Gene Copy Number Variation in Natural Populations of Plasmodium falciparum

Thesis

Gene Copy Number Variation in Natural Populations of *Plasmodium falciparum*

A thesis

submitted in fulfilment

of the requirements for the degree of

Doctor of Philosophy at

the Open University, UK

by

Joan Jebet Simam

KEMRI/Wellcome Trust Research Programme

March 2015

DATE OF SUBMISSION: 26 MARCH 2015
DATE OF AWARD: 16 SEPTEMBER 2015

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

Gene copy number variants (CNVs), which consist of gene deletions and amplifications contribute to the great diversity in the *Plasmodium falciparum* genome. CNVs may influence the expression of genes and hence may affect important parasite phenotypes such as virulence, drug resistance, persistence and transmissibility. The hypothesis underlying the studies in this thesis is that CNVs may be important for adaptation of the parasite to its variable environments. To investigate this hypothesis, a population wide survey of CNVs in 183 fresh field isolates from four populations with different transmission intensities was conducted. To detect CNVs, comparative genome hybridization was performed using a 70mer microarray. This is the first large scale survey for CNVs in natural populations of parasites. A total of 98 different CNVs, consisting of 225 genes, were identified. Various systematic aspects that could affect detection of CNVs were explored and the population of origin of the isolate was found to be the only factor that affects CNV detection. Some of these CNVs showed high differentiation in frequency between populations suggestive of the action of directional selection. Other CNVs showed no or low differentiation in frequencies between populations, indicative of action of neutral evolutionary processes. Validation of the CNVs identified using microarrays was done using whole genome sequencing. Very low concordance was observed between the CNVs identified by the two technologies. These differences may be attributed to technical and analytic differences between the two technologies. Furthermore, the effect of CNVs on gene expression levels was analysed. A number of CNVs were found to be significantly associated (positively or negatively) with the expression levels of genes located inside and also outside the CNVs.

Acknowledgements

I wish thank all the people who helped me through my PhD experience. First and foremost, I would like to express deepest gratitude to my first supervisor/director of studies, Assoc. Prof. Margaret Mackinnon, who has been extremely supportive throughout the period. I am extremely grateful for her guidance, motivation, mentorship, patience, immense expertise and knowledge that has contributed to great PhD experience. I would also like to sincerely thank my supervisor, Prof. Kevin Marsh, for insightful discussions and comments on my research project. I am grateful to my other supervisior, Assoc. Zbynek Bozdech, for my initial training in some lab experiments and providing the microarray slides used in the project.

I would like to thank my fellow members of Mackinnon's group; Mary Nyonda, Joyce Mwongeli and Dr. Martin Rono for the stimulating discussions and fun times we had in the lab. I am also very grateful for the help they offered me in the labs and the use of the transciptome data they generated. I am extremely grateful to Esther Kiragu, whom I tirelessly worked with in performing sequencing, Dr. Etienne De Villiers for consultations on sequence analysis, and Assoc. Prof. Pete Bull and Dr. Vandana Thathy for productive discussions on my project. The Pathogen, Vector and Human Biology Department at KEMRI/Wellcome Trust Research Programme offered a platform for continuous scientific growth.

I am forever grateful my family; my husband Noah Murbiy and daughter Ashlynn Chemutai for their unending support and understanding, my parents, Paul and Zipporah Simam for instilling in me the importance of hard work, education and discipline and my siblings for their continued support.

I would also like to acknowledge the funders of my PhD experience. My PhD studentship has been supported by The Wellcome Trust Strategic Award to KEMRI/Wellcome Trust Research

Programme. My PhD research project was supported by The Wellcome Trust Project Grant (088634) to Assoc. Prof. Mackinnon, Prof. Kevin Marsh and Assoc. Prof. Zbynek Bozdech.

Table of Contents

1 Chapter 1: Introduction 15

1.1 Background 15

1.1.1 Malaria epidemiology 15

1.1.2 Genetic diversity of *Plasmodium falciparum* 17

1.1.3 Gene copy number variation as a source of genetic diversity in *P. falciparum* . 19

1.2 Methods for detecting CNVs in *P. falciparum* 21

1.2.1 Whole genome/chromosome scan for CNVs 21

1.2.2 Targeted loci CNV detection 28

1.3 Mechanisms of CNV formation 30

1.3.1 Homologous recombination 30

1.3.2 Replicative mechanisms 33

1.4 Functional impact of CNVs 36

1.5 The role of CNVs in *P. falciparum* 41

1.6 Scope of the thesis 42

2 Chapter 2: Materials and methods 46

2.1 Population study of CNVs by microarrays 46

2.1.1 Study population 46

2.1.2 Sample processing 46

2.1.3 Genomic DNA extraction 47

2.1.4 Multiplicity of infection 48

2.1.5 Comparative genomic hybridization (CGH) 49

2.1.6 Microarray data analysis 54

2.1.7 Statistical analysis 56

2.2 Confirmation of CNVs by whole genome sequencing using the Personal Genome Machine (PGM™) Ion Torrent 58

2.2.1 Samples 58

2.2.2 Genomic DNA extraction 58

2.2.3 Quantification of the proportions of human and parasite DNA using real time PCR 58

2.2.4 Human DNA depletion by DNase treatment 60

2.2.5 Library preparation of gDNA for sequencing 62

2.2.6 Sequence data analysis 64

2.3 Transcriptome profiling using microarrays 65
2.3.1 RNA extraction 66
2.3.2 cDNA synthesis, amplification and hybridisation 66
2.3.3 Statistical analysis 68
3 Chapter 3: Population-wide survey of gene copy number variation in *P. falciparum* isolates from Africa 71
3.1 Introduction 71
3.2 Methods 72
3.2.1 Experimental strategy 72
3.2.2 Sample populations 74
3.2.3 Quality control of microarray data 75
3.2.4 Pre-processing of microarray data 75
3.2.5 CNV detection using Genome Alteration Detection Analysis (GADA) 76
3.2.6 CNV definition 76
3.2.7 Analysis of systematic effects on CNV detection 78
3.2.8 Assessing reproducibility of microarrays 78
3.2.9 Gene set enrichment analysis of CNVs 79
3.2.10 Population genetics analysis 79
3.3 Results 79
3.3.1 Quality of microarray data 79
3.3.2 Background correction and normalisation of microarray data 80
3.3.3 CNV detection using Genome Alteration Detection Analysis (GADA) 82
3.3.4 Reproducibility 82
3.3.5 Systematic effects on CNV detection probabilities 86
3.3.6 CNVs detected in field isolates 87
3.3.7 Population genetics of CNVs 87
3.3.8 Comparison with published literature 96
3.3.9 Functional gene set enrichment of CNVs 98
3.4 Discussion 100
4 Chapter 4: Confirmation of CNVs detected using microarray by whole genome sequencing 105
4.1 Introduction 105
4.2 Methods 106

4.2.1 Library preparation 106
4.2.2 Sequence data processing for CNV detection 107
4.3 Results 111
4.3.1 Sequence quality 111
4.3.2 Challenges in mapping of whole genome sequence data of *P. falciparum* 114
4.3.3 Gene copy number variation definition 117
4.3.4 CNVs detected by next-generation sequencing 118
4.3.5 CNVs detected by both microarrays and sequencing 121
4.3.6 CNVs exclusively detected in sequence data or microarray data 121
4.4 Discussion 126
5 Chapter 5: Effects of CNVs on gene expression levels 132
5.1 Introduction 132
5.2 Methods 133
5.2.1 Samples 133
5.2.2 Gene copy number variation 133
5.2.3 Gene expression data 133
5.2.4 Analysis of expression in relation to gene copy number 134
5.3 Results 135
5.3.1 Direct dosage effect of CNVs 135
5.3.2 Direction of CNV effects on expression 137
5.3.3 CNV effects on the expression of genes located outside CNV intervals 142
5.4 Discussion 145
6 General discussion 150
6.1 Study objectives 150
6.2 Key findings 151
6.2.1 CNVs are prevalent in natural populations of *P. falciparum* 151
6.2.2 CNVs may be under selection 155
6.2.3 Poor overlap between CNVs identified by microarrays and sequencing 156
6.2.4 CNVs affect the levels of transcription of genes within and outside CNV boundaries 157
6.3 Future directions 158
7 References 160
8 Appendix 178

8.1 Appendix 1.1. A summary of genome-wide studies of CNVs in *P. falciparum* using microarrays and next generation sequencing technologies. 178

8.2 Appendix 3.1. Overlap between CNVs identified in our study compared to published studies. 180

8.3 Appendix 3.2. Potentially novel CNVs 184

8.4 Appendix 4.2. Number of CNV genes detected and fraction of gene length identified to be copy number variable. 191

List of figures

Figure 1.1. The distribution of predicted *Plasmodium falciparum* parasite rates (*Pf*PR) in Africa in 2000 and 2010

Figure 1.2. An illustration of the approaches used to detect CNVs using next generation sequencing

Figure 1.3. Models of CNVs formation through non-allelic homologous recombination (NAHR)

Figure 1.4. Model for formation of CNVs by non-homologous end joining (NHEJ)

Figure 1.5. Models of CNVs formation by replicative mechanisms mediated by microhomology

Figure 2.1. Size of amplified DNA using random nonamers

Figure 2.2. Image of a hybridized array

Figure 2.3. Quantification of parasite and human DNA using the standard curve method

Figure 3.1. Overview of the process of CNV detection using microarrays

Figure 3.2. CNV detection using *GADA*

Figure 3.3. A schematic representation of the definition of CNVs breakpoints

Figure 3.4. Summary of quality of microarray data

Figure 3.5. Pre-processing of microarray data

Figure 3.6. Segmentation using GADA on a region on chromosome 13 (CNV: cnv13_473)

Figure 3.7. Effect of different stringency measures on CNV calling

Figure 3.8. Reproducibility of microarray data and CNV detection

Figure 3.9. Systematic effects on CNV prevalence

Figure 3.10. General properties of CNVs detected using microarrays

Figure 3.11. F_{ST} estimates obtained from pairwise population comparisons

Figure 3.12. Population frequencies of CNVs with top 4 highest F_{ST} values

Figure 3.13. Population frequencies of CNVs showing signs of selection in Kilifi population

Figure 3.14. Frequencies of CNVs showing signs of purifying selection in single populations

Figure 3.15. Overlap between CNVs identified in this study and published studies

Figure 4.1. Overview of the process of CNV detection using whole genome sequencing

Figure 4.2. Plots of quality scores and nucleotide content of sequenced reads

Figure 4.3. Effect of GC content and uniqueness of a region on sequence coverage

Figure 4.4. Normalisation of read coverage across samples

Figure 4.5. CNV calling using *cn.MOPS*

Figure 4.6. Genomic location of CNVs identified in CGH and sequence data in 18 isolates.

Figure 4.7. The number and length of CNV genes detected by five methods using *cn.MOPS*

Figure 4.8. Positive and negative predictive values (PPV and NPV) of CNVs genes detection in five genome sequences analysed by *cn.MOPS* and CGH (cn.MOPS) with CGH (GADA) as the gold standard

Figure 4.9. Amplification of GTP cyclohydrolase gene 1

Figure 4.10. A CNV identified in chromosome 9 using microarrays but not using sequencing

Figure 4.11. Deletion of *clag 3.1* gene (PF3D7_0302500) on chromosome 3

Figure 5.1. Correlation between CGH $\log_2$ ratio of 221 CNV genes within 95 CNVs and their corresponding $\log_2$ expression ratio

Figure 5.2. Direction of CNV effect on level of gene expression

Figure 5.3. Association between the CGH ratio and expression ratio of genes located in chromosome 9 region containing a deletion

Figure 5.4. CNV effect on expression of genes located outside the CNVs

Figure 5.5. Effect of CNV on expression of genes at a distance from CNV

Figure 5.6. Increased copy number of PF3D7_1248600 gene shown to affect expression of genes located on different chromosomes

List of tables

Table 1.1. A summary of published genome wide studies of CNVs in *P. falciparum*

Table 2.1. Thermocycling conditions for PCR amplification of Klenow reaction product

Table 2.2. Thermocycling conditions for qPCR assay for human and parasite DNA quantification

Table 2.3. Proportion of parasite DNA in samples determined by qPCR and also obtained from whole genome sequence data

Table 2.4. Proportion of parasite DNA sequences in whole genome sequence data of samples after DNase treatment

Table 2.5. Thermocycling conditions for PCR amplification of cDNA

Table 3.1. Characteristics of the four sample populations

Table 3.2. Table showing the kappa value of each replicate pair of samples

Table 3.3. Genes and gene functions of CNVs that exhibit the greatest population differentiation

Table 3.4. Genes and gene functions of CNVs that appear to be under positive selection in Kilifi populations

Table 3.5. Genes and gene functions of CNVs that appear to be under purifying selection in one of the four populations

Table 3.6. Functional gene categories showing significant enrichment

Table 4.1. Summary statistics of whole genome sequence data of 22 *P. falciparum* field isolates

Table 4.2. A list of genes CNV genes detected in CGH (GADA) and at least one sequence method with the fraction of the number of samples with CNV in a method that had the CNV detected in both the method and CGH (GADA)

Table 5.1. Genes in a CNV showing both negative and positive associations between copy CGH and expression data

Table 5.2. Genes showing direct dosage effect of CNVs on expression

List of abbreviations

BIR: Break induced replication

BWA: Burrows-Wheeler aligner

CBS: Circular binary segmentation

cDNA: complementary DNA

CGH: Comparative genomic hybridisation

cn.MOPS: Copy number estimation by a mixture of Poissons

CNVs: Copy number variations

DNA: Deoxyribonucleic acid

FoSTeS: Fork stalling and template switching

F_{ST}: F-statistics

GADA: Genome alteration detection analysis

gDNA: Genomic deoxyribonucleic acid

indels: Insertion and deletion

iRBCs: Infected red blood cells

MOI: Multiplicity of infection

NAHR: Non-allelic homologous recombination

NHEJ: Non-homologous end joining

PGM: Personal genome machine

qPCR: quantitative polymerase chain reaction

RNA: Ribonucleic acid

SNPs: Single nucleotide Polymorphisms

VSA: Variant surface antigen

WHO: World Health Organisation

Chapter 1

Introduction

1 Chapter 1: Introduction

1.1 Background

Malaria is a disease caused by a protozoan parasite, *Plasmodium*, and transmitted by infected female mosquitoes of the genus Anopheles to the human host during blood feeding. The species of *Plasmodium* that infect humans include *P. falciparum, P. vivax, P. malariae, P. knowlesi* and two *P. ovale* subspecies, i.e., *P. ovale curtisi and P. ovale wallikeri. P. falciparum* causes the most severe form of the disease in tropical regions. The parasite's life cycle involves both the mosquito vector and the human host. The parasite exists in both sexual and asexual forms and is known to undergo morphological and metabolic changes throughout its life cycle. The parasite genome is approximately 23Mbp in size, with nucleotide content of 80% AT. It is comprised of 14 chromosomes with approximately 5300 genes covering about half of the genome length (Gardner et al. 2002).

1.1.1 Malaria epidemiology

Malaria continues to be a leading contributor to the global burden of disease in developing countries (Snow et al. 2005, Noor et al. 2014). According to WHO report 2013, an estimated 198 million malaria cases occurred in 2013, with 90% of the 584 000 deaths occurring in Sub-Saharan Africa (WHO 2014). It is also estimated that 57% of the population in Africa is resident in regions with at least moderate to high malaria transmission (Figure 1) (Noor et al. 2014). Nonetheless, there is evidence, in some regions, of a decline in prevalence of parasite positive individuals over the period 1990 to 2010 (Figure 1) (Noor et al. 2014, Murray et al. 2012, WHO 2014). Some of this has been attributed to interventions including the distribution of insecticide-treated bed nets, the replacement of antimalarial drugs to which *P. falciparum* has developed resistance (e.g., sulfadoxine-pyrimethamine) with effective artemisinin

combination therapy, and better access to treatment (O'Meara et al. 2008, Phillips-Howard et al. 2003, Barnes et al. 2005).

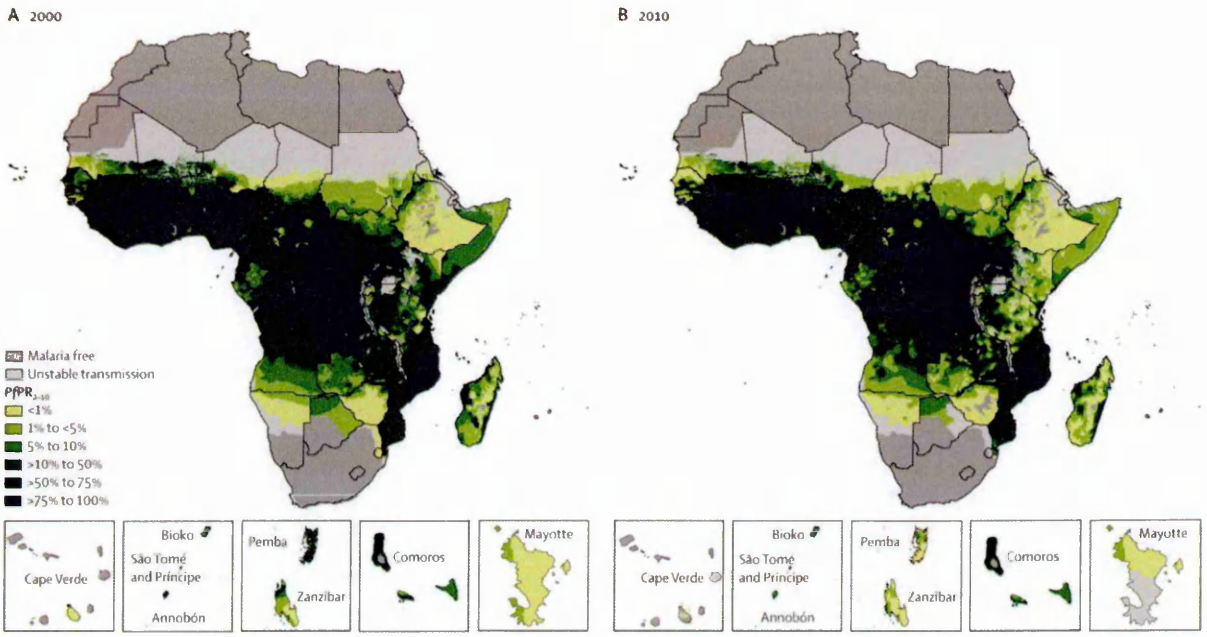


Figure 1.1. The distribution of predicted *Plasmodium falciparum* parasite rates (*Pf*PR) in Africa in 2000 and 2010.
P. falciparum parasite rates (PfPR) predicted from data obtained from surveys of the prevalence of *P. falciparum* infections standardized to the 2-10 years age group in the year A) 2000 and B) 2010. Darker shades represent regions with high parasite prevalence. Image reproduced from Noor et al. 2014 under the terms of the Creative Commons Attribution License.

However, in some regions the decline was observed before the introduction of interventions: thus the drivers of this decline are poorly understood (Okiro et al. 2010). Despite this general decline in malaria, there is ongoing high disease prevalence in many regions (Noor 2014), emergence of resistance to the mainstay drug, artemisinin, in the Greater Mekong Subregion (Mok et al. 2011, Miotto et al. 2013, Takala-Harrison et al. 2013, Dondorp et al. 2009, Noedl et al. 2008), and widespread insecticide resistance in vector populations (Corbel et al. 2012). Thus there is still a great need for additional tools for prevention and treatment in order to further decrease the burden of malaria.

1.1.2 Genetic diversity of *Plasmodium falciparum*

The malaria parasite is highly genetically diverse. This characteristic is thought to enable the parasite to survive in its changing environment, including escaping from specific antigenic immune responses and resisting antimalarial drug treatment (Volkman et al. 2007, Jeffares et al. 2007, Anderson, Patel and Ferdig 2009, Soulama et al. 2011, Mackinnon and Marsh 2010).

The forms of variation in the genome range from large structural variations to single nucleotide changes. The structural variations include chromosomal size polymorphisms, gene copy number variation (CNVs), inversions and translocations. Other forms of variation include small insertions and deletions (indels) and single nucleotide polymorphism (SNPs) (Volkman et al. 2007, Kidgell et al. 2006). Genomic variation arises as a result of mutations that occur spontaneously or are caused by mutagens e.g., reactive oxygen species (Branzei and Foiani 2008). Spontaneous mutations may be caused by errors during DNA replication that remain uncorrected and also failure in DNA repair mechanisms to restore a damaged DNA strand to its original state. Genetic diversity is also generated by genetic recombination that allows exchange of genetic segments affecting phenotype. In *P. falciparum*, recombination hotspots have been identified and are characterized by AT-rich repeats, a 12bp G/C-rich motif with 3bp repeat (Jiang et al. 2011) and occur mostly at chromosome ends (Mu et al. 2005). The rate of recombination is influenced by the diversity of parasite populations and malaria transmission intensities (Mu et al. 2010). These different types of genomic variation in a population are shaped by selection pressures, population isolation, genetic drift and migration (Mackinnon and Marsh 2010). Genetic drift is related to the effective population size. Populations with small effective population sizes experience more genetic drift compared to those with large sizes. The increase in the number of variants in a population is limited by the carrying capacity of the population.

A number of genomic variants have been linked to parasite phenotypes that are crucial for survival and spread, and which also contribute to disease pathogenesis. First, genetic variability in antigens expressed on the surface of the merozoite and on the infected red cell surface are thought to allow the parasite to evade host immune responses by expressing novel antigenic types (Amambua-Ngwa et al. 2012, Tetteh et al. 2009, Bull et al. 1998, Wright and Rayner 2014). Second, erythrocyte invasion, a key determinant of parasite replication rate, is known to involve multiple parasite ligands that are polymorphic in nature. Third, the parasite adhesion phenotype that results in obstruction of blood capillaries and prevent clearance of parasite by the spleen, is caused by a family of highly polymorphic genes referred to as *var* genes encoding *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) (Scherf, Lopez-Rubio and Riviere 2008). PfEMP1 mediates adhesion of infected erythrocytes to endothelial cells (cytoadherence) (Su et al. 1995) and uninfected erythrocytes (rosetting) and is also thought to be involved in adhesion of infected erythrocytes to platelets forming clumps (platelet-mediated clumping) (Rowe et al. 2009). Two other variant surface antigens, RIFINS and STEVOR are thought to be linked to the adherent phenotype (Niang et al. 2014). Fourth, production of gametocytes, the asexual forms of the parasite, a key determinant of malaria transmission has been shown to be lost, in some studies, during long term *in vitro* adaptation of some parasite lines and has been associated with a large deletion on chromosome 9 (Robinson et al. 2011, Cheeseman et al. 2009, Kidgell et al. 2006, Jiang et al. 2008b, Mackinnon et al. 2009, Mok et al. 2011, Bozdech et al. 2003a, Ribacke et al. 2007, Nair et al. 2008, Pologe and Ravetch 1988, Wellems et al. 1987, Nair et al. 2010, Day et al. 1993). Lastly, antimalarial drug resistance has been associated with increased copy number in genomic regions containing three genes (multi drug resistance 1 (*pfmdr1*), GTP cyclohydrolase 1 (*gch1*) and falcipain-2) and also SNPs in Kelch 13 gene, chloroquine resistance transporter gene and dihydrofolate reductase (dhfr) gene (Van Tyne et al. 2011,

Ariey et al. 2014, Miotto et al. 2013, Plowe, Kublin and Doumbo 1998, Singh and Rosenthal 2004, Fidock et al. 2000). From these studies, it is evident that the understanding of malaria parasite genetic diversity is crucial since it offers insight into genomic loci that affect important parasite phenotypes and hence are potential targets for vaccines, therapeutics and tools for monitoring of drug resistance.

1.1.3 Gene copy number variation as a source of genetic diversity in *P. falciparum*

The recent interest in CNVs in malaria parasites is driven by increasing evidence of their role in adaptation, evolution and disease in a number of organisms (Anderson et al. 2009, Henrichsen, Chaignat and Reymond 2009, Tam et al. 2009, Kirov et al. 2012, Craddock et al. 2010, Angstadt et al. 2013, Palli et al. 2013) and the advances in technologies that enable high-throughput genome-wide scans of CNVs (Mackinnon et al. 2009, Kidgell et al. 2006, Ribacke et al. 2007, Cheeseman et al. 2009, Jiang et al. 2008b). In *P. falciparum,* CNVs have been identified in long- and short-term cultured lines (Mackinnon et al. 2009, Ribacke et al. 2007, Mok et al. 2011, Carret et al. 2005, Kidgell et al. 2006, Jiang et al. 2008a, Cheeseman et al. 2009, Dharia et al. 2009, Samarakoon et al. 2011a, Samarakoon et al. 2011b, Bozdech, Mok and Gupta 2013) and also in fresh clinical isolates (Robinson et al. 2011, Ribacke et al. 2007). From these CNV studies, it is estimated that between 0.3% - 1% of the parasite's genome is subject to variation in gene copy number (Anderson et al. 2009), a fraction of the genome larger than that represented by SNPs (Volkman et al. 2007, Miotto et al. 2013). This implies potential significant effects on phenotypic variability. These CNV studies employed different platforms, parasite strains and CNVs calling strategies (detailed explation in section 1.2.1.3.1) and hence, as observed (Table 1), their results are expected to vary in the number and genomic location of CNVs. Furthermore, a number of CNVs have been found to typically arise during culture adaptation of parasite lines, thus introducing a further source of variation

in CNVs reported so far (Nair et al. 2010, Mackinnon et al. 2009). None of these studies has, on large scale, examined CNVs in natural populations of parasites (Appendix 1.1). Thus there is a need for large scale exploration of CNVs in fresh clinical isolates to determine the extent of these variants in the genome and their potential role in adaptive evolution of natural populations of parasites.

Table 1.1. A summary of published genome wide studies of CNVs in *P. falciparum*

Platform	Type	No. of CNV genes	No. of isolates studied	CNV detection analysis method	References
Microarrays	Spotted arrays (70mer)	144	1	Difference <50% of total intensity	(Bozdech et al. 2003a)
		82	9	limma (B-statistics)	(Ribacke et al. 2007)
		324	5	1.5 fold difference	(Mackinnon et al. 2009)
		138	6	R-GADA	(Mok et al. 2011)
	Affymetrix arrays (25mer)	177	7	% Signal reduction	(Carret et al. 2005)
		149	14	MOID algorithm	(Kidgell et al. 2006)
		390	4	Partek Genomic suite v6.3	(Jiang et al. 2008b)
		186	16	Threshold set at absolute $\log_2$ ratio > 1	(Cheeseman et al. 2009)
		~156	4	MOID algorithm	(Dharia et al. 2009)
	NimbleGen arrays	537	37	Nexus Copy Number 3.0 software	(Samarakoon et al. 2011a)
Next generation sequencing	Illumina	7	5	Read depth analysis & Paired end mapping	(Robinson et al. 2011)
	Roche 454	7	2	Read depth analysis	(Samarakoon et al. 2011a)
	Illumina	>100	5	Read depth analysis using 3 tools	(Sepulveda et al. 2013)

R-GADA: Genome Alteration Detection Analysis in R. MOID: Match-Only Integral Distribution. Limma: Linear Models for Microarray Analysis

1.2 Methods for detecting CNVs in *P. falciparum*

CNV detection molecular methods have rapidly evolved in the last decade enabling whole genome scans to the resolution level of single nucleotides. The methods are discussed below.

1.2.1 Whole genome/chromosome scan for CNVs

1.2.1.1 Pulse field gradient gel electrophoresis (PFGE)

The early evidence for structural variation in *P. falciparum* was the detection of chromosome size polymorphism in isolates using pulse field gradient (PFG) gel electrophoresis currently referred to as pulse field gel electrophoresis (PFGE) (Corcoran et al. 1986, Kemp et al. 1985, Van der Ploeg et al. 1985). Variation in chromosome lengths could be as a result of loss/addition of whole genes leading to change in the number of gene copies (CNVs) in the parasite genome. PFGE involves fractionation of chromosomal DNA in a 1.5% -2% agarose gel with a voltage applied across the gel that is constantly changing in direction (Schwartz and Cantor 1984) . This enables separation of the chromosomes based on their length ranging from 30 kilobases to 3000 kilobases. Chromosome sizes were estimated by comparison with known yeast chromosome molecular weight sizes.

1.2.1.2 Optical mapping

Optical mapping involves the development of ordered restriction maps of DNA fragments (Cai et al. 1995). The restriction map of the whole genome of *P. falciparum* was generated in 1999 by Lai and colleagues and provided a scaffold of the entire genome onto which contigs could be aligned (Lai et al. 1999). The starting material of the assay was either chromosomal DNA separated using pulse field gel electrophoresis (PFGE) or randomly fragmented DNA (250bp-3000kbp) mounted on a surface together with a sizing standard (Lai et al. 1999, Jing et al.

1999). The mounted DNA was then digested with restriction enzymes of choice (*Bam*H1 and *Nhe*1 used) to generate fragments that are stained with a fluorescent reagent that intercalates into DNA. The image was acquired using microscopy, processed and the maps constructed. The restriction maps were assembled to form restriction maps/contigs that were then mapped onto the available whole genome optical map. To detect CNVs, the full genome/contig restriction maps of test genomes were compared to a reference map which was generated *in silico* using available complete genome sequence (Riley et al. 2011). Riley and colleagues generated full maps of four cultured *P. falciparum* strains and were able to detect deletions, duplications and inversions with sizes ranging from 3.5kb to 78kb (Riley et al. 2011).

1.2.1.3 Microarrays

Completion of sequencing of the *P. falciparum* genome of strain 3D7 in 2002 (Gardner et al. 2002) facilitated the development of oligonucleotide arrays that have been used to perform comparative genome hybridization (CGH) (Bozdech et al. 2003a). CGH involves comparison of hybridization of a labelled test DNA and a reference DNA on a glass slide spotted with oligonucleotides. The first published array used to detect CNVs in *P. falciparum* was a 25-mer affymetrix array with 298,752 probes targeting approximately 5,179 genes (Kidgell et al. 2006, Carret et al. 2005). The major limitation of this array was the large quantity of DNA required for the experiment (15 μg) which would not be suitable to assay small quantities of genomic DNA (gDNA) obtained from clinical samples. However, this problem has now been overcome by the development of whole genome amplification methods that end at the log linear phase of amplification thus maintain the relative abundance of genomic content and also minimize PCR errors (Petalidis et al. 2003, Carret et al. 2005, Mackinnon et al. 2009).

Carret and others tested various whole genome amplification methods including multiple displacement amplification (MDA) and ligation-mediated PCR (LMP), to overcome this problem (Carret et al. 2005). LMP involves attachment of adapter sequence to fragmented gDNA and using a primer complementary to the adapter for amplification. MDA, which uses random hexamers, had PCR products that were more similar to the non-amplified gDNA as compared to LMP products. The CGH intensity signal of genes were observed to be similar between the hybridization experiments of MDA products and non-amplified gDNA.

A second 70-mer array developed by Bozdech and others containing 10,680 probes targeting 5,343 *P. falciparum* genes, has been used on both non-amplified and randomly amplified gDNA samples to detect CNVs in both cultured and fresh clinical samples (Ribacke et al. 2007, Bozdech et al. 2003a, Mackinnon et al. 2009, Mok et al. 2011, Hu et al. 2007) . Two additional studies have used a higher resolution array than the aforementioned two, the Affymetrix PFSANGER GeneChip, a 25mer tiling array with about 2.5 million probes spanning exons, coding regions, intron/exon junction and intergenic regions. Both studies used 10-12 μg of non-amplified DNA (Cheeseman et al. 2009, Jiang et al. 2008b).

The main advantages of microarrays are that the assay is cheaper compared to sequencing, the arrays can easily be updated for new probes, and less nucleic material is used, especially when amplification of gDNA is performed, compared to earlier technologies. The main drawbacks of the use of microarray technology in CNV detection include inability to detect rare variants in mixed parasite populations, the ability to assay only genomic regions that have been previously characterised, that correct mapping of variants in multi-gene families is impossible, that the presence of unknown polymorphisms in genes targeted by the probes affects the

hybridization results and that the boundaries of CNVs cannot be precisely mapped if probes are non-overlapping.

1.2.1.3.1 CNVs detection analysis methods for microarray data

A number of CNV detection tools for CGH data are available and differ by the type of array platform, input data (log_2 intensity ratios or absolute intensity values), the statistical models used, detection thresholds applied, ability to detect CNV boundaries, ploidy of genome and the way in which the reference genome is used. Some of the tools available that are applicable to my study, i.e., compatible with CGH data from two colour arrays (log_2 intensity ratio calculated against intensity for a reference genome), haploid genome analysis and that are freely available will be highlighted.

In an ideal two colour array CGH experiment, the presence of similar copies of a gene between the test and reference samples are indicated by log_2 intensity ratio, i.e., the difference between log_2 hybridization signal intensities of a test and reference sample. In the case of a 2-fold difference in gene copy number, the log_2 intensity ratio is ideally 1 or -1. However, due to noise in microarray data, there is uncertainty in the thresholds on which gene copy number values can be assigned.

One of the methods used for identifying CNVs in *P. falciparum* has been to apply thresholds on the log_2 intensities ratios and the minimum number of consecutive probes that meet a set threshold (Cheeseman et al. 2009) . The second method involves identifying array probes that statistically significantly differ in hybridization between isolates using B-statistics in *limma* package in R (Smyth 2004) then applying cut-offs on the number of consecutive probes and the direction of change (Ribacke et al. 2007). Other tools developed are based on segmentation

analysis (Pique-Regi, Caceres and Gonzalez 2010, Olshen et al. 2004) and the Smith-Waterman algorithm (Price et al. 2005).

In segmentation analysis, the $\log_2$ intensity ratio along a chromosome is partitioned into segments of equal copy number. The first segmentation algorithm developed for arrays is Circular Binary Segmentation (CBS) (Olshen et al. 2004). CBS recursively detects the points of change of copy number (start and end of segments) along a chromosome and statistically tests whether the means of segments (the mean of $\log_2$ intensity ratio of probes within the segments) bordering each other differ. A package was later developed in R known as DNACopy that uses CBS to detect CNVs (Venkatraman and Olshen 2007). A second algorithm, Genome Alteration Detection Algorithm (GADA), also available in R, represents $\log_2$ intensity ratio along a chromosome as piecewise constant (PWC) vector then uses Bayesian learning to identify change-points and segment means along a chromosome (Pique-Regi et al. 2010). It then tests for significance of the change-points using t-statistics based on the segment means and variance. GADA's performance was found to be faster than CBS and as accurate as CBS (Pique-Regi et al. 2010).

1.2.1.4 Next Generation Sequencing (NGS)

The exploration of CNVs in the parasite genome to single nucleotide resolution has been made possible by next generation sequencing (NGS). Whole genome and exome sequencing of single-end or paired-end reads has been extensively used to identify CNVs in various organisms (Mills et al. 2011a, Yalcin et al. 2011, Zichner et al. 2013). NGS has led to an increased accuracy of detecting the breakpoint junctions of CNVs thus making inference of mechanisms of CNV formation possible (Carvalho et al. 2013, Hermetz et al. 2014). Another

major advantage of sequencing over microarrays is its ability to explore a number of genomic variants without prior knowledge of the sequence.

1.2.1.4.1 Sequence data analysis tools for CNVs detection

The process of CNV detection from genome sequence date requires initial mapping of reads to a reference genome followed by either read depth analysis (normalized depth of coverage correlates with copy number), paired end mapping (distance between paired reads that are longer/shorter than expected), split-read analysis (reads that span CNVs breakpoint and appear to map on two locations on the reference genome) (Karakoc et al. 2012), or a combination of these approaches (Zhang et al. 2011). A summary of these analyses methods is illustrated in Figure 1.2. In addition, fine mapping of CNV boundaries can be achieved by assembly of the short reads, followed by mapping of contigs to the reference (Le Scouarnec and Gribble 2012). One of the challenges observed in these analyses methods is that when used on the same set of samples, some CNVs can be identified by one method and not the others (Alkan et al. 2009, Mills et al. 2011a).

CNV detection using read depth analysis involves calculation of read depth in non-overlapping window sizes in a chromosome, normalisation and application of statistical methods to identifying regions that have excessive or reduced coverage. Ideally, the depth of read coverage of a region is proportional to copy number. However, there exist factors other than copy number that affect sequence coverage. These include GC content and the ability of a region to be uniquely mapped (Quail et al. 2012, Ross et al. 2013). The regions with very low and very high GC-content have lower coverage compared to regions with intermediate GC content. Most of the CNV calling algorithms correct for these two effects and also perform normalisation of read depth across samples. These tools use a variety of statistical models and

segmentation analysis methods (Miller et al. 2011, Xie and Tammi 2009, Chiang et al. 2009, Abyzov et al. 2011, Xi et al. 2011, Ivakhno et al. 2010, Magi et al. 2011).

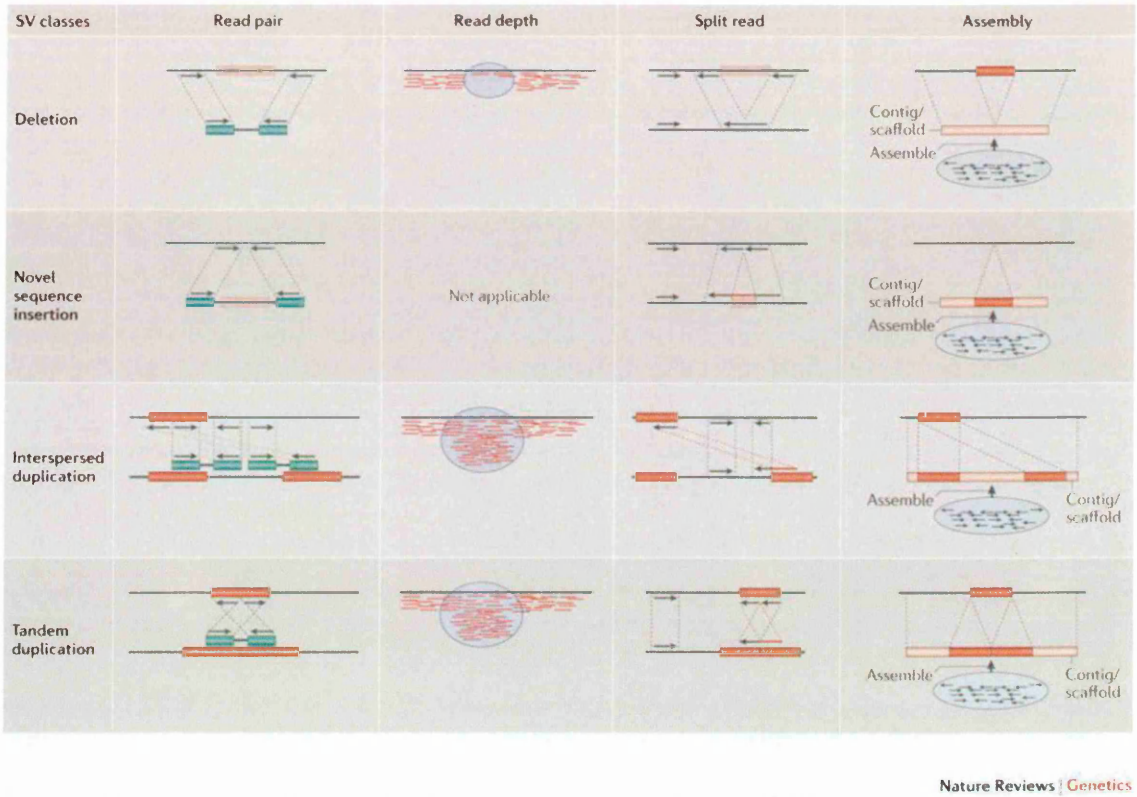


Figure 1.2. An illustration of the approaches used to detect CNVs using next generation sequencing.
CNVs, referred to as structural variation (SV) in the image, can be identified from reads pairs that map to the reference at distances shorter/longer than expected read length (first column), variation of normalized read depth at a locus compared to the expected (second column), reads that map to two locations may span a CNV breakpoint junction (third column) and lastly mapping of contigs generated from de novo assembly of reads may identify structural variations in the genome (fourth column) Image reproduced from Alkan, Coe and Eichler 2011 with permission from Nature Publishing Group.

1.2.1.4.2 Challenges of using short read sequence data to detect CNVs

The short reads obtained from NGS present computational challenges during alignment of the reads to a genome characterized by regions with repetitive sequences, of which *P. falciparum* is one (Treangen and Salzberg 2012, Gardner et al. 2002). The reads that map to the repetitive regions cannot be mapped uniquely and can either be filtered off, assigned to one of the

possible genomic locations at random or excluded. This still poses limitations on the downstream analysis since not all regions in the genome can be assayed, and in the instance of random alignment of non-unique reads, copy number estimates are inaccurate.

1.2.2 Targeted loci CNV detection

Laboratory methods for CNV identification at pre-decided ("targeted") gene loci have been used to validate results from whole genome scans and also to map CNV breakpoints. These methods are described below.

1.2.2.1 Southern Blot

Southern blot is a technique developed to identify a specific DNA sequence in a mixture of DNA fragments (Southern 1975). The fragmented DNA is separated by size in an agarose gel then denatured to later enable hybridization of a labelled probe of sequence similar to the region of interest. The denatured DNA is transferred from the gel to a membrane where hybridization of the labelled probe is performed, and the hybridization results imaged. Increased copy number is observed when there is increased signal in the blot. No signal indicates a deletion in the test sample compared to the control. The principle was applied to the development of microarrays.

1.2.2.2 Fluorescent in situ hybridization (FISH)

FISH is a method developed in the early 1980's used to detect specific DNA sequence in chromosomes in fixed cells (Langer-Safer, Levine and Ward 1982). A labelled probe, with sequence complementary to the region of interest, i.e., a CNV region, was generated and hybridized to fixed cells which were then visualized using a microscope (Ribacke et al. 2007).

Both southern blot and FISH are low-throughput, labour-intensive and require a large quantity of DNA, and hence are mostly used in small-scale studies.

1.2.2.3 Long range PCR

Long range PCR amplification of the large genomic regions (up to 40kbp) containing CNVs has been useful for precise mapping of breakpoint junctions of CNVs through sequencing of the PCR products (Carvalho et al. 2013, Winchester et al. 2008). CNVs are identified through differences in the size of PCR products in the test and reference samples. The major challenge of this method is that prior knowledge of the sequence in the region of interest is required in order to design primers.

1.2.2.4 Quantitative real time PCR (qPCR)

Real time PCR has almost completely replaced the methods described above owing to its application in high-throughput assays, its less laborious nature and the low quantities of DNA template required (Ribacke et al. 2007). To detect CNVs, primers against the regions of interest are designed and amplified by PCR in a quantitative assay using an endogenous gene (existing as single copy) as a control in order to normalize for variation in input gDNA quantity and PCR inhibitors in the samples. Two methods can be used to calculate copy number – the comparative Ct method (in which it is assumed that the efficiency of amplification of the test and endogenous genes are similar) or the relative standard curve method in which two standard curve generated for a single copy gene and a target gene and used to calculate the fold difference in quantity of copies between a test sample and a calibrator sample (assumed to contain a single copy of the gene of interest).

1.3 Mechanisms of CNV formation

The generation of gene copy number changes requires structural changes in the chromosome that enable joining of two separate DNA regions. Processes involved in joining these regions have been inferred from the DNA sequence at the point of joining, i.e. the CNV's breakpoint junction of the once separate regions (Hastings et al. 2009, Liu et al. 2012, Gu, Zhang and Lupski 2008). Investigations on the characteristics of the breakpoint junction have revealed that some are found in low copy repeat regions (LCRs) characterized by long stretches of sequence homology (>50bp), while others are present in regions of short sequence homology of 2bp-15bp (microhomology) (Gu et al. 2008, Liu et al. 2012, Hastings et al. 2009). These findings led to the proposition that CNVs may arise through homologous recombination at regions with extensive sequence homology enabled by DNA repair mechanisms. Second, short homology regions (microhomology) may be subject to template switches during replication thereby leading to CNV formation. These two broad mechanisms are further discussed below.

1.3.1 Homologous recombination

In eukaryotes, homologous recombination (HR) underlies DNA repair of double stranded breaks and nicks (Cahill, Connor and Carney 2006). HR also enables exchange of genetic material during meiosis leading to new combinations of DNA sequences (Amunugama and Fishel 2012). In a diploid organism, a successful repair event involves the use of alleles to repair a broken strand, restoring it to its original sequence. Crossover between non-alleles, a process known as non-allelic homologous recombination (NAHR), has been identified as one of mechanisms that generates structural changes in the genome (Gu et al. 2008, Hastings et al. 2009, Liu et al. 2012). NAHR is seen during repair of double stranded DNA breaks when unequal crossover occurs between non-homologous repeat regions. NAHR also occurs during

repair of collapsed replication forks due to a nick in the template DNA strand, a process termed break-induced replication (BIR). In this case, the exposed broken strand may bind to a non-homologous region with high sequence similarity. These processes may lead to deletions, duplications or inversions of genomic segments. The models of CNV formation by NAHR are illustrated below (Figure 1.3).

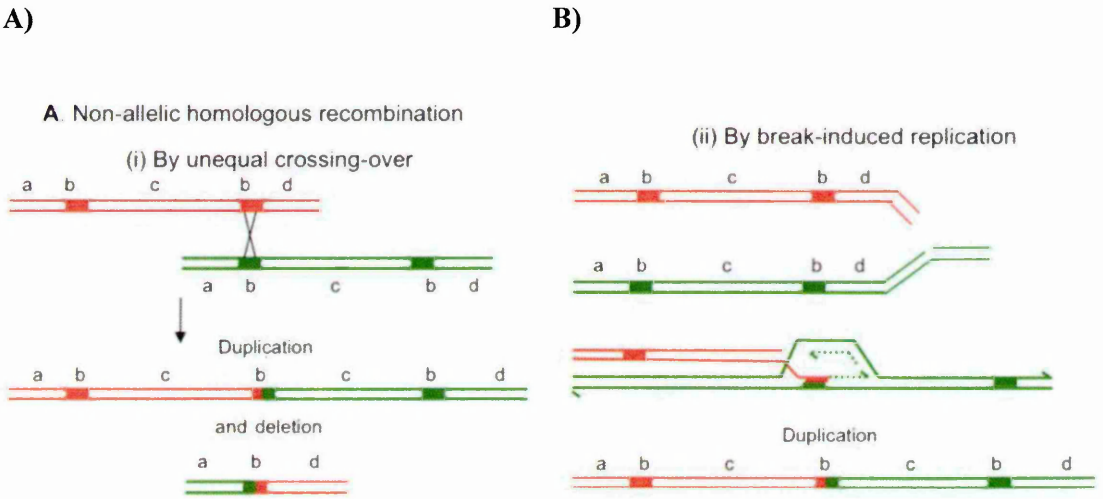


Figure 1.3. Models of CNVs formation through non-allelic homologous recombination (NAHR)
A) NAHR could occur through homologous recombination between non-allelic repeats regions, during meiosis or repair of double stranded DNA breaks resulting in deletions or duplications or B) during repair of collapsed DNA replication fork by a process called break-induced replication (BIR). Image reproduced from Hastings et al. 2009 with permission from Nature Publishing Group.

The malaria parasite DNA is exposed to damage by reactive oxygen and nitrogen species released during host immune defence (Nathan and Shiloh 2000) and heme metabolism. Errors during DNA replication also contribute to alteration of DNA sequence. The parasite, in its haploid state in the human host, uses homologous recombination as one of the mechanisms to repair double stranded DNA breaks, utilizing regions with high sequence similarity (mostly segmental duplications) of the genome (Kirkman, Lawrence and Deitsch 2014). NAHR can also occur when the parasite is in its diploid state in the mosquito. NAHR may be involved in

CNV generation in *P. falciparum* as for other eukaryotes since the parasite uses homologous recombination to repair double stranded DNA breaks (Samarakoon et al. 2011b). This is supported by the observed association of CNVs with segmental duplications (SDs) and repeat regions reported in *P. falciparum* (Cheeseman et al. 2009). Segmental duplications are segments of DNA in the genome, greater than 1kb, that exhibit high sequence similarity.

1.3.1.1 Non-homologous recombination (microhomology mediated mechanisms)

The presence of microhomology in some CNV breakpoint junctions has been reported in the human genome (Conrad et al. 2010a, Carvalho et al. 2009). There exist DNA repair pathways that do not require extensive sequence homology. Instead, they use short sequences of homology termed microhomology (1-25bp) or no homology and therefore increase the chances of genetic alterations (Symington and Gautier 2011). These mechanisms include non-homologous end joining (NHEJ) also referred to as canonical-NHEJ (C-NHEJ) and microhomology mediated end joining (MMEJ) sometimes referred to as alternative-NHEJ (Symington and Gautier 2011, Fattah et al. 2010). In eukaryotes, C-NHEJ involves binding of Ku heterodimer to broken ends, recruiting a protein kinase (DNA PK_{CS}) and forming a complex that leads to further recruitment of other proteins (Artemis, LIGIV, XLF and XRCC4) that facilitate ligation of the two ends. Chromosomal rearrangements occur as a result of joining of two strands that are from different genomic regions. MMEJ differs from C-NHEJ in the proteins that bind to the broken ends of DNA. The proteins involved in C-NHEJ have not been identified in *Plasmodium*. Instead, it is thought that the malaria parasite uses alternative-non homologous end joining (a-NHEJ) in addition to HR to repair DNA (Kirkman et al. 2014). The model of NHEJ is illustrated below (Figure 1.4).

1.3.2 Replicative mechanisms

The presence of microhomology at breakpoint junctions does not only provide evidence for NHEJ but also the involvement of DNA replication in chromosomal rearrangements (Lee, Carvalho and Lupski 2007, Zhang et al. 2009, Hastings et al. 2009). From studies in *Escherichia coli* and the human genome, it is proposed that template switching within a replication fork results in structural changes in the genome. The presence of short lengths of homology in the exposed lagging strand in a replication fork may result in formation of secondary structures that prevent replication of certain segments leading to deletion of sequences (Hastings et al. 2009). Deletions observed in regions that occur between two sites with sequence similarity and a length between them the size of an Okazaki (short fragments newly synthesized in the lagging template strand in a replication fork) fragment supports the proposition of slippage/template switch in DNA replication within the same replication fork (Figure 1.5) (Albertini et al. 1982, Hastings et al. 2009).

In some instances, formation of secondary structures may block continuation of synthesis of the lagging strand resulting in the exposed 3' end annealing to a different exposed single strand with short sequence similarity on a different replication fork (Zhang et al. 2009, Carvalho et al. 2013, Gu et al. 2008). This results in duplications, deletions or inversion of distant genomic regions, a mechanism referred to as Fork stalling and template switching (FoSTeS) (Figure 1.5).

A third replicative mechanism known as microhomology mediated break-induced replication (MMBIR) is proposed to occur when an exposed single stranded end from a collapsed replication fork (due to a nick in DNA template) anneals to any other single strand with

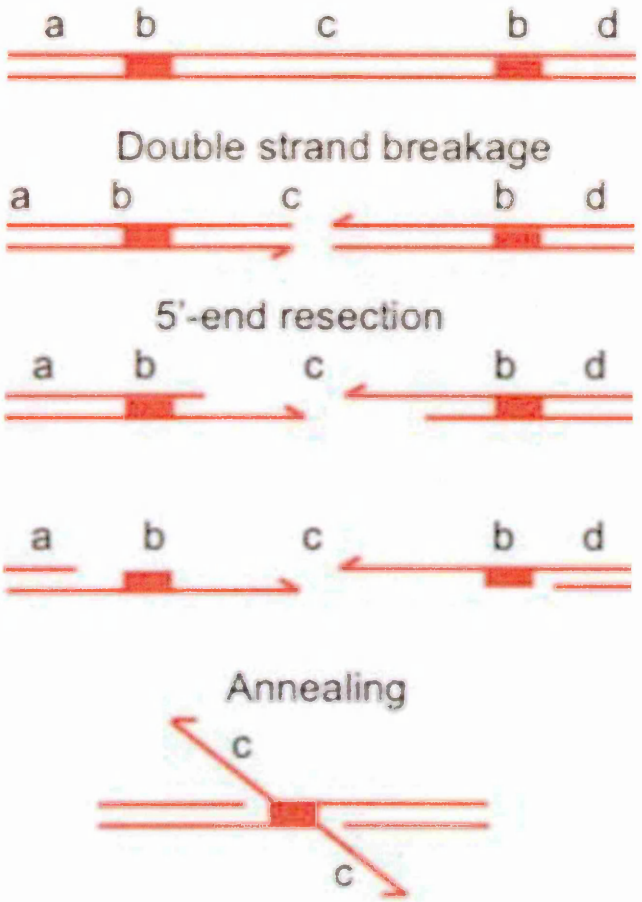


Figure 1.4. Model for formation of CNVs by non-homologous end joining (NHEJ). Double stranded breaks in the genome at region c is repaired by binding of the proteins involved in NHEJ to the exposed ends (c). This is followed by resection and ligation of broken ends. A deletion of region c is observed as a result of the exposed ends ligation to other exposed ends (b) with sequence similarity. Image reproduced from Hastings et al. 2009 with permission from Nature Publishing Group.

In *P. falciparum*, DNA secondary structures have been shown to associate with recombination sites of *var* genes (Sander et al. 2014). The secondary structures are thought to promote recombination during replication of DNA generating diversity of *var* genes that encode PfEMP1. PfEMP1 is a protein important for sequestration of the parasite to host endothelial cells and under antigenic variation to evade host immunity.

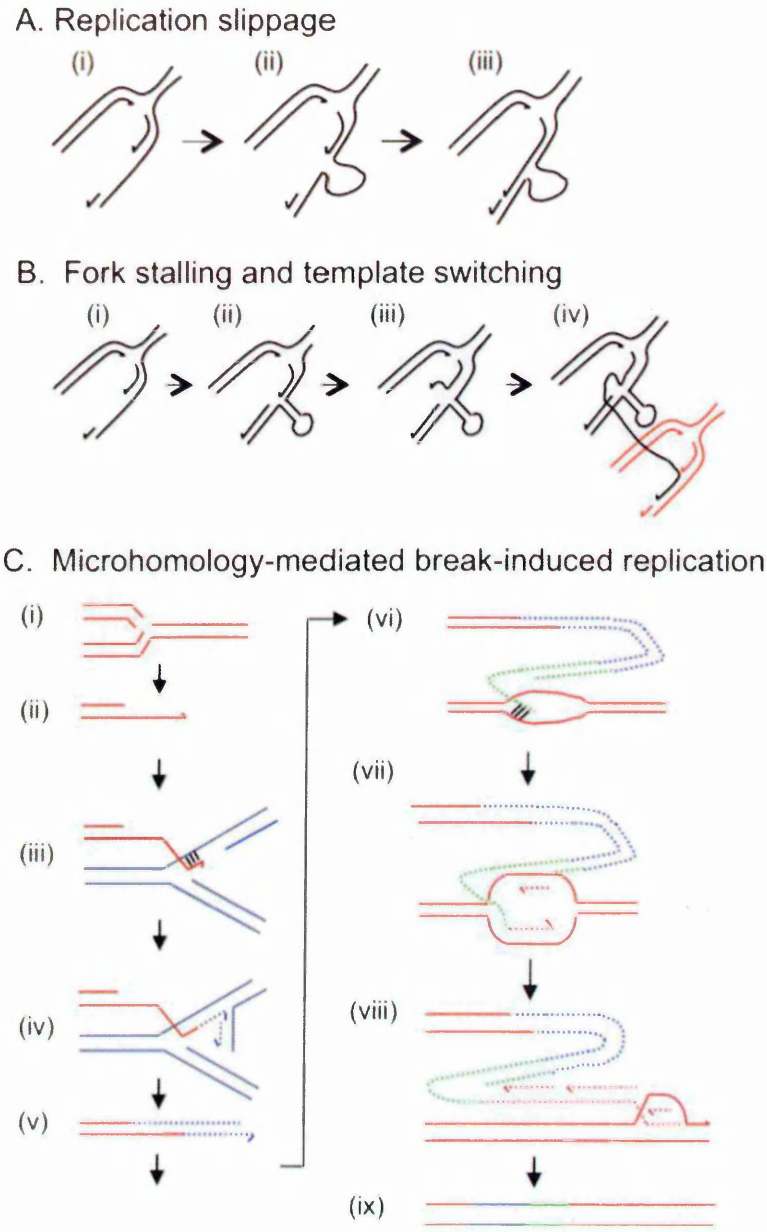


Figure 1.5. Models of CNVs formation by replicative mechanisms mediated by microhomology

Microhomology mediated mechanisms of CNV formation that occur during DNA replication include A) replication slippage as a result of secondary structure formed due to presence of regions of single stranded DNA with sequence similarity within a replication fork, B) fork stalling and template switching (FoSTeS) caused by formation of hairpins that hinder the progress of a replication fork resulting in invasion of the exposed single stranded DNA to another replication fork with short sequence similarity and finally C) MMBIR is proposed to occur in collapsed replication forks due its encounter of a nick in the DNA template. Image reproduced from Hastings et al. 2009 with permission from Nature Publishing Group.

1.4 Functional impact of CNVs

The impact of CNVs on phenotypic variation has been observed in a number of organisms (Henrichsen et al. 2009, Craddock et al. 2010, Hou et al. 2012, Anderson et al. 2009, Chen et al. 2012). This has been evaluated by studying the association between gene copy number and gene expression levels and also through associations of copy number with important phenotypes. Gene expression variation is commonly used as a proxy for phenotypic variation.

1.4.1.1 CNVs and gene expression

CNVs may affect gene expression levels through altering gene dosage, introducing changes in regions that regulate gene expression, and possibly by modifying the chromatin environment of a gene (Kleinjan and van Heyningen 2005). Studies on the effect of CNVs on gene expression levels in *P. falciparum* have revealed that certain CNVs influence the transcript levels of the genes within the CNV regions and also outside these regions (Mackinnon et al. 2009, Gonzales et al. 2008). Gene copy number has been shown in *P. falciparum* to both positively and negatively correlate with gene transcription levels suggesting a second mechanism in addition to the direct dosage effect (Mackinnon et al. 2009, Gonzales et al. 2008). One example of this is an amplification of a region in chromosome 5 containing the multidrug resistant gene 1 (*mdr1*) that has been identified as a 'hotspot' for gene expression regulation through gene expression quantitative trait loci (eQTL) analysis of the progeny of HB3 and Dd2 genetic cross (Gonzales et al. 2008). This region has been shown to associate with transcription levels of 269 genes. It was observed that an increase in copy number in progeny that inherited multiple copies from Dd2 resulted in subsequent increase in transcript levels of 85% of the 269 genes and decrease in transcript levels of 15% of the genes (Gonzales et al. 2008). Locus-specific studies on GTP cyclohydrolase 1 gene (*gch1*) amplification also report an increase in *gch1* expression with increased copy number. *gch1* amplification has

been indirectly associated with antifolate resistance in geographic regions that historically use antifolates as antimalarial drugs (Nair et al. 2008).

1.4.1.2 Influence of CNVs on phenotypic variation and parasite adaptation

Directional selection on *P. falciparum* parasites bearing CNVs associated with certain phenotypes has been explored in several studies as discussed below.

1.4.1.2.1 Drug resistance

Some of the CNVs associated with malaria treatment failure include the amplification of a locus in chromosome 5 containing the multidrug resistance gene (*mdr1)* (Gadalla et al. 2011, Lim et al. 2009) and a region in chromosome 12 containing the GTP cyclohydrolase (*gch1*) gene (Nair et al. 2008). Increased copy number of the region containing *mdr1* has been linked to resistance to mefloquine, halofantrine and artemisinin whereas reduced copy number is associated with chloroquine resistance (Sidhu et al. 2006, Price et al. 2004). Amplification of *gch1*, encoding an enzyme involved in the folate synthesis pathway, has been strongly associated with a mutation on dihydrofolate reductase (*dhfr-164L*) and marginally associated with mutations on *dihydropteroate synthase* (*dhps-A436S and dhps A581G))* (Nair et al. 2008, Robinson et al. 2011). DHFR and DHPS are enzymes that act downstream of GCH1 in the folate biosynthesis pathway and form targets of antifolate drugs. These mutations has been linked to resistance to antifolate malaria drugs and found at high frequency in Thailand (Sirawaraporn et al. 1997, Nair et al. 2008, Plowe et al. 1998). The *gch1* amplification has also been observed to exist at a higher frequency in Thailand, a region with historical use of antifolates as first-line treatment, with higher selection of resistant parasites than a second region, Laos, where antifolates were second-line treatment at the same period (Nair et al.

2008). It is thought that increased *gch1* copy number acts a compensatory mechanism for altered function of DHFR and DHPS enzymes as a result of the mutations (Kidgell et al. 2006).

1.4.1.2.2 Host immunity

Host immunity is thought to be one of the factors involved in selection of parasites bearing a deletion at the C-terminal of reticulocyte binding protein homolog 2b (*PfRh2b*) located in chromosome 13 (Ahouidi et al. 2010). The deleted region is thought to facilitate binding and erythrocyte invasion of the parasite. This polymorphism has been shown to exist at a higher frequency in Senegal, Tanzania and Malaysia as compared to Thailand and Brazil (Ahouidi et al. 2010). The difference in frequency is speculated to be as a result of difference in polymorphisms on the host erythrocyte receptors and origin of the deletion in Africa and its spread to the rest of the world. It was also observed that antibodies against the C-terminal existed in Senegal and Tanzania and increased with age, though the increase was not statistically significant. These antibodies are thought to bind to the C-terminal and possibly block sialic-acid independent erythrocyte invasion. Therefore, the deletion may be a mechanism by which the parasite evades host immunity, thus potentially explaining its existence at non-trivial frequencies (Ahouidi et al. 2010, Reiling et al. 2010). This evidence is inconclusive since the presence of the antibodies was not tested in the regions with low deletion frequencies (Thailand and Brazil). Furthermore, deletions in both homologs (*Rh2a* and *Rh2b*) have been observed in clinical isolates in other studies (Robinson et al. 2011). RH2a and RH2b proteins are involved in erythrocyte invasion through the sialic acid independent pathway (Sahar et al. 2011).

1.4.1.2.3 Adaptation to *in vitro* environments

Parasites bearing CNVs associated with cell invasion, cytoadherence, and gametogenesis among others, have been reported to be selected during *in vitro* culture propagation. These CNVs include large subtelomeric deletions in chromosome 2 and 9 (resulting in loss of a number of genes) that have been associated with loss of cytoadherence and loss of both cytoadherence and gametogenesis phenotypes respectively (Ribacke et al. 2007, Shirley et al. 1990, Biggs, Kemp and Brown 1989, Mackinnon et al. 2009, Kemp et al. 1992). Other CNVs observed to arise *in vitro* include an inversion of a region containing a gene encoding ring-infected erythrocyte surface antigen (RESA) coupled with a deletion upstream of the same gene observed in knobless parasites that do not express the *resa* gene (Pologe, de Bruin and Ravetch 1990). RESA is one of the proteins exported to the erythrocyte surface and thought to be involved in blocking additional merozoite invasion of an already infected erythrocyte, decreasing deformability of infected host erythrocytes containing the early ring stage of the parasite and also protecting the parasite during episodes of high fever of the host (Mills et al. 2007, Pei et al. 2007). Another important variant observed in cultured parasites is a deletion of skeleton binding protein 1 gene (*pfsbp1*) (Cheeseman et al. 2009), known to be involved in erythrocyte membrane protein 1 (PfEMP1) trafficking to the infected red blood cell surface (iRBC) (Maier et al. 2007). These gene copy number changes have not been observed in clinical samples and therefore may be involved in an otherwise costly mechanism that is important to the parasite for maintaining itself in a host.

In culture, the action of immunity and the need for sexual reproduction is diminished and thus selection for parasites with high growth rates may occur (Ribacke et al. 2007). High growth rates of parasites in culture have been linked to increased copy number of the gene encoding reticulocyte-binding protein 1 (RH1) (Ribacke et al. 2007). This CNV has only been seen in

some culture-adapted parasites and not in fresh clinical isolates and may be a relief of fitness cost associated with host immunity or host receptor polymorphisms(Nair et al. 2010, Stubbs et al. 2005). RH1 protein is involved in the sialic acid-dependent invasion of the merozoite into the host erythrocyte hence the gene amplification is thought to promote the parasite's rate of growth in culture.

Drug selection in the laboratory has led to generation of parasite lines resistant to certain drugs. Some of the drug resistance traits have been associated with gene copy number differences between the resistant and sensitive strains. These include amplification of the region containing *pfmdr1* (Sidhu et al. 2006, Price et al. 2004, Jiang et al. 2008a), which has also been reported in clinical isolates, increased copy number of *falcipain-2* and *falcipain-3* genes seen in parasite lines resistant to a drug that inhibits cysteine protease (Singh and Rosenthal 2004), and deletion of 15 genes (PF3D7_1000900 to PF3D7_1002100) in chromosome 10 in parasites bearing mutations in chloroquine-resistant transporter gene (*pfcrt*) exhibiting different drug resistance profiles (Jiang et al. 2008a). Another deletion of 23 genes in chromosome 14 was observed in two Dd2 clones selected for fosmidomycin resistance and not in the wild type parasite (Dharia et al. 2009). One of the genes deleted is Deoxy-D-xylulose 5-phosphate reductoisomerase (*pfdxr*) which encodes an enzyme involved in the isoprenoid synthesis in *P. falciparum* that is under inhibition by the drug. An amplification of dihydrofolate reductase (*dhfr*) gene was observed in parasites subjected to increased doses of pyrimethamine leading to development of resistance (Thaithong et al. 2001). Knowing whether CNVs directly cause or are in linkage with causal mutations of drug resistance will be useful in monitoring drug resistance and would also offer insight into improving treatments.

Overall, adaptation of parasites to laboratory culture has been shown to generate non-natural gene copy number changes, as seen by considerable lack of overlap in CNVs genes identified in fresh clinical and laboratory adapted strains in published studies. Therefore the studies of CNVs so far, most of which have been performed on culture adapted isolates, might not be informative for understanding the role of CNVs in natural populations. There is thus a need for large scale exploration of CNVs in fresh field isolates to determine the extent of these variants in the genome and their potential role in adaptive evolution of the natural population of parasites.

1.5 The role of CNVs in *P. falciparum*

The evidence, so far, points to the contribution of CNVs to the dynamic structure of the genome that enables the parasite to adapt to its changing environment. First, the high abundance of repeat regions and monomeric A/T tracts in the genome (Gardner et al. 2002, Samarakoon et al. 2011b) increases the chances of recombination events that lead to CNV formation that may constantly change the genome structure. The breakpoints of two CNVs, i.e., amplification of *mdr1* and *gch1* have been observed to consist of monomeric A/T tracts (Nair et al. 2007, Nair et al. 2008). Additional evidence for the contribution of CNVs to the dynamic genome structure is the observation of CNV regions with multiple copy number states, e.g., the number of copies of *gch1* has been found to vary from 1-11 copies in a population (Nair et al. 2008). Differences in CNVs observed between parasite parent lines and the progeny lines further illustrates the contribution of CNVs to the dynamic state of the parasite's genome (Samarakoon et al. 2011a). The growing evidence for the potential adaptive relevance of the CNV in the parasite warrants an investigation of the CNVs in nature and their potential adaptive role.

1.6 Scope of the thesis

Most studies of CNVs in *P. falciparum* have been on a small scale and performed in short-term or long-term cultured parasite lines which are not representative of natural populations. One of the reasons for the small number of clinical isolates previously studied is the low quantities of genetic material that are available for the assays: this issue has now been solved by use of whole genome amplification techniques. Furthermore, there has been growing evidence for the importance and impact of CNVs on disease and evolution in various organisms, simultaneous with remarkable improvements in technologies that facilitate high-throughput genome-wide scans for CNVs, thereby setting the stage for large-scale studies of CNVs in *P. falciparum*.

In this thesis, it is hypothesised that CNVs are important for the adaptation of the parasite to its variable environment. My goal is to establish the distribution of the CNVs in the parasite genome, determine frequency of CNVs in three *P. falciparum* populations in Eastern Africa, using microarrays, and identify the CNVs that may be under selection. Some of the differences among the three populations include variation in malaria transmission intensity (Noor et al. 2014, Okiro et al. 2010), antimalarial drug use (Amin et al. 2007), host genetics (Piel et al. 2010) and vector populations (Mwangangi et al. 2013). Apart from CNV identification, validation of the CNVs using a different technology, i.e., sequencing was performed. In addition, the influence of CNV on gene expression was investigated. The overall aim of this thesis is to identify and characterise CNVs in natural populations of parasites and determine whether they are involved in adaptation.

The methodology is described in Chapter 2. Chapter 3 presents results of a genome-wide screen for CNVs in four spatially and temporally separated populations of *P. falciparum* under

different transmission intensities using microarrays with a view to identifying the CNVs under selection. Described in this chapter is the analytic approach of identifying CNVs using microarrays, including an investigation of the effects of various parameter settings on CNV detectability in order to select the settings that best define CNVs from microarray data. In addition, a number of systematic effects relating to experimental design that affect CNV detection were explored in order to help isolate the effects of population differences – the main focus of the study - from methodological effects. The distribution of the CNVs in the genome is shown and the frequency of each in the populations determined. A number of functional gene groups were found to be enriched in the CNV gene list. Using population genetics analysis, CNVs with frequencies that appear to be highly differentiated between populations and potentially under selection were reported.

Chapter 4 focuses on confirmation of the CNVs detected using microarrays by whole genome sequencing of 22 of the 183 samples using PGM Ion Torrent machine. A summary of the sequence data output including quality and read coverage is given first. Read depth analysis is then used to detect CNVs which involves mapping of reads to a reference genome followed by the use of a CNV calling tool that identifies regions that differ in sequence coverage from normal. CNVs were detected in different target genomic regions in an effort to understand the effect of various aspects of the *P. falciparum* genome, e.g., low complexity regions, which may affect CNV detection. Some correspondence between microarrays and sequencing was observed. The chapter mostly highlights the challenges in using whole genome sequence data for CNV detection.

In Chapter 5, the effect of CNVs on expression of genes within the CNVs and those outside the CNV regions was investigated with a view to understanding the functional impact of CNVs in *P. falciparum.*

In Chapter 6, the main findings was summarised and their implications for future research in malaria parasite biology discussed with a view to design tools that are useful for the ultimate control of this disease.

Chapter 2

Materials and Methods

2 Chapter 2: Materials and methods

2.1 Population study of CNVs by microarrays

2.1.1 Study population

Patients with malaria attending healthcare facilities in three Eastern Africa regions were recruited for the study. Malaria positivity was determined by microscopy. The study populations include North Eastern Sudan (Gedaref, Kassab, Medani) and two populations, Kisumu (at western Kenya) and Kilifi (at the coast of Kenya with low, high and moderate malaria transmission intensities, respectively (Noor et al. 2014, Okiro et al. 2010). The study participants in Kilifi were recruited at two time points; (1994-1996) a period with higher transmission intensity and 2010 a period with lower transmission intensity (O'Meara et al. 2008). The participants from Kisumu and Sudan were recruited in 2008 and 2007 respectively. The populations also differ in antimalarial drug used (Amin et al. 2007), vector populations (Mwangangi et al. 2013) and host genetics (Piel et al. 2010). Ethical approval for the study was obtained from the Kenyan National Ethical Review Committee (SSC 1292). Written consent was obtained from parents/guardians of the study participants if they were below 14 years, or the participants themselves otherwise.

2.1.2 Sample processing

For samples used for Comparative Genome Hybridization (CGH) by microarray, 2-5 ml of blood was drawn by venepuncture from patients into sterile heparinised tubes (BD) and placed at 4-8^{0}C before further processing within 12 hours. To minimize contaminating human host genetic material, whole blood obtained from individuals was centrifuged at 440 x g for 5 minutes and plasma and buffy coat removed. An aliquot of the infected red blood cell pellet (iRBCs), between 30 µl- 200 µl, of each sample, was stored at -80°C for CGH after washing

the pellet in 1X Phosphate Buffered Saline (PBS) (Oxoid). For some of the samples collected from Sudan, the buffy coat was not removed prior to washing and storing.For samples used for DNA sequencing, the buffy coat was returned to the tube with the remaining iRBCs pellet onto which 5ml of incomplete culture medium was added, mixed and then layered on top of 3 ml of sterile Lymphoprep (Axis-Shield Pos AS) contained in a separate 15ml tube in order to separate PBMCs for other studies. The incomplete culture medium was made up of 500ml of RPMI 1640 (Gibco), 18.75ml of IM HEPES buffer (Gibco), 5ml of 2mM L-glutamine (Gibco), 1.25ml of 10mg/ml gentamicin (Gibco), 5ml of 20% glucose (Gibco), 3 ml of 1M NaOH. The tube was centrifuged at a speed of 440 x g for 20 min. The layer containing peripheral blood mononuclear cells (PBMCs) was removed. Granulocytes were then removed from the remaining iRBCs pellet by Plasmion (Bellon) flotation. An aliquot of 100-200 µl of these white cell-depleted iRBCs, to be used for whole genome sequencing, was stored in 1ml of glycerolyte (42.25% glycerol, 0.1M Sodium Acetate, 4mM KCl, NaH_2PO_4 pH 6.8 (Sigma-Adrich)) in cryovials (Thermo Fisher Scientific Inc.) in liquid nitrogen.

The reference parasite line used in CGH, a Kilifi laboratory adapted line, originated from a malaria patient at the Kilifi District Hospital (Mackinnon et al. 2009). The parasite line was maintained *in vitro* at 2% hematocrit in complete medium made up of incomplete medium, human red cells (O cells) and 10% human serum. The culture medium in the flasks was changed every other day and gassed with a gas mixture (BOC) containing 3% carbon dioxide, 1% oxygen and 96% nitrogen.

2.1.3 Genomic DNA extraction

Genomic DNA was isolated from Kilifi samples by lysis of 30 µl- 200 µl of iRBCs using saponin. The iRBCs were resuspended in 1X PBS to a final volume of 1 ml followed by

addition of 1 ml of 0.1% of saponin (Sigma) dissolved in 1X PBS to obtain a final concentration of 0.05% saponin (Sigma).The resuspension was incubated for 3 minutes at room temperature and centrifuged for 10 minutes at a speed of 1440 x g. The pellet was washed with 3 ml cold 1X PBS. Protein digestion and further lysis of the pellet was performed by incubation in 75ul Proteinase K (20 mg/ml) (Applied Biosystems) and 425 µl lysis buffer (containing 80mM EDTA (Sigma-Aldrich) at pH 8.0, 40mM Tris-HCl (Sigma) at pH 8.0 and 2% SDS (Sigma)) for 1 hour at 37^0C. Finally, separation of DNA from proteins was achieved by adding 500 µl of phenol chloroform (Invitrogen) to the lysed material and centrifugation at 11714 x g to separate the aqueous layer containing nucleic acid from the inorganic material. Genomic DNA (gDNA) was precipitated from the aqueous mixture using 500 µl isopropanol (Sigma-Aldrich) and 0.1 X the total volume (aqueous mixture and isopropanol) of 3M sodium acetate (Ambion) at -20 ^{0}C overnight. The precipitate was washed with 700 µl 75% ethanol (Sigma-Aldrich), dried and resuspended in 20 µl 1X TE buffer (Invitrogen). Sudan and Kisumu DNA samples were extracted from 100 µl blood pellet using the automated ABI PRISM 6100 Nucleic Acid PrepStation (Applied Biosystems) as described in the manufacturer's protocol. The extracted DNA was eluted in 100 µl BloodPrep DNA Elution solution 2 (Applied Biosystems). The presence and integrity of extracted gDNA was confirmed using a 1% agarose (Promega) gel electrophoresis.

2.1.4 Multiplicity of infection

The number of distinct parasite clones per isolate was determined by genotyping of *P. falciparum* merozoite surface antigen 2 (*msp2*) using a method developed by Liljander and others (Liljander et al. 2009). The procedure involves a nested PCR reaction with the primary PCR amplification of block 3 of *msp2* followed by a secondary PCR with fluorescent primers targeting the *msp2* allelic types FC27 and IC. Fragment analysis of the PCR product was

performed using capillary electrophoresis and the results analyzed using GeneMapper® Software version 4.0 (Applied Biosystems). One microliter of the DNA sample, extracted using different methods (described in section 2.1.3), was used in the PCR. The samples assayed had different parasite DNA quantities per microliter hence could lead to bias in the estimation of MOI.

2.1.5 Comparative genomic hybridization (CGH)

2.1.5.1 Microarray design

The microarray used for the study consisted of 70mer oligonucleotides (probes) spotted on a glass slide (Bozdech et al. 2003b). The design and printing of the array was done at Assoc. Prof. Zbynek Bozdech's laboratory, a collaborator, at Nanyang Technological University in Singapore. The probes on the array were designed using the available complete *P. falciparum* genome sequence of 3D7 parasite line (Gardner et al. 2002) to target conserved regions of approximately 5400 genes with an average of 2 probes per gene (Bozdech et al. 2003b). The CGH experiment was carried out in a microarray facility established in the KEMRI-Wellcome Trust Research Programme laboratories in Kilifi, Kenya.

2.1.5.2 Whole genome amplification

The CGH experiment was performed on 183 samples, with 8 of these samples with duplicate experiments. To increase the amount of DNA available for hybridization to the array, whole genome amplification using random nonamers was performed (Petalidis et al. 2003). The samples were randomized during amplification and hybridization experiments to minimize experimental variability. The whole genome PCR amplification involved three steps. First, random priming was performed on 100ng of gDNA (gDNA quantity measured using

Nanodrop 2000 (Thermo Scientific)) using random nonamers (SMART-Random) consisting of 9 random nucleotides with a 23 bp sequence tag (5'-AAGCAGTGGTATCAACGCAGAGTNNNNNNNNN-3'), obtained from Eurofins MWG, at 95^0C for 6 min. The second step, PCR extension of the randomly primed regions, involved addition of 2.5 units of Klenow fragment (5U/µl) (3′→5′ exo–) (New England Biolabs), 1 µl of 3X dilution of aa-dUTP/dNTP mix (consisting of 34 µl of 100mM dATP, 17 µl each of 100mM dCTP, dGTP, dTTP (New England Biolabs), 17 µl of 100mM amino acyl dUTP (aadUTP) (Biotium) and 11.3 µl of distilled water), 1 µl NEBuffer 2 (New England Biolabs) and 2.5 µl ddH_20. The Klenow fragment is a large fragment of DNA polymerase 1 that lacks the 3'-5' exonuclease activity hence facilitates incorporation of aadUTP that enables coupling of the Cy3 and Cy5 fluorescent dyes used for DNA labelling (GE Amersham). The thermocycling conditions for the Klenow mixture were; 25^0C for 10 min, 37^0C for 60 min and 75^0C for 20 min. The last step involved a 100µl PCR amplification reaction containing 10 units of DNA Taq Polymerase (5U/µl) (New England Biolabs), 6 µl of 100nM primer (SMART-Amplification) complementary to the tag sequence in the random primer (5'-AAGCAGTGGTATCAACGCAGAGT-3') (MWG), 10X Buffer, 1.5 µl of aa-dUTPs/dNTP mix and 2.5 µl of the Klenow mixture product. The PCR amplification ended at the linear phase, after 19 cycles, to maintain the relative abundance of DNA. The thermocycling conditions for this step are shown below (Table 2.1).

The PCR product was purified using QIAquick PCR Purification Kits (Qiagen). The purification procedure was provided with the kit and it involved addition of 500 µl of PB binding buffer to the PCR product then transferred to the spin column. This was followed by centrifugation at 15026 x g for 1 min and the flow-through discarded. 750 µl of PE buffer was added to the column, centrifuged for 1 minute and flow through discarded. Additional 1 min

Table 2.1. Thermocycling conditions for PCR amplification of Klenow reaction product

Temperature	Time	Number of cycles
95^{0}C	5 min	1
60^{0}C	1 min	1
68^{0}C	10 min	1
95^{0}C	30 sec	19
60^{0}C	30 sec	
68^{0}C	5min	
72^{0}C	5 min	1
4^{0}C	Hold	

centrifugation was performed to remove any remaining PE buffer. The column was then placed on a clean 1.5ml tube and 14 µl of Elution buffer added and incubated for 5 minutes. The spin column was spun for 1 minute to recover the purified PCR product. The DNA was then quantified using the Nanodrop 2000 (Thermo Scientific) and run on a 1% agarose gel. The size of the PCR product ranged between 100-200bp (Figure 2.1).

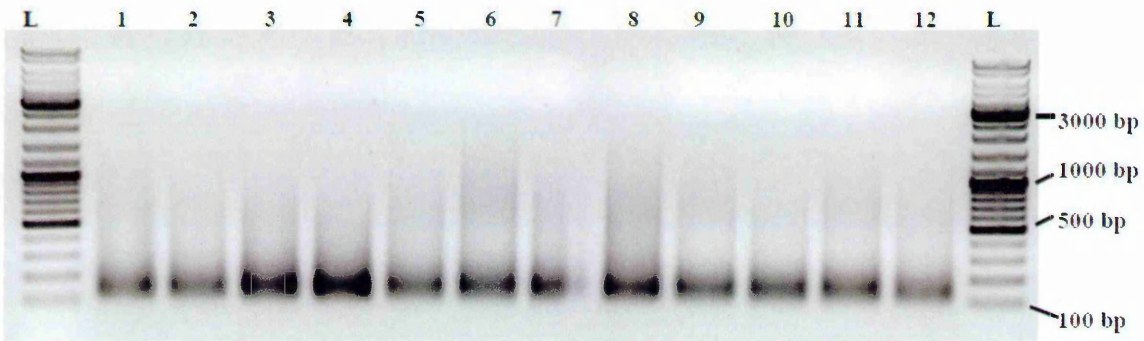


Figure 2.1. Size of amplified DNA using random nonamers
Image of 1% agarose gel containing PCR products from 12 samples (1-12) and sizes indicated by ladder (GeneRuler DNA Ladder Mix (Thermo Scientific) run in 2 wells (L). As expected,the majority of amplified fragments were between 100bp and 200bp.

2.1.5.3 Dye labeling

Cyanine dyes, Cy3 and Cy5 dyes, (GE Amersham) were coupled to reference and test DNA samples respectively. Dye coupling involved mixing of 10 µl of the purified PCR product, with a concentration of at least 3 µg, 2 µl of 0.5M $NaHCO_3$ (Sigma) and 2 µl each of Cy3 and

Cy5 dye, containing 3076 pmol of reactive dye, resuspended in $DMSO_4$ (Sigma). This was incubated at room temperature, in the dark, for 4 hours followed by removal of the uncoupled dye using the QIAquick PCR Purification Kit (Qiagen). The purification procedure was similar to that described in section 2.1.5.2 with changes in the volume of PB buffer added to 700 µl and Elution Buffer of 12 µl. The concentration of the purified labelled material was determined using the 'microarray' mode on Nanodrop 2000 (Thermo Scientific). The labelled material was stored at -20^{0}C awaiting hybridization.

2.1.5.4 Hybridization of sample and reference DNA

The printed slides were hydrated using vapour from 2X saline-sodium citrate (SSC) buffer (Sigma-Aldrich) for 20 sec and dried on a heated block at 100^{0}C. This was followed by UV crosslinking at an energy level of 80,000 microjoules per cm^2 in a HL-2000 HybriLinker™ System (UVP) for 1 min. The slides were then fully immersed in blocking buffer for 45 minutes at 42^{0}C under rotation to reduce non-specific binding to the glass slide. The blocking buffer contained 1% bovine serum albumin (PAA), 5X SSC buffer (Sigma-Aldrich) and 0.1% SDS (Sigma). The slides were then washed with distilled water and dried on a benchtop rotor with slide holders. The microarray chamber known as HX3 mixer (Nimblegen) was assembled on each slide, using a slide alignment tool, ready for sample loading.

Prior to loading the samples, a hybridization master mix, containing 3 µl of 20X SSC buffer, 0.45 µl of 10% SDS and 0.5 µl of 1M HEPES per sample, was prepared. A volume containing 2 µg of Cy3 labeled product, 2 µg of Cy5 labeled product and 3.95 µl of the hybridization mix was combined and topped up, with distilled water, to a final volume of 20 µl. This procedure was carried out in the dark. The tube containing the mixture was placed on a heating block at 100^0C for 5 min, to denature DNA, and allowed to cool at room temperature for 5 min. The mix was then loaded to the assembled slides and placed in a MAUI 12-bay hybridization station (BioMicro Systems) overnight (at least 12 hours) at 65^0C.

The hybridized slides were washed twice in an ArrayIT wash bucket with slide holder. The first wash was in a solution containing 15ml of 20X SSC buffer, 1.5ml of 10% SDS and 510ml of distilled water with magnetic stirring. The second wash solution contained 1.5 ml 20X SSC buffer and 520 ml of distilled water. The slides were then dried by centrifugation. All these procedures were performed in the dark since the dyes are photosensitive.

2.1.5.5 Image acquisition

The slides were scanned by the GenePix 4000B Microarray Scanner (Molecular Devices) using GenePix Pro 4.0 software (Molecular Devices) at optimal wavelengths ranging from 600-690 nm for the Cy5 dye and 350-420 nm for the Cy3 dye and at a resolution of 10 µm. The images were saved as TIFF files (Figure 2.2). Using GenePix Pro 4.0, a GenePix Array List (GAL) file, containing the sizes, positions and probe identifiers of the spots on the array, was overlaid on the images. By visual inspection of the images, poor quality spots and regions were marked as bad. The results were saved as GenePix Results (GPR) files.

2.1.6 Microarray data analysis

2.1.6.1 Pre-processing of microarray data

Analysis of the microarray data was performed using the linear models for microarray analysis (*limma*) package (Smyth 2005) in R. The GPR files were imported into R. The quality of the data was assessed using the flag values associated with each spot. Spots with flag values of less than zero, i.e., -100, -75 and -50 indicated as 'bad', 'absent' and 'not found' respectively were considered of poor quality. The GenePix 4.0 software flagged a spot as 'bad', if by manual inspection, it was marked of poor quality, 'absent' when the GAL file lacked a probe identifier associated with the spot or the identifier was indicated as 'empty' and lastly 'not found' if the spot had less than 6 pixels, or its size greatly differed from that in the GAL file.

Spots with the above flag values were weighted as zero and were not included in further analyses. To remove any experimental variation, the background intensity was subtracted from the foreground by a method known as 'normexp' in *limma* (Ritchie et al. 2007). The method uses a normal plus exponential convolution model where the normal distribution reflects the background intensities and the exponential distribution reflects the signal intensities. The background intensity was calculated from regions outside each spot while the foreground intensity was determined from within each spot. Normalization of the data within the array was performed using "robustspline" (Smyth and Speed 2003) in *limma*. This involves, for each array, normalization of $\log_2$ intensity ratios with respect to the overall intensity values in each array. This normalization minimizes the bias in $\log_2$ intensity ratios observed at low intensities. After this, between-array normalization was performed using the "quantile" method which standardizes the intensity means and inter-quartile ranges across arrays.

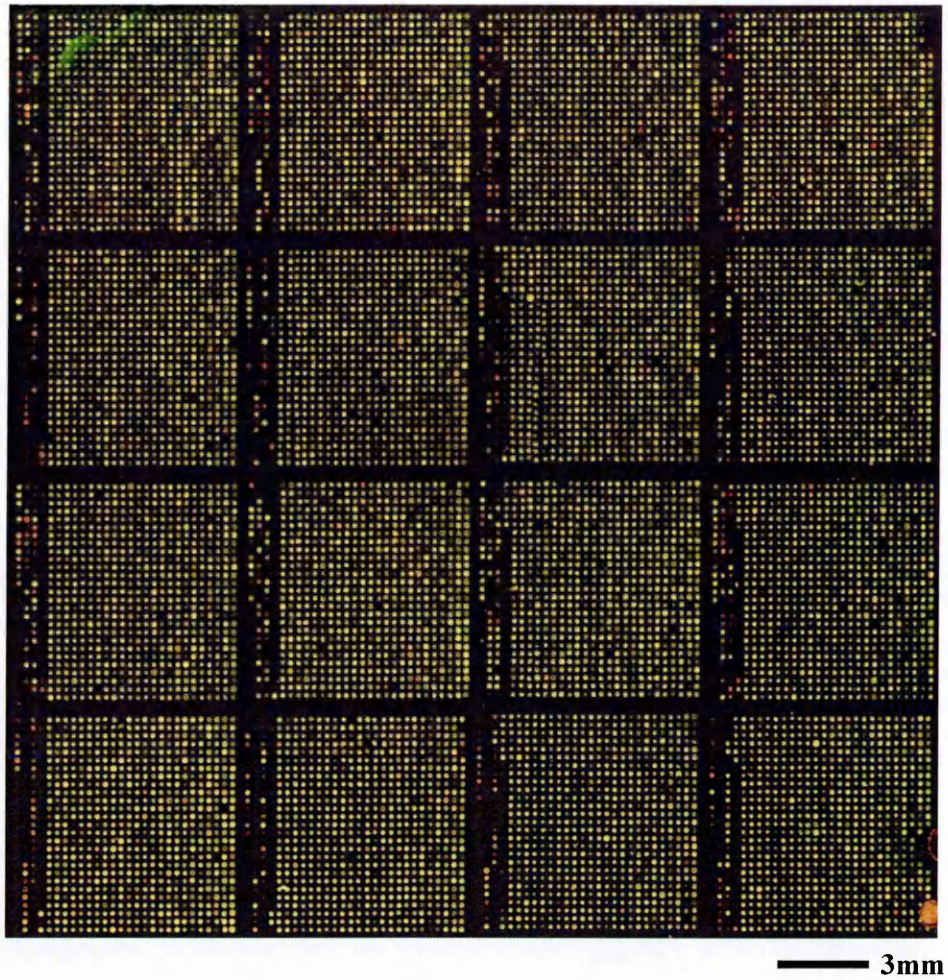


Figure 2.2. Image of a hybridized array
Image of a single array obtained using GenePix 4000B Microarray Scanner. The array contains 16 blocks with red, green and yellow spots. The yellow spots indicate equal hybridization of the test and reference to the oligos. The red spots indicate oligos with higher amounts of hybridization of the test DNA (Cy5 dye (red)) than the reference DNA (Cy3 dye (green)). The green spots indicate oligos with higher amount of hybridization of the reference DNA (Cy3 dye (green)) than the test DNA (Cy5 dye (red)).

2.1.6.2 Detection of gene copy number variation using R-GADA

To identify genomic regions that vary in copy number using microarray data, a R package called Genome Alteration Detection Analysis (*GADA*) was used (Pique-Regi et al. 2010). The input to this program is normalised microarray $\log_2$ intensity data and location of the probes and the output is the location and supporting statistics for contiguous segments of genome that

differ in log_2 intensity (amplitude) from their flanking segments. The program includes two steps; segmentation analysis and backward elimination. The first step involves the conversion of the normalized log_2 intensity ratios into piecewise constant vectors. It then employs a sparse Bayesian learning (SBL) model to identify the possible breakpoints and the segment average log_2 intensity ratio (amplitude). The learning is controlled by the array noise level that is automatically generated from the data and a second sparseness hyperparameter, aAlpha (aα) that is manually set and controls the number of segments generated. When aAlpha is set at a high value, the expected number of segments *a priori* is fewer that when aAlpha is low. In the second step, backward elimination, a statistical analysis, t-statistic, is calculated for each breakpoint detected based on the mean and variance of the segments. Breakpoints with low statistical significance, i.e., with t-statistic less than the set T, are removed. Also included in this step is the filtering of segments based on the minimum number of probes within the segment. Thus the stringency on CNV detection can be applied by altering T, aAlpha and amplitude (magnitude of difference in signal between the putative CNV and its neighbouring genes) parameters in the GADA package. It can further be altered by defining the proportion of probes per gene and genes per CNV required to define a CNV.

2.1.7 Statistical analysis

2.1.7.1 Systematic effects of CNVs detection

Once the CNVs were 'called', other study-specific systematic effects on their prevalence were analysed using mixed effects logistic regression models. The possible effects of population of origin, multiplicity of infection (MOI), haemoglobin, age of participant and parasitaemia (5 fixed effects), experimental batch, isolate (sample) and CNV on the probability of determining the presence or absence of a CNV (dependent variable) were explored. The model was fitted

to each CNV separately using the 'glmer' command with the family as 'binomial' in linear mixed-effects models using the *lme4* package in R. Because of the large number of levels in isolate and CNV, these factors, together with batch, were fit as random effects. Least squares means were calculated for each level of the fixed effects from the fitted models using the *lsmeans* package in R.

2.1.7.2 Functional gene set enrichment analysis of CNVs

Gene set enrichment analysis was performed on the list of genes identified to be copy number variable. The functional gene sets were obtained from Malaria Parasite Metabolic Pathways website: http://mpmp.huji.ac.il/ . The hypergeometric test was used to identify functional groups of genes that were enriched. Hypergeometric test calculates the probability of overlap between two gene lists, i.e., genes detected as CNV and those in a functional gene group. The 'phyper' function in the *stats* package in R was used. Functional groups with p-values of less than 0.05 were classified as statistically significantly enriched.

2.1.7.3 Population genetics analysis

The frequency of each CNV in each of the four populations was calculated. To characterize the between to within population variability in CNV frequencies, Weir and Cockerham F-statistics (F_{ST}) were calculated for each CNV using *hierfstat* as implemented in R (Goudet 2005).

2.2 Confirmation of CNVs by whole genome sequencing using the Personal Genome Machine (PGM™) Ion Torrent

2.2.1 Samples

Twenty-two out of the 183 samples with CGH data were selected for whole genome sequencing. These set of samples had all the 98 CNVs reported in chapter 3 among them. An aliquot of iRBCs in glycerolyte stored in liquid nitrogen, described in section 2.1.2, was used for whole genome sequencing.

2.2.2 Genomic DNA extraction

Genomic DNA extraction was performed in two batches. The first batch of 9 samples were centrifuged and then the glycerolyte was removed. Then gDNA extraction procedure using the saponin lysis method described in section 2.1.3 was used. The second batch were processed using the procedure described in section 2.2.4. The presence of DNA after extraction was confirmed by 1% agarose gel electrophoresis and the quantity of DNA was measured using a Qubit® 3.0 Fluorometer. The DNA quantities ranged between 4-25 ng/µl in a volume of 20µl.

2.2.3 Quantification of the proportions of human and parasite DNA using real time PCR

The SYBR green method using primers (below) against parasite DNA and human DNA was used to quantify the amount of DNA from each organism in each of the samples. The parasite primer set targeted the fructose-bisphosphate aldolase gene (PF3D7_1444800) using the forward and reverse primer sequences below. The primer set was used as an endogenous control in qPCR by Salanti and colleagues (Salanti et al. 2003).

PF3D7_1444800 forward primer: 5'-TGTACCACCAGCCTTACCAG-3'

PF3D7_1444800 reverse primer: 5'-TTCCTTGCCATGTGTTCAAT-3'

A commercially available human DNA primer targeting the beta hemoglobin gene: HUMAN hbb- Hs_HBB_1_SG QuantiTect Primer Assay (Qiagen) was used to amplify human DNA. The qPCR assay was performed in 15 µl reaction with 7.5 µl SYBR Green Master Mix (Applied Biosystems), forward and reverse primers each at a concentration of 500nM, 1 µl (2-40ng) of DNA template and 5 µl DNAse free water. The cycling conditions for qPCR amplification on a 7500 Real-Time PCR System (Applied Biosystems) are listed in Table 2.2.

Table 2.2. Thermocycling conditions for qPCR assay for human and parasite DNA quantification

Stage	Temperature	Time (min)	Cycles
Holding	50^{0}C	2.00	1
Holding	95^{0}C	10.00	1
	95^{0}C	0.01	45
Cycling	60^{0}C	1.00	
Melt Curve	95^{0}C	0.15	1
	60^{0}C	1.00	1

The standard curve method (Applied Biosystems, User bulletin #2) was used to quantify the parasite and human DNA quantities in each sample. The standards were generated from pure parasite and human DNA extracted from a lab adapted parasite line and peripheral blood mononuclear cells (PBMCs) respectively. The standards consisted of 10-fold serial dilution of the DNA with the highest concentration at 50 ng/µl and the lowest at 0.005 ng/µl. The DNA quantities were measured using Qubit® 3.0 Fluorometer. The five human and parasite standards were included in each plate of the qPCR assays. The parasite and human DNA quantities for each sample were interpolated from the standard curves (Figure 2.3). The

fraction of human and parasite DNA in the first 9 samples ranged from 5-71% (Table 2.3). This is comparable to the estimate obtained from next-generation DNA sequencing in most of the samples (Table 2.3).

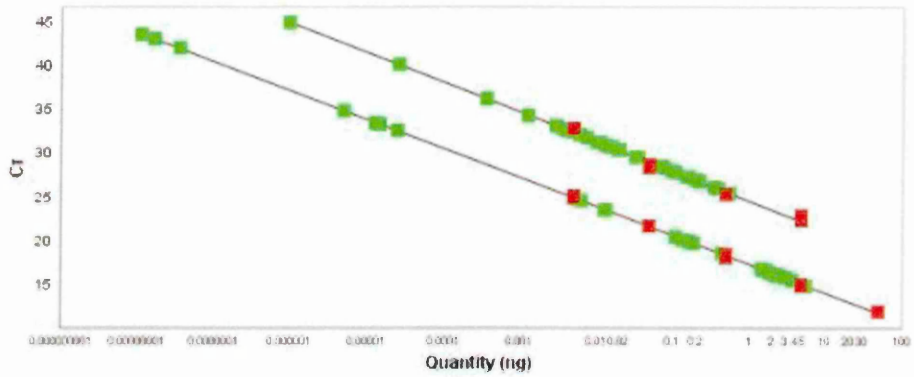


Figure 2.3. Quantification of parasite and human DNA using the standard curve method. Plot (from 7500 Real-Time PCR software) of the relationship between C_T (threshold cycle) value of each well on the Y axis and the quantity of DNA (ng) in the well. The two standard curves, human (upper line) and parasite (lower line), generated from the qPCR measurements of the standards (red points), are shown. The samples assayed are indicated in green. The quantities of DNA in the standards are known and that of the samples are interpolated from the standard curves. The human and parasite standard curves have slopes of -3.38 and -3.31 which correspond to amplification efficiencies of 97.5% and 100.4%, respectively (calculated by the 7500 Real-Time PCR software). The R^2 values were 0.99 in both cases. The efficiencies, slopes and R values were calculated by the 7500 Real-Time PCR software.

2.2.4 Human DNA depletion by DNase treatment

In the first batch of 9 samples parasite DNA consisted of, on average, 37% and 54% of the total DNA content as determined by sequencing and qPCR respectively (Table 2.3). As this is wasteful of sequencing power, a protocol was developed to further minimize the amount of human DNA present in the additional 13 samples to be sequenced. This protocol was based on the principle that storage of iRBCs in glycerolyte preserves the integrity of RBCs but not PBMCs (Farrugia et al. 1993). If PBMCs, but not iRBCs, are lysed upon thawing, addition of DNase to the cells in glycerolyte would be expected to lead to digestion of human DNA but not parasite DNA.

Table 2.3. Proportion of parasite DNA in samples determined by qPCR and also obtained from whole genome sequence data.

Sample Identifier	Parasitaemia par/µl (by microscopy)	Proportion of parasite DNA in sample by qPCR	Proportion of parasite sequences in sequencing output
pf1299	270,000	0.62	0.58
pf1349	74,000	0.55	0.45
pf1624	460,000	0.71	0.54
pf10676	280,000	0.43	0.42
pf10760	200,000	0.60	0.27
pf10814	6,600	0.05	0.13
pf10836	910,000	0.79	0.63
pf10724	290,000	NA	0.22
pfG013	NA	NA	0.05
Mean	600000	0.54	0.37

NA (first column) - No data on parasitemia due to unreadable microscopy slide of sample pfG013 .NA (second column) - Failed qPCR reaction. May be due to presence of PCR inhibitors The proportion of parasite sequences in sequencing output was calculated from the total sequence reads that mapped to 3D7 (as described in section 2.2.6) divided by the total sequence reads obtained for each sample.

The 13 samples stored in glycerolyte were centrifuged at 2156 x g for 3 min and glycerolyte removed. To the samples, 10X DNase buffer (Ambion) and 6 µl (12 units) of DNase (Ambion) enzyme were added and incubated at 37^0C for 30 minutes. The enzyme was inactivated by 2 min incubation, at room temperature, with addition of EDTA at a final concentration of 15mM. The samples were resuspended in 2.5mM EDTA in 1 X PBS to a volume of 1 ml. gDNA extraction was undertaken as described in section 2.1.3 using the saponin lysis method. The percent of parasite DNA sequences after whole genome sequencing ranged between 5 and 98% in the 13 samples (Table 2.4). Unfortunately, no controls were tested in the human DNA depletion experiment to assess its success. However, the observation of comparable proportions of parasite DNA between blood samples from Sudan (pfM007 and pfM004) and Kisumu (pfK007, pfK020, pfK071) that were not subjected to prior removal of

white blood cells (section 2.1.2) and Kilifi samples (pf1212, pf10770, pf1590, pf10820 and pf10495) that had white blood cells removed (section 2.1.2) is promising.

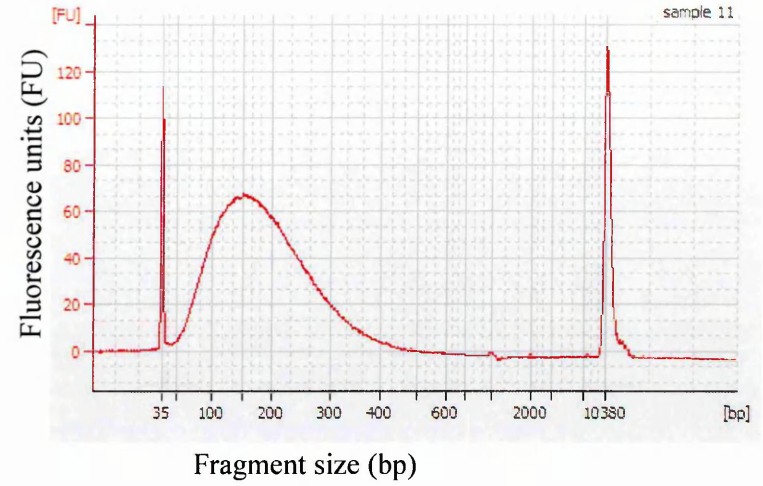
Table 2.4. Proportion of parasite DNA sequences in whole genome sequence data of samples after DNase treatment

Sample identifier	Proportion of parasite DNA sequences in sequence output
pf1212	0.93
pf10770	0.62
pf1590	0.18
pf1895	0.98
pfK007	0.90
pfK020	0.94
pfK071	0.43
pfM004	0.70
pfM007	0.96
pfK065	0.05
pf10820	0.87
pf10578	0.08
pf10495	0.98

2.2.5 Library preparation of gDNA for sequencing

Library preparation of gDNA was performed using the Ion Xpress™ Plus Fragment Library Kit (Part No. 441269) according to the Ion Torrent protocol. 20 ng -100 ng of gDNA was fragmented using Ion Shear™ Plus reagents (Part No. 441248) to a median fragment size of between 200-300bp. The fragmented product size was assessed using Agilent Bioanalyzer™ 2100 (Agilent Technologies, Inc) (Figure 2.4A). Adapters were ligated to the fragmented DNA, nick repaired and purified using Agencourt AMPure beads (Beckman Coulter). The target fragment size of approximately 330bp was selected for using E-Gel SizeSelect™ 2% agarose gel (Life technologies). Five cycles of PCR amplification of the size-selected DNA was performed (Figure 2.4B).

A)



B)

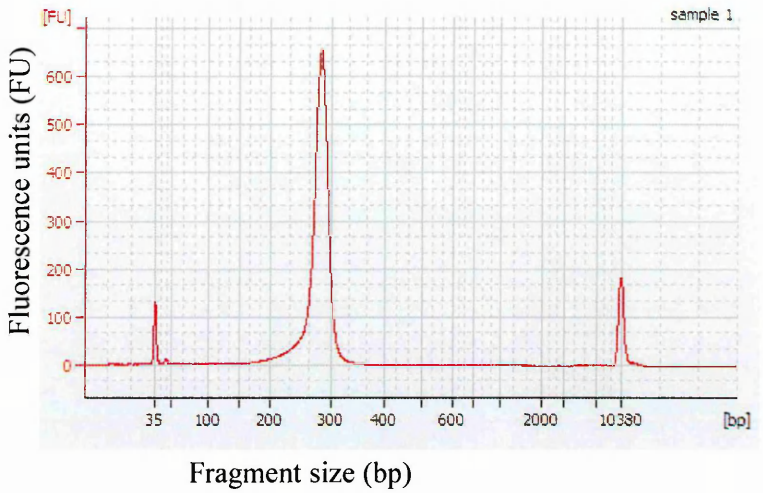


Figure 2.4. Distribution of DNA fragment sizes
Graphs from Agilent Bioanalyzer™ 2100 showing the distribution of fragment sizes A) after enzymatic shearing of gDNA of sample loaded on well 11 and read as sample 11 by the machine (pf1299) B) after size selection of fragmented DNA, for a 200bp library, and 5 cycles of PCR amplification of sample loaded on well 1 (pf1299). The fluorescence units (FU) on the y axis represents the quantity of DNA fragments and the sizes of the fragments, in base pairs (bp) are shown on the x axis. The peaks at 35bp and 10380 are the sizing marker peak.

The quantity of the amplified product for use in emulsion PCR on the Ion One Touch

Instrument (Life Technologies) was assessed using Ion Library Quantitation Kit on the 7500

real time PCR machine (Applied Biosystems). A volume of the amplified products that corresponded to a concentration of 13pM of molecules were taken to emulsion PCR step. Emulsion PCR involves clonal amplification of individual DNA templates on magnetic beads in droplets formed in water-oil emulsion. The Ion PGM OneTouch Two Template Kit (Part No. 4480974) was used in the emulsion PCR. After the PCR, the percentage of templated ion spheres was assessed using Ion Sphere Quality Control Kit in order avoid sequencing of samples with insufficient templated Ion Sphere Particles or samples with multiple templates per Ion Sphere Particle. Enrichment of Ion Spheres with templates was then carried out on the Ion OneTouch ES. The enriched templated spheres were loaded onto an Ion 318 Chip (Part No. 4469497) with a capacity of up to 2 Gbp and sequenced. Signal processing and base calling was performed in the Ion Torrent Server. Trimming of 3' end of reads based on base quality was turned off during base calling. The raw files in Standard Flowgram Format (SFF files) were converted to FastQ files using a plugin known as *FastQCreator* in the Torrent Server. The FastQ files were then exported for analysis.

2.2.6 Sequence data analysis

2.2.6.1 Mapping of reads to reference genome

The sequence reads contained in the FastQ files were mapped to a reference genome using the Burrows Wheeler Aligner *(BWA)* (Li and Durbin 2010). The reference genome used was 3D7 version 3, available from GeneDB (Logan-Klumpler et al. 2012). The BWA-MEM option suitable for mapping of reads greater than 70bp was used and output to files in Sequence Alignment/Map (SAM) format. These files were converted to compressed binary versions (BAM format) using SAMtools (Li et al. 2009). Alignments with low alignment quality

scores, i.e., less than 30 ,and reads mapping to multiple locations in the genome together with non-mapped reads were excluded from further analysis using SAMtools (Li et al. 2009).

2.2.6.2 CNV detection using sequence data

To detect CNVs in the sequence data, Copy Number estimation by Mixture Of PoissonS *(cn.MOPS),* in R, was used (Klambauer et al. 2012). Using *cn.MOPS* package, the BAM files were imported in R from which the read coverage across non-overlapping genomic windows was calculated. Normalization of read counts across all samples was performed using the 'quantile' method to correct for differences in sequencing depth of the samples. To identify CNVs, *cn.mops* employs a mixture of Poisson distributionsmodel at each window across the samples. The model assumes that the read coverage for a specific copy number across samples has a Poisson distribution and in instances of presence of different copy numbers, the model is a mixture of Poisson distributions. An initial CNV call is made when there exists variation in coverage across the samples, e.g., a $\log_2$ fold difference of less than -0.8 (loss) or greater than 0.8 (gain) between sample and the mean normalized coverage in the window. A second CNV call is made when consecutive initial calls along a chromosome are joined by segmentation analysis using the 'fastseg' method and called as a single CNV.

2.3 Transcriptome profiling using microarrays

Transcription profiling experiment was performed by Rono, M., Nyonda, M., Ngoi, J., Simam, J., Mackinnon, M. J. (unpublished). Below are the procedures that were undertaken in generation of transcriptome data.

2.3.1 RNA extraction

An aliquot of blood sample processed as described in section 2.1.2 was used for transcription profiling. Prior to storage, each fresh blood aliquot, 200 µl-1000 µl was subjected to laboratory culture to obtain parasites at all the 48 hr asexual stages, i.e., ring, trophozoite and schizont stages. Seven aliquots, of equal volume, were obtained from culture at an interval of 10 hours each and the iRBCs pellet stored in TRI Reagent® (Sigma-Aldrich) at -80^{0}C. I was involved in the culture of some of the parasites and set up of the sample database. To extract RNA from blood, the stored aliquot was thawed and TRI Reagent® (Sigma-Aldrich) was added to a total volume of 1000µl and mixed. 240µl of Chloroform (VWR) was then added mixed and centrifuged at 2000 x g for 10 minutes at 4^{0}C. The aqueous layer containing nucleic acid was separated, 500µl of Isopropanol (Sigma-Aldrich) added to it and stored at -20^{0}C overnight to enable precipitation of the nucleic acids. The sample was spun at a speed of 12000 x g for 1 hr at 4^{0}C and the supernatant removed. The pellet was then washed with 500µl of ice-cold 75% Ethanol (Sigma-Aldrich) and dried. The pellet was resuspended in 20µl of RNAsecure (Invitrogen) and incubated at 60^{0}C for 10 minutes. The RNA quality and quantity was assessed by Nanodrop and 1% Agarose gel.

2.3.2 cDNA synthesis, amplification and hybridisation

First, reverse transcription reaction was performed on 500ng of RNA. To a volume of RNA containing 500ng, 2 µl of a mixture of primers containing SMART-dT (MWG), SMART-Random (section 2.5.1.2), SMART-TS (MWG) at concentrations of 25µM, 25µM and 50µM respectively was added and the total reaction volume made up to 8 µl by adding distilled water. The primer sequences for SMART-dT and SMART-TS are shown below. The reaction mix was incubated at 65°C for 5 minutes and at 4°C for 10 minutes. Secondly, cDNA

synthesis was carried out using the SuperScript II Reverse Transciptase kit (Life technologies). It involved making a reaction mixture containing 4 µl of 5X First-Strand, 2 µl of 100mM DTT, 2 µl of 3mM dNTPs, 1 µl of 40U/ µl RNAse OUT (Invitrogen) and 2 µl distilled water to the first reaction. The mixture was mixed and incubated at room temperature for 2 minutes. 1 µl of 200U/µl Superscript II (Invitrogen) was added then incubated at 42°C for 50 minutes and then 70°C for 15 minutes. The third step involved amplification of the cDNA, a method similar to the amplification step of gDNA (section 2.1.5.2). 2 µl 5U/µl Taq polymerase (New England Biolabs), 10 µl 10X Taq thermopol buffer (New England Biolabs), 1.5µl aa-dUTP/dNTP mix (described in section 2.1.5.2), 6µl SMART-amplification primer (section 2.5.1.2), 5µl of cDNA synthesis product and 75.5 µl of ddH_2O were mixed and placed in a thermocycler under the following conditions below (Table 2.5).

SMART TS primer sequence: 5'-AAGCAGTGGTATCAACGCAGAGTACGCGGG - 3'

SMART dT primer sequence: 5'-AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTT TTTTTTTTTTTTTTTTTTTAGCN-3'

Table 2.5. Thermocycling conditions for PCR amplification of cDNA

Temperature	Time	Number of cycles
95^0C	5 min	1
60^0C	1 min	1
68^0C	10 min	1
95^0C	30 sec	22
60^0C	30 sec	
68 ^{0}C	5min	
72 ^{0}C	5 min	1
4 ^{0}C	Hold	

The reference parasite line used for hybridisation was an isolate similar to that of CGH experiment; Kilifi isolate adapted to lab culture, 'P4'. The amplified product used as reference

for microarrays was prepared by pooling, at equal concentrations, the RNA from different lifecycle stages of the reference parasite line prior to cDNA synthesis and amplification.

The amplified product was purified, labelled with dyes and hybridized using the procedure described in section 2.1.5.2, section 2.1.5.3 and section 2.1.5.4 respectively. The image acquisition and microarray data pre-processing was performed as described in sections 2.1.5.5 and 2.1.6.1, respectively.

2.3.3 Statistical analysis

The effect of CNVs on the expression of genes located within them was assessed by performing a Pearson correlation between the mean CGH ratio per gene and 'Mean' expression data obtained from Rono, M., Nyonda, M., Simam, J., Mwongeli, J., Mok, S., Marsh, K., Bozdech, Z. and Mackinnon, M.J. (unpublished work). The p-value for assessing significant differences from the expected correlation of zero under the null hypothesis of no relationship between gene copy number and expression levels was calculated using the *cor.test* function in R.

To investigate whether the correlation obtained from the analysis may have been by chance, a permutation test was performed. 100 random permutations of the expression data of each gene were generated and the correlation between mean CGH ratio of each gene and each of the 100 randomized expression data per gene was calculated. A test for significant differences between the distribution of observed correlations and distribution of correlations from the permuted data was applied using the Kolmogorov–Smirnov test implemented in the *ks.test* function in R.

To analyse the relationship between gene expression (of genes in the whole genome) and the CNVs detected, a linear regression model was applied using the *lm* function in R. The mean $\log_2$ expression ratio of each gene was the dependent variable while the copy number state (fitted as a fixed effect factor with levels for loss, gain or normal) was the independent variable. The significance of the difference in gene expression between samples with a gene copy number difference and those without was assessed from the p-value of the regression coefficient from the same model.

Chapter 3

Population-wide survey of gene copy number variation in *Plasmodium falciparum* isolates from Africa

3 Chapter 3: Population-wide survey of gene copy number variation in *P. falciparum* isolates from Africa

3.1 Introduction

Genome-wide scans for gene copy number variants (CNVs) in *P. falciparum* have led to the discovery that CNVs contribute to up to 1% of the genetic differences between parasite genomes (Anderson et al. 2009). CNVs have been linked to differences in gene expression among parasite strains (Mackinnon et al. 2009, Gonzales et al. 2008, Nair et al. 2008) and have also been associated with important phenotypes including drug resistance (Nair et al. 2008, Price et al. 2004, Dharia et al. 2009), erythrocyte invasion (Jiang et al. 2008b), cytoadherence and gametogenesis (Shirley et al. 1990, Biggs et al. 1989, Kemp et al. 1992). However, most of these surveys have taken place in laboratory-cultured isolates and only a small fraction of parasite isolates studied so far are direct from infected individuals. Certain CNVs are known to arise *in vitro* (Nair et al. 2010) and for this reason parasites subjected to *in vitro* culture for long periods of time are not a good representation of naturally occurring CNVs. The main goal of this study was to investigate the potential adaptive role of CNVs in nature through the largest CNV survey to date of natural populations of parasites.

A genome wide survey of CNVs in 183 parasite isolates from three populations in Eastern Africa with different malaria transmission intensities was conducted. The CNVs were detected by comparative genomic hybridization using a 70mer oligonucleotide microarray. First, the microarray data were subjected to normalisation and quality screening to remove effects of technical variation. This was followed by CNV calling using Genomic Alteration Detection Analysis (GADA) package in R. In order to choose suitable thresholds that would maximize the number and accuracy of CNVs detected, a range of parameters for quality filtering and

CNV calling and their impact on CNV detection were explored. In addition, to rule out any possible systematic bias, including sample and patient characteristics, other than population effects on the chances of detecting a CNV, a mixed effects logistic regression model was used to examine these potential influences on CNV detection. A summary of the CNVs identified including their genomic locations, gene content, sizes and frequencies in each populations was generated. Furthermore, population differentiation in CNV frequencies was determined, using Weir and Cockerham's F_{ST} estimates, to identify those CNVs that show strong differentiation between populations indicative of directional selection. Lastly, functional enrichment of gene groups in the genes located within the CNVs was investigated.

3.2 Methods

3.2.1 Experimental strategy

Comparative genomic hybridization (CGH) was performed in order to detect CNVs (described in section 2.1). DNA extracted from infected blood of 183 patients was amplified using random nonamers and competitively hybridized on an oligonucleotide array against a common reference genome. The reference parasite line originated from a patient resident in Kilifi, Kenya, and had undergone adaptation to *in vitro* culture (Mackinnon et al. 2009). The parasite line was in culture for approximately 100 cycles. Randomization of the samples was carried out during whole genome amplification and microarray hybridization to minimize any experimental bias. The 70mer oligonucleotide array used contained probes targeting approximately 5500 genes (Bozdech et al. 2003b). Each gene was targeted by, on average, 2 probes. The microarray data were pre-processed using linear models for microarray data *limma* in R (Smyth 2005) and CNVs were identified using GADA (Pique-Regi et al. 2010). The experimental design is illustrated below.

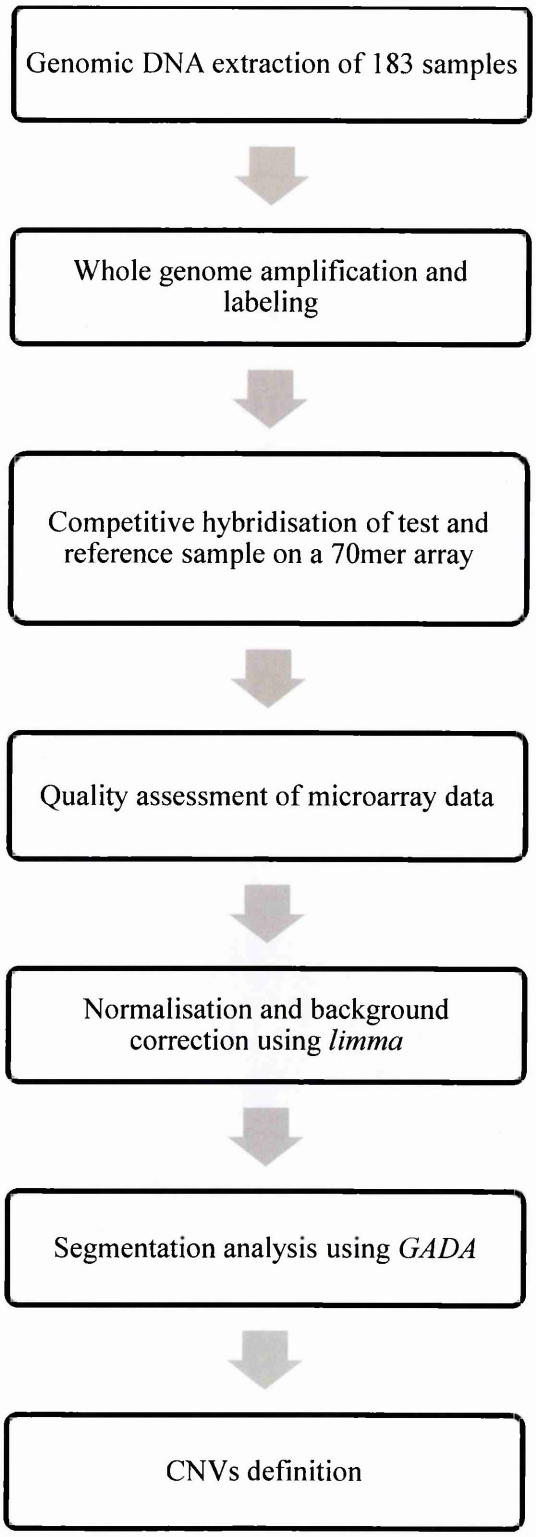


Figure 3.1. Overview of the process of CNV detection using microarrays.

3.2.2 Sample populations

A total of 183 *P. falciparum* isolates were obtained from individuals living in three geographical regions in Eastern Africa, i.e., Kilifi and Kisumu in Kenya and in Sudan, that differ in malaria transmission intensity (Noor et al. 2014, Okiro et al. 2010). The samples from Kilifi were obtained at two time points: 1994-1996, a period with higher transmission intensity, and 2010 when transmission intensity was lower (O'Meara et al. 2008). These two Kilifi populations will be referred to as "Kilifi pre-malaria decline" and "Kilifi post-malaria decline" respectively. Characteristics of each population are shown in Table 3.1. The age of the malaria patients significantly differed among the four populations (p-value < 0.05 by Mann-Whitney-Wilcoxon Test). Its effect on CNV prevalence is investigated in section 3.3.5.

Table 3.1. Characteristics of the four sample populations.

Population	Kisumu	Kilifi Pre-malaria decline	Kilifi Post-malaria decline	Sudan
Malaria transmission intensity	High	Moderate	Low	Low
Number of samples	49	33	49	52
Year of sample collection	2008	1994-1996	2010	2007
Median age in months (range)	36 (6-72)	30 (11-37)	53.5 (14-147)	84 (12-612)
$\log_{10}$ median parasitaemia (par/µl) (range)	5.3 (4.9-5.8)	4.7 (4.0-5.8)	5.1 (2.5-6.1)	5.1 (4.4-5.8)
Median haemoglobin in g/dl (range)	9.9 (5.2-15.2)	9.4 (5.3-13.2)	10.6 (3.4-12.1)	9.6 (3.2-12.9)
Median number of clones (range)	2 (1-6)	2 (1-7)	2 (1-5)	2 (1-5)
Monoclonal infections (percentage)	10.2	33.3	18.4	26.9

3.2.3 Quality control of microarray data

Quality of hybridisation of the spots on the microarray was measured by the consistency in red intensity (test sample) and green intensity (reference sample) across a spot. The correlation between red and green fluorescence intensity of each pixel in a spot was determined by the image acquisition software, GenePix. From this, the square of the correlation coefficient, termed 'Rgn R^2', with values between 0 and 1, was used as an indicator of its quality, with values of less than 0.6 flagged as 'bad'. In addition, microarray spots were also automatically and manually flagged as bad by visual inspection using the GenePix software (described in Chapter 2.1.6). Poor quality spots were assigned a weight of 0 and good quality spots given a weight of 1.

Furthermore, data from probes in the microarray targeting highly polymorphic gene families including variant surface antigens (*vars, rifins and stevors*) and also genes with known SNPs located within the probe sequences based on all SNPs data available in PlasmoDB version 9.1 (SNPs information available for 37 isolates) were excluded. These polymorphic probes constituted 17% of probes on the microarray. The presence of 7 mismatches was found to lead to, on average, a reduction of 64% of the microarray hybridisation intensity (Bozdech et al. 2003b) which could be called as a CNV. Probes targeting more than one gene location (6% of the total number of probes) were filtered out.

3.2.4 Pre-processing of microarray data

The total red and green fluorescence intensity of a spot, i.e., the foreground intensity, correlates with the amount of hybridization of the test sample and the reference sample respectively. Included in spot intensity measurement is signal from non-specific hybridization

known as background signal estimated from regions surrounding the spots by the GenePix software. The background intensity was subtracted from the foreground intensity using linear models for microarray data *limma* package in R (Smyth 2005). To further minimize technical variation as a result of poor hybridization in part of a slide due to a bubble or poor printing, within array normalisation of microarray data was performed using *limma* (Smyth 2005). Technical variation between arrays due to, for example, differences in printing quality and dye concentrations was adjusted for using the 'quantile' normalisation method in *limma*. Detailed explanation of these analysis methods are in section 2.1.6

3.2.5 CNV detection using Genome Alteration Detection Analysis (GADA)

To detect CNVs, Genome Alteration Detection Analysis (GADA) package in R was used (Pique-Regi et al. 2010). This performs segmentation analysis to detect adjacent regions in the genome that differ in gene copy number. The input to GADA is the normalised $\log_2$ intensity ratios for each probe ordered by genomic position. The output is a list of chromosomal segments which differ in their average $\log_2$ intensity ratios from those in adjacent regions (Figure 3.2). CNV calling is restricted by certain parameter threshold settings including the amplitude (A) of segment (the average $\log_2$ intensity ratios of probes in a segments), T value (the t-statistic above which breakpoints are excluded) and aAlpha ($a\alpha$) (a measure of the degree of segmentation). These are described in detail in chapter 2 section 2.1.6.2.

3.2.6 CNV definition

For the remainder of the analyses, the cut-offs at the segmentation analysis step in GADA were set at T= 3.5 and $a\alpha$=0.2, a low stringency, on the basis that prior knowledge of the expected degree of segmentation was unavailable and also to maximize on the number of

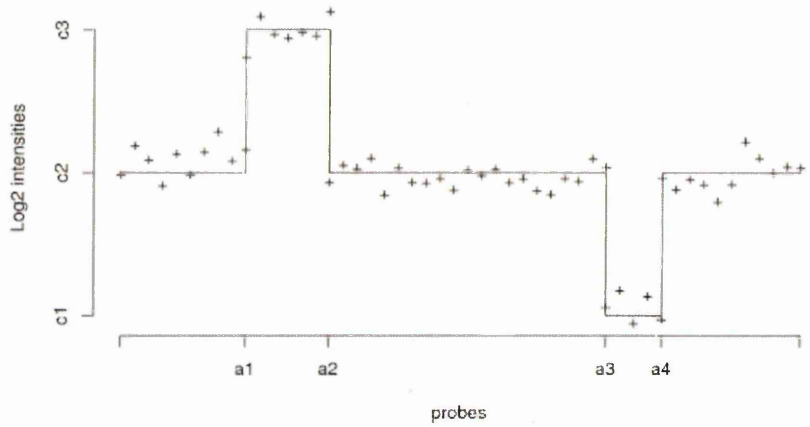


Figure 3.2. CNV detection using GADA
An illustration of CNV detection using GADA. The log_2 intensity ratio of probes (+ sign) ordered by genomic positions. The segments detected by GADA are indicated by horizontal lines. The amplitudes (A) of these segments indicated by log_2 intensity ratio of c1, c2 and c3. The breakpoints of the segments indicated by probes a1, a2, a3 and a4. Image reproduced from Pique-Regi et al. 2010.

CNVs identified. The stringency was increased by defining a segment to be copy number variable if it contained a minimum of two consecutive probes with an absolute average log_2 intensity ratio greater than 1, i.e., a two-fold increase or reduction in copy number. Due to measurement error in the data, the boundaries, i.e. the start and end of segments identified by GADA, may not be precise and thus appear to vary among isolates thus making it difficult to distinguish artefactual from real (Figure 3.3). Therefore, segments that overlapped in genomic position among samples, but had different start or end genomic locations, were considered to be the same CNV (Figure 3.3). Lastly, CNVs that appeared in less than 2% of the isolates studied were excluded from further analysis to allow for the possibility that they were artefactual. Each CNV was assigned a unique identifier that included the chromosome number, e.g., cnv13_473, indicates that the CNV is on chromosome 13 and has unique number identifier 473.

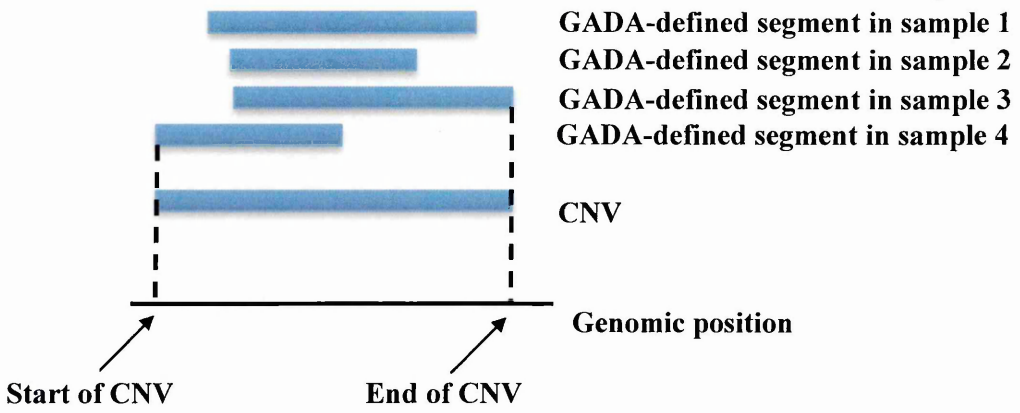


Figure 3.3. A schematic representation of the definition of CNVs breakpoints

3.2.7 Analysis of systematic effects on CNV detection

In order to determine whether there was systematic bias in the probability of detecting CNVs that may distort comparisons of CNV frequencies between populations, The effects of multiplicity of infection (MOI), parasitaemia, patient characteristics (age and haemoglobin), in addition to population, were analysed using generalized linear mixed effects model under a binomial model in the *lme4* package in R. The aforementioned factors were classified as fixed effects. Three further effects - experimental batch, sample and individual CNV - were also included in the model as random effects. The response variable was binary, i.e., presence/absence of a CNV. Least-squares means were estimated from the fitted model for each category of the fixed effects using the *lsmeans* package in R. The overall effect of each factor was determined using ANOVA.

3.2.8 Assessing reproducibility of microarrays

Included in the set of samples for this CNV survey were 8 isolates that had been independently assayed twice. The 8 isolates were randomly chosen. To assess reproducibility of the

microarray in identification of CNVs, the correlation between duplicates of the $\log_2$ intensity ratios of only the probes falling inside the detected CNVs was calculated for each of the 8 pairs. Also calculated was Cohen's kappa value that measures the agreement between two replicates.

3.2.9 Gene set enrichment analysis of CNVs

Gene set enrichment analysis was performed on the list of genes identified to be copy number variable. The functional gene sets were obtained from Malaria Parasite Metabolic Pathways website: http://mpmp.huji.ac.il/ . The hypergeometric test was used to identify functional groups of genes that were enriched. The 'phyper' function in the *stats* package in R was used. Functional groups with p-values of less than 0.05 were classified as statistically significantly enriched.

3.2.10 Population genetics analysis

The frequency of each CNV in each of the four populations was calculated. To characterize the between to within population variability in CNV frequencies, Weir and Cockerham F-statistics (F_{ST}) were calculated for each CNV using *hierfstat* as implemented in R (Goudet 2005).

3.3 Results

3.3.1 Quality of microarray data

A large proportion, 87%, of the microarray data was of good quality (Figure 3.4A). The majority of the spots had weights of 1 and Rgn R^2 values greater than 0.7. The proportion of

data finally retained for further analysis after applying quality filters described in section 3.2.3 was greater than 65% in most arrays (Figure 3.4B).

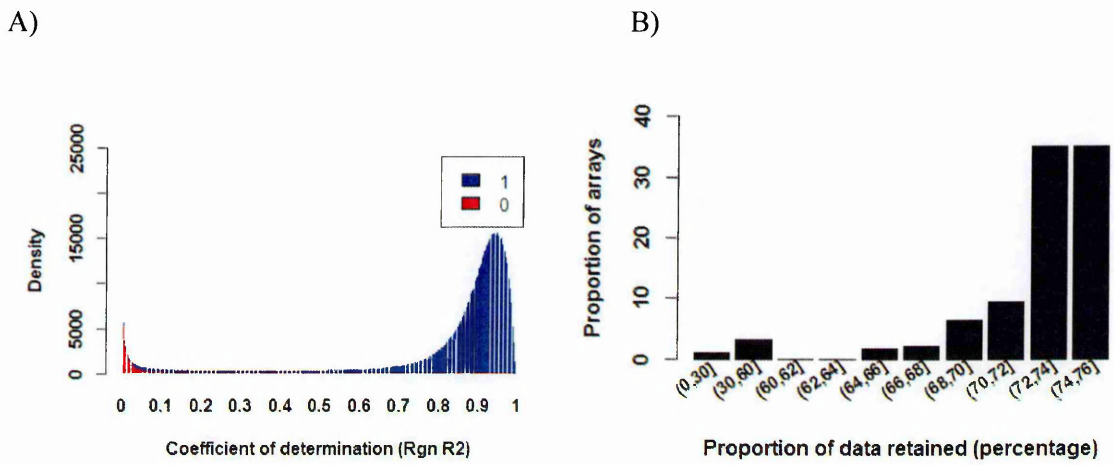


Figure 3.4. Summary of quality of microarray data
A) Histogram showing the distribution of Rgn R^2 and the weights (1, blue; 0, red) of spots in the array data generated from 183 samples. Rgn R^2 value of 0 and weight of 0 indicate the worst quality and 1 is the best quality data from a microarray spot. B) Distribution of the proportions of arrays (in percentage), y-axis, with the proportion (percentage), x-axis, of microarray data retained for further analysis after quality filtering and removal of probes targeting highly polymorphic gene families including variant surface antigens, and probes with known SNPs within their sequences.

3.3.2 Background correction and normalisation of microarray data

Technical variation among the arrays was minimized as follows. Firstly, the foreground fluorescence intensity values that are used for calculation of the $\log_2$ intensity ratios were corrected for background noise (Figure 3.5A vs. 3.5B). Secondly, bias in $\log_2$ intensity ratios (M) (M = $\log_2$ (Red intensity (test sample)/Green Intensity (reference)) observed at low average intensity values (A) (A= 1/2 $\log_2$ (Red*Green intensities) observed within an array was removed (Figure 3.5C vs. 3.5D) and lastly, between array normalisation was not necessary because it did not have major effect on the distribution of intensities between the

arrays after the data was corrected for background noise and within array normalisation performed (Figure 3.5E vs. 3.5F).

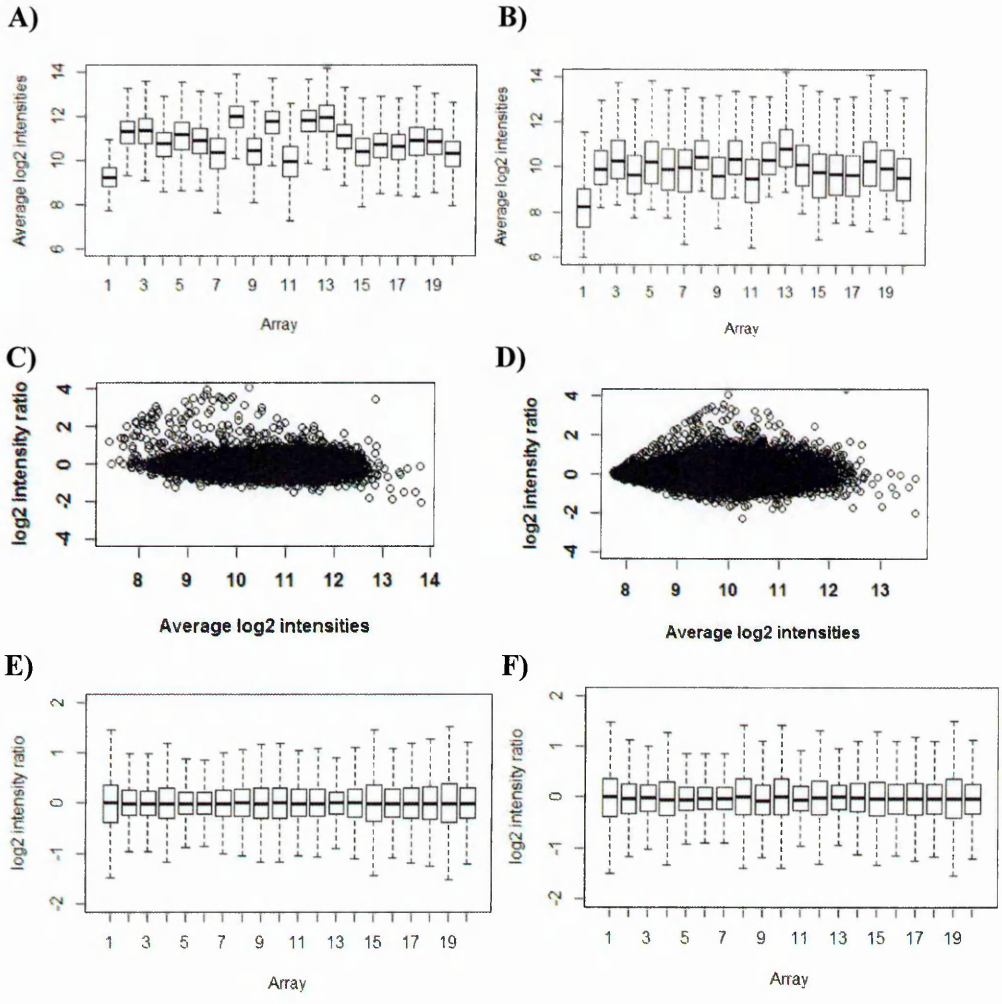


Figure 3.5. Pre-processing of microarray data.
Box plots of average $\log_2$ intensity values of a subset of 20 arrays of **A)** raw data **B)** background corrected data. Plot of relationship between $\log_2$ ratio (M) (Y axis) and average $\log_2$ intensity (A) (X axis) of a single array **C)** before within array normalisation **D)** after within array normalisation using the 'robustspline' method. Boxplot of the average intensity ratios (M) **E)** before between array normalisation and **F)** after within array normalisation using the 'quantile' method. The $\log_2$ intensity ratio (M) is calculated from the equation $M = \log_2$ (Red intensity/Green intensity). Average $\log_2$ intensity (A) is calculated as $A = 1/2 \log_2$ (Red*Green intensities), i.e., the geometric mean of the red and green intensities.

3.3.3 CNV detection using Genome Alteration Detection Analysis (GADA)

An example of a CNV segments generated in chromosome 13 using GADA in one of the samples is illustrated below (Figure 3.6). The number of probes within each segment and the amplitudes of the segments was different from the consecutive segment. Later, the effects of different cut-offs on the number of probes and amplitude size in defining a CNV were examined. In this study, the stringency in 'calling' CNVs at both the pre-input stage and at the GADA analysis stage was altered. At the pre-input stage, the threshold quality control filter, Rgn R^2, was varied and found that this parameter did not greatly affect the number of segments detected by GADA regardless of the chosen amplitude, A (Figure 3.7A). At the GADA analysis stage, increasing the absolute amplitude cut-off (Figure 3.7A) and the value of 'T' leads to fewer segments identified (Figure 3.7B).

By contrast, increasing aα (degree of segmentation) results in a decrease in the number segments detected (Figure 3.7B). Additionally, the cut-off can be set at the minimum number of probes within a segment for it to be considered as a CNV. Thus, as expected, the use of less stringent parameters results in more 'noisy' CNV 'calls'. On the other hand, use of high stringency parameters causes some CNVs to be missed. Thus there is an optimal stringency which ought to be determined.

3.3.4 Reproducibility

High correlation was observed between duplicates of the $\log_2$ intensity ratios of only the probes falling inside the detected CNVs with values ranging from 0.46 to 0.91 across duplicates (Figure 3.8). The highly correlated $\log_2$ intensity ratios were greater than 1 or less than -1 in both duplicates of a sample (Figure 3.8). Due to less accurate identification of CNV

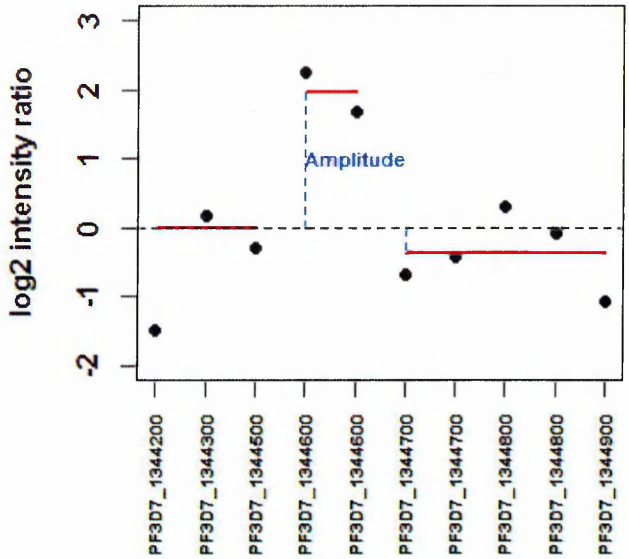
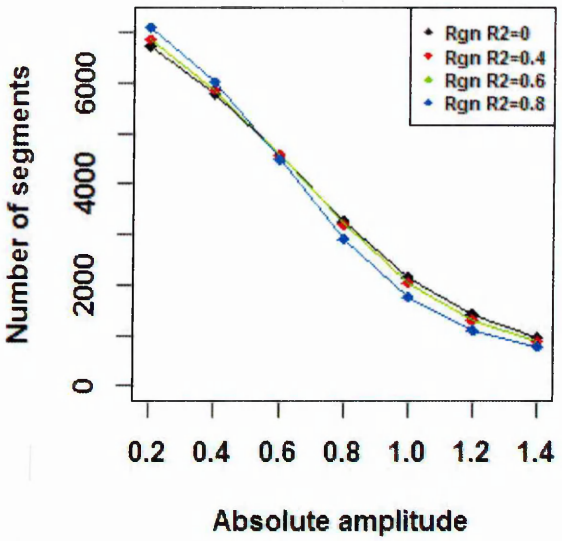


Figure 3.6. Segmentation using GADA on a region on chromosome 13 (CNV: cnv13_473) Illustration of segmentation on a region on chromosome 13 in a single isolate (pf10882). Each point is a $\log_2$ intensity ratio of a given probe. A cut-off of a minimum of 2 consecutive probes in a segment was set. The probes are ordered according to genomic location. On the X-axis are the genes targeted by the probes. The genes that appear more than once are targeted by more than one probe. The red horizontal lines are the CNV segments detected by GADA with their amplitudes indicated on the Y axis, delineated by blue dashed lines. This CNV was assigned the name "cnv13_473" which indicates that it is a CNV on chromosome 13 and has unique identifier 473.

breakpoints, some probes within CNVs (red points) that had $\log_2$ ratios between 0, 1 and -1 were observed (Figure 3.8). The reproducibility of CNV detection was assessed using Cohen's kappa statistic. The kappa values for the replicate pairs ranged between 0.1-0.5 (Table 3.2).

Poor concordance was observed in the CNVs detected between some replicate pairs (Kappa value <0.20). Fair (kappa value 0.21-0.40) and moderate (kappa value 0.41-0.60) concordance was observed in other replicate pairs. The eight isolates had multiple parasite clones. The low concordance observed may be as a result of noise in microarrays that contribute to uncertainty in the threshold for calling a CNV. A threshold of $\log_2$ ratio of 1 would miss calling some of the segments that have values close to 1, e.g., 0.9 that may be true CNVs.

A)



B)

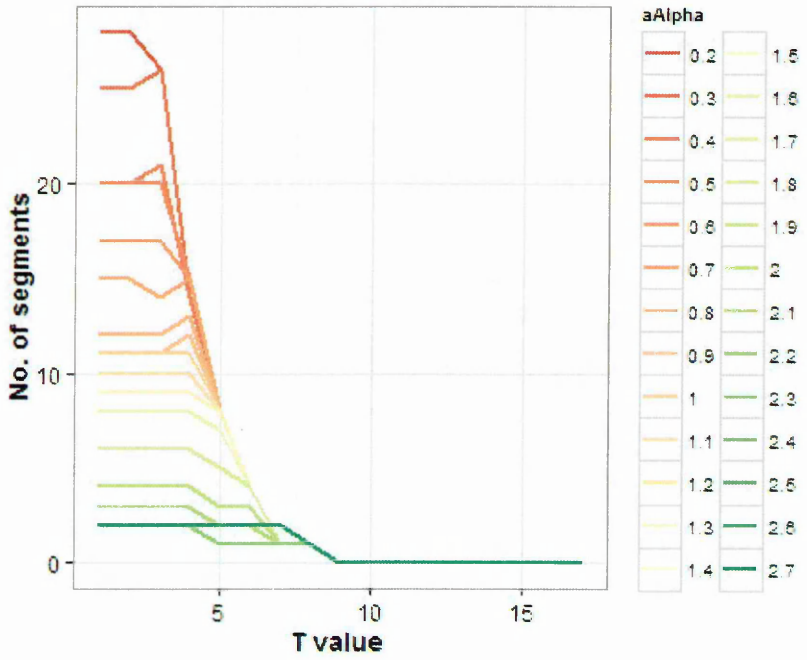


Figure 3.7. Effect of different stringency measures on CNV calling
A) The relationship between total number of segments called (y-axis) against average intensity values of the segments (x-axis) of 183 arrays. The different coloured lines represent different quality filters (Rgn R^2) on the data input to GADA. **B)** The effect of a combination of all the possible GADA settings of T (range from 1 to 17) and aα (range from 0.2 to 2.7) on the number of segments identified, with absolute amplitude A >0.8 and a minimum of 2 probes within the segment, of a single array.

Table 3.2. Table showing the kappa value of each replicate pair of samples.

Sample	kappa value
1	0.21
2	0.11
3	0.27
4	0.29
5	0.49
6	0.49
7	0.35
8	0.10

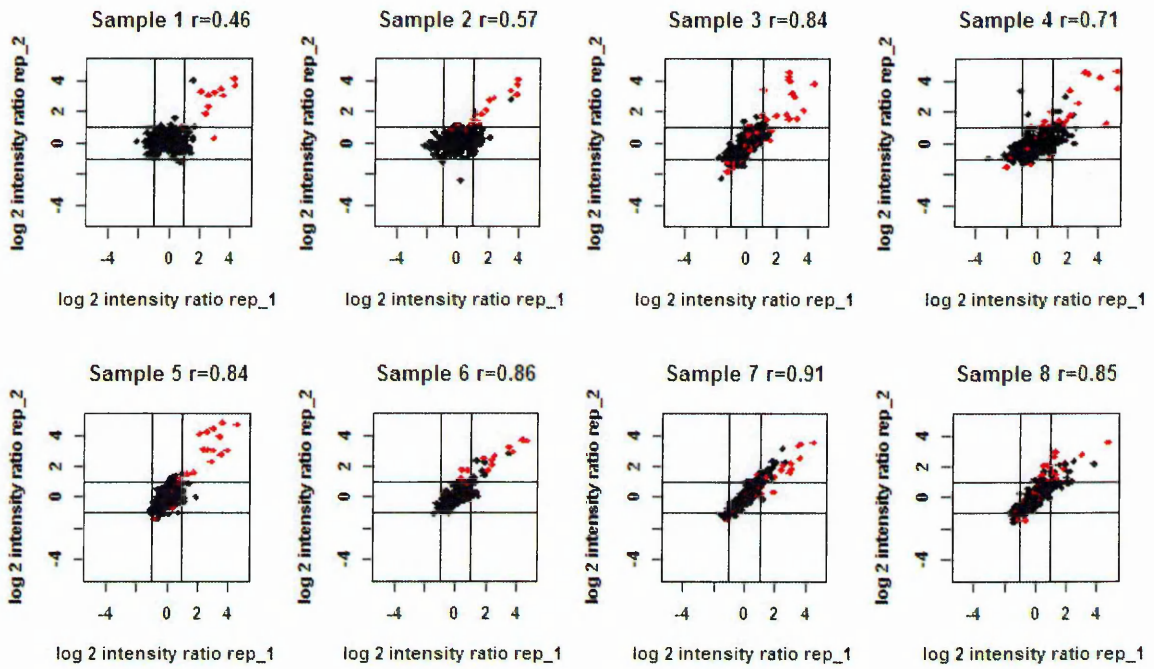


Figure 3.8. Reproducibility of microarray data and CNV detection
A) Plots showing the correlation between technical duplicates of 8 samples of $\log_2$ ratios for only the microarray probes within the total CNVs detected (red and black points) (352 out of approximately 8800 probes in the array). The correlation coefficient is indicated by the value 'r'. The red points are the $\log_2$ ratios of the probes within the CNVs identified in each individual sample and the black points are the $\log_2$ ratios of non-CNVs microarray probes.

3.3.5 Systematic effects on CNV detection probabilities

The only strong effect on the prevalence of CNV was population (p value < 0.001 by ANOVA). Specifically, across all CNVs, the prevalence of CNVs was lower in the Kilifi populations (1 and 2) than in Kisumu and Sudan (populations 3 and 4 respectively) (Figure 3.9). The lack of significance of the other factors, i.e., MOI, parasitaemia, haemoglobin and patient's age, on prevalence of CNVs (p value >0.05 by ANOVA) is reassuring because only the factor of interest, i.e., population, had significant effect on CNV detection probabilities. The presence of multiple genotypes (MOI) in an infection is a common phenomenon in *P. falciparum* infections and tends to vary with age (Ntoumi et al. 1995) and malaria transmission intensity (Konate et al. 1999). MOI may interfere with the average intensity in a given probe if a CNV is present in the sample and thereby mask its detection. The number of clones detected in the isolates ranged from 1 to 7 with the distribution of MOI varying in the four populations. The analysis here showed that MOI did not have an effect on the detectability of CNVs (p value > 0.05 by ANOVA).

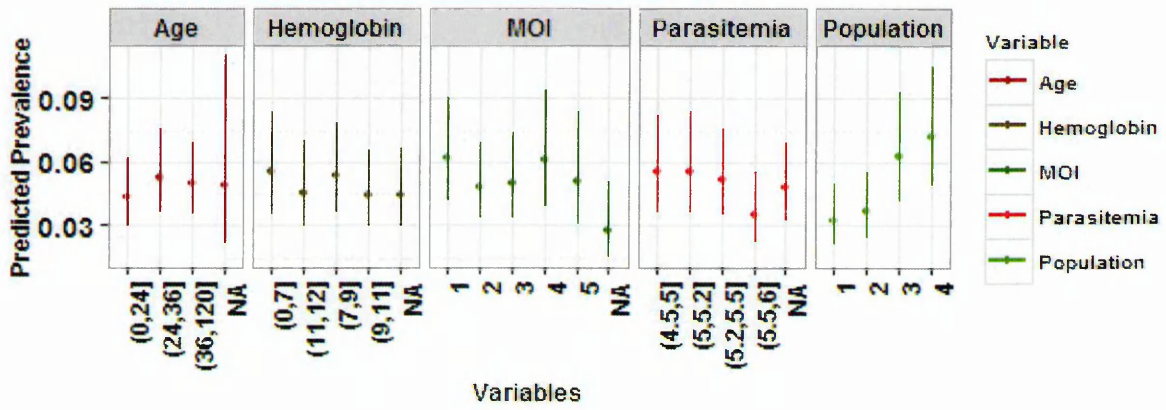


Figure 3.9. Systematic effects on CNV prevalence.
Least-squares means are indicated by points, with upper and lower 95% confidence limits represented by whiskers. Population 1 is Kilifi pre-malaria decline, population 2 is Kilifi Post-malaria decline, population 3 is Kisumu and population 4 is Sudan.

3.3.6 CNVs detected in field isolates

3.3.6.1 General characteristics of CNVs

A total of 98 different CNVs with minor allele frequency (MAF) greater than 2.2% (i.e., occurring in 4 or more isolates) were detected among the 183 field isolates studied. 34 of these were deletions and 64 were amplifications in gene copy number. These CNVs were distributed throughout the 14 chromosomes of the genome (Figure 3.10A). The number of CNVs per sample ranged from 1 to 20 with an average at 8 CNVs per isolate (Figure 3.10B). The majority of CNVs contained less than 3 genes within each, with the largest CNV on chromosome 9 consisting of approximately 20 genes (Figure 3.10C). The estimated CNV lengths were between 400bp and 90kb (Figure 3.10D).

3.3.7 Population genetics of CNVs

Most CNVs were observed at low frequencies in the population studied with majority of the CNVs occurring in less than 10% of the isolates (Figure 3.11A). To assess levels of population differentiation of CNVs frequencies, Weir and Cockerham F-statistics (F_{ST}) for each CNV were calculated for pairs of populations using the *hierfstat* package in R (Goudet 2005). The average estimates of F_{ST} are a sign of the background levels of population differences that are expected to arise through neutral processes including drift and population isolation and affect all loci in the genome to a similar manner (Anderson et al. 2005, Beaumont and Balding 2004, Akey et al. 2002). The average F_{ST} estimates across all CNVs ranged between 0.02 and 0.11 across the 6 pairwise population comparisons (Figure 3.11B). The greatest difference was exhibited by the comparison of Kilifi Pre-malaria decline and Sudan followed by Kilifi Pre-malaria decline and Kisumu (Figure 3.11B). The Kisumu vs. Sudan comparison showed the least average F_{ST}.

Outliers from these average F_{ST} estimates are CNVs that show unexpectedly low or high population differentiation thus potentially indicating that they are maintained by balancing and directional selection forces, respectively, that are locus specific (Figure 3.11B). These differences in CNV frequencies between populations may have arisen as a consequence of

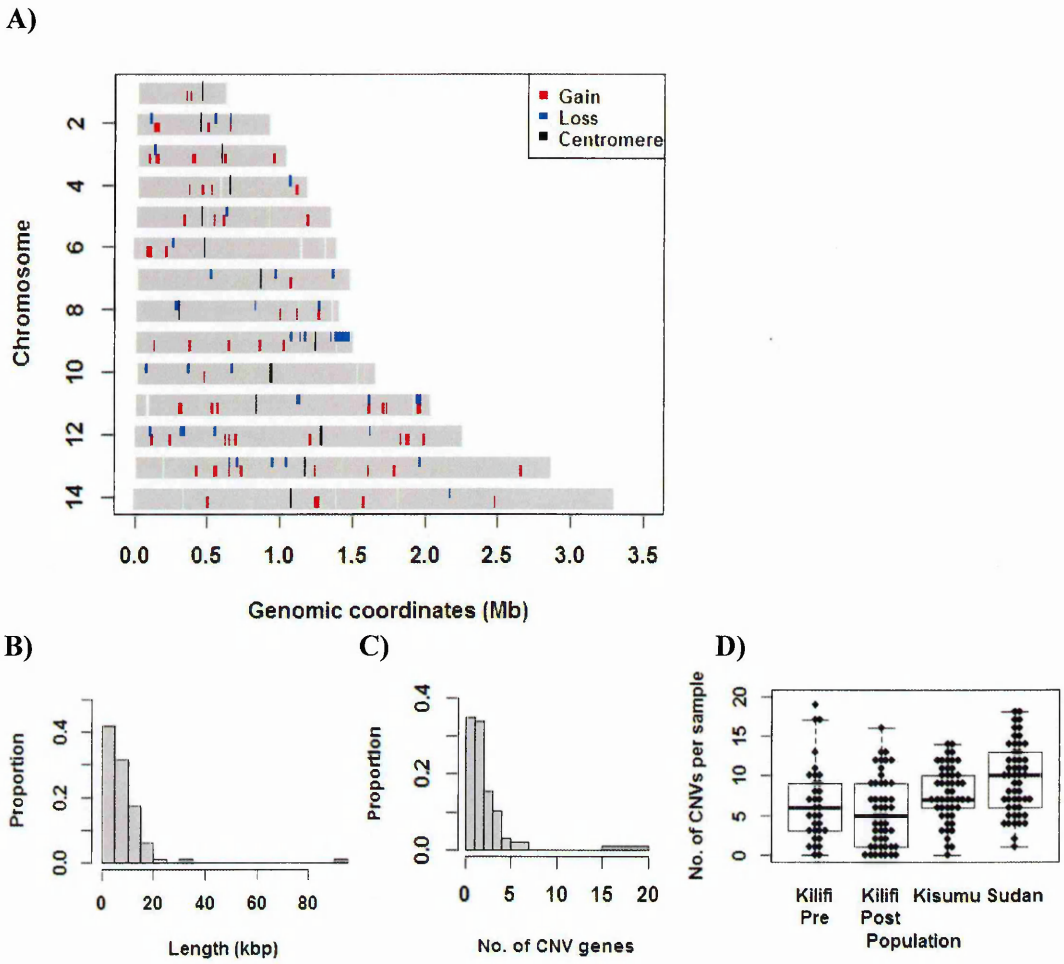
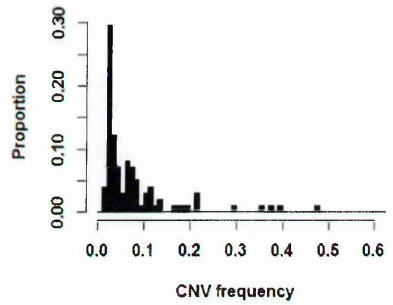


Figure 3.10. General properties of CNVs detected using microarrays.
A) A chromosome map showing the location of the CNVs in the 14 chromosomes of *P. falciparum* genome (deletions in blue and amplifications in red (MAF>2%). White vertical bars represent regions that are not targeted by the microarray probes. Black vertical bars are locations of centromeres. The distribution of the **B)** length of the CNVs (in kilobase pairs) **C)** number of genes per CNV and **D)** number of CNVs detected in each of the 183 samples in the four populations studied. Within each box, the horizontal line represents the median number of CNVs per sample, top and bottom boundaries show the 75th percentile and 25th percentile values respectively. The end of the whiskers show the minimum and maximum number of CNVs per sample and the data points above and below the end of the whiskers show the outliers.

local adaptation to population-specific selection pressures. The high differentiation may also be as a result of population bottlenecks, e.g., change in antimalarial use, vector populations among others and migration that leads to introduction on new variants to a population. These CNVs were classified into three groups, discussed below, based on the populations affected.

A)



B)

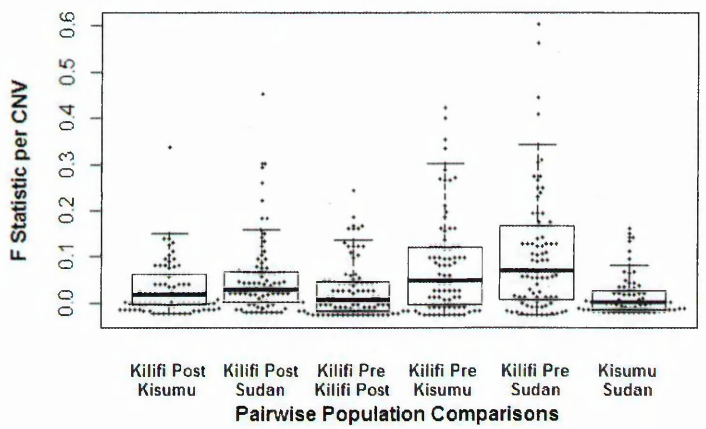


Figure 3.11. F_{ST} estimates obtained from pairwise population comparisons
A) The distribution of CNV frequencies in the study population B) Boxplot showing the distribution of F_{ST} estimates of each CNV obtained from six pairwise population comparisons, i.e., Kilifi Post and Kisumu, Kilifi Post and Sudan, Kilifi Post and Kilifi Pre, Kilifi Pre and Kisumu, Kilifi Pre and Sudan and Kisumu and Sudan. The horizontal line within each box represents the median F_{ST} estimates. The top and bottom boundaries of the box shows the 75^{th} percentile and 25^{th} percentile respectively. The end of the whiskers show the extreme values (minimum and maximum values) and the data points above and below the end of the whiskers show the outliers.

In the first category are four CNVs with the greatest population differentiation observed, i.e., top 7% in F_{ST} estimates, in the comparison between the two Kilifi populations and Sudan and between Kisumu and Kilifi Pre-malaria decline populations. These CNVs occur at lower frequencies in the Kilifi populations than in Sudan and Kisumu (Figure 3.12). These 4 CNVs show great variance in CNV frequencies between populations and thus are likely to indicate differential directional selection between populations or population bottlenecks. One weakness of the evidence of high differentiation is the small sample size of Kilifi Pre compared to the other populations that may affect the chance of CNV detection. They include three amplifications on chromosome 5, 12, and 9 and a deletion on chromosome 13. The frequencies of each of these CNVs were found to statistically significantly different in at least one of the populations at 95% confidence level using binomial test (prop.test function in R). The genes located within these CNVs and their functions are indicated in table 3.3.

Table 3.3. Genes and gene functions of CNVs that exhibit the greatest population differentiation

CNV identifier	No. of genes in CNV	Gene ID	Putative function
cnv5_101	3	PF3D7_0507900	Unknown
		PF3D7_0508000	Exposed to host immunity (Sanders et al. 2005)
		PF3D7_0508100	gene expression regulation (Cui et al. 2008)
cnv12_413	1	PF3D7_1248600	Unknown
cnv13_478	3	PF3D7_1348800, PF3D7_1348900, PF3D7_1349000	Unknown
cnv9_242	2	PF3D7_0908000	DNA repair (Nishino and Morikawa 2002)
		PF3D7_0908100	Unknown

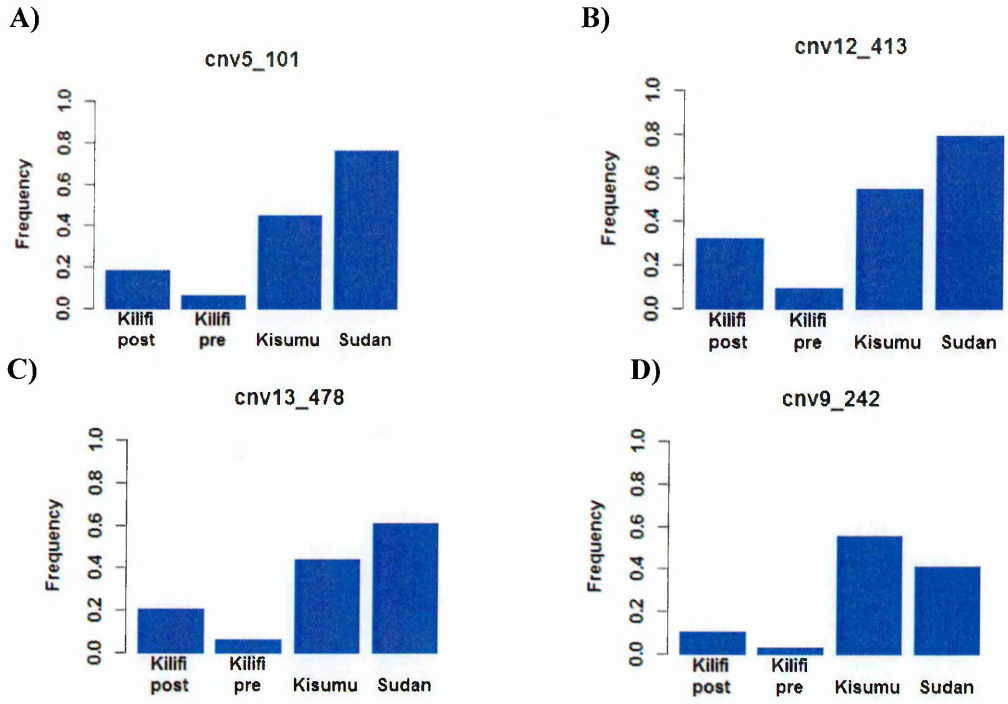


Figure 3.12. Population frequencies of CNVs with top 4 highest F_{ST} values
Barplot showing CNV frequencies in the four populations of 4 CNVs A) cnv5_101 on chromosome 5 consisting of PF3D7_0507900 - PF3D7_0508100 b) cnv12_413 consisting of PF3D7_1248600 on chromosome 12 c) cnv13_478 on chromosome 13 containing PF3D7_1348800- PF3D7_1348900and d) cnv9_242 on chromosome 9 consisting of PF3D7_0908000 and PF3D7_0908100PF3D7_0908100.

A second set of CNVs appear to be under positive selection in the Kilifi populations (F_{ST} >0.20) as compared to Kisumu or Sudan. They comprise deletions on chromosome 6 and 11 and amplifications on chromosome 3, 9 and 12 (Figure 3.13). The frequencies of each of these CNVs were found to be statistically significantly different in at least one of the populations. The genes found in these regions are shown in table 3.4.

A third set are CNVs that appear to be under purifying selection in one of the four populations (Figure 3.14). Seven CNVs were not detected in the Kilifi pre-malaria population but found to exist in the other three populations (Figure 3.14). The CNV frequencies were found to be

Table 3.4. Genes and gene functions of CNVs that appear to be under positive selection in Kilifi populations

CNV identifier	No. of genes in CNV	Gene ID	Putative function
cnv9_254	3	PF3D7_0925400	Gametocytogenesis (Li and Baker 1997)
		PF3D7_0925500	protection from reactive oxygen species (Nickel et al. 2006, Kehr et al. 2010)
		PF3D7_0925700	Transcriptional regulation (Cui and Miao 2010)
cnv3_051	1	PF3D7_0315200	ookinete motility and infectivity (Dessens et al. 1999, Templeton, Kaslow and Fidock 2000)
cnv12_388	3	PF3D7_1229400	modulation of host immunity (Augustijn et al. 2007)
		PF3D7_1229500	Protein trafficking (Mbengue et al. 2015)
		PF3D7_1229300	Unknown
cnv6_129	4	PF3D7_0606200	Protein degradation
		PF3D7_0606500	RNA splicing
		PF3D7_0606300	Unknown
		PF3D7_0606400	Unknown
cnv11_354		PF3D7_1148700	Exported to erythrocyte (Sargeant et al. 2006)
		PF3D7_1148800	Exported to erythrocyte (Sargeant et al. 2006)

statistically significantly different in at least one of the populations at 95% confidence level. The genes that were located in these CNVs and their functions are listed in table 3.5. The amplification of PF3D7_1149000 was at a lower frequency in Kisumu (high transmission) compared to the low transmission populations (Kilifi post-malaria decline and Sudan) and also in Kilifi pre-malaria decline compared to the low transmission period (Kilifi Post-malaria decline). A large deletion on chromosome 9, previously observed in laboratory adapted lines (Mackinnon et al. 2009, Cheeseman et al. 2009, Kidgell et al. 2006, Ribacke et al. 2007) and linked to loss of cytoadherence (Kemp et al. 1992, Trenholme et al. 2000) and gametocyte production (Day et al. 1993) was observed in 24 out of 183 isolates from all the populations

except the Kilifi pre-malaria decline population. A CNV on chromosome 7 (PF3D7_0710100 and PF3D7_0710200PF3D7_0710200) was found to be absent in Sudan only.

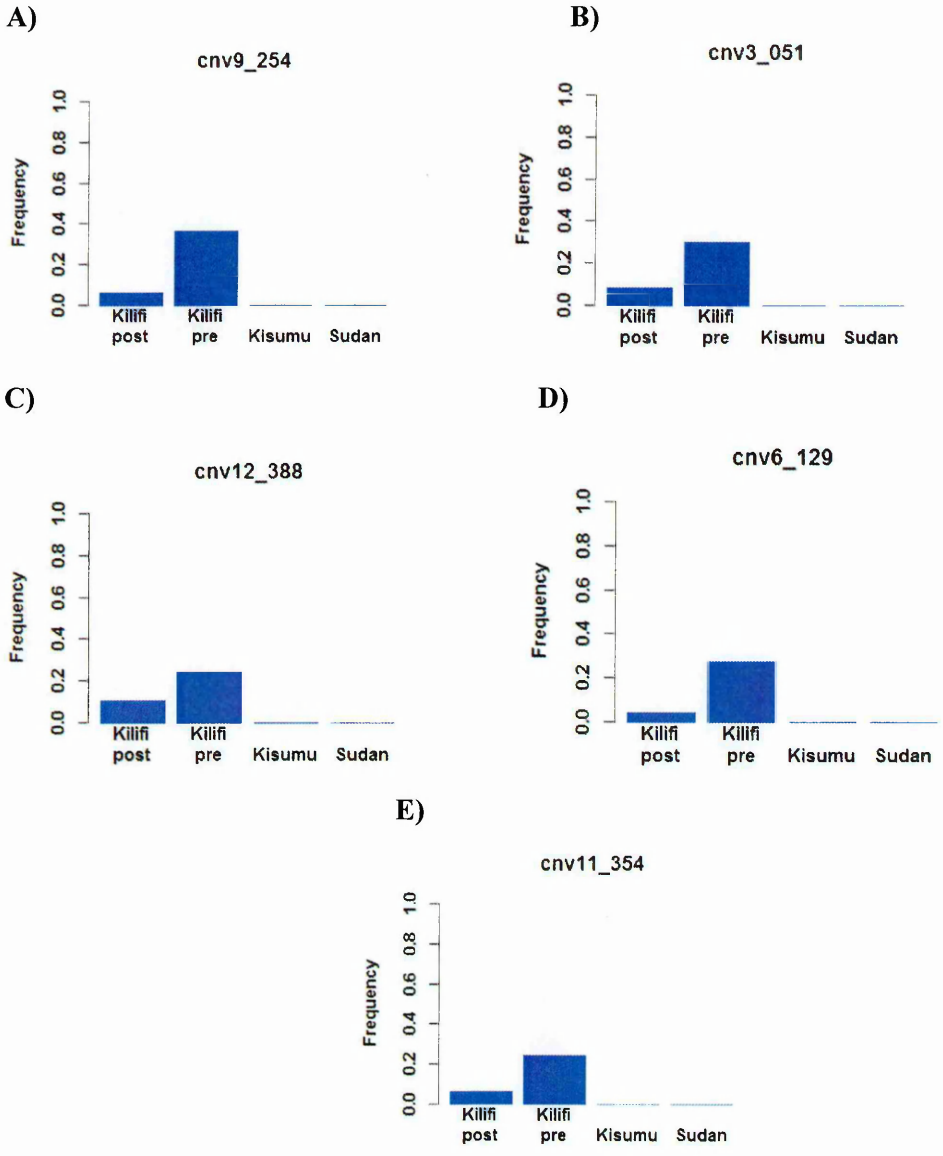
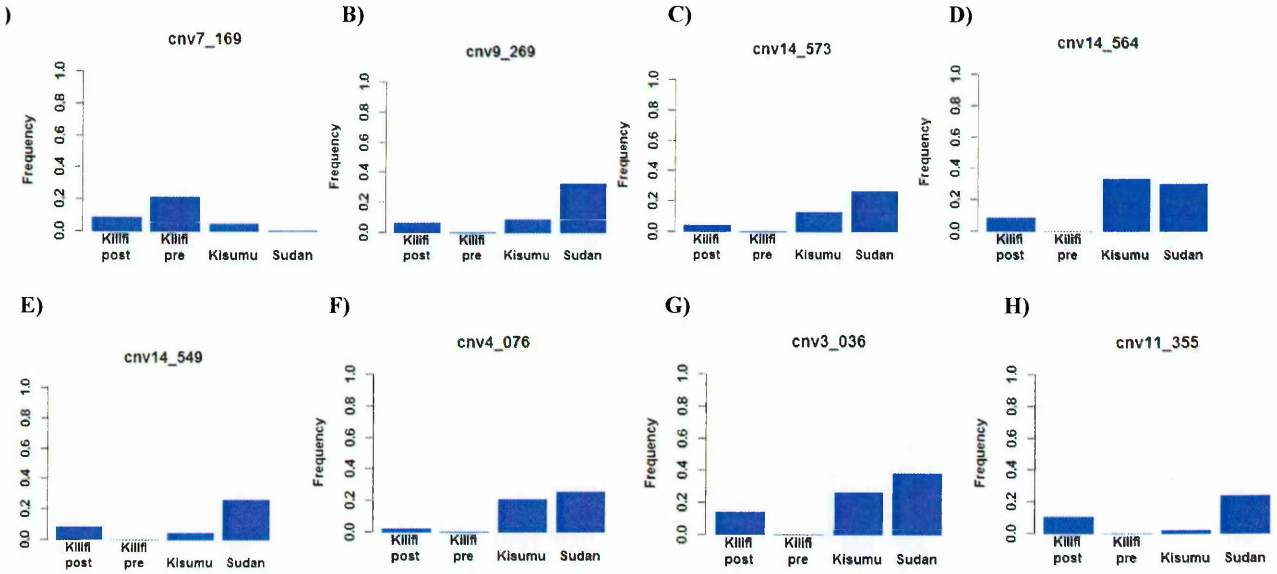


Figure 3.13: Population frequencies of CNVs showing signs of selection in Kilifi population
Barplots showing frequencies of CNV A) cnv9_254 on chromosome 9 containing PF3D7_0925400PF3D7_0925400 - PF3D7_0925700PF3D7_0925700 B) cnv3_051 on chromosome 3 containing PF3D7_0315200 C) cnv12_388 on chromosome 12 containing PF3D7_1229300 - PF3D7_1229500 D) cnv6_129 on chromosome 6 containing PF3D7_0606200 - PF3D7_0606500 and E) cnv11_354 on chromosome 11 containing PF3D7_1148700- PF3D7_1148900.

Table 3.5. Genes and gene functions of CNVs that appear to be under purifying selection in one of the four populations

CNV identifier	No. of genes in CNV	Gene ID	Putative function
cnv7_169	2	PF3D7_0710100, PF3D7_0710200	Unknown
cnv9_269	18	PF3D7_0935400	Gametocytogenesis(Eksi et al. 2012)
		PF3D7_0935500	Gametocyte stage (Silvestrini et al. 2010)
		PF3D7_0935600	Gametocytogenesis (Gardiner et al. 2005)
		PF3D7_0935700	Unknown
		PF3D7_0935800	Cytoadherence (Trenholme et al. 2000)
		PF3D7_0935900	Protein trafficking (Dixon et al. 2011)
		PF3D7_0936000	Exported protein (Spielmann et al. 2006b)
		PF3D7_0936100	Located at host-parasite interface (Spielmann et al. 2006a)
		PF3D7_0936400	Exported protein (Spielmann et al. 2006b)
		PF3D7_0936500	Host cell adhesion (Nacer et al. 2015)
		PF3D7_0936200, PF3D7_0936600, PF3D7_0936800, PF3D7_0936900, PF3D7_0937000, PF3D7_0937100	PHIST gene family (Sargeant et al. 2006)
		PF3D7_0936700, PF3D7_0937200	lysophospholipase
cnv14_573	2	PF3D7_1460700	A ribosomal subunit
		PF3D7_1460800	Unknown
cnv14_564	2	PF3D7_1452700, PF3D7_1452800	Unknown
cnv14_549	3	PF3D7_1438800	Unknown
		PF3D7_1438900	Antioxidative activity (Nickel et al. 2006)
		PF3D7_1439000	Copper transporter (Choveaux, Przyborski and Goldring 2012)
cnv4_076	3	PF3D7_0409500, PF3D7_0409700	Unknown
		PF3D7_0409600	DNA replication (Voss et al. 2002)
cnv3_036	3	PF3D7_0301600, PF3D7_0301700, PF3D7_0301800	Exported proteins (Sargeant et al. 2006)
cnv11_355	1	PF3D7_1149000	iRBCs remodelling (Glenister et al. 2009)



gure 3.14. Frequencies of CNVs showing signs of purifying selection in single populations

ırplots of frequencies of CNV A) cnv7_169 containing PF3D7_0710100- PF3D7_0710200 B) cnv9_269 containing PF3D7_0935400 - '3D7_0937200 C) cnv14_573 containing PF3D7_1460700- PF3D7_1460800 D) cnv14_564 containing PF3D7_1452700- PF3D7_1452800 E) v14_549 containing PF3D7_1438800- PF3D7_1439000 F) cnv4_076 containing PF3D7_0409500 - PF3D7_0409700 G) cnv3_036 containing '3D7_0301600 - PF3D7_0301800 and H) cnv11_355 containing PF3D7_1149000.

3.3.8 Comparison with published literature

There was considerable overlap of CNVs identified in our study with already published CNVs from genome-wide scans of *P. falciparum*. Out of the 225 genes located within the 98 CNVs detected, 20% of these genes (50 genes in 25 CNVs) have been previously identified to be copy number variable (Figure 3.10 and Appendix 3.1). These include a CNV on chromosome 11 containing an amplification of three adjacent genes (PF3D7_1148700PF3D7_1148700-PF3D7_1149000) identified in four other studies (Jiang et al. 2008b, Sepulveda et al. 2013, Samarakoon et al. 2011a, Cheeseman et al. 2009). and PF3D7_0423500, identified herein have also been previously identified (Samarakoon et al. 2011a). These genes encode three proteins thought to be exported to the erythrocyte surface including a gene belonging to the PHISTc gene family. A deletion of a second PHISTc gene (PF3D7_0202100) on chromosome 2 was also detected in three other studies (Cheeseman et al. 2009, Mackinnon et al. 2009, Carret et al. 2005). Two genes within a CNV, PF3D7_0423400 and are thought to play a role in erythrocyte invasion and anchoring of the cytoskeleton to the inner membrane (Bullen et al. 2009) respectively. An additional copy number variable gene PF3D7_0424400 also known as SURFIN 4.2 has been reported (Samarakoon et al. 2011a, Jiang et al. 2008b, Kidgell et al. 2006). The gene product was found to be exposed at the surface of an infected erythrocyte and merozoite hence is thought to be under selection by host immunity or involved in merozoite invasion (Winter et al. 2005, Chan, Fowkes and Beeson 2014).

Evidence of deletion of one or two members of the cytoadherence linked asexual gene i.e. *clag 3.1* and *clag 3.2* (Robinson et al. 2011) has been previously reported. In this study a deletion of *clag 3.1* was detected. A *clag 3* hybrid, made up of the *clag 3.2* sequence at 5' UTR and the

clag 3.1 sequence at the 3'UTR, has also been observed (Iriko et al. 2008). The *clag* gene family is thought to be involved in erythrocyte invasion (Kaneko et al. 2005) and cytoadherence (Trenholme et al. 2000) of the infected erythrocyte. Recently, *clag 3.1* has been implicated in transport of solutes across the infected erythrocyte membrane (Nguitragool et al. 2011).

Interestingly, a large deletion was observed on the right arm of chromosome 9 in field isolates, a phenomenon that has previously been observed only in laboratory adapted isolates (Mackinnon et al. 2009, Shirley et al. 1990, Kemp et al. 1992, Cheeseman et al. 2009, Ribacke et al. 2007, Kidgell et al. 2006) and thought to arise due to reduced selection pressures, e.g. host immunity, antimalarial drugs and the need for sexual reproduction, *in vitro*.

3.3.8.1 Novel CNVs

Seventy-five potentially novel CNVs out of the 98 CNVs detected (Appendix 3.2) were found. These included a gene *rad51* (PF3D7_1107400) coding for protein involved in homologous recombination during DNA repair and recombination events that may lead to DNA rearrangements that enable antigenic variation (PF3D7_1107400) (Bhattacharyya et al. 2005). Interestingly, the *rad51* human homolog has been found to be amplified in cancer cells (Mathews et al. 2011). Two additional genes adjacent to *rad51*, identified to vary in copy number, are thought to play a role in regulation of stimulation of translation (PF3D7_1107300) (Ochoa, Llinas and Singh 2011, Martineau et al. 2008) and protein folding (PF3D7_1107500). A number of novel CNVs containing genes with unknown functions were also detected (Appendix 3.2).

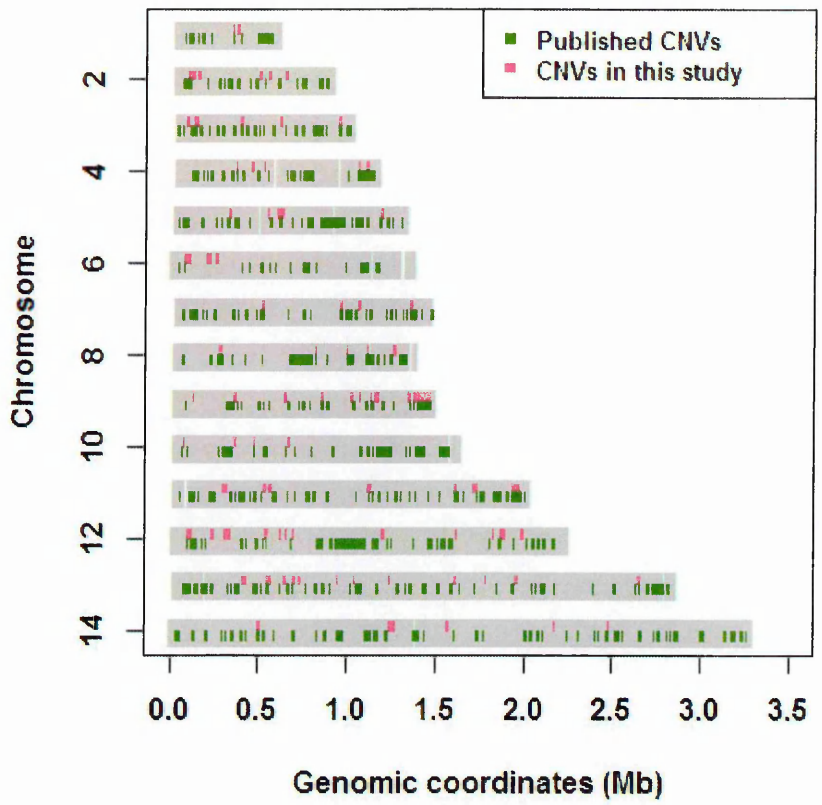


Figure 3.15. Overlap between CNVs identified in this study and published studies. Chromosome map showing the location of CNVs identified in this study (pink) and CNVs detected in earlier published work (green) in the 14 chromosomes of *P. falciparum* genome.

3.3.9 Functional gene set enrichment of CNVs

Enrichment of 15 out of 205 functional categories (obtained from the Malaria Parasite Metabolic Pathways website: http://mpmp.huji.ac.il/) of genes in the list of genes located within CNVs were found (Table 3.3). The groups that showed statistically significant enrichment (p-value < 0. 05), using the hypergeometric test, included genes coding for proteins that are exported to the surface of infected erythrocytes and may be exposed to selection pressure from the host immunity. These include PHISTs, Maurer cleft proteins and genes encoding proteins exported to the infected erythrocyte and classified as

Table 3.6. Functional gene categories showing significant enrichment

Gene set	No. of genes in gene set	No. of genes present in a CNV	P value
Maurer's clefts proteins	98	14	4.00E-05
Exported PHISTs	51	9	9.66E-05
Glycolysis	27	5	0.001266
ER to Golgi translocation and quality control	13	3	0.002385
Asparagine and Aspartate metabolism	8	2	0.004747
Rab proteins in intracellular traffic	16	3	0.005435
Pentose Phosphate Cycle	9	2	0.006877
Thioredoxin glutaredoxin and peroxiredoxin	18	3	0.008491
Cotranslational.cleavage of.N terminal. Methionine residues and N terminal acetylation	10	2	0.009489
Exported 'Immunoreactive' proteins (Crompton et al. 2010)	19	3	0.010368
Methionine and polyamine metabolism	20	3	0.012494
Tubulin and microtubules	20	3	0.012494
Utilization of phospholipids	42	5	0.012508
Anaphase promoting complex ubiquitin ligase	11	2	0.012603
Double strand break repair and homologous recombination	11	2	0.012603
Mitochondrial electron flow	21	3	0.01488
Mitochondrial antioxidant system	12	2	0.016232
Compartmentation of redox metabolism	24	3	0.023685
Histone chaperones	36	4	0.024691
Lipoic acid metabolism	6	1	0.028841
Skp1 Cullin.F box biquitin ligase	15	2	0.030278
Dolichol metabolism	7	1	0.039148
Ubiquinone metabolism	7	1	0.039148
Initiation of translation	41	4	0.040707
Biogenesis of cytochrome oxidase	17	2	0.042259

'immunoreactive' by Crompton and colleagues after probing 1,200 *P. falciparum* antigens for antibody reactivity using protein arrays (Crompton et al. 2010). This suggests that varying gene copy number is one of the different strategies the parasite uses to escape host immunity. Increased copy number would facilitate immune escape by increasing the chances of

mutations which could lead to increased antigenic diversity. Alternatively, some of these genes exposed to host immunity may be selected for sequence variability, and detected as CNVs but are not real CNVs because of the effect of sequence polymorphism on hybridisation. Significant enrichment was also observed for genes involved in key functional processes including glycolysis, DNA repair, transcriptional and translation regulation, intracellular trafficking and antioxidative processes.

3.4 Discussion

The results of this study confirm the high prevalence of CNVs in the *P. falciparum* genome and are consistent with earlier work (Mackinnon et al. 2009, Cheeseman et al. 2009, Kidgell et al. 2006, Ribacke et al. 2007, Mok et al. 2011, Dharia et al. 2009). On average 8 CNVs were observed per isolate and approximately 4.5 % of the genes in the genome were identified to be copy number variable. 20% of the CNV genes identified in this study have been previously reported in other studies. The results further highlight their high prevalence in natural population, something not previously studied. The CNVs found in nature appear to be distributed throughout the genome, contrary to some findings that CNVs tend to occur at the subtelomeric regions (Cheeseman et al. 2009, Jiang et al. 2008b). This discrepancy is most likely explained by exclusion from analysis in this study of probes targeting genes encoding variant surface antigens, mostly located in the subtelomeric regions and highly polymorphic, because of their reduced hybridization to the microarray and hence propensity to generate false CNV 'calls'.

The CNVs identified here are mostly small in size (average of 7.4 kilobase pairs and 2 genes) with the exception of a chromosome 9 deletion covering about 18 consecutive genes. This may indicate that large-sized CNVs could be at a selective disadvantage because of the cost

associated with the DNA replication of the extra copies and likely impact on epigenetics due to the large structural changes that they may cause. The amplifications outnumbered deletions at a ratio of 1.9:1. This deficit in deletions supports the idea that CNVs are adaptive since purifying selection would act to remove deletions which, if important to function, are expected to be deleterious.

The enrichment of key parasite functional gene groups within the CNVs, e.g., DNA replication, gametogenesis, cytoadherence, transport and DNA repair reflects the potential impact of CNVs on the survival of the parasite. Importantly, the observation of enrichment of CNV genes belonging to groups with environmental responsiveness functions, such as those that are exposed to host immunity and may be involved in immune escape and also those that respond to oxidative stress arising from glycolysis, could reflect an adaptive role of CNVs to the environment. None of the *P. falciparum* published CNV studies to date have investigated the functional enrichment of genes with CNVs. Further analysis of the functional impact of these CNVs is assessed in Chapter 5 by investigating the influence of CNVs on gene transcriptional levels.

Overall, low population frequencies of CNVs with most occurring in less than 10% of the 183 field isolates studied was observed. The abundance of low frequency CNVs may suggest that some variants may be favoured in a fraction of host or vector conditions, e.g., host red cell polymorphism and co-infection, offering short term adaptation that enabled the parasite to survive and transmit to the next host, after which they revert back to normal copy when the environmental pressure ceases to exist.

Applying population differentiation analysis by calculating F_{ST} estimates of CNV frequencies, CNVs that may be under selection pressure in the populations were identified. Among the CNVs that were found to be potentially adaptive (F_{ST} estimates of greater than 0.20) were genes that encode proteins involved in gene expression regulation (Cui et al. 2008), gametogenesis (Dessens et al. 1999, Templeton et al. 2000, Li and Baker 1997), protection from reactive oxygen species (Kehr et al. 2010, Nickel et al. 2006) and some exposed to host immunity (Sanders et al. 2005, Sargeant et al. 2006). The high F_{ST} observed between populations could be reflective of the differences in environmental selection between populations including vector populations, antimalarial drug used, ethnicity, host genetics and climatic conditions. The background levels of differentiation between populations, which may be due to neutral processes, were further established. Some of the CNVs are likely to exist in the population under no selective pressure. These variants may be present in the population until they become advantageous to the parasite when environmental conditions change. By having a flexible genome that allows for the random generation of genetic variation including CNVs, the parasite is able to adapt to current conditions and also even future changes in environmental conditions.

This study has some limitations. First, though high stringency was applied on calling CNVs by filtering out probes targeting highly polymorphic genes, probes with known SNPs within the probe sequences, poorly hybridizing probes and low frequency CNVs, the presence of additional sequence variation in field isolates, not able to be ruled out using these filters, is likely. The presence of sequence polymorphism is expected to cause false positives in CNV calls using microarray data. Second, the array could only assay genes present in the 3D7 parasite line since the array probes were designed against the complete 3D7 genome. Any gene that is absent in 3D7 could not be assayed by the array. Lastly, microarray data provides

low resolution of CNV breakpoints and hence the genes located at the start and the end of a CNV may be incorrectly called.

This study shows that CNVs contribute to genetic variation in the genome of the parasite in nature. It also provides evidence of the adaptive role of specific CNVs under current selection pressures in the populations. Population differences in specific CNV frequencies may be due to differences in environmental conditions including a change in the dominant vector species and feeding behaviour reported in Kilifi during the period of 1990-2010 (Mwangangi et al. 2013), differences in antimalarial drug used and differences in host immunity.

Chapter 4

Confirmation of CNVs detected using microarray by whole genome sequencing

4 Chapter 4: Confirmation of CNVs detected using microarray by whole genome sequencing

4.1 Introduction

Next generation sequencing has, in theory, the potential for high resolution identification and characterization of CNVs. It may also facilitate understanding of mechanisms of CNV formation through more accurate identification of CNV boundaries (Carvalho et al. 2009, Dittwald et al. 2013, Samarakoon et al. 2011b) and in the detection of translocation events to new positions in the genome (Ma et al. 2013). However, comparisons between CNVs detected using sequencing vs. microarrays have shown some discrepancies (Retterer et al. 2014, Sepulveda et al. 2013). Methodology for detecting CNVs in NGS data has not yet stabilised with many new methods for analysis being published in recent years (Zhao et al. 2013, Yoon et al. 2009, Xie and Tammi 2009, Miller et al. 2011, Abyzov et al. 2011, Magi et al. 2011, Boeva et al. 2012, Alkan, Coe and Eichler 2011). One of the strategies for detection of CNVs from sequencing data is based on the assumption that read depth in a genomic region correlates with genome copy number. However, certain biases including GC content and uniqueness (the number of times a sequence appears in the entire genome) of a region affect read depth thus making estimation of copy number from read-depth analysis problematic. Of particular relevance to *P. falciparum* are the long stretches of A and T nucleotides in its genome which pose a challenge to CNV identification for both the above reasons.

In *P. falciparum*, most studies of CNVs have used DNA microarrays to detect CNVs. Three recent studies, however, have used next-generation sequencing to detect CNVs in *P. falciparum* (Sepulveda et al. 2013, Robinson et al. 2011, Samarakoon et al. 2011b). Robinson and colleagues reported 7 CNV genes in 5 clinical samples using read depth and paired end analysis of Illumina sequence data. Using 7 publicly available whole genome sequence data,

i.e., 2 samples from Robinson and others and 5 laboratory lines, Sepulveda and others tested a CNV detection method they had developed against two other published tools. A third study identified 5 CNVs in 5 laboratory lines using read depth analysis on 454 sequence data (Samarakoon et al. 2011b) .

In Chapter 3, it was described that 98 different CNVs, containing a total of 225 genes, were identified in 183 field isolates using microarrays. In this chapter, whole genome sequencing of a subset of 22 of these isolates was performed with the aim of comparing the results of detection using sequencing vs. microarrays and confirming some of the CNVs. Analysis of sequence data involved mapping the sequence reads to a reference genome, calculating and normalising for read coverage, followed by identification of CNVs using the Copy Number estimation using the Mixture Of PoissonS (*cn.MOPS*) package in R (Klambauer et al. 2012).

4.2 Methods

4.2.1 Library preparation

Whole genome sequencing of 22 out of the 183 *P. falciparum* clinical isolates was performed using the Personal Genome Machine (PGM) Ion Torrent sequencing platform at the KEMRI-Wellcome Trust Research Programme laboratories in Kilifi, Kenya. These samples were selected based on the CNVs identified. All the 98 CNVs (identified in chapter 3) were represented in these samples sequenced. Ion Torrent 200bp library preparation involved enzymatic fragmentation of gDNA, ligation of adaptors to the fragments, selection of fragments of appropriate size, PCR amplification (5 cycles), clonal amplification of fragments (emulsion PCR) and finally performing the sequence run on an Ion Torrent 318 chip which has a capacity of up to 2Gbp (Chapter 2, section 2.2). To achieve successful runs, several quality control steps in the library preparation were undertaken including assessment of

fragment size and quantification of the proportion of templated ion spheres after emulsion PCR. Human DNA quantification of the samples was assessed prior to sequencing using quantitative real time PCR (qPCR). A summary of the experimental procedure is shown below (Figure 4.1).

4.2.2 Sequence data processing for CNV detection

4.2.2.1 Quality assessment and read mapping of sequence data

The sequence data were obtained from the Ion Torrent Suite software in FastQ format and the quality was assessed using *FastQC* software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The sequence reads were mapped to a reference genome from the 3D7 parasite line (Version 3) available at GeneDB (Logan-Klumpler et al. 2012). The mapping was performed using the *Burrows-Wheeler Aligner* (*BWA*) (Li and Durbin 2009) with the BWA-MEM option. This option is suitable for mapping reads longer than 70bp, and was implemented with the '–M' option for marking secondary read alignments (reads mapping to more than one location). Non-mapped reads, PCR duplicates and reads with mapping quality of below 30, i.e., non-uniquely mapped reads, were filtered out using *SAMtools* (Li et al. 2009). Filtering of reads mapping to more than one location would lead to exclusion from analysis of genes that are amplified in reference genome (3D7) but exist in different copy number state in the test genome..To minimize mapping errors that may result in bias in read coverage and hence accuracy of identification of CNVs, I performed analyses that targeted different sets of specific regions of the genome. First, I filtered out genes that are known to be highly variable including the variant surface antigens (VSAs) (*var, rifins* and *stevors*) (Gardner et al. 2002, Cheng et al. 1998) and low complexity noncoding regions, i.e., intergenic regions and introns: this targeted genome is referred to as the 'exome'. The second target genome was the 'probome', which consisted of the regions

targeted by the probes in the microarray used for the CGH experiment described in Chapter 3 (Bozdech et al. 2003b). Third, all the genes in the genome, including their introns weretargeted, filtering out the VSAs. This was referred to as 'genes'. Fourth, the whole

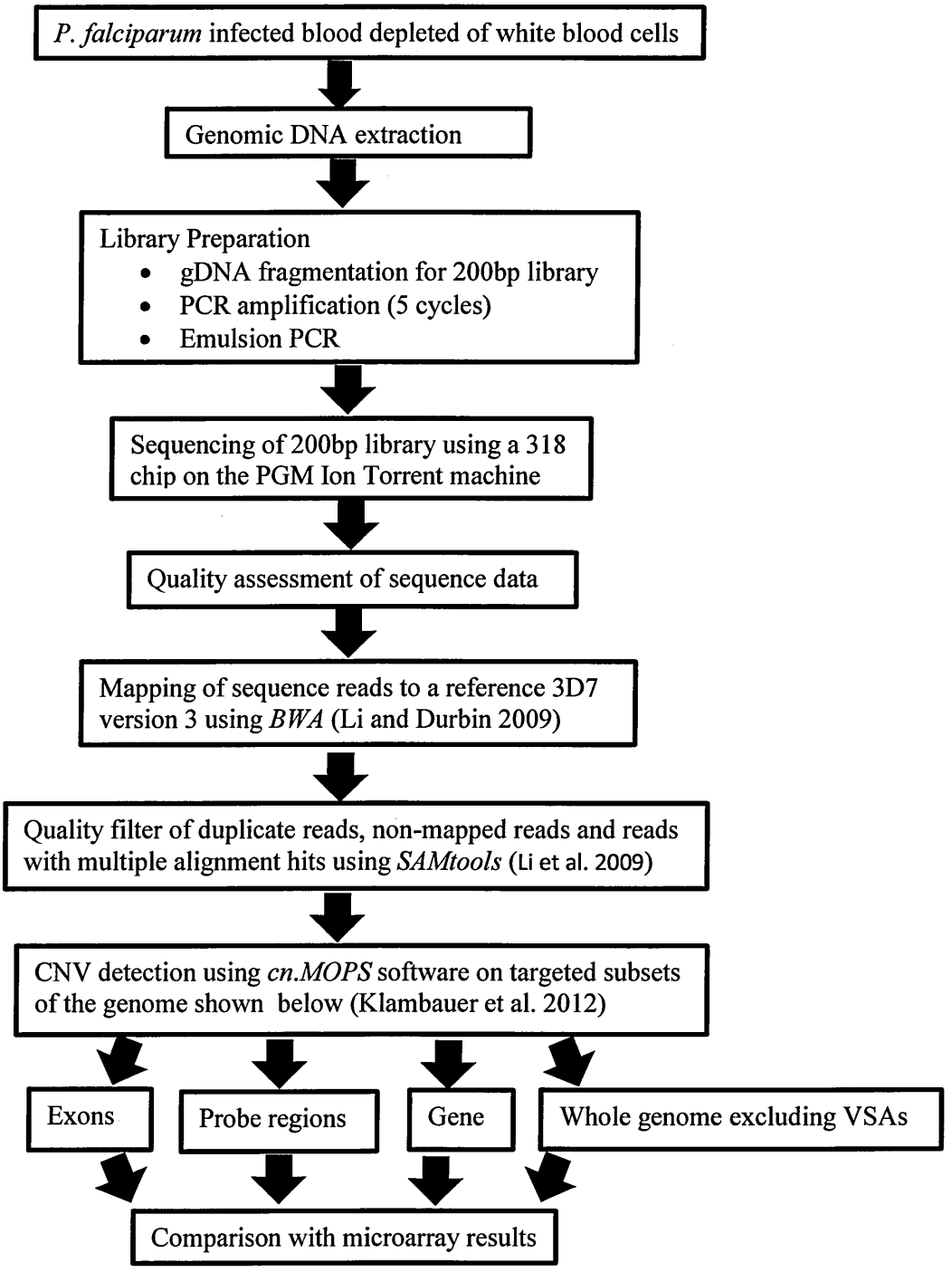


Figure 4.1. Overview of the process of CNV detection using whole genome sequencing

genome (genic and non-coding regions) was targeted with VSAs filtered out for comparison (the 'genome').

GC content and presence of repetitive sequence/non-unique regions have been previously reported to affect the read mapping in genomes known to be AT rich and also with repetitive regions (Quail et al. 2012, Ross et al. 2013). To examine these issues prior to the main analysis to detect CNVs, uniqueness and GC content were calculated using a function in the ReadDepth software (Miller et al. 2011). For this, the reference genome was split into 100bp sequences which were then mapped back to the reference genome. The number of times a sequence, if found in the genome, was calculated and then normalised to give a 'uniqueness' score of between 0 and 1 such that a score of 0 indicated a 100bp sequence that occurs more than four times in the genome and that of 1 indicated a sequence that appears only once. The GC content (the proportion of the G and C nucleotides) in the 100bp windows was also calculated using the same software.

4.2.2.2 Copy number estimation and segmentation

Read depth in the targeted regions was calculated in 100bp non-overlapping windows and normalised across the samples using the 'quantile' method in *cn.MOPS* package in R (Klambauer et al. 2012). The method adjusts the distribution of read depth in the samples such that they are similar. To identify CNVs, *cn.MOPS* applies a mixture of Poisson models at each window across the samples. The model assumes that the read coverage for a specific copy number across samples has a Poisson distribution and, in the case of there being different copy numbers among the samples, the model assumes a mixture of Poissons with different means. In this analysis, a copy number difference was considered by *cn.MOPS* to be present when the normalised read depth for a sample in a 100bp sequence window showed a fold-difference

from the mean read depth of all the samples for that window corresponding to a $\log_2$ value of greater than 0.8 (> 1.7-fold, i.e., a gain) or less than -0.8 (<0.6-fold, i.e., a loss). Thus the outputs of *cn.MOPS* were the segments identified to be copy number variable in each of the samples, their estimated copy number, their fold difference and their location in the genome. For subsequent analyses, the cut-off of defining a CNV was set at a minimum of 50% of the gene length called as CNV in 'gene', 'genome' and 'exome' methods and a cut-off of a minimum of 50% of the probes identified to be copy number variable in 'CGH (GADA)', 'CGH (cn.MOPS)' and 'probome' methods.

To ensure a fair comparison between CNVs identified in sequence and CGH data, *cn.MOPS* was also applied to $\log_2$ intensity ratio data from microarrays. The normalized $\log_2$ ratios were transformed to a scale similar to read depth by calculating the base 2 antilog of the $\log_2$ ratio and multiplying the result by 30 (average sequence coverage) (designated 'CGH (cn.MOPS)'). Finally, for comparison with the results from Chapter 3, the CGH data were reanalysed for CNVs using *GADA*, as in Chapter 3, except that the $\log_2$ ratio cut-off of 0.8/-0.8 was used instead of 1/-1 so as to be comparable to the analyses here using *cn.MOPS*. This analysis was designated 'CGH (GADA)'.

To evaluate the consistency between CNV calling using sequence and CGH data, for each CNV genes, the positive and negative predictive value (PPV and NPV) of each of the methods was calculated using CGH (GADA) as the 'gold standard'. PPV gives the proportion of CNVs detected in a method that are also detected in CGH (GADA). The NPV gives the proportion of CNVs not identified in a method that are not detected in CGH (GADA). The values of the two measures range between 0 and 1. A PPV of 1 indicates that all samples identified to have a specific CNV gene in one method were also identified to have the same CNV gene in CGH

(GADA) and 0 indicates that none of the samples identified to have a specific CNV gene in one method had CNVs detected in CGH (GADA). A NPV of 1 indicates that all samples without a CNV in one method did not have the CNV detected in CGH (GADA) and 0 indicates that all the samples without a CNV gene in one method had the CNV detected in CGH (GADA). PPV and NPV were used instead of sensitivity and specificity because they take into account the prevalence of the CNVs.

4.3 Results

4.3.1 Sequence quality

The majority of sequenced bases were of good quality with Phred quality scores of greater than 20, i.e., a 1 in 100 probability that a base is incorrectly called. The mean read length ranged between 160bp and 235bp in the 22 samples. The average genome coverage differed in the samples sequenced. The genome coverage ranged between 7.8 - 47.7 in the 18 samples analysed and from 1.6 - 5.1 in the four samples excluded from analysis (Table 4.1).
The quality of the bases decreased with increasing sequence read length above 200bp (Figure 4.2A), a characteristic commonly observed with most sequencing platforms. This phenomenon is thought to be due to some templates on the bead lagging behind or ahead in sequence of other templates on the bead as a result of lack of complete extension or leftover nucleotides in the well (Bragg et al. 2013, Salipante et al. 2014, Margulies et al. 2005). The average nucleotide content observed was 20% GC which is similar to the expected value in *P. falciparum* based on Sanger sequencing (Gardner et al. 2002) (Figure 4.2B) and the proportion of the nucleotide content varied greatly after the 250bp position. Four samples with the lowest average sequence coverage, due to presence of high proportion of human DNA sequences relative to parasite sequences, of up to 95% (Table 2.3 and 2.4), were excluded from further analysis. A summary of the sequence output for each of the samples is shown in Table 4.1

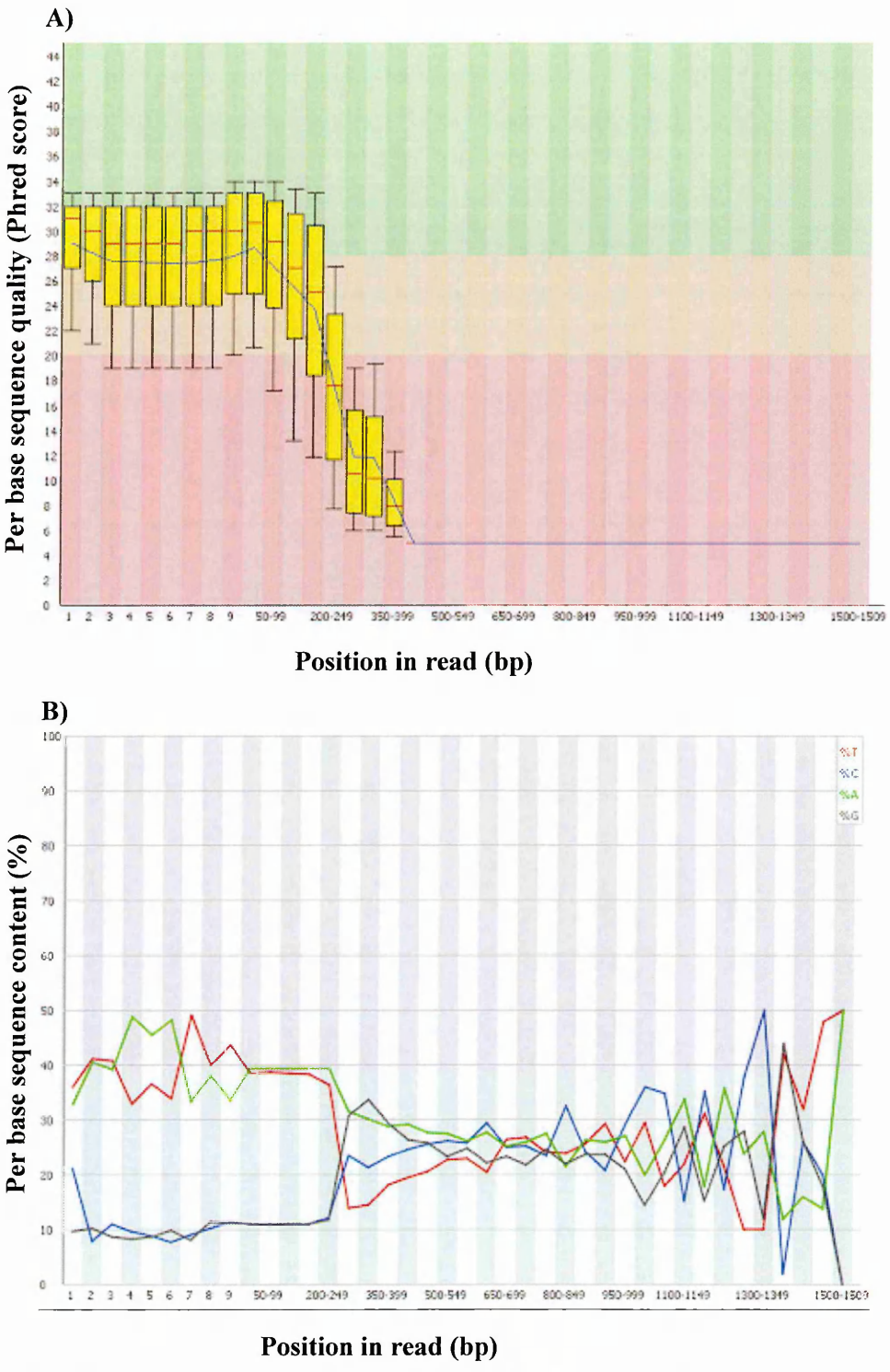


Figure 4.2. Plots of quality scores and nucleotide content of sequenced reads
A) Boxplots of quality scores (y-axis) by position in read (x-axis) in a samples. Yellows boxes show the inter-quartile range, 25th -75th, the red line-median, upper and lower whiskers are the 10th and 90th values and the blue line-mean. The poor quality bases are in red, good quality bases are in green and moderate quality in orange background. B) Mean proportion of each nucleotide (y-axis) by position in sequence reads (x-axis) in a sample.

able 4.1 Summary statistics of whole genome sequence data of 22 *P. falciparum* field isolates

Sample Identifier	Population of origin	Number of reads	Mean read length (bp)	Number of reads mapped to 3D7	Final read count after filter	Percent of genome with coverage values 0, >0-<=5, >5-<=10, >10-=<20 and >20				
						=0	=<5 >0	>5 <=10	>10 =<20	>20
pf1299	Kilifi Pre	5,814,179	192	3,343,322	2,634,034	8.4	11.7	10.1	17.4	52.4
pf1624	Kilifi Pre	4,768,102	181	2,584,197	1,903,046	12.0	15.3	11.7	24.0	37.0
pf1349	Kilifi Pre	6,117,359	200	2,752,578	2,172,391	10.0	13.2	11.1	20.7	45.0
pf10676	Kilifi Post	6,516,576	195	2,726,767	2,158,002	10.8	16.0	12.3	21.8	38.1
pf10724	Kilifi Post	7,212,289	235	1,579,785	1,199,707	10.8	19.8	17.2	37.2	14.9
pf10836	Kilifi Post	4,221,534	204	2,642,321	1,882,187	9.0	15.6	12.0	19.4	44.0
pf10760	Kilifi Post	4,469,713	206	1,192,910	984,949	14.2	27.0	25.5	31.0	2.3
pf1212	Kilifi Pre	6,151,974	188	5,715,908	4,868,604	9.4	12.4	7.9	11.0	59.3
pf10770	Kilifi Post	7,325,789	190	4,558,838	3,776,892	10.9	15.2	9.4	12.0	52.5
pf1590	Kilifi Pre	6,966,753	205	1,239,892	972,995	17.2	27.3	19.8	31.5	4.2
pf1895	Kilifi Pre	7,282,318	178	7,112,587	6,151,059	8.3	5.9	4.9	8.9	71.9
pfK007	Kisumu	6,673,824	183	6,021,604	5,068,212	7.6	11.6	8.3	12.0	60.5
pfK020	Kisumu	6,161,420	160	5,761,970	4,647,383	8.5	12.0	9.8	15.7	54.0
pfK071	Kisumu	6,661,003	189	2,882,086	2,365,500	13.7	22.1	10.3	12.4	41.5
pfM004	Sudan	6,674,924	204	4,651,427	3,953,732	8.8	6.8	6.0	11.7	66.7
pfM007	Sudan	6,718,825	192	6,430,355	5,428,351	9.9	7.9	5.2	8.8	6.8
pf10814	Kilifi Post	6,620,382	190	883,916	672,578	24.5	36.5	22.4	15.6	1.0
pfK065	Kisumu	6,699,194	200	332,760	209,049	35.3	61.5	3.2	0	0
pfG013	Sudan	6,716,879	162	366,073	237,561	3.5	61.5	3.3	0	0
pf10820	Kilifi Post	4,593,168	199	4,267,905	3,501,088	21.7	15.0	6.7	9.2	47.4
pf10578	Kilifi Post	5,690,616	233	496,494	375,707	55.5	36.0	5.1	2.0	1.5
pf10495	Kilifi Post	4,369,272	225	5,387.450	4,184,087	22.9	14.4	6.4	8.6	47.6

4.3.2 Challenges in mapping of whole genome sequence data of *P. falciparum*

As in previous studies (Quail et al. 2012, Ross et al. 2013), GC content in a region was found to positively affect read coverage at GC content below 50% (Figure 4.3A). This effect has been linked to the PCR amplification step during sequencing (Quail et al. 2012): in this study, 5 rounds of amplification per sample were performed. Low sequence complexity also affects the ability of reads to unambiguously map to the reference thus biasing read coverage estimates in these regions (Treangen and Salzberg 2012). In this study, low or no coverage in non-unique regions was observed since the reads that aligned to multiple locations on the genome were filtered out (Figure 4.3B).

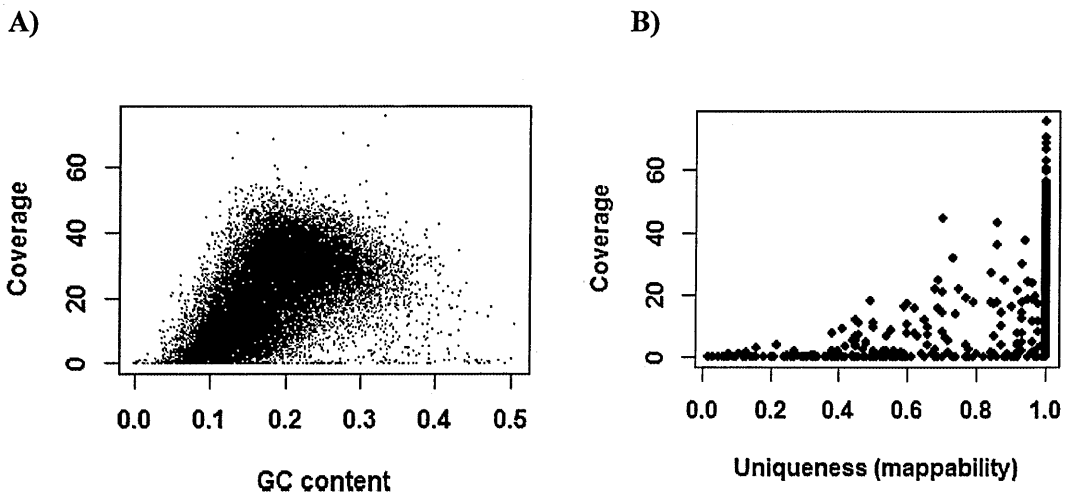


Figure 4.3. Effect of GC content and uniqueness of a region on sequence coverage
The relationships between A) GC content and coverage, and B) uniqueness and coverage calculated in 100bp non-overlapping windows. A uniqueness value of 1 indicates a 100bp sequence window that is not similar to any other region in the genome while a value of 0 represents a 100bp sequence that appears more than four times in different locations in the genome.

4.3.2.1 Normalisation of read depth across the samples

The read depth was normalised across the samples by applying the 'quantile' method in *cn.MOPS* package in R (Klambauer et al. 2012) . Quantile normalisation adjusts the distribution of read depth such that they are similar among the samples. Normalisation minimized the variation in read coverage that is due to the differences in number of reads sequenced among the samples (Figure 4.4A vs 4.4B).

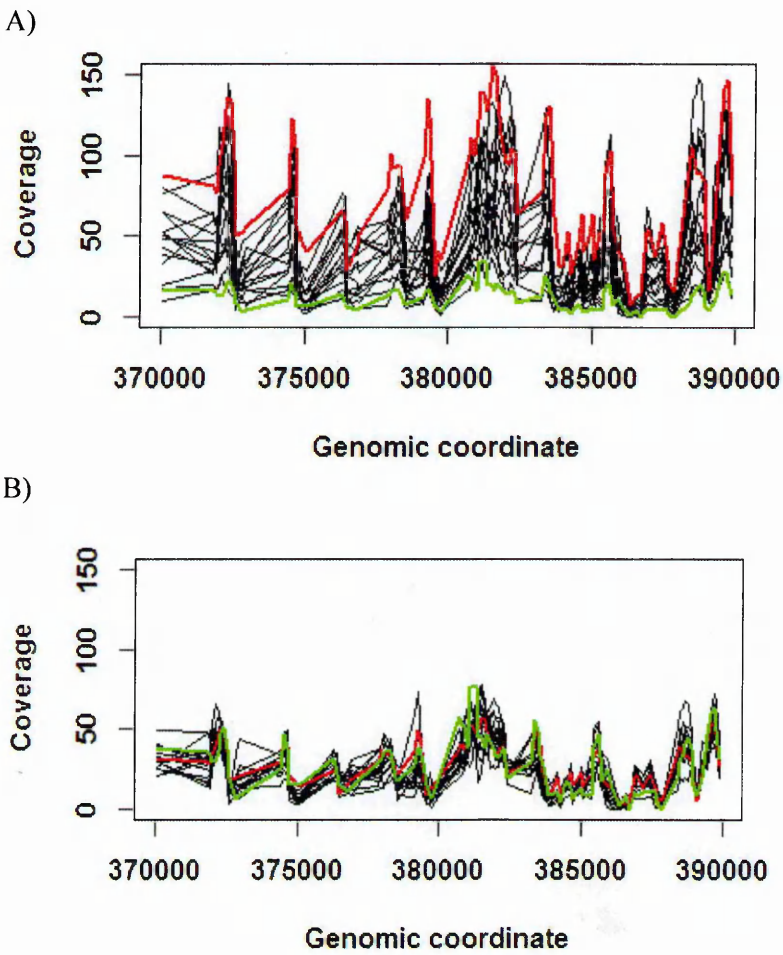


Figure 4.4. Normalisation of read coverage across samples
A) The raw coverage of 18 samples across a region in chromosome 1. The coverage of the sample with the highest overall coverage indicated by the red line and the coverage of the sample with the lowest overall coverage indicated in green. B) Normalised coverage of the same region on chromosome 1.

4.3.2.2 CNV detection using *cn.MOPS*

Genomic regions that vary in copy number were identified using *cn.MOPS*. The example illustrated in Figure 4.5 shows the normalised read coverage (black lines), the CNV segments detected (pink bars) in 11 out of the 18 samples and their genomic positions in the 'genes' method. The length and location of the CNV segments detected varied across the 11 samples (Figure 4.5). Only three samples had CNV segments spanning a whole gene. Most of the samples contained CNV segments that spanned very small portions of genes, and may be noise or short repeat regions located within genes. For this reason, a cut-off on the minimum length of CNV segments within a gene for it to be called a CNV was later applied.

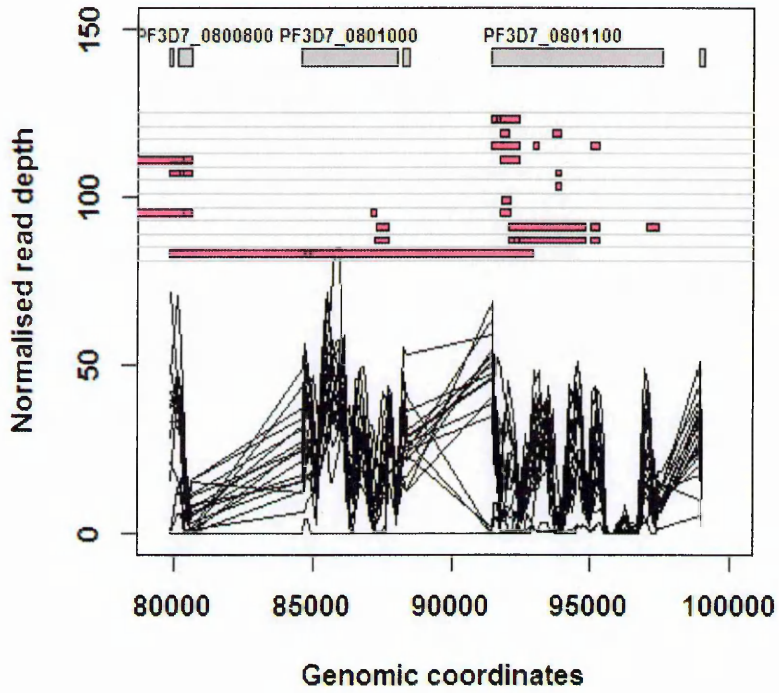


Figure 4.5. CNV calling using *cn.MOPS*
A region on chromosome 8 identified to contain CNV regions in the 'genes' method. The line plot shows the normalized read depth of the 18 samples. The CNV segments identified by *cn.MOPS* in 11 samples are shown by the pink horizontal bars. The three genes located in this region are shown by grey rectangles.

Overall, the total number and length of the CNV segments varied across the six methods (sequencing and CGH) and 18 samples (Appendix 4.1 A-F). The 'genome' had the highest number of segments owing to the inclusion of intergenic regions in the analysis (Appendix 4.1C). The 'probome' had the lowest number of segments among the sequence methods since only regions targeted by probes, ~ 10000 regions of 70bp width, were analysed (Appendix 4.1D). The number of CNV segments detected in the CGH data was higher when using the *GADA* method than *cn.MOPS* due to the inability of *cn.MOPS* to analyse data of probes with missing values in any one sample whereas GADA could analyse data of the same probes ignoring the sample with missing values (Appendix 4.1F vs. 4.1E). The number of probes analysed by GADA was 9615 whereas 4749 probes were analysed using *cn.MOPS.* The majority of the CNV segments were observed to be shorter than 500bp in the 'exome', 'genes' and 'genome' methods (Appendix 4.1A-C). The 'probome', 'CGH (GADA)', and 'CGH (cn.MOPS)' had most of the CNV segments above 1000bp (Appendix 4.1 D-F).

4.3.3 Gene copy number variation definition

The genes that lie in the CNV segments were identified and found to differ in number between the samples in a method and also between the 6 analysis methods (Appendix 4.2).The fraction of the gene lengths detected to be copy variable were calculated for each of the methods (Appendix 4.2). In the sequence-based methods, the majority of the genes had less than 20% of the gene length identified to be copy number variable except in the case of the 'probome' which had a majority greater than 80% (Appendix 4.2 A-D). The majority of CNV genes identified in CGH data had a copy number difference in at least 80% of the probes (Appendix 4.2E-F). For subsequent analysis, a CNV gene was defined as those with greater than 50% of their gene length/probes within identified CNV segments.

There were differences between the sequence and the probe based methods. First, a number of 100bp windows per gene were assessed sequence methods ('exome', 'genes' and 'genome') whereas on average two 70bp windows per gene were assessed in CGH data and 'probome' methods. Second, the total number of genes assayed by the two methods varied. A total of 5179 genes per genome were assayed in sequence methods ('exome', 'genes' and 'genome') whereas 5031genes were assayed in CGH methods ('GADA', 'probome') and 3121 genes in 'CGH (cn.MOPS)'.

4.3.4 CNVs detected by next-generation sequencing

The CNV genes identified in both sequence and microarray data , defining CNV genes as those with greater than 50% of gene length/probes in a gene identified to be copy number variable, are distributed throughout the parasite's genome (Figure 4.6). There was poor overlap in the CNV genes identified from the sequence and microarray methods (Figure 4.6). The reason for the higher number of CNV genes detected in CGH data using *GADA* than in Chapter 3 is that low stringency was used in CNV definition in this analysis. The number of different genes showing copy number gain, loss or both gain and loss ranged from 191 to 828 in the six methods (Figure 4.7A). Many CNV genes identified were exclusive to each method with approximately 300 CNV genes shared between any two methods (Figure 4.7B). The CNV genes identified using the three sequence methods, 'exome', 'genes' and 'genome', were significantly shorter (p-value < 0.0001 using t-test) than those identified in CGH data and the 'probome' method (Figure 4.7C). The length of the CNV genes identified on CGH data using the two methods ('cn.MOPS' and 'GADA') were statistically similar (p-value=0.37 using t-test). Also, the CNV genes length identified using 'probome' method was statistically similar to 'GADA' method (p-value 0.05).

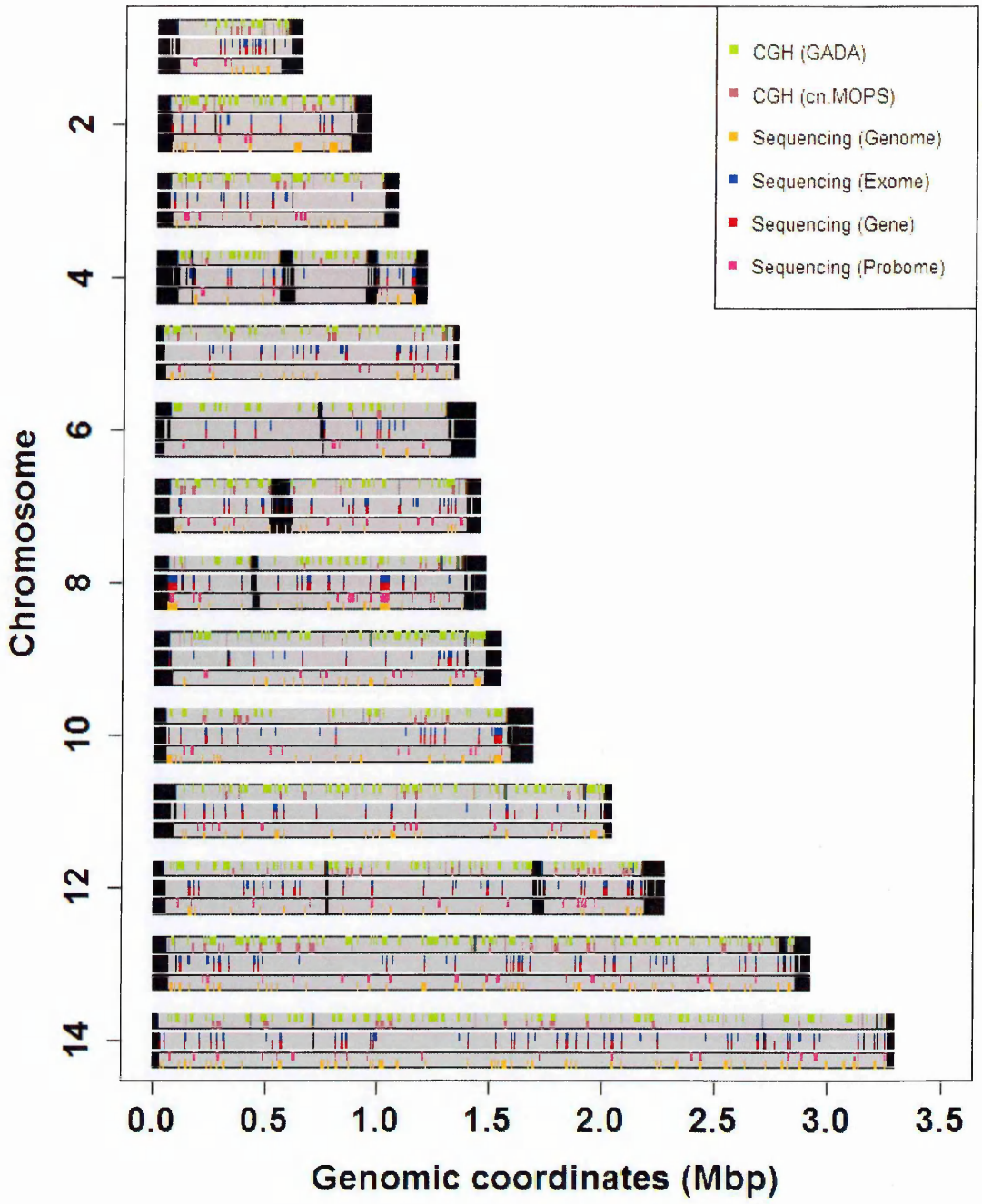


Figure 4.6. Genomic location of CNVs identified in CGH and sequence data in 18 isolates.

The location of CNV genes are shown as coloured vertical bars. The CNVs detected by CGH (GADA) method in green, CGH (cn,MOPS) in maroon, 'exome' in blue, 'genes' in red, 'probome' in magenta and 'genome' in orange. The regions targeted by the microarray probes, the exons and the full genome are shown in grey in the top, middle and bottom row of each chromosome, respectively. Regions excluded from the analysis are shown in black.

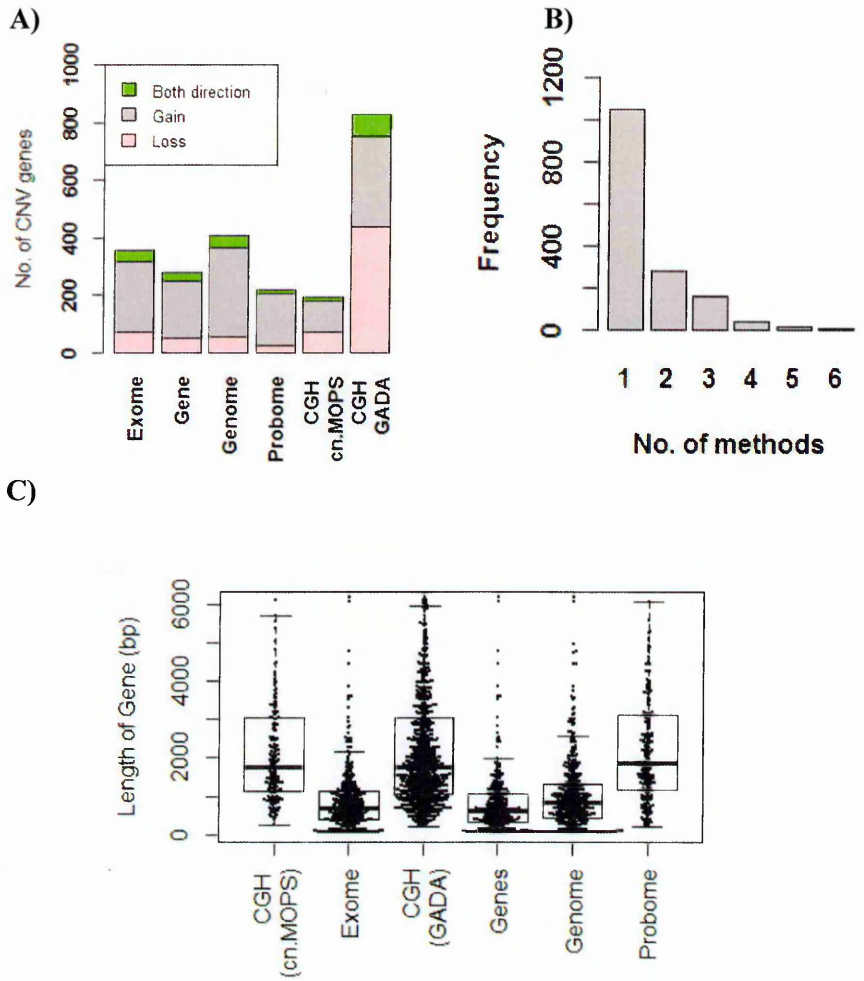


Figure 4.7. The number and length of CNV genes detected by five methods using *cn.MOPS.*
A) Barplot showing the number of CNV genes in six methods and the proportions of genes with increased copy number (gain) in grey and decreased copy number (loss) in pink and CNV genes with both gain and loss in green. B) Barplot showing the number of CNV genes exclusively detected in a single method, or detected in two or more methods. C) Boxplot showing the distribution of the length of CNV genes in the six methods.

4.3.4.1 Comparison between CNV genes identified in microarray and sequence data

The PPV and NPV of each method in detecting a CNV gene identified in CGH (GADA) were determined across the 18 samples. These calculations were based on less than 100 CNV genes because less than 100 CNV genes were detected by both CGH (GADA) and either one of the methods. Less than 10 CNV genes in each of the methods had a PPV of 1, i.e., all samples

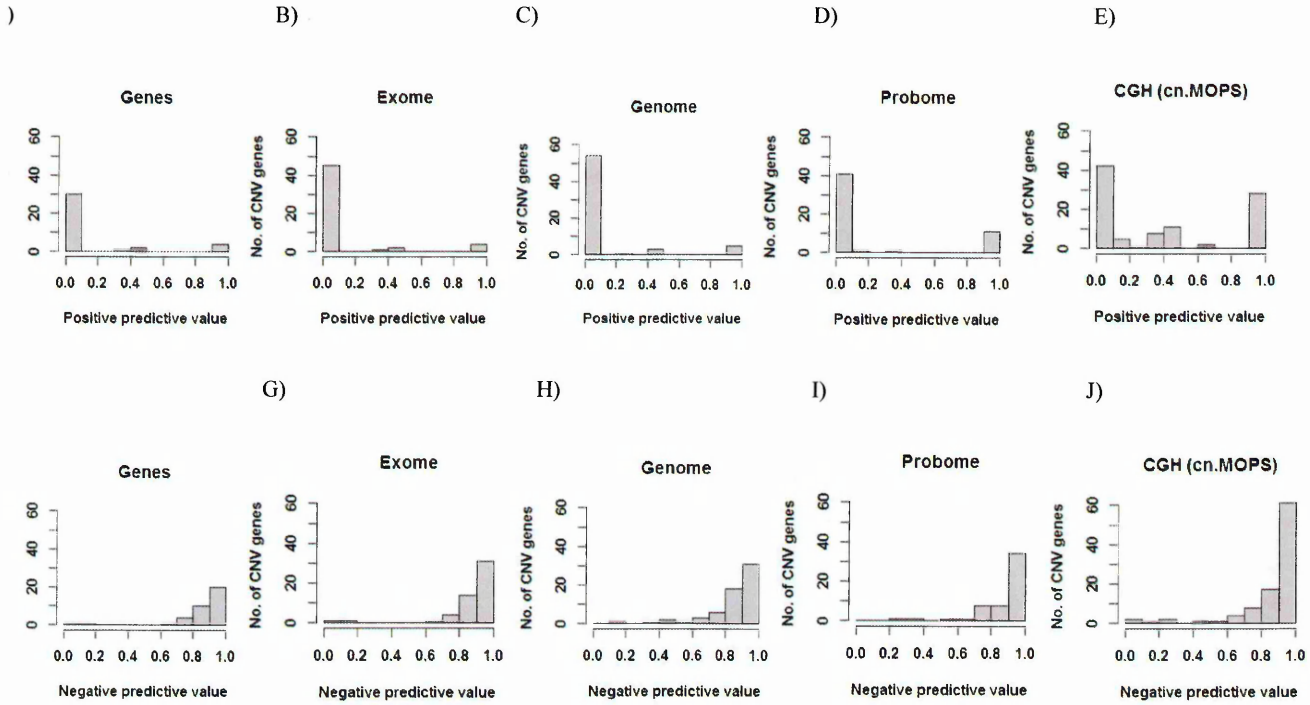
with the CNV in a method also had the CNV detected using CGH (GADA) (Figure 4.8 A, B, C, D and E). However, the majority of the CNV genes detected by each of the methods had low PPV of 0 and high NPV of 1. The low PPV indicates that the CNVs identified in one of the five methods were not identified in the same samples in CGH (GADA) (Figure 4.8 A, B, C, D and E). High NPV observed shows that CNV genes that were not identified in a method were also not detected in CGH (GADA) (Figure 4.8 F, G, H, I and J). CGH (cn.MOPS) was found to have a higher proportion of genes with high PPV than the other methods indicating that CNV genes that were detected in this method were also detected in CGH (GADA) (Figure 4.8H).

4.3.5 CNVs detected by both microarrays and sequencing

Examples of CNV genes identified in the same samples in both CGH (GADA) and sequence methods are listed below (Table 4.2). Some of the genes are involved in regulation of gene expression (PF3D7_0925700, PF3D7_1105000). One gene encodes proteins that form part of the 60S ribosomal subunit of the ribosome (PF3D7_1351400). Two genes are involved in gametogenesis, PF3D7_1038400 (Scherf et al. 1992) and PF3D7_0935400 (Eksi et al. 2012).

4.3.6 CNVs exclusively detected in sequence data or microarray data

One of the CNVs detected in sequence data and not in CGH data of the 18 sequenced samples was an amplification of three consecutive genes in chromosome 12 (Figure 4.9A). However, this amplification was observed in one sample of the 183 samples with CGH data that was not sequenced (Figure 4.8B). One of the genes in this CNV, PF3D7_1224000, known as GTP cyclohydrolase 1 (*gch1*) has been previous reported to be amplified in field populations of *P. falciparum* (Nair et al. 2008, Mackinnon et al. 2009, Robinson et al. 2011) as well as in culture



gure 4.8. Positive and negative predictive values (PPV and NPV) of CNVs genes detection in five genome sequences analysed by *.MOPS* and CGH (cn.MOPS) with CGH (GADA) as the gold standard. Histogram of the positive predictive value (PPV) CNV gene tection in the five method A) 'genes' B) 'exome' C) 'genome' D) 'probome' and E) CGH (cn.MOPS). The negative predictive value (NPV) of 'genes' G) 'exome' H) 'genome' I) 'probome' and J) CGH (cn.MOPS)

(Kidgell et al. 2006, Jiang et al. 2008b). The presence of up to 11 copies of the *gch1* identified in field isolates (Nair et al. 2008) is consistent with the extremely high sequence coverage of this region for the isolate displaying the amplification (Fig. 4.9A). The estimated copy number of the region in our analysis was 4. *gch1* codes for the first enzyme in the folate synthesis pathway. Its amplification has been observed to occur together with a SNP (*dhfr-164L*) present in an enzyme, dihydrofolate reductase (dhfr), which occurs in the same metabolic pathway and which is targeted by antifolate antimalarial drugs. The amplification is thought to be a compensatory mechanism as a result of presence of the antifolate resistance SNP mutation (Nair et al. 2008, Kidgell et al. 2006).

A large deletion at the right end of chromosome 9 observed in CGH data, in 3 of the 19 samples sequenced, was observed to contain sequence coverage in sequence data (Figure 4.10). However, four genes in the region (PF3D7_0935400, PF3D7_0936000, PF3D7_0936400 and PF3D7_0936500) were observed to be copy number variable by sequencing in reference to the mean of all the samples sequenced (Table 4.2). The reference sample in CGH experiment is known to have the deletion (Mackinnon et al. 2009), and therefore clinical isolates with similar copies to the reference, i.e., a $\log_2$ ratio close to zero, indicate a deletion in the samples. Of interest to this study of field isolates is that this deletion has previously only been detected in laboratory cultured lines and thought to be an adaptation mechanism to growth *in vitro* (Kemp et al. 1992, Shirley et al. 1990, Mackinnon et al. 2009).

Another interesting CNV gene, *clag 3.1* identified as PF3D7_0302500, belonging to the cytoadherence-linked asexual gene family (*clag* genes) was detected in one sequence method ('probome') and microarrays. One or both of the *clag 3* genes (PF3D7_0302200/*clag3.2* and

Table 4.2. A list of CNV genes detected in CGH (GADA) and at least one sequence method with the fraction of the number of samples with CNV in a method that had the CNV detected in both the method and CGH (GADA)

Gene	Product	Exome	Gene	Genome	Probome
PF3D7_0107200	Carbon catabolite repressor protein 4 putative				1/1
PF3D7_0302500	Cytoadherence linked asexual protein 3.1 (CLAG3.1)				1/1
PF3D7_0312000	Conserved Plasmodium protein unknown function	1/6	1/6		
PF3D7_0315200	Circumsporozoite-and TRAP-related protein (CTRP)				1/5
PF3D7_0422300	Alpha tubulin 2	1/5	1/5		1/5
PF3D7_0522800	G10 protein putative				1/1
PF3D7_0915300	Conserved Plasmodium protein unknown function				1/8
PF3D7_0925600	Zinc binding protein (Yippee) putative	2/7	1/7		
PF3D7_0925700	Histone deacetylase (HDAC1)			1/7	
PF3D7_0935400	Gametocyte development protein 1 (GDV1)				1/14
PF3D7_0936000	Ring-exported protein 2 (REX2)				1/3
PF3D7_0936400	Ring-exported protein 4 (REX4)			2/13	
PF3D7_0936500	Plasmodium exported protein unknown function			1/8	
PF3D7_1038400	Gametocyte-specific protein (Pf11-1)	1/16	1/16	1/16	
PF3D7_1105000	Histone H4 (H4)	5/7	5/7	3/7	5/7
PF3D7_1107100	Nucleic acid binding protein putative				1/1
PF3D7_1107400	Rad51 homolog (RAD51)	1/4	1/4	1/4	
PF3D7_1114200	GTPase activator putative			1/1	
PF3D7_1245200	Conserved Plasmodium membrane protein unknown				1/6
PF3D7_1333100	Conserved Plasmodium protein unknown function				1/3
PF3D7_1351400	60S ribosomal protein L17 putative			1/3	
PF3D7_1479000	Acyl-CoA synthetase (ACS1a)	1/1	1/1	1/1	1/1
Total		12	11	12	17

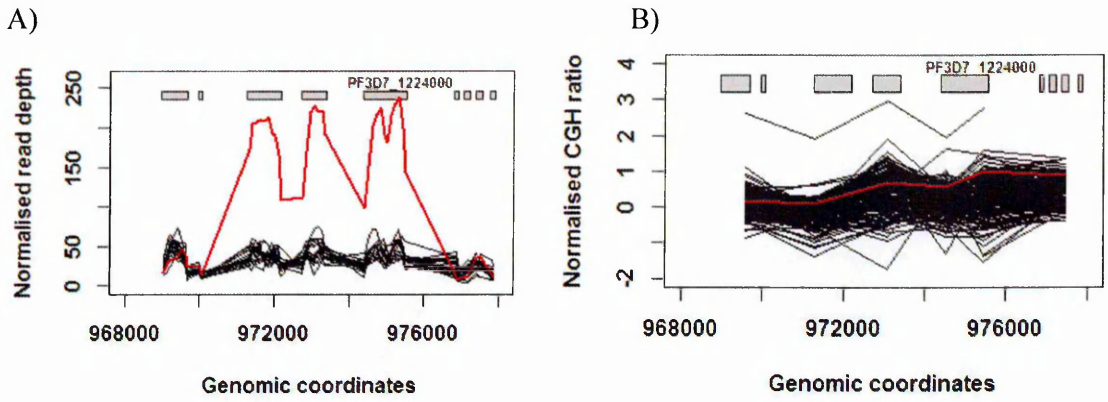


Figure 4.9. Amplification of GTP cyclohydrolase gene 1
A) Plot of normalised coverage of a region on chromosome 12 containing three CNV genes identified in the 'exome' method. B) Plot of the CGH ratio of the same region of all the 183 samples with CGH data. The coverage and CGH ratio of the sample with the amplification in among the sequenced isolates is indicated by the red line.

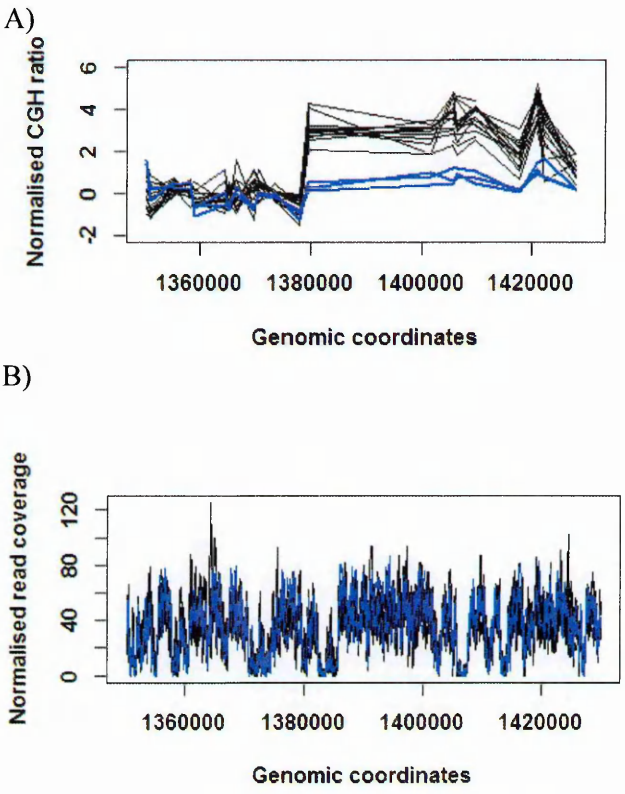


Figure 4.10. A CNV identified in chromosome 9 using microarrays but not using sequencing.
Plot of A) CGH ratio along a CNV region on chromosome 9. B) Plot of the normalised sequence coverage in the 'genome' method across the same region. The blue lines indicate CGH ratio/normalised coverage of samples with the chromosome 9 deletion observed in CGH data.

PF3D7_0302500) have been previously reported as CNVs in both clinical isolates (Robinson et al. 2011) and laboratory lines (Jiang et al. 2008b, Mackinnon et al. 2009, Iriko et al. 2008). However, in one of the studies, *clag 3.1* was reported to be amplified (Jiang et al. 2008b). The two *clag 3* genes are known to exhibit a sequence similarity of greater than 90% (Kaneko et al. 2005). Therefore, only a small proportion of each gene can be uniquely mapped to two genes in the reference genome. For the *clag 3.1* gene, CNV segments were detected using the sequence methods 'exome', 'gene' and 'genome', but for the above reason, the criterion of CNV segments constitution at least 50% of the gene length for it to be designated as a CNV gene were not met (Figure 4.11A). Closer inspection of the raw sequence coverage of only the regions that differed in sequence between the two genes shows that some of the samples had no coverage in these regions in the *clag 3.1* gene but had sequence coverage in *clag 3.2*, thus suggesting that there may be a true deletion of this gene (Figure 4.11 C vs. B).

4.4 Discussion

Array CGH has been widely used to detect CNVs in *P. falciparum* genome (Cheeseman et al. 2009, Kidgell et al. 2006, Ribacke et al. 2007, Mackinnon et al. 2009, Mok et al. 2011, Jiang et al. 2008b). Recently, there has been an attempt to apply whole genome sequencing to detect CNVs in the expectation that its higher resolution might yield more accurate results (Robinson et al. 2011, Samarakoon et al. 2011b, Sepulveda et al. 2013). A total of 7 CNV genes were identified in 5 clinical isolates in one of these studies, of which 2 have been previously reported (Robinson et al. 2011). In a second study, 4 out of 7 isolates sequenced had publicly available CGH data (Sepulveda et al. 2013). In the four laboratory parasite lines, the proportion of CNV hits (100bp windows detected to be copy number variable) detected in

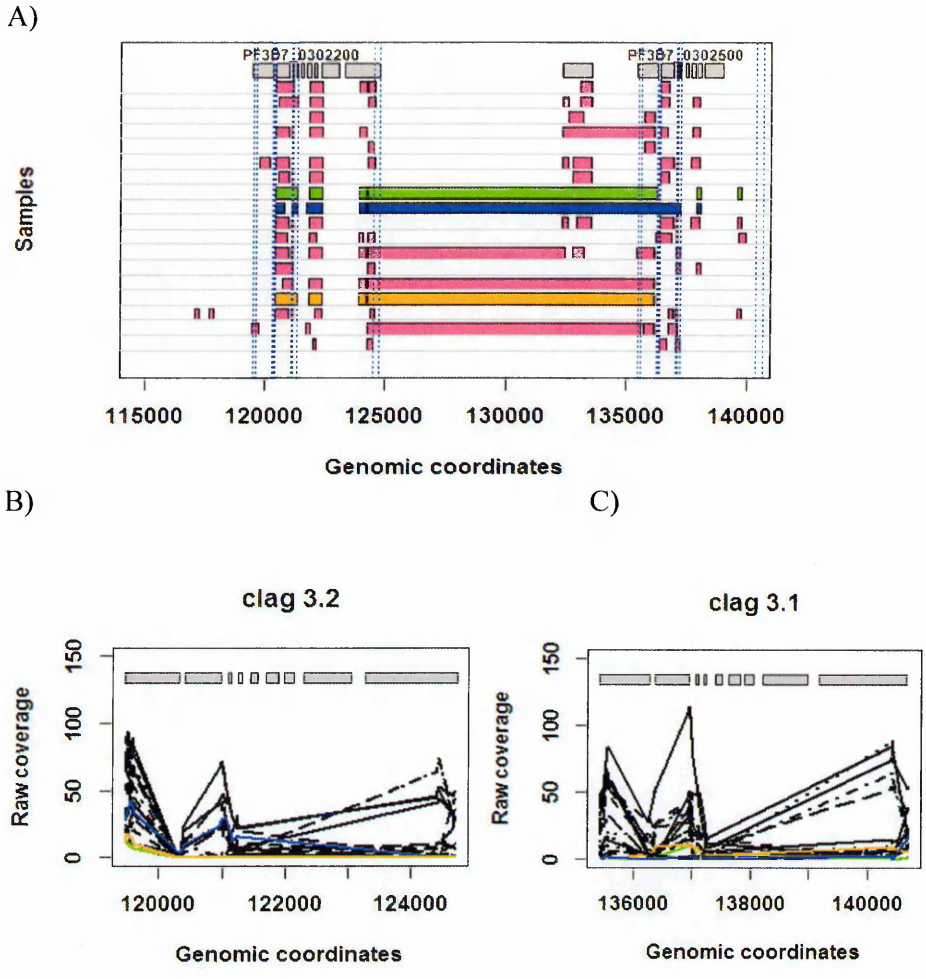


Figure 4.11. Deletion of *clag 3.1* gene (PF3D7_0302500) on chromosome 3
A) Plot of the CNV segments (deep pink, green, blue and orange rectangles) identified in the region containing the two *clag 3* genes, *clag 3.2* (PF3D7_0302200) and *clag 3.1* (PF3D7_0302500), on chromosome 3 in the 'exome' method. The blue vertical lines show the regions that have sequences unique to each gene. Plot of raw per base coverage of only the sequences unique to each gene B) the *clag 3.2* and C) *clag 3.1*. The samples with very low coverage/no coverage in the unique regions of *clag 3.1* are shown by green, orange and blue lines and may be real deletions. The CNV segments detected in these samples are shown in similar colours in A.

The main goal of this chapter was to confirm the CNVs identified using microarrays by whole genome sequencing. Comparison of the two technologies of CNV detection has its challenges. First, CNVs detection in *P. falciparum* using sequencing data is still new and the analysis remains a challenge hence complicates the comparison with microarrays. Second, most

analysis tools used in the two technologies were developed to suit each of the individual technologies, therefore comparing CNVs detected by different algorithms may be problematic. Third, the two technologies differ in the regions of the genomes targeted with sequencing covering the whole genome whereas microarrays target specific regions of genes. Fourth, the reference upon which copy number was calculated differ. In microarrays, the number of gene copies in the samples is in reference to a single parasite genome while in sequencing, the number of copies is in reference to the mean of all the samples. Lastly, the forms of 'noise' (measurement error) differ in the two technologies: whereas microarrays average out the signal, sequencing generates high base to base variability in read depth.

These challenges were overcome by employing various strategies in CNV detection analysis. First, regions with less accurate mapping in *P. falciparum*, i.e., low complexity regions (introns and intergenic regions) and highly polymorphic genes (*vars*, *rifins* and *stevors*), were excluded prior to CNV calling. The 'genome' target, with low complexity regions included in analysis, had the highest number of CNV segments detected in the samples ranging from 1000-3000 per sample (which did not correspond to high numbers of CNV genes) compared to the target genomes ('genes', 'exome', 'probome') range (5-1200 segments) which had the low complexity regions excluded. Most of these CNV segments in the 'genome' were located in the intergenic and intron regions.

Second, to analyse similar targets in the two technologies, a sequence genome target, 'probome', made up of the genomic regions targeted by the microarray probes were analysed in sequence data. Also, a target genome of only coding regions of the genome named the 'exome', which are targeted by the array, was included in the comparison. The number of

overlapping CNV genes detected in the samples was greatest between the 'probome' and CGH (GADA).

Differences in CNV detection in the two technologies is also contributed by the differences in the reference used in copy number estimation. In sequencing, copy number was calculated from the normalised coverage of the region in an isolate against the mean of the all the isolates whereas in microarrays it was calculated against one reference isolate. Another major difference in the microarrays and sequencing technology is measurement used. In microarrays, an average intensity signal is obtained for a 70bp sequence whereas in sequencing the read depth of each base is obtained. The read depth is biased by GC content and uniqueness of a region. The choice of window size and regions over which read depth is calculated and normalised to minimize variation is a critical step. Though the read depth, calculated in 100bp windows and normalised in 100bp windows, a number of very short fragments of genes/exons were detected as CNVs, reflective of the high amount of variability in sequence data. To minimize the technology-specific noise, a cut-off was set on the minimum length of sequence and the number of probes required for a gene to be called a CNV.

After accounting for some of these variables, concordance was found in 21 CNV genes with the 'probome' method proving to be the most concordant. This is because its genomic targets were similar to CGH (GADA).

Two clear examples of major discrepancies between microarray and sequencing were found. A large deletion in chromosome 9 was found by microarray in three isolates but sequencing data showed presence of sequence reads in these same samples. Second, a high copy number amplification of three consecutive genes in another region of chromosome 12, one of which

was *gch1,* was found in sequence data but no amplification was observed in the same sample by microarrays but was instead was observed in another sample by microarrays that was not sequenced. One possible reason for these observed differences is that microarrays have a limited dynamic range and, unlike sequencing, may not accurately quantify the massive deletions and amplifications. A further possibility is that the presence of subpopulations of parasites in the isolate with the CNV were undetected by microarray but became amplified during library preparation and visible by sequencing. These examples illustrate that one technology cannot replace the other yet for the purpose of CNV detection.

The two *clag 3* genes best illustrates the difficulty of detecting CNVs in multiple regions of sequence homology in the genome. Focussing on the unique regions of these genes is one way to overcome the mapping problem, though mapping in the short unique regions may still be affected by sequencing errors and the stringency applied during read mapping to the reference. Increasing the sequence read length coupled with de novo assembly may help solve this problem (Zhao et al. 2013, Nijkamp et al. 2012).

Improvement of CNV detection methods using sequence data is still needed. One of the challenges of comparing CNV detection methods is the lack of a gold standard. Every technology has its challenges and even the analysis tools used for CNV detection in each of the technologies have shown to yield different results (Pinto et al. 2011, Mills et al. 2011b).

Chapter 5

Effects of CNVs on gene expression levels

5 Chapter 5: Effects of CNVs on gene expression levels

5.1 Introduction

In *P. falciparum*, a number of CNVs have been identified: however, the functional impact of these CNVs on gene expression has not been widely studied. In many organisms, it has been shown that CNVs alter gene expression levels and hence affect phenotype (Henrichsen et al. 2009, Stranger et al. 2007). CNVs have been shown to affect the levels of expression of genes within the CNVs ('direct effects') and also genes located outside the CNVs ('indirect effects'). In *P. falciparum*, an example of a direct effect of a CNV is an amplification of the GTP-cyclohydrolase 1 (*gch1*) gene that leads to increased levels of *gch1* expression which has been associated with antifolate drug resistance (Nair et al. 2008). An example of an indirect and global effect of a CNV is the 'super CNV' on chromosome 5 comprising an amplification of the region containing the *P. falciparum* multidrug resistance gene 1(*Pfmdr1*) associated with multidrug resistance gene. This CNV was found to be either positively or negatively associated with the level of expression of 269 genes (Gonzales et al. 2008). A number of *P. falciparum* CNVs have also been observed to influence the levels of expression of genes within and outside CNV regions (Mackinnon et al. 2009).

In this chapter, the impact of gene copy number variation on the levels of gene expression was assessed. Gene expression/transcriptome data were generated by Rono M et al. (unpublished). The relationship between gene expression and CNVs discovered in the same set of isolates described in Chapters 3 and 4 was investigated. The direct effect of alteration of gene copy number on gene expression and the global effect of these CNVs were investigated by correlation and linear regression analyses of CGH data in relation to expression data.

5.2 Methods

5.2.1 Samples

Seventy-two isolates out of the 183 isolates used for CNV identification in Chapter 3 had expression data and were used in this analysis. The 72 parasite isolates selected had complete maturation *ex vivo* and good quality array data of all the parasite stages. These isolates were from Kilifi pre-malaria decline, Kilifi post-malaria decline, Kisumu and Sudan, as described in Chapter 2. Twelve out of the 72 isolates had sequence data (chapter 4).

5.2.2 Gene copy number variation

The CGH data ($\log_2$ CGH ratio) used in this Chapter were pre-processed as described in Chapter 3. The mean CGH ratio was calculated for each gene from CGH ratio of probes targeting the genes and used for association analysis. Ninety-five different CNVs comprising 221 genes were detected in the 72 samples with both CGH and transcriptome data. The mean CGH ratio of probes contained in the CNVs (mean amplitude of CNVs) provided an indication of the size and direction of gene copy number differences between the test and reference isolates.

5.2.3 Gene expression data

Transcriptome data were generated by Rono. M et al. using a similar microarray to that used in CGH (Chapter 2). Parasite isolates were obtained straight from the arm of infected individuals at the ring stage of development and were matured *ex vivo* to obtain samples at all the stages of the parasite's 48-hour intraerythrocytic development. This step is required since levels of expression of most of the parasite's genes are known to vary across this cycle (Bozdech et al. 2003a). I was involved in maturation experiments of some of the parasites and set up of

sample database. Preparation of cDNA and amplification from extracted RNA, at each of the parasite stages, were carried out (section 2.3.1 and 2.3.2). The amplified product was labelled and hybridized against a reference line, a Kilifi laboratory-adapted strain designated as 'P4', as for CGH. The reference cDNA consisted of a pool of cDNA from all the intraerythrocytic developmental stages of the parasite.

Prior to analysis, Loess curves were fitted to the data for each gene across all isolates and the residual values were computed. For each gene and for each isolate, these residual values were used to calculate the mean value across all stages of development ('Mean'), and the mean value of data 6 hours each side of the time of maximum expression of the gene ('Max'). These isolate means were then used for the analyses described below.

5.2.4 Analysis of expression in relation to gene copy number

The effect of CNVs on the expression of genes located within them was assessed by performing a Pearson correlation between the mean CGH ratio per gene and 'Mean' expression data of each gene across the 72 isolates. The p-value for assessing significant differences from the expected correlation of zero under the null hypothesis of no relationship between gene copy number and expression levels was calculated using the *cor.test* function in R.

To investigate whether the correlation obtained from the analysis may have been by chance, a permutation test was performed. 100 random permutations of the expression data of each gene were generated and the correlations between mean CGH ratio of each gene and each of the 100 randomized expression data per gene were calculated. A Kolmogorov–Smirnov test for

significant differences between the distribution of observed correlations and distribution of correlations from the permuted data was applied using the *ks.test* function in R.

To analyse the relationship between gene expression (of genes in the whole genome) and the CNVs detected, a linear regression model was applied using the *lm* function in R. The mean $\log_2$ expression ratio of each gene was the dependent variable while the copy number state (fit as a fixed effect factor with levels for loss, gain or normal) was the independent variable. The difference in gene expression between samples with a gene copy number difference and those without was indicated by the regression coefficient (effect size). The significance of the difference was assessed from the p-value of the regression coefficient from the same model.

5.3 Results

5.3.1 Direct dosage effect of CNVs

The relationships between gene copy number and expression of 221 genes located within the 95 CNVs was determined by calculating the correlation between the CGH ratio and the $\log_2$ expression ratio in the 72 samples. Both positive and negative correlations between expression and copy number were observed (Figure 5.1A). Significant correlations were detected in 41 CNV genes located in 30 different CNVs (p value < 0.05 and absolute correlation > 0.23) (Table 5.2). These represent direct dosage effects of CNVs on gene expression levels. To test whether the observed distribution of correlations may have been obtained by chance, the expression data for each gene were randomized, 100 times per gene, among the 72 isolates and the correlation re-calculated, 100 times per gene (Figure 5.1B). The observed (Fig. 5.1A) and empirical (from randomized data, Fig. 5.1B) were found to differ, thus showing that the observed correlations between CGH and expression data were not as

expected by chance (Kolmogorov–Smirnov test with p value < 2.2e-16). These distributions differed in that there were unexpected peaks at strong negative and positive correlation values of r = -0.3 and r = 0.2-0.3 respectively (Figure 5.1A).

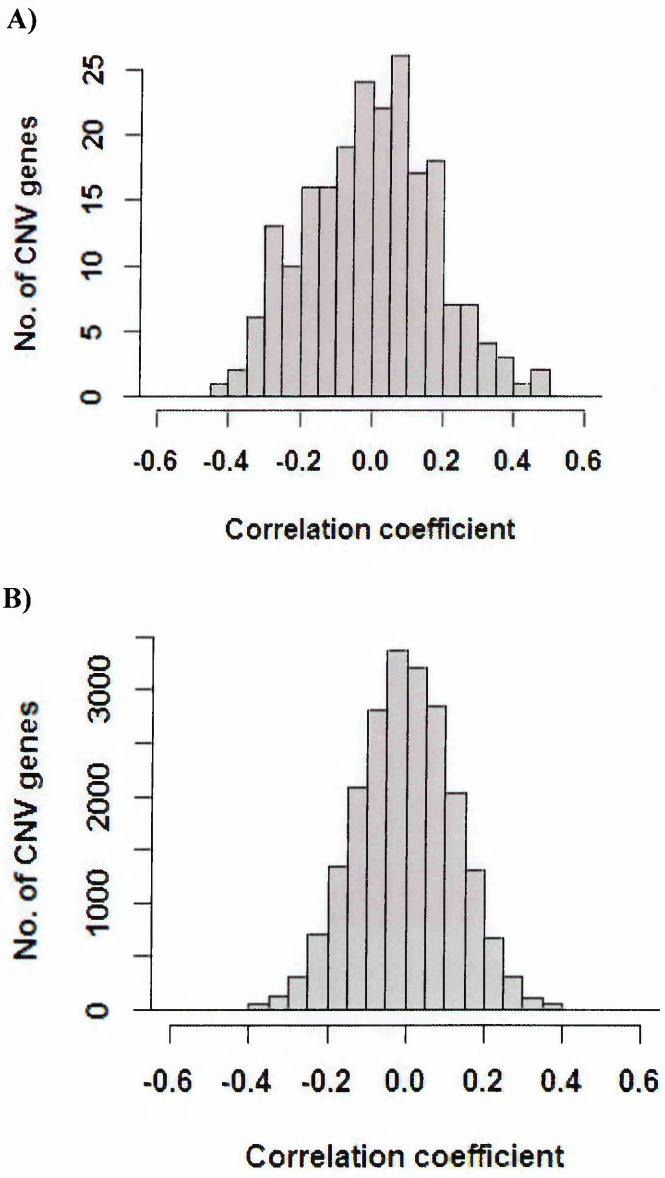


Figure 5.1. Correlation between CGH log_2 ratio of 221 CNV genes within 95 CNVs and their corresponding log_2 expression ratio
A) Histogram showing the distribution of correlation coefficient of CGH log_2 ratio and log_2 expression ratio of 221 CNV genes in 95 CNVs detected in 72 samples. B) Histogram showing the correlation between CGH data and 100 permutations per gene of expression data of the 221 genes.

5.3.2 Direction of CNV effects on expression

Some of the genes shown to have gains or losses in copy number by CGH showed correlations with gene expression that varied from strongly negative to strongly positive while others showed very low correlations (Figure 5.2A). For genes showing gains in copy number, there was an approximately equal number that showed significantly higher (i.e., positive correlations) vs. significantly lower expression levels (negative correlations) (15/31 vs. 16/31, p-value=1 using binomial test). For genes showing loss in copy number, slightly more genes showed significant positive relationships with expression (i.e., decreased expression) than significant negative relationships, though the difference in proportions between the two correlations was not significant (6/10 vs. 4/10, p value = 0.75 using binomial test).

Surprisingly, expression was observed in genes identified to be deleted (Figure 5.2A). This is possibly because some of the observed loss in gene copy number may not necessarily mean complete absence of the gene, but rather presence of reduced number of copies of the gene in the test isolates compared to the reference parasite line.

Also surprising is that, in some instances, genes located within the same CNV exhibited opposite directions of the CNV effect. An example of this is the largest CNV on chromosome 9 with a mean amplitude of 2.5 and containing the highest number of genes (15 genes) which showed both positive and negative correlations with expression of some of the genes contained within it. However only five genes with positive correlations were statistically significant (p value < 0.05) (Table 5.1).

Examples of genes showing positive correlations , i.e., an increase or decrease in copy number associated with an increase or decrease in gene expression levels, respectively, are genes

PF3D7_0508100 and PF3D7_1207000 (Figure 5.2D and E). An example of a CNV gene showing a negative correlation with expression upon loss of copy number is PF3D7_1149000 (Figure 5.2B). An example of a negative correlation upon gain in copy number is gene PF3D7_0925400 (Figure 5.2C). The direction of CNV effects for all the 41 CNV genes showing significant correlations between CGH and expression are shown in Table 5.2.

Table 5.1. Genes in a CNV showing both negative and positive associations between CGH and expression data

Gene ID	Correlation	P value
PF3D7_0935500	Positive	0.315833
PF3D7_0935600	Positive	0.010319
PF3D7_0935700	Positive	0.734123
PF3D7_0935800	Negative	0.507444
PF3D7_0935900	Positive	0.036364
PF3D7_0936000	Positive	0.027061
PF3D7_0936100	Positive	0.600683
PF3D7_0936400	Positive	0.028079
PF3D7_0936500	Negative	0.200308
PF3D7_0936700	Positive	0.460633
PF3D7_0936800	Positive	0.008986
PF3D7_0937000	Negative	0.172866
PF3D7_0937100	Negative	0.25637

Among the 41 genes that showed significant correlations between copy number and expression were 5 genes located in a region of chromosome 9 that contained a large deletion at low frequency in the populations. All the 5 genes within this CNV, PF3D7_0935600, PF3D7_0935900, PF3D7_0936000, PF3D7_0936400 and PF3D7_0936800, showed significant positive correlations between copy number and expression (Figure 5.3).

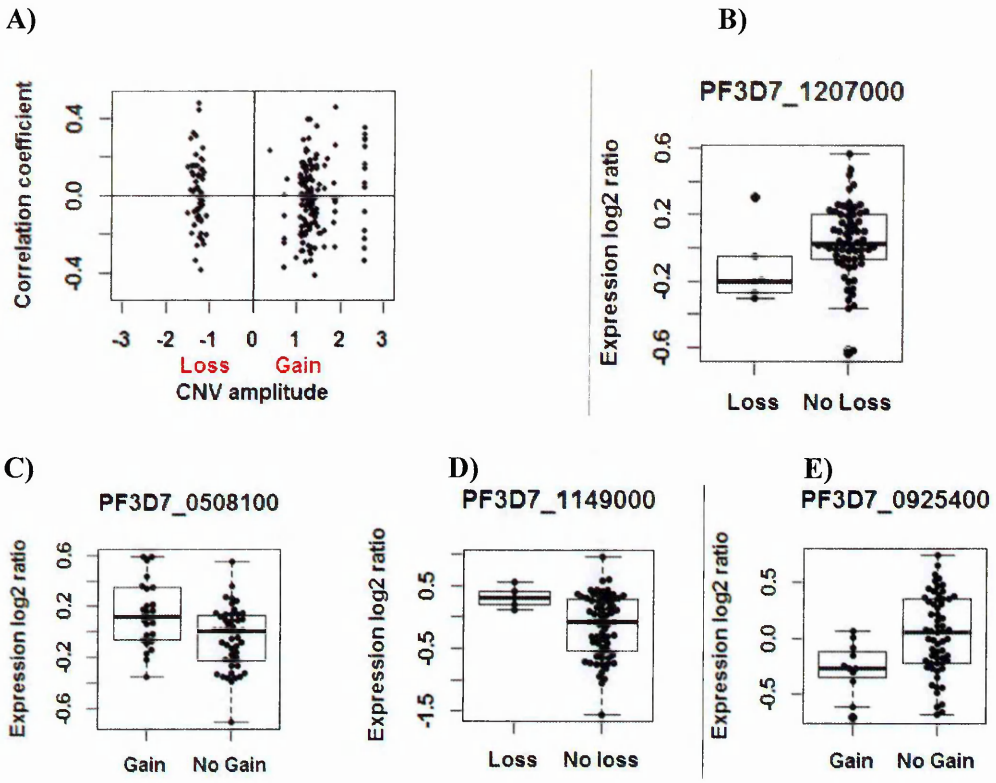


Figure 5.2. Direction of CNV effects on levels of gene expression
A) Plot showing relationship between the amplitude of 95 CNVs (mean $\log_2$ CGH ratio of probes within CNV) on the x axis and the correlation coefficient of each of the 221 CNV genes in the CNVs. Each spot represents the correlation coefficient obtained from the correlation between CGH and expression value in each of the 72 samples, of each CNV gene (x axis) and the amplitude of the CNV that contains the gene (y axis). Some of the 95 CNVs have more than two 2 genes located within them thus a single y value (amplitude) may have more than two different correlation coefficients (x axis). Box plots show 'Mean' gene expression data (a single 'mean' value, calculated from all the time points per gene, of an isolate (explained in section 2.3.3)) in samples with the CNV gene and those with normal copy number of the gene. A positive correlation between CGH and expression data was observed in gene PF3D7_1207000 **B)** and PF3D7_0508100 **C)** while a negative correlation was observed in PF3D7_1149000 **C)** and PF3D7_0925400 **D).**

This deletion has been previously reported in the reference parasite (Mackinnon et al. 2009): thus parasites bearing similar gene copies to the reference would have CGH $\log_2$ ratios close to zero, and those without the deletion will have higher CGH $\log_2$ ratios. The deletion was observed in 7 of the 72 isolates. Low levels of gene expression were observed in some of the

Table 5.2. Genes showing direct dosage effect of CNVs on expression

Gene ID	Protein Annotation	CNV state	Correlation
PF3D7_1129000	Spermidine synthase	Gain	Positive
PF3D7_1412400	Conserved protein, unknown function	Gain	Positive
PF3D7_1212300	Conserved protein, unknown function	Gain	Positive
PF3D7_1212400	Tetratricopeptide repeat family protein	Gain	Positive
PF3D7_0423400	Asparagine-rich protein	Gain	Positive
PF3D7_0936800	Plasmodium exported protein (PHISTc)	Gain	Positive
PF3D7_0935600	Gametocytogenesis-implicated protein	Gain	Positive
PF3D7_0322700	Conserved protein, unknown function	Gain	Positive
PF3D7_1329300	Chromatin assembly factor 1 subunit	Gain	Positive
PF3D7_0928000	Cytochrome c oxidase subunit 6B	Gain	Positive
PF3D7_0936000	Ring-exported protein 2	Gain	Positive
PF3D7_0936400	Ring-exported protein 4	Gain	Positive
PF3D7_0935900	Ring-exported protein 1	Gain	Positive
PF3D7_1128700	GPI-anchor transamidase	Gain	Positive
PF3D7_0515200	Conserved protein, unknown function	Gain	Positive
PF3D7_0925400	Protein phosphatase-beta	Loss	Positive
PF3D7_0925600	Zinc binding protein (Yippee)	Loss	Positive
PF3D7_1008900	Adenylate kinase	Loss	Positive
PF3D7_0822600	Protein transport protein SEC23	Loss	Positive
PF3D7_0826000	Conserved protein, unknown function	Loss	Positive
PF3D7_0804700	Conserved protein, unknown function	Loss	Positive
PF3D7_1143400	Translation initiation factor eIF-1A	Gain	Negative
PF3D7_1313000	Ubiquitin-like protein nedd8 homologue	Gain	Negative
PF3D7_1244800	Cytoplasmic translation machinery associated protein	Gain	Negative
PF3D7_1201900	Conserved protein, unknown function	Gain	Negative
PF3D7_0213600	Conserved protein, unknown function	Gain	Negative
PF3D7_1316800	Protein transport protein SEC20	Gain	Negative
PF3D7_1312800	Conserved protein, unknown function	Gain	Negative
PF3D7_1317300	Conserved protein, unknown function	Gain	Negative
PF3D7_0602200	MYND finger protein	Gain	Negative
PF3D7_0915300	Conserved protein, unknown function	Gain	Negative
PF3D7_0407100	Methyltransferase	Gain	Negative
PF3D7_0203100	Protein kinase	Gain	Negative
PF3D7_1245200	Conserved protein, unknown function	Gain	Negative
PF3D7_1128900	Conserved protein, unknown function	Gain	Negative
PF3D7_0302900	Exportin-1	Gain	Negative
PF3D7_1329200	Conserved protein, unknown function	Gain	Negative
PF3D7_1149000	Antigen 332, DBL-like protein	Loss	Negative
PF3D7_1207200	Conserved protein, unknown function	Loss	Negative
PF3D7_0309600	60S acidic ribosomal protein P2	Loss	Negative
PF3D7_1114800	Glycerol-3-phosphate dehydrogenase	Loss	Negative

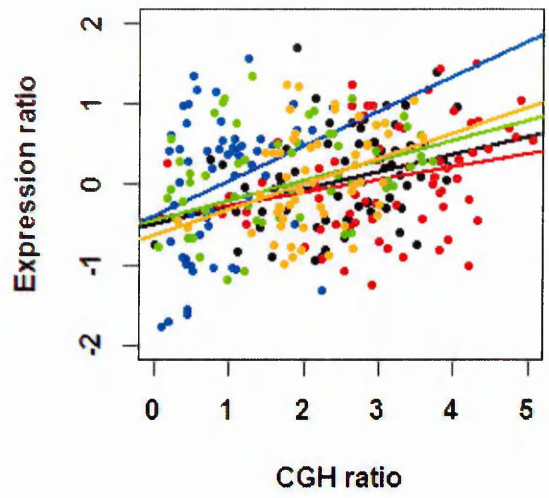


Figure 5.3. Association between the CGH ratio and expression ratio of genes located in chromosome 9 region containing a deletion
Plot of the relationship between CGH ratio (x axis) and expression ratio (y axis) of five genes PF3D7_0935600 (black), PF3D7_0935900 (red), PF3D7_0936000 (blue), PF3D7_0936400 (green) and PF3D7_0936800 (orange) located in the right arm of chromosome 9 contained within a CNV. The coloured linear regression lines fitted to the corresponding coloured data points are generated using *lm* function in R with the CGH ratio of the genes as the independent variable and the expression ratio as the dependent variable. All the fitted lines were found to be statistically significant (p value < 0.05).

isolates with CGH ratios close to zero or below 1 (i.e., with the deletion) (Figure 5.3). This CNV of interest, observed using CGH data (Chapter 3) but not confirmed using sequencing (Chapter 4), has previously been reported only in laboratory-adapted lines but not from clinical isolates. The low expression levels observed at low CGH ratios in some isolates may support the observation of a deletion of these genes. Alternatively, it may have resulted from a result of presence of polymorphism in the isolates which may interfere with microarray hybridization. This seems unlikely given that SNP-containing probes were excluded prior to CNV analysis (Chapter 3) and that all five genes contained within the region exhibited similar reductions in signal in both CGH and expression in some isolates.

5.3.3 CNV effects on the expression of genes located outside CNV intervals

Previously, CNVs have been associated with not only changes in expression of genes with altered copy number, but also genes outside the CNV boundaries (Mackinnon et al. 2009, Gonzales et al. 2008). In this study, the potential CNV regulation of expression of genes neighbouring the CNVs and also genes on different chromosomes was examined. The effect of the 95 CNVs on expression of approximately 4000 genes outside CNV regions was assessed using linear regression models. Gain and loss in copy number showed both negative and positive effects on expression of genes outside the CNVs (Figure 5.4A). For the majority of genes the level of expression was not affected by CNVs, indicated by high density observed at zero (Figure 5.4A). CNV effects on expression were more extreme for deletions than for amplifications, especially in causing a reduction in expression (Figure 5.4A). Five percent of these CNV effects on expression were found to be significant (log10 (p value) < -1.30) (Figure 5.4B).

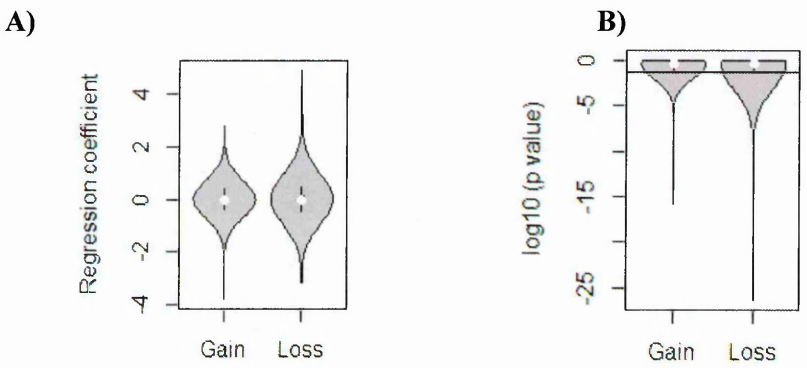


Figure 5.4. CNV effects on expression of genes located outside the CNVs
The violin plot shows the results from a linear regression model that was used to explore the relationship between changes in expression of approximately 4000 genes with changes in gene copy number in 95 CNVs. Violin plot showing the distribution of the **A)** regression coefficient of the CNVs (gain and loss) on gene expression compared to normal copy **B)** log10 (p value) indicating the significance of the CNV effect. The log10 of the p value of 0.05 is -1.30. In the violin plots, the median value is indicated by the white spot, and the interquartile range is indicated by the short black lines. The outer line shows the density estimations at each value.

5.3.3.1 Intra-chromosomal effects of CNVs on gene expression

CNV effects on expression were observed in some genes located in close proximity to the CNVs (Figure 5.5). The estimated proportions of genes neighbouring the CNVs that showed significant influence of the CNV on their expression were highest at 30-50kb and 100-200kb from CNV boundaries (defined as the end of the last gene and the start of the first gene in the CNV). The lowest proportion was at an interval of 10-20kb from the CNV boundaries.

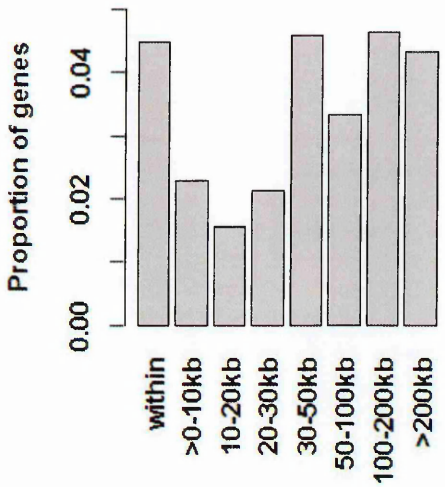


Figure 5.5. Effect of CNV on expression of genes at a distance from CNV
Barplot showing the proportion of genes as a function of their distance from the CNV boundaries that had expression levels significantly affected by CNVs. The CNV boundaries include the start of the first gene in the CNV and the end of the last gene in the CNV. Proportions were calculated as the fraction of those genes whose expression levels were found to be significantly affected by the CNV out of the number of genes located within these intervals for which data were available. The CNV effect was assessed using the linear regression model described above.

5.3.3.2 Inter-chromosomal effect of CNVs on expression

CNVs were also found to affect the levels of expressions of genes on different chromosomes. The example in Figure 5.6 shows expression values of genes in samples with and without a CNV (cnv12_413) on chromosome 12. The CNV consists of an amplification of PF3D7_1248600, a gene coding for a conserved Plasmodium protein of unknown function.

The CNV was found to significantly affect the expression of 170 genes located in all 14 chromosomes of the parasite genome. This CNV is located approximately 1Mbp downstream from another CNV (*gch1* amplification) region in chromosome 12 associated with expression of 269 genes (Gonzales et al. 2008).

The isolates with the CNV (cnv12_413) showed two patterns of expression of the genes under CNV regulation. Some of the genes had high expression (expression values >0) in some of the isolates while others showed low expression (expression values<0). The isolates without the amplification (columns under black bar) exhibited a single pattern of gene expression in most of the isolates that was similar to a subset of isolates with the CNV (the first 18 columns under green bar). These isolates with or without the CNV showing similar pattern of gene expression originate from Kilifi and Kisumu populations. The isolates with the CNV showing a different expression pattern (last 14 columns under green bar) originate from Sudan. Some of the genes showing the CNV effect on expression include 12 genes known to be involved in splicing of pre-mRNA and 11 genes that encode chaperones and their regulations (http://mpmp.huji.ac.il/). Pre-mRNA splicing has been identified as one of the mechanisms utilized by plants to regulate expression of stress responsive genes (Mastrangelo et al. 2012, Dubrovina, Kiselev and Zhuravlev 2013). In Arabidopsis, changes in the expression of a gene encoding a protein, Sm-like protein 5 (LSm5), involved in mRNA splicing, has been found to regulate splicing of stress-responsive genes which influence the ability of the plant to tolerate salt conditions (Cui et al. 2014). In *P. falciparum*, some of the genes, e.g., heat shock protein 70, that encode chaperones have been shown to be protective to the parasite under heat stress (Pesce et al. 2008, Shonhai et al. 2011). The difference in heat stress response between Sudan and the two Kenyan populations (Kilifi and Kisumu) may be due to difference in thermal climatic conditions that the human host occupies.

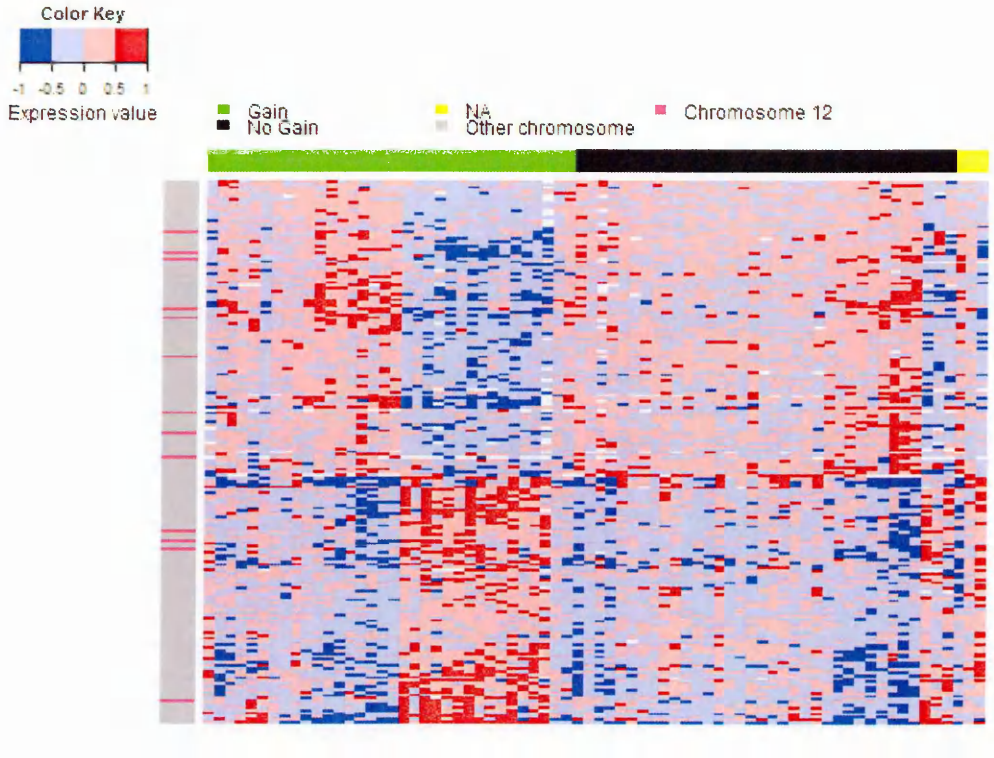


Figure 5.6. Increased copy number of PF3D7_1248600 gene shown to affect expression of genes located on different chromosomes
A heatmap showing the expression values of 183 genes with expression found to be significantly affected by cnv12_413 containing a single gene (PF3D7_1248600) on chromosome 12 using linear regression model. 170 of these genes were located on chromosomes other than (grey rows) the chromosome on which it was located (pink rows). The red colour shade indicates expression of greater than 0 while the blue colour indicates expression values below 0. The samples with the amplifications are indicated by the columns under the green bar, while the samples with normal copy are the columns below the black bar. A few samples had missing CGH data for this gene (columns under yellow bar).

5.4 Discussion

The impact of CNVs on gene expression was assessed in light of prior evidence that CNVs affect gene expression (Gonzales et al. 2008, Mackinnon et al. 2009) and play a role in phenotypic variation and adaptation (Nair et al. 2008, Chavchich et al. 2010) . Forty-one out of 221 CNV genes were found to exhibit direct dosage effects on gene expression. These CNV genes showed significant positive and negative correlations between CGH and expression data in both deletions and amplifications. The positive association between copy number and gene

expression is expected and commonly reported in studies. However, the observed negative correlations are unexpected under a model of direct gene dosage effects on expression although have been observed elsewhere (Stranger et al. 2007, Henrichsen et al. 2009). These negative correlations may be as a result of the CNV being in linkage with other regulatory variants that cause the change in gene expression (Stranger et al. 2007). They may also be explained by the presence of a feedback loop whereby the expression of additional copies of a gene promotes the expression of a repressor that reduces the level of gene expression or reduced copies of a gene induces the expression of a gene enhancer leading to increased expression (Henrichsen et al. 2009). The extra copies of a gene may also interfere with the steric conformation of the gene disrupting its access to transcription machinery (Henrichsen et al. 2009, Sexton et al. 2007). In cases where there was a lack of association between CGH and expression data there may be a different regulatory mechanism to gene dosage operating. In the case of amplified genes, copies of genes may be located in different regions of the genome and hence regulated through a different chromatin environment (Henrichsen et al. 2009): the microarray technology cannot distinguish between tandem and non-tandem duplications and thus would not detect such transpositions.

Some CNVs were observed to affect the expression of genes neighbouring or at a distance from CNVs. These include a large proportion of genes located at distances greater than 30kb from the CNVs boundaries and some genes located on different chromosomes. This is consistent with previous studies in *P. falciparum* showing that CNVs affect the expression of genes located outside as well as inside CNVs (Gonzales et al. 2008, Mackinnon et al. 2009). The observation of CNV effects on genes located at different distances from the CNV implies that CNVs may regulate other genes through various mechanisms. First, CNVs may alter the chromatin structure that is known to play a role in determining gene transcriptional activity

through epigenetic mechanisms (Cui and Miao 2010). The chromatin structure can be altered through modifications at the histone tails forming histone marks upon which effector proteins bind, or through histone variants, or through other chromatin remodellers that influence accessibility of DNA to transcription factors (Cui and Miao 2010, Duffy et al. 2012). The histone marks enable the recruitment of chromatin-associated proteins which maintain the chromatin active and repressive states. Amplifications or deletions that disrupt the histone marks may interfere with transcriptional activity in the chromatin environment. An example of this in the human genome is a reduction in copy number of the D4Z4 repeats that interferes with chromatin structure through loss of histone marks (Bodega et al. 2009, Zeng et al. 2009). In addition, genomic alterations that affect genes encoding proteins that bind to the modified histones and those involved in histone modifications may lead to global effects on expression (Kleinjan and van Heyningen 2005). In *P. falciparum,* several genes involved in histone modifications (HDACs, HATs) and those interacting with the modified histones (SET1, SET2, *Pf*HPI, PF3D7_1141800among others) have been identified and have been associated with either repressive or active states of the chromatin (Duffy et al. 2012, Cui and Miao 2010). In this study, genes encoding histone deacetylase 1 (HDAC1) and SET9 were observed to be copy number variable in both microarrays and sequencing and only microarrays, respectively. One of the genes under chromatin-mediated regulation is the *pfap2-g* (PF3D7_1222600) whose expression, by removal of the chromatin silencing histone marks, has been linked to reprogramming of the parasite's transcriptome promoting conversion of asexual forms of the parasite to sexual forms (gametocytes) (Kafsack et al. 2014). Some of the observed global CNV effects, in this study, may be as a result of alteration in chromatin structure that interferes with expression of such master regulator genes.

Furthermore, CNVs may affect the expression of genes at a distance through other methods including perturbation of certain functional pathways that result in changes in expression of genes involved in these pathways, interference with long-range interactions among promoters, *cis*-regulatory elements and their transcriptional units (Kleinjan and van Heyningen 2005) and lastly, disruption of interactions that form pockets of repressive and active nuclear compartments that control gene transcription (Fraser and Bickmore 2007).

In conclusion, the gene expression variation associated with CNVs may result in phenotypic diversity in parasite populations that enable the parasite to survive when environmental conditions suddenly change.

Chapter 6

General Discussion

6 General discussion

6.1 Study objectives

The malaria parasite is exposed to constantly changing vector and host environments and yet is able to thrive under these conditions. The parasite has the ability to respond to these changes by possessing multiple variants that have varied responses to environmental pressure in order to facilitate its survival. There is limited understanding of the mechanisms that facilitate the responsive nature of the parasite that enables it to adapt to its environment. Various factors in the parasite have been linked to regulation of genes that interact with the environment *in vitro*, or which are associated with adaptive phenotypes, and therefore may contribute to phenotypic plasticity in the parasite's natural setting. These include genetic variants that are thought to affect parasite phenotype, epigenetic mechanisms thought to be involved in expression of genes involved in immune evasion (Duraisingh et al. 2005, Freitas-Junior et al. 2005, Lopez-Rubio et al. 2007) and erythrocyte invasion (Cortes et al. 2007, Jiang et al. 2010) and the presence of master regulators of expression, e.g., the Apicomplexan *ap2* gene family of DNA-binding transcription factors (Balaji et al. 2005, De Silva et al. 2008). It is possible that the parasite has a large number of mechanisms that underlie its adaptability. An understanding of the factors, networks and interactions that underlie adaptive parasite phenotypes in their natural setting may be useful in development of effective therapeutics and vaccines.

Genetic and phenotypic variation has been widely studied as the basis for understanding malaria parasite biology. In the process, a vast number of SNPs, indels, CNVs and chromosome size polymorphism have been identified. Investigations of the relevance of these variants have been on the rise over the years. First, experiments involving generation of artificial mutants have aided in the understanding of various processes in the parasite including erythrocyte invasion, gametocytogenesis, drug resistance among others (Crosnier et

al. 2011, Ikadai et al. 2013, Triglia et al. 1998). Second, selection experiments have also contributed to identification of variants linked to particular phenotypes (Price et al. 2004, Singh and Rosenthal 2004, Nzila and Mwai 2010). Since genetic variation is the substrate for evolution, plasticity in the parasite genome may well underlie the phenotypic plasticity that enables parasites to thrive in variable environments. A number of CNVs have been shown to underlie adaptive evolution under drug pressure (Nair et al. 2008, Price et al. 2004, Sidhu et al. 2006, Singh and Rosenthal 2004, Dharia et al. 2009), and *in vitro* culture conditions in the laboratory (Nair et al. 2010, Biggs et al. 1989, Shirley et al. 1990, Kemp et al. 1992, Ribacke et al. 2007, Mackinnon et al. 2009). The study of CNVs in natural populations under variable natural selection pressures offers insight into their potential role in adaptation to different environments.

The hypothesis that CNVs underlie malaria parasite adaptation to different environmental conditions in nature was investigated. The specific aims included detection of CNVs in four parasite populations using microarrays, calculation of the CNV frequencies in these populations and tests for evidence of selection on them, confirmation of the detected CNVs using sequencing, and investigation of the impact of the CNVs on gene transcription levels.

6.2 Key findings

6.2.1 CNVs are prevalent in natural populations of *P. falciparum*

In this study, a genome-wide scan of CNVs in 183 field isolates from four populations, with different malaria transmission intensities, was conducted using microarrays. Ninety-eight different CNVs comprising 225 genes, located in the 14 chromosomes of the genome, were detected. The proportion of the genes in the genome targeted by the microarray and found to

be copy number variable was 4.5% (225/5032 genes). This proportion falls within the range of the fractions reported in other *P. falciparum* CNV studies, using microarrays, of 1.5% - 7% (Ribacke et al. 2007, Jiang et al. 2008b, Mackinnon et al. 2009, Dharia et al. 2009, Mok et al. 2011, Kidgell et al. 2006, Carret et al. 2005, Bozdech et al. 2003a). The difference observed in the studies may be due to differences in numbers of isolates studied (2-37 isolates), density of the arrays (6000 probes - 4.8 million probes) and stringency measures of CNV detection. Other studies on different organisms estimated the proportions of genomes with CNVs at 2.1% - 4.6% in the cattle genome (Bickhart et al. 2012, Hou et al. 2011), 3.7% - 12% in the human genome (Redon et al. 2006, Conrad et al. 2010b) and 2% in *Drosophila melanogaster* (Emerson et al. 2008). The length of the CNVs detected in this study ranged between 400bp and 90kb with a median of 6.7 kb. This is comparable to the range of 100bp to 107kb reported in previous studies of *P. falciparum* (Kidgell et al. 2006, Cheeseman et al. 2009).

The CNVs identified consist of amplifications and deletions at a ratio of 1.9:1. A greater fraction of amplifications than deletions has been commonly observed in different studies of *P. falciparum* (Ribacke et al. 2007) and other organisms (Redon et al. 2006). This may be as a result of stronger purifying selection on deletions that may be more deleterious than amplifications, assuming that the rates of formation of deletions and amplifications are similar. It may also reflect technical challenges in detecting deletions. CNVs showing both deletion and amplifications made up 5% (5/98) of the CNVs identified. This has also been observed in other genomes (Hou et al. 2011, Chen et al. 2012, Stranger et al. 2007). If real, these indicate that CNVs exist at a wide range of copy number in populations, and reflect high plasticity in that genomic segment. Alternatively, these may sometimes be an artefact due to the presence of sequence polymorphisms in the population. In this study, data from array probes with known SNPs within the probe regions were excluded from the analysis and that 1bp mismatch

out of 70bp probe sequence would not have a great effect on hybridisation intensity (Bozdech et al. 2003b) thus making this explanation unlikely.

Some of the CNVs reported in this study have been previously identified. Of all the CNVs published so far in *P. falciparum*, approximately 4% (50/1236 genes) were detected in this study. The low concordance with previous studies could be explained by differences in technical aspects, CNV calling algorithms, populations and numbers of isolates studied. It may also imply that there are many more CNVs yet to be discovered.

An interesting CNV found here, previously identified in only laboratory adapted lines and not in field isolates, is the large deletion on chromosome 9 containing 18 consecutive genes. This was identified in 13% (24/183) of the isolates studied of which 67% (14/24) originated from a single population (Sudan). This CNV appears to be real since its deletion is supported by data from probes in a large number of consecutive genes located within the CNV. However, upon analysis of this deletion by sequencing (Chapter 4), the region was observed to be covered by sequence reads in samples that were found to contain the deletion by microarrays. CNV calling using sequence data detected four of the 18 genes in the region as deleted, partially supporting the results by microarray, but not for the entire region. A deletion/amplification, in the case of sequencing, was defined as a region with coverage that was 1.74 fold (0.8 on the $\log_2$ scale) lower/higher than the mean sequence coverage of all the samples. The variability between genes in sequence coverage, inherent in sequencing data, may therefore have obscured the deletions in the other 14 genes. The signal may have been further obscured by the presence of subpopulations of parasites with and without the deletion within the isolates. Deletion of this region has been linked to loss of cytoadherence (Kemp et al. 1992, Biggs et al. 1989) and gametocytogenesis (Day et al. 1993). The parasites bearing this deletion are thought

to have a growth advantage since, so far, they appear to arise only *in vitro*, an environment where they are relieved of the cost that comes with the need for transmission and immune evasion. In the parasite's natural setting inside the host, the occurrence of this deletion in a subpopulation in nature may be beneficial for increased growth rate of the parasite in the short-term, i.e., within its host, but detrimental in the longer term because of its loss of transmissibility. The majority of isolates bearing the deletion were found in Sudan, a population with low transmission intensity and immunity. Reduced selection pressure from host immunity compared to the other populations may enable existence of parasites that are less virulent (Gandon et al. 2001), as characterized by loss of cytoadherence. Alternatively, increased selection pressure to prolong the infection through the long mosquito-free dry season in Sudan may bring this deletion to the fore.

Fifteen functional gene groups were found to be significantly enriched in the list of CNV genes detected. Some of these groups are involved in environmental responsiveness processes. The two most enriched groups of genes were those coding for Maurer's cleft proteins and the PHIST family of genes. The Maurer's cleft proteins consist of proteins that reside in the Maurer cleft, a parasite derived organelle in the iRBC cytoplasm, or exported to the iRBC membrane through the Maurer's cleft (Lanzer et al. 2006). PHIST family of genes are also thought to be exported to the iRBC surface (Sargeant et al. 2006). These two gene groups are exposed to the host immune responses and alteration of their copy numbers may be a mechanism for evasion of host immunity. Gain in gene copy number would increase the chances of random mutations that increase genetic diversity allowing for immune escape. Other enriched groups include those involved in glycolysis, an important mechanism for production of energy using glucose obtained from the host to support parasite multiplication and growth. Enrichment was observed in genes involved in antioxidative activity that relieves

the parasite of oxidative stress from processes including digestion of haemoglobin, production of energy in the mitochondria and host immune responses (Jortzik and Becker 2012).

6.2.2 CNVs may be under selection

I explored for signs of CNVs under selection by assessing the differentiation in CNV frequencies between populations using Weir and Cockerham's F_{ST} estimates. Twenty percent of the CNVs (19/ 95) were observed to show high differentiation between populations (F_{ST} >0.20) and hence may show signs of directional selection that result from population-specific pressures. The high differentiation may also be as a result of bottlenecks in the population, difference in the effective population sizes leading to variation in genetic drift between populations and migration of individuals. The CNVs that are show high differentiation between populations contain genes coding for proteins involved in gametocytogenesis, transcription regulation, DNA repair and proteins that interact with host immunity. Some of the differences in the populations that may contribute to differential selection include host genetics, host immunity, vector population densities and genetics, transmission intensities, drug use and co-infections. So far in *P. falciparum*, the only natural selection forces on CNVs that have been identified have been anti-malarial drugs (Nair et al. 2008) and host immunity (Ahouidi et al. 2010).

An example of a CNV with high differentiation between three population pairs (F_{ST} >0.35), the two Kilifi populations compared to Sudan and Kilifi-Pre compared to Kisumu, was cnv12_413 on chromosome 12 containing a single gene PF3D7_1248600. This CNV occurred at a higher frequency in Sudan and Kisumu than in the Kilifi populations (Chapter 3). The function of the protein encoded by the gene has not been experimentally tested in the laboratory. It is inferred to be involved in attachment of glycosylphosphatidylinositol (GPI)

anchor to proteins, based on its sequence. The GPI anchor is useful for attachment of proteins to the extracellular surface of the membranes (Gilson et al. 2006). The gene has also been identified to belong to sexual stage Gene Ontology (GO) biological process of response to external stimuli (Young et al. 2005). The sexual stage GO is found at http://chemlims.com/OPI/MServlet.ChemInfo?module=Go&act=find&act2=viewRecord&GO_Name__LikeSBE_0=%3D&GO_Name__TextSBE_1=GO:0009628&DataSet=21 . Interestingly, in Chapter 5, this deletion was found to significantly affect the levels of expression of 170 genes, both positively and negatively. Thus the existence of copy number variation at this locus could conceivably be the consequence of selection pressure on its regulatory effect on gametocyte production, a key component of parasite fitness.

Most of the CNVs showed low to moderate differentiation between populations (F_{ST}< 0.20). The presence of these CNVs could be as a result of neutral evolutionary processes, e.g., drift and population separation. These CNVs may be of no immediate relevance to the parasite, but might be considered to be on standby for future changes in environment that would render them advantageous.

6.2.3 Poor overlap between CNVs identified by microarrays and sequencing

A second technology - whole genome sequencing - that has been widely used for CNV studies in human genome and a few *P. falciparum* isolates was adopted to validate the CNVs identified using microarrays. Sequence data of 18 of the 183 isolates were generated and CNV calling performed. The CGH data of the 18 isolates were also reanalysed using the same methods for sequence data with adjustments to certain parameters (fold change, number of probes within CNV, generally at lower stringency than for CGH) in calling of CNVs in order to compare the results to those from sequence data. Three percent (21/828) of the CNV genes

detected using microarrays (*GADA* software) were confirmed by sequencing (*cn.MOPs* software).

The low concordance in results from the two different technologies could be as a result of technical and analytic differences between them. First, the read count data used for CNV detection in sequence data is affected by inefficient alignment of reads to the reference genome due to presence of sequencing errors, repeat regions, sequence polymorphisms and short read length. Second, the type of data analysed differ in the two technologies. CGH data are $\log_2$ intensity ratios (difference in intensity between test and reference genome) and CNVs were defined as regions showing greater or less than absolute $\log_2$ intensity ratio of 0.8. By contrast, sequencing data are normalised read counts per isolate and CNVs were defined as regions with fold-differences corresponding to $\log_2$ value of 0.8 from the mean read count. Third, microarrays have a cap on the maximum signal intensity that can be detected whereas this is not the case for read counts in sequencing. Fourth, microarray hybridization may be affected by the presence of sequence polymorphisms in natural populations that were not detected and hence not ruled out prior to analysis thus resulting in false positive calls. Given these differences between the two technologies, one cannot replace the other in CNV detection. Instead, they can be used to complement each other for comprehensive mapping of CNVs.

6.2.4 CNVs affect the levels of transcription of genes within and outside CNV boundaries

An integrated analysis using CGH data and transcriptome data of 72 isolates was performed to assess variation in gene expression that could be attributed to CNVs. Significant positive and negative correlations between genetic content (CGH) and expression (transcriptome) data for

genes identified to be located within CNVs by CGH. These indicate direct dosage effects of CNVs on gene expression. Negative correlations are unexpected and may be attributed to the involvement of other gene expression regulatory mechanisms. Amplified genes showed significant associations with expression more often than deleted genes suggesting that amplifications have greater impact on expression and parasite phenotype than deletions.

The effect of CNVs on the expression of the genes in the rest of the genome (i.e., outside the CNV boundaries) was assessed using linear regression models. It was surprising to find that most of the variation in gene expression significantly associated with CNVs were of genes not located in the CNVs. Significant effects on expression were observed in 5.7% of the relationships analysed (95 CNVs and expression of 4797 genes). The genes showing altered expression associated with a CNV were located in close proximity, at a distance or on different chromosomes from the CNVs. However, from these analyses, it is impossible to confidently conclude that the CNVs confer causal effects on CNV expression or are linked to other causal variants.

6.3 Future directions

This study has identified CNVs as a major source of variation in naturally occurring parasites. It has also revealed that CNVs affect phenotype, using gene expression as a proxy to phenotype, and also appear to be under natural selection. This sets the basis for understanding parasite adaptation in response to its natural environment (host immunity, vectors, host genetics among others) and control programmes, e.g., vaccines and therapeutics (Mackinnon and Marsh 2010, Gandon et al. 2001). The findings provide a lead in to better understanding of new regulatory mechanisms that might be able to be targeted by drugs or vaccines. For example the CNV on chromosome 12 (cnv12_413) that was associated with expression of 170

genes is a high priority candidate for future study because it may be one of the master regulators of gene expression. The candidate CNVs can in future be studied using transfection technology, e.g., the introduction of extra gene copy number or a deletion to better understand the mechanism through which they contribute to phenotypic variation.

However, most of the CNVs identified using microarrays could not be validated by sequencing. To improve this situation, better CNV detection tools for sequence data that are suitable for the *P. falciparum* genome are needed. Second, increased sequencing read length would be useful in overcoming problems associated with short reads mapping in repeat regions. Third, generation of a representative 'genome' of all the naturally occurring isolates that could be used to establish unique and conserved regions would improve accuracy of read depth calculation and hence CNV calling. Lastly, development of a CNV detection tool that could be applied to both sequence and CGH data would enable fair comparison between the technologies and pooling of information from different sources in order to improve the accuracy and power to detect novel CNVs.

7 References

Abyzov, A., A. E. Urban, M. Snyder & M. Gerstein (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res,* 21, 974-84.

Ahouidi, A. D., A. K. Bei, D. E. Neafsey, O. Sarr, S. Volkman, D. Milner, J. Cox-Singh, M. U. Ferreira, O. Ndir, Z. Premji, S. Mboup & M. T. Duraisingh (2010) Population genetic analysis of large sequence polymorphisms in *Plasmodium falciparum* blood-stage antigens. *Infect Genet Evol,* 10, 200-6.

Akey, J. M., G. Zhang, K. Zhang, L. Jin & M. D. Shriver (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Res,* 12, 1805-14.

Albertini, A. M., M. Hofer, M. P. Calos & J. H. Miller (1982) On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. *Cell,* 29, 319-28.

Alkan, C., B. P. Coe & E. E. Eichler (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet,* 12, 363-76.

Alkan, C., J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. O. Kitzman, C. Baker, M. Malig, O. Mutlu, S. C. Sahinalp, R. A. Gibbs & E. E. Eichler (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet,* 41, 1061-7.

Amambua-Ngwa, A., K. K. Tetteh, M. Manske, N. Gomez-Escobar, L. B. Stewart, M. E. Deerhake, I. H. Cheeseman, C. I. Newbold, A. A. Holder, E. Knuepfer, O. Janha, M. Jallow, S. Campino, B. Macinnis, D. P. Kwiatkowski & D. J. Conway (2012) Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genet,* 8, e1002992.

Amin, A. A., D. Zurovac, B. B. Kangwana, J. Greenfield, D. N. Otieno, W. S. Akhwale & R. W. Snow (2007) The challenges of changing national malaria drug policy to artemisinin-based combinations in Kenya. *Malar J,* 6, 72.

Amunugama, R. & R. Fishel (2012) Homologous recombination in eukaryotes. *Prog Mol Biol Transl Sci,* 110, 155-206.

Anderson, T. J., S. Nair, D. Sudimack, J. T. Williams, M. Mayxay, P. N. Newton, J. P. Guthmann, F. M. Smithuis, T. H. Tran, I. V. van den Broek, N. J. White & F. Nosten (2005) Geographical distribution of selected and putatively neutral SNPs in Southeast Asian malaria parasites. *Mol Biol Evol,* 22, 2362-74.

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

Angstadt, A. Y., A. Berg, J. Zhu, P. Miller, T. J. Hartman, S. M. Lesko, J. E. Muscat, P. Lazarus & C. J. Gallagher (2013) The effect of copy number variation in the phase II detoxification genes UGT2B17 and UGT2B28 on colorectal cancer risk. *Cancer,* 119, 2477-85.

Ariey, F., B. Witkowski, C. Amaratunga, J. Beghain, A. C. Langlois, N. Khim, S. Kim, V. Duru, C. Bouchier, L. Ma, P. Lim, R. Leang, S. Duong, S. Sreng, S. Suon, C. M. Chuor, D. M. Bout, S. Menard, W. O. Rogers, B. Genton, T. Fandeur, O. Miotto, P. Ringwald, J. Le Bras, A. Berry, J. C. Barale, R. M. Fairhurst, F. Benoit-Vical, O. Mercereau-Puijalon & D. Menard (2014) A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature,* 505, 50-5.

Augustijn, K. D., R. Kleemann, J. Thompson, T. Kooistra, C. E. Crawford, S. E. Reece, A. Pain, A. H. Siebum, C. J. Janse & A. P. Waters (2007) Functional characterization of the *Plasmodium falciparum* and P. berghei homologues of macrophage migration inhibitory factor. *Infect Immun,* 75, 1116-28.

Balaji, S., M. M. Babu, L. M. Iyer & L. Aravind (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res,* 33, 3994-4006.

Barnes, K. I., D. N. Durrheim, F. Little, A. Jackson, U. Mehta, E. Allen, S. S. Dlamini, J. Tsoka, B. Bredenkamp, D. J. Mthembu, N. J. White & B. L. Sharp (2005) Effect of artemether-lumefantrine policy and improved vector control on malaria burden in KwaZulu-Natal, South Africa. *PLoS Med,* 2, e330.

Beaumont, M. A. & D. J. Balding (2004) Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol,* 13, 969-80.

Bhattacharyya, M. K., S. Bhattacharyya nee Deb, B. Jayabalasingham & N. Kumar (2005) Characterization of kinetics of DNA strand-exchange and ATP hydrolysis activities of recombinant PfRad51, a *Plasmodium falciparum* recombinase. *Mol Biochem Parasitol,* 139, 33-9.

Bickhart, D. M., Y. Hou, S. G. Schroeder, C. Alkan, M. F. Cardone, L. K. Matukumalli, J. Song, R. D. Schnabel, M. Ventura, J. F. Taylor, J. F. Garcia, C. P. Van Tassell, T. S. Sonstegard, E. E. Eichler & G. E. Liu (2012) Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res,* 22, 778-90.

Biggs, B. A., D. J. Kemp & G. V. Brown (1989) Subtelomeric chromosome deletions in field isolates of *Plasmodium falciparum* and their relationship to loss of cytoadherence in vitro. *Proc Natl Acad Sci U S A,* 86, 2428-32.

Bodega, B., G. D. Ramirez, F. Grasser, S. Cheli, S. Brunelli, M. Mora, R. Meneveri, A. Marozzi, S. Mueller, E. Battaglioli & E. Ginelli (2009) Remodeling of the chromatin structure of the facioscapulohumeral muscular dystrophy (FSHD) locus and upregulation of FSHD-related gene 1 (FRG1) expression during human myogenic differentiation. *BMC Biol,* 7, 41.

Boeva, V., T. Popova, K. Bleakley, P. Chiche, J. Cappo, G. Schleiermacher, I. Janoueix-Lerosey, O. Delattre & E. Barillot (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics,* 28, 423-5.

Bozdech, Z., M. Llinas, B. L. Pulliam, E. D. Wong, J. Zhu & J. L. DeRisi (2003a) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum. PLoS Biol,* 1, E5.

Bozdech, Z., S. Mok & A. P. Gupta (2013) DNA microarray-based genome-wide analyses of Plasmodium parasites. *Methods Mol Biol,* 923, 189-211.

Bozdech, Z., J. Zhu, M. P. Joachimiak, F. E. Cohen, B. Pulliam & J. L. DeRisi (2003b) Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biol,* 4, R9.

Bragg, L. M., G. Stone, M. K. Butler, P. Hugenholtz & G. W. Tyson (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput Biol,* 9, e1003031.

Branzei, D. & M. Foiani (2008) Regulation of DNA repair throughout the cell cycle. *Nat Rev Mol Cell Biol,* 9, 297-308.

Bull, P. C., B. S. Lowe, M. Kortok, C. S. Molyneux, C. I. Newbold & K. Marsh (1998) Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nat Med,* 4, 358-60.

Bullen, H. E., C. J. Tonkin, R. A. O'Donnell, W. H. Tham, A. T. Papenfuss, S. Gould, A. F. Cowman, B. S. Crabb & P. R. Gilson (2009) A novel family of Apicomplexan glideosome-associated proteins with an inner membrane-anchoring role. *J Biol Chem,* 284, 25353-63.

Cahill, D., B. Connor & J. P. Carney (2006) Mechanisms of eukaryotic DNA double strand break repair. *Front Biosci,* 11, 1958-76.

Cai, W., H. Aburatani, V. P. Stanton, Jr., D. E. Housman, Y. K. Wang & D. C. Schwartz (1995) Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proc Natl Acad Sci U S A,* 92, 5164-8.

Carret, C. K., P. Horrocks, B. Konfortov, E. Winzeler, M. Qureshi, C. Newbold & A. Ivens (2005) Microarray-based comparative genomic analyses of the human malaria parasite *Plasmodium falciparum* using Affymetrix arrays. *Mol Biochem Parasitol,* 144, 177-86.

Carvalho, C. M., D. Pehlivan, M. B. Ramocki, P. Fang, B. Alleva, L. M. Franco, J. W. Belmont, P. J. Hastings & J. R. Lupski (2013) Replicative mechanisms for CNV formation are error prone. *Nat Genet,* 45, 1319-26.

Carvalho, C. M., F. Zhang, P. Liu, A. Patel, T. Sahoo, C. A. Bacino, C. Shaw, S. Peacock, A. Pursley, Y. J. Tavyev, M. B. Ramocki, M. Nawara, E. Obersztyn, A. M. Vianna-Morgante, P. Stankiewicz, H. Y. Zoghbi, S. W. Cheung & J. R. Lupski (2009) Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Hum Mol Genet,* 18, 2188-203.

Chan, J. A., F. J. Fowkes & J. G. Beeson (2014) Surface antigens of *Plasmodium falciparum*-infected erythrocytes as immune targets and malaria vaccine candidates. *Cell Mol Life Sci,* 71, 3633-57.

Chavchich, M., L. Gerena, J. Peters, N. Chen, Q. Cheng & D. E. Kyle (2010) Role of pfmdr1 amplification and expression in induction of resistance to artemisinin derivatives in *Plasmodium falciparum*. *Antimicrob Agents Chemother,* 54, 2455-64.

Cheeseman, I. H., N. Gomez-Escobar, C. K. Carret, A. Ivens, L. B. Stewart, K. K. Tetteh & D. J. Conway (2009) Gene copy number variation throughout the *Plasmodium falciparum* genome. *BMC Genomics,* 10, 353.

Chen, C., R. Qiao, R. Wei, Y. Guo, H. Ai, J. Ma, J. Ren & L. Huang (2012) A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. *BMC Genomics,* 13, 733.

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

Chiang, D. Y., G. Getz, D. B. Jaffe, M. J. O'Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, M. Meyerson & E. S. Lander (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods,* 6, 99-103.

Choveaux, D. L., J. M. Przyborski & J. P. Goldring (2012) A *Plasmodium falciparum* copper-binding membrane protein with copper transport motifs. *Malar J,* 11, 397.

Conrad, D. F., C. Bird, B. Blackburne, S. Lindsay, L. Mamanova, C. Lee, D. J. Turner & M. E. Hurles (2010a) Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet,* 42, 385-91.

Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer & M. E. Hurles (2010b) Origins and functional impact of copy number variation in the human genome. *Nature,* 464, 704-12.

Corbel, V., M. Akogbeto, G. B. Damien, A. Djenontin, F. Chandre, C. Rogier, N. Moiroux, J. Chabi, B. Banganna, G. G. Padonou & M. C. Henry (2012) Combination of malaria

vector control interventions in pyrethroid resistance area in Benin: a cluster randomised controlled trial. *Lancet Infect Dis,* 12, 617-26.

Corcoran, L. M., K. P. Forsyth, A. E. Bianco, G. V. Brown & D. J. Kemp (1986) Chromosome size polymorphisms in *Plasmodium falciparum* can involve deletions and are frequent in natural parasite populations. *Cell,* 44, 87-95.

Cortes, A., C. Carret, O. Kaneko, B. Y. Yim Lim, A. Ivens & A. A. Holder (2007) Epigenetic silencing of *Plasmodium falciparum* genes linked to erythrocyte invasion. *PLoS Pathog,* 3, e107.

Craddock, N., M. E. Hurles, N. Cardin, R. D. Pearson, V. Plagnol, S. Robson, D. Vukcevic, C. Barnes, D. F. Conrad, E. Giannoulatou, C. Holmes, J. L. Marchini, K. Stirrups, M. D. Tobin, L. V. Wain, C. Yau, J. Aerts, T. Ahmad, T. D. Andrews, H. Arbury, A. Attwood, A. Auton, S. G. Ball, A. J. Balmforth, J. C. Barrett, I. Barroso, A. Barton, A. J. Bennett, S. Bhaskar, K. Blaszczyk, J. Bowes, O. J. Brand, P. S. Braund, F. Bredin, G. Breen, M. J. Brown, I. N. Bruce, J. Bull, O. S. Burren, J. Burton, J. Byrnes, S. Caesar, C. M. Clee, A. J. Coffey, J. M. Connell, J. D. Cooper, A. F. Dominiczak, K. Downes, H. E. Drummond, D. Dudakia, A. Dunham, B. Ebbs, D. Eccles, S. Edkins, C. Edwards, A. Elliot, P. Emery, D. M. Evans, G. Evans, S. Eyre, A. Farmer, I. N. Ferrier, L. Feuk, T. Fitzgerald, E. Flynn, A. Forbes, L. Forty, J. A. Franklyn, R. M. Freathy, P. Gibbs, P. Gilbert, O. Gokumen, K. Gordon-Smith, E. Gray, E. Green, C. J. Groves, D. Grozeva, R. Gwilliam, A. Hall, N. Hammond, M. Hardy, P. Harrison, N. Hassanali, H. Hebaishi, S. Hines, A. Hinks, G. A. Hitman, L. Hocking, E. Howard, P. Howard, J. M. Howson, D. Hughes, S. Hunt, J. D. Isaacs, M. Jain, D. P. Jewell, T. Johnson, J. D. Jolley, I. R. Jones, L. A. Jones, et al. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature,* 464, 713-20.

Crompton, P. D., M. A. Kayala, B. Traore, K. Kayentao, A. Ongoiba, G. E. Weiss, D. M. Molina, C. R. Burk, M. Waisberg, A. Jasinskas, X. Tan, S. Doumbo, D. Doumtabe, Y. Kone, D. L. Narum, X. Liang, O. K. Doumbo, L. H. Miller, D. L. Doolan, P. Baldi, P. L. Felgner & S. K. Pierce (2010) A prospective analysis of the Ab response to *Plasmodium falciparum* before and after a malaria season by protein microarray. *Proc Natl Acad Sci U S A,* 107, 6958-63.

Crosnier, C., L. Y. Bustamante, S. J. Bartholdson, A. K. Bei, M. Theron, M. Uchikawa, S. Mboup, O. Ndir, D. P. Kwiatkowski, M. T. Duraisingh, J. C. Rayner & G. J. Wright (2011) Basigin is a receptor essential for erythrocyte invasion by *Plasmodium falciparum. Nature,* 480, 534-7.

Cui, L., Q. Fan, L. Cui & J. Miao (2008) Histone lysine methyltransferases and demethylases in *Plasmodium falciparum. Int J Parasitol,* 38, 1083-97.

Cui, L. & J. Miao (2010) Chromatin-mediated epigenetic regulation in the malaria parasite *Plasmodium falciparum. Eukaryot Cell,* 9, 1138-49.

Cui, P., S. Zhang, F. Ding, S. Ali & L. Xiong (2014) Dynamic regulation of genome-wide pre-mRNA splicing and stress tolerance by the Sm-like protein LSm5 in Arabidopsis. *Genome Biol,* 15, R1.

Day, K. P., F. Karamalis, J. Thompson, D. A. Barnes, C. Peterson, H. Brown, G. V. Brown & D. J. Kemp (1993) Genes necessary for expression of a virulence determinant and for transmission of *Plasmodium falciparum* are located on a 0.3-megabase region of chromosome 9. *Proc Natl Acad Sci U S A,* 90, 8292-6.

De Silva, E. K., A. R. Gehrke, K. Olszewski, I. Leon, J. S. Chahal, M. L. Bulyk & M. Llinas (2008) Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc Natl Acad Sci U S A,* 105, 8393-8.

Dessens, J. T., A. L. Beetsma, G. Dimopoulos, K. Wengelnik, A. Crisanti, F. C. Kafatos & R. E. Sinden (1999) CTRP is essential for mosquito infection by malaria ookinetes. *Embo j,* 18, 6221-7.

Dharia, N. V., A. B. Sidhu, M. B. Cassera, S. J. Westenberger, S. E. Bopp, R. T. Eastman, D. Plouffe, S. Batalov, D. J. Park, S. K. Volkman, D. F. Wirth, Y. Zhou, D. A. Fidock & E. A. Winzeler (2009) Use of high-density tiling microarrays to identify mutations globally and elucidate mechanisms of drug resistance in *Plasmodium falciparum. Genome Biol,* 10, R21.

Dittwald, P., T. Gambin, P. Szafranski, J. Li, S. Amato, M. Y. Divon, L. X. Rodriguez Rojas, L. E. Elton, D. A. Scott, C. P. Schaaf, W. Torres-Martinez, A. K. Stevens, J. A. Rosenfeld, S. Agadi, D. Francis, S. H. Kang, A. Breman, S. R. Lalani, C. A. Bacino, W. Bi, A. Milosavljevic, A. L. Beaudet, A. Patel, C. A. Shaw, J. R. Lupski, A. Gambin, S. W. Cheung & P. Stankiewicz (2013) NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Res,* 23, 1395-409.

Dixon, M. W., S. Kenny, P. J. McMillan, E. Hanssen, K. R. Trenholme, D. L. Gardiner & L. Tilley (2011) Genetic ablation of a Maurer's cleft protein prevents assembly of the *Plasmodium falciparum* virulence complex. *Mol Microbiol,* 81, 982-93.

Dondorp, A. M., F. Nosten, P. Yi, D. Das, A. P. Phyo, J. Tarning, K. M. Lwin, F. Ariey, W. Hanpithakpong, S. J. Lee, P. Ringwald, K. Silamut, M. Imwong, K. Chotivanich, P. Lim, T. Herdman, S. S. An, S. Yeung, P. Singhasivanon, N. P. Day, N. Lindegardh, D. Socheat & N. J. White (2009) Artemisinin resistance in *Plasmodium falciparum* malaria. *N Engl J Med,* 361, 455-67.

Dubrovina, A. S., K. V. Kiselev & Y. N. Zhuravlev (2013) The role of canonical and noncanonical pre-mRNA splicing in plant stress responses. *Biomed Res Int,* 2013, 264314.

Duffy, M. F., S. A. Selvarajah, G. A. Josling & M. Petter (2012) The role of chromatin in Plasmodium gene expression. *Cell Microbiol,* 14, 819-28.

Duraisingh, M. T., T. S. Voss, A. J. Marty, M. F. Duffy, R. T. Good, J. K. Thompson, L. H. Freitas-Junior, A. Scherf, B. S. Crabb & A. F. Cowman (2005) Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in *Plasmodium falciparum. Cell,* 121, 13-24.

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

Emerson, J. J., M. Cardoso-Moreira, J. O. Borevitz & M. Long (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science,* 320, 1629-31.

Farrugia, A., N. Shea, S. Knowles, R. Holdsworth, H. Piouronowski, D. Portbury & A. Romeo (1993) Cryopreservation of red blood cells: effect of freezing on red cell quality and residual lymphocyte immunogenicity. *J Clin Pathol,* 46, 742-5.

Fattah, F., E. H. Lee, N. Weisensel, Y. Wang, N. Lichter & E. A. Hendrickson (2010) Ku regulates the non-homologous end joining pathway choice of DNA double-strand break repair in human somatic cells. *PLoS Genet,* 6, e1000855.

Fidock, D. A., T. Nomura, A. K. Talley, R. A. Cooper, S. M. Dzekunov, M. T. Ferdig, L. M. Ursos, A. B. Sidhu, B. Naude, K. W. Deitsch, X. Z. Su, J. C. Wootton, P. D. Roepe & T. E. Wellems (2000) Mutations in the *P. falciparum* digestive vacuole transmembrane

protein PfCRT and evidence for their role in chloroquine resistance. *Mol Cell,* 6, 861-71.

Fraser, P. & W. Bickmore (2007) Nuclear organization of the genome and the potential for gene regulation. *Nature,* 447, 413-7.

Freitas-Junior, L. H., R. Hernandez-Rivas, S. A. Ralph, D. Montiel-Condado, O. K. Ruvalcaba-Salazar, A. P. Rojas-Meza, L. Mancio-Silva, R. J. Leal-Silvestre, A. M. Gontijo, S. Shorte & A. Scherf (2005) Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell,* 121, 25-36.

Gadalla, N. B., I. Adam, S. E. Elzaki, S. Bashir, I. Mukhtar, M. Oguike, A. Gadalla, F. Mansour, D. Warhurst, B. B. El-Sayed & C. J. Sutherland (2011) Increased pfmdr1 Copy Number and Sequence Polymorphisms in *Plasmodium falciparum* Isolates from Sudanese Malaria Patients Treated with Artemether-Lumefantrine. *Antimicrob Agents Chemother,* 55, 5408-11.

Gandon, S., M. J. Mackinnon, S. Nee & A. F. Read (2001) Imperfect vaccines and the evolution of pathogen virulence. *Nature,* 414, 751-6.

Gardiner, D. L., M. W. Dixon, T. Spielmann, T. S. Skinner-Adams, P. L. Hawthorne, M. R. Ortega, D. J. Kemp & K. R. Trenholme (2005) Implication of a *Plasmodium falciparum* gene in the switch between asexual reproduction and gametocytogenesis. *Mol Biochem Parasitol,* 140, 153-60.

Gardner, M. J., N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser & B. Barrell (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum. Nature,* 419, 498-511.

Gilson, P. R., T. Nebl, D. Vukcevic, R. L. Moritz, T. Sargeant, T. P. Speed, L. Schofield & B. S. Crabb (2006) Identification and stoichiometry of glycosylphosphatidylinositol-anchored membrane proteins of the human malaria parasite *Plasmodium falciparum. Mol Cell Proteomics,* 5, 1286-99.

Glenister, F. K., K. M. Fernandez, L. M. Kats, E. Hanssen, N. Mohandas, R. L. Coppel & B. M. Cooke (2009) Functional alteration of red blood cells by a megadalton protein of *Plasmodium falciparum. Blood,* 113, 919-28.

Gonzales, J. M., J. J. Patel, N. Ponmee, L. Jiang, A. Tan, S. P. Maher, S. Wuchty, P. K. Rathod & M. T. Ferdig (2008) Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. *PLoS Biol,* 6, e238.

Goudet, J. (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes,* 184-186.

Gu, W., F. Zhang & J. R. Lupski (2008) Mechanisms for human genomic rearrangements. *Pathogenetics,* 1, 4.

Hastings, P. J., J. R. Lupski, S. M. Rosenberg & G. Ira (2009) Mechanisms of change in gene copy number. *Nat Rev Genet,* 10, 551-64.

Henrichsen, C. N., E. Chaignat & A. Reymond (2009) Copy number variants, diseases and gene expression. *Hum Mol Genet,* 18, R1-8.

Hermetz, K. E., S. Newman, K. N. Conneely, C. L. Martin, B. C. Ballif, L. G. Shaffer, J. D. Cody & M. K. Rudd (2014) Large inverted duplications in the human genome form via a fold-back mechanism. *PLoS Genet,* 10, e1004139.

Hou, Y., G. E. Liu, D. M. Bickhart, M. F. Cardone, K. Wang, E. S. Kim, L. K. Matukumalli, M. Ventura, J. Song, P. M. VanRaden, T. S. Sonstegard & C. P. Van Tassell (2011) Genomic characteristics of cattle copy number variations. *BMC Genomics,* 12, 127.

Hou, Y., G. E. Liu, D. M. Bickhart, L. K. Matukumalli, C. Li, J. Song, L. C. Gasbarre, C. P. Van Tassell & T. S. Sonstegard (2012) Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. *Funct Integr Genomics,* 12, 81-92.

Hu, G., M. Llinas, J. Li, P. R. Preiser & Z. Bozdech (2007) Selection of long oligonucleotides for gene expression microarrays using weighted rank-sum strategy. *BMC Bioinformatics,* 8, 350.

Ikadai, H., K. Shaw Saliba, S. M. Kanzok, K. J. McLean, T. Q. Tanaka, J. Cao, K. C. Williamson & M. Jacobs-Lorena (2013) Transposon mutagenesis identifies genes essential for *Plasmodium falciparum* gametocytogenesis. *Proc Natl Acad Sci U S A,* 110, E1676-84.

Iriko, H., O. Kaneko, H. Otsuki, T. Tsuboi, X. Z. Su, K. Tanabe & M. Torii (2008) Diversity and evolution of the rhoph1/clag multigene family of *Plasmodium falciparum. Mol Biochem Parasitol,* 158, 11-21.

Ivakhno, S., T. Royce, A. J. Cox, D. J. Evers, R. K. Cheetham & S. Tavare (2010) CNAseg--a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics,* 26, 3051-8.

Jeffares, D. C., A. Pain, A. Berry, A. V. Cox, J. Stalker, C. E. Ingle, A. Thomas, M. A. Quail, K. Siebenthall, A. C. Uhlemann, S. Kyes, S. Krishna, C. Newbold, E. T. Dermitzakis & M. Berriman (2007) Genome variation and evolution of the malaria parasite *Plasmodium falciparum. Nat Genet,* 39, 120-5.

Jiang, H., N. Li, V. Gopalan, M. Zilversmit, V. S., N. V., L. J., M. J., H. K., H. B., Y. M., S. R., M. G., A. P., W. TE. & S. XZ. (2011) High recombination rates and hotspots in a *Plasmodium falciparum* genetic cross. *Genome Biol,* 12.

Jiang, H., J. J. Patel, M. Yi, J. Mu, J. Ding, R. Stephens, R. A. Cooper, M. T. Ferdig & X. Z. Su (2008a) Genome-wide compensatory changes accompany drug- selected mutations in the *Plasmodium falciparum* crt gene. *PLoS One,* 3, e2484.

Jiang, H., M. Yi, J. Mu, L. Zhang, A. Ivens, L. J. Klimczak, Y. Huyen, R. M. Stephens & X. Z. Su (2008b) Detection of genome-wide polymorphisms in the AT-rich *Plasmodium falciparum* genome using a high-density microarray. *BMC Genomics,* 9, 398.

Jiang, L., M. J. Lopez-Barragan, H. Jiang, J. Mu, D. Gaur, K. Zhao, G. Felsenfeld & L. H. Miller (2010) Epigenetic control of the variable expression of a *Plasmodium falciparum* receptor protein for erythrocyte invasion. *Proc Natl Acad Sci U S A,* 107, 2224-9.

Jing, J., Z. Lai, C. Aston, J. Lin, D. J. Carucci, M. J. Gardner, B. Mishra, T. S. Anantharaman, H. Tettelin, L. M. Cummings, S. L. Hoffman, J. C. Venter & D. C. Schwartz (1999) Optical mapping of *Plasmodium falciparum* chromosome 2. *Genome Res,* 9, 175-81.

Jortzik, E. & K. Becker (2012) Thioredoxin and glutathione systems in *Plasmodium falciparum. Int J Med Microbiol,* 302, 187-94.

Kafsack, B. F., N. Rovira-Graells, T. G. Clark, C. Bancells, V. M. Crowley, S. G. Campino, A. E. Williams, L. G. Drought, D. P. Kwiatkowski, D. A. Baker, A. Cortes & M. Llinas (2014) A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature,* 507, 248-52.

Kaneko, O., B. Y. Yim Lim, H. Iriko, I. T. Ling, H. Otsuki, M. Grainger, T. Tsuboi, J. H. Adams, D. Mattei, A. A. Holder & M. Torii (2005) Apical expression of three RhopH1/Clag proteins as components of the *Plasmodium falciparum* RhopH complex. *Mol Biochem Parasitol,* 143, 20-8.

Karakoc, E., C. Alkan, B. J. O'Roak, M. Y. Dennis, L. Vives, K. Mark, M. J. Rieder, D. A. Nickerson & E. E. Eichler (2012) Detection of structural variants and indels within exome data. *Nat Methods,* 9, 176-8.

Kehr, S., N. Sturm, S. Rahlfs, J. M. Przyborski & K. Becker (2010) Compartmentation of redox metabolism in malaria parasites. *PLoS Pathog,* 6, e1001242.

Kemp, D. J., L. M. Corcoran, R. L. Coppel, H. D. Stahl, A. E. Bianco, G. V. Brown & R. F. Anders (1985) Size variation in chromosomes from independent cultured isolates of *Plasmodium falciparum. Nature,* 315, 347-50.

Kemp, D. J., J. Thompson, D. A. Barnes, T. Triglia, F. Karamalis, C. Petersen, G. V. Brown & K. P. Day (1992) A chromosome 9 deletion in *Plasmodium falciparum* results in loss of cytoadherence. *Mem Inst Oswaldo Cruz,* 87 Suppl 3, 85-9.

Kidgell, C., S. K. Volkman, J. Daily, J. O. Borevitz, D. Plouffe, Y. Zhou, J. R. Johnson, K. Le Roch, O. Sarr, O. Ndir, S. Mboup, S. Batalov, D. F. Wirth & E. A. Winzeler (2006) A systematic map of genetic variation in *Plasmodium falciparum. PLoS Pathog,* 2, e57.

Kirkman, L. A., E. A. Lawrence & K. W. Deitsch (2014) Malaria parasites utilize both homologous recombination and alternative end joining pathways to maintain genome integrity. *Nucleic Acids Res,* 42, 370-9.

Kirov, G., A. J. Pocklington, P. Holmans, D. Ivanov, M. Ikeda, D. Ruderfer, J. Moran, K. Chambert, D. Toncheva, L. Georgieva, D. Grozeva, M. Fjodorova, R. Wollerton, E. Rees, I. Nikolov, L. N. van de Lagemaat, A. Bayes, E. Fernandez, P. I. Olason, Y. Bottcher, N. H. Komiyama, M. O. Collins, J. Choudhary, K. Stefansson, H. Stefansson, S. G. Grant, S. Purcell, P. Sklar, M. C. O'Donovan & M. J. Owen (2012) De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry,* 17, 142-53.

Klambauer, G., K. Schwarzbauer, A. Mayr, D. A. Clevert, A. Mitterecker, U. Bodenhofer & S. Hochreiter (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res,* 40, e69.

Kleinjan, D. A. & V. van Heyningen (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet,* 76, 8-32.

Konate, L., J. Zwetyenga, C. Rogier, E. Bischoff, D. Fontenille, A. Tall, A. Spiegel, J. F. Trape & O. Mercereau-Puijalon (1999) Variation of *Plasmodium falciparum* msp1 block 2 and msp2 allele prevalence and of infection complexity in two neighbouring Senegalese villages with different transmission conditions. *Trans R Soc Trop Med Hyg,* 93 Suppl 1, 21-8.

Lai, Z., J. Jing, C. Aston, V. Clarke, J. Apodaca, E. T. Dimalanta, D. J. Carucci, M. J. Gardner, B. Mishra, T. S. Anantharaman, S. Paxia, S. L. Hoffman, J. Craig Venter, E. J. Huff & D. C. Schwartz (1999) A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nat Genet,* 23, 309-13.

Langer-Safer, P. R., M. Levine & D. C. Ward (1982) Immunological method for mapping genes on Drosophila polytene chromosomes. *Proc Natl Acad Sci U S A,* 79, 4381-5.

Lanzer, M., H. Wickert, G. Krohne, L. Vincensini & C. Braun Breton (2006) Maurer's clefts: a novel multi-functional organelle in the cytoplasm of *Plasmodium falciparum*-infected erythrocytes. *Int J Parasitol,* 36, 23-36.

Le Scouarnec, S. & S. M. Gribble (2012) Characterising chromosome rearrangements: recent technical advances in molecular cytogenetics. *Heredity (Edinb),* 108, 75-85.
Lee, J. A., C. M. Carvalho & J. R. Lupski (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell,* 131, 1235-47.
Li, H. & R. Durbin (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25, 1754-60.
--- (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics,* 26, 589-95.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis & R. Durbin (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25, 2078-9.
Li, J. L. & D. A. Baker (1997) Protein phosphatase beta, a putative type-2A protein phosphatase from the human malaria parasite *Plasmodium falciparum. Eur J Biochem,* 249, 98-106.
Liljander, A., L. Wiklund, N. Falk, M. Kweku, A. Martensson, I. Felger & A. Farnert (2009) Optimization and validation of multi-coloured capillary electrophoresis for genotyping of *Plasmodium falciparum* merozoite surface proteins (msp1 and 2). *Malar J,* 8, 78.
Lim, P., A. P. Alker, N. Khim, N. K. Shah, S. Incardona, S. Doung, P. Yi, D. M. Bouth, C. Bouchier, O. M. Puijalon, S. R. Meshnick, C. Wongsrichanalai, T. Fandeur, J. Le Bras, P. Ringwald & F. Ariey (2009) Pfmdr1 copy number and arteminisin derivatives combination therapy failure in falciparum malaria in Cambodia. *Malar J,* 8, 11.
Liu, P., C. M. Carvalho, P. J. Hastings & J. R. Lupski (2012) Mechanisms for recurrent and complex human genomic rearrangements. *Curr Opin Genet Dev,* 22, 211-20.
Logan-Klumpler, F. J., N. De Silva, U. Boehme, M. B. Rogers, G. Velarde, J. A. McQuillan, T. Carver, M. Aslett, C. Olsen, S. Subramanian, I. Phan, C. Farris, S. Mitra, G. Ramasamy, H. Wang, A. Tivey, A. Jackson, R. Houston, J. Parkhill, M. Holden, O. S. Harb, B. P. Brunk, P. J. Myler, D. Roos, M. Carrington, D. F. Smith, C. Hertz-Fowler & M. Berriman (2012) GeneDB--an annotation database for pathogens. *Nucleic Acids Res,* 40, D98-108.
Lopez-Rubio, J. J., A. M. Gontijo, M. C. Nunes, N. Issar, R. Hernandez Rivas & A. Scherf (2007) 5' flanking region of var genes nucleate histone modification patterns linked to phenotypic inheritance of virulence traits in malaria parasites. *Mol Microbiol,* 66, 1296-305.
Ma, J., J. Stiller, P. J. Berkman, Y. Wei, J. Rogers, C. Feuillet, J. Dolezel, K. F. Mayer, K. Eversole, Y. L. Zheng & C. Liu (2013) Sequence-based analysis of translocations and inversions in bread wheat (Triticum aestivum L.). *PLoS One,* 8, e79329.
Mackinnon, M. J., J. Li, S. Mok, M. M. Kortok, K. Marsh, P. R. Preiser & Z. Bozdech (2009) Comparative transcriptional and genomic analysis of *Plasmodium falciparum* field isolates. *PLoS Pathog,* 5, e1000644.
Mackinnon, M. J. & K. Marsh (2010) The selection landscape of malaria parasites. *Science,* 328, 866-71.
Magi, A., M. Benelli, S. Yoon, F. Roviello & F. Torricelli (2011) Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic Acids Res,* 39, e65.
Maier, A. G., M. Rug, M. T. O'Neill, J. G. Beeson, M. Marti, J. Reeder & A. F. Cowman (2007) Skeleton-binding protein 1 functions at the parasitophorous vacuole membrane to traffic PfEMP1 to the *Plasmodium falciparum*-infected erythrocyte surface. *Blood,* 109, 1289-97.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley & J. M. Rothberg (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437, 376-80.

Martineau, Y., M. C. Derry, X. Wang, A. Yanagiya, J. J. Berlanga, A. B. Shyu, H. Imataka, K. Gehring & N. Sonenberg (2008) Poly(A)-binding protein-interacting protein 1 binds to eukaryotic translation initiation factor 3 to stimulate translation. *Mol Cell Biol,* 28, 6658-67.

Mastrangelo, A. M., D. Marone, G. Laido, A. M. De Leonardis & P. De Vita (2012) Alternative splicing: enhancing ability to cope with stress via transcriptome plasticity. *Plant Sci,* 185-186, 40-9.

Mathews, L. A., S. M. Cabarcas, E. M. Hurt, X. Zhang, E. M. Jaffee & W. L. Farrar (2011) Increased expression of DNA repair genes in invasive human pancreatic cancer cells. *Pancreas,* 40, 730-9.

Mbengue, A., E. Vialla, L. Berry, G. Fall, N. Audiger, E. Demettre-Verceil, D. Boteller & C. Braun-Breton (2015) New Export Pathway in *Plasmodium falciparum*-Infected Erythrocytes: Role of the Parasite Group II Chaperonin, PfTRiC. *Traffic,* 16, 461-75.

Miller, C. A., O. Hampton, C. Coarfa & A. Milosavljevic (2011) ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One,* 6, e16327.

Mills, J. P., M. Diez-Silva, D. J. Quinn, M. Dao, M. J. Lang, K. S. Tan, C. T. Lim, G. Milon, P. H. David, O. Mercereau-Puijalon, S. Bonnefoy & S. Suresh (2007) Effect of plasmodial RESA protein on deformability of human red blood cells harboring *Plasmodium falciparum. Proc Natl Acad Sci U S A,* 104, 9213-7.

Mills, R. E., K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. Lam, J. Leng, R. Li, Y. Li, C. Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stutz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll & J. O. Korbel (2011a) Mapping copy number variation by population-scale genome sequencing. *Nature,* 470, 59-65.

Mills, R. E., K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. Lam, J. Leng, R. Li, Y. Li, C. Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stutz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll & J. O. Korbel

(2011b) Mapping copy number variation by population-scale genome sequencing. *Nature,* 470, 59-65.

Miotto, O., J. Almagro-Garcia, M. Manske, B. Macinnis, S. Campino, K. A. Rockett, C. Amaratunga, P. Lim, S. Suon, S. Sreng, J. M. Anderson, S. Duong, C. Nguon, C. M. Chuor, D. Saunders, Y. Se, C. Lon, M. M. Fukuda, L. Amenga-Etego, A. V. Hodgson, V. Asoala, M. Imwong, S. Takala-Harrison, F. Nosten, X. Z. Su, P. Ringwald, F. Ariey, C. Dolecek, T. T. Hien, M. F. Boni, C. Q. Thai, A. Amambua-Ngwa, D. J. Conway, A. A. Djimde, O. K. Doumbo, I. Zongo, J. B. Ouedraogo, D. Alcock, E. Drury, S. Auburn, O. Koch, M. Sanders, C. Hubbart, G. Maslen, V. Ruano-Rubio, D. Jyothi, A. Miles, J. O'Brien, C. Gamble, S. O. Oyola, J. C. Rayner, C. I. Newbold, M. Berriman, C. C. Spencer, G. McVean, N. P. Day, N. J. White, D. Bethell, A. M. Dondorp, C. V. Plowe, R. M. Fairhurst & D. P. Kwiatkowski (2013) Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia. *Nat Genet.*

Mok, S., M. Imwong, M. J. Mackinnon, J. Sim, R. Ramadoss, P. Yi, M. Mayxay, K. Chotivanich, K. Y. Liong, B. Russell, D. Socheat, P. N. Newton, N. P. Day, N. J. White, P. R. Preiser, F. Nosten, A. M. Dondorp & Z. Bozdech (2011) Artemisinin resistance in *Plasmodium falciparum* is associated with an altered temporal pattern of transcription. *BMC Genomics,* 12, 391.

Mu, J., P. Awadalla, J. Duan, K. M. McGee, D. A. Joy, G. A. McVean & X. Z. Su (2005) Recombination hotspots and population structure in *Plasmodium falciparum*. *PLoS Biol,* 3, e335.

Mu, J., R. A. Myers, H. Jiang, S. Liu, S. Ricklefs, M. Waisberg, K. Chotivanich, P. Wilairatana, S. Krudsood, N. J. White, R. Udomsangpetch, L. Cui, M. Ho, F. Ou, H. Li, J. Song, G. Li, X. Wang, S. Seila, S. Sokunthea, D. Socheat, D. E. Sturdevant, S. F. Porcella, R. M. Fairhurst, T. E. Wellems, P. Awadalla & X. Z. Su (2010) *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nat Genet,* 42, 268-71.

Murray, C. J., L. C. Rosenfeld, S. S. Lim, K. G. Andrews, K. J. Foreman, D. Haring, N. Fullman, M. Naghavi, R. Lozano & A. D. Lopez (2012) Global malaria mortality between 1980 and 2010: a systematic analysis. *Lancet,* 379, 413-31.

Mwangangi, J. M., C. M. Mbogo, B. O. Orindi, E. J. Muturi, J. T. Midega, J. Nzovu, H. Gatakaa, J. Githure, C. Borgemeister, J. Keating & J. C. Beier (2013) Shifts in malaria vector species composition and transmission dynamics along the Kenyan coast over the past 20 years. *Malar J,* 12, 13.

Nacer, A., A. Claes, A. Roberts, C. Scheidig-Benatar, H. Sakamoto, M. Ghorbal, J. J. Lopez-Rubio & D. Mattei (2015) Discovery of a novel and conserved *Plasmodium falciparum* exported protein that is important for adhesion of PfEMP1 at the surface of infected erythrocytes. *Cell Microbiol,* 17, 1205-16.

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

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

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

Nathan, C. & M. U. Shiloh (2000) Reactive oxygen and nitrogen intermediates in the relationship between mammalian hosts and microbial pathogens. *Proc Natl Acad Sci U S A,* 97, 8841-8.

Nguitragool, W., A. A. Bokhari, A. D. Pillai, K. Rayavara, P. Sharma, B. Turpin, L. Aravind & S. A. Desai (2011) Malaria parasite clag3 genes determine channel-mediated nutrient uptake by infected red blood cells. *Cell,* 145, 665-77.

Niang, M., A. K. Bei, K. G. Madnani, S. Pelly, S. Dankwa, U. Kanjee, K. Gunalan, A. Amaladoss, K. P. Yeo, N. S. Bob, B. Malleret, M. T. Duraisingh & P. R. Preiser (2014) STEVOR is a *Plasmodium falciparum* erythrocyte binding protein that mediates merozoite invasion and rosetting. *Cell Host Microbe,* 16, 81-93.

Nickel, C., S. Rahlfs, M. Deponte, S. Koncarevic & K. Becker (2006) Thioredoxin networks in the malarial parasite *Plasmodium falciparum*. *Antioxid Redox Signal,* 8, 1227-39.

Nijkamp, J. F., M. A. van den Broek, J. M. Geertman, M. J. Reinders, J. M. Daran & D. de Ridder (2012) De novo detection of copy number variation by co-assembly. *Bioinformatics,* 28, 3195-202.

Nishino, T. & K. Morikawa (2002) Structure and function of nucleases in DNA repair: shape, grip and blade of the DNA scissors. *Oncogene,* 21, 9022-32.

Noedl, H., Y. Se, K. Schaecher, B. L. Smith, D. Socheat & M. M. Fukuda (2008) Evidence of artemisinin-resistant malaria in western Cambodia. *N Engl J Med,* 359, 2619-20.

Noor, A. M., D. K. Kinyoki, C. W. Mundia, C. W. Kabaria, J. W. Mutua, V. A. Alegana, I. S. Fall & R. W. Snow (2014) The changing risk of *Plasmodium falciparum* malaria infection in Africa: 2000-10: a spatial and temporal analysis of transmission intensity. *Lancet.*

Ntoumi, F., H. Contamin, C. Rogier, S. Bonnefoy, J. F. Trape & O. Mercereau-Puijalon (1995) Age-dependent carriage of multiple *Plasmodium falciparum* merozoite surface antigen-2 alleles in asymptomatic malaria infections. *Am J Trop Med Hyg,* 52, 81-8.

Nzila, A. & L. Mwai (2010) In vitro selection of *Plasmodium falciparum* drug-resistant parasite lines. *J Antimicrob Chemother,* 65, 390-8.

O'Meara, W. P., P. Bejon, T. W. Mwangi, E. A. Okiro, N. Peshu, R. W. Snow, C. R. Newton & K. Marsh (2008) Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya. *Lancet,* 372, 1555-62.

Ochoa, A., M. Llinas & M. Singh (2011) Using context to improve protein domain identification. *BMC Bioinformatics,* 12, 90.

Okiro, E. A., V. A. Alegana, A. M. Noor & R. W. Snow (2010) Changing malaria intervention coverage, transmission and hospitalization in Kenya. *Malar J,* 9, 285.

Olshen, A. B., E. S. Venkatraman, R. Lucito & M. Wigler (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics,* 5, 557-72.

Palli, D., P. Rizzolo, I. Zanna, V. Silvestri, C. Saieva, M. Falchetti, A. S. Navazio, V. Graziano, G. Masala, S. Bianchi, A. Russo, S. Tommasi & L. Ottini (2013) SULT1A1 gene deletion in BRCA2-associated male breast cancer: a link between genes and environmental exposures? *J Cell Mol Med,* 17, 605-7.

Pei, X., X. Guo, R. Coppel, S. Bhattacharjee, K. Haldar, W. Gratzer, N. Mohandas & X. An (2007) The ring-infected erythrocyte surface antigen (RESA) of *Plasmodium falciparum* stabilizes spectrin tetramers and suppresses further invasion. *Blood,* 110, 1036-42.

Pesce, E. R., P. Acharya, U. Tatu, W. S. Nicoll, A. Shonhai, H. C. Hoppe & G. L. Blatch (2008) The *Plasmodium falciparum* heat shock protein 40, Pfj4, associates with heat

shock protein 70 and shows similar heat induction and localisation patterns. *Int J Biochem Cell Biol,* 40, 2914-26.

Petalidis, L., S. Bhattacharyya, G. A. Morris, V. P. Collins, T. C. Freeman & P. A. Lyons (2003) Global amplification of mRNA by template-switching PCR: linearity and application to microarray analysis. *Nucleic Acids Res,* 31, e142.

Phillips-Howard, P. A., B. L. Nahlen, M. S. Kolczak, A. W. Hightower, F. O. ter Kuile, J. A. Alaii, J. E. Gimnig, J. Arudo, J. M. Vulule, A. Odhacha, S. P. Kachur, E. Schoute, D. H. Rosen, J. D. Sexton, A. J. Oloo & W. A. Hawley (2003) Efficacy of permethrin-treated bed nets in the prevention of mortality in young children in an area of high perennial malaria transmission in western Kenya. *Am J Trop Med Hyg,* 68, 23-9.

Piel, F. B., A. P. Patil, R. E. Howes, O. A. Nyangiri, P. W. Gething, T. N. Williams, D. J. Weatherall & S. I. Hay (2010) Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nat Commun,* 1, 104.

Pinto, D., K. Darvishi, X. Shi, D. Rajan, D. Rigler, T. Fitzgerald, A. C. Lionel, B. Thiruvahindrapuram, J. R. Macdonald, R. Mills, A. Prasad, K. Noonan, S. Gribble, E. Prigmore, P. K. Donahoe, R. S. Smith, J. H. Park, M. E. Hurles, N. P. Carter, C. Lee, S. W. Scherer & L. Feuk (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol,* 29, 512-20.

Pique-Regi, R., A. Caceres & J. R. Gonzalez (2010) R-Gada: a fast and flexible pipeline for copy number analysis in association studies. *BMC Bioinformatics,* 11, 380.

Plowe, C. V., J. G. Kublin & O. K. Doumbo (1998) *P. falciparum* dihydrofolate reductase and dihydropteroate synthase mutations: epidemiology and role in clinical resistance to antifolates. *Drug Resist Updat,* 1, 389-96.

Pologe, L. G., D. de Bruin & J. V. Ravetch (1990) A and T homopolymeric stretches mediate a DNA inversion in *Plasmodium falciparum* which results in loss of gene expression. *Mol Cell Biol,* 10, 3243-6.

Pologe, L. G. & J. V. Ravetch (1988) Large deletions result from breakage and healing of *P. falciparum* chromosomes. *Cell,* 55, 869-74.

Price, R. N., A. C. Uhlemann, A. Brockman, R. McGready, E. Ashley, L. Phaipun, R. Patel, K. Laing, S. Looareesuwan, N. J. White, F. Nosten & S. Krishna (2004) Mefloquine resistance in *Plasmodium falciparum* and increased pfmdr1 gene copy number. *Lancet,* 364, 438-47.

Price, T. S., R. Regan, R. Mott, A. Hedman, B. Honey, R. J. Daniels, L. Smith, A. Greenfield, A. Tiganescu, V. Buckle, N. Ventress, H. Ayyub, A. Salhan, S. Pedraza-Diaz, J. Broxholme, J. Ragoussis, D. R. Higgs, J. Flint & S. J. Knight (2005) SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res,* 33, 3455-64.

Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow & Y. Gu (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* 13, 341.

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C.

Lee, K. W. Jones, S. W. Scherer & M. E. Hurles (2006) Global variation in copy number in the human genome. *Nature,* 444, 444-54.

Reiling, L., J. S. Richards, F. J. Fowkes, A. E. Barry, T. Triglia, W. Chokejindachai, P. Michon, L. Tavul, P. M. Siba, A. F. Cowman, I. Mueller & J. G. Beeson (2010) Evidence that the erythrocyte invasion ligand PfRh2 is a target of protective immunity against *Plasmodium falciparum* malaria. *J Immunol,* 185, 6157-67.

Retterer, K., J. Scuffins, D. Schmidt, R. Lewis, D. Pineda-Alvarez, A. Stafford, L. Schmidt, S. Warren, F. Gibellini, A. Kondakova, A. Blair, S. Bale, L. Matyakhina, J. Meck, S. Aradhya & E. Haverfield (2014) Assessing copy number from exome sequencing and exome array CGH based on CNV spectrum in a large clinical cohort. *Genet Med.*

Ribacke, U., B. W. Mok, V. Wirta, J. Normark, J. Lundeberg, F. Kironde, T. G. Egwang, P. Nilsson & M. Wahlgren (2007) Genome wide gene amplifications and deletions in *Plasmodium falciparum. Mol Biochem Parasitol,* 155, 33-44.

Riley, M. C., B. C. Kirkup, Jr., J. D. Johnson, E. P. Lesho & C. F. Ockenhouse (2011) Rapid whole genome optical mapping of *Plasmodium falciparum. Malar J,* 10, 252.

Ritchie, M. E., J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway & G. K. Smyth (2007) A comparison of background correction methods for two-colour microarrays. *Bioinformatics,* 23, 2700-7.

Robinson, T., S. G. Campino, S. Auburn, S. A. Assefa, S. D. Polley, M. Manske, B. MacInnis, K. A. Rockett, G. L. Maslen, M. Sanders, M. A. Quail, P. L. Chiodini, D. P. Kwiatkowski, T. G. Clark & C. J. Sutherland (2011) Drug-resistant genotypes and multi-clonality in *Plasmodium falciparum* analysed by direct genome sequencing from peripheral blood of malaria patients. *PLoS One,* 6, e23204.

Ross, M. G., C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum & D. B. Jaffe (2013) Characterizing and measuring bias in sequence data. *Genome Biol,* 14, R51.

Rowe, J. A., A. Claessens, R. A. Corrigan & M. Arman (2009) Adhesion of *Plasmodium falciparum*-infected erythrocytes to human cells: molecular mechanisms and therapeutic implications. *Expert Rev Mol Med,* 11, e16.

Sahar, T., K. S. Reddy, M. Bharadwaj, A. K. Pandey, S. Singh, C. E. Chitnis & D. Gaur (2011) *Plasmodium falciparum* reticulocyte binding-like homologue protein 2 (PfRH2) is a key adhesive molecule involved in erythrocyte invasion. *PLoS One,* 6, e17102.

Salanti, A., T. Staalsoe, T. Lavstsen, A. T. Jensen, M. P. Sowa, D. E. Arnot, L. Hviid & T. G. Theander (2003) Selective upregulation of a single distinctly structured var gene in chondroitin sulphate A-adhering *Plasmodium falciparum* involved in pregnancy-associated malaria. *Mol Microbiol,* 49, 179-91.

Salipante, S. J., T. Kawashima, C. Rosenthal, D. R. Hoogestraat, L. A. Cummings, D. J. Sengupta, T. T. Harkins, B. T. Cookson & N. G. Hoffman (2014) Performance Comparison of Illumina and Ion Torrent Next-Generation Sequencing Platforms for 16S rRNA-Based Bacterial Community Profiling. *Appl Environ Microbiol,* 80, 7583-91.

Samarakoon, U., J. M. Gonzales, J. J. Patel, A. Tan, L. Checkley & M. T. Ferdig (2011a) The landscape of inherited and de novo copy number variants in a *Plasmodium falciparum* genetic cross. *BMC Genomics,* 12, 457.

Samarakoon, U., A. Regier, A. Tan, B. A. Desany, B. Collins, J. C. Tan, S. J. Emrich & M. T. Ferdig (2011b) High-throughput 454 resequencing for allele discovery and recombination mapping in *Plasmodium falciparum. BMC Genomics,* 12, 116.

Sander, A. F., T. Lavstsen, T. S. Rask, M. Lisby, A. Salanti, S. L. Fordyce, J. S. Jespersen, R. Carter, K. W. Deitsch, T. G. Theander, A. G. Pedersen & D. E. Arnot (2014) DNA

secondary structures are associated with recombination in major *Plasmodium falciparum* variable surface antigen gene families. *Nucleic Acids Res,* 42, 2270-81.

Sanders, P. R., P. R. Gilson, G. T. Cantin, D. C. Greenbaum, T. Nebl, D. J. Carucci, M. J. McConville, L. Schofield, A. N. Hodder, J. R. Yates, 3rd & B. S. Crabb (2005) Distinct protein classes including novel merozoite surface antigens in Raft-like membranes of *Plasmodium falciparum. J Biol Chem,* 280, 40169-76.

Sargeant, T. J., M. Marti, E. Caler, J. M. Carlton, K. Simpson, T. P. Speed & A. F. Cowman (2006) Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biol,* 7, R12.

Scherf, A., R. Carter, C. Petersen, P. Alano, R. Nelson, M. Aikawa, D. Mattei, L. Pereira da Silva & J. Leech (1992) Gene inactivation of Pf11-1 of *Plasmodium falciparum* by chromosome breakage and healing: identification of a gametocyte-specific protein with a potential role in gametogenesis. *EMBO J,* 11, 2293-301.

Scherf, A., J. J. Lopez-Rubio & L. Riviere (2008) Antigenic variation in *Plasmodium falciparum. Annu Rev Microbiol,* 62, 445-70.

Scherf, A. & D. Mattei (1992) Cloning and characterization of chromosome breakpoints of *Plasmodium falciparum*: breakage and new telomere formation occurs frequently and randomly in subtelomeric genes. *Nucleic Acids Res,* 20, 1491-6.

Schwartz, D. C. & C. R. Cantor (1984) Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell,* 37, 67-75.

Sepulveda, N., S. G. Campino, S. A. Assefa, C. J. Sutherland, A. Pain & T. G. Clark (2013) A Poisson hierarchical modelling approach to detecting copy number variation in sequence coverage data. *BMC Genomics,* 14, 128.

Sexton, T., D. Umlauf, S. Kurukuti & P. Fraser (2007) The role of transcription factories in large-scale structure and dynamics of interphase chromatin. *Semin Cell Dev Biol,* 18, 691-7.

Shirley, M. W., B. A. Biggs, K. P. Forsyth, H. J. Brown, J. K. Thompson, G. V. Brown & D. J. Kemp (1990) Chromosome 9 from independent clones and isolates of *Plasmodium falciparum* undergoes subtelomeric deletions with similar breakpoints in vitro. *Mol Biochem Parasitol,* 40, 137-45.

Shonhai, A., A. G. Maier, J. M. Przyborski & G. L. Blatch (2011) Intracellular protozoan parasites of humans: the role of molecular chaperones in development and pathogenesis. *Protein Pept Lett,* 18, 143-57.

Sidhu, A. B., A. C. Uhlemann, S. G. Valderramos, J. C. Valderramos, S. Krishna & D. A. Fidock (2006) Decreasing pfmdr1 copy number in *Plasmodium falciparum* malaria heightens susceptibility to mefloquine, lumefantrine, halofantrine, quinine, and artemisinin. *J Infect Dis,* 194, 528-35.

Silvestrini, F., E. Lasonder, A. Olivieri, G. Camarda, B. van Schaijk, M. Sanchez, S. Younis Younis, R. Sauerwein & P. Alano (2010) Protein export marks the early phase of gametocytogenesis of the human malaria parasite *Plasmodium falciparum. Mol Cell Proteomics,* 9, 1437-48.

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

Sirawaraporn, W., T. Sathitkul, R. Sirawaraporn, Y. Yuthavong & D. V. Santi (1997) Antifolate-resistant mutants of *Plasmodium falciparum* dihydrofolate reductase. *Proc Natl Acad Sci U S A,* 94, 1124-9.

Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol,* 3, Article3.

---. 2005. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor.*, ed. R. C. Gentleman, V. Dudoit, S. Irizarry, R. Huber, W., 397-420. New York: Springer.

Smyth, G. K. & T. Speed (2003) Normalization of cDNA microarray data. *Methods,* 31, 265-73.

Snow, R. W., C. A. Guerra, A. M. Noor, H. Y. Myint & S. I. Hay (2005) The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature,* 434, 214-7.

Soulama, I., J. D. Bigoga, M. Ndiaye, E. C. Bougouma, J. Quagraine, P. N. Casimiro, T. T. Stedman & S. B. Sirima (2011) Genetic diversity of polymorphic vaccine candidate antigens (apical membrane antigen-1, merozoite surface protein-3, and erythrocyte binding antigen-175) in *Plasmodium falciparum* isolates from western and central Africa. *Am J Trop Med Hyg,* 84, 276-84.

Southern, E. M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol,* 98, 503-17.

Spielmann, T., D. L. Gardiner, H. P. Beck, K. R. Trenholme & D. J. Kemp (2006a) Organization of ETRAMPs and EXP-1 at the parasite-host cell interface of malaria parasites. *Mol Microbiol,* 59, 779-94.

Spielmann, T., P. L. Hawthorne, M. W. Dixon, M. Hannemann, K. Klotz, D. J. Kemp, N. Klonis, L. Tilley, K. R. Trenholme & D. L. Gardiner (2006b) A cluster of ring stage-specific genes linked to a locus implicated in cytoadherence in *Plasmodium falciparum* codes for PEXEL-negative and PEXEL-positive proteins exported into the host cell. *Mol Biol Cell,* 17, 3613-24.

Stranger, B. E., M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavare, P. Deloukas, M. E. Hurles & E. T. Dermitzakis (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science,* 315, 848-53.

Stubbs, J., K. M. Simpson, T. Triglia, D. Plouffe, C. J. Tonkin, M. T. Duraisingh, A. G. Maier, E. A. Winzeler & A. F. Cowman (2005) Molecular mechanism for switching of *P. falciparum* invasion pathways into human erythrocytes. *Science,* 309, 1384-7.

Su, X. Z., V. M. Heatwole, S. P. Wertheimer, F. Guinet, J. A. Herrfeldt, D. S. Peterson, J. A. Ravetch & T. E. Wellems (1995) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell,* 82, 89-100.

Symington, L. S. & J. Gautier (2011) Double-strand break end resection and repair pathway choice. *Annu Rev Genet,* 45, 247-71.

Takala-Harrison, S., T. G. Clark, C. G. Jacob, M. P. Cummings, O. Miotto, A. M. Dondorp, M. M. Fukuda, F. Nosten, H. Noedl, M. Imwong, D. Bethell, Y. Se, C. Lon, S. D. Tyner, D. L. Saunders, D. Socheat, F. Ariey, A. P. Phyo, P. Starzengruber, H. P. Fuehrer, P. Swoboda, K. Stepniewska, J. Flegg, C. Arze, G. C. Cerqueira, J. C. Silva, S. M. Ricklefs, S. F. Porcella, R. M. Stephens, M. Adams, L. J. Kenefic, S. Campino, S. Auburn, B. MacInnis, D. P. Kwiatkowski, X. Z. Su, N. J. White, P. Ringwald & C. V. Plowe (2013) Genetic loci associated with delayed clearance of *Plasmodium falciparum* following artemisinin treatment in Southeast Asia. *Proc Natl Acad Sci U S A,* 110, 240-5.

Tam, G. W., R. Redon, N. P. Carter & S. G. Grant (2009) The role of DNA copy number variation in schizophrenia. *Biol Psychiatry,* 66, 1005-12.

Templeton, T. J., D. C. Kaslow & D. A. Fidock (2000) Developmental arrest of the human malaria parasite *Plasmodium falciparum* within the mosquito midgut via CTRP gene disruption. *Mol Microbiol,* 36, 1-9.

Tetteh, K. K., L. B. Stewart, L. I. Ochola, A. Amambua-Ngwa, A. W. Thomas, K. Marsh, G. D. Weedall & D. J. Conway (2009) Prospective identification of malaria parasite genes under balancing selection. *PLoS One,* 4, e5568.

Thaithong, S., L. C. Ranford-Cartwright, N. Siripoon, P. Harnyuttanakorn, N. S. Kanchanakhan, A. Seugorn, K. Rungsihirunrat, P. V. Cravo & G. H. Beale (2001) *Plasmodium falciparum*: gene mutations and amplification of dihydrofolate reductase genes in parasites grown in vitro in presence of pyrimethamine. *Exp Parasitol,* 98, 59-70.

Treangen, T. J. & S. L. Salzberg (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet,* 13, 36-46.

Trenholme, K. R., D. L. Gardiner, D. C. Holt, E. A. Thomas, A. F. Cowman & D. J. Kemp (2000) clag9: A cytoadherence gene in *Plasmodium falciparum* essential for binding of parasitized erythrocytes to CD36. *Proc Natl Acad Sci U S A,* 97, 4029-33.

Triglia, T., P. Wang, P. F. Sims, J. E. Hyde & A. F. Cowman (1998) Allelic exchange at the endogenous genomic locus in *Plasmodium falciparum* proves the role of dihydropteroate synthase in sulfadoxine-resistant malaria. *Embo j,* 17, 3807-15.

Van der Ploeg, L. H., M. Smits, T. Ponnudurai, A. Vermeulen, J. H. Meuwissen & G. Langsley (1985) Chromosome-sized DNA molecules of *Plasmodium falciparum. Science,* 229, 658-61.

Van Tyne, D., D. J. Park, S. F. Schaffner, D. E. Neafsey, E. Angelino, J. F. Cortese, K. G. Barnes, D. M. Rosen, A. K. Lukens, R. F. Daniels, D. A. Milner, Jr., C. A. Johnson, I. Shlyakhter, S. R. Grossman, J. S. Becker, D. Yamins, E. K. Karlsson, D. Ndiaye, O. Sarr, S. Mboup, C. Happi, N. A. Furlotte, E. Eskin, H. M. Kang, D. L. Hartl, B. W. Birren, R. C. Wiegand, E. S. Lander, D. F. Wirth, S. K. Volkman & P. C. Sabeti (2011) Identification and functional validation of the novel antimalarial resistance locus PF10_0355 in *Plasmodium falciparum. PLoS Genet,* 7, e1001383.

Venkatraman, E. S. & A. B. Olshen (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics,* 23, 657-63.

Volkman, S. K., P. C. Sabeti, D. DeCaprio, D. E. Neafsey, S. F. Schaffner, D. A. Milner, Jr., J. P. Daily, O. Sarr, D. Ndiaye, O. Ndir, S. Mboup, M. T. Duraisingh, A. Lukens, A. Derr, N. Stange-Thomann, S. Waggoner, R. Onofrio, L. Ziaugra, E. Mauceli, S. Gnerre, D. B. Jaffe, J. Zainoun, R. C. Wiegand, B. W. Birren, D. L. Hartl, J. E. Galagan, E. S. Lander & D. F. Wirth (2007) A genome-wide map of diversity in *Plasmodium falciparum. Nat Genet,* 39, 113-9.

Voss, T. S., T. Mini, P. Jenoe & H. P. Beck (2002) *Plasmodium falciparum* possesses a cell cycle-regulated short type replication protein A large subunit encoded by an unusual transcript. *J Biol Chem,* 277, 17493-501.

Wellems, T. E., D. Walliker, C. L. Smith, V. E. do Rosario, W. L. Maloy, R. J. Howard, R. Carter & T. F. McCutchan (1987) A histidine-rich protein gene marks a linkage group favored strongly in a genetic cross of *Plasmodium falciparum. Cell,* 49, 633-42.

WHO. 2014. *World Health Organization World Malaria Report 2014.* World Health Organization.

Winchester, L., D. F. Newbury, A. P. Monaco & J. Ragoussis (2008) Detection, breakpoint identification and detailed characterisation of a CNV at the FRA16D site using SNP assays. *Cytogenet Genome Res,* 123, 322-32.

Winter, G., S. Kawai, M. Haeggstrom, O. Kaneko, A. von Euler, S. Kawazu, D. Palm, V. Fernandez & M. Wahlgren (2005) SURFIN is a polymorphic antigen expressed on *Plasmodium falciparum* merozoites and infected erythrocytes. *J Exp Med,* 201, 1853-63.

Wright, G. J. & J. C. Rayner (2014) *Plasmodium falciparum* erythrocyte invasion: combining function with immune evasion. *PLoS Pathog,* 10, e1003943.

Xi, R., A. G. Hadjipanayis, L. J. Luquette, T. M. Kim, E. Lee, J. Zhang, M. D. Johnson, D. M. Muzny, D. A. Wheeler, R. A. Gibbs, R. Kucherlapati & P. J. Park (2011) Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci U S A,* 108, E1128-36.

Xie, C. & M. T. Tammi (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics,* 10, 80.

Yalcin, B., K. Wong, A. Agam, M. Goodson, T. M. Keane, X. Gan, C. Nellaker, L. Goodstadt, J. Nicod, A. Bhomra, P. Hernandez-Pliego, H. Whitley, J. Cleak, R. Dutton, D. Janowitz, R. Mott, D. J. Adams & J. Flint (2011) Sequence-based characterization of structural variation in the mouse genome. *Nature,* 477, 326-9.

Yoon, S., Z. Xuan, V. Makarov, K. Ye & J. Sebat (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res,* 19, 1586-92.

Young, J. A., Q. L. Fivelman, P. L. Blair, P. de la Vega, K. G. Le Roch, Y. Zhou, D. J. Carucci, D. A. Baker & E. A. Winzeler (2005) The *Plasmodium falciparum* sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol Biochem Parasitol,* 143, 67-79.

Zeng, W., J. C. de Greef, Y. Y. Chen, R. Chien, X. Kong, H. C. Gregson, S. T. Winokur, A. Pyle, K. D. Robertson, J. A. Schmiesing, V. E. Kimonis, J. Balog, R. R. Frants, A. R. Ball, Jr., L. F. Lock, P. J. Donovan, S. M. van der Maarel & K. Yokomori (2009) Specific loss of histone H3 lysine 9 trimethylation and HP1gamma/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD). *PLoS Genet,* 5, e1000559.

Zhang, F., M. Khajavi, A. M. Connolly, C. F. Towne, S. D. Batish & J. R. Lupski (2009) The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet,* 41, 849-53.

Zhang, Z. D., J. Du, H. Lam, A. Abyzov, A. E. Urban, M. Snyder & M. Gerstein (2011) Identification of genomic indels and structural variations using split reads. *BMC Genomics,* 12, 375.

Zhao, M., Q. Wang, Q. Wang, P. Jia & Z. Zhao (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics,* 14 Suppl 11, S1.

Zichner, T., D. A. Garfield, T. Rausch, A. M. Stutz, E. Cannavo, M. Braun, E. E. Furlong & J. O. Korbel (2013) Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing. *Genome Res,* 23, 568-79.

8 Appendix

8.1 Appendix 1.1. A summary of genome-wide studies of CNVs in *P. falciparum* using microarrays and next generation sequencing technologies.

Platform	No. of CNVs identified	Reference genome	Size of CNVs	Isolates	CNV detection criteria	Reference
70-mer microarray 4488 genes	144 CNV genes	3D7		1 lab line (HB3)	Probes with difference in red and green intensities of less than 50% of total (3D7) signal intensity. Exclude VSAs	(Bozdech et al. 2003a).
25-mer Affymetrix scrMalaria array. 5159 genes 260,596 probes	177 CNV genes 37 deleted genes	NF54		7 cultured isolates	Deletion defined as greater than 40% reduction in intensity of test relative to reference (NF54)	(Carret et al. 2005).
25 mer Affymetrix array	13 amplifications (~116 genes) 33 gene deleted	3D7		9 lab lines and 5 culture field isolates	Amplification detection cutoff at $\log_2$ratio > 0.7 Deletion detected using MOID algorithm.	(Kidgell et al. 2006).
70 mer microarray	50 genes amplified 32 genes deleted	3D7	980bp - 107K bp	2 fresh isolates and 7 lab lines	2 consecutive probes showing similar statistically significant ratio difference (B statistics in limma (Smyth 2005)). Change in one direction (increase/decrease), VSAs excluded.	(Ribacke et al. 2007).

PFSANGER GeneChip 2.2 million probes	390 CNV genes	3D7		4 cultured lines	Segmentation using Partek Genomic Suite v6.3 Segments 1.5 fold signal difference relative to 3D7 Minimum of 15 probes. >300bp	(Jiang et al. 2008b).
70 mer microarray 5224 genes	~324 CNV genes	3D7		5 short term culture isolates	1.5 fold in intensity difference between test samples and 3D7. Significant variation in $\log_2$ intensity between strains ($p<0.001$). Occurrence of CNV in at least two isolates.	(Mackinnon et al. 2009).
70-mer array	138 CNV genes	3D7		6 fresh clinical isolates	GADA (Pique-Regi et al. 2010).	(Mok et al. 2011).
Custom Tiling array 4.8 million probes 90% coding and 60% noncoding regions.	79 CNV regions	3D7	1.9kb p-83kbp	4 lab lines	Deletion detected by MOID algorithm. 10 unique probes per gene. Amplifications detected by performing z-test on windows testing whether $\log_2$ ratio >0. Z-score cut off of 18.	(Dharia et al. 2009).
454 sequencing	2 deletions 5 amplifications	3D7	2.5kb-160kb	2 lab lines	Read depth analysis	(Samarakoon et al. 2011a).
Illumina sequencing	5 deletions 2 amplification	3D7	300bp -20kb	5 fresh isolates	Read depth analysis (Yoon et al. 2009) and paired end mapping	(Robinson et al. 2011).

*MOID-Match-only Integral Distribution Algorithm

8.2 Appendix 3.1. Overlap between CNVs identified in our study compared to published studies.

CNV name	Gene ID	Chromosome	Gene name	Reference
cnv1_005	PF3D7_0108500	1	Hypothetical protein, conserved	(Cheeseman et al. 2009)
cnv2_013	PF3D7_0202000	2	Knob associated histidine-rich protein	(Jiang et al. 2008b, Samarakoon et al. 2011a, Samarakoon et al. 2011b, Pologe and Ravetch 1988, Scherf and Mattei 1992, Mackinnon et al. 2009, Ribacke et al. 2007, Carret et al. 2005)
	PF3D7_0202100	2	Plasmodium exported protein (PHISTc), unknown function	(Cheeseman et al. 2009, Mackinnon et al. 2009, Carret et al. 2005)
	PF3D7_0202200	2	Plasmodium exported protein, unknown function	(Mackinnon et al. 2009)
cnv3_043	PF3D7_0309300	3	N2227-like protein, putative	(Mackinnon et al. 2009)
	PF3D7_0309600	3	60S Acidic ribosomal protein P2	(Cheeseman et al. 2009)
cnv3_064	PF3D7_0322700	3		(Samarakoon et al. 2011b)
cnv4_091	PF3D7_0423400	4		(Samarakoon et al. 2011b)
	PF3D7_0423500	4		(Samarakoon et al. 2011b)
cnv4_092	PF3D7_0424400	4	hypothetical protein	(Samarakoon et al. 2011b, Jiang et al. 2008b)
cnv5_122	PF3D7_0529100	5	hypothetical protein, conserved	(Jiang et al. 2008b)
	PF3D7_0529200	5	sugar transporter, putative	(Mackinnon et al. 2009)
cnv7_169	PF3D7_0710200	7	hypothetical protein	(Samarakoon et al. 2011b, Jiang et al. 2008b, Mackinnon et al. 2009)

cnv7_181	PF3D7_0721000	7		(Samarakoon et al. 2011b)
cnv7_193	PF3D7_0730500	7		(Samarakoon et al. 2011b)
cnv8_201	PF3D7_0804700	8	conserved Plasmodium protein, unknown function	(Mackinnon et al. 2009)
cnv8_215	PF3D7_0829500PF3D7_0818000	8	RNA binding protein, putative	(Mackinnon et al. 2009)
cnv9_251	PF3D7_0921000	9	ubiquitin conjugating enzyme	(Mackinnon et al. 2009)
cnv9_269	PF3D7_0935400	9	cytoadherence-linked protein	(Mackinnon et al. 2009, Carret et al. 2005, Samarakoon et al. 2011a, Samarakoon et al. 2011b, Kidgell et al. 2006)
	PF3D7_0935500	9		(Mackinnon et al. 2009, Jiang et al. 2008b, Samarakoon et al. 2011a, Samarakoon et al. 2011b)
	PF3D7_0935600	9		(Cheeseman et al. 2009, Mackinnon et al. 2009, Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b)
	PF3D7_0935700	9		(Cheeseman et al. 2009, Mackinnon et al. 2009, Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b)
	PF3D7_0935800	9	cytoadherence linked asexual protein 9 (CLAG9)	(Cheeseman et al. 2009, Mackinnon et al. 2009, Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b)
	PF3D7_0935900	9		(Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b, Mackinnon et al. 2009)
	PF3D7_0936000	9	hypothetical protein	(Cheeseman et al. 2009, Mackinnon et al. 2009, Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b)

	PF3D7_0936100	9	hypothetical protein, conserved	(Cheeseman et al. 2009, Mackinnon et al. 2009, Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b)
	PF3D7_0936200	9	hypothetical protein	(Cheeseman et al. 2009, Mackinnon et al. 2009, Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b)
	PF3D7_0936400	9	hypothetical protein	(Cheeseman et al. 2009, Mackinnon et al. 2009, Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b)
	PF3D7_0936500	9	hypothetical protein	(Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b, Mackinnon et al. 2009)
	PFI1770w	9	hypothetical protein	(Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b)
	PF3D7_0936700	9	hypothetical protein	(Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b, Mackinnon et al. 2009)
	PF3D7_0936800	9	hypothetical protein	(Cheeseman et al. 2009, Mackinnon et al. 2009, Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b, Ribacke et al. 2007)
	PF3D7_0936900	9	hypothetical protein, conserved in *P. falciparum*	(Cheeseman et al. 2009, Mackinnon et al. 2009, Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b, Ribacke et al. 2007, Kidgell et al. 2006, Carret et al. 2005)
	PF3D7_0937000	9	Hypothetical hypothetical protein	(Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b, Mackinnon et al. 2009, Ribacke et al. 2007)
	PF3D7_0937100	9	hypothetical protein	(Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b, Mackinnon et al.

				2009, Ribacke et al. 2007, Cheeseman et al. 2009)
	PF3D7_0937200	9	enzyme, putative	(Samarakoon et al. 2011a, Samarakoon et al. 2011b, Jiang et al. 2008b, Mackinnon et al. 2009, Ribacke et al. 2007, Carret et al. 2005)
cnv11_335	PF3D7_1129000	11	spermidine synthase	(Mackinnon et al. 2009)
cnv11_344_1/2	PF3D7_1140500	11	myosin heavy chain subunit, putative	(Mackinnon et al. 2009)
cnv11_348	PF3D7_1143400	11	translation initiation factor eIF-1A, putative	(Mackinnon et al. 2009)
cnv11_354	PF3D7_1148700	11		(Samarakoon et al. 2011b, Sepulveda et al. 2013, Jiang et al. 2008b)
	PF3D7_1148800PF3D7_1148800	11	hypothetical protein	Jiang 2008, Samarakoon 2011 A, Cheeseman 2009
	PF3D7_1148900	11	hypothetical protein	(Jiang et al. 2008b, Samarakoon et al. 2011b)
	PF3D7_1149000	11	hypothetical protein	(Jiang et al. 2008b, Samarakoon et al. 2011b)
cnv11_355	PF3D7_1149000	11	antigen 332, putative	(Jiang et al. 2008b, Samarakoon et al. 2011b)
cnv12_375	PF3D7_1212400	12	Tetratricopeptide repeat family protein, putative	(Mackinnon et al. 2009)
cnv13_434	PF3D7_1313100	13	conserved Plasmodium protein, unknown function	(Mackinnon et al. 2009)
cnv13_441	PF3D7_1316800	13	sec20 homolog, putative	(Mackinnon et al. 2009)
cnv13_503	PF3D7_1366400	13		(Samarakoon et al. 2011b)
cnv14_516	PF3D7_1412400	14	hypothetical protein	(Jiang et al. 2008b)

8.3 Appendix 3.2. Potentially novel CNVs

CNV name	Gene ID	Description
cnv1_007	PF3D7_0109400	tubulin-specific chaperone a putative
	PF3D7_0109500	N-acetyltransferase putative
cnv2_014	PF3D7_0202500	early transcribed membrane protein 2
	PF3D7_0202600	conserved Plasmodium protein unknown function
cnv2_016	PF3D7_0203100	protein kinase putative
cnv2_023	PF3D7_0212400	conserved Plasmodium membrane protein unknown function
cnv2_024	PF3D7_0213600	conserved Plasmodium protein unknown function
cnv2_028_1/ cnv2_028_2	PF3D7_0215700	DNA-directed RNA polymerase II second largest subunit putative
cnv3_036	PF3D7_0301600	Plasmodium exported protein (hyp1) unknown function
	PF3D7_0301700	Plasmodium exported protein unknown function
	PF3D7_0301800	Plasmodium exported protein unknown function
cnv3_037	PF3D7_0302600	ABC transporter (TAP family) putative
	PF3D7_0302700	PFMNL-1 CISD1-like iron-sulfur protein putative
	PF3D7_0302800	conserved Plasmodium protein unknown function
	PF3D7_0302900	exportin 1 putative
cnv3_051	PF3D7_0315200	CSP and TRAP-related protein
cnv4_073	PF3D7_0407000	conserved Plasmodium protein unknown function
	PF3D7_0407100	conserved Plasmodium protein unknown function
	PF3D7_0407200	peptidyl-tRNA hydrolase PTH2 putative
cnv4_076	PF3D7_0409500	conserved Plasmodium protein unknown function
	PF3D7_0409600	replication protein A large subunit
	PF3D7_0409700	conserved Plasmodium protein unknown function
cnv4_078	PF3D7_0411700	conserved Plasmodium protein unknown function
cnv4_092	PF3D7_0424400	surface-associated interspersed gene 4.2 (SURFIN4.2)
cnv5_101	PF3D7_0507900	conserved Plasmodium protein unknown function

	PF3D7_0508000	6-cysteine protein
	PF3D7_0508100	SET domain protein putative
cnv5_106	PF3D7_0512600	Rab GTPase 1b
	PF3D7_0512700	orotate phosphoribosyltransferase
cnv5_108	PF3D7_0514300	aspartyl-tRNA synthetase putative
	PF3D7_0514500	conserved Plasmodium membrane protein unknown function
	PF3D7_0514600	ribose 5-phosphate epimerase putative
cnv5_109	PF3D7_0515000	RNA recognition motif putative
	PF3D7_0515100	rhomboid protease ROM9
	PF3D7_0515200	conserved Plasmodium protein unknown function
	PF3D7_0515300	phosphatidylinositol 3-kinase
cnv6_125	PF3D7_0602000	conserved Plasmodium protein unknown function
	PF3D7_0602100	ATP-dependent RNA helicase putative
	PF3D7_0602200	MYND finger protein putative
	PF3D7_0602300	conserved Plasmodium protein unknown function
	PF3D7_0602400	elongation factor G putative
	PF3D7_0602500	geranylgeranyltransferase putative
cnv6_127	PF3D7_0604800	RAP protein putative
	PF3D7_0604900	conserved Plasmodium protein unknown function
	PF3D7_0605000	mitochondrial ribosomal protein L24 precursor putative
	PF3D7_0605100	RNA binding protein putative
cnv6_129	PF3D7_0606200	ubiquitin conjugating enzyme E2 putative
	PF3D7_0606300	conserved Plasmodium protein unknown function
	PF3D7_0606400	conserved Plasmodium protein unknown function
	PF3D7_0606500	polypyrimidine tract binding protein putative
cnv7_185	PF3D7_0724100	conserved Plasmodium protein unknown function
	PF3D7_0724200	immunoglobulin-binding protein 1-related putative
cnv8_201	PF3D7_0805100	conserved Plasmodium protein unknown function

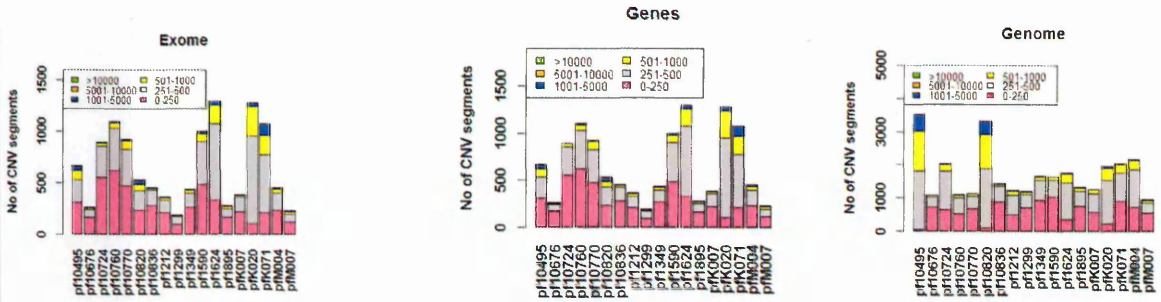
cnv8_222	PF3D7_0822600	Pfsec23
cnv8_227	PF3D7_0826000	conserved Plasmodium protein unknown function
	PF3D7_0825900	conserved Plasmodium protein unknown function
cnv8_233_1/ cnv8_233_2	PF3D7_0829500	conserved Plasmodium protein unknown function
cnv9_235	PF3D7_0902900	conserved Plasmodium protein unknown function
	PF3D7_0903000	conserved protein unknown function
	PF3D7_0903100	retrieval receptor for endoplasmic reticulum membrane proteins putative
cnv9_242	PF3D7_0908000	P1 nuclease putative
	PF3D7_0908100	conserved Plasmodium membrane protein unknown function
cnv9_249	PF3D7_0915200	conserved Plasmodium protein unknown function
	PF3D7_0915300	conserved Plasmodium protein unknown function
cnv9_254	PF3D7_0925400	protein phosphatase-beta
	PF3D7_0925500	thioredoxin-like protein 2
	PF3D7_0925600	zinc binding protein (Yippee) putative
	PF3D7_0925700	histone deacetylase
cnv9_255	PF3D7_0926400	monocarboxylate transporter putative
cnv9_259	PF3D7_0928000	cytochrome c oxidase putative
cnv9_262	PF3D7_0928900	guanylate kinase
	PF3D7_0929000	conserved Plasmodium protein unknown function
	PF3D7_0929100	conserved Plasmodium protein unknown function
	PF3D7_0929200	RNA binding protein putative
	PF3D7_0929300	conserved Plasmodium protein unknown function
cnv9_268	PF3D7_0934300	conserved Plasmodium protein unknown function
	PF3D7_0934400	transcription factor with AP2 domain(s) putative
cnv10_270	PF3D7_1001300	Plasmodium exported protein (PHISTa) unknown function
	PF3D7_1001400	alpha/beta hydrolase putative
cnv10_276	PF3D7_1008700	tubulin beta chain
	PF3D7_1008800	small subunit rRNA processing protein putative

	PF3D7_1008900	adenylate kinase
cnv10_279	PF3D7_1012500	phosphoglucomutase putative
cnv10_283	PF3D7_1016500	Plasmodium exported protein (PHISTc) unknown function
	PF3D7_1016700	Plasmodium exported protein (PHISTc) unknown function
cnv11_304	PF3D7_1107300	MIF4G domain containing protein
	PF3D7_1107400	Rad51 homolog
	PF3D7_1107500	prefoldin putative
cnv11_313	PF3D7_1113800	conserved Plasmodium membrane protein unknown function
	PF3D7_1113900	mitogen-activated protein kinase 2
cnv11_316	PF3D7_1114800	glycerol-3-phosphate dehydrogenase putative
	PF3D7_1114900	conserved Plasmodium protein unknown function
cnv11_347	PF3D7_1142600	60S ribosomal protein L35ae putative
	PF3D7_1142700	conserved protein unknown function
	PF3D7_1142800	conserved Plasmodium protein unknown function
	PF3D7_1142900	conserved Plasmodium protein unknown function
cnv12_359	PF3D7_1201700	conserved Plasmodium membrane protein unknown function
	PF3D7_1201800	cytochrome c oxidase assembly protein putative
	PF3D7_1201900	conserved protein unknown function
	PF3D7_1202000	ATP-dependent RNA helicase putative
cnv12_360	PF3D7_1202100	conserved Plasmodium protein unknown function
	PF3D7_1202200	mitochondrial phosphate carrier protein
cnv12_364	PF3D7_1205400	conserved Plasmodium protein unknown function
	PF3D7_1205500	conserved Plasmodium protein unknown function
cnv12_367	PF3D7_1207000	conserved Plasmodium protein unknown function
cnv12_368	PF3D7_1207200	conserved Plasmodium protein unknown function
cnv12_377	PF3D7_1214900	conserved Plasmodium membrane protein unknown function
	PF3D7_1215000	thioredoxin peroxidase 2
cnv12_378	PF3D7_1216200	glycerol-3-phosphate dehydrogenase putative

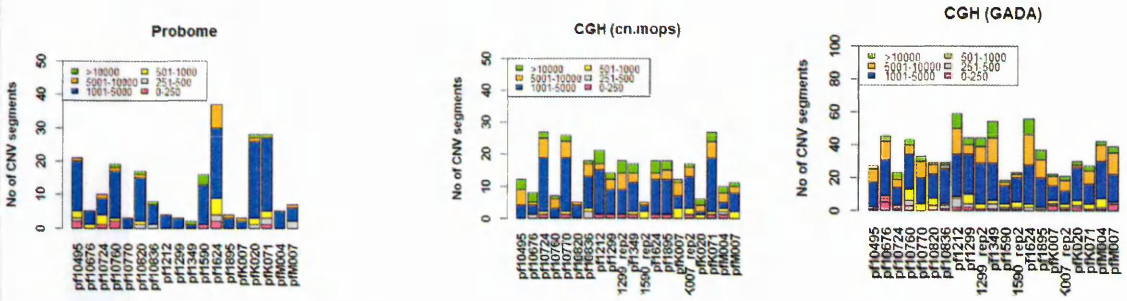
cnv12_379	PF3D7_1217500	conserved Plasmodium protein unknown function
	PF3D7_1217600	anaphase promoting complex subunit 10 putative
cnv12_388	PF3D7_1229300	conserved Plasmodium protein unknown function
	PF3D7_1229400	macrophage migration inhibitory factor
	PF3D7_1229500	T-complex protein 1
cnv12_398	PF3D7_1238900	protein kinase 2
cnv12_405	PF3D7_1243700	ubiquitin conjugating enzyme E2 putative
cnv12_408	PF3D7_1244700	conserved Plasmodium protein unknown function
cnv12_408	PF3D7_1244800	cytoplasmic translation machinery associated protein putative
cnv12_409	PF3D7_1245200	conserved Plasmodium membrane protein unknown function
cnv12_413	PF3D7_1248600	conserved Plasmodium protein unknown function
cnv13_428	PF3D7_1309000	conserved Plasmodium protein unknown function
	PF3D7_1309200	protein phosphatase 2c-like protein putative
	PF3D7_1309300	U4/U6 small nuclear ribonucleoprotein putative
	PF3D7_1309100	60S ribosomal protein L24 putative
cnv13_437_1/ cnv13_437_2	PF3D7_1315400	conserved Plasmodium protein unknown function
cnv13_442	PF3D7_1317300	conserved Plasmodium protein unknown function
	PF3D7_1317400	conserved Plasmodium protein unknown function
cnv13_447	PF3D7_1322300	translation initiation factor EIF-2B subunit related
cnv13_453	PF3D7_1324900	L-lactate dehydrogenase
	PF3D7_1325000	lsm6 homologue putative
cnv13_457	PF3D7_1329200	conserved Plasmodium protein unknown function
	PF3D7_1329300	chromatin assembly factor 1 subunit putative
cnv13_469	PF3D7_1340400	conserved Plasmodium protein unknown function
cnv13_473_1/ cnv13_473_2	PF3D7_1344600	lipoate synthase putative
cnv13_478	PF3D7_1348800	conserved Plasmodium membrane protein unknown function
	PF3D7_1349000	conserved Plasmodium protein unknown function
	PF3D7_1348900	conserved Plasmodium protein unknown function

cnv13_482	PF3D7_1351200	conserved Plasmodium protein unknown function
cnv14_517	PF3D7_1412500	actin II
cnv14_541	PF3D7_1431600	ATP-specific succinyl-CoA synthetase beta subunit putative
cnv14_543	PF3D7_1432200	conserved Plasmodium protein unknown function
cnv14_549	PF3D7_1438800	conserved Plasmodium protein unknown function
	PF3D7_1438900	thioredoxin peroxidase 1
	PF3D7_1439000	copper transporter putative
cnv14_564	PF3D7_1452700	U1 snRNA associated protein putative
	PF3D7_1452800	conserved Plasmodium protein unknown function
cnv14_573	PF3D7_1460700	60S ribosomal protein L27 putative
cnv14_573	PF3D7_1460800	conserved Plasmodium protein unknown function

A) B) C)

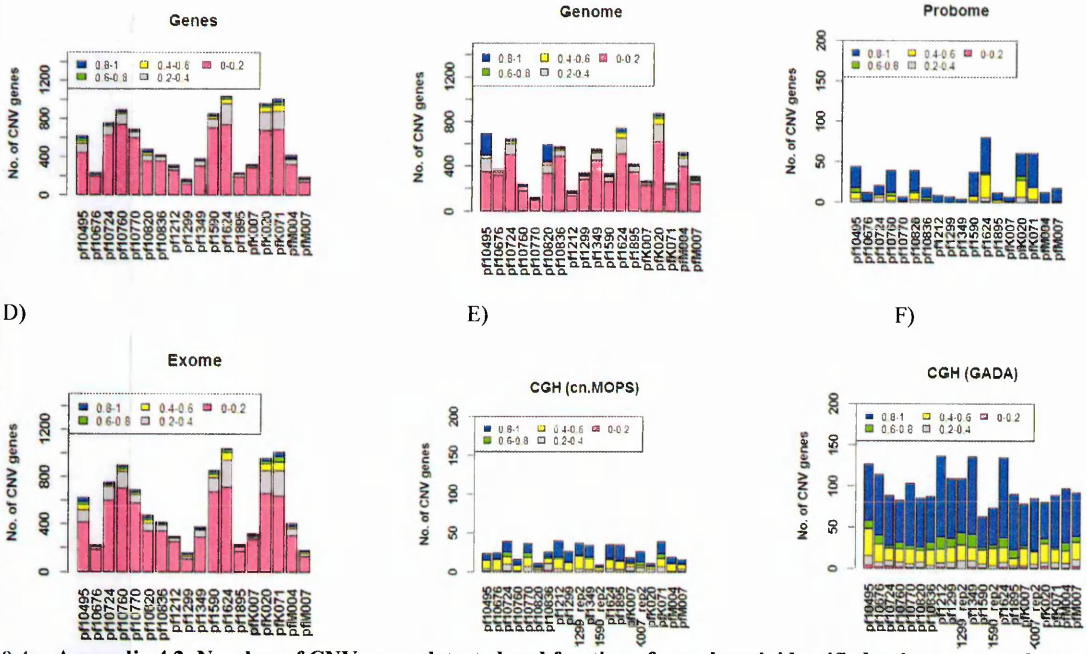


D) E) F)



Appendix 4.1. Number and length of CNV segments detected in the six analysis methods of CGH and sequence data
Barplots of the number of CNV segments detected in each of the 18 samples and the distribution of the length of the segments in the A) 'exome' B) 'genes' C) 'genome' D) 'probome' E) CGH (cn.MOPS) F) CGH (GADA). The lengths are indicated by red (0-250bp), grey (251-500bp), yellow (501-1000bp), blue (1001-5000bp), orange (5001-10000bp), green (>10000)

A) B) C)



D) E) F)

8.4 Appendix 4.2. Number of CNV genes detected and fraction of gene length identified to be copy number variable.
Barplot of the number of CNV genes identified in the six methods A) 'genes' B) 'genome' C) 'probome' E) 'exome' F) CGH (cn.MOPS) and G) CGH (GADA) with their corresponding fraction of gene length in CNV segments shown by different colours on bars 0-0.2 (pink), 0.2-0.4 (grey), 0.4-0.6 (yellow), 0.6-0.8 (green) and 0.8-1 (blue)