gives directions for future research.

5.1 GENOTYPING BY AMPLICON SEQUENCING

Amp-Seq of SNP polymorphic markers is increasingly used for genotyping. High multiplexing of samples for sequencing enables the use Amp-Seq even for large epidemiological studies. Amp-Seq genotyping is superior to genotyping of length-polymorphic markers by capillary electrophoresis, because of its increased sensitivity to detect minority clones and quantification of concurrently infecting clones [1]. The higher sensitivity of Amp-Seq genotyping resulted in a significant increase of mean MOI. However, no significant increase of mean $_{\mathrm{mol}}$FOI could be determined in longitudinal samples. The sample size of this study was small, so it is possible that the non-significant increase of $_{\mathrm{mol}}$FOI could be due to insufficient power in statistical analysis. Alternatively, $_{\mathrm{mol}}$FOI might be less prone to detectability of minority clones than MOI, as each clone might be at a relatively high density at least once during an infection.

Duration of infection is a key epidemiological parameter. The impact of higher sensitivity to detect minority clones on the duration of infection could not be studied in this thesis, because persisting infections with *P. falciparum* were mostly cleared by treatment. The children in the cohort study were aged 1-5 years and suffered regularly from clinical attacks of *P. vivax* or *P. falciparum* infections requiring treatment. Thus, natural clearance of infections could not be studied. However, during analysis of longitudinal samples some parasite clones were discovered earlier and/or for a longer period if more relaxed cut-off criteria were used (Chapter 3 Table 3). This observation might indicate an increased measured duration of infection when a more sensitive method is used for genotyping. A study on individuals with higher levels of acquired immunity, and thus less treatment, would be better suited to study the impact of higher sensitivity to detect minority clones on duration of infection. An example of such a study was conducted in Navrongo, Northern Ghana, which includes individuals of all ages [2]. Duration of infection and $_{\mathrm{mol}}$FOI was already extensively studied in this cohort by *msp2*-CE genotyping [2–6]. Re-analysis of the Navrongo study with Amp-Seq genotyping would show if a more sensitive genotyping method impacts duration of infection.

Analysis of individual parasites clone density over time regardless of sample MOI allows investigation of whether the duration of infection depends on parasite densities in the blood. For example, it could be that long-lasting infections show a lower clone density during an infection compared to infections which are cleared within a week of appearance in the blood. It could also be that infections with the same, or similar clones as observed in previous infections show lower clone density as compared to infections of unrelated clones. Furthermore, patterns of clone density fluctuations over time could also be studied. For example, it is unknown whether in multi-clone infections the density of each clone peaks at the beginning of the infection and slowly decreases. Finally, fitness of individual parasite clones in the human host could be studied, i.e. the density of clones of a specific genotype, or between-clone-competition in multi-clone infections. In summary, the ability to measure individual clone density in multi-clonal infections opens the field for a new set of epidemiological studies.

5.1.1 Comparative analysis of Amp-Seq and CE genotyping methods

When *msp2*-CE genotyping replaced *msp2*-RFLP genotyping, the authors found that genotyping by CE increases resolution and avoids subjectivity in analysing the readout [7]. A similar statement is made today about the replacement of CE with Amp-Seq. Amp-Seq promises advantages, like higher sensitivity and the possibility of increased standardisation of data analysis, but comes with new challenges. CE genotyping of microsatellites has the advantage that they are presumably not under selection pressure, and that the mutation rate of indels caused by DNA polymerase slippages is 10-fold higher than of base pair substitution, which are the origin of SNPs [8,9]. The main advantages of Amp-Seq genotyping lies in the higher sensitivity to detect minority clones at low within-host clone frequencies. The high sensitivity of Amp-Seq (Chapter 2) was achieved by removal of amplification artefacts, e.g. chimeric reads caused by incomplete primer extension and inhomologous re-annealing, or indels caused by polymerase slippage at stretches of homo-polymers. The specific removal of amplification artefacts permitted lower cut-off criteria for Amp-Seq genotyping than for CE genotyping. Such a specific removal of amplification artefacts is not possible for *msp2*-CE genotyping and consequently some minority clones were not detected, although were visible below the cut-off criteria in the background noise of longitudinal samples.

For the first time, sensitivity and false-discovery rate (FDR) to detect parasite clones in longitudinal samples could be estimated. Yet, the true composition of haplotypes within a field sample is unknown, because every genotyping method has limited detectability to detect minority clones. The following two factors can affect the estimates of sensitivity and FDR: (1) Additional missed false-negative haplotypes would lead to a lower sensitivity and FDR than currently estimated. (2) A false-positive haplotype that should have been classified as a true-positive would lead to a higher sensitivity and a lower FDR. For example, if a haplotype occurred only at a single time-point in an individual and if one of the replicates failed or the haplotype was detected below cut-off, then the haplotype would be classified as false-positive instead as true-positive. Such a situation was found 6 times in the *msp2*-CE genotyping data of Chapter 3. Therefore, FDR was not estimated for *msp2*-CE genotyping in Chapter 3. The FDR of *msp2*-CE would more likely estimate standardisation problems between different laboratory than the specificity of the *msp2*-CE genotyping method.

5.1.2 Technical considerations for assay development

A recent publication (Kou et al. 2016) claims that the sensitivity and specificity of Amp-Seq genotyping can be further improved by identifying ‘PCR duplicates’, i.e. amplified fragments which originate from the same template. The Amp-Seq genotyping technique presented in this thesis cannot identify ‘PCR duplicates’. A possibility to identify PCR duplicates is to integrate a molecular unique identifier (UID) consisting of a random nucleotide sequence of ~8 bp between a linker and a marker specific primer sequence of both nested amplicon primers [10] (Figure 1). Identification of ‘PCR duplicates’ permits calculation of consensus sequence of reads sharing the same UID. Based on the consensus sequence of a UID, amplification and sequencing errors of reads with the same UID can be identified and thus corrected.

UID guarantees that in standard PCR application sequence reads with identical UID originate from the same template. But it cannot be guaranteed that two different UID originated from different templates, because a template DNA is amplified multiple times during subsequent amplification cycles. Consequently, two different consensus sequences of different UIDs could still originate from the same template, if an amplification error was introduced. Thus, incorrect interpretation of UIDs used in standard PCR application could lead to false haplotype calls.

Another technique for Amp-Seq library preparation is the molecular inversion probes (MIP) techniques. MIPs guarantees that every UID originates from an original template [11,12]. MIPs are single-stranded DNA molecules consisting of a ~30 bp linker sequence flanked by ~20 bp target-specific sequence on both ends. The target-specific sequence hybridises to the target region (~100 bp in length), followed by a gap-filling and ligation step leading to a circularised DNA. The non-circularised fragments are then digested by exonucleases and the circularised fragments amplified with primers containing sequencing adapter, sample barcode and linker specific sequence at the 3' end. However, capture efficiency of MIPs is limited and >40'000 templates are required as input material for MIPs [13], which corresponds to a field sample with very high parasitaemia. Thus, MIP technique is not suited for genotyping of *Plasmodium* field samples.

In order to benefit from UID for error correction, consensus sequences retrieved from at least three reads with the same UID are required [14]. However, without MIP technique, the vast majority of sequence reads with identical UID occur only once or twice and cannot be considered in the analysis. This in turn limits the sensitivity to detect minority clones. Consequently, using UID for error correction does not necessary improve sensitivity and specificity of Amp-Seq genotyping, because some amplification errors might not be identified. But, UID can identify sequencing errors and would permit to use more error prone sequencing platforms for Amp-Seq genotyping, e.g. MinION (Oxford Nanopore Technologies). Furthermore, by using UID a lot can be learned about the variation in the sequencing data caused by sequence and amplification errors. With the gained knowledge, filtering of PCR artefacts and cut-off criteria for minority clone detection can likely be further optimized.

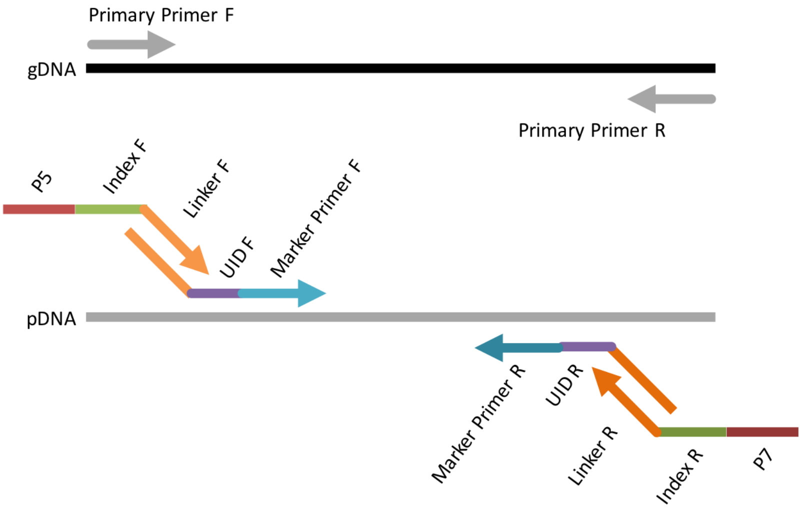


Figure 1: Design of Amp-Seq genotyping primers, including a molecular unique identifier (UID). Primary primers target the gene of interest. Primary PCR is followed by nested PCR using marker-specific primers that carry UID and linker sequences at their 5' ends. The primers for the final round of amplification target the F and R linker sequences. These primers carry sample-specific indices (barcodes) plus Illumina sequencing adapter P5 and P7 at their 5' ends.

The volume of a finger prick blood sample collected in the field is often limited to a maximum of 300-400 μl of whole blood. For typing of multiple loci, individual PCR reactions with ~4 μl template are required depending on the desired limit of detection. Therefore, multiplexing of several markers would be preferable. Multiplexing of more than eight different amplicons was tested during the development of the Amp-Seq genotyping method. However, optimizing the amplification reaction so that all amplicons were equally efficiently amplified was difficult. Therefore, multiplexing was limited to three different amplicons per primary PCR. The main difficulties for multiplexing were reduced amplification efficiency caused by dimer interactions, and unbalanced amplification caused by different amplicon sizes. Amplification of longer amplicons was much less efficient. It was impossible to completely prevent primer dimer interactions, because of the very low complexity of the genome of *P. falciparum* (average GC content of coding regions is 23.7%)[15].

The challenges of multiplexing marker of different fragment length can partly be overcome by using digital PCR platforms, such as the BioRad droplet digital PCR (ddPCR) system, or RainDance technology [16]. A digital PCR platform divides the PCR reaction in thousands of micro-droplets before amplification, thus each droplet represents a separate PCR reaction containing only a single template. This prevents or minimises direct template competition and also reduces formation of chimeric reads that are caused by incomplete primer extension and in-homologous re-annealing. Digital PCR platforms might also prevent the amplification bias for shorter fragments of length-polymorphic marker *msp2* [17].

The Amp-Seq library preparation protocol presented in this thesis, required an initial target enrichment by primary PCR. When using digital PCR, this target enrichment might be unnecessary, since all amplified fragments within a droplet come from the same template. It may even be possible to perform all three PCRs of Amp-Seq library preparation, i.e. primary PCR, nested PCR and sequencing adapter attachment, in a single digital PCR reaction. The concentration of target-specific primer could then be limited and thus primer dimer interactions reduced, as the sequencing adapter would amplify the fragment as soon as the first fragments carrying the linker sequence are present in the reaction (Chapter 2 Figure S4). Reducing the library preparation to a single digital PCR reaction would also reduce cross-sample contamination and reduce potential false positive haplotypes caused by carryover effects. However, the feasibility to perform Amp-Seq library preparation in a single digital PCR reaction must be experimentally proven.

The developed Amp-Seq genotyping laboratory protocols have further optimisation potential. Firstly, the final elongation step could be removed for all amplification steps. During final elongation, incompletely extended primers are elongated, which leads to chimeric reads. Depending on the marker sequence it can be very challenging to distinguish chimeric reads from true genotypes. Secondly, the equimolar pooling step is labourious and error prone. The pooling step of the different markers and samples could be simplified by using a procedure that captures only a limited amount of fragments, e.g. SequalPrep kit (Invitrogen) [18]. However, samples with low parasitaemia may not contain enough fragments to reach saturation of the capturing method. And thirdly, the final purification step could be modified by using magnetic baits carrying the P5 and P7 sequencing adapters (Chapter 2 Figure S4). This would improve the sequencing library quantification leading to optimal loading and cluster formation on the sequencing flow cell, and in turn increase the amount of sequencing output. Such sequencing adapter specific baits could also potentially be used as capturing method for the equimolar pooling step, thus simplifying the protocol further.

5.1.3 Considerations for marker selection

The new SNP polymorphic marker PF3D7_0104100 (*cpmp*) was discovered by scanning the unfiltered SNP list of WGS data from PNG that were part of the MalariaGEN project. The same scan was repeated on the filtered SNP list from the global MalariaGEN dataset [19]. The expected heterozygosity of *cpmp* was lower in the filtered SNP list. The reason for this lower genetic diversity was that many SNPs were removed from the list by the MalariaGEN filtering criteria, as the read coverage was not sufficient to pass the SNP filter criteria. A closer look at the read alignments showed indeed, a lower coverage in the SNP polymorphic region of *cpmp*. The decrease in sequence coverage of WGS data can be explained by too many mismatches to the reference sequence genome, caused for example by length-polymorphism of microsatellite, or by too many SNPs in a region similar to the size of a sequence read. It then depends on the used alignment parameter whether such reads are mapped to the reference genome. Most of the SNPs that were filtered out by the MalariaGEN analysis workflow were found to be true SNPs by the newly obtained Amp-Seq data. Less stringent criteria used by MalariaGEN might result in too many SNPs being called e.g. from length-polymorphism region. As a consequence of this observation, if decreased read coverage in the MalariaGEN data is observed for certain genes from samples of a distinct geographical region, a further in-depth analysis should be done. It is possible that SNPs were falsely excluded during filtering due to high local variation, i.e. too many SNPs compared to the reference genome.

During the analysis of the Amp-Seq data, additional criteria became obvious for future marker selection. The genetic diversity of marker *cpmp* was only slightly higher than that of marker *ama1*-D2 and *ama1*-D3, but marker *cpmp* contains many more SNPs (17, 11, and 48 SNPs, respectively). Many SNPs of marker *cpmp* were in close proximity, and thus were linked within a genotype, i.e. they showed a high linkage disequilibrium (LD). Those SNPs do not add information towards genetic diversity, but they increase the genetic distance between the haplotypes, leading to more robustness for haplotype calling in the presence of sequencing and amplification errors. In addition, chimera haplotypes can be easier identified in the presence of more SNPs. If those SNPs are equally distributed over the whole length of the amplicon, the probability of a resulting identical chimera haplotype in repeatedly genotyped samples is very low. Both characteristics were not given for the *ama1* markers, and as a result the identification of chimera haplotypes was much more difficult and often less conclusive for both *ama1* markers than for marker *cpmp*.

5.1.4 Considerations for haplotype calling

The decision to choose the swarm software for haplotype clustering instead of SeekDeep was mainly based on computational performance [20,21]. The runtime to cluster haplotypes with swarm software was much shorter than with SeekDeep. Furthermore, the clustering with swarm can be carried out on a personal computer, whereas a computer cluster with large working memory is needed for SeekDeep. It must be noted that SeekDeep was tested for data analysis in August 2015, before SeekDeep was published, and that in the meantime the performance of SeekDeep may have improved. A systematic comparison of clustering results of both methods was not carried out, because of the slow computational performance of SeekDeep. SeekDeep and swarm software use similar clustering approach and it is unlikely that the resulting haplotype clusters differ much between the methods.

Recently, the new method DEploid was published [22]. DEploid infers haplotypes from unlinked SNPs rather than by clustering of sequence reads. A preliminary analysis of our Amp-Seq genotyping data with DEploid showed that DEploid cannot always correctly infer a minority clone at a within-host haplotype frequency <1% in defined mixtures of *P. falciparum* strains HB3 and 3D7 (own unpublished data). Furthermore, DEploid cannot always infer haplotypes in samples with high MOI. However, DEploid detected two closely related clones correctly, whereas clustering by swarm could not differentiate the two clones based on a single marker. DEploid infers multi-locus haplotypes, which is not yet integrated in HaplotypeR. A systematic comparison of HaplotypeR and DEploid is still outstanding. The high quality strain mixtures used in this thesis would be ideally suited for such comparison.

The algorithm used in Chapter 3 to infer multi-locus haplotypes based on longitudinal field samples performed similar as DEploid without using longitudinal samples. It is therefore likely, that by using a combinatorial approach of linked and unlinked SNP information, as well as longitudinal sample information, the challenge of very complex clone mixture and low abundancy of minority clones could be overcome. Firstly, single locus haplotypes could be inferred by using the same approach as DEploid (unlinked SNP), by limiting the combinatorial search space to the sequence reads. Secondly, the inferred local haplotypes and within-host haplotype frequency can be used to infer the multi-locus haplotype by using the same approach as DEploid. Finally, local haplotype inferred from preceding or following bleeds can be included to define the final haplotype set.

5.2 Deconvolution of Mixed Stage Transcriptomes

Knowledge of the gene expression profiles and their regulation is very important for basic malaria research. It helps in identifying new drug targets, as well as understanding drug resistance mechanisms [23]. Stage-specific marker genes can be used to monitor changes in stage composition during IDC, for example after drug treatment. However, the study of gene expression in field samples is complicated by the presence of mixtures of different parasite stages in the human blood. Even in cultured parasites after tight synchronisation or enrichment of a specific stage, small fractions of other stages are found. In the past, many different deconvolution methods were developed for heterogeneous human samples [24]. Most of those methods infer the stage composition rather than stage-specific transcriptomes. Evaluation of a subset of methods with experimental mixed *P. falciparum* stages showed that analysis of such heterogeneous transcriptomes can be very challenging, especially when an increase in total RNA takes place simultaneously.

One deconvolution method was specifically developed to infer *P. falciparum* stage-specific transcriptomes and stage composition from field samples [25]. It was developed based on gene expression data from the affymetrix microarray platform. Application of this methodology to RNA-Seq data of experimentally mixed transcriptomes showed that neither inferred stage-specific transcriptomes (deconvolution method 'lm' in Chapter 4), nor inferred stage composition (deconvolution method 'qprog') agreed to stage-specific transcriptomes or mixture ratios used for experimentally mixed transcriptomes.

Measurements of gene expression by microarray and RNA-Seq differ in two important aspects. Firstly, microarray platforms measure abundance of RNA by hybridisation to gene specific probes. Each hybridisation represents an independent process, thus measurements of individual genes are not influenced by abundance of RNA from other genes. In contrast, in RNA-Seq, fragments of all genes are sequenced. Thus, sampling of each gene depends on the relative abundance of RNA in the sample, and thus on expression levels of other genes. Therefore, RNA-Seq is more prone to bias of changes in RNA composition between different samples. Secondly, microarray platforms measure continuous fluorescence intensity, whereas by RNA-Seq, read count is observed [26]. Therefore, the resulting different distribution of the data requires different statistical models.

In the present study, estimating stage-specific signatures by CPM normalisation and deconvolution with a negative binomial regression model (method used by edgeR), followed by selection of genes with a significant fold change (as measured by permutation tests) showed the best agreement to stage-specific transcriptomes (Chapter 4 Figure 2). In contrast to other deconvolution methods, normalisation was less important for deconvolution with edgeR. An explanation for observation could be that the log transformation of the mixed-stage transcriptomes (during negative binomial regression of edgeR) reduces the effect of the 18-fold increase in total RNA

Initial attempts to infer stage composition of experimental mixed-stage samples based on known stage-specific signatures were unsuccessful so far. Both tested deconvolution methods, Cibersort and qprog, overestimated the proportion of schizonts in the sample. However, if the estimated stage compositions were additionally adjusted by the fold increase in total RNA, then agreement to original stage compositions was much better.

This preliminary result holds the promise that if fold increase in total RNA can be included into a new method for deconvolution, it may be feasible to estimate stage compositions from RNA-Seq data.

An evaluation of different normalisation and deconvolution methods was performed to provide a proof of concept for the intended application of inferring the *P. vivax* gametocyte transcriptome after enrichment of gametocytes from patient samples. Currently the transcriptome of *P. vivax* gametocytes is largely unknown. Existing knowledge about *P. vivax* gametocytes was gained through orthologous genes that are present in *P. falciparum* and *P. vivax*, e.g. *pvs25* and *pvs28*. However, gametocytogenesis of *P. falciparum* and *P. vivax* differ [27,28], and gametocyte-specific genes only found in the genome of *P. vivax* cannot be identified by comparative analysis of *P. falciparum* and *P. vivax*. Thus, the transcriptome needs to be inferred from gametocyte enriched field samples with known stage composition.

During the course of this thesis, initial attempts to infer the transcriptome of *P. vivax* gametocytes by deconvolution with edgeR were made (these preliminary data were not presented in results section). No significant differentially expressed genes could be identified. An explanation for this failure is that the parasite stage compositions determined by light microscopy might not be correct. In contrast to *P. falciparum*, where gametocytes show a characteristic shape, *P. vivax* gametocyte and trophozoite look similar (Figure 2). Stage compositions of the same sample by two different expert microscopists showed disagreement in gametocyte and trophozoite counts (unpublished data). Stage count of asexual and sexual parasites could be facilitated by indirect fluorescent antibody labelling of a known gametocyte-specific protein, such as the sexual stage antigen s16 [29,30]. However, this assay would only differentiate between gametocytes and asexual parasites, allowing deconvolution to infer genes that are up or down regulated in gametocyte only, but would not permit inferring of ring-, trophozoite- or schizont-specific genes.

Differential gene expression analysis of stage-specific transcriptomes from highly synchronized *P. falciparum* cultures showed that known schizont-specific genes were also upregulated in ring stage transcriptome. This upregulation can be explained by a fraction of schizont stage parasites, which was also found in the highly synchronised ring stage sample (Chapter 4 Figure S1). This gene upregulation was no longer observed in our data after additional purification of the synchronised ring stage sample, yielding a sample of highly pure ring stages. The same observation was made with known ring stage-specific genes, which were upregulated in the highly synchronised schizont stage sample. But, the upregulation caused by contaminating ring stage parasites was smaller than the upregulation caused by contaminating schizont stage parasites. The different influence of contaminating stages can be explained by the 18-fold increase in total RNA during the development from ring to schizont stages.

Single cell RNA sequencing (scRNA-Seq) platforms offer another solution to the problem of stage-heterogeneity in field isolates, e.g. Fluidigm C1, DropSeq, 10x Chromium and single cell FACS sorting. scRNA-Seq characterizes a defined single cell, rather than an average of the gene expression of individual cells as in RNA-Seq. Recently, the first scRNA-Seq transcriptome of the erythrocytic cycle of *P. falciparum* was published showing the dynamics of stage development in cultured parasites [31]. However, scRNA-Seq comes with new

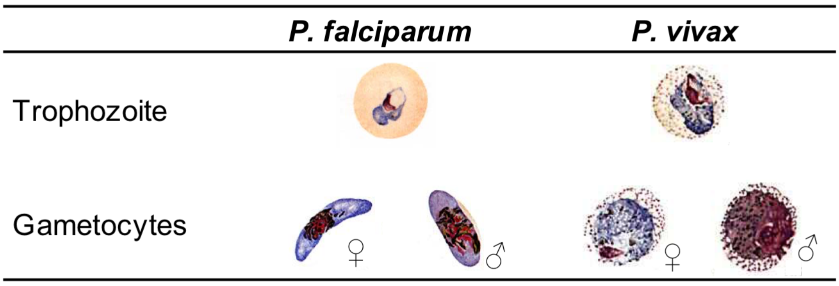


Figure 2: Illustration of trophozoite and gametocyte of *P. falciparum* and *P. vivax* (image source Coatney et al. 1971).

challenges. For example, only transcripts of few genes at low coverage are sequenced per cell, showing only the ‘tip of the iceberg’ of gene expression in a cell. Additionally, scRNA-Seq is prone to sampling artefacts, as amplification of reverse transcribed RNA is required to achieve sufficient material for sequencing. And finally, analysis of scRNA data is complicated by a lot of missing data. Future studies will show how far scRNA-seq can resolve difficulties encountered with mixed parasite stages.

5.3 CONCLUSIONS

Next generation sequencing (NGS) permits fundamentally novel approaches to study *Plasmodium* parasites and will continue to shape malaria research. The unique challenges of *Plasmodium* parasite field samples for NGS data analysis require specifically developed tools, in particular for data analysis. This thesis provides solutions for genotyping and analysing *P. falciparum* samples containing a mixture of parasite clones, as well as a new highly sensitive Amp-Seq genotyping assay, including a novel, highly diverse marker, *cpmp*. Furthermore, this thesis provides a strategy on how to best infer stage-specific gene expression from samples containing a mixtures of *Plasmodium* parasite developmental stages.

Amp-Seq genotyping permits quantification of individual genotypes within a human host and thus to study individual parasite clone densities. This novel molecular epidemiological parameter opens new possibilities to study malaria epidemiology and might help to answer open questions about parasite fitness. For example, parasite densities might help to explain the difference between short and long duration of infection [2]. Furthermore, fitness of individual parasite genotypes could be studied when resources are limited due to superinfection or when natural immunity is acquired.

Deconvolution of mixed-stage field samples into stage-specific transcriptomes is a crucial method to analyse gene expression data from field samples. Field isolates are the main source of material for RNA-Seq experiments if no *in vitro* culture system is available, e.g. in the case of *P. vivax*, or when the interaction of the parasite with a clinical phenotype of the host is of interest. So far, the study of stage-specific gene expression of parasites carrying a specific phenotype, e.g. artemisinin resistance, in field samples was greatly hampered by the mixture of development stages [23]. Deconvolution permits analysis of stage-specific gene expression of isolates containing mixed stages, when the proportions of parasite developmental stages in the sample is known, e.g. through stage counts by microscopy.

This thesis shows that malaria epidemiology can greatly benefit from next generation sequencing technologies. Further development of sequencing technology will simplify laboratory procedures, as well as data analysis. It might be expected that sequencing of field samples will be as common in future as performing a PCR is today.

REFERENCES

1. Juliano JJ, Porter K, Mwapasa V, et al. Exposing malaria in-host diversity and estimating population diversity by capture-recapture using massively parallel pyrosequencing. Proc Natl Acad Sci U S A. **2010**; 107(46):20138–43.

2. Bretscher MT, Maire N, Felger I, Owusu-Agyei S, Smith T. Asymptomatic Plasmodium falciparum infections may not be shortened by acquired immunity. Malar J. BioMed Central; **2015**; 14(1):294.

3. Felger I, Maire M, Bretscher MT, et al. The Dynamics of Natural Plasmodium falciparum Infections. Gosling RD, editor. PLoS One. Public Library of Science; **2012**; 7(9):e45542.

4. Bretscher MT, Maire N, Chitnis N, Felger I, Owusu-Agyei S, Smith T. The distribution of Plasmodium falciparum infection durations. Epidemics. Elsevier B.V.; **2011**; 3(2):109–118.

5. Sama W, Owusu-Agyei S, Felger I, Vounatsou P, Smith T. An immigration-death model to estimate the duration of malaria infection when detectability of the parasite is imperfect. Stat Med. **2005**; 24(21):3269–88.

6. Sama W, Owusu-Agyei S, Felger I, Dietz K, Smith T. Age and seasonal variation in the transition rates and detectability of Plasmodium falciparum malaria. Parasitology. **2006**; 132(Pt 1):13–21.

7. Falk N, Maire N, Sama W, et al. Comparison of PCR-RFLP and Genescan-based genotyping for analyzing infection dynamics of Plasmodium falciparum. Am J Trop Med Hyg. **2006**; 74(6):944–50.

8. Anderson TJC, Su X-Z, Roddam A, Day KP. Complex mutations in a high proportion of microsatellite loci from the protozoan parasite Plasmodium falciparum. Mol Ecol. **2000**; 9(10):1599–1608.

9. Hamilton WL, Claessens A, Otto TD, et al. Extreme mutation bias and high AT content in Plasmodium falciparum. Nucleic Acids Res. **2017**; 45(4):1889–1901.

10. Kou R, Lam H, Duan H, et al. Benefits and challenges with applying unique molecular identifiers in next generation sequencing to detect low frequency mutations. PLoS One. **2016**; 11(1):1–15.

11. Niedzicka M, Fijarczyk A, Dudek K, Stuglik M, Babik W. Molecular Inversion Probes for targeted resequencing in non-model organisms. Sci Rep. Nature Publishing Group; **2016**; 6(October 2015):24051.

12. Hardenbol P, Banér J, Jain M, et al. Multiplexed genotyping with sequence-tagged molecular inversion probes. Nat Biotechnol. **2003**; 21(6):673–678.

13. Lau HY, Palanisamy R, Trau M, Botella JR. Molecular inversion probe: a new tool for highly specific detection of plant pathogens. PLoS One. **2014**; 9(10):e111182.

14. Beerenwinkel N, Günthard HF, Roth V, Metzner KJ. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbiol. **2012**; 3(September):329.

15. Gardner MJ, Hall N, Fung E, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature. **2002**; 419(6906):498–511.

16. Mamanova L, Coffey AJ, Scott CE, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. **2010**; 7(2):111–118.

17. Messerli C, Hofmann NE, Beck H-P, Felger I. Critical Evaluation of Molecular Monitoring in Malaria Drug Efficacy Trials and Pitfalls of Length-Polymorphic Markers. Antimicrob Agents Chemother. **2017**; 61(1):AAC.01500-16.

18. Harris JK, Sahl JW, Castoe TA, Wagner BD, Pollock DD, Spear JR. Comparison of normalization methods for construction of large, multiplex amplicon pools for next-generation sequencing. Appl Environ Microbiol. **2010**; 76(12):3863–8.

19. MalariaGEN Plasmodium falciparum Community Project. Genomic epidemiology of artemisinin resistant malaria. Elife. **2016**; 5:1–29.

20. Mahé F, Rognes T, Quince C, Vargas C de, Dunthorn M. Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ. **2015**; 3:e1420.

21. Hathaway NJ, Parobek CM, Juliano JJ, Bailey JA. SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. Nucleic Acids Res. Oxford University Press; **2017**; (December):1–13.

22. Zhu SJ, Almagro-Garcia J, McVean G. Deconvolution of multiple infections in Plasmodium falciparum from high throughput sequencing data. Bioinformatics. **2017**; .

23. Mok S, Ashley EA, Ferreira PE, et al. Drug resistance. Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance. Science. **2015**; 347(6220):431–5.

24. Mohammadi S, Zuckerman N, Goldsmith A, Grama A. A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues. Proc IEEE. **2017**; 105(2):340–366.

25. Joice R, Narasimhan V, Montgomery J, et al. Inferring Developmental Stage Composition from Gene Expression in Human Malaria. Przytycka TM, editor. PLoS Comput Biol. **2013**; 9(12):e1003392.

26. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. **2010**; 11(3):R25.

27. Bousema T, Drakeley C. Epidemiology and infectivity of Plasmodium falciparum and Plasmodium vivax gametocytes in relation to malaria control and elimination. Clin Microbiol Rev. **2011**; 24(2):377–410.

28. Coatney G, Collins W, Warren M, Contacos P. The Primate Malarias. Atlanta: Center for disease control and prefention; 1971.

29. Moelans II, Meis JF, Kocken C, Konings RN, Schoenmakers JG. A novel protein antigen of the malaria parasite Plasmodium falciparum, located on the surface of gametes and sporozoites. Mol Biochem Parasitol. **1991**; 45(2):193–204.

30. Bruce MC, Carter RN, Nakamura K, Aikawa M, Carter R. Cellular location and temporal expression of the Plasmodium falciparum sexual stage antigen Pfs16. Mol Biochem Parasitol. **1994**; 65(1):11–22.

31. Poran A, Nötzel C, Aly O, et al. Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. Nature. Nature Publishing Group; **2017**; 551(7678):95–99.

APPENDIX

During the course of this PhD additional contribution to a project has been made:

Integrated transcriptomic, proteomic and epigenomic analysis of *Plasmodium vivax* salivary-gland sporozoites
Vivax Sporozoites Consortium

– Manuscript is published in bioRxiv Jun 7, 2017, DOI: https://doi.org/10.1101/145250 –
– Manuscript is submitted to Journal of PLOS Neglected Tropical Diseases –

Integrated transcriptomic, proteomic and epigenomic analysis of *Plasmodium vivax* salivary-gland sporozoites.

Vivax Sporozoite Consortium* (Ivo Muller[1,2,3], Aaron R. Jex[1,3,4], Stefan H. I. Kappe[5], Sebastian A. Mikolajczak[5], Jetsumon Sattabongkot[7], Rapatbhorn Patrapuvich[6], Scott Lindner[8], Erika L. Flannery[5], Cristian Koepfli[1], Brendan Ansell[4], Anita Lerch[1], Kristian E. Swearingen[5], Robert L. Moritz[9], Michaela Petter[10,11], Michael Duffy[10], Vorada Chuenchob[5]).

***Group authorship – all authors are equal contributors (order per author contributions section below).**

1. Population Health and Immunity Division, The Walter and Eliza Hall Institute for Medical Research, 1G Royal Parade, Parkville, Victoria, 3052, Australia.
2. Malaria: Parasites & Hosts Unit, Institut Pasteur, 28 Rue de Dr. Roux, 75015, Paris, France.
3. Department of Medical Biology, The University of Melbourne, Victoria, 3010, Australia.
4. Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Corner of Park and Flemington Road, Parkville, Victoria, 3010, Australia.
5. Center for Infectious Disease Research, 307 Westlake Avenue North, Suite 500, Seattle, WA 98109, USA;
6. Department of Global Health, University of Washington, Seattle, WA 98195, USA.
7. Mahidol Vivax Research Center, Faculty of Tropical Medicine, Mahidol University, Bangkok 10400, Thailand.
8. Department of Biochemistry and Molecular Biology, Center for Malaria Research, Pennsylvania State University, University Park, PA 16802, USA.
9. Institute for Systems Biology, Seattle, WA, 98109, USA.
10. Department of Medicine Royal Melbourne Hospital, The Peter Doherty Institute, The University of Melbourne, 792 Elizabeth Street, Melbourne, Victoria 3000, Australia.
11. Institute of Microbiology, University Hospital Erlangen, Erlangen 91054, Germany

Abstract

Background: *Plasmodium vivax* is the key obstacle to malaria elimination in Asia and Latin America, largely attributed to its ability to form resilient 'hypnozoites' (sleeper-cells) in the host liver that escape treatment and cause relapsing infections. The decision to form hypnozoite is made early in the liver infection and may already be set in sporozoites prior to invasion. To better understand these early stages of infection, and the potential mechanisms through which the development may be pre-programmed, we undertook a comprehensive transcriptomic, proteomic and histone epigenetic characterization of *P. vivax* sporozoites.

Results: Our study highlights the loading of the salivary-gland sporozoite with proteins required for cell traversal and invasion and transcripts for infection of and development within hepatocytes. We characterise histone epigenetic modifications in the *P. vivax* sporozoite and explore their role in regulating transcription. This work shows a close correlation between H3K9ac marks and transcriptional activity, with H3K4me3 and H3K9me3 appearing to act as general markers of euchromatin and heterochromatin respectively. We also identify the remarkable transcriptional silence in the (sub)telomeres and discuss potential roles of AP2 transcription factors, specifically ApiAP2-SP and L in regulating this stage.

Conclusions: Collectively, these data indicate the sporozoite as a tightly programmed stage primed to infect the human host and identifies key targets to be further explored in liver stage models.

Background

Malaria is among the most significant infectious diseases impacting humans globally, with 3.3 billion people at risk of infection, 381 million suspected clinical cases and up to ~660,000 deaths attributed to malaria globally in 2014 [1]. Two major parasite species contribute to the vast majority of human malaria, *Plasmodium falciparum* and *P. vivax*. Historically, *P. falciparum* has attracted the majority of global attention, due to its higher contribution to morbidity and mortality. However, *P. vivax* is broadly distributed, more pathogenic than previously thought, and is recognised as the key obstacle to malaria elimination in the Asia-Pacific and Americas [2]. Unlike *P. falciparum*, *P. vivax* can establish long-lasting 'sleeper-cells' (= hypnozoites) in the host liver that emerge weeks, months or years after the primary infection (= relapsing malaria) [3]. Primaquine is the only approved drug that prevents relapse. However, the short half-life, long dosage regimens and incompatibility of primaquine with glucose-6-phosphate-dehydrogenase deficiency (which requires pre-screening of recipients [4]) makes it unsuitable for widespread use. As a consequence, *P. vivax* is overtaking *P. falciparum* as the primary cause of malaria in a number of co-endemic regions [5]. Developing new tools to diagnose, treat and/or prevent hypnozoite infections is considered one of the highest priorities in the malaria elimination research agenda [6].

When *Plasmodium* sporozoites are deposited by an infected mosquito, they likely traverse the skin cells, enter the blood-stream and are trafficked to the host liver, as has been shown in rodent malaria parasites [7]. Upon reaching the liver, sporozoites traverse Kupffer and endothelial cells to reach the parenchyma, moving through several hepatocytes before invading a final hepatocyte suitable for liver stage development [7, 8]. Within hepatocytes, these parasites replicate, and undergo further development and differentiation to produce tens of thousands of merozoites that emerge from the liver and infect red blood cells. However, *P. vivax* sporozoites are able to commit to two distinct developmental fates within the hepatocyte: they either immediately continue development as replicating schizonts and establish a blood infection, or delay replication and persist as hypnozoites. Regulation of this major development fate decision is not understood and this represents a key gap in current knowledge of *P. vivax* biology and control.

The sporozoites' journey from skin deposition to hepatocytes takes less than a few minutes [9]. It has been hypothesized that *P. vivax* sporozoites exist within an inoculum as replicating 'tachysporozoites' and relapsing 'bradysporozoites [10] and that these subpopulations may have distinct a developmental fate as schizont or hypnozoites, thus contributing to their relapse phenotype [10-12]. This observation is supported by the stability of different hypnozoite phenotypes in *P. vivax* infections of liver-chimeric mouse models [13]. Sporozoites prepare for mammalian host infection while still residing in the mosquito salivary glands. Studies using rodent malaria parasites have identified genes [14], that are transcribed in sporozoites but translationally repressed (i.e., present as transcript but un- or under-represented as protein), via RNA-binding proteins [15], and ready for just-in-time translation after the parasites infection of the mammalian host [13, 16]. Translational repression (i.e., the blocking of translation of present and retained transcripts) and other mechanisms of epigenetic control may contribute to the *P. vivax* sporozoite fate decision and hypnozoite formation, persistence and activation. Supporting this hypothesis, histone methyltransferase inhibitors stimulate increased activation of *Plasmodium cynomolgi* hypnozoites in macaque hepatocytes [17, 18]. Epigenetic control of stage development is further evidenced in *Plasmodium* through chromatin structure controlling expression of PfAP2-G, a specific transcription factor that, in turn, regulates gametocyte (dimorphic sexual stages) development in blood-stages [19]. It is well documented that *P. vivax* hypnozoite activation patterns stratify with climate and geography [11] and recent modelling suggests transmission potential selects for hypnozoite phenotype [20]. Clearly the ability for *P. vivax* to dynamically regulate hypnozoite formation and relapse phenotypes in response to high or low transmission periods in different climate conditions would confer a significant evolutionary advantage.

Unfortunately, despite recent advances [21] current approaches for *in vitro P. vivax* culture do not support routine maintenance in the laboratory and tools to directly perturb gene function are not established. This renders studies on *P. vivax*, particularly its sporozoites and liver stages, exceedingly difficult. Although *in-vitro* liver stage assays and humanised mouse models are being developed [13], they cannot yet support 'omics analysis of *P. vivax* liver stage dormancy. Recent characterization [22] of liver-stage (hypnozoites and schizonts) of *P. cynomolgi* (a related and relapsing parasite in macaques) provides valuable insight, but investigations in *P. vivax* directly are clearly needed. The systems analysis of *P. vivax* sporozoites that reside in the mosquito salivary glands and are poised for transmission and liver infection offer a key opportunity to gain insight into *P. vivax* infection. To date, such characterization of *Plasmodium vivax* sporozoites is limited [23], and only one recent study, of *P. falciparum* [24], has undertaken exploration of epigenetic regulation in sporozoites of any *Plasmodium* species. Here, we present a detailed characterization of the *P. vivax* sporozoite transcriptome, proteome and epigenome and use these data to better understand this key infective stage and the role of sporozoite programming in invasion and infection of the human host, and development within the host liver.

Results and Discussion

We quantified transcript abundance for 5,714 *P. vivax* genes (4,991 with a mean transcript per million (TPM) count ≥ 1.0) at a mean estimated abundance of 175.1 TPM (Additional File 1: Figure S1 and Additional File 2: Table S1) for *P. vivax* sporozoites isolated from *Anopheles dirus* salivary glands using the recently completed *P. vivax* P01 assembly and gene models (see methods). For ease of reference, where one-to-one orthologs are established between the P01 and previous *P. vivax* (Sal1) reference, we use the Sal1 gene names in text (both the P01 and Sal1 gene names are provided for all genes in the supplementary information). Mosquito infections were generated by membrane feeding of blood samples taken from *P. vivax* infected patients in western Thailand (n = 9). Among the most highly transcribed genes in the infectious sporozoite stage are *csp* (circumsporozoite protein), five *etramps* (early transcribed membrane proteins), including *uis3* (up-regulated in infective sporozoites), *uis4* and *lsap-1* (liver stage associated protein 1), a variety of genes involved in cell transversal and initiation of invasion, including *celtos* (cell traversal protein for ookinetes and sporozoites), *gest* (gamete egress and sporozoite traversal protein), *spect1* (sporozoite protein essential for cell traversal) and *siap-1* (sporozoite invasion associated protein), and genes associated with translational repression (*alba1*, *alba4* and *Puf2*). Collectively, these genes account for >1/3^rd^ of all transcription in the sporozoite. We found moderate agreement ($R^2 = 0.35$; Additional File 1: Figure S2) between our RNA-seq data and previous microarray data for *P. vivax* sporozoites [23]. Improved transcript detection and quantitation is expected with the improved technical resolution of RNA-seq over microarray. Supporting this, we find higher correlation between RNA-seq data from *P. vivax* and *P. falciparum* (single replicate sequenced herein for comparative purposes) sporozoite datasets ($R^2 = 0.42$), compared to either species relative to published microarray data (Additional File 1: Figure S2). Although microarray supports the high transcription in sporozoites of genes such as *uis4*, *csp*, *celtos* and several other *etramps*, 27% and 16% of the most abundant 1% of transcribed genes in our sporozoite RNA-seq data are absent from the top decile or quartile respectively in the existing *P. vivax* sporozoite microarray data [23]. Among these are genes involved in early invasion/hepatocyte development, such as *lsap-1*, *celtos*, *gest* and *siap-1*, or translational repression (e.g., *alba-1* and *alba-4*); orthologs of these genes are also in the top percentile of transcripts in RNA-seq (see [24] and Additional File 2: Table S2) and (see [25] and Additional File 2: Table S3) and previous microarray data [26, 27] for *P. falciparum* and *P. yoelii* sporozoites respectively, suggesting many are indeed more abundant than previously characterized.

Transcription in P. vivax *relative to other plasmodia*

To gain insight into species-specific aspects of the *P. vivax* transcriptome, we qualitatively compared these data with available data from *P. falciparum* and *P. yoelii* sporozoites (single replicate only) for 4,220 and 4,067 single-copy orthologs (SCO) (transcribed at ≥ 1 TPM in *P. vivax* infectious sporozoites) shared with *P. falciparum* (Additional File 2: Table S3) and with both *P. falciparum* and *P. yoelli* (Additional File 2: Table S4) respectively. Genes highly transcribed in salivary-gland sporozoites of all three species include *celtos*, *gest*, *trap*, *siap1*, *spect1* and *puf2*. There are 696 *P. vivax* genes shared as orthologs between *P. vivax* P01 and *P. vivax* Sal1 lacking a defined SCO in *P. falciparum* or *P. yoelli* transcribed at a mean of ≥ 1 TPM in *P. vivax* salivary-gland sporozoites (Additional File 2: Table S5). Prominent among these are *vir* (n=25) and *Pv-fam* (41 fam-e, 16 fam-b, 14 fam-a, 8 fam-d and 3 fam-h) genes, as well as, hypothetical proteins or proteins of unknown function (n=212) and, interestingly, a number of 'merozoite surface protein' 3 and 7 homologs (n=5 of each). Both *msp3* and *msp7* have undergone significant expansion in *P. vivax* relative to *P. falciparum* and *P. yoelii* [28] and may have repurposed functions in sporozoites. In addition, there are 69 *P. vivax* P01 genes lacking a defined ortholog in *P. vivax* Sal1, *P. falciparum* or *P. yoelii* transcribed at ≥ 1 TPM in infectious *P. vivax* sporozoites; most of which are *Plasmodium* interspersed repeat (PIR) genes [28] found in telomeric regions of the P01 assembly and likely absent from the Sal1 assembly but present in the Sal1 genome.

P. vivax sporozoite transcriptional enrichment

To comprehensively identify sporozoite enriched transcripts, we compared the *P. vivax* sporozoite transcriptome (Additional File 2: Table S6) to RNA-seq data for *P. vivax* blood-stages [29] (the only other RNA-seq data presently available for *P. vivax*; Fig. 1 and Additional File 1: Figures S3-5). We identified 1,672 up (Additional File 2: Table S7) and 1,958 down-regulated (Additional File 2: Table S8) transcripts (FDR ≤ 0.05; minimum 2-fold change in Counts per Million (CPM)) and next explored patterns among these differentially transcribed genes (DTGs) by protein family (Fig. 1C and Additional File 2: Table S9) and Gene Ontology (GO) classifications (Additional File 2: Table S10). RNA recognition motifs (RRM-1 and RRM-6) and helicase domains (Helicase-C and DEAD box helicases) are over-represented (p-value <0.05) among sporozoite-enriched transcripts, consistent with translational repression through ribonucleoprotein (RNP) granules [30]. Transcripts encoding nucleic acid binding domains, such as bromodomains (PF00439; which can also bind lysine-acetylated proteins), zinc fingers (PF13923) and EF hand domains (PF13499) are also enriched in sporozoites. Included among these proteins are a putative ApiAP2 transcription factor (PVX_083040) and a homologue of the *Drosophila* zinc-binding protein 'Yippee' (PVX_099695). Thrombospondin-1 like repeats (TSR: PF00090) and von Willebrand factor type A domains (PF00092) are enriched in sporozoites as well. In sporozoites, *P. falciparum* genes enriched in TSR domains are important in invasion of the mosquito salivary gland (e.g., *trap*) and secretory vesicles released by sporozoites upon entering the vertebrate host (e.g., *csp*) [31]. By comparison, genes up-regulated in blood-stages are enriched for *vir* gene domains (PF09687 and PF05796), Tryptophan-Threonine-rich *Plasmodium* antigens (PF12319; which are associated with merozoites [32]), markers of cell-division (PF02493),[33] protein production/degradation (PF00112, PF10584, PF00152, PF09688 and PF00227) and ATP metabolism (PF08238 and PF12774). 47 of the 343 transcripts unique to *P. vivax* sporozoites relative to *P. falciparum* or *P. yoelii* are enriched in sporozoites compared to *P. vivax* blood stages. Nine of these are in the top decile of transcription, and include a Pv-fam-e (PVX_089880), a Pf-fam-b homolog (PVX_001710) and 7 proteins of unknown function. A further nine have an ortholog in *P. cynomolgi* (which also forms hypnozoites) but not the closely related *P. knowlesi* (which does not form hypnozoites) and include '*msp7*'-like (PVX_082685, PVX_082650 and PVX_082670) and '*msp3*'-like (PVX_097705) and Pv-fam-e genes (PVX_001100, PVX_089860 and PVX_089810), a serine-threonine protein kinase (PVX_081395) and a RecQ1 helicase homolog (PVX_099345). Notably, the *P. cynomolgi* ortholog of PVX_081395, PCYB_021650, is transcriptionally enriched in hypnozoites relative to replicating schizonts [22], indicating a target of significant interest when considering hypnozoite formation and/or biology.

Translational repression machinery

In *Plasmodium*, translational repression regulates key life-cycle transitions coinciding with switching between the mosquito and the mammalian host (either as sporozoites or gametocytes) [30]. For example, although *uis4* is the most abundant transcript in the infectious sporozoite ([23, 27]; Additional File 2: Table S1), UIS4 is translationally repressed in this stage [15] and only expressed after hepatocyte invasion [34]. In sporozoites, it is thought that PUF2 binds to mRNA transcripts and prevents their translation [25], and SAP1 stabilises the repressed transcripts and prevents their degradation [34]. Consistent with this, *Puf2* and *SAP1* are among the more abundant *P. vivax* transcripts enriched in the sporozoite relative to blood-stages. Indeed, *Puf2* is among the top percentile of transcripts in infectious sporozoites. However, our data implicate other genes, many already known to be involved in translational repression in other *Plasmodium* stages and other protists [30], that may act in *P.*

vivax sporozoites. Notable among these are *alba-2* and *alba-4*, both of which are among the top 2% of genes transcribed in sporozoites and ~14 to 20-fold more highly transcribed in sporozoites relative to blood-stages. In addition, *P. vivax* sporozoites are enriched for genes encoding RRM-6 RNA helicase domains. Intriguing among these genes are HoMu (homolog of Musashi) and ptbp (polypyrimidine tract binding protein). Musashi is a master regulator of eukaryotic stem cell differentiation through translational repression [35] and HoMu localizes with DOZI and CITH in *Plasmodium* gametocytes [36]. PTBP is linked to mRNA stability, splice regulation and translational initiation [37] and may perform a complementary role to SAP1.

Translational repression in *P. vivax* sporozoites

More than 700 genes have been identified as being translationally repressed in *Plasmodium berghei* ('rodent malaria') gametocytes based on DOZI pulldowns [38]. In contrast, translationally repressed genes have not been characterized in sporozoites in a comprehensive way. As a step in addressing this, we analysed the *P. vivax* sporozoite proteome (Additional File 1: Figure S6 and Additional File 2: Table S11) by mass spectrometry and identified peptide signals for 2,640 proteins. Among the most highly expressed proteins in sporozoites were those associated with the apical complex (AMA1, GAMA, RON12, RON3, RON5), motility / cell traversal (MYOSIN A, PLP1, TRAP, SIAP1, GEST, SPECT1, CELTOS) and the inner membrane complex (ISP1/3, IMC1a, e, g, h, m and k), which has a key role in motility and invasion [39]. We identified 2,402 *P. vivax* genes transcribed in the sporozoite (TPKM > 1) for which no protein expression was detected. In considering genes that may be translational repressed (i.e., transcribed but not translated) in the *P. vivax* sporozoite, we confine our observations to those transcripts representing the top decile of transcript abundance to ensure their lack of detection as proteins was not due to limitations in the detection sensitivity of the proteomic dataset. Notably, ~1/3rd of transcripts in the top decile of transcriptional abundance (n = 173 of 558) in *P. vivax* sporozoites were not detectable as peptides in multiple replicates (Additional File 2: Table S12). Of these 173 putatively repressed transcripts, 156 and 154 have orthologs in *P. falciparum* and *P. yoelii* respectively, with 89 and 118 of these also not detected as proteins in *P. falciparum* and *P. yoelii* salivary-gland sporozoites [40] despite being identified as transcribed in these stages (see [24, 25]; Additional File 2: Tables S2-4). In addition, a number of genes (e.g., *uis4*) are expressed in infectious *P. vivax* sporozoites at levels many fold lower than their transcription might indicate (bottom quartile of protein expression, compared with top decile of transcript abundance). While each putatively repressed transcript will require validation, this system level approach is supported by immunofluorescent imaging (Additional File 1: Figure S7) of UIS4 and LISP1 (one known and one proposed here as translationally repressed in *P. vivax* sporozoites) relative to TRAP and BiP (which are both transcribed and expressed as protein in the *P. vivax* sporozoite; Additional File 2: Table S12).

Development within the host hepatocyte

Following cell traversal and hepatocyte invasion, *P. vivax* sporozoites establish their intracellular niche, which includes modification of the parasitophorous vacuole membrane (PVM) and the parasite then proceeds to replicate as a liver stage. UIS3 and UIS4 are resident PVM-proteins and are the best characterized proteins under translational repression by Puf2/SAP1 in infectious sporozoites [41], both of which are essential for liver stage development [14]. In the present study, *uis4* represents 18.8% of transcripts but just 0.06% of proteins in the sporozoites. Similarly, *uis3* is the 7th most abundant transcript in sporozoites, but represented only by a single peptide count in one proteomic replicate. In addition to *uis3* and *uis4*, genes involved in liver stage development and under apparent translational repression in the *P. vivax* sporozoites include *lsap1* (liver stage associated protein 1), *zipco* (ZIP domain-containing protein), several other *etramps* (PVX_118680, PVX_003565, PVX_088870 and PVX_086915), *pv1* (parasitophorous vacuole protein 1) and *lisp1* and *lisp2* (PVX_085550 and PVX_000975). The *lisp1* gene is an intriguing find, and may have an altered role in *P. vivax* liver stages (Additional File 1: Figure S7). In *P. berghei*, *lisp1* is essential for rupture of the PVM during liver stage development allowing release of the merozoite into the host blood stream. *Pv-lisp1* is ~350-fold and ~1,350-fold more highly transcribed in *P. vivax* sporozoites compared to sporozoites of either *P. falciparum* or *P. yoelli* (see Additional File 2: Table S4). Also notable among translationally repressed genes in sporozoites is a putative 'Yippee' homolog (PVX_099695). Yippee is a DNA-binding protein that, in humans (YPEL3), suppresses cell growth [42]. Its specific function in *Plasmodium*, either in parasite development or on the host interactions, is not yet known. However, that Yippee-like proteins suppress cell growth/division and appear to be regulated through histone acetylation [43] is intriguing in the context of a potential role in *P. vivax* hypnozoite developmental arrest.

The *P. vivax* ortholog (PVP01_1016100; no corresponding ortholog is identified in the *P. vivax* Sal1 assembly) of the *P. cynomolgi* AP2 transcription factor, PCYB_102390, which was recently designated AP2-Q (i.e., 'quiescent') due to its enriched transcription in *P. cynomolgi* hypnozoites [22], is also detectable as transcripts but not proteins in *P. vivax* sporozoites. This may support a specific role for this transcription factor in hypnozoites. However, as Pv-AP2-Q is transcribed at an abundance (~50 TPM) at or below which ~≥50% of *P. vivax* genes are detectable as transcripts but not as proteins, the lack of detected AP2-Q protein could as likely result from the detection sensitivity of the proteomics data-set as from translation repression. Furthermore, while AP2-Q is proposed in *P. cynomolgi* as a possible hypnozoite marker in part due to its presence in *P. cynomolgi*, *P. vivax* and *P. ovale* (all of which generate hypnozoites) and reported absence from other *Plasmodium* species [22]. However, orthologs of this gene are also identified in PlasmoDB for several non-hypnozoite producing *Plasmodium* species, such as *P. knowlesi*, *P. gallinaceum* and *P. inui*, raising questions in regard to its function in these parasite species.

Lastly, while *Plasmodium* species lack a classical Golgi body, some genes (e.g., *golgi reassembly stacking protein*) functioning in protein transport between the Golgi body and the endoplasmic reticulum have been repurposed for vesicular transport and protein secretion during invasion [44]. Noting this, several homologs of genes associated with cycling of proteins between the Golgi body and the ER in other eukaryotes, including COPI-associated protein (PVX_100850), a putative STF2 (PVX_116780) and Got1 (PVX_090050) appear under translational repression in *P. vivax* sporozoites. Interestingly, in liver cells, the membrane of the parasitophorous vacuole, in which *Plasmodium* resides, often associates with the host cell ER and Golgi apparatus and may exploit this association to hijack host secretory pathways [45]. This may represent a key mechanism underpinning development in hepatocytes meriting further study.

Apoptosis-inhibition

Also notable among genes apparently translationally repressed in sporozoites are two putative Bax1 (Bcl-2 associated X protein) inhibitors (PVX_117470 and PVX_101315). Bax1 dimerizes with Bcl-2 to promote intrinsic apoptosis, leading to destruction of the mitochondrial membrane, caspase release and cell death. Bax1 inhibitor is a component of the cell stress response to prevent Bax1 from prematurely triggering cell death. When Bax1 is blocked, Bcl-2 switches from a cell-death to a pro-survival/anti-apoptotic role [46]. Intriguingly, specific suppression of mitochondrial-induced apoptosis has been demonstrated in liver-cells infected with *P. yoelii* [47] and this anti-apoptotic signal is blocked by Bcl-2 family inhibitors [48]. Orthologs of both *P. vivax* encoded Bax1 inhibitors are found in all *Plasmodium* species, suggesting a conserved function across the genus. Nonetheless, it is attractive to contemplate a potential role for these genes in promoting survival of host hepatocytes following the initial parasite invasion. Notably, the *P. cynomolgy* orthology of PVX_101315, PCYB_147290, is ~2-fold enriched in transcript abundance in schizonts compared to hypnozoites, which may indicate a role in repressing hepatocyte cell death during parasite replication rather than extending its life-span during parasite dormancy. This is to be explored.

Potential binding motif for Pv-Puf2

Research in *Toxoplasma gondii*, has identified a repetitive UGU motif in coding regions of translationally repressed genes bound by *Tg*-Puf2 [49] and, presumably, mediating repression. A similar UGU motif has been identified in the 3'UTR of *P. falciparum* transcripts (e.g., pfs25 and pfs28) and shown to bind PfPUF2 leading to their translational repression [50]. The binding motif for *Pv*-PUF2 has not been described. We found one motif (AGAT[TAC]G; Additional File 1: Figure S8) over-represented in coding regions of putatively repressed sporozoite transcripts relative to similarly highly transcribed but also translated genes e-value: $1.9e^{-9}$). We note the complementarity between AGAT and UGUA, however no over-represented motifs were detected in the 3'UTRs of these genes. Intriguingly, translational repression of *uis4* in *P. berghei* does not require the UTR [15]. It may be that the location of the *Puf2*-binding motif is somewhat flexible in *Plasmodium* and other apicomplexan species. We also identified a similarly over-represented motif ([GT]CGTC[CT]) within 500bp upstream of putatively repressed genes (p-value: 2.2e-9). It is possible this motif is a binding site for an as yet unattributed transcription factor co-ordinating genes destined for translational repression in the sporozoite. This motif is comparable to the [AG]C[AG]TGC motif identified for Pf-AP2-Sp [24], a transcription factor that is required for sporozoite development in *P. berghei* [51], and transcriptionally enriched in *P. falciparum* [24] and *P. vivax* (Additional File 2: Table S7) sporozoites relative to oocysts or blood stages respectively.

Histone modifications in *P. vivax* sporozoites

No epigenetic data are currently available for any *P. vivax* life-cycle stage. Studies of *P. falciparum* blood-stages have identified the importance of histone modifications as a primary epigenetic regulator [52, 53] and characterized key markers of heterochromatin ($H3K9me^3$) and euchromatin/transcriptional activation ($H3K4me^3$ and H3K9ac). Recently, these marks have been explored with the maturation of *P. falciparum* sporozoites in the mosquito [24]. Here, we characterize these marks in *P. vivax* sporozoites and assess their relationship to transcript abundance. Clearly this is of particular interest as a potential mechanism for dynamic regulation of sporozoite development in human hepatocytes. We identified 1,506, 1,999 and 5,262 ChIP-seq peaks stably represented in multiple *P. vivax* sporozoite replicates and associated with $H3K9me^3$, H3K9ac and $H3K4me^3$ histone marks respectively (Fig. 2). Peak width, spacing and stability differed with histone mark type (Additional File 1: Figures S9 and S10). $H3K4me^3$ peaks were significantly broader (mean width: 1,985 bp) than H3K9 peaks, and covered the greatest breadth of the genome; 36.0% of all bases were stably associated with $H3K4me^3$ marks. This mark was also most stable among replicates, with just ~16% of bases associated with an $H3K4me^3$ not supported by more than one biological replicate. By comparison $H3K9me^3$ marks were narrowest (mean width: 796 bp) and least stable, with 46% of bases associated with this mark supported by just one replicate. Consistent with observations in *P. falciparum* $H3K9me^3$ 'heterochromatin' marks primarily clustered in telomeric and subtelometric regionsv (Additional File 1: Figure S11). In contrast, the 'euchromatin' / transcriptionally open histone marks, $H3K4me^3$ and H3K9ac clustered around genic regions and did not overlap with regions under $H3K9me^3$ suppression. Both $H3K9me^3$ and $H3K4me^3$ marks were reasonably uniformly distributed (mean peak spacing ~500bp for each) within their respective regions of the genome. In contrast, H3K9ac peaks were spaced farther apart (mean: ~2kb), but also with a greater variability in spacing (likely reflecting their association with promoter regions [54]). The instability of $H3K9me^3$ may reflect its use in *Plasmodium* for regulating variegated expression of contingency genes from multigene families whose members have overlapping and redundant functions [55] and confer phenotypic plasticity [56].

Genes under histone regulation

We explored an association between these histone marks and the transcriptional behaviour of protein coding genes (Fig. 2 and Additional File 2: Tables S13-17). 485 coding genes stably intersected with an $H3K9me^3$ mark; all are located near the ends of the chromosomal scaffolds (i.e., are (sub)telomeric). On average, these genes are transcribed at ~30 fold lower levels (mean <3 TPKMs) than genes not stably intersected by H3K9me^3 marks. These data clearly support the function of this mark in transcriptional silencing. This is largely consistent with observations in *P. falciparum* sporozoites [24], however, we observe no instances of genes that are stably marked by H3K9me^3 and moderately or highly transcribed regardless. Whether this relates to differences in epigenetic control between the species is not clear. We note that (sub)telomeric genes are overall transcriptionally silent in *P. vivax* sporozoites relative to blood-stages (Fig. 2a and 2b and Additional File 2: Tables S18-20). Consistent with observations in *P. falciparum* [52], the bulk of these genes include complex protein families, such as *vir* and *Pv-fam* genes, which function primarily in blood-stages. Also notable among the genes are several reticulocyte-binding proteins, including RBP2, 2a, 2b and 2c. Strikingly, we find no exceptions to this trend in our data, indicating the (sub)telomeres are remarkably transcriptionally silent in the sporozoite stage. By comparison, H3K4me^3 marks are stably associated with the Transcription Start Site (TSS) and/or 5' UTRs of 3,677 genes. We also identified 1,284 coding genes stably associated with an H3K9ac mark within 1kb or the TSS, with 179 of these genes stably marked also by H3K4me^3. The average transcription of these genes is 116, 180 and 199 TPKMs respectively (39, 60 and 66-fold higher than H3K9me^3 marked genes). These data support the role of these marks in transcriptional activation, the lower abundance of H3K4me^3 marker, compared with H3K9ac or H3K9ac and H3K4me^3 marked genes suggest these marks work synergistically and that H3K9ac is possibly the better single mark indicator of transcriptional activity in *P. vivax*. This is consistent with recent observations in *P. falciparum* sporozoites [24].

Interestingly, H3K9ac-marked genes ranged in transcriptional activity from the most abundantly transcribed genes to many in the lower 50% and even lowest decile of transcription. This suggests more contributes to transcriptional activation in *P. vivax* than, simply, gene accessibility through chromatin regulation. Specific activation by a transcription factor (e.g., ApiAP2s [57]) is the most obvious candidate. To explore this, we compared upstream regions (within 1kb of the TSS or up to the 3' end of the next gene upstream, whichever was less) of highly (top 10%) and lowly (bottom 10%) transcribed H3K9ac marked genes for over-represented sequence motifs that might coincide with known ApiAP2 transcription factor binding sites [58]. We identified these based on the location of the nearest stable H3K9ac peak relative to the transcription start site for each gene (Additional File 1: Figure S12). In most instances, these peaks were within 100bp of the TSS and, consistent with data from *P. falciparum* [54], *P. vivax* promoters appear to be no more than a few hundred to a maximum of 1000 bp upstream of the TSS. Exploring these regions, we identified two over-represented motifs: TGTACMA (e-value 2.7e^{-2}) and ATATTTH (e-value 3.3e^{-3}) (Fig. 2D). TGTAC is consistent with the known binding site for *Pf*-AP2-G, which regulates sexual differentiation in gametocytes [59], but its *P. vivax* ortholog (PVX_123760) is neither highly transcribed nor expressed in sporozoites. It may be that some genes encoding this domain are active in both sporozoites and gametocytes, but regulated by different mechanisms in each stage. Alternatively, this motif may represent a binding site for another, as yet uncharacterized transcription factor (e.g., PVX_083040). ATATTTH is similar to the binding motif for *Pf*-AP2-L (AATTTCC), a transcription factor that is important for liver stage development in *P. berghei* [60]. In contrast to AP2-G, *Pv*-AP2-L (PVX_081180) is in the top 10% of transcription and expression in *P. vivax* sporozoites and enriched relative to blood-stages. In *P. vivax* sporozoites, the ATATTTH motif is associated with a number of highly transcribed genes, including *lisp1* and *uis2-4*, known to be regulated by AP2-L in *P. berghei* [60] as well as many of the most highly transcribed, H3K9ac marked genes, including two *etramps* (PVX_086815 and PVX_088870), several RNA-binding proteins, including *Puf2*, *ddx5* and a dead-box helicase (PVX_123240), as well as one of the putative *bax1* inhibitors (PVX_101315). Interestingly, a number of highly transcribed and translationally repressed genes associated with the ATATTTH motif, including *uis4*, *siap2* and *pv1*, are not stably marked by H3K9ac in all replicates (i.e., there is significant variation in the placement of the H3K9ac peak or their presence/absence among replicates for these genes). It may be that additional histone modifications, for example H3K27me or H2 or H4 modifications, are involved in regulating transcription of these genes. Certainly H2A.Z, which is present in *P.*

falciparum, and controls temperature responses in plants [61] is intriguing as a potential mark regulating sporozoite fate in *P. vivax* considering the association between hypnozoite activation rate and climate [11].

Conclusions

We provide the first comprehensive study of the transcriptome, proteome and epigenome of infectious *Plasmodium vivax* sporozoites and the only study to integrate 'omics investigation of the sporozoite of any *Plasmodium* species. These data support the proposal that the sporozoite is a highly-programmed stage that is primed for invasion of and development in the host hepatocyte. Translational repression clearly plays a major role in shaping this stage, with many of the genes proposed here as being under translational repression are involved in hepatocyte infection and early liver-stage development. We highlight a major role for RNA-binding proteins, including PUF2, ALBA2/4 and, intriguingly, 'Homologue of Musashi' (HoMu). Noting that HoMu uses translational repression to regulate, in *Drosophila*, stem cell, and, in *Plasmodium*, gametocyte differentiation, it is intriguing to contemplate its potential role in setting liver-stage developmental fate. Identifying the sporozoite transcripts regulated by HoMu and other RNA binding proteins should be a key priority. As should in-depth comparative analysis using similar approaches of differences between/among relapsing and non-relapsing *Plasmodium* species, as well as, *P. vivax* field isolates with distinct, hypnozoite phenotypes. Our study provides a key foundation for understanding the early stages of hepatocyte infection and the developmental switch between liver trophozoite and hypnozoite formation. Importantly, it is a major first step in rationally prioritizing targets underpinning liver-stage differentiation for functional evaluation in humanized mouse and simian models for relapsing *Plasmodium* species and identifying novel avenues to understand and eradicate liver-stage infections.

Methods

Material collection, isolation and preparation

Nine field isolates (PvSpz-Thai 1 to 9), representing symptomatic blood-stage malaria infections were collected as venous blood (20 mL) from patients presenting at malaria clinics in Tak and Ubon Ratchatani provinces in Thailand. Each isolate was used to establish, infections in *Anopheles dirus* colonized at Mahidol University (Bangkok) by membrane feeding [13], after14-16 days post blood feeding, ~3-15 million sporozoites were harvested per field isolate from the salivary glands of up to 1,000 of these mosquitoes as per [62] and shipped in preservative (trizol (RNA/DNA) or 1% paraformaldehyde (DNA for ChiP-seq)) to the Walter and Eliza Hall Institute (WEHI).

Transcriptomics sequencing and differential analysis

Upon arrival at WEHI, messenger RNAs were purified from an aliquot (~0.5-1 million sporozoites) of each *P. vivax* field isolate as per [29] and subjected to RNA-seq on Illumina NextSeq using TruSeq library construction chemistry as per the manufacturer's instructions. Raw reads for each RNA-seq replicate are available through the Sequence Read Archive (XXX-XXX). Sequencing adaptors were removed and low quality reads trimmed and filtered using Trimmomatic v. 0.36 [63]. To remove host contaminants, processed reads were aligned, as single-end reads, to the *Anopholes dirus* wrari2 genome (VectorBase version W1) using Bowtie2[64] (--very-sensitive preset). All non-host reads were then aligned to the manually curated transcripts of the *P. vivax* P01 genome (http://www.genedb.org/Homepage/PvivaxP01) using RSEM [65] (pertinent settings: --bowtie2 --bowtie2-sensitivity-level very_sensitive --calc-ci --ci-memory 10240 --estimate-rspd --paired-end). Transcript abundance for each gene in each replicate was calculated by RSEM as raw count, posterior mean estimate expected counts (pme-EC) and transcripts per million (TPM).

Transcriptional abundance in *P. vivax* sporozoites was compared qualitatively (by ranked abundance) with previously published microarray data for *P. vivax* salivary-gland sporozoites [23]. As a further quality control, these RNA-seq data were compared also with previously published microarray data for *P. falciparum* salivary-gland sporozoites [26], as well as RNA-seq data from salivary-gland sporozoites generated here for *P. falciparum* (single replicate generated from *P. falciparum* 3D7 lab cultures isolated from *Anopholes stephensi* and processed as above) and previously published for *P. yoelii* [25]. RNA-seq data from these additional *Plasmodium* species were (re)analysed from raw reads and transcriptional abundance for each species was determined (raw counts and pme-EC and TPM data) as described above using gene models current as of 04-10-2016 (PlasmoDB release v29). Interspecific transcriptional behaviour was qualitatively compared by relative ranked abundance in each species using TPM data for single copy orthologs (SCOs; defined in PlasmoDB) only, shared between *P. vivax* and *P. faliciparum* or shared among *P. vivax*, *P. falciparum* and *P. yoelii*.

To define sporozoite-enriched transcripts, we remapped raw reads representing early (18-24 hours post-infection (HPI)), mid (30-40 HPI) and late (42-46 HPI) *P. vivax* blood-stage infections recently published by Zhu *et al* [29] to the *P. vivax* P01 transcripts using RSEM as above. All replicate data was assessed for mapping metrics, transcript saturation and other standard QC metrics using QualiMap v 2.1.3 [66]. Differential transcription between *P. vivax* salivary-gland sporozoites and mixed blood-stages [29] was assessed using pme-EC data in EdgeR [67] (differential transcription cut-off: ≥ 2-fold change in counts per million (CPM) and a False Discovery Rate (FDR) ≤ 0.05). Pearson Chi squared tests were used to detect over-represented Pfam domains and Gene Ontology (GO) terms among differentially transcribed genes in sporozoites (Bonferroni-corrected $p < 0.05$), based on gene annotations in PlasmoDB (release v29).

Proteomic sequencing and quantitative analysis

Aliquots of ~10^7 salivary-gland sporozoites were generated from PvSpz-Thai1 and PvSpz-Thai6 isolates, purified on an Accudenz gradient per [62] and shipped on dry ice (protein) to the Center for Infectious Disease Research (CIDR). These cells were lysed in 2x Sample Buffer and their proteins separated by SDS-PAGE per [40]. For the whole proteome analysis, each gel was run out 52 mm and cut into 27-29 fractions using a grid cutter (Gel Company, San Francisco, CA). Pooled peptides in each gel fraction were reduced in dithiothreitol / ammonium bicarbonate, and digested for 4.5 hours at 36 °C in 6.25 ng/mL trypsin under vortex at 700 RPM. The supernatant was recovered and peptides were extracted by incubating the gel in 2% (v/v) acetonitrile/1% (v/v) formic acid. Supernatant after three extractions was combined with the digest supernatant, evaporated to dryness in a rotary vacuum, and reconstituted in HPLC loading buffer consisting of 2% (v/v) acetonitrile/0.2% (v/v) trifluoroacetic acid. Nanoflow liquid chromatography (nanoLC) was performed using an Agilent 1100 nano pump with electronically controlled split flow or a Proxeon Easy nLC. Peptides were separated on a column with an integrated fritted tip (360 µm outer diameter (O.D.), 75 µm inner diameter (I.D.), 15 µm I.D. tip; New Objective) packed in-house with a 20 cm bed of C18 (Dr. Maisch ReproSil-Pur C18-AQ, 120 Å, 3 µm; Ammerbuch-Entringen, Germany). Tandem mass spectrometry (MS/MS) was performed with an LTQ Velos Pro-Orbitrap Elite (Thermo Fisher Scientific). Two nanoLC-MS technical replicates were performed for each fraction, with roughly half the available sample injected for each replicate. The mass spectrometry data generated for this manuscript, along with the search parameters, analysis parameters and protein databases can be downloaded from PeptideAtlas (www.peptideatlas.org) using the identifier #####.

Mass spectrometer output files were converted to .mZML format using MSConvert version 2.2.0 (whole proteome data) or 3.0.5533 (surface-labeled data) [68] and searched with X!Tandem [69] version 2013.06.15.1 JACKHAMMER and Comet version 2015.02 rev.0.[70] MS/MS data were analyzed using the Trans-Proteomic Pipeline[71] version 4.8.0 PHILAE. Peptide spectrum matches (PSM) generated by each search engine were analyzed separately with PeptideProphet [72] and combined in iProphet.[73] Protein identifications were inferred with ProteinProphet [74]. In the case that multiple proteins were inferred at equal confidence by a set of peptides, the inference was counted as a single identification and all relevant protein IDs were listed. Only proteins with ProteinProphet probabilities corresponding to a model-estimated false discovery rate (FDR) less than 1.0 % were reported. Spectra were searched against a protein sequence database comprised of *P. vivax* P01 (version 29, www.plasmodb.org), *An. stephensi* SDA 500 (version 1.3, www.vectorbase.org), and a modified version of the common Repository of Adventitious Proteins (version 2012.01.01, www.thegpm.org/cRAP) with the Sigma Universal Standard Proteins removed and the LC calibration standard peptide [Glu-1] fibrinopeptide B appended. Label-free proteomics methods based on spectral counts (SpC) were used to identify proteins that were significantly more abundant in labeled samples compared to unlabeled controls. The SpC for a given protein in a given biological replicate was taken as the number of PSM used by ProteinProphet to make the protein inference. All SpC values were increased by one in order to give all proteins non-zero SpC values for log-transformation [75]. The spectral abundance factor (SAF) for a given protein was calculated as the quotient of the SpC and the protein's length and natural log-transformed to ln(SAF) [76]. For a more detailed description of the proteome data collection process and analysis please refer to manuscript by Swearingen *et al (submitted)*.

To identify genes likely under translational repression in the *P. vivax* sporozoite, we examined these data for genes that were highly transcribed (top 10 percentile) but for which we could find no evidence of protein expression in any sporozoite replicate. In addition, we conducted abundance ranked comparisons between the mean transcriptional abundance of each *P. vivax* gene in sporozoites (see above) and the mean quantitative abundance of its protein in our expressional data. Genes were sorted on the differential between their relative transcription and relative expression ranking to identify highly transcribed genes with substantially lower expression relative to their transcriptional abundance.

Salivary-gland sporozoite and liver-stage immunofluorescence assays (IFAs)

IFAs were performed as per [13]. Liver stages were obtained from 10µm formalin fixed paraffin embedded day 7 liver stages generated previously [13] from FRG knockout huHep mice;[13] these were deparaffinized prior to staining. Fresh salivary-gland sporozoites were fixed in acetone per [13]. All cells were incubated twice for 3 minutes in Xylene, then 100% Ethanol, and finally once for 3 minutes each in 95%, 70%, and 50% Ethanol. The cells were rinsed in DI water and permeabilized immediately in 1XTBS, containing Triton X-100 and 30% hydrogen peroxide. The cells were blocked in 5% milk in 1XTBS. The hepatocytes were stained overnight with a rabbit polyclonal LISP1 antibody (A), a rabbit polyclonal UIS4 antibody (B), and a rabbit polyclonal BIP antibody (C) in blocking buffer. The cells were washed with 1XTBS and the primary antibodies were detected with goat anti-rabbit Alexa Fluor 488 antibody (Life Technologies). The cells were washed in 1XTBS. The hepatocytes were rinsed in KMNO4 and washed in 1XTBS. The cells were incubated in DAPI for 5 minutes.

Histone ChIP sequencing and analysis

Aliquots of 2 – 6 million freshly isolated sporozoites were fixed with 1% paraformaldehyde for 10 min at 37°C and the reaction subsequently quenched by adding glycine to a final concentration of 125 mM. After three washes with PBS, sporozoite pellets were stored at -80°C and shipped to Australia. Nuclei were released from the sporozoites by dounce homogenization in lysis buffer (10 mM Hepes pH 7.9, 10 mM KCl, 0.1 mM EDTA, 0.1 mM EDTA, 1 mM DTT, 1x EDTA-free protease inhibitor cocktail (Roche), 0.25% NP40). Nuclei were pelleted by centrifugation at 21,000 g for 10 min at 4°C and resuspended in SDS lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris pH 8.1, 1x EDTA-free protease inhibitor cocktail). Chromatin was sheared into 200–1000 bp fragments by sonication for 16 cycles in 30 sec intervals (on/off, high setting) using a Bioruptor (Diagenode) and diluted 1:10 in ChIP dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris pH 8.1, 150 mM NaCl). Chromatin was precleared for 1 hour with protein A/G sepharose (4FastFlow, GE Healthcare) equilibrated in 0.1% BSA in ChIP dilution buffer. Chromatin from 3 x 10^5 nuclei was taken aside as input material. Chromatin from approximately 3 x 10^6 sporozoite nuclei was used for each ChIP. ChIP was carried out over night at 4°C with 5 µg of antibody

(H3K9me3 (Active Motif), H3K4me3 (Abcam), H3K9ac (Upstate), H4K16ac (Abcam)) and 10 µl each of equilibrated protein A and G sepharose beads (4FastFlow, GE Healthcare). After washes in low-salt, high-salt, LiCl, and TE buffers (EZ-ChIP Kit, Millipore), precipitated complexes were eluted in 1% SDS, 0.1 M $NaHCO_{3.}$ Cross-linking of the immune complexes and input material was reversed for 6 hours at 45°C after addition of 500 mM NaCl and 20 µg/ml of proteinase K (NEB). DNA was purified using the MinElute® PCR purification kit (Qiagen) and paired-end sequenced on Illumina NextSeq using TruSeq library construction chemistry as per the manufacturer's instructions. Raw reads for each ChIP-seq replicate are available through the Sequence Read Archive (XXX-XXX).

Fastq files were checked for quality using fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and adapter sequences were trimmed using cutadapt [77]. Paired end reads were mapped to the *P. vivax* P01 strain genome annotation using Bowtie2 [64]. The alignment files were converted to Bam format, sorted and indexed using Samtools [78]. ChIP peaks were called relative to input using MACS2[79] in paired end mode with a q value less than or equal to 0.01. Peaks and peak summits were converted to sorted BED files. Bedtools intersect[80] was used to identify genes that intersected H3K9me3 peaks and Bedtools closest was used to identify genes that were closest to and downstream of H3K9ac and H3K4me3 peak summits.

Sequence motif analysis

Conserved sequence motifs were identified using the program DREME [81]. Only genes in the top decile of transcription showing no evidence of protein expression in multiple salivary-gland sporozoite replicates were considered as putatively translationally repressed (n = 170). We queried coding regions and regions upstream of the transcriptional start site (TSS) for each gene, defined by Zhu *et al* [29] and/or predicted here from all RNA-seq data using the Tuxedo suite [82], for enriched sequence motifs in comparison to 170 genes found to be in the top decile of both transcriptional and expressional abundance in the same sporozoite replicates. In searching for motifs associated with highly transcribed genes with stable H3K9ac marks within 1kb of the TSS (or up to the 3' end of the next gene upstream), we compared H3K9ac marked genes in the top decile of transcription to the same number of H3K9ac marked genes in the bottom decile of transcription. In both instances, an e-value threshold of 0.05 was considered the minimum threshold for statistical significance.

Author contributions: Study design and development: Ivo Muller[1,2,3] (IM), Aaron R. Jex[1,3,4] (AJ), Stefan H. I. Kappe[5] (SK) and Sebastian A. Mikolajczak[5] (SM); Parasite collection and sporozoite production and purification: Jetsumon Sattabongkot[7] (JSP), Rapatbhorn Patrapuvich[6] (RP), SK, SM, Scott Lindner[8] (SL) and Erika L. Flannery[5] (EF); DNA/RNA isolation and sequence library preparation: Cristian Koepfli[1] (CK) and EF; Transcriptomics analysis: AJ, Brendan Ansell[4] (BA) and Anita Lerch[1] (AL); Proteomics analysis: Kristian Swearingen[5] (KS), Robert Moritz (RM)[9] SL, SM and EF; ChIP-seq preparation and analyses: Michaela Petter[10] (MP) and Michael Duffy[10] (MD); Immunofluorescence assays: Vorada Chuenchob[5]; Data integration and interpretation: AJ, IM, EF, SM and SK; Manuscript preparation: AJ, IM, SM, SK, SL, and EF.

Author affiliations: 1. Population Health and Immunity Division, The Walter and Eliza Hall Institute for Medical Research, 1G Royal Parade, Parkville, Victoria, 3052, Australia; 2. Malaria: Parasites & Hosts Unit, Institut Pasteur, 28 Rue de Dr. Roux, 75015, Paris, France; 3. Department of Medical Biology, The University of Melbourne, Victoria, 3010, Australia; 4. Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Corner of Park and Flemington Road, Parkville, Victoria, 3010, Australia; 5. Center for Infectious Disease Research, 307 Westlake Avenue North, Suite 500, Seattle, WA 98109, USA; 6. Department of Global Health, University of Washington, Seattle, WA 98195, USA; 7. Mahidol Vivax Research Center, Faculty of Tropical Medicine, Mahidol University, Bangkok 10400, Thailand; 8. Department of Biochemistry and Molecular Biology, Center for Malaria Research, Pennsylvania State University, University Park, PA 16802, USA. 9. Institute for

Systems Biology, Seattle, WA, 98109, USA. 10. Department of Medicine Royal Melbourne Hospital, The Peter Doherty Institute, The University of Melbourne, 792 Elizabeth Street, Melbourne, Victoria 3000, Australia. 11. Institute of Microbiology, University Hospital Erlangen, Erlangen 91054, Germany.

Acknowledgements: The authors acknowledge funding from the National Health and Medical Research Council (NHMRC, APP1021544, 1043345, & 1092789), the Australian Research Council (ARC), the Victorian State Government Operational Infrastructure Support and Australian Government National Health and Medical Research Council Independent Research Institute Infrastructure Support Scheme, the Ian Potter Foundation, the National Institute of Health, the Bill and Melinda Gates Foundation, the US Department of Defense (W81XWH-15-1-0249) and the Office of the Assistant Secretary of Defence for Health Affairs through the Peer Reviewed Medical Research Program (PRMRP).

Competing Interests: The authors declare that no author of this manuscript has a competing financial or non-financial interest related to this work.

References

1. Organization WH: World Malaria Report 2015. WHO, Geneva. 2015.
2. Feachem RG, Phillips AA, Hwang J, Cotter C, Wielgosz B, Greenwood BM, et al: Shrinking the malaria map: progress and prospects. Lancet. 2010;376(9752):1566-78.
3. Price RN, Douglas NM, Anstey NM: New developments in *Plasmodium vivax* malaria: severe disease and the rise of chloroquine resistance. Curr Opin Infect Dis. 2009;22(5):430-5.
4. Baird KJ: Malaria caused by *Plasmodium vivax*: recurrent, difficult to treat, disabling, and threatening to life - averting the infectious bite preempts these hazards. Pathogens Global Health. 2013;107475-9.
5. Sattabongkot J, Tsuboi T, Zollner GE, Sirichaisinthop J, Cui L: *Plasmodium vivax* transmission: chances for control? Trends Parasitol. 2004;20(4):192-8.
6. Mueller I, Galinski MR, Baird JK, Carlton JM, Kochar DK, Alonso PL, et al: Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite. Lancet Infect Dis. 2009;9(9):555-66.
7. Lindner SE, Miller JL, Kappe SH: Malaria parasite pre-erythrocytic infection: preparation meets opportunity. Cell Microbiol. 2012;14(3):316-24.
8. Mota MM, Pradel G, Vanderberg JP, Hafalla JC, Frevert U, Nussenzweig RS, et al: Migration of *Plasmodium* sporozoites through cells before infection. Science. 2001;291(5501):141-4.
9. Shin SC, Vanderberg JP, Terzakis JA: Direct infection of hepatocytes by sporozoites of *Plasmodium berghei*. J Protozool. 1982;29(3):448-54.
10. Lysenko AJ, Beljaev A, Rybalka V: Population studies of *Plasmodium vivax*: 1. The theory of polymorphism of sporozoites and epidemiological phenomena of tertian malaria. Bulletin WHO. 1977;55(5):541.
11. White NJ: Determinants of relapse periodicity in *Plasmodium vivax* malaria. Malar J. 2011;10297.
12. Price RN, Tjitra E, Guerra CA, Yeung S, White NJ, Anstey NM: Vivax malaria: neglected and not benign. Amer J Trop Med Hyg. 2007;77(6 Suppl):79-87.
13. Mikolajczak SA, Vaughan AM, Kangwanrangsan N, Roobsoong W, Fishbaugher M, Yimamnuaychok N, et al: *Plasmodium vivax* liver stage development and hypnozoite persistence in human liver-chimeric mice. Cell Host Microbe. 2015;17(4):526-35.

14. Mueller A-K, Camargo N, Kaiser K, Andorfer C, Frevert U, Matuschewski K, et al: *Plasmodium* liver stage developmental arrest by depletion of a protein at the parasite–host interface. Proc Natl Acad Sci U S A. 2005;102(8):3022-7.

15. Silvie O, Briquet S, Muller K, Manzoni G, Matuschewski K: Post-transcriptional silencing of UIS4 in *Plasmodium berghei* sporozoites is important for host switch. Mol Microbiol. 2014;91(6):1200-13.

16. Mackellar DC, O'Neill MT, Aly AS, Sacci JB, Jr., Cowman AF, Kappe SH: *Plasmodium falciparum* PF10_0164 (ETRAMP10.3) is an essential parasitophorous vacuole and exported protein in blood stages. Eukaryot Cell. 2010;9(5):784-94.

17. Dembele L, Franetich JF, Lorthiois A, Gego A, Zeeman AM, Kocken CH, et al: Persistence and activation of malaria hypnozoites in long-term primary hepatocyte cultures. Nat Med. 2014;20(3):307-12.

18. Malmquist NA, Moss TA, Mecheri S, Scherf A, Fuchter MJ: Small-molecule histone methyltransferase inhibitors display rapid antimalarial activity against all blood stage forms in *Plasmodium falciparum*. Proc Natl Acad Sci U S A. 2012;109(41):16708-13.

19. Josling GA, Llinas M: Sexual development in *Plasmodium* parasites: knowing when it's time to commit. Nat Rev Microbiol. 2015;13(9):573-87.

20. White MT, Karl S, Battle KE, Hay SI, Mueller I, Ghani AC: Modelling the contribution of the hypnozoite reservoir to *Plasmodium vivax* transmission. Elife. 2014;3.

21. Roobsoong W, Tharinjaroen CS, Rachaphaew N, Chobson P, Schofield L, Cui L, et al: Improvement of culture conditions for long-term in vitro culture of *Plasmodium vivax*. Malaria J. 2015;14(1):1.

22. Cubi R, Vembar SS, Biton A, Franetich JF, Bordessoulles M, Sossau D, et al: Laser capture microdissection enables transcriptomic analysis of dividing and quiescent liver stages of *Plasmodium* relapsing species. Cell Microbiol. 2017.

23. Westenberger SJ, McClean CM, Chattopadhyay R, Dharia NV, Carlton JM, Barnwell JW, et al: A systems-based analysis of *Plasmodium vivax* lifecycle transcription from human to mosquito. PLoS Negl Trop Dis. 2010;4(4):e653.

24. Gomez-Diaz E, Yerbanga RS, Lefevre T, Cohuet A, Rowley MJ, Ouedraogo JB, et al: Epigenetic regulation of *Plasmodium falciparum* clonally variant gene expression during development in *Anopheles gambiae*. Sci Rep. 2017;740655.

25. Lindner SE, Mikolajczak SA, Vaughan AM, Moon W, Joyce BR, Sullivan WJ, Jr., et al: Perturbations of *Plasmodium* Puf2 expression and RNA-seq of Puf2-deficient sporozoites reveal a critical role in maintaining RNA homeostasis and parasite transmissibility. Cell Microbiol. 2013;15(7):1266-83.

26. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, et al: Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. Genome Res. 2004;14(11):2308-18.

27. Mikolajczak SA, Silva-Rivera H, Peng X, Tarun AS, Camargo N, Jacobs-Lorena V, et al: Distinct malaria parasite sporozoites reveal transcriptional changes that cause differential tissue infection competence in the mosquito vector and mammalian host. Mol Cell Biol. 2008;28(20):6196-207.

28. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al: Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature. 2008;455(7214):757-63.

29. Zhu L, Mok S, Imwong M, Jaidee A, Russell B, Nosten F, et al: New insights into the *Plasmodium vivax* transcriptome using RNA-Seq. Sci Rep. 2016;620498.

30. Kramer S: RNA in development: how ribonucleoprotein granules regulate the life cycles of pathogenic protozoa. WIR: RNA. 2014;5(2):263-84.

31. Tucker RP: The thrombospondin type 1 repeat superfamily. Int J Biochem Cell Biol. 2004;36(6):969-74.

32. Ntumngia FB, Bouyou-Akotet MK, Uhlemann AC, Mordmuller B, Kremsner PG, Kun JF: Characterisation of a tryptophan-rich *Plasmodium falciparum* antigen associated with merozoites. Mol Biochem Parasitol. 2004;137(2):349-53.
33. Gubbels MJ, Vaishnava S, Boot N, Dubremetz JF, Striepen B: A MORN-repeat protein is a dynamic component of the *Toxoplasma gondii* cell division apparatus. J Cell Sci. 2006;119(Pt 11):2236-45.
34. Aly AS, Lindner SE, MacKellar DC, Peng X, Kappe SH: SAP1 is a critical post-transcriptional regulator of infectivity in malaria parasite sporozoite stages. Mol Microbiol. 2011;79(4):929-39.
35. Okano H, Imai T, Okabe M: Musashi: a translational regulator of cell fate. J Cell Sci. 2002;115(7):1355-9.
36. Cui L, Lindner S, Miao J: Translational regulation during stage transitions in malaria parasites. Annals N Y Acad Sci. 2015;1342(1):1-9.
37. Lasko P: Gene regulation at the RNA layer: RNA binding proteins in intercellular signaling networks. Sci STKE. 2003;179RE6.
38. Guerreiro A, Deligianni E, Santos JM, Silva PA, Louis C, Pain A, et al: Genome-wide RIP-Chip analysis of translational repressor-bound mRNAs in the *Plasmodium* gametocyte. Genome Biol. 2014;15(11):493.
39. Kappe SH, Buscaglia CA, Bergman LW, Coppens I, Nussenzweig V: Apicomplexan gliding motility and host cell invasion: overhauling the motor model. Trends in parasitology. 2004;20(1):13-6.
40. Lindner SE, Swearingen KE, Harupa A, Vaughan AM, Sinnis P, Moritz RL, et al: Total and putative surface proteomics of malaria parasite salivary gland sporozoites. Mol Cell Proteomics. 2013;12(5):1127-43.
41. Silvie O, Briquet S, Müller K, Manzoni G, Matuschewski K: Post-transcriptional silencing of UIS4 in *Plasmodium berghei* sporozoites is important for host switch. Molecular microbiology. 2014;91(6):1200-13.
42. Kelley KD, Miller KR, Todd A, Kelley AR, Tuttle R, Berberich SJ: YPEL3, a p53-regulated gene that induces cellular senescence. Cancer Res. 2010;70(9):3566-75.
43. Tuttle R, Simon M, Hitch DC, Maiorano JN, Hellan M, Ouellette J, et al: Senescence-associated gene YPEL3 is downregulated in human colon tumors. Ann Surg Oncol. 2011;18(6):1791-6.
44. Struck NS, de Souza Dias S, Langer C, Marti M, Pearce JA, Cowman AF, et al: Re-defining the Golgi complex in *Plasmodium falciparum* using the novel Golgi marker PfGRASP. J Cell Sci. 2005;118(Pt 23):5603-13.
45. Graewe S, Stanway RR, Rennenberg A, Heussler VT: Chronicle of a death foretold: *Plasmodium* liver stage parasites decide on the fate of the host cell. FEMS Microbiol Rev. 2012;36(1):111-30.
46. Bruchhaus I, Roeder T, Rennenberg A, Heussler VT: Protozoan parasites: programmed cell death as a mechanism of parasitism. Trends Parasitol. 2007;23(8):376-83.
47. Albuquerque SS, Carret C, Grosso AR, Tarun AS, Peng X, Kappe SH, et al: Host cell transcriptional profiling during malaria liver stage infection reveals a coordinated and sequential set of biological events. BMC genomics. 2009;10(1):1.
48. Kaushansky A, Metzger PG, Douglass AN, Mikolajczak SA, Lakshmanan V, Kain HS, et al: Malaria parasite liver stages render host hepatocytes susceptible to mitochondria-initiated apoptosis. Cell Death Dis. 2013;4e762.
49. Liu M, Miao J, Liu T, Sullivan WJ, Cui L, Chen X: Characterization of TgPuf1, a member of the Puf family RNA-binding proteins from *Toxoplasma gondii.* Parasites & vectors. 2014;7(1):1.

50. Miao J, Fan Q, Parker D, Li X, Li J, Cui L: Puf mediates translation repression of transmission-blocking vaccine candidates in malaria parasites. PLoS Pathog. 2013;9(4):e1003268.

51. Yuda M, Iwanaga S, Shigenobu S, Kato T, Kaneko I: Transcription factor AP2-Sp and its target genes in malarial sporozoites. Mol Microbiol. 2010;75(4):854-63.

52. Lopez-Rubio J-J, Mancio-Silva L, Scherf A: Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. Cell Host Microbe. 2009;5(2):179-90.

53. Duffy MF, Selvarajah SA, Josling GA, Petter M: Epigenetic regulation of the *Plasmodium falciparum* genome. Brief Funct Genomics. 2014;13(3):203-16.

54. Cui L, Miao J, Furuya T, Li X, Su XZ, Cui L: PfGCN5-mediated histone H3 acetylation plays a key role in gene expression in *Plasmodium falciparum*. Eukaryot Cell. 2007;6(7):1219-27.

55. Guizetti J, Scherf A: Silence, activate, poise and switch! Mechanisms of antigenic variation in *Plasmodium falciparum*. Cell Microbiol. 2013;15(5):718-26.

56. Rovira-Graells N, Gupta AP, Planet E, Crowley VM, Mok S, de Pouplana LR, et al: Transcriptional variation in the malaria parasite *Plasmodium falciparum*. Genome Res. 2012;22(5):925-38.

57. De Silva EK, Gehrke AR, Olszewski K, León I, Chahal JS, Bulyk ML, et al: Specific DNA-binding by apicomplexan AP2 transcription factors. Proc Natl Acad Sci U S A. 2008;105(24):8393-8.

58. Painter HJ, Campbell TL, Llinás M: The Apicomplexan AP2 family: integral factors regulating *Plasmodium* development. Mol Biochem Parasitol. 2011;176(1):1-7.

59. Kafsack BF, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, Campino SG, et al: A transcriptional switch underlies commitment to sexual development in human malaria parasites. Nature. 2014;507(7491):248.

60. Iwanaga S, Kaneko I, Kato T, Yuda M: Identification of an AP2-family protein that is critical for malaria liver stage development. PLoS One. 2012;7(11):e47557.

61. Boden SA, Kavanova M, Finnegan EJ, Wigge PA: Thermal stress effects on grain yield in Brachypodium distachyon occur via H2A.Z-nucleosomes. Genome Biol. 2013;14(6):R65.

62. Kennedy M, Fishbaugher ME, Vaughan AM, Patrapuvich R, Boonhok R, Yimamnuaychok N, et al: A rapid and scalable density gradient purification method for *Plasmodium* sporozoites. Malar J. 2012;11421.

63. Bolger AM, Lohse M, Usadel B: Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.

64. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9.

65. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12(1):323.

66. Okonechnikov K, Conesa A, Garcia-Alcalde F: Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32(2):292-4.

67. Nikolayeva O, Robinson MD: edgeR for differential RNA-seq and ChIP-seq analysis: an application to stem cell biology. Methods Mol Biol. 2014;115045-79.

68. Kessner D, Chambers M, Burke R, Agus D, Mallick P: ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics. 2008;24(21):2534-6.

69. Craig R, Beavis RC: TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004;20(9):1466-7.

70. Eng JK, Jahan TA, Hoopmann MR: Comet: an open-source MS/MS sequence database search tool. Proteomics. 2013;13(1):22-4.

71. Deutsch EW, Mendoza L, Shteynberg D, Slagel J, Sun Z, Moritz RL: Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. Proteomics Clin Appl. 2015;9(7-8):745-54.

72. Keller A, Nesvizhskii AI, Kolker E, Aebersold R: Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002;74(20):5383-92.

73. Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, et al: iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol Cell Proteomics. 2011;10(12):M111 007690.

74. Nesvizhskii AI, Keller A, Kolker E, Aebersold R: A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003;75(17):4646-58.

75. Hendrickson EL, Xia Q, Wang T, Leigh JA, Hackett M: Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics. Analyst. 2006;131(12):1335-41.

76. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP: Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. J Proteome Res. 2006;5(9):2339-47.

77. Martin M: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17(1):pp. 10-2.

78. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al: The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

79. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al: Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

80. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.

81. Bailey TL: DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011;27(12):1653-9.

82. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562-78.

Figures

Fig. 1 Differential transcription between *Plasmodium vivax* salivary-gland sporozoites and blood-stages. **a** BCV plot showing separation between blood-stage (black) and salivary-gland sporozoite (red) biological replicates. **b** Volcano plot of distribution of fold-changes (FC) in transcription between blood-stages and salivary-gland sporozoites relative to statistical significance threshold (False Discovery Rate (FDR) ≤ 0.05). Positive FC represents enriched transcription in the sporozoite stage. **c** Mirror plot showing pFam domains statistically significantly (FDR ≤ 0.05) over-represented in salivary-gland sporozoite enriched (red) or blood-stage enriched (black) transcripts. Scale bar truncated for presentation. * - 55 PRESAN domains are in this dataset. ** - 99 Vir domains are in this dataset.

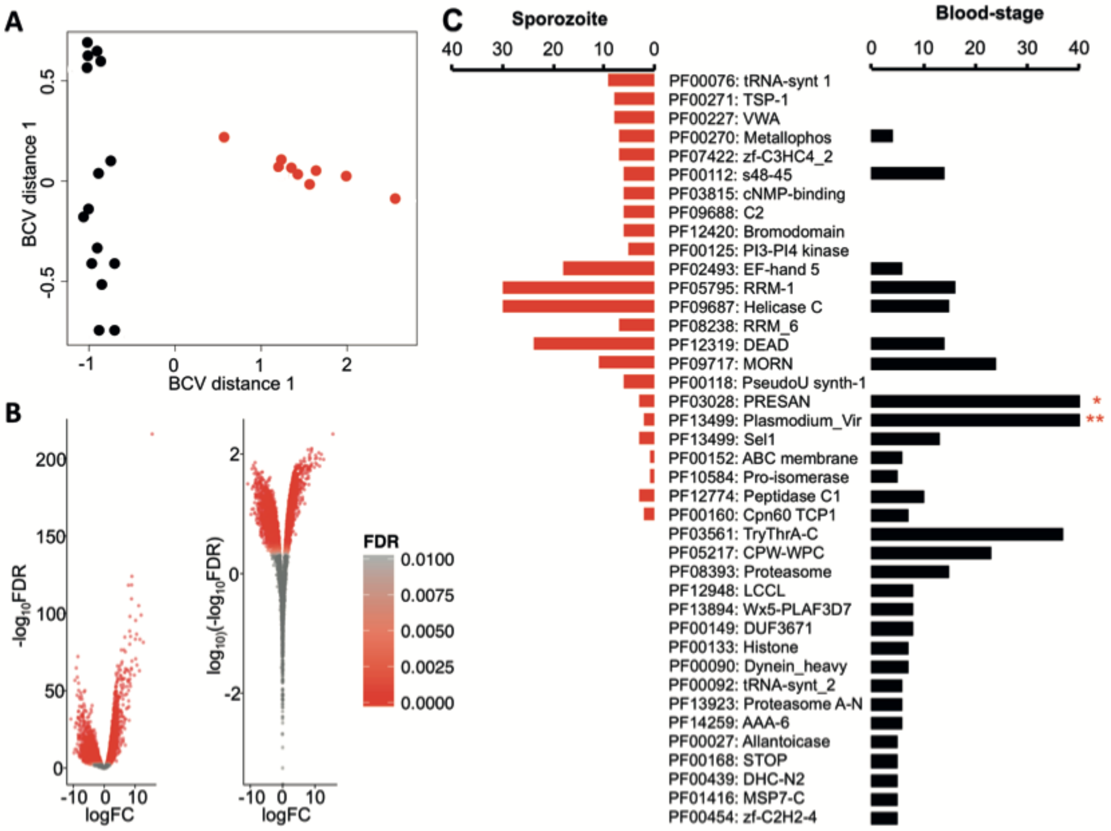
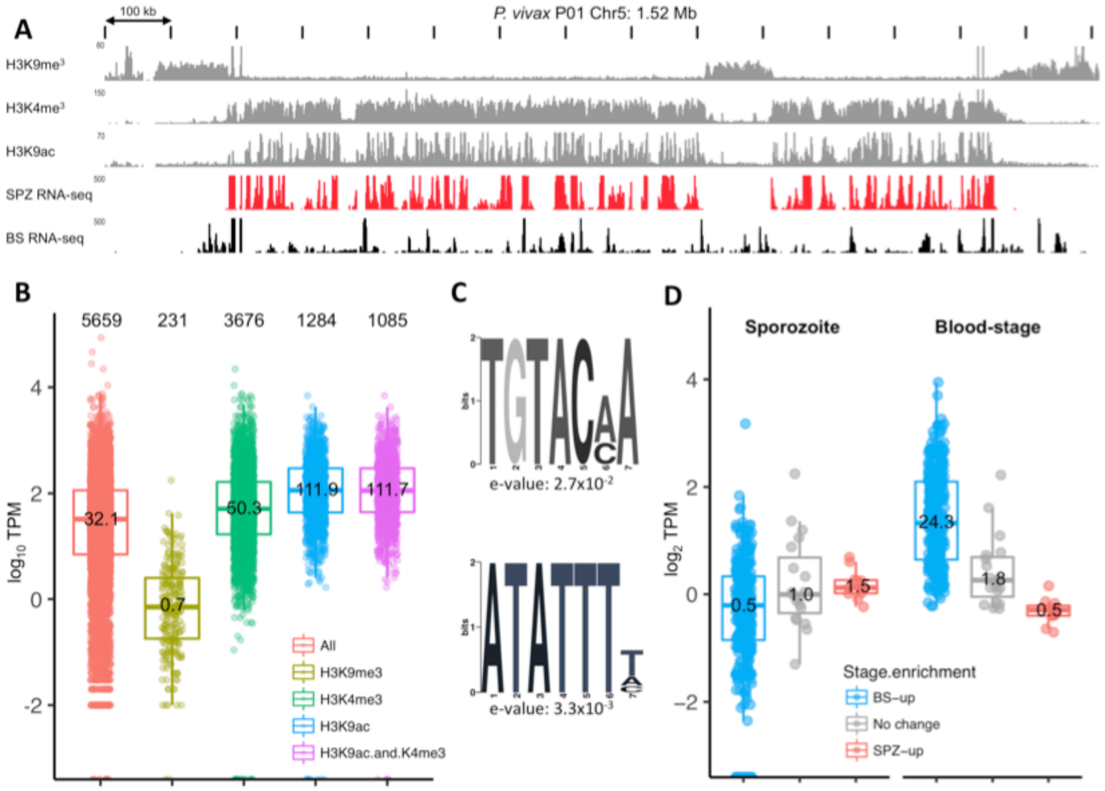


Fig. 2 Histone epigenetics relative to transcriptional behaviour in salivary-gland sporozoites. **a** Representative H3K9me^3, H3K4me^3 and H3K9ac ChIP-seq data (grey) from a representative chromosome (*P. vivax* P01 Chr5) relative to mRNA transcription in salivary-gland sporozoites (black) and blood-stages (black). Small numbers to top left of each row show data range. **b** Salivary-gland sporozoite transcription relative to nearest stable histone epigenetic marks. Numbers at the top of the figure represent total genes included in each category. Numbers within in box plot represent mean transcription in transcripts per million (TPM). **c** Sequence motifs enriched within 1kb upstream of the Transcription Start Site of highly transcribed (top 10%) relative to lowly transcribed genes associated with H3K9ac marks in salivary-gland sporozoites. **d** Relative transcription of (sub)telomeric genes in *P. vivax* salivary-gland sporozoites and blood-stages categorized by gene sets enriched in blood-stages (blue), salivary sporzoites (red) or not stage enriched (grey). Numbers in each box show mean transcription in TPM.


A
100 kb
P. vivax P01 Chr5: 1.52 Mb
H3K9me³
H3K4me³
H3K9ac
SPZ RNA-seq
BS RNA-seq
B
log10 TPM
All
H3K9me3
H3K4me3
H3K9ac
H3K9ac.and.K4me3
C
bits
TGTACAA
e-value: 2.7x10⁻²
ATATTT
e-value: 3.3x10⁻³
D
Sporozoite
Blood-stage
log2 TPM
Stage.enrichment
BS−up
No change
SPZ−up