Detection of heterogeneous copy number variation reveals the dynamics of adaptation in malaria parasites

Shiwei Liu
Sichuan, China

M.S., Shanghai Jiao Tong University, Shanghai, China, 2017
B.S., Shanghai Jiao Tong University, Shanghai, China, 2014

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
Dec 2022

Acknowledgement

First and foremost, I am extremely grateful to my supervisor, Prof. Jennifer L. Guler for her invaluable advice, continuous support, and patience during my PhD study. Within the past 5.5 years, she provided me the best environment and resources to learn to be a good researcher. Her immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. She mentored me with her patience and good heart even when I was frustrated and struggled. The greatest lesson I learned from Jenny is to never give up and her advice will always guide me in my future career and life.

I would also like to thank all the members of my dissertation committee, including Dr. Martin Wu, Dr. Aakrosh Ratan, Dr. Eyleen J. O'Rourke, Dr. John N. Campbell, and Dr. Michael J. McConnell for their support on my study. I would like to thank Dr. Martin Wu. As my first reader, he provided valuable advice on my research and support for Nanopore sequencing projects. I also want to thank Dr. Michael J. McConnell, especially for all the guidance and advice on the single cell sequencing project. I also would like to express my special appreciation to Dr. John N. Campbell for his help with the lab equipment and reagents. I also appreciate Eyleen J. O'Rourke always provides detailed and valuable advice for my research. To sum up, I appreciate having such a great committee to guide me through my graduate school.

Next, I would like to express my gratitude to everyone from the laboratory of Dr. William Petri Jr for their helpful discussions and insight for my research. Their opinions are influential in shaping my experiment methods and critiquing my results.

I would like to express my thanks to all the undergraduate students I have worked with, including Jane Kim, Nnenna Ene, Mary Lewis Simpson. Without their help and dedication to the research projects, the success of the experiments would not have been possible.

I also want to express my appreciation to all my colleagues in the laboratory. Without their help and coordination, it would not have been possible to make such progress. Particularly, I would like to express my thanks to Michelle Warthan for cell culturing and always taking care of the lab members. I want to thank Adam Huckaby, Jennifer Mcdaniels and Audrey Brown for teaching me lab techniques and providing valuable advice on research. I would like to extend my sincere thanks to Noah Brown, Anush Aryal and Aleksander Luniewski for their help and support on the research.

Lastly, I would like to thank my friends, family for encouragement and support all through my studies. It is their kind help and support that have made my study and life in the UVA a wonderful time. Without their tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

Table of Content

List of Figures

List of Supplementary Figures

List of Tables

List of Supplementary Tables (see Supplement Tables in additional excel file)

1 CHAPTER I: Introduction

Malaria is an infectious disease caused by unicellular parasitic protozoans in the genus *Plasmodium* (Rich et al., 2009). *Plasmodium falciparum* is the deadliest species of *Plasmodium* causing human infection (Rich et al., 2009). In the past two decades of the global fight against malaria through the widespread distribution of bed nets, drugs, and insecticides, there are historic reductions in malaria incidence and mortality since the year 2000 (Cohen et al., 2022). However, the progress of malaria elimination has stalled in recent years (World Health Organization, 2021). Malaria continues to be a global health problem, infecting 241 million people and causing 627,000 deaths in 2020 globally (World Health Organization, 2021). One of the main factors contributing to the stalled progress is the adaptive nature of the *Plasmodium* parasite, which has led to increasing resistance to antimalarial drugs. Resistance to the frontline treatment, artemisinin-based combination therapies (ACTs), has been widespread throughout Southeast Asia and potential signs of resistance have been reported in a few countries in Africa (Woodrow and White, 2017; Uwimana et al., 2020; Balikagala et al., 2021). Thus, it is critical to understand the molecular mechanism of adaptation in *Plasmodium falciparum* parasites to develop new interventions to alleviate the global health burden caused by malaria.

1.1 Understanding the adaptive nature of malaria parasite

Malaria is a life-threatening disease caused by the protozoan *Plasmodium* parasite. Infection with *P. falciparum*, the most fatal human malaria parasite, caused 627,000 deaths in 2020 (World Health Organization, 2021). The clinical symptoms of malaria occur during the erythrocytic cycle; parasites invade human red blood cells (RBCs), go through multiple rounds of asexual reproduction, mature from early stage to late stage, and finally burst out from the RBC to begin the cycle again. During the early stage, parasites possess a single haploid genome but during the late stage, they possess an average of 16 copies of their haploid genome. Antimalarial drugs often target this blood stage of the parasite life cycle. Due to a lack of an effective vaccine, antimalarials are critical weapons against malaria. However, their efficacy is threatened by the frequent emergence and spread of antimalarial resistance in *P. falciparum*. Today, resistance has been reported for all known clinically used antimalarials. Even though antimalarial resistance could arise independently in other regions (e.g., South America), most antimalarial resistance emerges in Southeast Asia (Blasco et al., 2017). Artemisinin resistance, the current frontline antimalarial, was first reported in Western Cambodia; since then, resistant parasites have spread across Southeast Asia and recently have been detected in South China and India (Woodrow and White, 2017; Uwimana et al., 2020; Balikagala et al., 2021). It is necessary to understand how resistance arises and the acquirement of genetic contributors to devise approaches for resistance prevention.

1.2 Copy number variations in malaria parasite and its significance in adaptation

Importance of copy number variations in malaria parasite

Extensive genetic diversity is known to contribute to antimalarial resistance. Sources of genomic variation in *P. falciparum* include single nucleotide polymorphisms and structural variations. Single nucleotide polymorphisms (SNPs) that disrupt drug binding or facilitate drug removal are implicated in many antimalarial resistance examples (Rosenthal, 2013). Structural variations are the result of chromosomal rearrangements, including inversions and balanced translocations or genomic imbalances (insertions, deletions, amplification). Copy number variations (CNVs), a type of DNA structural variations involving changes in copy number of a genomic segment, have also been implicated in *P. falciparum* resistance (Guler et al., 2013). In general, the fraction of *P. falciparum*'s genome subject to CNVs is greater than the fraction represented by SNPs (Miles et al., 2016). Amplifications, or increases in the copy number of a genomic segment, that can increase the expression of target proteins or drug efflux pumps by altering gene dosage are known to be associated with antimalarial resistance in both laboratory and field parasites (Simam et al., 2018a). For example, amplifications of *pfmdr1* and *pfgch1* genes drive the resistance to mefloquine and antifolates drugs, respectively (Simam et al., 2018a). Gene amplifications and codon mutations are both random events, but codon mutations are much less frequent in *P. falciparum* parasites (Preechapornkul et al., 2009). For example, the amplification of *pfmdr*1 has been suggested to be a frequent event, and the rate of occurrence of *pfmdr*1 duplication (1: 10^8) is much higher than that of point mutations within codons, which occurs at a much lower frequency (1:10^{14}) (White et al., 2003).

Formation of copy number variations in *Plasmodium* parasite

Even though most spontaneous mutations occur randomly, it is well accepted that the mutation rate of copy number variations, can increase or decrease under certain stress in human cells, yeast, and bacteria (Arlt et al., 2009; Kondrashov, 2012; Hull et al., 2017). *P. falciparum*, an organism with the highest genome AT content compared to all other *Plasmodium* species and eukaryotes (Hamilton et al., 2017), may also use this strategy to quickly adapt to highly variable environments (e.g., distinct antimalarials, mosquito, and human host environments). CNVs are generated during the repair of double-strand break (DSB) DNA damage. DNA damage originates from a range of sources, like reactive oxygen species generated by metabolism, free radicals which are often produced after uptake of antimalarial drugs such as chloroquine or artemisinin, and DNA replication errors (Matthews et al., 2018). Despite the destructive potential of DSBs, their role in the parasite can also be beneficial. The fidelity of DSB repair can be lax enough to allow for sufficient genetic variation. The high-fidelity homologous recombination (HR) and error-prone non-homologous end joining (NHEJ) are the two major repair pathways for DSB repair in eukaryotes (San Filippo et al., 2008; Lieber, 2010). The HR pathway is encoded in the *P. falciparum* genome but surprisingly none of the components of the canonical non-homologous end joining (C-NHEJ) pathway were identified in *P. falciparum* (Gardner et al., 2002). Alternative error-prone microhomology-mediated end joining (MMEJ) has been demonstrated to occur in *P. falciparum* (Kirkman et al., 2014). Even though microhomology mediated break induced replication (MMBIR, erroneous DNA replication) is not characterized in *P. falciparum*, there is evidence indicating the usage of MMBIR in *P. falciparum* (Huckaby et al., 2018). During the erythrocytic stage, *P. falciparum* parasites start with a haploid genome and thus must rely upon non-HR repair pathways (e.g., MMEJ and MMBIR) during early stages due to the lack of homologous sequences required for HR repair. The long adenine or thymine (A/T) tracks commonly found at the CNV junctions in *P. falciparum* could be used as microhomology for microhomology mediated repair pathways (Guler et al., 2013). A recent study by our lab showed that genome sequence features proximal to A/T tracks, such as DNA hairpins, likely trigger CNV events across the genome and provided some mechanistic insight into their formation (Huckaby et al., 2018). The microhomology (A/T-track) mediated repair pathways used for CNV generation may be uniquely matched with the high AT content of the *P. falciparum* genome, which averages 81% AT but can reach upwards of ~90% in introns and intergenic regions. Thus, we expect the genome of *P. falciparum* genome is prone to generate CNVs.

The evolution to antimalarial resistance through copy number variations

Given the presence of over 10^{11} parasites within infected patients, we think copy number variations, particularly amplifications of a genomic segment are likely an essential step during the parasite's adaptation to countless antimalarials (Guler et al., 2013). We hypothesize that de novo amplicons (new amplified regions) are being randomly generated in the haploid genome of *P. falciparum*, but antimalarial stress may also stimulate the formation of amplicons along the genome. As shown in **Figure 1.1**, if one of the amplicons contains a drug target gene, in the long term, point mutations can accumulate in one of the extra target gene copies, while having a gene copy with normal function can allow point mutations more likely to persist in the cell population and adapt to a changing environment (i.e. antimalarial drug treatment). Because extra copies of genes often carry strong fitness costs than point mutations due to increased cellular burden for DNA replication and alterations of metabolic flux due to differing levels of enzyme expression (Huckaby et al., 2018). Also, SNPs in drug binding pocket can completely abrogate drug action. Thus, genome amplifications may eventually be lost in favor of the beneficial SNPs (**Figure 1.1**). Based on our hypothesis, it is critical to understand how genome amplification events are formed. Of particular importance, how and when do these amplification events occur. Are de novo amplicons constantly being generated or are there certain conditions that induce their formation (i.e. sub-lethal stress from antimalarial exposure). Also, how is the genome amplification rate altered under stress? These questions are important for understanding the role of genome amplification in parasite's adaptation to antimalarials. Thus, it is important to detect these de novo amplification events that are only present in a low frequency in the parasite populations.

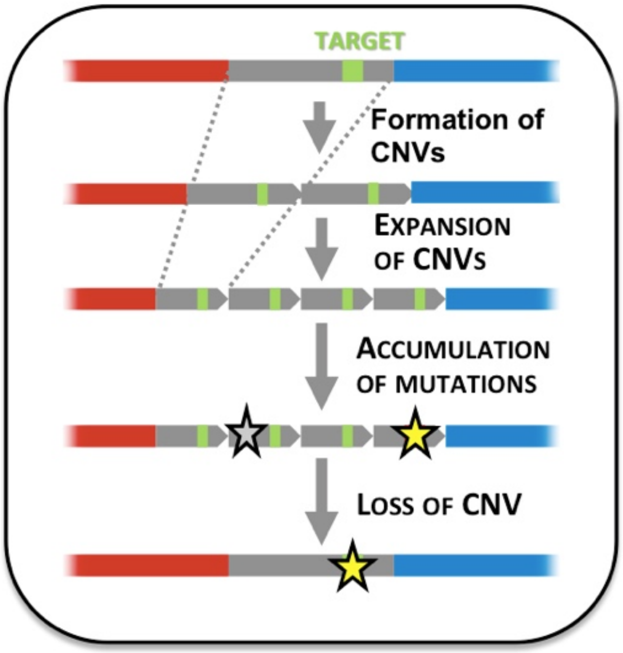


Figure 1.1 Model of CNV-facilitated resistance development in *Plasmodium falciparum*
New CNVs are generated randomly across the genome of *P. falciparum*. Under selection, beneficial CNVs can be further amplified. In the long term, SNPs may accumulate on extra gene copies and CNVs may be lost in favor of SNPs for higher level of resistance.

1.3 The necessity of developing new tools for detecting heterogeneous CNVs in malaria parasites

The limitation of traditional methods for CNV detection

Several traditional methods have been used for detecting copy number variations, including PCR-based assays (Quantitative PCR (qPCR), droplet digital PCR (ddPCR)), microarray-based hybridizations (aCGH), and whole genome sequencing techniques. Although PCR-based assays are simple and at low cost, they are limited to specific genes. While microarrays detect genetic variations with much higher numbers of probes, the high AT content and abundance of repeat elements in the parasite genome still limit proper detection in the noncoding regions. Whole genome sequencing of bulk DNA provides more complete genetic information across the genome (Cantsilieris et al., 2013). However, these methods are not capable of detecting rare copy number variation events unless they occur at a high frequency in a heterogeneous population (Lauer et al., 2018). To detect low-frequency genetic variations, previous studies cloned cells by limiting dilution from a cell population before sequencing, a method might allow the detection of some of the genetic variations, but it is labor intensive and time-consuming (>200 days of laboratory culture time) with limited through-put (Jett et al., 2020). In addition, during the long period of laboratory culture, mutations can continue to arise and disappear in each cell. Instead, single-cell analysis can be effective for addressing these issues and providing a more in-depth understanding of the genetic variations in cell populations. As it starts from only one cell, single-cell analysis could reveal information about individual cells to detect rare CNVs in *P. falciparum* parasites. However, next-generation sequencing typically requires 1ng to 1μg of genomic DNA, which corresponds to 10^5 to 10^8 *P. falciparum* genomes. Therefore, accurate and robust amplification of a single parasite genome (~25 femtograms) is needed prior to single cell sequencing.

The challenges for developing single parasite sequencing pipeline

There are many DNA amplification methods available including multiple displacement amplification (MDA)-based and non-MDA-based methods. Versions of MDA have been widely employed for the amplification of the *Plasmodium* parasite genome (Oyola et al., 2012; Nair et al., 2014; Sundararaman et al., 2016). In a MDA-based approach, *P. falciparum* specific primers designed based on common motifs in the genome were used (instead of random amplification primers) to amplify a minimum of ~20 copies of *P. falciparum* genome (~500 femtograms) and yield coverage of ~70% of the genome (Sundararaman et al., 2016). This method shows relatively good genome coverage but is not sensitive enough for single parasite genomes. Only one group has successfully amplified the genomes of isolated single parasites (Nair et al., 2014; Trevino et al., 2017). Their earlier work, which compared multiple amplification kits, determined analysis parameters, and obtained between 30-80% genome coverage (at ≥10×) for clinical and laboratory single *P. falciparum* parasites (Nair et al., 2014). In a later study from this group, additional protocol modifications achieved near-complete genome coverage (~90%) for both laboratory and clinical samples (Trevino et al., 2017; Nkhoma et al., 2018). So far, their analysis has been focused on the identification of SNPs, presumably because the chimeric reads generated in MDA lead to severe disruptions in CNV detection (Hou et al., 2015). In addition, the amplification bias generated by MDA is known to be sequence-dependent and not reproducible along the genome from cell to cell, resulting in noisy CNV measurements and ineffective normalization steps (Huang et al., 2015). While a non-MDA-based method called multiple annealing and looping-based amplification cycling (MALBAC) have been also used for single cell sequencing. MALBAC amplification consists of a linear pre-amplification step and an exponential amplification step. This method has the unique feature of quasi-linear amplification, which reduces the sequence-dependent bias exacerbated by exponential amplification. However, MALBAC is not free from sequence-dependent bias, it is reported to have a slight amplification bias toward GC-rich regions (Huang et al., 2015). Unlike MDA, this bias is reproducible along the genome from cell to cell (Zong et al., 2012). Therefore, signal normalization for CNV noise reduction can be effective. For example, MALBAC offers high CNV accuracy after signal normalization with a reference genome following single cell sequencing analysis of human cells (Zong et al., 2012). However, the extremely imbalanced base composition (80.6% AT content) and small size of the *P. falciparum* genome (23Mb haploid, which is 2400X smaller than the ~3234Mb diploid human genome) push the limits of MALBAC. Additionally, degenerate MALBAC primers are more likely to melt from high-AT regions of the *P. falciparum* genome during the elevated extension temperature. Thus, we optimized the MALBAC amplification method for the *P. falciparum* parasite in this study.

Besides a robust amplification method, reliable CNV calling after sequencing is also essential for this study. Generally, CNV analysis tools detect CNVs through four distinct approaches on short sequencing reads, including 1) paired-end mapping strategies through discordantly mapped reads (unexpected mapping distance and orientation of paired reads; 2) split-read-based approaches by incomplete mapping (substrings of reads mapping to different genomic locations; 3) read-depth based approaches via detecting the border of consecutive windows/bins with increased or decreased read counts (assuming there is a correlation between depth of coverage and the copy number of a genome region); 4) assembly-based approaches by mapping and comparing contigs to the reference genome. Read depth (RD) can detect the exact number of CNVs, while paired-read (RP) and split-read (SR) can only report the position of the potential CNVs but not the counts (Tattini et al., 2015). In addition, RD can work better on large-size CNVs, which are hard to detect by RP and SR (Yoon et al., 2009). LUMPY is a probabilistic model integrating any or all of the three different signals (RP, SR, and RD) from a single sample (Layer et al., 2014). By integrating evidence and probabilities from different signals (RP, SR, RD), LUMPY determines the type of CNVs (e.g. duplication and deletion) and the breakpoint interval (a pair of genomic coordinates that are adjacent in a sample genome but not in a reference genome) (Layer et al., 2014). In general, the probability of the location and variety (e.g. duplication and deletion) of a breakpoint is affected by the distribution of library fragment length, read quality and read alignment quality of the sequenced sample, and the distribution of read count along the genome. Although RP, SR, and RD-based approaches have been used in CNV analysis of bulk *P. falciparum* genome sequencing (Huckaby et al., 2018), several challenges need to be addressed for single parasite sequencing datasets. Firstly, a lower fraction of the genome is covered with single cell sequencing compared to bulk sequencing, which makes split-read, and paired-end approaches less effective. Secondly, whole-genome amplification introduces biases that markedly distort read counts, including failure to amplify entire segments. Thirdly, repetitive regions of the genome (e.g. subtelomeric regions) lead to the artificial inflation of read counts ("bad bins").

Most single cell CNV studies have been done in human cells using read-depth-based approach (Zhang et al., 2013; Cai et al., 2014; Knouse et al., 2016; Chronister et al., 2019). Read-depth

CNV analysis generally includes the following steps: 1) aligning reads to reference genome and counting read-depth in predefined windows/bins; 2) normalizing read counts to remove potential biases, mainly due to GC content and repetitive regions (Janevski et al., 2012); 3) using segmentation algorithm to identify a contiguous set of windows sharing the same number of CNVs; 4) predicting the statistical significance of the calls and filtering called CNVs (Zhao et al., 2013). By utilizing various strategies or tools for binning, normalization, segmentation, and filtering in read-depth CNV analysis, several single cell CNV studies have improved their CNV calling accuracy and limited false positive CNV calls (Zhang et al., 2013; Cai et al., 2014; Knouse et al., 2016; Chronister et al., 2019). A recent study detected rare megabase-scale CNVs in human neuron cells through optimized read depth-based single cell CNV analysis (Chronister et al., 2019). In their study, they tuned the parameter combinations of the DNAcopy segmentation for single cell data and effectively eliminated false positive CNV calls with filtering strategies (Chronister et al., 2019). They also applied their optimized method in other published datasets by adjusting filtering cutoffs. As mentioned earlier, the amplification bias from MALBAC can be reduced via normalization to a reference sample. As proof of principle, one study detected known CNVs in MALBAC-amplified single cancer cells after comparing and normalizing the read coverage to a MALBAC-amplified blood cell (Zong et al., 2012). With bulk sequencing of human cells, read-depth methods allowed the detection of CNVs as small as 500bp in 37X read depth (Miller et al., 2011). However, due to the difficulties in discovering real CNVs at single cell level, most studies limit CNV detection to sub-megabase or mega-base sizes by extending bin size in read-depth analysis in human cells (Huang et al., 2015). In this study, we utilized and combined Lumpy (based on split-read and discordant-read) and Ginkgo (based on read depth) to limit the detection of false positive CNVs in single parasite DNA.

Nanopore sequencing is an alternative method for heterogeneous CNV detection

As we mentioned above, bulk DNA sequencing via short read sequencing cannot detect variants that present at a very low frequency in a cell population, and whole genome amplification of single parasite DNA introduces amplification bias that may complicate CNV detection. Instead, third generation sequencing technology, Nanopore single molecule long read sequencing, provides an alternative to detect sub-populational CNVs using bulk DNA from many cells. Importantly, the genome of *P. falciparum* has extremely high AT content (80.6%) and has many repetitive regions. As a result, many regions of the parasite genome have many short tandem repeats and other low complexity sequences unusually abundant in both coding and non-coding regions (Gardner et al., 2002). The detection of CNVs has been challenging via short read sequencing, which includes many limitations in the access to high AT content regions, resolution of complex regions of the genome, repetitive regions where short reads will not map uniquely, and difficulty in detecting large structural variation. While as long reads can span the low complexity and repetitive regions, detection of structural variants (e.g. large segmental duplications) through Nanopore long reads is possible. Especially, with Nanopore single molecule sequencing, each read is originated from a single DNA fragment in the bulk DNA. It is possible for researchers to start to exam the clonal heterogeneity of pathogens in Nanopore long reads.

1.4 Introductory remarks

In this dissertation, I use the malaria parasite, *Plasmodium falciparum,* as a model organism to understand the adaptation strategies in single cellular organisms and demonstrate multiple approaches to detect heterogeneous CNVs in the parasite genome. With these new approaches, we can improve our knowledge about the molecular mechanisms behind adaptation (resistance acquisition) in malaria parasites. More specifically, I identified a CNV hotspot that may compensate for the fitness cost of a resistance-conferring gene amplification (**Chapter II**), detected novel rare CNVs in parasite populations with or without environment stress using Nanopore long reads (**Chapter III**), and described the experimental and bioinformatics pipelines of single cell sequencing to identify heterogeneous CNVs (**Chapter IV**). I also combined Illumina short reads and Nanopore long reads to generate new genome assemblies for laboratory parasite lines to improve the accuracy of reads mapping and CNV detection (**Chapter V**). Lastly, I discussed how we can combine single cell sequencing and Nanopore single molecular long read sequencing together to detect various sizes of heterogeneous CNVs in parasite populations and how this impacts our understanding of resistance evolution in malaria parasites (**Chapter VI**). Altogether, this dissertation emphasizes the variety of methods to detect heterogeneous CNVs, the CNV dynamics in evolving malaria parasites, and the importance of understanding genome adaptation in single cellular organisms.

2 CHAPTER II: Expansion of GTP cyclohydrolase I copy number in malaria parasites resistant to a pyrimidine biosynthesis inhibitor

The study presented in this chapter is prepared for publication using the title:

Expansion of GTP cyclohydrolase I copy number in malaria parasites resistant to a pyrimidine biosynthesis inhibitor

Manuscript Author list: Shiwei Liu[1], Emily R. Ebel[2], Jane Kim[1], Nnenna Ene[1], Thomas Werner Anthony Braukmann[3], Ellen Yeh[3], Elizabeth S. Egan[2], Jennifer L. Guler[1]

[1]University of Virginia, Department of Biology, Charlottesville, VA, USA

[2]Stanford University, Departments of Pediatrics and Microbiology & Immunology, Stanford, CA, USA

[3]Stanford University, Departments of Pathology and of Microbiology & Immunology, Stanford, CA, USA

Statement of contribution and acknowledgement

S.L and J.L.G. designed the study and wrote the manuscript. S.L., N.E. and J.K performed the experiments and data analysis. E.R.E developed the Shiny application and provided instructions for using the application. E.R.E. and E.S.E. edited and reviewed the manuscript. E.Y. and T.W.A.B. conducted the experiment described in the Supplementary Method and wrote the Supplementary Method. All authors contributed to the article and approved the submitted version.

Our thanks to Michelle Warthan and Noah Brown (University of Virginia) for experimental support, Martin Wu and Mercedes Campos-Lopez (University of Virginia) for Oxford Nanopore sequencing support, and the laboratories of Dr. William Petri Jr (University of Virginia) for their helpful discussions and insight. We are grateful for David Fidock (Columbia University), Dyann F. Wirth (Harvard University) for sharing sequencing data.

2.1 Abstract

Changes in the copy number of large genomic regions, termed copy number variations or CNVs, are an important adaptive strategy for malaria parasites. Numerous CNVs across the *Plasmodium falciparum* genome contribute directly to drug resistance or impact the fitness of this protozoan parasite. CNVs that encompass the dihydroorotate dehydrogenase (DHODH) gene confer resistance to antimalarials that target this enzyme in the pyrimidine biosynthesis pathway (i.e. DSM1). Compounds in this class of inhibitors are currently in clinical trials for the treatment of malaria. During the characterization of DSM1-resistant parasite lines with DHODH CNVs, we identified an additional CNV that encompasses 3 genes (~5 kb) including GTP cyclohydrolase I (GCH1 amplicon). While this locus has been implicated in the increased fitness of antifolate-resistant parasites, GCH1 CNVs had not previously been reported to contribute to resistance to other antimalarials. Here, we further explored the association between GCH1 and DHODH copy numbers. We visualized single long reads and directly quantified the number of tandem GCH1 amplicons in a parental line versus a DSM1-selected line. We found that the GCH1 amplicons share a consistent structure. However, we detected more reads that encompassed a higher number of amplicons in the resistant (up to 7 amplicons) compared to the parental line (3 amplicons). To better understand the implications of this result, we evaluated variation at this locus across multiple short- and long-read data sets collected from various parasite lines. Based on our analysis of parasites resistant to other DHODH inhibitors (DSM265, DSM267, and DSM705), GCH1 is not likely to contribute directly to resistance; however, higher numbers of the GCH1 amplicon are associated with increased DHODH copies and may compensate for changes in the metabolism of resistant parasites. This is supported by the direct connection between folate and pyrimidine metabolism, which together contribute to nucleic acid

biosynthesis. This study highlights the importance of studying this relationship further as DHODH inhibitors move closer to clinical approval.

2.2 Introduction

Malaria is a disease caused by the protozoan *Plasmodium* parasite. *Plasmodium falciparum* is the leading cause of human malaria deaths (Rich et al., 2009). Due to the lack of effective vaccines against malaria infection, antimalarial drugs are the primary approach for malaria treatment (Casares et al., 2010). However, drug efficacy is mitigated by the frequent emergence of antimalarial-resistant parasites (Blasco et al., 2017).

Changes in the copy number of large genomic regions, termed copy number variations, or CNVs, are an important adaptive strategy for malaria parasites (Kidgell et al., 2006; Conway, 2007; Hyde, 2007; Ribacke et al., 2007; Nair et al., 2008; Cheeseman et al., 2009; Bopp et al., 2013; Guler et al., 2013; Menard and Dondorp, 2017). Numerous CNVs across the *P. falciparum* genome contribute directly to drug resistance or impact parasite fitness (Hyde, 2007; Ribacke et al., 2007; Nair et al., 2008; Guler et al., 2013). Amplification, one type of CNV with increased copy number, plays an essential role in the evolution of resistance to various antimalarials (Foote et al., 1989; Wilson et al., 1993; Hyde, 2007; Ribacke et al., 2007; Cheeseman et al., 2009; Guler et al., 2013; Heinberg et al., 2013; Osei et al., 2018). As one example, amplification of the dihydroorotate dehydrogenase (DHODH) gene in the *P. falciparum* genome confers resistance to DHODH inhibitors (i.e. DSM1) in parasites propagated in vitro (Guler et al., 2013). DHODH is an important enzyme in the *P. falciparum* pyrimidine biosynthesis pathway that contributes resources for nucleic acid synthesis (Phillips and Rathod, 2010; Mandt et al., 2019). DHODH amplicons presumably increase transcription and translation of the drug target to directly impact drug sensitivity (Guler et al., 2013).

In another example, amplification of the GTP cyclohydrolase 1 (GCH1) gene increases the fitness of clinical parasite populations that are antifolate resistant (i.e. pyrimethamine and sulfadoxine) (Kidgell et al., 2006; Ribacke et al., 2007; Nair et al., 2008; Osei et al., 2018). GCH1 is the first enzyme in the folate biosynthesis pathway and increased levels of this enzyme likely increases flux to compensate for the fitness costs of the resistance-conferring dihydropteroate synthase (DHPS) and dihydrofolate synthase (DHFS) mutations (Kidgell et al., 2006; Nair et al., 2008; Heinberg et al., 2013). Although the contribution of GCH1 amplification to antifolate resistance is well studied, this CNV has not been reported to contribute to resistance to antimalarials targeting other pathways (Heinberg et al., 2013; Kümpornsin et al., 2014; Heinberg and Kirkman, 2015).

Typically, the gene copy number is studied using widely accessible high coverage short read sequencing (Guler et al., 2013; Herman et al., 2014; Manary et al., 2014; Cowell et al., 2018; Huckaby et al., 2018). However, this approach has limitations including non-unique mapping in repetitive regions, the inability to resolve complex genomic regions, and the overall difficulty in detecting structural variations (Alkan et al., 2011; Treangen and Salzberg, 2012; Kosugi et al., 2019). These challenges are exacerbated by the high AT content of the *P. falciparum* genome (Beghain et al., 2016; Miles et al., 2016). Long read technologies such as Oxford Nanopore sequencing have the potential to span low complexity and repetitive regions to better represent structural variation (Cretu Stancu et al., 2017; Sedlazeck et al., 2018a, 2018b; Ho et al., 2020).

Moreover, the single molecule sequencing allows examination of clonal heterogeneity of a parasite population (Belikova et al., 2020; Rugbjerg et al., 2021).

In this study, we identified a positive association between GCH1 and DHODH copy number from previously acquired short read sequencing data. To explore this association further, we performed long read sequencing and directly observed the expansion of the GCH1 amplicon. Using single long read visualization, we also determined that the structure and orientation of the amplicon was preserved during expansion. When we evaluated short read data from parasite lines that were resistant to other DSM-based compounds, we did not detect increases in GCH1 copy number. This result indicates that GCH1 does not contribute directly to resistance; however, our observations, as well as the biochemical connection between folate and pyrimidine biosynthesis, suggest that increased copies of GCH1 may facilitate the acquisition of increased DHODH copy number under certain selection conditions. Further study of the relationship between these two genomic loci is important considering the imminent use of DHODH inhibitors to treat clinical malaria.

2.3 Materials and Methods

DSM1 and parasite clones

DSM1 is a triazolopyrimidine antimalarial that specifically and potently inhibits the *P. falciparum* dihydroorotate dehydrogenase enzyme of pyrimidine biosynthesis (Phillips et al., 2008). DSM1-resistant parasites were previously selected according to the scheme depicted in (**Figure 2.1**) (Guler et al., 2013). In this study, we simplified the naming scheme to represent low (L), moderate (M), and high (H) levels of resistance to DSM1 (**Figure 2.1A**).

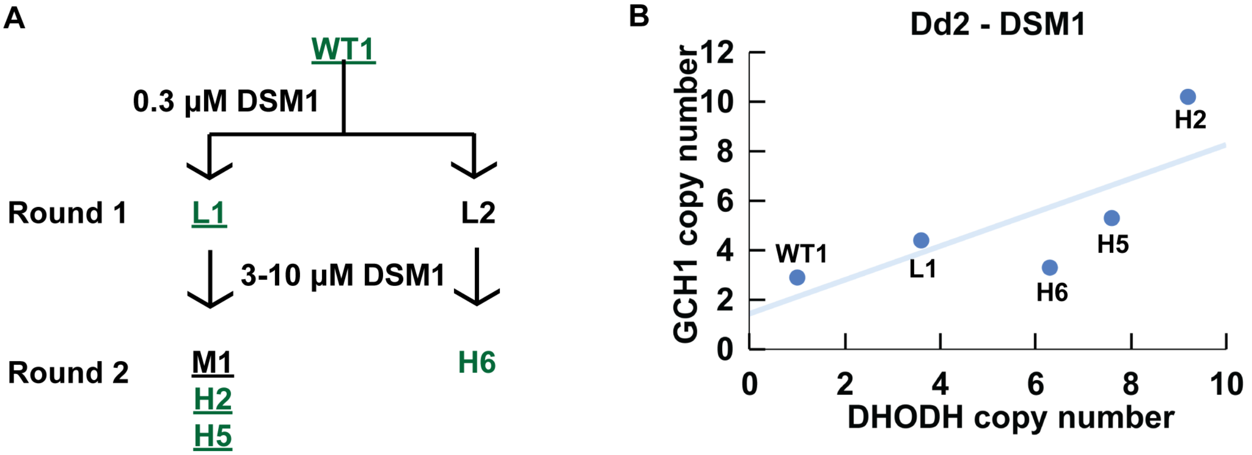


Figure 2.1 GCH1 copy number increase is positively correlated with DHODH copy number in one family of DSM1 resistant parasites

(A) Schematic depicting DSM1 selections, as presented previously (Guler et al. PLoS Pathogens 2013). Green: Illumina short read sequenced lines. Underline: modern lines confirmed by ddPCR analysis. Wild-type (WT1, *Dd2*) *P. falciparum* was selected with DSM1 in two steps; the first step selected for low-level (L) resistance and the second step selected for moderate- (M) or high-level (H) resistance. DSM1 EC50 values are as follows: L1 (1 μM), L2 (0.9 μM), M1 (7.2 μM), H2 (85 μM), H5 (56 μM), H6 (49 μM). All values were previously reported in and clone names adapted from (Guler et al., 2013). **(B)** Relationship between GCH1 and DHODH copy number in DSM1 selected parasites as quantified using short read data from Guler et al. 2013. A trendline was added to show the relationship between GCH1 and DHODH copy numbers.

Parasite culture

We thawed erythrocytic stages of *P. falciparum* (Dd2, MRA-150; 3D7, MRA-102, Malaria Research and Reference Reagent Resource Center, BEI Resources and DSM1 resistant clones as highlighted in **Figure 2.1A**, a generous gift from Pradipsinh Rathod, University of Washington) from frozen stocks and maintained them as previously described (Haynes et al., 1976). Briefly, we grew parasites at 37 °C in vitro at 3% hematocrit (serotype A positive human erythrocytes, Valley Biomedical, Winchester, VA) in RPMI 1640 medium (Invitrogen, USA) containing 24 mM NaHCO3 and 25 mM HEPES and supplemented with 20% human type A positive heat inactivated plasma (Valley Biomedical, Winchester, VA) in sterile, sealed flasks flushed with 5% O2, 5% CO2, and 90% N2 (Guler et al., 2013). We maintained the cultures with media changes every other day and sub-cultured them as necessary to keep parasitemia below 5%. We determined all parasitemia measurements using SYBR green-based flow cytometry (Bei et al., 2010). We routinely tested cultures using the LookOut Mycoplasma PCR Detection Kit (Sigma-Aldrich, USA) to confirm negative Mycoplasma status.

DNA extraction for long read sequencing

We lysed asynchronous *P. falciparum*-infected erythrocytes with 0.15% saponin (Sigma-Aldrich, USA) for 5 min at room temperature and washed them three times with 1× PBS (diluted from 10× PBS Liquid Concentrate, Gibco, USA). We then lysed parasites with 0.1% Sarkosyl Solution (Bioworld, bioPLUS, USA) in the presence of 1 mg/ml proteinase K (from Tritirachium album, Sigma-Aldrich, USA) overnight at 37 °C. We first extracted nucleic acids with phenol/chloroform/isoamyl alcohol (25:24:1) pH 8.0 (Sigma-Aldrich, USA) three times using 1.5 ml light Phase lock Gels (5Prime, USA), then further extracted nucleic acids with chloroform twice using 1.5 ml light Phase lock Gels (5Prime, USA). Lastly, we precipitated the DNA with ethanol using the standard Maniatis method (Maniatis et al., 1989). To obtain high molecular weight genomic DNA, we avoided any pipetting during the extraction, transferred solutions by directly pouring it from one tube to another, and mixed solutions by gently inverting the tubes.

Oxford Nanopore long read sequencing and analysis

We subjected 1 µg of high molecular weight genomic DNA from each sample to library preparation for Oxford Nanopore sequencing following the Nanopore Native barcoding genomic DNA protocol (version: NBE_9065_v109_revAB_14Aug2019) with 1x Ligation Sequencing kit (SQK-LSK109, Oxford Nanopore Technologies, Oxford, UK). We performed DNA repair and end preparation using NEBNext FFPE DNA Repair Mix (New England Biolabs, Ipswich, MA, USA) and NEBNext End Repair/dA-Tailing Module (New England Biolabs, Ipswich, MA, USA). We cleaned the A-tailed fragments using 0.9X AMPure XP beads (Beckman Coulter, High Wycombe, UK). We then ligated barcodes to the end-prepped DNA using the Native Barcoding Expansion 1–12 kit (EXP-NBD104, Oxford Nanopore Technologies, Oxford, UK) and Blunt/TA Ligase Master Mix (New England Biolabs, Ipswich, MA, USA). We cleaned the barcoded samples using 0.9X AMPure XP beads. Then we pooled barcoded samples in equimolar ratios and subjected them to an adaptor ligation step, using the Adapter Mix II from the Native Barcoding Expansion 1–12 kit and NEBNext Quick Ligation Reaction Buffer (New England Biolabs, Ipswich, MA, USA) as well as Quick T4 DNA Ligase (New England Biolabs, Ipswich, MA, USA). After adaptor ligation, we cleaned the library using AMPure XP beads. We quantified the adapter-ligated and barcoded DNA using a Qubit fluorimeter (Qubit 1X dsDNA High Sensitivity Assay Kit, Life Technologies, Carlsbad, CA). We sequenced the WT1 and initial H2 libraries using a R9.4.1 flow cell (FLO-MIN106D, Oxford Nanopore Technologies, Oxford, UK) on MinION (Oxford Nanopore Technologies, Oxford, UK) and another H2 library using the R10 flow cell (FLO-MIN111) on MinION. To obtain the maximum number of reads, we ran both flow cells for 48 h (controlled and monitored using the MinKNOW software (3.6.5)).

For base calling and demultiplexing of the Nanopore sequencing reads, we used Guppy (version 3.4.5+fb1fbfb) with the parameter settings "-c dna_r9.4.1_450bps_hac.cfg --barcode_kits "EXP-NBD104" -x auto" for samples sequenced with R9.4.1 flow cell and "-c dna_r10_450bps_hac.cfg -x auto" for the sample sequenced with R10 flow cell. We checked the read length and read quality using "Nanoplot" (version 1.0.0) (see Supplementary Table 2.1). We trimmed the adapters with "qcat" (version 1.1.0) (Oxford Nanopore Technologies) and filtered the reads with a cutoff "length ≥ 500 and Phred value ≥ 10" using the program " filtlong version 0.2.0" (https://github.com/rrwick/Filtlong). To estimate the coverage of sequencing reads in each sample, we aligned the filtered reads to *Plasmodium falciparum* 3D7 reference genome using "minimap2" (version 2.17) (Li, 2018). "QualiMap" (version 2.2.1) (García-Alcalde et al., 2012) was used to calculate the coverage of the aligned reads (see Supplementary Table 2.1).

Shiny analysis of long reads

To visualize structural variants in the parasite genome, we used a custom R Shiny script to plot the arrangement of reference gene segments on individual Nanopore reads. Briefly, we defined a target region in the 3D7 reference genome (chromosome 12: 932916 bp – 999275 bp) that contained 3 genes in GCH1 amplicon and 11 flanking genes. We extracted the reference sequences of these genes and subsequent intergenic regions, then split these sequences into fragments of 500-1000 base pairs. We compared these fragments to individual Nanopore reads using BLAST (Ye et al., 2006). We used the BLAST output as input for a custom R script, which drew rectangles representing homology between the defined genes (y-axis) and each individual read (x-axis). The percent identity required to draw a homologous rectangle was allowed to vary between reads, which varied in quality, using a slider in the Shiny app. To filter out long reads with potentially spurious hits to gene fragments, we also used BLAST to compare Nanopore reads to the reference genome and removed reads with <90% identity to chromosome 12. To compare the mean copy number of GCH1 amplicon between WT1 and H2 reads covering GCH1 as well as the read length of these reads between WT1 and H2, we performed a one-way ANOVA in Microsoft Excel with Alpha value of 0.05.

Short read sequencing analysis and CNV detection

We analyzed CNVs in Illumina short read datasets of *P. falciparum* parasites selected by three DSM antimalarial drugs (DSM1, DSM265, DSM267, and DSM705, Supplementary Table 2.2) (Guler et al., 2013; Mandt et al., 2019; Palmer et al., 2021). We first processed and mapped the reads to reference genome as previously described (Huckaby et al., 2018; Liu et al., 2021). Briefly, we trimmed Illumina adapters from reads with BBDuk tool in BBMap (version 38.57) (Bushnell 2016). We aligned each fastq file to the 3D7 *P. falciparum* reference genome with Speedseq (version 0.1.0) through BWA-MEM alignment (Chiang et al., 2015). We discarded the reads with low-mapping quality score (below 10) and removed duplicated reads using Samtools (version 1.10) (Li et al., 2009). We analyzed split and discordant reads from the mapped reads using LUMPY in Speedseq to determine the location and length of the previously reported

GCH1, DHODH and multidrug resistance protein 1 (MDR1) CNVs (**Supplementary Table 2.2**) (Layer et al., 2014). For read-depth analysis, we further filtered the mapped reads using a mapping quality score of 30. To determine the copy number of the GCH1, DHODH and MDR1 CNVs, we used CNVnator (version 0.4.1) with a bin size of 100 bp; the optimal bin size was chosen to detect GCH1 CNVs in all analyzed samples (Abyzov et al., 2011).

Droplet Digital PCR

Prior to Droplet Digital (dd) PCR, we digested DNA with restriction enzyme RsaI (Cut Site: GT/AC) following the manufacturer's instructions (New England Biolabs, Ipswich, MA, USA) in 37°C incubation for one hour. We selected the restriction enzyme RsaI to cut outside of the ddPCR amplified regions of desired genes and separate copies of CNVs to be distributed into droplets. We diluted the digested DNA for ddPCR reactions. We performed ddPCR using ddPCR Supermix for Probe (no dUTP, Bio-Rad Laboratories, California, USA) with DNA input 0.1 ng (in duplicate per sample), 0.025 ng (in duplicate per sample) as previously described (McDaniels et al., 2021). The primers and probes used in reactions are included in Supplementary Table 2.3. The PCR protocol for the probe-based assay was 95°C for 10 min, followed by 40 rounds of 95°C for 30 sec and 60°C for 1 min. Seryl-tRNA synthetase and calcium-transporting ATPase (ATP6) served as single-copy reference genes on chromosome 7 and chromosome 1 respectively; dihydroorotate dehydrogenase (DHODH) and GTP cyclohydrolase 1 (GCH1) are multi-copy genes (Supplementary Table 2.3). We performed droplet generation and fluorescence readings per the manufacturer's instructions. For each reaction, we required a minimum number of 10,000 droplets to proceed with analysis. We calculated the ratio of positive droplets containing a single- (ATP6) or multi-copy gene (GCH1, DHODH) versus a single-copy gene (Seryl-tRNA synthetase) using the Quantasoft analysis software (QuantaSoft Version 1.7, BioRad Laboratories) and averaged between independent replicates.

2.4 Results

Through analysis of short read data from a family of parasites selected with DSM1, originally presented in (Guler et al., 2013) (**Figure 2.1A**), we noticed a positive association between GCH1 and DHODH copy number (**Figure 2.1B**). This association was confirmed using a second method, droplet digital PCR, on analogous parasite lines that had recently been propagated in our laboratory; GCH1 copy number trends higher as DHODH copy number increases (**Table 2.1**). Correlation coefficient is not calculated due to the small sample size (n=5) and dependence among the lines.

Table 2.1 Positive GCH1: DHODH association is validated using Digital Droplet PCR on modern parasite lines.

Line	Sample type	Average copy number (SD)*			GCH1 CN relative to parent	DHODH CN relative to parent
		GCH1	DHODH	ATP6		
WT1	Parent	2.5 (0.2)	0.8 (0.1)	1.1 (0.0)	1.0	1.0
L1	Round 1 selection	3.9 (0.3)	3.2 (0.1)	1.0 (0.1)	1.6	4
M1	Round 1 selection	3.9 (0.1)	5.2 (0.3)	1.0 (0.1)	1.6	6.5
H2	Round 2 selection	4.6 (0.2)	6.0 (0.3)	1.1 (0.1)	1.8	7.5
H5	Round 2 selection	4.3 (0.3)	7.0 (0.4)	1.0 (0.1)	1.7	8.8

*The average copy number is calculated by comparing ddPCR signal to a single copy gene signal (Seryl tRNA synthetase, PF3D7_0717700). N=4. ATP6 copy number is expected to be 1 in all parasite lines.

We conducted long read sequencing to more precisely define GCH1 copy number and amplicon structure in the parental line versus one DSM1-selected line (WT1 and H2, **Supplementary**

Table 2.1). We directly visualized single reads using an app that represents gene segments of individual Nanopore reads (see **Materials and Methods**). Small amplicons, like those including GCH1, are especially conducive to this approach because long reads span multiple copies of the amplicon as well as both up and downstream regions. Read visualization showed that two 3-gene amplicons, separated by an inversion of the same gene set, were conserved between the parental and H2 lines (**Figure 2.2A and B**). This amplicon structure was reported previously in the WT1 (Dd2) parental line (**Figure 2.2C**) (Kidgell et al., 2006).

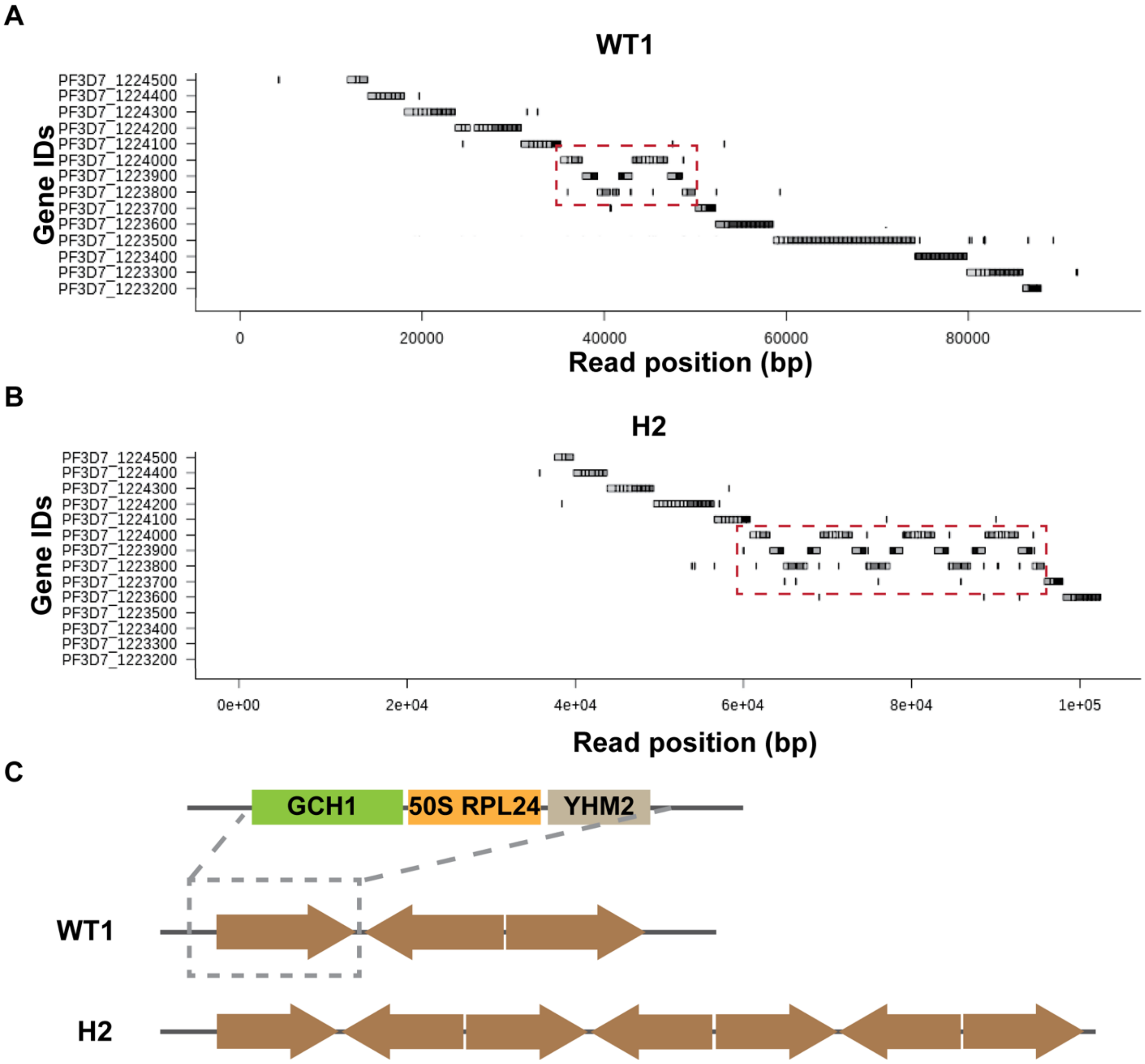


Figure 2.2 Long-read visualization shows GCH1 amplicon has the same boundaries and structure in a DSM1 resistant parasite line

(A) and **(B)** Representative images from the Shiny app comparing the GCH1 amplicon in WT1 and H2 reads (alignment to the *3D7* reference genome with no GCH1 amplicon represented). Red dashed square: GCH1 amplicon. Each gene sequence was split into <=500 bp fragments and blasted against individual Nanopore reads (darker: genic regions, lighter: intergenic regions). **(C)** Diagram of amplicon orientation in GCH1 amplicon for WT1 and H2 parasite lines. The three genes within the GCH1 amplicon unit include PF3D7_1224000 (GTP cyclohydrolase I, GCH1), PF3D7_1223900 (50S ribosomal protein L24, putative, 50S RPL24), and PF3D7_1223800 (citrate/oxoglutarate carrier protein, putative, YHM2).

Using read visualization, we also manually recorded the number of GCH1 amplicon units (depicted in **Figure 2.2C**) from both spanning and non-spanning reads (**Figure 2.3A, Supplementary Table 2.4**). We consistently detected 3 amplicon units per read from the parental line. However, we observed more reads that encompassed a higher number of GCH1 amplicon units in the H2 selected line (up to 7 units). If only considering spanning reads, that is those covering both upstream and downstream of the amplicon, the difference between WT1 and H2 mean copy number is significantly different (F(1,10)=[75], p = 5.84E-06) revealed by a one-way ANOVA. A one-way ANOVA also revealed that there was not a significant difference in mean read length between WT1 and H2 (F(1,34)=[1.07], p = 0.31, **Figure 2.3B**), indicating that a skewed read distribution was not contributing to this difference. All groups of amplicons from the H2 line began with a set of 3 amplicon units, as observed in the parental line, followed by groups of 2 amplicon units where one was inverted and the other was in normal orientation (**Figure 2.2C**).

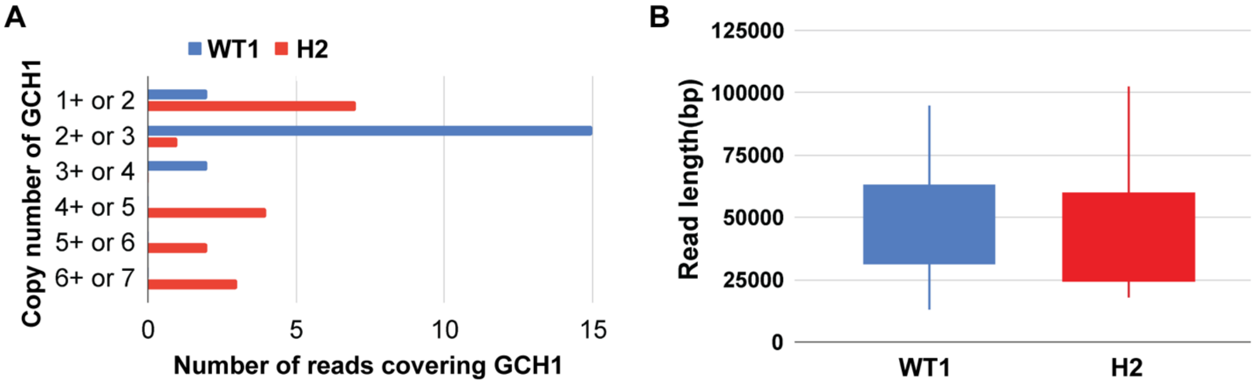


Figure 2.3 Quantification of long reads displays greater variability and increase in GCH1 amplicon number in DSM1 resistant parasites
Copy number **(A)** and read length **(B)** distributions from Nanopore long reads (>=5kb) covering the GCH1 amplicon in WT1 and H2 parasite line. A one-way ANOVA revealed that there was not a significant difference in mean read length between WT1 and H2 (F (1,34)=[1.07], p = 0.31).

To better understand whether variation in copy number of the GCH1 locus was common in other laboratory-adapted parasite lines, we evaluated several additional long read-based datasets (**Supplementary Table 2.4**, Supplementary Method, Vembar et al., 2016). In general, different parasite lines exhibited different GCH1 amplicon sizes as expected (i.e. Dd2 versus 3D7, as reported previously (Kidgell et al., 2006)) but the amplicon copy number was relatively stable across diverse datasets; we detected ~10-25% of reads above the expected copy number for each respective line. This contrasts with the H2 line that had >50% of reads that depicted amplicon copy numbers greater than the expected copy number for a Dd2-derived line.

To investigate whether an increase in GCH1 amplicons was contributing to the resistance of DHODH inhibitors in general, we evaluated the copy number of GCH1 amplicons from parasites selected with other DSM derivatives (Mandt et al., 2019; Palmer et al., 2021). Contrary to DSM1-selected parasites, we did not detect increases in GCH1 amplicon number in parasites resistant to DSM265, DSM267, or DSM705 (compared to parental Dd2 or 3D7 parasite lines,

Figure 2.4A-2.4C, Supplementary Table 2.5). For this analysis, we only included resistant parasite lines that carried DHODH amplicons. When we compiled the data for each parasite line together, those that harbored more DHODH amplicons generally maintained higher numbers of GCH1 amplicons and this association reached significance in Dd2-derived parasite lines but not 3D7-derived parasite lines (**Figure 2.4D and 2.4E**).

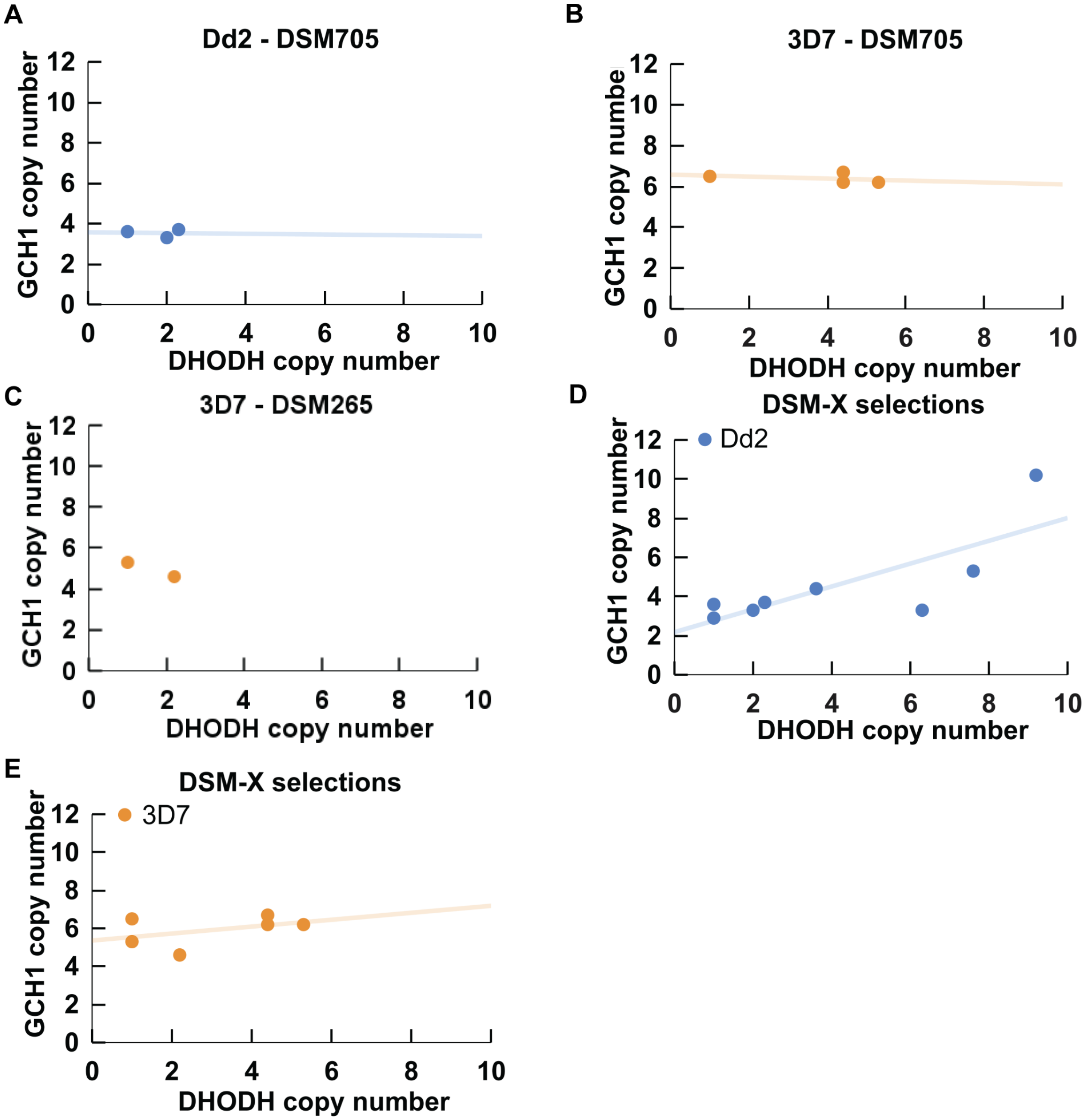


Figure 2.4 Parasite lines with more DHODH amplicons generally have more GCH1 amplicons

(A) Relationship between GCH1 and DHODH copy number in *Dd2* parent line and DSM705 selected lines. **(B)** Relationship between GCH1 and DHODH copy number in *3D7* parent line and DSM705 selected lines. **(C)** Relationship between GCH1 and DHODH copy number in *3D7* parent line and DSM265 selected lines. **(D)** Relationship between GCH1 and DHODH copy number in *Dd2* parent line and DSM1, DSM705, DSM265-selected lines. Data from DSM1 selected lines are represented in Figure 1B as well. **(E)** Relationship between GCH1 and DHODH copy number in *3D7* parent line and DSM1, DSM705, DSM265-selected lines. Correlation coefficients are not calculated due to small sample sizes and dependence among DSM1 selected lines. Trendlines were added to show the relationship between GCH1 and DHODH copy numbers.

2.5 Discussion

We found that GCH1 copy number is positively associated with DHODH copy number in parasites resistant to DSM1. This association was initially detected using both short-read sequencing as well as an accurate PCR-based method ddPCR (**Figures 2.1B and 2.4**, **Table 2.1**); however, limitations in each of these methods contribute to inaccuracies in these results. For example, because the GCH1 amplicon is small (~5kb), copy number analysis using short reads requires small stretches of the genome to be combined, or binned, together; smaller-sized bins in general leads to higher levels of variation in CNV calling and potentially false differences. Furthermore, DNA fragmentation and restriction digestion can alter the distribution of DNA fragments into oil droplets during ddPCR, thus limiting copy number quantification. Additionally, both methods result in an average value for the entire population of parasites.

We, therefore, utilized long-read sequencing combined with a custom visualization tool to directly quantify the copy number of this locus and assess the structure of the GCH1 amplicon (**Figures 2.2 and 2.3**). Each long read represents a DNA strand from a single parasite genome and therefore, this approach is an accurate way to visualize copy number heterogeneity across a population of haploid parasites. We considered whether our detection of more GCH1 amplicons on reads from the H2 line (**Figure 2.3A**) was related to Nanopore sample preparation or natural variation during in vitro culture. We excluded differences in sample preparation because the parental and H2 long read sequencing runs had similar N50s and read length distributions (**Supplementary Table 2.4, Figure 2.3B**). While we did observe a low level of variation at the GCH1 locus in different parasite lines that were grown independently (~10-25% of reads), the variation in copy number in the H2 line was well above this level (>50% of reads,

Supplementary Table 2.5). This observation, combined with the visualization of up to 7 tandem amplicons in a single read (**Figure 2.2B**), provided direct evidence for GCH1 amplicon expansion in the DSM1 resistance context.

GCH1 amplicons have previously only been associated with antifolate resistance (Kidgell et al., 2006; Nair et al., 2008; Heinberg et al., 2013; Kümpornsin et al., 2014; Heinberg and Kirkman, 2015; Osei et al., 2018). An increase in flux through the folate biosynthesis pathway alleviates fitness effects of mutations that confer pyrimethamine and sulfadoxine resistance (Kidgell et al., 2006; Nair et al., 2008; Kümpornsin et al., 2014; Heinberg and Kirkman, 2015; Osei et al., 2018). A similar contribution to resistance to DHODH inhibitors would not be surprising, given the close connection of the folate and pyrimidine biosynthesis pathways (**Figure 2.5**); they both contribute to nucleotide biosynthesis and converting dUMP to dTMP requires conversion of N5, N10-methylene-THF to DHF. We speculate that a change in GCH1 copy number arose serendipitously during DSM1 selection and further increases were beneficial for parasite fitness, thus increased copies of GCH1 may facilitate the acquisition of increased DHODH copy numbers.

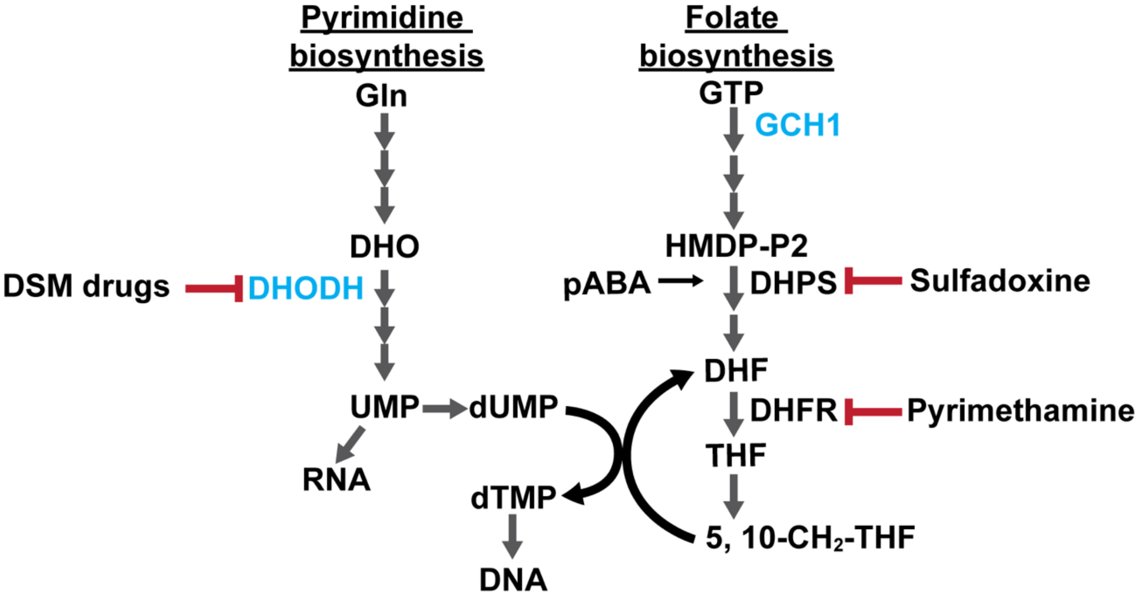


Figure 2.5 The connection between pyrimidine and folate biosynthesis pathways
Enzymes with gene copy number variations are indicated in blue. Gln: Glutamine; DHO: Dihydroorotate; UMP: Uridine monophosphate; dUMP: Deoxyuridine monophosphate; dTMP: deoxythymidine monophosphate; GTP: Guanosine-5'-triphosphate; DHPS: Dihydropteroate synthase; DHFR: Dihydrofolate reductase; DHF: Dihydrofolate; THF: Tetrahydrofolate; HMDP-P2: 6-hydroxymethyl-7, 8-dihydropterin diphosphate; pABA: para-amino-benzoic acid; 5, 10-CH2-THF: 5,10-Methylenetetrahydrofolate.

Although not required for resistance to all DHODH inhibitors (**Figure 2.4A-C, Supplementary Table 2.5**), the positive association in DSM1-selected parasites indicates that GCH1 amplicons may be especially favorable under specific conditions. For example, DSM1 selections were performed in media supplemented with 20% human serum as opposed to AlbuMAX II (**Supplementary Table 2.5**) (Guler et al., 2013). It is possible that the presence of folate precursors in different media formulations changes the parasite's dependence on folate biosynthesis and thus, GCH1 flux. Levels of para-amino benzoic acid (pABA) can vary widely in human serum while the common AlbuMAX serum replacement contains no additional pABA (Chulay et al., 1984; Watkins et al., 1985; Wang et al., 1986; Salcedo-Sora et al., 2011; Valenciano et al., 2019). These environments likely exert different selective pressures on parasites during drug selection as reported previously (Kumar et al., 2021); higher pABA levels in human serum may drive positive selection for higher GCH1 copy numbers and hold extra benefits by contributing to both folate and pyrimidine biosynthesis (**Figure 2.5**). Of note, our observation of some level of variation in standard parasite lines (**Supplementary Table 2.4**) suggests that the growth environment could drive changes at this locus.

Another factor that could contribute to GCH1 evolution is genetic background. The Dd2 parasite line (and its parent line, W2) were isolated in Southeast Asia (Oduola et al., 1988; Wellems et al., 1990), where antifolates were widely used. Consequently, W2 and Dd2 carry 5 mutations in folate biosynthesis enzymes (DHPS and DHFR). The 3D7 line originated in Africa (Walliker et al., 1987) and is wild type at these loci. In our studies, we observed a positive association of copy number between DHODH and GCH1 in Dd2 versus 3D7 parasite lines (**Figure 2.4D and 2.4E**), indicating that mutant backgrounds may rely more heavily on the GCH1 CNV to alleviate fitness effects for resistance to both antifolates and DHODH inhibitors.

Importantly, the DSM265 antimalarial is currently being evaluated for the treatment of clinical malaria (Mandt et al., 2019). Antifolate resistance and GCH1 CNVs are widespread in clinical isolates (Kidgell et al., 2006; Ribacke et al., 2007; Nair et al., 2008; Osei et al., 2018), thus, necessitating further evaluation of the contribution of GCH1 in parasites that are resistant to DHODH inhibitors.

3 CHAPTER III: Copy number variation detection via long read sequencing reveals the dynamics of adaptation in malaria parasites

Manuscript authors: Shiwei Liu[1], Emily R. Ebel[2], Mary Lewis Simpson[1], Aleksander Luniewski[3], Nnenna Ene[1], Elizabeth S. Egan[2], Jennifer L. Guler[1*]

[1]University of Virginia, Department of Biology, Charlottesville, VA, USA

[2]Stanford University, Departments of Pediatrics and Microbiology & Immunology, Stanford, CA, USA

[3] Pomeranian Medical University, Department of Medicine, Szczecin, Poland

Statement of contribution and acknowledgement

Shiwei Liu and Jennifer Guler designed the study and Shiwei Liu wrote the chapter. Shiwei Liu, Mary Lewis Simpson, Jane Kim and Aleksander Luniewski performed the experiments and data analysis. Emily Ebel (Stanford University) developed the Shiny application and provided support for using the application.

Our thanks to Michelle Warthan (University of Virginia) for experimental support, Martin Wu (University of Virginia) for Oxford Nanopore sequencing support, and the laboratories of Dr. William Petri Jr (University of Virginia) for their helpful discussions and insight.

3.1 Abstract

Genetic diversity is an important driver for adaptation to environmental changes in malaria parasites. Changes in copy number of genomic segments, termed copy number variations, are a type of adaptive structural variations in malaria parasites. Numerous copy number variations have been identified in parasites from clinical infections and laboratory selections. They are often directly implicated in drug resistance or parasite fitness. We do not know whether random spontaneous copy number variations arise constantly across the genome or are stimulated when parasites are under stress. Regardless of the mechanism, de novo copy number variations in minor parasite populations allow quick adaptation during environmental changes but remain undetected by traditional detection methods. In this study, we used single molecule Nanopore sequencing of laboratory cultured parasite lines with or without stress treatment to understand more about the frequency of de novo structural variations encompassing genes across various chromosomes. We suggest that stress may induce structural changes including amplifications (increased copies of genes), inverted amplification and breakpoints of unknown structural variations, underscoring the importance of copy number variations in driving the emergence of genetic heterogeneity, but further experiments are needed to confirm this result. Importantly, this is the first direct observation of de novo copy number variations in a malaria parasite population and enhanced our understanding of genome evolution through copy number variations, and ultimately inspires new strategies to block the development of antimalarial resistance in this deadly organism.

3.2 Introduction

Importance of genetic diversity within a cell population

A cell population with a great diversity is likely to be able to adapt and survive under environment changes. Cell diversity could come from genetic variations and phenotypical variations through various molecular mechanisms such as stochasticity in molecular processes, asymmetrical cell division, cell-cell interaction, epigenetic modification responses in microbial organisms like bacteria, yeast, and malaria parasites (Carey et al., 2018). In malaria parasites, epigenetic variations, direct transcriptional changes, and genetic variations are indicated to be involved in the response of parasites to environmental changes (Jiang et al., 2010; Cortés et al., 2012; Manske et al., 2012; Kafsack et al., 2014; Miles et al., 2016; Deitsch and Dzikowski, 2017). Epigenetic variations and directed transcriptional changes are more dynamic in response to environmental changes in the malaria parasites, while dogma dictates that genetic variations are accumulated over a long period of time and enable natural selection of parasites under environmental changes (Llorà-Batlle et al., 2019). However, the dynamics of genetic variations in a changing environment in parasites could be underestimated. One reason for this is that the majority of *P. falciparum* genetic variations have been identified by analyzing bulk DNA using traditional methods like PCR, microarray, targeted short reads sequencing, and whole genome short reads sequencing, where the detected genetic variations are present in the majority of parasites in the population (Nair et al., 2010; Gadalla Nahla B. et al., 2011; Tan et al., 2011; Miles et al., 2016; Silva et al., 2018; Duanguppama et al., 2019). Thus, we need new tools to detect heterogeneous genetic variations to better understand the adaptation of malaria parasites.

Copy number variations in malaria parasites

Copy number variations are known to drive rapid adaptation in response to stress and changes in the environment in many organisms, such as yeast, and bacteria (Riesenfeld et al., 1997;

Hastings et al., 2000; Rosenberg, 2001; Cirz et al., 2005; Gresham et al., 2010; Shor et al., 2013; Hull et al., 2017). Amplifications, or increases in the copy number of a genomic segment, that can increase the expression of target proteins or drug efflux pumps by altering gene dosage are known to be associated with antimalarial resistance in both laboratory and field malaria parasites (Foote et al., 1989; Wilson et al., 1993; Kidgell et al., 2006; Conway, 2007; Ribacke et al., 2007; Guler et al., 2013; Heinberg and Kirkman, 2015; Cheeseman et al., 2016; Menard and Dondorp, 2017). The formation of copy number variations is likely an essential step during the parasite's adaptation to countless antimalarials. Besides altering the gene dosage, the creation of redundant gene copies through copy number variations can also facilitate the accumulation of SNPs (Lynch and Conery, 2000; Kondrashov et al., 2002; Kondrashov, 2012; Guler et al., 2013)**.** This is due to the extra gene copies that may release mutational constraints. More specifically, point mutations can accumulate in some of the gene copies, while keeping a gene copy with normal function can reduce the negative selection of antimalarial-resistance-conferring mutations in normal conditions (Lynch and Conery, 2000; Kondrashov et al., 2002; Kondrashov, 2012; Guler et al., 2013). Studies observing co-occurrence of both types of mutations in *Plasmodium* provide evidence that copy number variations appear to eventually be lost in favor of SNPs (Thaithong et al., 2001; Rottmann et al., 2010; Phillips et al., 2015). In addition, previous studies found repeat sequences and secondary structures across the parasite's genome are likely involved in the generation of copy number variations (Huckaby et al., 2018). Copy number variations can be generated during the DNA repair process of double-strand break (Lee Andrew H. et al., 2014). Environmental changes, such as antimalarial stress or DNA synthesis inhibitor (Aphidicolin), can lead to increased DNA damage, thus potentially stimulate the generation of copy number variations (Inselburg and Banyal, 1984; Rothkamm et al., 2003). With the repetitive and extremely AT-rich genome of *P. falciparum* parasites, copy number variations are likely the major genetic drivers for parasite adaptation and studying the dynamics of copy number variations in malaria parasites is critical to understand the evolution of parasite adaptation.

The necessity of Nanopore sequencing for detecting heterogeneous copy number variations

Our group has developed a single cell short reads sequencing pipeline for detecting heterogenous copy number variations (Liu et al., 2021). However, the whole genome amplification from single parasite DNA introduces background noise for copy number variation detection. Currently, there is no tool to validate copy number variations detected from single cell samples. Thus, in this study, we explored the use of single molecule Nanopore sequencing to detect heterogeneous copy number variations. The genome of *P. falciparum* has extremely high AT content (80.6%) and contains many repetitive regions (Gardner et al., 2002). As a result, many regions of the parasite genome show unusually abundant short tandem repeats and other low complexity sequences in both coding and non-coding regions (Gardner et al., 2002). The detection of copy number variation using short reads sequencing has been challenging. It is difficult to access repetitive regions where short reads cannot map uniquely and detect complex structural variations (Treangen and Salzberg, 2012). Instead, long reads from third generation sequencing can span these low complexity and repetitive regions, and enable the detection of structural variants (e.g., segmental duplications) (Ho et al., 2020). Especially, with Nanopore single molecule sequencing, as every read is originated from a single DNA fragment, researchers can start to examine the clonal heterogeneity of pathogens via copy number variation detection (Greninger Alexander L. et al., 2018; Li et al., 2018; Girgis et al., 2021). For example, Nanopore sequencing has been used in bacteria to investigate heterogeneity of specific antibiotic resistance genes (Girgis et al., 2021).

With the advance in third-generation sequencing technologies, we can detect copy number variations at single molecule resolution. The Nanopore sequencing can produce long reads (largest reads up to ~1Mb in this study), which can span entire copy number variation repeats. Because the DNA library preparation of Nanopore sequencing used in this study is PCR-free, the difference in copy numbers of genes in individual Nanopore reads will reflect the copy number variations that exist in the cell population in which the DNA is sequenced (Greninger Alexander L. et al., 2018; Girgis et al., 2021). In this study, we performed Nanopore sequencing in 4 laboratory parasite lines to test whether there are any heterogeneous copy number variations in these parasites. We also tested whether Aphidicolin (double-strand break inducer) can increase the rate of copy number variations and whether sublethal DSM-1 (an antimalarial drug in clinical development) can increase the rate of copy number variations. Through single molecule analysis using Nanopore reads, we can detect the basal rate of copy number variations and understand whether stress can affect copy number variation formation in the parasite's genome. Through the detection of heterogeneous copy number variations by Nanopore sequencing in malaria parasites, we enhanced our understanding of the level of genetic diversity in malaria parasites and how parasites respond to antimalarial stress. Such progress can help better understand genome evolution through copy number variations and ultimately help develop new strategies to block the development of resistance in this deadly organism.

3.3 Materials and Methods

Parasite culture

We thawed erythrocytic stages of *P. falciparum* (Dd2, MRA-150 (Dd2-A sample: cultured during 2020 and Dd2-B sample: cultured during 2022); 3D7, MRA-102 (3D7 sample: cultured during 2020); HB3, MRA-155 (HB3 sample: cultured during 2020)), Malaria Research and Reference Reagent Resource Center, BEI Resources and DSM1 resistant clone H2 as mentioned in Chapter II) from frozen stocks and maintained them as previously described (Haynes et al., 1976). Briefly, we grew parasites at 37 °C in vitro at 3% hematocrit (serotype A positive human erythrocytes, Valley Biomedical, Winchester, VA) in RPMI 1640 medium (Invitrogen, USA) containing 24 mM NaHCO3 and 25 mM HEPES and supplemented with 0.5% AlbuMAX (TM) II Lipid-Rich Media (Gibco, USA) and 36.73 µM Hypoxanthine (Sigma-Aldrich, USA) in sterile, sealed flasks flushed with 5% O2, 5% CO2, and 90% N2 (Guler et al., 2013). We maintained the cultures with media changes every other day and sub-cultured them as necessary to keep parasitemia below 5%. We determined all parasitemia measurements using SYBR green-based flow cytometry (Bei et al., 2010). To check the parasite culture for Mycoplasma contamination, we routinely tested cultures using the LookOut Mycoplasma PCR Detection Kit (Sigma-Aldrich, USA) to confirm negative Mycoplasma status.

Parasite stress treatment

To test whether antimalarial drug stress can affect the formation of de novo copy number variations, we cultured Dd2-B parasites as mentioned in the **Parasite culture** in the **Materials and Methods** (during June 2022) and split the cultured parasites into two samples. One sample was treated with 1µM DSM-1 (10x EC50) dissolved in 10µl Dimethyl sulfoxide (DMSO, 0.1%)

for 12 hours and the other sample was treated with 10µl DMSO (0.1%) for 12 hours. Parasites from the two samples were washed in 1X PBS (diluted from 10× PBS Liquid Concentrate, Gibco, USA), then re-cultured in Albumax-based media for 20 hours. DSM1 is a triazolopyrimidine antimalarial that specifically and potently inhibits the *P. falciparum* dihydroorotate dehydrogenase enzyme of pyrimidine biosynthesis (Phillips et al., 2008). A previous study used DSM1 for drug selection and observed resistant parasites after weeks of continuous culture under drug selection (0.3uM of DSM-1 for 17-88 days; 1-10uM of DSM1 for 7-26 days) (Guler et al., 2013). Dihydroorotate dehydrogenase (DHODH) gene amplification was identified as the copy number variation conferring resistance (Guler et al., 2013). In this study, since DSM-1 is a slow-acting drug and requires at least 48 hours to show a lethal effect on the parasites (Sanz et al., 2012), we do not expect the time of treatment and re-culturing used is long enough to cause rapid reduction in parasitemia by killing the parasites. After 20 hours of re-culturing, the parasites were harvested at 1.9% (Dd2-B DSM1) and 4.9% (Dd2-B DMSO-1) parasitemia with 18.5% (Dd2-B DSM1) and 72.4% (Dd2-B DMSO-1) of early-stage parasites (**Table 3.1**).

Aphidicolin is a drug inhibiting DNA synthesis and inducing double strand break, thus has been used to reversibly block DNA synthesis and maturation of trophozoite and schizonts during the asexual life cycle (Inselburg and Banyal, 1984). To test whether Aphidicolin (24x EC50) can affect the formation of de novo copy number variations, we treated parasites with 4.4µM Aphidicolin; this compound has been used for parasite synchronization (Inselburg and Banyal, 1984) and we previously confirmed that this concentration pauses replication (data not shown). Thus, this sublethal treatment is not likely to directly kill the parasites but instead cause replicative stress on the parasites. We cultured Dd2-B parasites as mentioned in the **Parasite culture** in the **Materials and Methods** (during July 2022), and split the cultured parasites into two samples, one sample was treated with 4.4µM Aphidicolin dissolved in 15µl DMSO (0.15%) for 12 hours and the other sample was treated with 15µl DMSO (0.15%) for 12 hours alone. Then we washed both samples in 1X PBS, and recultured the parasites in Albumax-based media. After 14 hours of reculturing, we measured the parasitemia percentage and early-stage parasite percentage for direct comparison between the two samples (**Table 3.1**). After 18 hours of reculturing, the Aphidicolin-treated parasites (Dd2-B Aphidicolin) were harvested with 5.1% parasitemia, 62.1% of early-stage parasites and 80.2% mitochondria membrane potential. After 40 hours of reculturing (longer reculturing time was used to obtain more late-stage parasites than DSM1-DMSO sample), the DMSO treated parasites (Dd2-B DMSO-2) were harvested with 3.5% parasitemia and 35.5% of early-stage parasites, and 93.1% mitochondria membrane potential.

Table 3.1 Parasitemia and early-stage parasite percentage before and after stress treatment

Samples	Pre-treatment		Post-12hr-treatment		Reculturing for 14 hrs		At harvest*	
	Parasitemia %	Early stage %	Parasitemia %	Early stage %	Parasitemia %	Early stage %	Parasitemia %	Early stage %
Dd2-B DMSO-1	1.4	96.3	1.4	51.0	NA	NA	4.9	72.4
Dd2-B DSM1	1.6	94.0	1.6	75.8	NA	NA	1.9	18.5
Dd2-B DMSO-2	1.1	93.0	1.2	51.7	2.2	62.0	3.5	35.5
Dd2-B Aphidicolin	1.1	93.0	1.1	49.3	0.9	15.7	5.1	62.1

NA: not measured
*Dd2-B DMSO-1 and Dd2-B DSM1: harvested after 20 hours of reculturing
*Dd2-B DMSO-2: harvested after 40 hours of reculturing, Dd2-B Aphidicolin: harvested after 18 hours of reculturing

DNA extraction for long read sequencing

We lysed asynchronous *P. falciparum*-infected erythrocytes with 0.15% saponin (Sigma-Aldrich, USA) for 5 min at room temperature and washed them three times with 1× PBS (diluted from 10× PBS Liquid Concentrate, Gibco, USA). We then lysed parasites with 0.1% Sarkosyl Solution (Bioworld, bioPLUS, USA) in the presence of 1 mg/ml proteinase K (from Tritirachium album, Sigma-Aldrich, USA) overnight at 37 °C. We first extracted nucleic acids with phenol/chloroform/isoamyl alcohol (25:24:1) pH 8.0 (Sigma-Aldrich, USA) three times using 1.5 ml light Phase lock Gels (5Prime, USA), then further extracted nucleic acids with chloroform twice using 1.5 ml light Phase lock Gels (5Prime, USA). Lastly, we precipitated the DNA with ethanol using the standard Maniatis method (Maniatis et al., 1989). To obtain high molecular weight genomic DNA, we avoided any pipetting during the extraction, transferred solutions by directly pouring it from one tube to another, and mixed solutions by gently inverting the tubes.

Oxford Nanopore long read sequencing and analysis

For Dd2-A, 3D7, HB3, H2 parasite DNA, we subjected 1 μg of high molecular weight genomic DNA from each sample to library preparation for Oxford Nanopore sequencing following the Nanopore Native barcoding genomic DNA protocol (version: NBE_9065_v109_revAB_14Aug2019) with 1x Ligation Sequencing kit (SQK-LSK109, Oxford Nanopore Technologies, Oxford, UK). We performed DNA repair and end preparation using NEBNext FFPE DNA Repair Mix (New England Biolabs, Ipswich, MA, USA) and NEBNext End Repair/dA-Tailing Module (New England Biolabs, Ipswich, MA, USA). We cleaned the A-tailed fragments using 0.9X AMPure XP beads (Beckman Coulter, High Wycombe, UK). We then ligated barcodes to the end-prepped DNA using the Native Barcoding Expansion 1–12 kit

(EXP-NBD104, Oxford Nanopore Technologies, Oxford, UK) and Blunt/TA Ligase Master Mix (New England Biolabs, Ipswich, MA, USA). We cleaned the barcoded samples using 0.9X AMPure XP beads. Then we pooled barcoded samples in equimolar ratios and subjected to an adaptor ligation step, using the Adapter Mix II from the Native Barcoding Expansion 1–12 kit and NEBNext Quick Ligation Reaction Buffer (New England Biolabs, Ipswich, MA, USA) as well as Quick T4 DNA Ligase (New England Biolabs, Ipswich, MA, USA). After adaptor ligation, we cleaned the library using AMPure XP beads. We quantified the adapter ligated and barcoded DNA using a Qubit fluorimeter (Qubit 1X dsDNA High Sensitivity Assay Kit, Life Technologies, Carlsbad, CA). We sequenced the Dd2-A, 3D7, HB3 and initial H2 libraries using a R9.4.1 flow cell (FLO-MIN106D, Oxford Nanopore Technologies, Oxford, UK) on MinION (Oxford Nanopore Technologies, Oxford, UK) and another H2 library using the R10 flow cell (FLO-MIN111) on MinION. To obtain the maximum number of reads, we ran both flow cells for 48 h (controlled and monitored using the MinKNOW software (3.6.5)).

For Dd2-B parasites treated with DMSO or DSM-1, Aphidicolin, we sequenced the four samples with Nanopore ultra-long sequencing kit (SQK-ULK001, Oxford Nanopore Technologies, Oxford, UK) to obtain long reads. We subjected ~30 µg of high molecular weight genomic DNA from Dd2-B samples to library preparation using the Nanobind Library Prep-Ultra Long Sequencing protocol (LBP-ULN-001, PacBio, USA). We fragmented the DNA using FRA Dilution Buffer (FDB) and Fragmentation Mix (FRA) and added adapter using Rapid Adapter F in the Nanopore ultra-long sequencing kit. The library was then cleaned with a Nanobind disk from the Nanobind UL Library Prep Kit (NB-900-601-01, PacBio, USA). Gentle pipetting with wide-bore tips is applied to minimize DNA fragmentation. Each Dd2-B parasite library is separated into three loadings and sequenced using a R9.4.1 flow cell (FLO-MIN106D, Oxford Nanopore Technologies, Oxford, UK) on MinION (Oxford Nanopore Technologies, Oxford, UK) for 72 h (24h + 24h +24h).

For base calling and demultiplexing of the Nanopore sequencing reads, we used Guppy (version 3.4.5+fb1fbfb) with the parameter settings “-c dna_r9.4.1_450bps_hac.cfg --barcode_kits "EXP-NBD104" -x auto” for samples (Dd2-A, 3D7, HB3, H2) sequenced with R9.4.1 flow cell and “-c dna_r10_450bps_hac.cfg -x auto” for the sample (H2) sequenced with R10 flow cell. The sequenced data of H2 on R9.4.1 and R10 flow cell was then pooled for downstream analysis. For Dd2-B samples, we did base calling with Guppy (version 3.4.5+fb1fbfb) using the parameter settings “-c dna_r9.4.1_450bps_hac.cfg -x auto” and sequenced with R9.4.1 flow cell. We checked the read length and read quality using “Nanoplot” (version 1.0.0). We trimmed the adapters with “qcat” (version 1.1.0) (Oxford Nanopore Technologies) and filtered the reads with a cutoff “length ≥ 500 and Phred value ≥ 10” using the program “filtlong version 0.2.0” (https://github.com/rrwick/Filtlong). To estimate the coverage of sequencing reads in each sample, we aligned the filtered reads to *Plasmodium falciparum* 3D7 reference genome using “minimap2” (version 2.17) (Li, 2018). “QualiMap” (version 2.2.1) (García-Alcalde et al., 2012) was used to calculate the coverage of the aligned reads.

Shiny analysis of long reads

To visualize structural variants in the parasite genome, we used a custom R Shiny script to plot the arrangement of reference gene segments on individual Nanopore reads. Sub-telomere and telomere regions with hypervariable genes were excluded from the analysis (Otto et al., 2018).

Briefly, we extracted the reference sequences from 3D7 reference genome of genes and subsequent intergenic regions cross the whole genome, then split these sequences into fragments of 500-1000 base pairs (blocks). We compared these fragments to individual Nanopore reads using BLAST (Ye et al., 2006). We used the BLAST output as input for a custom R script, which drew rectangles representing homology between the defined genes (y-axis) and every read (x-axis). The percent identity required to draw a homologous rectangle was allowed to vary between reads, which varied in quality, using a slider in the Shiny app. To filter out long reads with potentially spurious hits to gene fragments, we also used BLAST to compare Nanopore reads to the reference genome and removed reads with <90% identity to the reference genome. To obtain reads that are likely showing structural variants, we filtered the blasted reads by the detection of repeated gene sequences or the detection of breakpoints on the reads. Specifically, for reads with repeated gene sequences, we kept reads with at least one gene block blasted to two or more different positions on a Nanopore read. To filter reads with breakpoints for structural variants, we ordered the hit results by gene IDs and kept reads showing direction changes of gene ID order along the reads. When there are no breakpoints, each read should only have either increasing gene IDs or decreasing gene IDs across the read; when there is a change in the direction, the gene ID order will switch from increasing to decreasing or vice versa. We determined whether a read spanning a copy number variation by checking the presence of genes outside of the amplicon, if the genes outside of the amplicon were present, we are able to determine the exact number of copies of this variant (as indicated by “Span: Yes”), while if the genes outside of the amplicon were not present in both ends, we do not know the copy number of this variant (as indicated by “Span: No”).

3.4 Results

Improvement of Nanopore sequencing read length and genome coverage

To assess the feasibility of using Nanopore sequencing to generate long reads to span copy number variations in the malaria parasite genome, we sequenced 4 laboratory parasite lines including Dd2 (Dd2-A sample), HB3, H2, 3D7 using a PCR-free library preparation kit (Ligation Sequencing Kit) and AMPure XP beads during library preparation (see **Materials and Methods**). Later, with the advance in Nanopore sequencing technology, we also sequenced another set of Dd2 parasites (Dd2-B) with or without DSM-1 or aphidicolin treatment using the Nanopore Ultra Long Sequencing Kit and Nanobind disk (see **Materials and Methods**) to obtain longer reads (**Table 3.2**).

Table 3.2 Sequencing samples and read depth

Sample	Library preparation kit	DNA purification method	Total reads (>500bp, q10)	Read length N50	Mean read length	Largest read length	Nucleotide identity	Genome coverage
Dd2-A	Ligation sequencing kit (SQK-LSK109)	AMPure XP	150,489	15,574 bp	10,902 bp	190,160 bp	93%	29 X
HB3			96,057	18,893 bp	12,480 bp	251,827 bp	93%	23 X
H2			162,539	11,654 bp	8,189 bp	232,569 bp	93%	22 X
3D7			111,074	17,587 bp	12,126 bp	256,681 bp	93%	23 X
Average			-	15,927 bp	10,924 bp	190,160 bp	93%	24 X
Dd2-B DMSO-1	Ultra-Long DNA Sequencing Kit (SQK-ULK001)	Nanobind UL library prep kit	205,635	90,255 bp	39,724 bp	986,050 bp	91%	364 X
Dd2-B DSM1			151,754	79,544 bp	45,712 bp	628,852 bp	92%	309 X
Dd2-B DMSO-2			246,403	71,194 bp	32,673 bp	749,269 bp	92%	358 X
Dd2-B Aphidicolin			388,784	32,995 bp	13,228 bp	320,993 bp	93%	226 X
Average			-	68,497 bp	32,834 bp	671,291 bp	92%	314 X

As shown in Table 3.2, Dd2-A resulted in 150,489, reads containing 1,640,716,826 bases, while Dd2-B sample resulted in 205,635 reads with 8,168,694,985 bases after filtering reads by size (reads > 500bp) and quality (Albacore-generated score of q10). On average, the mean read length of Dd2-A, HB3, H2, 3D7 samples (10,924 bp) is shorter than that of all the Dd2-B samples (32,834 bp). The average N50 of reads increased from 15,927 bp in Dd2 (Dd2-A), HB3, H2, 3D7 samples to 68,497 bp in all the Dd2-B samples. Notably, all 4 Dd2-B samples (average: 671,291 bp) show a much longer maximum read length than the Dd2-A, HB3, H2, 3D7 samples (average: 190,160 bp). We measured the Nanopore read accuracy by alignment to *Plasmodium* Dd2 reference genome. Reads from each sample were mapped to the reference genome and showed on average similar nucleotide identity across both sample sets (93% nucleotide identity for Dd2-A, HB3, H2, 3D7 samples and 92% for Dd2-B samples, **Table 3.2**). Mapping all filtered reads to the Dd2 reference genome resulted in on average 24X (Dd2-A, HB3, H2, 3D7 samples) and 314X (Dd2-B samples) coverage (**Table 3.2**).

Detection of known copy number variations in 4 laboratory cultured parasite lines

To detect known copy number variations across the genome of *Plasmodium* parasites, we visualized how gene blocks are represented on individual reads using Shiny app (see **Materials and Methods**). In the 4 laboratory cultured parasite lines (Dd2-A, HB3, H2, 3D7 samples), we detected copy number variations that are previously identified and validated in other studies (Miles et al., 2016). As shown in **Table 3.3**, we identified reads that span the whole GTP cyclohydrolase I (GCH1) copy number variation in Dd2 (**Figure 3.1A, Table 3.3**), 3D7 (**Figure 3.1B, Table 3.3**). Since the size of the GCH1 copy number variation in HB3 is much larger (161kb), we only identified reads that cover the breakpoint of the GCH1 copy number variation

(**Figure 3.1C, Table 3.3**), but the breakpoint genes match the ones found in other studies, indicating the size of copy number variation is the same as that of other studies (Miles et al., 2016). We also identified reads covering the breakpoint of multidrug resistant protein 1 (MDR1) amplification with similar copy number variation size as previous studies (**Figure 3.1D, Table 3.3**). Lastly, we also identified reads covering the breakpoint of DHODH amplification with similar size of copy number variation as a previous study (Guler et al., 2013) (**Figure 3.1E, Table 3.3**). Thus, the Nanopore long read visualization analysis can accurately detect the whole structures of small copy number variations (i.e., copy number variations with 2kb, 5kb repeat sequences) and the breakpoints of large copy number variations (i.e., copy number variations with 82kb, 74kb, 161kb repeat sequences).

Table 3.3 Known copy number variations detected by Nanopore read Shiny App analysis

Lines	Copy number variation	Chromosome	Start (Gene ID)	End (Gene ID)	Repeat size (kb)	Span	Copies	Number of reads	Proportion
HB3	*GCH1*	12	PF3D7_1223400	PF3D7_1227200	161	No	>=1	28	100.0%
3D7	*GCH1*	12	PF3D7_1224000	PF3D7_1224000	2	Yes	4	8	47.1%
						No	>=1	3	17.6%
						No	>=2	2	11.8%
						No	>=3	2	11.8%
						No	>=4	2	11.8%
Dd2-A	*GCH1*	12	PF3D7_1223800	PF3D7_1224000	5	Yes	3	10	52.6%
						No	>=1	2	10.5%
						No	>=2	5	26.3%
						No	>=3	2	10.5%
	MDR1	5	PF3D7_0521900	PF3D7_0523200	82	No	>=2	33	94.3%
						No	>=3	2	5.7%
H2	*GCH1*	12	PF3D7_1223800	PF3D7_1224000	5	Yes	5	1	5.9%
						Yes	7	1	5.9%
						No	>=1	7	41.2%
						No	>=2	1	5.9%
						No	>=4	3	17.6%
						No	>=5	2	11.8%
						No	>=6	2	11.8%
	MDR1	5	PF3D7_0521900	PF3D7_0523200	82	No	>=2	10	100.0%
	DHODH	6	PF3D7_0601900	PF3D7_0603700	74	No	>=2	69	95.8%
						No	>=3	3	4.2%

*Span (No) indicates the Nanopore read only detected the breakpoint of the copy number variation and is not long enough to span the whole copy number variation

*Span (Yes) indicates the Nanopore read is long enough to cover the whole copy number variation reference

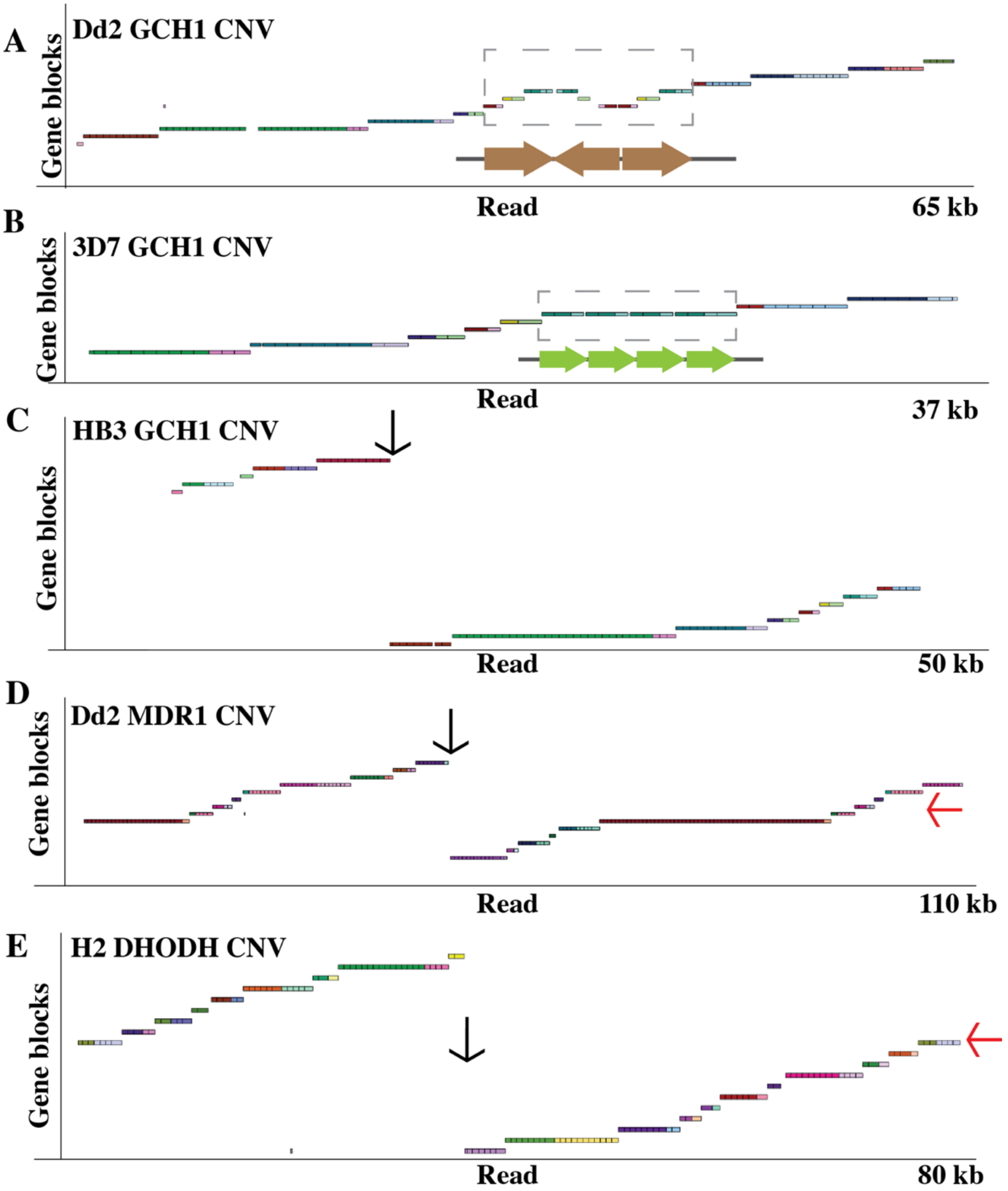


Figure 3.1 Detection of known copy number variations in 4 laboratory cultured parasite lines

(**A**) GCH1 amplification in Dd2-A spanned by a Nanopore read. The brown arrow indicates the orientation of the copy number variation repeats (each repeat includes 3 genes). (**B**) GCH1 amplification in 3D7 spanned by a Nanopore read. The green arrow indicates the orientation of the copy number variation repeats (each repeat includes 1 gene). (**C**) The breakpoint of GCH1 amplification in HB3 covered by a Nanopore read. The black arrow points to the breakpoint. (**D**) The breakpoint of MDR1 amplification in Dd2-A covered by a Nanopore read. The black arrow points to the breakpoint. (**E**) The breakpoint of DHODH amplification in H2 covered by a Nanopore read. The black arrow points to the breakpoint. The red arrow points to the genes with increased copy number.

Detection of de novo copy number variations in 4 laboratory cultured parasite lines

Since Dd2-A, HB3, H2, 3D7 parasites were originally cloned in the laboratory culture, we expect most of the cells in each line show similar known copy number variation profiles but also have some rare copy number variations that were previously not detected in the core genome. To detect de novo copy number variations across the genome of *Plasmodium* parasites, we visualized how gene blocks are represented on individual reads using Shiny app (see **Materials and Methods**). Overall, we observed 182 (Dd2-A), 37 (HB3), 49 (H2) and 90 (3D7) pyramid-like structures of genes (symmetrical inverted amplifications) in sequenced reads. The second repeat of the reads often showed lower base calling accuracy than the first repeat (**Figure 3.2A**), which was also observed in a previous study (Spealman et al., 2020) and indicated that they are caused by the secondary structures formed by inverted duplications interfering with the translocation of DNA through the sequencing pore. In this study, the pyramid structures (symmetrical inverted amplifications) were observed across all 14 chromosomes (**Table 3.4**). Their ubiquitous nature is supported by their identification in Nanopore sequencing datasets from other labs (personal communication, Emily Ebel). Most copy number variations previously detected in *P. falciparum* parasites are head-to-tail tandem amplifications (Huckaby et al., 2018). Such structures could be transient complex DNA structure generated during DNA repair of double-strand break or stalled DNA replication (see **Discussion**).

Table 3.4 Count of pyramid structures across chromosomes in all samples

Sample	Chr* 1	Chr 2	Chr 3	Chr 4	Chr 5	Chr 6	Chr 7	Chr 8	Chr 9	Chr 10	Chr 11	Chr 12	Chr 13	Chr 14	Total	Proportion#
Dd2-A	2	5	3	8	14	9	15	11	23	6	11	9	14	36	166	0.42%
HB3	3	1	2	5	3	4	7	2	1	3	0	0	2	2	35	0.15%
C710	3	4	0	2	2	3	0	1	6	4	5	4	8	7	49	0.27%
3D7	3	1	5	7	4	6	4	5	3	8	4	12	13	15	90	0.31%
Dd2-B DMSO-1	0	5	2	2	1	3	4	6	2	5	3	1	6	10	50	0.04%
Dd2-B DSM-1	7	10	12	12	14	22	19	15	20	21	26	29	35	49	291	0.35%
Dd2-B DMSO-2	13	20	27	19	22	25	20	28	21	44	51	43	61	57	451	0.37%
Dd2-B Aphidicolin	13	25	40	46	67	59	54	56	63	70	75	93	129	152	942	0.56%

*Chr: Chromosome
#Proportion is calculated by dividing the total of pyramids to the total blast hit reads to 3D7 reference genome in each sample

We also observed 8 (Dd2-A, **Figure 3.3B-2D**), 3 (HB3), 3(H2) and 7 (3D7) inverted amplifications, which are slightly different from the pyramid structures (**Table 3.5**). Since these inverted amplifications often lack repeats for some genes, are not always symmetrical, and have similar base calling accuracy in both repeats, it is possible that they are true copy number variations that are present in the genome prior to sequencing (**Table 3.5**).

Table 3.5 Summary of de novo copy number variations detected in *Plasmodium* parasites by Nanopore long reads

Type of structural variations	Parasite lines			
	Dd2-A	HB3	C710	3D7
Amplification	1	0	0	0
Inverted amplification (non-symmetrical)	8	3	3	7
Breakpoints of unknown structural variation	1	1	1	1
Total	10	4	4	8

Amplifications that contribute to important malaria phenotypes typically sit in tandem, head-to-tail orientation in the genome (Guler et al., 2013; Huckaby et al., 2018). Read visualization, which can display the breakpoint between the two copies, showed 1 de novo tandem amplification event in Dd2-A sample (**Figure 3.3A, Table 3.5**). We also observed reads displaying the breakpoint of a distinct unknown structural variation in each sample (Dd2-A, HB3, H2, 3D7, **Figure 3.2B-3.2D, Figure 3.3E**). We were not able to visualize the entire regions, due to the limited read length; these unknown structural variations could potentially be inversions, deletions, or amplification events.

With sequencing coverage of 22X to 29X across this set of samples, all the detected de novo copy number variations are only supported by 1 read in each sample (see **Table 3.6**), while other reads spanning the same genes show no copy number variation. None of the de novo copy number variation locations were shared across different parasite lines. The genes contained within these regions include DNA topoisomerase, kinase, guanyl-nucleotide exchange factor, microtubule binding motor protein, ubiquitin-protein ligase, chaperone, RNA helicase, ribosomal protein, and some uncharacterized proteins, which are proteins involved in wide biological processes like DNA replication, signal transduction, biosynthesis, protein metabolism, RNA metabolism (**Table 3.6**).

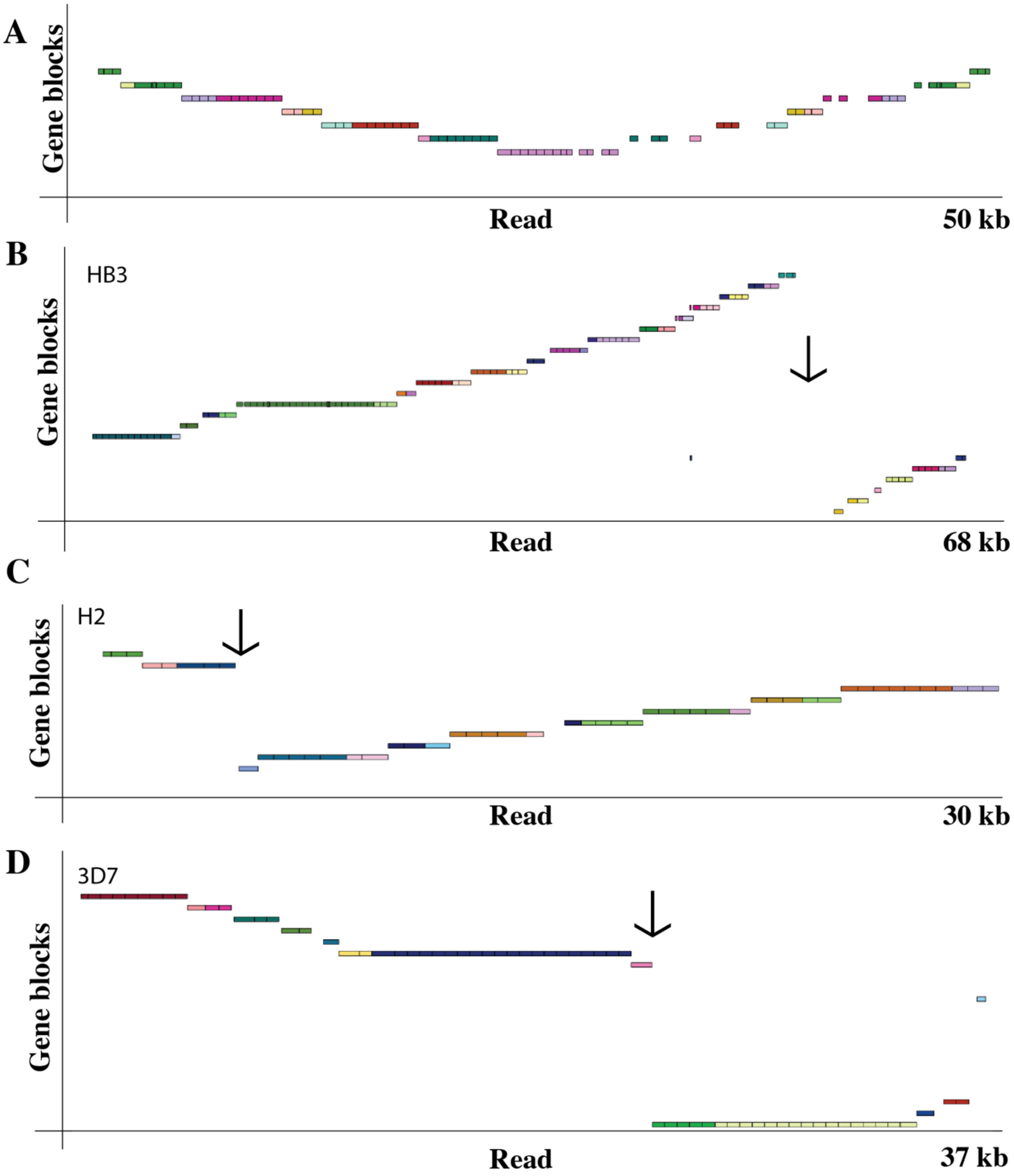


Figure 3.2 Visualization of de novo copy number variations detected in Plasmodium parasites by Nanopore long reads

(**A**) An inverted amplification showing a symmetrical pyramid-like structure detected by a Nanopore read. (**B**) The de novo structural variant detected by breakpoints in HB3 sample. The black arrow points to the breakpoint. (**C**) The de novo structural variant detected by breakpoints in H2 sample. The black arrow points to the breakpoint. (**D**) The de novo structural variant detected by breakpoints in 3D7 sample. The black arrow points to the breakpoint.

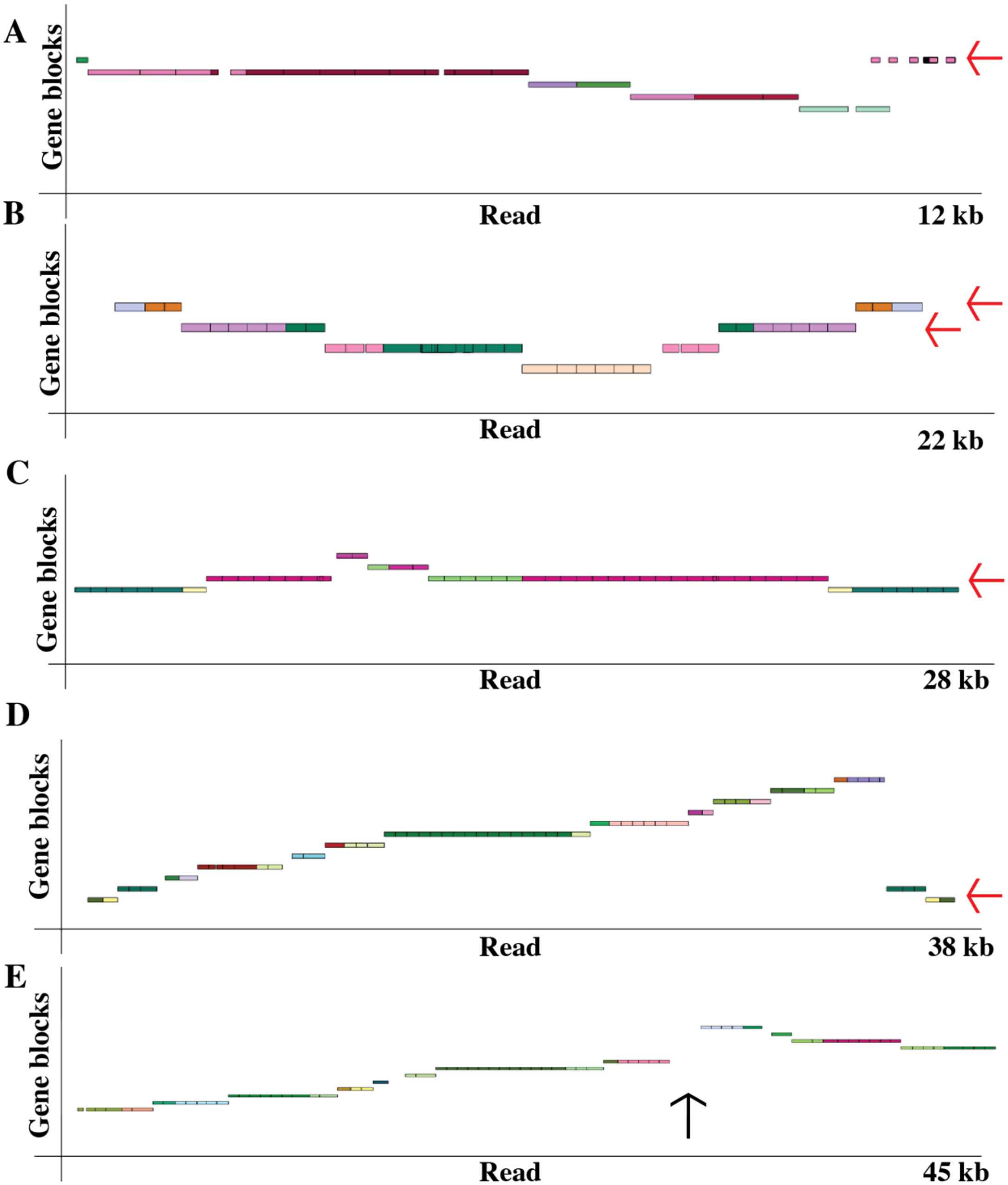


Figure 3.3 Visualization of de novo copy number variations detected in Dd2-A Plasmodium parasites by Nanopore long reads

(**A**) The de novo amplification event detected in Dd2-A sample. The red arrow points to the gene with an increased copy number. (**B**)-(**D**) De novo inverted amplification events (non-symmetrical) detected in Dd2-A sample. The red arrow points to the gene with an increased copy number. (**E**) The de novo structural variant detected by breakpoints in Dd2-A sample. The black arrow points to the breakpoint.

Table 3.6 Gene information of detected de novo copy number variations in Dd2-A, HB3, H2, 3D7 samples

Sample	Type of variants	Copy number	Chr	Start (Gene ID)	Start (Gene name)	End (Gene ID)	End (Gene name)	Number of support read
Dd2-A	Amplification	>=2	9	Pf3D7_0919900	Regulator of chromosome condensation-PP1-interacting protein	Pf3D7_0919900	Regulator of chromosome condensation-PP1-interacting protein	1
	Inverted amplification (non-symmetrical)	>=2	4	Pf3D7_0424700	Serine/threonine protein kinase, FIKK family	Pf3D7_0424900	PRESAN domain-containing protein	1
			5	Pf3D7_0510500	Topoisomerase I	Pf3D7_0510500	Topoisomerase I	1
			11	Pf3D7_1146700	Kinesin-like protein	Pf3D7_1146800	Uncharacterized protein	1
			12	Pf3D7_1221000	Histone-lysine N-methyltransferase, H3 lysine-4 specific	Pf3D7_1221000	Histone-lysine N-methyltransferase, H3 lysine-4 specific	1
			13	Pf3D7_1346400	VPS13 domain-containing protein, putative	Pf3D7_1347400	Uncharacterized protein	1
			13	Pf3D7_1361200	Uncharacterized protein MAL13P1.304	Pf3D7_1361300	Uncharacterized protein	1
			14	Pf3D7_1412400	Uncharacterized protein	Pf3D7_1412400	Uncharacterized protein	1
			14	Pf3D7_1443100	Uncharacterized protein	Pf3D7_1443100	Uncharacterized protein	1
	Breakpoint of unknown structural variation	NA	10	Pf3D7_1003500	40S ribosomal protein S20e, putative	Pf3D7_1004000	60S ribosomal protein L13, putative	1
HB3	Inverted amplification (non-symmetrical)	>=2	7	Pf3D7_0705600	RNA helicase, putative	Pf3D7_0705900	ATP synthase subunit C, putative	1
			9	Pf3D7_0926300	Protein kinase, putative	Pf3D7_0926800	Uncharacterized protein	1
			14	Pf3D7_1437900	HSP40, subfamily A	Pf3D7_1438000	Eukaryotic translation initiation factor 2A	1
	Breakpoint of unknown structural variation	NA	7	Pf3D7_0720500	Uncharacterized protein	Pf3D7_0722200	Rhoptry-associated leucine zipper-like protein 1	1
H2	Inverted amplification (non-symmetrical)	>=2	6	Pf3D7_0624800	Uncharacterized protein	Pf3D7_0624800	Uncharacterized protein	1
			9	Pf3D7_0904900	Copper-transporting ATPase	Pf3D7_0904900	Copper-transporting ATPase	1
			13	Pf3D7_1360700	E3 SUMO-protein ligase PIAS, putative	Pf3D7_1360700	E3 SUMO-protein ligase PIAS, putative	1
	Breakpoint of unknown structural variation	NA	8	Pf3D7_0811400	Uncharacterized protein	Pf3D7_0810500	Protein phosphatase PPM7, putative	1
3D7	Inverted amplification (non-symmetrical)	>=2	7	Pf3D7_0713700	Uncharacterized protein	Pf3D7_0713900	Uncharacterized protein	1
			7	Pf3D7_0718700	Uncharacterized protein	Pf3D7_0719100	ATP synthase F0 subunit a-like protein, putative	1
			10	Pf3D7_1033600	Pre-mRNA-splicing factor CEF1, putative	Pf3D7_1033600	Pre-mRNA-splicing factor CEF1, putative	1
			11	Pf3D7_1131300	Uncharacterized protein	Pf3D7_1131300	Uncharacterized protein	1
			12	Pf3D7_1220700	Uncharacterized protein	Pf3D7_1220700	Uncharacterized protein	1

			14	Pf3D7_1422300	DnaJ protein, putative	Pf3D7_1422400	Uncharacterized protein	1
			14	Pf3D7_1442900	Protein transport protein SEC7, putative	Pf3D7_1443100	Uncharacterized protein	1
	Breakpoint of unknown structural variation	NA	14	Pf3D7_1468100	Kelch domain-containing protein, putative	Pf3D7_1469500	Uncharacterized protein	1

*NA indicates we are unable to determine the copy number of variations

Stress may induce complex structure in malaria parasites

To test whether stress from antimalarial drugs (DSM-1) or double-strand break inducer (Aphidicolin) can affect the generation of copy number variations, we also compared the de novo copy number variations detected in untreated Dd2-B samples and treated Dd2-B samples. To understand the impact of the treatments on the parasites, we initially compared the parasitemia and parasite stage in treated and untreated samples to evaluate the effect of the treatment. The DSM-1 treated sample showed a similar parasitemia percentage, but a higher percentage (Dd2-B DSM-1: 75.8%; Dd2-B DMSO-1: 51%) of early-stage parasites compared to the untreated sample, indicating DSM-1 slowed the progression of parasite life cycle (**Table 3.1**). After re-culturing the parasite for 20 hours, treated parasites exhibited a lower parasitemia level (Dd2-B DSM-1: 1.9%; Dd2-B DMSO-1: 4.9%) and lower percentage of early-stage parasites (Dd2-B DSM-1: 18.5%; Dd2-B DMSO-1: 72.4%). In other words, most of the untreated parasites progressed to a new invasion round while the treated parasites mostly remained as late stages, indicating DSM-1 slowed the progression of the life cycle of the parasites. Similarly, after removing Aphidicolin treatment and 14-hour of re-culturing, treated parasites exhibited a lower parasitemia (Dd2-B Aphidicolin: 0.9%; Dd2-B DMSO-2: 2.2%) and a lower percentage of early-stage parasites (Dd2-B Aphidicolin: 15.7%; Dd2-B DMSO-2: 62%) (**Table 3.1**). Thus, Aphidicolin treatment also appeared to slow the progression of the life cycle of the parasites and limited new erythrocyte invasion.

After sequencing, we noticed the Dd2-B parasite contained two main populations of parasites with different genetic variations (point mutations and copy number variations) instead of one main population of parasites. As shown in **Table 3.7**, the untreated samples, Dd2-B DMSO-1 and Dd2-B DMSO-2 show both reference genome alleles (3D7) and alternative alleles for known resistant SNPs in Dd2 parasite line (Runtuwene et al., 2018). MDR1 and GCH1 amplification (copy number >1) are known copy number variations in the Dd2 parasite line (**Table 3.3**), however, we detected many reads containing only one copy of either MDR1 or GCH1 (**Table 3.8**). Considering both mixed alleles in many loci and loss of MDR1 and GCH1 amplification in many reads, it is likely we have a non-clonal cell line with two main sub-populations of cells for the Dd2-B sample. Such sub-populations are not detected in the Dd2-A sample (**Table 3.7 and Table 3.8**). Thus, the comparison between treated and untreated samples might not reflect a stimulation by stress but instead, the selection for the favorable sub-population.

Table 3.7 Allele frequency of Dd2-B samples with or without stress treatment

Gene	Chromosome	POS	REF (3D7)	ALT (DD2)	Dd2-B DMSO-1		Dd2-B DSM-1		Dd2-B DMSO-2		Dd2-B Aphidicolin		Dd2-A	
					REF	ALT	REF	ALT	REF	ALT	REF	ALT	REF	ALT
PfATPase6	Pf3D7_01_v3	265941	C	T	92%	6%	92%	7%	75%	24%	56%	42%	5%	95%
	Pf3D7_01_v3	266480	A	T	4%	94%	4%	94%	3%	95%	2%	97%	4%	91%
PfMRP1	Pf3D7_01_v3	465296	C	T	95%	5%	96%	4%	75%	24%	58%	42%	0%	100%
	Pf3D7_01_v3	466034	T	G	96%	3%	95%	4%	76%	23%	60%	37%	4%	93%
	Pf3D7_01_v3	467351	A	G	96%	4%	96%	4%	76%	23%	58%	42%	7%	93%
	Pf3D7_01_v3	468893	T	A	96%	4%	96%	4%	78%	22%	59%	41%	6%	94%
DHFR-TS	Pf3D7_04_v3	748239	A	T	96%	3%	95%	4%	82%	17%	61%	39%	18%	82%
	Pf3D7_04_v3	748262	T	C	96%	3%	94%	4%	79%	18%	61%	38%	5%	90%
	Pf3D7_04_v3	748410	G	A	86%	4%	93%	6%	77%	22%	58%	40%	0%	100%
PfMDR1	Pf3D7_05_v3	958145	A	T	92%	7%	94%	6%	66%	33%	63%	35%	0%	100%
	Pf3D7_05_v3	958146	A	T	97%	3%	96%	4%	78%	21%	62%	37%	**38%**	**62%**
PfCRT1	Pf3D7_07_v3	404407	G	T	86%	5%	92%	5%	76%	20%	53%	43%	0%	100%
	Pf3D7_07_v3	404836	C	G	93%	3%	93%	6%	77%	19%	52%	45%	10%	90%
	Pf3D7_07_v3	405362	A	G	94%	5%	93%	6%	79%	20%	54%	45%	0%	100%
	Pf3D7_07_v3	405600	T	C	94%	6%	92%	7%	82%	17%	55%	44%	0%	90%
	Pf3D7_07_v3	405838	G	T	93%	4%	92%	6%	77%	19%	54%	44%	5%	95%
PfDHPS	Pf3D7_08_v3	549682	C	T	89%	5%	89%	6%	68%	24%	58%	40%	4%	92%
	Pf3D7_08_v3	550212	G	T	90%	1%	94%	2%	72%	20%	62%	35%	0%	100%
Average*					93%	4%	94%	5%	76%	22%	58%	41%	6%	93%

*The averages and standard deviation were calculated by excluding one outlier loci (Pf3D7_01_v3: 266,480)

Table 3.8 Copy number frequency of MDR1 and GCH1 copy number variations in Dd2-B samples with or without stress treatment

CNVs	Span	CN*	Dd2-B DMSO-1		Dd2-B DSM-1		Dd2-B DMSO-2		Dd2-B Aphidicolin		Dd2-A	
			# of reads	Proportion	# of reads	Proportion	# of reads	Proportion	# of reads	Proportion	# of reads	Proportion
MDR 1 CNV	Yes	1	80	24.8%	48	18.9%	45	11.7%	0	0.0%	0	0.0%
	Yes	2	0	0.0%	0	0.0%	0	0.0%	1	0.3%	0	0.0%
	Yes	3	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
	No	>=1	226	70.2%	192	75.6%	272	70.6%	217	63.6%	0	0.0%
	No	>=2	13	4.0%	14	5.5%	59	15.3%	120	35.2%	33	94.3%
	No	>=3	3	0.9%	0	0.0%	9	2.3%	3	0.9%	2	5.7%
GCH1 CNV	Yes	1	256	91.4%	234	89.7%	189	66.1%	78	41.9%	0	0.0%
	Yes	2	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
	Yes	3	4	1.4%	4	1.5%	55	19.2%	42	22.6%	10	52.6%
	Yes	4	0	0.0%	0	0.0%	0	0.0%	1	0.5%	0	0.0%
	No	>=1	18	6.4%	23	8.8%	27	9.4%	32	17.2%	2	10.5%
	No	>=2	2	0.7%	0	0.0%	6	2.1%	20	10.8%	5	26.3%
	No	>=3	0	0.0%	0	0.0%	9	3.1%	13	7.0%	2	10.5%

*CN: Copy Number

Despite this unexpected finding, we continued our analysis of copy number variations in treated and untreated samples to gain some insight into potential effects and justify future experiments. Similar to the previous set of samples, we detected many pyramid-like gene structures (symmetrical inverted amplifications) across all 14 chromosomes (**Table 3.4**). They are present in both treated and untreated Dd2-B samples, but at an elevated level in the treated samples (~ 2-fold increment in DSM1 treated sample, ~ 6-fold increment in Aphidicolin treated sample). We also detected more “complex” pyramid structures in DSM-1 treated (16) and Aphidicolin (5) treated samples (**Figure 3.4A, 3.4B, 3.4C**), which were not detected in untreated samples (**Table 3.9**). The “complex” pyramid structures often include more than one inverted amplification of the same or different genes.

Despite higher sequencing coverage of 226X to 364X in the all Dd2-B samples, most of the de novo copy number variations detected are still only supported by few reads (**Table 3.10**), while other reads spanning the same genes show no copy number variations. Interestingly, we detected a tandem amplification event in Dd2-B DSM-1 sample (**Figure 3.4D**). We also detected an unknown structural variation with a different breakpoint in a nearby locus (**Figure 3.4E, Table 3.10**, colored in green), indicating that this locus might be prone to form copy number variations. We also have now detected multiple copy number variations covered by >1 reads across the different data sets indicating true rare copy number variations (**Table 3.10**, colored in blue, gray, and yellow). While we previously have detected reads that only cover breakpoints between tandem copies (**Figure 3.2B**), with the improved read length, we have now also observed evidence of a true tandem amplification event with coverage of the breakpoint as well as a duplicated sets of genes in Dd2-B DMSO-2 sample (**Figures 3.5A, 3.5B**), the breakpoint of this copy number variation is also present in Dd2-B Aphidicolin sample (**Table 3.10**, colored in yellow). In another example, we detected a tandem amplification in Dd2-B DSM-1, Dd2-B DMSO-2, Dd2-B Aphidicolin samples showing at least 2 or 3 repeats (**Figure 3.5C, Table 3.10,** colored in blue).

Overall, we detected more copy number variations in treated Dd2-B samples while untreated Dd2-B samples exhibited similar numbers **(Table 3.9)**. Specifically, we detected 27 copy number variations (4 amplification, 6 inverted amplification, 16 complex pyramid and 2 breakpoints of unknown structural variations) in Dd2-B DSM-1 sample compared to the 3 copy number variations (inverted amplification) detected in the Dd2-B DMSO-1 sample (**Table 3.9**).

Similarly, we detected 25 copy number variations (3 amplification, 8 inverted amplification, 5 complex pyramid and 9 breakpoints of unknown structural variations) in Dd2-B Aphidicolin sample compared to the 6 copy number variations (2 amplification, 4 inverted amplification) detected in the Dd2-B DMSO-2 sample (**Table 3.9)**. Interestingly, there are more reads with breakpoints with unknown structural variations in Dd2-B Aphidicolin sample (9 reads) than its paired Dd2-B DSM-1 sample (2 reads), see **Table 3.9, Figure 3.5D, 3.5E**. Finally, there are fewer reads with complex pyramid structures in Dd2-B Aphidicolin sample (5 reads) than in the other treated sample Dd2-B DSM-1 sample (16 reads), see **Table 3.9, Figure 3.5D, 3.5E**.

As in our initial study above, the de novo copy number variations are rare and heterogeneous within each sample (**Table 3.10**). The genes in these de novo copy number variations cross all 14 chromosomes in the *P. falciparum* genome, encoding a variety of proteins, such as some uncharacterized proteins, transcription factor, histone modifying enzyme, DNA metabolism protein, membrane traffic protein, cell adhesion molecule, nucleotide kinase and so on, which are involved in wide biological processes like DNA metabolism, RNA metabolism, protein metabolism, protein folding, intracellular signal transduction, transcription regulation, microtubule binding (**Table 3.10**).

Table 3.9 Summary of de novo copy number variations detected in Dd2-B samples with or without stress treatment by Nanopore long reads

Type of structural variations	**Dd2-B DMSO-1**	**Dd2-B DSM-1**	**Dd2-B DMSO-2**	**Dd2-B Aphidicolin**
Amplification	0	4	2	3
Inverted amplifications (non-symmetrical)	3	6	4	8
Complex pyramid	0	16	0	5
Breakpoints of unknown structural variations	0	2	0	9
Total	3	27	6	25

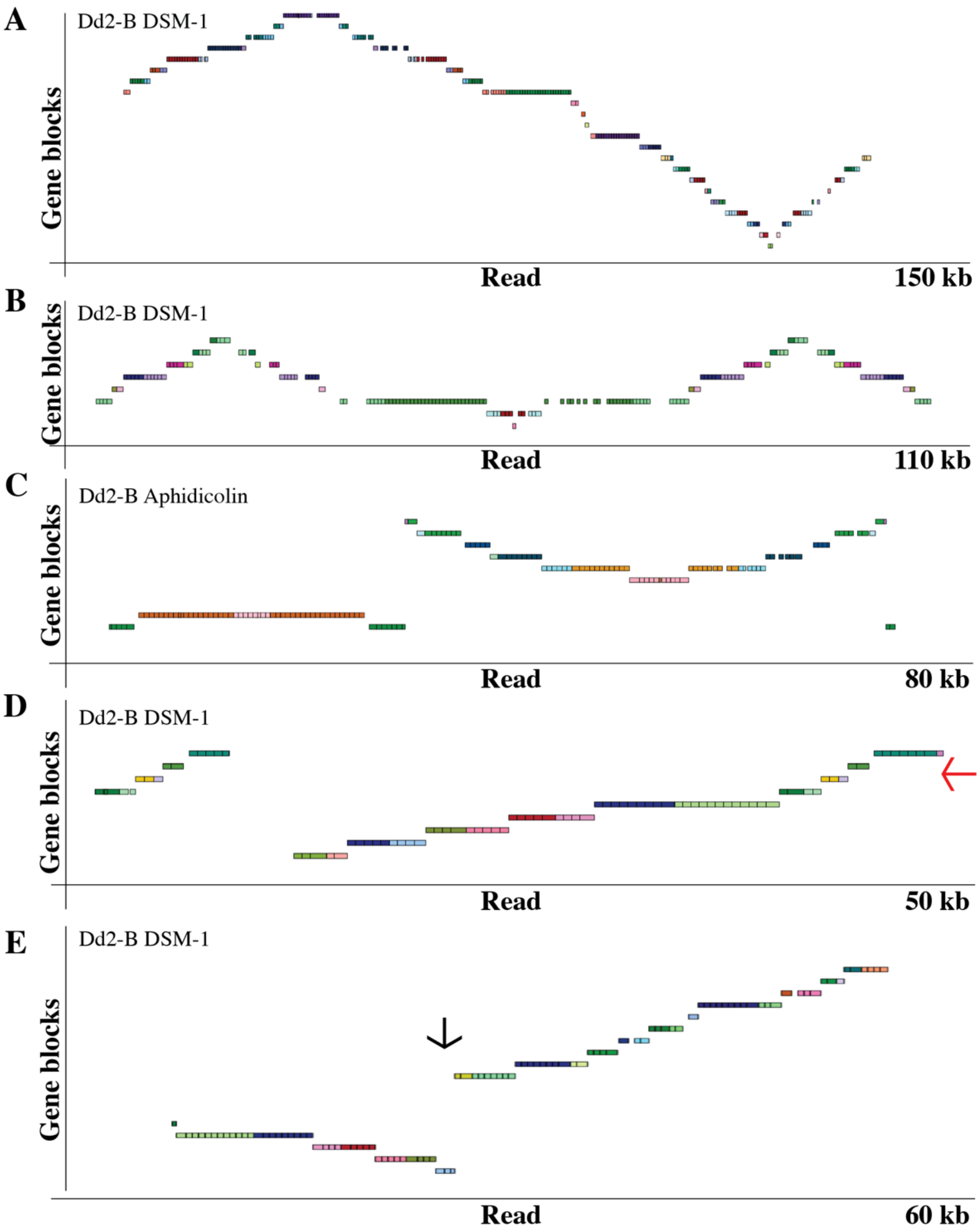


Figure 3.4 Visualization of de novo copy number variations detected in Dd2-B samples by Nanopore long reads

(**A**) A complex pyramid structure detected in Dd2-B DSM-1 sample. (**B**) A complex pyramid structure detected in Dd2-B DSM-1. (**C**) A complex pyramid structure detected in Dd2-B Aphidicolin sample. (**D**) A de novo amplification event detected in Dd2-B DSM-1 sample. The red arrow points to the gene with an increased copy number. (**E**) An unknown structural variation covering the breakpoint close to the breakpoint in panel D. The black arrow points to the breakpoint of the copy number variation.

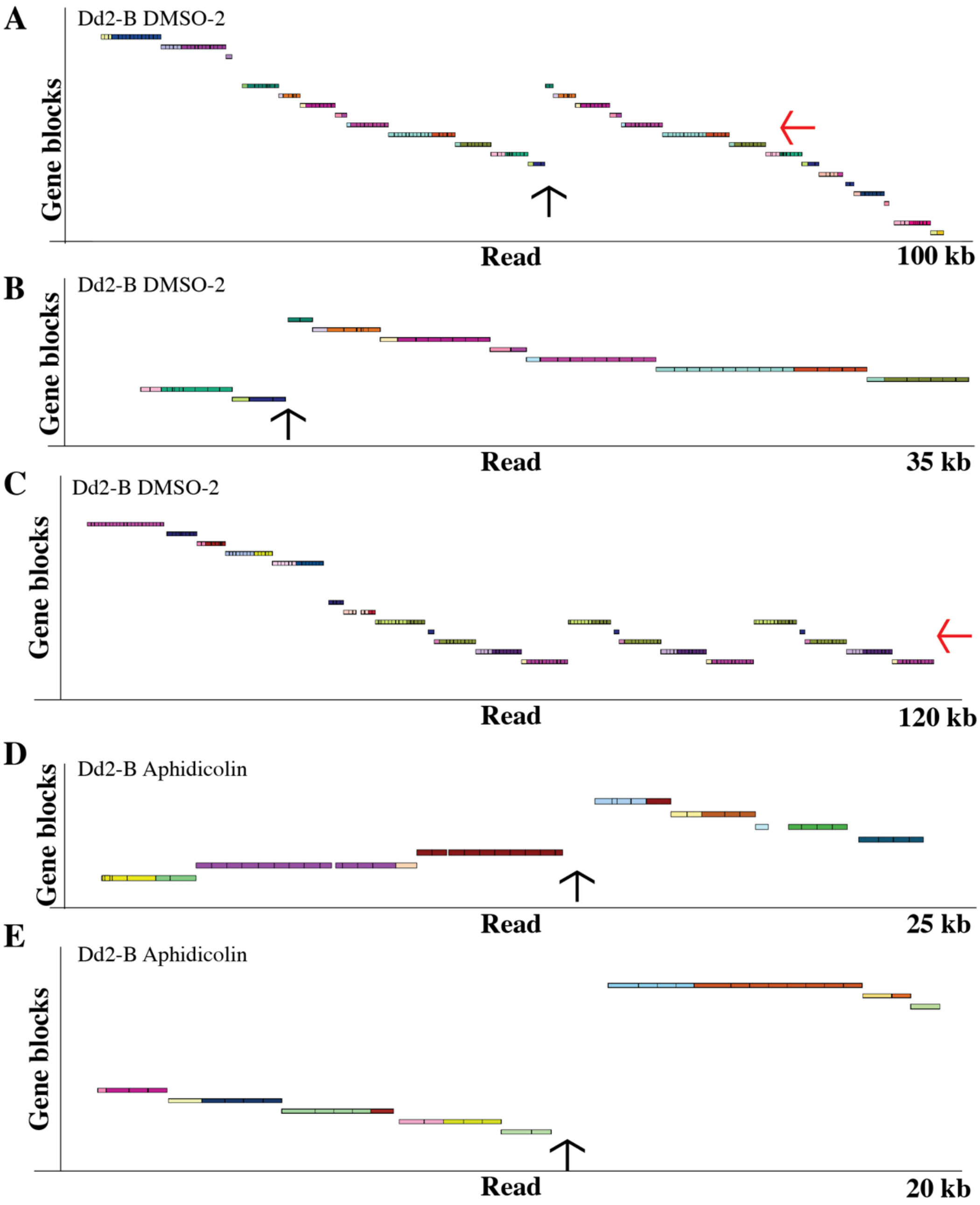


Figure 3.5 Visualization of de novo copy number variations detected in Dd2-B (aphidicolin treatment) samples by Nanopore long reads

(**A**)-(**B**) A de novo amplification event detected in Dd2-B DMSO-2 sample by 2 reads. The red arrow points to the gene with an increased copy number. The black arrow points to the breakpoint of the structural variation. (**C**) One tandem amplification detected in Dd2-B DMSO-2 sample. The red arrow pointed to the genes with increased copies. (**D**) One de novo structural variant detected in Dd2-B Aphidicolin sample. The black arrow points to the breakpoint of the structural variation. (**E**) One de novo structural variant detected by breakpoints in Dd2-B Aphidicolin sample. The black arrow points to the breakpoint of the structural variation.

Table 3.10 Gene information of detected de novo copy number variations in Dd2-B samples with or without stress treatment

Sample	Type of variants	Copy number	Chr	Start (Gene ID)	Start (Gene name)	End (Gene ID)	End (Gene name)	Number of supporting read
Dd2-B DMSO-1	Inverted amplifications	>=2	13	Pf3D7_1328300	Uncharacterized protein	Pf3D7_1329000	DNA-directed RNA polymerase subunit	1
			13	Pf3D7_1338900	Serine/threonine protein kinase, putative	Pf3D7_1339200	tRNA Proline	1
			13	Pf3D7_1357100	Elongation factor 1-alpha	Pf3D7_1357900	Pyrroline-5-carboxylate reductase, putative	1
Dd2-B DSM-1	Amplification	>=2	7	Pf3D7_0724900	Kinesin-19, putative	Pf3D7_0725300	Uncharacterized protein	1
			11	Pf3D7_1114900	Uncharacterized protein	Pf3D7_1115400	Cysteine proteinase falcipain 3	1
			13	Pf3D7_1344400	Uncharacterized protein	Pf3D7_1344700	Uncharacterized protein	1
			2	Pf3D7_0214700	Uncharacterized protein	Pf3D7_0214900	Rhoptry neck protein 6	1
	Inverted amplification	>=2	7	Pf3D7_0717600	Uncharacterized protein	Pf3D7_0718400	Mitochondrial ribosomal protein S8, putative	1
			11	Pf3D7_1122600	Uncharacterized protein	Pf3D7_1123000	Uncharacterized protein	1
			5	Pf3D7_0521900	Uncharacterized protein	PF3D7_0523200	Heptatricopeptide repeat-containing protein, putative	1
			3	Pf3D7_0323400	Rh5-interacting protein	Pf3D7_0323700	U4/U6.U5 tri-snRNP-associated protein 1, putative	1
			4	Pf3D7_0418000	Uncharacterized protein	Pf3D7_0418900	Uncharacterized protein	1
			13	Pf3D7_1312450	Apical ring associated protein 1, putative	Pf3D7_1313200	Methionyl-tRNA formyltransferase, putative	1
	Complex pyramid	>=2	7	Pf3D7_0718100	Exported serine/threonine protein kinase	Pf3D7_0719600	60S ribosomal protein L11a, putative	1
			9	Pf3D7_0930700	Uncharacterized protein	Pf3D7_0931300	Uncharacterized protein	1
			11	Pf3D7_1126800	RNA-binding protein, putative	Pf3D7_1127800	TFIIS central domain-containing protein, putative	1
			14	Pf3D7_1412600	Deoxyhypusine synthase	Pf3D7_1413000	Uncharacterized protein	1
			10	Pf3D7_1002200	Tryptophan-rich antigen 3	Pf3D7_1002500	Uncharacterized protein	1
			13	Pf3D7_1335900	Thrombospondin-related anonymous protein	Pf3D7_1336800	Nuclear movement protein, putative	1
			11	Pf3D7_1146500	Leucine-rich repeat protein	Pf3D7_1147200	Tubulin--tyrosine ligase, putative	1
			12	Pf3D7_1251300	dTMP kinase	Pf3D7_1251700	Tryptophan--tRNA ligase	1
			11	Pf3D7_1101900	erythrocyte membrane protein 1	Pf3D7_1105200	WD repeat-containing protein WRAP73, putative	1
			11	Pf3D7_1137100	Mitochondrial ribosomal protein S9, putative	Pf3D7_1139100	RNA-binding protein, putative	1

			6	Pf3D7_0604600	DNA helicase, putative	Pf3D7_0605800	Probable DNA repair protein RAD50	1
			13	Pf3D7_1334100	Uncharacterized protein	Pf3D7_1335400	Reticulocyte-binding protein 2 homolog a	1
			14	Pf3D7_1429200	AP2 domain transcription factor AP2-O3, putative	Pf3D7_1429400	rRNA (Adenosine-2'-O-)-methyltransferase, putative	1
			4	Pf3D7_0418300	Uncharacterized protein	Pf3D7_0419600	Ran-specific GTPase-activating protein 1, putative	1
			6	Pf3D7_0614200	Cytosolic Fe-S cluster assembly factor NAR1, putative	Pf3D7_0615000	Uncharacterized protein	1
			7	Pf3D7_0721800	Uncharacterized protein	Pf3D7_0722200	Rhoptry-associated leucine zipper-like protein 1	1
	Breakpoints of unknown structural variations	NA	13	Pf3D7_1344000	Aminomethyltransferase, putative	Pf3D7_1344800	Aspartate carbamoyltransferase	1
			11	Pf3D7_1110300	G-patch domain-containing protein	Pf3D7_1111700	Uncharacterized protein	1
Dd2-B DMSO-2	Amplification	3	7	Pf3D7_0724900	Kinesin-19, putative	Pf3D7_0725300	Uncharacterized protein	1
		2	4	Pf3D7_0410600	Uncharacterized protein	Pf3D7_0411400	DEAD box ATP-dependent RNA helicase, putative	1
	Inverted amplification	>=2	4	Pf3D7_0418800	MOLO1 domain-containing protein, putative	Pf3D7_0419400	Uncharacterized protein	1
			6	Pf3D7_0609600	Uncharacterized protein	Pf3D7_0610900	Transcription elongation factor SPT5, putative	1
			1	Pf3D7_0104400	4-hydroxy-3-methylbut-2-enyl diphosphate reductase, apicoplast	Pf3D7_0104600	Uncharacterized protein	1
			11	Pf3D7_1145400	Dynamin-like protein	Pf3D7_1145900	Uncharacterized protein	1
	Breakpoints of unknown structural variation	NA	4	Pf3D7_0410600	Uncharacterized protein	Pf3D7_0411400	DEAD box ATP-dependent RNA helicase, putative	1
Dd2-B Aphidicolin	Amplification	>=2, >=4	7	Pf3D7_0724900	Kinesin-19, putative	Pf3D7_0725300	Uncharacterized protein	2
		>=2	9	Pf3D7_0923200	Nitric oxide synthase, putative	Pf3D7_0923300	Perforin-like protein 3	1
			7	Pf3D7_0703900	Uncharacterized protein	Pf3D7_0703900	Uncharacterized protein	1
	Inverted amplification	>=2	6	Pf3D7_0631200	erythrocyte membrane protein 1	Pf3D7_0631200	erythrocyte membrane protein 1	1
			9	Pf3D7_0928800	Serine/threonine protein kinase, putative	Pf3D7_0928800	Serine/threonine protein kinase, putative	1
			5	Pf3D7_0522000	Uncharacterized protein	Pf3D7_0524800	Ubiquitin fusion degradation protein 1, putative	5
			4	Pf3D7_0409300	Methyltransferase, putative	Pf3D7_0409600	Replication protein A1, large subunit	1
			12	Pf3D7_1232200	Dihydrolipoyl dehydrogenase	Pf3D7_1232200	Dihydrolipoyl dehydrogenase	1

			3	Pf3D7_0315500	Mitochondrial ribosomal protein L29/L47, putative	Pf3D7_0316000	Microneme associated antigen	1
			8	Pf3D7_0805300	C2H2-type domain-containing protein	Pf3D7_0805300	C2H2-type domain-containing protein	1
			10	Pf3D7_1013800	Uncharacterized protein	Pf3D7_1014300	SPRY domain-containing protein, putative	1
	Complex pyramid	>=2	6	Pf3D7_0611400	SWIB/MDM2 domain-containing protein	Pf3D7_0611900	Lsm12, putative	1
			14	Pf3D7_1413400	30S ribosomal protein S9, putative	Pf3D7_1413800	2-(3-amino-3-carboxypropyl)histidine synthase subunit 1	1
			9	Pf3D7_0926000	Protein kinase, putative	Pf3D7_0926900	Replication termination factor, putative	1
			11	Pf3D7_1115200	Histone-lysine N-methyltransferase SET7	Pf3D7_1115800	Uncharacterized protein	1
			11	Pf3D7_1121300	Tyrosine kinase-like protein	Pf3D7_1123000	Uncharacterized protein	1
	Breakpoints of unknown structural variation	NA	9	Pf3D7_0914600	Transcription elongation factor 1 homolog	Pf3D7_0916000	Major facilitator superfamily domain-containing protein, putative	1
			13	Pf3D7_1313400	DEAD box helicase, putative	Pf3D7_1314700	Pinin/SDK/MemA domain-containing protein, putative	1
			12	Pf3D7_1224600	Holocytochrome c-type synthase	Pf3D7_1225200	Uncharacterized protein	1
			13	Pf3D7_1364800	DNA-directed RNA polymerases I, II, and III subunit RPABC1, putative	Pf3D7_1365700	SNARE associated Golgi protein, putative	1
			1	Pf3D7_0103600	ATP-dependent RNA helicase, putative	Pf3D7_0104000	Thrombospondin-related sporozoite protein	1
			4	Pf3D7_0410600	Uncharacterized protein	Pf3D7_0411400	DEAD box ATP-dependent RNA helicase, putative	2
			8	Pf3D7_0804700	Uncharacterized protein	Pf3D7_0805300	C2H2-type domain-containing protein	1
			7	Pf3D7_0730700	tRNA Threonine	Pf3D7_0731300	PRESAN domain-containing protein	1
			12	Pf3D7_1206600	DNA-directed RNA polymerase subunit beta	Pf3D7_1207100	Small subunit rRNA processing factor, putative	1

Note: genes detected multiple times were marked by the same color

3.5 Discussion

In this study, we compared two Nanopore sequencing library preparation methods and improved the read length of Nanopore reads for better copy number variation detection. We sequenced laboratory *Plasmodium falciparum* parasite lines with and without DSM-1 or Aphidicolin stress treatment. By analyzing the sequenced Nanopore reads using the Shiny Application, we identified previously known and de novo copy number variations in the parasites. Importantly, we made several interesting observations in this pilot experiment and discussed their implications as below.

Improvement of Nanopore sequencing read length

In this study, we compared two PCR-free library preparation methods for Nanopore sequencing. We found that the use of the Nanopore Ultra-long sequencing kit and Nanobind disk (average read length: 32 kb) improved the read length of the sequencing compared to the use of the Nanopore Ligation sequencing kit and AMPure XP beads (average read length 11 kb) (**Table 3.2**). This is because: 1) the Ultra-long sequencing kit (one DNA purification step) includes less DNA purification steps compared to the Ligation sequencing kit (three DNA purification steps); 2) the Nanobind disk is designed to use nanostructured silica to protect DNA from shearing and facilitate washing to generate largest DNA based on the product description; 3) it requires vigorous pipetting to elute ultra-high molecular weight DNA from the AMPure XP beads, which consequently shears DNA. As shown in **Table 3.3**, we detected the whole structure of GCH1 amplification (2 kb or 5 kb repeat sequences), but not that of the MDR1 amplification (82 kb repeat sequences) or DHODH amplification (74 kb repeat sequences) in parasites DNA prepared by the Ligation sequencing kit and AMPure XP beads. Longer read length could improve the detection of large copy number variations by covering a larger portion of their structures. However, due to lack of direct comparison of the same parasite DNA with the same copy number variations prepared by the two different library preparation methods, we could not determine whether longer read length can allow the better detection of large copy number variations.

Proof of principle for detecting de novo structural changes

We sequenced 4 laboratory *Plasmodium falciparum* parasite lines (Dd2-A, HB3, H2, 3D7) that are originally cloned in the laboratory culture. With the long read visualization analysis method, we identified known copy number variations (GCH1, MDR1, DHODH) in these parasites (**Table 3.3**). More specifically, we accurately detected the whole structures of small copy number variations (GCH1 amplicon with 2kb or 5kb repeat sequences) and the breakpoints of large copy number variations with 82kb (MDR1 in Dd2), 74kb (DHODH in H2) and 161 kb (GCH1 in HB3) repeat sequences that have been found in previous studies (Miles et al., 2016, Huckaby et al., 2018). Thus, this confirms the validity of the Nanopore sequencing, our analysis method, and provides confidence for detecting novel structural changes in the parasites.

We identified de novo amplifications, inverted amplifications, and breakpoints of unknown structural variations, which have not been previously identified in the 4 laboratory lines (Dd2-A, HB3, H2, 3D7) (**Table 3.5** and **Table 3.6**). Since these de novo structural changes are only detected by one read at ~ 24X coverage (**Table 3.6**), while other reads spanning the same genes show no copy number variation, these events are likely rare and arise randomly across the parasite populations. We also identified these same categories of structural variations in Dd2-B parasites with or without DSM-1 or Aphidicolin treatment (**Table 3.9**, **Table 3.10**). Most of the de novo copy number variations detected in Dd2-B parasites are still only supported by few reads at 226X to 364X coverage (**Table 3.10**), providing further evidence for their de novo generation across the genome.

The nature of pyramid structures detected in the Nanopore reads

We observed many pyramid-like structures of genes (symmetrical inverted amplifications) across the genome in sequenced reads (**Table 3.4**). It is unclear whether these pyramid structures are caused by Nanopore sequencing error or real structural changes.

It is possible that pyramids are caused by transient structural changes in the parasite genome. A previous study suggests that secondary structures are often formed within repetitive DNA sequences that can pair out of register after double-stranded DNA is denatured, resulting in the misalignment of the two strands. Particularly, secondary structures can form on a single-strand DNA end created by replication fork or during the repair of double-strand break (Kaushal et al., 2019). In addition, the previous study indicates that if the misalignment is not corrected, expansions or contractions in repeat length will result (Kaushal et al., 2019). In our study, based on the observation of pyramid structures, it is possible that the secondary structure formed during DNA replication or DNA repair can lead to the formation of double-strand DNAs with a hairpin-capped DNA end (as shown in the **Figure 3.6A, 3.6B**) can cause the pyramid structures in Nanopore sequencing reads. Further validation using PCR-based method targeting the breakpoints of the pyramid structures is necessary to confirm these structural changes.

It is also possible that the pyramids are caused by Nanopore sequencing error. The generation of inverted copies of sequenced DNA during Nanopore sequencing has not been reported in the literature. However, it is possible this type error is associated with the high AT-content and prevalent repetitive sequences of the *P. falciparum* genome.

Interestingly, the second repeat of these pyramid structures often showed lower base calling accuracy than the first repeat (**Figure 3.2A**), which has been observed in another study (Spealman et al., 2020). It has been suggested that the higher error rate in the second copy of inverted amplifications is likely caused by the secondary structures formed by inverted amplifications interfering with the translocation of DNA through the sequencing pore (Spealman et al., 2020).

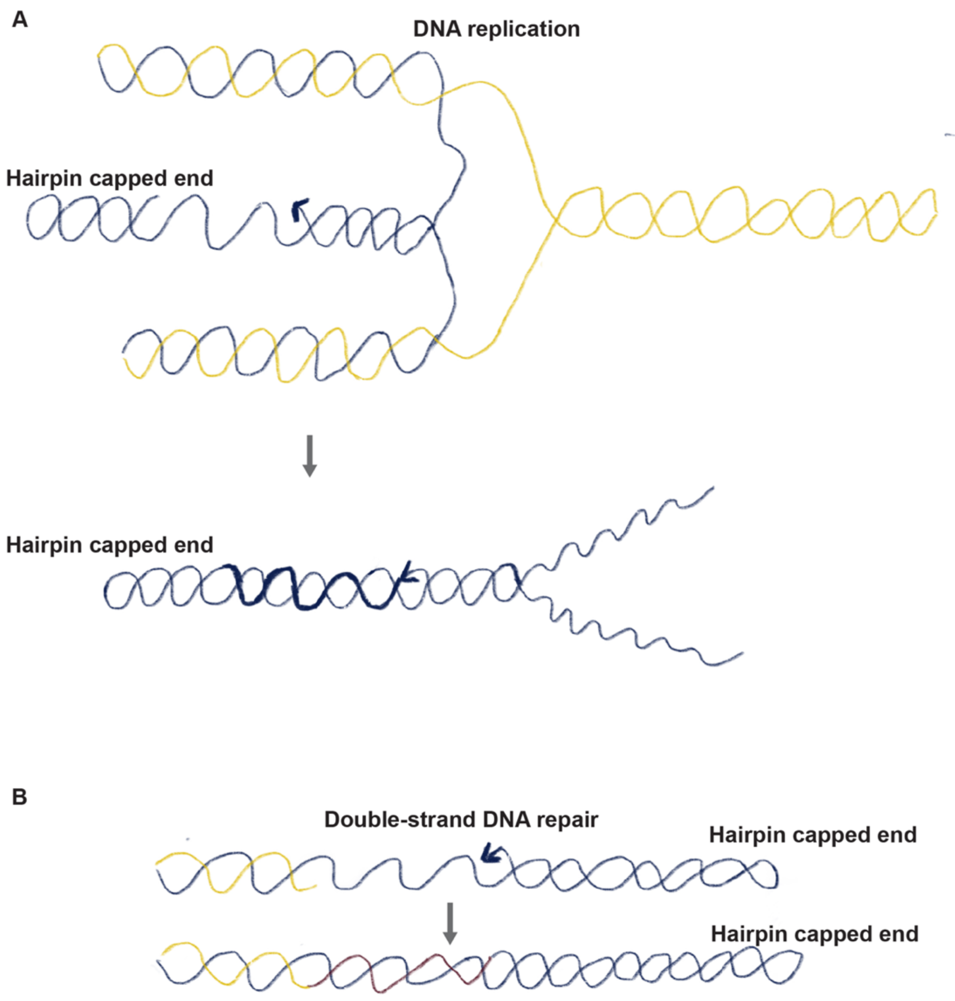


Figure 3.6 Illustration of the formation of inverted duplication during DNA replication and DSB repair
(**A**) DNA secondary structure can form during DNA replication on a single-strand DNA end created by fork reversal and the ligation of these sequences can result in inverted duplication sequences with a hairpin capped end. (**B**) Secondary structure can also form when DNA is single-strand during DNA repair of double-strand break. And the ligation between these sequences can lead to an inverted duplication sequence with hairpin capped end. Image is adapted from Kaushal et al., 2019.

Interestingly, we found that there are more pyramid structures (symmetrical inverted amplifications, **Figure 3.2A, Table 3.4**), and more complex-pyramid structures (including more than one inverted amplification of the same or different genes **Figure 3.4A-C**, **Table 3.9**) in treated samples. Given what we know about DSM1 and Aphidicolin (see below), this observation indicates that the formation of these structures may be related to replicative stress or DNA repair of double strand break in vivo.

Potential factors contributing to novel structural changes in parasites under stress

With the treatment of sublethal DSM-1 and Aphidicolin stress, we observed an increased rate of copy number variation within the parasite samples (~9-fold increase in the DSM-1 treated sample and ~ 4-fold increase in the Aphidicolin treated sample). There are several factors that could potentially contribute to this effect.

Firstly, the drug stress might select the sub-population with more copy number variations within the Dd2-B sample. We identified two main sub-populations of cells within the Dd2-B samples by allele frequency analysis at specific genes (**Table 3.7**), which limited our ability to interpret the difference between the treated and untreated samples. One sub-population carries more 3D7 alleles (sensitive to multiple antimalarial drugs), while the other sub-population carries more Dd2 alleles (resistant to multiple antimalarial drugs) (**Table 3.7**). Since the proportion of two sub-populations changed after Aphidicolin treatment based on allele frequency analysis (58%: 41% in Aphidicolin treated sample; 76%: 22% in DMSO control sample), it is possible that Aphidicolin selected the sub-population with a shorter maturation-blocking time, or with more alleles that enhance parasite fitness or erythrocyte invasion, or other beneficial traits. Further analysis on the overlap between genes with allele frequency changes and genes with copy number changes is needed to investigate the selection effect on structural changes by Aphidicolin. We did not observe an obvious change in the sub-population proportion in DSM-1 treated sample by allele frequency analysis (**Table 3.7**), thus it is possible DSM-1 did not exhibit a selection effect.

Secondly, the changing proportion of the two sub-populations under Aphidicolin treatment could also be due to random fluctuations over time. We expanded the same Dd2-B parasite stock and cultured parasites for the experiment. The Dd2-B-DMSO-1 and Dd2-B-DMSO-2 samples are equivalent, other than the time at which they were harvested (4 weeks apart), yet they exhibited different proportions of subpopulations. Random fluctuations could also be contributing to the difference between the Aphidicolin treated and DMSO control sample since they were re-cultured for different lengths of time (18 hours compared to 40 hours, respectively). Thus, it is also possible that the increased structural changes in the treated sample are caused by random fluctuations during in vitro culture. More biological repeats of the experiment sequenced separately may help to determine if random variation is contributing to the differences in increased formation of copy number variation.

Thirdly, differences in the detection of structural variants in the genome could be stage-dependent. Of note, the early-stage percentage of the parasites between treated and untreated samples were different at the time of harvest (**Table 3.1).** Early stage parasites have not yet begun to replicate their genome (Matthews et al., 2018). As shown in a previous study, early-stage parasites cannot reconstitute chromosome content as efficient as multinuclear late-stage parasites after exposure to ionizing radiation that leads to DNA breakage (Lee Andrew H. et al., 2014). Thus, the compact nature of their chromatin of early stage parasites may impact the formation of secondary structure, activity of repair pathways, and perhaps formation of structural variants (Lee Andrew H. et al., 2014). In our experiments, we did not identify supporting evidence for a stage-effect; the two untreated control samples had drastically different levels of early-stage parasites (72.4% and 35.5%, **Table 3.1**), yet their overall number of structural variations were very similar (3 and 6, **Table 3.9**). Similarly, the two treated samples varied in their early-stage percentage (18.5% and 62.1%, **Table 3.1**) and exhibited a high frequency of structural variation (27 and 25, **Table 3.9**). Future experiments could control for the possibility of a stage effect more specifically by ensuring that all samples are relatively similar in stage.

Finally, it is possible that the treatments are truly inducing structural changes in the parasite genomes. Aphidicolin induces double-strand breaks on AT-rich DNA sequences and copy number variations could be introduced during DNA repair (Inselburg and Banyal, 1984). DSM-1 is a drug that targets the biosynthesis of pyrimidines and required for DNA synthesis (Phillips et al., 2008) and replication stress is also a contributor to DNA breakage (Gupta et al., 2016). Since the structural variants that we observed after treatment were broadly distributed across the genome with many repetitive AT-rich sequences (Huckaby et al., 2018), there were not specific structural changes related to these pathways detected (**Table 3.10**) and it is unclear whether there is a stage-effect as discussed above, it is also possible that both treatment conditions induced random and novel structural changes. To understand whether the formation of these novel structural changes is related to replication stress or double-strand break, we can treat the parasites with a combination of double-strand break inducing drugs and drugs inhibiting DNA repair process that would be needed during the formation in future experiments. The targets of the drugs inhibiting DNA repair process can be global protein synthesis or specific double-strand break repair pathways as suggested in a previous study in bacteria (Amarh and Arthur, 2019).

Drugs such as Mefloquine, Emetine, Halofuginone etc., have been suggested to inhibit protein synthesis in *P. falciparum* parasites (Tamaki Fabio et al., 2022). Drugs such as Butylphenyl-dGTP and 7-acetoxypentyl-(3, 4 dichlorobenzyl) guanine targeting essential enzymes (such as DNA polymerase δ) in double-strand break repair pathways, thus can be used to inhibit the DNA repair process in *P. falciparum* parasites (Vasuvat et al., 2016).

To summarize, we cannot determine the precise cause of our observation of increased structural changes in treated samples based on current data. The detection of two main sub-population of cells was an unexpected observation, but it provides information about how the treatments impacted the parasite population. This information will be useful as we consider experimental design for future experiments. It will be essential that we use a homogeneous parasite line, employ multiple biological replicates sequenced separately, and include parasite samples before and after treatment to better understand the role of drug stress in the evolution of copy number variations.

Methods to validate the detected copy number variations from Nanopore reads

We identified several reads covering breakpoints of de novo amplifications and unknown structural variations. To validate whether we are detecting real copy number variations, we can perform PCR to confirm the existence of the breakpoints of the copy number variations using the DNA from each sample. Alternatively, we can also sequence the parasites with Illumina short-read sequencing to check if we can detect split reads or discordant reads that covering the breakpoint of the de novo copy number variations (Layer et al., 2014). However, it will be challenging to detect rare de novo copy number variations covered by a single read since we use bulk DNA for these validation methods and PCR amplification or short-read sequencing lack the sensitivity to detect rare genetic variations (Cantsilieris et al., 2013). The chance of success for these validation methods will be higher for loci that are detected by more than one read or in multiple parasite lines (**Table 3.10**). In addition, we also plan to use our single cell sequencing pipeline to evaluate the impact of treatments on de novo CNV formation, see **Chapter 4**.

In this study, we used single molecule Nanopore sequencing of laboratory cultured parasite lines with or without stress treatment to assess the frequency of de novo structural variations encompassing genes across various chromosomes. This is the first study enabling direct visualization of copy number variations on Nanopore reads and discovery of rare heterogeneous copy number variations in malaria parasites. This new method allows us to study the molecular process of the generation of copy number variation, understand genome evolution through copy number variations and the dynamics of parasite adaptation under stressed conditions. With the understanding of the dynamics of copy number variation in the malaria parasites, this study can provide new strategies to block the development of resistance in this deadly organism. Lastly, an extension of this study can also help understand the evolutionary role of copy number variations in the adaptation of diverse organisms.

4 CHAPTER IV: Single cell sequencing of the small and AT-skewed genome of malaria parasites for heterogeneous copy number variations detection

The method presented in this chapter was published on *Genome Med* 13, 75 (2021) using the title:

Single cell sequencing of the small and AT-skewed genome of malaria parasites

https://doi.org/10.1186/s13073-021-00889-9

Publication Author List: Shiwei Liu, Adam C. Huckaby, Audrey C. Brown, Christopher C. Moore, Ian Burbulis, Michael J. McConnell, Jennifer L. Güler

Statement of contribution and acknowledgement

MJM and JLG conceived of the project. SL, ACH, and JLG designed the experiments. IB and MJM provided WGA protocols, advice, and equipment (CellRaft AIR System) throughout the project. ACB and CCM procured and processed clinical samples from the University of Virginia Medical Center. SL conducted all the experiments. SL analyzed the data, with support from ACH and JLG. SL and JLG wrote the manuscript. ACH, ACB, CCM, IB, and MJM edited the manuscript. All authors read and approved the final manuscript.

Our thanks to Michelle Warthan for parasite culturing support, AnhThu Nguyen (Biology Genomics Core, University of Virginia) for sequencing support, and all members of Dr. Jennifer Guler's laboratory, as well as the laboratories of Dr. William Petri Jr and Barbara Mann at University of Virginia, for their helpful discussions and insight.

4.1 Abstract

Single cell genomics is a rapidly advancing field; however, most techniques are designed for mammalian cells. We present a single cell sequencing pipeline for an intracellular parasite, *Plasmodium falciparum*, with a small genome of extreme base content. Through optimization of a quasi-linear amplification method, we target the parasite genome over contaminants and generate coverage levels allowing detection of minor genetic variants. This work, as well as efforts that build on these findings, will enable detection of parasite heterogeneity contributing to *P. falciparum* adaptation. Furthermore, this study provides a framework for optimizing single cell amplification and variant analysis in challenging genomes.

4.2 Introduction

Malaria is a life-threatening disease caused by protozoan *Plasmodium* parasites. *P. falciparum* causes the greatest number of human malaria deaths (Rich et al., 2009). The clinical symptoms of malaria occur when parasites invade human erythrocytes and undergo rounds of asexual reproduction by maturing from early forms into late-stage forms and bursting from erythrocytes to begin the cycle again (Matthews et al., 2018). In this asexual cycle, parasites possess a single haploid genome during the early stages; rapid genome replication during subsequent stages leads to an average of 16 genome copies per individual (Matthews et al., 2018).

Due to lack of effective vaccines, antimalarial drugs are required to treat malaria. However, drug efficacy is mitigated by the frequent emergence of resistant populations (Blasco et al., 2017). Both single nucleotide polymorphisms (SNPs) and copy number variations (CNVs, the amplification or deletion of a genomic region) contribute to antimalarial resistance in *P. falciparum* (Kidgell et al., 2006; Conway, 2007; Hyde, 2007; Ribacke et al., 2007; Nair et al., 2008; Cheeseman et al., 2009; Bopp et al., 2013; Guler et al., 2013; Menard and Dondorp, 2017). It is important to assess genetic diversity within parasite populations to better understand the mechanisms of rapid adaption to antimalarial drugs and other selective forces. These studies are often complicated by multi-clonal infections and limited parasite material from clinical isolates.

Recent studies have begun to overcome these limitations for SNP analysis; methods including leukocyte depletion (Venkatesan et al., 2012), selective whole genome amplification (WGA) of parasite DNA (Ibrahim et al., 2020), hybrid selection with RNA baits (Melnikov et al., 2011) and single cell sequencing of *P. falciparum* parasites (Trevino et al., 2017; Jett et al., 2020) help enrich parasite DNA, determine genetic diversity, and understand the accumulation of SNPs in long term culture. However, the study of genetic diversity in early-stage parasites on a single cell level remain challenging (Trevino et al., 2017); the lack of alternative single cell approaches for *P. falciparum* parasites impedes the validation of SNP results by parallel investigations (Lähnemann et al., 2020).

The dynamics of CNVs in evolving populations are not well understood. One reason for this is that most of the *P. falciparum* CNVs have been identified by analyzing bulk DNA from selected parasites, where CNVs are present in the majority of parasites (Price et al., 2004; Kidgell et al., 2006; Ribacke et al., 2007; Heinberg et al., 2013; Ravenhall et al., 2019). However, many low frequency CNVs undoubtedly remain undetected. There is speculation that these low-frequency CNVs are either deleterious or offer no advantages for parasite growth or transmission (Cheeseman et al., 2016; Ravenhall et al., 2019) but orthogonal methods to verify genome dynamics within the population are needed. Recent investigations in other organisms have analyzed single cells to detect low frequency CNVs within heterogeneous populations (Macaulay and Voet, 2014; Wang and Navin, 2015; Gawad et al., 2016; Lauer et al., 2018; Wang et al., 2018, 2020).

Single cell-based approaches provide a significant advantage for detecting rare genetic variants (SNPs and CNVs) 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. Therefore, WGA is required to generate sufficient DNA quantities for these analyses. Several WGA approaches have been reported and each has advantages and disadvantages for different applications; however, most have been optimized for mammalian cell analysis (Hughes et al., 2014; Neves et al., 2014; Wang et al., 2014; Hou et al., 2015; Deleye et al., 2017; Vitak et al., 2017; Burbulis et al., 2018; Duan et al., 2018; Rohrback et al., 2018; Chronister et al., 2019). Because WGA leads to high levels of variation in read abundance across the genome, CNV analysis in the single cell context is especially challenging. Previous approaches have been tuned specifically for CNV detection in mammalian genomes, which are generally hundreds of kilobases to megabases in size (Navin et al., 2011; Zong et al., 2012; McConnell et al., 2013; Campbell et al., 2015; Fu et al., 2015; Ning et al., 2015).

To date, the detection of CNVs in single *P. falciparum* parasites using whole genome sequencing has not been achieved. The application of existing WGA methods is complicated by the parasite's small genome size and extremely imbalanced base composition (23Mb haploid genome with 19.4% GC-content (Gardner et al., 2002)). Each haploid parasite genome contains 25 femtograms of DNA, which is ~280-times less than the ~6400Mb diploid human genome. Therefore, an effective *P. falciparum* WGA method must be both highly sensitive and able to handle the extreme base composition. One WGA method, multiple displacement amplification (MDA), has been used to amplify single *P. falciparum* genomes with near complete genome coverage (Trevino et al., 2017; Nkhoma et al., 2020). These studies successfully detected SNPs in single parasite genomes but did not report CNV detection, which is possibly disrupted by low genome coverage uniformity (Huang et al., 2015) and the generation of chimeric reads by MDA (Lasken and Stockwell, 2007), as well as the relatively small size of CNVs in *P. falciparum* (<100kb) (Huckaby et al., 2018; Simam et al., 2018a).

Multiple annealing and looping-based amplification cycles (MALBAC) is another WGA method that exhibits adequate uniformity of coverage, which was advantageous for detecting CNVs in single human cells (Zong et al., 2012). MALBAC has the unique feature of quasi-linear pre-amplification, which reduces the bias associated with exponential amplification (Zong et al., 2012). However, standard MALBAC is less tolerant to AT-biased genomes, unreliable with low DNA input, and prone to contamination (de Bourcy et al., 2014; Oyola et al., 2014; Ning et al., 2015). Thus, optimization of this WGA method is necessary for *P. falciparum* genome analysis.

In this study, we developed a single cell sequencing pipeline for *P. falciparum* parasites, which included efficient isolation of single infected erythrocytes, an optimized WGA step inspired by MALBAC, and a 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. We assessed amplification bias first using a PCR-based approach and then by sequencing. We evaluated genome coverage breadth and coverage uniformity, as well as amplification reproducibility. Furthermore, we combined two approaches to limit false positives for CNV detection and applied stringent filtering steps for SNP detection in single cell genomes. This work, as well as efforts that build on these findings, will enable the detection of parasite-to-parasite heterogeneity to clarify the role of genetic variations in the adaptation of *P. falciparum*. Furthermore, this study provides a framework for the optimization of single cell whole genome amplification and CNV/SNP analysis in other organisms with challenging genomes.

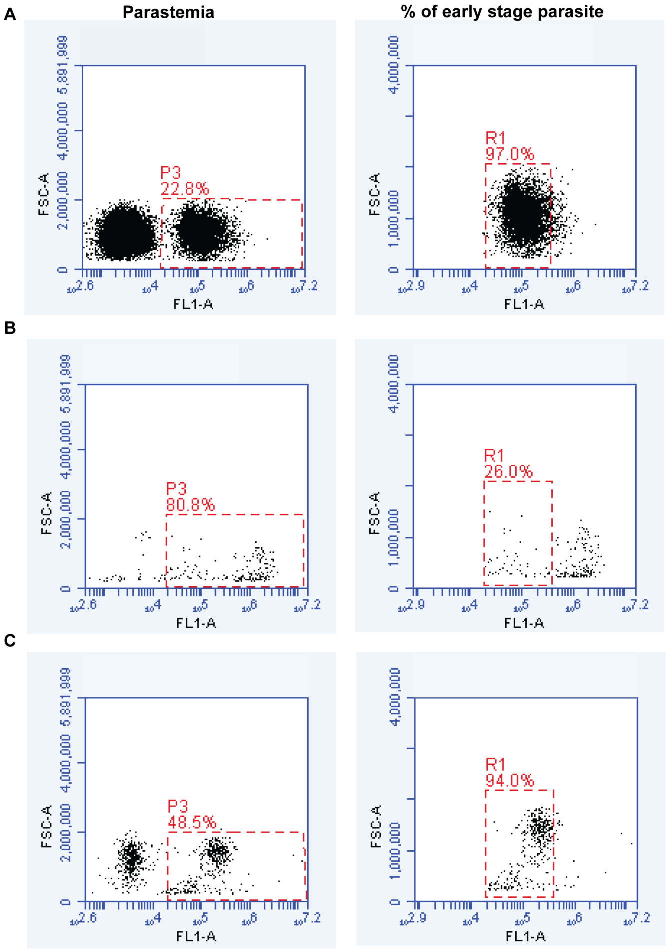
4.3 Methods

Parasite Culture

We freshly thawed erythrocytic stages of *P. falciparum* (*Dd2*, MRA-150, Malaria Research and Reference Reagent Resource Center, BEI Resources) from frozen stocks and maintained them as previously described (Haynes et al., 1976). Briefly, parasites were grown in *vitro* at 37°C in solutions of 3% hematocrit (serotype A positive human erythrocytes, Valley Biomedical, Winchester, VA) in RPMI 1640 (Invitrogen, USA) medium containing 24 mM $NaHCO_3$ and 25 mM HEPES, and supplemented with 20% human type A positive heat inactivated plasma (Valley Biomedical, Winchester, VA) in sterile, plug-sealed flasks, flushed with 5% O_2, 5% CO_2, and 90% N_2 (Guler et al., 2013). We maintained the cultures with media changes every other day and sub-cultured them as necessary to keep parasitemia below 5%. All parasitemia measurements were determined by SYBR green based flow cytometry (Bei et al., 2010). Cultures were routinely tested using the LookOut Mycoplasma PCR Detection Kit (Sigma-Aldrich, USA) to confirm negative infection status.

Clinical Sample Collection

We obtained parasites from an infected patient admitted to the University of Virginia Medical Center with clinical malaria. The patient had a recent history of travel to Sierra Leone, a malaria-endemic country, and *P. falciparum* infection was clinically determined by a positive rapid diagnostic test and peripheral blood smear analysis. We obtained the sample of 1.4% early-stage parasites within 24h of phlebotomy, incubated in the conditions described in *Parasite Culture* for 48 hours and washed the sample 3 times with RPMI 1640 HEPES to decrease levels of white blood cells. To fully evaluate our amplification method in the presence of heavy human genome contamination, we did not perform further leukodepletion. We set aside some of the sample for bulk DNA preparation (see *Bulk DNA Extraction*). Using another portion of the sample, we enriched for parasite-infected erythrocytes using SLOPE (Streptolysin-O Percoll) method (Brown et al., 2020), which increased the parasitemia from 1.4% to 48.5% (**Figure S4.1**). We then isolated the single *P. falciparum*-infected erythrocytes using the CellRaft AIR™System (Cell Microsystems, Research Triangle Park, NC) as detailed in *Parasite Staining and Isolation*.



Supplementary Figure 4.1 Confirmation of staging for enriched parasite samples

Parasitemia (proportion of infected erythrocytes; P3 gate, left plots) and proportion of early stage parasites (R1 gate, right plots) is shown for each sample. **A. Early stage laboratory parasite samples.** Early stage parasites were enriched by harvesting the flow-through of the MACS column, which contained parasite stages that have not accumulated hemozoin, a paramagnetic crystal, along with uninfected cells. The SLOPE method was used to deplete the sample of uninfected cells to yield a high parasitemia (22.8%), predominantly early-staged population (97.0%) for single cell isolation. The flow cytometry run collected 30,485 total events. **B. Late stage laboratory parasite samples.** Late stage parasites were enriched by collecting the bound fraction from the MACS column, which contained parasite stages with high levels of paramagnetic hemozoin. This high parasitemia (80.8%), predominantly late stage population (74%), was used for single cell isolation. Late-stage parasites were further selected by higher fluorescence (due to increased DNA content and mitochondrial size) on the CellRaft microscope during isolation (see **Figure 4.1A**). The flow cytometry run collected 1357 total events. **C. Early stage clinical parasite samples**. Whole blood collected from a *P. falciparum*-infected patient was stored in sodium citrate for 23 hours in the hospital and incubated in RPMI for 48 hours in the laboratory (see *Methods* for details). The SLOPE method was used to deplete the sample of uninfected cells to yield a high parasitemia (48.5%), predominantly early-staged population (94.0%, confirmed by microscopy) for single cell isolation. The flow cytometry run collected 1314 total events.

Bulk DNA Extraction

We lysed asynchronous *P. falciparum*-infected erythrocytes with 0.15% saponin (Sigma-Aldrich, USA) for 5min and washed them with 1x PBS (diluted from 10x PBS Liquid Concentrate, Gibco, USA). We then lysed parasites with 0.1% Sarkosyl Solution (Bioworld, bioPLUS, USA) in the presence of 1mg/ml proteinase K (from *Tritirachium album*, Sigma-Aldrich, USA) overnight at 37°C. We extracted nucleic acids with phenol/chloroform/isoamyl alcohol (25:24:1) pH 8.0 (Sigma-Aldrich, USA) using 2ml light Phase lock Gels (5Prime, USA). Lastly, we precipitated the DNA with ethanol using the standard Maniatis method (Maniatis et al., 1989).

Parasite Staining and Isolation

For late-stage parasite samples, we obtained laboratory *Dd2* parasite culture with a starting parasitemia of 1.7% (60% early-stage parasites). We separated late stage *P. falciparum*-infected erythrocytes from non-paramagnetic early stages using a LS column containing MACS® microbeads (Miltenyi Biotec, USA, (Ribaut et al., 2008)). After elution of bound late-stage parasite, the sample exhibited a parasitemia of 80.8% (74.0% late-stage parasites, **Figure S4.1**). For early-stage parasites, we obtained laboratory *Dd2* parasites culture with a starting parasitemia of 3% (46% early-stage parasites). We harvested the non-paramagnetic early stages parasites which were present in the flow-through of the LS column containing MACS® microbeads. Next, we enriched the infected erythrocytes using the SLOPE method, which preferentially lysed uninfected erythrocytes (Brown et al., 2020). The final parasitemia of enriched early-stage parasites was 22.8% (97.0% early-stage parasites, **Figure S4.1**). To differentiate *P. falciparum*-infected erythrocytes from remaining uninfected erythrocytes or cell debris, we stained the stage specific *P. falciparum*-infected erythrocytes with both SYBR green and MitoTracker Red CMXRos (Invitrogen, USA). We then isolated single *P. falciparum*-infected erythrocytes using the CellRaft AIR™ System (Cell Microsystems, Research Triangle Park, NC). We coated a 100-micron single reservoir array (CytoSort Array and CellRaft AIR user manual, CELL Microsystems) with Cell-Tak Cell and Tissue Adhesive (Corning, USA) following the manufacture's recommendations. Then, we adhered erythrocytes on to the CytoSort array from a cell suspension of ~20,000 cells in 3.5mL RPMI 1640 (Invitrogen, USA) with AlbuMAX II Lipid-Rich BSA (Thermo Fisher Scientific, USA) and Hypoxanthine (Sigma-Aldrich, USA). Lastly, we set up the AIR™ System to automatically transfer the manually selected single infected erythrocytes (SYBR+, Mitotracker+) into individual PCR tubes.

Steps to Limit Contamination

We suspended individual parasite-infected erythrocytes in freshly prepared lysis buffer, overlaid them with one drop (approx. 25µl) of mineral oil (light mineral oil, BioReagent grade for molecular biology, Sigma Aldrich, USA), and stored them at -80°C until WGA. We amplified DNA in a clean positive pressure hood located in a dedicated room, using dedicated reagents and pipettes, and stored them in a dedicated box at -20°C. We wore a new disposable lab coat, gloves and a face mask during reagent preparation, cell lysis, and WGA steps. We decontaminated all surfaces of the clean hood, pipettes, and tube racks with DNAZap (PCR DNA Degradation Solutions, Thermo Fisher Scientific, USA), followed by Cavicide (Metrex Research, Orange, CA), and an 80% ethanol rinse prior to each use. We autoclaved all tubes, tube racks and the waste bin on a dry vacuum cycle for 45min. Finally, we used sealed sterile filter tips, new nuclease-free water (Qiagen, USA) for each experiment, and filtered all salt solutions through a 30mm syringe filter with 0.22µm pore size (Argos Technologies, USA) before use in each experiment.

Whole Genome Amplification

Standard MALBAC**:** The MALBAC assay was originally designed for human cells (Zong et al., 2012). This approach involved making double stranded DNA copies of genomic material using random primers that consist of 5 degenerate bases and 27 bases of common sequence. These linear cycles are followed by exponential amplification via suppression PCR. Here, we transferred individual cells into sterile thin-wall PCR tubes containing 2.5µl of lysis buffer that yielded a final concentration of 25mM Tris pH 8.8 (Sigma-Aldrich, USA), 10mM NaCl (BAKER ANALYZED A.C.S. Reagent, J.T.Baker, USA), 10mM KCl (ACS reagent, Sigma-Aldrich, USA), 1mM EDTA (molecular biology grade, Promega, USA), 0.1% Triton X-100 (Acros Organics, USA), 1mg/ml Proteinase K (*Tritirachium album,* Sigma-Aldrich, USA). After overlaying one drop of mineral oil, we lysed cells at 50°C for 3h and inactivated the proteinase at 75°C for 20min, then 80°C for 5min before maintaining at 4°C. We added 2.5µl of amplification buffer to each sample to yield a final concentration of 25mM Tris pH 8.8 (Sigma-Aldrich, USA), 10mM $(NH_4)_2SO_4$ (Molecular biology grade, Sigma-Aldrich, USA), 8mM $MgSO_4$ (Fisher BioReagents, Fisher Scientific), 10mM KCl (ACS reagent, Sigma-Aldrich, USA), 0.1% Triton X-100 (Acros Organics, USA), 2.5mM dNTP's (PCR grade, Thermo Fisher Scientific, USA), 1M betaine (PCR Reagent grade, Sigma-Aldrich, USA) and 0.667µM of each random primer (5'GTGAGTGATGGTTGAGGTAGTGTGGAGNNNNNTTT 3', and 5'GTGAGTGATGGTTGAGGTAGTGTGGAGNNNNNGGG 3') ordered from Integrated DNA Technologies, USA. To denature DNA, we heated samples to 95°C for 3min and snap-cooled on an ice slush before gently adding 0.5µl of enzyme solution (8,000 U/ml *Bst* DNA Polymerase Large Fragment, New England Biolabs, USA, in 1X amplification buffer) into the aqueous droplet.

We thermo-cycled samples (Bio-Rad, USA) holding at 4°C and heated according to the following cycles: 10°C – 45s, 15°C – 45s, 20°C – 45s, 30°C – 45s, 40°C – 45s, 50°C – 45s, 64°C – 10min, 95°C – 20s. The samples were immediately snap-cooled on an ice slush and held for at least 3min to maintain the DNA in a denatured state for the next round of random priming. We added another 0.5µl of enzyme solution and mixed thoroughly with a pipette on ice as above. We placed the samples back into the 4°C thermo-cycler and heated according to the cycles listed above with an additional 58°C step for 1min before once again cooling on an ice slush for 3min. We repeated the addition of enzyme mix (as above) and performed additional rounds of amplification cycles (as above, including the 58°C step). Once completed, we placed the samples on ice and supplemented with cold PCR master mix to yield 50µl with the following concentrations: 0.5µM Common Primer (5'GTGAGTGATGGTTGAGGTAGTGTGGAG3', Integrated DNA Technologies, USA), 1.0mM dNTPs (PCR grade, Thermo Fisher Scientific, USA), 6.0mM $MgCl_2$ (Molecular biology, Sigma-Aldrich, USA), 1X Herculase II Polymerase buffer and 1X Herculase II polymerase (Agilent Technologies, USA). We immediately thermo-cycled samples with the following temperature-time profile: 94°C – 40s, 94°C – 20s, 59°C – 20s, 68°C – 5min, go to step two for several times, and an additional extended at 68°C – 5min, and finally, a hold at 4 °C. For comparison, we used 18/19 linear cycles and 17 exponential cycles for single parasite genomes amplified by the standard MALBAC protocol.

Optimized MALBAC

We made the following modifications to standard MALBAC to produce our improved method. **1)** We froze cells at -80°C until usage because freeze-thaw enhanced cell lysis as previously reported (Trevino et al., 2017); **2)** We removed betaine from the amplification buffer because it improved amplification of GC-rich sequences (Jensen et al., 2010), which are infrequent in *P. falciparum* genomes (**Supplementary Table 4.1**); **3)** We used a single random primer where the GC-content of the degenerate bases were 20% instead of 50% (5'GTGAGTGATGGTTGAGGTAGTGTGGAGNNNNNTTT 3') at final concentration of 1.2µM; **4)** We reduced the volume of the random priming reaction by added only 0.29µl of 2X amplification buffer to the lysed samples and 0.13µl of enzyme solution to the aqueous droplet each amplification cycle; **5)** We added additional random priming cycles over prior MALBAC studies for a total of 18 (for late stage parasites) or 19 (for early stage parasites) cycles; **6)** We reduced the total volume of exponential amplification from 50µl to 20µl and increased the number of exponential amplification cycles from 15 to 17; **7)** We verified the presence of high molecular weight DNA products in the samples before purifying nucleic acids by Zymo DNA Clean & Concentrator-5 (ZYMO Research).

Further Optimized MALBAC

More recently, we made the following modifications to optmized MALBAC to further improve our method. **1)** modifying the pre-amplification random primer by adding 5 more degenerate bases with 20% GC-content to limit the potentially preferential annealing towards more GC-balanced DNA through the common sequence in front of degenerate bases (5'GTGAGTGATGGTTGAGGTAGTGTGGAGNNNNNNNNNNTTT 3'); 2) replacing Bst

DNA Polymerase Large Fragment with Bsu DNA Polymerase Large Fragment with a lower reaction temperature (37°C) to improve the amplification on AT-rich sequences; 3) the first pre-amplification cycle is changed to the following: 10°C – 45s, 15°C – 45s, 20°C – 45s, 30°C – 45s, 37°C – 10 min, 50°C – 45s, 64°C – 45s, 95°C – 20s; while the rest of the pre-amplification used the following condition: 10°C – 45s, 15°C – 45s, 20°C – 45s, 30°C – 45s, 37°C – 10min, 50°C – 45s, 64°C – 45s, 95°C – 20s, 58°C– 1min ; **4)** We added additional random priming cycles over prior optimized MALBAC studies for a total of 20 (for late stage parasites) or 21 (for early stage parasites) cycles.

Integrating robotic pipetting into single cell sequencing pipeline

To improve the throughput of sequencing and limit contamination due to manual pipetting, we integrated robotic pipetting (Mosquito LV) into our single cell sequencing pipeline. Since full-skirted 96 well plated is needed for Mosquito LV, we changed our cell isolation method to FACS sorting (Sony SH800) from CellRaft AIR™ System, which is incompatible with full-skirted 96 well plate. We stained *P. falciparum*-infected erythrocytes with both SYBR green and MitoTracker Red CMXRos (Invitrogen, USA) and dilute the cell to 1 x 10^7 cells/ml. The diluted cells were then sorted into 96-well plate (Armadillo high performance 96 well plate, Thermo Scientific, USA) containing 2.5 µl MALBAC lysis buffer in each well using Sony SH800 cell sorter (Sony Biotechnology) with 100μm cell sorting chip following the manufacturer's instructions. The same MALBAC amplification process is used except utilizing robotic pipetting during MALBAC pre-amplification cycles instead of hand pipetting. Mosquito LV (SPT Labtech, Royston, UK) was used to pipetting the 0.13 µl enzyme solution in the MALBAC pre-amplification cycles. Specifically, for the pre-amplification, the enzyme solution was set as source in column 12 at deck location P2, while the destination plate was set at deck location P4. The tips were set to change after each pipetting and the changing location is on deck location P3 column 6. To test whether the Mosquito LV pipetting works for our single cell sequencing pipeline, we sequenced 3 HB3 single cell samples and 4 Dd2 single cell samples.

Pre-Sequencing Quality Assessment

We assayed 6 distinct genomic loci across different chromosomes to determine variations in copy number following the whole genome amplification step. We included this step, which employs highly sensitive droplet digital PCR (ddPCR, QX200 Droplet Digital PCR system, Bio-Rad, USA), to identify samples that exhibited more even genome coverage prior to short-read sequencing. The sequence of primers and probes are described in **Supplementary Table 4.2** (Pickard et al., 2003; Perandin et al., 2004; Guler et al., 2013). Each ddPCR reaction contained 5µl of DNA (0.3ng/µl for single cell samples), 10µl ddPCR Supermix for Probes (without dUTP), primers and probes with the final concentration in **Supplementary Table 4.2**, and sterile H_2O to bring the per-reaction volume to 22µl. We prepared droplets with the PCR mixture following the manufacture's protocol: 95°C – 10 min; 40 cycles of 95°C – 30s, 60°C – 60s, and an infinite hold at 4°C. After thermal cycling, we counted positive droplets using the Bio-Rad QX200 Droplet Reader (Bio-Rad, USA). We analyzed data through QuantaSoft (Bio-Rad, USA). For each gene, a no template control (sterile water, NTC) and a positive control (0.025ng *Dd2* genomic DNA) are included in each ddPCR run. Following ddPCR, we calculated the "uniformity score" using the locus representation of the 6 genes: *seryl tRNA synthetase* (gene-1, PF3D7_0717700)*, heat shock protein 70* (gene-2, PF3D7_0818900)*, dihydrofolate reductase* (gene-3, PF3D7_0417200)*, lactate dehydrogenase* (gene-4, PF3D7_1324900)*, 18S* *ribosomal RNA* (gene-5, PF3D7_0112300, PF3D7_1148600, PF3D7_1371000)*,* and *multi-drug resistance transporter 1* (*Pfmdr1*, gene-6, PF3D7_0523000) in the amplified DNA sample relative to non-amplified DNA using the following equation:

$$
\begin{aligned}
Uniformityscore = {} & \frac{gene1}{gene2}+\frac{gene1}{gene3}+\frac{gene1}{gene4}+\frac{gene1}{gene5}+\frac{gene1}{gene6}+\frac{gene2}{gene3}+\frac{gene2}{gene1}+\frac{gene2}{gene4}+\frac{gene2}{gene5} \\
& +\frac{gene2}{gene6}+\frac{gene3}{gene4}+\frac{gene3}{gene1}+\frac{gene3}{gene2}+\frac{gene3}{gene5}+\frac{gene3}{gene6}+\frac{gene4}{gene5}+\frac{gene4}{gene1}+\frac{gene4}{gene2}+\frac{gene4}{gene3}+\frac{gene4}{gene6} \\
& +\frac{gene5}{gene1}+\frac{gene5}{gene2}+\frac{gene5}{gene3}+\frac{gene5}{gene4}+\frac{gene5}{gene6}+\frac{gene6}{gene1}+\frac{gene6}{gene2}+\frac{gene6}{gene3}+\frac{gene6}{gene4}+\frac{gene6}{gene5}
\end{aligned}
$$

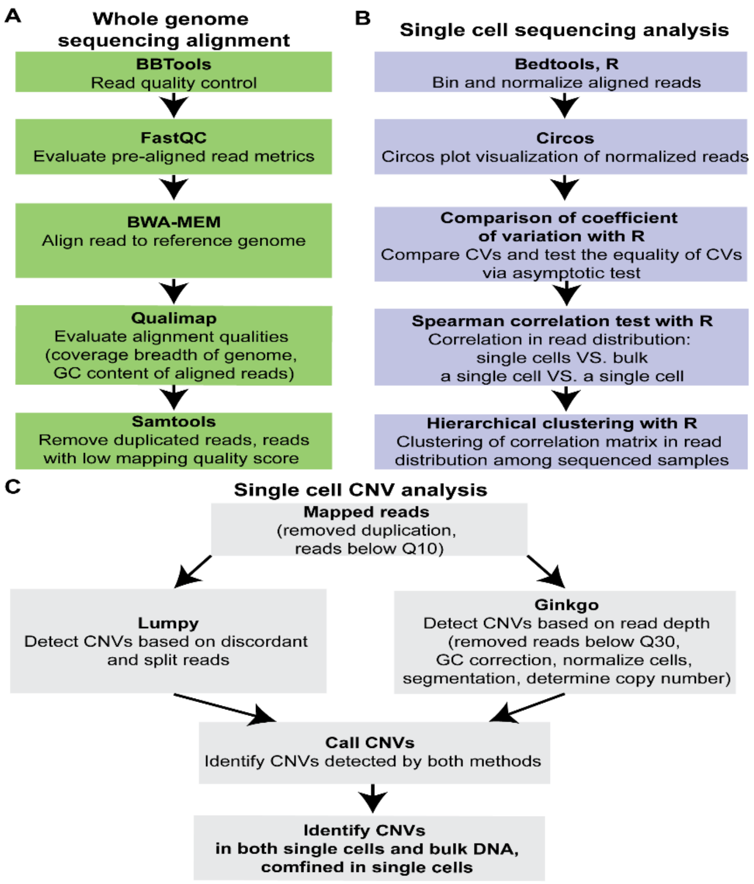
When certain loci were over- or under-represented in the amplified sample, this score increased above the perfect representation of the genome; a uniformity score of 30 indicates that all genes are equally represented. We calculated the locus representation from the absolute copies of a gene measured by ddPCR from 1ng of amplified DNA divided by the absolute copies from 1ng of the bulk DNA control (Dean et al., 2002). We only included samples in which all six genes were detected by ddPCR. The relative copy number of the *Pfmdr1*, which was amplified in the *Dd2* parasite line (Cowman et al., 1994, 1), was also used to track the accuracy of amplification. We calculated this value by dividing the ddPCR-derived absolute copies of *Pfmdr1* by the average absolute copies of the 6 assayed loci (including *Pfmdr1*, normalized to a single copy gene*)*. To confirm the efficiency of ddPCR detection as a pre-sequencing quality control step, we determined the strength of association based on the pattern of concordance and discordance between the ddPCR detection and the sequencing depth of the 5 gene targets with Kendall rank correlation (*18S ribosomal RNA* was excluded from correlation analysis due to the mapping of non-unique reads). We then calculated the correlation coefficient (**Supplementary Table 4.3**). When the level of ddPCR detection corresponded to the sequencing depth in at least 3 of the 5 gene targets (a correlation coefficient of >0.6), we regarded the two measurements as correlated.

Short-Read Sequencing

We fragmented MALBAC amplified DNA (>1ng/µL, 50µL) using Covaris M220 Focused Ultrasonicator in microTUBE-50 AFA Fiber Screw-Cap (Covaris, USA) to a target size of 350bp using a treatment time of 150s. We determined the fragment size range of all sheared DNA samples (291bp-476bp) with a Bioanalyzer on HS DNA chips (Agilent Technologies, USA). We used the NEBNext Ultra DNA Library Prep Kit (New England Biolabs, USA) to generate Illumina sequencing libraries from sheared DNA samples. Following adaptor ligation, we applied 3 cycles of PCR enrichment to ensure representation of sequences with both adapters and the size of the final libraries range from 480bp to 655bp. We quantified the proportion of adaptor-ligated DNA using real-time PCR and combined equimolar quantities of each library for sequencing on 4 lanes of an Illumina Nextseq 550 using 150bp paired end cycles. We prepared the sequencing library of clinical bulk DNA as above but sequenced it on an Illumina Miseq using 150bp paired end sequencing.

Sequencing Analysis

We performed read quality control and sequence alignments essentially as previously described (Huckaby et al., 2018) (**Figure S4.2A**). The codes are accessible through Github: https://github.com/Pfal-analysis/Single-cell-sequencing-data. Briefly, we removed Illumina adapters and PhiX reads, and trimmed MALBAC common primers from reads with BBDuk tool in BBMap (Bushnell, 2016). To identify the source of DNA reads, we randomly subsetted 10,000 reads from each sample by using the reformat tool in BBMap and blasted reads in nucleotide database using BLAST+ remote service. We aligned each fastq file to the hg19 human reference genome and kept the unmapped reads (presumably from *P. falciparum*) for analysis. Then, we aligned each fastq file to the *3D7 P. falciparum* reference genome with Speedseq (Chiang et al., 2015). We discarded the reads with low-mapping quality score (below 10) and duplicated reads using Samtools (Li et al., 2009). To compare the coverage breadth (the percentage of the genome that has been sequenced at a minimum depth of one mapped read (Sims et al., 2014)) between single cell samples, we extracted mappable reads from BAM files using Samtools and randomly downsampled to 300,000 reads using the reformat tool in BBMap. This level is dictated by the sample with the lowest number of mappable reads (ENM, **Supplementary Table 4.4**). We calculated the coverage statistics using Qualimap 2.0 (García-Alcalde et al., 2012) for the genic, intergenic, and whole genome regions.


A
Whole genome sequencing alignment
BBTools
Read quality control
FastQC
Evaluate pre-aligned read metrics
BWA-MEM
Align read to reference genome
Qualimap
Evaluate alignment qualities (coverage breadth of genome, GC content of aligned reads)
Samtools
Remove duplicated reads, reads with low mapping quality score
B
Single cell sequencing analysis
Bedtools, R
Bin and normalize aligned reads
Circos
Circos plot visualization of normalized reads
Comparison of coefficient of variation with R
Compare CVs and test the equality of CVs via asymptotic test
Spearman correlation test with R
Correlation in read distribution:
single cells VS. bulk
a single cell VS. a single cell
Hierarchical clustering with R
Clustering of correlation matrix in read distribution among sequenced samples
C
Single cell CNV analysis
Mapped reads
(removed duplication, reads below Q10)
Lumpy
Detect CNVs based on discordant and split reads
Ginkgo
Detect CNVs based on read depth (removed reads below Q30, GC correction, normalize cells, segmentation, determine copy number)
Call CNVs
Identify CNVs detected by both methods
Identify CNVs
in both single cells and bulk DNA, comfined in single cells

Supplementary Figure 4.2 Bioinformatic analysis of sequencing reads
A. Whole genome sequencing analysis and alignment. Alignment of reads started with BBTools to remove low quality bases, adapter sequences, trim common sequence of MALBAC primer and verify correct pairing of reads. The resulting "clean" paired reads were evaluated by FastQC. After passing read quality control, BWA-MEM was used to align "clean" paired reads to *3D7 Plasmodium falciparum* reference genome. Qualimap was then used to evaluate the alignments for coverage breadth, GC-content, and mapping quality. Duplicated reads and reads with mapping quality score below 10 were removed for downstream analysis by Samtools. **B. Single cell sequencing analysis steps.** The reference genome was divided into 20kb bins by Bedtools. Read counts was calculated in every 20kb bins and normalized by the mean read count with Bedtools and R. Circos was utilized to visualize the distribution of normalized read counts over 14 chromosomes. **C. CNV analysis steps.** LUMPY were used for CNV detection with at least two supporting reads. After mapped reads were further filtered by a mapping quality score of 30, Ginkgo was used to detect CNVs based on read depth across 1kb, 5kb, 8kb, 10kb bins; steps included normalization, GC content correction, independent segmentation, and copy number determination. CNVs were called by identifying those that are shared between LUMPY and Ginkgo calls (see *Methods* for details).

To understand where the primers of MALBAC amplification are annealing to the genome, we overlaid information on the boundaries of genic or intergenic regions with the mapping position of reads containing the MALBAC primer common sequence. To do so, we kept the MALBAC common primers in the sequencing reads, filtered reads and aligned reads as in the above analysis. We subsetted BAM files for genic and intergenic regions using Bedtools, searched for the MALBAC common primer sequence using Samtools, and counted reads with MALBAC common primer using the pileup tool in BBMap (**Supplementary Table 4.5**).

We conducted single cell sequencing analysis following the steps in **Figure S4.2B.** We compared the variation of normalized read abundance (log10 ratio) at different bin sizes using boxplot analysis (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 then divided the *P. falciparum* genome into non-overlapping 20 kb bins using Bedtools (Quinlan and Hall, 2010). The normalized read abundance was the mapped reads of each bin divided by the total average reads in each sample. To show the distribution of normalized read abundance along the genome, we constructed circular coverage plots using Circos software and ClicO FS (Krzywinski et al., 2009; Cheong et al., 2015). To assess uniformity of amplification, we calculated the coefficient of variation of normalized read abundance by dividing the standard deviation by the mean and multiplying by 100 (Chen et al., 2017) and analyzed the equality of coefficients of variation using the R package “cvequality” version 0.2.0 (B. Marwick and K. Krishnamoorthy). We employed correlation coefficients to assess amplification reproducibility as previous studies (Chen et al., 2018). Due to presence of non-linear correlations between some of the samples, we used Spearman correlation for this analysis. We removed outlier bins if their read abundance was above the highest point of the upper whisker (Q3 + 1.5×interquartile range*)* or below the lowest point of the lower whisker (Q1-1.5×interquartile range) in each sample. Then, we subsetted remaining bins present in all samples to calculate the correlation coefficient using the R package "Hmisc" version 4.3-0 (Harrell F. E.). We visualized Spearman correlations, histograms and pairwise scatterplots of normalized read abundance using "pairs.panels" in the "psych" R package. We then constructed heatmaps and hierarchical clustering of Spearman correlation coefficient with the "gplots" R package version 3.0.1.1 (Gregory R. Warnes et al.). Additionally, to estimate the chance of random primer annealing during MALBAC pre-amplification cycles (likely affected by the GC content of genome sequence), we generated all possible 5-base sliding windows with 1 base step-size in the *P. falciparum* genome and calculated the GC-content of the 5-bases windows using Bedtools (**Supplementary Table 4.1**).

Copy Number Variation Analysis

We conducted CNV analysis following the steps in **Figure S4.2C**. To ensure reliable CNV detection, our CNV analysis is limited to the core genome, as defined previously (Otto et al., 2018). Specifically, we excluded the telomeric, sub-telomeric regions and hypervariable *var* gene clusters, due to limited mappability of these regions. For discordant/split read analysis, we used LUMPY (Layer et al., 2014) in Speedseq to detect CNVs (>500 bp) with at least two supporting reads in each sample (**Supplementary Table 4.6**). For read-depth analysis, we further filtered the mapped reads using a mapping quality score of 30. We counted the reads in 1kb, 5kb, 8kb, 10kb bins by Bedtools and used Ginkgo (Garvin et al., 2015) to normalize (by dividing the count in each bin by the mean read count across all bins), correct the bin read counts for GC bias, independently segment (using a minimum of 5 bins for each segment), and determine the copy number state in each sample with a predefined ploidy of 1 (**Supplementary Table 4.7**). The quality control steps of Ginkgo were replaced by the coefficient of variation of normalized read count used in this study to assess uniformity in each cell. Lastly, we identified shared CNVs from the two methods when one CNV overlapped with at least 50% of the other CNV and vice versa (50% reciprocal overlap). We calculated precision of CNV detection in single cell genome by dividing the number of true CNVs (same as those detected in the bulk sample) by the total number of CNVs. We calculated sensitivity by dividing the number of true CNVs by 3 (total number of true CNVs in the bulk sample).

Single Nucleotide Polymorphism Analysis

We conducted SNP analysis following the MalariaGen *P. falciparum* Community Project V6.0 pipeline (MalariaGEN et al., 2021; MalariaGEN P. *falciparum* Community Project V6.0 pipeline) based on GATK best practices (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013). We first applied GATK's Base Quality Score Recalibration using default parameters. We used GATK's HaplotypeCaller to detect potential SNPs in BAM files and genotyped them using GATK's CombineGVCFs and GenotypeGCVFs. We ran GATK's VariantRecalibrator using previously validated SNP set from the *Pf*-Crosses variant set as a training set (Miles et al., 2016).We then applied GATK's ApplyRecalibration to assign each SNP a VQSLOD quality score, which uses a machine learning approach to assess the probability that raw SNPs are true variants based on the training set. Higher VQSLOD scores indicate higher quality SNPs; filtering SNPs by "VQSLOD score > 0" has been applied to variant detection studies using the GATK pipeline (Hamilton et al., 2017; MalariaGEN et al., 2021), whereas

VQSLOD score > 6 is recommended to further improve SNP accuracy in *P. falciparum* specifically *(MalariaGEN et al., 2021)*. We calculated precision by dividing the number of called SNP variants with the same genotype as the standard data set (SNPs detected in the *Dd2* bulk sample) by the total number of SNP variants called in each single cell sample. We calculated sensitivity by dividing the number of called SNP variants with the same genotype as the standard SNPs in single cell samples by the number in the bulk standard SNPs at three different stringency levels: VQSLOD score > 0, VQSLOD score > 6, and VQSLOD score > 6 with read depth > 10. We only included bi-allelic SNPs (loci with either the wild type or one mutant type allele) from the core genome in this analysis (MalariaGEN et al., 2021). We also evaluated the detection of SNPs in resistant genes of the *Dd2* parasite line. We successfully detected 16 out of 17 resistant SNPs in the bulk sample at VQSLOD >6; the one remaining SNP failed to pass the filtering step (VQSLOD = 3.77) so we excluded it from all single cell analyses. We further filtered novel SNPs in single cell samples by removing those that exhibited multiple alleles (mixed allele SNPs). We utilized SnpEff (Cingolani et al., 2012) to annotate VCF files and used VIVA (v0.4.0) (Tollefson et al., 2019) to generate heatmaps to illustrate the relationship between SNP calling and read depth.

4.4 Results

Plasmodium falciparum **genomes from single-infected erythrocytes are amplified by MALBAC.**

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 4.1A**). We collected parasites from either an *in vitro*-propagated laboratory line (*Dd2*) 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 *Dd2* parasite 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 microscopy-based CellRaft Air system (**Figure 4.1B**), which has the benefit of low capture volume (minimum: 2μl) and visual confirmation of parasite stages. Following isolation, using the standard MALBAC protocol (termed non-optimized MALBAC), we successfully amplified 3 early (ENM) and 4 late stage (LNM) laboratory *Dd2* parasite samples. 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 *Dd2* parasite samples as well as 4 clinical samples (COM) (**Supplementary Table 4.8**). Compared to standard MALBAC, our optimized protocol has a lower reaction volume, more amplification cycles, and a modified pre-amplification random primer (see *Methods* for more details). Using this method, we successfully amplified 43% of the early and 90% of the late stage laboratory *Dd2* parasite samples and 100% of the clinical samples (see post-amplification yields in **Supplementary Tables 4.8 and 4.9**).

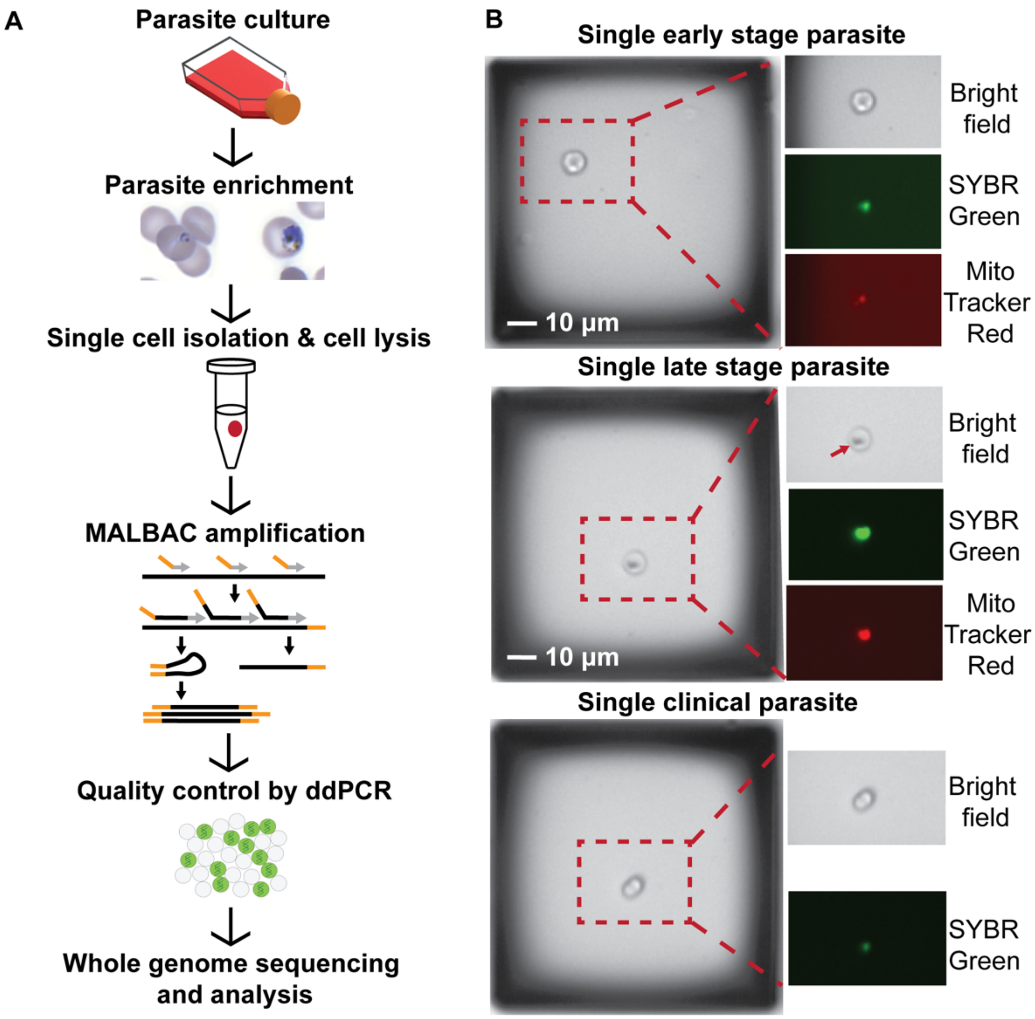


Figure 4.1 Single *P. falciparum*-infected erythrocytes are isolated, amplified, and sequenced A. Experimental workflow. Parasites are grown *in vitro* in human erythrocytes or isolated from infected patients. To limit the number of uninfected erythrocytes in the sample, infected cells are enriched using column and gradient-based methods (see *Methods*). Individual early-stage (left image) and late-stage (right image) parasite-infected erythrocytes were automatically isolated into PCR tubes using the CellRaft AIR System (Cell Microsystems, see panel B). All cells were lysed and amplified by MALBAC. MALBAC uses a combination of common (orange) and degenerate (grey) primers to amplify the genome. The quality of amplified genomes was assessed prior to library preparation and sequencing using droplet digital (dd)PCR; DNA is partitioned into individual droplets to measure gene copies. Suitable samples were Illumina sequenced and analyzed as detailed in Figure S4.2. **B. Parasite stage visualization on the CellRaft AIR System using microscopy (10X magnification).** Enriched early and late stage parasite-infected erythrocytes at low density 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 (indicates 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 10µm.

Pre-sequencing quality control step identifies samples with more even genome amplification.

We assessed the quality of WGA products from single cells using droplet digital PCR (ddPCR) to measure the copy number of known single and multi-copy genes dispersed across the *P. falciparum* genome (6 genes in total including *Pfmdr1*, which is present at ~3 copies in the *Dd2* laboratory parasite line). Using this sensitive quantitative method, along with calculation of a "uniformity score" which reflects both locus dropout and over-amplification, we were able to select genomes that had been more evenly amplified; a low uniformity score and accurate copy number values indicated a genome that has been amplified with less bias (see *Methods* for details on score calculation and **Supplementary Table 4.10** for raw data). This quality control step was important to reduce unnecessary sequencing costs during single cell studies.

When we analyzed differences between amplified samples by optimized MALBAC (17 EOM samples and 14 LOM samples processed for ddPCR evaluation) and non-optimized MALBAC (3 ENM and 4 LNM samples), we found that samples amplified with the optimized protocol were generally more evenly covered than those from the standard method (**Table 4.1**). Specifically, one ENM sample lacked detection of any of the target genes (likely due to heavy contamination from non-parasite DNA) and other ENM and LNM samples consistently showed over-amplification of a set of 2 genes (*P. falciparum* seryl-tRNAsynthetase and 18S ribosomal RNA; **Supplementary Table 4.10**). Therefore, due to evidence of a high level of bias in the majority of ENM/LNM samples, we selected the ENM and LNM samples (one each) with the lowest level of ddPCR-based bias for sequencing. We also used ddPCR results to select 13 EOM and 10 LOM samples for sequencing (**Supplementary Table 4.8**). Overall, selected samples had lower average uniformity scores (i.e. 248 and 1012 for selected and unselected EOMs, respectively, see **Table 4.1**). For clinical parasite samples, 3 out of 4 COM samples showed a lack of ddPCR detection on at least one parasite gene; thus, we were not able to calculate uniformity scores for these samples and the amplification of clinical genomes was likely more skewed than laboratory samples (**Table 4.1**).

Table 4.1 Pre-sequencing quality control by droplet digital PCR

Result	Sample type	MALBAC type	Sample name (#)	Pre-sequencing ddPCR assessment	
				Uniformity score AVG (SD)*	PfMDR1 CN AVG (SD)
Sequenced	Single cell	Optimized	EOM (13)	248 (202)	2.6 (0.8)
			LOM (10)	118 (69)	2.2 (1.3)
			COM (2)	369 (-)	1.9 (0.8)
		Non-optimized	ENM (1)	18519 (-)	0.2 (-)
			LNM (1)	13121 (-)	0.1 (-)
	Bulk	N/A	*Dd2*_Bulk (1)	30	2.7
			Clinical_Bulk (1)	-	-
Not Sequenced	Single cell	Optimized	EOM (4)	1012 (195)	3.7 (3.9)
			LOM (4)	775 (683)	2.8 (2.1)
			COM (2)	-^ (-)	4.7 (6.6)
		Non-optimized	ENM (2)	13689 (-)	0 (-)
			LNM (3)	1578 (-)	0.1 (0.1)

EOM: Early stage single parasites amplified by optimized MALBAC; LOM: Late stage single parasites amplified by optimized MALBAC; COM: Clinical single parasites amplified by optimized MALBAC; ENM: Early stage single parasites amplified by non-optimized MALBAC; LNM: Late stage single parasites amplified by non-optimized MALBAC.

Both standard and optimized MALBAC-amplified parasite genomes were short-read sequenced alongside a matched bulk DNA control (**Table 4.1**). To confirm the efficiency of ddPCR detection as a pre-sequencing quality control step, we calculated the correlation between ddPCR quantification and the sequencing depth at these specific loci in each sample. We found that the ddPCR-derived gene copy concentration was correlated with sequencing coverage of the corresponding genes in many samples (**Supplementary Table 4.3**, 17 out of 28 samples with a Kendal rank correlation coefficient >= 0.6), confirming the validity of using ddPCR detection as a quality control step prior to sequencing.

Optimized MALBAC limits contamination of single cell genomes.

After read quality control steps, we mapped the reads to the *P. falciparum 3D7* reference genome (see *Methods* and **Figure S4.2** for details). We first assessed the proportion of contaminating reads in our samples; NCBI Blast results showed that the majority of non-*P. falciparum* reads were of human origin (**Figure 4.2A**). The mean proportions of human reads in EOM samples (6.6%, SD of 3.2%) and LOM samples (4.3%, SD of 2.9%) were similar to that in the control bulk sample (7.4%, **Figure 4.2A**); in fact, a majority of optimized MALBAC samples were lower than the bulk level (14/23). Conversely, the proportion of human reads in ENM and LNM samples were substantially higher (81.8% and 18.9%, respectively). As shown in other studies (Auburn et al., 2011; Oyola et al., 2013), the clinical bulk DNA (81.9%) contained a much higher level of human contamination than the laboratory *Dd2* bulk DNA (7.4%). However, we found that the proportion of the human contaminating DNA in the two single cell COM samples was considerably lower (58.8% and 65.5%). The second most common source of contaminating reads was from bacteria such as *Staphylococcus* and *Cutibacterium*. The ENM sample exhibited a ~10-fold increase in the proportion of bacterial reads over averaged EOM samples (8.2% versus 0.8%, respectively) whereas the LNM samples showed the same proportion of bacterial reads as the averaged LOM samples (0.2%). These results indicated that the optimized MALBAC protocol exhibits lower amplification bias towards contaminating human and bacterial DNA in *P. falciparum* samples.

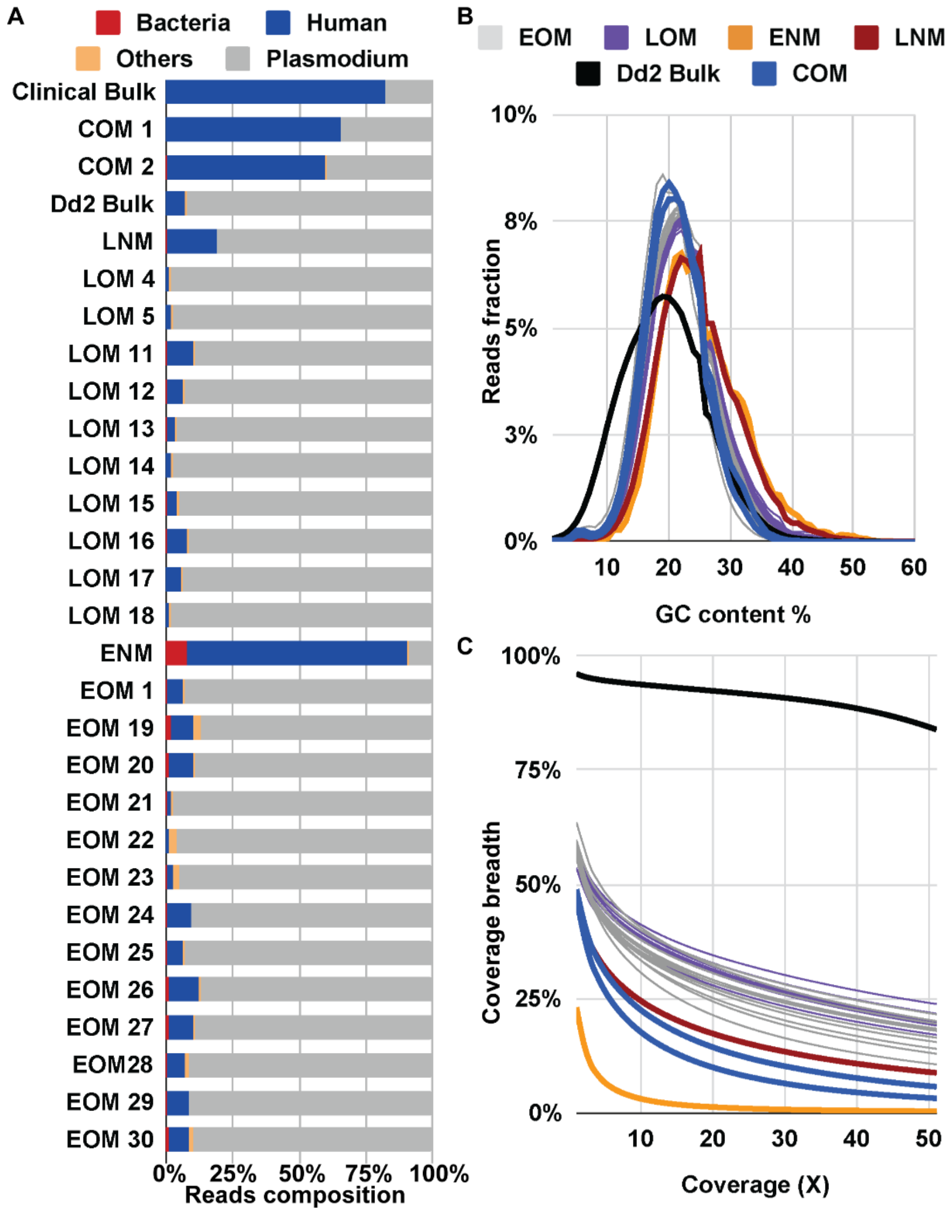


Figure 4.2 Sequencing statistics show benefits of optimized MALBAC
A. Contribution of reads based on organism type. A subset of 10,000 reads from each sample were randomly selected for BLAST to identify sources of DNA. Color representation: bacteria (red); human (blue); other organisms (orange); *Plasmodium* (grey). **B. GC-content of *P. falciparum* mapped reads.** GC-content of reads was calculated by Qualimap. Color representation: EOM (grey): Early stage single parasites amplified by optimized MALBAC; LOM (purple): Late stage single parasites amplified by optimized MALBAC; ENM (orange): Early stage single parasites amplified by non-optimized MALBAC; LNM (dark red): Late stage single parasites amplified by non-optimized MALBAC; *Dd2* bulk genomic DNA (black); COM samples (blue): Clinical single parasites amplified by optimized MALBAC. Clinical Bulk genomic DNA is not shown here due to <1% of the genome being covered by at least one read. **C. Fraction of *P. falciparum* genome covered by at least 1 read.** Color representations are the same as described in panel B.

Amplification bias and uniformity is altered in single cell genomes.

To further assess the optimized MALBAC protocol, we evaluated GC-bias at several steps of our pipeline (i.e. WGA, library preparation, and the sequencing platform). Analysis of the bulk genome control (without WGA) indicated that there was little GC-bias introduced by the library preparation, sequencing, or genome alignment steps; the GC-content of mapped reads from bulk sequencing data is 18.9% (**Table 4.2**), which was in line with the GC-content (19.4%) of the reference genome (Gardner et al., 2002). We then compared values from single cell samples to those from the appropriate bulk control to evaluate the GC-bias caused by MALBAC amplification (**Figure 4.2B**). The average GC-content of all EOM (21.4%), LOM (22.4%), and COM (20.7%) samples was within 1-3.5% of the bulk controls from laboratory *Dd2* and clinical samples (18.9% and 19.7%, respectively, **Table 4.2**). However, the average GC-content of ENM and LNM samples was 6.1% and 5.4% greater than that of the bulk control; this result is consistent with the high GC preference of the standard protocol (Hou et al., 2015; Ning et al., 2015). ENM and LNM samples also showed a greater proportion of mapped reads with high GC-content (>30%) than EOM, LOM, and bulk DNA samples (**Figure 4.2B**).

Table 4.2 Average GC-content and coverage breadth of sequenced samples

Reads	Sample name (#)	Average of mean coverage (X)	Average GC content	Average coverage breadth		
				Whole genome	Genic regions	Intergenic regions
All mappable reads	EOM (13)	37.54	21.4%	57.9%	78.0%	27.8%
	LOM (10)	43.10	22.4%	57.3%	79.0%	25.0%
	COM (2)	9.54	20.7%	48.0%	67.7%	18.5%
	ENM (1)	1.47	25.0%	23.0%	34.4%	6.1%
	LNM (1)	20.43	24.3%	47.4%	67.9%	16.9%
	*Dd2*_Bulk (1)	75.83	18.9%	96.1%	97.0%	94.9%
	Clinical_Bulk (1)	0.03	19.7%	0.3%	0.3%	0.2%
Down-sampled*	EOM (13)	1.66	21.4%	30.9%	47.2%	6.7%
	LOM (10)	1.69	22.4%	32.1%	49.8%	5.8%
	COM (2)	1.66	20.8%	31.1%	47.0%	7.5%
	ENM (1)	1.33	25.2%	21.7%	32.9%	5.0%
	LNM (1)	1.62	24.3%	26.2%	40.3%	5.1%
	*Dd2*_Bulk (1)	1.85	18.8%	76.8%	80.6%	71.2%

*Down-sampling is to 300,000 mappable reads based on the sample with the lowest number of mappable reads (ENM).

Since GC-bias during the amplification step can limit which areas of the genome are sequenced, we assessed the genome coverage of MALBAC-amplified samples. The coverage breadth of single cell samples increased by 34.9% in early stage samples (**Figure 4.2C**, orange-ENM to grey-EOM lines) and by 9.9% for late stage samples following optimization (**Figure 4.2C**, red-LNM to purple-LOM lines, see values in **Table 4.2**). Despite just a single ENM and LNM sample for comparison, the variation of coverage breadth across all EOM/LOM samples is low (**Supplementary Table 4.4**, SD of 1.9%), indicating that differences between the two methods are substantial. This pattern of differences is conserved despite random down-sampling of reads to the same number per sample (300,000; **Table 4.2**).

Even though optimized MALBAC showed less bias towards GC-rich sequences, it was still problematic for highly AT-rich and repetitive intergenic regions (mean of 13.6% GC-content). The fraction of intergenic regions covered by reads was only 27.8% for EOM samples and

25.0% for LOM samples on average. When we excluded intergenic regions, the fraction of genic regions of the genome covered by at least one read reached an average of 78.0% and 79.0% for EOM and LOM samples (**Table 4.2**). Conversely, the coverage of both intergenic and genic regions was substantially lower for the non-optimized samples. Coverage of the *P. falciparum* genome in the clinical bulk sample was very low due to heavy contamination with human reads (0.3% of the genome was covered by at least one read). This was much lower than that from the 2 COM samples (an average of 48%, **Figure 4.2C** and **Table 4.2**).

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 4.3A, Figure S4.3 and S4.4A**). 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 (**Figure S4.5**). To quantitatively measure this variation, we normalized the read abundance per bin in each sample by dividing the raw read counts with the mean read counts per 20kb bin (**Figure 4.3B, Figure S4.3C**). Here, the bulk control displayed the smallest range of read abundance for outlier bins (blue circles) and lowest interquartile range (IQR) value of non-outlier bins (black box, **Figure 4.3B, Figure S4.3C**), 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 4.3B, Figure S4.3C**). In addition, the read abundance variation of COM samples was similar to EOM or LOM samples (**Figure 4.3B, Figure S4.4B**). Due to the extremely low coverage of the clinical bulk sample, the read abundance variation was much higher than all other samples (**Figure 4.3B, Figure S4.4B**).

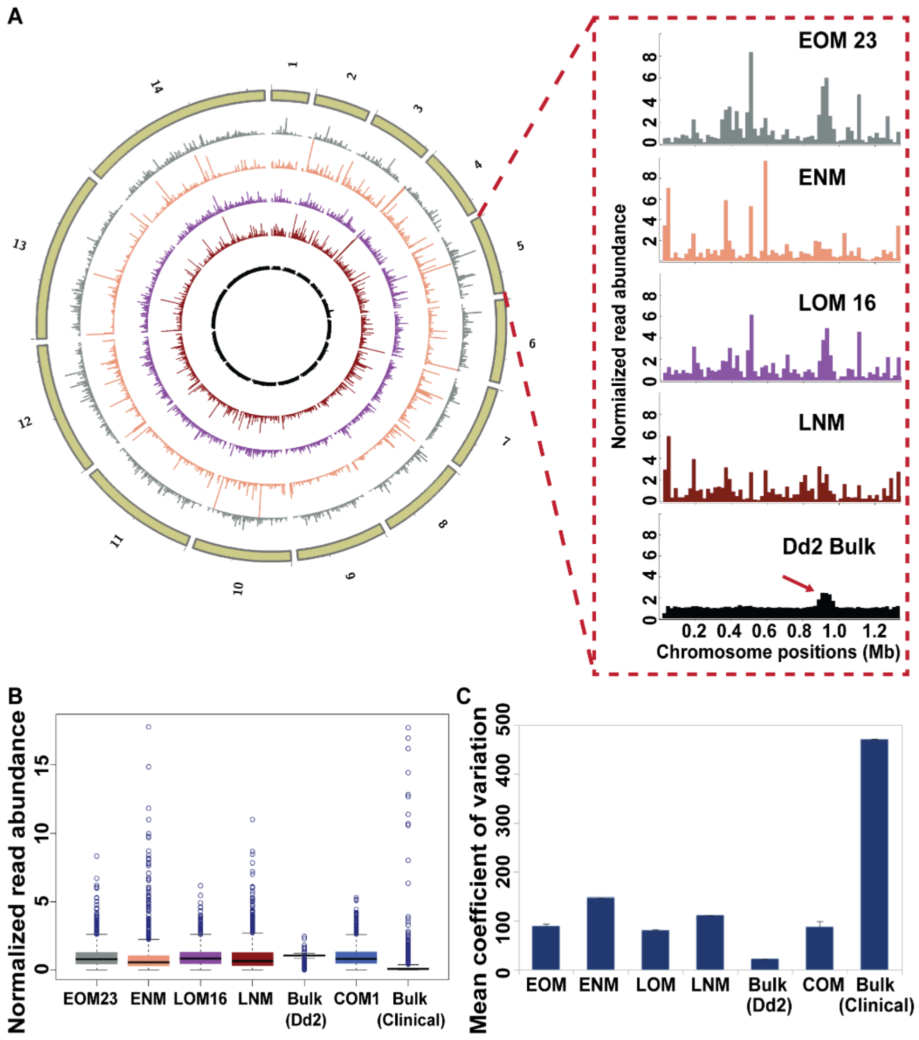
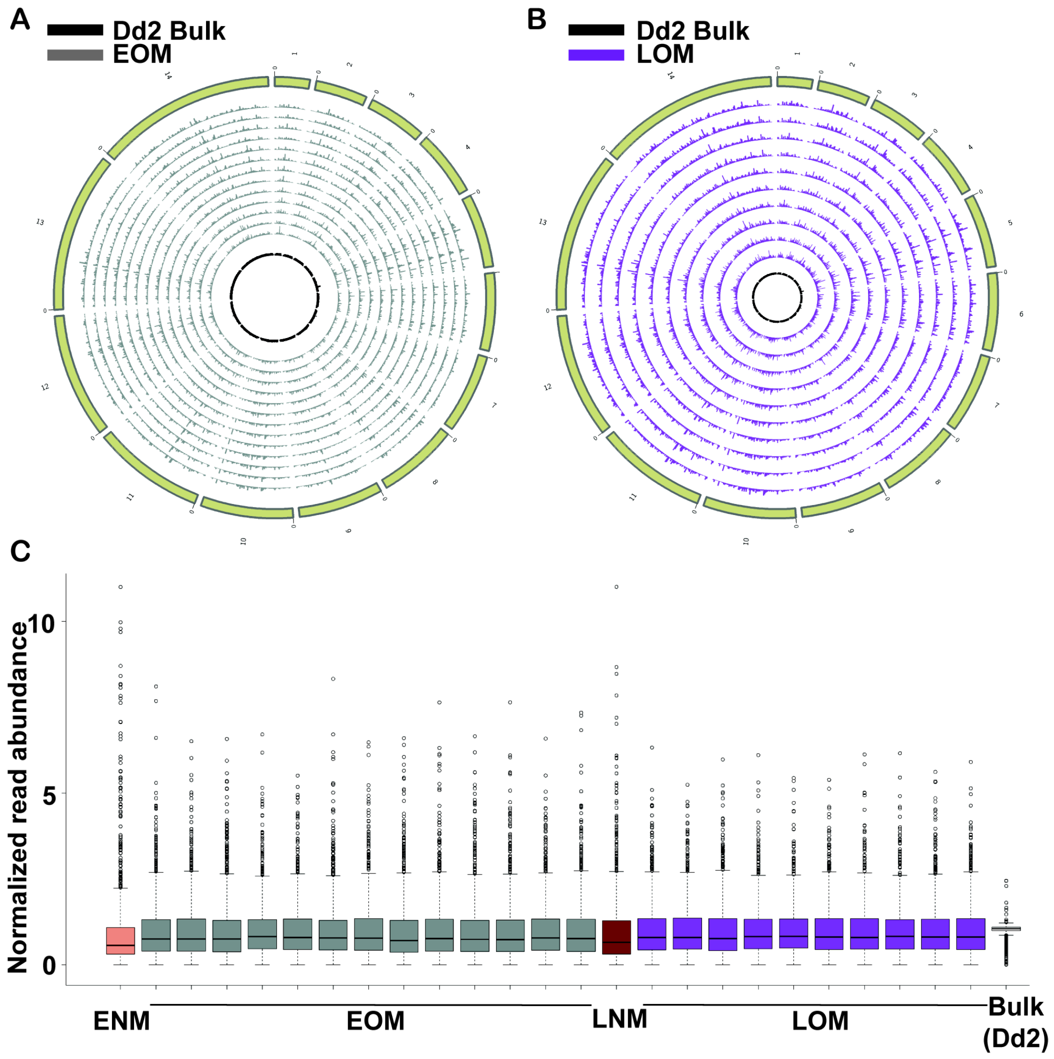


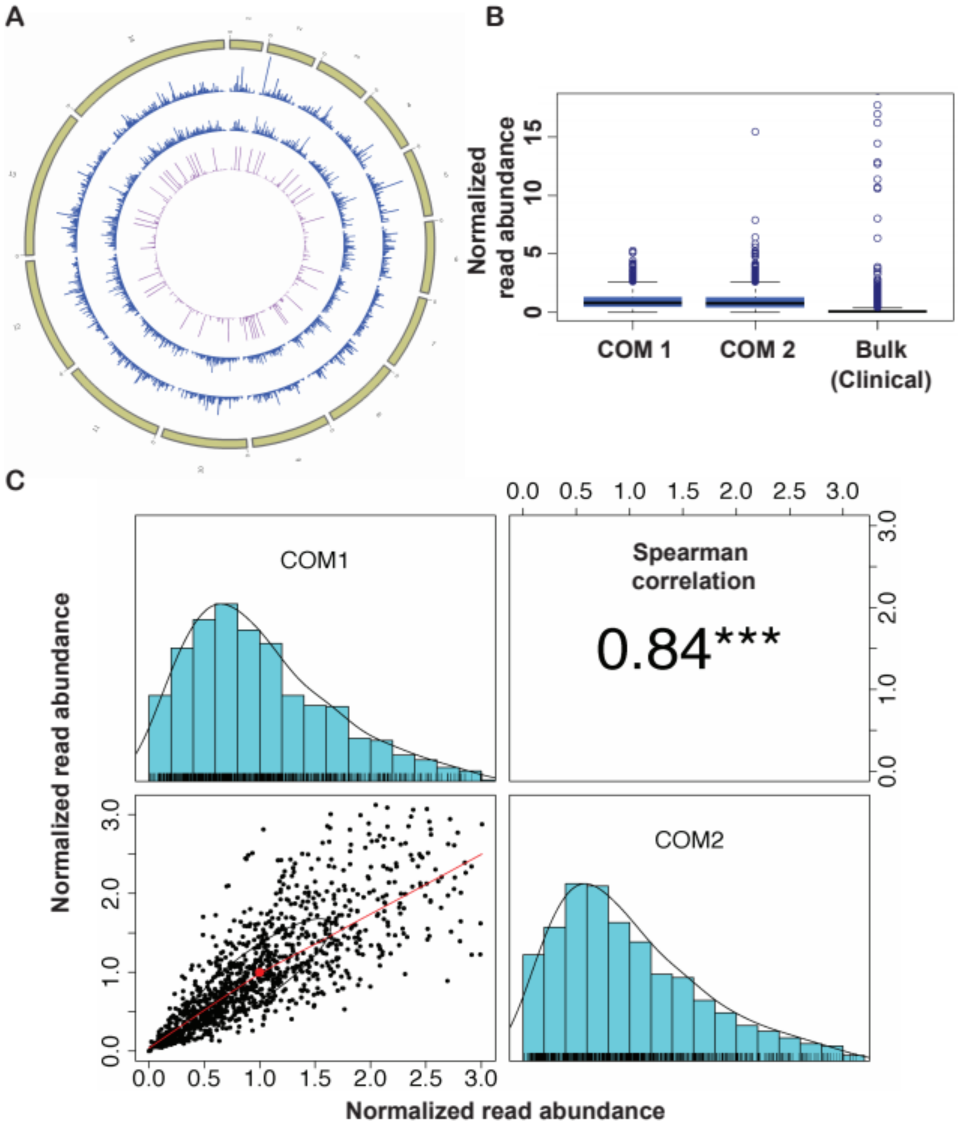
Figure 4.3 Samples amplified by optimized MALBAC display improved uniformity of read abundance

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 contains a known CNV (*Pfmdr1* amplification, arrow on *Dd2* bulk sample). **B. Distribution of normalized read abundance values for all bins.** The boxes were drawn from Q1 (25th percentiles) to Q3 (75th percentiles) with a horizontal line drawn in the middle to denote the median of normalized read abundance for each sample. 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 depicted with circles. One sample from each type is represented (see all samples in Figure S4.3C). **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.



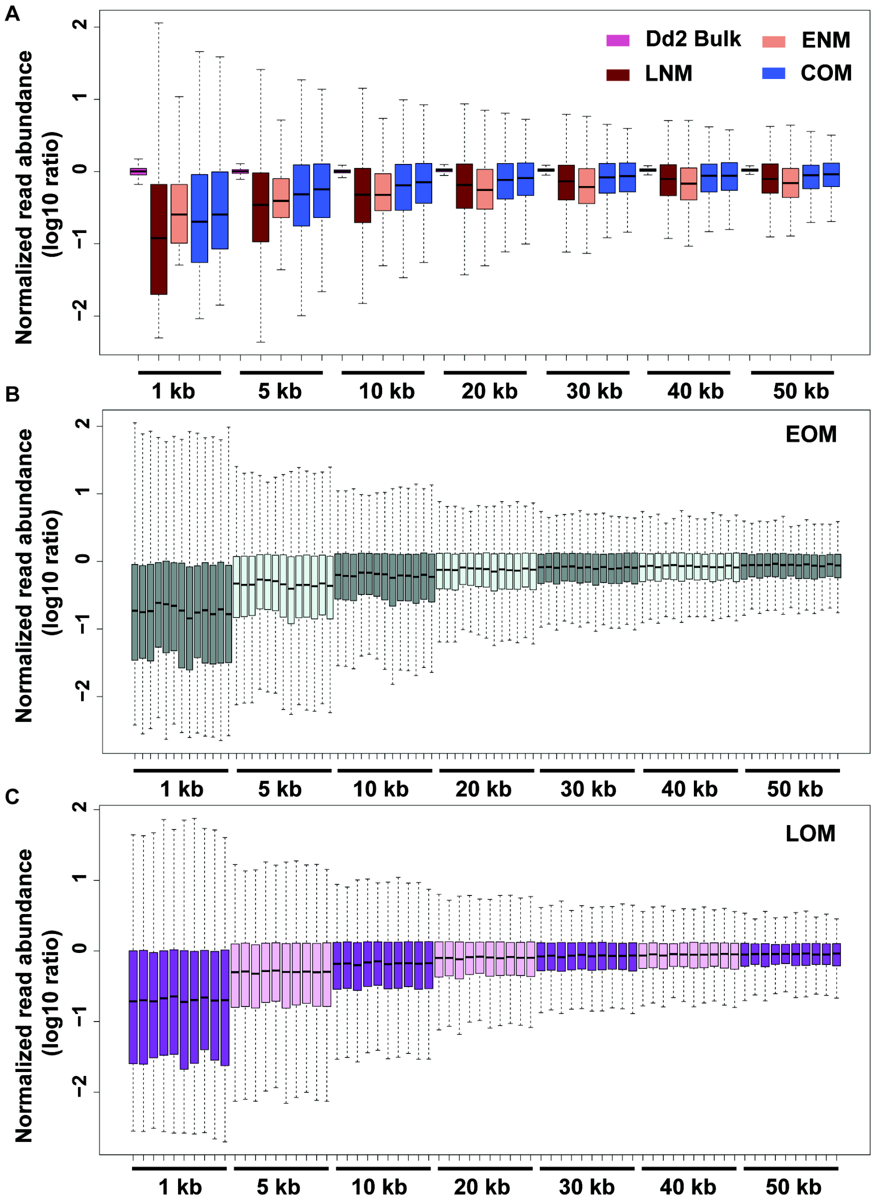
Supplementary Figure 4.3 Uniformity of read abundance across the whole genome of all EOM and LOM samples

A & B. Normalized read abundance across the genome. Distribution of read abundance in 20kb bins in all EOM (A) and LOM (B) samples is shown. Read counts in each bin were normalized by the mean read count over the whole genome for each sample. The circles from outside to the inside represent: chromosomes 1:14 (tan); 13 EOM samples (grey) or 10 LOM samples (purple); Bulk genomic DNA (black). **C. Distribution of normalized read abundance values for all bins**. Normalized read abundance per 20kb bins with outliers (circles) is represented.



Supplementary Figure 4.4 Uniformity of read abundance across the whole genome and correlation analysis for clinical samples

A. Normalized read abundance across the genome. Distribution of read abundance in 20kb bins of clinical samples is shown. Read counts in each bin were normalized by the mean read count over the whole genome for each sample. The circles from outside to the inside represent: chromosome 1:14 (green); 2 COM sample (blue); 1 Clinical bulk genomic DNA (pink). **B. Distribution of normalized read abundance values for all bins.** Box plot representing the normalized read abundance per 20kb bins, outliers are illustrated by blue dots. C. **Paired panels for 1X1 matrices represent Spearman correlation, histogram and pairwise scatterplot among the normalized read abundance (20kb bins) of the COM1 and COM2 sample.** Outlier bins were removed, see *Methods* for outlier identification. The Spearman correlation coefficient is listed above the diagonal, and stars indicate the p-value at the levels of 0.1 (no star), 0.05 (*), 0.01 (**), and 0.001 (***). The histograms on the diagonal show the distribution of normalized read abundance in each sample. The scatter plots include a fitted line through the locally smoothed regression and correlation ellipses (an ellipse around the mean with the axis length reflecting one standard deviation of the x and y variables).



Supplementary Figure 4.5 Distribution of normalized read counts in various bins sizes 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).

We then calculated the mean coefficient of variation (CV) for read abundance in the different sample types (**Table 4.3, Figure 4.3C, and Supplementary Table 4.11**). Following normalization for coverage, the mean CV from the EOM/LOM samples was closer to the CV of the bulk sample than ENM/LNM samples (89/79% versus 22% versus 147/111%, respectively). Once again, the limited standard deviation in these measurements indicates that CV differences represent alterations of read uniformity in each sample type (**Table 3, Supplementary Table 12)**. In support of improved uniformity with optimized MALBAC, the CV value of COM samples was similar to EOM and LOM samples (**Table 3, Figure 4.3C**).

Table 4.3 Coefficient variation of normalized read abundance in each sample type

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.

Optimized MALBAC exhibits reproducible coverage of single cell genomes.

To better assess the amplification patterns across the genomes, we compared the distribution of binned normalized reads from single cell samples to the bulk control using a correlation test (as performed in other single cell studies (Hou et al., 2015; Zhang et al., 2017)). This analysis revealed that amplification patterns of optimized EOM and LOM samples were slightly correlated with the bulk control (mean Spearman correlation coefficient of 0.27 and 0.25, respectively, **Supplementary Table 4.13**), while the non-optimized samples were not correlated (ENM: 0.05 and LNM: 0.07) (**Figure 4.4A**).

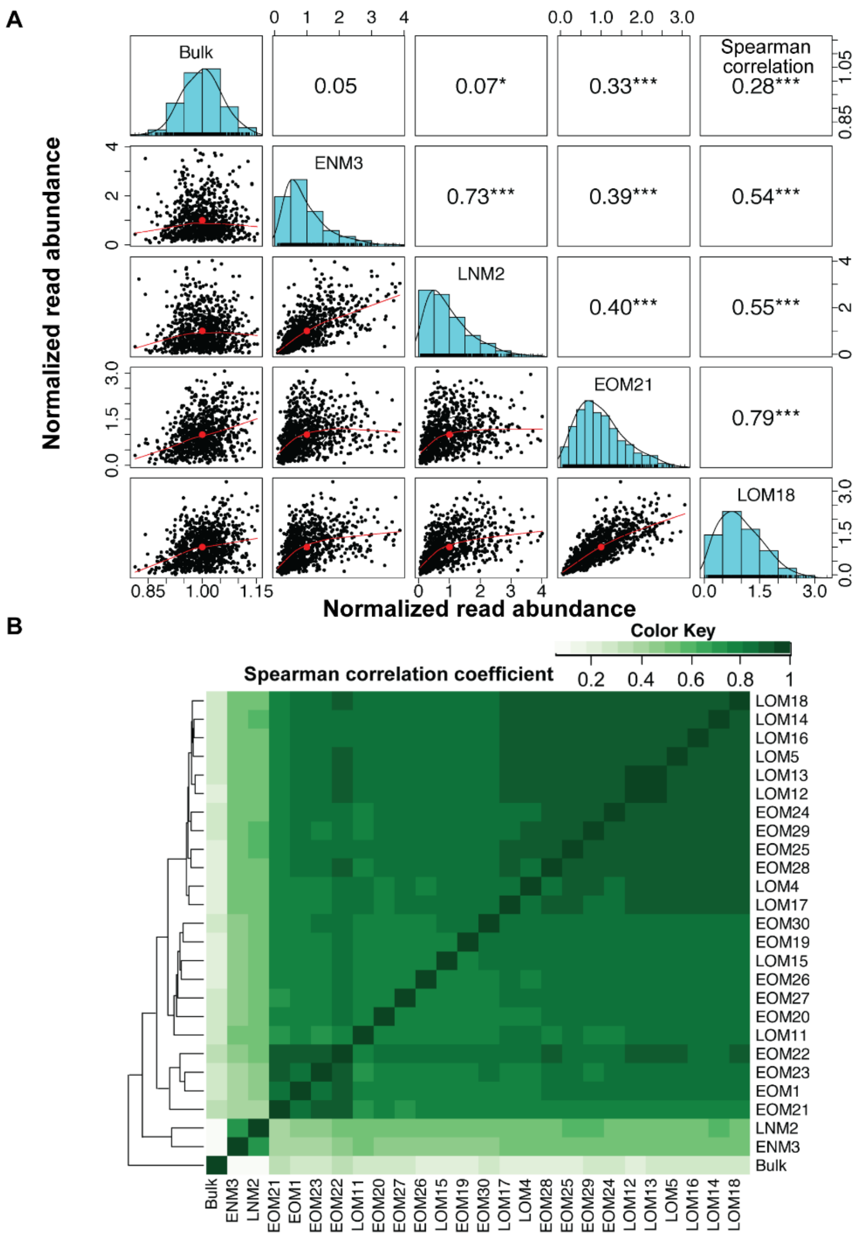


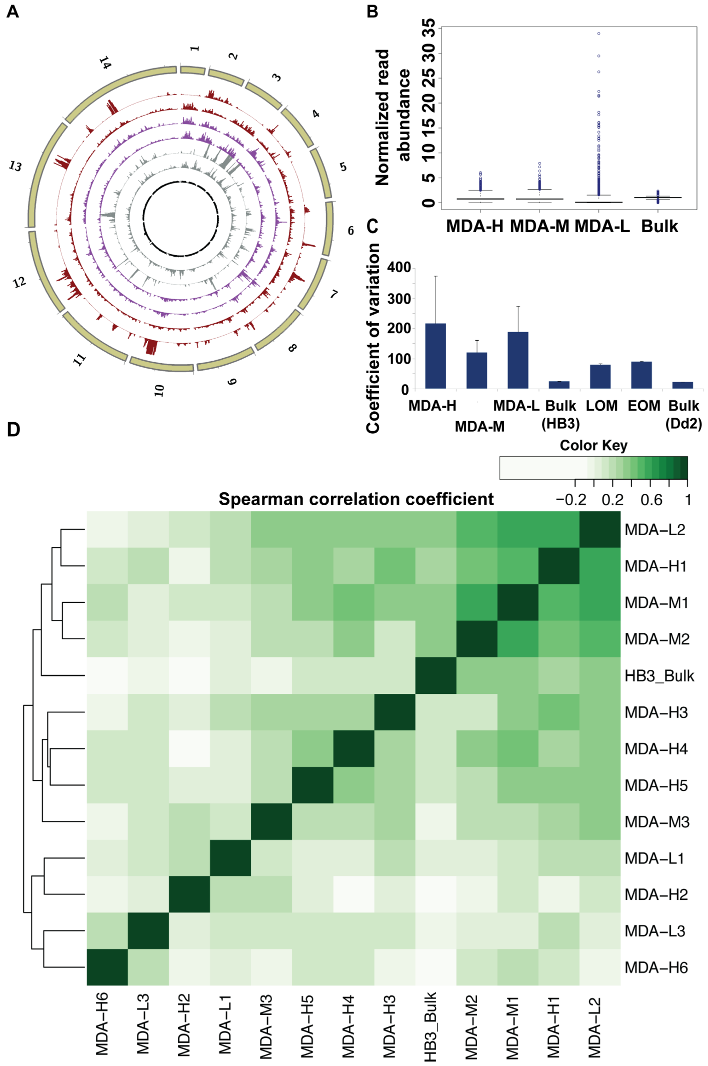
Figure 4.4 Correlations show reproducibility of amplification pattern by optimized MALBAC

A. Paired panels for 5X5 matrices represent Spearman correlation, histogram and pairwise scatterplot among the normalized read abundance of the *Dd2* Bulk, ENM, LNM, and one of each EOM and LOM samples. Outlier bins were removed prior to this analysis (see *Methods* for outlier identification). The Spearman correlation coefficients of each pair are listed above the diagonal, and stars indicate the p-value at levels of 0.1 (no star), 0.05 (*), 0.01 (**), and 0.001 (***). The histograms on the diagonal shows the distribution of normalized read abundance in each sample. The bivariate scatter plots, below the diagonal, depict the fitted line through locally smoothed regression and correlation ellipses (an ellipse around the mean with the axis length reflecting one standard deviation of the x and y variables). **B. Spearman correlation coefficients between sequenced samples.** The hierarchical clustering heatmap was generated using Spearman correlation coefficients of normalized read abundance. The color scale indicates the degree of correlation (white, correlation= 0; green, correlation > 0).

To quantify the reproducibility of read distribution between single cell samples amplified by MALBAC, we compared Spearman correlation coefficients. The read abundance across all single cell samples was highly correlated; two individual EOM or LOM samples had a mean correlation coefficient of 0.83 and 0.88 respectively (**Figure 4.4B**). When we expanded our analysis to calculate the correlation of binned normalized reads between all 26 sequenced samples (**Supplementary Table 4.13**) and hierarchically clustered the Spearman correlation coefficient matrix between these samples, all 23 optimized single cell samples (EOM and LOM) clustered with a mean Spearman correlation coefficient of 0.84 (**Figure 4.4B**). In addition, the two COM samples were correlated with each other (Spearman correlation coefficient of 0.84) (**Figure S4.4**C). This correlation indicated high reproducibility of normalized read distribution across the genomes that were amplified by optimized MALBAC. Within the large cluster, two LOM samples (LOM12 and LOM13) displayed the highest correlation (0.94, **Figure 4.4B**).

Reproducible coverage with lower variation is the main benefit of MALBAC over MDA-based amplification of single cell genomes.

We compared our data to that from a MDA-based study because this is the only other method that has been used to amplify single *Plasmodium* genomes ((Trevino et al., 2017), **Figure S4.6**). This study sorted individual infected erythrocytes with high (H), medium (M) and low (L) DNA content corresponding to late, mid, and early stage parasites, applied MDA-based WGA to single erythrocytes, and sequenced the DNA products. The authors measured a similar amplification success rate in early (L) stage samples as our study (MDA: 50% by DNA yield, MALBAC: 43% by DNA yield) yet slightly improved success rates for late (H) stage samples (MDA: 100%, MALBAC: 90%, **Supplementary Table 4.8 and Supplementary Table 4.9**). In light of experimental differences between the two studies (**Supplementary Table 4.14**), we analyzed data from the twelve MDA samples using our exact analysis pipeline and parameters (six MDA-H and three of each MDA-M and -L samples) and confined our comparison of the data to a few metrics: 1) coefficient of variation of read abundance, 2) coverage breadth, and 3) correlation between samples (see below).



Supplementary Figure 4.6 Uniformity of coverage and correlation analysis in MDA-amplified single cell samples

Samples for this analysis were from Trevino et al., 2017 and analyzed using our pipeline (**Figure S4.2**). **A. Normalized read abundance across the genome.** Distribution of read abundance in 20kb bins of MDA-amplified samples is shown. The circles from outside to the inside represent: chromosome 1:14 (tan); two MDA-H samples (dark red, HB3 parasite with high DNA content amplified by MDA); two MDA-M sample (purple, HB3 parasite with medium DNA content amplified by MDA); two MDA-L samples (grey, parasite with low DNA content amplified by MDA); one HB3 Bulk sample. **B. Distribution of normalized read abundance values for all bins in MDA amplified samples.** MDA-H: 1 representative sample; MDA-M: 1 representative sample; MDA-L: 1 representative sample from (Trevino et al., 2017). **C. Comparison of coefficient of variation of normalized read abundance between MDA and MALBAC amplified single parasites.** The average and SD (error bars) of coefficient of variation of all samples from each type are represented (MDA-H: 6 samples; MDA-M: 3 samples; MDA-L: 3 samples; HB3 Bulk: 1 sample; LOM: 10 samples; EOM: 13 samples; Dd2 Bulk: 1 sample). **D. Spearman correlation coefficient between MDA amplified samples.** The color scale indicates the degree of correlation (white, low correlation; green, high correlation).

MDA is known to produce artifacts that impair CNV detection (Arriola et al., 2007; Corneveaux et al., 2007; Lasken and Stockwell, 2007). While MALBAC-amplified genomes exhibited a consistent amplification pattern (**Figure S4.3A and S4.3B**), the MDA-amplified genomes showed more variation across cells (**Figure S4.6A**). We also detected higher variation in normalized read abundance in the MDA-H samples (compared to MDA-L and -M samples, **Figure S4.6B**), which was not consistent with the report that the MDA method amplifies high DNA content better than parasites with lower DNA content (Trevino et al., 2017). Even though the bulk DNA controls used in both studies showed similar CVs (24% versus 22%), the MDA-amplified samples displayed a higher CV than MALBAC-amplified single cell samples regardless of the parasite stage (a mean of 186% versus 85%, respectively, **Table 4.3, Supplementary Table 4.11 and Supplementary Table 4.15**). As expected based on MALBAC's limited coverage of intergenic regions (**Table 4.2**), MDA amplified samples displayed a higher coverage breadth cross the genome, especially in the intergenic regions (**Supplementary Table 4.16**). Additionally, the correlation between MDA-amplified cells (mean correlation coefficient: 0.20; **Supplementary Table 4.17, Figure S4.6D**) was much lower than that between our optimized MALBAC-amplified cells (mean correlation coefficient: 0.84; **Supplementary Table 4.13**, **Figure 4.4B**); this finding confirms prior observations that MDA exhibits a more random amplification pattern than MALBAC (Chen et al., 2014).

Copy number variation analysis is achievable in MALBAC-amplified single cell genomes. To detect CNVs with confidence, we employed both discordant/split read detection and read-depth based methods with strict parameters. We used LUMPY to detect paired reads that span CNV breakpoints or have unexpected distances/orientations (requiring a minimum of 2 supporting reads). We also used a single cell CNV analysis tool, Ginkgo, to segment the genome based on read depth across bins of multiple sizes and determine copy number of segments (requiring a minimum of 5 consecutive bins). We regarded the CNVs detected by the two methods as the same if one CNV overlapped with at least half of the other CNV and vice versa (50% reciprocal overlap). Using this approach, we first identified a "true set" of CNVs from the bulk *Dd2* DNA sample (**Table 4.4**, 3 CNVs on 3 different chromosomes). One of the true CNVs was identified previously in this parasite line (the large *Pfmdr1* amplification on chromosome 5, (Cowman et al., 1994, 1)); another true CNV occurs in an area of the genome that is reported to commonly rearrange in laboratory parasites ((Scherf et al., 1992), the *Pf11-1* amplification of chromosome 10).

Table 4.4 True CNVs detected in the Dd2 bulk genome

Name	Chr.	Start Pos.	Size (bp)	Type	Support read*		Start Pos.	Size (bp)	Copy number detected by Ginkgo** in different bin sizes				Mappability^
					Discordant read	Split read			1kb	5kb	8kb	10kb	
Pfmdr1	5	888316	81935	DUP	53	0	888000	82000	2	2	Nd	Nd	1
Pf11-1	10	1524527	18472	DUP	29	1	1520000	28000	4	5	N/A	N/A	0.86
Pf332	11	1956623	8719	DUP	0	8	1953000	13000	4	N/A	N/A	N/A	0.92

*Detected by LUMPY based on discordant/split read detection, minimum number of supporting reads is 2.

**For Ginkgo analysis, the minimum bin number of segmentation is 5.

^For comparison, the mean mappability of the core genome is 0.99 and the mean mappability telomere/subtelomere regions including *var* gene clusters is 0.65.

DUP, duplication; N/A, not applicable because the target CNVs will not be detected as the bin size (>= 5 x bin size) is larger than the size of the target CNVs. Nd, not detected in the specified bin size.

With a set of true CNVs in hand, we assessed our ability to identify them in the single cell samples amplified by MALBAC and explored parameters that impacted their detection. As expected, each CNV detection method exhibited differences in the ability to identify true CNVs, which is likely due to a number of factors including CNV size, genomic neighborhood, and sequencing depth (Pirooznia et al., 2015). For example, using discordant/split read analysis, we were able to readily identify the *Pf11-1* amplification in the majority of samples (21 of 25 samples, **Supplementary Table 4.18**). This method was less successful in identifying the *Pfmdr1* amplification (only 3 optimized MALBAC samples in total, **Supplementary Table 4.18**). For read-depth analysis, the success of true CNV detection was heavily dependent on the bin size (**Supplementary Table 4.18**). If we selected the lowest bin size (1kb) in which it was possible to detect the smallest of the true CNVs (13kb), we were able to readily identify the *Pfmdr1* amplification in all samples (**Supplementary Table 4.18**). As we increased the bin size, the number samples with *Pfmdr1* detection decreased, only optimized MALBAC samples were represented, and the copy number estimate in single cells approached the bulk control

(**Supplementary Table 4.7 and S18**). The other two true CNVs were only detected at the 1kb bin size in a minority of samples (6 total, **Supplementary Table 4.18**).

When we assessed true CNVs that overlapped between the two methods, we were able to improve the precision and sensitivity of CNV detection in five single cell samples (**Supplementary Table 4.19**) and detect at least one CNV in each (3 EOM and 2 LOM samples out of 25 total cells, **Table 4.5**). Notably, in one sample, EOM 23, the *Pfmdr1* amplification was detected in bin sizes of up to 10kb at a copy number similar to the bulk control (**Table 4.5**). Besides the CNVs conserved with the bulk sample, we also detected unique CNVs that were only identified in the single cell samples. In general, the CNVs detected by both discordant/split read and read depth analyses were spread across all chromosomes except chromosome 9, predominantly confined to optimized MALBAC samples, and were only detected at 1kb read depth bin sizes (**Supplementary Table 20**).

Table 4.5 True CNVs detected in single cell samples

Sample name	CNV name	Start Position	Size (bp)	Supporting reads		Start Position	Size (bp)	Copy number detected by Ginkgo in different bin sizes			
				Discordant read	Split read			1kb	5kb	8kb	10kb
LOM 5	*Pfmdr1*	891390	34069	0	2	907000	28000	9	Nd	N/A	N/A
LOM 16	*Pf11-1*	1542335	3836	0	3	1543000	5000	3	N/A	N/A	N/A
EOM 23	*Pfmdr1*	889899	79890	3	3	888000	82000	4	6	5	5
EOM 26	*Pf11-1*	1542335	3836	0	5	1543000	5000	4	N/A	N/A	N/A
EOM 29	*Pf11-1*	1539158	5639	4	0	1541000	7000	3	N/A	N/A	N/A

"N/A" indicates the target CNVs will not be detected as the bin size (>= 5 bin size) is larger than the size of the target CNVs. Nd, not detected in the specified bin size.

High quality-SNPs are detected in MALBAC-amplified single cell genomes.

Firstly, to understand the accuracy of SNP detection in MALBAC-amplified genomes, we estimated the precision and sensitivity of SNP detection in single cells by treating those from the *Dd2* bulk sample as standard SNP set. We performed this analysis with increasing stringency levels (VQSLOD score > 0; VQSLOD score > 6; VQSLOD score > 6 with read depth > 10, **Table 4.6**) in order to calibrate with previous SNP studies and evaluate the impact of read depth on SNP identification. In the *Dd2* bulk sample, 18369 SNPs were detected with VQSLOD score > 0, while 13168 SNPs were detected with VQSLOD score > 6 and read depth > 10; the later number is more consistent with the number of SNPs identified in previous studies of *Dd2 P. falciparum* (Volkman et al., 2007). Similarly, as we increased the stringency level, fewer SNPs were detected for each single cell sample and sensitivity decreased, indicating increased false negatives for SNP detection. The precision of SNP detection, however, increased from 65% (VQSLOD score > 0) to 92% (VQSLOD score > 6) and 99% (VQSLOD score > 6/read depth > 10) in EOM samples; the same trend was observed for LOM samples (**Table 4.6**). The best balance of precision and sensitivity for SNP detection in single cells was achieved at the level of VQSLOD score > 6. Even though the sensitivity for SNP detection is only 46% (EOMs) and 47% (LOMs) in individual single cells at this stringency level, we observed up to 72% sensitivity when we pooled optimized single cell samples (13 EOMs and 10 LOMs, **Figure S4.7**).

We also evaluated the detection of 16 known drug resistance SNPs from the *Dd2* bulk sample (**Supplementary Table 4.21**). When we pooled all single cell samples (EOMs and LOMs), we detected 13 of the 16 resistance SNPs (**Supplementary Table 4.22**); the 3 remaining SNPs were not identified due to a lack of coverage at these sites in single cell genomes (**Additional File 3:**

SNPs detected in all samples). As expected, the sensitivity of SNP detection was much lower in non-*Dd2* patient-isolated COM samples (15.37%, VQSLOD score > 6) when compared to that in the *Dd2*-derived EOM samples (46.22%).

Since the *Dd2* parasites that we used in this study were not recently cloned, there is a possibility of detecting novel SNPs that have arisen in the population over time in laboratory culture (Jett et al., 2020). After removing any mixed allele calls and applying the highest stringency level (VQSLOD score > 6, read depth > 10), we identified 124 novel SNPs in the single cell samples that were not present in the *Dd2* bulk sample (**Additional file 4: Single cell novel SNPs**). These loci affected 226 genes on all 14 chromosomes of the parasite genome (**Supplementary Table 4.23**), representing genes involved in the biosynthesis of antibiotics, Ac/N-end rule pathway, purine metabolism, thiamine metabolism, and aminoacyl-tRNA biosynthesis (**Supplementary Table 4.24**).

New optimizations of MALBAC further improved coverage breath and amplification uniformity

More recently, we further improved our MALBAC protocol (termed further-optimized MALBAC) by modifying the pre-amplification random primer and replacing Bst DNA Polymerase Large Fragment with Bsu DNA Polymerase Large Fragment with a lower reaction temperature (see "Methods" for more details). We applied this new version of MALBAC (further-optimized MALBAC) to amplify early-stage single parasite samples (early stage further optimized MALBAC, EFOM). When comparing to the previously optimized MALBAC from the same batch of sequencing run, we found the Bsu DNA polymerase used in pre-amplification shows higher coverage breadth (34.03% in Bsu amplified sample, 26.51% in Bst amplified sample) and coefficient of variation of read abundance (65 in Bsu amplified sample, 126 in Bst amplified sample) than Bst polymerase under the same coverage level (**Table 4.6**). The optimized random primer version 2 shows higher coverage breadth in intergenic regions than the optimized random primer version 1. Thus, the further optimized MALBAC shows overall improved coverage breadth and lower coefficient of variations of normalized read abundance than the optimized MALBAC in the same sequencing run.

Table 4.6 Comparison between optimized and further optimized MALBAC sequencing results

Studies	Pre-amplification		Coverage (X)	GC %	Coverage breadth			
	Polymerase	Random primer			Whole genome	Genic regions	Intergenic regions	Coefficient of Variation (CV) of normalized read abundance (20kb bin)
Optimized MALBAC (Dd2, early stage)	Bst	optimized random primer version 1*	1.58	24.73	26.51%	41.67%	3.95%	126
Further optimized MALBAC (Dd2, early stage)	Bst	optimized random primer version 2**	1.71	23.07	29.25%	45.41%	5.24%	109
	Bst	optimized random primer version 2	1.70	22.58	22.24%	32.75%	6.57%	131
	Bsu	optimized random primer version 1	1.68	22.97	34.03%	50.22%	9.90%	65
	Bsu	optimized random primer version 2	1.70	22.66	35.32%	51.98%	10.52%	78
	Bsu	optimized random primer version 2	1.68	22.43	35.84%	52.38%	11.23%	69

***optimized random primer version 1:**
5′GTGAGTGATGGTTGAGGTAGTGTGGAGNNNNNTTT 3′ (20% GC for NNNNN)
****optimized random primer version 2:**
5′GTGAGTGATGGTTGAGGTAGTGTGGAGNNNNNNNNNNTTT 3′ (20% GC for NNNNNNNNNN)

Integrating robotic pipetting into single cell sequencing pipeline

To improve single cell sequencing throughput and limit contamination due to manual pipetting, we integrated robotic pipetting into our single cell sequencing pipeline (see Methods). We also changed our cell isolation method to FACS sorting (Sony SH800) from CellRaft AIR System, since the FACS sorting sorts cells into full-skirted 96 well plates, which is compatible with the

Mosquito LV robotic pipetting system. To test whether the Mosquito LV pipetting works for our single cell sequencing pipeline, we sequenced 3 HB3 single cell samples and 4 Dd2 single cell samples (**Table 4.7**). As shown in **Table 4.7**, we successfully used Mosquito LV to conduct automatic pipetting during the pre-amplification cycles and amplified 7 single cell samples.

Table 4.7 Sequencing results of single cells amplified using liquid handler

Studies	Pre-amplification		Coverage (X)	GC %	Coverage breadth			
	Polymerase	**Random primer**			**Whole genome**	**Genic regions**	**Intergenic regions**	**Coefficient of Variation (CV) of normalized read abundance (20kb bin)**
HB3, N=3	Bsu and Bst*	optimized random primer version 2	1.6X	24.60%	26%	40%	5%	80
Dd2, N=4	Bsu and Bst^	optimized random primer version 2	1.5X	25.20%	25%	40%	4%	127

*17 cycles with Bsu polymerase and 2 cycles with Bst polymerase; ^17 cycles with Bsu polymerase and 6 cycles with Bst polymerase

4.5 Discussion

This study is the first to optimize the standard MALBAC protocol for single cell sequencing of a genome with extreme GC-content (*P. falciparum*: 19.4%). We showed that this optimized method can reliably amplify early-stage parasite genomes, which contain <30 femtograms of DNA per sample. Single cell experiments are innately very sensitive to contaminating DNA from other organisms and we detected a lower proportion of human and bacteria DNA in MALBAC-amplified samples, which impacted overall coverage of the *P. falciparum* genome. Furthermore, we showed that this method reduced GC-bias to impact the breadth and uniformity of genome amplification. Finally, with these single cell genomes, we were able to explore the detection of CNVs and SNPs to study parasite-to-parasite heterogeneity.

MALBAC Volume and Cycles

MALBAC amplification has been used in studies of human cells, where each diploid genome harbors ~7 picograms of DNA (Zong et al., 2012). In this study, we used the MALBAC method to successfully amplify femtogram levels of DNA from single *P. falciparum* parasites. Reducing the total reaction volume (from 50µl to 20µl) and increasing the number of amplification cycles (pre-amplification: from 5 to 18-19; exponential: from 15 to 17) likely contributed significantly to this increased sensitivity. Both modifications were important. Initially the lower sample volume reduced the overall DNA yield and this was reversed using increased amplification cycles. These modifications provide additional benefits including reduced contaminating reads and experimental costs. Importantly, these simple steps can be applied to the MALBAC amplification of small genomes or genomes with skewed GC-content from other organisms such as bacteria (Wang et al., 2016). For example, studies of *Mycoplasma capricolum* (GC-poor) (Ohkubo et al., 1987)*, Rickettsia prowasekii* (GC-poor) (Andersson et al., 1998)*,* and *Borrelia burgdorferi* (GC-poor) (Fraser et al., 1997)*, Entamoeba* histolytica (GC-poor) (Lorenzi et al., 2010)*, Micrococcus luteus* (GC-rich) (Ohama et al., 1990) could be improved using this method.

Primers and Coverage Bias

The modification of the primer was essential for the successful amplification of the AT-rich *P. falciparum* genome. This change was meant to prevent the preferential amplification of GC-rich sequences as observed for human and rat single cell genomes (Hou et al., 2015; Ning et al., 2015). We achieved a coverage breadth across *P. falciparum* genic regions (a mean of 21.7% GC-content) of as high as ~80% (**Table 4.2**) by specifically altering the base content of the degenerate 5-mer of MALBAC pre-amplification primer from 50% to 20% GC-content. The initial priming step is crucial for whole genome amplification and controlling this step can limit amplification bias (Lasken, 2013). Indeed, 5-mers with ~20% GC-content across the *P. falciparum* genome are 2- and 6-fold more common than those with 40% and 60% GC-content, respectively (**Supplementary Table 4.1**). This difference indicated that annealing of the optimized MALBAC primer based on the degenerate bases was more specific for the parasite's genome than the standard MALBAC primer. Interestingly, during this study we observed a preferential amplification of genic over intergenic regions (**Table 4.2**), which may be explained by lower percentage of 5-mers with 20% GC-content in intergenic regions than in genic regions (**Supplementary Table 4.1**). Furthermore, when we searched for reads that contained the MALBAC common sequence (see *Methods* and **Supplementary Table 4.5**) to identify WGA binding sites across the *P. falciparum* genome, we found that binding sites were predominantly located in the genic regions (**Supplementary Table 4.5**); this result indicated that there was an issue with primer annealing in intergenic regions, which may be caused by a high predicted rate of DNA secondary structure formation across these regions of the *P. falciparum* genome (Huckaby et al., 2018). The polymerase used in the MALBAC linear amplification steps (*Bst* large fragment) exhibits strand displacement activity, which presumably allows resolution of secondary structure (Viguera et al., 2001; Ignatov et al., 2014). However, a longer extension time may be required for amplification of repetitive DNA sequence, either during linear or exponential steps.

Parasite and Contaminating Genomes

The standard MALBAC method was reported to display the most favorable ratio of parasite DNA amplification over human DNA when compared to other common WGA methods (Srisutham et al., 2020). Out steps to optimize MALBAC (reduced volume and increased cycle numbers) not only enhanced the amplification of the small parasite genome, but also likely improved the sensitivity to amplify contaminating non-parasite DNA. Nevertheless, in many samples, optimized MALBAC yielded lower proportions of contaminating DNA than the bulk sample (**Figure 4.2A**). We speculate that this effect was once again due to our modification of the GC-content of the degenerate bases of the primer; this alteration limited the preferential amplification of contaminating DNA with higher GC content (observed during standard MALBAC), thus improving the representation of parasite DNA in the overall WGA product.

The major contaminating DNA source that we detected in our samples was from humans (**Figure 4.2A**). This was not surprising given that, in our experimental system, the culture and host environments are rich in human DNA (Oyola et al., 2013; Carey et al., 2018). It is also possible that human DNA was introduced during the single cell isolation or WGA steps. The former situation is a larger issue for clinical parasite isolates due to the abundance of white blood cells that contribute to extracellular DNA when they decay outside of the host (Waldvogel Abramowski et al., 2018). Indeed, we observed more human DNA in clinical bulk and single cell samples (an increase of ~11-fold over laboratory-derived *Dd2* bulk and single cell samples, respectively). The massive level of contamination in the clinical bulk sample and limited coverage of the parasite genome (0.3%) was exacerbated by 1) the omission of a stringent leukodepletion step that is routinely employed to limit host cell contamination (Manske et al., 2012; Jacob et al., 2014) and 2) the lower overall sequencing output of that particular run (**Supplementary Table 4.4**).

The second most common source of contaminating DNA was bacteria (**Figure 4.2A**). WGA approaches are known to occasionally amplify residual bacterial DNA associated with commercial polymerases (Rand and Houck, 1990; Woyke et al., 2011; Salter et al., 2014; Kil et al., 2015) or other reagents (McFeters et al., 1993; Nogami et al., 1998; Kulakov et al., 2002). Since this contaminant was absent in the bulk DNA control and increased in early stage parasite samples (representing an average of 0.8% of EOM reads compared to 0.2% for LOM samples), bacterial DNA may also have been introduced during single parasite isolation and WGA steps. While we took precautions to limit this occurrence (see *Methods*), minimizing the reaction volume could further reduce this source of contamination.

Early and Late Stage Parasites

Depending on when a novel CNV or SNP arises (i.e. early or late in replication), each parasite stage holds advantages for its detection. If the mutation arises in the first round of replication and is copied into most of the genomes of a late stage parasite, having multiple genomes will be advantageous for detection. If the mutation arises later in replication, it will be present in few of the genomes; therefore, averaging across the genomes, as with bulk analysis, will limit its detection. Since only one haploid genome is present in an early stage parasite, the sensitivity for detecting rare/unique CNVs/SNPs within parasite populations is favored.

For this reason, staging of parasites in this study was important. We performed stage-specific enrichment before single cell isolation and confirmed that the majority of parasites were the desired stage using flow cytometry (see *Methods*, **Figure S4.1**, 97% for early stage enrichment and 74% for late stage enrichment). Furthermore, during selection of cells by microscopy (before automated collection by the IsoRaft instrument), we confirmed the expected fluorescence intensities for each stage; early stage parasites had a significantly smaller genome and mitochondrion size compared to late state (as in **Figure 4.1B**). However, differences in preparation of samples may have impacted our parasite stage comparisons. While all late stage samples were isolated, lysed and amplified in the same batch under the same conditions, early stage samples were processed in three separate batches (**Supplementary Table 4.11**). Despite conserved methods and good concordance in CV between all samples (**Supplementary Table 4.11**), minor differences could have contributed to variations in the amplification steps.

Differences in genome analysis results from optimized MALBAC samples provided further confidence that the parasites were of the expected stage. Firstly, late stage parasites showed a higher WGA success rate than early stage parasites (90% versus 43%, **Supplementary Table 4.9**). This result was explained simply by the presence of extra genomes in the late stage samples (~16n versus 1n) and was consistent with a previous study that used MDA-based amplification methods (Trevino et al., 2017). Late stage parasites also displayed better uniformity of read abundance (**Table 4.3**), indicating less amplification bias because fewer regions were missed when more genomes were present. Additionally, there were fewer overall contaminating reads found in late stage parasites than early stage parasites (5.1% versus 8.6%). Once again, this was likely due to a higher ratio of parasite DNA to contaminating DNA in the late stage samples.

Despite these differences, we observed similar coverage breadth and Spearman correlation coefficients of read abundance for both early and late stage MALBAC-amplified parasites (**Table 4.2 and Supplementary Table 4.13**). This was contrary to the MDA study in single *P. falciparum* parasites that found a higher breadth of genome coverage from the late stage parasites (Trevino et al., 2017). Our findings confirm that the pattern of amplification across the genome is determined by the binding of the optimized MALBAC primers and not the parasite developmental form.

CNV Analysis and Related Considerations

Sequencing at very high depth improves the detection of low frequency CNVs in bulk samples, but the sensitivity is limited to large-scale CNVs present in > 5% cells (Zhang and Vijg, 2018). Other analysis methods that rely on the detection of reads that span CNV junctions (i.e. split reads or discordant reads) have improved the sensitivity and specificity of CNV detection (Zhang et al., 2011), but continue to struggle with minor allele detection. For single cell analysis, the high level of MALBAC amplification reproducibility (i.e. the same regions are over- and under-amplified across multiple genomes), that we and others have observed, is especially advantageous for CNV detection. This is because amplification bias can be normalized across cells, as has been successfully performed for human cells (Zong et al., 2012; Hou et al., 2013). Unfortunately, cross-sample normalization was not possible in our study due to the use of a single laboratory parasite line that includes known CNVs (*Dd2*). Instead, as described below, we combined a read-depth based tool (Ginkgo (Garvin et al., 2015)) with a split/discordant read-based method (LUMPY (Layer et al., 2014)) to improve the accuracy of CNV detection (as in (Pirooznia et al., 2015)).

We observed a large number of raw CNVs detected in single cell genomes by each individual method (**Supplementary Table 4.6-**LUMPY **and Supplementary Table 4.7-**Ginkgo) and the precision and sensitivity of each method was low (**Supplementary Table 4.19).** These initial results may be explained by a number of possibilities, including those that are both biological in nature as well as artifacts of our analysis methods. From a biological perspective, these calls can represent large CNVs that are known to exist in the bulk sample (i.e. *Pfmdr1* and *Pf11.1*, **Table 4.5**) as well as the abundance of small CNVs that may be present in a minor part of the population (unique CNVs, **Supplementary Table 4.20**). Because prior *P. falciparum* CNV analyses were confined to bulk DNA sequencing, our view of minor variants in parasite populations is limited. The recent discovery of *P. falciparum* extrachromosomal DNA that is derived from regions of the genome that harbor CNVs (McDaniels et al., 2021) suggests that there are cellular pathways that could contribute to cell-to-cell variations in CNV boundaries and dynamics (i.e. perhaps through the excision and reintegration of extrachromosomal DNA). While the differences in start position of true CNVs from single cell genomes (**Tables 4.4 and 4.5**) could represent true minor variants, they could also be due to analysis artifacts that contribute to excess CNV calls and inaccuracies in estimating boundaries. For example, raw LUMPY results exhibit redundancy due to slightly varied boundaries and sizes of the same CNV. Additionally, parameters of read-depth based approaches like Ginkgo (i.e. bin sizes and the requirements for consecutive bins) can alter CNV calling; 1kb bins may heavily reflect coverage variation in the genome and have a high level of false positives while larger bin sizes may miss smaller CNVs.

In an effort to limit false positives and these uninformative variations, we combined approaches by retaining calls that overlapped between the two approaches (see *Methods*). In support of this recommendation, the combined approach limited the number of overall CNV calls and improved the precision in some single cell samples (**Supplementary Table 4.19**).

Even when using the combined approach, we observed variations in the boundaries and copy numbers of the true CNVs in single cell samples (**Table 4.4 and Table 4.5**). For example, in *Dd2* parasites, the *Pfmdr1* CNV is ~82kb but in single cell samples, it is called as ~30kb with later starting position. This difference is most likely due to uneven coverage across these large CNVs in single cell samples; some regions accurately reflect the CNV where others do not. Importantly, as we increased the bin size, the uniformity of read count improves (**Figure S4.5**), which impacts CNV identification (i.e. the *Pfmdr1* amplification is found in fewer single cell genomes and the copy number estimate approaches that of the bulk control, **Supplementary Table 4.7 and Supplementary Table 4.18**). Thus, efforts to improve the uniformity of read coverage, genome coverage breadth and the potential for cross-sample normalization will improve our ability to accurately detect CNVs. Overall, it is notable that we can detect CNVs in some single cell genomes (<100kb, **Table 4.5**) that are below the current resolution of CNV detection from single cell genomes amplified with common WGA methods (>>100kb to Mb) (Navin et al., 2011; Zong et al., 2012; McConnell et al., 2013; Campbell et al., 2015; Ning et al., 2015). Our CNV analysis capabilities will improve with expanded numbers and genomic diversity; the inclusion of parasite lines with different CNV profiles will greatly facilitate the removal of reproducible amplification bias and increase the detection of conserved and unique CNVs of all sizes.

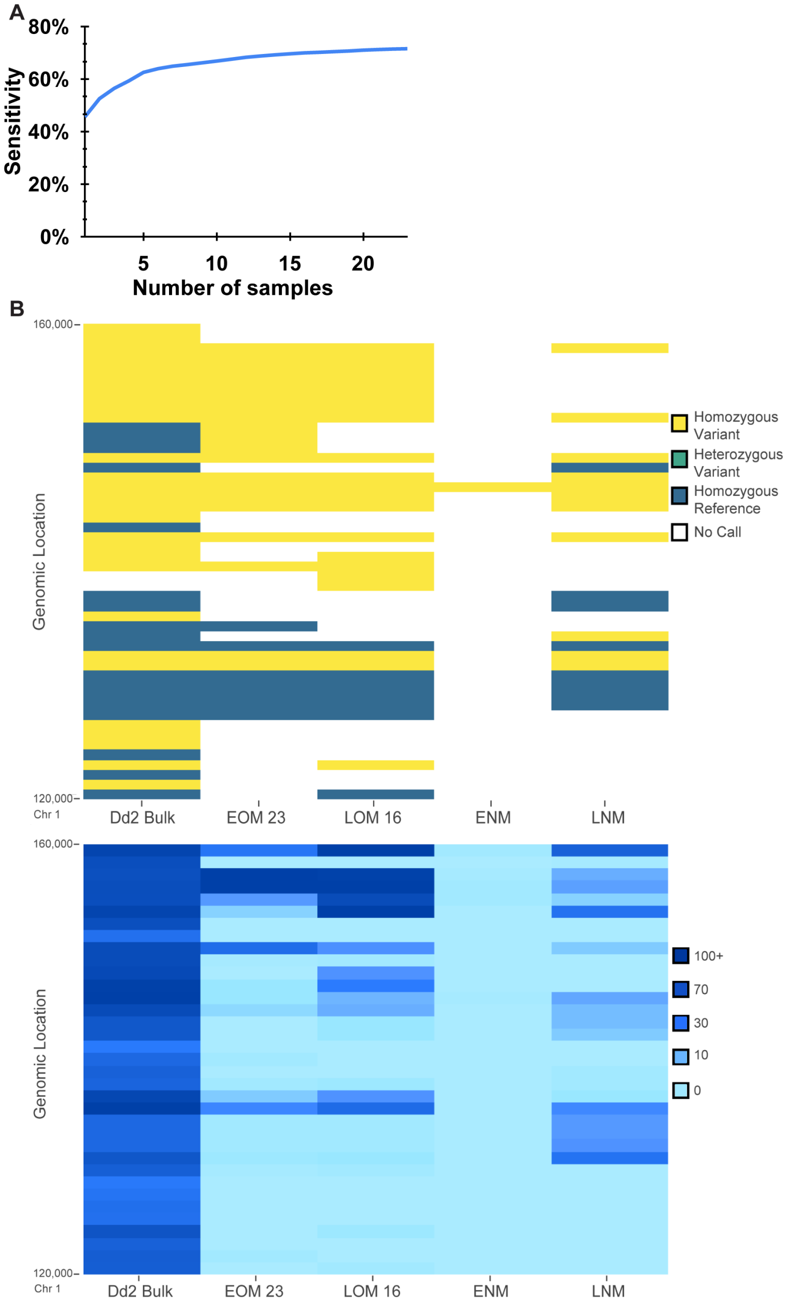
Standard and Novel SNP Analysis

By combining stringent SNP filtering strategies (i.e. VQSLOD and read depth cutoffs), we increased the precision for SNP calling in single cell samples and detected 72% of SNPs identified in the bulk sample and 13 of 16 known resistance SNPs (**Table 4.8, Supplementary Table 4.22, Figure S4.7A**). We also detected a number of novel SNPs across our single cell samples (**Additional File 4: Single cell novel SNPs VCF file**). On average, this is ~13 SNPs per genome, since many of the 124 novel SNPs are shared among genomes (46 shared SNPs). We are not able to compare this rate to that estimated from other studies (Bopp et al., 2013; Claessens et al., 2014; McDew-White et al., 2019; Jett et al., 2020) because our bulk sample was not cloned prior to single cell isolation and culturing days are unknown. Although we take precautions to limit divergence (i.e. parasite lines are only grown for limited amounts of time and periodically cloned), we do not know the complete life history of the *Dd2* lines because they come from a general repository. However, when we assessed SNPs from multiple *Dd2* lines using the current pipeline (see *Methods*, VQSLOD>6), we identified 146 SNPs in the short reads used to generate the Dd2 reference genome (Otto et al., 2018) that were not present in our bulk sample, indicating some divergence occurs in samples that are independently propagated.

Table 4.8 SNP detection in sequenced samples

Variant Filtering Conditions	Sample name	Number of SNPs	Precision	Sensitivity
VQSLOD > 0	EOM (13)	12734	65.25%	45.09%
	LOM (10)	12730	67.16%	46.40%
	ENM (1)	2269	84.93%	10.49%
	LNM (1)	9235	67.29%	33.83%
	Dd2 Bulk (1)	18369	-	-
	COM (2)	8851	31.55%	15.22%
VQSLOD > 6	EOM (13)	6917	91.75%	46.22%
	LOM (10)	6990	92.40%	47.05%
	ENM (1)	1375	96.58%	9.68%
	LNM (1)	4902	95.17%	33.99%
	Dd2 Bulk (1)	13725	-	-
	COM (2)	3162	66.69%	15.37%
VQSLOD > 6, Read Depth > 10	EOM (13)	3937	99.48%	29.74%
	LOM (10)	4360	99.41%	32.92%
	ENM (1)	252	97.62%	1.87%
	LNM (1)	2575	98.87%	19.33%
	Dd2 Bulk (1)	13168	-	-
	COM (2)	1021	79.28%	6.13%

*Precision and sensitivity are calculated by using SNPs detected in *Dd2* bulk sample as the standard SNP set.



Supplementary Figure 4.7 Sensitivity of SNP detection in pooled single cells and association between SNP calls and read depth

A. Pooling and subsampling of single cell samples shows a plateau of SNP discovery as numbers of cells increase. Gradually increasing the number of single cells samples and pooling single cell samples (13 EOMs and 10 LOMs) allows the detection of up to 72% of the SNPs sites from *Dd2* bulk sample. **B. Relationship between number of SNPs and read depth in single cell samples.** Positions of SNP sites (top panel) in the *Dd2* bulk, EOM23, LOM16, ENM, LNM samples (VQSLOD >6) are compared to the read depth at these locations (bottom panel). Chromosome 1 between 120,000bp – 160,000bp is illustrated by VIVA.

Limitations, Scope, and Future Efforts

One limitation in our comparison between standard and optimized MALBAC-amplified samples was that we only sequenced a single standard MALBAC sample from each parasite stage. However, during our studies we evaluated a total of 7 independent non-optimized samples using ddPCR (3 ENM and 4 LNM) and detected multiple instances of allelic dropout and heavy skewing of the copy number of a known CNV (**Table 4.1 and Supplementary Table 4.10**). These results indicated extreme bias coverage and high levels of contaminating DNA, which made sequencing of these samples futile. Nevertheless, evaluating specific genes is not equivalent to sequencing a whole genome. Thus, while we have adapted MALBAC for amplifying single *P. falciparum* parasite genomes, further studies are required to rigorously evaluate the differences between standard and optimized MALBAC.

A second limitation of our study was our inability to directly compare MALBAC results to those produced using MDA. Our studies specifically sought to adapt MALBAC for amplification of the *Plasmodium* genome; therefore, we did not perform MDA on our samples in parallel. However, in order to gain some insight into the performance of the two WGA methods on the *P. falciparum* genome, we performed limited comparisons with data from a previous MDA-based study (**Figure 4.4B, Figure S4.3 versus S4.6; Table 4.2 versus Supplementary Table 4.16; Supplementary Table 4.11 versus Supplementary Table 4.15; Supplementary Table 4.13 versus Supplementary Table 4.17**). Direct comparisons were constrained by the use of distinct parasite lines (*HB3* vs *Dd2*) and single cell preparation pipelines but the results emphasized the strengths and weaknesses of each method. While MDA is known to exhibit lower single nucleotide amplification error and acquired overall higher genome coverage in *P. falciparum* genome (Trevino et al., 2017) (**Supplementary Table 4.16**), it is not suitable for CNV detection

(Lasken and Stockwell, 2007). MALBAC, on the other hand, can provide high quality SNP identification following strict filtering steps ((Chen et al., 2014) and **Table 4.8**) and the reproducible amplification pattern (**Figure 4.4B**) can be beneficial for both CNV and SNP detection (see below).

Another limitation is related to the lack of MALBAC coverage across certain genomic regions (~40% of the overall genome, ~70% of the intergenic regions, **Table 4.4**), which impacts the detection of genetic variations in these locations. Low read depth resulted in the failure to detect SNPs (**Figure S4.7B**) and variation of coverage leads to inaccuracies in the size and boundaries of CNVs (see *CNV Analysis and Related Considerations*). However, the reproducible pattern of genome coverage by MALBAC provides some advantages. First, as mentioned above, we can exploit this feature to normalize across diverse samples to minimize noise and improve CNV detection; any improvements in the coverage of intergenic regions and uniformity will also impact CNV identification through increased detection of discordant/split reads and more accurate read-depth calling. Second, the consistent coverage pattern allows us to predict a defined set of SNPs that can be consistently detected across pooled single cell samples with a given coverage level.

Finally, we specifically recognize the limitations of our CNV analysis pipeline. First, we confined our assessments to duplications and deletions (**Supplementary Table 4.19)** but have not evaluated other types of structural variations that may also be important for adaptation. Second, we acknowledge that our CNV analysis on single cell genomes is not yet robust (see *CNV Analysis and Related Considerations* above). We also recognize that there is a tradeoff between sensitivity and precision during CNV analysis; accepting the possibility of false positives allows maximal sensitivity to detect novel CNVs. Ultimately, the benefit of single cell genomics is the discovery of minor variants that provide insight into the dynamics of adaptation. While we would not consider individual CNVs identified in our current analysis to be particularly informative, studies assessing relative CNV levels (under condition 1 vs 2) would likely yield informative results using the current methods. To achieve consistent and robust CNV calling, we require a combination of improvements in both amplification methods and analysis tools (as proposed above). This study can be used as a springboard for such advancements.

4.6 Conclusions

It is notable that we can successfully amplify a small, base-skewed genome and detect genetic variations on a single cell level. Our modifications of reaction volume, cycle number, and GC-content of degenerate primers will expand the use of MALBAC-based approaches to organisms not previously accessible because of small genome size or skewed base content. Furthermore, these changes can reduce amplification of undesired contaminating genomes in a sample. The reproducible nature of this WGA method, combined with new genome analysis tools, will reduce the effect of amplification bias when conducting large scale single cell analysis and enhance our ability to explore genetic heterogeneity in the form of both SNPs and CNVs. Thus, we expect this approach to broadly improve study of mechanisms of genetic adaptation in a variety of organisms.

5 CHAPTER V: Genome assembly using Nanopore long reads and Illumina short reads

Author List: Shiwei Liu, Nnenna Ene, Jennifer Guler

Statement of contribution and acknowledgement

Shiwei Liu (University of Virginia) and Jennifer Guler (University of Virginia) designed the study and Shiwei Liu wrote the chapter. Shiwei Liu, Nnenna Ene (University of Virginia) performed the experiments and data analysis.

Our thanks to Michelle Warthan (University of Virginia) for experimental support, Martin Wu (University of Virginia) for Oxford Nanopore sequencing support, and the laboratories of Dr. William Petri Jr (University of Virginia) for their helpful discussions and insight.

5.1 Abstract

Antimalarial drugs play a vital role in curtailing malaria, a human disease caused by *Plasmodium* parasites. However, the efficacy of these drugs is undermined by the acquisition of antimalarial drug resistance in malarial parasites. *Plasmodium falciparum*, the deadliest malaria parasite causing human infection, has developed resistance to every clinical antimalarial drug. Previous studies have implicated genomic variations in the acquisition of antimalarial resistance, including copy number variations (CNVs) where parasites accumulate extra copies of a genomic segment. However, the detection of CNVs has been challenging due to a lack of a complete and accurate genomic assembly. The genome of *P. falciparum* is extremely AT-rich (80.6%) and contains many repetitive sequences, especially in sub-telomeric regions, which are difficult to assemble using only Illumina short reads. In addition, various parasite strains used in vitro studies require their reference genome assemblies for accurate read mapping and CNV detections in these parasites. Long read sequencing methods hold promise to provide more accurate assembly of the parasite genome and resolve CNVs but few groups have used this approach for *Plasmodium* parasites. In this study, we sequenced and assembled the *P. falciparum* genome by combining Nanopore long reads and Illumina short reads to improve the accuracy of reads alignment and downstream CNV analyses.

5.2 Introduction

Despite a lot of efforts to eliminate malaria, this disease continues to be a major and growing global health problem. Antimalarial drugs play a vital role in curtailing malaria, but the efficacy of antimalarial drugs is undermined by the acquisition of antimalarial drug resistance in malaria parasites. Malaria is caused by a protozoan parasite called *Plasmodium*. *Plasmodium falciparum*, the deadliest malaria parasite causing human infection, has developed resistance to every clinical antimalarial drug. Previous studies have implicated genomic variations in the acquisition of antimalarial resistance, including copy number variations (CNVs) where parasites accumulate extra copies of a genomic segment. However, the detection of CNVs has been challenging due to the lack of a complete and accurate genomic assembly.

The first *Plasmodium* genome sequence was published in 2002 for the strain 3D7 using shotgun sequencing (Gardner et al., 2002). Genome sequences for several other primate *Plasmodium* species have been generated (Carlton et al., 2008; Pain et al., 2008; Tachibana et al., 2012; Otto et al., 2014). The 23Mb haploid genome of *P. falciparum* consists of 14 chromosomes and includes 5300 genes. The genome is also extremely AT-rich (80.6%) and contains many repetitive sequences, especially in sub-telomeric regions, which are difficult to assemble with only short reads. Recently, a few studies performed PacBio long read sequencing on *Plasmodium* parasites, de novo assembled the genome, and obtained more complete genome assemblies (Chien Jung-Ting et al., 2016; Vembar et al., 2016; Bryant et al., 2018; Zhang et al., 2021). However, PacBio sequencing is expensive, while Nanopore sequencing is more cost-effective for generating ultra-long reads. With Nanopore ultra-long reads longer than a Megabase (Mb), more continuous whole genome assemblies have been established for several organisms (Giordano et al., 2017; Eccles et al., 2018; Miller et al., 2018; Tyson et al., 2018).

Various parasite strains used in vitro studies are from the different genetic backgrounds and require their own reference genome assemblies for accurate read mapping and CNV detections. Long read sequencing methods hold promise to provide more accurate assembly of the sequenced genomes and resolve CNVs but few groups have used this approach for *Plasmodium* parasites. In this study, we tested Nanopore PCR-free sequencing kit, optimized high molecular weight DNA extraction protocol, sequenced and assembled the *P. falciparum* genome by combining Nanopore long reads and Illumina short reads to improve the accuracy of reads alignment and downstream CNV analyses.

5.3 Materials and Methods

Parasite culture

We thawed erythrocytic stages of *P. falciparum* (*Dd2*, MRA-150, Malaria Research and Reference Reagent Resource Center, BEI Resources) from frozen stocks and maintained them as previously described (Haynes et al., 1976). Briefly, we grew parasites at 37 °C in vitro at 3% hematocrit (serotype A positive human erythrocytes, Valley Biomedical, Winchester, VA) in RPMI 1640 medium (Invitrogen, USA) containing 24 mM NaHCO3 and 25 mM HEPES, and supplemented with 20% human type A positive heat inactivated plasma (Valley Biomedical, Winchester, VA) in sterile, sealed flasks flushed with 5% O_2, 5% CO_2, and 90% N_2 (Guler et al., 2013). We maintained the cultures with media changes every other day and sub-cultured them as necessary to keep parasitemia below 5%. We determined all parasitemia measurements using SYBR green-based flow cytometry (Bei et al., 2010). We routinely tested cultures using the LookOut Mycoplasma PCR Detection Kit (Sigma-Aldrich, USA) to confirm negative Mycoplasma status.

DNA preparation for Nanopore sequencing

We lysed asynchronous *P. falciparum*-infected erythrocytes with 0.15% saponin (Sigma-Aldrich, USA) for 5 min at room temperature and washed them three times with 1× PBS (diluted from 10× PBS Liquid Concentrate, Gibco, USA). We then lysed parasites with 0.1% Sarkosyl Solution (Bioworld, bioPLUS, USA) in the presence of 1 mg/ml proteinase K (from *Tritirachium album*, Sigma-Aldrich, USA) overnight at 37 °C. We first extracted nucleic acids with phenol/chloroform/isoamyl alcohol (25:24:1) pH 8.0 (Sigma-Aldrich, USA) three times using 1.5 ml light Phase lock Gels (5Prime, USA), then further extracted nucleic acids with chloroform twice using 1.5 ml light Phase lock Gels (5Prime, USA). Lastly, we precipitated the DNA with ethanol using the standard Maniatis method (Maniatis et al., 1989). To obtain high molecular weight genomic DNA, we avoided any pipetting during the extraction, transferred solutions by directly pouring it from one tube to another, and mixed solutions by gently inverting the tubes.

We subjected 1 µg of high molecular weight genomic DNA from Dd2 sample to library preparation for Oxford Nanopore sequencing following the Nanopore Native barcoding genomic DNA protocol (version: NBE_9065_v109_revAB_14Aug2019) with 1x Ligation Sequencing kit (SQK-LSK109, Oxford Nanopore Technologies, Oxford, UK). We performed DNA repair and end preparation using NEBNext FFPE DNA Repair Mix (New England Biolabs, Ipswich, MA, USA) and NEBNext End Repair/dA-Tailing Module (New England Biolabs, Ipswich, MA, USA). We cleaned the A-tailed fragments using 0.9X AMPure XP beads (Beckman Coulter, High Wycombe, UK). We then ligated barcodes to the end-prepped DNA using the Native Barcoding Expansion 1–12 kit (EXP-NBD104, Oxford Nanopore Technologies, Oxford, UK) and Blunt/TA Ligase Master Mix (New England Biolabs, Ipswich, MA, USA). We cleaned the barcoded samples using 0.9X AMPure XP beads. Then we pooled barcoded samples in equimolar ratios and subjected them to an adaptor ligation step, using the Adapter Mix II from the Native Barcoding Expansion 1–12 kit and NEBNext Quick Ligation Reaction Buffer (New England Biolabs, Ipswich, MA, USA) as well as Quick T4 DNA Ligase (New England Biolabs, Ipswich, MA, USA). After adaptor ligation, we cleaned the library using AMPure XP beads. We quantified the adapter-ligated and barcoded library using a Qubit fluorimeter (Qubit 1X dsDNA High Sensitivity Assay Kit, Life Technologies, Carlsbad, CA). We sequenced the Dd2 library using a flow cell (R9.4.1, FLO-MIN106D, Oxford Nanopore Technologies, Oxford, UK) on MinION (Oxford Nanopore Technologies, Oxford, UK). To obtain the maximum number of reads, we ran the flow cell for 48 h (controlled and monitored using the MinKNOW software (3.6.5)).

Sequencing data analysis

For the Dd2 sample, we conducted base calling and demultiplexing of the Nanopore sequencing reads, using Guppy (version 3.4.5+fb1fbfb) with the parameter settings "-c dna_r9.4.1_450bps_hac.cfg --barcode_kits "EXP-NBD104" -x auto". We checked the read length and read quality using "Nanoplot" (version 1.0.0). We trimmed the adapters with "qcat" (version 1.1.0) (Oxford Nanopore Technologies) and filtered the reads with a cutoff "length ≥ 500 and Phred value ≥ 10" using the program " filtlong version 0.2.0" (https://github.com/rrwick/Filtlong). To estimate the coverage of sequencing reads in each sample, we aligned the filtered reads to the *Plasmodium falciparum* Dd2 reference genome using "minimap2" (version 2.17) (Li, 2018). "QualiMap" (version 2.2.1) (García-Alcalde et al., 2012) was used to calculate the coverage of the aligned reads.

Genome sequence assembly

To identify the suitable assembly tools for *Plasmodium* genome using nanopore long reads. We first assembled the reads with both long-read-only methods (Flye, Miniasm, Canu) and hybrid methods of nanopore long reads and Illumina short reads (Haslr, Wengan). The Flye, Miniasm, Canu assembled genomes were polished first by Racon (Vaser et al. 2017) using nanopore long reads for 4 iterations, then Medaka (https://github.com/nanoporetech/medaka) polished once using nanopore long reads, then Pilon polished (Walker et al. 2014) using Illumina short reads for 2 iterations following polishing pipelines in previous studies with default setting (Dmitriev et al., 2021; McCartney et al., 2021; Vandenbogaert et al., 2022). The Haslr, Wengan assembled genomes were Pilon polished (Walker et al. 2014) using Illumina short reads for 2 iterations as shown in previous studies (Díaz-Viraqué et al., 2021; Bessette et al., 2022). QUAST (Gurevich et al., 2013) was used to compare the assemblies with the published reference genome (Dd2, v46) from PlasmoDB. BUSCO (Simão et al., 2015) was used to evaluate assembly quality for all genomes.

5.4 Results

High molecular weight DNA sequencing

To confirm the size range of the high molecular DNA of Dd2 parasites extracted by PCI method (see ***Materials and Methods***), we ran the DNA through Pulsed-filed gel electrophoresis. As shown in **Figure 5.1**, we see most of the DNA is larger than 50 kb. We sequenced the high molecular weight DNA on MinION, which generated 150,489 reads with read length above 500 bp and read quality above Q10. The mean read length is 10,902 bp, the N50 of the reads is 15,574bp and the maximum read length is 190,160bp. To estimate how many reads are from the

Dd2 parasite DNA, we mapped the reads to the Dd2 reference genome. As shown in **Table 5.1**, the nanopore reads show 29X genome coverage and 93% nucleotide identity to the Dd2 reference genome.

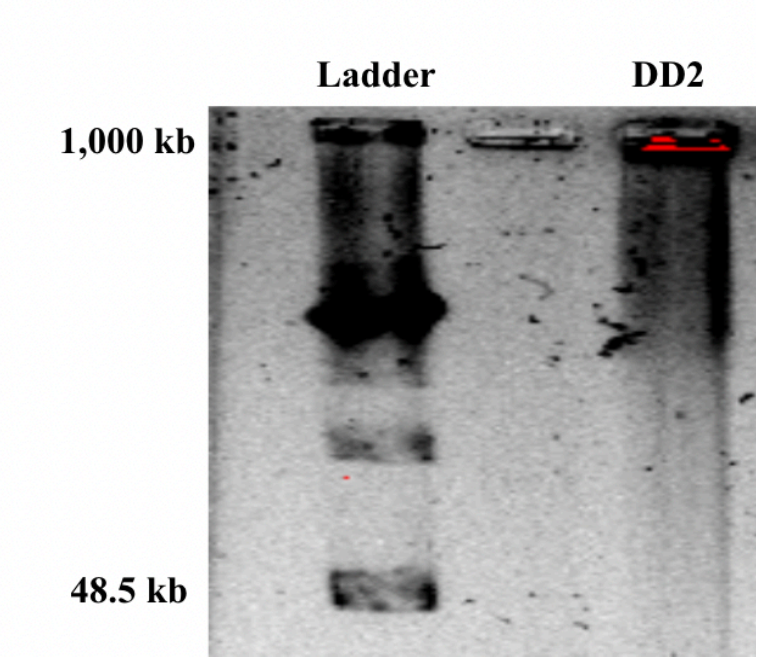


Figure 5.1 Pulsed-filed gel electrophoresis of Dd2 high molecular weight DNA

Table 5.1 Nanopore sequencing reads length and coverage

Sample	Total reads (>500bp, Q10)	Read length N50	Mean read length	Maximum read length	Nucleotide identity	Genome coverage
Dd2	150,489.00	15,574 bp	10,902 bp	190,160 bp	93%	29 X

Table 5.2 Comparison of assembly tools using Dd2 sample

Assembly methods		**Aligned contigs**	**Genome fraction**	**GC %**	**Genomic features**	**Partial genome features**	**Contig N50 (bp)**	**Mis-assemblies**	**Mismatches/ 100kb**	**indels/ 100kb**
Hybrid	Haslr	54	96.76%	18.93%	38076	161	1,456,389	2	9	99
	Wengan	62	95.18%	19.07%	37014	170	840,047	10	14	101
Long reads	Flye	27	99.70%	19.90%	38673	39	1,699,979	23	13	113
	Miniasm	18	98.50%	19.36%	38429	12	1,679,125	14	13	113
	Canu	58	97.85%	21.69%	37756	114	634,783	27	13	116
Reference	-	16	-	19.19%	38792	-	˜1.7Mb	-	-	-

Comparison of assembly tools

To obtain an accurate genome assembly from Nanopore long reads, long-read-only assembly tools (Flye, Miniasm, Canu) and hybrid assembly tools (Haslr, Wengan) were used to generate the genome of Dd2 sample. All assemblies were evaluated and compared to the genome of Dd2 reference genome using the quality assessing tool QUAST (**Table 5.2**). Overall, Flye and Miniasm show the lowest number of aligned contigs (27 for Flye and 18 for Miniasm) and the highest contig N50 length (1,699,979 bp for Flye and 1,679,125 bp for Miniasm), indicating these two tools generate more continuous contigs (**Table 5.2**). Even though Haslr assembly shows 54 aligned contigs, the numbers of differently assembled sequences, mismatches and indels are lower than genomes assembled by other tools when compared to the published Dd2 reference genome (**Table 5.2**). Overall, Miniasm assembly shows fewer differently assembled sequences, mismatches and indels than the Flye assembly when compared to the reference genome (**Table 5.2**). Thus, Miniasm might be more suitable for genome assembly of the *Plasmodium* genome at 29 X coverage.

Improvement of assembly by polishing using Nanopore long reads and Illumina short reads

To improve the accuracy of genome assembly, we performed read polishing by a well-known long-read polishing tool combination (Racon and Medaka) or Illumina short reads (Pilon) (see ***Materials and Methods***). All polished assemblies were evaluated and compared to the genome of Dd2 reference genome using the quality assessing tool QUAST (**Table 5.3**). As shown in **Table 5.3**, the raw assembly only has 8.4% of genome coverage, while the racon polished assembly shows significantly improved genome coverage (98.5%). This is because most of the raw assembly contigs are not aligned the reference genome due to high error rate before polishing, while longer and more accurate contigs were aligned to the reference genome after racon polishing step. The number of aligned contigs also improved from 15 to 18, and the number of mismatches and indels per 100kb significantly decreased after Racon polishing for 4 iterations (**Table 5.3**). The Medaka polishing step also improved the assembly by decreasing the number of mismatches and indels per 100kb (**Table 5.3**). Short reads are known to have a lower error rate than Nanopore long reads. After polishing by Illumina short reads via Pilon, the number of mismatches and indels per 100kb decreased significantly (**Table 5.3**).

Table 5.3 Improvement of genome assembly by polishing steps

Assembly/ polish		Total contigs	Aligned contigs	Genome fraction	GC %	Genomic features	Partial genome features	Contig N50 (bp)	Differently assembled sequences	Mismatches/ 100kb	indels/ 100kb
Miniasm		30	15	8.40%	19.91%	3105	229	1,644,391	2	846	2250
+Racon	Polish by long reads	30	18	98.50%	19.36%	38459	12	1,677,796	15	56	543
+Medaka		41	18	98.50%	19.36%	38439	6	1,676,194	13	37	336
+Pilon	Polish by short reads	41	18	98.50%	19.36%	38429	12	1,679,164	14	16	118
+Pilon		41	18	98.50%	19.36%	38429	12	1,679,125	14	13	113

Table 5.4 Comparison with previous assemblies using different sequencing technologies

	Sanger sequencing		**Illumina sequencing**		**454 pyrosequencing**		SMRT sequencing	Nanopore sequencing
Parasite strain	Dd2	HB3	NP-3D7-S	NP-3D7-L	7C126	SC05	3D7	Dd2
Average Read length (bases)	600–700		36	76	3,000 (paired-end)		12,130	10,902
Number of contigs	4,511	2,971	26,920	22,839	9,452	9,597	21	18
N50 contig size (kb)	11.6	20.6	1.5	1.6	3.3	3.3	1,710	1,679
Largest contig (kb)	79.2	111.9	29.1	24	36.7	34.4	3,290	3,252
Number of assembled bases (Mb)	19.5	23.4	19	21.1	20.8	21.1	23.6	22.7
Average coverage	×7.8	×7.1	×43	×64	×33	×36	×94	x29

Table adapted from Shruthi Sridhar Vembar et al., 2016

The polished Miniasm assembly was then evaluated by BUSCO score using the 3642 conserved genes within *Plasmodium* lineage. Based on the pilon polished assembly, the gene completeness is 95.6%, 58 genes were fragmented, and 104 genes were missing. We compared our Nanopore long read assembly to other de novo assemblies of the *P. falciparum* genome that were generated using data from Roche pyrosequencing, Sanger shotgun sequencing or Illumina-based sequencing by synthesis of different *P. falciparum* strains (**Table 5.4**). Even though the read coverage is low compared to the SMRT sequencing assembly, our assembly contained the least number of contigs, with a contig N50 of 1.68 Mb and the largest contig size of 3.3 Mb that is similar to the SMRT assembly assembled at a much higher read coverage (94X).

5.5 Summary

The advancement of long reads sequencing provides new opportunities to resolve repetitive sequences for improving whole genome assembly. However, it is challenging to directly use Nanopore reads for whole genome assembly due to the relatively high error rate of Nanopore sequencing reads. In this study, we have tested the feasibility of combining Nanopore long reads and Illumina short reads to generate better whole genome assemblies for the genome of *Plasmodium* parasites. We report sequencing and assembly of *Plasmodium falciparum* genomes with an NG50 of 1.68 Mb using unamplified DNA and nanopore reads followed by long-read and short-read consensus improvement. At 29X coverage, we have produced a contiguous assembly that shows a similar N50 value as the SMRT assembly (N50: 1.7 Mb) generated from a much higher read coverage (94X). We report that the combining of long-read polishing and short-read (Illumina) correction reduced the number of differently assembled sequences, mismatches and indels when compared with the previous assembly. As Nanopore sequencing is much more economically feasible than PacBio SMRT sequencing, Nanopore sequencing and long-read assembly provide an economical way to generate continuous assembly of malaria parasite genome. However, due to the presence of many differently assembled sequences when compared to the previous assembly (Dd2 reference genome), experimental validations (such as PCR) are necessary to confirm these differently assembled sequences are not sequencing or assembly error but improvements over the previous assembly. Additionally, the mismatches and indels in the generated assembly are more than a few bases per 100kb, indicating higher sequencing depth might be needed to improve the accuracy of the assembly.

6 CHAPTER VI: Conclusion

6.1 Summary of this dissertation

In this dissertation, I introduced different methods and approaches to study the heterogenous copy number variations that drive the adaptation of the malaria parasite, *Plasmodium falciparum*, to environmental changes like antimalarial drug and other stressed conditions. I described a study to understand the level of diversity in copy number variations of specific genes and genes across the whole genome in *Plasmodium* parasites.

In Chapter II, I discovered an additional CNV that encompasses 3 genes (~5 kb) including GTP cyclohydrolase I (GCH1 amplicon) in DSM1 (antimalarial drug currently under clinical trial) resistant parasite lines with resistant DHODH CNVs. While this locus has been previously implicated in the increased fitness of antifolate-resistant parasites, GCH1 CNVs had not previously been reported to contribute to resistance to other antimalarials. I explored the association between GCH1 and DHODH copy numbers. Through the visualization of single Nanopore long reads and directly quantified the number of tandem GCH1 amplicons in a parental line versus a DSM1-selected line. I found that the GCH1 amplicons share a consistent structure, but I detected more reads that encompassed a higher number of amplicons in the resistant line (up to 7 amplicons) compared to the parental line (3 amplicons). By evaluating variations at this locus across multiple short- and long-read data sets collected from various parasite lines, I concluded that GCH1 is not likely directly contributing to DSM-1 resistance but may compensate for changes in the metabolism of resistant parasites with increased copy number of DHODH.

In Chapter III, I expanded the Nanopore long read analysis used in Chapter II to whole genome analysis instead of limiting to one specific CNV. With the genome-wide single molecule long read analysis, I identified heterogeneous de novo copy number variations in laboratory cultured parasite lines, which are originally cloned in the laboratory and should share similar genetic profiles. The existence of heterogeneous de novo copy number variations provides direct evidence to our hypothesis that the *Plasmodium falciparum* parasites use copy number variations as an important adaptive strategy to survive under changing environments, because minority subclones with de novo CNVs can allow quick adaptation during environmental changes. I also detected increased de novo structural changes for parasites under DSM-1 or Aphidicolin sub-lethal stress treatment, which indicates stress might potentially stimulate CNVs and enhance the adaptation in malaria parasites, but further experiments are needed to confirm these results.

In Chapter IV, I described a single cell sequencing pipeline for an intracellular parasite, *Plasmodium falciparum*, with a small genome size and extreme base content for the genome. Through optimization of a quasi-linear amplification method, I targeted the parasite genome over contaminants and generated coverage levels allowing the detection of minor genetic variants. This pipeline enables detection of parasite heterogeneity contributing to the adaptation of *P. falciparum* parasites and can potentially be conducted at a high throughput level after our optimization described in Chapter IV.

In Chapter V, I sequenced and assembled the *P. falciparum* genome by combining Nanopore long reads and Illumina short reads to improve the accuracy of reads alignment and downstream CNV analyses. With this genome assembly, we can now improve the sequence accuracy in the repetitive regions in the reference genome and resolve CNVs more accurately in the *Plasmodium* genome.

6.2 Discussion and Future directions

Using single cell sequencing and Nanopore sequencing together for detecting heterogeneous CNVs

To understand CNV evolution in malaria parasites, I have pioneered the single parasite genomics method to study CNVs and developed the Nanopore sequencing protocol to detect heterogeneous CNVs. The two methods have their pros and cons, using the results from both methods together can improve the accuracy of heterogeneous CNVs detection. The pros of single cell sequencing pipeline include identifying all the detected CNVs in individual cells and modifying the throughput of experiments easily. While the cons of single cell sequencing pipeline are the detection of false positive CNVs due to whole genome amplification bias, the mapping problem of short reads in repetitive regions. In addition, due to the amplification bias, robust computational tools are necessary to reduce the detection of false positives and improve the detection accuracy for small CNVs, which are more difficult to detect than large CNVs (>= 50kb. In comparison, it is difficult to know whether two different CNVs detected by Nanopore sequencing are from the same cell. The level of diversity to be detected in copy numbers of genes is limited by read-depth. But the pros of Nanopore sequencing are detecting the breakpoints of both small and large CNVs, and direct visualization of the whole structure of copy number variations covered by reads. In addition, the read length of Nanopore long reads is continuously being improved with the advance of Nanopore sequencing technology, thus it is possible to visualize the whole structure of large CNVs. Thus, the two methods can be complementary to each other and validate the results from each other.

Understand the difference among malaria parasites through CNV detection using our new methods

In Chapter III, we identified many heterogeneous de novo copy number variations in four laboratory cultured *Plasmodium falciparum* parasite lines. Similarly, we can expand such analysis to compare the difference in CNVs among parasites from different origins, parasites from different laboratory culture or clinical isolates and parasites species with different GC content in the genomes to better understand the contribution of these factors to the formation of CNVs.

Southeast Asia has given rise to several antimalarial-resistant strains of the malaria parasite. It was earlier hypothesized that Southeast Asian strains associated with multi-drug resistance exhibited a hypermutability phenotype compared to non-Southeast Asian counterparts, enabling the former to acquire mutations and new antimalarial resistance traits at an accelerated rate. Evidence both for and against this hypothesis has been found (Rathod et al., 1997; Brown et al., 2015). Recent studies indicate that a mild mutator phenotype may provide a greater overall benefit for multi-drug-resistant parasites in Southeast Asia in terms of generating antimalarial resistance without incurring detrimental fitness costs (Lee and Fidock, 2016). Most of these studies have focused on the mutation rate of single nucleotides, while another study found that large clinical parasite populations (South East Asia and Africa parasites) show lower CNV frequency and carry smaller CNV than South America clinical parasites with a small population size using SNP-CNV microarray (Cheeseman et al., 2016). Nevertheless, the previous study did not analyze the correlation between CNV frequency or size and antimalarial resistance for the distinct geographical populations. Now, with the single cell sequencing method and Nanopore sequencing method, we can start to understand the copy number variation rate in parasites from different origins with or with antimalarial presence. More specifically, we can start to examine whether Southeast Asia multi-drug resistant strains display higher copy number variation rates compared to non-Southeast Asian counterparts.

It is estimated a large proportion (~5%) of *P. falciparum* genome exhibits CNV (Carret et al., 2005; Cheeseman et al., 2016). Many CNV studies have been based on parasites grown and selected in vitro (both short- and long-term experiments) (Kidgell et al., 2006; Dharia et al., 2009; Samarakoon et al., 2011). The relevance of identified CNVs from lab-cultured parasites to clinical parasites is currently unclear. The deletion of genes only required for parasite's survival in vivo has been detected following long-term *in vitro* culture (Simam et al., 2018). These genes include those involved in the formation of gametocytes, transmission to new hosts via mosquitoes, and cytoadherence to evade host immunity response (Kemp et al., 1992; Alano et al., 1995). The amplification of reticulocyte-binding protein 1 encoding gene (*RH1*) is an in-vitro-associated CNV, with the function of enhancing the asexual replication rate in vitro. Nevertheless, the cytoadherence-associated and gametocyte-linked deletions and *RH1* amplification have not been identified in field isolates of *P. falciparum* (Mackinnon et al., 2009; Nair et al., 2010). Amplification of *pfmdr1* associated with multi-drug resistance has been implicated in both in vitro and clinical studies (Sidhu et al., 2006). For the gene encoding GTP cyclohydrolase 1 (*GCH1*), amplifications have only been found in field isolates and have not been observed in parasites maintained in vitro under low doses of pyrimethamine. In contrast, *DHFR* amplification has only been found in parasites grown in vitro (Heinberg et al., 2013). However, many CNVs not detected in laboratory parasites and carrying unknown clinical or adaptive significance have been discovered in the global surveys of field populations (Cheeseman et al., 2016). Thus, it is possible that laboratory and clinical parasite strains carry distinct basal and stress-triggered amplification profiles that affect their ability to acquire antimalarial resistance. As CNVs are known to emerge extremely rapidly during laboratory selection, clinical isolates might have lower CNV rates than laboratory clones and genes affected by CNVs may be different. With the single cell sequencing method and Nanopore sequencing method, we can also investigate the difference of copy number variation rates of laboratory and clinical *P. falciparum* parasites and obtain a better understanding of what genomic features might trigger amplification and possible mechanisms relevant both in vivo and in vitro.

As we mentioned in Chapters III and IV, the AT-rich genome (80.6% AT content) of *P. falciparum* might influence its CNV rate (Huckaby et al., 2018). Thus, we can also use the single cell sequencing method and Nanopore sequencing method to investigate and compare the CNV rate in *Plasmodium falciparum* with the CNV rate in parasites carrying a more balanced genome content (i.e. *P. knowlesi* with 61.2% AT content). It has been estimated that the *P. knowlesi* genome contains many long monomeric A/T tracks that are important for CNV formation (Huckaby et al., 2018), thus we might also observe many de novo CNVs in *P. knowlesi* genome. Such investigation will add to our knowledge about how specific genome features (i.e., A/T tracks) affect the formation of CNVs.

The consequences of increased copy number of GCH1

In Chapter II, the detection of additional GCH1 CNV in DSM-1 resistant parasites stresses the importance of compensating genetic mutations in the adaptation of malaria parasites. A change in GCH1 copy number arose serendipitously during DSM1 selection and further increases were beneficial for parasite fitness, thus increased copies of GCH1 may facilitate the acquisition of increased resistant DHODH copy numbers. With the common presence of increased copy number of GCH1 in clinical parasite populations, such association increases concerns for new drug development, as parasites with GCH1 CNV background may develop resistance to new drugs like DSM-1 quickly.

As the first study suggesting that GCH1 copy number could contribute to the fitness of drug-resistant parasites outside of the sulfadoxine/pyrimethamine context. It is necessary to validate the role of GCH1 in resistance development to pyrimidine biosynthesis inhibitors. The next step for the study will be understanding the nature of this connection using experiments. Specifically, we can genetically modify the copy number of the GCH1 copy number and compare the fitness differences among the modified parasites carrying various GCH1 copy numbers but same DHODH copy number by grow them in the same cell culture.

Understanding resistance evolution in malaria parasites

Almost half of the global population is at risk for malaria and approximately half a million people die from this disease annually. There is no widely available, highly effective malaria vaccine and many drugs such as chloroquine and artemisinin can no longer be used in certain areas of the world due to resistance (White et al., 2014; Stokes et al., 2021). Drug resistance is one of world’s most urgent health challenges; it limits our ability to control many life-threatening diseases, like cancer and malaria, which affect populations worldwide. Decades of research have focused on identifying new drug targets to replace those that are ineffective due to resistance, but this a costly and slow endeavor that has yielded few success stories. My dissertation research is to understand how malaria parasites adapt to environment changes, such as antimalarial drug treatment. The understanding of parasite evolution through copy number variations positions us to target key processes to overcome resistance evolution.

Previous studies in our laboratory have made significant progress in understanding how parasites change their genome to acquire stable drug resistance (Huckaby et al., 2018; McDaniels et al., 2021). Particularly, previous studies have discovered that CNVs, or extra copies of large regions of the genome, play an important role in resistance evolution and the high AT-content of the parasite's genome contributes to CNV formation. My dissertation research focused on devising methods to sensitively measure gene copy number, evaluate genomic heterogeneity on a single cell level (Liu et al., 2021), and improving our knowledge about the mechanisms behind adaptation (resistance acquisition) in malaria parasites. With these new knowledge and tools, we can investigate our hypothesis of parasite adaptation through copy number variations, which can potentially allow us to develop new strategies to block the processes that contribute to resistance development in malaria parasites in the future. With the understanding of resistance development, we may extend the effectiveness of current and future antimalarial drugs, which will improve overall health for people from malaria-endemic regions, slow the spread of resistant parasites to new areas, and save considerable resources. CNV evolution is also important contributor to drug resistance in other microbes and cancer (LaFleur Michael D. et al., 2006;

Sansregret et al., 2018; Vallette et al., 2019; Todd and Selmecki, 2020); therefore, this research has the potential to impact the control of numerous global public health threats.

7 Reference

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21, 974–984. doi:10.1101/gr.114876.110.

Alano, P., Roca, L., Smith, D., Read, D., Carter, R., and Day, K. (1995). *Plasmodium falciparum*: parasites defective in early stages of gametocytogenesis. Experimental parasitology 81, 227–235.

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. Nature Reviews Genetics 12, 363–376. doi:10.1038/nrg2958.

Amarh, V., and Arthur, P. K. (2019). DNA double-strand break formation and repair as targets for novel antibiotic combination chemotherapy. Future Sci OA 5, FSO411–FSO411. doi: 10.2144/fsoa-2019-0034.

Andersson, S. G. E., Zomorodipour, A., Andersson, J. O., Sicheritz-Pontén, T., Alsmark, U. C. M., Podowski, R. M., et al. (1998). The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature 396, 133–140. doi: 10.1038/24094.

Arlt, M. F., Mulle, J. G., Schaibley, V. M., Ragland, R. L., Durkin, S. G., Warren, S. T., et al. (2009). Replication Stress Induces Genome-wide Copy Number Changes in Human Cells that Resemble Polymorphic and Pathogenic Variants. The American Journal of Human Genetics 84, 339–350. doi: 10.1016/j.ajhg.2009.01.024.

Arriola, E., Lambros, M. B. K., Jones, C., Dexter, T., Mackay, A., Tan, D. S. P., et al. (2007). Evaluation of Phi29-based whole-genome amplification for microarray-based comparative genomic hybridisation. Laboratory Investigation 87, 75–83. doi: 10.1038/labinvest.3700495.

Auburn, S., Campino, S., Clark, T. G., Djimde, A. A., Zongo, I., Pinches, R., et al. (2011). An Effective Method to Purify *Plasmodium falciparum* DNA Directly from Clinical Blood Samples for Whole Genome High-Throughput Sequencing. PLOS ONE 6, e22213. doi: 10.1371/journal.pone.0022213.

B. Marwick and K. Krishnamoorthy Cvequality: Tests for the Equality of Coefficients of Variation from Multiple Groups. R software package version 0.2.0. Available at: https://github.com/benmarwick/cvequality [Accessed October 1, 2019].

Balikagala, B., Fukuda, N., Ikeda, M., Katuro, O. T., Tachibana, S.-I., Yamauchi, M., et al. (2021). Evidence of Artemisinin-Resistant Malaria in Africa. N Engl J Med 385, 1163–1171. doi: 10.1056/NEJMoa2101746.

Beghain, J., Langlois, A.-C., Legrand, E., Grange, L., Khim, N., Witkowski, B., et al. (2016). *Plasmodium* copy number variation scan: gene copy numbers evaluation in haploid genomes. Malaria Journal 15, 206. doi:10.1186/s12936-016-1258-x.

Bei, A. K., Desimone, T. M., Badiane, A. S., Ahouidi, A. D., Dieye, T., Ndiaye, D., et al. (2010). A flow cytometry-based assay for measuring invasion of red blood cells by *Plasmodium falciparum*. Am J Hematol 85, 234–237. doi:10.1002/ajh.21642.

Belikova, D., Jochim, A., Power, J., Holden, M. T. G., and Heilbronner, S. (2020). "Gene accordions" cause genotypic and phenotypic heterogeneity in clonal populations of Staphylococcus aureus. Nature Communications 11, 3526. doi:10.1038/s41467-020-17277-3.

Bessette, M., Ste-Croix, D. T., Brodeur, J., Mimee, B., and Gagnon, A. (2022). Population genetic structure of the carrot weevil (Listronotus oregonensis) in North America. Evolutionary applications 15, 300–315.

Blasco, B., Leroy, D., and Fidock, D. A. (2017). Antimalarial drug resistance: linking *Plasmodium falciparum* parasite biology to the clinic. Nature medicine 23, 917–928. doi:10.1038/nm.4381.

Bopp, S. E. R., Manary, M. J., Bright, A. T., Johnston, G. L., Dharia, N. V., Luna, F. L., et al. (2013). Mitotic Evolution of *Plasmodium falciparum* Shows a Stable Core Genome but Recombination in Antigen Families. PLOS Genetics 9, e1003293. doi:10.1371/journal.pgen.1003293.

Brown, A. C., Moore, C. C., and Guler, J. L. (2020). Cholesterol-dependent enrichment of understudied erythrocytic stages of human *Plasmodium* parasites. Scientific Reports 10, 4591. doi: 10.1038/s41598-020-61392-6.

Brown, T. S., Jacob, C. G., Silva, J. C., Takala-Harrison, S., Djimdé, A., Dondorp, A. M., et al. (2015). *Plasmodium falciparum* field isolates from areas of repeated emergence of drug resistant malaria show no evidence of hypermutator phenotype. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases 30, 318–322. doi: 10.1016/j.meegid.2014.12.010.

Bryant, J. M., Baumgarten, S., Lorthiois, A., Scheidig-Benatar, C., Claës, A., and Scherf, A. (2018). De Novo Genome Assembly of a *Plasmodium falciparum* NF54 Clone Using Single-Molecule Real-Time Sequencing. Genome Announc 6, e01479-17. doi: 10.1128/genomeA.01479-17.

Burbulis, I. E., Wierman, M. B., Wolpert, M., Haakenson, M., Lopes, M.-B., Schiff, D., et al. (2018). Improved molecular karyotyping in glioblastoma. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 811, 16–26. doi: 10.1016/j.mrfmmm.2018.06.002.

Bushnell B. BBMap. Available at: http://sourceforge.net/projects/bbmap (University of California, Berkeley, 2016) [Accessed December 1, 2021].

Cai, X., Evrony, G. D., Lehmann, H. S., Elhosary, P. C., Mehta, B. K., Poduri, A., et al. (2014). Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. Cell reports 8, 1280–1289. doi: 10.1016/j.celrep.2014.07.043.

Campbell, I. M., Shaw, C. A., Stankiewicz, P., and Lupski, J. R. (2015). Somatic mosaicism: implications for disease and transmission genetics. Trends Genet 31, 382–392. doi: 10.1016/j.tig.2015.03.013.

Cantsilieris, S., Baird, P. N., and White, S. J. (2013). Molecular methods for genotyping complex copy number polymorphisms. Genomics 101, 86–93. doi: 10.1016/j.ygeno.2012.10.004.

Carey, M. A., Covelli, V., Brown, A., Medlock, G. L., Haaren, M., Cooper, J. G., et al. (2018). Influential Parameters for the Analysis of Intracellular Parasite Metabolomics. mSphere 3, e00097-18. doi: 10.1128/mSphere.00097-18.

Carlton, J. M., Adams, J. H., Silva, J. C., Bidwell, S. L., Lorenzi, H., Caler, E., et al. (2008). Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature 455, 757–763.

Carret, C. K., Horrocks, P., Konfortov, B., Winzeler, E., Qureshi, M., Newbold, C., et al. (2005). Microarray-based comparative genomic analyses of the human malaria parasite *Plasmodium falciparum* using Affymetrix arrays. Molecular and Biochemical Parasitology 144, 177–186. doi: 10.1016/j.molbiopara.2005.08.010.

Casares, S., Brumeanu, T.-D., and Richie, T. L. (2010). The RTS,S malaria vaccine. Vaccine 28, 4880–4894. doi:10.1016/j.vaccine.2010.05.033.

Cheeseman, I. H., Gomez-Escobar, N., Carret, C. K., Ivens, A., Stewart, L. B., Tetteh, K. K. A., et al. (2009). Gene copy number variation throughout the *Plasmodium falciparum* genome. BMC Genomics 10, 353–353. doi:10.1186/1471-2164-10-353.

Cheeseman, I. H., Miller, B., Tan, J. C., Tan, A., Nair, S., Nkhoma, S. C., et al. (2016). Population Structure Shapes Copy Number Variation in Malaria Parasites. Molecular biology and evolution 33, 603–620. doi: 10.1093/molbev/msv282.

Chen, C., Xing, D., Tan, L., Li, H., Zhou, G., Huang, L., et al. (2017). Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). Science 356, 189. doi: 10.1126/science.aak9787.

Chen, D., Zhen, H., Qiu, Y., Liu, P., Zeng, P., Xia, J., et al. (2018). Comparison of single cell sequencing data between two whole genome amplification methods on two sequencing platforms. Sci Rep 8, 4963–4963. doi: 10.1038/s41598-018-23325-2.

Chen, M., Song, P., Zou, D., Hu, X., Zhao, S., Gao, S., et al. (2014). Comparison of Multiple Displacement Amplification (MDA) and Multiple Annealing and Looping-Based Amplification Cycles (MALBAC) in Single-Cell Sequencing. PLOS ONE 9, e114520. doi: 10.1371/journal.pone.0114520.

Cheong, W.-H., Tan, Y.-C., Yap, S.-J., and Ng, K.-P. (2015). ClicO FS: an interactive web-based service of Circos. Bioinformatics 31, 3685–3687. doi: 10.1093/bioinformatics/btv433.

Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., et al. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods 12, 966–968. doi:10.1038/nmeth.3505.

Chien Jung-Ting, Pakala Suman B., Geraldo Juliana A., Lapp Stacey A., Humphrey Jay C., Barnwell John W., et al. (2016). High-Quality Genome Assembly and Annotation for *Plasmodium* coatneyi, Generated Using Single-Molecule Real-Time PacBio Technology. Genome Announcements 4, e00883-16. doi: 10.1128/genomeA.00883-16.

Chronister, W. D., Burbulis, I. E., Wierman, M. B., Wolpert, M. J., Haakenson, M. F., Smith, A. C. B., et al. (2019). Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. Cell Reports 26, 825-835.e7. doi: 10.1016/j.celrep.2018.12.107.

Chulay, J. D., Watkins, W. M., and Sixsmith, D. G. (1984). Synergistic Antimalarial Activity of Pyrimethamine and Sulfadoxine against *Plasmodium falciparum* In Vitro. The American Journal of Tropical Medicine and Hygiene 33, 325–330. doi:10.4269/ajtmh.1984.33.325.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80–92. doi: 10.4161/fly.19695.

Cirz, R. T., Chin, J. K., Andes, D. R., de Crécy-Lagard, V., Craig, W. A., and Romesberg, F. E. (2005). Inhibition of mutation and combating the evolution of antibiotic resistance. PLoS biology 3, e176.

Claessens, A., Hamilton, W. L., Kekre, M., Otto, T. D., Faizullabhoy, A., Rayner, J. C., et al. (2014). Generation of Antigenic Diversity in *Plasmodium falciparum* by Structured Rearrangement of Var Genes During Mitosis. PLOS Genetics 10, e1004812. doi: 10.1371/journal.pgen.1004812.

Cohen, J. M., Okumu, F., and Moonen, B. (2022). The fight against malaria: Diminishing gains and growing challenges. Science Translational Medicine 14, eabn3256. doi: 10.1126/scitranslmed.abn3256.

Conway, D. J. (2007). Molecular epidemiology of malaria. Clin Microbiol Rev 20, 188–204. doi:10.1128/CMR.00021-06.

Corneveaux, J. J., Kruer, M. C., Hu-Lince, D., Ramsey, K. E., Zismann, V. L., Stephan, D. A., et al. (2007). SNP-based chromosomal copy number ascertainment following multiple displacement whole-genome amplification. BioTechniques 42, 77–83. doi: 10.2144/000112308.

Cortés, A., Crowley, V. M., Vaquero, A., and Voss, T. S. (2012). A View on the Role of Epigenetics in the Biology of Malaria Parasites. PLOS Pathogens 8, e1002943. doi: 10.1371/journal.ppat.1002943.

Cowell, A. N., Istvan, E. S., Lukens, A. K., Gomez-Lorenzo, M. G., Vanaerschot, M., Sakata-Kato, T., et al. (2018). Mapping the malaria parasite druggable genome by using in vitro evolution and chemogenomics. Science 359, 191–199. doi:10.1126/science.aan4472.

Cowman, A. F., Galatis, D., and Thompson, J. K. (1994). Selection for mefloquine resistance in *Plasmodium falciparum* is linked to amplification of the pfmdr1 gene and cross-resistance to halofantrine and quinine. Proc Natl Acad Sci U S A 91, 1143–1147. doi: 10.1073/pnas.91.3.1143.

Cretu Stancu, M., van Roosmalen, M. J., Renkens, I., Nieboer, M. M., Middelkamp, S., de Ligt, J., et al. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nature Communications 8, 1326. doi:10.1038/s41467-017-01343-4.

de Bourcy, C. F. A., De Vlaminck, I., Kanbar, J. N., Wang, J., Gawad, C., and Quake, S. R. (2014). A Quantitative Comparison of Single-Cell Whole Genome Amplification Methods. PLOS ONE 9, e105585. doi: 10.1371/journal.pone.0105585.

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci USA 99, 5261. doi: 10.1073/pnas.082089499.

Deitsch, K. W., and Dzikowski, R. (2017). Variant gene expression and antigenic variation by malaria parasites. Annu Rev Microbiol 71, 625–641.

Deleye, L., Tilleman, L., Vander Plaetsen, A.-S., Cornelis, S., Deforce, D., and Van Nieuwerburgh, F. (2017). Performance of four modern whole genome amplification methods for copy number variant detection in single cells. Sci Rep 7, 3422–3422. doi: 10.1038/s41598-017-03711-y.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43, 491–498. doi: 10.1038/ng.806.

Dharia, N. V., Sidhu, A. B. S., Cassera, M. B., Westenberger, S. J., Bopp, S. E., Eastman, R. T., et al. (2009). Use of high-density tiling microarrays to identify mutations globally and elucidate mechanisms of drug resistance in *Plasmodium falciparum*. Genome Biology 10, R21. doi: 10.1186/gb-2009-10-2-r21.

Díaz-Viraqué, F., Greif, G., Berná, L., and Robello, C. (2021). “Nanopore long read DNA sequencing of protozoan parasites: Hybrid genome assembly of Trypanosoma cruzi,” in Parasite Genomics (Springer), 3–13.

Dmitriev, A. A., Pushkova, E. N., Novakovskiy, R. O., Beniaminov, A. D., Rozhmina, T. A., Zhuchenko, A. A., et al. (2021). Genome sequencing of fiber flax cultivar atlant using oxford nanopore and Illumina platforms. Frontiers in genetics 11, 590282.

Duan, M., Hao, J., Cui, S., Worthley, D. L., Zhang, S., Wang, Z., et al. (2018). Diverse modes of clonal evolution in HBV-related hepatocellular carcinoma revealed by single-cell genome sequencing. Cell Research 28, 359–373. doi: 10.1038/cr.2018.11.

Duanguppama, J., Mathema, V. B., Tripura, R., Day, N. P. J., Maxay, M., Nguon, C., et al. (2019). Polymorphisms in Pvkelch12 and gene amplification of Pvplasmepsin4 in *Plasmodium vivax* from Thailand, Lao PDR and Cambodia. Malaria Journal 18, 114. doi: 10.1186/s12936-019-2749-3.

Eccles, D., Chandler, J., Camberis, M., Henrissat, B., Koren, S., Le Gros, G., et al. (2018). De novo assembly of the complex genome of Nippostrongylus brasiliensis using MinION long reads. BMC biology 16, 1–18.

Foote, S. J., Thompson, J. K., Cowman, A. F., and Kemp, D. J. (1989). Amplification of the multidrug resistance gene in some chloroquine-resistant isolates of P. *falciparum*. Cell 57, 921–930. doi:10.1016/0092-8674(89)90330-9.

Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., et al. (1997). Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature 390, 580–586. doi: 10.1038/37551.

Fu, Y., Li, C., Lu, S., Zhou, W., Tang, F., Xie, X. S., et al. (2015). Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. Proc Natl Acad Sci USA 112, 11923. doi: 10.1073/pnas.1513988112.

Gadalla Nahla B., Adam Ishag, Elzaki Salah-Eldin, Bashir Sahar, Mukhtar Izdihar, Oguike Mary, et al. (2011). Increased pfmdr1 Copy Number and Sequence Polymorphisms in *Plasmodium falciparum* Isolates from Sudanese Malaria Patients Treated with Artemether-Lumefantrine. Antimicrobial Agents and Chemotherapy 55, 5408–5411. doi: 10.1128/AAC.05102-11.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., et al. (2012). Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics 28, 2678–2679. doi:10.1093/bioinformatics/bts503.

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

Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G. S., Hicks, J., et al. (2015). Interactive analysis and assessment of single-cell copy-number variations. Nature methods 12, 1058–1060. doi: 10.1038/nmeth.3578.

Gawad, C., Koh, W., and Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. Nature Reviews Genetics 17, 175.

Giordano, F., Aigrain, L., Quail, M. A., Coupland, P., Bonfield, J. K., Davies, R. M., et al. (2017). De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. Scientific reports 7, 1–10.

Girgis, H. S., DuPai, C. D., Lund, J., Reeder, J., Guillory, J., Durinck, S., et al. (2021). Single-molecule nanopore sequencing reveals extreme target copy number heterogeneity in arylomycin-resistant mutants. Proceedings of the National Academy of Sciences 118, e2021958118. doi: 10.1073/pnas.2021958118.

Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, et al. Gplots: Various R programming tools for plotting data. R package version 3.0.1.1. Available at: https://cran.r-project.org/web/packages/gplots/index.html [Accessed October 1, 2019].

Greninger Alexander L., Roychoudhury Pavitra, Makhsous Negar, Hanson Derek, Chase Jill, Krueger Gerhard, et al. (2018). Copy Number Heterogeneity, Large Origin Tandem Repeats, and Interspecies Recombination in Human Herpesvirus 6A (HHV-6A) and HHV-6B Reference Strains. Journal of Virology 92, e00135-18. doi: 10.1128/JVI.00135-18.

Gresham, D., Usaite, R., Germann, S. M., Lisby, M., Botstein, D., and Regenberg, B. (2010). Adaptation to diverse nitrogen-limited environments by deletion or extrachromosomal element formation of the GAP1 locus. Proceedings of the National Academy of Sciences 107, 18551–18556.

Guler, J. L., Freeman, D. L., Ahyong, V., Patrapuvich, R., White, J., Gujjar, R., et al. (2013). Asexual Populations of the Human Malaria Parasite, *Plasmodium falciparum*, Use a Two-Step Genomic Strategy to Acquire Accurate, Beneficial DNA Amplifications. PLOS Pathogens 9, e1003375. doi:10.1371/journal.ppat.1003375.

Gupta, D. K., Patra, A. T., Zhu, L., Gupta, A. P., and Bozdech, Z. (2016). DNA damage regulation and its role in drug-related phenotypes in the malaria parasites. Sci Rep 6, 23603–23603. doi: 10.1038/srep23603.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075. doi: 10.1093/bioinformatics/btt086.

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

Harrell F. E. Hmisc: Harrell miscellaneous (R package Version 4.3-0). Available at: https://CRAN.R-project.org/package=Hmisc [Accessed May 1, 2019].

Hastings, P., Bull, H. J., Klump, J. R., and Rosenberg, S. M. (2000). Adaptive amplification: an inducible chromosomal instability mechanism. Cell 103, 723–731.

Haynes, J. D., Diggs, C. L., Hines, F. A., and Desjardins, R. E. (1976). Culture of human malaria parasites *Plasmodium falciparum*. Nature 263, 767–769. doi:10.1038/263767a0.

Heinberg, A., and Kirkman, L. (2015). The molecular basis of antifolate resistance in *Plasmodium falciparum*: looking beyond point mutations. Ann N Y Acad Sci 1342, 10–18. doi:10.1111/nyas.12662.

Heinberg, A., Siu, E., Stern, C., Lawrence, E. A., Ferdig, M. T., Deitsch, K. W., et al. (2013). Direct evidence for the adaptive role of copy number variation on antifolate susceptibility in *Plasmodium falciparum*. Molecular microbiology 88, 702–712. doi:10.1111/mmi.12162.

Herman, J. D., Rice, D. P., Ribacke, U., Silterra, J., Deik, A. A., Moss, E. L., et al. (2014). A genomic and evolutionary approach reveals non-genetic drug resistance in malaria. Genome Biol 15, 511–511. doi:10.1186/PREACCEPT-1067113631444973.

Ho, S. S., Urban, A. E., and Mills, R. E. (2020). Structural variation in the sequencing era. Nature Reviews Genetics 21, 171–189. doi:10.1038/s41576-019-0180-9.

Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., Huang, J., et al. (2013). Genome Analyses of Single Human Oocytes. Cell 155, 1492–1506. doi: 10.1016/j.cell.2013.11.040.

Hou, Y., Wu, K., Shi, X., Li, F., Song, L., Wu, H., et al. (2015). Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. GigaScience 4. doi: 10.1186/s13742-015-0068-3.

Huang, L., Ma, F., Chapman, A., Lu, S., and Xie, X. S. (2015). Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. Annu. Rev. Genom. Hum. Genet. 16, 79–102. doi: 10.1146/annurev-genom-090413-025352.

Huckaby, A. C., Granum, C. S., Carey, M. A., Szlachta, K., Al-Barghouthi, B., Wang, Y.-H., et al. (2018). Complex DNA structures trigger copy number variation across the *Plasmodium falciparum* genome. Nucleic Acids Research 47, 1615–1627. doi:10.1093/nar/gky1268.

Hughes, A. E. O., Magrini, V., Demeter, R., Miller, C. A., Fulton, R., Fulton, L. L., et al. (2014). Clonal architecture of secondary acute myeloid leukemia defined by single-cell sequencing. PLoS Genet 10, e1004462–e1004462. doi: 10.1371/journal.pgen.1004462.

Hull, R. M., Cruz, C., Jack, C. V., and Houseley, J. (2017). Environmental change drives accelerated adaptation through stimulated copy number variation. PLOS Biology 15, e2001333. doi: 10.1371/journal.pbio.2001333.

Hyde, J. E. (2007). Drug-resistant malaria - an insight. FEBS J 274, 4688–4698. doi:10.1111/j.1742-4658.2007.05999.x.

Ibrahim, A., Diez Benavente, E., Nolder, D., Proux, S., Higgins, M., Muwanguzi, J., et al. (2020). Selective whole genome amplification of *Plasmodium* malariae DNA from clinical samples reveals insights into population structure. Scientific Reports 10, 10832. doi: 10.1038/s41598-020-67568-4.

Ignatov, K. B., Barsova, E. V., Fradkov, A. F., Blagodatskikh, K. A., Kramarova, T. V., and Kramarov, V. M. (2014). A strong strand displacement activity of thermostable DNA polymerase markedly improves the results of DNA amplification. BioTechniques 57, 81–87. doi: 10.2144/000114198.

Inselburg, J., and Banyal, H. S. (1984). *Plasmodium falciparum*: Synchronization of asexual development with aphidicolin, a DNA synthesis inhibitor. Experimental Parasitology 57, 48–54. doi: 10.1016/0014-4894(84)90061-4.

Jacob, C. G., Tan, J. C., Miller, B. A., Tan, A., Takala-Harrison, S., Ferdig, M. T., et al. (2014). A microarray platform and novel SNP calling algorithm to evaluate *Plasmodium falciparum* field samples of low DNA quantity. BMC Genomics 15, 719–719. doi: 10.1186/1471-2164-15-719.

Janevski, A., Varadan, V., Kamalakaran, S., Banerjee, N., and Dimitrova, N. (2012). Effective normalization for copy number variation detection from whole genome sequencing. BMC genomics 13 Suppl 6, S16–S16. doi: 10.1186/1471-2164-13-S6-S16.

Jensen, M. A., Fukushima, M., and Davis, R. W. (2010). DMSO and Betaine Greatly Improve Amplification of GC-Rich Constructs in De Novo Synthesis. PLOS ONE 5, e11024. doi: 10.1371/journal.pone.0011024.

Jett, C., Dia, A., and Cheeseman, I. H. (2020). Rapid emergence of clonal interference during malaria parasite cultivation. bioRxiv, 2020.03.04.977165. doi: 10.1101/2020.03.04.977165.

Jiang, L., López-Barragán, M. J., Jiang, H., Mu, J., Gaur, D., Zhao, K., et al. (2010). Epigenetic control of the variable expression of a *Plasmodium falciparum* receptor protein for erythrocyte invasion. Proceedings of the National Academy of Sciences 107, 2224–2229.

Kafsack, B. F., Rovira-Graells, N., Clark, T. G., Bancells, C., Crowley, V. M., Campino, S. G., et al. (2014). A transcriptional switch underlies commitment to sexual development in malaria parasites. Nature 507, 248–252.

Kaushal, S., and Freudenreich, C. H. (2019). The role of fork stalling and DNA structures in causing chromosome fragility. Genes Chromosomes Cancer 58, 270–283. doi: 10.1002/gcc.22721.

Kemp, D., Thompson, J., Barnes, D., Triglia, T., Karamalis, F., Petersen, C., et al. (1992). A chromosome 9 deletion in *Plasmodium falciparum* results in loss of cytoadherence. Memórias do Instituto Oswaldo Cruz 87, 85–89.

Kidgell, C., Volkman, S. K., Daily, J., Borevitz, J. O., Plouffe, D., Zhou, Y., et al. (2006). A Systematic Map of Genetic Variation in *Plasmodium falciparum*. PLOS Pathogens 2, e57. doi:10.1371/journal.ppat.0020057.

Kil, E.-J., Kim, S., Lee, Y.-J., Kang, E.-H., Lee, M., Cho, S.-H., et al. (2015). Advanced loop-mediated isothermal amplification method for sensitive and specific detection of Tomato chlorosis virus using a uracil DNA glycosylase to control carry-over contamination. Journal of Virological Methods 213, 68–74. doi: 10.1016/j.jviromet.2014.10.020.

Kirkman, L. A., Lawrence, E. A., and Deitsch, K. W. (2014). Malaria parasites utilize both homologous recombination and alternative end joining pathways to maintain genome integrity. Nucleic Acids Research 42, 370–379. doi: 10.1093/nar/gkt881.

Knouse, K. A., Wu, J., and Amon, A. (2016). Assessment of megabase-scale somatic copy number variation using single-cell sequencing. Genome research 26, 376–384. doi: 10.1101/gr.198937.115.

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. Proceedings. Biological sciences 279, 5048–5057. doi: 10.1098/rspb.2012.1108.

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Selection in the evolution of gene duplications. Genome biology 3, 1–9.

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., and Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biology 20, 117. doi:10.1186/s13059-019-1720-5.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. Genome Res 19, 1639–1645. doi: 10.1101/gr.092759.109.

Kulakov, L. A., McAlister, M. B., Ogden, K. L., Larkin, M. J., and O'Hanlon, J. F. (2002). Analysis of bacteria contaminating ultrapure water in industrial systems. Appl Environ Microbiol 68, 1548–1555. doi: 10.1128/aem.68.4.1548-1555.2002.

Kumar, S., Li, X., McDew-White, M., Reyes, A., Delgado, E., Sayeed, A., et al. (2021). Bulk segregant analysis reveals environment × genotype interactions determining malaria parasite growth. bioRxiv, 2020.09.12.294736. doi: 10.1101/2020.09.12.294736.

Kümpornsin, K., Modchang, C., Heinberg, A., Ekland, E. H., Jirawatcharadech, P., Chobson, P., et al. (2014). Origin of Robustness in Generating Drug-Resistant Malaria Parasites Single-molecule nanopore sequencing reveals extreme target copy number heterogeneity in arylomycin-resistant mutants. Molecular Biology and Evolution 31, 1649–1660. doi:10.1093/molbev/msu140.

LaFleur Michael D., Kumamoto Carol A., and Lewis Kim (2006). Candida albicans Biofilms Produce Antifungal-Tolerant Persister Cells. Antimicrobial Agents and Chemotherapy 50, 3839–3846. doi: 10.1128/AAC.00684-06.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biology 21, 31. doi: 10.1186/s13059-020-1926-6.

Lasken, R. S. (2013). Single-cell sequencing in its prime. Nature Biotechnology 31, 211–212. doi: 10.1038/nbt.2523.

Lasken, R. S., and Stockwell, T. B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnology 7, 19. doi: 10.1186/1472-6750-7-19.

Lauer, S., Avecilla, G., Spealman, P., Sethia, G., Brandt, N., Levy, S. F., et al. (2018). Single-cell copy number variant detection reveals the dynamics and diversity of adaptation. PLOS Biology 16, e3000069. doi: 10.1371/journal.pbio.3000069.

Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biology 15, R84. doi:10.1186/gb-2014-15-6-r84.

Lee Andrew H., Symington Lorraine S., and Fidock David A. (2014). DNA Repair Mechanisms and Their Biological Roles in the Malaria Parasite *Plasmodium falciparum*. Microbiology and Molecular Biology Reviews 78, 469–486. doi: 10.1128/MMBR.00059-13.

Lee, A. H., and Fidock, D. A. (2016). Evidence of a mild mutator phenotype in Cambodian *Plasmodium falciparum* malaria parasites. PloS one 11, e0154166.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. doi:10.1093/bioinformatics/bty191.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, R., Chen, K., Chan, E. W. C., and Chen, S. (2018). Resolution of dynamic MDR structures among the plasmidome of Salmonella using MinION single-molecule, long-read sequencing. Journal of Antimicrobial Chemotherapy 73, 2691–2695. doi: 10.1093/jac/dky243.

Lieber, M. R. (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. Annual review of biochemistry 79, 181–211. doi: 10.1146/annurev.biochem.052308.093131.

Liu, S., Huckaby, A. C., Brown, A. C., Moore, C. C., Burbulis, I., McConnell, M. J., et al. (2021). Single-cell sequencing of the small and AT-skewed genome of malaria parasites. Genome Medicine 13, 75. doi:10.1186/s13073-021-00889-9.

Llorà-Batlle, O., Tintó-Font, E., and Cortés, A. (2019). Transcriptional variation in malaria parasites: why and how. Briefings in Functional Genomics 18, 329–341. doi: 10.1093/bfgp/elz009.

Lorenzi, H. A., Puiu, D., Miller, J. R., Brinkac, L. M., Amedeo, P., Hall, N., et al. (2010). New assembly, reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information. PLoS Negl Trop Dis 4, e716–e716. doi: 10.1371/journal.pntd.0000716.

Lynch, M., and Conery, J. S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. Science 290, 1151. doi: 10.1126/science.290.5494.1151.

Macaulay, I. C., and Voet, T. (2014). Single Cell Genomics: Advances and Future Perspectives. PLOS Genetics 10, e1004126. doi: 10.1371/journal.pgen.1004126.

Mackinnon, M. J., Li, J., Mok, S., Kortok, M. M., Marsh, K., Preiser, P. R., et al. (2009). Comparative transcriptional and genomic analysis of *Plasmodium falciparum* field isolates. PLoS pathogens 5, e1000644–e1000644. doi: 10.1371/journal.ppat.1000644.

MalariaGEN P. *falciparum* Community Project V6.0 pipeline Available at: ftp://ngs.sanger.ac.uk/production/malaria/pfcommunityproject/Pf6/Pf_6_extended_methods.pdf [Accessed January 5, 2021].

MalariaGEN, Ahouidi, A., Ali, M., Almagro-Garcia, J., Amambua-Ngwa, A., Amaratunga, C., et al. (2021). An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples. Wellcome Open Res 6, 42–42. doi: 10.12688/wellcomeopenres.16168.1.

Manary, M. J., Singhakul, S. S., Flannery, E. L., Bopp, S. E. R., Corey, V. C., Bright, A. T., et al. (2014). Identification of pathogen genomic variants through an integrated pipeline. BMC Bioinformatics 15, 63. doi:10.1186/1471-2105-15-63.

Mandt, R. E. K., Lafuente-Monasterio, M. J., Sakata-Kato, T., Luth, M. R., Segura, D., Pablos-Tanarro, A., et al. (2019). In vitro selection predicts malaria parasite resistance to dihydroorotate dehydrogenase inhibitors in a mouse infection model. Science Translational Medicine 11, eaav1636. doi:10.1126/scitranslmed.aav1636.

Maniatis, T., Sambrook, J., and Fritsch, E. F. (1989). Molecular cloning: a laboratory manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

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

Matthews, H., Duffy, C. W., and Merrick, C. J. (2018). Checks and balances? DNA replication and the cell cycle in *Plasmodium*. Parasites & Vectors 11, 216. doi: 10.1186/s13071-018-2800-1.

McCartney, A. M., Hilario, E., Choi, S.-S., Guhlin, J., Prebble, J. M., Houliston, G., et al. (2021). An exploration of assembly strategies and quality metrics on the accuracy of the rewarewa

(Knightia excelsa) genome. Molecular Ecology Resources 21, 2125–2144. doi: 10.1111/1755-0998.13406.

McConnell, M. J., Lindberg, M. R., Brennand, K. J., Piper, J. C., Voet, T., Cowing-Zitron, C., et al. (2013). Mosaic copy number variation in human neurons. Science 342, 632–637. doi: 10.1126/science.1243472.

McDaniels, J. M., Huckaby, A. C., Carter, S. A., Lingeman, S., Francis, A., Congdon, M., et al. (2021). Extrachromosomal DNA amplicons in antimalarial-resistant *Plasmodium falciparum*. Mol Microbiol 115, 574–590. doi:10.1111/mmi.14624.

McDew-White, M., Li, X., Nkhoma, S. C., Nair, S., Cheeseman, I., and Anderson, T. J. C. (2019). Mode and Tempo of Microsatellite Length Change in a Malaria Parasite Mutation Accumulation Experiment. Genome Biology and Evolution 11, 1971–1985. doi: 10.1093/gbe/evz140.

McFeters, G. A., Broadaway, S. C., Pyle, B. H., and Egozy, Y. (1993). Distribution of bacteria within operating laboratory water purification systems. Appl Environ Microbiol 59, 1410–1415.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297–1303. doi: 10.1101/gr.107524.110.

Melnikov, A., Galinsky, K., Rogov, P., Fennell, T., Van Tyne, D., Russ, C., et al. (2011). Hybrid selection for sequencing pathogen genomes from clinical samples. Genome Biology 12, R73. doi: 10.1186/gb-2011-12-8-r73.

Menard, D., and Dondorp, A. (2017). Antimalarial Drug Resistance: A Threat to Malaria Elimination. Cold Spring Harb Perspect Med 7, a025619. doi:10.1101/cshperspect.a025619.

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

Miller, C. A., Hampton, O., Coarfa, C., and Milosavljevic, A. (2011). ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. PloS one 6, e16327–e16327. doi: 10.1371/journal.pone.0016327.

Miller, D. E., Staber, C., Zeitlinger, J., and Hawley, R. S. (2018). Highly contiguous genome assemblies of 15 Drosophila species generated using nanopore sequencing. G3: Genes, Genomes, Genetics 8, 3131–3141.

Nair, S., Miller, B., Barends, M., Jaidee, A., Patel, J., Mayxay, M., et al. (2008). Adaptive Copy Number Evolution in Malaria Parasites. PLOS Genetics 4, e1000243. doi:10.1371/journal.pgen.1000243.

Nair, S., Nkhoma, S. C., Serre, D., Zimmerman, P. A., Gorena, K., Daniel, B. J., et al. (2014). Single-cell genomics for dissection of complex malaria infections. Genome research 24, 1028–1038. doi: 10.1101/gr.168286.113.

Nair, S., Nkhoma, S., Nosten, F., Mayxay, M., French, N., Whitworth, J., et al. (2010). Genetic changes during laboratory propagation: copy number At the reticulocyte-binding protein 1 locus of *Plasmodium falciparum*. Mol Biochem Parasitol 172, 145–148. doi: 10.1016/j.molbiopara.2010.03.015.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature 472, 90–94. doi: 10.1038/nature09807.

Neves, R. P. L., Raba, K., Schmidt, O., Honisch, E., Meier-Stiegen, F., Behrens, B., et al. (2014). Genomic High-Resolution Profiling of Single CKpos/CD45neg Flow-Sorting Purified

Circulating Tumor Cells from Patients with Metastatic Breast Cancer. Clin. Chem. 60, 1290. doi: 10.1373/clinchem.2014.222331.

Ning, L., Li, Z., Wang, G., Hu, W., Hou, Q., Tong, Y., et al. (2015). Quantitative assessment of single-cell whole genome amplification methods for detecting copy number variation using hippocampal neurons. Scientific Reports 5, 11415.

Nkhoma, S. C., Trevino, S. G., Gorena, K. M., Nair, S., Khoswe, S., Jett, C., et al. (2018). Resolving within-host malaria parasite diversity using single-cell sequencing. bioRxiv, 391268. doi: 10.1101/391268.

Nkhoma, S. C., Trevino, S. G., Gorena, K. M., Nair, S., Khoswe, S., Jett, C., et al. (2020). Co-transmission of Related Malaria Parasite Lineages Shapes Within-Host Parasite Diversity. Cell Host & Microbe 27, 93-103.e4. doi: 10.1016/j.chom.2019.12.001.

Nogami, T., Ohto, T., Kawaguchi, O., Zaitsu, Y., and Sasaki, S. (1998). Estimation of Bacterial Contamination in Ultrapure Water: Application of the Anti-DNA Antibody. Anal. Chem. 70, 5296–5301. doi: 10.1021/ac9805854.

Oduola, A. M. J., Weatherly, N. F., Bowdre, J. H., and Desjardins, R. E. (1988). *Plasmodium falciparum*: Cloning by single-erythrocyte micromanipulation and heterogeneity in vitro. Experimental Parasitology 66, 86–95. doi:10.1016/0014-4894(88)90053-7.

Ohama, T., Muto, A., and Osawa, S. (1990). Role of GC-biased mutation pressure on synonymous codon choice in Micrococcus luteus a bacterium with a high genomic GC-content. Nucleic Acids Research 18, 1565–1569. doi: 10.1093/nar/18.6.1565.

Ohkubo, S., Muto, A., Kawauchi, Y., Yamao, F., and Osawa, S. (1987). The ribosomal protein gene cluster of Mycoplasma capricolum. Molecular and General Genetics MGG 210, 314–322. doi: 10.1007/BF00325700.

Osei, M., Ansah, F., Matrevi, S. A., Asante, K. P., Awandare, G. A., Quashie, N. B., et al. (2018). Amplification of GTP-cyclohydrolase 1 gene in *Plasmodium falciparum* isolates with the quadruple mutant of dihydrofolate reductase and dihydropteroate synthase genes in Ghana. PLOS ONE 13, e0204871. doi:10.1371/journal.pone.0204871.

Otto, T. D., Böhme, U., Sanders, M., Reid, A., Bruske, E. I., Duffy, C. W., et al. (2018). Long read assemblies of geographically dispersed *Plasmodium falciparum* isolates reveal highly structured subtelomeres. Wellcome Open Res 3, 52–52. doi: 10.12688/wellcomeopenres.14571.1.

Otto, T. D., Rayner, J. C., Böhme, U., Pain, A., Spottiswoode, N., Sanders, M., et al. (2014). Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. Nature communications 5, 1–9.

Oxford Nanopore Technologies qcat (version 1.1.0). Available at: https://github.com/nanoporetech/qcat [Accessed June 1, 2021].

Oyola, S. O., Gu, Y., Manske, M., Otto, T. D., O'Brien, J., Alcock, D., et al. (2013). Efficient depletion of host DNA contamination in malaria clinical sequencing. J Clin Microbiol 51, 745–751. doi: 10.1128/JCM.02507-12.

Oyola, S. O., Manske, M., Campino, S., Claessens, A., Hamilton, W. L., Kekre, M., et al. (2014). Optimized whole-genome amplification strategy for extremely AT-biased template. DNA Res 21, 661–671. doi: 10.1093/dnares/dsu028.

Oyola, S. O., Otto, T. D., Gu, Y., Maslen, G., Manske, M., Campino, S., et al. (2012). Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. BMC Genomics 13, 1–1. doi: 10.1186/1471-2164-13-1.

Pain, A., Böhme, U., Berry, A. E., Mungall, K., Finn, R., Jackson, A., et al. (2008). The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature 455, 799–803.

Palmer, M. J., Deng, X., Watts, S., Krilov, G., Gerasyuto, A., Kokkonda, S., et al. (2021). Potent Antimalarials with Development Potential Identified by Structure-Guided Computational Optimization of a Pyrrole-Based Dihydroorotate Dehydrogenase Inhibitor Series. Journal of Medicinal Chemistry 64, 6085–6136. doi:10.1021/acs.jmedchem.1c00173.

Perandin, F., Manca, N., Calderaro, A., Piccolo, G., Galati, L., Ricci, L., et al. (2004). Development of a real-time PCR assay for detection of *Plasmodium falciparum*, *Plasmodium vivax*, and *Plasmodium* ovale for routine clinical diagnosis. J Clin Microbiol 42, 1214–1219. doi: 10.1128/jcm.42.3.1214-1219.2004.

Phillips, M. A., and Rathod, P. K. (2010). *Plasmodium* dihydroorotate dehydrogenase: a promising target for novel anti-malarial chemotherapy. Infect Disord Drug Targets 10, 226–239. doi:10.2174/187152610791163336.

Phillips, M. A., Gujjar, R., Malmquist, N. A., White, J., El Mazouni, F., Baldwin, J., et al. (2008). Triazolopyrimidine-based dihydroorotate dehydrogenase inhibitors with potent and selective activity against the malaria parasite *Plasmodium falciparum*. J Med Chem 51, 3649–3653. doi:10.1021/jm8001026.

Phillips, M. A., Lotharius, J., Marsh, K., White, J., Dayan, A., White, K. L., et al. (2015). A long-duration dihydroorotate dehydrogenase inhibitor (DSM265) for prevention and treatment of malaria. Science translational medicine 7, 296ra111-296ra111.

Pickard, A. L., Wongsrichanalai, C., Purfield, A., Kamwendo, D., Emery, K., Zalewski, C., et al. (2003). Resistance to antimalarials in Southeast Asia and genetic polymorphisms in pfmdr1. Antimicrob Agents Chemother 47, 2418–2423. doi: 10.1128/aac.47.8.2418-2423.2003.

Pirooznia, M., Goes, F. S., and Zandi, P. P. (2015). Whole-genome CNV analysis: advances in computational approaches. Front Genet 6, 138–138. doi: 10.3389/fgene.2015.00138.

Preechapornkul, P., Imwong, M., Chotivanich, K., Pongtavornpinyo, W., Dondorp, A. M., Day, N. P. J., et al. (2009). *Plasmodium falciparum* pfmdrl amplification, mefloquine resistance, and parasite fitness. Antimicrobial Agents and Chemotherapy 53, 1509–1515. doi: 10.1128/AAC.00241-08.

Price, R. N., Uhlemann, A.-C., Brockman, A., McGready, R., Ashley, E., Phaipun, L., et al. (2004). Mefloquine resistance in *Plasmodium falciparum* and increased pfmdr1 gene copy number. Lancet 364, 438–447. doi: 10.1016/S0140-6736(04)16767-6.

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

Rand, K. H., and Houck, H. (1990). Taq polymerase contains bacterial DNA of unknown origin. Molecular and Cellular Probes 4, 445–450. doi: 10.1016/0890-8508(90)90003-I.

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

Ravenhall, M., Benavente, E. D., Sutherland, C. J., Baker, D. A., Campino, S., and Clark, T. G. (2019). An analysis of large structural variation in global *Plasmodium falciparum* isolates identifies a novel duplication of the chloroquine resistance associated gene. Scientific Reports 9, 8287. doi: 10.1038/s41598-019-44599-0.

Ribacke, U., Mok, B. W., Wirta, V., Normark, J., Lundeberg, J., Kironde, F., et al. (2007). Genome wide gene amplifications and deletions in *Plasmodium falciparum*. Molecular and Biochemical Parasitology 155, 33–44. doi:10.1016/j.molbiopara.2007.05.005.

Ribaut, C., Berry, A., Chevalley, S., Reybier, K., Morlais, I., Parzy, D., et al. (2008). Concentration and purification by magnetic separation of the erythrocytic stages of all human *Plasmodium* species. Malaria Journal 7, 45. doi: 10.1186/1475-2875-7-45.

Rich, S. M., Leendertz, F. H., Xu, G., LeBreton, M., Djoko, C. F., Aminake, M. N., et al. (2009). The origin of malignant malaria. Proc Natl Acad Sci U S A 106, 14902–14907. doi:10.1073/pnas.0907740106.

Riesenfeld, C., Everett, M., Piddock, L., and Hall, B. G. (1997). Adaptive mutations produce resistance to ciprofloxacin. Antimicrobial agents and chemotherapy 41, 2059–2060.

Rohrback, S., April, C., Kaper, F., Rivera, R. R., Liu, C. S., Siddoway, B., et al. (2018). Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. Proc Natl Acad Sci USA 115, 10804. doi: 10.1073/pnas.1812702115.

Rosenberg, S. M. (2001). Evolving responsively: adaptive mutation. Nature Reviews Genetics 2, 504–515.

Rosenthal, P. J. (2013). The interplay between drug resistance and fitness in malaria parasites. Molecular Microbiology 89, 1025–1038. doi: 10.1111/mmi.12349.

Rothkamm, K., Kruger, I., Thompson, L. H., and Löbrich, M. (2003). Pathways of DNA double-strand break repair during the mammalian cell cycle. Molecular and cellular biology 23, 5706–5715.

Rottmann, M., McNamara, C., Yeung, B. K. S., Lee, M. C. S., Zou, B., Russell, B., et al. (2010). Spiroindolones, a potent compound class for the treatment of malaria. Science 329, 1175–1180. doi: 10.1126/science.1193225.

Rugbjerg, P., Dyerberg, A. S. B., Quainoo, S., Munck, C., and Sommer, M. O. A. (2021). Short and long-read ultra-deep sequencing profiles emerging heterogeneity across five platform Escherichia coli strains. Metabolic Engineering 65, 197–206. doi:10.1016/j.ymben.2020.11.006.

Runtuwene, L. R., Tuda, J. S. B., Mongan, A. E., Makalowski, W., Frith, M. C., Imwong, M., et al. (2018). Nanopore sequencing of drug-resistance-associated genes in malaria parasites, *Plasmodium falciparum*. Scientific Reports 8, 8286. doi: 10.1038/s41598-018-26334-3.

Salcedo-Sora, J. E., Ochong, E., Beveridge, S., Johnson, D., Nzila, A., Biagini, G. A., et al. (2011). The Molecular Basis of Folate Salvage in *Plasmodium falciparum*: CHARACTERIZATION OF TWO FOLATE TRANSPORTERS*. Journal of Biological Chemistry 286, 44659–44668. doi:10.1074/jbc.M111.286054.

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

Samarakoon, U., Gonzales, J. M., Patel, J. J., Tan, A., Checkley, L., and Ferdig, M. T. (2011). The landscape of inherited and de novo copy number variants in a *plasmodium falciparum* genetic cross. BMC Genomics 12, 457. doi: 10.1186/1471-2164-12-457.

San Filippo, J., Sung, P., and Klein, H. (2008). Mechanism of Eukaryotic Homologous Recombination. Annu. Rev. Biochem. 77, 229–257. doi: 10.1146/annurev.biochem.77.061306.125255.

Sansregret, L., Vanhaesebroeck, B., and Swanton, C. (2018). Determinants and clinical implications of chromosomal instability in cancer. Nature Reviews Clinical Oncology 15, 139–150. doi: 10.1038/nrclinonc.2017.198.

Sanz, L. M., Crespo, B., De-Cózar, C., Ding, X. C., Llergo, J. L., Burrows, J. N., et al. (2012). *P. falciparum* In Vitro Killing Rates Allow to Discriminate between Different Antimalarial Mode-of-Action. PLOS ONE 7, e30949. doi: 10.1371/journal.pone.0030949.

Scherf, A., Carter, R., Petersen, C., Alano, P., Nelson, R., Aikawa, M., et al. (1992). Gene inactivation of Pf11-1 of *Plasmodium falciparum* by chromosome breakage and healing: identification of a gametocyte-specific protein with a potential role in gametogenesis. EMBO J 11, 2293–2301.

Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018a). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nature Reviews Genetics 19, 329–346. doi:10.1038/s41576-018-0003-4.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., et al. (2018b). Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15, 461–468. doi:10.1038/s41592-018-0001-7.

Shor, E., Fox, C. A., and Broach, J. R. (2013). The yeast environmental stress response regulates mutagenesis induced by proteotoxic stress. PLoS genetics 9, e1003680.

Sidhu, A. B. S., Uhlemann, A.-C., Valderramos, S. G., Valderramos, J.-C., Krishna, S., and Fidock, D. A. (2006). Decreasing pfmdr1 copy number in *plasmodium falciparum* malaria heightens susceptibility to mefloquine, lumefantrine, halofantrine, quinine, and artemisinin. J Infect Dis 194, 528–535. doi: 10.1086/507115.

Silva, S. R., Almeida, A. C. G., da Silva, G. A. V., Ramasawmy, R., Lopes, S. C. P., Siqueira, A. M., et al. (2018). Chloroquine resistance is associated to multi-copy pvcrt-o gene in *Plasmodium vivax* malaria in the Brazilian Amazon. Malaria Journal 17, 267. doi: 10.1186/s12936-018-2411-5.

Simam, J., Rono, M., Ngoi, J., Nyonda, M., Mok, S., Marsh, K., et al. (2018a). Gene copy number variation in natural populations of *Plasmodium falciparum* in Eastern Africa. BMC Genomics 19, 372–372. doi: 10.1186/s12864-018-4689-7.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics 15, 121–132. doi: 10.1038/nrg3642.

Spealman, P., Burrell, J., and Gresham, D. (2020). Inverted duplicate DNA sequences increase translocation rates through sequencing nanopores resulting in reduced base calling accuracy. Nucleic Acids Research 48, 4940–4945. doi: 10.1093/nar/gkaa206.

Srisutham, S., Suwannasin, K., Mathema, V. B., Sriprawat, K., Smithuis, F. M., Nosten, F., et al. (2020). Utility of *Plasmodium falciparum* DNA from rapid diagnostic test kits for molecular analysis and whole genome amplification. Malaria Journal 19, 193. doi: 10.1186/s12936-020-03259-9.

Stokes, B. H., Dhingra, S. K., Rubiano, K., Mok, S., Straimer, J., Gnädig, N. F., et al. (2021). *Plasmodium falciparum* K13 mutations in Africa and Asia impact artemisinin resistance and parasite fitness. eLife 10, e66277. doi: 10.7554/eLife.66277.

Sundararaman, S. A., Plenderleith, L. J., Liu, W., Loy, D. E., Learn, G. H., Li, Y., et al. (2016). Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. Nature Communications 7, 11078.

Tachibana, S.-I., Sullivan, S. A., Kawai, S., Nakamura, S., Kim, H. R., Goto, N., et al. (2012). *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. Nature genetics 44, 1051–1055.

Tamaki Fabio, Fisher Fabio, Milne Rachel, Terán Fernando Sánchez-Román, Wiedemar Natalie, Wrobel Karolina, et al. (2022). High-Throughput Screening Platform To Identify Inhibitors of Protein Synthesis with Potential for the Treatment of Malaria. Antimicrobial Agents and Chemotherapy 66, e00237-22. doi: 10.1128/aac.00237-22.

Tan, J. C., Miller, B. A., Tan, A., Patel, J. J., Cheeseman, I. H., Anderson, T. J., et al. (2011). An optimized microarray platform for assaying genomic variation in *Plasmodium falciparum* field populations. Genome Biology 12, R35. doi: 10.1186/gb-2011-12-4-r35.

Tattini, L., D'Aurizio, R., and Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Frontiers in Bioengineering and Biotechnology 3, 92. doi: 10.3389/fbioe.2015.00092.

Thaithong, S., Ranford-Cartwright, L. C., Siripoon, N., Harnyuttanakorn, P., Kanchanakhan, N. S., Seugorn, A., et al. (2001). *Plasmodium falciparum*: gene mutations and amplification of dihydrofolate reductase genes in parasites grown in vitro in presence of pyrimethamine. Experimental parasitology 98, 59–70.

Todd, R. T., and Selmecki, A. (2020). Expandable and reversible copy number amplification drives rapid adaptation to antifungal drugs. eLife 9, e58349. doi: 10.7554/eLife.58349.

Tollefson, G. A., Schuster, J., Gelin, F., Agudelo, A., Ragavendran, A., Restrepo, I., et al. (2019). VIVA (VIsualization of VAriants): A VCF File Visualization Tool. Scientific Reports 9, 12648. doi: 10.1038/s41598-019-49114-z.

Treangen, T. J., and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics 13, 36–46. doi:10.1038/nrg3117.

Trevino, S. G., Nkhoma, S. C., Nair, S., Daniel, B. J., Moncada, K., Khoswe, S., et al. (2017). High-Resolution Single-Cell Sequencing of Malaria Parasites. Genome Biol Evol 9, 3373–3383. doi: 10.1093/gbe/evx256.

Tyson, J. R., O'Neil, N. J., Jain, M., Olsen, H. E., Hieter, P., and Snutch, T. P. (2018). MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. Genome research 28, 266–274.

Uwimana, A., Legrand, E., Stokes, B. H., Ndikumana, J.-L. M., Warsame, M., Umulisa, N., et al. (2020). Emergence and clonal expansion of in vitro artemisinin-resistant *Plasmodium falciparum* kelch13 R561H mutant parasites in Rwanda. Nature Medicine 26, 1602–1608. doi: 10.1038/s41591-020-1005-2.

Valenciano, A. L., Fernández-Murga, M. L., Merino, E. F., Holderman, N. R., Butschek, G. J., Shaffer, K. J., et al. (2019). Metabolic dependency of chorismate in *Plasmodium falciparum* suggests an alternative source for the ubiquinone biosynthesis precursor. Scientific Reports 9, 13936. doi:10.1038/s41598-019-50319-5.

Vallette, F. M., Olivier, C., Lézot, F., Oliver, L., Cochonneau, D., Lalier, L., et al. (2019). Dormant, quiescent, tolerant and persister cells: Four synonyms for the same target in cancer. Biochemical Pharmacology 162, 169–176. doi: 10.1016/j.bcp.2018.11.004.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43, 11.10.1-11.10.33. doi: 10.1002/0471250953.bi1110s43.

Vandenbogaert, M., Kwasiborski, A., Gonofio, E., Descorps-Declère, S., Selekon, B., Nkili Meyong, A. A., et al. (2022). Nanopore sequencing of a monkeypox virus strain isolated from a pustular lesion in the Central African Republic. Scientific Reports 12, 10768. doi: 10.1038/s41598-022-15073-1.

Vasuvat, J., Montree, A., Moonsom, S., Leartsakulpanich, U., Petmitr, S., Focher, F., et al. (2016). Biochemical and functional characterization of *Plasmodium falciparum* DNA polymerase δ. Malar J 15, 116–116. doi: 10.1186/s12936-016-1166-0.

Vembar, S. S., Seetin, M., Lambert, C., Nattestad, M., Schatz, M. C., Baybayan, P., et al. (2016). Complete telomere-to-telomere de novo assembly of the *Plasmodium falciparum* genome through long-read (>11 kb), single molecule, real-time sequencing. DNA Res 23, 339–351. doi:10.1093/dnares/dsw022.

Venkatesan, M., Amaratunga, C., Campino, S., Auburn, S., Koch, O., Lim, P., et al. (2012). Using CF11 cellulose columns to inexpensively and effectively remove human DNA from *Plasmodium falciparum*-infected whole blood samples. Malaria Journal 11, 41. doi: 10.1186/1475-2875-11-41.

Viguera, E., Canceill, D., and Ehrlich, S. D. (2001). In vitro replication slippage by DNA polymerases from thermophilic organisms. Journal of Molecular Biology 312, 323–333. doi: 10.1006/jmbi.2001.4943.

Vitak, S. A., Torkenczy, K. A., Rosenkrantz, J. L., Fields, A. J., Christiansen, L., Wong, M. H., et al. (2017). Sequencing thousands of single-cell genomes with combinatorial indexing. Nat Methods 14, 302–308. doi: 10.1038/nmeth.4154.

Volkman, S. K., Sabeti, P. C., DeCaprio, D., Neafsey, D. E., Schaffner, S. F., Milner, D. A., et al. (2007). A genome-wide map of diversity in *Plasmodium falciparum*. Nature Genetics 39, 113–119. doi: 10.1038/ng1930.

Waldvogel Abramowski, S., Tirefort, D., Lau, P., Guichebaron, A., Taleb, S., Modoux, C., et al. (2018). Cell-free nucleic acids are present in blood products and regulate genes of innate immune response. Transfusion 58, 1671–1681. doi: 10.1111/trf.14613.

Walliker, D., Quakyi, I. A., Wellems, T. E., McCutchan, T. F., Szarfman, A., London, W. T., et al. (1987). Genetic Analysis of the Human Malaria Parasite *Plasmodium falciparum*. Science 236, 1661–1666. doi:10.1126/science.3299700.

Wang, P., Brobey, R. K. B., Horii, T., Sims, P., Hyde, J. E., Schapira, A., et al. (1986). Utilization of exogenous folate in the human malaria parasite *Plasmodium falciparum* and its critical role in antifolate drug synergy The Susceptibility of *Plasmodium falciparum* to Sulfadoxine and Pyrimethamine: Correlation of in Vivo and in Vitro Results Synergistic Antimalarial Activity of Pyrimethamine and Sulfadoxine against *Plasmodium falciparum* In Vitro. Molecular Microbiology 32, 239–245. doi:10.4269/ajtmh.1986.35.239.

Wang, R., Lin, D.-Y., and Jiang, Y. (2020). SCOPE: A Normalization and Copy-Number Estimation Method for Single-Cell DNA Sequencing. Cell Systems 10, 445-452.e6. doi: 10.1016/j.cels.2020.03.005.

Wang, X., Chen, H., and Zhang, N. R. (2018). DNA copy number profiling using single-cell sequencing. Brief Bioinform 19, 731–736. doi: 10.1093/bib/bbx004.

Wang, Y., and Navin, N. E. (2015). Advances and Applications of Single-Cell Sequencing Technologies. Molecular Cell 58, 598–609. doi: 10.1016/j.molcel.2015.05.005.

Wang, Y., Gao, Z., Xu, Y., Li, G., He, L., and Qian, P. (2016). An evaluation of multiple annealing and looping based genome amplification using a synthetic bacterial community. The Chinese Society of Oceanography 35, 131–136. doi: https://doi.org/10.1007/s13131-015-0781-x.

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature 512, 155–160. doi: 10.1038/nature13600.

Watkins, W. M., Sixsmith, D. G., Chulay, J. D., and Spencer, H. C. (1985). Antagonism of sulfadoxine and pyrimethamine antimalarial activity in vitro by p-aminobenzoic acid, p-aminobenzoylglutamic acid and folic acid. Molecular and Biochemical Parasitology 14, 55–61. doi:10.1016/0166-6851(85)90105-7.

Wellems, T. E., Panton, L. J., Gluzman, I. Y., do Rosario, V. E., Gwadz, R. W., Walker-Jonah, A., et al. (1990). Chloroquine resistance not linked to mdr-like genes in a *Plasmodium falciparum* cross. Nature 345, 253–255. doi:10.1038/345253a0.

White, N. J., Pongtavornpinyo, W., other, and other (2003). The de novo selection of drug-resistant malaria parasites. Proceedings. Biological sciences 270, 545–554. doi: 10.1098/rspb.2002.2241.

White, N. J., Pukrittayakamee, S., Hien, T. T., Faiz, M. A., Mokuolu, O. A., and Dondorp, A. M. (2014). Malaria. The Lancet 383, 723–735. doi: 10.1016/S0140-6736(13)60024-0.

Wilson, C. M., Volkman, S. K., Thaithong, S., Martin, R. K., Kyle, D. E., Milhous, W. K., et al. (1993). Amplification of pfmdr1 associated with mefloquine and halofantrine resistance in *Plasmodium falciparum* from Thailand. Molecular and Biochemical Parasitology 57, 151–160. doi:10.1016/0166-6851(93)90252-S.

Woodrow, C. J., and White, N. J. (2017). The clinical impact of artemisinin resistance in Southeast Asia and the potential for future spread. FEMS Microbiology Reviews 41, 34–48. doi: 10.1093/femsre/fuw037.

World Health Organization (2021). World Malaria Report 2021.

Woyke, T., Sczyrba, A., Lee, J., Rinke, C., Tighe, D., Clingenpeel, S., et al. (2011). Decontamination of MDA Reagents for Single Cell Whole Genome Amplification. PLOS ONE 6, e26161. doi: 10.1371/journal.pone.0026161.

Ye, J., McGinnis, S., and Madden, T. L. (2006). BLAST: improvements for better sequence analysis. Nucleic Acids Research 34, W6–W9. doi:10.1093/nar/gkl164.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. Genome research 19, 1586–1592. doi: 10.1101/gr.092981.109.

Zhang, C., Oguz, C., Huse, S., Xia, L., Wu, J., Peng, Y.-C., et al. (2021). Genome sequence, transcriptome, and annotation of rodent malaria parasite *Plasmodium* yoelii nigeriensis N67. BMC Genomics 22, 303. doi: 10.1186/s12864-021-07555-9.

Zhang, C., Zhang, C., Chen, S., Yin, X., Pan, X., Lin, G., et al. (2013). A Single Cell Level Based Method for Copy Number Variation Analysis by Low Coverage Massively Parallel Sequencing. PLOS ONE 8, e54236. doi: 10.1371/journal.pone.0054236.

Zhang, L., and Vijg, J. (2018). Somatic Mutagenesis in Mammals and Its Implications for Human Disease and Aging. Annu Rev Genet 52, 397–419. doi: 10.1146/annurev-genet-120417-031501.

Zhang, X., Liang, B., Xu, X., Zhou, F., Kong, L., Shen, J., et al. (2017). The comparison of the performance of four whole genome amplification kits on ion proton platform in copy number variation detection. Bioscience Reports 37. doi: 10.1042/BSR20170252.

Zhang, Z. D., Du, J., Lam, H., Abyzov, A., Urban, A. E., Snyder, M., et al. (2011). Identification of genomic indels and structural variations using split reads. BMC Genomics 12, 375–375. doi: 10.1186/1471-2164-12-375.

Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC bioinformatics 14 Suppl 11, S1–S1. doi: 10.1186/1471-2105-14-S11-S1.

Zong, C., Lu, S., Chapman, A. R., and Xie, X. S. (2012). Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. Science 338, 1622. doi: 10.1126/science.1229164.