2015-04-15

Systematic Analysis of Duplications and Deletions in the Malaria Parasite P. falciparum: A Dissertation

Derrick K. DeConti
University of Massachusetts Medical School

SYSTEMATIC ANALYSIS OF DUPLICATIONS AND DELETIONS IN THE MALARIA PARASITE *P. FALCIPARUM*

A Dissertation Presented

By

Derrick K. DeConti

Submitted to the Faculty of the

University of Massachusetts Graduate School of Biomedical Sciences, Worcester

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

April 15, 2015

Bioinformatics and Integrative Biology

SYSTEMATIC ANALYSIS OF DUPLICATIONS AND DELETIONS IN THE MALARIA PARASITE P. FALCIPARUM

A Dissertation Presented By

Derrick K. DeConti

The signatures of the Dissertation Committee signify completion and approval as to style and content of the Dissertation.

[redacted]

Jeffrey Bailey, M.D., Ph.D., Thesis Advisor

[redacted]

Douglas Golenbock, M.D., Member of Committee

[redacted]

Ann Moormann, Ph.D., Member of Committee

[redacted]

Evgeny Rogaev, Ph.D., Member of Committee

[redacted]

V. Ann Stewart, D.V.M., Ph.D., Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee.

[redacted]

Zhiping Weng, Ph.D., Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the School.

[redacted]

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

Interdisciplinary Graduate Program

April 15, 2015

Acknowledgements

I would like to thank my mentor and thesis advisor, Dr. Jeffrey A. Bailey, for his support and mentorship through my scientific career. His expertise, advice, and patience helped create a environment for scientific discourse and experimentation.

I would like to thank my thesis research advisor committee members, Dr. Zhiping Weng, Dr. Doug Golenbock, Dr. Ann Moormann, and Dr. Evgeny Rogaev. Their mentorship of my thesis work helped me progress my research to this body of work.

I would like to thank my collaborators at the Harvard School of Public Health and the Broad Institute, Dr. Daniel Neafsey, Dr. Sarah Volkman, and Dr. Dyan Wirth. Without their collaboration and help, I would not have been able to test my methods on a natural population.

I would like to thank my collaborators at UNC and Duke, Dr. Steve Taylor, Dr. Jonathan Juliano, and Christian Parobek, for providing opportunities to exercise my skills, knowledge, and expertise on scientific endeavors outside the domain of my thesis.

I would like to thank past and present lab members for their experience, skills, mentorship, and camaraderie, Dr. Ozkan Aydemir, Dr. Richard Lambrecht, Nicholas Blouin, Yasin Kaymaz, Guang Xu, and Nicholas Hathaway.

Finally, I would like to thank my family, friends, and loved ones for all their support during my thesis work, whether that was through their sympathies, patience, camaraderie, or welcome diversions.

Abstract

Duplications and deletions are a major source of genomic variation. Duplications, specifically, have a significant impact on gene genesis and dosage, and the malaria parasite *P. falciparum* has developed resistance to a growing number of anti-malarial drugs via gene duplication. It also contains highly duplicated families of antigenically variable allelic genes. While specific genes and families have been studied, a comprehensive analysis of duplications and deletions within the reference genome and population has not been performed. We analyzed the extent of segmental duplications (SD) in the reference genome for *P. falciparum*, primarily by a whole genome self alignment. We discovered that while 5% of the genome identified as SD, the distribution within the genome was partition clustered, with the vast majority localized to the subtelomeres. Within the SDs, we found an overrepresentation of genes encoding antigenically diverse proteins exposed to the extracellular membrane, specifically the *var*, *rifin*, and s*tevor* gene families. To examine variation of duplications and deletions within the parasite populations, we designed a novel computational methodology to identify copy number variants (CNVs) from high throughput sequencing, using a read depth based approach refined with discordant read pairs. After validating the program against *in vitro* lab cultures, we analyzed isolates from Senegal for initial tests into clinical isolates. We then expanded our search to a global sample of 610 strains from Africa and South East Asia, identifying 68 CNV regions. Geographically, genic CNV were found on average in less than 10% of the population, indicating that CNV are rare. However, CNVs at high frequency were almost exclusively duplications associated with known drug resistant CNVs. We also identified the novel biallelic duplication of the *crt* gene – containing both the chloroquine resistant and sensitive allele. The synthesis of our SD and CNV analysis indicates a CNV conservative *P. falciparum* genome except where drug and human immune pressure select for gene duplication.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

array CGH – array comparative genomic hybridization

BAM – binary sequence alignment/map; compressed container of short read alignment information

BED – browser extensible data; tab delimited text file that defines a annotation track

CNV – copy number variant

gch – GTP-cyclohydrolase gene; increased gene dosage related to anti-folate resistance

GO – gene ontology

GWAS – genome wide association study

gdv – gametocyte development protein

mdr1 – multidrug resistance gene; involved in resistance to multiple anti-malarial drugs

qPCR – quantitative polymerase chain reaction

SAM – sequence alignment/map; container of short read alignment information

SD – segmental duplication

Preface

The work presented in Chapter II is in preparation for publication.

The work presented in Chapter III is in preparation for publication. This work was in collaboration with Kate M. Fernandez[1], Dr. Sarah K. Volkman[1], Dr. Dyann F. Wirth[1], and Dr. Daniel E. Neafsey[2]. They have provided the sequencing data and *in vitro* cultures for the 33 Senegal P. falciparum samples upon which we first tested our computational methods on a natural population.

The work presented in Chapter IV is in preparation for publication.

[1] Harvard School of Public Health, Department of Immunology and Infectious Disease, Cambridge, MA

[2] The Broad Institute, Cambridge, MA

CHAPTER I: Introduction

1.1 Genomic Duplications and Deletions

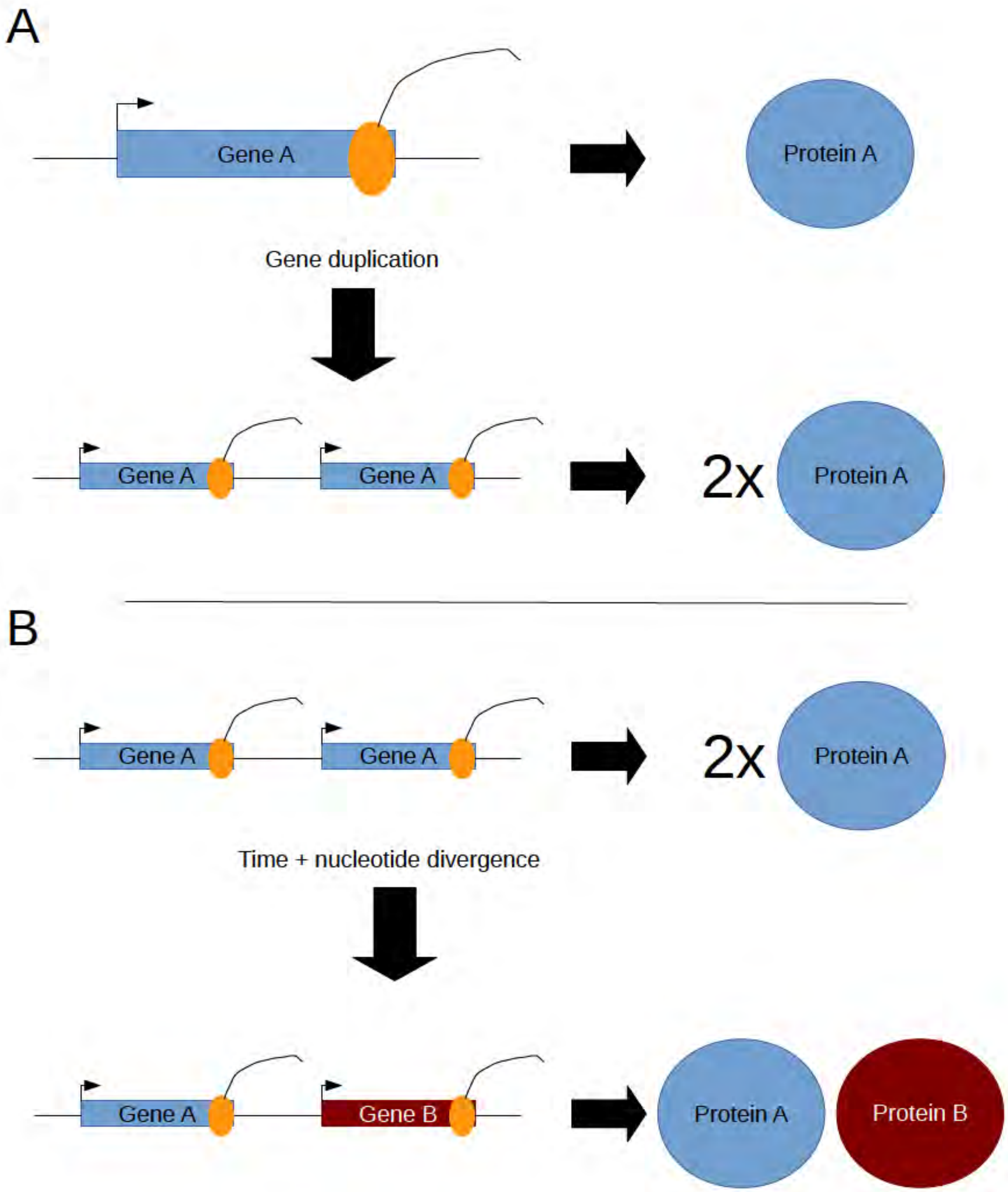
Duplications and deletions are the gain and loss of nucleotide sequence in the genome. Genomic duplications are a significant means of genetic adaptation found through all organisms. The duplication and deletion of genic content is conducive to rapid gene dosage control, genesis of new gene function, and diversification of function within a gene family. However, the change in gene expression from duplications and deletions is not without consequence, and can be subject to significant selective pressures for fixation or removal from the population. This results in genetic copy changes that are transient and highly variable within the population of the organism, resulting in copy number variation.

During evolution, genes are often subject to copy number changes, whether a duplication or deletion of the gene [1]. The size of the region affected can range from single genes to entire chromosomes [1, 2]. A copy number change has immediate short term ramifications, specifically on gene dosage, but over time, can result in new functions arising as mutations occur [3]. While deletions usually have deleterious effects and are rarely propagated, duplications have a number of possible evolutionary end points [4]. The evolutionary fate of duplications is dependent on the selective pressures on the gene. In situations where increased protein production confers an advantage, gene duplication is a simple means to provide increased protein production, as it doubles the available transcribable nucleotide sequence. It is also a more rapid means to control gene expression than through nucleotide substitution as the gene duplication rate has been found to be significantly higher than the per nucleotide substitution rate [5]. This has been a significant area of investigation due the effect increased gene dosage has on pathogen drug resistance.

A primary effect of gene duplication is increased gene dosage (Figure 1.1) [6]. Given positive selection, the duplication can become fixed in the population, and the copies will both remain highly conserved [7]. Changes in gene dosage is an important means of mutability and adaptability for organisms [8]. Commonly, it is a means for pathogenic resistance to chemotherapy, such as the duplication of *bla*$_{\text{SHV-11}}$ gene in *Klebsiella pneumoniae* for increased amoxicillin resistance [9]. The increased gene dosage can also be present to compensate for another deleterious mutation. For example, resistance to actinonin, a peptide deformylase inhibitor, arises via mutations to methionyl-tRNA formyltransferase and incurs a large fitness disadvantage in the absence of actinonin [10]. In the absence of actinonin pressure, duplications are rapidly removed from the gene pool due to incurred fitness disadvantages. However, gene duplications to *metZ* and *metW* genes have been shown to mitigate the fitness disadvantage of actinonin resistance [11]. Gene dosage change by duplication is a rapid, compared to nucleotide substitution, and effective reaction to various selective pressures.

The balance of gene dosage is typically at an effective equilibrium, and so the genome is resistant to many changes in gene regulation. Therefore, if the increased gene dosage has overall a deleterious effect, it results in the rapid removal of the duplication from the population [3]. With little to no selective pressure to maintain the gene duplication, the fate of duplicated genes is loss from the population due to deleterious

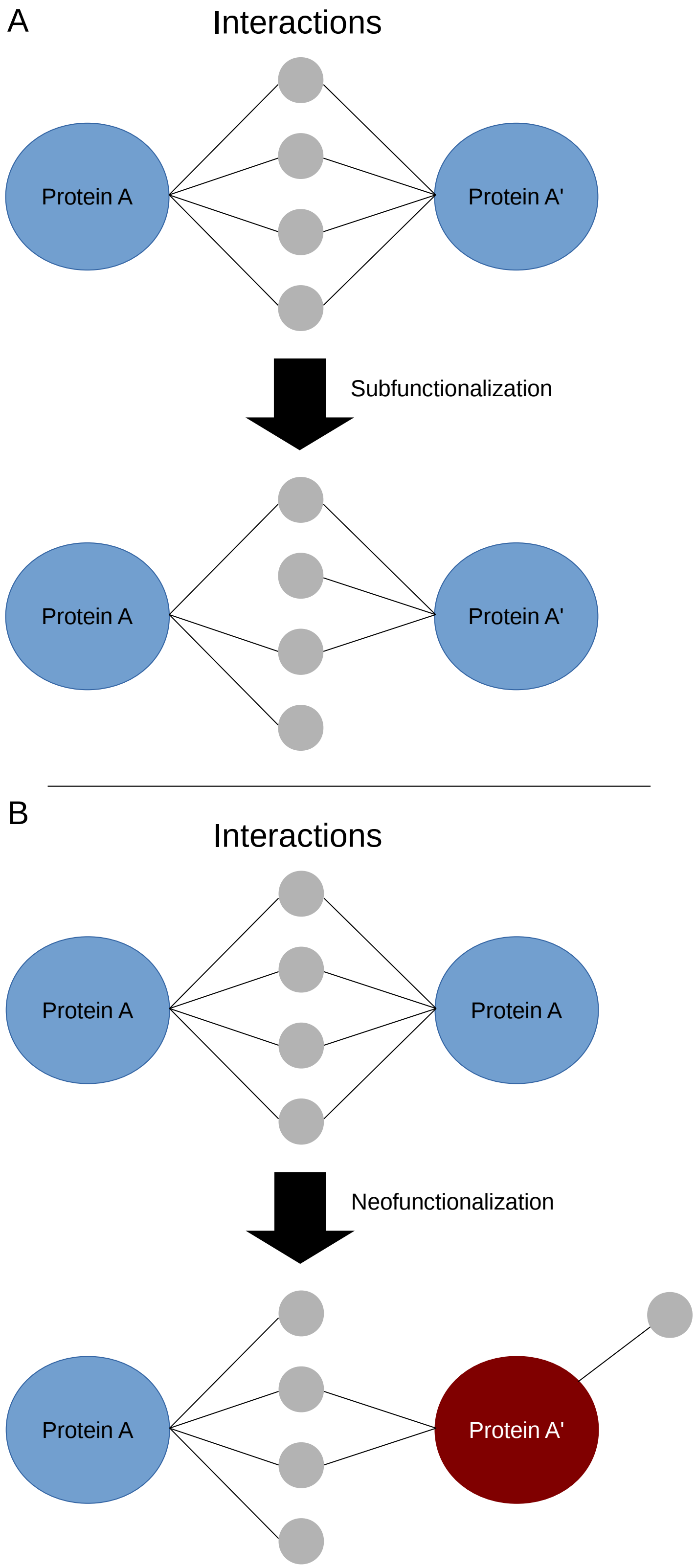
Figure 1.1 The effects of gene duplication. A) Gene duplication results in increased gene dosage, as the duplication of a gene directly results in increased protein production of the gene. B) Over time, one copy of a gene duplication may accrue enough mutations to produce a protein with a new function. This results in a reversion to previous gene dosage levels and an increase in potential functionality.


A
Gene A
Protein A
Gene duplication
Gene A
Gene A
2x
Protein A
B
Gene A
Gene A
2x
Protein A
Time + nucleotide divergence
Gene A
Gene B
Protein A
Protein B

selection or loss through random drift [12]. If a gene is maintained, often silencing mutations occur that result in the pseudogenization of the gene. More rarely relaxed selective pressures on the gene dosage change can maintain the duplication for the gain of advantageous mutations and new function [13–15].

A new function can evolve by one of two courses: subfunctionalization or neo-functionalization (Figure 1.2) [16]. In subfunctionalization, accumulation of mutations results in the combination of both copies providing the original and necessary function [16–18]. This decouples the domains or the original gene, resulting in dependent evolutionary trajectories and further specialization of functions. Neofunctionalization is the more common fate for duplicated genes that are not silenced [15]. In the case of neofunctionalization, one copy attains new functionality separate from the original copy [15, 19]. This provides an important avenue for genetic adaptation [4]. Evidence indicates that the new copy, if not silenced, will undergo rapid and asymmetric mutation [13, 15, 20]. Recent evidence supports neofunctionalization as the more common means of new gene specialization and gain of function over subfunctionalization [13]. This indicates an importance in the maintenance of original gene copy function. Neofunctionalization can result in large duplicated gene families with diverse or redundant functionality [1, 21]. This is especially evident in gene families undergoing diversifying selection, such as the human olfactory receptor contingent [22, 23]. It is also a common means to increase antigen diversity in pathogens, including the malaria parasite *Plasmodium falciparum*. *P. falciparum*'s ability to chronically infect the human is thought to account for a large repertoire of diverse antigens to counter the human immune response [24]. Increased

Figure 1.2 Subfunctionalization and neofunctionalization of gene duplication. A) The equivalent loss of function rate between two duplicated genes results subfunctionalization. The two duplicate genes have specialized in various functions of the ancestral gene. B) The asymmetric loss of function between two duplicate genes and the gain of new function of one of the duplicates is the typical course of action for neofunctionalization.


A
Interactions
Protein A
Protein A'
Subfunctionalization
Protein A
Protein A'
B
Interactions
Protein A
Protein A
Neofunctionalization
Protein A
Protein A'

diversity and redundancy in function can provide a greater fitness advantage.

Numerous studies have investigated the extent of duplications and deletions to better understand the impact of the subsequent genetic diversity on fitness [25–29]. Genomic duplications are an avenue of genetic adaptation found in all types of organisms. They provide a means for rapid gene dosage control, genesis of new gene function, and diversification of function within a gene family. However, the genetic copy changes are also very transient and so are affected strongly by purifying selection [3]. This makes detection of these copy number changes difficult.

1.2 Detection of Duplications and Deletions

Duplications and deletions are difficult phenomena to detect and for which to delineate lineage. However, current sequencing efforts have made it possible to detect ancient and recent copy number changes [30, 31]. With the advent of complete high quality reference genomes [32, 33], comparison within the reference genome and between species has made the delineation of duplications and deletions possible [34, 35]. Array comparative genomic hybridization (array CGH) and high throughput sequencing technologies have also made it possible to detect genome wide copy number variation [30]. Detection of these duplications and deletions aids in understanding past and current evolutionary pressures [36].

Duplications and deletions leave behind evidence of their existence in the genome – the sequence itself. Segmental duplications are blocks of normal genomic DNA that occur more than once within the genome, such as the tracts of duplicated sequence composing much of the mammalian repertoire of olfactory genes [37]. Segmental duplications are identified by regions of high similarity sequence within the genome. We can infer regions of duplication and fixation from the divergence of sequence, with greater divergence related to greater likelihood of fixation. Segmental duplications can point to sequence in the genome that may still be undergoing variation in the greater population. For example, mammalian genomes have shown a large expansion of olfactory genes, but in primates the extent of duplication of olfactory genes has been decreasing [36]. Segmental duplications are most easily via self-alignment of sequence from the reference genome, either with amino acid or nucleotide sequence [38–40]. However, only duplications can be reliably detected. Without related species for comparison, deletions of genomic regions leave little trace of their occurrence. Species specific duplication is difficult to determine from more ancient duplication without cross comparison between genomes [41]. High similarity sequence between species can identify potential orthologous genes between the species and paralogous genes within the species. Examination of what genes are duplicated is important for understanding the evolutionary pressures upon duplications after speciation [42]. Using sequence data from related species, we can also detect deletion events in our species of interest [36].

Genomes are not static, and the continual production of duplications and deletions means that genomes are constantly in flux. The resultant variation caused by periodic genomic duplication and deletion is referred to as copy number variation. These are duplications and deletions that have not become fixed throughout the population, and so vary between individual genomes. These genetic copy change events are in the midst of either fixation for gene dosage advantages, silencing, or neo- and subfunctionalization [43]. Currently, the capabilities exist to rapidly detect copy number variations genome-wide between individuals [44]. This can be done either with array CGH or high throughput sequencing, making it possible to survey entire populations for copy number variation. Surveying populations for copy number variation can identify variants between individuals, potentially highlighting a fitness advantage to these copy number variations [45].

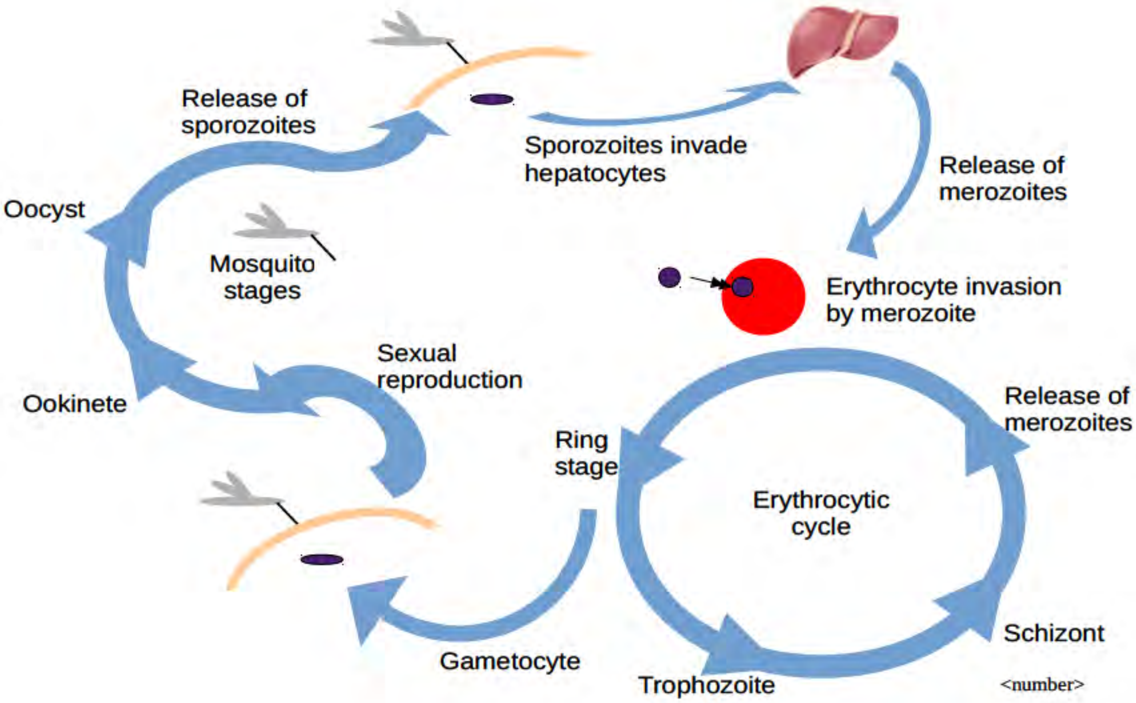
Currently, there are methods for the detection of duplications and deletions, both ancient and recent. These methods involve either self-alignment of the reference genome or read depth analysis of multiply aligned sequences against the reference genome [31, 46, 47]. The analysis of these duplications and deletions provides important insight into the evolutionary pressures at work. These insights are not only important to understand evolutionary forces at work on the human genome [48], but also play an important role in pathogenic genomes [49]. Of great interest is the pathogenic parasite *P. falciparum*. The sheer scale and rapidity of population *P. falciparum* turnover results in duplications and deletions having an important and rapid role in adaptation. Genetic duplications are extensively used by pathogenic agents for diversification of whole gene families for the purpose of evasion of the human immune system, *P. falciparum* included. In addition, the mounting selective pressure from chemotherapeutic therapies has increased the role of gene dosage control by genetic copy number variation in *P. falciparum* [50, 51]. This has spurred increased efforts to rapidly survey *P. falciparum* populations to 1) understand the extent of known drug resistance-related CNVs and 2) identify previously unknown drug resistance-related CNVs.

1.3 *Plasmodium falciparum*

In *P. falciparum*, a human malaria parasite, the duplication and deletion of genes is a major means of adaptation [52]. Malaria is a global epidemic, responsible for an estimated 219 million cases and 660,000 deaths a year [53]. It primarily affects tropical areas and developing nations. Greater than 90% of malaria related mortalities are caused by *Plasmodium falciparum* [53]. In addition to the mortality totals it inflicts, it also significantly impacts local economies, both due to the health care burden and the loss of work time at all ages from recurrent infection [54]. This species is distributed in the tropics worldwide, but the majority of cases occur in Sub-Saharan Africa. Children tend to suffer disproportionately from the most serious clinical syndromes of *P. falciparum* [53].

The life cycle of *P. falciparum* involves two hosts, both the mosquito vector and the human host (Figure 1.3). Both aspects of its life cycle involve points of significant duplication of the genome, indicating points of potential recombination and therefore duplication and deletion of parts of the genome. There are four points in the life cycle of *P. falciparum* where asexual duplication of the genome occurs to produce numerous progeny, in addition to sexual reproduction between a male and female gametes. Upon uptake of blood from a human host of *P. falciparum*, the sexual reproductive cycle occurs between parasites. Male gametocytes undergo further duplication of the genome to 8N ploidy, before subsequently dividing into eight haploid gametes [55]. Fertilization

Figure 1.3 Life cycle of the malaria parasite P. falciparum.


Release of sporozoites
Sporozoites invade hepatocytes
Release of merozoites
Oocyst
Mosquito stages
Erythrocyte invasion by merozoite
Sexual reproduction
Ookinete
Ring stage
Release of merozoites
Erythrocytic cycle
Schizont
Gametocyte
Trophozoite
<number>

between a male and female gamete results in a diploid zygote, which will undergo genome duplication to become tetraploid ookinetes [56]. Eventual differentiation of the zygote to an oocyst and then the sporogony of thousands of sporozoites [57]. From gametocyte to sporozoite, the *P. falciparum* parasite traverses from the mosquito midgut to the salivary glands [58]. The sporozoites infect the human host from the saliva of the mosquito during a blood meal [59].

Upon infection by mosquito bite, the sporozoite infects hepatocytes, starting the life cycle of the *P. falciparum* parasite in the human host [59]. Within the hepatocytes, the sporozoite matures further until schizogony, resulting in serial duplications of the genome and subsequent release of thousands of merozoites upon lysis of the hepatocyte [56]. The merozoites infect erythrocytes to enter the erythrocyte stage of the *P. falciparum* life cycle. During this erythrocytic cycle, the parasite undergoes asexual reproduction during schizogony to produce about 16 duplicated genomes [60]. These schizonts eventually divide into individual merozoites to reinfect a new erythrocyte and begin the cycle anew [61].

Untreated, the erythrocytic stage is the longest stage in the *P. falciparum* life cycle. Despite treatment being widely available and prescribed, asymptomatic malaria still remains a major reservoir of long term *P. falciparum* infection [62]. The long term infection of *P. falciparum* in the erythrocytic cycle, which can exceed 3 years, requires significant diversity of extracellularly exposed parasite proteins [63]. These proteins are targeted by the human immune system and so are under significant selective pressure by the human immune system for diversification. The proteins during the erythrocytic cycle that have the most time exposed extracellularly are found in large gene families with significant nucleotide diversity, and therefore significant amino acid diversity [64]. This may be a prominent reason for the constant recurrent infections – the high antigen diversity from the large and heavily duplicated gene families of under strong selective pressure for diversity from evasion of the human immune system [64]. The life cycle of the malaria parasite involved numerous points of genomic duplication and recombination, each of which provides the necessary conditions establishment of duplication within the haploid genome.

1.4 Duplications and Deletions in *Plasmodium falciparum*

While many chemotherapeutic solutions have been developed over the years, developing drug resistance has materialized for current mainstream therapies, many via duplication of particular alleles [65–67]. Genetic duplications have been shown to be important in the arms race against human immune system and chemotherapeutic agents. Large gene families undergo diversifying selection for antigenic evasion [68]. There have been numerous studies on gene duplications for increased drug resistance [50, 69, 70]. Large deletions of entire arms of chromosomes have been identified as adaptation to *in vitro* culturing of the parasites [71]. However, the extent of duplication fixation and variation between strains has not been systematically investigated.

Large gene families within the *P. falciparum* genome are borne of ancient duplication events [64]. The consecutive duplications of an ancestral gene were generated out of necessity for diversification or novelty of function. One of the major examples of this is the *var* gene family [72]. Proteins coded by these genes are inserted in the erythrocyte membrane and are responsible for the subsequent binding to the endothelial lining, thereby sequestering the infected erythrocyte in post-capillary venules [73]. As deformability of the erythrocyte decreases during maturation of the *P. falciparum* parasite during infection of the erythrocyte, the sequestration of the parasite into post-capillary venules by the *var* genes prevents splenic filtration of infected erythrocytes and parasite destruction [74]. There are a number of similar gene families, such as *rifin* and *stevor* genes [75], which have functions in other aspects of the parasite life cycle. s*tevor* genes are implicated in the rosetting of erythrocytes and mediation of merozoite invasion [76]. While there are numerous studies analyzing the sequence similarity within these gene families [77, 78], little has been done to holistically examine the duplication profile of these genes. Most work has excluded surrounding genes and sequence. All these gene families exist within the same regions of the genome, primarily the subtelomeres [79]. As gene duplications rarely occur for singular genes, multiple genes are often duplicated [80]. The extent of duplication in *P. falciparum* has not been studied systematically. We perform a systematic analysis of copy number variation and segmental duplications in the fully sequenced genome of *P. falciparum*, identifying high sequence identity regions throughout the genome by a BLAST-like self-alignment. Gene ontology trends, GC% bias, and the potential means of sequence diversification between the large gene families are explored.

In this work, segmental duplications identified from a high quality reference genome denote duplications found within the reference genome. We performed a self-alignment based process of the *P. falciparum* 3D7 reference genome to identify these segmental duplications. From this, we identified regions of high identity and likely duplications. While many of these duplications may be fixed in the overall *P. falciparum* population, many may not be fixed. Duplications and deletions are a continual process, so there may be significant variation of gene duplication and deletion between strains of *P. falciparum* [70, 80]. These copy number variations are typically short term gene duplications as adaptations against selective pressure. The most prominent examples have been in copy number variation linked to drug resistance. PFE1150w, PFL1155w, and PFD0830w are all gene whose amplification has been linked to drug resistance [50, 69, 70]. Multiple classes of drugs are affected by these gene amplifications including anti-folates and 4-amino-quinolones. In addition, there are known copy number variations that are side effects of adaptation to *in vitro* cell culture of parasites, having roles in growth fitness [52, 71].

Many of the copy number variations identified were through select quantitative PCR experiments on likely gene targets [69, 81, 82]. There have been attempts to provide a genome wide systematic test for copy number variations – via array comparative genomic hybridization (array CGH) and high throughput sequencing [83–86]. However, the difficulty of working with a highly AT biased genome composition with significant tracts of simple sequence has made surveying the extent of copy number variation throughout the natural *P. falciparum* population a difficult endeavor. Most genome wide analysis involves *in vitro* cell cultured lab strains, both with array CGH and next gen sequencing. Despite numerous methods to leverage high throughput sequencing for copy number variation detection in *P. falciparum* [84–86], there have been no published experiments to identify copy number variation from a large cohort of field isolates of *P. falciparum*. This is partly due to the significant AT bias of the genome, which mitigates the ability to specifically identify copy number variation from high throughput sequencing. To identify copy number variation in natural populations of *P. falciparum*, we developed a custom computational methodology to leverage high throughput sequencing data to sensitively and specifically identify copy number variations in the genome. We tested this program on a publicly available sample of unique *P. falciparum* strains from Senegal to test the validity of our method [87]. In addition, we investigate the genome wide analysis of copy number variation from over 600 publicly available high throughput sequencing datasets from Africa and South East Asia: Burkina Faso, Cambodia, Gambia, Ghana, Guinea, Kenya, Mali, Senegal, and Thailand [87–91]. We determine that copy number variation in P. falciparum is rare within the population, and we also discovered a number of novel copy number variations - with potential drug resistance implications or growth adaptations.

CHAPTER II: Segmental Duplications in *P. falciparum*

This section is a version of a manuscript currently under review:

Derrick K. DeConti[1], Jeffrey A. Bailey[1,2]

[1] Program in Bioinformatics and Integrative Biology, UMass Medical School, Worcester, MA

[2] Department of Medicine, UMass Medical School, Worcester, MA

2.1 Introduction

Duplication of genomic DNA is a long recognized route for the creation of novel genes [92]. Duplications can be categorized based on various metrics such as length, location, gene content, or mechanism of transposition. A primary categorization is often performed by length, ranging from duplication of the entire chromosomal complement (whole genome duplication or polyploidy), through duplication of a single chromosome (aneuploidy), to duplication of small portions of chromosomes (segmental duplication), which can range from hundreds to millions of bases in size. Segmental duplications are also often described on the basis of their functional content (gene, partial gene, or exon) or in terms of the relative location (intrachromosomal versus intrachromosomal and tandem versus interspersed). Segmental duplications can occur through multiple mechanisms, of which aberrant non-allelic homologous recombination, or breakage and repair events through non-homologous end joining are the most common [93–95].

Initial studies of gene evolution through duplication have focused on the role of large-scale whole genome duplications that left signatures of massive gains and alterations in the total gene complement [92, 96, 97]. While further studies have shown strong evidence in both yeast and vertebrates for ancient rounds of whole genome duplication [96, 98, 99], the advent of the Human Genome Project and high-quality reference genomes [32] also have led to the renewed recognition of segmental duplications as the most prevalent, recent, and continuous source of gene evolution [100, 101]. Segmental duplications can have differing biological consequences. An immediate impact from any duplication event of a gene or other functional element is that the activity or dosage can be increased [6]. Segmental duplication followed by sequence divergence allows for the evolution of new and diverse functions [15]. This is of particular importance for countering the breadth of environmental and pathogen exposures. For example, human genes within segmental duplications are highly enriched for histocompatability antigens [102]. Alternatively, duplications in pathogens are often associated with virulence or diversification of extracellularly exposed proteins for immune evasion [29, 98, 103]. In pathogens, duplications also provide an avenue for rapid gene dosage response to drug pressure for which *P. falciparum* has multiple examples (e.g. *mdr1*, *gch*)[70, 104]. Thus, duplications provide plasticity in an organism allowing for the evolution of new or diversified responses to an everchanging environment.

Segmental duplications also create dynamic regions within the genome due to high sequence identity that can promote misalignment and non-allelic homologous recombination (NAHR), resulting in further duplication, deletion, or potentially more complex arrangements [105, 106]. Misalignment can equally lead to gene conversion, a phenomenon implicated in increasing the diversity within segmental duplications [107]. Thus, duplicated regions in the genome can rapidly evolve and diversify in response to evolutionary pressure.

Duplicated regions have practical consequences for genome analysis as they are more difficult to assemble. Even with high-quality reference genomes, improper alignment of reads because of duplicated high similarity regions have effects on the downstream analysis causing problems with correctly identifying SNPs, copy number variants, etc. In addition, graph based *de novo* genome assemblies with next gene sequencing can collapse at these regions of high similarity, preventing contiguous assembly of these regions of the genome. Knowing what regions contain segmental duplications and the scale of similarity can inform future experiments for adjusting downstream analysis to account for the segmental duplications.

In malaria, segmental duplication plays an important role within the causative pathogen, *Plasmodium*. Malaria is a disease with an estimated 219 million cases per year and an estimated 660,000 deaths [53]. Similar to many other pathogens – *Plasmodium spp.*, piroplasms, coccidians and *Cryptosporidium spp.* - duplicated gene families of Apicomplexa are predominantly found in the subtelomeric and telomeric regions of the chromosomes and have a predilection toward antigenic variation for immune evasion [25]. There have been extensive studies into the structure of the subtelomeric and telomeric regions of the *P. falciparum* genome resulting in a general structure of the subtelomeric regions [33, 108–110]. Of particular interest within these regions have been the large gene families of extracellularly exposed proteins involved in cytoadherence – *rifins, stevors,* and *vars* [75, 111, 112]. The *var* genes are singly expressed genes involved in the endothelial sequestration of the parasite during blood stage [113]. These genes have been studies extensively for their antigenic variation [78, 111]. They show significant copy number variation between strains and within monoclonal cultures – both *in vivo* and *in vitro* [111]. The molecular mechanism of variability and duplication of the *var* genes has been attributed to gene conversion [31, 114].

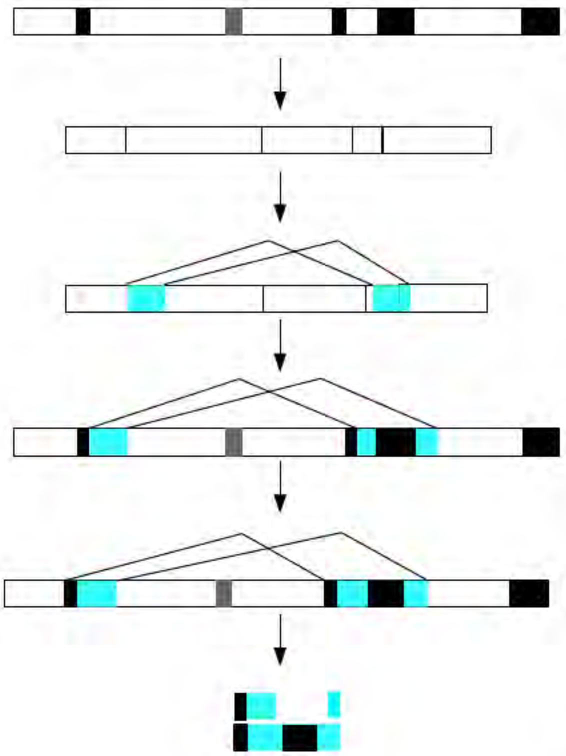
While many insights have been gained from such targeted analyses, a complete picture of the pattern and nature of segmental duplications within *P. falciparum* can be of great biological and practical use. Such an analysis will allow us to better understand the genomic constraints and evolutionary forces at play. Here we apply established methods for genome-wide detection of segmental duplications and characterize their salient properties and patterns.

2.2 Methods

Segmental duplication detection

We refined and applied two well-established methods for detecting segmental duplications, whole-genome alignment comparison (WGAC) and whole-genome shotgun sequence detection (WSSD), which have both been used extensively in larger eukaryotic genomes [31, 47, 84, 102]. Both methods make use of a high quality reference genome. For both WGAC and WSSD, we used *P. falciparum* 3D7 reference genome from PlasmoDB v9.3. Our WGAC analysis for segmental duplication detection employs openly available alignment software, sequence analysis software, and custom Perl programs modified from previous work as outlined in Figure 2.1 and detailed in Bailey et al. [31]. The overall methodology for any genome consists of (1) removing high copy repeats that represent simple tandem repeats and transposable elements, (2) local alignment of the entire "repeat-free" genome to itself to detect similar regions above a given identity and length, (3) reinsertion of the high copy repeats, and (4) refining (trimming/extending) the termini of alignments to locate the most accurate end point. If

Figure 2.1 Overview of segmental duplication identification by self-alignment. Tandem repeats and simple repeat sequence were identified with Tandem Repeat Finder and RepeatMasker. The identified repeat sequences were spliced from the genome, and LASTZ was used to identify high identity alignments with a custom substitution matrix from this modified genome. Repeat sequences were spliced back into the genome. With custom scripts, we heuristically extended alignment to incorporate spliced repeat sequences. Finally, a global alignment of the pairwise alignments was used to more accurately delineate alignments.



Identify tandem and simple repeats

Remove identified repeats leaving pututatively unique sequence

Identify high identity alignments with LASTZ local alignments

Insert previously masked elements

Refine alignment extension heuristically

Final global alignment

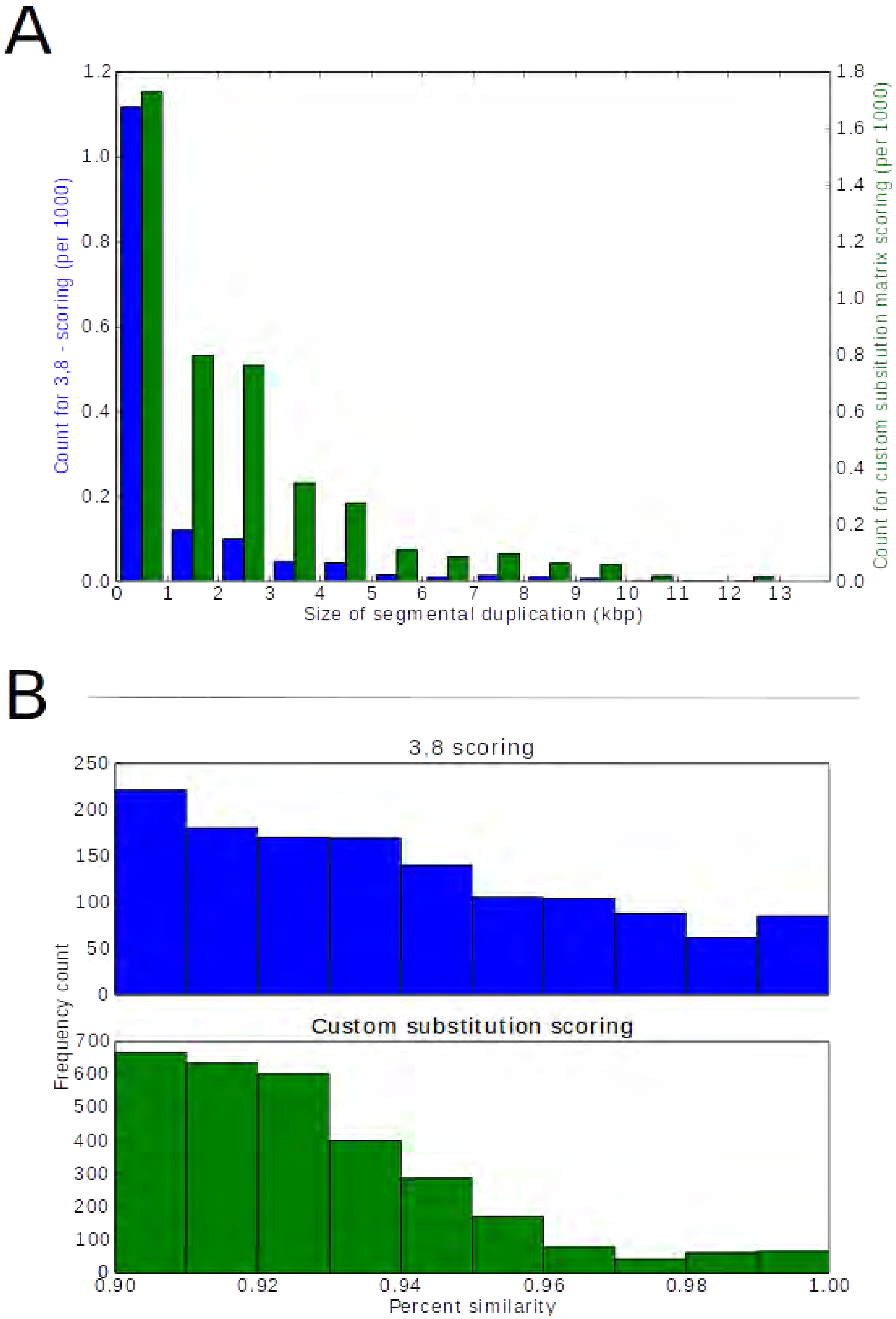
unrecognized transposable elements are not a contaminating issue, then the final alignments represent a database of recent segmental duplications (Figure 2.1). For this work, WGAC methods were specifically updated and modified to improve pipeline speed, robustness, as well as to address specific characteristics of the *P. falciparum* genome – smaller size, abundance of tandem repeats, and highly elevated AT content (GC % 18%) [33]. Specifically for this analysis, the *P. falciparum* genome (version PlasmoDB 9.3) was analyzed for tandem repeats and low-complexity regions using Tandem Repeat Finder (version 4.04) [115] and RepeatMasker version 3.2.9 [116], respectively. The lack of known transposable elements within the genome abrogated the need for their characterization with RepeatMasker. Large-tandem repeats (period 50-350 and copies ≥5 via Tandem Repeat Finder), and low-complexity sequence (100bp regions ≥87% AT or ≥89% GC with a 30bp stretch of 29 AT or GC nucleotides via RepeatMasker) were spliced out of the chromosomes and local alignments were generated within and between all chromosomes. Improvements to the previous methods included replacing slower NCBI BLAST with LASTZ [117] and using a custom scoring matrix to account for the GC bias during the alignment process (Table 2.1). This scoring matrix was calculated using the described methods [118] from a representative subset of initial alignments >250 bp and >88% identity generated using a 3,8 match/mismatch LASTZ scoring. The optimized matrix produced a greater median segmental duplication length (1,480 bp) compared to non-custom LASTZ scoring parameters (222 bp). The custom match parameters also detected more alignments and with higher average identities (Figure 2.2).

Coordinates of LASTZ alignments lacking repeats were translated back into

Table 2.1 LASTZ custom substitution matrix. Self-alignments from LASTZ with flat match/mis-match scoring between >80% identity and >250 bp were used to create the custom substitution matrix. Log-odds scores between the paralogs were calculated and averaged to generate the custom LASTZ substitution matrix.

		Match Base			
		A	**T**	**C**	**G**
	A	2	-9	-7	-13
Query	**T**	-9	13	-1	-5
Base	**C**	-7	-1	11	-8
	G	-12	-5	-10	4

Figure 2.2 LASTZ custom substitution scoring outperforms flat match/mis-match scoring. A) The custom substitution matrix produces significantly more pairwise alignments as alignment size grows as compared to a flat match/mis-match scoring system. B) The custom substitution matrix identifies a significantly higher number of low percent identity alignments than the flat match/mis-match scoring system, though both are comparable at identifying high identity regions.


A
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Count for 3,8 - scoring (per 1000)
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Count for custom subsitution matrix scoring (per 1000)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
Size of segmental duplication (kbp)
B
3,8 scoring
250
200
150
100
50
0
Frequency count
Custom substitution scoring
700
600
500
400
300
200
100
0
0.90
0.92
0.94
0.96
0.98
1.00
Percent similarity

normal genomic coordinates, effectively reinserting the previously removed tandem and simple sequence repeats. The alignment end points were refined to more accurately determine the extent of segmental duplication which might terminate within an adjacent tandem or simple sequence repeat. This was accomplished heuristically through iterative extension of the global alignment up to 2 kbp and redetermination of the alignment end point. This was iterated until end point convergence. After refinement, final optimal global alignments were kept if ≥250bp and ≥90% bp identity. Lastly, to yield our final analysis set, pairwise alignments with juxtaposed and properly oriented and ordered copies were merged across up to 2 kb gaps, in order to more completely capture likely segmental duplication events even if subsequent large insertions or deletions have occurred over time within the individual copies.

WSSD is a method exploiting whole-genome shotgun sequence reads as a random sample of the genome such that increasing density of reads in a region relative to the genome average directly correlates with increased number of copies. Illumina whole-genome shotgun sequence for the reference genome strain 3D7 [119] were mapped with bowtie2 to the PlasmoDB 9.3 3D7 reference genome assembly [33, 120]. Sequence files for PCR-free Illumina GAII 3D7 libraries can be accessed at SRA archive SRP056541. In separate analyses, alignments were performed both with single best and multiple (up to 100) best placements within the genome. Samtools was used to determine read depth per base position in the genome [121]. To remove GC sequencing biases, we applied a correction factor for each 100bp window baed on a LOESS correction of read depth against GC% [122]. Regions of potential duplication were defined by tiling windows of

1,000bp with a median read depth ≥2 and tandem repeat content <70%. Windows were merged to generate contiguous regions of elevated read depth indicative of more than one copy within the genome. These regions were compared to detect probable assembly errors, either the false positive or false negative identification of segmental duplications from WGAC.

Descriptive statistics of pairwise alignments

All identified pairwise alignments ≥90% identity and ≥250 bp were analyzed either at the level of pairwise alignment or based on non-redundant coverage of the genome. We determined a measure for the most recent duplication event by the highest percent identity pairwise alignment present in each base in the genome. Due to the high nucleotide diversity between some genes with segmental duplications, particularly the *var* genes, a requirement for the complete intersection of a gene with an alignment did not accurately capture the extent of genic content within segmental duplications. Many genes had little to no self-alignment with any other regions of the genome, yet the flanking regions of these genes show evidence of duplication. Therefore genes with any portion of their coding regions or ±50 bp of the flanking regions of the gene having an aligned pair were cataloged. Gene information was obtained from PlasmoDB gff files – PlasmoDB 9.3. Genes included in the segmental duplications were analyzed in GOSTAT to determine under and overrepresented gene ontology (GO) terms [123]. Visualization of segmental duplications was performed using Parasight [124].

Analysis of segmental duplications around *var* and *rifin* genes

Pairwise alignments often did not include the interior of the *var* genes, which is due to the gene-families high nucleotide diversity [64]. We analyzed the pairwise alignments on either side of the *var* gene (up to 50 bp away from the gene itself) to extend the pairwise alignment. The pairwise alignments flanking the *var* gene were identified, and examined for whether the pairwise alignments paired with the same gene elsewhere in the genome. This would identify the extent of whole gene duplication and gene conversion within the *vars*. The same methods were also applied to the *rifin* gene contingent.

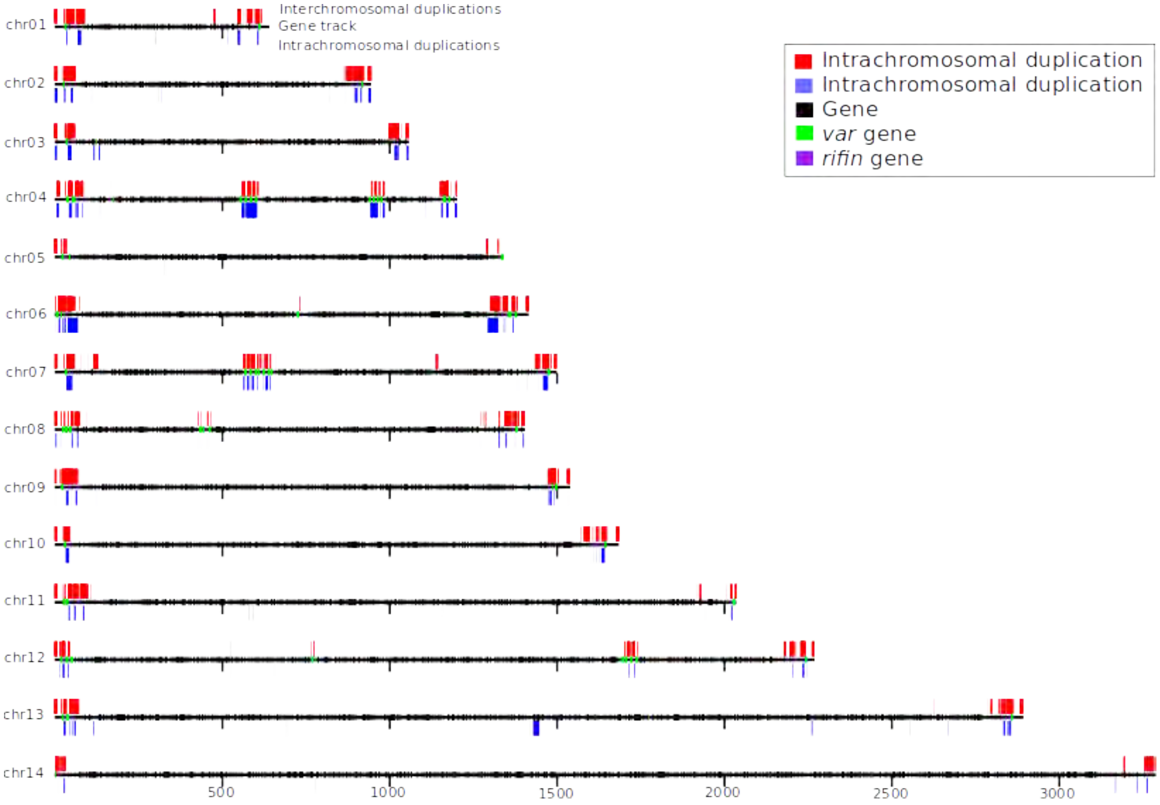
2.3 Results

Detection of pairwise alignments representing segmental duplications

We further optimized the whole genome alignment comparison (WGAC) to detect segmental duplications within the reference *P. falciparum* 3D7 genome. WGAC is a well-established method which initially seeds on putatively unique sequence by removing high-copy repeats from the sequence leaving putatively unique sequence. Repeats are then reinserted and alignment edges are then refined to capture the precise extent of any pairwise alignments. Global alignments are then calculated to provide the most accurate measures of pairwise identity. A total of 2,579 pairwise alignments (median length of 2,005 bp) were found at an alignment threshold of ≥90% identity and ≥250 bp. These alignments covered 5.93% (1,380,021bp / 23,292,622bp) of the genome (Figure 2.3).

The WSSD analysis is sensitive for highly-similar segmental duplications [102,

Figure 2.3 Overview of segmental duplications in the genome. We identified duplicated sequence as alignments ≥250 bp in size and ≥90% sequence identity and aligned them to the genome. The map shows the overlay of segmental duplications on the genome, split into interchromosomal and intrachromosomal segmental duplications). Interchromosomal (red) duplications are depicted above the chromosome. Intrachromosomal (blue) duplications are depicted below the chromosome. Gene content is demarcated in black along the chromosome, with exceptions for var genes (green) and rifin genes (purple). Tick marks are at every 500 kb interval.


chr01
Interchromosomal duplications
Gene track
Intrachromosomal duplications
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12
chr13
chr14
500
1000
1500
2000
2500
3000
🟥 Intrachromosomal duplication
🟦 Intrachromosomal duplication
■ Gene
🟩 *var* gene
🟪 *rifin* gene

125] and was highly concordant with the WGAC alignments ≥95% identity. There was one exception greater than 250 bp (Figure 2.4) where WSSD detected a known duplication of gch1 (PFL1155w) related to anti-folate resistance [70]. Overall, WSSD supports that the 3D7 genome is a high quality and accurate reference genome assembly in respect to duplication.

Genomic distribution of segmental duplications

Segmental duplications localized predominantly to the telomeric and subtelomeric regions of almost every chromosome (Figure 2.4). The vast majority of duplication outside of the subtelomeric regions were associated with clusters of extracellularly exposed genes under human immune pressure. Overall, only 129 kb (0.6% of the genome) of the duplicated sequence were not associated with either the subtelomeres or extracellularly located proteins. Thus, given this highly skewed distribution, the amount of segmental duplications between chromosomes appeared to correlate with the size of the subtelomeric regions and cytoadherence complement rather than the overall chromosomal length. This can be seen in that the amount of segmental duplication content per chromosome remained relatively similar across chromosomes and was not correlated with chromosome size (Figure 2.5). However, a chi-squared test showed a statistically significant difference in the goodness of fit for duplicated space per chromosome (p-value = 8.79×10^{-70}) indicating some variability. Chromosome 5 stands out as it was over 2 standard deviations ($\sigma = 2.11$) from the mean duplicated space per

Figure 2.4 Comparison of WSSD to WGAC for segmental duplication identification. A map of all high read depth regions in the genome >1,000 bp and <70% tandem repeat content as determined by WSSD (above chromosome in red) is compared to our WGAC method at >95% bp similarity (below chromosome in blue).

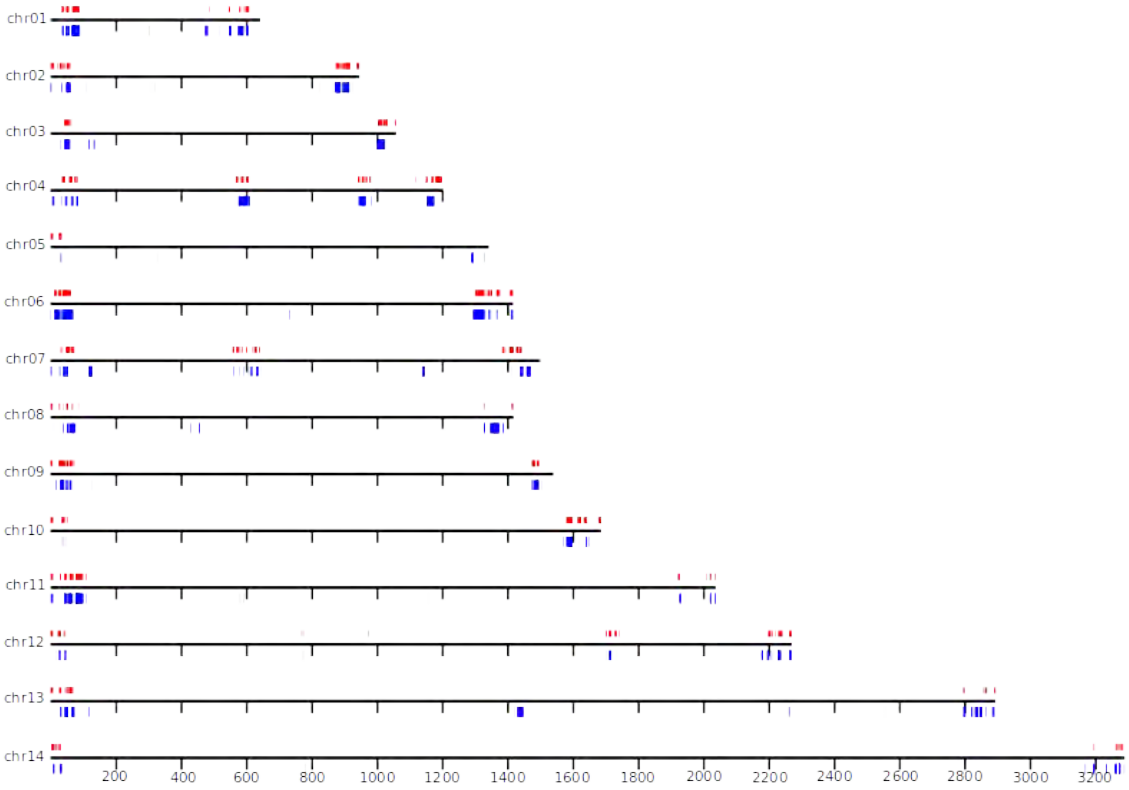
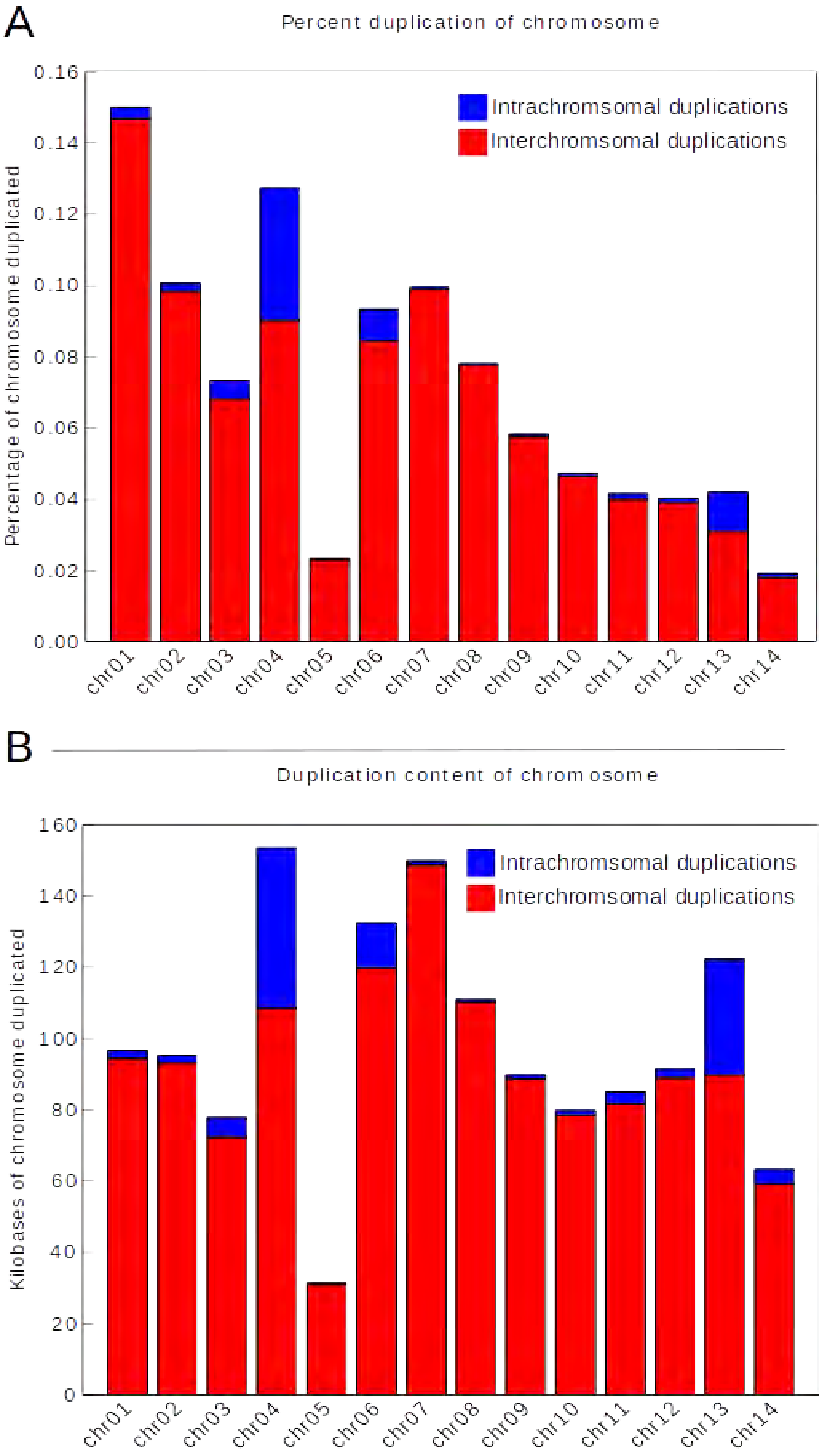

chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12
chr13
chr14
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200

Figure 2.5 Duplication content between chromosomes. A) The cumulative lengths of the highest pairwise identity at all loci of duplication - i.e. reducing to the unique highest identity pairwise alignments – also indicates small variation between the various chromosomes. B) Duplicated fraction of the chromosome. Chromosome 4 and 13 stand out for their relative enrichment of intrachromosomal duplication – caused by the high identity tandemly duplicated clusters of *var* genes internal to the chromosome. The duplicated fraction of the chromosome decreases with chromosome size, indicating the duplicated space is unlinked with chromosome size.


A
Percent duplication of chromosome
Percentage of chromosome duplicated
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Intrachromsomal duplications
Interchromsomal duplications
chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12
chr13
chr14
B
Duplication content of chromosome
Kilobases of chromosome duplicated
160
140
120
100
80
60
40
20
0
Intrachromsomal duplications
Interchromsomal duplications
chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12
chr13
chr14

chromosome. Most other chromosomes (9 chromosomes) were within 1 standard deviation of the mean, with all other chromosomes under 2 standard deviations.

We examined the relative rates of intra versus interchromosomal duplication to determine any biases for intrachromosomal duplications versus interchromosomal duplications, as tandem duplication of genomic sequence is common. We performed a binomial test assuming the probability for an intrachromosomal duplication to be dependent on the number of chromosomes (expected probability = 7.14%), as the duplicated space was approximately the same across chromosomes. The binomial test for all identified segmental duplications showed there was no statistically significant difference between intrachromosomal and interchromosomal duplications (observed probability = 7.10%, p-value = 0.9694). However, a binomial test for high identity segmental duplications (≥98%) showed a statistically significant difference with an observed probability of 16.4% (p-value = 3.73×10^{-4}), indicating that the majority of all segmental duplications had no bias toward intra- or interchromosomal duplications but the high identity duplications were enriched for tandem duplications.

Gene content of segmental duplications

There were 466 genes and pseudogenes out of 5,772 in the genome that intersected regions of identified segmental duplications. 81 of those were *var* genes or pseudogenes and 164 were *rifins*, comprising the majority of genes in their respective families – 101 and 185 respectively. While often overlooked, *rifins* actually represent the most abundant gene family [112] and contain many recent highly-identical duplications.

In addition to the the vars and rifins, another gene family has been associated with the subtelomeric regions – *stevors* [75]. Previous reports with unfixed copy number variations indicate that nucleotide diversity was elevated in duplications [83]. Our analysis verifies this claim as the enrichment of these antigenic genes in segmental duplications resulted in increased nucleotide diversity within segmental duplications. From a collated list of nucleotide diversity of genes, we determined that genes located within segmental duplications have elevated nucleotide diversity (two sided t-test, t-value = 2.271, p = 0.0237) [126].

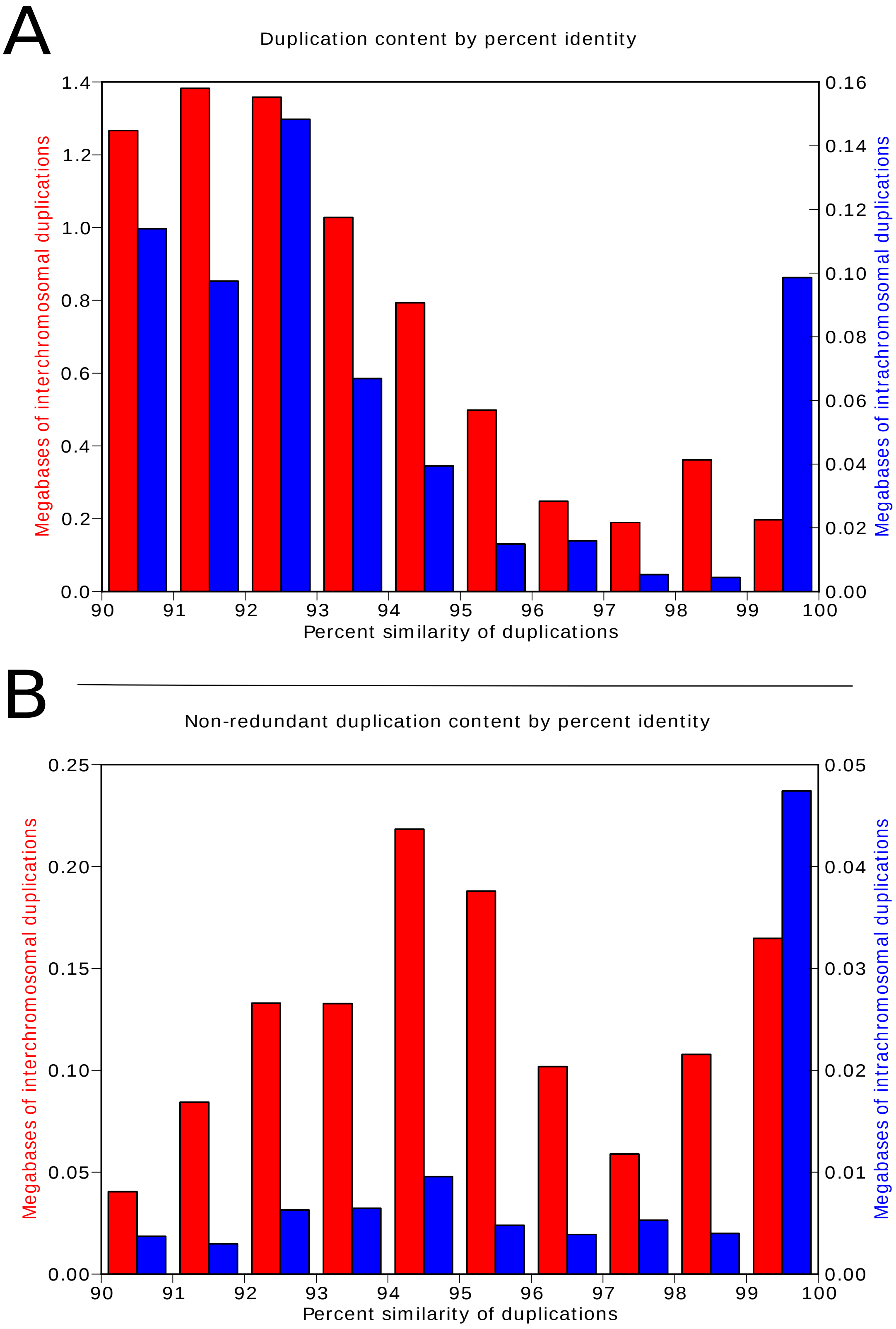
Outside of the sub-telomeric regions and internal *var* clusters, we identified 17 genic intrachromosomal pairwise alignments and 10 genic interchromosomal pairwise alignments (only one pair required to be outisde sub-teolmeric regions). Of these 27, 25 had pairwise alignments that completely spanned the gene(s). The largest group of genes to be duplicated were 10 rRNAs. This is in accordance with the expansion of rRNAs and subsequent divergence for life cycle specific expression and function [127]. Other genes including reticulocyte binding proteins, falcipain 2, elongation factor 1-α, var trafficking, SERA, CLAG, and ubiquitin. PFL0585w, did not have complete coverage of pairwise alignment, but it showed a significant number of pairwise alignments all along its genic content. As a polyubiquitin, it was successive tandem duplications of the ubuquitin domain in Pf13_0346.

Consistent with their localization to the subtelomeres, segmental duplications had a lower mean GC% of 18.74% compared to the genomic mean of 19.4% (one sample two-sided t-test against genome mean, p =7.97e-10). After a duplication event, mutations will have caused paralogs to diverge [13]. If mutations rates were regular, the percent identity could be used as a surrogate of relative age and could provide insight into the tempo of segmental duplication over time. When alignments were categorized by percent identity, there was an abundance of pairwise alignments at lower identity (Figure 2.6). However, when only the highest identity pairwise per loci were examined, which provides a better correlate of events when accounting for repeated duplications of the same loci, the quantity of duplicated sequence appeared more uniform across levels of percent identity. This suggested that the drivers and processes of duplication has been relatively continuous rather than punctuated process.

Extracellularly exposed genes overrepresented in segmental duplications

We performed ontological analysis of the genes intersecting with segmental duplications. Of the overrepresented GO terms, most were related to extraorganismal interaction - i.e. extra-organismal space, interaction between organisms, extracellular, etc - while GO terms related to normal cell function are underrepresented - i.e. metabolic activity, cell signaling. Excluding vars, rifins, and stevors genes, segmentally duplicated genes had no over or under representation of GO terms. 46 of these 100 genes were hypothetical proteins with no known function, however 13 were rRNA associated. Of those 100 genes, 70 of them overlapped high identity segmental duplications, indicating a reduction in relaxed selective pressure on these genes as compared to genes in the subtelomeric regions.

Figure 2.6 Duplication content by percent similarity. The cumulative lengths of pairwise alignments per 1% ranges in identity were binned for A) all pairwise alignments and B) the highest pairwise identity at all loci of duplication – i.e. the reducing to unique highest identity pairwise alignments. The shape of the graphs indicate that the rate of duplication is has likely been relatively consistent over time based on the fact that many pairwise alignments represent repeated duplication of the same region.

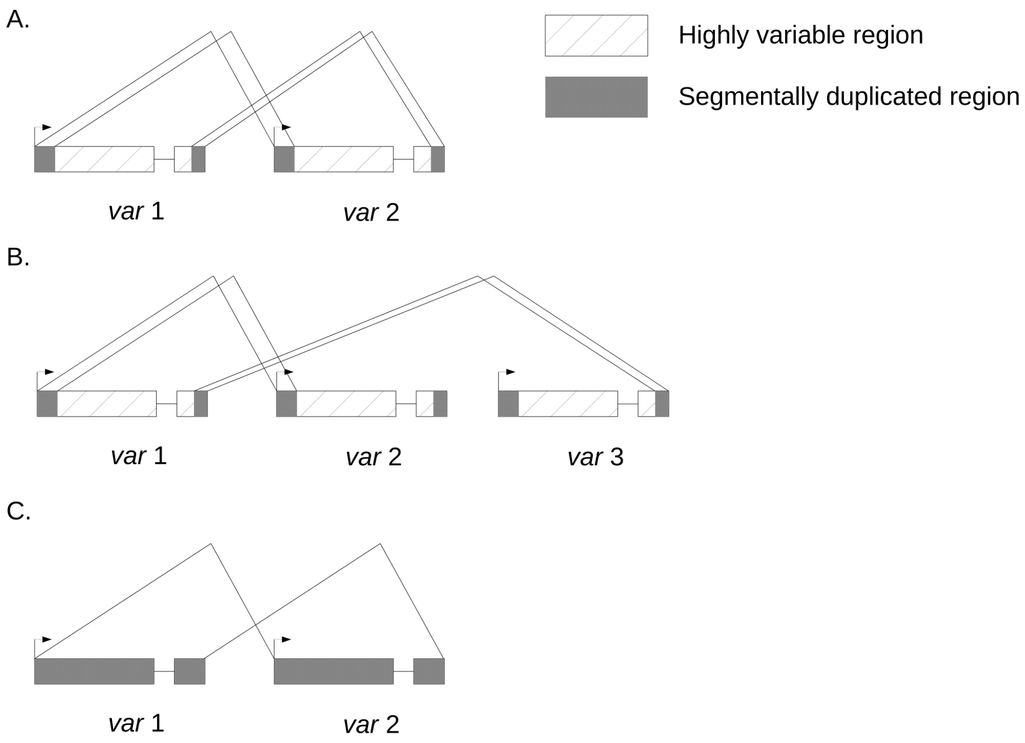

A
Duplication content by percent identity
Megabases of interchromosomal duplications
Megabases of intrachromosomal duplications
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
90
91
92
93
94
95
96
97
98
99
100
Percent similarity of duplications
B
Non-redundant duplication content by percent identity
Megabases of interchromosomal duplications
Megabases of intrachromosomal duplications
0.25
0.20
0.15
0.10
0.05
0.00
0.05
0.04
0.03
0.02
0.01
0.00
90
91
92
93
94
95
96
97
98
99
100
Percent similarity of duplications

The nature of *var* and *rifin* gene segmental duplications

The majority of *var* genes lacked pairwise alignments that spanned the entire gene. This appeared to be a consequence of high nucleotide diversity within exon 1, which could be a consequence of rapid divergence through positive selection or a consequence of an elevated rate of gene conversion. To examine these possibilities, we focused on the highest identity alignments across each *var* gene (Figure 2.7). This revealed that most *var* genes have divergent pairwise alignments flanking the genes – i.e. each side of the *var* gene best identifies with a different *var* gene. Of the 81 var genes, only 29 (35%) genes had flanking pairwise alignments match to the same *var* gene (Figure 2.7a), and only 5 (6%) genes had pairwise alignments that spanned the entire gene (Figure 2.7c). The genes lacking matching alignments consisted of 39 (48%) genes (Figure 2.7b) where flanking alignments mapped to different var genes in the genome, and 12 genes where the other end did not have evidence of a matching gene.

The other large gene family, *rifins*, were analyzed similarly. While rifins show similar evidence of a high rate of duplication, with 90% of *rifin* genes having pairwise alignments intersecting and 100% of *var* genes. However, the *rifin* genes were more likely to have a spanning pairwise alignment, 33% of *rifin* genes had a spanning pairwise alignment as opposed to 6% of *var* genes. When looking at the highest identity alignment for the flanking of *rifin* genes, show similar levels of cross mapping of alignments with 106 (58%) of the *rifin* genes had the highest pairwise alignment on either end of the *rifin* pair to different *rifin* genes, while only 55 (30%) genes had alignments on either side that matched to same *rifin* gene. This pattern suggests a strong role for non-duplicative

Figure 2.7 Model of possible duplication patterns for virulence genes. Virulence related genes, *vars* and *rifins*, display three distinct patterns of duplication across their respective gene families. A) A spanning duplication of the virulence gene 1, but time and high nucleotide variation of the genes themselves has prevented an spanning alignment of the duplication. B) Gene conversion within virulence gene 1 has resulted in a hybrid of both virulence gene 2 and virulence gene 3 ancestry. This results in the highest identity location of pairwise alignments on either side of virulence gene 1 to point to different virulence genes. C) A complete high identity pairwise alignment of entire an entire virulence gene.


A.
Highly variable region
Segmentally duplicated region
var 1
var 2
B.
var 1
var 2
var 3
C.
var 1
var 2

processes such as conversion or telomeric exchanges hybridizing in the genesis of these duplications relative to strict duplication followed by rapid divergence.

2.4 Discussion

We analyzed the *P. falciparum* reference 3D7 strain P. falciparum genome for segmental duplications via whole genome alignment comparison. We validated the sensitivity of the whole genome alignment comparison with a whole genome shotgun sequence detection approach. The analysis with whole genome alignment comparison highlighted that the vast majority of segmentally duplicated sequence was located within the sub-telomeric regions of the genome or within *var* gene clusters. Consequently, genes within the segmental duplications were overrepresented for extracellularly exposed genes, particularly the *var* and *rifin* genes, which appeared based on our analysis, to be evolving through duplicative mechanisms combined with partial exchange methods, such as gene conversion, creating genes of chimeric origin. Overall, the segmental duplications have been sequestered mainly in the sub-telomeric space of the genome leaving a core genome that is relatively staid and lacking in significant standing duplication.

Our systematic analysis of segmental duplications in the *P. falciparum* genome confirmed that the subtelomeric regions have been highly duplicated, and the region is the major source of segmental duplications in the genome. The vast majority of genes in the segmental duplications were related to antigen presenting proteins. Except for tandem duplication in non-subtelomeric *var* clusters, segmental duplications did not show any significant preference for intrachromosomal duplication. This seemed to indicate a mostly random duplication of genes through relatively unbiased subtelomeric interaction.

Analysis of pairwise alignments can be challenging to interpret because they do not correlate one to one, as the pairwise representation increase rapidly with repeated duplication of a given sequence. Direct examination of the pairwise alignments shows a prominent abundance below 95% identity. This abundance markedly decreases when only the most recent segmental duplication in a region are examined. This suggests that there *vars* and *rifins* are likely under constant amplification with genes being gained and lost. These genes appear to be continually evolving based on the broad distribution of pairwise identities and that antigenic diversification has likely been a more continuous, rather than punctuated, process *vis-a-vis* host immune evasion. Additionally, there appears to be significant amounts of conversion or hybridization taking place, in addition to diversifying mutation.

Conversely, segmental duplications are nearly absent from the rest of the genome. There is little trace of duplication of genes outside the subtelomeric regions or genes related to antigenic diversity and immune evasion. Many of these genes are of unknown function and may represent unrecognized genes encoding extracelluarly exposed proteins. Duplication of biochemical and cellular process genes appears minimal nor does evidence of non-functional duplication. Given active duplication elsewhere and evidence that drug resistance can form through duplication, this suggests that duplications are poorly tolerated within the genome and rapidly lost compared to organisms where genomic duplications can be maintained despite loss of function. The few duplications that do occur unrelated to cytoadherence and extracellular exposure are limited to intrachromosomal tandem duplications, whether singly genic or non-genic tandem repeat elements. The rRNA related genes benefit from duplication because of increased gene dosage and life cycle specific expression and function. The SERA family of genes, where the individual genes have evolved specialized function during various life stages, have expanded for diversification of function. The identification of the duplication of these genes can aid in understanding the biochemical and functional roles these proteins have and the selective pressures placed on them. However, the staid nature of the genome suggests that a duplication is likely of functional consequence – particularly for a recent duplication that is still unfixed and copy number variant (CNV) within the population.

CHAPTER III: Copy Number Variation Detection from High Throughput Sequencing in a Natural Population of *P. falciparum*

This section is a version of a manuscript currently being drafted:

Derrick K. DeConti[1], Kate M. Fernandez[2], Sarah K. Volkman[2], Dyann F. Wirth[2], Daniel E. Neafsey[3], and Jeffrey A. Bailey[1,4]

[1] Program in Bioinformatics and Integrative Biology, UMass Medical School, Worcester, MA

[2] Harvard School of Public Health, Department of Immunology and Infectious Disease, Cambridge, MA

[3] The Broad Institute, Cambridge, MA

[4] Department of Medicine, UMass Medical School, Worcester, MA

3.1 Introduction

Copy number variation represents duplications and deletions ranging from hundreds of bases to megabases in size that are unfixed in the population [43]. Copy number variation can be a means of rapid gene dosage adjustment for organisms and novel gene genesis [4, 5]. In the human genome, copy number variants (CNVs) represent a significant source of genomic variation resulting in wide range effects underlying normal phenotypes as well as disease [128]. While the previous segmental duplication analysis often identifies ancient and fixed duplications, copy number variation in the population of *Plasmodium falciparum* may represent the latest avenues of adaptation to recent or current selective pressures. Aspects of parasite physiology impacted by copy number variation can include parasite growth rate modifications, increased antigenic diversity, and increased metabolic rates [25]. Significant interest in the field of copy number variation in *P. falciparum* has been engendered by the discovery of multiple gene amplifications directly related to drug resistance, particularly duplications of *mdr1* and *gch* [50, 70]. In addition to drug resistance, copy number variants offer avenues for changes to parasite fitness by adapting gene dosage to alter growth rate, differentiation rate, nutrient metabolism, etc. These changes can all be in response to geographic selective pressures, whether by human genetic adaptation, government regulation and protocol, changes in drug regime, or seasonal weather changes.

Copy number variation has yet to be fully explored in *P. falciparum* despite its known importance in drug resistance and antigen diversity [129]. Since then, multiple studies have investigated the extent of copy number variation within the parasite genome. The first genome wide studies utilized array comparative genomic hybridization [83]. The advent of next generations sequencing has rapidly decreased the cost, increased the amount of data, and improved the ease of preparation for genome wide sequencing of *P. falciparum* [88, 130]. There are now numerous methods for the analysis of copy number variation leveraging whole genome sequencing [85, 86]. All employ combinations of read depth, discordant read or read pair mapping to a reference genome or *de novo* assembly. These methods are not always easily applied to *P. falciparum* given its low GC content which can lead to extremely biased sequencing and read mapping [119]. Given these specific challenges, there has been additional effort to develop and tune algorithms to detect copy number specifically in *P. falciparum* [84–86]. However, to date these methods have only been applied to laboratory strains and the true extent of copy number variation is unclear due to the fact that (1) there is concern that many copy number variants may be driven by culture adaptation [131] and (2) the copy number detection often has had low specificity and so a relatively high false discovery rate. Given this, we have implemented a computational methodology written in Python to utilize read depth to identify copy number variation combined with confirmatory discordant read pairs. In combination, this provides reasonable sensitivity and specificity. We then apply it to publicly available Illumina libraries of natural isolates consisting of 33 unique isolates of *P. falciparum* from three sites in Senegal over the time period of 2004-2009 [87].

3.2 Methods

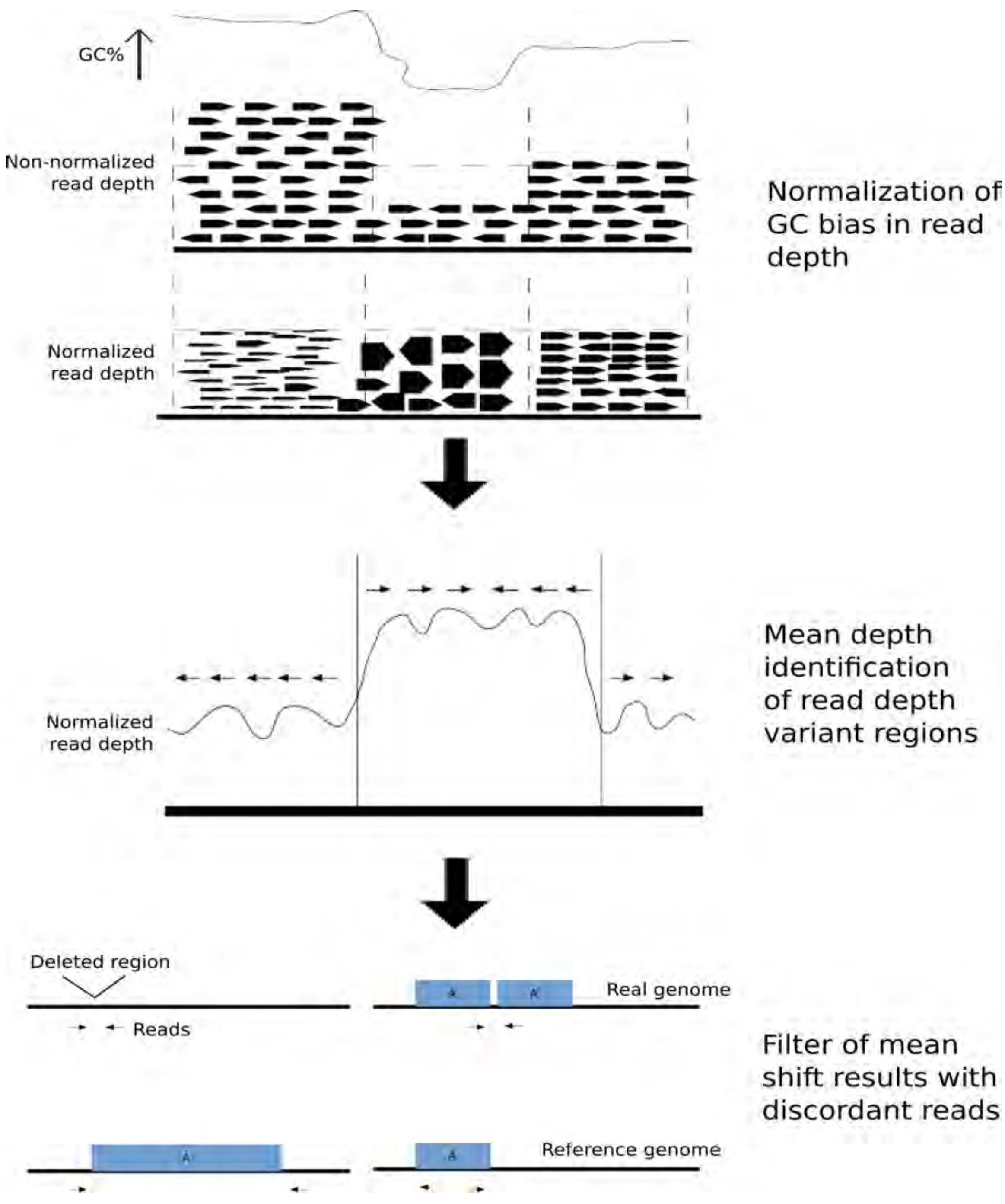
Data collection

DNA from in vitro cultures of 3D7, FCB and 106/1 were sequenced by non-amplified paired end Illumina GAII sequencing utilizing PCR free library construction (SRA: PRJNA279397) [119]. We also used 33 publicly available paired end sequencing libraries from Senegal demographics study by H. Chang et al. [87] Samples represented isolates from three different villages - Pikine, Velingara and Thiès - between the years 2004-2009 with either GAII or HiSEQ2000 sequencing technology (sequence data can be accessed at SRA: SRP018047).

Overview of CNV detection

We developed a custom suite of Python and Java programs to identify copy number variants (Figure 3.1) and optimized it for use within the *P. falciparum* genome with the goal of having sensitivity and specificity down to the gene and exon level. The methodology combined two well-described metrics. First, initial regions were identified based on either greater than or less than expected read depth after correction for GC-biases in the sequencing. Second, putative regions of high or low read depth were confirmed based on the presence of discordant read pairs. Combining these two methods in succession provided the improved specificity. Given the highly divergent nature of the subtelomeric regions and particularly the var genes, where read placement on the reference genome often fails due to high levels of divergence, we excluded these regions from our analysis. Accurate delineation of copy number variants within these regions

Figure 3.1 Overview of CNV detection pipeline. GC-biased read depth is normalized on a per read basis. A mean shift algorithm is used to identify signals of variant read depth by local mean minima and maxima. Mean shift identified variants in read depth are verified by discordant reads near mean shift identified breakpoints (spanning reads for deletions; inverted reads for duplications).


GC%
Non-normalized read depth
Normalized read depth
Normalization of GC bias in read depth
Normalized read depth
Mean depth identification of read depth variant regions
Deleted region
Reads
A
A'
Real genome
A
A
Reference genome
Filter of mean shift results with discordant reads

will likely require better assembly techniques as well as longer reads.

At the cores of our read depth analysis, we employed the mean shift algorithm from CNVnator by Abyzov *et al.* [84] which we reimplemented and modified. As input for sequencing depth, we also developed a per read pair read correction method to correct GC biases in the next-generation sequencing at the level of the read pair. Read-based rather than window-based correction allowed us to more accurately compensate for the sharp transitions in GC content between coding and non-coding sequence in *P. falciparum*. To normalize the depth by a per read basis, we assigned a normalized depth value to individual reads based on that read product GC%, by a ratio of observed versus expected reads at each 2% interval of GC%. Read pairs within an over-represented GC% bin were down-weighted, while under-represented pairs received proportionally more weight. Modification involved reversing the order of bandwidth search - large (512,000 bp) to small (400 bp) windows. An iterative t-test, with a p-value of 0.05, was used for each test of bandwidth for potential read depth variants. After identification of variant regions, we used a Bonferroni correction for all iterative tests for each bandwidth tests during the mean shift segmentation. Regions passing cutoffs for copy number variation were than examined for supporting discordant reads indicative of true copy number variants. A minimum of two discordant read pairs that correlated with the defined copy number variant region were required – inverted reads for duplications (minimum distance between pairs of 300 bp) and paired reads with insert lengths greater than 1000 bp for deletions. In order to detect copy number variants that might transition into the subtelomeric regions, we included detected copy number variants proximal to the subtelomeric regions – as defined by non-syntenic regions and var genes in the middle of the chromosome – that passed a stricter threshold for deletions, i.e. a mean read depth value of less than 0.05.

GC normalization of read depth

We aligned paired end read libraries with Bowtie 2 using default alignment settings [120]. Sequence alignment/map (SAM) files from Bowtie 2 were converted to a binary sequence alignment/map (BAM) file and sorted by coordinate with samtools [121]. To determine the expected values for read depth on GC bias, we determined the GC% for all possible read products (i.e. the sequence fragment) from 50 bp to 1,000 bp and their frequency within the genome. We notated the number of read products (by product size and then GC%), simulating an expected rate of read product GC% and size.

We input coordinate sorted BAM files into a custom python program to determine the normalization factor to apply to each read. Aligned reads were assigned to deciles of product size. Expected read product sizes and their GC% counts were similarly assigned to the same deciles. Within deciles of read product size, we counted reads by GC% for every 2% GC window for both expected and observed read products. For every 2% GC window at each product size decile, we normalized the read product counts by the total number of read products at that product size decile. This was performed on both observed and expected read counts. We calculated the correction factor for reads at each 2% GC window within its product size decile as the normalized expected count divided by the normalized observed count.

Based on this, we assigned each read the appropriate correction factor for its read product GC% and read product size. Given these corrected reads, we summed the correction factors of all reads aligned to a given loci for all base pair positions in the genome (the pileup file).

Mean shift analysis of normalized read depth

We ran GC bias corrected read depth pileup file through a modified mean shift algorithm based. By chromosome, we iterated over the entire genome from bandwidths 512,000 bp to 400 bp, halving bandwidth per iteration. We then iterate over the new array of mean read depth, given current bandwidth, until finding a segment that passes minimum requirement change in the read depth of the mean relative to the overall average (≤0.4 read depth or ≥1.4 read depth). Regions with local significant by t-test (p-value = 0.05) given the variance of the read depth across the chromosome are kept. After determining mean shifts across the genome, p-values were subjected to Bonferroni correction given all possible bandwidths across the chromosome. Segments meeting all statistical criteria representing putative copy number variant regions were outputted to a browser extensible data (BED) file.

Discordant read pair intersection analysis and breakpoint trimming

We searched through the coordinate sorted BAM file for discordant reads. We denoted read pairs with a product size greater than 400 bp into a bed file to compare mean shift identified deletions (Figure 3.1). We also denoted everted read pairs of product size greater than 300 bp to compare mean shift consistent identified duplications (Figure

3.1). We then intersected the mean shift identified copy number variation calls with the discordant reads. For valid intersection, we required all read placements to be within a specified distance from either side of the estimated boundaries – 20% of total copy number variation size for all variants less than 2,000 bp and 100 bp for all variants greater than or equal to 2,000 bp. We then required every mean shift identified copy number variation calls to have a minimum of two discordant reads pairs – everted reads for duplications and spanning reads for deletions – to validly intersect.

To refine the breakpoints for copy number variant regions, we used the innermost reads from discordant pairs as inference of the maximal true breakpoint. We adjusted the estimated copy number variation breakpoint to be within 50 bp of the minimum and maximum positions of all intersected discordant reads for that copy number variation, which essentially defines the maximal size of the breakpoint of the copy number variant region.

3.3 Results

Copy number detection within well-characterized laboratory strains

Initial mean shift identification of 3D7, at the p-value of 0.05 (subsequently adjusted by Bonferroni correction) for testing for variance within mean shift bandwidths, resulted in the identification of one copy number variant region compared to the reference genome: *gch* (PF3D7_1224000). The mean shift identified duplicated region of *gch* coincided with two pairs of inverted reads. This singular genic copy number variation is a known duplication in the 3D7 strain of *P. falciparum* from which the reference genome is based, but was not assembled [132]. From further testing of the mean shift copy number variation identification on both the FCB and 106/1 laboratory strains, we identified from the mean shift on read depth 108 copy number variants outside the sub-telomeric regions at a testing threshold bandwidth p-value < 0.05 and segment size ≥ 400bp. As these were well-characterized laboratory strains, the 108 identified mean shift variants represented numerous false positives. With more stringent p-value thresholds, the number of identified copy number variants significantly decreased. When taken separately, the discordant read analysis displayed lower specificity, identifying over 2,000 regions with discordant read pairs per library. Additionally, there was great variability between library preparations for both GC bias (Figure 3.2) and the number of mean shift calls (Figure 3.3). The high rate of false positives with either method alone suggested the need to combine these metrics to improve specificity. The best combination was simply to use the discordant read placements as confirmation of read depth based CNV calls. Doing this, we identified only 11 copy number variants in the 3D7, FCB, and 106/1 strains – representing 5 copy number variation regions (Table 3.1).

Four of the five copy number variant regions had been established in laboratory strains and were genic. There were no known CNVs within these strains that were not detected suggesting the algorithm has good sensitivity to detect typical genic CNVs. We determined that 3 instances of genic copy number duplications were related to drug resistant alleles – *gch*, *mdr1*, *dhfr*. The duplications had significant support from inverted read pairs (Figure 3.4). These genes have been well documented as copy number variant [87]. The two deleted regions on chromosome 9 for both FCB and 106/1 are known

Figure 3.2 High variability of GC bias between sequencing libraries. Histograms of GC% between sequencing library replicates display the extent of variability between sequencing library preparations. Orange and blue lines represent two technical replicates of Illumina GAII sequencing libraries from a *in vitro* 3D7 lab culture DNA isolate.

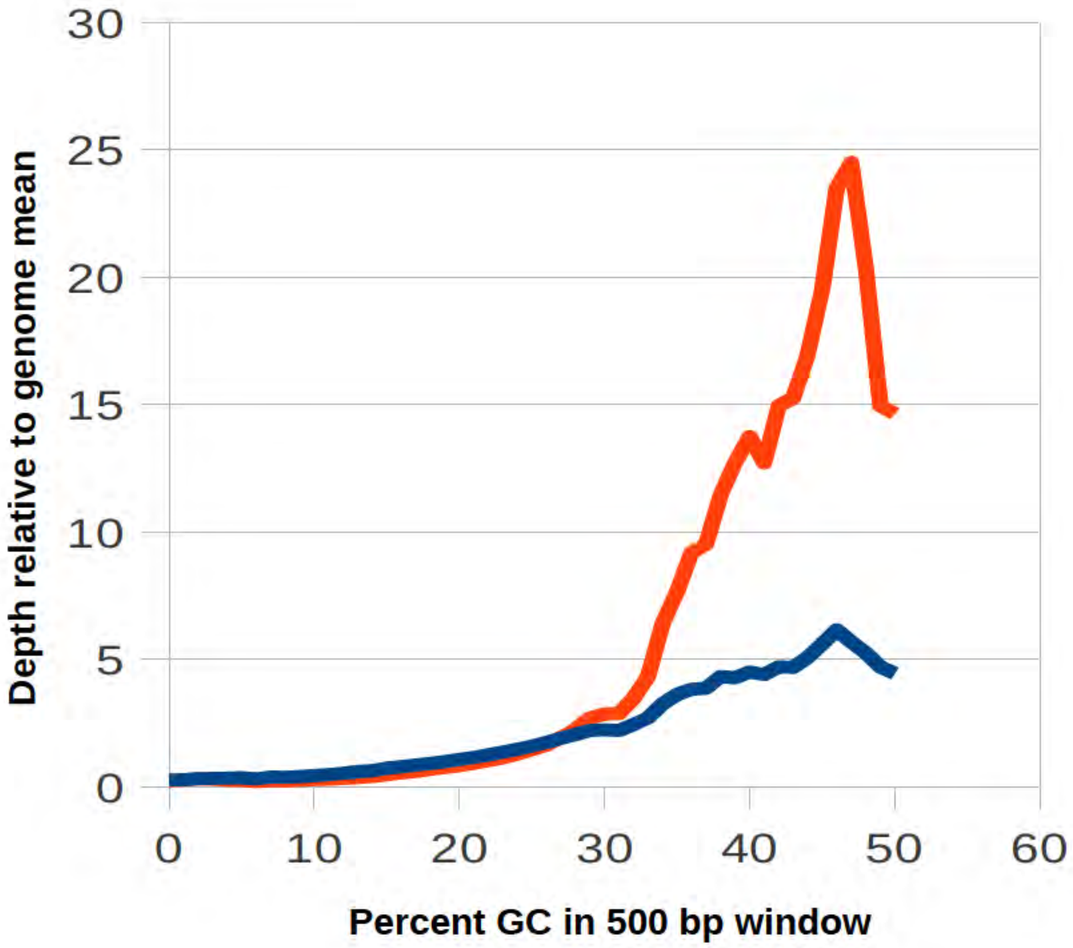

30
25
20
15
10
5
0
Depth relative to genome mean
0
10
20
30
40
50
60
Percent GC in 500 bp window

Figure 3.3 High variability read depth based mean shift CNV identification between sequencing libraries. We detected orders of magnitude difference in the CNV calls by our read depth mean shift method in three *in vitro* cultured laboratory strains: 3D7, 106/1, and FCB. While more stringent p-values reduced the number of false positive CNV calls, FCB and 106/1 still maintained high numbers of false positive CNV calls.

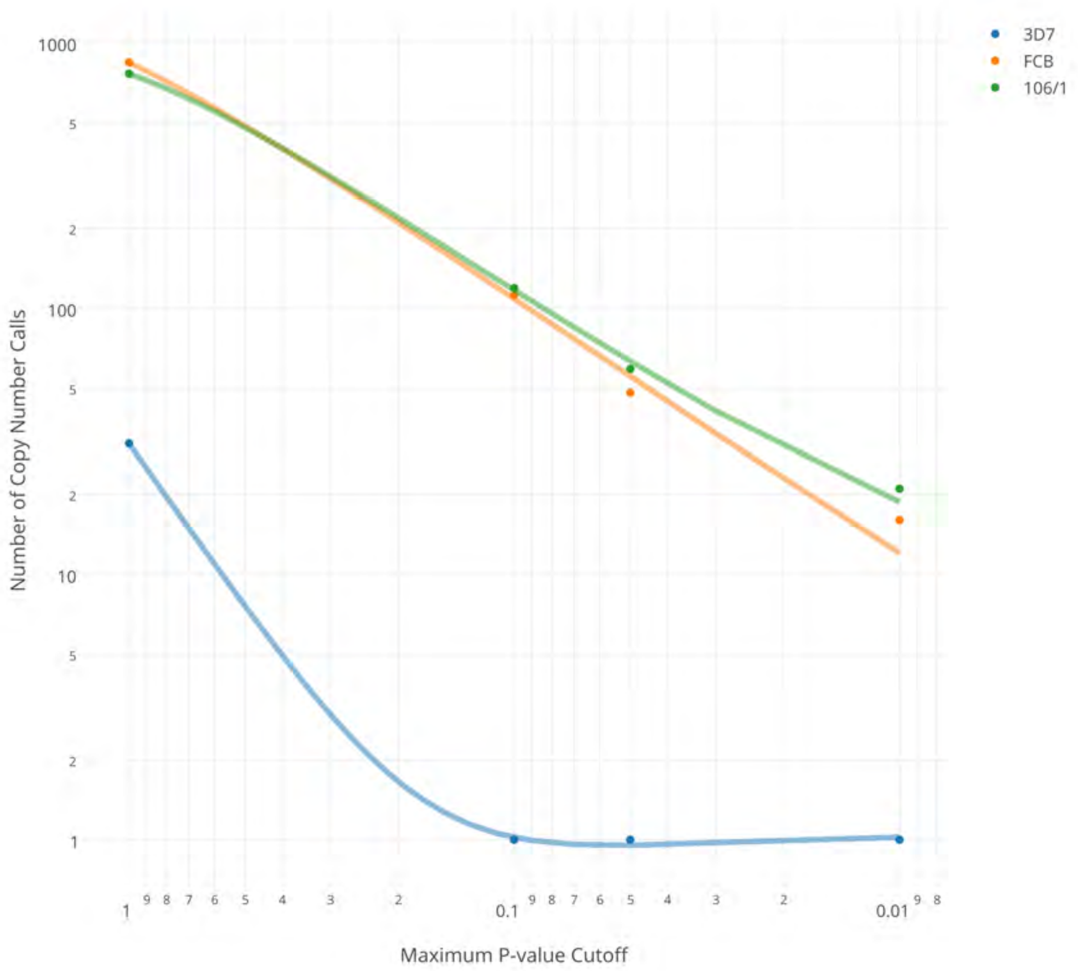
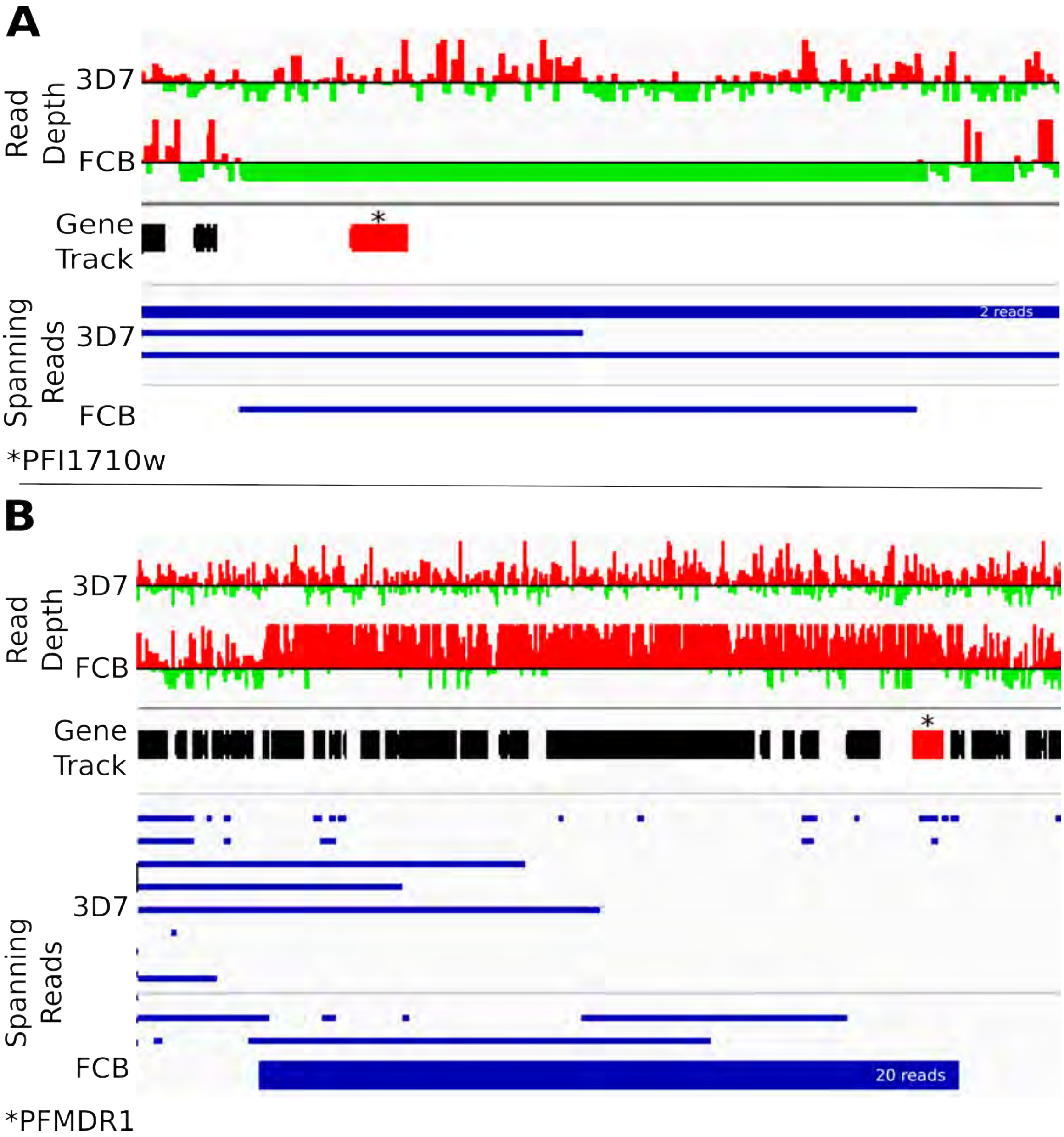

3D7
FCB
106/1
Number of Copy Number Calls
1000
5
2
100
5
2
10
5
2
1
1
9
8
7
6
5
4
3
2
0.1
9
8
7
6
5
4
3
2
0.01
9
8
Maximum P-value Cutoff

Table 3.1 Identified CNV from *in vitro* cultured lab strains. We identified 5 CNVs from *in vitro* cultured lab strains 3D7, 106/1, and FCB. Two were well-studied CNVs of drug resistance associated alleles: *gch* and *mdr1*. We also identified another novel drug resistance associated CNV in *dhfr*. In addition, the chromosome 9 CNV is well-known as an *in vitro* culture adaptation.

Strain	Chrom	Position	Length	Presumed relevant gene	Copy Number
106/1	4	732101 - 783400	51299	pfdhfr1	1.91
FCB	5	868901 - 964600	95699	pfmdr1	2.09
106/1	5	869001 - 952700	83699	pfmdr1	1.65
FCB	7	1293001 - 1295500	2499	-	0
106/1	9	1374001 - 1396100	22099	pfgdv1	0
FCB	9	1374401 - 1396000	21599	pfgdv1	0
FCB	9	1459502 – 1541735	82233	-	0
106/1	9	1459502 – 1541735	82233	-	0
106/1	12	946301 - 980600	34299	pfgch1	4.24
FCB	12	961501 - 980600	19099	pfgch1	2.90
3D7	12	974301 - 975900	1599	pfgch1	3.67

Figure 3.4 Intersection of read depth and discordant reads identifies CNV of PFE1150w and PFI1710w. A) Read depth of the FCB copy number variation on chromosome 5 over the PFE1150w (*mdr*) locus was verified with the presence of 20 inverted reads. Above, the read depth is displayed centered around a normalized read depth for both 3D7 and FCB. Below are the alignment of inverted read pairs that overlap the duplicated region. B) Read depth of the FCB copy number variation on chromosome 9 of the PFI1710w (*gdv*) locus was verified with the presence of spanning reads. Above, the read depth is displayed centered around a normalized read depth for both 3D7 and FCB. Below are the alignment of appropriate orientation read pairs with a read product length greater than 400 bp.

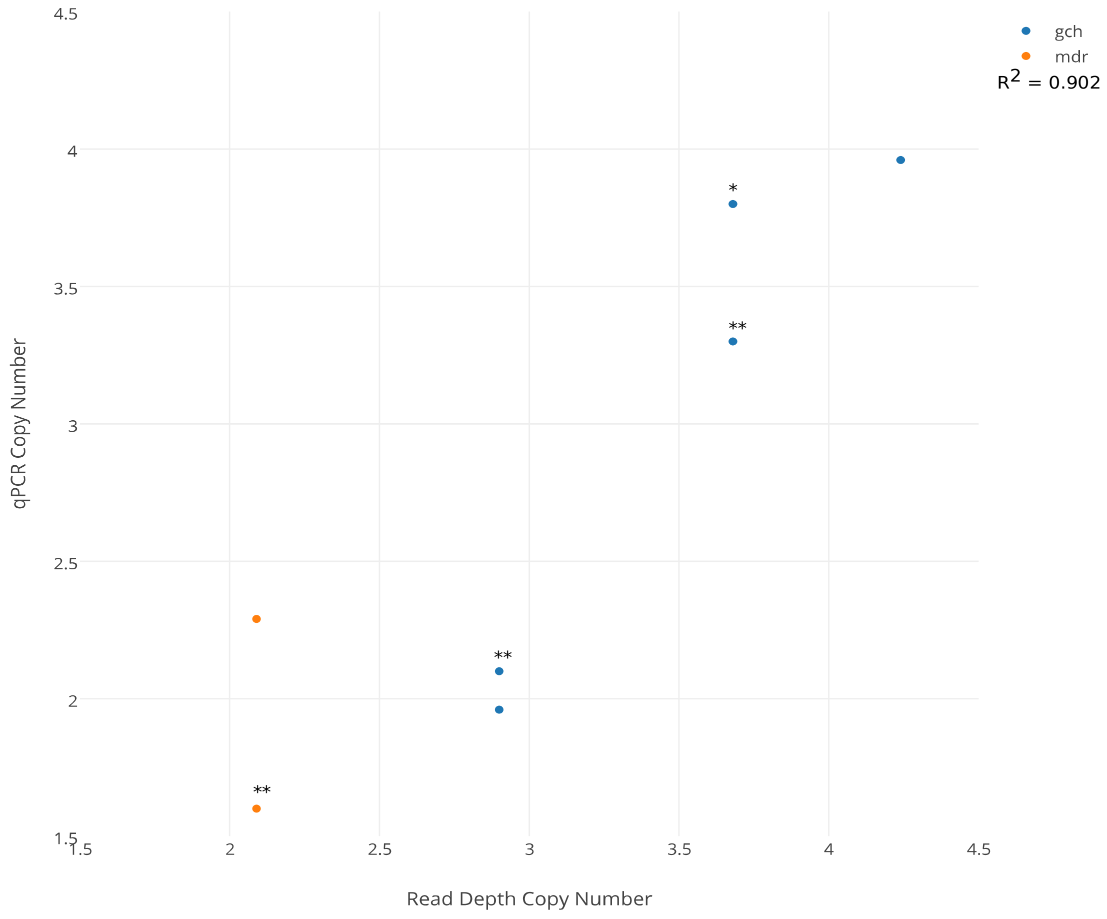

A
Read Depth
3D7
FCB
Gene Track
*
Spanning Reads
3D7
2 reads
FCB
*PFI1710w
B
Read Depth
3D7
FCB
Gene Track
*
Spanning Reads
3D7
FCB
20 reads
*PFMDR1

deletions that are common occurrences within *in vitro* cultured strains [84]. This deletion has been shown to be stable in two isoforms as either a single large deletion or a two part deletion with a single genic deletion of *gdv* (PFI1710w) and the rest of the subtelomeric region after some variable amount of intervening sequence – of which both 106/1 and FCB were the latter (Figure 3.4) [71, 133]. Our read depth based copy number calls accurately estimated the copy number of duplications. We compared of our read depth copy number estimations against our own qPCR results and those in the literature (Figure 3.5) [71]. We found that our estimated copy number calls for read depth correlated strongly with the qPCR results ($R^2 = 0.902$).

Copy number detection within Senegal strains

To examine the extent of copy number variation in a natural popuation, 33 paired end *P. falciparum* samples from Senegal were analyzed for copy number variation with our computation methodology. Ignoring copy number variants located within the subtelomeric regions or internal *var* clusters where reads poorly map to the reference genome, we identified 1445 potential copy number variant calls from the read depth based mean shift at a testing threshold with a p-value of 0.05. A small random sample of potential copy number variation targets outside the subtelomeric regions were queried by qPCR, but were proven to be false positive (data not shown) confirming a high false positive rate in a purely read depth based method. Inspection of the read depth at these copy number variation regions indicated the potential for regional biases between libraries. This further highlighted the need for a secondary metric as unaccounted for

Figure 3.5 Accurate estimation of copy number via read depth based CNV detection. We compared our copy number estimates for duplications to qPCR results from both our lab and the literature (R^2 = 0.902). Denoted points were qPCR results from: *Kidgel *et al.*, 2006 [132]; ** Kiwuwa *et al.*, 2012 [134].


4.5
4
3.5
3
2.5
2
1.5
qPCR Copy Number
1.5
2
2.5
3
3.5
4
4.5
Read Depth Copy Number
● gch
● mdr
R^2 = 0.902
*
**
**
**

biases apart from read depth cause library variability. After filtering for the mean shift calls with discordant read pairings, eight copy number variant regions were discovered across the isolates in non-subtelomeric regions. Only two of the copy number variants were duplications, and six of the eight involved at least partial genic content (Appendix Table 1). Only one of the duplicated copy number variants were genic.

Five of the copy number variants were recurrent, with four of the five being deletions and three of the five having genic content. The chromosome 2 and 9 deletions of the telomeric arms are known adaptations to in vitro cultures. Considering this, these copy number variants, along with the telomeric arm deletion copy number variation on chromosome 1, may be predominantly a result of the short term *in vitro* culture adaptation prior to library preparation. All three were recurrent deletions.

Antigenic genes were the largest contributor in the genic copy number variants. Excluding the three copy number variants suspected of in vitro culture adaptation, we identified a total of 20 genes in all other copy number variants. Nine of these genes had unknown function. Another nine were antigen associated, with six being merozoite protein in the chromosome 10 copy number variation. Additionally, there was little evidence for any bias in GC% for copy number variants. The mean GC% for the copy number variants was 20.0%, whereas the genomic mean is 19.34% GC [71]. Only the non-genic duplication on chromosome six had a significant difference in GC% at 10% GC. However, the copy number variants showed a small but statistically significant bias (two-sided t-test, p-value = $8.3x10^{-3}$) for increased tandem repeat content, with the mean tandem repeat content of the copy number variants being 25.02% of the copy number variation as opposed to the total tandem repeat content of the genome at 14.93%. This implies that significant tandem repeat content is not a factor in creating copy number variants.

3.4 Discussion

We created an improved computational method for identifying copy number variation from next gen sequencing technology that combined both a read depth based metric and a discordant read pair metric. A single metric alone fails given the high-sensitivity needed within a genome with a paucity of copy number. Testing it on laboratory strains with known copy number variants, we were able to accurately identify those known copy number variations. The read depth of the natural strains of 33 paired end sequenced *P. falciparum* strains from Senegal showed systemic, variable, and regional biases, resulting in false positive copy number variation. Our process of improved mean shift depth analysis combined with discordant read pairs, greatly improved specificity, removing many of these false positives. Overall, we discovered significant evidence for only 26 copy number variants, of which only nine were genic, suggesting natural populations of *P. falciparum* are relatively conserved, excusing the subtelomeric regions.

The systemic biases discovered in the Senegalese samples show that a purely read depth based approach to copy number variation discovery may be highly fraught with false positives. These biases also showed little relation to the overall variability of read depth in the sample. As we did not see the phenomenon in the *in vitro* lab samples, there is some variable between source DNA quality, library preparation or sequencing that can cause these systemic biases. Guarding against such biases will be important for all future investigations. Using a secondary metric independent of read depth effectively mitigated the effect of the systemic bias dramatically increasing our specificity. This alleviates the need to investigate the cause for the systemic bias and further normalization of the data. Our method can be applied to current sequencing efforts without additional sequencing or laboratory testing.

Our analysis of the Senegal population of *P. falciparum* indicates a very quiescent population, as concerns copy number variation. Discounting the large chromosomal arm deletions, there were only three copy number variants with genic content. Of those genes affected, antigenic genes were the common element. We found no amplification of genes with potential known effects on drug resistance. While gene amplifications for known drug resistance alleles have been shown to be uncommon, the relative lack of copy number variation is unsurprising [132, 134]. This may be a regional effect due to sudden bottlenecking or from a vast effective population – to the random exclusion or dilution of copy number variation propagation respectively. The majority of samples sequenced and analyzed for copy number variation analysis were from 2008-2010, after significant inroads were made by Senegalese government for malaria control – reporting an estimated 41% drop in confirmed malaria cases between 2008 and 2009 [33]. R. Daniels *et al.* (2013) further supports this idea from their determination that over this time frame there had been a significant increase in propagation of clonal populations of *P. falciparum* in Senegal [135], which was shown to not be occurring in neighboring regions

[136]. This evidence points to a limited pool of genetic diversity of *P. falciparum* in Senegal. Alternatively, the changing landscape of drug use in the region may be at cause for the lack of copy number variation variety.

In addition, recent drug efforts within Senegal may have produced differing selective pressures on copy number variation propagation, as they have with particular resistance associated allele frequencies [137]. Without analysis of other populations of *P. falciparum*, we cannot be certain that relative paucity of copy number variation results are not a peculiarity of this population or if it is truly indicative of extent of variation throughout the global *P. falciparum* population. However, our analysis of a few laboratory strains suggests that it is most likely a global phenomenon, consistent with our duplication analysis and comparative analyses, that central regions of the chromosome are very staid. With our methods for highly specific copy number variation identification from next gen sequencing technology, an investigation of the global population of *P. falciparum* for copy number variation is the next logical step.

CHAPTER IV: Global Analysis of Copy Number Variation in *P. falciparum*

4.1 Introduction

Little is known of the extent of copy number variation in *Plasmodium falciparum*. Many studies have investigated the impact and significance of previously identified copy number variants. These copy number variants were identified via qPCR and genetic cross screens for identification of drug resistance factors [91]. These specific copy number variants were further investigated to determine the various isoforms and copy number of copy number variants within populations of *P. falciparum* [136–138]. Until recently, the study of copy number variation has been performed after determination of its role in drug resistance. However, a few studies have attempted to capture genome-wide copy number variants by array CGH or through next gen sequencing [69, 81, 82]. Most work has focused on lab strains and little has been done to gain a comprehensive study of copy number variation globally.

The investigation of copy number variants in *P. falciparum* to date has not been sufficient to understanding either the extent of coverage or the extent of its effects. Lab strains are not representative of the *P. falciparum* population as a whole. In particular, the bias for study of drug resistance, both in strains and copy number variants, ignores large regions of the potential adaptive benefit of copy number variants for the parasites. A more comprehensive study of copy number variants in a natural population can identify gene targets under selective pressure other than drug resistance, such as adaptations to human immune pressure or external factors like bed netting and insect repellent. In addition, *in vitro* culture introduces its own biases for adaptation, as evidenced by recurrent independent deletions of large regions of the genome in lab strains [70, 80]. Therefore, we have potentially only studied a microcosm of the copy number variation in *P. falciparum*.

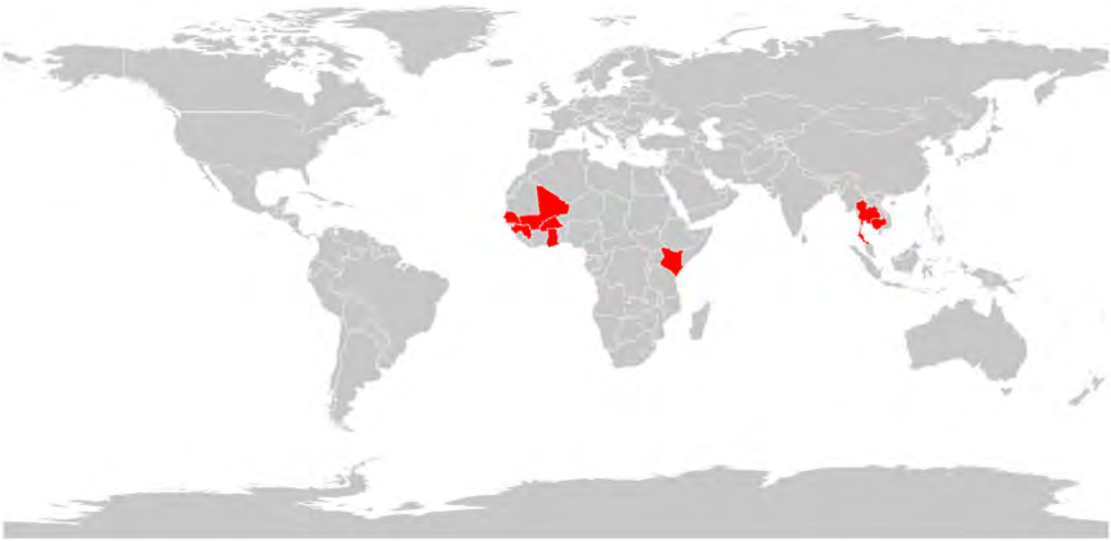
Our previous work has been shown to rapidly and specifically identify copy number variants, both *in vitro* and in a natural population. The nature of next gen sequencing provides a rapid and cost-effective means to broadly identify genomic variation, i.e. sing nucleotide polymorphism, structural variation, and now copy number variation [83–86]. The natural progression would be to extend our technology to a global data set. Large scale sequencing studies of *P. falciparum* are becoming common and a wealth of sequencing data is accumulating around the globe. We use our methodology for copy number variation detection from high throughput sequencing to investigate copy number variants across these publicly available sequencing datasets from a large global sample of *P. falciparum*.

4.2 Methods

Sample acquisition and quality control

The set of whole genome shotgun sequences of *P. falciparum* isolates was gathered from multiple published papers and publicly available sequence data: Burkina Faso, Gambia, Ghana, Mali, Cambodia, Thailand [71]; Guinea [44]; Gambia [88]; Kenya [139]; Senegal [90] (Figure 4.1). All samples had been pair end Illumina sequenced. We merged samples with multiple sequencing libraries prior to normalization. We had previously determined that of the samples with multiple libraries there was minimal

Figure 4.1 Map of *P. falciparum* sample origin.



difference in read depth variation between GC normalization on individual libraries and merged libraries. We filtered any sequencing libraries with less than 70% genomic coverage or less than 10x coverage, as both criteria are required for initial read depth analysis. We tested the remaining sequencing libraries for multiplicity of infection with estMOI [89]. No sampled showed greater than a cumulative 5% secondary infection insuring that multiple strain infections would not confound our analysis. The final cleaned data set had 610 unique samples. From this, we can refer to each sample as being a unique strain as concerns copy number variation and major allele identification.

Copy number variation identification

We analyzed all sequencing libraries for copy number variation as previously established. We normalized read depth for GC bias on a per read basis. Then we analyzed the normalized read depth for copy number variation with a mean shift algorithm. We removed any copy number calls that intersected with the subtelomeric regions or var clusters due to an inability to adequately account for significant sequence diversity between paralogs. These regions were effectively removed with the secondary metric of read discordant reads: inverted read pairs for duplications and spanning reads for deletions.

Principal component analysis for CNV subgroup analysis

To look for population structure associated with particular CNVs, we partitioned the isolates to those having or lacking a particular copy number variant and assessed the SNPs by principal components. Specficially, we examined all SNPs from PlasmoDB version 9.3. SNPs were called by GATK version 2.8.1 with local indel realignment and verified against the PlasmoDB SNP database [140]. For each chromosome, an array of GATK-called SNPs limited to the sites in PlasmoDB version 9.3 per sample was built for principal component analysis in R with prcomp. Analyzed samples included only regions positive for the copy number variants: Cambodia for plasmepsin duplication; Burkina Faso, Ghana and Guinea for *crt* duplication; Thailand and Cambodia for both *mdr1* and *gch* duplications.

4.3 Results

Identified copy number variants

From 610 samples, we identified 68 copy number variant regions – 54 of which were duplications and 61 of which were genic (Appendix Table 1). We determined that 61 of the copy number variants intersected a gene, however only 33 fully encompassed a gene (for at least one isoform). In addition, we determined that 22 of the 68 copy number variants were recurrent between strains. Within these 22 copy number variants, 16 fully encompassed a gene. When considering the individual copy number variants in each strain, only 17 instances of unique genic copy number variants were present. Alternatively, the recurrent deletions accounted for 144 individual instances of genic copy number variants. Of the 22 recurrent copy number variants, we determined that copy number variants were not necessarily confined to a particular country or region. Nearly half of the recurrent copy number variants were restricted to a single region (West Africa, East Africa, South East Asia) with 10 of the 22 recurrent deletions spread between multiple regions. It was even rarer to find a recurrent copy number variant confined to a single country, with only 6 copy number variants remaining confined to a single location. However, copy number variants between continents were more limited. This may highlight differences in selective pressures between continents.

The frequency of strains with genic copy number variants was highly variable both between copy number variants and geographic regions, ranging from 2% to 54% (Figure 4.2). Strains averaged 0.50 copy number variants, with a range between 0 and 5 copies per strain and a standard deviation of 0.82. However, Cambodia and Thailand both had significant percentages of their population with genic copy number variants, with 40% and 55% respectively. African samples had a much lower frequency of copy number variation. All African samples were under 10% frequency, except Ghana at 16%. However, the high frequency of copy number variants in South East Asia is mostly due to two recurrent copy number variants of *mdr1* and *gch*. In Cambodia, 42 out of 62 strains with copy number variants were due to *mdr1* or *gch*, while Thailand had all samples containing a copy number variant with either a *mdr1* or *gch* duplication. The *gch* duplication accounted for 6 of the 12 strains in Ghana with copy number variants.

Across the genome, copy number variants showed no bias for particular chromosomes, nor any pattern to location of the copy number variation (Figure 4.3). We found that the copy number variants also did not show biases toward GC%, with a mean of 18.8% GC in all copy number variants as compared to the genomic mean of 19.4%. While we noticed an increase in tandem repeat content of the copy number variants, however the difference in mean tandem repeat content of 18.4% for copy number variants

Figure 4.2 Frequency plot of genic CNVs. The frequency plot of genic CNVs indicates that most genic CNVs are present at a low level in the population. However, we discovered a number of genic CNVs that are highly prevalent in the population – particularly *gch* and *mdr1* duplications.

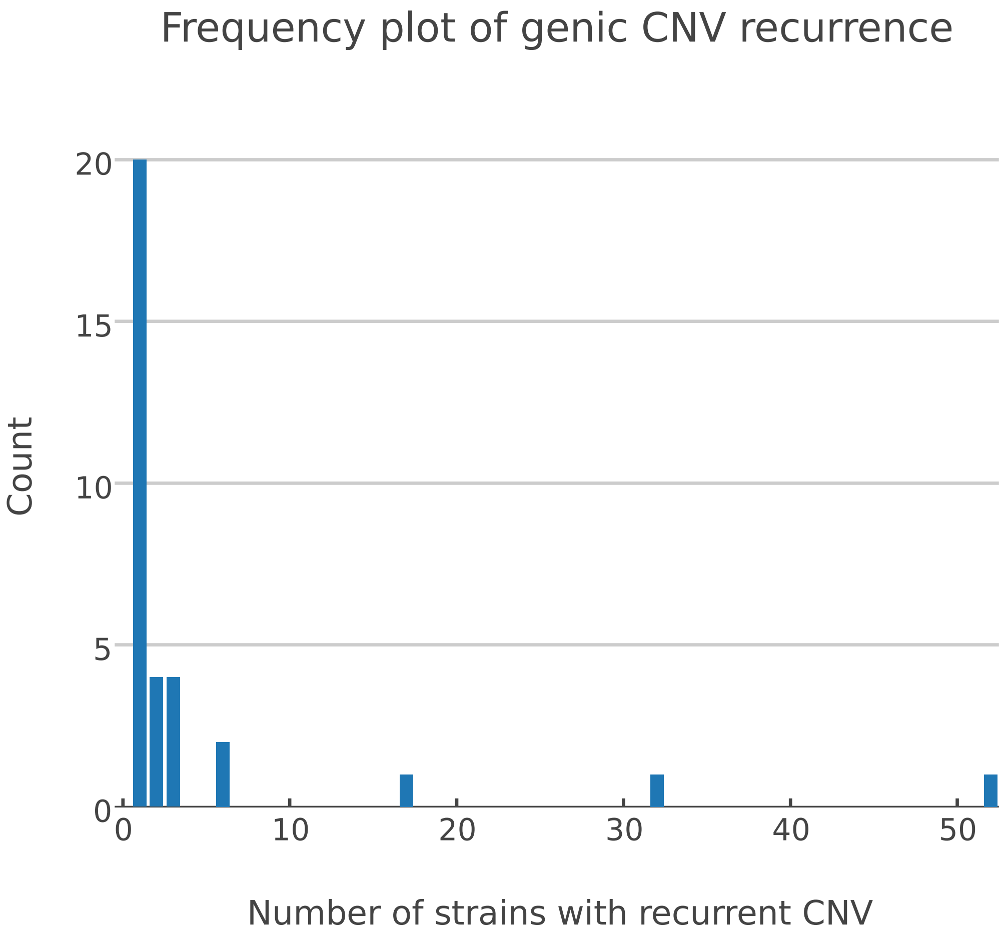
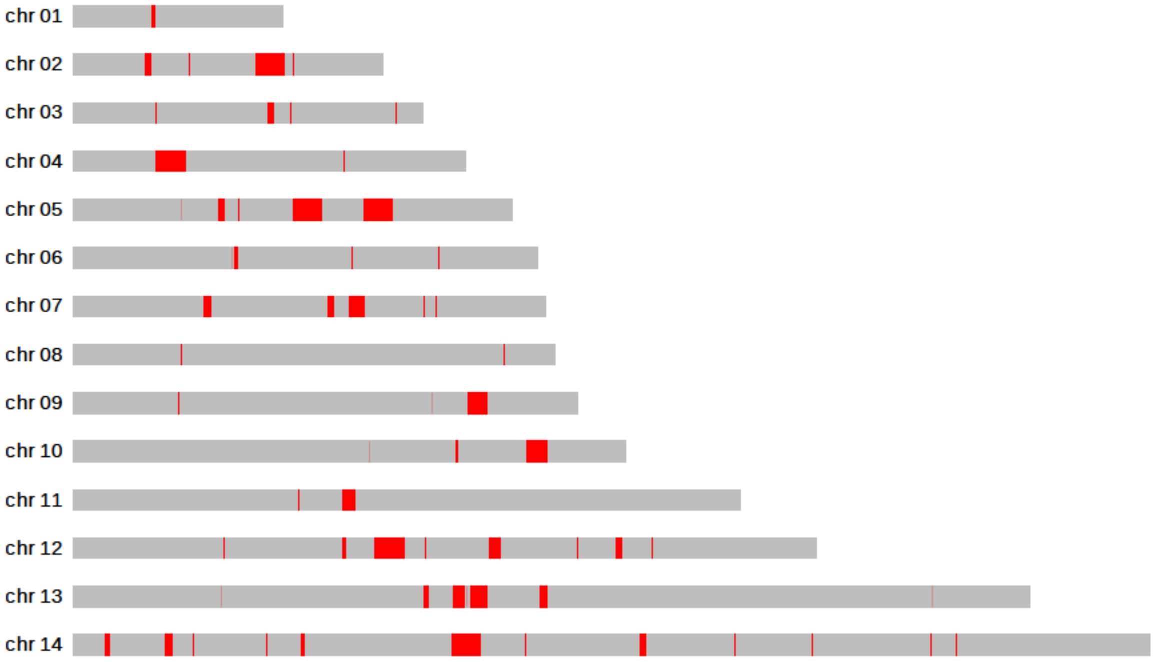

Frequency plot of genic CNV recurrence
Count
0
5
10
15
20
0
10
20
30
40
50
Number of strains with recurrent CNV

Figure 4.3 Overview of CNV loci throughout the genome. The distribution of CNV loci indicates that there is likely no bias for particular regions or loci in the genome (not considering the subtelomeres).


chr 01
chr 02
chr 03
chr 04
chr 05
chr 06
chr 07
chr 08
chr 09
chr 10
chr 11
chr 12
chr 13
chr 14

against the genomic mean of 14.9% was not statistically significant (double sided t-test, p-value = 0.081). We determined that there were likely no biases between GC% and tandem repeat content related to copy number variation size as there was little difference between the unweighted means and the copy number variation size weighted means (weighted GC% mean = 20.1%, weighted tandem repeat percentage = 19.5%).

The distribution of copy number from the various isoforms of all detected duplications indicated that most have a copy number of between 2 to 3, with the number of duplications with high copy number rapidly dropping past 3 copies (Figure 4.4). However, the copy number did reach as high as 14 for *gch*. The recurrence frequency of copy number variants indicated that the majority recurred between less than five strains. However, the range of frequency was wide, ranging up to 73 strains for a deletion and 28 strains for a duplication (Figure 4.2).

Copy number variation size was skewed toward smaller sizes (Figure 4.5). The majority of copy number variants were under 30 kbp, and the vast majority under 60 kbp. The small selection of copy number variants greater than 60 kbp were all large deletions of the subtelomeric arms of chromosomes. These deletions are considered *in vitro* culture adaptations, so *in vivo* there is strong pressure to limit large copy number variants [71, 133].

Genes within copy number variants

We discovered 201 genes full encapsulated within a copy number variant regions, of which 79 (39%) were of unknown function. Between recurrent and single copy number variants, the majority of genic copy number variants were recurrent, with 132

Figure 4.4 Copy number of genic duplications indicate preference for minimal duplication. The frequency of higher copy numbers drops drastically after 4 copies, with most duplications found at two copies. However, there are a few high copy duplications, with a notable example being *gch* at 11 copies.

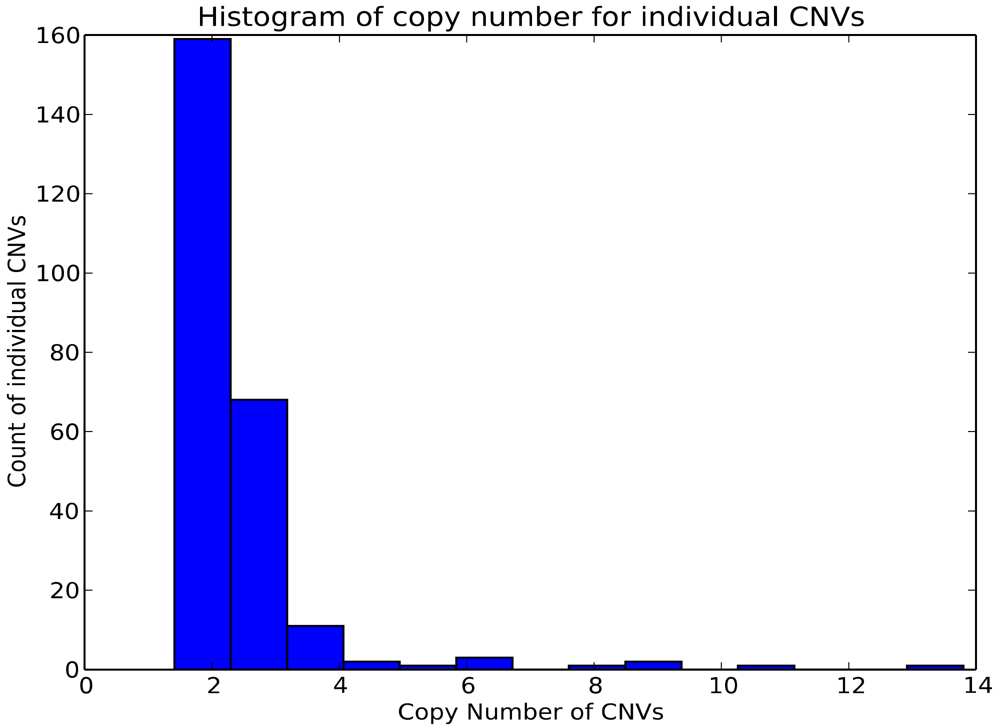
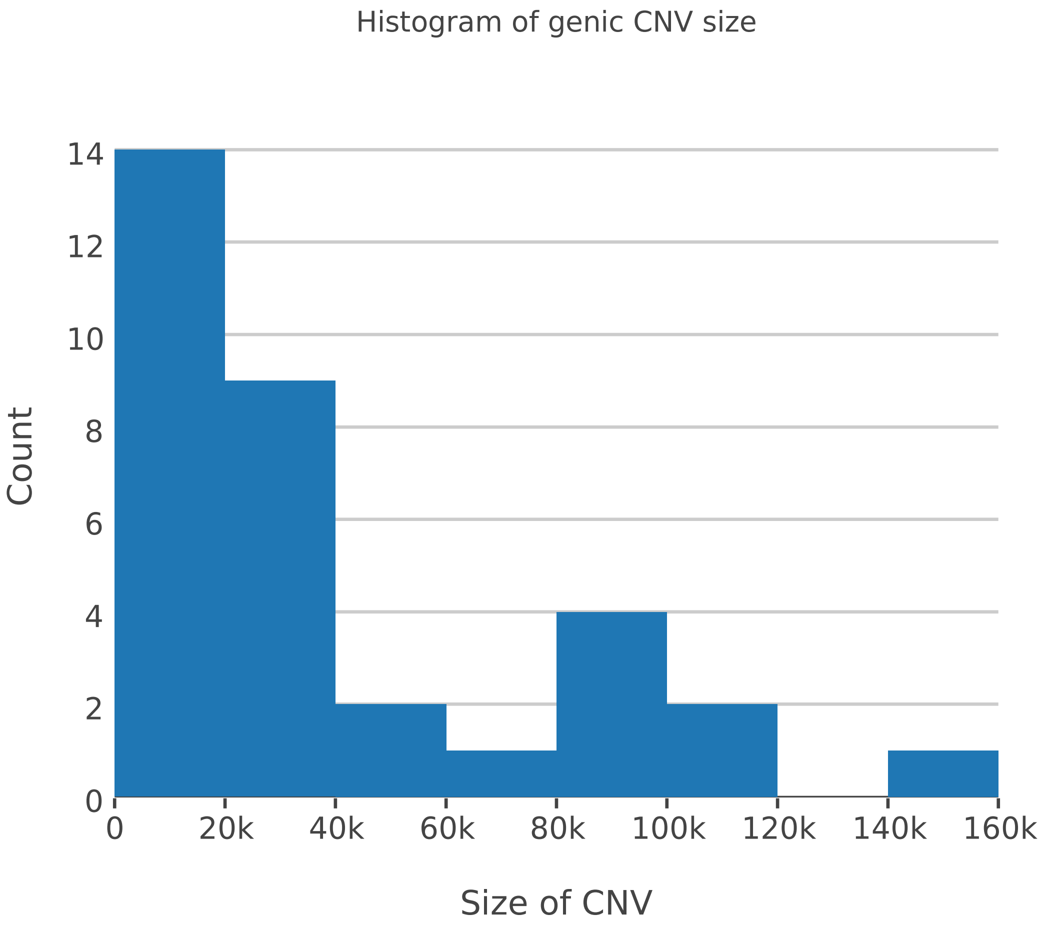

Histogram of copy number for individual CNVs
Count of individual CNVs
Copy Number of CNVs
0
20
40
60
80
100
120
140
160
0
2
4
6
8
10
12
14

Figure 4.5 Majority of CNVs are under 40 kbp. The histogram of genic CNV size indicates that most CNVs prefer smaller sizes – under 40 kbp. There are few CNVs larger than 60 kbp, and those are deletions of entire subtelomeres of chromosomes.


Histogram of genic CNV size
Count
14
12
10
8
6
4
2
0
0
20k
40k
60k
80k
100k
120k
140k
160k
Size of CNV

genes in duplications (from 16 CNVs) and 69 genes in deletions (from 6 CNVs). Alternatively, 122 genes were found in single strain duplications (from 39 CNVs) and only 13 genes in single strain deletions (from 6 CNVs). However, of those 6 recurrent deletions, 3 were previously identified as copy number variants in the Senegal strains that are potentially *in vitro* culture adaptations – the loss of chromosome 9 subtelomere, the loss of chromosome 2 subtelomere, and potentially the loss of chromosome 1 subtelomere. These are deletion events that have not been seen directly *in vivo*, but are frequently produced over long term *in vitro* culture of an isolate [71, 133]. These deletions often result in loss of gametocytogenesis, loss of cytoadherence, and increased growth rate *in vitro* [71, 131, 141, 142].

We found multiple recurrences of two well-known and characterized copy number variants of *mdr1* and *gch*. Overlay of the copy number variants of *mdr1* and *gch* (Figure 4.6 and 4.7) compare similarly to previous studies on break points and copy number of those specific genes in a natural population of P. falciparum [87]. We identified the *mdr1* gene solely in South East Asia, however the *gch* duplication was identified in Ghana in addition to South East Asia. Both copy number variants had a high frequency, with the *mdr1* duplication found in 32 strains and the *gch* duplication recurring in 57 strains. Without information on the sampling protocols or timeline of sample collection, we cannot ascertain whether the disparity in geographic location of the two copy number variants was due to a bias in sample collection timing between studies or an effect of differing drug regimes between regions. Previous studies have shown that the *mdr1* duplication exists at a low frequency in West Africa [143].

Figure 4.6 Multiple isoforms of *mdr1* duplication identified. We conservatively identified 5 separate isoforms of the *mdr1* duplication by breakpoint, with each isoform having variable copy number. Copy number ranged from 2-3 copies.

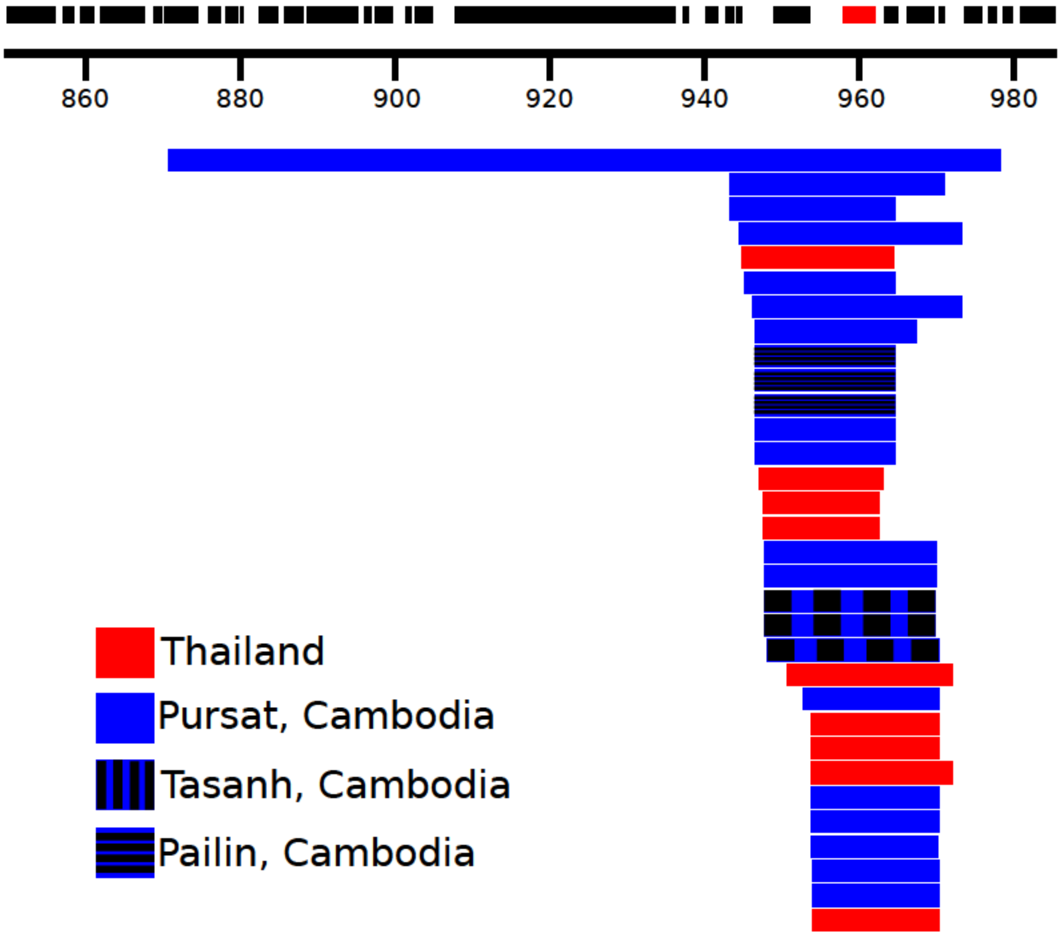
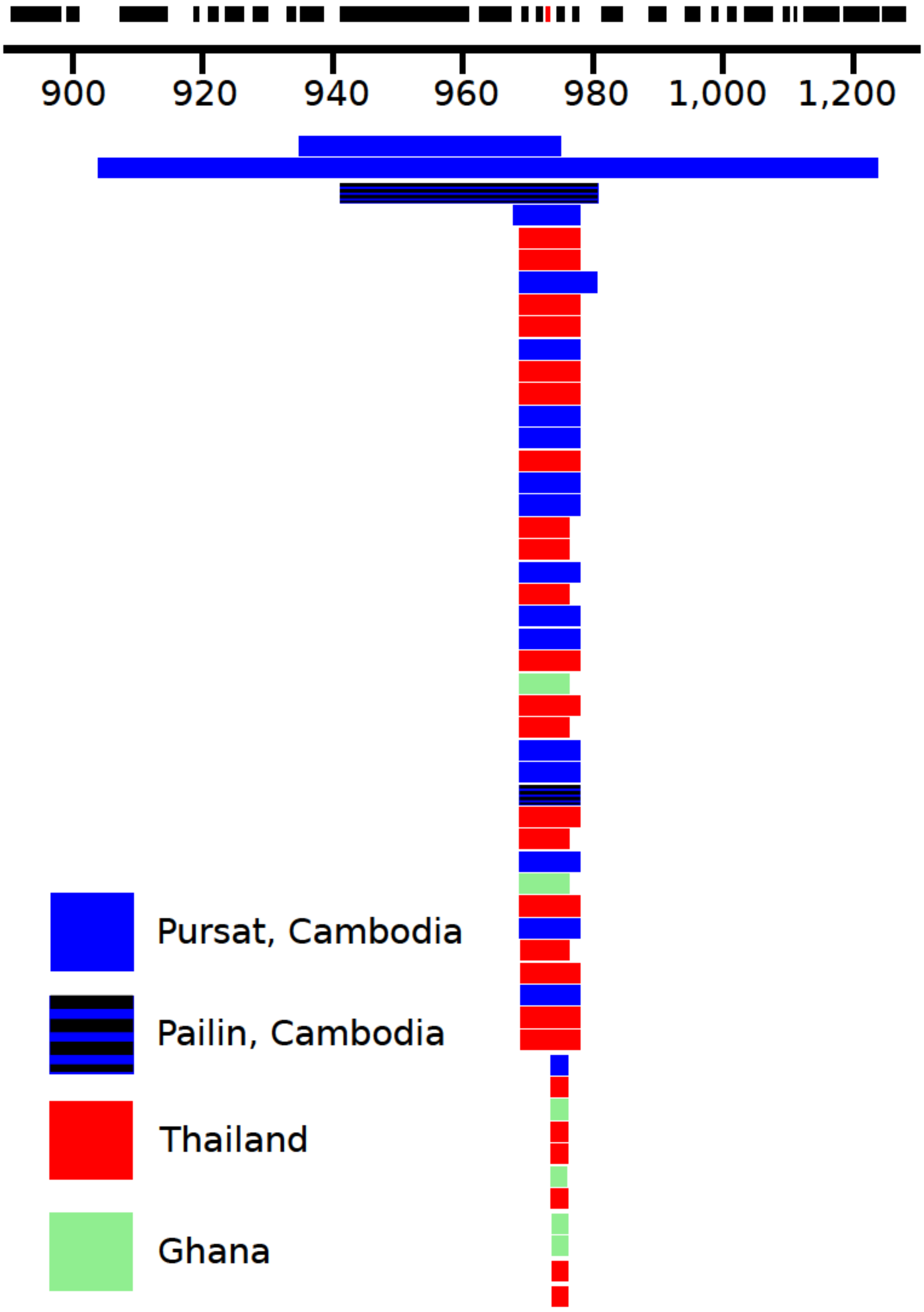

860
880
900
920
940
960
980
Thailand
Pursat, Cambodia
Tasanh, Cambodia
Pailin, Cambodia

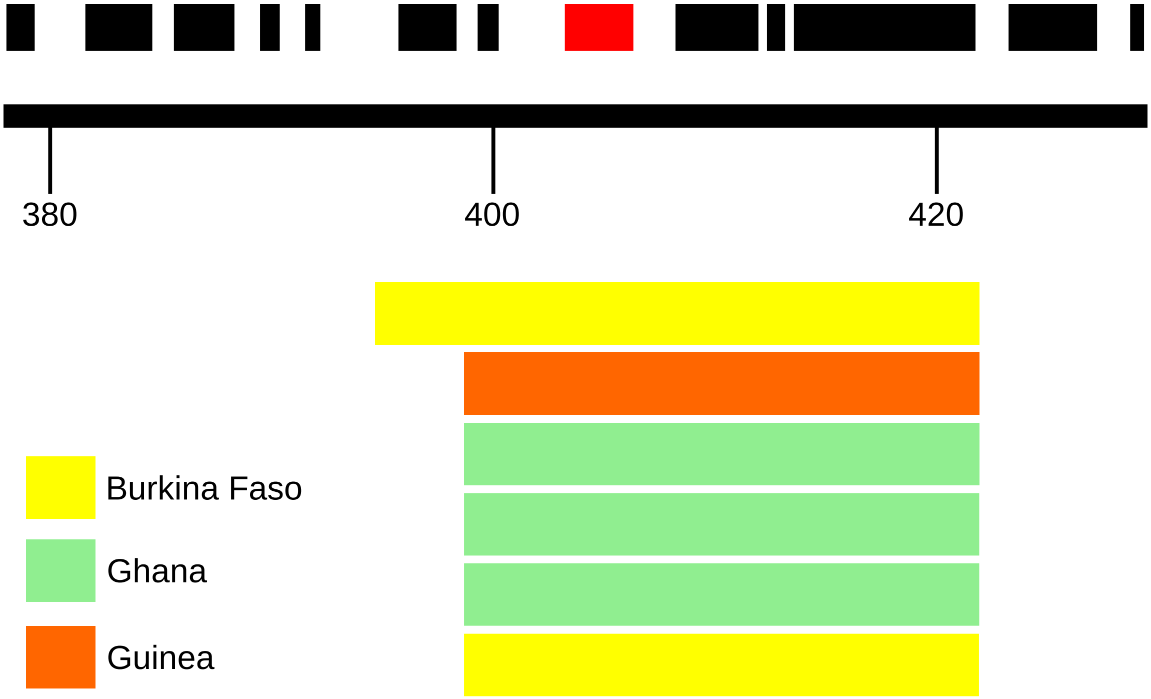
Figure 4.7 Multiple isoforms of *gch* duplication identified. We conservatively identified 6 separate isoforms of the *gch* duplication by breakpoint, with each isoform having variable copy number. The copy number ranged from 2-11 copies.


900
920
940
960
980
1,000
1,200
Pursat, Cambodia
Pailin, Cambodia
Thailand
Ghana

In addition, we discovered many novel copy number variants. Of particular note was the novel copy number variation of another drug resistance associated allele – *crt* (Figure 4.8). This copy number variation was identified in West Africa, spanning multiple countries. The *crt* copy number variation was a duplication with a copy number of two and identified six times between Burkina Faso, Ghana, and Guinea.

Of the 201 genes identified to be copy number variant, 79 had unknown function. Gene ontological testing of the other 122 identified genes with GOstat found no over- or underrepresentation of gene ontology terms [123]. Between different isoforms of a recurrent copy number variation, a set of intersecting genes could potentially highlight the gene under selection in the copy number variation. While this reduced set did not elucidate the causative allele under selection for all copy number variants, a literature search of potential targets informed us of possible hypotheses for select copy number variants. Duplication of plasmepsins II and III (PF3D7_1408100 and PF3D7_1408000) on chromosome 14 was present in 17 strains, all in Cambodia. These plasmepsins form a complex that is involved in the hemoglobin-to-hemozoin process in the parasite's food vacuole [144, 145]. Chloroquine's mechanism of action is via interruption of the hemoglobin-to-hemozoin process to allow free heme to form, which is highly toxic for the parasite [146, 147]. The duplication of topoisomerase I (PF3D7_0510500) also had implications with chloroquine resistance in the literature. Chloroquine has been shown to be a catalytic inhibitor of human topoisomerase I [148]. While topoisomerase I in *P. falciparum* has significant sequence divergence from human topoisomerase I, there is still the potential for similar chloroquine interaction considering there is likely high

Figure 4.8 Identification of single isoform of *crt* duplication. We identified a novel CNV of the *crt* gene. Only one isoform was discovered, all at a copy number of 2.


380
400
420
Burkina Faso
Ghana
Guinea

conservation of the catalytic site. However, most genes involved in copy number variants had little literature on their importance or the functional implications of copy number variation.

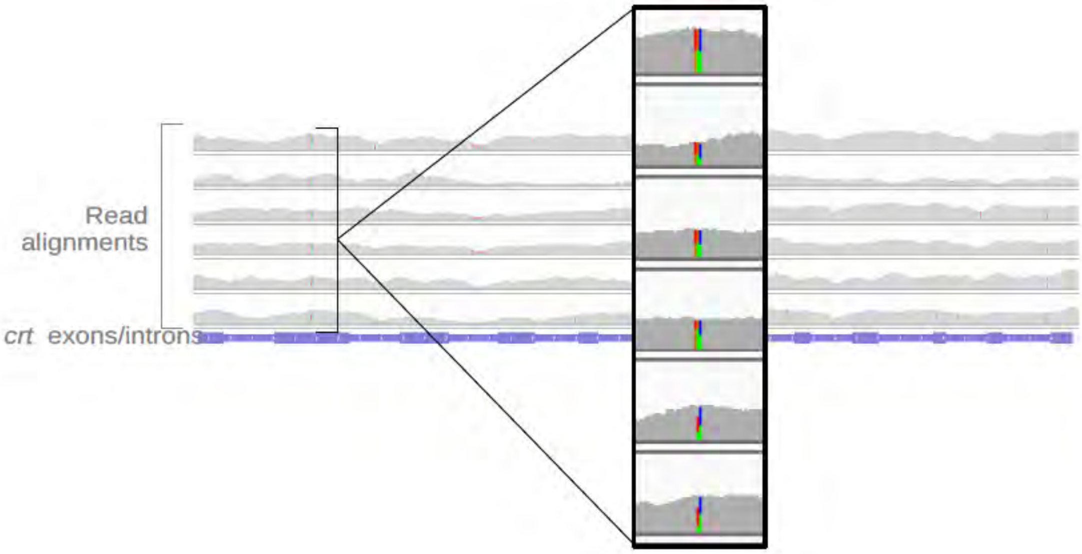
Biallelic *crt* duplication

Upon closer inspection of the *crt* duplication, we discovered that the duplication was biallelic in that we identified both the chloroquine sensitive and chloroquine resistant alleles in the duplication (Figure 4.9). This phenomenon was not seen in any other copy number variation. We found that all SNPs within the copy number variation of *crt* approached 50% frequency in the read depth. From paired read linkage, we could verify that neighboring SNPs were in linkage within individual copies of *crt*. This lends credence to the idea that two separate alleles were present in the copy number variation.

SNP analysis identified nine non-synonymous mutations in exonic regions of *crt*. Included in the mutations is the K76T mutation considered a cornerstone for chloroquine resistance [70, 80]. However the K76T mutation alone is insufficient [149]. In total, all the identified biallelic SNPs in the *crt* duplication correspond to SNPs in the *crt* allele in the chloroquine resistant lab strain Dd2, except for one position at 405,600 bp that did not have a Dd2 linked SNP or were the chloroquine sensitive wild type allele [150, 151].

We determined that recombination had occurred within the copy number variation itself, creating divergence between strains sharing the copy number variation. Breakpoint analysis of the *crt* copy number variation with Pindel (version 0.2.5) found exact breakpoints of 398,871 to 421,765 bp on chromosome 7 in sample 43 from Burkina Faso

Figure 4.9 Biallelic duplication of *crt* gene. We discovered that SNPs within the the *crt* gene were biallelic, with the alternative allele ranging from 30-70% of the read depth. SNPs corresponded with a chloroquine sensitive allele and a chloroquine resistant allele. Aligned reads are visualized along *crt*, with non-reference bases highlighted along the gene. Zooming in on a particular SNP loci displays identification of biallelic SNPs.


Read alignments
crt exons/introns

[152, 153]. These Pindel defined breakpoints correlated well with our mean shift and discordant read identified maximal-span breakpoints of 398,679 to 421,905. Both Pindel identified breakpoints located to a long poly-A stretch (>10 bp). We analyzed SNPs within the copy number variation and flanking regions to identify divergence between the six recurrences of the copy number variation (Figure 4.10). We determined that the shared haplotype diverges upstream at 378,904 bp – about 20 kbp upstream of the breakpoint. However, the shared haplotype ends 3' within the copy number variation itself at 414,615 bp.

Principal component analysis of SNPs located on chromosome 7 for strains in Burkina Faso, Ghana, and Guinea indicated that there was no subgrouping of strains with the *crt* duplication (Figure 4.11), consistent with their observance in multiple countries. Principal components 1 and 2 were still affected by geographic effects. However, at principal components 3 and 4, where geography had much less impact, variation within the *crt* duplicated strains was in line with the rest of the West African population. The lack of a separate subgroup of *crt* duplicated strains reinforces the identification of only a ~40 kbp region of non-divergence around the copy number variant. We can conclude that the duplication has been in the population for some time, and has undergone significant recombination around the locus, occurring over time as this duplication spread within West Africa.

4.4 Discussion

This study is the first comprehensive analysis of copy number variation across the

Figure 4.10 Consensus sequence diverges within the *crt* CNV. By tracking the consensus SNP around the *crt* locus, we determined the boundary of divergence between the isolates' CNVs. The breakpoints of the *crt* CNV are 398,679 bp and 421,905 bp. The divergence boundaries were at 386,728 bp and 414,614 bp.

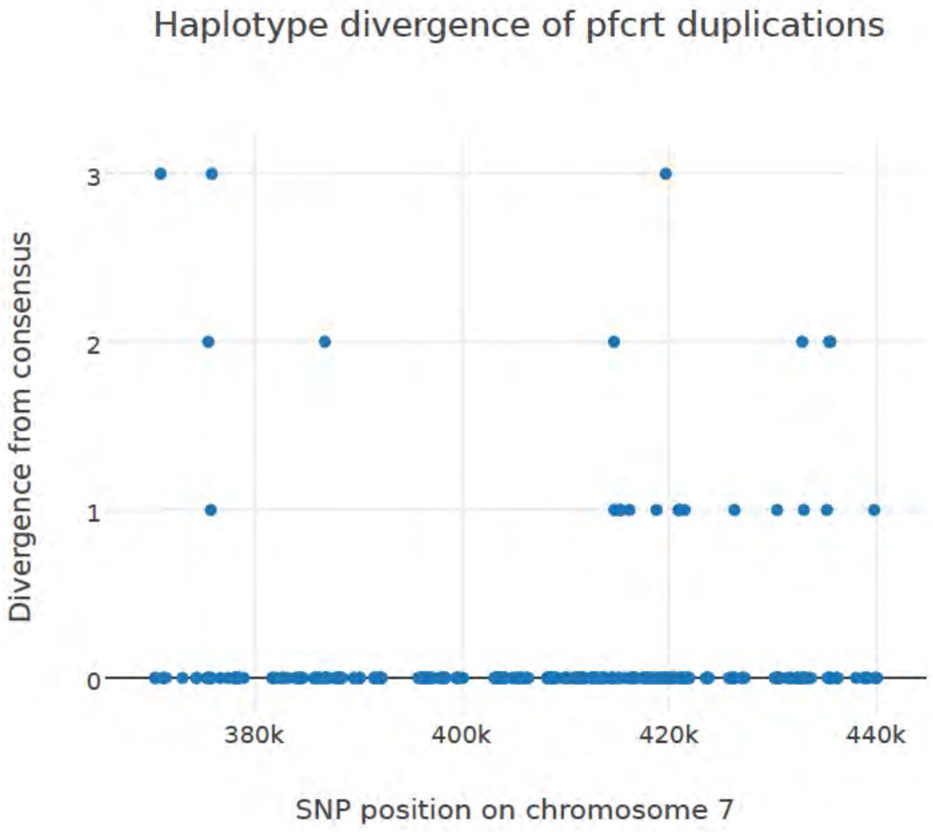
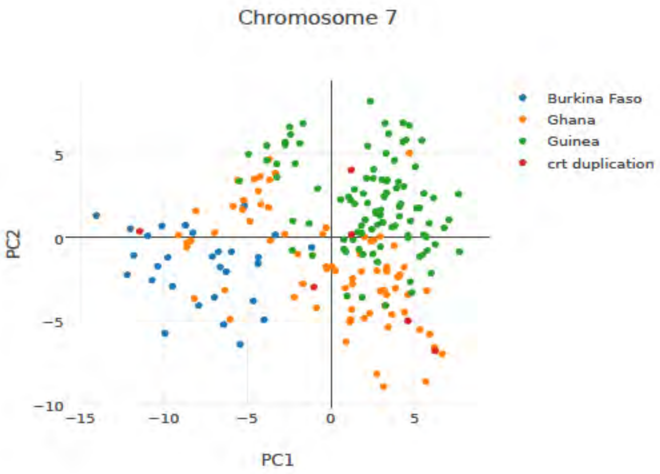
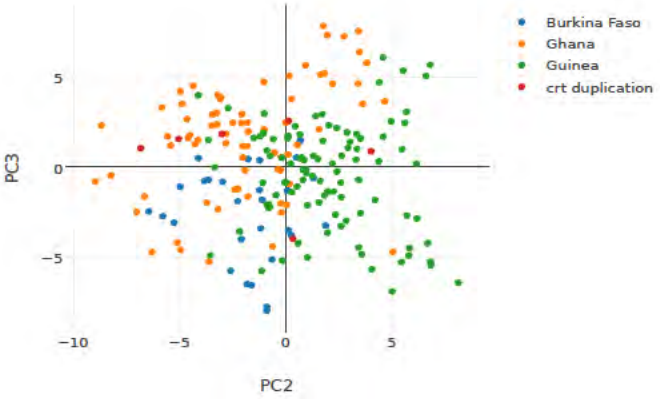
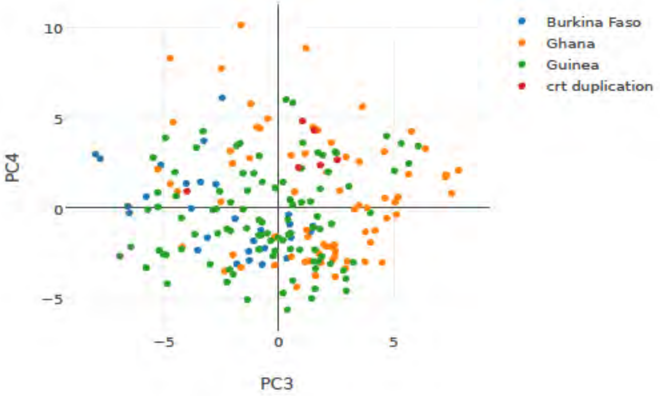

Haplotype divergence of pfcrt duplications
Divergence from consensus
0
1
2
3
380k
400k
420k
440k
SNP position on chromosome 7

Figure 4.11 PCA of SNPs across chromosome 7 between geographic regions with *crt* duplication. Plots of principal components identify significant effect from geographic distribution on SNPs for chromosome 7. *crt* duplication positive strains are shown to be not unique enough a subpopulation to separate, indicating significant recombination has occurred around the duplication.


Chromosome 7
Burkina Faso
Ghana
Guinea
crt duplication
PC2
5
0
−5
−10
−15
−10
−5
0
5
PC1


Burkina Faso
Ghana
Guinea
crt duplication
PC3
5
0
−5
−10
−5
0
5
PC2


Burkina Faso
Ghana
Guinea
crt duplication
PC4
10
5
0
−5
−5
0
5
PC3

global population of *P. falciparum*. We have established a baseline for copy number variation in the core genome as compared to the 3D7 reference genome. Our analysis indicates minimal copy number variation in the non-subtelomeric regions of the genome. We found that copy number variation is a rare occurrence in *P. falciparum*. Individual strains averaged 0.5 copy number variants. However, the dearth of copy number variation was most pronounced in Africa. Strains from South East Asia have a high rate of copy number variation, with up to 50% of strains being positive for a copy number variation. The contributing factor to this was the high frequency of either *gch* or *mdr1* duplications. Outside of those two copy number variants, South East Asia had similar rates of copy number variation as Africa. From this, we can conclude that copy number variation in the non-subtelomeric regions of the genome is rare, unless strong selective pressures elevate the frequency of the copy number variant in the population. In this case, significant drug pressure may be increasing the frequency of copy number variation in South East Asia.

From our analysis, we identified numerous novel copy number variants. By literature research, a number of these copy number variant genes are impacted by chloroquine in some manner – topoisomerase I and the plasmepsins being among those genes. In addition, we identified a novel copy number variant of *crt*, which has known function regarding chloroquine resistance. We discovered a single isoform of the *crt* copy number variant present in six strains between Burkina Faso, Guinea and Ghana. Additionally, we determined that the duplication was biallelic, containing both chloroquine resistant and chloroquine sensitive alleles. This potentially has significant importance in future drug design and discovery. Strategies for combination therapy of drugs that target both chloroquine resistant alleles and chloroquine sensitive alleles, taking advantage of the functional constraint on *crt* mutations, could be seriously compromised by this biallelic copy number variation. The biallelic duplication also potentially provides an adaptation to the fitness disadvantage that chloroquine resistance typically confers in the absence of drug [154, 155]. This might have effects on the improved retention of chloroquine resistance or resistance to other antimalarials in the population by reducing the disadvantageous effects of the chloroquine resistant allele.

From bacteria to fruit flies and humans, *P. falciparum* has comparatively low frequency of copy number variation [5]. Given the low frequency of copy number variation in *P. falciparum*, it stands to reason that there is significant purifying selection against the retention of copy number variation into the genome. The only cases of high frequency genic copy number variants in the population are due to drug resistance-associated alleles, such as *mdr1* and *gch*. Given this, the observed existing copy number variants likely confer some functional benefit providing an adaptive benefit. Our study has provided multiple novel targets to investigate for the functional consequences of these duplications. As drug resistance played a prominent role in a number of identified copy number variants, study into these other copy number variants may provide additional avenues for the evolution of drug resistance.

CHAPTER V: Discussion

5.1 Systematic analysis of duplications and deletions in *P. falciparum*

Duplications and deletions are an important avenue of adaptation and evolution. They are a significant route for the diversification of genes, creation of new genes, and the regulation of gene dosage. In *P. falciparum*, these avenues are extensively used for adaptation. Large gene families of genes encoding extracellularly exposed proteins have undergone significant duplication and divergence to improve amino acid diversity for immune evasion. In addition, copy number variation (duplication) of the particular drug resistance-related genes has provided increased expression and resultant increased drug resistance in *P. falciparum*.

In our studies, we have attempted to systematically identify and analyze these duplication and deletions in the genome. From our analysis of segmental duplications, we have observed the compartmentalization of the genome with respect to the presence of duplication. We determined the rarity of duplications in the core of the genome, where few genes showed evidence of past duplication. From this, we can infer that there is strong selective pressure to conserve the core genomic genes and function leading to removal of duplications barring strong selective advantage. However, we identified significant duplication of genes in the subtelomeres. The majority of genes duplicated were genes under strong selective pressure for human immune evasion. This fits with the inference that the parasite uses the subtelomeres to duplicate and recombine its repertoire of extracellularly exposed genes, thereby diversifying these genes for human immune evasion.

To rapidly and specifically identify these variable duplications and deletions between strains from high throughput sequencing data, we created a novel computational method to identify copy number variants in the non-subtelomeric regions of the genome from high throughput sequencing. This method increased the specificity of copy number variation identification from high throughput sequencing libraries of clinical isolates. The method involved GC bias normalization of read depth followed by detection combining mean shift detection of regions and validation of these identified copy number variants by discordant read pairs in the sequencing library. We applied this computational method to publicly available high throughput sequencing libraries of 610 clinical isolates of *P. falciparum* from Africa and South East Asia. We observed a rarity of copy number variation, with only high frequency for copy number variants correlating to known drug resistance. Additionally, these high frequency duplications occurred mostly in South East Asia, indicating that regional selective pressures have a significant impact on the retention of copy number variants. We identified numerous novel genic copy number variants, one of which was a biallelic duplication of *crt*. This duplication is intriguing as it contains both a chloroquine resistant and a chloroquine sensitive allele. However, overall, this global analysis indicates that copy number variation in the non-subtelomeric regions is a rare event, however drug resistance is a major factor in the frequency and genesis of these copy number variants.

5.2 Genes of extracellularly exposed proteins are highly duplicated in *P. falciparum*

Our segmental duplication analysis highlighted the extent of gene duplication for genes encoding extracellularly exposed proteins. The overrepresentation of antigenic genes under human immune pressure to the exclusion of other factors has resulted in a highly segregated genome, with segmental duplications and the subsequent antigenic factor primarily isolated to the subtelomeric regions, with the rest of the genome containing few segmental duplications. Meanwhile, significant selective pressures for neogenesis and diversification of protein coding sequence is manifested in the significant duplication and high nucleotide diversity of the subtelomeric genes.

This is not a circumstance unique to *Plasmodium falciparum*, but is common in multiple species in the *Plasmodium* clade [156, 157]. However, the exact repertoire of genes encoding extracellularly exposed proteins under human immune pressure that are duplicated differs between species [158]. We can infer that the specific genes are not responsible for the genesis of new duplications, but rather that the subtelomeric regions are regions suitable for promotion of duplication and diversification – allowing for telomeric exchange without disrupting meiotic stability. Selective pressures for increased antigenic diversity have promoted the co-opting of these regions as engines of diversification for particular gene families unique to the selective pressures of each plasmodium species [159–162].

Significant resources and study has been invested into the var gene family due to its relation to severe symptoms of malaria [160, 162]. However, our analysis of segmental duplications indicate that other gene families may have equally important impact on the *P. falciparum* antigenic diversity and cytoadherence capabilities. The *rifin* gene family has a larger contingent of genes than the *var* gene family, and shows similar rates of recombination. Neither gene family showed strong evidence for recent whole gene duplications that has not already undergone significant nucleotide divergence.

In addition, the relative lack of segmental duplications outside the subtelomeric regions provides insight into the purpose of copy number variants outside the subtelomeric regions and the reason for the bias toward a heavily duplicated subtelomere. The lack of duplication of the core genome indicates a strong purifying selection against duplication of genes not encoding extracellularly exposed proteins under human immune pressure, as few have become fixed, resulting in maintenance of gene dosage balance. This indicates that future copy number variants in the core genome are likely to be present as adaptations for gene dosage. In particular, the gene dosage plays a pivotal role in adaptation to drug resistance or metabolic necessity. However, the pressure to maintain high diversity in antigen presenting proteins requires significant recombination and duplication of genes. This has resulted in the parasite co-opting a duplication favorable region, the subtelomeres, to continually duplicate and recombine its large genes familes of genes encoding extracellularly exposed proteins under human immune pressure while preserving the conserved state of the rest of the genome.

5.3 Segmental duplications and high throughput sequencing

Currently, the detection of copy number variation in the population requires a high quality reference genome. However, our analysis of segmental duplications in the genome highlight problematic regions for current means of analysis. It affects *de novo* graph-based assemblies due to collapse of high identity regions into single contigs.

Additionally, high identity duplicated regions are problematic for read alignment because current alignment technologies cannot discern accurate placement of sequence-identical regions, resulting in misplacement of reads during alignment. In addition, high recombination rates within the subtelomeric regions create a unique subtelomeric genome between strains, thereby mitigating the relational accuracy of the reference genome subtelomeric regions to other strains in the species.

Both the effects on *de novo* assembly and accurate read placement have significant consequences for current copy number variation detection. High identity regions prevent complete genome reproduction of *de novo* genome assemblies, reducing the likelihood of detecting deletions and high identity duplications via *de novo* assembly. In addition, the effects on read placement accuracy have effects on copy number variation detection methods that require accurate read placement. Improper read placement due to duplication and recombination will hamper methods involving both discordant reads and read depth. For this reason, our analysis of copy number variation in *P. falciparum* ignored nearly all regions we identified through our segmental duplication analysis. The high rate of recombination and copy number variation in these regions prevents simply adjusting for nucleotide diversity and will require accurate assembly. Future improvements to sequencing technology and *de novo* assembly will be required for accurate copy number variation assessment in these regions.

5.4 Rapid and efficient copy number variation detection

Our novel computational method for copy number variation detection from next gen sequencing is the first study of copy number variation in a large natural population. Considering our work with segmental duplications and the limitations of the methods being applied, we could not investigate copy number variation in the regions identified as segmental duplications. The long tracts of simple, high AT repeat sequence between genes and the high nucleotide diversity of genes in the segmental duplications make accurate read placement and *de novo* assembly of the region difficult. However, the majority of the genome is still available to investigate. The ability to identify copy number variants in a natural population is a major step forward in understanding the role of copy number variation in *P. falciparum*.

The method we designed is rapid and specific. It also has a simple requirement of a single next gene sequencing library. Though it has only been tested with Illumina sequencing at various read sizes, theoretically it should work with a number of next gen sequencing platforms. Additionally, the program has modest read depth requirements – working at 10x coverage. The low read depth requirement allows multiplexing of multiple samples, further reducing the cost of surveying copy number variation across a population. The rapidly plunging cost of next gen sequencing and the modest requirements of our copy number detection methodology bodes well for the future of copy number variation surveillance around the globe. We now have the capability to take census of current copy number variation in *P. falciparum* and identify potential adaptations via copy number variation from genome-wide scans, reducing the necessity of targeted copy number variation detection methods.

5.5 Extent of copy number variation in P. falciparum

From our study of lab strains alone, we found a significant fraction of the copy number variants were correlated with drug resistance – *mdr1*, *gch*, and *dhfr*. However, our investigation into a global population of *P. falciparum* identified that drug resistance still plays an important role in copy number variation. While our our initial foray into a natural population produced results that hinted at a staid and relatively copy number invariant genome compared to many other organisms [5], the expansion of scope to multiple other countries and geographic regions reinforced the idea that copy number variation may be rare. From 610 samples across the globe, only 68 copy number variants were identified. When filtering for solely genic copy number variants, only 33 copy number variants incorporated the entirety of a gene. However, the frequency of individual instances of copy number variation was dominated by *mdr1* and *gch* duplications. Furthermore, the frequency of these two copy number variants are regionally specific, with nearly all identified in strains from South East Asia. Potentially, this could highlight major differences in drug pressure between the two regions, in that South East Asia has much higher drug pressure to have such high representation of drug resistance related copy number variants.

The extent of copy number variation in P. falciparum is significantly impacted by drug pressure. The frequency of copy number variation remained low (<10%) unless confronted with strong drug pressure. Nearly 50% of samples in South East Asia had at least one copy number variation, and that copy number variation was predominantly *mdr1* or *gch*. However, there was little overlap between the two copy number variants. This indicates that there is strong selective pressure to maintain the gene dosage balance of the majority of the core genome, i.e. the genomic regions aside from the telomeres and subtelomeres. *S. cerevisiae* display a lack of compensation for increased protein levels from gene duplication [163]. Comparatively, *P. falciparum*'s lack of duplication indicate that nutrient scarcity may be a strong driving force for selection against gene duplication. Exacerbating this selection against increased protein levels is the relative lack of epigenetic regulation. Except for the subtelomeric regions, most of the genome is euchromatin [164–166]. This results in a lack of specific regulation of gene expression by epigenetic factors. This lack of epigenetic control and low tolerance for protein level increases results in a strong purifying selection against genic duplication.

5.6 Future direction for the investigation of duplications in P. falciparum

The stark contrast between the prolific duplication and deletion of the subtelomeric regions of the genome and the paucity of duplications and deletions in the rest of the genome indicates a strong purifying selection on the non-subtelomeric regions of the genome. It highlights that the observed copy number variation within the core genome likely has significant functional consequence important to the parasite fitness. This study identifies numerous such genes whose copy number variation has potential functional consequence. Future research will hopefully identify the specific gene within any copy number variant as the gene under selection. From there, studies to understand the potential functional significance will provide insight to the effects of these copy number variants – affecting competitive growth advantages over other parasites, adaptation to human genetics, or drug resistance.

One copy number variant in particular will require significant research into its effects – the duplication of *crt*. This duplication was bialleic for both chloroquine resistant and chloroquine sensitive alleles of *crt*. The duplication potentially confers significant compensatory adaptation to drug resistance. While chloroquine resistant alleles confer significant growth disadvantages to the parasite, a duplication with both sensitive and resistant alleles should have both the advantage of chloroquine resistance and mitigated growth inhibition from said chloroquine resistant allele. The independent verification of its existence in the population would prove the sensitivity and power of our computational methods. Also, testing the hypothetical benefits of the biallelic duplication *in vitro* for IC50 drug assays and growth rate assays would verify the potential advantages of the duplication. This has effects on future anti-malarial protocols, potentially eliminating the viability of re-introduction of chloroquine as this duplication could rapidly sweep across the population. It also has serious ramification in current and future drug design strategies, as relying on the functional constraints for mutation of the *crt* gene may be a strategic dead end.

With the high throughput sequence data and our identification of copy number variants, future studies can identify the genesis of copy number variants. Multiple copy number variants showed great variability in breakpoint and copy number. Study of the haplotype and tracing historical recombination could illuminate the origin of these copy number variants. Currently, it is uncertain whether many copy number variants had multiple independent genesis events or whether the multiple isoforms of a copy number variation resulted from a single progenitor. A comprehensive analysis of haplotype between copy number variant isoforms, both in break point and copy number, could elucidate the genesis of various copy number variant isoforms. In addition, our method can be used as a survey tool to rapidly and efficiently detect copy number variation in a population of P. falciparum. This allows study of demographic evolution of copy number variation in the population and the identification of novel copy number variants as they arise. Both the understanding of the genesis of copy number variants, their proliferation, and consistent survey of populations for copy number variation can help monitor for emerging drug resistance and inform best practices for anti-malarial regimens.

APPENDIX

Appendix Table 1 Identified Copy Number Variants

Chrom	Position	Length	Copy number	Sampling site	Sample ID
chr01	238445 - 248349	9,905	2.16	Gambia	374
chr01	549500 - 640851	91,352	0.00	Senegal	SenP27.02
chr01	563800 - 640851	77,052	0.00	Senegal	SenT10.04D10
chr01	551800 - 554000	2,201	0.00	Senegal	SenT32.09
chr02	220981 - 238374	17,394	2.02	Cambodia (Ratanakiri)	286
chr02	355068 - 356081	1,014	2.02	Thailand	796
chr02	555458 - 662357	106,900	1.66	Gambia	403
chr02	671115 - 672530	1,416	2.01	Thailand	766
chr02	835300 - 947102	111,803	0.00	Senegal	SenT090.09
chr02	863500 - 947102	83,603	0.00	Senegal	SenT135.09
chr03	128244 - 134634	6,391	0.13	Ghana	477
chr03	251527 - 252914	1,388	3.84	Thailand	766
chr03	663859 - 664651	793	2.21	Gambia	374
chr03	984410 - 985091	682	2.61	Cambodia (Pursat)	136
chr04	250530 - 349000	98,471	1.95	Gambia	407
chr05	388213 - 390121	1,909	0.41	Ghana	477
chr05	441610 - 461650	20,041	1.48	Cambodia (Pursat)	218
chr05	446993 - 462053	15,061	1.56	Ghana	574
chr05	448740 - 460238	11,499	1.70	Cambodia (Ratanakiri)	248
chr05	505687 - 506365	679	2.28	Ghana	477
chr05	670907 - 770083	99,177	1.83	Gambia	407
chr05	870787 - 978366	107,580	3.58	Cambodia (Pursat)	122

chr05	943270 - 970993	27,724	2.03	Cambodia (Pursat)	126
chr05	943344 - 964658	21,315	2.11	Cambodia (Pursat)	168
chr05	944490 - 973296	28,807	2.07	Cambodia (Pursat)	178
chr05	944894 - 964526	19,633	3.15	Thailand	811
chr05	945099 - 964624	19,526	2.00	Cambodia (Pursat)	114
chr05	946149 - 973288	27,140	1.91	Cambodia (Pursat)	150
chr05	946483 - 967340	20,858	3.08	Cambodia (Pursat)	127
chr05	946485 - 964658	18,174	3.11	Cambodia (Pailin)	56
chr05	946487 - 964647	18,161	2.94	Cambodia (Pailin)	63
chr05	946488 - 964640	18,153	3.09	Cambodia (Pailin)	79
chr05	946488 - 964652	18,165	3.07	Cambodia (Pursat)	207
chr05	946558 - 964629	18,072	2.03	Cambodia (Pursat)	101
chr05	947121 - 963187	16,067	2.10	Thailand	778
chr05	947579 - 962565	14,987	2.02	Thailand	774
chr05	947636 - 962557	14,922	2.82	Thailand	721
chr05	947773 - 969928	22,156	3.23	Cambodia (Pursat)	153
chr05	947774 - 969926	22,153	2.09	Cambodia (Pursat)	131
chr05	947781 - 969907	22,127	1.91	Cambodia (Tasanh)	313
chr05	947785 - 969917	22,133	2.01	Cambodia (Tasanh)	337
chr05	948114 - 970296	22,183	1.47	Cambodia (Tasanh)	315
chr05	950684 - 972006	21,323	3.12	Thailand	759
chr05	952701 - 970350	17,650	2.13	Cambodia (Pursat)	118
chr05	953745 - 970361	16,617	3.04	Thailand	769
chr05	953747 - 970349	16,603	2.12	Thailand	764
chr05	953794 - 972127	18,334	3.82	Thailand	756
chr05	953827 - 970299	16,473	2.94	Cambodia (Pursat)	116
chr05	953828 - 970293	16,466	3.14	Cambodia (Pursat)	112
chr05	953843 - 970259	16,417	2.33	Cambodia (Pursat)	109
chr05	953870 - 970299	16,430	3.04	Cambodia (Pursat)	102
chr05	953871 - 970285	16,415	2.62	Cambodia (Pursat)	103
chr05	953894 - 970289	16,396	2.69	Thailand	812
chr05	962655 - 963279	625	2.41	Kenya	ERS010407
chr06	485029 - 485675	647	2.06	Gambia	PA0068-C
chr06	492315 - 503056	10,742	1.62	Thailand	778
chr06	849764 - 850377	614	2.40	Mali	677
chr06	1117358 - 1119225	1,868	2.04	Senegal	SenT033.09

chr06	1117364 - 1119201	1,838	2.11	Guinea	PA0186-C
chr06	1117368 - 1119193	1,826	2.03	Senegal	SenT086.09
chr06	1117369 - 1119203	1,835	2.40	Guinea	PA0245-C
chr06	1117371 - 1119206	1,836	2.89	Guinea	PA0208-C
chr06	1117373 - 1119156	1,784	1.94	Guinea	PA0140-C
chr06	1117376 - 1119225	1,850	2.13	Senegal	SenT148.09
chr06	1117379 - 1119205	1,827	3.07	Guinea	PA0180-C
chr06	1117380 - 1119204	1,825	2.02	Guinea	PA0234-C
chr06	1117381 - 1119204	1,824	1.97	Senegal	SenT142.09
chr06	1117385 - 1119187	1,803	2.23	Guinea	PA0157-C
chr06	1117385 - 1119193	1,809	2.20	Guinea	PA0169-C
chr06	1117397 - 1119202	1,806	2.44	Thailand	730
chr06	1117399 - 1119209	1,811	2.35	Cambodia (Ratanakiri)	290
chr06	1117399 - 1119202	1,804	2.83	Gambia	367
chr06	1117399 - 1119211	1,813	1.98	Mali	679
chr06	1117399 - 1119204	1,806	2.09	Cambodia (Pailin)	73
chr06	1117399 - 1119202	1,804	2.06	Thailand	769
chr06	1117399 - 1119186	1,788	3.24	Thailand	765
chr06	1117400 - 1119206	1,807	2.94	Cambodia (Pursat)	202
chr06	1117400 - 1119200	1,801	2.38	Thailand	772
chr06	1117400 - 1119207	1,808	2.19	Cambodia (Ratanakiri)	269
chr06	1117400 - 1119199	1,800	1.92	Thailand	764
chr06	1117400 - 1119016	1,617	2.24	Cambodia (Pursat)	150
chr06	1117400 - 1119198	1,799	2.12	Thailand	767
chr06	1117400 - 1119185	1,786	2.60	Ghana	590
chr06	1117401 - 1119202	1,802	1.95	Cambodia (Pursat)	163
chr06	1117401 - 1119195	1,795	2.27	Cambodia (Pursat)	199
chr06	1117401 - 1119201	1,801	2.05	Cambodia (Pursat)	166
chr06	1117401 - 1119191	1,791	2.00	Cambodia (Tasanh)	315
chr06	1117401 - 1119199	1,799	2.13	Thailand	760
chr06	1117402 - 1119206	1,805	2.02	Cambodia (Pursat)	227
chr06	1117402 - 1119202	1,801	2.01	Cambodia (Pailin)	56
chr06	1117402 - 1119204	1,803	2.60	Cambodia (Pursat)	178
chr06	1117402 - 1119200	1,799	2.09	Cambodia (Pursat)	225
chr06	1117402 - 1119202	1,801	2.25	Thailand	775
chr06	1117403 - 1119187	1,785	2.04	Cambodia (Tasanh)	318
chr06	1117403 - 1119195	1,793	1.99	Thailand	759
chr06	1117403 - 1119204	1,802	2.01	Thailand	768
chr06	1117403 - 1119188	1,786	2.02	Cambodia (Tasanh)	316
chr06	1117404 - 1119201	1,798	2.30	Thailand	763

chr06	1117404 - 1119191	1,788	2.09	Cambodia (Tasanh)	328
chr06	1117404 - 1119200	1,797	2.24	Cambodia (Pursat)	204
chr06	1117404 - 1119203	1,800	2.37	Cambodia (Pursat)	203
chr06	1117404 - 1119187	1,784	2.01	Cambodia (Tasanh)	335
chr06	1117404 - 1119201	1,798	2.27	Cambodia (Tasanh)	331
chr06	1117405 - 1119200	1,796	1.97	Cambodia (Pursat)	180
chr06	1117405 - 1119198	1,794	1.97	Thailand	753
chr06	1117405 - 1119204	1,800	2.14	Cambodia (Pursat)	169
chr06	1117406 - 1119174	1,769	2.26	Cambodia (Pursat)	205
chr06	1117406 - 1119186	1,781	4.02	Cambodia (Tasanh)	325
chr06	1117407 - 1119199	1,793	2.28	Cambodia (Tasanh)	336
chr06	1117407 - 1119206	1,800	2.21	Cambodia (Pursat)	164
chr06	1117407 - 1119175	1,769	2.36	Cambodia (Tasanh)	314
chr06	1117408 - 1119197	1,790	2.01	Cambodia (Tasanh)	312
chr06	1117412 - 1119128	1,717	2.17	Cambodia (Tasanh)	327
chr06	1117413 - 1119198	1,786	1.99	Thailand	762
chr06	1117418 - 1119019	1,602	2.30	Thailand	778
chr06	1117418 - 1119131	1,714	1.91	Senegal	SenP51.02
chr06	1117424 - 1119151	1,728	2.51	Cambodia (Tasanh)	326
chr06	1117434 - 1119149	1,716	2.26	Guinea	PA0214-C
chr06	1117438 - 1119192	1,755	2.87	Cambodia (Pailin)	76
chr06	1117484 - 1119158	1,675	2.16	Cambodia (Pursat)	105
chr06	1117485 - 1119195	1,711	2.26	Guinea	PA0225-C
chr06	1117486 - 1119130	1,645	2.17	Cambodia (Pursat)	104
chr06	1117487 - 1119162	1,676	3.04	Cambodia (Pursat)	115
chr06	1117488 - 1119162	1,675	3.13	Cambodia (Pursat)	108
chr06	1117489 - 1119152	1,664	2.22	Cambodia (Pursat)	110
chr06	1117598 - 1119095	1,498	2.09	Cambodia (Tasanh)	333
chr06	1117624 - 1119095	1,472	2.12	Ghana	565
chr06	1117746 - 1119167	1,422	2.34	Ghana	525
chr06	1117766 - 1119181	1,416	2.11	Guinea	PA0219-C
chr06	1117827 - 1119194	1,368	2.36	Cambodia (Ratanakiri)	262
chr07	394657 - 421919	27,263	1.91	Burkina Faso	43
chr07	398667 - 421914	23,248	1.91	Guinea	PA0193-C
chr07	398674 - 421912	23,239	1.83	Ghana	590
chr07	398676 - 421917	23,242	2.06	Ghana	520
chr07	398678 - 421912	23,235	1.98	Ghana	489
chr07	398679 - 421905	23,227	2.21	Burkina Faso	47

chr07	779132 - 796453	17,322	1.42	Cambodia (Pursat)	227
chr07	838352 - 888689	50,338	1.83	Gambia	403
chr07	1071568 - 1072232	665	2.21	Kenya	ERS010453
chr08	1314610 - 1315703	1,094	2.01	Gambia	PA0092-C
chr09	321465 - 322688	1,224	2.00	Gambia	372
chr09	1095691 - 1096396	706	2.91	Mali	661
chr09	1196177 - 1272137	75,961	1.43	Gambia	PA0100-C
chr09	1226756 - 1251976	25,221	1.44	Cambodia (Pursat)	226
chr09	1240368 - 1251255	10,888	1.49	Cambodia (Pursat)	218
chr09	1377600 - 1541735	164,136	0.00	Senegal	SenT10.04D10
chr09	1379800 - 1541735	161,936	0.00	Senegal	SenT137.09
chr09	1385500 - 1541735	156,236	0.00	Senegal	SenT002.09
chr09	1387200 - 1541735	154,536	0.00	Senegal	SenT113.09
chr09	1387700 - 1541735	154,036	0.00	Senegal	SenT021.09
chr09	1469100 - 1541735	72,636	0.00	Senegal	SenT15.04
chr09	1384812 - 1385890	1,079	0.22	Cambodia (Pursat)	168
chr09	1397601 - 1399000	1,400	0.18	Senegal	SenP27.02
chr09	1397755 - 1399113	1,359	0.07	Cambodia (Pursat)	101
chr09	1397756 - 1399122	1,367	0.23	Cambodia (Pursat)	105
chr09	1397758 - 1399124	1,367	0.24	Cambodia (Pursat)	104
chr09	1397762 - 1399109	1,348	0.31	Cambodia (Pursat)	110
chr09	1397763 - 1399115	1,353	0.16	Cambodia (Pursat)	117
chr09	1397764 - 1399111	1,348	0.13	Cambodia (Pursat)	113
chr09	1397764 - 1399114	1,351	0.19	Cambodia (Pursat)	116
chr09	1397768 - 1399118	1,351	0.25	Cambodia (Pursat)	115
chr09	1397790 - 1399132	1,343	0.02	Thailand	807
chr09	1397792 - 1399116	1,325	0.12	Cambodia (Pursat)	108
chr09	1397796 - 1399114	1,319	0.21	Cambodia (Pursat)	111
chr09	1397801 - 1399100	1,300	0.22	Senegal	SenT15.04
chr09	1397812 - 1399134	1,323	0.15	Thailand	796
chr09	1397824 - 1399117	1,294	0.14	Cambodia (Pursat)	109
chr09	1397839 - 1399122	1,284	0.18	Cambodia (Pursat)	107

chr09	1397843 - 1399125	1,283	0.23	Thailand	803
chr09	1397901 - 1399000	1,100	0.11	Senegal	SenT127.09
chr09	1397901 - 1399000	1,100	0.17	Senegal	SenT231.08
chr09	1397901 - 1399000	1,100	0.13	Senegal	SenT74.08
chr09	1397901 - 1399000	1,100	0.09	Senegal	SenT135.09
chr09	1397901 - 1399200	1,300	0.13	Senegal	SenT123.09
chr09	1397901 - 1399200	1,300	0.18	Senegal	SenT130.09
chr09	1397901 - 1399200	1,300	0.20	Senegal	SenT032.09
chr09	1397901 - 1399300	1,400	0.20	Senegal	SenT230.08
chr09	1397901 - 1399300	1,400	0.22	Senegal	SenV42.05
chr09	1407201 - 1408300	1,100	0.40	Senegal	SenT128.09
chr10	903354 - 904074	721	2.52	Cambodia (Pursat)	159
chr10	1170378 - 1174324	3,947	2.43	Gambia	374
chr10	1376686 - 1440190	63,505	3.04	Senegal	SenT151.09
chr10	1424265 - 1425505	1,241	2.07	Gambia	PA0071-C
chr11	461088 - 493610	32,523	0.16	Gambia	403
chr11	687757 - 688583	827	2.08	Mali	678
chr11	814857 - 868533	53,677	1.41	Gambia	PA0101-C
chr11	821011 - 845503	24,493	1.60	Mali	677
chr11	1054820 - 1059223	4,404	0.16	Burkina Faso	44
chr11	1054851 - 1059280	4,430	0.43	Mali	666
chr11	1054935 - 1059247	4,313	0.44	Mali	676
chr12	461869 - 462612	744	2.37	Kenya	ERS010455
chr12	822219 - 830842	8,624	3.04	Cambodia (Ratanakiri)	290
chr12	903997 - 1023825	119,829	1.81	Cambodia (Pursat)	239
chr12	934736 - 975016	40,281	1.59	Cambodia (Pursat)	121
chr12	936352 - 936993	642	7.90	Kenya	ERS017455
chr12	941097 - 980769	39,673	4.29	Cambodia (Pailin)	80
chr12	967737 - 978066	10,330	1.91	Cambodia (Pursat)	171
chr12	968639 - 978077	9,439	2.69	Thailand	763

chr12	968640 - 978085	9,446	10.50	Thailand	711
chr12	968643 - 980646	12,004	2.91	Cambodia (Pursat)	168
chr12	968643 - 978084	9,442	5.80	Thailand	757
chr12	968646 - 978080	9,435	1.91	Thailand	767
chr12	968647 - 978079	9,433	1.93	Cambodia (Pursat)	167
chr12	968649 - 978078	9,430	1.73	Thailand	755
chr12	968649 - 978073	9,425	3.48	Thailand	766
chr12	968650 - 978077	9,428	1.81	Cambodia (Pursat)	231
chr12	968651 - 978079	9,429	1.67	Cambodia (Pursat)	208
chr12	968656 - 978087	9,432	2.20	Thailand	710
chr12	968659 - 978078	9,420	1.69	Cambodia (Pursat)	159
chr12	968661 - 978072	9,412	1.94	Cambodia (Pursat)	211
chr12	968672 - 976316	7,645	6.09	Thailand	762
chr12	968676 - 976313	7,638	2.89	Thailand	760
chr12	968677 - 978083	9,407	1.69	Cambodia (Pursat)	216
chr12	968678 - 976310	7,633	1.95	Thailand	753
chr12	968678 - 978079	9,402	1.98	Cambodia (Pursat)	165
chr12	968693 - 978077	9,385	1.85	Cambodia (Pursat)	244
chr12	968696 - 978076	9,381	2.91	Thailand	730
chr12	968698 - 976310	7,613	2.97	Ghana	478
chr12	968699 - 978077	9,379	4.58	Thailand	716
chr12	968701 - 976319	7,619	6.18	Thailand	761
chr12	968702 - 978079	9,378	1.95	Cambodia (Pursat)	169
chr12	968704 - 978076	9,373	1.82	Cambodia (Pursat)	197
chr12	968709 - 978076	9,368	2.52	Cambodia (Pailin)	76
chr12	968711 - 978073	9,363	2.03	Thailand	733
chr12	968720 - 976305	7,586	2.22	Thailand	714
chr12	968729 - 978068	9,340	2.14	Cambodia (Pursat)	120
chr12	968730 - 976273	7,544	1.98	Ghana	525
chr12	968773 - 978045	9,273	3.00	Thailand	805
chr12	968777 - 978045	9,269	2.69	Cambodia (Pursat)	111
chr12	968779 - 978041	9,263	2.77	Thailand	809
chr12	968779 - 976281	7,503	3.91	Thailand	803
chr12	968780 - 978042	9,263	3.26	Cambodia (Pursat)	104
chr12	968794 - 978040	9,247	2.85	Thailand	810
chr12	968884 - 978063	9,180	3.16	Thailand	720
chr12	973602 - 974374	773	2.24	Guinea	PA0177-C
chr12	973608 - 976162	2,555	2.23	Cambodia (Pursat)	152
chr12	973609 - 976160	2,552	2.77	Thailand	769
chr12	973611 - 976157	2,547	2.05	Ghana	541

chr12	973612 - 976139	2,528	8.84	Thailand	778
chr12	973612 - 976162	2,551	3.01	Thailand	764
chr12	973613 - 976144	2,532	9.25	Thailand	774
chr12	973613 - 975920	2,308	2.23	Ghana	543
chr12	973615 - 974272	658	2.21	Kenya	ERS010641
chr12	973615 - 976154	2,540	2.02	Ghana	540
chr12	973621 - 976147	2,527	1.97	Ghana	565
chr12	973625 - 974362	738	1.94	Guinea	PA0201-C
chr12	973628 - 976117	2,490	2.45	Thailand	759
chr12	973639 - 974364	726	2.07	Guinea	PA0158-C
chr12	973645 - 976138	2,494	2.52	Thailand	775
chr12	1073984 - 1075014	1,031	1.99	Cambodia (Pailin)	55
chr12	1267185 - 1309576	42,392	1.57	Gambia	PA0100-C
chr12	1268398 - 1296032	27,635	1.47	Gambia	PA0101-C
chr12	1274036 - 1293832	19,797	1.66	Cambodia (Pursat)	121
chr12	1542242 - 1542923	682	2.06	Kenya	ERS010452
chr12	1657036 - 1680085	23,050	1.59	Ghana	477
chr12	1767965 - 1769423	1,459	2.26	Ghana	477
chr13	452231 - 453006	776	4.03	Kenya	ERS017455
chr13	965987 - 969833	3,847	0.24	Ghana	459
chr13	1068874 - 1084103	15,230	1.45	Cambodia (Pursat)	221
chr13	1154958 - 1183794	28,837	1.46	Cambodia (Pursat)	221
chr13	1162411 - 1181184	18,774	1.62	Cambodia (Pursat)	159
chr13	1200973 - 1201602	630	5.92	Kenya	ERS017455
chr13	1212379 - 1264030	51,652	1.90	Gambia	407
chr13	1423943 - 1449811	25,869	2.06	Kenya	ERS010454
chr13	1430177 - 1449822	19,646	2.23	Kenya	ERS010455
chr13	1428865 - 1429987	1,123	0.34	Gambia	PA0068-C

chr13	1428908 - 1429972	1,065	0.32	Mali	688
chr13	1428953 - 1429975	1,023	0.31	Ghana	515
chr13	1429053 - 1429993	941	0.27	Ghana	575
chr13	1429079 - 1429975	897	0.34	Burkina Faso	13
chr13	2623768 - 2624384	617	13.80	Kenya	ERS017455
chr14	95794 - 109934	14,141	2.12	Cambodia (Pursat)	150
chr14	282835 - 300646	17,812	1.86	Cambodia (Pursat)	165
chr14	282842 - 300646	17,805	1.88	Cambodia (Pursat)	169
chr14	289091 - 298886	9,796	3.12	Cambodia (Pursat)	198
chr14	289396 - 298896	9,501	2.32	Cambodia (Pursat)	129
chr14	289401 - 298894	9,494	3.58	Cambodia (Pursat)	194
chr14	289401 - 298889	9,489	1.99	Cambodia (Pursat)	199
chr14	289402 - 298900	9,499	1.68	Cambodia (Pursat)	180
chr14	289402 - 298898	9,497	1.68	Cambodia (Pursat)	218
chr14	289405 - 298895	9,491	1.74	Cambodia (Pursat)	227
chr14	289407 - 298885	9,479	2.62	Cambodia (Tasanh)	320
chr14	289407 - 298894	9,488	1.74	Cambodia (Pursat)	215
chr14	289408 - 298875	9,468	2.06	Cambodia (Tasanh)	318
chr14	289409 - 298886	9,478	1.71	Cambodia (Pursat)	173
chr14	289409 - 298869	9,461	2.01	Cambodia (Tasanh)	336
chr14	289411 - 298888	9,478	1.83	Cambodia (Pursat)	163
chr14	289417 - 298882	9,466	1.83	Cambodia (Tasanh)	317
chr14	289421 - 298893	9,473	1.81	Cambodia (Tasanh)	325
chr14	367404 - 368300	897	1.94	Kenya	ERS017458
chr14	589978 - 592404	2,427	2.61	Guinea	PA0200-C
chr14	695480 - 704797	9,318	1.55	Thailand	775
chr14	1099777 - 1101133	1,357	0.45	Mali	677
chr14	1152175 - 1259839	107,665	1.73	Gambia	403
chr14	1185390 - 1186146	757	2.37	Gambia	382
chr14	1185425 - 1186197	773	2.94	Burkina Faso	7

chr14	1733015 - 1748645	15,631	1.99	Gambia	374
chr14	1734101 - 1735000	900	0.24	Senegal	SenT021.09
chr14	2020371 - 2021382	1,012	2.74	Kenya	ERS010454
chr14	2257914 - 2259309	1,396	2.78	Cambodia (Pursat)	125
chr14	2620036 - 2620833	798	2.14	Gambia	374
chr14	2698050 - 2699156	1,107	2.44	Thailand	766

REFERENCES

1. Li W, Gu Z, Cavalcanti ARO, Nekrutenko A: **Detection of gene duplications and block duplications in eukaryotic genomes.** *J. Struct. Funct. Genomics* 2003, **3**:27–34.

2. Glasauer SMK, Neuhauss SCF: **Whole-genome duplication in teleost fishes and its evolutionary consequences.** *Mol. Genet. Genomics* 2014, **289**:1045–60.

3. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290**:1151–5.

4. Ohno S: *Evolution by gene duplication*. Springer-Verlag; 1970.

5. Katju V, Bergthorsson U: **Copy-number changes in evolution: rates, fitness effects and adaptive significance.** *Front. Genet.* 2013, **4**:273.

6. Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Pääbo S, Rocchi M, Eichler EE: **A genome-wide comparison of recent chimpanzee and human segmental duplications.** *Nature* 2005, **437**:88–93.

7. Konrad A, Teufel AI, Grahnen J a, Liberles D a: **Toward a general model for the evolutionary dynamics of gene duplicates.** *Genome Biol. Evol.* 2011, **3**:1197–209.

8. Andersson DI, Hughes D: **Gene amplification and adaptive evolution in**

bacteria. *Annu. Rev. Genet.* 2009, **43**:167–95.

9. Duvernay C, Coulange L, Dutilh B, Dubois V, Quentin C, Arpin C: **Duplication of the chromosomal blaSHV-11 gene in a clinical hypermutable strain of Klebsiella pneumoniae.** *Microbiology* 2011, **157**:496–503.

10. Guillon JM, Mechulam Y, Schmitter JM, Blanquet S, Fayat G: **Disruption of the gene for Met-tRNA(fMet) formyltransferase severely impairs growth of Escherichia coli.** *J. Bacteriol.* 1992, **174**:4294–301.

11. Nilsson AI, Zorzet A, Kanth A, Dahlström S, Berg OG, Andersson DI: **Reducing the fitness cost of antibiotic resistance by amplification of initiator tRNA genes.** *Proc. Natl. Acad. Sci. U. S. A.* 2006, **103**:6976–81.

12. Watterson GA: **On the time for gene silencing at duplicate Loci.** *Genetics* 1983, **105**:745–66.

13. Wagner A: **Asymmetric functional divergence of duplicate genes in yeast.** *Mol. Biol. Evol.* 2002, **19**:1760–8.

14. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I: **Multicopy suppression underpins metabolic evolvability.** *Mol. Biol. Evol.* 2007, **24**:2716–22.

15. Pich I Roselló O, Kondrashov F a: **Long-term asymmetrical acceleration of protein evolution after gene duplication.** *Genome Biol. Evol.* 2014, **6**:1949–55.

16. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations.** *Genetics* 1999, **151**:1531–45.

17. Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L: **Gene families: the taxonomy of protein paralogs and chimeras.** *Science* 1997, **278**:609–14.

18. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154**:459–73.

19. Assis R, Bachtrog D: **Neofunctionalization of young duplicate genes in Drosophila.** *Proc. Natl. Acad. Sci. U. S. A.* 2013, **110**:17409–14.

20. Pegueroles C, Laurie S, Albà MM: **Accelerated evolution after gene duplication: a time-dependent process affecting just one copy.** *Mol. Biol. Evol.* 2013, **30**:1830–42.

21. Rensing S a: **Gene duplication as a driver of plant morphogenetic evolution.** *Curr. Opin. Plant Biol.* 2014, **17**:43–8.

22. Fuchs T, Glusman G, Horn-Saban S, Lancet D, Pilpel Y: **The human olfactory subgenome: from sequence to structure and evolution**. *Hum. Genet.* 2001, **108**:1–13.

23. Nguyen D-Q, Webber C, Ponting CP: **Bias of selection on human copy-**

number variants. *PLoS Genet.* 2006, **2**:e20.

24. Kraemer SM, Smith JD: **A family affair: var genes, PfEMP1 binding, and malaria disease.** *Curr. Opin. Microbiol.* 2006, **9**:374–80.

25. DeBarry JD, Kissinger JC: **A survey of innovation through duplication in the reduced genomes of twelve parasites.** *PLoS One* 2014, **9**:e99213.

26. Marques-Bonet T, Kidd JM, Ventura M, Graves T a, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton L a, Alkan C, Aksay G, Girirajan S, Siswara P, Chen L, Cardone MF, Navarro A, Mardis ER, Wilson RK, Eichler EE: **A burst of segmental duplications in the genome of the African great ape ancestor.** *Nature* 2009, **457**:877–81.

27. Lenormand T, Guillemaud T, Bourguet D, Raymond M: **Appearance and sweep of a gene duplication: adaptive response and potential for new functions in the mosquito Culex pipiens**. *Evolution (N. Y).* 1998, **52**:1705–1712.

28. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark R a, Anderson S a, O’connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK: **The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility.** *Science* 2005, **307**:1434–40.

29. Wickstead B, Ersfeld K, Gull K: **The small chromosomes of Trypanosoma**

brucei involved in antigenic variation are constructed around repetitive palindromes. *Genome Res.* 2004, **14**:1014–24.

30. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays.** *Nat. Genet.* 1999, **23**:41–6.

31. Bailey J a, Yavor AM, Massa HF, Trask BJ, Eichler EE: **Segmental duplications: organization and impact within the current human genome project assembly.** *Genome Res.* 2001, **11**:1005–17.

32. Lander ES, Linton LM, Birren B, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860–921.

33. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan M-S, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DMA, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419**:498–511.

34. McGrath CL, Gout J, Johri P, Doak TG, Lynch M: **Differential retention and**

divergent resolution of duplicate genes following whole-genome duplication. *Genome Res.* 2014, **24**:1665–75.

35. Hughes AL, Verra F: **Malaria parasite sequences from chimpanzee support the co-speciation hypothesis for the origin of virulent human malaria (Plasmodium falciparum).** *Mol. Phylogenet. Evol.* 2010, **57**:135–43.

36. Hughes GM, Teeling EC, Higgins DG: **Loss of olfactory receptor function in hominin evolution.** *PLoS One* 2014, **9**:e84714.

37. Young JM, Trask BJ: **The sense of smell: genomics of vertebrate odorant receptors.** *Hum. Mol. Genet.* 2002, **11**:1153–60.

38. Giannuzzi G, D'Addabbo P, Gasparro M, Martinelli M, Carelli FN, Antonacci D, Ventura M: **Analysis of high-identity segmental duplications in the grapevine genome.** *BMC Genomics* 2011, **12**:436.

39. She X, Cheng Z, Zöllner S, Church DM, Eichler EE: **Mouse segmental duplication and copy number variation.** *Nat. Genet.* 2008, **40**:909–14.

40. Li W, Liu W, Wei H, He Q, Chen J, Zhang B, Zhu S: **Species-specific expansion and molecular evolution of the 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGR) gene family in plants.** *PLoS One* 2014, **9**:e94172.

41. Zhang Q, Su B: **Evolutionary origin and human-specific expansion of a**

cancer/testis antigen gene family. *Mol. Biol. Evol.* 2014, **31**:2365–75.

42. Cuadrat RRC, da Serra Cruz SM, Tschoeke DA, Silva E, Tosta F, Jucá H, Jardim R, Campos MLM, Mattoso M, Dávila AMR: **An orthology-based analysis of pathogenic protozoa impacting global health: an improved comparative genomics approach with prokaryotes and model eukaryote orthologs.** *OMICS* 2014, **18**:524–38.

43. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll S a, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C: **Copy number variation: new insights in genome diversity.** *Genome Res.* 2006, **16**:949–61.

44. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z: **Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives.** *BMC Bioinformatics* 2013, **14 Suppl 1**:S1.

45. Pinto D, Marshall C, Feuk L, Scherer SW: **Copy-number variation in control population cohorts.** *Hum. Mol. Genet.* 2007, **16 Spec No**:R168–73.

46. Bailey J a, Church DM, Ventura M, Rocchi M, Eichler EE: **Analysis of segmental duplications and genome assembly in the mouse.** *Genome Res.* 2004, **14**:789–801.

47. Saitoh S, Aoyama H, Akutsu M, Nakano K, Shinzato N, Matsui T: **Genomic sequencing-based detection of large deletions in Rhodococcus rhodochrous strain B-276.** *J. Biosci. Bioeng.* 2013, **116**:309–12.

48. Gazave E, Darré F, Morcillo-Suarez C, Petit-Marty N, Carreño A, Marigorta UM, Ryder OA, Blancher A, Rocchi M, Bosch E, Baker C, Marquès-Bonet T, Eichler EE, Navarro A: **Copy number variation analysis in the great apes reveals species-specific patterns of structural variation.** *Genome Res.* 2011, **21**:1626–39.

49. Conrad B, Antonarakis SE: **Gene duplication: a drive for phenotypic diversity and cause of human disease.** *Annu. Rev. Genomics Hum. Genet.* 2007, **8**:17–35.

50. Sidhu ABS, Uhlemann A-C, Valderramos SG, Valderramos J-C, Krishna S, Fidock D a: **Decreasing pfmdr1 copy number in plasmodium falciparum malaria heightens susceptibility to mefloquine, lumefantrine, halofantrine, quinine, and artemisinin.** *J. Infect. Dis.* 2006, **194**:528–35.

51. Heinberg A, Siu E, Stern C, Lawrence E a, Ferdig MT, Deitsch KW, Kirkman L a: **Direct evidence for the adaptive role of copy number variation on antifolate susceptibility in Plasmodium falciparum.** *Mol. Microbiol.* 2013, **88**:702–12.

52. Nair S, Nkhoma S, Nosten F, Mayxay M, French N, Whitworth J, Anderson T: **Genetic changes during laboratory propagation: copy number At the reticulocyte-binding protein 1 locus of Plasmodium falciparum.** *Mol. Biochem. Parasitol.* 2010, **172**:145–8.

53. World Health Organization: *World Malaria Report: 2010*. Geneva,: 2010:94.

54. Kioko UM: **Economic Burden of Malaria on Subsistence Crop Production in Kenya**. 2013, **1**:1–20.

55. Sinden RE, Canning EU, Bray RS, Smalley ME: **Gametocyte and gamete development in Plasmodium falciparum.** *Proc. R. Soc. London. Ser. B, Biol. Sci.* 1978, **201**:375–99.

56. Gerald N, Mahajan B, Kumar S: **Mitosis in the Human Malaria Parasite Plasmodium falciparum.** *Eukaryot. Cell* 2011.

57. Garcia JE, Puentes A, Patarroyo ME: **Developmental biology of sporozoite-host interactions in Plasmodium falciparum malaria: implications for vaccine design.** *Clin. Microbiol. Rev.* 2006, **19**:686–707.

58. Beier JC: **Malaria parasite development in mosquitoes.** *Annu. Rev. Entomol.* 1998, **43**:519–43.

59. Silvie O, Franetich J-F, Rénia L, Mazier D: **Malaria sporozoite: migrating for a living.** *Trends Mol. Med.* 2004, **10**:97–100; discussion 100–1.

60. Bannister LH, Hopkins JM, Fowler RE, Krishna S, Mitchell GH: **A Brief Illustrated Guide to the Ultrastructure of Plasmodium falciparum Asexual Blood Stages**. 2000, **4758**:427–433.

61. Margos G, Bannister LH, Dluzewski a. R, Hopkins J, Williams IT, Mitchell

GH: **Correlation of structural development and differential expression of invasion-related molecules in schizonts of Plasmodium falciparum**. *Parasitology* 2004, **129**:273–287.

62. Stresman GH, Stevenson JC, Ngwu N, Marube E, Owaga C, Drakeley C, Bousema T, Cox J: **High levels of asymptomatic and subpatent Plasmodium falciparum parasite carriage at health facilities in an area of heterogeneous malaria transmission intensity in the Kenyan highlands.** *Am. J. Trop. Med. Hyg.* 2014, **91**:1101–8.

63. Ashley E a, White NJ: **The duration of Plasmodium falciparum infections.** *Malar. J.* 2014, **13**:500.

64. Scherf A, Lopez-Rubio JJ, Riviere L: **Antigenic variation in Plasmodium falciparum.** *Annu. Rev. Microbiol.* 2008, **62**:445–70.

65. Ibraheem ZO, Abd Majid R, Noor SM, Sedik HM, Basir R: **Role of Different Pfcrt and Pfmdr-1 Mutations in Conferring Resistance to Antimalaria Drugs in Plasmodium falciparum.** *Malar. Res. Treat.* 2014, **2014**:950424.

66. Heinberg A, Kirkman L: **The molecular basis of antifolate resistance in Plasmodium falciparum: looking beyond point mutations.** *Ann. N. Y. Acad. Sci.* 2015:1–9.

67. Straimer J, Gnädig NF, Witkowski B, Amaratunga C, Duru V, Ramadani AP,

Dacheux M, Khim N, Zhang L, Lam S, Gregory PD, Urnov FD, Mercereau-Puijalon O, Benoit-Vical F, Fairhurst RM, Ménard D, Fidock DA: **Drug resistance. K13-propeller mutations confer artemisinin resistance in Plasmodium falciparum clinical isolates.** *Science* 2015, **347**:428–31.

68. Barry AE, Leliwa-Sytek A, Tavul L, Imrie H, Migot-Nabias F, Brown SM, McVean GA V, Day KP: **Population genomics of the immune evasion (var) genes of Plasmodium falciparum.** *PLoS Pathog.* 2007, **3**:e34.

69. Inselburg J, Bzik DJ, Horii T: **Pyrimethamine resistant Plasmodium falciparum: overproduction of dihydrofolate reductase by a gene duplication.** *Mol. Biochem. Parasitol.* 1987, **26**:121–34.

70. Nair S, Miller B, Barends M, Jaidee A, Patel J, Mayxay M, Newton P, Nosten F, Ferdig MT, Anderson TJC: **Adaptive copy number evolution in malaria parasites.** *PLoS Genet.* 2008, **4**:e1000243.

71. Shirley MW, Biggs BA, Forsyth KP, Brown HJ, Thompson JK, Brown G V, Kemp DJ: **Chromosome 9 from independent clones and isolates of Plasmodium falciparum undergoes subtelomeric deletions with similar breakpoints in vitro.** *Mol. Biochem. Parasitol.* 1990, **40**:137–45.

72. Zilversmit MM, Chase EK, Chen DS, Awadalla P, Day KP, McVean G: **Hypervariable antigen genes in malaria have ancient roots.** *BMC Evol. Biol.* 2013,

13:110.

73. Biggs BA, Anders RF, Dillon HE, Davern KM, Martin M, Petersen C, Brown G V: **Adherence of infected erythrocytes to venular endothelium selects for antigenic variants of Plasmodium falciparum.** *J. Immunol.* 1992, **149**:2047–54.

74. David PH, Hommel M, Miller LH, Udeinya IJ, Oligino LD: **Parasite sequestration in Plasmodium falciparum malaria: spleen and antibody modulation of cytoadherence of infected erythrocytes.** *Proc. Natl. Acad. Sci. U. S. A.* 1983, **80**:5075–9.

75. Cheng Q, Cloonan N, Fischer K, Thompson J, Waine G, Lanzer M, Saul A: **stevor and rif are Plasmodium falciparum multicopy gene families which potentially encode variant antigens.** *Mol. Biochem. Parasitol.* 1998, **97**:161–76.

76. Niang M, Bei AK, Madnani KG, Pelly S, Dankwa S, Kanjee U, Gunalan K, Amaladoss A, Yeo KP, Bob NS, Malleret B, Duraisingh MT, Preiser PR: **STEVOR is a Plasmodium falciparum erythrocyte binding protein that mediates merozoite invasion and rosetting.** *Cell Host Microbe* 2014, **16**:81–93.

77. Frank M, Kirkman L, Costantini D, Sanyal S, Lavazec C, Templeton TJ, Deitsch KW: **Frequent recombination events generate diversity within the multi-copy variant antigen gene families of Plasmodium falciparum.** *Int. J. Parasitol.* 2008, **38**:1099–109.

78. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrbass U, Wellems TE, Scherf A: **Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum.** *Nature* 2000, **407**:1018–22.

79. Scherf A, Figueiredo LM, Freitas-Junior LH: **Plasmodium telomeres: a pathogen's perspective.** *Curr. Opin. Microbiol.* 2001, **4**:409–14.

80. Nair S, Nash D, Sudimack D, Jaidee A, Barends M, Uhlemann A-C, Krishna S, Nosten F, Anderson TJC: **Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites.** *Mol. Biol. Evol.* 2007, **24**:562–73.

81. Foote SJ, Thompson JK, Cowman AF, Kemp DJ: **Amplification of the multidrug resistance gene in some chloroquine-resistant isolates of P. falciparum.** *Cell* 1989, **57**:921–30.

82. Wilson CM, Serrano AE, Wasley A, Bogenschutz MP, Shankar AH, Wirth DF: **Amplification of a gene related to mammalian mdr genes in drug-resistant Plasmodium falciparum.** *Science* 1989, **244**:1184–6.

83. Cheeseman IH, Gomez-Escobar N, Carret CK, Ivens A, Stewart LB, Tetteh KK a, Conway DJ: **Gene copy number variation throughout the Plasmodium falciparum genome.** *BMC Genomics* 2009, **10**:353.

84. Abyzov A, Urban AE, Snyder M, Gerstein M: **CNVnator: An approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing.** *Genome Res.* 2011.

85. Sepúlveda N, Campino SG, Assefa SA, Sutherland CJ, Pain A, Clark TG: **A Poisson hierarchical modelling approach to detecting copy number variation in sequence coverage data.** *BMC Genomics* 2013, **14**:128.

86. Manary MJ, Singhakul SS, Flannery EL, Bopp SE, Corey VC, Bright AT, McNamara CW, Walker JR, Winzeler E a: **Identification of pathogen genomic variants through an integrated pipeline.** *BMC Bioinformatics* 2014, **15**:63.

87. Chang H-H, Park DJ, Galinsky KJ, Schaffner SF, Ndiaye D, Ndir O, Mboup S, Wiegand RC, Volkman SK, Sabeti PC, Wirth DF, Neafsey DE, Hartl DL: **Genomic sequencing of Plasmodium falciparum malaria parasites from Senegal reveals the demographic history of the population.** *Mol. Biol. Evol.* 2012, **29**:3427–39.

88. Miotto O, Almagro-Garcia J, Manske M, Macinnis B, Campino S, Rockett K a, Amaratunga C, Lim P, Suon S, Sreng S, Anderson JM, Duong S, Nguon C, Chuor CM, Saunders D, Se Y, Lon C, Fukuda MM, Amenga-Etego L, Hodgson AVO, Asoala V, Imwong M, Takala-Harrison S, Nosten F, Su X-Z, Ringwald P, Ariey F, Dolecek C, Hien TT, Boni MF, Thai CQ, Amambua-Ngwa A, Conway DJ, Djimdé A a, Doumbo OK, Zongo I, Ouedraogo J-B, Alcock D, Drury E, Auburn S, Koch O, Sanders M, Hubbart C, Maslen G, Ruano-Rubio V, Jyothi D, Miles A, O'Brien J, Gamble C, Oyola SO, Rayner

JC, Newbold CI, Berriman M, Spencer CC a, McVean G, Day NP, White NJ, Bethell D, Dondorp AM, Plowe C V, Fairhurst RM, Kwiatkowski DP: **Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia.** *Nat. Genet.* 2013, **45**:648–55.

89. Wendler JP, Okombo J, Amato R, Miotto O, Kiara SM, Mwai L, Pole L, O’Brien J, Manske M, Alcock D, Drury E, Sanders M, Oyola SO, Malangone C, Jyothi D, Miles A, Rockett K a, MacInnis BL, Marsh K, Bejon P, Nzila A, Kwiatkowski DP: **A genome wide association study of Plasmodium falciparum susceptibility to 22 antimalarial drugs in Kenya.** *PLoS One* 2014, **9**:e96486.

90. Amambua-Ngwa A, Tetteh KK a, Manske M, Gomez-Escobar N, Stewart LB, Deerhake ME, Cheeseman IH, Newbold CI, Holder A a, Knuepfer E, Janha O, Jallow M, Campino S, Macinnis B, Kwiatkowski DP, Conway DJ: **Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites.** *PLoS Genet.* 2012, **8**:e1002992.

91. Mobegi VA, Loua KM, Ahouidi AD, Satoguina J, Nwakanma DC, Amambua-Ngwa A, Conway DJ: **Population genetic structure of Plasmodium falciparum across a region of diverse endemicity in West Africa.** *Malar. J.* 2012, **11**:223.

92. Ohno S, Wolf U, Atkin NB: **Evolution from fish to mammals by gene duplication.** *Hereditas* 1968, **59**:169–87.

93. Eichler EE, Sankoff D: **Structural dynamics of eukaryotic chromosome evolution.** *Science* 2003, **301**:793–7.

94. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps K a, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll S a, Altshuler D a, Peiffer D a, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson D a, Mullikin JC, Wilson RK, Bruhn L, Olson M V, Kaul R, Smith DR, Eichler EE: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453**:56–64.

95. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AWC, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME: **Origins and functional impact of copy number variation in the human genome.** *Nature* 2010, **464**:704–12.

96. Hughes MK, Hughes a L: **Evolution of duplicate genes in a tetraploid animal, Xenopus laevis.** *Mol. Biol. Evol.* 1993, **10**:1360–9.

97. Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997, **387**:708–13.

98. Poláková S, Blume C, Zárate JA, Mentel M, Jørck-Ramberg D, Stenderup J, Piskur J: **Formation of new chromosomes as a virulence mechanism in yeast Candida glabrata.** *Proc. Natl. Acad. Sci. U. S. A.* 2009, **106**:2688–93.

99. Jackson AP: **Evolutionary consequences of a large duplication event in Trypanosoma brucei: chromosomes 4 and 8 are partial duplicons.** *BMC Genomics* 2007, **8**:432.

100. Kasahara M: **The 2R hypothesis: an update.** *Curr. Opin. Immunol.* 2007, **19**:547–52.

101. Bailey J a, Eichler EE: **Primate segmental duplications: crucibles of evolution, diversity and disease.** *Nat. Rev. Genet.* 2006, **7**:552–64.

102. Bailey J a, Gu Z, Clark R a, Reinert K, Samonte R V, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: **Recent segmental duplications in the human genome.** *Science* 2002, **297**:1003–7.

103. Hu G, Wang J, Choi J, Jung WH, Liu I, Litvintseva AP, Bicanic T, Aurora R, Mitchell TG, Perfect JR, Kronstad JW: **Variation in chromosome copy number influences the virulence of Cryptococcus neoformans and occurs in isolates from AIDS patients.** *BMC Genomics* 2011, **12**:526.

104. Cowman AF, Galatis D, Thompson JK: **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.* 1994, **91**:1143–7.

105. Potocki L, Chen KS, Park SS, Osterholm DE, Withers M a, Kimonis V, Summers a M, Meschino WS, Anyane-Yeboa K, Kashork CD, Shaffer LG, Lupski JR: **Molecular mechanism for duplication 17p11.2- the homologous recombination reciprocal of the Smith-Magenis microdeletion.** *Nat. Genet.* 2000, **24**:84–7.

106. Shaw CJ, Bi W, Lupski JR: **Genetic proof of unequal meiotic crossovers in reciprocal deletion and duplication of 17p11.2.** *Am. J. Hum. Genet.* 2002, **71**:1072–81.

107. Dumont BL, Eichler EE: **Signals of historical interlocus gene conversion in human segmental duplications.** *PLoS One* 2013, **8**:e75949.

108. Vernick KD, McCutchan TF: **Sequence and structure of a Plasmodium falciparum telomere.** *Mol. Biochem. Parasitol.* 1988, **28**:85–94.

109. De Bruin D, Lanzer M, Ravetch J V: **The polymorphic subtelomeric regions of Plasmodium falciparum chromosomes contain arrays of repetitive sequence elements.** *Proc. Natl. Acad. Sci. U. S. A.* 1994, **91**:619–23.

110. Scotti R, Pace T, Roca L, Frontali C: **Subtelomeric structure of Plasmodium falciparum chromosomes.** *Mem. Inst. Oswaldo Cruz* 1994, **89 Suppl 2**:31–2.

111. Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt J a, Peterson DS,

Ravetch J a, Wellems TE: **The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes.** *Cell* 1995, **82**:89–100.

112. Kyes S a, Rowe J a, Kriek N, Newbold CI: **Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with Plasmodium falciparum.** *Proc. Natl. Acad. Sci. U. S. A.* 1999, **96**:9333–8.

113. Pongponratn E, Riganti M, Punpoowong B, Aikawa M: **Microvascular sequestration of parasitized erythrocytes in human falciparum malaria: a pathological study.** *Am. J. Trop. Med. Hyg.* 1991, **44**:168–75.

114. Fawcett J a, Innan H: **The role of gene conversion in preserving rearrangement hotspots in the human genome.** *Trends Genet.* 2013, **29**:561–8.

115. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res.* 1999, **27**:573–80.

116. **RepeatMasker** [http://repeatmasker.org].

117. Harris R: *Improved pairwise alignment of genomic DNA.* 2007.

118. Chiaromonte F, Yap VB, Miller W: **Scoring pairwise genomic sequence alignments.** *Pac. Symp. Biocomput.* 2002, **126**:115–26.

119. Kozarewa I, Ning Z, Quail M a, Sanders MJ, Berriman M, Turner DJ:

Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* 2009, **6**:291–5.

120. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat. Methods* 2012, **9**:357–9.

121. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078–9.

122. Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, Antonacci F, Ventura M, Prado-Martinez J, Marques-Bonet T, Eichler EE: **Evolution and diversity of copy number variation in the great ape lineage.** *Genome Res.* 2013:1373–1382.

123. Beissbarth T, Speed TP: **GOstat: find statistically overrepresented Gene Ontologies within a group of genes.** *Bioinformatics* 2004, **20**:1464–5.

124. **Parasight** [http://baileylab.umassmed.edu/parasight].

125. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs R a, Eichler EE: **Personalized copy number and segmental duplication maps using next-generation sequencing.** *Nat. Genet.* 2009, **41**:1061–7.

126. Mu J, Awadalla P, Duan J, McGee KM, Keebler J, Seydel K, McVean GAT, Su X: **Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome.** *Nat. Genet.* 2007, **39**:126–30.

127. Waters AP, Syin C, McCutchan TF: **Developmental regulation of stage-specific ribosome populations in Plasmodium.** *Nature* 1989, **342**:438–40.

128. Fiegler H, Redon R, Andrews D, Scott C, Andrews R, Carder C, Clark R, Dovey O, Ellis P, Feuk L, French L, Hunt P, Kalaitzopoulos D, Larkin J, Montgomery L, Perry GH, Plumb BW, Porter K, Rigby RE, Rigler D, Valsesia A, Langford C, Humphray SJ, Scherer SW, Lee C, Hurles ME, Carter NP: **Accurate and reliable high-throughput detection of copy number variation in the human genome.** *Genome Res.* 2006, **16**:1566–74.

129. Anderson TJC, Patel J, Ferdig MT: **Gene copy number and malaria biology.** *Trends Parasitol.* 2009, **25**:336–43.

130. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, O’Brien J, Djimde A, Doumbo O, Zongo I, Ouedraogo J-B, Michon P, Mueller I, Siba P, Nzila A, Borrmann S, Kiara SM, Marsh K, Jiang H, Su X-Z, Amaratunga C, Fairhurst R, Socheat D, Nosten F, Imwong M, White NJ, Sanders M, Anastasi E, Alcock D, Drury E, Oyola S, Quail M a, Turner DJ, Ruano-Rubio V, Jyothi D, Amenga-Etego L, Hubbart C, Jeffreys A, Rowlands K, Sutherland C, Roper C, Mangano V, Modiano D, Tan JC, Ferdig MT, Amambua-Ngwa A, Conway DJ, Takala-Harrison S, Plowe C V, Rayner JC, Rockett

K a, Clark TG, Newbold CI, Berriman M, MacInnis B, Kwiatkowski DP: **Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing.** *Nature* 2012, **487**:375–9.

131. Kemp DJ, Thompson J, Barnes DA, Triglia T, Karamalis F, Petersen C, Brown G V, Day KP: **A chromosome 9 deletion in Plasmodium falciparum results in loss of cytoadherence.** *Mem. Inst. Oswaldo Cruz* 1992, **87 Suppl 3**:85–9.

132. Kidgell C, Volkman SK, Daily J, Borevitz JO, Plouffe D, Zhou Y, Johnson JR, Le Roch K, Sarr O, Ndir O, Mboup S, Batalov S, Wirth DF, Winzeler E a: **A systematic map of genetic variation in Plasmodium falciparum.** *PLoS Pathog.* 2006, **2**:e57.

133. Bourke PF, Holt DC, Sutherland CJ, Currie B, Kemp DJ: **Positional cloning of a sequence from the breakpoint of chromosome 9 commonly associated with the loss of cytoadherence.** *Ann. Trop. Med. Parasitol.* 1996, **90**:353–7.

134. Kiwuwa MS, Byarugaba J, Wahlgren M, Kironde F: **Detection of copy number variation and single nucleotide polymorphisms in genes involved in drug resistance and other phenotypic traits in P. falciparum clinical isolates collected from Uganda.** *Acta Trop.* 2013, **125**:269–75.

135. Menard S, Morlais I, Tahar R, Sayang C, Mayengue PI, Iriart X, Benoit-Vical F, Lemen B, Magnaval J-F, Awono-Ambene P, Basco LK, Berry A: **Molecular**

monitoring of Plasmodium falciparum drug susceptibility at the time of the introduction of artemisinin-based combination therapy in Yaoundé, Cameroon: implications for the future. *Malar. J.* 2012, **11**:113.

136. Roll Back Malaria: *ImPact serIes Focus on senegal*. 2010.

137. Daniels R, Chang H-H, Séne PD, Park DC, Neafsey DE, Schaffner SF, Hamilton EJ, Lukens AK, Van Tyne D, Mboup S, Sabeti PC, Ndiaye D, Wirth DF, Hartl DL, Volkman SK: **Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal.** *PLoS One* 2013, **8**:e60780.

138. Duah NO, Matrevi S a, de Souza DK, Binnah DD, Tamakloe MM, Opoku VS, Onwona CO, Narh C a, Quashie NB, Abuaku B, Duplessis C, Kronmann KC, Koram K a: **Increased pfmdr1 gene copy number and the decline in pfcrt and pfmdr1 resistance alleles in Ghanaian Plasmodium falciparum isolates after the change of anti-malarial drug treatment policy.** *Malar. J.* 2013, **12**:377.

139. Mobegi V a, Duffy CW, Amambua-Ngwa A, Loua KM, Laman E, Nwakanma DC, Macinnis B, Aspeling-Jones H, Murray L, Clark TG, Kwiatkowski DP, Conway DJ: **Genome-wide analysis of selection on the malaria parasite Plasmodium falciparum in West African populations of differing infection endemicity.** *Mol. Biol. Evol.* 2014:1–37.

140. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A,

Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res.* 2010, **20**:1297–303.

141. Trenholme KR, Gardiner DL, Holt DC, Thomas E a, Cowman a F, Kemp DJ: **clag9: A cytoadherence gene in Plasmodium falciparum essential for binding of parasitized erythrocytes to CD36.** *Proc. Natl. Acad. Sci. U. S. A.* 2000, **97**:4029–33.

142. Eksi S, Morahan BJ, Haile Y, Furuya T, Jiang H, Ali O, Xu H, Kiattibutr K, Suri A, Czesny B, Adeyemo A, Myers TG, Sattabongkot J, Su X, Williamson KC: **Plasmodium falciparum gametocyte development 1 (Pfgdv1) and gametocytogenesis early gene identification and commitment to sexual development.** *PLoS Pathog.* 2012, **8**:e1002964.

143. Assefa S a, Preston MD, Campino S, Ocholla H, Sutherland CJ, Clark TG: **estMOI: estimating multiplicity of infection using parasite deep sequencing data.** *Bioinformatics* 2014, **30**:1292–4.

144. Banerjee R, Liu J, Beatty W, Pelosof L, Klemba M, Goldberg DE: **Four plasmepsins are active in the Plasmodium falciparum food vacuole, including a protease with an active-site histidine.** *Proc. Natl. Acad. Sci. U. S. A.* 2002, **99**:990–5.

145. Chugh M, Sundararaman V, Kumar S, Reddy VS, Siddiqui WA, Stuart KD, Malhotra P: **Protein complex directs hemoglobin-to-hemozoin formation in**

Plasmodium falciparum. *Proc. Natl. Acad. Sci. U. S. A.* 2013, **110**:5392–7.

146. Hempelmann E: **Hemozoin biocrystallization in Plasmodium falciparum and the antimalarial activity of crystallization inhibitors.** *Parasitol. Res.* 2007, **100**:671–6.

147. Olafson KN, Ketchum M a, Rimer JD, Vekilov PG: **Mechanisms of hematin crystallization and inhibition by the antimalarial drug chloroquine.** *Proc. Natl. Acad. Sci. U. S. A.* 2015:1–6.

148. Sorensen M, Sehested M, Jensen PB: **pH-dependent regulation of camptothecin-induced cytotoxicity and cleavable complex formation by the antimalarial agent chloroquine.** *Biochem. Pharmacol.* 1997, **54**:373–80.

149. Pascual A, Fall B, Wurtz N, Fall M, Camara C, Nakoulima A, Baret E, Diatta B, Wade B, Briolant S, Pradines B: **Plasmodium falciparum with multidrug resistance 1 gene duplications, Senegal.** *Emerg. Infect. Dis.* 2013, **19**:814–5.

150. Djimdé A, Doumbo OK, Cortese JF, Kayentao K, Doumbo S, Diourté Y, Coulibaly D, Dicko A, Su XZ, Nomura T, Fidock DA, Wellems TE, Plowe C V: **A molecular marker for chloroquine-resistant falciparum malaria.** *N. Engl. J. Med.* 2001, **344**:257–63.

151. Sidhu ABS, Verdier-Pinard D, Fidock DA: **Chloroquine resistance in Plasmodium falciparum malaria parasites conferred by pfcrt mutations.** *Science*

2002, **298**:210–3.

152. Fidock DA, Nomura T, Talley AK, Cooper RA, Dzekunov SM, Ferdig MT, Ursos LM, Sidhu AB, Naudé B, Deitsch KW, Su XZ, Wootton JC, Roepe PD, Wellems TE: **Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance.** *Mol. Cell* 2000, **6**:861–71.

153. Wootton JC, Feng X, Ferdig MT, Cooper RA, Mu J, Baruch DI, Magill AJ, Su X-Z: **Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum.** *Nature* 2002, **418**:320–3.

154. Osman ME, Mockenhaupt FP, Bienzle U, Elbashir MI, Giha H a: **Field-based evidence for linkage of mutations associated with chloroquine (pfcrt/pfmdr1) and sulfadoxine-pyrimethamine (pfdhfr/pfdhps) resistance and for the fitness cost of multiple mutations in P. falciparum.** *Infect. Genet. Evol.* 2007, **7**:52–9.

155. Ord R, Alexander N, Dunyo S, Hallett R, Jawara M, Targett G, Drakeley CJ, Sutherland CJ: **Seasonal carriage of pfcrt and pfmdr1 alleles in Gambian Plasmodium falciparum imply reduced fitness of chloroquine-resistant parasites.** *J. Infect. Dis.* 2007, **196**:1613–9.

156. Volkman SK, Sabeti PC, DeCaprio D, Neafsey DE, Schaffner SF, Milner DA, Daily JP, Sarr O, Ndiaye D, Ndir O, Mboup S, Duraisingh MT, Lukens A, Derr A, Stange-Thomann N, Waggoner S, Onofrio R, Ziaugra L, Mauceli E, Gnerre S, Jaffe DB,

Zainoun J, Wiegand RC, Birren BW, Hartl DL, Galagan JE, Lander ES, Wirth DF: **A genome-wide map of diversity in Plasmodium falciparum.** *Nat. Genet.* 2007, **39**:113–9.

157. Oduola AM, Weatherly NF, Bowdre JH, Desjardins RE: **Plasmodium falciparum: cloning by single-erythrocyte micromanipulation and heterogeneity in vitro.** *Exp. Parasitol.* 1988, **66**:86–95.

158. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z: **Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads.** *Bioinformatics* 2009, **25**:2865–71.

159. Sharkey A, Langsley G, Patarapotikul J, Mercereau-Puijalon O, McLean AP, Walliker D: **Chromosome size variation in the malaria parasite of rodents, Plasmodium chabaudi.** *Mol. Biochem. Parasitol.* 1988, **28**:47–54.

160. Sheppard M, Thompson JK, Anders RF, Kemp DJ, Lew AM: **Molecular karyotyping of the rodent malarias Plasmodium chabaudi, Plasmodium berghei and Plasmodium vinckei.** *Mol. Biochem. Parasitol.* 1989, **34**:45–52.

161. Dore E, Pace T, Picci L, Pizzi E, Ponzi M, Frontali C: **Dynamics of telomere turnover in Plasmodium berghei.** *Mol. Biol. Rep.* 1994, **20**:27–33.

162. Del Portillo HA, Fernandez-Becerra C, Bowman S, Oliver K, Preuss M, Sanchez CP, Schneider NK, Villalobos JM, Rajandream MA, Harris D, Pereira da Silva

LH, Barrell B, Lanzer M: **A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax.** *Nature* 2001, **410**:839–42.

163. Springer M, Weissman JS, Kirschner MW: **A general lack of compensation for gene dosage in yeast.** *Mol. Syst. Biol.* 2010, **6**:368.

164. Duffy MF, Selvarajah S a, Josling G a, Petter M: **Epigenetic regulation of the Plasmodium falciparum genome.** *Brief. Funct. Genomics* 2014, **13**:203–16.

165. Ponts N, Fu L, Harris EY, Zhang J, Chung D-WD, Cervantes MC, Prudhomme J, Atanasova-Penichon V, Zehraoui E, Bunnik EM, Rodrigues EM, Lonardi S, Hicks GR, Wang Y, Le Roch KG: **Genome-wide mapping of DNA methylation in the human malaria parasite Plasmodium falciparum.** *Cell Host Microbe* 2013, **14**:696–706.

166. Gupta AP, Chin WH, Zhu L, Mok S, Luah Y-H, Lim E-H, Bozdech Z: **Dynamic epigenetic regulation of gene expression during the life cycle of malaria parasite Plasmodium falciparum.** *PLoS Pathog.* 2013, **9**:e1003170.