Mechanisms and genomic features of copy number variation in malaria parasites

Adam Chase Huckaby

Charlottesville, Virginia

Bachelors of Arts, University of Colorado – Boulder 2012

A Dissertation presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Department of Biology

University of Virginia
April 2020

Acknowledgements

I would like to express my deep appreciation and gratitude to my advisor, Dr. Jennifer Guler, for her patient guidance and mentorship throughout my journey to my PhD. She has always held me to the highest standards and pushed me to be not only a better scientist but a better person and I can't thank her enough for always believing in me.

I am also deeply appreciative of my committee members, Drs. Christopher Deppmann, Yuh-hwa Wang, Michael McConnell, and Martin Wu for their guidance, thought-provoking questions and suggestions, and encouragement over the years. I looked forward to each and every meeting for the opportunity to engage with them and hone my research.

I would also like to recognize all of my amazing lab members and collaborators over the years. Basel al-Bargouthi, Drs. Karol Szlachta, Jennifer McDaniels and Maureen Carey, Audrey Brown, Shiwei Liu, Michelle Warthan, Shaun Spisak, Claire Granum, Arian Azizi, Sabrina Lingeman, Vincent Covelli, Jessica Cooper and many more for making every day working on this project and working in lab a day to remember. This work couldn't have been done without your input and support. On a similar note, I'd also like to thank Drs. William Petri, Barbara Mann, Girija Ramakrishnan and the whole Petri group for contributing ideas and feedback on my work and encouraging me to push my scientific boundaries.

Thank also should go to Drew Grainger, Mark Dombrovskiy, Katie Owsiany, all of my friends and colleagues in the UVA Biology Department, the Dotabros, my friends from the BIMS Core Course, my friends from UVA Honor, and so many more for helping me make a home away from Colorado at UVA and helping me to stay sane through my grad school journey. And last but certainly not least, I would also like to thank my family for always being there for me when I needed it and at least pretending as if you understood my science rants. Your unwavering support and patience with me through my long journey to this final degree has meant more than you can know.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

1 Introduction

Malaria is a disease that has evaded eradication despite decades of concentrated research and pooling of the world's resources. The protozoan parasite that causes malaria, *Plasmodium*, has developed resistance to every drug used thus far and innovative approaches are needed to finally eliminate this disease. Basic research, research into the fundamental underpinnings of observed phenomena, has been a goal of the malaria field since its discovery. Unfortunately, this parasite is particularly challenging to study due to its complex life-cycle, relative intractability to genetic studies, and the fact that it is a small intracellular parasite that requires stringent isolation to avoid host contamination. However, using recent technical advances to revisit old observations has led to newfound abilities to answer basic biological questions for *Plasmodium.*

The goal of my dissertation research has been to investigate the genetic mechanisms of evolution of *Plasmodium,* the parasite that causes malaria*.* It has been known for decades that a common mechanism of adaptation for *P. falciparum (*one species) is through genome structural variations such as DNA copy number variations. Copy number variations, gene duplications in particular, have been demonstrated to contribute to antimalarial drug resistance, the creation of new genes, changing gene functions, and even the creation of new species.

For this reason, I have sought to identify a mechanism conserved across *Plasmodium* species that they utilize as the first step in the development of copy number variations. In order to explore copy number variation and evolution in *Plasmodium*, my focus has primarily been on developing and adapting computational methods to analyze whole genome sequencing data and compare genome features between parasites. If we can identify the DNA repair mechanisms and sequences that are prone to mutation, we can potentially formulate a way of blocking the development of resistance to new antimalarials. These studies are key to avoiding the development of future antimalarial drug resistance.

Chapter 2: Background

2 Background

In this chapter, I summarize general background information on the parasite that causes malaria, *Plasmodium*. I discuss the epidemiology, pathology, and the life-cycle of *Plasmodium spp*. While discussing the life-cycle, I comment on methods of evolution that the parasite utilizes to rapidly adapt to challenges. I also discuss various drug treatments and their failure rates. Copy number variations are an under-appreciated mechanism of evolution that has profound effects on both genome evolution and selection that we are only now gaining the tools to thoroughly investigate.

Plasmodium species infecting humans

Malaria is caused by members of the *Plasmodium* species, which belong to the parasitic Apicomplexan phylum so named due to their possession of the apicoplast, an endosymbiotic, non-photosynthetic plastid. Other members of the Apicomplexans cause common human diseases such as Babesiosis, Cryptosporidiosis, and Toxoplasmosis. *Plasmodium* are intracellular, eukaryotic pathogens with two different hosts: a blood-feeding insect and a vertebrate. There are >2000 different species of *Plasmodium* that infect a wide-range of hosts including mammals, birds, reptiles, and amphibians. Human malaria is caused by six different species of *Plasmodium* parasites. The species that cause malaria in humans are *P. falciparum*, *P. vivax*, *P. knowlesi*, *P. malariae*, *P. ovale wallikeri*, and *P. ovale curtisi (*https://www.cdc.gov/malaria/about/disease.html).

Plasmodium falciparum is the primary cause of death by malaria and is by far the most well-studied Plasmodium species. *P. falciparum* is thought to have transitioned from gorillas to humans approximately 10000 years ago and is the likely cause of ancient records of malaria as early as 4000BCE [1]. *Plasmodium vivax* is less virulent than *P. falciparum* but was one of the first species to be discovered in 1886. *P. vivax* is the leading cause of recurrent malaria and can also cause severe symptoms[2]. *Plasmodium knowlesi* is a zoonotic species that has drawn increased research attention in the last 10 years. Zoonosis is an infectious disease that

transitions from infecting non-human animals to humans. Evidence for human infection by *P. knowlesi* was identified as early as 1932. It is most closely related to *P. vivax* and appears to have diverged 18 to 34 million years ago [3]. There have been no reports of transmission of *P. knowlesi* from humans back to mosquitoes.

Plasmodium malariae, along with *P. vivax,* was one of the first species of malarial parasite to be described by Camillo Golgi in 1886. It has a long incubation period ranging from 16-59 days and can result in life-long infections [4, 5]. Unfortunately, it is understudied due to its benign symptoms and cases are thought to be underreported. One possible reason for this is that *P. malariae* and *knowlesi* are frequently confused due to similarity in their appearance under the microscope.

Plasmodium ovale is the rarest type of human infecting malaria that has recently been discovered to consist of two subspecies and is a fascinating example of sympatric speciation [5, 6]. Sympatry is where a subspecies has evolved from a living ancestor in the same region but are unable to produce progeny with each other. The two subspecies are *Plasmodium ovale wallikeri* and *Plasmodium ovale curtisi* which are thought to have diverged 1 to 3.5 million years ago. *Plasmodium ovale* is also an understudied species of *Plasmodium.*

Malaria epidemiology

Malaria was estimated to have caused between 2-5% of all deaths in the 20th century, but we have made much progress in its eradication [1]. The total number of malaria cases in 2000 was estimated to have been ~262 million with 839,000 deaths [7]. However, by 2018 we managed to reduce that burden to an estimated 228 million cases worldwide with an estimated 405,000 deaths. The majority of deaths continue to occur in the elderly, pregnant women, and children. Children under 5 comprised 67% of all deaths in 2018 [7]. It is still estimated that ~3 billion people around the world are at risk of contracting malaria, with the vast majority of cases occurring in Africa (**Fig. 2.1**).

Figure 2.1 - Map of malaria case incidence rate (cases per 1000 population at risk) by country, 2018.

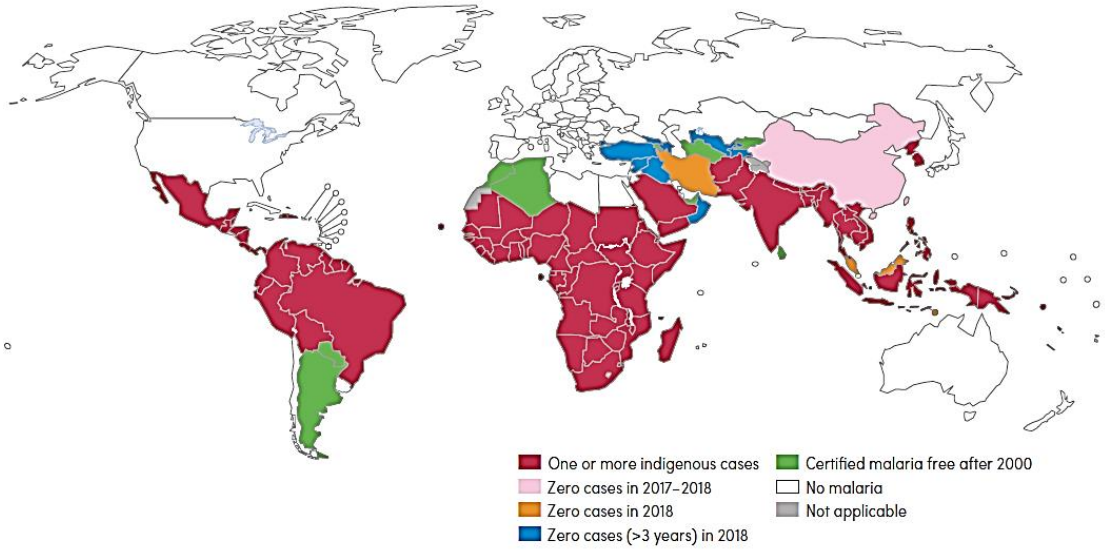


Image from World Health Organization 2019 report.
White indicates which countries have no malaria, green are countries that have recently eradicated malaria or have <0.1 cases per 1000 pop., and progressively darker colors indicate countries with higher incidence with countries in Africa >250 cases per 1000 pop.

Africa experienced 94% of cases in 2018 [8], but 49 endemic countries reported fewer than 10,000 cases each and an additional 27 countries reported fewer than 100 cases each (**Fig. 2.1**). Paraguay, Uzbekistan, Algeria, and Argentina eliminated malaria in 2018 and China, El Salvador, Iran, Malaysia, and Timor-Leste are close to elimination (**Fig. 2.1**, [7]). Total eradication appears to be stalling as the incidence rate of malaria globally has remained virtually the same since 2014 with approximately 57 cases per 1000 population at risk (**Fig. 2.1**, [7]).

Plasmodium falciparum is the species that causes the most morbidity and mortality. It causes 99.7% of cases in Africa, 50% in South-east Asia, 71% in the Eastern Mediterranean, and 65% in the Western Pacific [7]. Africa continues to have the vast majority of malaria cases with 213 million cases in 2018 (93% of all cases compared to 3.4% and 2.1% of cases in South-east Asia and the Eastern Mediterranean respectively).

Plasmodium vivax is endemic to virtually all of the same countries as *P. falciparum* and puts approximately 2.5 billion people at risk of infection [9]. The total number of cases world-wide was estimated to be 13.8 million cases in 2014. In

2018, *P. vivax* accounted for 53% of the cases in the South-east Asia region, 47% of cases in India, and 75% of malaria cases in the Americas [7].

Plasmodium knowlesi is the zoonotic malaria species found solely in South-east Asia [10]. This isolation is due to the geographic restriction of its primary hosts, the long-tailed and pig-tailed macaques. There has been an increase in reported cases of *P. knowlesi* in the past 15 years, but this is likely due to increased awareness and better detection methods [11].

The epidemiology of the other malarial species that infect humans is less well understood. *P. malariae* is most common in sub-Saharan Africa and the southwest Pacific but is also detected in Asia, the Middle East, and the Americas [5]. The total number of cases throughout the world is difficult to estimate but is likely to be relatively low compared to *P. falciparum* and *P. vivax*. *Plasmodium ovale* is endemic to Africa, New Guinea, Indonesia, the Philippines, the Middle East, India, and South-east Asia [5]. However, it is also thought to be relatively uncommon compared to *P. falciparum* and *P. vivax*.

Symptoms of malaria

The major distinctive symptom of malaria is cyclical chills followed by fever. This cycle presents with 24, 48, or 72-hour increments depending on the species of malarial parasite: 24 hours for *P. knowlesi,* 48 hours for *P. falciparum, vivax,* and *ovale,* and 72 hours for *P. malariae*. Other common symptoms include vomiting, fever, and headaches. However, the most severe symptoms that can lead to death are cerebral malaria, pulmonary edema, organ failure (kidney, liver, or spleen), and infrequently, hypoglycemia. A final major symptom is anemia, which can have long-lasting and profound effects on the health of children. Of the estimated 24 million children infected with *P. falciparum* in 2018 in sub-Saharan Africa, approximately 1.8 million of them likely had severe anemia. *P. vivax* cases typically have low blood-stage parasitemia and are frequently asymptomatic. Furthermore, people infected with *P. vivax* can relapse weeks to months later due to dormant liver-stage parasites known as hypnozoites. *P. vivax* cases typically have low blood-stage parasitemia and are frequently asymptomatic, however severe symptoms of *P. vivax* are similar to

those of *P. falciparum [9]*. *P. knowlesi* symptoms are similar to those of *P. falciparum* and *vivax [10].* Symptoms of *P. malariae* are generally much milder than other species, but it still causes the same chill and fever patterns and is associated with nephrotic syndrome from long-term infections.

Plasmodium life cycle and evolution

Plasmodium species are characterized by an extremely complex life-cycle in which they replicate through two different methods. The first is schizogony, within erythrocytes (red blood cells) which is a form of asexual replication. Schizogony involves the fission of a single cell with multiple nuclei to form daughter cells with a single set of chromosomes (haploid) that subsequently reinvade other erythrocytes. *Plasmodium spp.* also undergo sexual replication through the formation of a diploid zygote within mosquitoes. This complex life-cycle allows *Plasmodium* parasites to rapidly evolve and adapt to various challenges and methods of selection (**Fig. 2.2**, [12, 13]. Below, I walk through each step of the *Plasmodium* life cycle and how it contributes to their rapid evolution and adaptation.

Figure 2.2: Malaria life cycle.

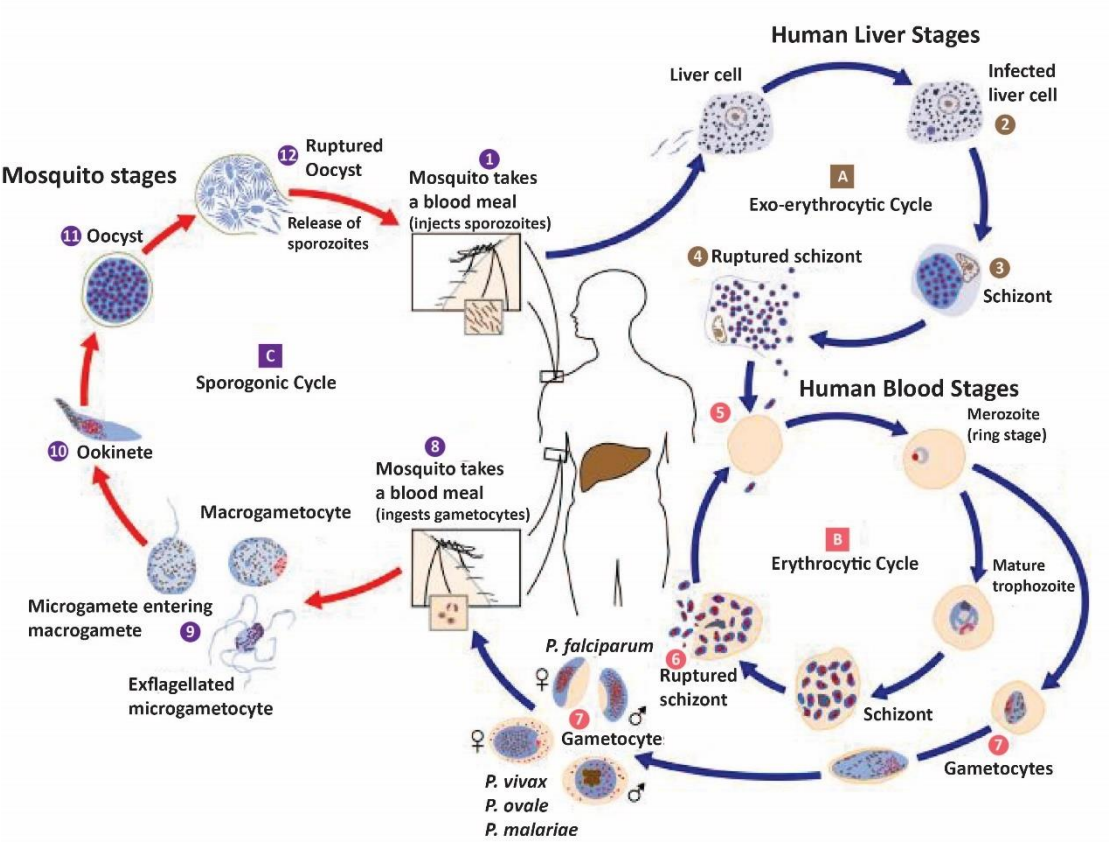


Figure adapted from https://www.cdc.gov/malaria/about/biology/

To begin, when an infected mosquito bites a human host, typically fewer than 1×10^2 *Plasmodium* sporozoites are injected into the bloodstream from the salivary glands of the mosquito (**Fig. 2.2, step 1**) [14, 15]. These sporozoites then travel to the liver to go through the exo-erythrocytic cycle, where they infect a single liver cell and replicate to create 10,000-30,000 merozoite progeny over the course of 7-10 days with no symptoms (**Fig 2.2A, steps 2-4**) [14]. During this stage, there is very little selective pressure imposed on the parasites as few current antimalarial drugs target the liver stage and the parasites are largely invisible to the immune system and therefore little need to adapt. Eventually, the liver cell ruptures to release merozoites into the bloodstream. This begins the erythrocytic cycle, in which *Plasmodium* parasites asexually reproduce for multiple rounds within red blood cells (**Fig. 2.2B, steps 5-7**).

At this stage, the *Plasmodium* merozoites invade erythrocytes to create what is known as the "ring stage." The ring stage is named as such because several hours after invasion, merozoites resemble a ring under the microscope. Once parasites

progress into the trophozoite stage, they replicate their DNA for multiple rounds within the red blood cells and finally become schizonts with multiple copies (up to 20) of their genome present within a single cell. The schizonts then rupture out of the red blood cell and release haploid merozoites to continue the cycle (**Fig 2.2B, step 6**). This red blood cell rupturing is what causes most visible symptoms of malaria. The intra-erythrocytic cycle will continue to expand the parasite population up to 10^8 – 10^{12} parasites before symptoms manifest [14].

Under stress, approximately 1-2% of the merozoites will commit to the formation of gametocytes in a process that takes ~10-12 days (**Fig. 2.2B, step 7** [14]). Each merozoite that commits to sexual reproduction can become either all male (micro) or all female (macro) gametes [16]. Once taken up by the mosquito host in a blood meal, male gametes are activated in the mosquito's gut to undergo a process known as exflagellation which involves 3 rounds of DNA replication and the creation of flagella (**Fig. 2.2C, step 9**). This replication creates 8 copies of the genome in less than 20 minutes and is one of the fastest known forms of DNA replication in eukaryotes [17]. DNA replication does not occur in female gametocytes but activation in the mosquito gut causes the macrogametes to leave the human red blood cell and fuse with the male gamete and subsequent meiosis occurs within the oocyst [18]. After fusion and meiosis, the oocyst becomes an ookinete (characterized by the ability to move), which then burrows into the midgut of the mosquito where it undergoes meiosis and ruptures to generate haploid parasites once again (**Fig 2.2C, steps 10/11**). These haploid parasites undergo sporogony to generate sporozoites. After generation of sporozoites, the oocyst ruptures and sporozoites travel to the salivary glands of the vector where they they are ready to be injected back into humans during a blood meal via a bite (**Fig. 2.2C, Steps 12 and 1** [14]).

Malaria is the quintessential example of the Red Queen Hypothesis that details the co-evolution of parasites and their hosts. The hypothesis derives its name from a quote in Lewis Carroll's *Through the Looking-Glass* by the Red queen to Alice in which she explains, "Now, here, you see, it takes all the running you can do, to keep in the same place." *Plasmodium spp.* are not only in an evolutionary arms race with their mosquito hosts but also with their vertebrate host. In order to adapt to

these dual challenges, *Plasmodium's* life cycle has allowed it to take an evolutionary path which is hypothesized to differ from normal population genetics in order to adapt quickly to virtually all possible stressors [19, 20].

P. falciparum's complicated life-cycle expands its population rapidly (up to 20x per cell cycle) and mutates readily during intraerythrocytic phases, which is then purified through natural selection from both mosquito and human host selective factors [20]. A natural consequence of this life cycle is that *Plasmodium* generates a large, heterogeneous population that allows for bet-hedging, a population-level survival strategy that maintains individuals with lower fitness that may be more fit and able to survive if the environment changes [19, 21-24]. One study estimated that in a person with 0.01% parasitamemia (% of erythrocytes infected by a *Plasmodium* parasite), they expect "~6 million base pair substitutions, 55 million indels, and 4 million newly created mosaic var exon 1 sequences" to be created every 2 days. when combined with ~262 million cases per year, this is a possibly staggering genetic reservoir and when combined with the removal of deleterious mutations through its complex life-cycle might explain the cause of *Plasmodium's* astounding ability to adapt to different host organisms, immune challenges, and drug treatments [7, 25].

Antimalarial treatment and drug resistance

Drugs have been used to treat malaria for millenia with the first drug treatment in ~168BCE [1]. A tincture using the *Artemisia annua* plant was used in ancient China and eventually led to the discovery of the current frontline antimalarial artemisinin by Youyou Tu in 1972. Quinine, which is derived from bark of the chinchoa tree, has been used to treat malaria since the 1600's and led to the discovery of chloroquine and other quinoline derivatives. While there have been effective drugs developed to combat malaria, there has not been an effective vaccine [1]. While there have been effective drugs developed to combat malaria, an effective vaccine has proved difficult to develop [1]. Malaria vaccines have low efficacy for multiple reasons; the current RTS,S vaccine only has 50% efficacy for adults and 25% efficacy for infants [7]. *Plasmodium* is an intracellular parasite and only spends a brief portion of time in the bloodstream before it reinvades new red blood cells, which makes removal by the immune system difficult. *Plasmodium* also

has very large gene families for antigenic variation and cell adhesion that frequently recombine which help explain the difficulties in creating a vaccine for malaria.

Drug resistance is a major challenge for the elimination of malaria. *P. falciparum* has developed drug resistance to virtually every drug used in the field thus far (**Table 2.1, Fig. 2.3**).

Figure 2.3 – Timeline of antimalarial resistance.

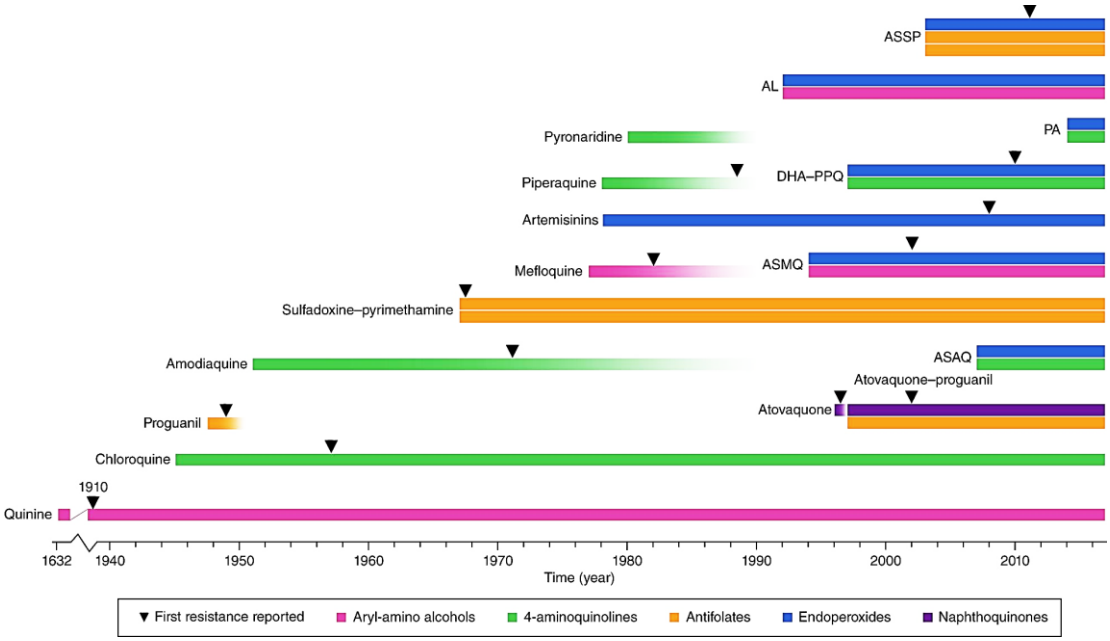


Figure from Blasco et al, Nature 2017[14].
Colors indicate the class of drug utilized and black triangles denote the first reported resistance. Single bars are monotherapies, double bars dual therapies , and triple bars are triple therapies. Quinine first had partial resistance in the early 20th century and was later replaced by chloroquine.

Table 2.1: Antimalarials used in the field and associated resistance.

Drug	Life-stage Target	Mechanism of action	Usage	Resistance reported
Chloroquine	Blood stage	Hemozoin synthesis inhibition	Treatment and prophylaxis	Yes [26]
Amodioquine	Blood stage	Hemozoin synthesis inhibition	Treatment	Yes [26]
Mefloquine	Blood stage	Hemozoin synthesis inhibition	Treatment or prophylaxis	Yes [26]
Piperaquine	Blood stage	Hemozoin synthesis inhibition	Treatment	Yes [26]
Primaquine	Liver and gametocyte stages	Oxidative damage, mechanism unknown	Treatment of P. vivax and ovale, transmission prevention of P. falciparum and P. vivax	Contested [27]

Lumefantrine	Blood stage	Hemozoin synthesis inhibition	Treatment of Plasmodium	Yes [28]
Halofantrine	Blood stage	Hemozoin synthesis inhibition	Treatment of Plasmodium	Yes [29]
Atovaquone	Liver and blood stages	Cytochrome bc1 inhibition	Treatment or prophylaxis	Yes [30]
Artemesinin derivatives	Liver and blood stages	Oxidative damage, mechanism unknown	Treatment	Yes [31, 32]
Sulphadoxine	Blood stage	DHPS inhibition	Treatment	Yes [33]
Pyrimethamine	Blood stage	DHFR inhibition	Treatment	Yes [34, 35]
Proguanil	Blood stage	DHFR inhibition	Treatment or prophylaxis	Yes [36]
Cycloguanil	Blood stage	DHFR inhibition	Treatment or prophylaxis	Yes [36]
Doxycycline	Blood stage	Protein translation in apicoplast	Prophylaxis	Yes [37]
Clindamycin	Blood stage	Protein translation in apicoplast	Prophylaxis	Yes [38]
Fosmidomycin	Blood stage	Protein translation in apicoplast	Prophylaxis	Yes [39]

Artemisinin in combination with other partner drugs is the current frontline treatment. However, there have been many recent reports of resistance to artemisinin, and resistance to the partner drugs already exists around the world. These resistant parasites are found primarily in South-east Asia. Molecular markers of artemisinin resistance have been found in Bangladesh, India, Myanmar, Thailand, Vietnam, Cambodia, and many other countries in the region [14].

Failure rates of the current front-line antimalarials for *P. falciparum* were greater than 10% for regions in South-east Asia and as high as 93% in Thailand. Artemesinin combination therapies (ACTs) utilizing artesunate, artemether, and dihydroartemesinin all have reported failure around the world. Artesunate-sulfadoxine-pyramethamine had high failure rates in Somalia and Sudan [7]. In Africa, artemether-lumefantrine, artesunate-amodiaquine, and dihydroartemisinin-piperaquine are still over 98% efficacious, and treatment with first-line antimalarials is still largely efficacious in the Americas. However, mutants in PfKelch13, the major gene responsible for artemisinin resistance, have been found around the world with significant prevalence (>5%) in Guyana, Papua New Guinea, and Rwanda.

Other species of *Plasmodium* parasites appear to have far less antimalarial drug resistance. *Plasmodium vivax* treatment remains efficacious in South-east Asia with less than 10% failure. Chloroquine treatment of *Plasmodium vivax* in Myanmar and Timor-Leste has significant failure rates at >10%, and Thailand has ~93% failure of chloroquine. Thus far, there have been no reports of *Plasmodium knowlesi, malariae,* or *ovale* antimalarial drug resistance. These species are now being studied with greater frequency, and we may see increased reports of resistance with these investigations.

Novel approaches to finding new drugs and targets are needed. One such approach was the creation of a library of novel compounds with proven efficacy that would be freely given to both malaria biologists and scientists studying other apicomplexans [40]. Another group then put *P. falciparum* under continuous treatment with these drugs to determine if they could develop resistance [41]. After whole genome sequence analysis of the drug resistance parasites, the resistance associated mutations found were not only single nucleotide polymorphisms, but also copy number variations in which a segment of the genome was amplified or deleted.

Comparative genomics and copy number variations

Comparative genomics is a method of studying evolution and adaptation by comparing gene content, linkage, and direct sequences. Comparisons are frequently represented as a phylogenetic tree, which demonstrates the relatedness of sequences in two organisms and groups based on the most common recent ancestor [42]. Another approach is the direct comparison of DNA sequences and blocks of sequence between two species. If the species are closely related, their genes are likely to be syntenic, which is when their genes and sequence motifs are grouped in a conserved, linear pattern [43]. Comparative genomics and synteny can be challenging for several reasons. Heritable large-scale changes in the overall structure between genomes can mix up sequences or the chromosomal order of genes. High quality sequence assemblies are also necessary but difficult to create due to repetitive sequences and structural variations. The overall goal of comparative

genomics is elucidation of the mechanisms of genome evolution and is therefore an important tool to study *Plasmodium* evolution.

Structural variations are changes to the overall architecture or structure of a genome compared to a reference (**Fig. 2.4**).

Figure 2.4 – CNVs are a subset of structural variations

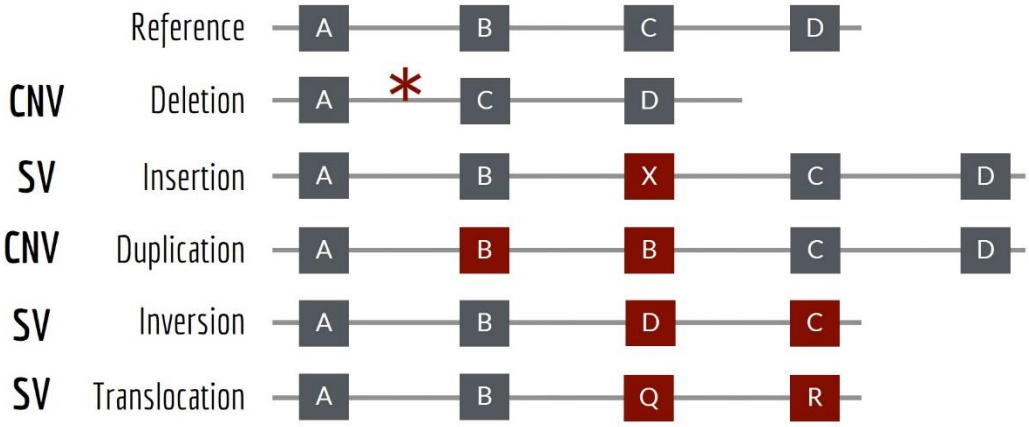


Figure adapted from a slide by Aaron Quinlan, University of Utah

By definition, copy number variations (CNVs) and structural variations are a form of genome comparison as they are increases or decreases in a segment of DNA compared to another genome. Types of structural variations include deletions, insertions, duplications, inversions, and translocations (**Fig. 2.4**). Deletions are the loss of a segment of DNA and insertions are the creation of a completely new segment of DNA. Duplications can be subdivided into tandem and interspersed duplications which differ in whether the segments are duplicated right next to each other or into another location in the genome. Inversions swap the order of segments of DNA on a single chromosome while maintaining their original location. Finally, translocations are the swapping of segments of DNA from one location to a completely separate location, either on the same chromosome or on another. CNVs are a subset of structural variations that create a change in the number of copies of a particular segment of a genome. CNVs include deletions, duplications, or amplifications of segments of DNA to >2 copies. They can be broadly grouped into two categories: small sequence repeats (bi-nucleotide or tri-nucleotide) or larger sequence repeats that can include parts of genes or even multiple genes (**Fig 2.4,** [44]**)**. CNVs, particularly whole gene amplifications, are becoming increasingly

important areas of study for human diseases and evolution in general as they allow rapid adaptation and evolution by allowing one copy of a gene to freely mutate while the other performs its original function [45, 46].

Methods for identifying copy number variations

CNVs have been a known form of genetic variance for many years [47]. Investigation of CNVs began with traditional cytogenetics [47]. Karyotyping was one of the original methods of DNA comparison which examines the size, shape, and number of chromosomes for abnormalities [47]. FISH, or fluorescent *in situ* hybridization, was the next major method of discovery [48]. FISH involves the labeling of segments of DNA with fluorescent probes for subsequent visualization under a microscope. CNVs are detected by using FISH by increased or decreased levels of fluorescence.

Methods for identifying CNVs then progressed to comparative genomic hybridization (CGH) microarrays [49]. In this approach, DNA sequence probes were hybridized to a microarray composed of many different known DNA sequences fixed to the array to create a testable library. For comparison, two genomes are then fragmented and labeled with different fluorescent colors. They are then hybridized to the microarray in equal DNA quantities to compete for binding to the DNA sequences fixed to microarray. Amplifications or deletions are identified through comparison of the binding of the two genomes to the microarray. Normal sequences would fluoresce as an equal mixture of colors of the two genomes. Amplifications would fluoresce primarily with the color of the test genome and deletions fluoresce primarily with the color of the reference genome. While this was a major step forward in the discovery of copy number variations, it still had severe limitations including only being able to detect the sequences included in the microarray design and limited resolution of boundaries and orientation of amplifications.

The current most common technology for identifying CNVs and other structural variations is the usage of Illumina short-read sequencing [50]. Illumina sequencing involves pairs of short-reads in an expected orientation (facing each other on opposite DNA strands), which can be used to identify the boundaries and

orientation of structural variations with greater sensitivity and resolution than previous technologies (**Fig. 2.5**).

Figure 2.5 – Comparison between de novo assembly, short-read, and long-read mapping approaches to identify structural variants

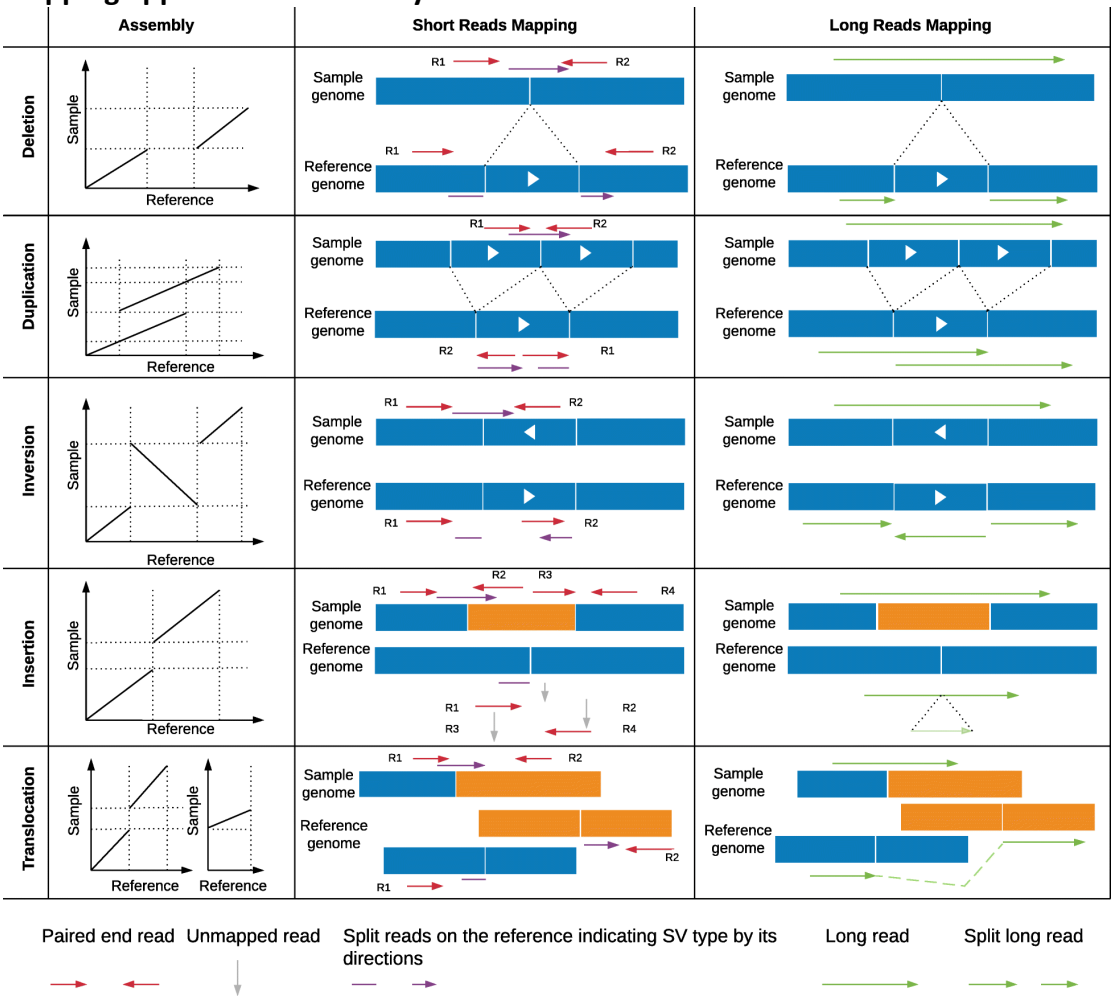


Image from Mahmoud et al. 2019 [51]. No changes made, used under the Creative Commons License - http://creativecommons.org/licenses/by/4.0/

For short-read-based mapping approaches, paired-end (red) and split reads (purple) are typically used to decipher the type size and location. In addition, the coverage can be used to improve the detection of deletions and duplications. Long-read-based mapping approaches typically leverage the alignment patterns of long reads (green) to detect the different types of SVs.

The first approach that can be utilized with short-read technologies is *de novo* assembly, which involves iteratively overlapping and matching short-reads to build a completely new sequence and then compare that sequence with a reference genome to identify structural variants (**Fig 2.5,** [51]). The next approach is the analysis of the orientation of paired-reads after mapping them to the reference genome [50]. Different orientations give different signatures of structural variants

(**Fig. 2.5**). Another method of analyzing the short-reads is the identification of split-reads, in which a portion of the read maps to one location in a genome and the other portion of the read maps to another location [50]. The final method that can be used to identify copy number variations is read-depth analysis, the quantifies reads mapped to a particular segment of the genome as either overrepresented or underrepresented if there are amplifications or deletions in the genome, respectively [50, 51].

Newer “third generation” sequencing technologies such as Oxford Nanopore and Pacific Biosciences are the most promising methods for identifying structural variations. These technologies create reads that are significantly longer than Illumina and can be up to 2.2Mb in length [52]. Long reads are much more likely to span the junction of a structural variant and therefore give the most confidence when identifying new structural variants. In the future, these technologies will expand our capability to accurately and sensitively identify structural and copy number variations in *Plasmodium*.

Chapter 3: Complex DNA structures trigger copy number variation across the *Plasmodium falciparum* genome

The following text, figures, and tables have been adapted from Huckaby et al. 2018 [53].

3 Complex DNA structures trigger copy number variation across the *Plasmodium falciparum* genome

Synopsis

Antimalarial resistance is a major obstacle in the eradication of the human malaria parasite, *Plasmodium falciparum.* Genome amplifications, a type of DNA copy number variation (CNV), facilitate overexpression of drug targets and contribute to parasite survival. Long monomeric A/T tracks are found at the breakpoints of many *Plasmodium* resistance-conferring CNVs. We hypothesize that other proximal sequence features, such as DNA hairpins, act with A/T tracks to trigger CNV formation. By adapting a sequence analysis pipeline to investigate previously reported CNVs, we identified breakpoints in 35 parasite clones with near single base-pair resolution. Using parental genome sequence, we predicted the formation of stable hairpins within close proximity to all future breakpoint locations. Especially stable hairpins were predicted to form near five shared breakpoints, establishing that the initiating event could have occurred at these sites. Further in-depth analyses defined characteristics of these 'trigger sites' across the genome and detected signatures of error-prone repair pathways at the breakpoints. We propose that these two genomic signals form the initial lesion (hairpins) and facilitate microhomology-mediated repair (A/T tracks) that lead to CNV formation across this highly repetitive genome. Targeting these repair pathways in *P. falciparum* may be used to block adaptation to antimalarial drugs.

Introduction

Major efforts have succeeded in eradicating malaria in North America and Europe, but have largely failed in Southeast Asia and Africa [54]. Some of the remaining challenges include a lack of accessible treatments and the widespread development of drug resistance. *Plasmodium falciparum*, the protozoan parasite

that causes the most severe form of malaria and the majority of malaria deaths, has developed resistance to all drug interventions thus far [55]. Single nucleotide polymorphisms (SNPs) are the most commonly studied genetic contribution to antimalarial drug resistance. However, chromosomal size polymorphisms, including copy number variations (CNVs) that encompass the genes of antimalarial targets or drug transporters, also play a key role in parasite survival [56].

CNVs often carry strong fitness costs due to increased cellular burden for DNA replication and alterations of metabolic flux due to differing levels of enzyme expression [57]. However, it has been proposed that in many organisms, including *P. falciparum*, the creation of redundant gene copies facilitates the accumulation of SNPs [58-61]. Studies observing both types of mutations in *Plasmodium* provide evidence that CNVs appear to eventually be lost in favor of SNPs [62-64].

Two CNVs associated with clinical antimalarial resistance encompass the genes encoding the multiple drug resistance protein 1 (*pfmdr1*) and GTP-cyclohydrolase 1 (*gch1*) [28, 65-69]. Additionally, a number of resistance-associated CNVs across many chromosomes were detected in the *P. falciparum* genome following laboratory selections with novel antimalarials [61, 62, 66, 70-77]. CNVs have also been detected in clinical *P. vivax* isolates [70, 78-81], providing evidence that this form of adaptation is not confined to *P. falciparum*.

Mechanisms leading to CNVs in *Plasmodium* are currently unknown. Due to a lack of significant sequence homology surrounding the CNV breakpoints, homologous recombination is not likely to be involved in the process. The most compelling evidence of a shared mechanism is the presence of long monomeric A/T tracks at CNV boundaries [61, 69, 79, 82, 83]. In other organisms, there is precedence for polymerase pausing and DNA double-stranded breaks (DSBs) at long mononucleotide repeats or AT/TA dinucleotide repeats [84-87]. However, in-depth characterization of multiple independently generated CNVs on chromosome 6 indicates an additional signal present that triggers amplification [61]. Specifically, two distinct CNVs were found to share a common boundary on one end, an event that is highly unlikely to occur by chance. The A/T track at this shared breakpoint is not significantly longer (37bp long compared to a mean of 33bp for all CNVs included

in our analysis) and thus other factors must be driving this repeat occurrence. Abnormal DNA structures, including hairpins and stem-loops, have also been implicated in replication fork stalling and DSBs in yeast and humans [88-93]. Therefore, we investigated whether sequences proximal to CNV breakpoints across the highly A/T-rich *P. falciparum* genome are enriched in these DNA structures.

Here, we present evidence that DNA hairpin formation is likely an initiating event in the generation of CNVs in *P. falciparum*. First, we adapted a CNV-calling pipeline to achieve near single base pair resolution to study laboratory acquired CNVs in 35 total resistant parasite clones selected with eight different antimalarials (19 parasite clones with distinct CNVs). Sequence analysis of sensitive parent genomes (before CNV generation, termed *pre-CNV*) confirmed that long A/T tracks are found at nearly all breakpoint locations and identified four additional shared breakpoints (5 in total). Computational predictions revealed stable hairpin structures in close proximity to all pre-CNV breakpoint locations. Especially stable hairpins sat close to the shared breakpoints, providing further support for a role of hairpin structures in alterations of copy number. We defined the relationship between these genomic features on a genome-wide scale and this association provided a map of CNV-capable sites available to the parasite during adaptation to countless antimalarials. These 'trigger sites' are found broadly throughout the parasite genome and would facilitate adaptation to most selective forces. Lastly, in-depth analysis of breakpoints in resistant clones (termed *post-CNV*) suggests the action of two repair pathways that utilize the A/T tracks as short stretches of homology. These findings contribute to a growing model of the mechanisms that lead to enhanced generation of CNVs across highly repetitive genomes.

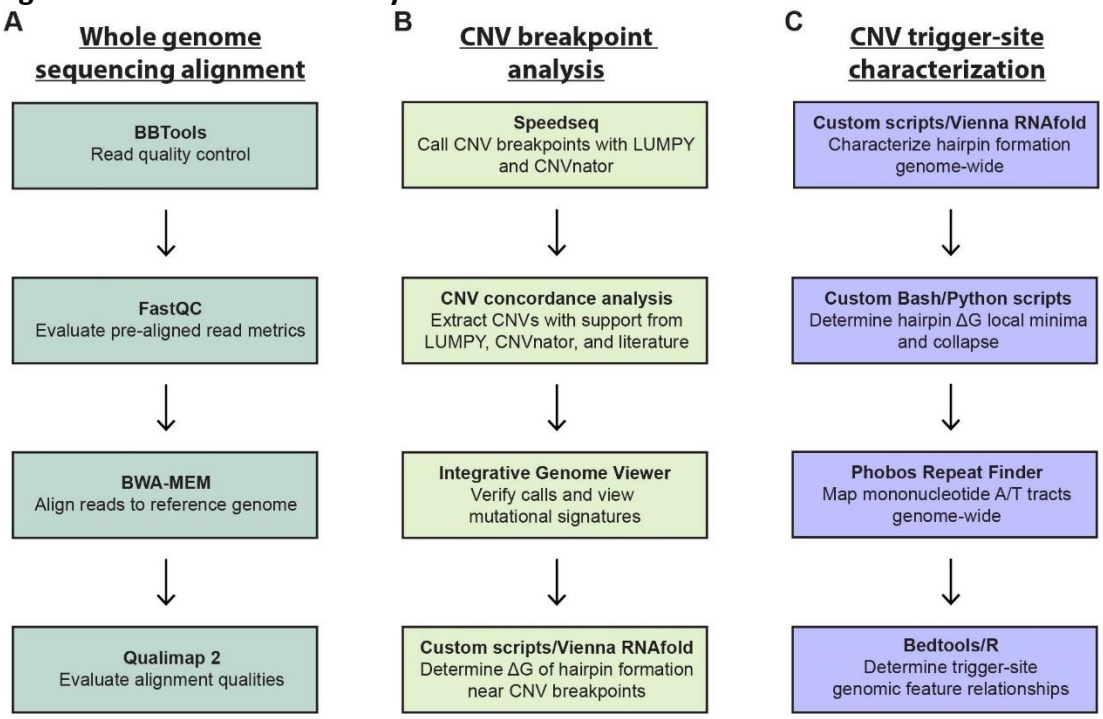
Materials and methods

Collection of genomic and breakpoint sequences.

We analyzed whole genome sequencing data to identify CNVs from in vitro haploid erythrocytic *P. falciparum* parasites that were selected with a number of different antimalarials (see details on parent and resistant clones, antimalarial target, chromosome, CNV sizes, and accession numbers in **Table 3.1** and **Table 3.2**,

[41, 61, 68, 94]). For clarity of procedures, we present a flow chart of our overall analysis methods (**Fig. 3.1**).

Figure 3.1: Bioinformatic analysis of *Plasmodium* CNVs.



A. Alignment of whole genome sequencing reads starts with BBTools to remove low quality bases or adapter sequences and verify correct pairing of reads. The resulting "clean" paired reads are evaluated by FastQC for overrepresented sequences, per base read qualities, and read length distributions. After passing read quality control, BWA-MEM is used to align "clean" paired reads to the *3d7 Plasmodium falciparum* reference genome. Qualimap 2 is then used to evaluate the alignments for mean/median read depth, paired read insert distributions, and mapping quality. **B.** After passing mapping quality control, Speedseq is used to call structural variants and CNVs with support from LUMPY, CNVnator, and positions from previous reports. The Integrative Genome Viewer is then used to manually verify CNV calls and evaluate mutational signatures such as read-pair orientation, CNV breakpoint sequences (i.e. A/T tract length), and proximal sequence changes that arise during CNV formation. Sequences windows around verified CNV breakpoints are extracted using a combination of Bash and Python scripts to create 50bp sliding windows with a 1bp shift and submitted to Vienna RNAfold for stable hairpin prediction. **C.** For genome-wide analysis, Vienna RNAfold is used to evaluate hairpin formation across all chromosomes (excluding subtelomeric/telomeric regions 50kb from the ends). Custom Bash/Python scripts are used to find local hairpin minima to find "stable hairpin forming regions". Phobos Repeat Finder is used on the same sequences to map mononucleotide A/T tracts. After mapping mononucleotide A/T tracts and stable hairpin forming regions, Bedtools and R are used to determine trigger-site feature relationships.

Briefly, low quality bases and adapter sequences from Illumina-based whole genome sequencing of both the parent and resistant clones (**Table 3.1**) were removed using BBTools (version 35.82, https://sourceforge.net/projects/bbmap/). Uncorrectable errors were assigned low quality scores and cleaned reads were

evaluated using FastQC to check per base read qualities, sequence duplication levels, overrepresented sequences, and read length distributions [95] (**Fig. 3.1A**).

For whole genome sequencing alignments, BWA-MEM was utilized to align reads with default settings to the *3d7* reference genome (PlasmoDB release 32, **Fig. 3.1B**) [96]. Alignment quality of the resulting bam files were evaluated for mean read depth, mean mapping quality, and quartiles of paired read insert-size using Qualimap 2 (**Table 3.2**) [97]. Breakpoints of the CNVs, or locations where DNA recombination occurred to generate genome amplifications, were identified by adapting the Speedseq pipeline [98]. We used the CNVnator algorithm for automated read depth analysis and copy number estimation, the LUMPY algorithm for split-read and discordant read pair analysis, and a Bayesian analytical method to genotype structural variants and call precise breakpoints [99, 100] (see more details below). CNVnator utilizes a read depth mean-shift approach to copy number variation detection and applies additional corrections including those for GC-content bias of Illumina sequencing; for this analysis, we used default settings to calculate read depth in 100bp bins. This was recommended in the CNVnator manuscript for 30x and 100x coverage, which is the range observed in our analysis. The Speedseq pipeline extracts discordant read-pairs and split-reads that can be visualized to determine CNV orientation and type (i.e. inversion or translocation). LUMPY takes the discordant read-pairs and split-reads and calculates probability distributions of breakpoints spanning a putative DNA structural variant. As discordant read-pair and split-read analysis give greater breakpoint resolution than read depth, the resulting LUMPY breakpoint locations were evaluated for sample quality scores (>100), quantity of supporting reads (>3), and significant overlap with amplification boundaries from CNVnator and the published data (Table 3.5). CNV calls were both manually verified and visualized using IGV 2.4.10 to determine CNV type and observe mutational signatures near CNV breakpoints [101]. These breakpoint locations were used to obtain consensus sequence (1kb in total, 500bp upstream (5'end) and downstream (3'end)) from the parent line for secondary structure predictions (*pre-CNV,* see below). For clones in which whole genome sequencing was not available

(DSM1-E and -F), previously published sequence from PCR-amplification across the breakpoint was used to pinpoint precise breakpoint locations [61].

Calculating the likelihood of DNA hairpin formation.

The probability of hairpin structure formation across the desired regions was predicted essentially as previously described [102, 103]. In brief, 50bp windows were selected by shifting by 1bp across a 2kb stretch of sequence surrounding the *pre-CNV* breakpoint position in the parent genome. 50bp windows were chosen to ensure hairpin formation was possible within the Okazaki initiation zone during replication. The size of the Okazaki initiation zone is not known in *Plasmodium* but it is expected to be in the same range as other eukaryotes (300 to 1000bp [104]). Next, the Gibbs free energy (ΔG), which predicts the stability of the sequence folding on itself, was determined for each window using Vienna 2.1.9 folding prediction software with Mathews 2004 DNA folding parameters and G-quadruplexes, GU pairing, and lonely base pairs were disallowed [105]. Lonely base pairs are helices in a hairpin or stem-loop that are composed of only 1bp and do not stack on other base pairs. These structures are not energetically favorable and cannot form and are therefore excluded from analyses. During this analysis, each 50bp window was counted as a separate possible hairpin. Initially this analysis was confined to sequences from the parent genome *prior* to CNV generation (the *pre-CNV* breakpoint position). Predictions were subsequently performed on sequences from *post-CNV* breakpoint locations from resistant clones.

Defining stable hairpins.

Due to a non-normal distribution of predicted hairpin ΔG values, the ΔG cutoff of stable hairpins was determined using a randomization method: sequence from each chromosome was randomly shuffled using the EMBOSS shuffleseq function to maintain overall A/T composition and hairpins were again predicted [106]. In this analysis, 50kb of sequence on either chromosome end was trimmed to avoid highly repetitive telomeric sequences. The value of the resulting top 3% of shuffled hairpins was used as the stability cut-off for all analyses (-5.8 kCal/mol); sequences with values below this cutoff indicated a high probability of a ‘stable’ structure forming. This value is consistent with that utilized in previous *P. falciparum*

investigations [103]. Furthermore, this value is similar to the top 5% of non-shuffled hairpins (ΔG of -5.5kcal/mol in our analysis), a threshold utilized in secondary structure studies of other organisms [107].

Determining the mean ΔG profile.

The mean ΔG of folding in close proximity to CNV breakpoints (shared or all) was determined by setting the end of the A/T track breakpoint to distance zero and calculating the mean ΔG for each 50bp window as the sequence is shifted by 1bp. The 95% confidence interval of each position was calculated and then plotted using Graphpad PRISM 7 (www.graphpad.com). For comparison with sequences not associated with CNVs, this process was repeated with 36 randomly chosen A/T tracks between 20-40 bp in length from intergenic regions across the genome. This length was chosen for random analysis because these A/T tracks are similar to those associated with CNV breakpoints (mean of 33bp, **Table 3.6** and see *Evaluation of A/T track lengths across the genome*). Each random A/T track position was chosen using a random number generator to pick a line number from the bed file of all A/T tracks of this size across the genome (excluding telomeres). Due to unequal sample sizes and a non-normal distribution, the level of significance in differences was calculated using the Wilcoxon-Mann-Whitney test.

Evaluation of A/T track lengths across the genome.

A/T tracks were identified with the Phobos Repeat Finder [108], which mapped the locations and lengths of long monomeric A/T tracks >9bp across the *3d7* genome (**Fig. 3.1C**). The level of significance in differences between the two data sets was again calculated using Wilcoxon-Mann-Whitney test. This length of track was chosen based on a previous study that showed that those above 9bp were overrepresented on *P. falciparum* chromosome 2 [109]. To determine if A/T tracks were observed solely due to the high A/T content of *P. falciparum* (80.6%), we calculated the probability of observing different A/T tracks lengths based purely on nucleotide composition. Frequencies of monomeric A/T tracks of length N were calculated as follows.

The observed frequency of A and T tracks of length N were obtained using the following equation:

$$f_N^{obs} = \frac{C_N^{obs}}{l_{seq}}$$

where C_N^{obs} is the observed number of monomeric tracks of length N and l_{seq} is the length of the sequence. For each A or T track observed with length N, the corresponding expected frequency of A and T tracks was obtained from the following equation:

$$f_N^{exp} = \left(f_A^{obs}\right)^N (1 - f_A^{obs})^2 + \left(f_T^{obs}\right)^N (1 - f_T^{obs})^2$$

where f_i^{obs} is the observed frequency of any base pair i which corresponds to the overall percent base composition.

Maximum expected length for each chromosome was found using the following formula:

$$N_{exp} = \frac{\log(\frac{1}{l_{seq}(1 - f_A^{obs})^2})}{\log(f_A^{obs})} + \frac{\log(\frac{1}{l_{seq}(1 - f_T^{obs})^2})}{\log(f_T^{obs})}$$

Investigating genome-scale A/T track-hairpin relationships.

In order to assess the hairpin and A/T track relationship on a larger scale, hairpins across the entire genome were predicted as described above. Where indicated, analyses were confined to tracks >20bp as this reflects the lengths of A/T tracks found at observed CNV breakpoints (**Table 3.6**). The relationship between hairpins and long A/T tracks was then determined in genic and intergenic regions separately. This was accomplished by taking gene annotations from the *3d7* reference genome and extracting A/T tracks from regions within or outside of gene annotations utilizing the '*intersect*' and '*subtract*' bedtools functions, respectively (**Fig. 3.1C**). Distance between genic or intergenic A/T tracks to the nearest stable hairpin (either upstream or downstream) was then calculated using the 'c*losest*' function in bedtools [110]. For this analysis, the positions of the local minima of hairpins had to be identified. First, we extracted all hairpins below our significance threshold (-5.8 kCal/mol, see *Defining stable hairpins*). Then, for each set of windows

with contiguous positions below this threshold, we identified the window with the most negative value and created a data subset with these minima. If there were multiple contiguous windows with the same value, all matching windows were extracted and used for analysis. The level of significance in differences were calculated using the Wilcoxon-Mann-Whitney test. Visualization of the frequency of lengths of the A/T tracks compared to distance to stable hairpins was performed using ggplot2 in R version 3.2.4 [111, 112]. The Kolmogorov-Smirnov non-parametric test was used to compare the equality of intergenic and genic distributions.

RESULTS

CNV breakpoint features are conserved in *Plasmodium falciparum*.

We obtained sequence from *P. falciparum* clones that had been selected for resistance to novel antimalarials i*n vitro* [41, 61, 68, 113] (**Table 3.1**).

Table 3.1: Summary of CNV characteristics used in our analysis.

Antimalarial	Parent clone	Clones	Putative gene amplified (chromosome)	Amp. sizes	Data source	Accession reference
DSM1	Dd2	C	Dihydroorotate dehydrogenase (6)	~70kb	[1]	SRX326516
		D		~95kb		SRX326519
		E		~34kb		N/A
		F		~39kb		N/A
		Parent		N/A		SRX326518
Halofuginone	Dd2	HFGRII	Prolyl–tRNA synthetase (12)	~30kb	[2]	SRX158283
		HFGRIII				SRX200273
		Parent		N/A		SRX738616
MMV029272	3d7	R2B2	ABC transporter I family member, putative (1)	~62kb	[3]	SRX2479359
		R2C9				SRX2479247
		R3E7				SRX2479252
		R3F10				SRX2479375
		Parent		N/A		SRX2479354
MMV019662	3d7	1C4	Lipid/sterol:H+ symporter (1)	~95kb		SRX2479223
		2B6				SRX2479224
		2F6				SRX2479226
		3G6				SRX2479265
		F7				SRX2479256
		2D6				SRX2479372
		3B6		~99kb		SRX2479268
		1F4		~52kb		SRX2479340
		1F9				SRX2479347
		33XC3		~35kb		SRX2479331
		3C3		~51kb		SRX2479355
		3F10				SRX2479219
		2G6		~41kb		SRX2479399
		2G9				SRX2479357
		Parent		N/A		SRX2479243
MMV028038	3d7	2E3	Lipid/sterol:H+ symporter (1)	~51kb		SRX2479393
		2F10				SRX2479204
		3E9				SRX2479235
		3F5				SRX2479392
		1E10		~41kb		SRX2479244
		1E3				SRX2479242
		Parent		N/A		SRX2479243
MMV08149	Dd2	1B2	Unknown (10, 12)	~18kb, ~30kb		SRX1561330
		Parent		N/A		SRX5161067
Cladosporin	Dd2	CladoA	Lysyl tRNA Synthetase, (13)	~58kb		SRX2479289
		CladoB		~50kb		SRX2479338

		CladoC		~35kb		SRX2479378
		Parent		N/A		SRX2479309
Primaquine	Dd2	PQA11	Patatin-like phospholipase, putative (10)	~18kb		SRX2479288
		Parent		N/A		SRX2479263

N/A = whole genome sequencing not available (For DSM1 clones, CNVs were determined by PCR across breakpoints and microarrays).

After read alignment and CNV calling using an adapted Speedseq pipeline with stringent quality controls (see *Materials and Methods*), we selected sequence from 35 parasite clones that displayed high confidence CNV breakpoints for further analysis (**Table 3.2** and **Table 3.3**).

Table 3.2: *Plasmodium falciparum* CNV locations used in this study.

Clone	Data Source	CNV Chr.	CNV Start (bp w/ 95% confidence interval)	CNV End (bp w/ 95% confidence interval)	# of supporting genomes
DSM1C	**Guler et al., 2013**	**6**	**79,067 ± 0**	**152,482 ± 0**	**1**
DSM1D		**6**	**64,578± 0**	**158,152± 0**	**1**
DSM1-E		**6**	**118,425[@]**	**153,231[@]**	**1**
DSM1-F		**6**	**113,523[@]**	**152,482[@]**	**1**
HFGRII	**Herman et al., 2014**	**12**	**587,623 ± 61**	**612,922 ± 3**	**1**
HFGRIII		**12**	**589,189 ± 5**	**621,909 ± 1**	**1**
CladoA	**Manary et al., 2014**	**13**	**2,000,221 ± 11**	**2,058,842 ± 1**	**1**
CladoB		**13**	**2,004,915 ± 2**	**2,055,159 ± 1**	**1**
CladoC		**13**	**2,000,213 ± 4**	**2,022,803 ± 0**	**1**
PQA11	**Cowell et al., 2018**	**10**	**290,655 ± 0**	**308,771 ± 2**	**1**
F7		**1**	**264,317 ± 0**	**359,349 ± 0**	**6**
3B6		**1**	**264,317 ± 1**	**362,912 ± 0**	**1**
1F4		**1**	**321,511 ± 5**	**373,058 ± 9**	**2**
2G9		**1**	**321,511 ± 2**	**362,912 ± 10**	**2**
1E3		**1**	**321,511 ± 2**	**362,913 ± 9**	**2**
33XC3		**12**	**1,733,591 ± 3**	**1,768,713 ± 3**	**1**
3C3		**12**	**1,718,154 ± 3**	**1,769,038 ± 3**	**6**
R2B2		**3**	**782,909 ± 45**	**845,526 ± 0**	**4**
1B2		**10**	**285,731 ± 27**	**315,681 ± 3**	**1**
		12	**1,549,855 ± 5**	**1,567,426 ± 1**	

[@]Whole genome sequencing is not available for these two clones. Analysis was performed using locations identified by PCR sequencing.

Table 3.3: Alignment statistics and mapping quality.

Clone	# of mapped reads	% mapped of total reads	Mean coverage (reads/bp ± std. dev)	Mean Mapping Quality*	Median Insert Size*	Mean Coverage 2kb around breakpoint regions (reads/bp ± std. dev)	Mean Coverage 100bp around breakpoint regions (reads/bp ± std. dev)
DSM1-C	23,606,598	98.73	96.2 ± 68	57.0	308	465.8 ± 435.6	315.4 ± 249.9
DSM1-D	58,986,651	97.95	210.2 ± 121.7	54.82	261	872.3 ± 737.1	125,592.0 ± 256,594.6
HFGRII	35,595,585	98.02	143.9 ± 63.4	56.4	144	158.2 ± 37.9	97.1 ± 25.0
HFGRIII	35,477,690	98.18	142.4 ± 91.6	54.46	150	255.4 ± 105.2	180.0 ± 65.9
CladoA	21,978,885	100	54.7 ± 34.8	55.4	321	118.5 ± 94.2	58.8 ± 30.9
CladoB	31,695,884	100	79.5 ± 54.2	55.6	349	152.3 ± 120.4	127.2 ± 68.9
CladoC	39,472,609	100	99.3 ± 64.9	55.8	294	237.2 ± 209.0	117.0 ± 89.0
PQA11	11,056,363	100	47.8 ± 76.5	57.0	267	60.7 ± 26.5	45.9 ± 22.3
F7	26,916,655	100	111.5 ± 141.4	57.8	242	127..0 ± 70.9	73.2 ± 33.4
3B6	15,482,302	100	63.5 ± 331.6	57.8	238	60.1 ± 49.7	30.3 ± 12.0
1F4	20,833,721	100	85.0 ± 95.9	57.8	227	118.1 ± 65.8	55.9 ± 33.3
2G9	15,622,275	100	61.6 ± 77.2	57.8	182	81.3 ± 52.2	48.4 ± 29.5
1E3	28,662,532	100	106.7 ± 127.2	57.7	122	117.0 ± 85.9	63.8 ± 39.9
33XC3	20,214,893	100	80.1 ± 74.7	57.6	160	67.7 ± 31.2	60.8 ± 16.8
3C3	24,656,913	100	97.1 ± 139.9	57.8	152	84.0 ± 111.1	38.59 ± 25.5
R2B2	21,492,697	100	89.2 ± 184.0	57.7	238	114.7 ± 72.3	23.8 ± 11.0
1B2ch10	24,513,373	90.23	85.9 ± 97.3	58.4	250	104.2 ± 61.8	62.5 ± 20.3
1B2ch12	24,513,373	90.23	85.9 ± 97.3	58.4	250	94.9 ± 78.7	39.7 ± 9.5

Clones with non-unique CNVs			Mean coverage (reads/bp ± std. dev)	Mean Mapping Quality*	Median Insert Size	Mean Coverage 2kb around breakpoint regions (reads/bp ± std. dev)	Mean Coverage 100bp around breakpoint regions (reads/bp ± std. dev)
R2C9	19,750,338	100	81.7 ± 179.8	57.7	255	127.7 ± 89.9	24.3 ± 13.5
R3E7	21,127,436	100	72.0 ± 75.1	57.8	213	107.3 ± 64.5	28.2 ± 13.7
R3F10	27,855,320	100	92.6 ± 97.3	57.6	160	141.0 ± 95.4	26.2 ± 12.8
1C4	13,804,219	100	56.9 ± 94.3	56.3	220	78.0 ± 67.3	73.6 ± 47.0
2B6	21,906,529	100	91.1 ± 127.2	57.8	253	99.2 ± 69.2	57.7 ± 30.9
2F6	25,244,172	100	104.6 ± 120.4	57.8	222	119.6 ± 68.9	84.7 ± 29.1
3G6	21,155,622	100	87.2 ± 303.6	57.8	223	90.3 ± 53.3	69.0 ± 26.3
2D6	20,427,288	100	84.9 ± 88.0	57.9	247	91.6 ± 62.5	51.7 ± 19.3
1F9	33,525,957	100	132.0 ± 178.8	57.8	159	149.6 ± 85.9	77.4 ± 43.9
3F10	21,739,212	100	80.5 ± 64.5	57.8	176	107.8 ± 115.7	45.0 ± 17.8
2E3	11,170,025	100	87.4 ± 101.3	57.7	128	21.5 ± 22	11.2 ± 2.8
2F10	24936046	100	86.7 ± 55.6	57.7	144	104.4 ± 96.3	52.2 ± 20.8
3E9	27,252,221	100	92.9 ± 70.6	57.6	132	98.0 ± 95.9	57.8 ± 29.7
3F5	29,612,259	100	114.6 ± 138.7	57.7	141	107.4 ± 97.8	53.87 ± 25.4
2G6	20,240,247	100	80.2 ± 104.4	57.8	193	114.0 ± 60.4	63.9 ± 42.1
1E10	24,613,192	100	91.2 ± 114.9	57.6	122	98.3 ± 71.2	56.8 ± 31.5

*Mean mapping quality was determined excluding 50kb from each end of chromosomes to avoid telomeric DNA, max value is 60. Median insert sizes are the median distance between mapped forward and reverse reads.

Due to improved resolution, breakpoint locations were identified primarily through discordant- and split-read analysis (extracted by LUMPY). This analysis identified 19 distinct CNVs for a total of 33 CNV breakpoints: 5 were conserved between different CNVs in multiple parasite clones (termed “shared” breakpoints) and 28 were unique to their respective CNV (**Table 3.4**). In total, these breakpoints

had a median of 27 supporting split and discordant reads (range of 3 to 1025 reads, **Table 3.5**).

Table 3.4: Hairpin stability and distance relationships at CNV breakpoints

Breakpoint	ΔG of Closest Hairpin	Track-Hairpin Distance*	Hairpin Forming Sequence
DSM1F/C_3	-10.9	88	Inverted repeat
CladoA/C_5	-9.1	222	Inverted repeat
F7/3B6_5	-8.1	218	AT dinucleotide
1F4/1E3_5	-10.2	104	AT dinucleotide
3B6/1E3_3	-10.2	194	AT dinucleotide
Mean of shared	**-9.7 ± 1.0**	**165 ± 58**	NA
DSM1C_5	-6.7	216	Inverted repeat
DSM1D_5	-9.7	49	Inverted repeat
DSM1D_3	-7.1	2	Inverted repeat
DSM1E_5@	-6.3	234	AT dinucleotide
DSM1E_5@	-5.8	424	Inverted repeat
DSM1F_5@	-8.4	172	AT dinucleotide
HFGRII_5	-6.1	0	Inverted repeat
HFGRII_3	-7.3	2	Inverted repeat
HFGRIII_5	-7.1	21	Inverted repeat
HFGRIII_3	-6.7	137	AT dinucleotide
CladoA_3	-7.9	59	AT dinucleotide
CladoB5	-6.6	105	Inverted repeat
CladoB3	-6.1	14	AT dinucleotide
CladoC3	-8.3	92	AT dinucleotide
PQA11_5	-13.2	234	AT dinucleotide
PQA11_3	-8.7	212	AT dinucleotide
1F4_3	-8.9	268	AT dinucleotide
2G9_5	-13.1	104	AT dinucleotide
2G9_3	-10.2	194	AT dinucleotide
33XC3_5**	-13.1	118	AT dinucleotide
33XC3_3**	-9	30	Inverted repeat
3C3_5	-6.8	342	Inverted repeat
3C3_3	-9	261	Inverted repeat
R2B2_5	-7.3	2	Inverted repeat
R2B2_3	-10.7	95	AT dinucleotide
1B2ch10_5	-8.4	0	AT dinucleotide

1B2ch10_3	-8.3	123	AT dinucleotide
1B2ch12_5*	-8	2	AT dinucleotide
1B2ch12_3*	-7.9	171	AT dinucleotide
Mean of all	**-8.6 ± 2.0**	**132.6 ± 106.3**	NA

*Track-hairpin distance was calculated to the nearest stably predicted hairpin. NA, not applicable. _5, upstream breakpoint. _3, downstream breakpoint. @sequences derived from PCR across breakpoints. Distances of 0 have the A/T track breakpoint participating in hairpin formation. **Utilize A/T dinucleotides as the breakpoint rather than A/T tracks.

Table 3.5: Variant statistics and confidence.

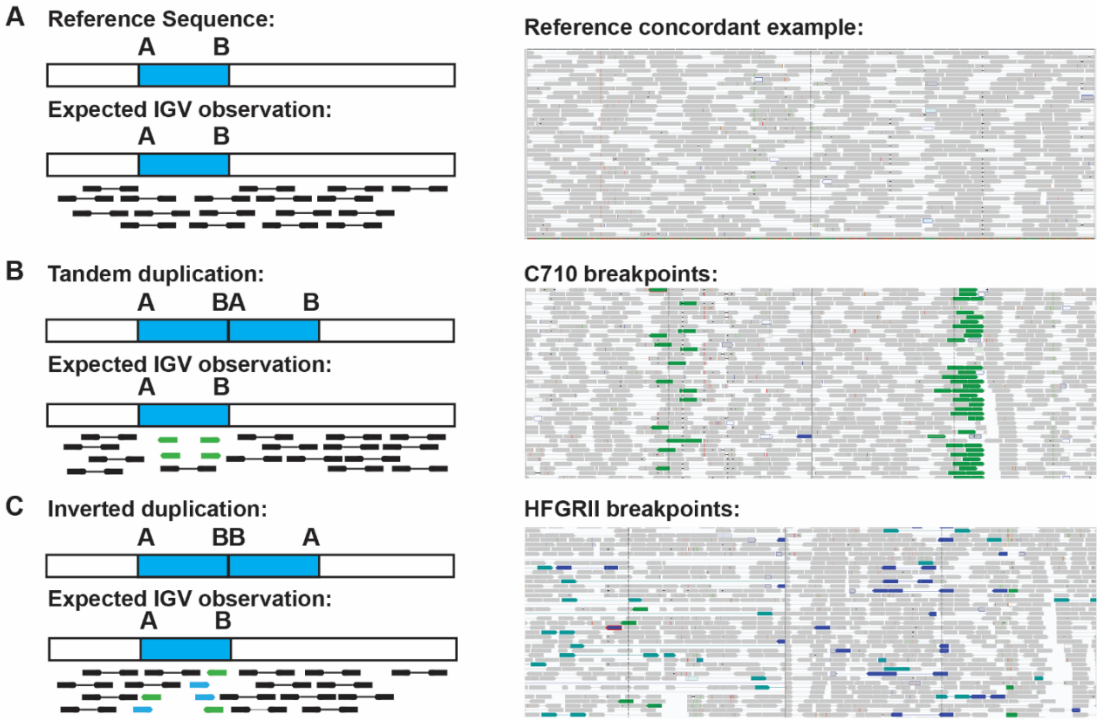
Clone	**Orientation of amplification**	**LUMPY Sample Quality**	**LUMPY PE/SR Support**	**CNVnator Start**	**CNVnator End**	**CNVnator Copy #**
DSM1-C	Tandem	18620.94	1025/0	79101	152500	7.2
DSM1-D	Tandem	8595.33	32/0	64501	158200	5.8
HFGRII	Inverted	174.29	3/0	N/A	N/A	N/A
HFGRIII	Tandem	902.33	44/0	575001	621900	2.0
CladoA	Tandem	2257.46	129/0	2000301	2058400	5.3
CladoB	Tandem	5542.93	330/0	2005701	2055100	5.0
CladoC	Tandem	7587.09	445/0	2000201	2022800	5.0
PQA11	Tandem	957.11	59/0	290001	308800	2.9
F7	Tandem	700.79	29/4	264301	359400	2.0
3B6	Tandem	338.98	39/3	264301	359300	2.2
1F4	Tandem	1574.1	11/0	321501	372900	2.3
2G9	Tandem	964.9	39/3	321501	360300	2.4
1E3	Tandem	1221.36	48/1	321601	362600	2.2
33XC3	Tandem	143.15	13/1	1733601	1768700	2.1
3C3	Tandem	307.73	7/1	1726001	1767900	2.4
R2B2	Tandem	179.06	12/0	782801	857600	2.1
1B2ch10	Tandem	231.77	22/0	285701	315700	2.4
1B2ch12	Tandem	528.74	33/0	1549901	1567200	2.3
Supporting Clones	**Orientation of amplification**	**LUMPY Quality Score**	**LUMPY PE/SR Support**	**CNVnator Start**	**CNVnator End**	**CNVnator Copy #**
R2C9	Tandem	459.26	24/0	783001	857600	3.0
R3E7	Tandem	241.39	15/0	782901	856300	2.1
R3F10	Tandem	208.27	11/0	783001	857600	2.0
1C4	Tandem	707.66	29/0	266201	359400	2.0
2B6	Tandem	709.47	30/1	264401	356400	2.1
2F6	Tandem	467.2	19/2	264301	359400	2.1
3G6	Tandem	437.43	17/1	266201	359400	2.1
2D6	Inverted	443.03	19/3	266201	359300	2.0
1F9	Tandem	1766.92	74/0	321601	372900	2.2
2G6	Tandem	1323.84	56/1	321601	364800	2.3
1E10	Tandem	969.48	38/1	321601	342600	2.2
3F10	Tandem	351.89	5/2	1718201#	1770000#	2.2
2E3	Tandem	474.66	5/1	1718201#	1768000#	2.0
2F10	Tandem	106.51	5/0	1718201	1768000	2.1
3E9	Tandem	419.7	7/1	1718201	1768100	2.0
3F5	Tandem	548.68	3/1	1718201	1768100	2.0

Amplification orientation was determined by comparing paired-end sequencing read-mate orientation and strand (Fig. S1). LUMPY sample qualities have no theoretical maximum but >100 are considered high quality calls. PE/SR= paired-end and split-read support respectively. CNVnator was unable to call read depth analysis but visual inspection of bam file showed increase in coverage indicating presence of CNV. #Clones had contiguous duplication calls from CNVnator that were combined for the overall amplification.

Read depth changes (detected by CNVnator) further confirmed these general breakpoint locations and the orientation of reads confirmed the tandem duplications at these sites (**Fig. 3.2**).

Confidence in this analysis was bolstered by overall read depth and quality scores determined for each sequenced genome. Read depth across each chromosome, excluding telomeric regions, was >40-fold (median of 87-fold); coverage across CNV breakpoints was similar with a median of 107-fold for 2kb surrounding breakpoints and a median of 57-fold for 100bp surrounding breakpoints (**Table 3.3**). The mean mapping quality scores across the genome was 57 (out of a maximum score of 60 [96]).

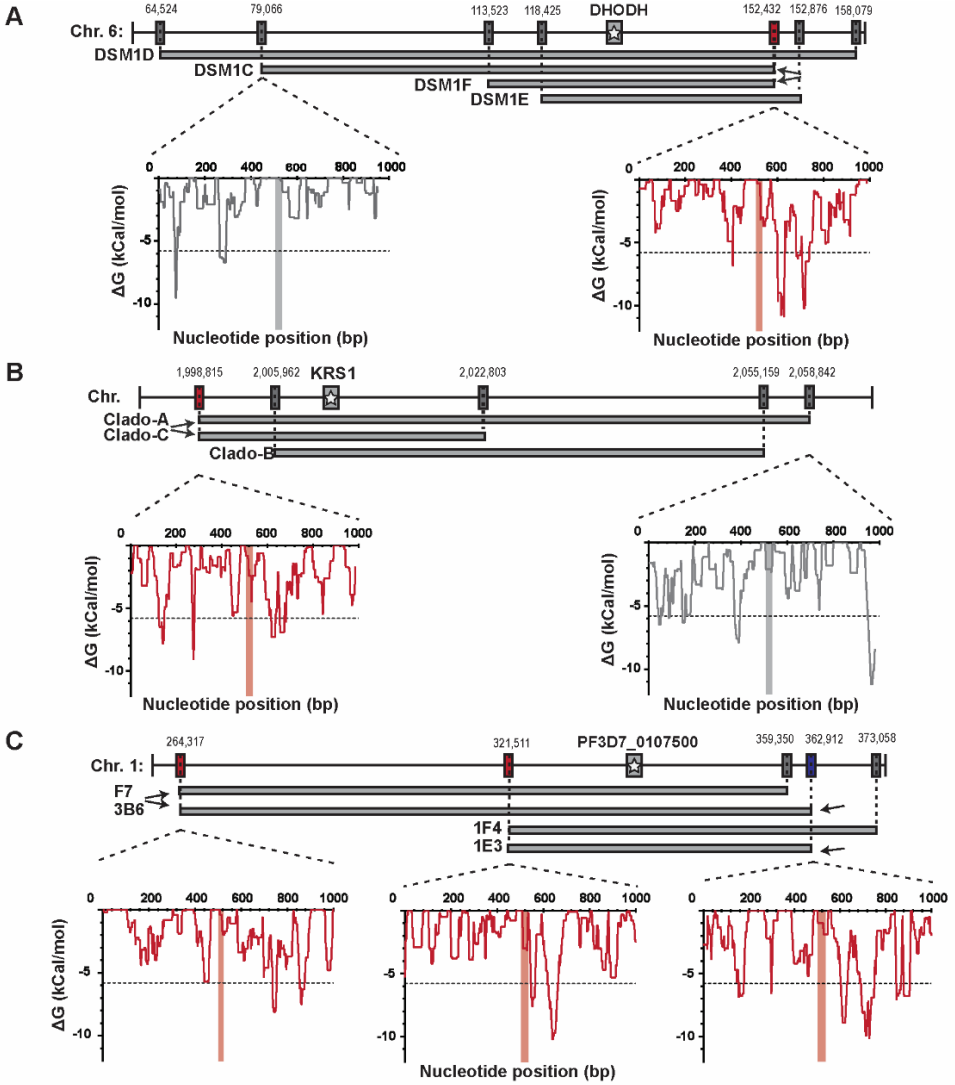
Figure 3.2: Discordant read orientation of duplications.



A. Reads aligning to the reference genome are colored based on read orientation and shown as pairs in IGV version 2.4.10. If reads match the reference sequence, they are expected to be gray and face towards each other as in the reference concordant example. **B.** If reads are found in a tandem duplication with respect to the reference sequence, they are colored green and face away from each other as in the C710 breakpoint example. These reads are shown with their pairs at their respective breakpoints and the insert sizes correspond to the size of the duplication. **C.** If reads are found in an inverted duplication with respect to the reference sequence, they are colored both blue and teal and are found facing each other and overlapping.

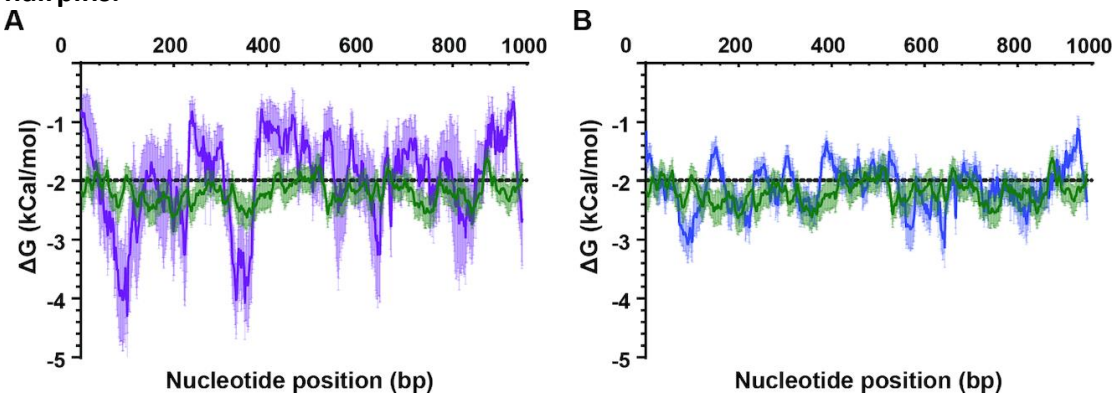
To determine whether DNA hairpins were associated with CNV breakpoints in *P. falciparum*, we went to the locations of the *shared* breakpoints in the *pre-CNV* parent genome. Two kilobases of proximal sequence were used to predict the probability of secondary structure formation nearby; a ΔG of less than -5.8kCal/mol indicated a high probability of a 'stable' structure forming from this sequence window (see *Materials and Methods*). From this focused analysis, we invariably detected extremely stable hairpins (the top 0.2% most stable structures across the entire genome, mean ΔG of <-9.7 kCal/mol) within a few hundred base pairs of the *shared* breakpoint A/T tracks (**Fig. 3.3A-C**, mean distance of 165bp +/- 58bp, **Table 3.2**). Stable hairpin structures were predicted to form by inverted repeats and AT dinucleotides present in the analyzed sequence (**Table 3.2**). In all cases, multiple stable hairpins were detected in close proximity to the shared breakpoints (see **Fig. 3.3,** where multiple peaks reach or fall below the dotted line); it was not clear which structure was contributing to CNV formation (i.e. the closest one or the most stable one). We therefore used this data to investigate whether there was a critical track-hairpin distance; we determined the mean ΔG at each bp traveling away from the A/T tracks for the 5 shared breakpoints. When we compared this profile with that from random A/T tracks across the genome that do not participate in CNV formation (see *Materials and Methods* for details about these sequences were chosen), we detected a ΔG minima for the shared breakpoints at a distance of ~80bp and ~360bp (**Fig. 3.4A,** $p < 0.05$ for both). This analysis provided evidence that stable hairpins within *very* close proximity (<400bp) to the breakpoint A/T track contributed to CNV formation.

Figure 3.3: Highly stable DNA hairpins are found near pre-CNV boundaries.



Resistant clones from various selections exhibit a range of CNV sizes but all have long A/T track breakpoints on their upstream and downstream ends. Shared breakpoints are indicated with arrows and depicted in red (boxes and plots); unique breakpoints are shown for comparison and depicted in gray (boxes and plots). The insets show the ΔG of folding for each 50bp window across 1kb of sequence surrounding the A/T track breakpoint (vertical red/grey bar at 500bp). The dotted line demarks the threshold for stable hairpin formation (ΔG of -5.8kCal/mol, see *Materials and Methods* for how this was defined). **A.** Each CNV from DSM1 resistant parasites (C, D, E, and F) encompasses the gene for the target dihydroorotate dehydrogenase (DHODH, grey bar with star). The shared 3' breakpoint from clones C and F is indicated (arrows). **B.** Each CNV from Cladosporin resistant parasites (A, B, and C) encompasses the gene for the target lysyl-tRNA synthetase (KRS1, grey bar with star). The shared 3' breakpoint from CladoA and CladoC is indicated (arrows). **C.** Each CNV from the MMV019662 and MMV028038 resistant parasites (1F4, 2G6, 3B6, and 2B6) encompasses the gene target Pf3D7_0107500, a member of the resistance-nodulation-division transporter family (grey bar with star). The shared breakpoints are indicated (arrows).

Figure 3.4: Mean free energy profiles highlight a critical distance for stable hairpins.



The mean ΔG of folding in close proximity to shared (A) and all (B) CNV breakpoints is plotted. This was done by setting the A/T track breakpoint at a distance of 0bp and calculating the mean ΔG for each window of 50bp as the sequence is shifted by 1bp (A: purple line, and B: blue line). As a comparator, the mean ΔG profile of 36 randomly chosen A/T tracks not associated with CNV formation (20-40 bp in length) was plotted (green line, see characteristics in *Materials and Methods*). Mean values with 95% confidence interval are shown. Shared breakpoints are DSM1C/F_3, CladoA/C_5, F7/3B6_5, 1F4/1E3_5, and 3B6/1E3_3 (see **Table 2).**

We extended our analysis to the remainder of the high quality CNV breakpoints identified in the above analysis (Supplementary Table S1). Although less pronounced than with the shared breakpoints, the mean ΔG profile for all CNV breakpoints indicated that the most stable structure is within ~400bp (**Fig. 3.4B**). Minima were identified at similar distances from the breakpoints and were significantly stronger than random A/T tracks ($p < 0.05$). In line with this result, stably predicted hairpins were found in very close proximity to all CNV breakpoints (mean hairpin distance of 133 $\pm$ 106bp, mean ΔG of -8.6kCal/mol). Overall, 42% of breakpoints had a highly stable structure within 100bp of the A/T track breakpoint, 60% within 150bp distance, and all but one within 400bp (**Table 3.4**). These proximal structures were frequently composed of inverted repeats or AT dinucleotide repeats (**Table 3.4**).

As has been noted before, the majority of CNV breakpoints occurred at very long A/T tracks (>20bp, **Table 3.6**). There were a few exceptions; AT dinucleotide repeats sat at both junctions for 33XC3 and 1B2 ch10 and an imperfect A/T track was found on the 3' end of the 1F4 clone (88% pure T's).

Table 3.6: Comparison of A/T track breakpoint length pre- and post-CNV formation.

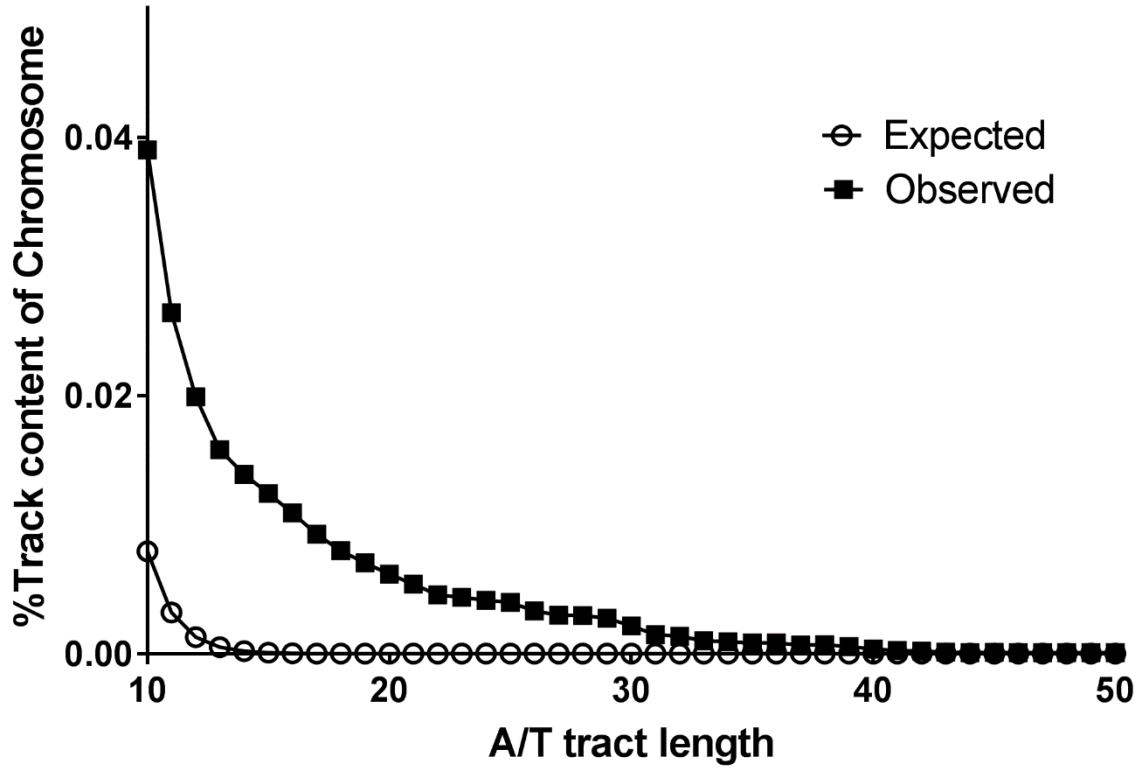
Shared Breakpoint	Pre-CNV A/T track length (bp)	Post-CNV A/T track length (bp)	% change	# of supporting split-reads	Mean phred score of split-read bases
DSM1F/C_3	37	31	-16	30	60
CladoA/C_5	40	ND	ND	ND	ND
F7/3B6_5	24	29	+21	2	60
1F4/1E3_5	33	29	-12	1	60
3B6/1E3_3	35	29	-18	3	60
Average	**34**	**30**	**-6**	**14**	**60**
Unique Breakpoint	**Pre-CNV A/T track length (bp)**	**Post-CNV A/T track length (bp)**	**% change**		
DSM1C_5	29	31	+7	30	60
DSM1D_5	38	20	-47	1	60
DSM1D_3	28	20	-29	1	60
DSM1E_5@	21	15	-29	ND	ND
DSM1E_5@	36	15	-58	ND	ND
DSM1-F@	32	25	-22	ND	ND
HFGRII_5	31	ND	ND	ND	ND
HFGRII_3	N/A^	N/A^	N/A^	N/A^	N/A^
HFGRIII_5	41	31	-24	2	60
HFGRIII_3	41	31	-24	2	60
CladoA_3	32	ND	ND	ND	ND
CladoB5	40	ND	ND	ND	ND
CladoB3	38	ND	ND	ND	ND
CladoC3	27	ND	ND	ND	ND
PQA11_5	37	26	-30	10	60
PQA11_3	26	26	0	10	60
1F4_3	25*	ND	ND	ND	ND
2G9_5	33	29	-12	3	60
2G9_3	35	29	-18	3	60
33XC3_5	N/A^	N/A^	N/A^	N/A^	N/A^
33XC3_3	N/A^	N/A^	N/A^	N/A^	N/A^
3C3_5	19	18	-5	3	60
3C3_3	34	18	-47	3	60
R2B2_5	24	ND	ND	ND	ND
R2B2_3	26	ND	ND	ND	ND
1B2ch10_5	35	ND	ND	ND	ND
1B2ch10_3	N/A^	ND	ND	ND	ND
1B2ch12_5	30	29	-3	3	60
1B2ch12_3	24	29	+21	3	60
Average	**32**	**25**	**-17**	**7**	**60**

Post-CNV A/T track length was determined through split-reads from whole genome sequencing data. ND = not determined due to absence of split-reads mapped across breakpoints. N/A^ = AT dinucleotide repeats instead of A/T tracks, * = imperfect A/T track repeat

CNV breakpoint features are enriched in intergenic regions.

We noted previously that CNV breakpoints are more often found in intergenic than genic regions [61]. To explore this further, we expanded our analysis across these two regions of the *P. falciparum* genome. Specifically, we investigated 1) the quantity and length of A/T tracks, 2) the propensity for DNA hairpin formation, as measured by ΔG of folding, and 3) the relationship between these two features. First, when compared to expected numbers, long A/T tracks (>9bp) were highly enriched across the genome (**Fig 3.5**, $p < 0.01$ for A/T tracks > 9bp).

Figure 3.5: Expected vs observed frequency of long A/T tracks.



Frequency of (# tracks observed/chromosome length) for varying A/T tract lengths on all chromosomes. For equations used in calculation, see *Materials and Methods*.

When comparing genic to intergenic regions of the genome, we found about twice as many long A/T tracks in intergenic sequences than genic (42,026 in intergenic versus 19,408 in genic*,* **Table 3.7,** $p < 0.001$). A more striking difference was observed if the quantity of very long A/T tracks (>20bp) were compared (~4-fold increase: 9509 in intergenic regions and 2410 in genic, **Table 3.7**, $p < 0.001$). Second, we predicted a greater number of stable structures (ΔG <-5.8 kCal/mol) in intergenic compared to genic regions (37,439 intergenic and 23,442 genic, **Table 3.7,** $p < 0.05$) and an increase in the mean hairpin strength of these *stable* hairpins (-7.56 kCal/mol

for intergenic compared to -7.23 kCal/mol for genic, $p < 0.01$). Lastly, we found that the distance between A/T tracks and hairpins differed greatly between genic and intergenic regions. The mean track-hairpin distance when considering long A/T tracks was 99bp in intergenic regions and 277bp in genic regions (**Table 3.7**, $p < 10^{-13}$). This trend was conserved when considering very long A/T tracks (mean of 104bp distance in intergenic and 163bp in genic, $p < 10^{-6}$).

Table 3.7: Quantification of A/T track frequency, hairpin frequency, and distance relationships across the genome.

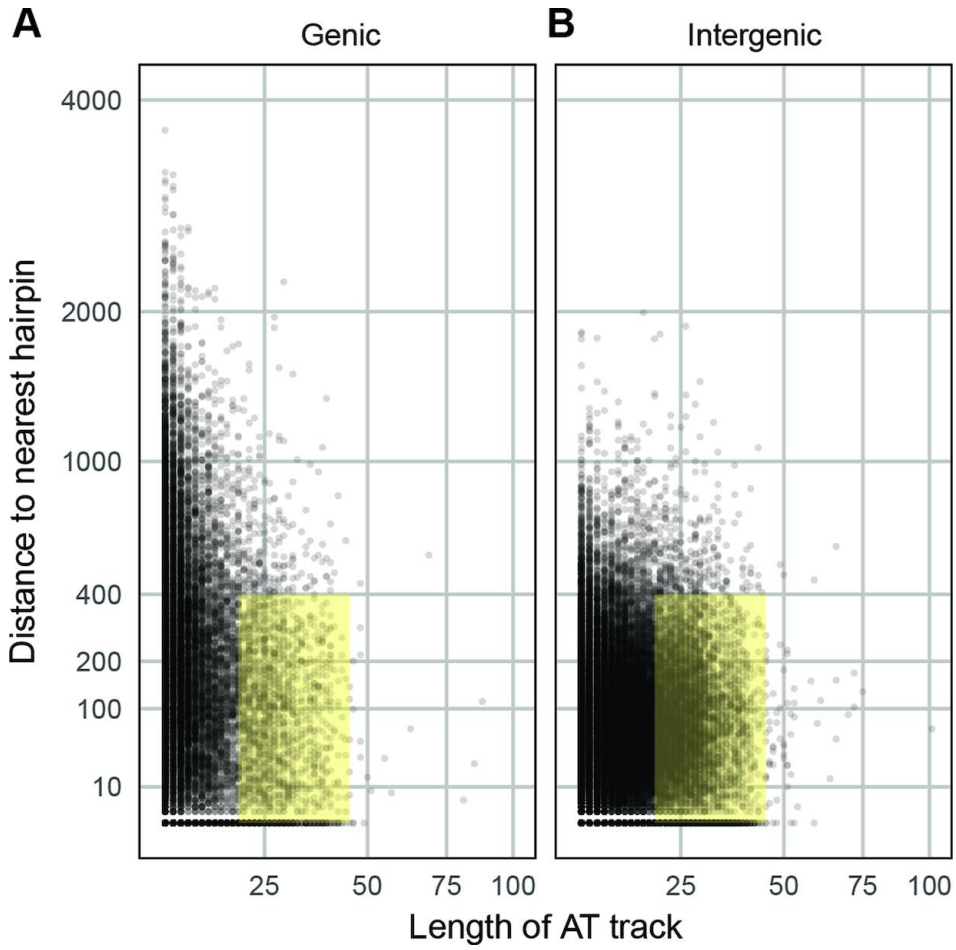
	A/T Tracks >9bp		A/T Tracks >20bp		Stable Hairpin Minima*		Mean Distance (bp): A/T tracks >9bp		Mean Distance (bp): A/T tracks >20bp	
	Genic	**Intergenic**	**Genic**	**Intergenic**	**Genic**	**Intergenic**	**Genic**	**Intergenic**	**Genic**	**Intergenic**
Chr. 1	434	1223	70	250	566	958	269	103.0	145.3	101.3
Chr. 2	742	1629	93	371	915	1484	254.1	99.6	160.6	115.4
Chr. 3	962	1824	126	401	1117	1568	249.5	98.8	143.8	105.9
Chr. 4	1040	2005	106	460	1270	1734	318.8	107.3	195.5	110.9
Chr. 5	1085	2404	131	576	1307	2085	287.1	92.5	143.0	101.9
Chr. 6	1155	2365	143	526	1447	2449	299.8	92.1	193.5	95.9
Chr. 7	1292	2350	153	532	1558	2037	302.6	102.7	190.9	105.8
Chr. 8	1274	2722	146	631	1562	2466	274.6	99.4	173.6	105.8
Chr. 9	1286	2977	167	675	1543	2604	234.3	103.3	120.3	100.3
Chr. 10	1336	3160	175	710	1657	3009	266.0	99.4	152.2	101.7
Chr. 11	1746	3821	219	858	2060	3477	273.9	95.3	170.1	100.5
Chr. 12	1935	4169	239	989	2273	3655	272.1	97.7	151.1	98.2
Chr. 13	2336	5321	283	1197	2870	4690	280.7	102.9	177.6	102.7
Chr. 14	2785	6056	359	1333	3297	5223	282.6	100.9	175.7	106.2
Total	**19408**	**42026**	**2410**	**9509**	**23442**	**37439**	**277.3**	**99.2**	**163.8**	**103.8**

*Stable hairpin minima were determined by identifying the most stably predicted structure, most negative ΔG. If contiguous windows had the same minimum, the windows were combined into the same structure feature for calculations. Distances between A/T tracks and stable hairpin minimum were calculated from the edge of A/T tracks to the edge of stable hairpin minima.

By visualizing these distributions on a whole genome scale, the disparities between the two genomic regions and the close A/T track-hairpin association in intergenic regions are emphasized (**Fig. 3.6A and B,** Kolmogorov-Smirnov test, $p < 10^{-15}$). Due to the characteristics of these features that are associated with observed CNV breakpoints, we propose that there is an optimal range for A/T track lengths (~20-40 bp) and track-hairpin distances (<400 bp) (yellow highlight in **Fig. 3.6A-B**). We defined genome positions with these characteristics as CNV "trigger sites": those locations that are competent to generate CNVs. Using these parameters, there are

9,130 intergenic and 2,222 genic trigger sites across the *P. falciparum* genome (corresponds to 19.0% of intergenic and 9.6% of genic A/T tracks).

Figure 3.6: Stable hairpins near long A/T tracks are overrepresented in the *P. falciparum* genome.



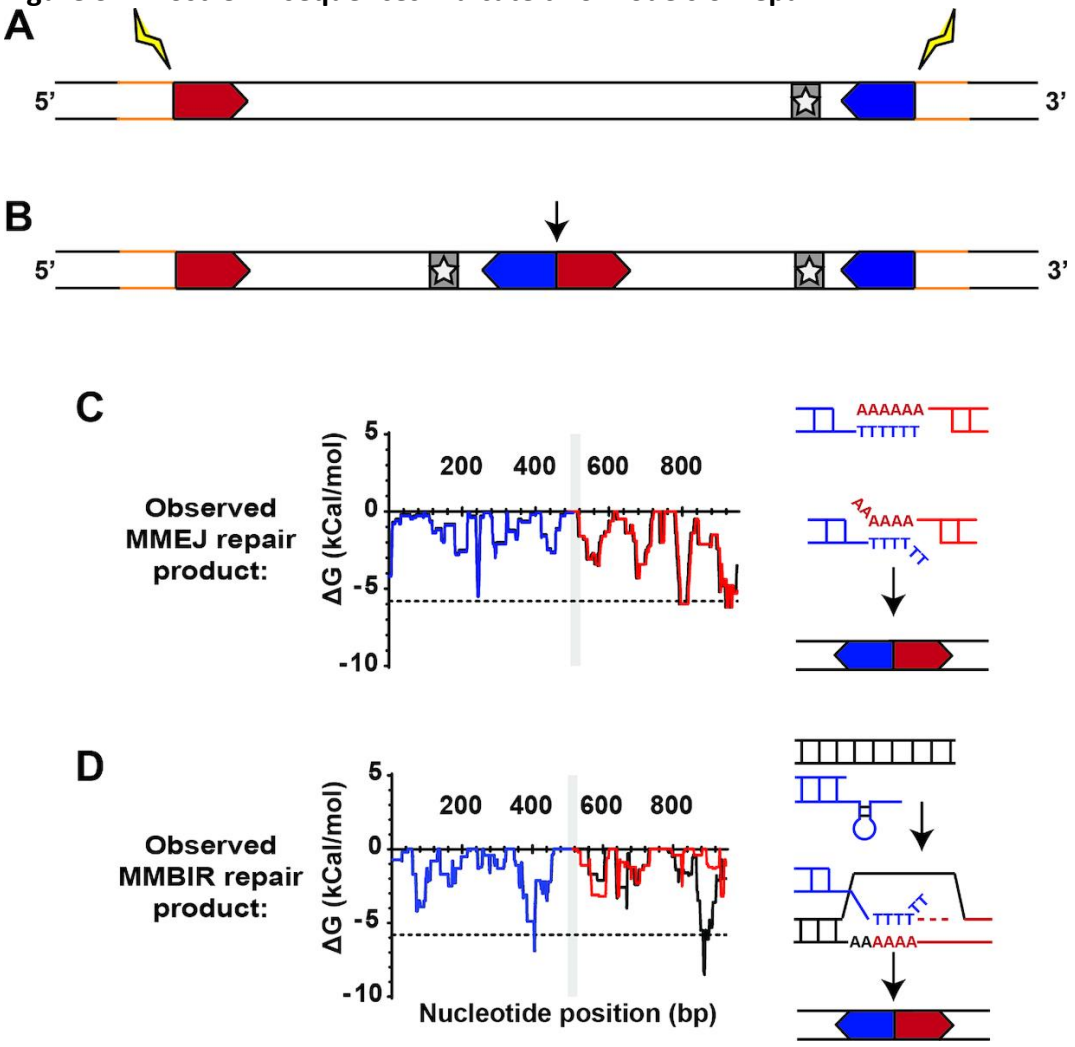
The distribution of absolute distances between long A/T tracks (>9bp) and the nearest stable hairpin (<-5.8 kCal/mol) for intergenic sequences (**A**) and genic sequences (**B**) in the *3d7* genome. The yellow highlight indicates the critical ranges noted in our analysis: A/T tracks between 20 and 40bp in length (the range detected in our analysis of CNV breakpoints, see supplementary table S4) and distance of <400bp (the distance limit for the most highly stable structures identified in mean profiles, **Fig. 2**). All plots exclude absolute distance values >4000bp (few data points fell beyond this distance).

Identifying DNA repair pathways utilized in CNV formation.

The above analysis was performed using parent sequence *prior* to CNV formation (*pre-CNV*, **Fig. 3.7A**). In order to pinpoint which repair pathways may be acting in this process, we also studied the sequence from resistant clones *after* CNV formation (*post-CNV*, **Fig. 3.7B**). This was accomplished by comparing *pre-* and *post-*CNV sequences from two sources, when available: PCR sequence of the A/T track breakpoint (for two DSM1 resistant clones) and split-reads from breakpoint

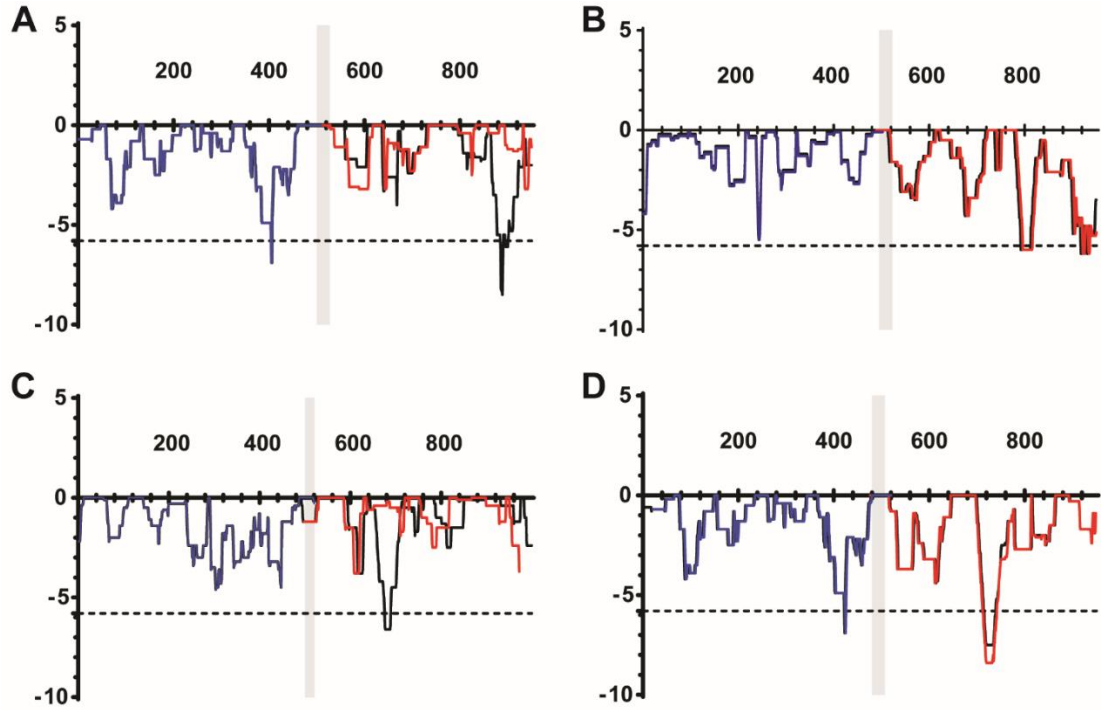
alignment sequences (for another 14 clones). We found that the *post*-CNV A/T track lengths were 16.6 ± 19.0% shorter than the *pre*-CNV lengths (Supplementary Table S4, $p < 0.01$). Despite the almost ubiquitous shortening of the breakpoint A/T track, hairpin predictions using post-CNV sequence from DSM1 resistant clones yielded a pattern similar to that of pre-CNV sequence due to a general lack of mutations surrounding the A/T tracks (**Fig. 3.7C, Fig. 3.8B and D**). In two exceptions (of 7 *post-CNV* breakpoints analyzed), a novel stable hairpin was generated (**Fig. 3.7D**, **Fig. 3.8A and C**), indicating sequence changes following CNV generation. Analysis of deep sequencing reads at these locations further confirmed these findings (unpublished data). These two different patterns suggest the action of multiple repair pathways in CNV generation (see *Discussion*).

Figure 3.7: Post-CNV sequences indicate two models of repair.



A and B. Steps leading to the generation of a novel junction after CNV formation. A. Upstream (5', red chevron) and downstream (3', blue chevron) sequences in the parent clone undergo recombination (yellow bolt), amplifying the genome surrounding the target gene (gray bar with star). Sequence outside of the amplified region is indicated in yellow. **B.** Following recombination, a tandem duplication with two copies of the target gene and a novel junction at the upstream and downstream sequence (arrow) is formed. Sequence outside of the amplicon is conserved (yellow). **C and D. Use of hairpin prediction at the novel junction to identify signatures of repair pathways. C.** Hairpin prediction pattern is conserved at the novel junction, indicating action of microhomology-mediated end joining (MMEJ, red/blue: predicted error-free repair, black: observed sequence, plot shown for DSM1 resistant D clone, see Fig. 3.8 for DSM1 resistant F clone). Repair via MMEJ occurs through resection, A/T track exposure, and annealing of two complementary genomic locations. The method of repair does not affect upstream and downstream sequence but may remove nucleotides from the A/T track. **D**. Hairpin prediction pattern is altered at the junction/novel downstream hairpins and mismatched locations indicate action of microhomology-mediated break induced replication (MMBIR, red/blue: predicted error-free repair, black: observed sequence, plot shown for DSM1 resistant C clone, see Fig. 3.8 for DSM1 resistant E clone). Repair via MMBIR uses error prone replication that induces mutations around the A/T track to resolve a stalled replication fork (arrow). This likely occurs through A/T track invasion at another genomic location for CNV generation and resolution.

Figure 3.8: Hairpin stability at novel junctions created by the generation of CNVs.



Hairpin stability (ΔG) across 1kb of sequence at novel junctions created by the generation of CNVs (see **Fig. 3.7B**). Red and blue lines indicate *predicted* error-free repair utilizing pre-CNV sequence, black lines demark *observed* post-CNV sequence. Conserved junctions from D and F clones (panels **B** and **D**, respectively) indicate MMEJ action (see Fig. 4. Novel junctions created post-CNV from C and E clones (panels **A** and **C**, respectively) indicate MMBIR action (also see **Fig. 3.7C** and **D**). Significant hairpins fall below the dotted black line (see methods for details on cut-off, -5.8 kCal/mol). The location of the A/T track at upstream and downstream breakpoints are indicated with vertical grey bars.

DISCUSSION

CNVs are an established contributor to clinical antimalarial resistance [56, 66, 69, 76, 83, 114, 115]. From conservative estimates on wild parasite populations, as much as 6% of *P. falciparum* genes are encompassed within CNVs [66]. It is important to note that this estimate is distinct from laboratory selections because it quantifies stable CNVs that persist following purifying selection in the mosquito or human parasite stages. Recent laboratory selections have shown that CNVs are as frequently observed as non-synonymous SNPs within *in vitro* selected *P. falciparum* clones [41]. However, CNVs affect more total base pairs and are distributed across all chromosomes [41, 66]. This broad distribution is somewhat unique. CNVs are often biased to certain chromosomes in organisms as diverse as rice [116], rats [117], cattle [118], and humans [119]. However, organisms that show vast phenotypic

diversity and high selective pressures appear to have a broader CNV distribution (such as dogs [120] and mice [121]).

Here, we took a novel approach to dissect CNV generation across the genome of the protozoan parasite; we performed in-depth bioinformatic analysis of sequences found at known CNV breakpoints across all chromosomes. In doing so, we gained an understanding of DNA features and molecular pathways that can trigger CNV formation. We and others have postulated that CNV formation is the initial step in *P. falciparum* that leads to the accumulation of high level, stable, resistance-conferring SNPs [61, 122]. This hypothesis is consistent with the role of CNVs as an adaptation strategy that is broadly relevant to the parasite as well as other organisms [66, 122-125].

Shared CNV breakpoints reveal a model of CNV formation

High quality deep sequencing of parasites from several controlled laboratory selections provided a unique opportunity to study CNV formation in the *P. falciparum* genome (**Table 3.2**). Three characteristics facilitated these studies: 1) the availability of sequence from parent clones (prior to selection or *pre-CNV*) allowed for analysis of the native genome architecture at the position of the future CNV breakpoint, 2) sequence from resistant clones (*post-CNV*) allowed for mechanistic studies on the pathways that enacted the change, and 3) breakpoints that occurred more than once in independent selections (or 'shared' breakpoints) allowed us to identify features that likely contribute to CNV formation.

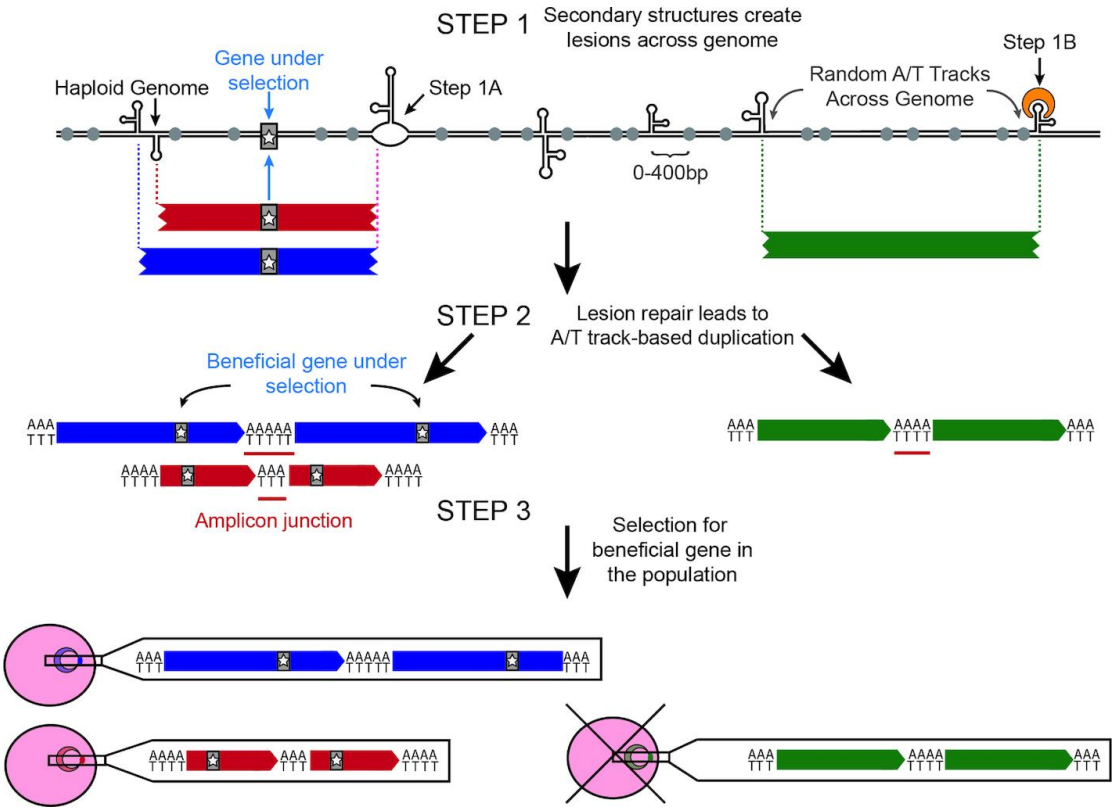
Overall, five shared breakpoints were detected in our analysis; due to their occurrence, we speculated that there was an additional CNV signal beyond the almost ubiquitous A/T track present at these locations. Indeed, secondary structure predictions identified extremely stable hairpins in close proximity to these shared breakpoints (**Fig. 3.3**, **Fig. 3.4**, **Table 3.4**). The specific hairpins identified in this analysis were more stable than 99.8% of hairpins predicted across the genome (~23.5 million structures overall) or the top 8% of *stable* hairpins (~61,000 structures with ΔG of <-5.8kCal/mol in total). This finding increased our confidence that hairpins within close proximity to the breakpoint A/T track were of importance. Structure predictions on the remaining unique CNV breakpoints displayed a similar

profile with a mean ΔG in the top 12% of stably predicted hairpins across the genome.

DNA hairpins and other secondary structures have been implicated in mechanisms of immune evasion by *P. falciparum* [93, 103, 126]. Additionally, such structures are known to cause problems during DNA replication in other organisms: they result in higher levels of replication fork collapse and DNA breakage [44, 92] and hairpin-binding proteins can stimulate recombination at these sites [127-129]. When repaired erroneously, these events can lead to the formation of CNVs [44, 102, 103, 130].

In light of these previous studies and our results, we propose a model of CNV generation (**Fig. 3.9**): DNA hairpins in close proximity to long A/T tracks throughout the *P. falciparum* genome have the propensity to create DSBs by replication fork collapse (Step 1A) or cleavage by hairpin-binding proteins (Step 1B). These DSBs are subsequently repaired in a non-faithful manner to create CNVs (Step 2). Resulting amplifications are initially rare throughout *P. falciparum* populations but then undergo selection to remove deleterious CNVs and promote the maintenance of beneficial CNVs (Step 3).

Figure 3.9: Model of copy number variation development and selection in *P. falciparum*.



In Step 1, DNA hairpins trigger double strand breaks throughout the *P. falciparum* genome presumably by either halting replication fork progression (**Step 1A**) or recognition by hairpin-binding proteins (**Step 1B**). In **Step 2**, long A/T tracks (grey circles) within 400bp of the double strand break are utilized as microhomology for error-prone repair pathways to generate CNVs (blue, red, and green bars). CNV breakpoints (vertical dotted lines) are generated semi-randomly across the genome but more stable hairpins are more likely to generate recurrent breakpoints (purple dotted line). *De novo* CNVs can either contain beneficial genes (grey bar with star) or those unrelated to the selection. New CNVs are generated frequently and could randomly occur throughout the highly repetitive *P. falciparum* genome (green bar), but may increase under selective pressure (see *Discussion*). In **Step 3**, selection (i.e. drug or fitness effects) enriches for beneficial CNVs (blue and red parasites) and purges deleterious CNVs (green parasite) from the population.

CNV trigger sites are enriched within intergenic regions

We detected elevated numbers of long A/T tracks (>9bp) and stable hairpins (>-5.8kCal/mol) in intergenic regions when compared to genic regions of the *P. falciparum* genome (**Table 3.7**). Furthermore, we identified a closer track-hairpin relationship in intergenic regions (**Table 3.7**, **Fig. 3.4**) and a corresponding enrichment in trigger sites (defined as A/T tracks between 20-40 bp in length within 400 bp of a stable hairpin, which occurs for 19.0% of intergenic A/T tracks). These data indicate that there may be a selective benefit of their association in non-coding regions of the genome. We hypothesize that one such benefit includes increased

CNV generation and thus, increased adaptability especially in the face of antimalarial selection. In support of this hypothesis, the presence of CNV trigger sites across the genome poises every potential drug target for amplification (**Fig. 3.4** and **Fig. 3.9**). It is interesting to speculate that characteristics of CNV trigger sites could contribute to the observation that some clones develop resistance *in vitro* more readily than others [61, 131]. This would be the first time that DNA sequence itself, as opposed to the regulation of specific repair proteins [115, 132], has been implicated in the ability of *P. falciparum* to develop resistance.

Potential DNA repair mechanisms leading to CNV formation in Plasmodium falciparum

Through the analysis of *post-CNV* sequences, we detect evidence for two DNA repair pathways acting in the generation of *P. falciparum* CNVs: microhomology-mediated end joining (MMEJ, [133-135]) and break-induced repair (MMBIR, [136, 137]). The ubiquitous shortening of long A/T tracks after CNV generation as well as several single nucleotide insertions after repair implicates MMEJ, which can cause deletions with and without small insertions (clones D and F, Fig. 4C, Supplementary Fig. S4 and Supplementary Table S2). Alternatively, the presence of short repeat expansions points to MMBIR, which has not been characterized in *P. falciparum* (clones C and E, **Fig. 3.7D**). Nucleotide addition is a common consequence of fork slippage during replication-mediated repair processes [134, 136-138]. Fork slippage is also a hallmark of an alternate and possibly unique pathway to *P. falciparum,* synthesis-dependent MMEJ, which appears to be a mixture of MMEJ and MMBIR [134].

One major influence on the use of microhomology-mediated pathways (MMEJ and MMBIR) verses homologous recombination is the distance of DNA resection (*e.g.* the distance from DNA lesion to homologous sequence used for repair). For example, short-range resection biases repair towards microhomology-mediated pathways and extensive resection biases repair towards homologous recombination [139, 140]. Furthermore, when excluding homologous recombination, short resection distances of <50bp are more likely to lead to MMEJ as a means of repair and longer distances <250bp are more likely enacted by MMBIR [141]. Our

CNV 'trigger site' model suggests an important role for the A/T track-hairpin distance (**Fig. 3.9**); we speculate that the span of sequence between each component could reflect the resection distance for either of these two repair pathways. Given the proposed 400bp distance limit (**Fig. 3.4)**, there are 9,130 intergenic and 2,222 genic trigger sites capable of being utilized by these pathways (**Table 3.7** and **Fig. 3.4**). Although our study only assessed amplifications, repair of DSB breaks at these sites can lead to deletions as well; further investigation is required to understand the mechanisms involved in the generation of deletions as well as how they contribute to the adaptability of the parasite.

Homologous recombination is highly active in the parasite [68, 74, 134, 135]; what then leads to the use of these error-prone pathways for repair? We propose that antimalarial treatment, which causes metabolic stress, skews repair towards MMEJ and MMBIR in *P. falciparum*. Microhomology-mediated pathways in other organisms have been shown to exhibit increased activity when cells are under stress [142-144]. For example, under normal conditions in mammalian cells, RAD51 inhibits MMBIR activity and facilitates the use of homologous recombination for DSB repair [145]. However, RAD51 is downregulated during hypoxic stress in tumors, dNTP depletion as well as the starvation response in *E. coli* and cancer, and replication stress in humans [143, 145-149]. Future studies on the levels of key repair proteins will be required to see if this is the case in *P. falciparum*.

Overall, we propose that a close A/T track-hairpin relationship in the *P. falciparum* genome leads to the utilization of error prone microhomology-mediated pathways. These events lead to enhanced generation of CNVs and adaptability of this parasite under selective pressure. Further investigation of these mechanisms may identify DNA repair pathways that can be targeted to limit parasite adaptability.

Chapter 4: CNV trigger sites are conserved in *Plasmodium spp.*

In this section, Claire Granum analyzed the probability of the formation of long homopolymeric A/T and G/C tracks in *P. vivax*

4 CNV trigger sites are conserved in *Plasmodium spp.*

SYNOPSIS

Genome amplifications, a type of DNA copy number variation (CNV), are a common method of adaptation of *Plasmodium falciparum* in response to drug treatment and other selective factors. We previously found evidence that long monomeric A/T tracks are found at the breakpoints of many *Plasmodium* resistance-conferring CNVs and that nearby DNA hairpins act as the mechanism to trigger CNV formation. *P. falciparum* is extremely A/T rich and thus the utilization of these features might be expected however not all *Plasmodium* species are as A/T rich and we wondered if this might be conserved. We applied our previous analysis pipeline to analyze known CNVs in two other *Plasmodium* species*, P. vivax* and *P. knowlesi*. We found that the breakpoints of CNVs were also located within long monomeric A/T tracts. Furthermore, we found that long monomeric A/T tracts were enriched within *Plasmodium* genomes regardless of their overall genome A/T content. The evolutionary conservation of trigger site features at CNVs and their overrepresentation in different *Plasmodium* species reinforces our previous trigger site model and stresses the need for further investigation of the molecular mechanisms of CNV creation utilized by *Plasmodium*.

INTRODUCTION

With the onset of COVID-19 this year, as well SARS, MERS, and other recent zoonotic diseases, it is increasingly important to understand the mechanisms of evolution for infectious diseases. Malaria is a disease caused by six species of *Plasmodium* that has been with humans for millennia [1-6]. *Plasmodium falciparum* is the species that causes the most morbidity and mortality [7]. It is highly adaptable and has developed resistance to every drug we have used in the field thus far, however different *Plasmodium* species have varying infection rates and symptom severity. *Plasmodium falciparum* appears best able to adapt to selective factors from

human intervention but the reasons for this are largely unknown. *Plasmodium vivax* causes the second most cases after *P. falciparum* but has exhibited far less drug resistance [7]. Previous studies found that *P. knowlesi* is an ancient zoonosis from macaque monkeys but there are two hypothesized factors limiting its zoonosis: its obligate mosquito vector is currently limited to South-east Asia and its cell surface invasion ligands are not efficient in human red blood cell invasion [11, 150]. It is hypothesized that the expressed ratio of ligands limits invasion efficiency due to copy number variation [150].

Copy number variation, particularly through gene duplication and deletion, is an important evolutionary strategy for many organisms [45, 58, 59, 151, 152]. Gene duplication is key for the evolution of new genes and as a strategy for increasing expression of the gene, but it also has been shown to facilitate the accumulation of SNPs [58-61]. Studies observing both types of mutations in *Plasmodium* provide evidence that CNVs appear to eventually be lost in favor of SNPs [62-64]. The contribution of CNVs to drug resistance and general adaptation in *P. falciparum* is very well established. Two CNVs associated with clinical antimalarial resistance encompass the genes encoding the multiple drug resistance protein 1 (*pfmdr1*) and GTP-cyclohydrolase 1 (*gch1*) [28, 65-69]. Selection with novel drugs under laboratory conditions frequently result in resistance-associated CNVs [41, 61, 62, 66, 70-77]. CNVs are also a commonly observed method of adaptation by *P. vivax* [70, 78-81]. All *Plasmodium* species have large families of genes involved in host cell invasion and immune system evasion, which frequently undergo copy number variation and recombination [153-155].

In our previous research, we identified genomic features involved in the creation of CNVs in *P. falciparum* [53]. Long homopolymeric A/T tracks (20-40bp in length) were found at virtually every CNV breakpoint. We also identified stable hairpins in close proximity to CNV breakpoints that were likely the initiating lesion. These features were everywhere through the *P. falciparum* genome due to it's overall 80.6% A/T content. However, we wondered whether these features might be utilized in other *Plasmodium* species that are not as A/T-rich such as *P. vivax* or *P. knowlesi* and if the features were present in an equally A/T-rich species, *P. relictum*.

In our investigation, we utilized whole genome sequencing to study the development of copy number variations in different *Plasmodium* species and the sequences utilized in their creation. Furthermore, we hypothesize that there are differences in our previous CNV trigger site model between different species. IF these features are conserved between species, the DNA repair pathways and proteins that utilize them to generate CNVs would likely serve as a novel antimalarial target to block the creation of CNVs and thereby slow or block development of drug resistance.

MATERIALS AND METHODS

Collection of genomic and breakpoint sequences.

In order to compare breakpoint data from *Plasmodium spp.,* we combined three different data sources: 1) data previously generated from CNV analysis of *P. falciparum* whole genome, 2) newly gathered CNV breakpoints from whole sequencing data of clinical *P. vivax* samples with known drug resistance associated CNVs, and 3) CNV breakpoints from the whole genome sequencing analysis of the YH1 laboratory strain of *P. knowlesi* that was adapted to human blood (**Table 4.1**, [150, 156]).

We used similar methods to those outlined in our previous investigation of drug resistance-associated amplifications in *P. falciparum* [53]. Bases with low quality scores and adapters were removed using BBTools (version 35.82, https://sourceforge.net/projects/bbmap/). Uncorrectable errors were assigned low quality scores and the resulting cleaned reads were evaluated using FastQC to check per base read qualities, sequence duplication levels, overrepresented sequences, and read length distributions as previously [95]. Reads were aligned to *P. falciparum 3d7, P. vivax P01,* and *P. knowlesi Strain H* reference genomes (PlasmoDB release 46) by BWA-MEM with default settings [96]. Alignment quality of the resulting bam files were evaluated for mean read depth, mean mapping quality, and quartiles of paired read insert-size using Qualimap 2 [97].

CNV breakpoints were identified as previously described using the Speedseq pipeline which utilizes a Bayesian analytical method for genotyping and precise calls from LUMPY for split-read and discordant read-pair analysis and from CNVnator for

read depth analysis [53, 98-100]. LUMPY breakpoint locations were used as the final breakpoint location after evaluation for sample quality scores (>100), quantity of supporting reads (>3), and significant overlap with amplification boundaries from CNVnator and the previously published data. We then used SURVIVOR to merge the resulting calls and determine which samples shared CNVs by requiring 95% overlap, a minimum of two supporting samples that agree on the type (duplication, deletion etc), on the strand, and have a minimum length of 2500bp and a maximum of 100,000bp [157]. The resulting duplications were visually inspected using IGV 2.4.10 to confirm the position and determine amplification orientation [101].

Determining the probability of homopolymeric track formation

To determine the probability for the formation of long homopolymeric tracks based upon genome composition (A/T vs G/C), we calculated the probability of observing different tracks lengths based purely on nucleotide composition. Frequencies of monomeric tracks of length N were calculated as follows. The observed frequency of A, T, G, and C tracks of length N were obtained using the following equation:

$$f_N^{obs} = \frac{C_N^{obs}}{l_{seq}}$$

where C_N^{obs} is the observed number of monomeric tracks of length N and l_{seq} is the length of the sequence. For each track observed with length N (in this case A and T), the corresponding expected frequency of tracks was obtained from the following equation given:

$$f_N^{exp} = \left(f_A^{obs}\right)^N (1 - f_A^{obs})^2 + \left(f_T^{obs}\right)^N (1 - f_T^{obs})^2$$

where f_i^{obs} is the observed frequency of any base pair *i* which corresponds to the overall percent base composition.

Maximum expected length for each chromosome was found using the following formula:

$$N_{exp} = \frac{\log(\frac{1}{l_{seq}(1 - f_A^{obs})^2})}{\log(f_A^{obs})} + \frac{\log(\frac{1}{l_{seq}(1 - f_T^{obs})^2})}{\log(f_T^{obs})}$$

Calculating the likelihood of DNA hairpin formation.

The probability of hairpin structure formation across the desired regions was predicted as previously described [53, 102, 103]. To summarize, 50bp windows of sequence were generated by shifting by 1bp across a 2kb stretch of sequence surrounding the *pre-CNV* breakpoint position in the parent genome. 50bp windows were chosen to ensure hairpin formation was possible within the Okazaki initiation zone during replication. The size of the Okazaki initiation zone is not known in *Plasmodium* but it is expected to be in the same range as other eukaryotes (300 to 1000bp [104]). Next, the Gibbs free energy (ΔG), which predicts the stability of the sequence folding on itself, was determined for each window using Vienna 2.1.9 folding prediction software with Mathews 2004 DNA folding parameters and G-quadruplexes, GU pairing, and lonely base pairs were disallowed [105]. Lonely base pairs are helices in a hairpin or stem-loop that are composed of only 1bp and do not stack on other base pairs. These structures are not energetically favorable and cannot form and are therefore excluded from analyses.

Defining stable hairpins.

Due to a non-normal distribution of predicted hairpin ΔG values, the ΔG cutoff of stable hairpins for each respective *Plasmodium* genome was determined as previously using a randomization method: sequence from each chromosome was randomly shuffled using the EMBOSS shuffleseq function to maintain overall A/T composition and hairpins were again predicted [106]. In this analysis, 50kb of sequence on either chromosome end was trimmed to avoid highly repetitive telomeric sequences and the mitochondria, apicoplast, and unplaced contigs were excluded. The value of the resulting top 3% of shuffled hairpins was used as the stability cut-off for each respective genome; sequences with values more negative than this cutoff indicated a high probability of a ‘stable’ structure forming.

To quantify how many stable hairpins were expected in a genome, the local minima of hairpins had to be identified. First, we extracted all hairpins below our significance threshold. Then, for each set of windows with contiguous positions below this threshold, we identified the window with the most negative value and created a data subset with these minima. If there were multiple contiguous windows

with the same value, they were collapsed to form a single hairpin forming minima location.

Evaluation of trigger site features across the genome

For this analysis, telomeres/subtelomeric regions, the mitochondria, apicoplast, and unplaced contigs were excluded from analysis and thus only core chromosomal sequences were analyzed. 50kb was trimmed off the end of each chromosome to remove telomeres and subtelomeres as previously. A/T tracks were identified with the Phobos Repeat Finder [108], which mapped the locations and lengths of long monomeric A/T tracks >9bp across the respective genomes. This length of track was chosen based on our previous finding, as well as others, that demonstrated that >9bp sequences were overrepresented in *Plasmodium falciparum* genomes [53, 109]. To determine if this value was appropriate for other *Plasmodium* species, we also calculated the probability of observing A/T track lengths of a given length for the less A/T rich *P. vivax* P01 reference genome utilizing the approach we previously applied and determined that 9bp was still an appropriate cut-off [53].

Syntenic trigger site comparison

A final method of comparison was to identify chromosomes and sequences with high synteny in order to more directly compare trigger site density within homologous locations. Previous studies have also performed this comparison but not with the precise versions of reference genomes in this analysis and thus synteny was analyzed using the progress MAUVE alignment algorithm within Geneious with default settings to compare all chromosomes between the *Plasmodium* species outlined above as well as *P. relictum [158, 159]*. The default settings include full alignment, automatic seed weight calculation, automatic calculation of minimum locally collinear block (LCB) scores, and the computation of LCBs. From this step, the chromosomes with the highest synteny were identified for direct comparison.

RESULTS

***Plasmodium spp.* genomes contents are highly variable**

In our previous analysis, we identified drug resistance associated CNVs in *P. falciparum* and found that virtually every CNV breakpoint was found in an long A/T track 20-40bp and hypothesized that this may be an evolutionarily conserved feature of *Plasmodia* [53]. To begin this investigation, we first compared the reference genomes and overall genome content of different *Plasmodium spp.: P. falciparum, P. vivax, P. knowlesi,* and *P. relictum* (**Table 4.1**).

Table 4.1 - *Plasmodium* genome composition comparison

Plasmodium species	Host species	Full Genome Size (Mb)	Contigs/Chroms	Sequencing for reference	AT Content	Protein Coding Genes	# of Gene Orthologs
P. falciparum	Human	23.33	14/14	Shotgun + Sanger	80.6%	5460	5458
P. vivax	Human	29.05	226/14	Illumina	60.2%	6830	6660
P. knowlesi	Macaque/human	24.4	128/14	Shotgun + Sanger	61.4%	5483	5319
P. relictum	Avian	22.61	498/14	Illumina	81.7%	5138	5108

Reference genomes: *P. falciparum = 3d7, P. vivax = P01, P. knowlesi = PknH, P. relictum = SGS1.*

In addition to *P. falciparum,* we chose two human-infective species of malaria, *P. vivax* and *P. knowlesi*, which both have genome A/T contents closer to 60% and *P. relictum* that is slightly more A/T-rich (**Table 4.1**). While the reference genomes are in differing states of completion, each species had 14 chromosomes, an apicoplast, and mitochondrial genomes (**Table 4.1**) [160-162]. *P. falciparum* had the most complete genome with zero unplaced contigs, which was accomplished in 2002 through the efforts of a large consortium to shotgun sequence the whole genome from plasmid clones [160]. *P. knowlesi* was also accomplished using shotgun sequencing of plasmid clones whereas the other two species were assembled using Illumina short-read sequencing with various read-lengths and assembly strategies [161, 162]. The final genome completion status was inferred based upon the number of unplaced contigs, protein coding genes, and gene orthologs. Despite different genome sizes and assembly methods, *P. falciparum, P. knowlesi,* and *P. vivax* had similar numbers of protein coding genes and orthologs (**Table 4.1**). *P. vivax* was the

only species which was significantly larger and had more predicted protein coding genes and orthologs.

CNV breakpoint junction sequences occur at long A/T tracks in *P. falciparum, vivax,* and *knowlesi*

After initial genome characterization where we sought to appreciate broad genomic differences, we utilized our previous CNV analysis pipeline to gain precise resolution of the breakpoints for validated gene amplifications in *P. vivax* and *P. knowlesi* genomes. For *P. vivax*, we analyzed a data set that included 46 high quality clinical isolates from various regions of the world [156]. These isolates contained 1) an MDR1 amplification associated with resistance to chloroquine, 2) an amplification of the PVP01_1468200 gene that may be involved in merozoite invasion, and 3) an amplification of the Duffy-Binding Protein (DBP) that may improve efficient invasion of Malagasy individuals [70, 79, 80, 163].

From this dataset, we identified 9 clinical isolates with the known CNVs described above that were supported by both CNVnator and LUMPY [156] (**Table 4.2**). There were 5 supporting isolates for the DBP duplication, 2 supporting isolates for the MDR1 amplification, and 2 isolates for the PVP01_1468200 duplication. It is interesting to note that the CNVs we identified differed greatly from the original data source in both their location and even chromosomes. This was entirely due to our usage of the P01 reference genome as opposed to the less accurate *PvSal1* reference genome (**Table 4.2**) [156, 164].

Table 4.2 – *Plasmodium* CNV locations

Plasmodium Species	Gene	CNV Chr.	CNV Start (±CI)	CNV End (±CI)	Supporting Isolates	Source
P. vivax	DBP	06	980473 ±1	987840 ±0	ERR111718 ERR111719 ERR111729 ERR111732 SRR828416	[156]
	MDR1	10	468190 ±0	506357 ±0	ERR111717 ERR111721	
	PVP01_1468200	14	2903557 ±0	2907109 ±0	ERR054084 ERR054085	
P. knowlesi	DBP	6	996,794 ±7	1,047,723 ±10	SRR3135172	[150]

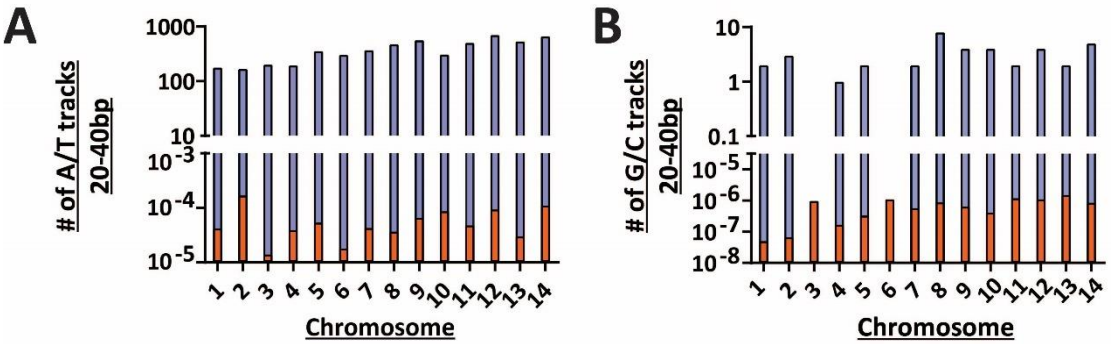
Reference genomes are: *P. falciparum = 3d7, P. vivax = P01, P. knowlesi = PknH.* ±CI = confidence interval. CNVs share 95% overlap and same strand support.

We also analyzed the *P. knowlesi* YH1 strain which has a DBP alpha duplication identified on chromosome 6 [150]. We previously reported that of 19 unique CNVs identified in *P. falciparum* clones with 33 unique breakpoint sequences, all but 4 were found in long A/T tracks [53]. Here we identified 6 unique junctions in *P. vivax* and 2 unique CNV junctions in *P. knowlesi*. Every single *P.* vivax breakpoint junction occurred at a long A/T track (mean length of 18bp) and the DBPalpha breakpoints for *P. knowlesi* were 18bp A/T tracks on either end. This means that virtually every CNV we have identified in *Plasmodia* species utilizes a long A/T track as its junction sequence.

A/T tracks are overrepresented genome-wide in all analyzed *Plasmodium spp.*

After identifying that CNV breakpoints were once again found within long A/T tracks, we next investigated the expected and observed number of long A/T tracks within the *P. vivax* genome. We chose the *P. vivax* genome because it has the lowest A/T content of all genomes that we analyzed and is composed of 60.2% A/T (**Table 4.1**). We calculated the probability of formation as previously and found the probability of observing A/T tracks of 7bp or longer for *P. vivax P01* is <0.1% (data not shown). Based upon our previous analysis with *P. falciparum,* we continued to analyze A/T tracks >9bp in length [53]. We also found that A/T tracks in the range of 20-40bp (which is the length we previously hypothesized were involved in CNV formation) are vastly overrepresented in the *P. vivax* genome with ~10^7 more observed than would be expected for virtually every chromosome (**Fig. 4.1, Table 4.3**). We also found that long G/C tracks were overrepresented. However, there were very few GC tracks between 20-40bp on any chromosome (with 0 on chromosomes 3 and 6) and these sequences are not expected to play a role in CNV formation (**Fig. 4.1**).

Figure 4.1 – Track length for *P. vivax P01* expected vs observed



The probability of observing a homopolymeric track of a given length (orange) was calculated based upon the base composition of each chromosome (see Methods). The observed number of homopolymeric tracks 20-40bp long (blue) was determined using Phobos Repeat Finder.

Table 4.3 – A/T track lengths per chromosome for *P. falciparum 3d7, P. vivax P01, P. knowlesi Strain H,* and *P. relictum SGS1*

Chromosome	A/T Tracks >9bp				A/T Tracks >20bp			
	P. falciparum	*P. vivax*	*P. knowlesi*	*P. relictum*	*P. falciparum*	*P. vivax*	*P. knowlesi*	*P. relictum*
Chr. 1	1657	1240	1324	2275	320	158	437	79
Chr. 2	2371	1140	1064	2185	464	141	306	70
Chr. 3	2786	1077	1591	1828	527	159	445	59
Chr. 4	3045	1309	1745	2275	566	162	538	67
Chr. 5	3489	2291	1166	2379	707	290	317	71
Chr. 6	3520	1433	1694	2562	669	261	502	62
Chr. 7	3642	2320	2558	4391	685	301	792	127
Chr. 8	3996	2483	3148	4435	777	398	866	127
Chr. 9	4263	3247	3850	5919	842	478	1048	174
Chr. 10	4496	2106	2325	3874	885	266	670	115
Chr. 11	5567	3007	3886	6272	1077	415	1131	174
Chr. 12	6104	4709	5680	9005	1228	596	1545	258
Chr. 13	7657	3087	4514	6884	1480	443	1249	177
Chr. 14	8841	4733	5640	8967	1692	560	1447	244
Total	**61434**	**34182**	**40185**	**63251**	**11919**	**4628**	**11293**	**1804**

For these calculations, the apicoplast, mitochondria, unplaced contigs, and telomeres were excluded.

Finally, we analyzed long A/T tracks on a chromosome-by-chromosome and genome-wide basis, we found that *P. falciparum* had the most A/T tracks >9bp with 61,434 genome-wide; despite being significantly less A/T-rich, *P. knowlesi* had 40,185 and *P. vivax* with 34,182 A/T tracks >9bp (**Table 4.3**). However, there were two interesting observations. The first was that *P. knowlesi* had almost as many A/T

tracks >20bp genome-wide as *P. falciparum* (11,293 versus 11919, **Table 4.3**). It is important to note that *P. knowlesi Strain H* is composed of only 61.4% A/T as opposed to *P. falciparum 3d7* which is composed of 80.6% A/T. The second interesting observation was that two species with the highest A/T content (*P. falciparum* and *P. relictum*) had similar quantities of A/T tracks >9bp but there were significantly fewer A/T tracks >20bp in *P. relictum.* It is difficult to compare the species on a chromosome by chromosome basis as each chromosome is a different length and due to differences in synteny, each chromosome may not be directly comparable. However, we address this later by identifying syntenic chromosomes and specifically comparing them.

***P. vivax* is predicted to form the most stable hairpins**

After investigating A/T tracks, we next investigated the formation of stable hairpins which are the second feature involved in our trigger sites. Using the same methods as previously outlined, we found that the overall genome-wide mean ΔG was highest for *P. falciparum* and lowest for *P. vivax* (after excluding unplaced contigs, the apicoplast, the mitochondria, and telomeres). This was partially expected based on the difference in overall A/T content as G/C bonds are stronger than A/T bonds (**Table 4.4**).

Table 4.4 – A/T track lengths per chromosome for *P. falciparum 3d7, P. vivax P01, P. knowlesi Strain H,* and *P. relictum SGS1*.

Plasmodium species	Hairpin Minima			
	Genome-wide Mean ΔG *	ΔG cutoff stable hairpins (kcal/mol)	Stable Hairpin Collapsed Minima *	Mean ΔG of Minima *
P. falciparum	-1.4	-5.8	60881	-7.43
P. vivax	-3.31	-7.2	135513	-8.76
P. knowlesi	-2.54	-6.2	91202	-7.65
P. relictum	-1.89	-5	64215	-6.20

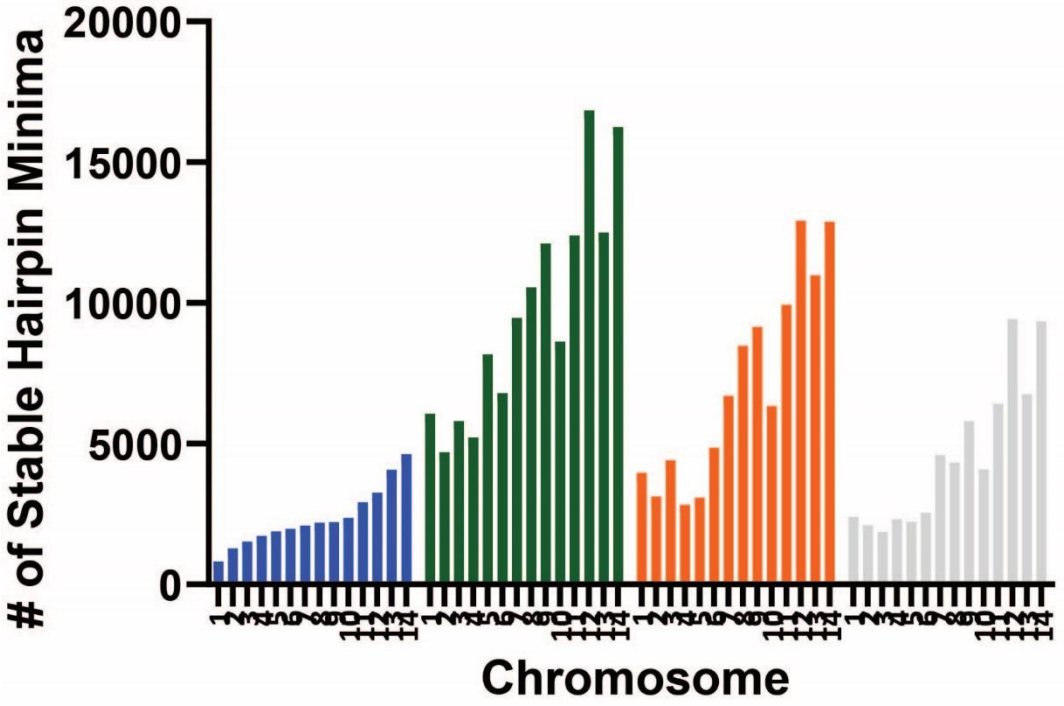
For all calculations, the apicoplast, mitochondria, unplaced contigs, and telomeres were excluded

After utilizing our previous method of determining stable hairpins, we found that the cutoff for formation of stable 50bp hairpins was -5.8 kcal/mol in *P. falciparum,* -7.2 kcal/mol in *P. vivax*, 6.2 kcall/mol in *P. knowlesi,* and -5 in *P. relictum* (**Table 4.4)**. In order to determine the number of stable hairpin forming regions, we

found local stable hairpin minima and found that *P. vivax* is predicted to form the most stable hairpins with ~135,000 genome-wide as opposed to ~61,000, ~91,000, and ~64,000 for *P. falciparum, P. knowlesi,* and *P. relictum* respectively (**Table 4.4, Fig 4.2**).

When analyzed on a chromosome by chromosome basis, several trends stand out. The *Plasmodium spp.* follow the same general trends for each chromosome as they do genome-wide with *P. falciparum* having the least A/T tracks per chromosome and *P. vivax* the most (**Fig. 4.2**). An interesting observation we made was that the number of stable hairpins increases linearly by chromosome for *P. falciparum* but jumps between chromosomes for *P. vivax* and *P. knowlesi*.

Figure 4.2 – Stable hairpin collapsed minima per chromosome

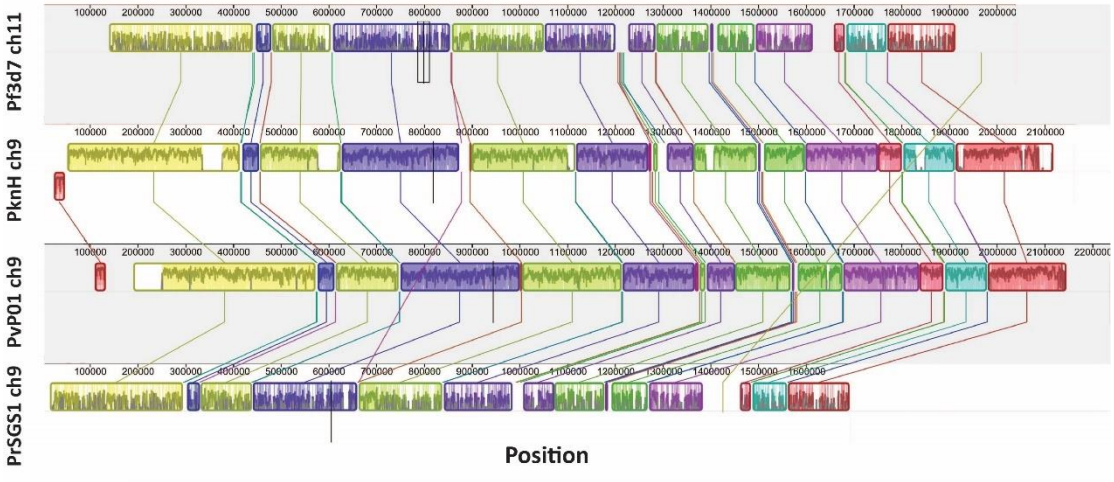


Stable hairpin minima were found in *P. falciparum 3d7* (blue), *P. vivax P01* (green), *P. knowlesi Strain H* (orange), and *P. relictum SGS1* (gray). For these calculations, the apicoplast, mitochondria, unplaced contigs, and telomeres were excluded. *P. falciparum* chromosomes are numbered by size and other species are numbered based upon synteny between species.

However, it is important to remember that the synteny and homology between species of *Plasmodium* is not perfect and each chromosome may not be a perfect match for the same chromosome number in another species. For this reason, we next determined syntenic chromosomes for a more direct comparison using the MAUVE algorithm which maps conserved blocks of homology (**Fig. 4.3**). It had

previously been reported that *P. falciparum* chromosome 11 was highly syntenic to chromosomes 9 for *P. knowlesi* and *P. vivax* and we confirmed this with our MAUVE analysis as well as showed that *P. relictum* chromosome 9 is also highly syntenic (**Fig. 4.3,** [153, 161]). We further confirmed that *P. vivax* and *P. knowlesi* had near 1-to-1 synteny and that both parasites had less synteny with *P. falciparum* and *P. relictum* (**Fig. 4.3**). When comparing these chromosomes, the three human infective species were relatively the same length but *P. relictum* was significantly shorter (**Table 4.5**).

Figure 4.3: Comparison of syntenic *Plasmodium* DNA matches genome-wide trigger site trends.



***P. falciparum 3d7* chromosome 11, *P. knowlesi* chromosome 9, *P. vivax* P01 chromosome 9, and *p. relictum* chromosome 9 are highly syntenic.** All chromosomes of the four species were analyzed using the MAUVE algorithm to find blocks of conserved DNA and each conserved block is given a particular color (red, yellow, dark purple, green, orange, and purple). Connecting lines point to boundaries of conserved blocks.

Table 4.5 – *Plasmodium* syntenic chromosome comparison

Plasmodium species chromosome	Ch. Size (Mb)	# of genes	A/T tracks >9bp *	A/T tracks >20bp *	Stable hairpin minima *
P. falciparum 3d7, Ch. 11	2.04	525	5567	1077	5780
P. vivax P01, Ch. 9	2.24	516	3451	575	12116
P. knowlesi Strain H, Ch. 9	2.16	480	3850	1048	9151
P. relictum SGS1, Ch. 9	1.69	453	5919	174	5805

*Apicoplast, mitochondria, unplaced contigs, and telomeres excluded

All of the species had similar numbers of genes. The previously observed genome-wide trends continued in these syntenic chromosomes.

The trends we observed genome-wide for both A/T tracks and stable hairpin minima were conserved with *P. falciparum* having the most long A/T tracks and *P. vivax* having the most stable hairpins. As a final means of comparison, we determined the distance in between the previously defined trigger site features of long A/T tracks and stable hairpins and found that the distance between the two features was shortest in *P. knowlesi* and longest in *P. falciparum*. This is somewhat expected based on the total number of stable hairpins and A/T tracks which would force the density to be higher in *P. knowlesi* and *P. vivax* than in *P. falciparum*. However, based upon our previous model *P. falciparum* would still have the most possible trigger sites due solely to having the most long A/T tracks.

Discussion

A/T track breakpoints are conserved in clinically relevant *Plasmodium spp.*

Based upon our previous analysis, we developed a trigger site model in *P. falciparum* that facilitates the creation of copy number variations [53]. We developed this model through the identification of two primary features at virtually every breakpoint of CNV junction: long A/T tracks and stable hairpins. In this study, we investigated whether this trigger site model contributes to CNV generation in two other clinically relevant *Plasmodium spp.* (*P. vivax* and *P. knowlesi*) through the sensitive and specific investigation of known CNVs [150, 156]. We found that every CNV had long A/T track as their breakpoints (**Table 4.2**). *P. vivax* CNVs on different chromosomes and from different areas of the world all utilized A/T tracks as microhomology in the creation of their CNVs. Furthermore, a lab adapted clone of *P. knowlesi* YH1 also utilized A/T tracks for the creation of the duplication that facilitated the invasion of human red blood cells (**Table 4.2**) [150]. Through the use of SURVIVOR and stringent requirements for CNV identification (see Materials and Methods), we identified multiple amplifications and deletions to corroborate the usage of A/T tracks in the creation of CNVs in *Plasmodium.* Automated tools such as SURVIVOR allow us to more efficiently analyze CNVs and gather greater quantities of

known trigger sites to expand our model. Finally, it would be interesting to see if *P. relictum* or *Plasmodium spp.* that infect other organisms such as rodent also utilized A/T tracks as their breakpoints.

Unfortunately, we were not able to perform the same stable hairpin analysis near CNV breakpoints as we performed previously [53]. Without the direct parent of an isolate, we are unable to determine the importance of stable hairpins in the formation of these CNVs. We had previously compared parent and clone genomes at CNV breakpoints, we were able to identify that the hairpin causing the DNA break was sometimes conserved and was sometimes lost during the repair process. Without direct parent and daughter samples this process is impossible to accomplish. The parasite undergoes many rounds of replication and mutation in the host and during this process it may lose stretches of sequence that have directly contributed to CNV formation.

Trigger site features are overrepresented in all *Plasmodium spp.*

We then investigated these trigger site features genome-wide in *P. vivax, P. knowlesi,* and *P. relictum.* Despite being considerably less A/T-rich, both *P. vivax* and *P. knowlesi* have far more long A/T tracks than would be expected (**Table 4.1**, **Fig. 4.1**, and **Table 4.3**). It was interesting to find that *P. knowlesi* had almost as many A/T tracks between 20-40bp in length as *P. falciparum* (**Table 4.3**). In juxtaposition, *P. relictum* had the A/T tracks >20bp despite being the most A/T-rich. The reasons for these differences likely have interesting biological causes and consequences for the formation of CNVs.

When investigating the formation of stable hairpins, we found that *P. vivax* was predicted to form both the most stable hairpins and the most stable hairpins genome-wide (**Table 4.4** and **Fig. 4.2**). In order to more directly compare sequences, we identified syntenic chromosomes where the order of blocks of homologous DNA were conserved between species (**Fig. 4.3**). The trends that we observed genome-wide held up for these sequences as well and may therefore be directly comparable between genes (**Table 4.5**). An interesting example would be the direct comparison of sequences surrounding the MDR1 gene, which is known to confer resistance to antimalarials in *P. falciparum and P. vivax* but has not thus far been identified as a

means of drug resistance in *P. knowlesi*. We have speculated previously that ubiquitous nature of the CNV trigger sites in *P. falciparum* contribute to its ability to rapidly adapt to various stressors and through this mechanism, it is possible that *P. knowlesi* exhibits a similar propensity. However, a firm conclusion in this regard would require laboratory selections with both species under various forms of stress. In order to perform the same rigorous CNV analysis and identify the repair pathways utilized in the creation of their CNVs that we did for *P. falciparum,* we would need three things: to obtain parent and daughter sequences for newly formed CNVs, to identify shared breakpoints in order to prove the importance of stable hairpins as a source of DNA breaks, and to expand the numbers of CNVs to see the exact frequency of A/T track usage in the other *Plasmodium* parasites. Future studies will investigate the mutational signatures found at *de novo* CNVs in other *Plasmodium* species in order to investigate their conservation of DNA repair pathways.

Chapter 5: Adaptation of novel computational methods to investigate *Plasmodium falciparum* biology

For the single cell sequencing section, some of the following text, figures, and tables have been adapted from Liu et al. 2020, [165]. For the extrachromosomal DNA section, some of the text, figures, and tables have been adapted from McDaniels, Jennifer M. et al. "The generation of extra-chromosomal DNA amplicons in antimalarial resistant *Plasmodium falciparum*", in preparation for Molecular Microbiology.

The work in this chapter was done in collaboration with and augmented by the work of several coauthors. Specifically, Shiwei Liu conducted all single cell experimental work, the read depth analysis for bins >10kb, and all analysis of variance in Figures 5.1 and 5.3. William Chronister analyzed read depth for the binning data ≤10kb to predict CNVs in Figure 5.5. Jennifer McDaniels conducted all ecDNA experimental work and expanded analysis of the sac3 super-amplicon in the sequencing data in Figures 5.6 and 5.7. Contributions are detailed beneath figures.

5 Adaptation of computational methods to Plasmodium falciparum

SYNOPSIS

P. falciparum is a challenging organism in which to perform genetic studies for a number of reasons. This intracellular parasite has a very biased (80.6% A/T), small genome at 23Mb and ~25 femtograms of DNA per genome copy. In addition, sample preparations are heavily contaminated with human host DNA from circulating cells and cell-free DNA [166]. Combined, these factors make isolation of parasite DNA for whole genome sequencing of single parasites and auxiliary forms of DNA (such as extrachromosomal DNA) especially difficult. In this chapter, I present novel applications and investigations of whole genome sequencing to investigate the *Plasmodium* genome and copy number variations. Specifically, I discuss our efforts to identify copy number variations within individual parasites to investigate heterogeneity within populations. Finally, I briefly discuss the computational investigation of *P. falciparum* extrachromosomal DNA.

Single cell sequencing to investigate *P. falciparum* copy number variations heterogeneity

INTRODUCTION

Population heterogeneity is an important strategy that *Plasmodium* utilizes to improve survival and functionality; having a diverse population enables a better chance of survival under novel stressors. However, results from whole genome sequencing represent an average of populations of cells and can only find mutations above a certain frequency in a population. Single cell sequencing is the best method of investigating cryptic populations hidden within bulk samples. We are attempting to use single cell sequencing to answer two major outstanding questions regarding *P. falciparum* CNVs. With what frequency do CNVs arise in *P. falciparum* under stress*?* Are there hidden copy number variations to be found within populations? The

present methods of identifying CNVs in *P. falciparum* utilize Illumina short-read sequencing of bulk DNA. The two current best methods of identifying CNVs from Illumina whole genome sequencing are read depth analysis and split-read/discordant read pair analysis (as used in [53]). Both approaches have great difficulty in detecting rare variants.

Due to the limitations of bulk sequence analysis, recent investigations have analyzed single cells to appreciate low frequency CNVs across heterogeneous populations of yeast, mouse, and human cells [167-172]. This approach provides a significant advantage for detecting rare genetic variants by no longer deriving an average signal from large quantities of cells. However, short read sequencing requires nanogram to microgram quantities of genomic material for library construction, which is many orders of magnitude greater than the genomic content of individual *Plasmodium* cells (femtogram quantities). Therefore, whole genome amplification (WGA) is required to generate sufficient DNA quantities. Several WGA approaches have been reported and each has advantages and disadvantages for different applications [173-175] but most were optimized for mammalian cell analysis [172, 174, 176-183].

Few studies have attempted to detect genetic variations in single *P. falciparum* parasites. One WGA method, multiple displacement amplification or MDA, has been used to amplify single *P. falciparum* genomes with near complete genome coverage [50, 51]. These studies successfully detected single nucleotide polymorphisms between single parasites; however, MDA is less useful for CNV detection because analysis is disrupted by low genome coverage uniformity and the generation of chimeric reads [175, 184].

Multiple annealing and looping-based amplification cycling (MALBAC) is another WGA method that exhibits improved uniformity over MDA, which is advantageous for detecting CNVs in single cells [185]. MALBAC has the unique feature of quasi-linear pre-amplification, which reduces the bias associated with exponential amplification [185]. However, standard MALBAC has been reported to be less tolerant to AT-biased genomes, unreliable with low DNA input, and prone to

contamination [186-188]. Thus, before using MALBAC to amplify the *P. falciparum* genome, optimization of this WGA method is necessary.

The ability to analyze the resulting Illumina whole genome sequencing data largely depends on the ability to align Illumina paired-end sequences to a reference genome. Finding unique mapping locations for the reads is frequently the goal for these analyses. However, each reference genome presents a unique challenge as it has its own sequence bias, repetitive sequences, and assembly quality. Gaps in the reference genome will lead to unmappable reads, sequence bias (high A/T or G/C content) can influence the ability of libraries to accurately capture the sequence, and repetitive sequences can either prevent unique mapping of reads or present a challenge for amplification by polymerases.

Both read depth and discordant/split-read analysis rely upon the ability to uniquely map reads to the reference genome. However, various alignment algorithms handle this challenge in different manners and therefore a standardized pre-processing step may be useful. One common approach to this problem is the utilizing the “mappability” of a genome. Regions of a genome with high “mappability” would tend to produce uniquely mapped reads and could theoretically be used for normalization or masking procedures in copy number variation analysis. This is most frequently calculated purely based upon a given reference genome and read length without mismatches.

In this study, we present a single cell sequencing pipeline for *P. falciparum* parasites, which includes efficient isolation of single infected erythrocytes, an optimized WGA method inspired by MALBAC, and a sensitive method of assessing sample quality prior to sequencing. We tested our pipeline on erythrocytes infected with laboratory-reared parasites as well as patient-isolated parasites with heavy human genome contamination. Genome amplification using our optimized protocol showed increased genome coverage, better coverage uniformity, and strong amplification reproducibility. These improvements will enable the detection of parasite-to-parasite heterogeneity to clarify the role of genetic variations, such as CNVs, in the adaptation of *P. falciparum*. These improvements also provide a

framework for the optimization of single cell amplification in other organisms with challenging genomes.

MATERIALS AND METHODS

Some of the methods detailed in this section have been adapted and abbreviated from our publication on improved single cell amplification of *P. falciparum* [165]. The full parasite culturing, amplification, sequencing methods, as well as expanded coverage analysis for this data are detailed in Liu et al. 2020 [165].

Selected sequencing analysis for this chapter

Sequencing quality control and alignments were performed essentially as previously described [53]. Briefly, we removed Illumina adapters and PhiX reads, and trimmed primers from reads in each fastq file with the BBDuk tool in BBMap [189]. We then aligned each fastq file to the hg19 human reference genome to remove human contamination and kept the unmapped reads (presumably from *P. falciparum*) for further analysis [189]. Each "cleaned" fastq file was then aligned to the 3D7 *P. falciparum* reference genome with Speedseq [98]. Reads with low-mapping quality score (below 10) and duplicated reads were discarded using Samtools [190]. Qualimap 2.0 was used to analyze genome-wide coverage statistics [97]. We compared the variation of normalized read abundance (log10 ratio) at different bin sizes using boxplot analysis (**Figure 5.1**, R version 3.6.1) and determined the bin size of 20 kb using the plateau of decreasing variation of normalized read abundance (log10 ratio) when increasing bin sizes. We also investigated the breadth of coverage for genic and intergenic regions of the genome using the gff files which give gene locations from PlasmoDB to create bedfiles to select regions (genic and intergenic) of the genome for further analysis (PlasmoDB release 44). Read coverage from the entire genome was then divided into non-overlapping 20 kb bins using Bedtools and normalized by dividing each bin by the total average reads in each sample [110]. Finally, we calculated the coefficient of variation of normalized read abundance by dividing the standard deviation by the mean and multiplying by 100 [175, 191], then and analyzed the equality of coefficients of variation by R package "cvequality" version 0.2.0 [192].

Figure 5.1: Distribution of normalized read counts in various bin sizes

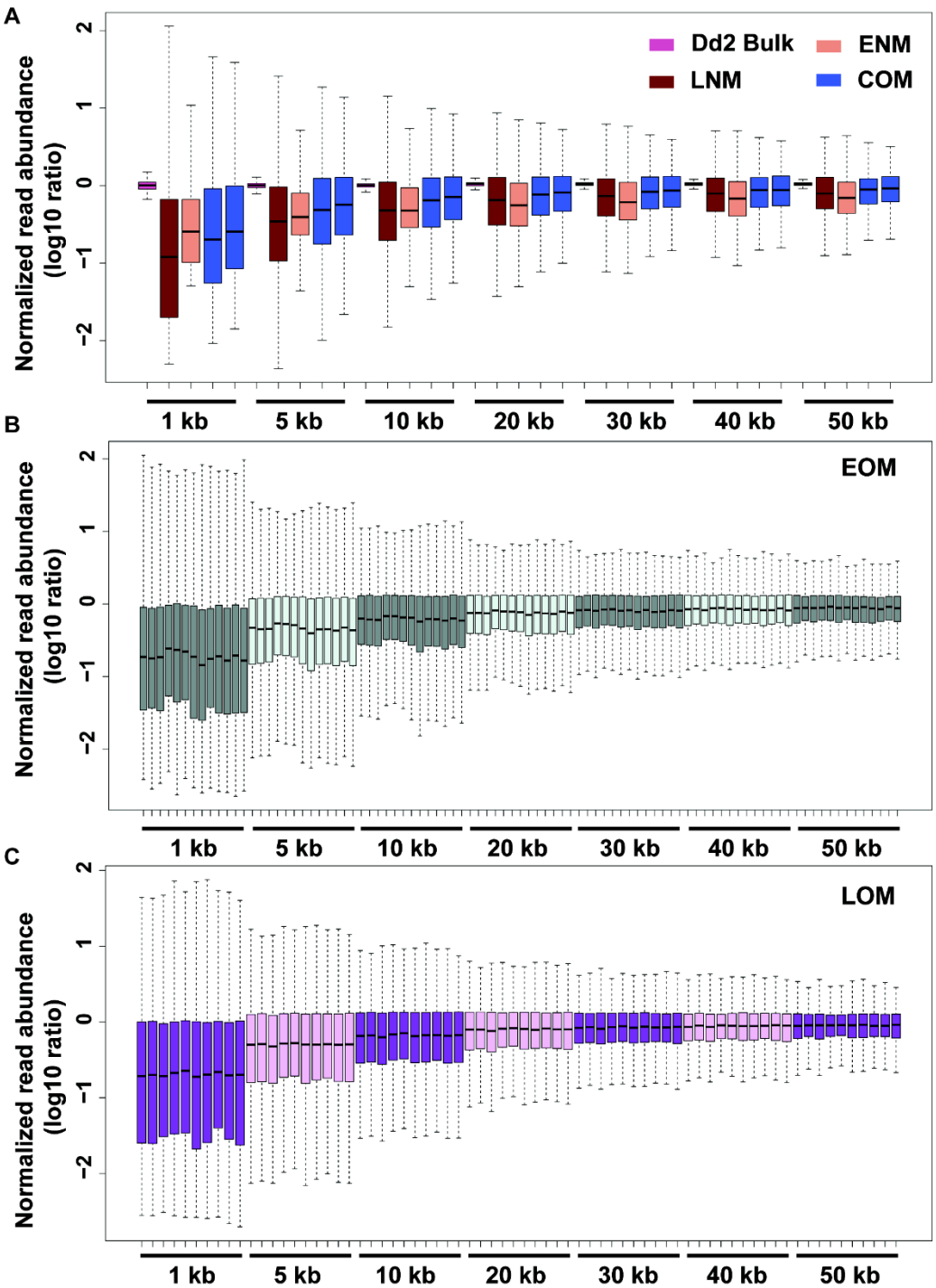


Figure from Liu et al. 2020, [165]. AH generated files for analysis and analyzed data for 1kb-10kb, SL analyzed 20kb-50kb and generated the figure.

The Log10 ratios of normalized read abundance in 1-50kb (at intervals of 5 and 10kb) are showed for sequenced samples. The boxes indicate Q1 (25th percentiles) to Q3 (75th percentiles) with a horizontal line drawn in the middle to denote the median. Outliers, above the highest point of the upper whisker (Q3 + 1.5×IQR) or below the lowest point of the lower whisker (Q1-1.5×IQR), are not displayed. **A. Distribution of normalized read counts in various bins sizes for select sample types.** Dd2 Bulk (purple), ENM (pink, 1 samples), LNM (maroon, 1 samples), COM (blue, 2 samples) samples. **B. Distribution of normalized read counts in various bin sizes for all EOM samples.** EOM (green,

n=13). **C. Distribution of normalized read counts in various bin sizes for LOM samples.** LOM (purple, n=10).

Calculating *P. falciparum* mappability

After preliminary coverage analysis and normalization, we determined that further methods for normalization were needed to attempt identification of CVNs. For this analysis, I analyzed the *Plasmodium falciparum* 3d7 reference genome, release 42. I utilized the GEM mappability suite to create "mappability" tracks of the *P. falciparum 3d7* reference genome (release 42) by first creating a GEM index of the genome and then computing the mappability for 50, 100, 150, and 300bp read lengths [193]. These GEM mappability files were then converted to wig files using the gem-2-wig program in the GEM suite of tools then visualized on IGV for broad characterization.

CNV analysis of *P. falciparum* single cell sequencing data

The "cleaned" alignments generated in the previous analysis were used for this section. The samples to be analyzed included a known *P. falciparum* CNV on chromosome 5 that includes the MDR1 gene. This CNV was known to occur from approximately 880,000 to 970,000 on chromosome 5.

For read depth analysis of these cells, we initially utilized bin sizes from 5kb through 10kb, but only the data for 7kb and 10kb are shown. This range was chosen as a trade-off between bin variation and the ability to detect medium to large CNVs (15kb or more). Read counts for each bin were determined by Bedtools V2.17.0 coverageBed [110]. To avoid read count bias arising from GC content, bins were grouped into 20 even quantiles of GC content using R. Counts were then normalized within those 20 groups; the median read count for the bin group (with outlier read counts excluded) was normalized to 1 and all other read counts were divided by the median read count. Outliers were considered greater than median + 4 MADs or less than median - 4 MADs. The 100-mer mappability track that I previously generated was then used to normalize each bin for probability of being able to map a read to each location within that bin. Single cell segmentation for CNVs was accomplished

on the normalized bin data using the R package DNAcopy with the parameters alpha = 0.0001, undo.SD = 0, and min.width = 5 [194]. Split-read and discordant read analysis of alignment data was conducted as previously using the Speedseq pipeline in order to identify a known amplification present within the cells [98].

RESULTS

Optimized MALBAC improves single cell *Plasmodium falciparum*-infected erythrocytes read coverage.

For the full results of this publication, see Liu et al. [165]. Our single cell sequencing pipeline for *P. falciparum* parasites included stage-specific parasite enrichment, isolation of single infected erythrocytes, cell lysis, whole genome amplification, pre-sequencing quality control, whole genome sequencing, and analysis steps (**Figure 5.2A**).

We collected parasites from either an *in vitro*-propagated laboratory line or from a blood sample of an infected patient (referred to as 'laboratory' and 'clinical' parasites, respectively). This allowed us to test the efficiency of our procedures on parasites from different environments with varying amounts of human host DNA contamination. Furthermore, for laboratory samples, we isolated both early (1n) and late (~16n) stage parasite-infected erythrocytes to evaluate the impact of parasite DNA content on the performance of WGA. For single cell isolation, we used the automatic microscopy-based CellRaft Air system (**Figure 5.2B**), which has the benefit of low volume capture procedures (minimum: 2µl).

Following isolation, we successfully amplified 3 early and 4 late stage individual cells from laboratory samples using the standard MALBAC protocol (termed non-optimized MALBAC, ENM and LNM respectively). We also applied a version of MALBAC that we optimized for the small AT-rich *P. falciparum* genome (termed optimized MALBAC) to 42 early (EOM) and 20 late stage (LOM) laboratory samples and 4 clinical samples. Post-amplification and purification DNA yields were detectable in all single cell samples, indicating successful amplification. Compared to standard MALBAC, our optimized protocol had a lower reaction volume, more amplification cycles, and used a modified pre-amplification random primer (see

Methods in Liu et al. 2020 publication for more details). Using this method, we successfully amplified 43% of the early and 90% of the late stage laboratory samples and 100% of the clinical samples.

Figure 5.2: Single *P. falciparum*-infected erythrocytes are isolated, amplified, and sequenced.

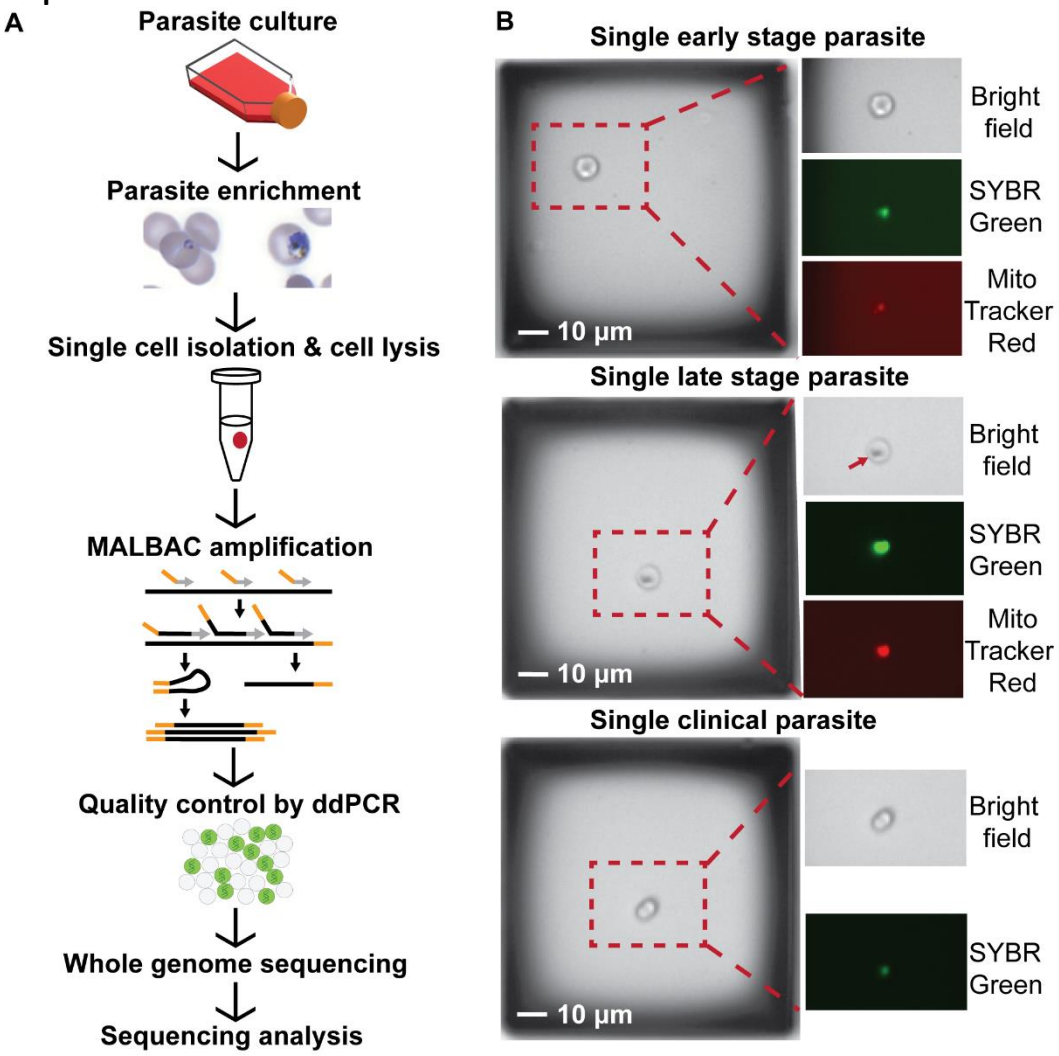


Figure from Liu et al. 2020, [165]. AH developed protocol for single cell isolation and application of non-optimized MALBAC to *Plasmodium*, SL optimized MALBAC, performed quality control, whole genome sequencing, and generated figure.

A. **Experimental workflow.** Parasite cultures are obtained and then enriched using column and gradient-based methods (see Liu et al 2020). Individual parasite-infected erythrocytes (see panel B) were automatically isolated into PCR tubes using the CellRaft AIR System (Cell Microsystems). All samples were lysed by combining a freeze–thaw step and treatment with a detergent prior to MALBAC amplification. MALBAC uses a combination of common (orange) and degenerate (grey) primer sequences to amplify the genome. The quality of amplified genomes was assessed prior to library preparation and sequencing using Droplet Digital (dd)PCR. **B. Parasite stage visualization on the CellRaft AIR System using microscopy** (10X magnification). Parasite-infected erythrocytes were seeded into microwells to yield only a single cell per well (left image of each group), and identified with SYBR green and Mitotracker Red staining (parasite DNA and mitochondrion, respectively). Early stage parasites exhibited lower fluorescence due to their smaller size and late stage parasites had noticeable dark spots (arrow) due to the accumulation of hemozoin pigment. Scale bar represents 10 µm.

Optimized MALBAC improves uniformity for single cell samples.

To investigate the uniformity of read abundance distributed over the *P. falciparum* genome, we divided the reference genome into 20kb bins and plotted the read abundance in these bins over the 14 chromosomes (**Figure 5.3A**). For this analysis, we selected a 20kb bin size based on its relatively low coverage variation compared to smaller bin sizes and similar coverage variation as the larger bin sizes. To quantitatively measure this variation, we calculated the normalized read abundance per bin in each sample (by dividing the raw read counts with the mean read counts per 20kb bin, **Figure 5.3B**). Indeed, the bulk control displayed the smallest range of read abundance for outlier bins (blue circles, **Figure 5.3B**) and lowest interquartile range (IQR) value of non-outlier bins (black box, **Figure 5.3B**), indicating less bin-to-bin variation in read abundance. Both EOM and LOM samples exhibited a smaller range of normalized read abundance in outlier bins than ENM and LNM samples (**Figure 5.3B**). In addition, the read abundance variation of COM samples was similar to EOM or LOM samples (**Figure 5.3B**). Finally, due to the extremely low coverage of the clinical bulk sample, the read abundance variation was much higher than all other samples (**Figure 5.3B**).

We then calculated the mean coefficient of variation (CV) for read abundance in the different sample types to compare the variation of read abundance in all sequenced samples (**Table 5.1, Figure 5.3C**).

Table 5.1: Coefficients of variation of normalized read abundance in each sample

Sample name	Mean Coefficient of Variation (CV)	SD
Dd2 Bulk (1)	22	-
ENM (1)	147	-
EOM (13)	89	4
LNM (1)	111	-
LOM (10)	79	2
COM (2)	87	12
Clinical Bulk (1)	472	-

SD, standard deviation.

Figure 5.3: Samples amplified by optimized MALBAC display improved uniformity of read abundance.

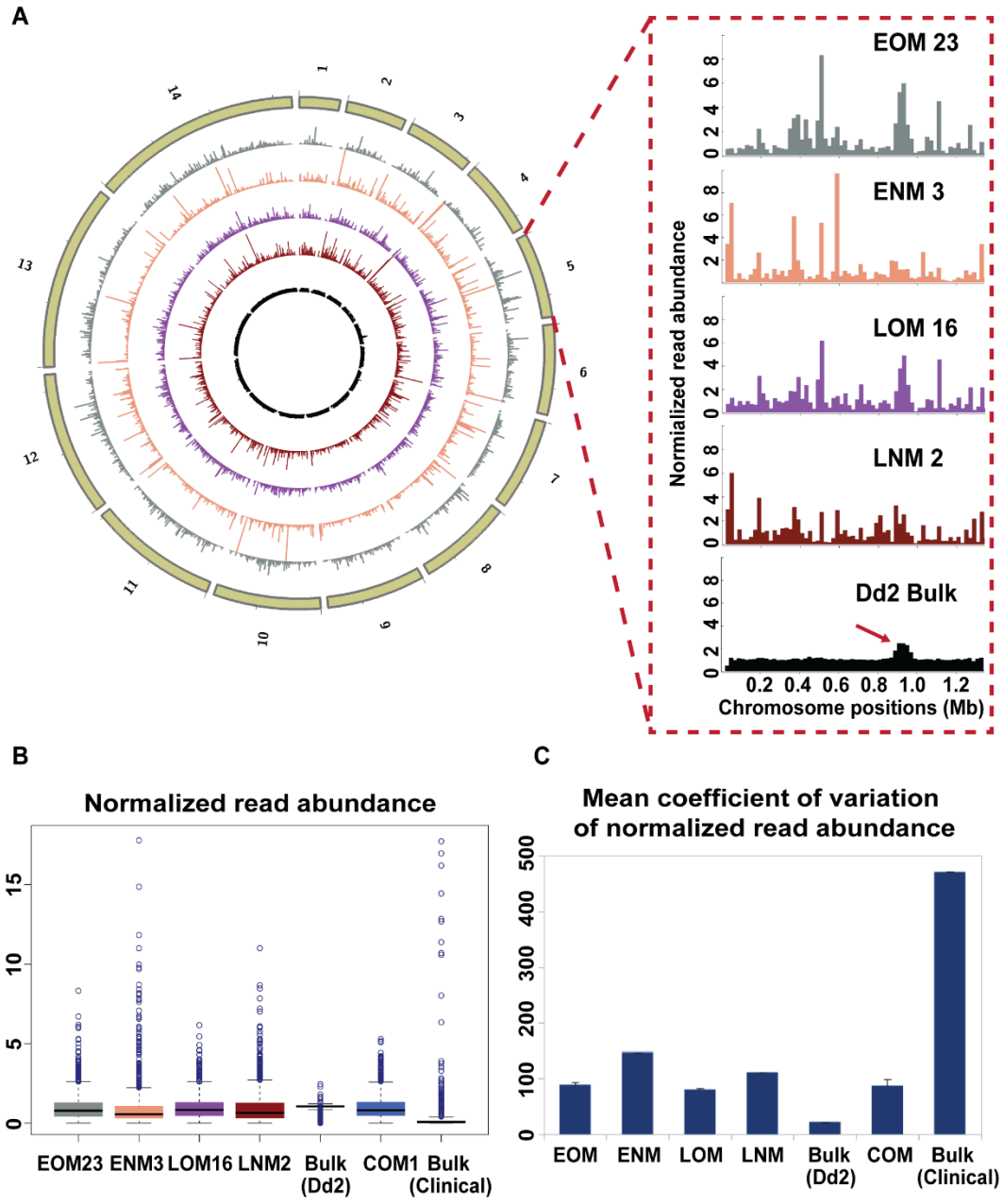


Figure from Liu et al. 2020, [165]. AH generated files for read-depth analysis and GC characteristics. SL did read-depth normalization and generated the figure.

A. Normalized read abundance across the genome. The reference genome was divided into 20kb bins and read counts in each bin were normalized by the mean read count in each sample. The circles of the plot represent (from outside to inside): chromosomes 1 to 14 (tan); one EOM sample (#23, grey); one ENM sample (#3, orange); one LOM sample (#16, purple); one LNM sample (#2, dark red); Dd2 bulk genomic DNA (black). The zoomed panel shows the read distribution across chromosome 5, which has a known copy number variation (arrow on Dd2 bulk sample). **B. Distribution of normalized read abundance values for all bins.** The midline represents the median normalized read abundance for each sample. Error bars represent the 25th (Q1) and 75th (Q3) percentiles. Outliers, identified by either 1.5×IQR (interquartile range) or more above Q3, or 1.5×IQR or more below Q1, are depicted with circles. One sample from each type is represented (see all samples in Figure S2 and Figure S3C). **C. Coefficient of variation of normalized read abundance.** The average and SD (error bars) coefficient of variation for

all samples from each type is represented (EOM: 13 samples; ENM: 1 sample; LOM: 10 samples; LNM: 1 sample; Dd2 Bulk: 1 sample; COM: 2 samples; Clinical Bulk: 1 sample). See *Methods* for calculation.

The CV from the ENM sample was significantly different when compared to the CV of each EOM sample (147% versus a mean of 89%, respectively, p value <0.01, **Table 5.1**). Similarly, the LNM-CV was significantly different when compared to the CV of each LOM sample (111% versus a mean of 79%, respectively, p value <0.01, **Table 5.1**). These data showed improvement in levels of read uniformity across the genome when using optimized MALBAC over the standard protocol. In support of this finding, the CV value of COM samples were similar to EOM and LOM samples (**Table 5.2, Figure 5.2C**).

Table 5.2 – Average coverage of sequenced samples

MALBAC type	Single cell	Sample type (#)	Coverage breadth		
			Whole genome	Genic coverage	Intergenic coverage
Optimized	Yes	EOM (13)	57.9%	78.0%	27.8%
		LOM (10)	57.3%	79.0%	25.0%
		COM (2)	48.0%	67.7%	18.5%
Non-optimized		ENM (1)	23.0%	34.4%	6.1%
		LNM (1)	47.4%	67.9%	16.9%
NA	No	Dd2_Bulk gDNA (1)	96.1%	97.0%	94.9%
		Clinical Bulk gDNA (1)	0.3%	0.3%	0.2%

Known *P. falciparum* CNVs are detected by single cell sequencing.

After preliminary read depth analysis and alignment characterization, we then attempted to find both rare and known CNVs within individual cells. The first attempt utilized our previous analysis pipeline to identify split-reads and discordant reads [53]. With sufficient depth, these reads would ignore any potential biases in amplification from MALBAC and would allow us to find copy number variations with high resolution. As a first investigation, we sequenced parasites with a known copy number variation on chromosome 5 and attempted to detect it in cells amplified by non-optimized MALBAC and our optimized MALBAC amplification. This approach allowed us to detect the CNV not only from bulk sequencing but also in 5/13 early stage parasites that were amplified with our optimized MALBAC procedure (EOM, **Table 5.3**).

Table 5.3 – MDR1 detection by discordant read pair and split-read analysis.

Condition	Proportion of samples MDR1 detected* (positive detection/total)	Mean Read Support (n)
Bulk	**1/1**	**67 (1)**
NM	**0/2**	**-**
EOM	**5/13**	**2.4 (5)**
LOM	**1/11**	**2 (1)**

*Full amplicon is detected

Unfortunately, we only detected the known amplification in sequencing data from 1 out of 11 late-stage cells amplified by our optimized MALBAC (LOM, **Table 5.3**). A particularly interesting finding from this analysis was the existence of two distinct known CNVs (MDR1) within the individual cells. The known copy number variation is found on chromosome 5 from ~880,000 to ~970,000bp. Of the EOM cells in which we identified the known CNV, the CNV was identified from ~889,900 to ~969,800 in two cells (19 and 23, **Table 5.4**) and from ~888,330 to ~970,200 in three cells (21, 25, and 27, **Table 5.4**).

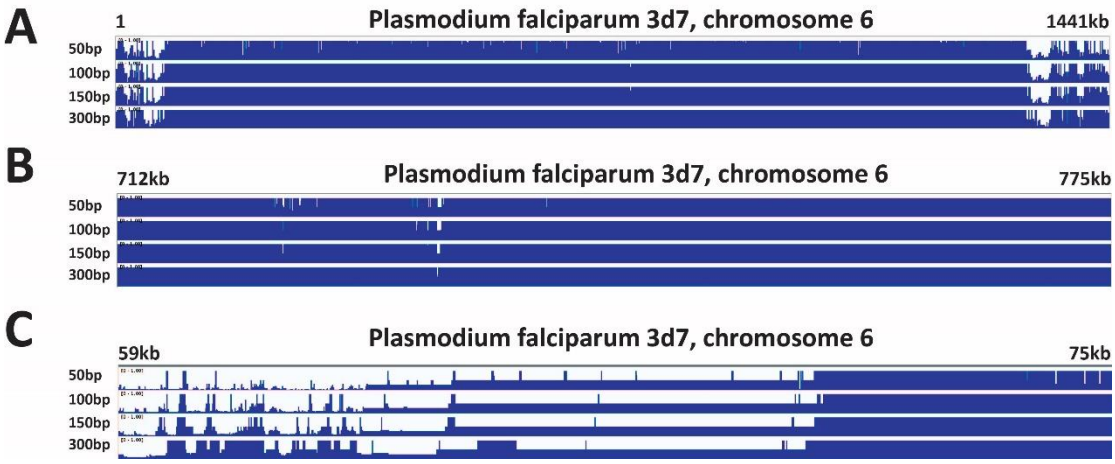
Table 5.4 – Discordant and split-read analysis identifies two distinct versions of MDR1 within the EOM population.

MALBAC condition	Cell	CNV Length	Chromosome	Start position (bp)	End Position (bp)	# of supporting reads (PE+SR)
EOM	19	79875	Pf3D7_05_v3	889928	969803	3 (2+1)
EOM	21	81831	Pf3D7_05_v3	888328	970159	1 (1+0)
EOM	23	79890	Pf3D7_05_v3	889899	969789	6 (3+3)
EOM	25	81921	Pf3D7_05_v3	888329	970250	1 (1+0)
EOM	27	81905	Pf3D7_05_v3	888346	970251	1 (1+0)

We next utilized a read depth analysis pipeline developed for sparsely sequenced individual human neurons [172]. However, our sequencing data was significantly deeper, performed on a much smaller genome with different sequence characteristics, and was done with a different amplification protocol. Thus, modifications were made including mappability calculations, GC-normalization methods, and overall binning. The mappability for the *Plasmodium falciparum 3d7* reference genome was largely very high with near perfect mappability in core genome sections on the interior of chromosomes (**Fig. 5.4A and B**). Regions of low mappability were found at telomeres and large gene families involved in antigenic variation as was expected (**Fig. 5.4C**). 50bp reads were still mostly mappable but

mappability increased as read length increased again as expected with standard read lengths of ≥ 100bp being highly mappable (**Fig 5.4**). It is important to note that this program estimates mappability for individual reads and does not account for mismatches allowed by alignment programs (caused by SNPs or through sequencing errors) and also does not take into account the reality of paired-end sequencing which more easily allows unique mappings.

Figure 5.4 - The *P. falciparum* core genome is mappable for reads greater than 50bp long.

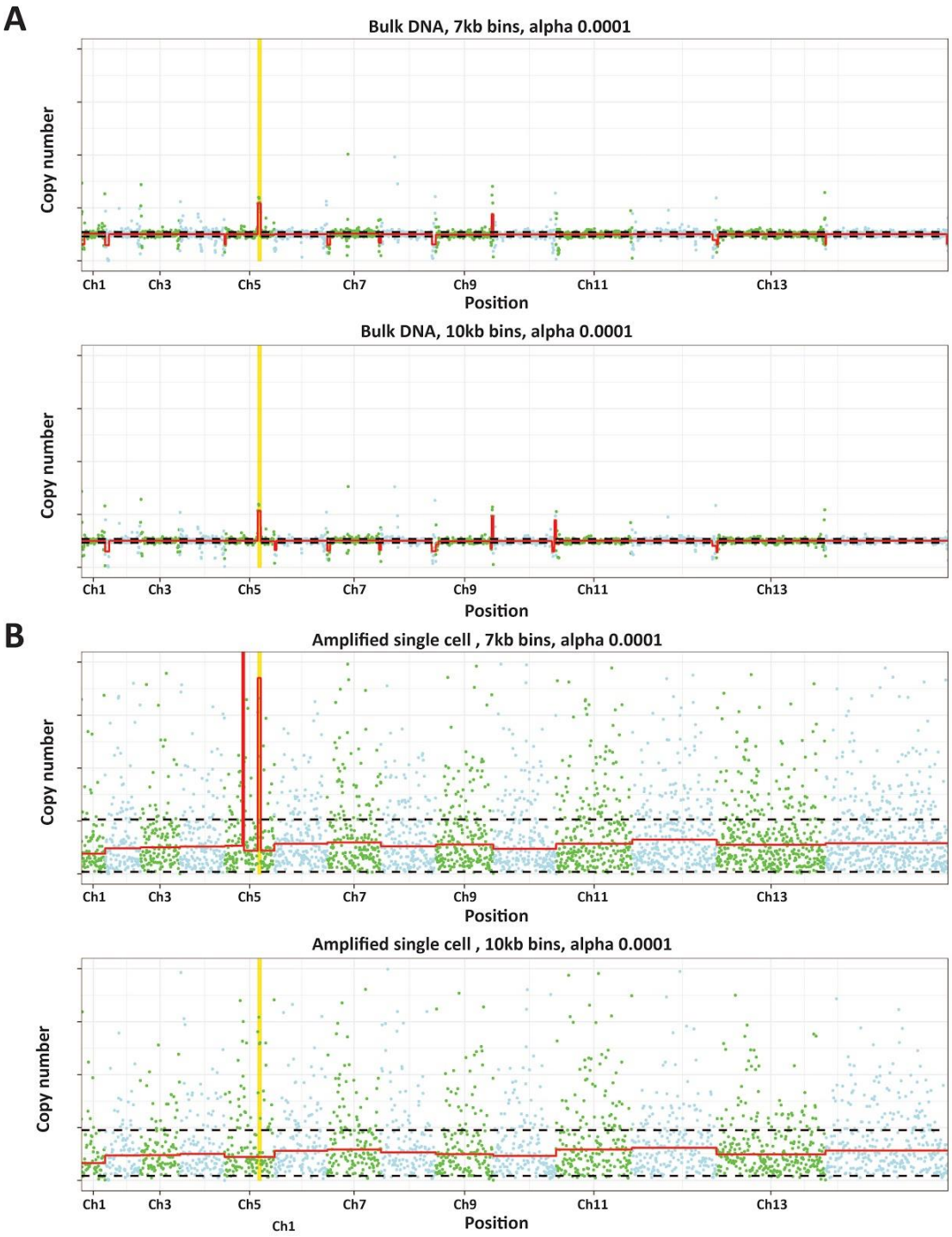


AH performed all work related to this figure.

Mappability file for Pf3d7 chromosome 6 were calculated using GEM mappability suite for 50, 100, 150, and 300bp read length. Files were converted to wig format and visualized on IGV. Probabilities are displayed in blue between 0 and 1 with the max height observed as 1. **A.** Pf3d7 chromosome 6. **B.** Core genome subsection of Pf3d7 chromosome 6. **C.** Telomeric subsection of Pf3d7 chromosome 6.

After these modifications, the known CNV was easy to detect in bulk sequencing regardless of bin size and significance cutoff (**Fig. 5.5A**). Cell 23 was one of the most promising cells as we were able to identify the known CNV through discordant/split-read approaches as well as read depth analysis (**Table 5.5**, **Fig. 5.4A**). Even under strict significance cutoffs, we were able to detect the known CNV with 7kb bins (**Fig. 5.5B**). However, our ability to detect the known CNV varied between bin sizes as we were unable to detect the CNV with 10kb bins (**Fig. 5.5B**). It is interesting to note that this analysis pipeline also identified another novel CNV on chromosome 5 in several cells that was not present in bulk sequencing (**Fig. 5.5)**. However, further mathematical analysis is needed to confidently say that these novel CNVs are the result of true signal and not due to bias in amplification.

Figure 5.5 – A known CNV is identifiable in bulk DNA and some single cells.



AH generated all bin files, mappability files, and GC bins. WC applied single cell analysis pipeline and generated the images used in this figure.

A. Bulk DNA analysis in 7kb and 10kb bins with stringent alpha 0.0001. Odd chromosomes are shown in green and even chromosomes in blue. Each bin for their respective size (7kb or 10kb) is graphed based upon its copy number state on the y-axis. The red line represents the segmentation output from DNAcopy. The vertical yellow bar is the position of the known CNV. **B. Amplified single cell 23 for both 7kb and 10kb bins.**

DISCUSSION

Discordant read pair and split-read analysis is a promising approach to the identification of CNVs from single cell sequencing data if relatively high coverage of single cells can be achieved (**Table 5.4** and **Table 5.5**). Discordant read pairs and split-reads identified the known CNV in almost half of the EOM samples and with further improvements to our amplification method, this is likely to improve. Furthermore, we identified that two separate versions of the known CNV within the single cell samples based upon these methods (**Table 5.4**), which is extremely promising for future studies. Two separate versions indicate that there is heterogeneity in either the creation of the CNVs or in their fate. CNVs typically carry an associated fitness cost either from changing the expression of off-target genes or from the necessity of greater DNA synthesis. If CNVs are slowly shortened to only include the necessary gene or mutation, the evolutionary advantage of the CNV would be maintained but the fitness cost would be removed. Given the relatively small differences between breakpoints (~200-300bp on either end), the ability to distinguish between these CNVs would be lost within bulk sequencing and would likely be judged as the same CNV with less breakpoint resolution.

Read depth analysis of the single cell sequencing data is also promising as we were able to detect the known CNV in cells with strict significance cutoffs as well as potentially novel CNVs (**Fig. 5.5B**). It is encouraging to note that the amplification reproducibility for our methods is fairly high (**Fig. 5.4**). This is especially helpful for read depth analysis as it allows approaches like cross-sample normalization to succeed by controlling for all noise regardless of source [168, 195]. The major challenge for continued read depth analysis is the remaining relatively high variance in coverage, likely due to sparse coverage in intergenic regions of the genome (**Table 5.3**). One possibility for future analysis is to exclude intergenic regions and solely analyze coverage in genic regions (similar to exome sequencing analysis). However, future improvements in the amplification method to improve coverage in intergenic regions and analytical methods to control for background noise in our methodology will allow us to confidently identify novel CNVs from individual parasites.

Computational investigation of *P. falciparum* extrachromosomal DNA

INTRODUCTION

Plasmodium's ability to adapt to different stressors through heterogeneity has previously been discussed; however, there are other mechanisms that could contribute to its flexibility. One major contributor that is frequently overlooked is extra-chromosomal DNA (ecDNA). The contribution of gene duplication has been previously discussed in other chapters but ecDNA constitutes another source of extra copies of DNA. ecDNA has been reported to contribute to the fitness of such diverse organisms as the *Leishmania* and *Trypansoma* parasites, yeast, human cancer, mammalian cells [136, 196-203].

Our lab began studying the genetic development of drug resistance with a set of parasites that had developed drug resistance to under controlled selection with the novel antimalarial, DSM1 [61]. The parasites first duplicated and then later amplified a segment of their genome to facilitate overexpression of the drug target and thus titrate out the drug. An interesting result was that in separate distinct clones selected in separate cultures developed different DNA breakpoints. Several interesting observations of these clones led to the hypothesis that the parasites were creating extrachromosomal DNA (ecDNA) as a further means of drug resistance. When they were subjected to higher concentrations of DSM1, the parasites further amplified the drug resistance associated CNV to as many as 10 chromosomal copies but each clone maintained the same sequence boundaries for each new copy. This demonstrated that increased CNV copies were not created *de novo* each time but were instead the amplification of the previous copies. More importantly, the drug resistance was not directly proportional to the apparent CNV copy number. When measured by qPCR, the H1 clone (high level resistance), had ~10x higher EC50 than another clone with the same observed copy number and we therefore hypothesized the existence of ecDNA.

Jennifer McDaniels definitively identified ecDNA in the highly resistant H1 clones using complex electrophoresis-based purification and multiple highly sensitive

DNA analysis methods including droplet digital PCR. These methods identified resistance-conferring genes outside of the chromosomes. Enzymatic digestions of the ecDNA showed that there were two separate structures of ecDNA but we attempted to investigate their precise sequence using whole genome sequencing.

After Jennifer McDaniels put in vast efforts to isolate, characterize, and prepare the ecDNA for sequencing I applied our previous whole genome sequencing analytical methods to see if we could find identify features unique within the ecDNA samples. Analysis has proven challenging but we present evidence of specific differences in the whole genome sequencing results.

METHODS AND MATERIALS

This text has been adapted from Jennifer McDaniels et. al, in preparation to only include the whole genome sequencing analysis that I performed. For the full experimental methodology and further information on the ecDNA, please consult the full paper, McDaniels et al. in preparation.

Methods for this analysis were previously reported in (Huckaby et al., 2019). For whole genome sequencing analysis, we used BBtools to trim adapter sequences and remove low-quality reads [189]. Remaining unmapped reads that did not align to the *P. falciparum* Dd2 genome were then blasted (NCBI database) to determine % contamination with sequences that aligned to other organisms using Geneious [158, 204]. After evaluation, human and bacterial read contamination were removed by aligning the reads to the human hg19 reference genome and top 3 bacterial genomes from BLAST using BBMAP version 38.33 [189, 204]. BBmap alignment options included a minimum of 95% identity, max indels of 3, a minimum of two seed hits, with quick match and fast modes enabled. Unmapped reads from this step were used for subsequent alignment and CNV analysis.

Lastly, to determine the orientation of the amplicon (tandem or reverse tandem duplication) discordant reads were visually inspected at the breakpoints using IGV 2.4.10 [101]. Two algorithms, CNVnator, which call CNVs using read depth, and LUMPY, which calls CNVs using discordant reads, were then used to further evaluate copy number variations after BWA-MEM was used to align reads with

default settings to the Dd2 genome (PlasmoDB release 42) [96, 99, 100]. QualiMap 2 was also used to evaluate the mapping quality of reads and number of reads that aligned to the *Plasmodium* genome [97]. Analysis of LUMPY was used to determine the location and length of the DSM1 CNV using paired-ends and split read alignments. CNVnator was used to confirm the location of the CNV and an estimation of copy number was provided by read depth analysis using 1000bp.

RESULTS

When comparing read coverage across the *dhodh* amplicon relative to that from the entire chromosome 6, we observed very high enrichment of the amplicon in the gel-incompetent material (mean of ~170-fold, **Table 5.5**) and conservation of the amplicon boundaries with those from genomic DNA (**Figure 5.6**).

Table 5.5: Summary of coverage enrichment at known CNVs

Samples		H1 genomic DNA	H1 gel- incompetent DNA*
Chromosome 6 coverage	**Minus *dhodh* amplicon**	6.5x	3.6x
	***dhodh* amplicon only**	79x	604x**
	Estimated CN	12	170**
Mitochondrial genome coverage	**MT-CYB coverage**	286.6x	683.5x
	Estimated CN	32	76
Chromosome 5 coverage	**Minus *mdr1* amplicon**	8.6x	5x
	***mdr1* amplicon only**	29x	14x
	Estimated CN	**3**	**3**
Chromosome 12 coverage	**Minus *gch1 amplicon***	8.7x	6.2x
	***gch1* amplicon only**	54x	1.5x
	Estimated CN	**6**	**0**

CN, copy number; dhodh, dihydroorate dehydrogenase; published copies 8-10 [61]*; mt-cyb*, *cytochrome b*; published copies 20-150 (Lane et al., 2018); *mdr1*, *multidrug resistance protein 1*, published copies 2-3 (Triglia et al., 1991); *gch1, GTP cyclohydrolase 1* published copies 2 (Anderson et al., 2009). *This sample was isolated form the loading well of a PFGE gel and amplified using a DNA amplification kit to generate enough material for sequencing. **This amplicon includes the super-peak region detailed in Supplemental Table 2**.** Without this region, the estimated CN is 15 copies.

Figure 5.6: In-depth investigation of H1 gDNA and gel-incompetent DNA revealed shared *dhodh* amplicon boundaries and a super-peak unique to gel-incompetent DNA.

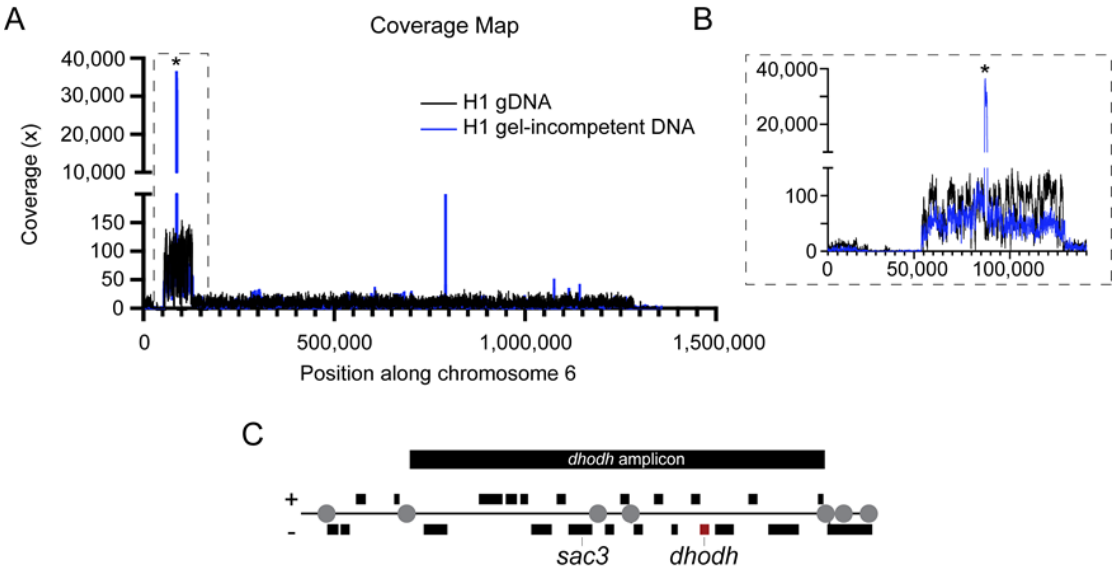


Figure from McDaniels et al. in prep. AH generated files for coverage analysis and analyzed amplicon boundaries/initial coverage, JM continued coverage analysis and generated figure.

A. Coverage map of gDNA (black) compared to gel-incompetent DNA (blue). Enrichment of region partially spanning *sac3 domain-containing protein, putative (sac3)* (PlasmoDB ID: PF3D7_0602600) is overrepresented at 36,636-fold (asterisk, super-peak). **B.** Exact boundaries of the amplicon are the same in both DNA samples. Dashed box are enlarged from panel A. **C**. Location of *dhodh* (red box) and *sac3* within the full *dhodh* amplicon (black box, top) adapted from Guler et al. 2013. Coverage was analyzed using Integrative Genomics Viewer Software (IGV 2.4.10). H1, highly resistant clone; Dashed box, ~70kb *dhodh* amplicon; grey circles, location of A/T tracks; unidentified black boxes, genes.

Additionally, we discovered an A/T-rich (88.2%), 714bp sequence found specifically within the *dhodh* amplicon of the gel-incompetent ecDNA; this region is drastically over-enriched (**Figure 5.6**). We termed this the "super-peak" due to a maximum coverage of >36,000-fold and a mean coverage of >25,000-fold (**Figure 5.6**). The mean coverage of the full ~70kb *dhodh* amplicon including the super-peak is 604x (**Table 5.5**) and excluding the super-peak is 55x (**Table 5.5, see footnote**). Initially, we suspected that the extremely high coverage at this region is due to an artifact of the DNA amplification method; if we exclude this particular region, we estimated an expected number of *dhodh* amplicons (~15 copies, **Table 5.5**). Indeed, other small regions of the genome are over-amplified in the well-derived sample (**Figure 5.6A**), although not to the same extent (mean of ~240-fold). Analysis of discordant reads at this location revealed that copies of the amplified region that

make up the super-peak are arranged in a tandem head-to-tail orientation (**Figure 5.7A** and **C**).

Figure 5.7. Orientation of discordant reads at *sac3* super-peak position and *dhodh* amplicon is indicative of tandem duplication.

Figure from McDaniels et al. in prep. AH generated the files for read-pair analysis and set-up visualization in IGV. JM generated the figure.

A. Schematic of a tandem duplication which illustrates paired-end reads pointing outwards. **B.** Schematic of an inverted duplication which illustrates paired reads pointing in the same direction. **C.** IGV image of paired reads at the super-peak. The super-peak includes the *sac3 domain-containing, putative protein* found on chromosome 6 at position 86,429 - 87,143bp. H1 gel-incompetent DNA is sequenced and the paired-ends are aligned to the WT1 reference genome. **D**. IGV image of paired reads of the *dhodh* breakpoints matches previously reported tandem duplication (Guler et al., 2013). Due to size of the amplicon, boundaries of reads are shown in a split screen. Reads were analyzed using Integrative Genomics Viewer Software (IGV 2.4.10) and does not depict the total reads at those locations. Colored arrows, discordant reads; green arrows depict tandem duplications and blue arrows depicts inverted duplications.

This result is similar to the known orientation of chromosomal copies of the *dhodh* amplicon ([61], **Figure 5.7D**). However, we do not detect this pattern in other over-amplified regions (data not shown), likely because MDA can create chimeric reads or randomly connected sequences due to template switching during high polymerase processivity [184]. Due to the extreme level of over-amplification and read orientation across this region, the super-peak is likely to represent a sequence that was present prior to WGA steps and therefore, may hold biological significance. A targeted analysis of this region in the non-amplified sample is precluded by the high A/T content of this region (88.2%), which makes the design of specific PCR primers impractical.

DISCUSSION

From these studies, we identified that the chromosomal amplicon is precisely conserved within the ecDNA as there are no observable differences between their boundaries and (**Figure 5.6** and **Figure 5.7**). Unfortunately, these shared features add

to the difficulty in purifying ecDNA away from the genome and the recognition of unique features. If we are able to better purify the ecDNA away from the genome, there are novel methods that we could potentially utilize such as stringent *de novo* assembly using the BBMap Tadpole tool and then comparing the resulting contigs. If this result works, we may be able to generate the circular ecDNA sequences [189]. Another possibility is kmer feature identification to directly compare the profiles of reads against each other [205].

Deep sequencing of this region did reveal one feature that was unique to ecDNA: the Sac3 super-peak. We found that this was a small highly A/T-rich region of the amplicon was greatly overrepresented in well-derived material (asterisk, **Figure 5.6B, 5.6C,** and **Table 5.6**, termed the super-peak). This region encompassed a portion of the upstream UTR and 5' end of the gene for the SAC3 domain-containing protein (PlasmoDB gene ID: PF3D7_0602600 [206]. This peak is unlikely to be an artifact of amplification due to the fact that chimeras created through MDA amplification are typically in an inverted orientation and there are too many supporting reads for this feature (**Figure 5.7**).

An interesting observation is that the super-peak is extremely A/T-rich and a tandem duplication orientation. These features are also found in *Plasmodium* replication origins and centromeres [160, 207-210]. It has been previously identified that short 400-500bp A/T-rich sequences can serve as replication origins [207, 208]. If there is an origin of replication or even multiple origins within the ecDNA this might explain the super-peak but does not explain why we don't see increased coverage of sequences surrounding it or the resistance conferring *dhodh* gene. Another extremely A/T-rich genome features is the centromere. However, all of the work done on *Plasmodium* centromeres currently show an average 4-4.5kb length with a 2-2.5kb extremely A/T-rich repeat [211]. Jennifer McDaniels has speculated that the super-peak sequence may serve as a binding scaffold to increase the translation rate of a nearby protein target, such as SAC3 itself or nearby DHODH [212]. Alternatively, this sequence may have a role in the maintenance of the ecDNA element. Past studies linked the stability of transfected episomes in malaria parasites with A/T-rich centromere-like elements, which increases the efficiency of mitotic

segregation and it is possible that the super-peak could perform a similar function in ecDNA [209, 211].

Chapter 6: Conclusions and future directions

6 Conclusions and future directions

CNV creation mechanisms in *Plasmodium spp.*

Malaria is in an evolutionary arms-race with humans and has had the strongest influence on the evolution of the human genome in recent history [213]. Many features of the *Plasmodium* life-cycle lend themselves to evolvability and thus it is critical to understand the mechanisms by which *Plasmodium spp.* evolve. *Plasmodium falciparum,* the primary cause of malaria-related morbidity and mortality, has been able to adapt to every antimalarial drug thus far. Previous work by our lab indicated that CNVs may be the first step in the development of drug antimalarial resistance [61]. In order to investigate this possibility, our research turned to identifying genome features associated with DNA damage and the subsequent DNA repair mechanisms utilized by *Plasmodium spp.*

Our research began with the identification of novel genetic features utilized by *P. falciparum* to create CNVs (**Chapter 3**). I adapted a whole genome sequencing analysis pipeline that gave us the ability to identify previously known CNV breakpoints with high sensitivity and specificity with near single base-pair resolution. Several previous observations had found that the breakpoints of drug-resistance associated CNVs were found in long homopolymeric A/T tracks but our study identified that virtually all of them were found within these sequences in a particular orientation. However, A/T tracks have not previously been shown to a source of DNA breakage and thus we worked on identifying the initial cause of DNA double-stranded breaks. Alternative DNA structures including hairpins and stem-loops were previously shown to cause DNA double-stranded breaks and facilitate the formation of CNVs in other organisms and were implicated in recombination of genes involved in cell adhesion and immune system evasion in *Plasmodium falciparum*. We confirmed the presence of highly stable alternative DNA structures in close proximity to the A/T track breakpoints. These two features at a particular distance indicate a CNV “trigger site” profile. Both stable hairpins and long A/T tracks are overrepresented in the *Plasmodium falciparum* genome and we hypothesize that they explain why *P. falciparum* has adapted to every antimalarial drug thus far.

We next investigated the CNV trigger site model in three other species of *Plasmodium:* the second most clinically relevant species *P. vivax,* the human-infective zoonotic species *P. knowlesi¸* and the A/T-rich avian malaria *P. relictum* (**Chapter 4**). By applying our previous analysis pipeline, we identified that the breakpoints of CNVs in *P. vivax* and *P. knowlesi* were also found in long A/T tracks. When investigating trigger sites on a genome-wide scale, we found that long A/T tracks are overrepresented in all four species of *Plasmodium.* Even given the 80.6% A/T content of the *P. falciparum* genome, long A/T tracks were overrepresented. Despite being closer to ~60% A/T, *P. knowlesi* and *P. vivax* were also enriched in trigger site features. These genome-wide observations were conserved when looking at syntenic chromosomes as well.

All of these data point towards an evolutionarily conserved CNV generation mechanism within the *Plasmodium* species. We propose that antimalarial treatment, which causes metabolic stress, skews DNA repair towards two repair pathways in *Plasmodium spp.:* microhomology-mediated end joining (MMEJ) and microhomology-mediated break induced replication (MMBIR). In *P. falciparum* clones, we identified the possible hallmarks of both MMEJ (deletions and single nucletotide insertions within the long A/T track breakpoint) and MMBIR (short repeat expansions in close proximity to the breakpoints) which indicates replication fork slippage during a replication mediated repair process (**Chapter 3**). MMBIR is capable of causing both of these features and is more likely to be the causative DNA repair pathway. A recent publication indicates that all four species we investigated possess the exact proteins (Rev1 and Polymerase Zeta) that are thought to drive MMBIR whereas *Plasmodium* species that infect rodents do not [214, 215]. In order to further investigate this mechanism of CNV creation in *Plasmodium*, there are several possible avenues forward.

Another method of studying these trigger sites and mechanisms of CNV creation is through expanding the identification of CNVs and our comparative genomics to other *Plasmodium spp.* Unfortunately, there are currently no known CNVS and few whole genome sequencing data sets of *P. relictum* however we would expect CNV breakpoints to also be found in long A/T tracks at our trigger sites in this

species. It would especially interesting to investigate CNVs and the genome of rodent malaria to determine if they also are enriched in trigger site features. These comparisons would bolster evidence for the conservation of a novel CNV creation mechanism in *Plasmodium spp*.

Another area of expansion is the investigation of the other type of CNV, deletions. My previous work only assessed amplifications, repair of DSB breaks at these sites is equally likely to lead to deletions. I hypothesize that the same trigger point features contribute to genomic deletions which are also critical to the adaptability of the parasite. Future investigations of deletions of parent and daughter *P. falciparum* clones will add breakpoints and stable hairpins to expand our trigger site model of gene duplication. The inclusion of deletions into our analysis will also afford us more opportunities to identify the usage of specific DNA repair pathways in repairing the breakpoints as we did previously (**Chapter 3**).

The most definitive method of identifying DNA repair pathways and expanding our trigger site analysis is through specifically mapping DNA double-stranded breaks. Previous studies in other organisms have shown that aphidicolin can induce replicative repair mechanisms in the absence of NHEJ, which all *Plasmodium* species lack [134, 144, 216]. We have preliminary data that not only does aphidicolin inhibit DNA replication in *P. falciparum* but it also increases DNA breakage. We believe that many of these break sites would be found near or within long A/T tracks. We are currently adapting a modified DSBcapture protocol created by collaborators to map DNA DSBs which will provide an unbiased method of investigating our trigger site model under different conditions including aphidicolin treatment, antimalarial treatment, and nutrient deprivation which are all known to enhance DNA replication associated damage. Thus far I have developed a protocol to isolate high-molecular weight (HMW) DNA >50kb in length from untreated *P. falciparum* and demonstrated our ability to identify a single break site within the bulk DSBcapture prep. The final steps in this analysis are the creation of an Illumina sequencing library and subsequent identification of break-site peaks. Through the identification of CNV trigger sites and DNA repair pathways utilized by *Plasmodium*

spp., we may be able to block a critical evolutionary strategy and the development of antimalarial drug resistance.

Computational investigation of CNV heterogeneity in malaria

A remaining question is the extent of *Plasmodium* CNV heterogeneity. The previously mentioned study that estimated the creation of "~6 million base pair substitutions, 55 million indels, and 4 million newly created mosaic var exon 1 sequences every 2 days" indicates that there is large reservoir of hidden heterogeneity. I helped identify two novel sources of heterogeneity in *P. falciparum:* CNVs hidden within bulk sequencing and extrachromosomal DNA (**Chapter 5**). The identification of these sources of heterogeneity and other structural variants such as inversions, translocations, and complex mutations require accurate reference genomes for comparison, specialized computational approaches, and further automation. It is a well-known fact that the identification of CNVs is more robust than other structural variants and there are several approaches that may help bridge this gap in the malaria field.

Up until recently there was only a single reference genome for *P. falciparum* and the reference genomes for other species are incomplete (**Chapter 4**). The field of long-read sequencing is not only promising for CNV identification but also *de novo* assembly of near complete malaria genomes [217]. My protocol for isolating HMW DNA is highly applicable to this and is already being utilized by our lab for long-read sequencing and *de novo* assembly. Long read sequencing is also useful for the identification of CNVs also may be useful in the identification of hidden CNVs within bulk DNA samples.

Another factor that would greatly facilitate identification of CNVs and heterogeneity is the creation of a database of known "gold-standard" structural variants. This approach has been utilized in the study of human CNVs and facilitates the creation and optimization of structural variation detection tools [46, 100]. The combination of different existing structural variant identification tools is also promising. We utilized an analysis pipeline that combined a read-depth approach and split-read/discordant read-pair approach and obtained improved data. Other

newer pipelines combine multiple versions of these approaches as well as *de novo* assembly which would be helpful for *Plasmodium* analysis [218-220].

7 REFERENCES

1. Institute of Medicine Committee on the Economics of Antimalarial, D., in *Saving Lives, Buying Time: Economics of Malaria Drugs in an Age of Resistance*, K.J. Arrow, C. Panosian, and H. Gelband, Editors. 2004, National Academies Press (US): Washington (DC).
2. Vogel, G., *The forgotten malaria.* Science, 2013. **342**(6159): p. 684-7.
3. Garrido-Cardenas, J.A., et al., *Plasmodium genomics: an approach for learning about and ending human malaria.* Parasitol Res, 2019. **118**(1): p. 1-27.
4. Collins, W.E. and G.M. Jeffery, *Plasmodium malariae: parasite and disease.* Clinical microbiology reviews, 2007. **20**(4): p. 579-592.
5. Mueller, I., P.A. Zimmerman, and J.C. Reeder, *Plasmodium malariae and Plasmodium ovale--the "bashful" malaria parasites.* Trends in parasitology, 2007. **23**(6): p. 278-283.
6. Sutherland, C.J., et al., *Two nonrecombining sympatric forms of the human malaria parasite Plasmodium ovale occur globally.* J Infect Dis, 2010. **201**(10): p. 1544-50.
7. Organization, W.H., *World malaria report 2019.* 2019.
8. Cibulskis, R.E., et al., *Malaria: Global progress 2000 - 2015 and future challenges.* Infectious diseases of poverty, 2016. **5**(1): p. 61-61.
9. Howes, R.E., et al., *Global Epidemiology of Plasmodium vivax.* The American journal of tropical medicine and hygiene, 2016. **95**(6 Suppl): p. 15-34.
10. Singh, B. and C. Daneshvar, *Human infections and detection of Plasmodium knowlesi.* Clinical microbiology reviews, 2013. **26**(2): p. 165-184.
11. Karunajeewa, H. and J. Berman, *Is the Epidemiology of Plasmodium knowlesi Changing, and What Does This Mean for Malaria Control in Southeast Asia?* Clinical Infectious Diseases, 2019. **70**(3): p. 368-369.
12. Talman, A.M., et al., *Gametocytogenesis: the puberty of Plasmodium falciparum.* Malar J, 2004. **3**: p. 24.

13. Walliker, D., R. Carter, and S. Morgan, *Genetic recombination in Plasmodium berghei.* Parasitology, 1973. **66**(02): p. 309-320.
14. Blasco, B., D. Leroy, and D.A. Fidock, *Antimalarial drug resistance: linking Plasmodium falciparum parasite biology to the clinic.* Nat Med, 2017. **23**(8): p. 917-928.
15. Beier, J.C., *MALARIA PARASITE DEVELOPMENT IN MOSQUITOES.* Annual Review of Entomology, 1998. **43**(1): p. 519-543.
16. Smith, T.G., D. Walliker, and L.C. Ranford-Cartwright, *Sexual differentiation and sex determination in the Apicomplexa.* Trends in Parasitology, 2002. **18**(7): p. 315-323.
17. Janse, C.J., et al., *DNA synthesis in gametocytes of Plasmodium falciparum.* Parasitology, 1988. **96 (Pt 1)**: p. 1-7.
18. Guttery, D.S., A.A. Holder, and R. Tewari, *Sexual Development in <italic>Plasmodium</italic>: Lessons from Functional Analyses.* PLoS Pathog, 2012. **8**(1): p. e1002404.
19. Neafsey, D.E., et al., *The malaria parasite Plasmodium vivax exhibits greater genetic diversity than Plasmodium falciparum.* Nat Genet, 2012. **44**(9): p. 1046-50.
20. Chang, H.-H., et al., *Malaria life cycle intensifies both natural selection and random genetic drift.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(50): p. 20129-20134.
21. Creasey, A., et al., *Genetic Diversity of Plasmodium falciparum Shows Geographical Variation.* The American Journal of Tropical Medicine and Hygiene, 1990. **42**(5): p. 403-413.
22. Wootton, J.C., et al., *Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum.* Nature, 2002. **418**(6895): p. 320-323.
23. Fecchio, A., et al., *Climate variation influences host specificity in avian malaria parasites.* Ecology Letters, 2019. **22**(3): p. 547-557.
24. Llorà-Batlle, O., E. Tintó-Font, and A. Cortés, *Transcriptional variation in malaria parasites: why and how.* Briefings in Functional Genomics, 2019. **18**(5): p. 329-341.

25. Hamilton, W.L., et al., *Extreme mutation bias and high AT content in Plasmodium falciparum.* Nucleic acids research, 2017. **45**(4): p. 1889-1901.

26. Foley, M. and L. Tilley, *Quinoline antimalarials: mechanisms of action and resistance and prospects for new agents.* Pharmacol Ther, 1998. **79**(1): p. 55-87.

27. Reddy, P. and J.P. Flaherty, *Plasmodium vivax malaria relapses after primaquine prophylaxis.* Emerg Infect Dis, 2006. **12**(11): p. 1795-6.

28. Sidhu, A.B.S., et al., *Decreasing pfmdr1 Copy Number in Plasmodium falciparum Malaria Heightens Susceptibility to Mefloquine, Lumefantrine, Halofantrine, Quinine, and Artemisinin.* The Journal of infectious diseases, 2006. **194**(4): p. 528-535.

29. Nateghpour, M., S.A. Ward, and R.E. Howells, *Development of halofantrine resistance and determination of cross-resistance patterns in Plasmodium falciparum.* Antimicrobial agents and chemotherapy, 1993. **37**(11): p. 2337-2343.

30. Cottrell, G., et al., *Emergence of resistance to atovaquone-proguanil in malaria parasites: insights from computational modeling and clinical case reports.* Antimicrob Agents Chemother, 2014. **58**(8): p. 4504-14.

31. Dondorp, A.M., et al., *Artemisinin resistance in Plasmodium falciparum malaria.* N Engl J Med, 2009. **361**(5): p. 455-67.

32. Noedl, H., et al., *Evidence of artemisinin-resistant malaria in western Cambodia.* N Engl J Med, 2008. **359**(24): p. 2619-20.

33. Anderson, T.J.C. and C. Roper, *The origins and spread of antimalarial drug resistance: Lessons for policy makers.* Acta Tropica, 2005. **94**(3): p. 269-280.

34. Corredor, V., et al., *Origin and dissemination across the Colombian Andes mountain range of sulfadoxine-pyrimethamine resistance in Plasmodium falciparum.* Antimicrob Agents Chemother, 2010. **54**(8): p. 3121-5.

35. Mita, T., *Origins and spread of pfdhfr mutant alleles in Plasmodium falciparum.* Acta Trop, 2010. **114**(3): p. 166-70.

36. Krudsood, S., et al., *Efficacy of atovaquone-proguanil for treatment of acute multidrug-resistant Plasmodium falciparum malaria in Thailand.* Am J Trop Med Hyg, 2007. **76**(4): p. 655-8.

37. Briolant, S., et al., *Susceptibility of Plasmodium falciparum Isolates to Doxycycline Is Associated with pftetQ Sequence Polymorphisms and pftetQ and pfmdt Copy Numbers.* The Journal of Infectious Diseases, 2010. **201**(1): p. 153-159.

38. Dharia, N.V., et al., *Genome scanning of Amazonian Plasmodium falciparum shows subtelomeric instability and clindamycin-resistant parasites.* Genome Res, 2010. **20**(11): p. 1534-44.

39. Armstrong, C.M., et al., *Resistance to the antimicrobial agent fosmidomycin and an FR900098 prodrug through mutations in the deoxyxylulose phosphate reductoisomerase gene (dxr).* Antimicrob Agents Chemother, 2015. **59**(9): p. 5511-9.

40. Spangenberg, T., et al., *The Open Access Malaria Box: A Drug Discovery Catalyst for Neglected Diseases.* PLOS ONE, 2013. **8**(6): p. e62906.

41. Cowell, A.N., et al., *Mapping the malaria parasite druggable genome by using in vitro evolution and chemogenomics.* Science, 2018. **359**(6372): p. 191-199.

42. Darling, A.C.E., et al., *Mauve: multiple alignment of conserved genomic sequence with rearrangements.* Genome research, 2004. **14**(7): p. 1394-1403.

43. Passarge, E., B. Horsthemke, and R.A. Farber, *Incorrect use of the term synteny.* Nature Genetics, 1999. **23**(4): p. 387-387.

44. Hastings, P.J., et al., *Mechanisms of change in gene copy number.* Nat Rev Genet, 2009. **10**(8): p. 551-564.

45. Ohno, S., *Evolution by gene duplication*. 1970, Berlin, New York,: Springer-Verlag. xv, 160 p.

46. Sudmant, P.H., et al., *An integrated map of structural variation in 2,504 human genomes.* Nature, 2015. **526**(7571): p. 75-81.

47. Ferguson-Smith, M.A., *History and evolution of cytogenetics.* Mol Cytogenet, 2015. **8**: p. 19.

48. Bauman, J.G., et al., *A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA.* Exp Cell Res, 1980. **128**(2): p. 485-90.

49. Kallioniemi, A., et al., *Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.* Science, 1992. **258**(5083): p. 818-21.

50. Pirooznia, M., F.S. Goes, and P.P. Zandi, *Whole-genome CNV analysis: advances in computational approaches.* Frontiers in Genetics, 2015. **6**(138).

51. Mahmoud, M., et al., *Structural variant calling: the long and the short of it.* Genome Biology, 2019. **20**(1): p. 246.

52. Payne, A., et al., *BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files.* Bioinformatics, 2018. **35**(13): p. 2193-2198.

53. Huckaby, A.C., et al., *Complex DNA structures trigger copy number variation across the Plasmodium falciparum genome.* Nucleic Acids Research, 2018. **47**(4): p. 1615-1627.

54. Carter, R. and K.N. Mendis, *Evolutionary and Historical Aspects of the Burden of Malaria.* Clinical Microbiology Reviews, 2002. **15**(4): p. 564-594.

55. Corey, V.C., et al., *A broad analysis of resistance development in the malaria parasite.* Nature Communications, 2016. **7**: p. 11901.

56. Foote, S.J., et al., *Amplification of the multidrug resistance gene in some chloroquine-resistant isolates of P. falciparum.* Cell, 1989. **57**(6): p. 921-930.

57. Heinberg, A., et al., *Direct evidence for the adaptive role of copy number variation on antifolate susceptibility in Plasmodium falciparum.* Molecular microbiology, 2013. **88**(4): p. 702-712.

58. Lynch, M. and J.S. Conery, *The Evolutionary Fate and Consequences of Duplicate Genes.* Science, 2000. **290**(5494): p. 1151-1155.

59. Kondrashov, F.A., et al., *Selection in the evolution of gene duplications.* Genome Biol, 2002. **3**(2): p. Research0008.

60. Kondrashov, F.A., *Gene duplication as a mechanism of genomic adaptation to a changing environment.* Proceedings of the Royal Society B: Biological Sciences, 2012.

61. Guler, J.L., et al., *Asexual populations of the human malaria parasite, Plasmodium falciparum, use a two-step genomic strategy to acquire accurate, beneficial DNA amplifications.* PLoS Pathog, 2013. **9**(5): p. e1003375.

62. Rottmann, M., et al., *Spiroindolones, a potent compound class for the treatment of malaria.* Science, 2010. **329**(5996): p. 1175-80.

63. Phillips, M.A., et al., *A long-duration dihydroorotate dehydrogenase inhibitor (DSM265) for prevention and treatment of malaria.* Science Translational Medicine, 2015. **7**(296): p. 296ra111-296ra111.

64. Thaithong, S., et al., *Plasmodium falciparum: gene mutations and amplification of dihydrofolate reductase genes in parasites grown in vitro in presence of pyrimethamine.* Exp Parasitol, 2001. **98**(2): p. 59-70.

65. Triglia, T., et al., *Amplification of the multidrug resistance gene pfmdr1 in Plasmodium falciparum has arisen as multiple independent events.* Molecular and Cellular Biology, 1991. **11**(10): p. 5244-5250.

66. Cheeseman, I.H., et al., *Gene copy number variation throughout the Plasmodium falciparum genome.* BMC Genomics, 2009. **10**: p. 353-353.

67. Kidgell, C., et al., *A systematic map of genetic variation in Plasmodium falciparum.* PLoS Pathog, 2006. **2**(6): p. e57.

68. Bopp, S.E.R., et al., *Mitotic Evolution of Plasmodium falciparum Shows a Stable Core Genome but Recombination in Antigen Families.* PLoS Genet, 2013. **9**(2): p. e1003293.

69. Nair, S., et al., *Adaptive copy number evolution in malaria parasites.* PLoS Genet, 2008. **4**(10): p. e1000243.

70. Ribacke, U., et al., *Genome wide gene amplifications and deletions in Plasmodium falciparum.* Molecular and Biochemical Parasitology, 2007. **155**(1): p. 33-44.

71. Nair, S., et al., *GENETIC CHANGES DURING LABORATORY PROPAGATION: COPY NUMBER AT THE RETICULOCYTE BINDING PROTEIN 1 LOCUS OF PLASMODIUM FALCIPARUM.* Molecular and biochemical parasitology, 2010. **172**(2): p. 145-148.

72. Banyal, H.S. and J. Inselburg, *Plasmodium falciparum: induction, selection, and characterization of pyrimethamine-resistant mutants.* Exp Parasitol, 1986. **62**(1): p. 61-70.

73. Cowman, A.F., D. Galatis, and J.K. Thompson, *Selection for mefloquine resistance in Plasmodium falciparum is linked to amplification of the pfmdr1 gene and cross-resistance to halofantrine and quinine.* Proceedings of the National Academy of Sciences of the United States of America, 1994. **91**(3): p. 1143-1147.

74. Crabb, B.S. and A.F. Cowman, *Characterization of promoters and stable transfection by homologous and nonhomologous recombination in Plasmodium falciparum.* Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(14): p. 7289-7294.

75. Price, R.N., et al., *Mefloquine resistance in Plasmodium falciparum and increased pfmdr1 gene copy number.* Lancet, 2004. **364**(9432): p. 438-447.

76. Dharia, N.V., et al., *Use of high-density tiling microarrays to identify mutations globally and elucidate mechanisms of drug resistance in Plasmodium falciparum.* Genome Biol, 2009. **10**(2): p. R21.

77. Singh, A. and P.J. Rosenthal, *Selection of cysteine protease inhibitor-resistant malaria parasites is accompanied by amplification of falcipain genes and alteration in inhibitor transport.* J Biol Chem, 2004. **279**(34): p. 35236-41.

78. Cheeseman, I.H., et al., *Population Structure Shapes Copy Number Variation in Malaria Parasites.* Molecular Biology and Evolution, 2015.

79. Auburn, S., et al., *Genomic Analysis Reveals a Common Breakpoint in Amplifications of the Plasmodium vivax Multidrug Resistance 1 Locus in Thailand.* The Journal of Infectious Diseases, 2016. **214**(8): p. 1235-1242.

80. Menard, D., et al., *Whole Genome Sequencing of Field Isolates Reveals a Common Duplication of the Duffy Binding Protein Gene in Malagasy Plasmodium vivax Strains.* PLoS Neglected Tropical Diseases, 2013. **7**(11): p. e2489.

81. Gunalan, K., et al., *Role of Plasmodium vivax Duffy-binding protein 1 in invasion of Duffy-null Africans.* Proc Natl Acad Sci U S A, 2016. **113**(22): p. 6271-6.

82. Samarakoon, U., et al., *The landscape of inherited and de novo copy number variants in a plasmodium falciparum genetic cross.* BMC Genomics, 2011. **12**: p. 457-457.

83. Nair, S., et al., *Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites.* Mol Biol Evol, 2007. **24**(2): p. 562-73.

84. Zhang, H. and C.H. Freudenreich, *An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in S. cerevisiae.* Mol Cell, 2007. **27**(3): p. 367-79.

85. Shah, S.N., et al., *DNA structure and the Werner protein modulate human DNA polymerase delta-dependent replication dynamics within the common fragile site FRA16D.* Nucleic Acids Res, 2010. **38**(4): p. 1149-62.

86. Burrow, A.A., et al., *Secondary structure formation and DNA instability at fragile site FRA16B.* Nucleic Acids Res, 2010. **38**(9): p. 2865-77.

87. Walsh, E., et al., *Mechanism of Replicative DNA Polymerase Delta Pausing and a Potential Role for DNA Polymerase Kappa in Common Fragile Site Replication.* Journal of molecular biology, 2013. **425**(2): p. 232-243.

88. Zheng, G.X., et al., *Torsionally tuned cruciform and Z-DNA probes for measuring unrestrained supercoiling at specific sites in DNA of living cells.* J Mol Biol, 1991. **221**(1): p. 107-22.

89. Cromie, G.A., et al., *Palindromes as substrates for multiple pathways of recombination in Escherichia coli.* Genetics, 2000. **154**(2): p. 513-522.

90. Rogers, F.A. and M.K. Tiwari, *Triplex-induced DNA damage response.* Yale J Biol Med, 2013. **86**(4): p. 471-8.

91. van Kregten, M. and M. Tijsterman, *The repair of G-quadruplex-induced DNA damage.* Exp Cell Res, 2014. **329**(1): p. 178-83.

92. Mirkin, E.V. and S.M. Mirkin, *Replication Fork Stalling at Natural Impediments.* Microbiology and Molecular Biology Reviews : MMBR, 2007. **71**(1): p. 13-35.

93. Stanton, A., et al., *Recombination events among virulence genes in malaria parasites are associated with G-quadruplex-forming DNA motifs.* BMC Genomics, 2016. **17**(1): p. 859.

94. Herman, J.D., et al., *A genomic and evolutionary approach reveals non-genetic drug resistance in malaria.* Genome Biology, 2014. **15**(11): p. 511.

95. Andrews, S. *FastQC: a quality control tool for high throughput sequence data*. 2010; Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

96. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

97. Okonechnikov, K., A. Conesa, and F. García-Alcalde, *Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data.* Bioinformatics, 2016. **32**(2): p. 292-294.

98. Chiang, C., et al., *SpeedSeq: ultra-fast personal genome analysis and interpretation.* Nat Methods, 2015. **12**(10): p. 966-8.

99. Abyzov, A., et al., *CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing.* Genome Res, 2011. **21**(6): p. 974-84.

100. Layer, R.M., et al., *LUMPY: a probabilistic framework for structural variant discovery.* Genome Biol, 2014. **15**(6): p. R84.

101. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Brief Bioinform, 2013. **14**(2): p. 178-92.

102. Dillon, L.W., et al., *Role of DNA secondary structures in fragile site breakage along human chromosome 10.* Human Molecular Genetics, 2013. **22**(7): p. 1443-1456.

103. Sander, A.F., et al., *DNA secondary structures are associated with recombination in major Plasmodium falciparum variable surface antigen gene families.* Nucleic Acids Research, 2014. **42**(4): p. 2270-2281.

104. Balakrishnan, L. and R.A. Bambara, *Okazaki fragment metabolism.* Cold Spring Harb Perspect Biol, 2013. **5**(2).

105. Gruber, A.R., et al., *The Vienna RNA Websuite.* Nucleic Acids Research, 2008. **36**(Web Server issue): p. W70-W74.

106. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: the European Molecular Biology Open Software Suite.* Trends Genet, 2000. **16**(6): p. 276-7.

107. Thys, R.G., et al., *DNA Secondary Structure at Chromosomal Fragile Sites in Human Disease.* Current Genomics, 2015. **16**(1): p. 60-70.

108. Mayer, C., *Phobos Repeat Finder*. 2006-2010.

109. Dechering, K.J., et al., *Distinct frequency-distributions of homopolymeric DNA tracts in different genomes.* Nucleic Acids Research, 1998. **26**(17): p. 4056-4062.

110. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.

111. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2009: Springer Publishing Company, Incorporated. 216.

112. R Development Core team, *R: A language and environment for statistical computing*. 2016, R Foundation for Statistical Computing: Vienna, Austria.

113. Manary, M.J., et al., *Identification of pathogen genomic variants through an integrated pipeline.* BMC Bioinformatics, 2014. **15**: p. 63.

114. Brown, T., et al., *Molecular surveillance for drug-resistant Plasmodium falciparum in clinical and subclinical populations from three border regions of Burma/Myanmar: cross-sectional data and a systematic review of resistance studies.* Malar J, 2012. **11**: p. 333.

115. Lee, A.H. and D.A. Fidock, *Evidence of a Mild Mutator Phenotype in Cambodian Plasmodium falciparum Malaria Parasites.* PLoS ONE, 2016. **11**(4): p. e0154166.

116. Yu, P., et al., *Genome-wide copy number variations in Oryza sativa L.* BMC Genomics, 2013. **14**: p. 649.

117. Guryev, V., et al., *Distribution and functional impact of DNA copy number variation in the rat.* Nat Genet, 2008. **40**(5): p. 538-45.

118. Fadista, J., et al., *Copy number variation in the bovine genome.* BMC Genomics, 2010. **11**: p. 284.

119. Zarrei, M., et al., *A copy number variation map of the human genome.* Nat Rev Genet, 2015. **16**(3): p. 172-83.

120. Nicholas, T.J., et al., *A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog.* BMC Genomics, 2011. **12**: p. 414.

121. Locke, M.E., et al., *Genomic copy number variation in Mus musculus.* BMC Genomics, 2015. **16**: p. 497.

122. Anderson, T.J., J. Patel, and M.T. Ferdig, *Gene copy number and malaria biology.* Trends Parasitol, 2009. **25**(7): p. 336-43.

123. Hendrickson, H., et al., *Amplification-mutagenesis: evidence that "directed" adaptive mutation and general hypermutability result from growth with a selected gene amplification.* Proc Natl Acad Sci U S A, 2002. **99**(4): p. 2164-9.

124. Roth, J.R. and D.I. Andersson, *Amplification-mutagenesis--how growth under selection contributes to the origin of genetic diversity and explains the phenomenon of adaptive mutation.* Res Microbiol, 2004. **155**(5): p. 342-51.

125. Elde, N.C., et al., *Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses.* Cell, 2012. **150**(4): p. 831-41.

126. Harris, L.M., et al., *G-Quadruplex DNA Motifs in the Malaria Parasite Plasmodium falciparum and Their Potential as Novel Antimalarial Drug Targets.* Antimicrob Agents Chemother, 2018. **62**(3).

127. Ma, Y., et al., *Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination.* Cell, 2002. **108**(6): p. 781-94.

128. Chiruvella, K.K., et al., *Biochemical Characterization of Kat1: a Domesticated hAT-Transposase that Induces DNA Hairpin Formation and MAT-Switching.* Sci Rep, 2016. **6**: p. 21671.

129. Brázda, V., et al., *Cruciform structures are a common DNA feature important for regulating biological processes.* BMC Molecular Biology, 2011. **12**: p. 33-33.

130. Carvalho, C.M., et al., *Replicative mechanisms for CNV formation are error prone.* Nat Genet, 2013. **45**(11): p. 1319-26.

131. Rathod, P.K., T. McErlean, and P.-C. Lee, *Variations in frequencies of drug resistance in Plasmodium falciparum.* Proceedings of the National Academy of Sciences, 1997. **94**(17): p. 9389-9393.

132. Gupta, D.K., et al., *DNA damage regulation and its role in drug-related phenotypes in the malaria parasites.* Scientific Reports, 2016. **6**: p. 23603.

133. Ottaviani, D., M. LeCain, and D. Sheer, *The role of microhomology in genomic structural variation.* Trends in Genetics, 2014. **30**(3): p. 85-94.

134. Kirkman, L.A., E.A. Lawrence, and K.W. Deitsch, *Malaria parasites utilize both homologous recombination and alternative end joining pathways to maintain genome integrity.* Nucleic Acids Research, 2014. **42**(1): p. 370-379.

135. Lee, A.H., L.S. Symington, and D.A. Fidock, *DNA Repair Mechanisms and Their Biological Roles in the Malaria Parasite Plasmodium falciparum.* Microbiology and Molecular Biology Reviews, 2014. **78**(3): p. 469-486.

136. Hastings, P.J., G. Ira, and J.R. Lupski, *A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation.* PLoS Genet, 2009. **5**(1): p. e1000327.

137. Zhang, F., et al., *The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans.* Nature genetics, 2009. **41**(7): p. 849-853.

138. Verdin, H., et al., *Microhomology-Mediated Mechanisms Underlie Non-Recurrent Disease-Causing Microdeletions of the FOXL2 Gene or Its Regulatory Domain.* PLOS Genetics, 2013. **9**(3): p. e1003358.

139. Symington, L.S. and J. Gautier, *Double-Strand Break End Resection and Repair Pathway Choice.* Annual Review of Genetics, 2011. **45**(1): p. 247-271.

140. Mimitou, E. and S. LS, *DNA end resection: many nucleases make light work.* DNA Repair (Amst.), 2009. **8**: p. 983.

141. Chung, W.-H., et al., *Defective Resection at DNA Double-Strand Breaks Leads to De Novo Telomere Formation and Enhances Gene Targeting.* PLoS Genetics, 2010. **6**(5): p. e1000948.

142. Galhardo, R.S., P.J. Hastings, and S.M. Rosenberg, *Mutation as a Stress Response and the Regulation of Evolvability.* Critical Reviews in Biochemistry and Molecular Biology, 2007. **42**(5): p. 399-435.

143. Scanlon, S.E. and P.M. Glazer, *Multifaceted control of DNA repair pathways by the hypoxic tumor microenvironment.* DNA Repair (Amst), 2015. **32**: p. 180-9.

144. Arlt, M.F., et al., *Replication Stress Induces Genome-wide Copy Number Changes in Human Cells that Resemble Polymorphic and Pathogenic Variants.* American Journal of Human Genetics, 2009. **84**(3): p. 339-350.

145. Bindra, R.S., et al., *Down-regulation of Rad51 and decreased homologous recombination in hypoxic cancer cells.* Mol Cell Biol, 2004. **24**(19): p. 8504-18.

146. Bristow, R.G. and R.P. Hill, *Hypoxia and metabolism. Hypoxia, DNA repair and genetic instability.* Nat Rev Cancer, 2008. **8**(3): p. 180-92.

147. Slack, A., et al., *On the mechanism of gene amplification induced under stress in Escherichia coli.* PLoS Genet, 2006. **2**(4): p. e48.

148. Mannava, S., et al., *Depletion of Deoxyribonucleotide Pools Is an Endogenous Source of DNA Damage in Cells Undergoing Oncogene-Induced Senescence.* The American Journal of Pathology, 2013. **182**(1): p. 142-151.

149. Bhattacharya, S., et al., *RAD51 interconnects between DNA replication, DNA repair and immunity.* Nucleic Acids Research, 2017. **45**(8): p. 4590-4605.

150. Dankwa, S., et al., *Ancient human sialic acid variant restricts an emerging zoonotic malaria parasite.* Nat Commun, 2016. **7**: p. 11187.

151. Magadum, S., et al., *Gene duplication as a major force in evolution.* J Genet, 2013. **92**(1): p. 155-61.

152. Andersson, D.I., J. Jerlström-Hultqvist, and J. Näsvall, *Evolution of new functions de novo and from preexisting genes.* Cold Spring Harbor perspectives in biology, 2015. **7**(6): p. a017996.

153. Frech, C. and N. Chen, *Genome Comparison of Human and Non-Human Malaria Parasites Reveals Species Subset-Specific Genes Potentially Linked to Human Disease.* PLOS Computational Biology, 2011. **7**(12): p. e1002320.

154. Carlton, J.M., et al., *Comparative genomics of the neglected human malaria parasite Plasmodium vivax.* Nature, 2008. **455**(7214): p. 757-763.

155. Liu, X., et al., *In-depth comparative analysis of malaria parasite genomes reveals protein-coding genes linked to human disease in Plasmodium falciparum genome.* BMC Genomics, 2018. **19**(1): p. 312.

156. Diez Benavente, E., et al., *Genomic variation in Plasmodium vivax malaria reveals regions under selective pressure.* PloS one, 2017. **12**(5): p. e0177134-e0177134.

157. Jeffares, D.C., et al., *Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast.* Nature communications, 2017. **8**: p. 14061-14061.

158. Kearse, M., et al., *Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.* Bioinformatics, 2012. **28**(12): p. 1647-9.

159. Darling, A.E., B. Mau, and N.T. Perna, *progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement.* PLOS ONE, 2010. **5**(6): p. e11147.

160. Gardner, M.J., et al., *Genome sequence of the human malaria parasite Plasmodium falciparum.* Nature, 2002. **419**(6906): p. 10.1038/nature01097.

161. Auburn, S., et al., *A new Plasmodium vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes.* Wellcome open research, 2016. **1**: p. 4-4.

162. Pain, A., et al., *The genome of the simian and human malaria parasite Plasmodium knowlesi.* Nature, 2008. **455**(7214): p. 799-803.

163. Gupta, A., G. Thiruvengadam, and S.A. Desai, *The conserved clag multigene family of malaria parasites: essential roles in host-pathogen interaction.* Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy, 2015. **18**: p. 47-54.

164. Hupalo, D.N., et al., *Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax.* Nat Genet, 2016. **48**(8): p. 953-8.

165. Liu, S., et al., *Single cell sequencing of the small and AT-skewed genome of malaria parasites.* bioRxiv, 2020: p. 2020.02.21.960039.

166. Meddeb, R., et al., *Quantifying circulating cell-free DNA in humans.* Sci Rep, 2019. **9**(1): p. 5220.

167. Lauer, S., et al., *Single-cell copy number variant detection reveals the dynamics and diversity of adaptation.* PLoS Biol, 2018. **16**(12): p. e3000069.

168. Wang, R., D.-Y. Lin, and Y. Jiang, *SCOPE: a normalization and copy number estimation method for single-cell DNA sequencing.* bioRxiv, 2019: p. 594267.

169. Gawad, C., W. Koh, and S.R. Quake, *Single-cell genome sequencing: current state of the science.* Nat Rev Genet, 2016. **17**(3): p. 175-88.

170. Hodzic, E., *Single-cell analysis: Advances and future perspectives.* Bosn J Basic Med Sci, 2016. **16**(4): p. 313-314.

171. Wang, Y. and N.E. Navin, *Advances and applications of single-cell sequencing technologies.* Mol Cell, 2015. **58**(4): p. 598-609.

172. Chronister, W.D., et al., *Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex.* Cell Rep, 2019. **26**(4): p. 825-835.e7.

173. Estévez-Gómez, N., et al., *Comparison of single-cell whole-genome amplification strategies.* bioRxiv, 2018: p. 443754.

174. Hou, Y., et al., *Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing.* Gigascience, 2015. **4**: p. 37.

175. Huang, L., et al., *Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications.* Annu Rev Genomics Hum Genet, 2015. **16**: p. 79-102.

176. Deleye, L., et al., *Performance of four modern whole genome amplification methods for copy number variant detection in single cells.* Sci Rep, 2017. **7**(1): p. 3422.

177. Duan, M., et al., *Diverse modes of clonal evolution in HBV-related hepatocellular carcinoma revealed by single-cell genome sequencing.* Cell Res, 2018. **28**(3): p. 359-373.

178. Hughes, A.E., et al., *Clonal architecture of secondary acute myeloid leukemia defined by single-cell sequencing.* PLoS Genet, 2014. **10**(7): p. e1004462.

179. Neves, R.P., et al., *Genomic high-resolution profiling of single CKpos/CD45neg flow-sorting purified circulating tumor cells from patients with metastatic breast cancer.* Clin Chem, 2014. **60**(10): p. 1290-7.

180. Paolillo, C., et al., *Detection of Activating Estrogen Receptor Gene (ESR1) Mutations in Single Circulating Tumor Cells.* Clin Cancer Res, 2017. **23**(20): p. 6086-6093.

181. Rohrback, S., et al., *Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing.* Proc Natl Acad Sci U S A, 2018. **115**(42): p. 10804-10809.

182. Burbulis, I.E., et al., *Improved molecular karyotyping in glioblastoma.* Mutat Res, 2018. **811**: p. 16-26.

183. Zahn, H., et al., *Scalable whole-genome single-cell library preparation without preamplification.* Nat Methods, 2017. **14**(2): p. 167-173.

184. Lasken, R.S. and T.B. Stockwell, *Mechanism of chimera formation during the Multiple Displacement Amplification reaction.* BMC Biotechnol, 2007. **7**: p. 19.

185. Zong, C., et al., *Genome-wide detection of single-nucleotide and copy-number variations of a single human cell.* Science, 2012. **338**(6114): p. 1622-6.

186. Oyola, S.O., et al., *Optimized whole-genome amplification strategy for extremely AT-biased template.* DNA Res, 2014. **21**(6): p. 661-71.

187. de Bourcy, C.F., et al., *A quantitative comparison of single-cell whole genome amplification methods.* PLoS One, 2014. **9**(8): p. e105585.

188. Ning, L., et al., *Quantitative assessment of single-cell whole genome amplification methods for detecting copy number variation using hippocampal neurons.* Sci Rep, 2015. **5**: p. 11415.

189. Bushnell, B., *BBMap short read aligner, and other bioinformatic tools.* 2020.

190. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

191. Chen, C., et al., *Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI).* Science, 2017. **356**(6334): p. 189-194.

192. Marwick, B.a.K.K., *cvequality: Tests for the Equality of Coefficients of Variation from Multiple Groups.* 2019. **R software package version 0.2.0**.

193. Derrien, T., et al., *Fast Computation and Applications of Genome Mappability.* PLOS ONE, 2012. **7**(1): p. e30377.

194. Venkatraman, E. and A. Olshen, *DNAcopy: A Package for analyzing DNA copy data.* 2010.

195. Wang, X., H. Chen, and N.R. Zhang, *DNA copy number profiling using single-cell sequencing.* Briefings in bioinformatics, 2018. **19**(5): p. 731-736.

196. Dillon, L.W., et al., *Production of Extrachromosomal MicroDNAs Is Linked to Mismatch Repair Pathways and Transcriptional Activity.* Cell Rep, 2015. **11**(11): p. 1749-59.

197. Albertson, D.G., *Gene amplification in cancer.* Trends Genet, 2006. **22**(8): p. 447-55.

198. McGill, J.R., et al., *Double minutes are frequently found in ovarian carcinomas.* Cancer Genet Cytogenet, 1993. **71**(2): p. 125-31.

199. Moller, H.D., et al., *Extrachromosomal circular DNA is common in yeast.* Proc Natl Acad Sci U S A, 2015. **112**(24): p. E3114-22.

200. Beverley, S.M., et al., *Unstable DNA amplifications in methotrexate-resistant Leishmania consist of extrachromosomal circles which relocalize during stabilization.* Cell, 1984. **38**(2): p. 431-9.

201. Wagner, W. and M. So, *Identification of a novel large extrachromosomal DNA (LED) in the Trypanosomatidae.* Mol Microbiol, 1992. **6**(16): p. 2299-308.

202. Wu, S., et al., *Circular ecDNA promotes accessible chromatin and high oncogene expression.* Nature, 2019. **575**(7784): p. 699-703.

203. Verhaak, R.G.W., V. Bafna, and P.S. Mischel, *Extrachromosomal oncogene amplification in tumour pathogenesis and evolution.* Nat Rev Cancer, 2019. **19**(5): p. 283-288.

204. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.

205. Zielezinski, A., et al., *Benchmarking of alignment-free sequence comparison methods.* Genome Biology, 2019. **20**(1): p. 144.

206. Wengelnik, K., W. Daher, and M. Lebrun, *Phosphoinositides and their functions in apicomplexan parasites.* Int J Parasitol, 2018. **48**(7): p. 493-504.

207. Agarwal, M., et al., *Identification and characterization of ARS-like sequences as putative origin(s) of replication in human malaria parasite Plasmodium falciparum.* Febs j, 2017. **284**(16): p. 2674-2695.

208. Matthews, H., C.W. Duffy, and C.J. Merrick, *Checks and balances? DNA replication and the cell cycle in Plasmodium.* Parasit Vectors, 2018. **11**(1): p. 216.

209. Iwanaga, S., et al., *Functional identification of the Plasmodium centromere and generation of a Plasmodium artificial chromosome.* Cell Host Microbe, 2010. **7**(3): p. 245-55.

210. Singh, D., S. Chaubey, and S. Habib, *Replication of the Plasmodium falciparum apicoplast DNA initiates within the inverted repeat region.* Mol Biochem Parasitol, 2003. **126**(1): p. 9-14.

211. Verma, G. and N. Surolia, *Centromere and its associated proteins-what we know about them in Plasmodium falciparum.* IUBMB Life, 2018. **70**(8): p. 732-742.

212. Davies, H.M., et al., *Repetitive sequences in malaria parasite proteins.* FEMS Microbiol Rev, 2017. **41**(6): p. 923-940.

213. Kwiatkowski, D.P., *How malaria has affected the human genome and what human genetics can teach us about malaria.* Am J Hum Genet, 2005. **77**(2): p. 171-92.

214. Sakofsky, Cynthia J., et al., *Translesion Polymerases Drive Microhomology-Mediated Break-Induced Replication Leading to Complex Chromosomal Rearrangements.* Molecular Cell, 2015. **60**(6): p. 860-872.

215. Siao, M.C., et al., *Evolution of Host Specificity by Malaria Parasites through Altered Mechanisms Controlling Genome Maintenance.* mBio, 2020. **11**(2).

216. Arlt, M.F., et al., *De novo CNV formation in mouse embryonic stem cells occurs in the absence of Xrcc4-dependent nonhomologous end joining.* PLoS Genet, 2012. **8**(9): p. e1002981.

217. Otto, T.D., et al., *Long read assemblies of geographically dispersed Plasmodium falciparum isolates reveal highly structured subtelomeres.* Wellcome open research, 2018. **3**: p. 52-52.

218. Becker, T., et al., *FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods.* Genome Biology, 2018. **19**(1): p. 38.

219. English, A.C., et al., *Assessing structural variation in a personal genome-towards a human reference diploid genome.* BMC genomics, 2015. **16**(1): p. 286-286.

220. Mohiyuddin, M., et al., *MetaSV: an accurate and integrative structural-variant caller for next generation sequencing.* Bioinformatics (Oxford, England), 2015. **31**(16): p. 2741-2744.