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

Derrick K. DeConti
University of Massachusetts Medical School

Let us know how access to this document benefits you.
Follow this and additional works at: https://escholarship.umassmed.edu/gsbs_diss

Part of the Bioinformatics Commons, Computational Biology Commons, Genetics Commons, and the Genomics Commons

Repository Citation

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in GSBS Dissertations and Theses by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
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.

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

Douglas Golenbock, M.D., Member of Committee

Ann Moormann, Ph.D., Member of Committee

Evgeny Rogaev, Ph.D., Member of Committee

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.

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.

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 an 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.
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 stegov 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 \textit{crt} gene – containing both the chloroquine resistant and sensitive allele. The synthesis of our SD and CNV analysis indicates a CNV conservative \textit{P. falciparum} genome except where drug and human immune pressure select for gene duplication.
# Table of Contents

Acknowledgements............................................................................................................ iii

Abstract................................................................................................................................v

List of Figures......................................................................................................................x

List of Tables..................................................................................................................... xii

List of Abbreviations........................................................................................................ xiii

Preface.............................................................................................................................. xiv

CHAPTER I: Introduction................................................................................................... 1
  1.1 Genomic Duplications and Deletions........................................................................2
  1.2 Detection of Duplications and Deletions...............................................................9
  1.3 *Plasmodium falciparum*.......................................................................................12
  1.4 Duplications and Deletions in *Plasmodium falciparum*......................................16

CHAPTER II: Segmental Duplications in *P. falciparum*...............................................20
  2.1 Introduction............................................................................................................. 21
  2.2 Methods...................................................................................................................24
    Segmental duplication detection..............................................................................24
    Descriptive statistics of pairwise alignments.........................................................33
    Analysis of segmental duplications around *var* and *riffin* genes......................34
  2.3 Results..................................................................................................................... 34
    Detection of pairwise alignments representing segmental duplications...............34
Genomic distribution of segmental duplications......................................................37
Gene content of segmental duplications.................................................................42
Extracellularly exposed genes overrepresented in segmental duplications..............44
The nature of var and rifin gene segmental duplications.........................................47

2.4 Discussion...............................................................................................................50

CHAPTER III: Copy Number Variation Detection from High Throughput Sequencing in
a Natural Population of P. falciparum........................................................................53

3.1 Introduction.............................................................................................................54

3.2 Methods..................................................................................................................56
Data collection..........................................................................................................56
Overview of CNV detection.....................................................................................56
GC normalization of read depth.............................................................................60
Mean shift analysis of normalized read depth.......................................................61
Discordant read pair intersection analysis and breakpoint trimming.......................61

3.3 Results....................................................................................................................62
Copy number detection within well-characterized laboratory strains......................62
Copy number detection within Senegal strains.......................................................72

3.4 Discussion...............................................................................................................76

CHAPTER IV: Global Analysis of Copy Number Variation in P. falciparum..............79

4.1 Introduction.............................................................................................................80

4.2 Methods..................................................................................................................81
Sample acquisition and quality control....................................................................81
List of Figures

Figure 1.1 The effects of gene duplication........................................................................4
Figure 1.2 Subfunctionalization and neofunctionalization of gene duplication..............7
Figure 1.3 Life cycle of the malaria parasite *P. falciparum*...........................................13
Figure 2.1 Overview of segmental duplication identification by self-alignment..............25
Figure 2.2 LASTZ custom substitution scoring outperforms flat match/mis-match scoring ............................................................................................................................................30
Figure 2.3 Overview of segmental duplications in the genome........................................35
Figure 2.4 Comparison of WSSD to WGAC for segmental duplication identification.....38
Figure 2.5 Duplication content between chromosomes...................................................40
Figure 2.6 Duplication content by percent similarity.......................................................45
Figure 2.7 Model of possible duplication patterns for *var* and *rifin* genes...............48
Figure 3.1 Overview of CNV detection pipeline.............................................................57
Figure 3.2 High variability of GC bias between sequencing libraries..............................64
Figure 3.3 High variability read depth based mean shift CNV identification between sequencing libraries.............................................................................................................66
Figure 3.4 Intersection of read depth and discordant reads identifies CNV of PFE1150w and PFI1710w....................................................................................................................70
Figure 3.5 Accurate estimation of copy number via read depth based CNV detection....73
Figure 4.1 Map of *P. falciparum* sample origin.............................................................82
Figure 4.2 Frequency plot of genic CNVs.................................................................87
Figure 4.3 Overview of CNV loci throughout the genome........................................89
Figure 4.4 Copy number of genic duplications indicate preference for minimal duplication........................................................................................................................92
Figure 4.5 Majority of CNVs are under 40 kbp..........................................................94
Figure 4.6 Multiple isoforms of mdr1 duplication identified........................................97
Figure 4.7 Multiple isoforms of gch duplication identified...........................................99
Figure 4.8 Identification of single isoform of crt duplication.................................102
Figure 4.9 Biallelic duplication of crt gene...............................................................105
Figure 4.10 Consensus sequence diverges within the crt CNV...............................108
Figure 4.11 PCA of SNPs across chromosome 7 between geographic regions with crt duplication.........................................................................................................................110
List of Tables

Table 2.1 LASTZ custom substitution matrix.................................................................28
Table 3.1 Identified CNV from in vitro cultured lab strains........................................68
Appendix Table 1 Identified Copy Number Variants..................................................126
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 \textit{in vitro} cultures for the 33 Senegal \textit{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.
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 neofunctionalization (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 in 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.
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.
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. stevor 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. DeConti¹, Jeffrey A. Bailey¹,²

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 histocompatibility 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 studied 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 putatively 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.
<table>
<thead>
<tr>
<th>Query Base</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>-9</td>
<td>-7</td>
<td>-13</td>
</tr>
<tr>
<td>T</td>
<td>-9</td>
<td>13</td>
<td>-1</td>
<td>-5</td>
</tr>
<tr>
<td>C</td>
<td>-7</td>
<td>-1</td>
<td>11</td>
<td>-8</td>
</tr>
<tr>
<td>G</td>
<td>-12</td>
<td>-5</td>
<td>-10</td>
<td>4</td>
</tr>
</tbody>
</table>
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.
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 $\geq$250bp and $\geq$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 based 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 $\geq 2$ 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 $\geq 90\%$ identity and $\geq 250$ 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 \textit{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 $\pm 50$ 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 $\geq 250$ bp in size and $\geq 90\%$ 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.
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 = 8.79x10^{-70}) indicating some variability. Chromosome 5 stands out as it was over 2 standard deviations (σ = 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).
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.
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 = 3.73x10^-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 outside 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 ubiquitin 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

B

Non-redundant duplication content by percent identity
The nature of \textit{var} and \textit{rifin} gene segmental duplications

The majority of \textit{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 \textit{var} gene (Figure 2.7). This revealed that most \textit{var} genes have divergent pairwise alignments flanking the genes – i.e. each side of the \textit{var} gene best identifies with a different \textit{var} gene. Of the 81 \textit{var} genes, only 29 (35\%) genes had flanking pairwise alignments match to the same \textit{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 \textit{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, \textit{rifins}, were analyzed similarly. While rifins show similar evidence of a high rate of duplication, with 90\% of \textit{rifin} genes having pairwise alignments intersecting and 100\% of \textit{var} genes. However, the \textit{rifin} genes were more likely to have a spanning pairwise alignment, 33\% of \textit{rifin} genes had a spanning pairwise alignment as opposed to 6\% of \textit{var} genes. When looking at the highest identity alignment for the flanking of \textit{rifin} genes, show similar levels of cross mapping of alignments with 106 (58\%) of the \textit{rifin} genes had the highest pairwise alignment on either end of the \textit{rifin} pair to different \textit{rifin} genes, while only 55 (30\%) genes had alignments on either side that matched to same \textit{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.
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 extracellularly 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. DeConti¹, Kate M. Fernandez², Sarah K. Volkman², Dyann F. Wirth², Daniel E. Neafsey³, and Jeffrey A. Bailey¹,⁴

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).
Normalization of GC bias in read depth

Mean depth identification of read depth variant regions

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 then 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
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 GAI1 sequencing libraries from an in vitro 3D7 lab culture DNA isolate.
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.
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.
<table>
<thead>
<tr>
<th>Strain</th>
<th>Chrom</th>
<th>Position</th>
<th>Length</th>
<th>Presumed relevant gene</th>
<th>Copy Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>106/1</td>
<td>4</td>
<td>732101 - 783400</td>
<td>51299</td>
<td>pfdhfr1</td>
<td>1.91</td>
</tr>
<tr>
<td>FCB</td>
<td>5</td>
<td>868901 - 964600</td>
<td>95699</td>
<td>pfmdr1</td>
<td>2.09</td>
</tr>
<tr>
<td>106/1</td>
<td>5</td>
<td>869001 - 952700</td>
<td>83699</td>
<td>pfmdr1</td>
<td>1.65</td>
</tr>
<tr>
<td>FCB</td>
<td>7</td>
<td>1293001 - 1295500</td>
<td>2499</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>106/1</td>
<td>9</td>
<td>1374001 - 1396100</td>
<td>22099</td>
<td>pfgdv1</td>
<td>0</td>
</tr>
<tr>
<td>FCB</td>
<td>9</td>
<td>1374401 - 1396000</td>
<td>21599</td>
<td>pfgdv1</td>
<td>0</td>
</tr>
<tr>
<td>FCB</td>
<td>9</td>
<td>1459502 - 1541735</td>
<td>82233</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>106/1</td>
<td>9</td>
<td>1459502 - 1541735</td>
<td>82233</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>106/1</td>
<td>12</td>
<td>946301 - 980600</td>
<td>34299</td>
<td>pfgch1</td>
<td>4.24</td>
</tr>
<tr>
<td>FCB</td>
<td>12</td>
<td>961501 - 980600</td>
<td>19099</td>
<td>pfgch1</td>
<td>2.90</td>
</tr>
<tr>
<td>3D7</td>
<td>12</td>
<td>974301 - 975900</td>
<td>1599</td>
<td>pfgch1</td>
<td>3.67</td>
</tr>
</tbody>
</table>
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.
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 population, 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].
\[ R^2 = 0.902 \]
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
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. Specifically, 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 \textit{crt} duplication; Thailand and Cambodia for both \textit{mdr1} and \textit{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

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).
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 \textit{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 \textit{in vitro} culture adaptations, so \textit{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.
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.
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 \textit{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 \textit{in vivo}, but are frequently produced over long term \textit{in vitro} culture of an isolate [71, 133]. These deletions often result in loss of gametocytogenesis, loss of cytoadherence, and increased growth rate \textit{in vitro} [71, 131, 141, 142].

We found multiple recurrences of two well-known and characterized copy number variants of \textit{mdr1} and \textit{gch}. Overlay of the copy number variants of \textit{mdr1} and \textit{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 \textit{P. falciparum} [87]. We identified the \textit{mdr1} gene solely in South East Asia, however the \textit{gch} duplication was identified in Ghana in addition to South East Asia. Both copy number variants had a high frequency, with the \textit{mdr1} duplication found in 32 strains and the \textit{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 \textit{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.
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.
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 GOSstat 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.
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 \textit{crt} duplication**

Upon closer inspection of the \textit{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 \textit{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 \textit{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 \textit{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 \textit{crt} duplication correspond to SNPs in the \textit{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 \textit{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 *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.
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 \textit{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 \textit{crt} duplicated strains was in line with the rest of the West African population. The lack of a separate subgroup of \textit{crt} duplicated strains reinforces the identification of only a \sim 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

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.
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 families 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 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 \textit{crt}. This duplication was bialleic for both chloroquine resistant and chloroquine sensitive alleles of \textit{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 \textit{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 \textit{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

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Position</th>
<th>Length</th>
<th>Copy number Sampling site</th>
<th>Sample ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr01</td>
<td>238445 - 248349</td>
<td>9,905</td>
<td>2.16 Gambia</td>
<td>374</td>
</tr>
<tr>
<td>chr01</td>
<td>549500 - 640851</td>
<td>91,352</td>
<td>0.00 Senegal</td>
<td>SenP27.02</td>
</tr>
<tr>
<td>chr01</td>
<td>563800 - 640851</td>
<td>77,052</td>
<td>0.00 Senegal</td>
<td>SenT10.04D10</td>
</tr>
<tr>
<td>chr01</td>
<td>551800 - 554000</td>
<td>2,201</td>
<td>0.00 Senegal</td>
<td>SenT32.09</td>
</tr>
<tr>
<td>chr02</td>
<td>220981 - 238374</td>
<td>17,394</td>
<td>2.02 Cambodia (Ratanakiri)</td>
<td>286</td>
</tr>
<tr>
<td>chr02</td>
<td>355068 - 356081</td>
<td>1,014</td>
<td>2.02 Thailand</td>
<td>796</td>
</tr>
<tr>
<td>chr02</td>
<td>555458 - 662357</td>
<td>106,900</td>
<td>1.66 Gambia</td>
<td>403</td>
</tr>
<tr>
<td>chr02</td>
<td>671115 - 672530</td>
<td>1,416</td>
<td>2.01 Thailand</td>
<td>766</td>
</tr>
<tr>
<td>chr02</td>
<td>835300 - 947102</td>
<td>111,803</td>
<td>0.00 Senegal</td>
<td>SenT09.09</td>
</tr>
<tr>
<td>chr02</td>
<td>863500 - 947102</td>
<td>83,603</td>
<td>0.00 Senegal</td>
<td>SenT135.09</td>
</tr>
<tr>
<td>chr03</td>
<td>128244 - 134634</td>
<td>6,391</td>
<td>0.13 Ghana</td>
<td>477</td>
</tr>
<tr>
<td>chr03</td>
<td>251527 - 252914</td>
<td>1,388</td>
<td>3.84 Thailand</td>
<td>766</td>
</tr>
<tr>
<td>chr03</td>
<td>663859 - 664651</td>
<td>793</td>
<td>2.21 Gambia</td>
<td>374</td>
</tr>
<tr>
<td>chr03</td>
<td>984410 - 985091</td>
<td>682</td>
<td>2.61 Cambodia (Pursat)</td>
<td>136</td>
</tr>
<tr>
<td>chr04</td>
<td>250530 - 349000</td>
<td>98,471</td>
<td>1.95 Gambia</td>
<td>407</td>
</tr>
<tr>
<td>chr05</td>
<td>388213 - 390121</td>
<td>1,909</td>
<td>0.41 Ghana</td>
<td>477</td>
</tr>
<tr>
<td>chr05</td>
<td>441610 - 461650</td>
<td>20,041</td>
<td>1.48 Cambodia (Pursat)</td>
<td>218</td>
</tr>
<tr>
<td>chr05</td>
<td>446993 - 462053</td>
<td>15,061</td>
<td>1.56 Ghana</td>
<td>574</td>
</tr>
<tr>
<td>chr05</td>
<td>448740 - 460238</td>
<td>11,499</td>
<td>1.70 Cambodia (Ratanakiri)</td>
<td>248</td>
</tr>
<tr>
<td>chr05</td>
<td>505687 - 506365</td>
<td>679</td>
<td>2.28 Ghana</td>
<td>477</td>
</tr>
<tr>
<td>chr05</td>
<td>670907 - 770083</td>
<td>99,177</td>
<td>1.83 Gambia</td>
<td>407</td>
</tr>
<tr>
<td>chr05</td>
<td>870787 - 978366</td>
<td>107,580</td>
<td>3.58 Cambodia (Pursat)</td>
<td>122</td>
</tr>
<tr>
<td>Chromosome</td>
<td>Start - End</td>
<td>Length</td>
<td>Distance</td>
<td>Location</td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>---------</td>
<td>----------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>chr05</td>
<td>943270 - 970993</td>
<td>27,724</td>
<td>2.03</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>943344 - 964658</td>
<td>21,315</td>
<td>2.11</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>944490 - 973296</td>
<td>28,807</td>
<td>2.07</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>944894 - 964526</td>
<td>19,633</td>
<td>3.15</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr05</td>
<td>945099 - 964624</td>
<td>19,526</td>
<td>2.00</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>946149 - 973288</td>
<td>27,140</td>
<td>1.91</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>946483 - 967340</td>
<td>20,858</td>
<td>3.08</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>946485 - 964658</td>
<td>18,174</td>
<td>3.11</td>
<td>Cambodia (Pailin)</td>
</tr>
<tr>
<td>chr05</td>
<td>946487 - 964647</td>
<td>18,161</td>
<td>2.94</td>
<td>Cambodia (Pailin)</td>
</tr>
<tr>
<td>chr05</td>
<td>946488 - 964640</td>
<td>18,153</td>
<td>3.09</td>
<td>Cambodia (Pailin)</td>
</tr>
<tr>
<td>chr05</td>
<td>946488 - 964652</td>
<td>18,165</td>
<td>3.07</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>946558 - 964629</td>
<td>18,072</td>
<td>2.03</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>947121 - 963187</td>
<td>16,067</td>
<td>2.10</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr05</td>
<td>947579 - 962565</td>
<td>14,987</td>
<td>2.02</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr05</td>
<td>947636 - 962557</td>
<td>14,922</td>
<td>2.82</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr05</td>
<td>947773 - 969928</td>
<td>22,156</td>
<td>3.23</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>947777 - 969926</td>
<td>22,153</td>
<td>2.09</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>947781 - 969907</td>
<td>22,127</td>
<td>1.91</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr05</td>
<td>947785 - 969917</td>
<td>22,133</td>
<td>2.01</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr05</td>
<td>948114 - 970296</td>
<td>22,183</td>
<td>1.47</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr05</td>
<td>950684 - 972006</td>
<td>21,323</td>
<td>3.12</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr05</td>
<td>952701 - 970350</td>
<td>17,650</td>
<td>2.13</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>953745 - 970361</td>
<td>16,617</td>
<td>3.04</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr05</td>
<td>953747 - 970349</td>
<td>16,603</td>
<td>2.12</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr05</td>
<td>953794 - 972127</td>
<td>18,334</td>
<td>3.82</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr05</td>
<td>953827 - 970299</td>
<td>16,473</td>
<td>2.94</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>953828 - 970293</td>
<td>16,466</td>
<td>3.14</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>953843 - 970259</td>
<td>16,417</td>
<td>2.33</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>953870 - 970299</td>
<td>16,430</td>
<td>3.04</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr05</td>
<td>953871 - 970285</td>
<td>16,415</td>
<td>2.62</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>962655 - 963279</td>
<td>625</td>
<td>2.41</td>
<td>Kenya</td>
</tr>
<tr>
<td>chr06</td>
<td>485029 - 485675</td>
<td>647</td>
<td>2.06</td>
<td>Gambia</td>
</tr>
<tr>
<td>chr06</td>
<td>492315 - 503056</td>
<td>10,742</td>
<td>1.62</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>849764 - 850377</td>
<td>614</td>
<td>2.40</td>
<td>Mali</td>
</tr>
<tr>
<td>chr06</td>
<td>1117358 - 1119225</td>
<td>1,868</td>
<td>2.04</td>
<td>Senegal</td>
</tr>
<tr>
<td>Chr 6</td>
<td>Start - End</td>
<td>Length</td>
<td>Dist</td>
<td>Country</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>--------</td>
<td>------</td>
<td>---------</td>
</tr>
<tr>
<td>chr06</td>
<td>1117364 - 1119201</td>
<td>1,838</td>
<td>2.11</td>
<td>Guinea</td>
</tr>
<tr>
<td>chr06</td>
<td>1117368 - 1119193</td>
<td>1,826</td>
<td>2.03</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr06</td>
<td>1117369 - 1119206</td>
<td>1,836</td>
<td>2.89</td>
<td>Guinea</td>
</tr>
<tr>
<td>chr06</td>
<td>1117373 - 1119156</td>
<td>1,784</td>
<td>1.94</td>
<td>Guinea</td>
</tr>
<tr>
<td>chr06</td>
<td>1117376 - 1119225</td>
<td>1,850</td>
<td>2.13</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr06</td>
<td>1117379 - 1119205</td>
<td>1,827</td>
<td>3.07</td>
<td>Guinea</td>
</tr>
<tr>
<td>chr06</td>
<td>1117380 - 1119204</td>
<td>1,825</td>
<td>2.02</td>
<td>Guinea</td>
</tr>
<tr>
<td>chr06</td>
<td>1117381 - 1119204</td>
<td>1,824</td>
<td>1.97</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr06</td>
<td>1117385 - 1119187</td>
<td>1,803</td>
<td>2.33</td>
<td>Guinea</td>
</tr>
<tr>
<td>chr06</td>
<td>1117385 - 1119193</td>
<td>1,809</td>
<td>2.20</td>
<td>Guinea</td>
</tr>
<tr>
<td>chr06</td>
<td>1117397 - 1119202</td>
<td>1,806</td>
<td>2.44</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>1117399 - 1119209</td>
<td>1,811</td>
<td>2.35</td>
<td>Cambodia(Ratanakiri)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117399 - 1119202</td>
<td>1,804</td>
<td>2.83</td>
<td>Gambia</td>
</tr>
<tr>
<td>chr06</td>
<td>1117399 - 1119211</td>
<td>1,813</td>
<td>1.98</td>
<td>Mali</td>
</tr>
<tr>
<td>chr06</td>
<td>1117399 - 1119204</td>
<td>1,806</td>
<td>2.09</td>
<td>Cambodia(Pailin)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117399 - 1119202</td>
<td>1,804</td>
<td>2.06</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>1117399 - 1119186</td>
<td>1,788</td>
<td>3.24</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>1117400 - 1119206</td>
<td>1,807</td>
<td>2.94</td>
<td>Cambodia(Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117400 - 1119200</td>
<td>1,801</td>
<td>2.38</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>1117400 - 1119207</td>
<td>1,808</td>
<td>2.19</td>
<td>Cambodia(Ratanakiri)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117400 - 1119199</td>
<td>1,800</td>
<td>1.92</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>1117400 - 1119016</td>
<td>1,617</td>
<td>2.24</td>
<td>Cambodia(Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117400 - 1119198</td>
<td>1,799</td>
<td>2.12</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>1117400 - 1119185</td>
<td>1,786</td>
<td>2.60</td>
<td>Ghana</td>
</tr>
<tr>
<td>chr06</td>
<td>1117401 - 1119202</td>
<td>1,802</td>
<td>1.95</td>
<td>Cambodia(Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117401 - 1119195</td>
<td>1,795</td>
<td>2.27</td>
<td>Cambodia(Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117401 - 1119201</td>
<td>1,801</td>
<td>2.05</td>
<td>Cambodia(Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117401 - 1119191</td>
<td>1,791</td>
<td>2.00</td>
<td>Cambodia(Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117401 - 1119199</td>
<td>1,799</td>
<td>2.13</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>1117402 - 1119206</td>
<td>1,805</td>
<td>2.02</td>
<td>Cambodia(Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117402 - 1119202</td>
<td>1,801</td>
<td>2.01</td>
<td>Cambodia(Pailin)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117402 - 1119204</td>
<td>1,803</td>
<td>2.60</td>
<td>Cambodia(Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117402 - 1119200</td>
<td>1,799</td>
<td>2.09</td>
<td>Cambodia(Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117402 - 1119202</td>
<td>1,801</td>
<td>2.25</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>1117403 - 1119187</td>
<td>1,785</td>
<td>2.04</td>
<td>Cambodia(Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117403 - 1119195</td>
<td>1,793</td>
<td>1.99</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>1117403 - 1119204</td>
<td>1,802</td>
<td>2.01</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>1117403 - 1119188</td>
<td>1,786</td>
<td>2.02</td>
<td>Cambodia(Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117404 - 1119201</td>
<td>1,798</td>
<td>2.30</td>
<td>Thailand</td>
</tr>
<tr>
<td>Chr</td>
<td>Position</td>
<td>Length</td>
<td>Distance</td>
<td>Location</td>
</tr>
<tr>
<td>-----</td>
<td>----------</td>
<td>--------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>chr06</td>
<td>1117404 - 1119191</td>
<td>1,788</td>
<td>2.09</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117404 - 1119200</td>
<td>1,797</td>
<td>2.24</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117404 - 1119203</td>
<td>1,800</td>
<td>2.37</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117404 - 1119187</td>
<td>1,784</td>
<td>2.01</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117404 - 1119201</td>
<td>1,798</td>
<td>2.27</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117405 - 1119200</td>
<td>1,796</td>
<td>1.97</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117405 - 1119198</td>
<td>1,794</td>
<td>1.97</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117405 - 1119204</td>
<td>1,800</td>
<td>2.14</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117406 - 1119174</td>
<td>1,769</td>
<td>2.26</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117406 - 1119186</td>
<td>1,781</td>
<td>4.02</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117407 - 1119199</td>
<td>1,793</td>
<td>2.28</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117407 - 1119206</td>
<td>1,800</td>
<td>2.21</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117407 - 1119175</td>
<td>1,769</td>
<td>2.36</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117408 - 1119197</td>
<td>1,790</td>
<td>2.01</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117412 - 1119128</td>
<td>1,717</td>
<td>2.17</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117413 - 1119198</td>
<td>1,786</td>
<td>1.99</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>1117418 - 1119019</td>
<td>1,602</td>
<td>2.30</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr06</td>
<td>1117418 - 1119131</td>
<td>1,714</td>
<td>1.91</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr06</td>
<td>1117424 - 1119151</td>
<td>1,728</td>
<td>2.51</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117434 - 1119149</td>
<td>1,716</td>
<td>2.26</td>
<td>Guinea</td>
</tr>
<tr>
<td>chr06</td>
<td>1117438 - 1119192</td>
<td>1,755</td>
<td>2.87</td>
<td>Cambodia (Pailin)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117484 - 1119158</td>
<td>1,675</td>
<td>2.16</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117485 - 1119195</td>
<td>1,711</td>
<td>2.26</td>
<td>Guinea</td>
</tr>
<tr>
<td>chr06</td>
<td>1117486 - 1119130</td>
<td>1,645</td>
<td>2.17</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117487 - 1119162</td>
<td>1,676</td>
<td>3.04</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117488 - 1119162</td>
<td>1,675</td>
<td>3.13</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117489 - 1119152</td>
<td>1,664</td>
<td>2.22</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117598 - 1119095</td>
<td>1,498</td>
<td>2.09</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr06</td>
<td>1117624 - 1119099</td>
<td>1,472</td>
<td>2.12</td>
<td>Ghana</td>
</tr>
<tr>
<td>chr06</td>
<td>1117746 - 1119167</td>
<td>1,422</td>
<td>2.34</td>
<td>Ghana</td>
</tr>
<tr>
<td>chr06</td>
<td>1117766 - 1119181</td>
<td>1,416</td>
<td>2.11</td>
<td>Guinea</td>
</tr>
<tr>
<td>chr06</td>
<td>1117827 - 1119194</td>
<td>1,368</td>
<td>2.36</td>
<td>Cambodia (Ratanakiri)</td>
</tr>
<tr>
<td>chr07</td>
<td>394657 - 421919</td>
<td>27,263</td>
<td>1.91</td>
<td>Burkina Faso</td>
</tr>
<tr>
<td>chr07</td>
<td>398667 - 421914</td>
<td>23,248</td>
<td>1.91</td>
<td>Guinea</td>
</tr>
<tr>
<td>chr07</td>
<td>398674 - 421912</td>
<td>23,239</td>
<td>1.83</td>
<td>Ghana</td>
</tr>
<tr>
<td>chr07</td>
<td>398676 - 421917</td>
<td>23,242</td>
<td>2.06</td>
<td>Ghana</td>
</tr>
<tr>
<td>chr07</td>
<td>398678 - 421912</td>
<td>23,235</td>
<td>1.98</td>
<td>Ghana</td>
</tr>
<tr>
<td>chr07</td>
<td>398679 - 421905</td>
<td>23,227</td>
<td>2.21</td>
<td>Burkina Faso</td>
</tr>
<tr>
<td>Chromosome</td>
<td>Start Position</td>
<td>End Position</td>
<td>Base Pair Count</td>
<td>Country</td>
</tr>
<tr>
<td>------------</td>
<td>---------------</td>
<td>--------------</td>
<td>----------------</td>
<td>--------</td>
</tr>
<tr>
<td>chr07</td>
<td>779132</td>
<td>796453</td>
<td>17,322</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr07</td>
<td>838352</td>
<td>888689</td>
<td>50,338</td>
<td>Gambia</td>
</tr>
<tr>
<td>chr07</td>
<td>1071568</td>
<td>1072232</td>
<td>665</td>
<td>Kenya</td>
</tr>
<tr>
<td>chr08</td>
<td>1314610</td>
<td>1315703</td>
<td>1,094</td>
<td>Gambia</td>
</tr>
<tr>
<td>chr09</td>
<td>321465</td>
<td>322688</td>
<td>1,224</td>
<td>Gambia</td>
</tr>
<tr>
<td>chr09</td>
<td>1095691</td>
<td>1096396</td>
<td>706</td>
<td>Mali</td>
</tr>
<tr>
<td>chr09</td>
<td>1196177</td>
<td>1272137</td>
<td>75,961</td>
<td>Gambia</td>
</tr>
<tr>
<td>chr09</td>
<td>1226756</td>
<td>1251796</td>
<td>25,221</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1240368</td>
<td>1251255</td>
<td>10,888</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1377600</td>
<td>1541735</td>
<td>164,136</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1379800</td>
<td>1541735</td>
<td>161,936</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1385500</td>
<td>1541735</td>
<td>156,236</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1387200</td>
<td>1541735</td>
<td>154,536</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1387700</td>
<td>1541735</td>
<td>154,036</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1469100</td>
<td>1541735</td>
<td>72,636</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1384812</td>
<td>1385890</td>
<td>1,079</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1397601</td>
<td>1399000</td>
<td>1,400</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1397755</td>
<td>1399113</td>
<td>1,359</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1397756</td>
<td>1399122</td>
<td>1,367</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1397758</td>
<td>1399124</td>
<td>1,367</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1397762</td>
<td>1399109</td>
<td>1,348</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1397763</td>
<td>1399115</td>
<td>1,353</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1397764</td>
<td>1399111</td>
<td>1,348</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1397764</td>
<td>1399114</td>
<td>1,351</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1397768</td>
<td>1399118</td>
<td>1,351</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1397790</td>
<td>1399132</td>
<td>1,343</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr09</td>
<td>1397792</td>
<td>1399116</td>
<td>1,325</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1397796</td>
<td>1399114</td>
<td>1,319</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1397801</td>
<td>1399100</td>
<td>1,300</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1397812</td>
<td>1399134</td>
<td>1,323</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr09</td>
<td>1397824</td>
<td>1399117</td>
<td>1,294</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr09</td>
<td>1397839</td>
<td>1399122</td>
<td>1,284</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>Chromosome</td>
<td>Position Range</td>
<td>Length</td>
<td>Fraction</td>
<td>Country/Region</td>
</tr>
<tr>
<td>------------</td>
<td>----------------</td>
<td>--------</td>
<td>----------</td>
<td>----------------</td>
</tr>
<tr>
<td>chr09</td>
<td>1397843 - 1399125</td>
<td>1,283</td>
<td>0.23</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr09</td>
<td>1397901 - 1399000</td>
<td>1,100</td>
<td>0.11</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1397901 - 1399000</td>
<td>1,100</td>
<td>0.17</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1397901 - 1399000</td>
<td>1,100</td>
<td>0.13</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1397901 - 1399000</td>
<td>1,100</td>
<td>0.09</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1397901 - 1399200</td>
<td>1,300</td>
<td>0.13</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1397901 - 1399200</td>
<td>1,300</td>
<td>0.18</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1397901 - 1399200</td>
<td>1,300</td>
<td>0.20</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1397901 - 1399300</td>
<td>1,400</td>
<td>0.20</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1397901 - 1399300</td>
<td>1,400</td>
<td>0.22</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr09</td>
<td>1407201 - 1408300</td>
<td>1,100</td>
<td>0.40</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr10</td>
<td>903354 - 904074</td>
<td>721</td>
<td>2.52</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr10</td>
<td>1170378 - 1174324</td>
<td>3,947</td>
<td>2.43</td>
<td>Gambia</td>
</tr>
<tr>
<td>chr10</td>
<td>1376686 - 1440190</td>
<td>63,505</td>
<td>3.04</td>
<td>Senegal</td>
</tr>
<tr>
<td>chr10</td>
<td>1424265 - 1425505</td>
<td>1,241</td>
<td>2.07</td>
<td>Gambia</td>
</tr>
<tr>
<td>chr11</td>
<td>461088 - 493610</td>
<td>32,523</td>
<td>0.16</td>
<td>Gambia</td>
</tr>
<tr>
<td>chr11</td>
<td>687757 - 688583</td>
<td>827</td>
<td>2.08</td>
<td>Mali</td>
</tr>
<tr>
<td>chr11</td>
<td>814857 - 868533</td>
<td>53,677</td>
<td>1.41</td>
<td>Gambia</td>
</tr>
<tr>
<td>chr11</td>
<td>821011 - 845503</td>
<td>24,493</td>
<td>1.60</td>
<td>Mali</td>
</tr>
<tr>
<td>chr11</td>
<td>1054820 - 1059223</td>
<td>4,404</td>
<td>0.16</td>
<td>Burkina Faso</td>
</tr>
<tr>
<td>chr11</td>
<td>1054851 - 1059280</td>
<td>4,430</td>
<td>0.43</td>
<td>Mali</td>
</tr>
<tr>
<td>chr11</td>
<td>1054935 - 1059247</td>
<td>4,313</td>
<td>0.44</td>
<td>Mali</td>
</tr>
<tr>
<td>chr12</td>
<td>461869 - 462612</td>
<td>744</td>
<td>2.37</td>
<td>Kenya</td>
</tr>
<tr>
<td>chr12</td>
<td>822219 - 830842</td>
<td>8,624</td>
<td>3.04</td>
<td>Cambodia (Ratanakiri)</td>
</tr>
<tr>
<td>chr12</td>
<td>903997 - 1023825</td>
<td>119,829</td>
<td>1.81</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr12</td>
<td>934736 - 975016</td>
<td>40,281</td>
<td>1.59</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr12</td>
<td>936352 - 936993</td>
<td>642</td>
<td>7.90</td>
<td>Kenya</td>
</tr>
<tr>
<td>chr12</td>
<td>941097 - 980769</td>
<td>39,673</td>
<td>4.29</td>
<td>Cambodia (Pailin)</td>
</tr>
<tr>
<td>chr12</td>
<td>967737 - 978066</td>
<td>10,330</td>
<td>1.91</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr12</td>
<td>968639 - 978077</td>
<td>9,439</td>
<td>2.69</td>
<td>Thailand</td>
</tr>
<tr>
<td>Chr</td>
<td>Start - End</td>
<td>Length</td>
<td>Population/Location</td>
<td>Code</td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
<td>--------</td>
<td>-----------------------</td>
<td>------</td>
</tr>
<tr>
<td>chr12</td>
<td>968640 - 978085</td>
<td>9,446</td>
<td>10.50 Thailand</td>
<td>711</td>
</tr>
<tr>
<td>chr12</td>
<td>968643 - 980646</td>
<td>12,004</td>
<td>2.91 Cambodia (Pursat)</td>
<td>168</td>
</tr>
<tr>
<td>chr12</td>
<td>968643 - 978084</td>
<td>9,442</td>
<td>5.80 Thailand</td>
<td>757</td>
</tr>
<tr>
<td>chr12</td>
<td>968646 - 978080</td>
<td>9,435</td>
<td>1.91 Thailand</td>
<td>767</td>
</tr>
<tr>
<td>chr12</td>
<td>968647 - 978079</td>
<td>9,433</td>
<td>1.93 Cambodia (Pursat)</td>
<td>167</td>
</tr>
<tr>
<td>chr12</td>
<td>968649 - 978078</td>
<td>9,430</td>
<td>1.73 Thailand</td>
<td>755</td>
</tr>
<tr>
<td>chr12</td>
<td>968649 - 978073</td>
<td>9,425</td>
<td>3.48 Thailand</td>
<td>766</td>
</tr>
<tr>
<td>chr12</td>
<td>968650 - 978077</td>
<td>9,428</td>
<td>1.81 Cambodia (Pursat)</td>
<td>231</td>
</tr>
<tr>
<td>chr12</td>
<td>968651 - 978079</td>
<td>9,429</td>
<td>1.67 Cambodia (Pursat)</td>
<td>208</td>
</tr>
<tr>
<td>chr12</td>
<td>968656 - 978083</td>
<td>9,432</td>
<td>2.20 Thailand</td>
<td>710</td>
</tr>
<tr>
<td>chr12</td>
<td>968659 - 978078</td>
<td>9,420</td>
<td>1.69 Cambodia (Pursat)</td>
<td>159</td>
</tr>
<tr>
<td>chr12</td>
<td>968661 - 978072</td>
<td>9,412</td>
<td>1.94 Cambodia (Pursat)</td>
<td>211</td>
</tr>
<tr>
<td>chr12</td>
<td>968672 - 976316</td>
<td>7,645</td>
<td>6.09 Thailand</td>
<td>762</td>
</tr>
<tr>
<td>chr12</td>
<td>968676 - 976313</td>
<td>7,638</td>
<td>2.89 Thailand</td>
<td>760</td>
</tr>
<tr>
<td>chr12</td>
<td>968677 - 978083</td>
<td>9,407</td>
<td>1.69 Cambodia (Pursat)</td>
<td>216</td>
</tr>
<tr>
<td>chr12</td>
<td>968678 - 976310</td>
<td>7,633</td>
<td>1.95 Thailand</td>
<td>753</td>
</tr>
<tr>
<td>chr12</td>
<td>968678 - 978079</td>
<td>9,402</td>
<td>1.98 Cambodia (Pursat)</td>
<td>165</td>
</tr>
<tr>
<td>chr12</td>
<td>968693 - 978077</td>
<td>9,385</td>
<td>1.85 Cambodia (Pursat)</td>
<td>244</td>
</tr>
<tr>
<td>chr12</td>
<td>968696 - 978076</td>
<td>9,381</td>
<td>2.91 Thailand</td>
<td>730</td>
</tr>
<tr>
<td>chr12</td>
<td>968698 - 976310</td>
<td>7,613</td>
<td>2.97 Ghana</td>
<td>478</td>
</tr>
<tr>
<td>chr12</td>
<td>968699 - 978077</td>
<td>9,379</td>
<td>4.58 Thailand</td>
<td>716</td>
</tr>
<tr>
<td>chr12</td>
<td>968701 - 976319</td>
<td>7,619</td>
<td>6.18 Thailand</td>
<td>761</td>
</tr>
<tr>
<td>chr12</td>
<td>968702 - 978079</td>
<td>9,378</td>
<td>1.95 Cambodia (Pursat)</td>
<td>169</td>
</tr>
<tr>
<td>chr12</td>
<td>968704 - 978076</td>
<td>9,373</td>
<td>1.82 Cambodia (Pursat)</td>
<td>197</td>
</tr>
<tr>
<td>chr12</td>
<td>968709 - 978076</td>
<td>9,368</td>
<td>2.52 Cambodia (Pailin)</td>
<td>76</td>
</tr>
<tr>
<td>chr12</td>
<td>968711 - 978073</td>
<td>9,363</td>
<td>2.03 Thailand</td>
<td>733</td>
</tr>
<tr>
<td>chr12</td>
<td>968720 - 976305</td>
<td>7,586</td>
<td>2.22 Thailand</td>
<td>714</td>
</tr>
<tr>
<td>chr12</td>
<td>968729 - 978068</td>
<td>9,340</td>
<td>2.14 Cambodia (Pursat)</td>
<td>120</td>
</tr>
<tr>
<td>chr12</td>
<td>968730 - 976273</td>
<td>7,544</td>
<td>1.98 Ghana</td>
<td>525</td>
</tr>
<tr>
<td>chr12</td>
<td>968773 - 978045</td>
<td>9,273</td>
<td>3.00 Thailand</td>
<td>805</td>
</tr>
<tr>
<td>chr12</td>
<td>968777 - 978045</td>
<td>9,269</td>
<td>2.69 Cambodia (Pursat)</td>
<td>111</td>
</tr>
<tr>
<td>chr12</td>
<td>968779 - 978041</td>
<td>9,263</td>
<td>2.77 Thailand</td>
<td>809</td>
</tr>
<tr>
<td>chr12</td>
<td>968779 - 976281</td>
<td>7,503</td>
<td>3.91 Thailand</td>
<td>803</td>
</tr>
<tr>
<td>chr12</td>
<td>968780 - 978042</td>
<td>9,263</td>
<td>3.26 Cambodia (Pursat)</td>
<td>104</td>
</tr>
<tr>
<td>chr12</td>
<td>968794 - 978040</td>
<td>9,247</td>
<td>2.85 Thailand</td>
<td>810</td>
</tr>
<tr>
<td>chr12</td>
<td>968884 - 978063</td>
<td>9,180</td>
<td>3.16 Thailand</td>
<td>720</td>
</tr>
<tr>
<td>chr12</td>
<td>973602 - 974374</td>
<td>773</td>
<td>2.24 Guinea</td>
<td>PA0177-C</td>
</tr>
<tr>
<td>chr12</td>
<td>973608 - 976162</td>
<td>2,555</td>
<td>2.23 Cambodia (Pursat)</td>
<td>152</td>
</tr>
<tr>
<td>chr12</td>
<td>973609 - 976160</td>
<td>2,552</td>
<td>2.77 Thailand</td>
<td>769</td>
</tr>
<tr>
<td>chr12</td>
<td>973611 - 976157</td>
<td>2,547</td>
<td>2.05 Ghana</td>
<td>541</td>
</tr>
<tr>
<td>Chromosome</td>
<td>Start Position</td>
<td>End Position</td>
<td>Length (bp)</td>
<td>p-value</td>
</tr>
<tr>
<td>------------</td>
<td>----------------</td>
<td>--------------</td>
<td>------------</td>
<td>---------</td>
</tr>
<tr>
<td>chr12</td>
<td>973612</td>
<td>976139</td>
<td>2,528</td>
<td>8.84</td>
</tr>
<tr>
<td>chr12</td>
<td>973612</td>
<td>976162</td>
<td>2,551</td>
<td>3.01</td>
</tr>
<tr>
<td>chr12</td>
<td>973613</td>
<td>976144</td>
<td>2,532</td>
<td>9.25</td>
</tr>
<tr>
<td>chr12</td>
<td>973613</td>
<td>975920</td>
<td>2,308</td>
<td>2.23</td>
</tr>
<tr>
<td>chr12</td>
<td>973615</td>
<td>974272</td>
<td>658</td>
<td>2.21</td>
</tr>
<tr>
<td>chr12</td>
<td>973615</td>
<td>976154</td>
<td>2,540</td>
<td>2.02</td>
</tr>
<tr>
<td>chr12</td>
<td>973621</td>
<td>976147</td>
<td>2,527</td>
<td>1.97</td>
</tr>
<tr>
<td>chr12</td>
<td>973625</td>
<td>974362</td>
<td>738</td>
<td>1.94</td>
</tr>
<tr>
<td>chr12</td>
<td>973628</td>
<td>976117</td>
<td>2,490</td>
<td>2.45</td>
</tr>
<tr>
<td>chr12</td>
<td>973639</td>
<td>974364</td>
<td>726</td>
<td>2.07</td>
</tr>
<tr>
<td>chr12</td>
<td>973645</td>
<td>976138</td>
<td>2,494</td>
<td>2.52</td>
</tr>
<tr>
<td>chr12</td>
<td>1073984</td>
<td>1075014</td>
<td>1,031</td>
<td>1.99</td>
</tr>
<tr>
<td>chr12</td>
<td>1267185</td>
<td>1309576</td>
<td>42,392</td>
<td>1.57</td>
</tr>
<tr>
<td>chr12</td>
<td>1268398</td>
<td>1296032</td>
<td>27,635</td>
<td>1.47</td>
</tr>
<tr>
<td>chr12</td>
<td>1274036</td>
<td>1293832</td>
<td>19,797</td>
<td>1.66</td>
</tr>
<tr>
<td>chr12</td>
<td>1542242</td>
<td>1542923</td>
<td>682</td>
<td>2.06</td>
</tr>
<tr>
<td>chr12</td>
<td>1657036</td>
<td>1680085</td>
<td>23,050</td>
<td>1.59</td>
</tr>
<tr>
<td>chr12</td>
<td>1767965</td>
<td>1769423</td>
<td>1,459</td>
<td>2.26</td>
</tr>
<tr>
<td>chr13</td>
<td>452231</td>
<td>453006</td>
<td>776</td>
<td>4.03</td>
</tr>
<tr>
<td>chr13</td>
<td>965987</td>
<td>969833</td>
<td>3,847</td>
<td>0.24</td>
</tr>
<tr>
<td>chr13</td>
<td>1068874</td>
<td>1084103</td>
<td>15,230</td>
<td>1.45</td>
</tr>
<tr>
<td>chr13</td>
<td>1154958</td>
<td>1183794</td>
<td>28,837</td>
<td>1.46</td>
</tr>
<tr>
<td>chr13</td>
<td>1162411</td>
<td>1181184</td>
<td>18,774</td>
<td>1.62</td>
</tr>
<tr>
<td>chr13</td>
<td>1200973</td>
<td>1201602</td>
<td>630</td>
<td>5.92</td>
</tr>
<tr>
<td>chr13</td>
<td>1212379</td>
<td>1264030</td>
<td>51,652</td>
<td>1.90</td>
</tr>
<tr>
<td>chr13</td>
<td>1423943</td>
<td>1449811</td>
<td>25,869</td>
<td>2.06</td>
</tr>
<tr>
<td>chr13</td>
<td>1430177</td>
<td>1449822</td>
<td>19,646</td>
<td>2.23</td>
</tr>
<tr>
<td>chr13</td>
<td>1428865</td>
<td>1429987</td>
<td>1,123</td>
<td>0.34</td>
</tr>
<tr>
<td>Chromosome</td>
<td>Start - End</td>
<td>Length</td>
<td>Proportion</td>
<td>Country</td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>--------</td>
<td>------------</td>
<td>------------------</td>
</tr>
<tr>
<td>chr13</td>
<td>1428908 - 1429972</td>
<td>1,065</td>
<td>0.32</td>
<td>Mali</td>
</tr>
<tr>
<td>chr13</td>
<td>1428953 - 1429993</td>
<td>1,023</td>
<td>0.31</td>
<td>Ghana</td>
</tr>
<tr>
<td>chr13</td>
<td>1429079 - 1429975</td>
<td>897</td>
<td>0.34</td>
<td>Burkina Faso</td>
</tr>
<tr>
<td>chr13</td>
<td>2623768 - 2624384</td>
<td>617</td>
<td>13.80</td>
<td>Kenya</td>
</tr>
<tr>
<td>chr14</td>
<td>95794 - 109934</td>
<td>14,141</td>
<td>2.12</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>282835 - 300646</td>
<td>17,812</td>
<td>1.86</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>282842 - 300646</td>
<td>17,805</td>
<td>1.88</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>289091 - 298886</td>
<td>9,796</td>
<td>3.12</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>289396 - 298896</td>
<td>9,501</td>
<td>3.22</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>289401 - 298894</td>
<td>9,494</td>
<td>3.58</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>289401 - 298889</td>
<td>9,489</td>
<td>1.99</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>289402 - 298900</td>
<td>9,499</td>
<td>1.68</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>289402 - 298898</td>
<td>9,497</td>
<td>1.68</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>289405 - 298895</td>
<td>9,491</td>
<td>1.74</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>289407 - 298885</td>
<td>9,479</td>
<td>2.62</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr14</td>
<td>289407 - 298894</td>
<td>9,488</td>
<td>1.74</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>289408 - 298875</td>
<td>9,468</td>
<td>2.06</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr14</td>
<td>289409 - 298886</td>
<td>9,478</td>
<td>1.71</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>289409 - 298869</td>
<td>9,461</td>
<td>2.01</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr14</td>
<td>289411 - 298888</td>
<td>9,478</td>
<td>1.83</td>
<td>Cambodia (Pursat)</td>
</tr>
<tr>
<td>chr14</td>
<td>289417 - 298882</td>
<td>9,466</td>
<td>1.83</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr14</td>
<td>289421 - 298893</td>
<td>9,473</td>
<td>1.81</td>
<td>Cambodia (Tasanh)</td>
</tr>
<tr>
<td>chr14</td>
<td>367404 - 368300</td>
<td>897</td>
<td>1.94</td>
<td>Kenya</td>
</tr>
<tr>
<td>chr14</td>
<td>589978 - 592404</td>
<td>2,427</td>
<td>2.61</td>
<td>Guinea</td>
</tr>
<tr>
<td>chr14</td>
<td>695480 - 704797</td>
<td>9,318</td>
<td>1.55</td>
<td>Thailand</td>
</tr>
<tr>
<td>chr14</td>
<td>1099777 - 1101133</td>
<td>1,357</td>
<td>0.45</td>
<td>Mali</td>
</tr>
<tr>
<td>chr14</td>
<td>1152175 - 1259839</td>
<td>107,665</td>
<td>1.73</td>
<td>Gambia</td>
</tr>
<tr>
<td>chr14</td>
<td>1185390 - 1186146</td>
<td>757</td>
<td>2.37</td>
<td>Gambia</td>
</tr>
<tr>
<td>chr14</td>
<td>1185425 - 1186197</td>
<td>773</td>
<td>2.94</td>
<td>Burkina Faso</td>
</tr>
<tr>
<td>chr14</td>
<td>Start - End</td>
<td>Length</td>
<td>Distance (kbp)</td>
<td>Distance (Gbp)</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------</td>
<td>--------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td></td>
<td>1733015 - 1748645</td>
<td>15,631</td>
<td>1.99</td>
<td>1.99</td>
</tr>
<tr>
<td></td>
<td>1734101 - 1735000</td>
<td>900</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>2020371 - 2021382</td>
<td>1,012</td>
<td>2.74</td>
<td>2.74</td>
</tr>
<tr>
<td></td>
<td>2257914 - 2259309</td>
<td>1,396</td>
<td>2.78</td>
<td>2.78</td>
</tr>
<tr>
<td></td>
<td>2620036 - 2620833</td>
<td>798</td>
<td>2.14</td>
<td>2.14</td>
</tr>
<tr>
<td></td>
<td>2698050 - 2699156</td>
<td>1,107</td>
<td>2.44</td>
<td>2.44</td>
</tr>
</tbody>
</table>
REFERENCES


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


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


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.


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


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


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


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.


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


104. Cowman AF, Galatis D, Thompson JK: **Selection for mefloquine resistance in Plasmodium falciparum is linked to amplification of the pfmdr1 gene and cross-


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


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


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


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


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


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


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


