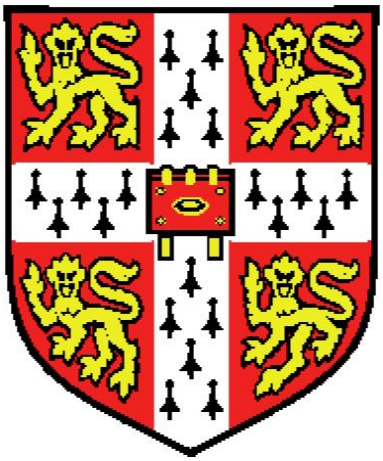
De novo assembly of the *var* multi-gene family in clinical samples of *Plasmodium falciparum*



Samuel Ayalew Assefa

Wolfson College

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

February 2013

Declaration

I hereby declare that this dissertation is my own work and contains nothing that is the outcome of work done in collaboration with others, except where specifically stated here and in the text.

Sequence data used in this thesis was generated at the Sanger Institute by Research and Development and Sequencing production teams.

None of the work presented has been previously submitted for the purpose of obtaining another degree. This dissertation does not exceed 60,000 words in length, as required by the School of Biological Sciences.

Samuel A. Assefa

February 2013

To my family

Acknowledgements

I am greatly indebted to my supervisors Matt Berriman and Chris Newbold for giving me the opportunity to conduct this research, their patience, guidance and support. I thank my university advisor Simon Tavare and members of my thesis committee Richard Durbin and Gavin Wright. Special thanks to Alex Bateman for his support as the director of studies and valuable advise; Julian Rayner for his enthusiasm and encouragement; Annabel and Christiana for their hard work with the PhD programme and their reminders.

I thank members of the parasite genomics team especially Thomas Otto for his help at various levels of the project; Mandy for helping with sequencing and libraries; Lia, Magdalena, Martin, Adam and James for their help with proof reading; Jacqui and her team for their Informatics support.

I am grateful for the support of Dominic Kwiatkowski his team for making the sequence data for the clinical samples used in Chapters 3 and 5 available. Special thanks to Magnus, Bronwyn, Olivia and Susy. I thank the Sequencing Production and Research and Development teams who have generated all the sequence data used in this thesis. Many thanks to the IT team, especially Tom Turner for their support during my PhD.

Many thanks to David, the trustees and supporters of the African Childrens Educational Trust (A-CET), who were instrumental in opening the first door that led to this PhD. I would also like to thank a number of people who have been supportive at various levels: Ed Johnson of Wolfson College for the lovely office space and interesting conversations; Pete Bull of the Kemri-wellcome Trust research

unit for hosting a productive work visit in Kilifi; Susy & Taane for countless dinner/movie nights and their support; members of my PhD year group especially Chinerye, Madushi, Jenn, Sathish and Luca; and the various people who offered to help during my RSI days.

Finally, I thank my family for their prayers and support, especially my brother Esubalew for his help with typesetting; and the Wellcome trust for the generous scholarship towards my PhD research.

Abstract

In the malaria parasite *Plasmodium falciparum*, *Pf*EMP1 (*Plasmodium falciparum* erythrocyte membrane protein 1) is a protein that is exported to the surface of infected human red blood cells and encoded by ∼60 *var* genes. *Pf*EMP1 plays a crucial role in parasite virulence and pathogenesis. It is also a target of host protective antibody responses that are avoided by the parasite by transcriptional switches between members of the *var* gene family resulting in antigenic variation of the surface expressed *Pf*EMP1.Thousands of malaria patient samples are being sequenced at the Sanger Institute's Malaria Programme to identify common polymorphisms but paradoxically some of the most variable sequences, such as *var* genes, are intractable due to high levels of polymorphism. Our understanding of *var* diversity in natural populations is thus limited to the DBLα domain, a conserved 300-400 bp region found in the majority of *var* genes studied so far. This thesis describes novel approaches developed to assemble full-length *var* genes from short reads of the Illumina sequencing platform.

The first part details an evaluation of existing assembly approaches through a comparative assessment of representative assembly tools. The results suggest that assembly of *var* genes in clinical samples using current methods is not practicaldue to a combination of factors including inherent sequence features (eg. high A+T content, low complexity, repeats and duplicates) and technical issues that affect quality of the raw sequence (eg. sequencing errors and uneven coverage). An alternative assembly strategy based on conserved sequence motifs was developed to address limitations of existing methods.

The second part investigates applications of short read sequencing to understand mechanisms of *var* gene diversity. Analysis of sequences from five progeny of the first genetic cross in *Plasmodium falciparum* between clones 3D7 and HB3 revealed evidence of ectopic-recombination as a mechanism for *var* gene diversity.

In the third and final part of the thesis, the iterative assembly approach developed in the first part is applied to a global collection of ∼800 clinical isolates resulting in the first and largest collection of full-length sequences for ∼50,000 *var*-contigs. Assembly results of *var* genes from these clinical samples were shown to have a higher repertoire-completeness (i.e. the number of contigs identified as *var* genes was close to the expected number of *var* genes), and contiguity (i.e. contig N50 size, largest contig size and open reading frame sizes were comparable with the expected values from previously completed genomes such as 3D7). Such availability of full-length *var* genes is a major progress towards understanding the population structure and diversity of *var* genes in natural populations with

Preliminarily analysis of *var*-contigs based on nucleotide and amino acid similarities (Chapter 5) revealed distinct clusters of highly conserved *var*-contigs within and between populations with percent-identities of up to 100% over their full length (i.e a match length of ∼5-10 kb). The validity of these continent-transcending *var*-contigs was confirmed by looking at the sizes of open reading frames and aligning short reads back to *var*-contigs. Potential reasons for such continent-transcending *var*-contigs are explored in Chapter 6. These observations were surprising and potentially interesting as the majority of continent-transcending *var*-contigs were members of a group of *var* genes that are known to be associated with severe malaria.

Contents

List of Figures

List of Tables

Abbreviations

aa	amino acid
ACT	Artemis Comparision Tool
AT (A+T)	Adenine and Thiamine
BAM	Binary Alignment and Map
bp	base pair
CTV	Continent-transcending *var*-contig
DBL	Duffy binding-like
DNA	Deoxyribonucleic acid
DSB	double strand break
DSBR	double-strand break repair
G+C	Guanine and Cytocine
GA	Genome analyser
kb	Killo base
Mb	Mega base
nt	nucleotide
ORF	Open reading frame

PCR	Polymerase chain reaction
PfEMP1	*Plasmodium falciparum* erythrocyte membrane protein 1
RAM	Random Access Memory
SAM	Sequence alignment and Map
SDSA	Synthesis Dependent Strand Annealing
SNPs	Single nucleotide polymorphisms
TAR	transformation-associated recombination

Chapter 1

Introduction

1.1 Malaria

1.1.1 Overview

Malaria is a disease that is caused by *Plasmodium* parasites and one of the oldest enemies of humanity, with a recorded history of its symptoms dating back for ∼4,000 years (Neghina et al., 2010). It is believed to have originated in Africa in the early days of human history and migrated to Euroasia as humans travelled to look for a better life. Malaria found a more stable reservoir of infection as early humans adopted a new lifestyle involving farming and agriculture, which allowed them to settle in a single area for longer periods of time (Webb, 2008).

Plasmodium falciparum (P. falciparum) is the deadliest species of human malaria parasites and claims over a million lives every year, of which nearly 80% are children under the age of five. A recent study has also shown an increase in number of adult deaths due to malaria, in both African and non-African countries (Murray et al., 2012). The majority of cases (∼60%) and deaths (>80%) occur in regions of sub-Saharan Africa where the disease continues to have a significant impact. There is an estimated ∼1.3% average annual reduction in economic growth for those countries with the highest disease burden (Greenwood, 2005; RBM, 2010; WHO, 2011). Four other species of *Plasmodium: P. vivax*, *P. ovale*, *P. malariae* and *P. knowlesi*, are also known to infect humans, although rarely cause fatalities. *P. knowlesi* is a zoonotic species that primarily infects

macaques, but has also been shown to cause severe infections in humans mainly in South East Asia (Pain et al., 2008). *P. falciparum* is transmitted by the female *Anopheles* mosquito, which is the definitive host.

Despite the current global efforts towards malaria elimination, over 200 million cases were estimated in 2010 (WHO, 2011) and nearly half of the world's population remains at risk (Figure 1.1). Insecticide and drug resistance are the two major threats in malaria control (Anderson, 2009; Cheeseman et al., 2012; Ranson et al., 2009).

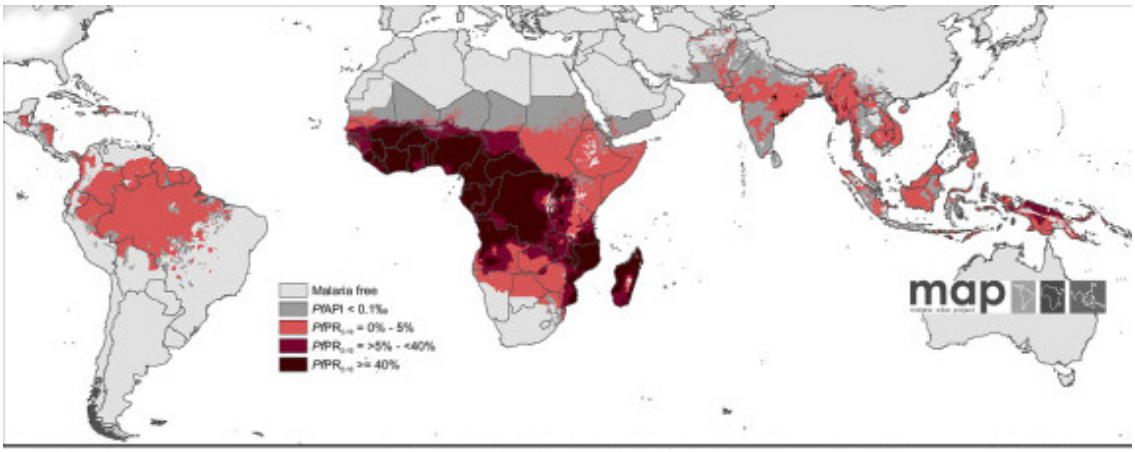


Figure 1.1: The spatial distribution of *P. falciparum* malarial endemicity in 2010. The map shows endemicity predictions based on *P. falciparum* Parasite Rates (*Pf*PR). Predictions were categorized as low risk *Pf*PR_2-10 $\leq$ 5% light red; intermediate risk *Pf*PR_2-10>5% to <40%, medium red; and high risk *Pf*PR_2-10 $\geq$ 40%, dark red. The rest of the land area was defined as unstable risk (medium grey areas, where *Pf*API<0.1 per 1,000 pa) or no risk (light grey). Adapted from (Gething et al., 2011).

1.1.2 The Parasite: *P. falciparum*

The genome

The *P. falciparum* genome was one of the first eukaryotic pathogen genomes to be sequenced, despite the challenges encountered due to its high A+T content (80.6% in protein coding regions and ~90% in non-coding regions). The 3D7

clone of *P. falciparum*, that is culturable throughout the bloodstream stages of its life cycle was chosen for the sequencing project (Gardner et al., 2002). The genome sequence revealed a total of ∼23.3 Mb organized into 14 chromosomes ranging in size from ∼640 kb to ∼3.3 Mb. The current annotation of the genome has ∼5,500 genes including 145 pseudogenes (Bohme, U. personal communication). A large proportion of the proteins lack similarity with known proteins from other organisms, suggesting a *Plasmodium*-specific role (Gardner et al., 2002). The presence of multi-copy gene families was also confirmed, and their actual number revealed, from the genome sequence. The three major gene families with confirmed and predicted expression properties on infected red blood cells (iRBCs) include the *var*, *rifn* (repetitive interspersed family, (Kyes, 1999)) and *stevor* (subtelomeric variable open reading frame, (Cheng et al., 1998)) multigene families. These genes are mainly located in subtelomeric regions and accounted for ∼7% of the genes in genome. *Var* genes are the focus of this thesis, as such; an overview of their organization, function and association with disease phenotypes is given in the following sections.

Life Cycle

Human malaria parasites require both the mosquito and human hosts to complete their life cycle (Figure 1.2). Development within the human host is initiated by a mosquito bite that releases sporozoites from the mosquito's salivary glands (Figure 1.2,A). Malaria parasites migrate to the liver where they invade, differentiate and multiply in hepatocyte liver cells. After ∼6 days, hepatocytes burst releasing merozoites into the blood stream (Figure 1.2, #3/4).

Once in the blood stream (Figure 1.2B), merozoites force themselves to enter red blood cells (RBCs) using a gliding motion (Tilley et al., 2011). The process creates a vacuole (*Parasitophorous Vacuole*) that expands to accommodate as the parasite asexually multiplies inside the infected red blood cell (iRBC) (Baumeister et al., 2009). Initially, parasites assume flat disc-like structures leading to the "ring stage" in their development, and further develop into "mature trophozoites" where they actively modify iRBCs in order to evade the host immune system and avoid clearance by the spleen (Haldar and Mohandas, 2007; Maier et al., 2009; Pasternak and Dzikowski, 2009). Formation of schizonts signals the

final stage of the intra-erythrocytic develpoment cycle, where parasites differentiate producing 16-32 new merozoites that develop inside the parasitophorous vacuole until ∼48 hr post invasion. Rupture of Schizonts releases merozoites that are able to infect new RBCs.

Some merozoites develop into male and female gametocytes (Figure 1.2, #7), ready to be taken by a mosquito during a blood meal. In the "mosquito stages" of the development (Figure 1.2C), fertilization of gametocytes within the mosquito midgut results in formation of ookinetes, which in turn traverse the gut wall to form oocysts. Oocysts rupture resulting in the release of sporozoites, which migrate to salivary glands of the mosquito.

Disease pathology and symptoms of malaria such as fever, headache, chills and muscle ache are a result of parasite development within the human blood stages (Figure 1.2B), as liver stages are asymptomatic (Miller et al., 2002). Severe complications of infection include anemia, respiratory distress and cerebral malaria in highly endemic areas or single or multi-organ failure in areas of very low endemicity (Andrej Trampuz, 2003; Miller et al., 2002; Milner et al., 2008). The virulence of *P. falciparum* is partly explained by the ability of mature parasites to export proteins to the surface of iRBCs in order to modify the iRBC and mediate adhesion to a variety of host cell types (MacPherson et al., 1985; Miller et al., 2002) as described in the following sections.

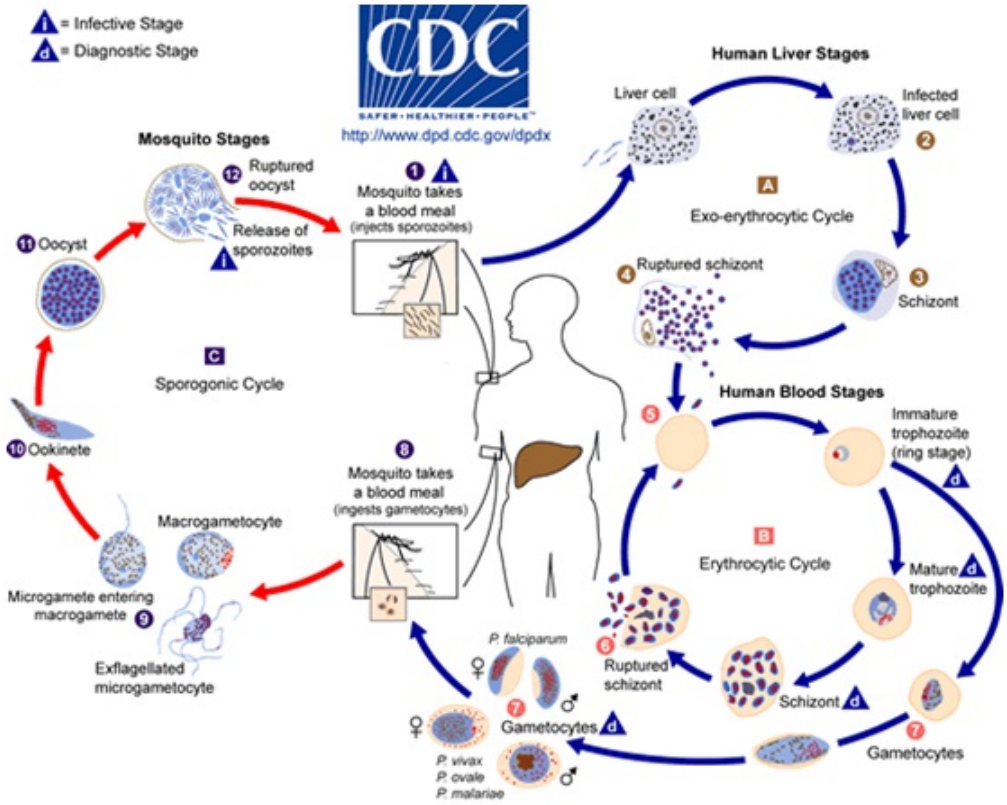


Figure 1.2: The life cycle of the human malaria parasite *P. falciparum* within two hosts (Taken from CDC: `http://www.cdc.gov/malaria/about/biology`)

1.2 *Plasmodium falciparum* erythrocyte membrane protein 1 (*Pf*EMP1)

Approximately 18 hours post red cell infection, *P. falciparum* parasites express variants of *Pf*EMP1 on the surface of iRBCs (Leech et al., 1984). It was shown that variants of this large surface protein (∼200-400 kDa) are encoded by ∼60 members of the *var* gene family (Baruch et al., 1995; Smith et al., 1995; Su et al., 1995). *Pf*EMP1 is regarded as a major virulence factor for two reasons; it is responsible for cytoadherence (Hughes et al., 2009; Newbold et al., 1999), and it undergoes antigenic variation (Newbold, 1999; Scherf et al., 1998; Sue Kyes et al., 2001) as a means of immune evasion.

1.2.1 Cytoadherence

Cytoadherence is a process in which parasite proteins expressed on the surface of iRBCs mediate binding to a number of host cell receptors (Biggs, 1990). Parasites begin to alter the surface of the iRBC after 12-14 hours post-infection. Modifications important for the parasite include changes in the shape of iRBCs resulting in an increased rigidity that restricts the ease of movement in the blood stream. An increased permeability of the cell membranes also occurs in order to allow acquisition of nutrients. As parasites continue to mature, further changes occur including appearance of electron dense protrusions (knobs) on the surface of iRBCs (Figure 1.3). Knobs are mainly composed of proteins known as knob-associated histidine rich proteins (KAHRP) (Sharma, 1991). These knobs (Figure 1.3) and the proteins exposed on the surface of iRBCs play a crucial role in pathogenesis (Fairhurst et al., 2012; Pasternak and Dzikowski, 2009).

The extracellular adhesive domains of *Pf*EMP1 (see later for a detailed description) bind to human endothelial cell receptors such as the scavenger receptor protein "Cluster of Differentiation 36" (CD36) and "Intercellular adhesion molecule 1" (ICAM-1) (Flick and Chen, 2004). As a result, iRBCs adhere to endothelial cells lining the small vasculature, do not circulate in the blood, and therefore avoid clearance by the spleen. When parasites are sequestered

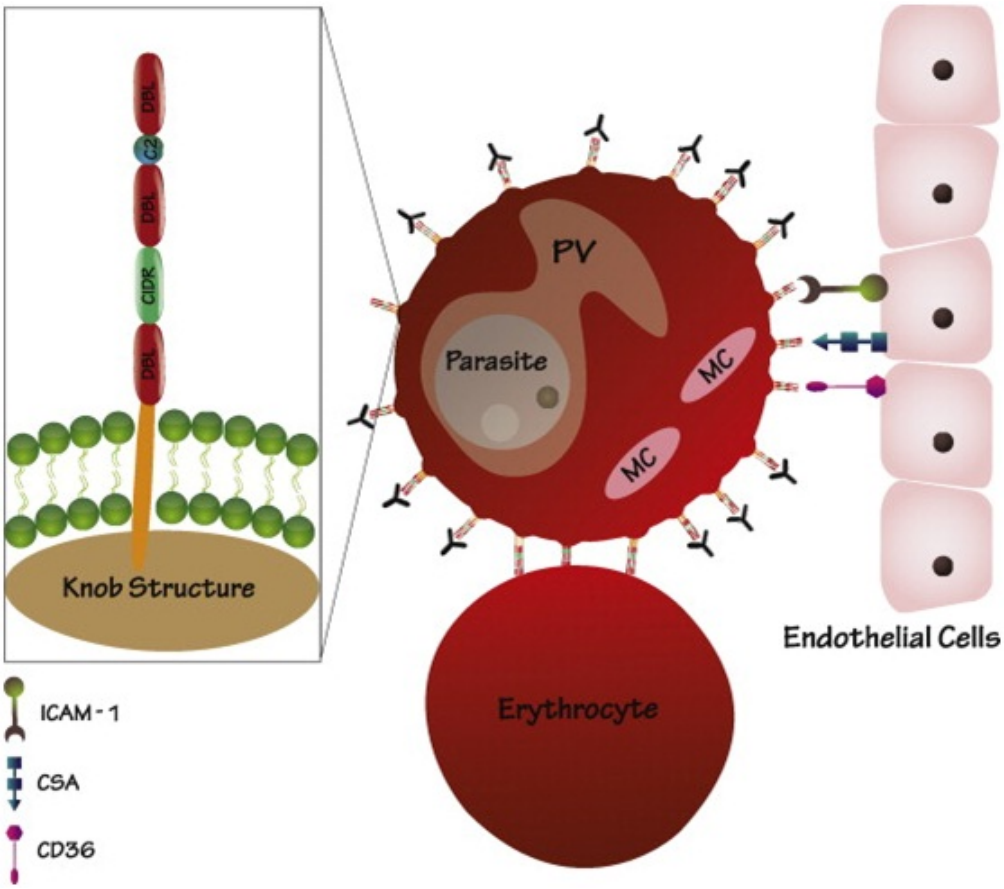


Figure 1.3: *Pf*EMP1 mediates adhesion of the infected RBC. *Pf*EMP1 is expressed by the malaria parasite *P. falciparum* on the knobs formed on the surface of infected erythrocytes. The variable extracellular regions DBLs and CIDR mediate adhesion through binding to several endothelial receptors such as CD36, ICAM1 and CSA. In addition, *Pf*EMP1 mediates adhesion to uninfected erythrocytes forming rosettes. (PV, parasitophorous vacuole; MC, Maurer's cleft). Taken from Pasternak and Dzikowski (2009)

in particular organs (such as the brain) then organ function is compromised and specific syndromes (such as cerebral malaria) may result. In addition, some iRBCs also bind to uninfected RBCs (a phenomenon known as rosetting) and form clumps (autoagglutination) a phenotype that has been associated with severe disease (Miller et al., 2002; Montgomery et al., 2007; Rowe et al., 2009)

1.2.2 Antigenic variation

Antigenic variation is used by a number of pathogens to continually change the antigenic epitopes that are exposed to the immune system (Sue Kyes et al., 2001). A common characteristic of parasites that undergo antigenic variation is the expression of new sub-populations of antigens at regular intervals in order to prolong the duration of infection.

First reported in the monkey malaria parasite *P. knowlesi* (using rhesus macques), antigenic variation was then observed in other malaria causing parasites such as *P. fragile* (in toque monkeys), *P. chabaudi* (in rodents) and later in *P. falciparum* (Sue Kyes et al., 2001).

The parasite's success in evading the host immune system is primarily attributed to its ability to effectively maintain diversity via antigenic variation. *P. falciparum* achieves such variation within a single genotype by transcriptional switching between different *var* genes (Recker et al., 2011; Roberts et al., 1992; Scherf et al., 1998). As a result, the parasite causes a chronic infection through reduction or avoidance of detection by the host immune system (Chookajorn et al., 2008; Frank and Deitsch, 2006; Kraemer and Smith, 2006; Scherf et al., 2008).

1.2.3 Regulation of gene expression

In order for the process of antigenic variation to be advantageous to the parasite it is evidently essential that the repertoire of potential variants are mostly hidden from the host at any one time. Thus *var* genes are expressed in a mutually exclusive fashion such that at any one time each parasite is only expressing one member of the family (Scherf et al., 2008). This mutually exclusive expression

of the ∼60 *var* genes is believed to be regulated by mechanisms that involve genetic factors and epigenetic elements, including histone modification and organization of *var* genes around the nuclear periphery (Kyes et al., 2007; Scherf et al., 2008). Regulation at the genetic level is believed to involve the use of two promoter regions, the first in the upstream regions of exon 1 of *var* genes and the second in the introns (Calderwood et al., 2003; Swamy et al., 2011; Voss et al., 2006). Epigenetic factors on the other hand involve modification of chromatin around *var* genes in order to maintain an epigenetic memory of the genes activated during successive cycles of cell division. Activated *var* genes are particularly associated with a loosely-packed chromatin with modifications including acetylated histone 3 lycine 9 (H3K9ac), di-methylated H3K4 (H3K4me2) and tri-methylated H3K4 (H3K4me3). Conversely, promoter regions of non-activated genes are surrounded by a tightly packed chromatin with tri-methylated H3K9 (H3K9me3) modifications (Chookajorn et al., 2007; Enderes et al., 2011; Hernandez-Rivas et al., 2010).

The localisation of genes around the nuclear periphery is also believed to provide additional epigenetic mechanisms for a mutually exclusive expression (Freitas-Junior et al., 2000). Clusters of silent genes are primarily located in the nuclear periphery surrounded by heterochromatin and only one gene moves to a different area that facilitates transcription (Dzikowski et al., 2007). The organisation of *var* genes in silent regions leads to bouquet like structures that facilitate recombination and gene conversion events, thus contributing to the generation of immense diversity in the *var* repertoire (Freitas-Junior et al., 2000).

1.2.4 Organization, chromosomal location and grouping of *var* genes

1.2.4.1 *Var* gene organization and Protein architecture

The first complete view of a repertoire of *var* genes was obtained from the genome of the reference isolate 3D7 (Gardner et al., 2002). Later, *var* genes from the IT and HB3 genomes were also compared with 3D7, revealing a comparable number of ∼60 protein coding genes per genome (Kraemer et al., 2007). In

the 3D7 genome, an additional ∼30 genes are annotated as pseudogenes and truncated exons of *Pf*EMP1. Although not fully understood, there is increasing evidence for pseudogene transcription (Otto et al., 2010b).

Var genes have a common structure that constitutes two exons. The first exon (∼3 to 9.4 kb) encodes the highly variable extracellular region of *Pf*EMP1. Exon 2 is shorter than exon 1 and encodes a semi-conserved intracellular region and a trans-membrane domain. Introns separating the two exons have a high A+T content and a size of up to ∼1.2 kb.

The protein encoded by exon 1 is composed of an N-terminal segment (NTS), variable numbers of the Duffy Binding like (DBL) and Cysteine Rich Interdomain Region (CIDR) domains. Although the order in which these domains appear is highly variable, the NTS is always found at the beginning of the protein followed by DBL and CIDR domains. Because of the continuous exposure of these domains to the immune system and their recombinogenic nature, they are highly polymorphic (Taylor et al., 2000b). Further classification of the DBL and CIDR domains using few conserved residues reflects the high sequence diversity in the family. DBL domains are subdivided in to seven classes: $\alpha, \alpha_1, \beta, \delta, \varepsilon, \gamma$ and χ (six groups were defined by the genome project; the group α was later sub classified into α and α1) (Kraemer et al., 2007). Similarly, CIDR domains have four sub categories: $\alpha, \alpha_1, \beta, \gamma$ (the initial definition only contained α and not-α). The DBLα domain is the most abundant and found in almost all *var* genes next to the N-terminal segment. The overall most conserved *var* sequence can be found within the part of exon 1 that encodes DBLα (Kraemer et al., 2007; Kyes et al., 2007). Taylor and colleagues (Taylor et al., 2000a) described a set of universal primers that were able to amplify this conserved region of DBLα from the majority of *var* genes. These primers are routinely used to rapidly identify *var* genes in culture-adapted and clinical isolates.

The second exon has higher A+T content and greater sequence conservation than the first exon. The protein encoded by exon 2 is composed of a semi-conserved acidic terminal segment (ATS), which was occasionally used in detecting *var* genes before universal *var* primers from the DBLα domain were adopted. Despite the extreme diversity within and between *var* genes of *P.*

falciparum isolates studied so far (Barry, 2007; Bull et al., 2005; Chen et al., 2011; Fowler et al., 2002; Trimnell et al., 2006), most *var* genes have a relatively conserved head structure composed of NTS-DBLα-CIDR1 domains (Kraemer and Smith, 2006). Depending on the total number of domains, *var* genes could also be described as either short (with 2 to 4 domains) or long (with over 5 domains). There is a higher degree of variability in the number and order of domains that constitute each gene. A total of $\sim$17 different architectures were originally described based on the *var* repertoire of the 3D7 genome (Figure 4).

A comparative study of *var* genes in three lab isolates 3D7, IT and HB3 subsequently revealed a total of 31 architecture types (Kraemer et al., 2007). Although *var* genes could have a comparable number and order of domains, the sequence similarity between genes even with the same architecture is extremely low. The most conserved domain, DBLα, has a similarity of up to 50%. Of the 31 different architectures defined in the three isolates (3D7, IT and HB3), only seven were found to overlap. A recent analysis of *var* genes from seven genomes by Rask and colleagues (Rask et al., 2010) used a combination of phylogenetic trees and an iterative detection of homology blocks to define regions with high similarity (homology blocks) and Domain Cassettes (DC) in *Pf*EMP1 sequence. Domain Cassettes are conserved sequence elements defined by grouping genes that share a similar order of domain architectures. A total of 23 domain cassettes were identified from $\sim$400 *Pf*EMP1 sequences in the seven genomes. This study has improved domain boundaries and provided some unit of defining associations with disease phenotypes.

1.2.4.2 Chromosomal locations of *var* genes and their transcription

Subtelomeric regions are the most unstable parts of the genome, with recombination rates predicted to be approximately ten times higher than those of core regions (Taylor et al., 2000b). The location of $\sim$60% of *var* genes in subtelomeric regions may thus play an important role in maintaining a genetic diversity essential for the parasite's survival (Gardner et al., 2002). Telomeric ends contain one to three *var* genes per chromosome except in Chromosome 14, which only contains a pseudogene. The remaining 40% of *var* genes are located in central

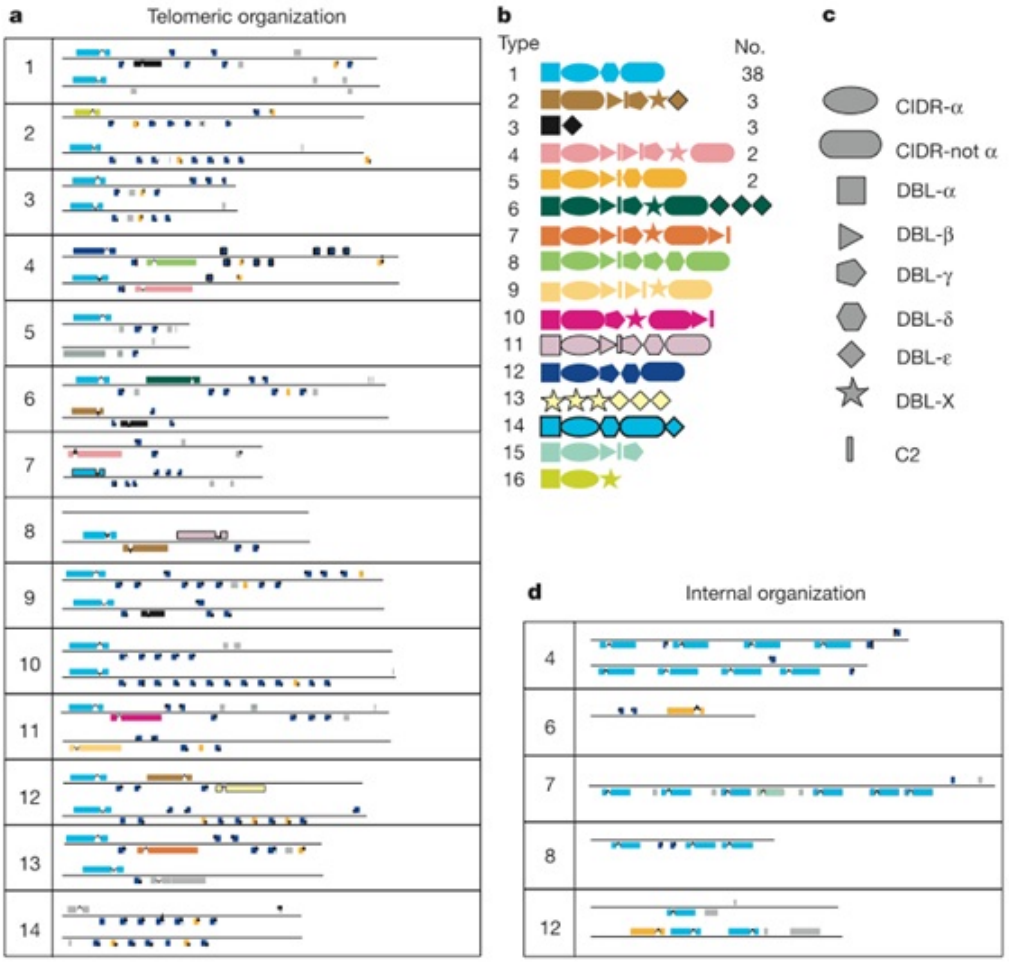


Figure 1.4: Organization of multi-gene families in *P. falciparum*. a, Telomeric regions of all chromosomes showing the relative positions of members of the multi-gene families: *rif* (blue) *stevor* (yellow) and *var* (colour coded as indicated; see b and c). Grey boxes represent pseudogenes or gene fragments of any of these families. The left telomere is shown above the right. Scale: ∼0.6 mm =1 kb. b, c, *var* gene domain structure. *Var* genes contain three domain types: DBL, of which there are six sequence classes; CIDR, of which there are two sequence classes; and conserved 2 (C2) domains (see text). The relative order of the domains in each gene is indicated (c). *Var* genes with the same domain types in the same order have been colour coded as an identical class and given an arbitrary number for their type (b) and the total number of members of each class in the genome of *P. falciparum* clone 3D7. d), Internal multi-gene family clusters. Key as in a. (Taken from Gardner et al. (2002))

clusters on chromosomes 4, 7, 8 and 12. Other multi-copy gene families that are located in subtelomeres and known to be expressed on the surface of iRBCs include *rif* and *stevor* genes, although their function is not fully understood (Jemmely et al., 2010).

A detailed survey of the location and direction of transcription of *var* genes showed that the majority of subtelomeric genes are positioned tail-to-tail separated by one or more members of the *rif* gene family. A few remaining genes assume a head-to-head or head-to-tail configuration (i.e. transcribed towards and away from each other respectively). Conversely, central *var* genes are almost always found in tandem arrays in a head-to-tail orientation (Gardner et al., 2002).

Var genes of the subtelomeres are located in close proximity to telomeric repeat units, specifically the six telomere-associated repeat elements (TARE 1 to TARE 6). A repeat unit known as rep20 is found next to *var* genes and its association with higher rates of recombination has been proposed (Jiang et al., 2011).

1.2.4.3 Grouping and classification of *var* genes

In order to better understand the sequence diversity, evolution and their association with disease phenotype, there have been several attempts to classify *var* genes into a small number of groups. In addition to domain architecture and chromosomal locations, sequences of upstream regions of *var* genes (Ups) were also used to group *var* genes (Lavstsen et al., 2003). Based on their sequence similarity Ups regions are grouped into four types: UpsA, UpsB, UpsC and UpsE (Kraemer et al., 2007). The original classification also contained another group (UpsD) that was subsequently merged with UpsA (UpsA2).

A combination of Ups sequence similarity, chromosome location and direction of transcription were used to define three major (A, B and C) and two intermediate (B/A and B/C) groups of *var* genes (Gardner et al., 2002; Kraemer et al., 2007). Type A and B *var* genes are both located in subtelomeric regions with upstream sequences of UpsA and UpsB respectively. However, they are transcribed in opposite directions (Type B to the center and Type A to telom-

eres). Type C represents *var* genes in the central regions of chromosomes with UpsC flanking sequence. The intermediate group B/A contains genes that are transcribed towards the telomeres (similar to Type B) but they are located towards the centromere end of subtelomeric regions. The remaining genes of type B/C are located in central regions with a flanking region (Ups) similar to UpsB sequences.

A different approach of *var* grouping that utilises DBLα tags was developed by Bull and colleagues (Bull et al., 2007). This approach first grouped sequences based on the number of cysteines into Cys2, Cys4 and CysX, representing DBLα sequences that contain two, four and other (one, three, five or six) residues. Four positions of limited variability (PoLVs), each containing four amino acid residues were then defined and used together with cysteine counts to classify sequences into six groups (groups 1 to 6). Although the classification was dependent on a small portion of the gene, the result of this typing method was highly comparable with existing groups defined using similarity of flanking (Ups) sequences and chromosome location. Sequences in groups 1 to 3 contain two cysteines (Cys2) and correspond with Types A, B/A and B. On the other hand, sequences in groups 4 and 5 were Cys4 types and correspond with B/A, B, B/C and C. The remaining CysX sequences of group 6 mainly correspond with type C and a small number of B/C and B genes.

1.2.4.4 Association with disease phenotypes

In vitro experiments conducted on parasites taken from mild and severe malaria infections have shown distinct interactions of *Pf*EMP1 domains and endothelial receptors. For example, binding between CD36 and CIDRα domains is observed in most cases of mild malaria, whereas binding of some DBLβc2 domains to ICAM1 was associated with severe complications (See Kraemer and Smith (2006) for a review). Pregnancy-associated malaria is a well studied example of cytoadherence and its complications (Fried and Duffy, 1996). It could lead to low birth weight and premature delivery in endemic areas as iRBCs sequester in the placenta of pregnant women that have not been exposed to antigens encoded by the gene var2CSA (Salanti, 2004). Ligands of var2CSA bind to

low-sulfate forms of Chondroitin Sulfate A (CSA) receptors in the placenta. Despite the occurrence of pregnancy-associated malaria during first pregnancy, immunity is quickly acquired for subsequent pregnancies. The high sequence conservation of var2CSA compared to other members of the family has raised hopes of a possible vaccine for pregnancy associated malaria (Chen, 2007; Rogerson et al., 2007).

Earlier studies conducted on African children from regions with variable levels of malaria transmission revealed that immunity to non-cerebral cases of severe malaria could be acquired at a young age with one to three infections. The time required for such immunity to develop may vary depending on transmission intensity, taking up to five years in areas of low transmission. Conversely, immunity to mild malaria takes longer and may never be achieved as parasites maintain a large repertoire of antigens that continue to evolve rapidly (Bull et al., 1998, 2000; Gupta et al., 1999). These observations have motivated a number of studies that aimed to find specific *Pf*EMP1 types that are highly expressed during severe and mild malaria (Ariey et al., 2001; Bull et al., 2005; Cham et al., 2010; Falk et al., 2009; Jensen, 2004; Kaestli et al., 2004, 2006; Kalmbach et al., 2010; Kirchgatter and Portilo, 2002; Kyriacou et al., 2006; Lavstsen et al., 2005; Montgomery et al., 2007; Nielsen et al., 2002; Normark et al., 2007; Rottmann et al., 2006).

A number of studies have shown that *var* genes that belong to groups Type A and B/A are associated with severe malaria infections (Falk et al., 2009; Kyriacou et al., 2006; Rottmann et al., 2006; Warimwe, 2009). However, binding properties of iRBCs especially in cerebral malaria remain poorly understood. Recently, three independent studies published in the same issue of the journal Proceedings of National Academy of Sciences (Avril et al., 2012; Claessens et al., 2012; Lavstsen et al., 2012) identified a group of mainly Type A *var* genes that are associated with severe cases of malaria including cerebral malaria. Genes that contained Domain Cassettes 8 and 13, as defined by Rask and colleagues (Rask et al., 2010) were found to bind to brain endothelial cells. Domain Cassette 8 sequences were of the groups A and B/A, while Domain Cassette 13 were primarily Type A *var* genes. Despite the significance of such findings, most studies still use the DBLα domain due to the difficulty of obtaining full length

sequence information. There is thus a need for quickly obtaining full length information of genes and transcripts for a better understanding of *Pf*EMP1's association with severe disease phenotypes. Such understanding may facilitate the search for intervention, especially for severe malaria infections (Chen, 2007; Hviid, 2011). It is however important to note that due to the high polymorphism in *var* genes, a great deal of further research is required to establish the feasibility of such interventions for example in the form of a 'severe malaria vaccine'.

1.2.5 Polymorphism, sequence diversity and mechanisms of generating diversity in *var* genes

Polymorphism and sequence diversity

Despite the importance of understanding the diversity of *var* genes in natural populations, our knowledge is still limited primarily for two reasons. Firstly, the highly recombinant nature of *P. falciparum* parasites in general and *var* genes in particular makes analysing and describing diversity of the gene family in natural populations extremely difficult (Bull et al., 2008; Conway, 1999; Gardner et al., 2002; Kraemer et al., 2007). Secondly, the highly polymorphic nature of *var* genes has hindered the possibility of obtaining full length regions of the repertoire using existing laboratory based methods (eg. PCR based amplification and capture of *var* genes). Universal *var* primers developed to amplify the DBLα region (Taylor et al., 2000a) are used by all except two (Kraemer et al., 2007; Rask et al., 2010) of the previous studies on *var* gene diversity. Recently, a modified transformation-associated recombination (TAR) cloning method was used to capture telomeric and central *var* gens (Gaida et al., 2011). Although this constitutes a step forward towards analysing full length genes compared to using DBLα tags, its application with large scale studies of natural populations is very limited.

A few studies have however tried to explore the diversity of *var* genes within these limitations. Evidence from all the studies points to the presence of extreme diversity in the *var* gene family (Barry, 2007; Bull et al., 2008; Chen et al., 2011; Mugasa et al., 2012; Ozarkar et al., 2009) with a very low sequence similarity between *Pf*EMP1 domains within and between isolates (typically below 50%).

Although initially such extreme levels of diversity may appear to be a result of random and potentially unlimited recombination between *var* genes (Barry, 2007), the idea of a recombination hierarchy (Kraemer and Smith, 2003) was later confirmed revealing two recombinationally isolated groups (Type A and non-type A) (Bull et al., 2008; Kraemer et al., 2007). Such a hierarchy is believed to restrict recombination possibilities between domains in different groups.

An overview of mechanisms used to generate *var* diversity

Understanding the mechanisms employed by parasites to generate such immense diversity in natural populations is one of the least explored topics in the area of *var* genes. The complex lifecycle of the parasite involves multiple stages of cell division in both the human and mosquito hosts, thus making such studies extremely challenging. Diversity at the basic molecular level is mainly generated by three major processes: mutation, homologous recombination and non-homologous gene conversion.

Mutation

Mutation is known as the ultimate source of genetic diversity as it is the only process capable of creating new sequences while the other two primarily shuffle existing sequence fragments. The extent of changes may vary from substitutions of a single nucleotide base and small insertions/deletions (indels) to large scale complex changes. Single nucleotide polymorphisms (SNPs) are the commonly studied events of mutation in natural populations of *P. falciparum*. The most recent study by Manske and colleagues (Manske et al., 2012) presented SNPs from a global collection of clinical isolates. Although new insights on population structure and the extent of polymorphism were obtained from the study, highly variable regions such as *var* genes were excluded due to the difficulty of reliably aligning short reads in these regions.

Homologous recombination (HR) and Gene conversion (GC)

HR refers to the reciprocal exchange of genetic material between two allelic regions of the genome. Although HR does not create new sequences as in the case of mutation, it plays a crucial role during both meiosis and mitosis. In

meiosis, the eukaryotic HR machinery is activated following a double strand break (DSB) as a result of actions of the enzyme spo11 (San Filippo et al., 2008). The main functions of HR include allowing proper segregation of chromosomes during cell division and generating diversity in progeny by providing a means of genetic exchange. In mitosis, HR is primarily used to repair DSBs due to damages and as a result of stalled replication forks. Enzymes responsible for the proper pairing of homologous regions during HR include Rad51 and Dmc1. The Double Strand Break Repair (DSBR) model was the first attempt to understand the process of HR (Lieber, 2010; Szostak et al., 1983) . Despite its limitations in explaining mitotic events where the majority of DSB repairs do not result in homologous cross-overs, the DSBR is still widely accepted model for meiotic HR. The Synthesis Dependent Strand Annealing (SDSA) model was proposed to account for mitotic non-cross over events (Figure 1.5). The HR machinery in *P. falciparum* and the genes involved are not well understood. A Rad51 homologue in *P. falciparum*, *Pf*Rad51, was the first Rad51 gene to be characterised in apicomplexan parasites(Kantibhattacharyya et al., 2004). Conversely, gene conversion is a non-reciprocal transfer of genetic material between non allelic (ectopic) regions where a homologous sequence from a donor region is used to replace a damaged region (acceptor).

In *P. falciparum*, three genetic cross experiments were used to better understand mechanisms and rates of recombination. The first genetic cross was made in 1987 between clones 3D7 and HB3 (Walliker et al., 1987). The two other crosses were later reported: HB3 with DD2 (wellems et al., 1990) and GB4 with 7G8 (Hayton et al., 2008). Such studies require a complex experimental setup to capture the complete life cycle in the mosquito (where sexual development and homologous recombination takes place) and a mammalian model organism (where asexual reproduction takes place). Both homologous recombination and gene conversion are believed to generate *var* gene diversity (Frank et al., 2008; Freitas-Junior et al., 2000; Taylor et al., 2000b). However, there is yet no evidence of mitotic ectopic gene conversion events.

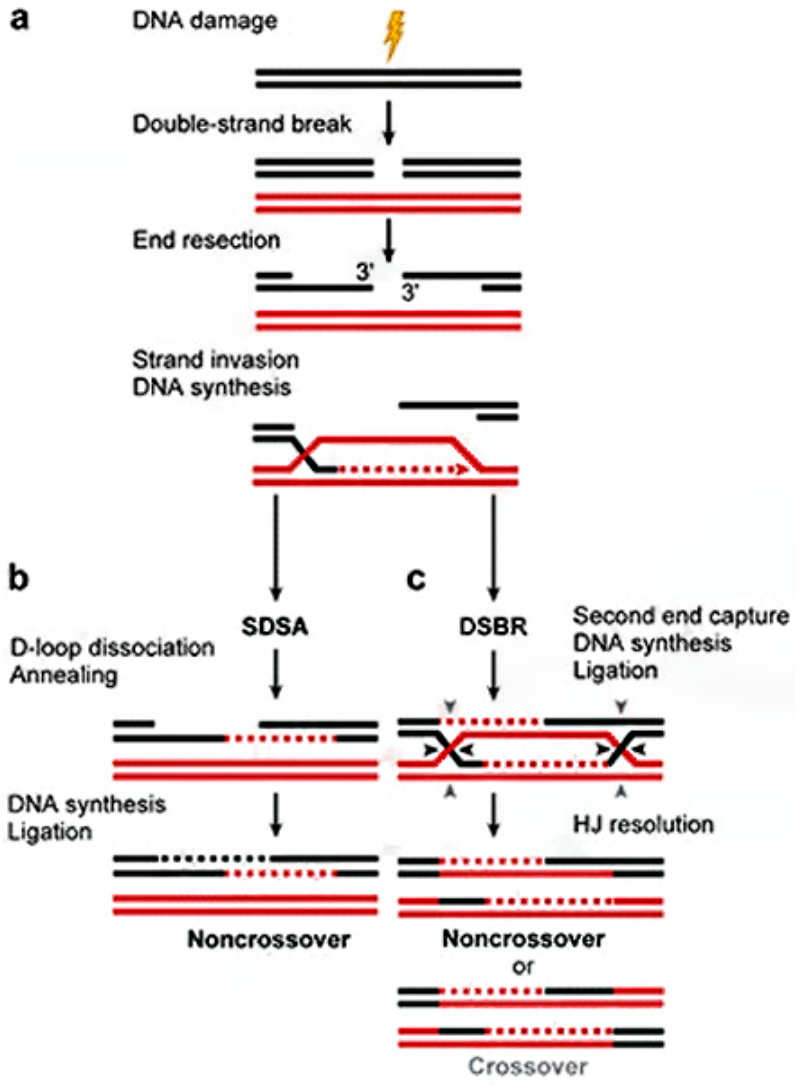


Figure 1.5: Pathways of DNA double-strand break repair by homologous recombination. Double-strand breaks (DSBs) can be repaired by distinctive homologous recombination (HR) pathways, such as synthesis-dependent strand annealing (SDSA) and double-strand break repair (DSBR). (a) After DSB formation, the DNA ends are resected to yield 3′ single-strand DNA (ssDNA) overhangs, which become the substrate for the HR protein machinery to execute strand invasion of a partner chromosome. After a successful homology search, strand invasion occurs to form a nascent D-loop structure. DNA synthesis then ensues. (b) In the SDSA pathway, the D loop is unwound and the freed ssDNA strand anneals with the complementary ssDNA strand that is associated with the other DSB end. The reaction is completed by gap-filling DNA synthesis and ligation. Only noncrossover products are formed. (c) Alternatively, the second DSB end can be captured to form an intermediate that harbours two Holliday junctions (HJ)s, accompanied by gap-filling DNA synthesis and ligation. The resolution of HJs by a specialized endonuclease can result in either noncrossover ((*horizontal triangles*) or crossover products ((*vertical triangles*).(Taken from San Filippo et al. (2008))

1.3 New frontiers and existing challenges

Advances in DNA sequencing technologies have enabled a better understanding of the parasite biology via '*re-sequencing*' and '*de novo sequencing*' applications (Bentley, 2006; Mardis, 2008). Re-sequencing is where a reference sequence is available and the aim is comparison of the new sequences (whole genome or parts of a genome) with the existing sequence. Alternatively, *de novo* sequencing does not assume the presence of a reference sequence. It is therefore applicable in the sequencing of a new genetic material or when the material is significantly different from what is already known.

Re-sequencing of parasites sampled from clinical patients is being used to understand the diversity and structures of natural populations. Nonetheless, existing challenges of drug resistance, insecticide resistance and lack of effective vaccine continue to pose major threats to malaria control efforts. DNA sequencing technologies may help us monitor outbreaks and emergence of drug resistance more effectively than traditional methods (Le Roch et al., 2012)

1.4 Next generation sequencing technologies and short read assembly

1.4.1 Second generation sequencing technologies

The most commonly used second generation instruments such as those provided by Roche (454 platform `http://www.roche.com`) and Illumina (GA1, GA2, HiSeq, MiSeq platforms: `htp://www.illumina.com`) are able to sequence millions of DNA fragments using a highly parallel Sequencing-by-synthesis process. In this thesis, the Illimina's GA2, HiSeq and MiSeq platforms were used to sequence laboratory clones and clinical isolates of *P. falciparum*.

Illumina sequencing

The Illumina sequencing technology uses cyclic reversible termination chemistry described by (Bentley et al., 2008). The length of short sequences (reads) generated by the platform is determined by the number of cycles. The sequencing-

by-synthesis process used by Illumina involves three main steps: library preparation, cluster generation and sequencing.

Initially, the sample DNA is fragmented using nebulization or sonication followed by an end-repairing step to generate blunt-ended fragments. Addition of a single nucleotide Adenine (A) base to the 3′ end of both DNA strands produces A-tailed fragments that are ready to be ligated with sequencing adaptors. The adaptors have a 3′ Thymine (T) overhang that complements the A-tails of template fragments. The final stages of library preparation include size-selection and quantification of fragments that contain the sequencing adaptor. The sequencing library is then transferred to flow cells. Flow cells contain oligonucleotides that are complementary to the sequencing adaptors ligated at the end of the templates such that template fragments bind to the surface where both cluster generation and sequencing take place. Paired-end reads are generated by sequencing the template from both ends, resulting in reads that are separated by a known fragment size. Illumina's flow cells contain eight lanes that are capable of taking independent samples (libraries) or up to 96 multiplexed libraries.

The cluster generation step first denatures fragments into single stranded templates followed by cycles of a bridge-amplification step, where each template is clonally amplified resulting in thousands of templates per cluster and millions of clusters per lane. Each cluster corresponds to a single template molecule.

The cyclic sequencing process involves the use of four fluorescently labeled nucleotides (dNTPs), which are added to the growing chain of sequenced (synthesised) bases for each template of a cluster. After each incorporation, the intensity of the fluorescent dye is measured via imaging techniques and used to determine the exact base for each cluster. Finally, the terminator containing the fluorophore are removed to allow another cycle of incorporation.

The Illumina system is distributed with a base calling software, Bustard, that analyses the signal from each template in a cluster in order to determine the correct base call after each incorporation cycle. Three major issues were reported to affect accuracy of base calls in the Illumina system (Erlich et al., 2008; Wei-Chun Kao, 2009). Firstly, incorporation of multiple bases per cycle or missed-

incorporation may result in longer (leading) or shorter (lagging) synthesised strands than the majority of the templates. As the frequency of templates that are affected by such phasing issues increases, the accuracy of base calls reduces. Secondly, with an increase in cycle number, the intensity of fluorescent signals decays resulting higher error rates towards the ends of reads. Finally, the cross-talk effect may also cause substitution errors towards read-ends. For these reasons, the maximum read length of the Illumina data presented here is 100 bp (HiSeq). In addition, studies in our group have shown a higher concentration of Illumina sequencing errors immediately after a homopolymer tract (Otto et al., 2010a).

1.4.2 Short read assembly

1.4.2.1 Overview of algorithms

Sequences obtained from the Sanger/Capillary sequencing platforms had fewer fragments covering larger stretches of genomic regions. The advantages of such data include the ease of reconstructing the genome due to longer reads spanning across difficult regions such as repeats. Data storage and run-time memory requirements are also minimal during the assembly process. Despite the high cost and slow sequencing time, read lengths of up to 1 kb were obtained from Sanger sequencing methods (Shendure and Ji, 2008). The Illumina platform on the other hand generates millions of short reads ($\sim$100 bp) at a fraction of the cost. Although the significant increase in the number of reads could provide additional benefits in terms of enhancing the sequence information for a given region (i.e. coverage), it also introduces new challenges of data storage and sequence analysis especially in alignment and assembly of short reads. *De novo* assembly of sequence fragments is among the most difficult computational problems (also known as Non-deterministic polynomial hard/NP-hard problems) that do not currently have efficient solutions (Myers, 1995). The most common formalisation of the assembly problem is to consider short reads as strings and solve the problem of 'shortest common superstring' where the aim is to find the minimal superstring that contains all fragments (short strings). Because such approximations are known to be computationally intractable (NP

hard), a number of assembly (Li et al., 2009b, 2010; Simpson, 2009; Zerbino and Birney, 2008) and scaffolding (Boetzer et al., 2011; Dayarian et al., 2010; Pop, 2003) tools apply algorithms that generate the best possible solution within an acceptable time.

Grouping assembly tools

Assembly tools could be grouped as greedy or graph-based based on the underlying strategy employed to process sequence fragments (Narzisi and Mishra, 2011).

Greedy algorithms have an incremental search strategy where they identify overlaps and sequentially extend sequences starting from the overlap with the highest score. Assembly tools developed for long capillary reads such as TIGR (Sutton et al., 1995), CAP3 (Huang and Madan, 1999), PCAP (Huang, 2003) and Phusion (Mullikin and Ning, 2002) employ a greedy searching algorithm and fall into this category.

Graph-based assembly tools use a string graph to represent overlapping sequence fragments prior to generating optimal solutions for the assembly problem (i.e. finding a path that passes through all nodes). Two main approaches have been used to construct overlap graphs in sequence assembly tools. The first method uses an overlap-layout-consensus approach where the vertices are full length sequences (reads) and edges represent overlaps between them. Due to the requirement for an all-against-all pairwise similarity check between input reads, this method was mainly used by assemblers developed for longer sequences such as Atlas (Havlak et al., 2004), ARACHNE (Batzoglou et al., 2002) and Celera (Myers et al., 2000). The Edena (Hernandez, 2008) short read assembler also used overlap-layout-consensus methods together with a graph-cleaning step that removes erroneous nodes before the assembly. Removing of nodes that are believed to be spurious is also known as pruning the assembly graph. SGA (String Graph Assembler) was the most recent short read assembler to use an overlap graph (Simpson and Durbin, 2012) with efficient data compression in order to handel large genomes. The second approach considers sub-fragments of length k (*k-mer*s) instead of the full length of reads to construct the assembly graph. In most short read assembly tools, reads are first broken

into overlapping *k-mer*s that are then represented by edges on a de Bruijn graph (Figure 1.6.d).

Implementations of the Eulerian path approach include EulerSR (Chaisson MJ, 2008), Velvet (Zerbino and Birney, 2008), ABySS (Simpson, 2009) and SOAP-denovo (Li et al., 2009b; Ruiqiang Li, 2010). A *k-mer*-based approach is generally preferred over full-length overlap graphs for high throughput sequence data due to its potential in dealing with large sequence data. Such efficiency in storage and processing is achieved by representing identical fragments (eg. repeats) in a single node (i.e. overlapping *k-mer*s) on the assembly graph. Although the number of nodes is effectively minimised, the significant number of connections between shared *k-mer*s adds complexity to the assembly graph. Memory requirements and quality of the final assembly vary with choice of *k-mer* size. While larger *k-mer* values require less run time memory (Random Access Memory, RAM) and could generate high quality assembly, a higher read coverage is also required.

Repeats pose a significant challenge in assembling of short reads. De Bruijn graph representations of sequences collapse repeat units into a single node that has multiple incoming and outgoing connections with other nodes (Treangen and Salzberg, 2011). Such sequences are the primary cause of ambiguities in searching for the single shortest path that connects all nodes and result in a highly fragmented assembly. Kingsford et al. (2010) show the limitations of short reads in reconstructing genomes using graphs constructed from complete bacterial chromosomes. Although the data are far from ideal in terms of representing real sequencing output, it provides an insight into the theoretically achievable (upper limit) quality of assemblies. Genomes of eukaryotes have higher degrees of complexity due to their size, repeats and presence of gene families that have identical stretches of sequences (Pop, 2009; Pop and Salzberg, 2008).

In order to overcome repetitive regions, assemblers need to make use of paired-end read information where a sequencing template is sequenced from both ends resulting in two reads in opposite orientation and separated by a known fragment size. In addition to the increase in sequence data, paired-end (PE) sequencing provides a way of jumping over difficult regions. Such evi-

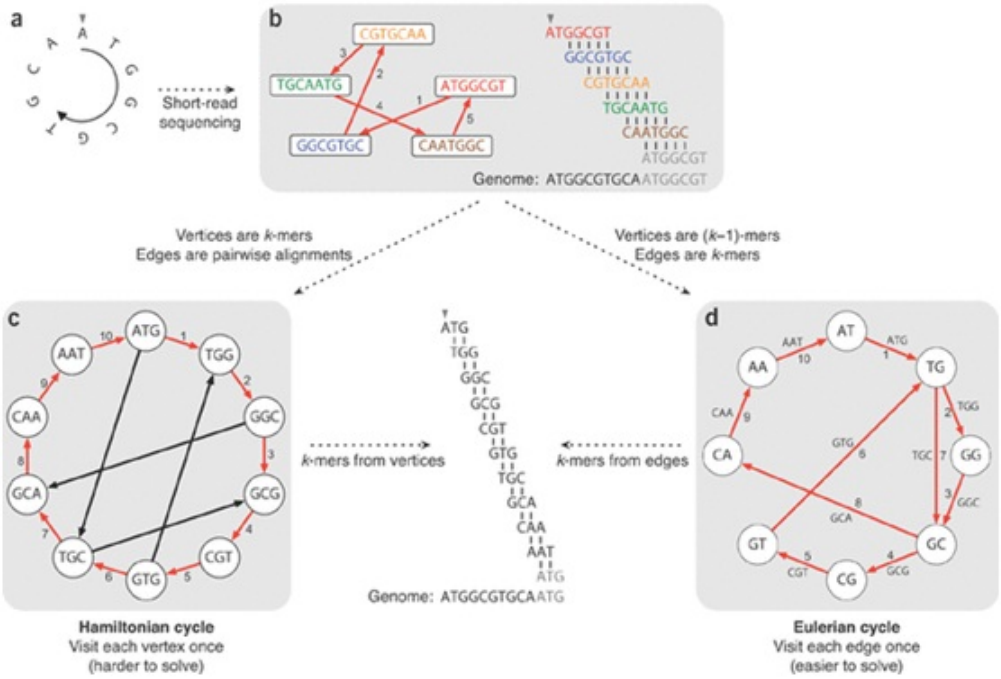


Figure 1.6: Example of graph-based assembly approaches (a) An example of a small circular genome. (b) Most assemblers developed to assemble reads from traditional Sanger sequencing platforms represent reads as nodes in a graph, and edges represented alignments between reads. Walking along a Hamiltonian cycle by following the edges in numerical order allows for a reconstruction of the circular genome by combining alignments between successive reads. At the end of the cycle, the sequence wraps around to the start of the genome. The repeated part of the sequence is grayed out in the alignment diagram. (c) An alternative assembly technique first splits reads into all possible *k-mer*s: with *k* = 3, ATGGCGT comprises ATG, TGG, GGC, GCG and CGT. Following a Hamiltonian cycle (indicated by red edges) allows one to reconstruct the genome by forming an alignment in which each successive *k-mer* (from successive nodes) is shifted by one position. This procedure recovers the genome but does not scale well to large graphs. (d) Modern short-read assembly algorithms construct a de Bruijn graph by representing all *k-mer* prefixes and suffixes as nodes and then drawing edges that represent *k-mer*s having a particular prefix and suffix. For example, the *k-mer* edge ATG has prefix AT and suffix TG. Finding an Eulerian cycle allows one to reconstruct the genome by forming an alignment in which each successive *k-mer* (from successive edges) is shifted by one position. This generates the same cyclic genome sequence without performing the computationally expensive task of finding a Hamiltonian cycle. (Taken from Compeau et al. (2011))

dence is valuable in resolving ambiguous overlaps due to repetitive regions in subtelomeres and multi-gene families such as *var* genes. Short read assemblers that did not support paired reads such as SSAKE (Warren et al., 2006), VCAKE (Jeck, 2007), SHARCGS (Dohm et al., 2007) and Edena (Hernandez, 2008) had limited applications in assembling *var* genes. On the other hand, assemblers including Velvet (Zerbino and Birney, 2008), EulerSR (Chaisson MJ, 2008), AllPaths (Butler, 2008), ABySS (Simpson, 2009) and SOAPdenovo (Li et al., 2009b; Ruiqiang Li, 2010) supported paired reads and hence were better equipped to deal with multi-gene families (Imelfort and Edwards, 2009). It is however important to mention that none of the current assembly tools use PE reads from the initial stages of constructing the overlap graph. The first tool to take advantage of read pairs from the beginning was published nearly four years after the introduction of PE sequencing protocols (Pop, 2009). Despite the importance of such approaches, one reason for the slow pace of development is the difficulty of defining a pair of *k-mer*s (from the forward and reverse reads) when the exact distance between them is unknown.

In addition to inherent features of genomes such as G+C content and repeats, random and systematic sequencing errors contribute to challenges in assembly of short reads. Second generation sequencing technologies are prone to different types of systematic errors. For instance, the 454 platform has issues with homopolymer tracks (Prabakaran et al., 2011) while Illumina's sequencing platforms are reported to have mismatch errors (Abnizova et al., 2012) and errors found downstream of homopolymeric tracts (Otto et al., 2010a). Identifying sequencing errors and accounting for their bias is an active area of research in such topics as indexing, aligning of short reads (Frith et al., 2010; Giladi et al., 2010), and *de novo* assembly (Zhao et al., 2010).

Sequencing errors lead to changes in graph topology causing erroneous structures such as tips, bubbles and chimeric connections (Zerbino and Birney, 2008). Errors at read-ends result in a chain of nodes where one of the ends has no connections leading to the formation of tips. Bubbles on the other hand, may result from overlaps between neighbouring tips or due to errors in the middle of reads. In the velvet assembler, tips and bubbles are first detected and removed in the assembly process. Chimeric connections are more complicated

as they are not easy to detect from the topology of the graph and may be caused by genuine biological events such as polymorphism.

The problem of scaffolding

The assembly process rarely produces a single complete sequence (for each chromosome) from short read fragments. Instead, the target sequence is represented in a set of contiguous fragments also known as contigs. During or after the assembly process, additional information from read-pairs (usually refers to standard fragment libraries) and mate-pairs (some providers use this term referring to long insert libraries) is used to assign contigs to scaffolds. Scaffolding involves determining the relative order of contigs and estimating gap sizes between them based on availability of mate-pairs that connect separate contigs. As in the case of assembly, the complexity of scaffolding is increased due to contig-ends that could be joined with multiple other contigs resulting in an NP-hard problem (Huson et al., 2002). A simplified approach is often used which involved the implementation of greedy heuristic algorithms to resolve ambiguities (Huson et al., 2002; Kim et al., 2008; Pop, 2003). However, such simplifications lead to mis-scaffolding in gene families due to the presence of highly identical segments. Although most short read assemblers include built-in scaffolding options, there is a growing number of stand alone scaffolding tools such as Bambus (Pop, 2003) and SSPACE (Boetzer et al., 2011) that focus on post assembly genome improvements steps. These tools may have a better performance than built-in scaffolders of short read assemblers due to the availability of options that could be tuned by the user.

1.5 Overview of thesis

The Illumina platform is being used to sequence thousands of parasite genomes at the Sanger Institute. However, the high A+T content and the frequent presence of multiple genotypes within an infection make it extremely challenging to reliably map or *de novo* assemble short reads in subtelomeric regions where most of the *var* genes are located. This thesis is therefore focused on the development

of algorithms to assemble *var* genes from second generation sequencing reads of clinical samples.

Chapter 2 describes tests performed to evaluate how well existing assembly tools reconstruct *var* genes from short reads of the Illumina platform. A new iterative assembly approach developed to assemble *var* genes is presented in Chapter 3. Chapter 4 describes application of short read sequencing to understand mechanisms used by parasites to generate new *var* genes. Finally, Chapter 5 presents applications of the new assembly approach (developed in Chapter 3) to a global collection of clinical samples. Concluding remarks and future directions are detailed in Chapter 6.

Chapter 2

Evaluating existing short read assembly methods

2.1 Introduction

Advances in next generation sequencing technologies have significantly improved read length, yield and quality of whole genome shotgun sequencing data. There has also been a considerable development of algorithms and software to deal with the problem of piecing together sequence fragments into a longer contiguous sequence. However, assembly attempts of whole genome *P. falciparum* sequence data are faced with unique challenges due to the high A+T content (∼80% in coding and ∼90% in non-coding regions) (Gardner et al., 2002). *De novo* assembly of *P. falciparum* genomes in general and subtelomeric gene families in particular is thus extremely challenging due to the base composition bias and presence of repeat sequences. *Var* genes are highly polymorphic and composed of mosaic blocks that have regions of high similarity. The quality of the raw sequence data is often affected by systematic errors from the sequencing instruments. Systematic errors and inherent sequence features are thus expected to have a significant impact on the assembly of *var* genes.

The aim of this part of the thesis (year 1) was to see if it was possible to assemble *var* genes using existing short read assembly tools. In this chapter, I will evaluate the feasibility of existing approaches to assemble *var* genes

followed by investigation of potential reasons for poor quality assembly in *var* genes.

2.2 Methods

2.2.1 Library preparation and sequencing

Library preparation and sequencing of samples used in this thesis were done by the Research and Development and Sequencing Production teams at the Sanger Institute. The protocols used for library preparation and sequencing are briefly summarised below.

Library preparation
Standard genomic DNA library preparation
DNA samples were initially quantified on the Invitrogen Qubit and then fragmented using the Covaris Adaptive Focused Acoustics technology (fragment sizes of 200-300 bp and 300-400 bp). End-repairing of fragments and creation of blunt-ends were done using T4 and Klenow DNA polymerases, and T4 polynucleotide kinase respectively. This was followed by A-tailing, addition of a single 3′ A nucleotide to the repaired ends using Klenow exo- and dATP. Standard adapters were then ligated according to the manufacturers guidelines. Size selection of ligated fragments was done using Agencourt AMPure XP beads. The libraries were then enriched by 8 cycles of PCR and quantified using Agilent Bioanalyser chip and Kapa Illumina SYBR Fast qPCR kit.

PCR-free Genomic DNA library (NoPCR) preparation
The PCR-free library preparation protocol (Kozarewa et al., 2009) was developed to minimize the effect of amplification artifacts especially in genomes with G+C bias such as *P. falciparum*. However, it also requires more input DNA. This method uses similar protocols of DNA qualification, shearing, end-repairing and A-tailing. However, instead of the standard adapters, NoPCR adapters were ligated (containing primer sites for sequencing and flowcell surface annealing) according to the Amplification-free Illumina sequencing

protocol (Kozarewa et al., 2009). Subsequent steps of size selection and DNA extraction are similar to the standard library preparation protocol above. Both standard and PCR-free libraries were used in this chapter (Table 2.1). Clinical samples used in Chapters 3 and 5 were all prepared using the PCR-free protocol.

Cluster generation and Sequencing

Initially, in order to allow hybridization of template strands to adaptors that are attached on the flowcell, libraries were denatured with sodium hydroxide and diluted in a hybridisation buffer. Cluster amplification was performed on the Illumina cluster station (changed to Illumina cBOT after April 2010) using the V4 cluster generation kit following the manufacturer's protocol. Cluster density was measured using SYBRGreen to determine whether a flowcell had enough DNA for sequencing. This was followed by consecutive linearization, blocking and hybridisation of R1 and R2 sequencing primers for (i.e. for the forward and reverse reads). Sequencing-by-synthesis was then performed for 75 to 150 cycles depending on the sequencing instrument. These steps were performed using proprietary reagents according to manufacturer's recommended protocol (`https://icom.illumina.com/`)

Samples used in this chapter

Laboratory adapted and cultured clinical isolates of *P. falciparum* were used with the aim of evaluating and optimising existing short read assembly methods in polymorphic gene families, specifically the *var* gene family. Due to the availability of a complete and nearly base-perfect reference genome, re-sequencing of the 3D7 isolate provides an ideal benchmarking standard for new sequencing technologies and protocols. DNA for the 3D7 isolate was obtained from Prof. Chris Newbold's lab in Oxford.

2.2.2 Choice of short read assemblers

Velvet (Zerbino and Birney, 2008) was one of the first short read assemblers to popularise de Bruijn graphs. It is easy to install and run with a growing community of users. The major limitation is the large amount of Random Access

	Libraries	Insert size(bp)	Library type	Read length(bp)	Machine	Cov. (X)
3D7	S_i200	200	Standard	76	GAII	58
	NP_i200.1	200	NoPCR	76	GAII	163
	NP_i200.2	200	NoPCR	76	GAII	184
	HS_i500.2	500	NoPCR	75	HiSeq2000	607
	NP_i500	500	NoPCR	76	GAII	338
	HS_i500.1	500	NoPCR	75	HiSeq2000	724
	MS_i3k.1	3000	Large insert	75	MiSeq	34
	MS_i3k.2	3000	Large insert	150	MiSeq	20
	MS_i3k.3	3000	Large insert	150	MiSeq	70
	MS_i3k.4	3000	Large insert	150	MiSeq	64
Field samples	F1	200	Standard	76	GAII	
	F2	200	Standard	76	GAII	
	F3	200	Standard	76	GAII	
	F4	200	Standard	76	GAII	
	F5	200	Standard	76	GAII	
P. reichenowi	*Pr*	200	Standard	76	GAII	

Table 2.1: Samples used to evaluate existing assembly approaches and determine error rates of the Illumina sequencing technology in *P. falciparum*. Libraries were named according to the library preparation protocol (S for standard, NP for NoPCR/PCR-free; HS for HiSeq and MS for MiSeq) and the insert sizes of the libraries (eg. i200 for 200 bp library)

Memory (RAM) required for large genome assemblies. However, assembly of *var* genes and the *P. falciparum* genome are within the limits of Velvet's memory requirement. Velvet was therefore a good starting point for the purpose of comparisons. Abyss and SOAPdenovo were especially designed to address Velvet's limitation with large genomes. Li and colleagues (Li et al., 2010; Ruiqiang Li, 2010) reported better assembly results for SOAPdenovo compared to Abyss using both small (*Ecoli*) and large (African human) genomes. Velvet (version 1.1.04) and SOAPdenovo (version 1.05) were therefore chosen to represent existing short read assemblers with the intention of identifying a program that could generate high quality assembly for polymorphic gene families in *P. falciparum*.

2.2.3 Assembly benchmarking using error free in silico reads

Error free reads were generated using a Python script written by Martin Hunt in our laboratory (*simulate_pe_reads.py*). Reads were simulated assuming a Gaussian distribution of fragment sizes given a mean fragment size, standard deviation of the distribution and the average read coverage. This script is used to generate in silico reads in this thesis unless stated otherwise.

To compare the performance of Velvet and SOAPdenovo, paired reads of length 76 bp (coverage=80x; mean fragment size=200 bp, standard deviation=30) were simulated from a total of 93 *var* genes of the 3D7 genome (including pseudogenes and truncated exon 2 sequences). The two assemblers were compared on the number and quality of contigs they produced. To obtain the best result for each assembler prior to comparison, the two assemblers were first optimised by running each tool at *k-mer* sizes of 51, 55, 61, 65 and 71. Ideally, the best assembler will generate the least number of contigs, with high N50 values, a total number of bases as close to the expected length of the target sequence (i.e. $\sim$450 kb for *var* genes in the 3D7 genome), and the fewest errors. Assembly quality was assessed based on the coverage of contigs and the repertoire completeness of *var* genes as described below.

2.2.3.1 Accuracy and coverage of contigs

ABACAS (Algorithm Based Automatic Contiguation of Assembled Contigs) was used to assess the quality of assembled contigs (Assefa et al., 2009). If a good quality reference sequence is available, comparing assembled contigs to the reference provides the best measure of completeness and accuracy. A similar approach is used during the process of finishing genomes whereby contigs are aligned to a reference sequence of a close or distant relative with the aim of establishing a relative order of contigs using regions of conserved synteny. However, there was no readily available software to automate the process. I originally developed ABACAS to address this issue by rapidly aligning contigs to a reference genome in order to determine the order and orientation of contigs relative to a reference genome. A multi-FASTA file of contigs was first aligned as nucleotide (*option -p nucmer*) or six-frame translated amino acid sequence

(*option -p promer*). Based on the alignment output, contigs are then ordered and orientated to generate the new reference, termed the pseudo-molecule. Output files include an ordered FASTA file, a feature file and comparison files for visualisation in the Artemis comparison tool (ACT) (Carver et al., 2005). The comparison file displays an ordered list of contigs that were aligned to the reference genome. The proportion of each contig aligned to the reference (i.e. percent coverage) is reported alongside the percent identity of the match.

Additional use-cases of ABACAS such as checking the quality of assembled contigs became apparent during the course of this project. Misassemblies were therefore identified by looking for unordered contigs of length above 1 kb and ordered contigs with a coverage and percent identity values lower than 95%. Despite the reliability of this approach, its application in the absence of a reference sequence is very limited. It was thus necessary to develop an alternative method to determine the completeness of *var* gene repertoires in a wide range of assembly projects including *de novo* assembly of clinical samples.

2.2.3.2 Estimating *var* gene repertoire completeness

As described in Chapter 1, the DBLα domain is present in nearly all *var* genes of the 3D7 genome and other sequenced isolates (Kraemer et al., 2007; Rask et al., 2010). The number of contigs that have a complete DBLα domain were counted using a customised Perl-script (*getDBL.pl*). The script was initially written for DBLα sequence tags by Dr. Pete Bull of Kemri-Wellcome Research Unit, Kilifi Kenya and improved to look for DBLα start and end motifs in all the six frames instead of one frame each from the forward and reverse strands. Outputs of the script include amino acid sequence file of DBLα tags and a FASTA-format file of contigs that contain DBLα. Counting such contigs will therefore rsult in the closest approximation of repertoire completeness that is robust in clinical samples. Although the var2CSA genes (gene that encodes a *Pf*EMP1 variant responsible for pregnancy associated malaria) do not contain the DBLα domain, they are highly conserved and could easily be identified using their 3D7 homologues.

2.2.4 Evaluating the velvet assembler on real and simulated reads

In order to evaluate Velvet on real data, a PCR-free library of 3D7 (Table 2.1, libraries NP_i200.1 and NP_i200.2) was sequenced on Illimina's GAII platform. One aim of this analysis was to investigate the quality of assembled contigs while decreasing the dataset from the whole-genome to a chromosome and finally to reads that belong to *var* genes.

2.2.4.1 Whole genome assembly

First, to find optimal results, a whole genome *de novo* assembly was done by varying the *k-mer* size and coverage cutoff values. A *k-mer* of 61 and coverage cutoff dynamically determined by Velvet (*i.e. coverage_cutoff=auto*) resulted in the best assembly results as determined by the highest N50 and least number of contigs. Other parameters were kept to the default settings. In order to investigate the effect of poor quality reads on assembly, a filtered set of reads was generated by aligning raw reads to the 3D7 reference genome (version 2.1.4) using Bowtie (Langmead et al., 2009) (*using the -v alignment mood, -v 2*) allowing a maximum of two mismatches. Bowtie is a memory efficient and fast short read alignment tool that uses the Burrows-Wheeler Transformation (BWT) to index the reference genome. The version of Bowtie used in this thesis (version 1) did not support gapped alignment, which contributed to its increased speed compared with other BWT based aligners such as BWA (Li and Durbin, 2009).

Filtered reads were then assembled using a similar set of parameters as the raw reads (*velveth -k=61; velvetg -cov_cutoff auto*).

2.2.4.2 Chromosome 1 assembly

To further investigate the effect of errors and reducing the dataset to a single chromosome, sequences that aligned to chromosome 1 were obtained from the filtered set of reads generated in the previous section (library NP_i200.1). These reads were assembled using velvet (*velveth -k 61; velvetg -cov_cutoff auto*). In addition, error free synthetic reads were evenly generated with a similar number

of reads and mean fragment size as the real data (mean=166 bp, standard deviation=35). A second set of synthetic reads was generated by mimicking the coverage of real data over chromosome 1 followed by a random introduction of mismatch errors. Both sets of simulated reads were assembled using similar parameters as the real data (*velveth -k=61; velvetg -cov_cutoff auto*).

2.2.4.3 Assembly of *var* genes

In order to evaluate the performance of Velvet on *var* genes from the real data, reads that aligned to *var* genes of the 3D7 genome were obtained from a PCR-free library (Table 2.1, Library NP_i200.1) and assembled (*Velvet, k-mers 25 - 65, cov-cutoff=auto*). Potential reasons for a poor quality assembly of *var* genes were investigated in two steps.

First, raw reads were aligned to a concatemer of 3D7 *var* genes with the aim of assessing the effect of sequencing errors and uneven coverage. The alignment output was stored in the Binary Alignment/Map (BAM) (Li et al., 2009a) format and visualised in ACT using the BamView utility (Carver et al., 2010). A graphical representation of read coverage and SNPs (errors) was used to look at their correlations with assembly quality.

Second, to look at the effect of repeats in assembly, shared sequences within *var* genes of the 3D7 genome were identified using a pairwise blast search (*all-against-all blast; blastn, -F F, -e=1x10*$^{-3}$). Regions of genes that have a perfect match with a length above the fragment size of the library (200 bp) were identified as repeat regions. Reads that aligned to such regions were excluded to generate a second set of *var* reads. Assembly of the two sets were compared by looking at the underlying de Bruijn graphs which were visualised using a python script contributed to the Velvet package by Paul Harrison (*graph2.py*) .

2.2.5 Evaluating mapping based assembly approaches

To assess the feasibility of a reference guided assembly approach, short reads from the 3D7 clone, three field samples and *Plasmodium reichenowi (Pr)* (Table 2.1) were aligned to the 3D7 reference genome (BWA; *version 0.5.5, default parameters*). Reads that aligned to *var* genes in proper pairs (i.e. read pairs aligned in the

correct orientation and within the expected insert size) were counted for each gene. The number of mapped reads per thousand bases (kb) was used as a measure of mappability over *var* genes and computed by normalizing the number of properly aligned paired reads over a gene by the length of the gene.

2.2.6 Sequencing errors

A total of 10 genomic DNA libraries were used for error profiling of Illumina's GA2, HiSeq and MiSeq instruments (Table 2.1). In order to assess the improvements and changes in error rates, libraries from early Illumina (GAII) runs and the latest MiSeq runs were used. Raw reads from the reference genome were aligned to the most recent version of the genome (version 3) using Bowtie1 allowing a maximum of three mismatches. The output file was processed using a purposely written Perl-script to identify mismatch/error positions and find error rates at low ($Q5$), medium ($Q15$) and high ($Q25$) quality thresholds. Quality values represent a confidence score assigned by Illumina's base calling algorithm (Bustard) which assigns quality scores (Q) based on an expected error probability P such that :

$$Q_{\text{Solexa prior to v.1.3}} = -10log_{10}(P/(1-P)) \tag{2.1}$$

or

$$Q_{\text{Illumina v.1.3+}} = -10log_{10}(P). \tag{2.2}$$

It is not unusual to observe incorrect base calls or unknown bases (Ns) with high quality score. It was therefore important to look at sequencing errors at low and high quality cutoff values. First, the overall error rate (E) for the forward and reverse reads were computed over the full length of each read R at a quality cutoff Q as follows:

$$ER = (\textit{mismatched bases above } Q\ )/(\textit{mapped bases above } Q) \tag{2.3}$$

Similarly, error rates were computed for each position P on the population of forward and reverse reads at a quality cutoff Q as follows:

$$ERP = (\textit{mismatched bases above } Q \textit{ at } P) / (\textit{mapped bases above } Q \textit{ at } P) \quad (2.4)$$

Finally, substitution profiles were obtained for the 12 mismatch types (A-C, A-G, A-T, C-A, C-G, C-T, G-A, G-C, G-T, T-A, T-C and T-G) by computing the average number of mismatches that exhibited a given patterns across all cycles.

2.3 Results

The standard approach to any assembly problem would be to use *de novo* or reference guided assembly on all fragments (reads). This chapter presents tests performed to evaluate the feasibility of reconstructing *var* genes from short reads using existing assembly tools. A comparison of two representative assemblers, Velvet and SOAPdenovo, is presented using synthetic reads simulated from *var* genes followed by further evaluations on real and simulated data. An investigation into the potential reasons of low quality assembly is also presented with a focus on errors specific to *P. falciparum* sequences.

2.3.1 Comparing *de novo* assembly tools

A total of 507,812 read-pairs were simulated from *var* genes of the 3D7 reference isolate to compare SOAPdenovo and Velvet. To account for the variability in assembly quality with changes in *k-mer* size, both assemblers were first run with *k-mer* sizes of 51, 61, 65 and 71. Optimal assembly values were obtained at a *k-mer* size of 65 for both assemblers (Figure 2.1). Assembly results from *k-mer* sizes of smaller and larger than 65 were highly fragmented. In addition to generating the highest number of contigs, a k-mer of 71 resulted in the shortest size for the Largest-contig (Table 2.2). The highest contig N50 size (5,713 bp) was obtained from the Velvet assembly with 252 contigs compared to SOAPdenovo's N50 of 2,346 bp and 1,398 contigs.

The number of contigs aligned to the reference set of 93 *var* genes (by ABACAS) was higher in Velvet (72% vs SOAPdenovo's 45%) (Table 2.3). No

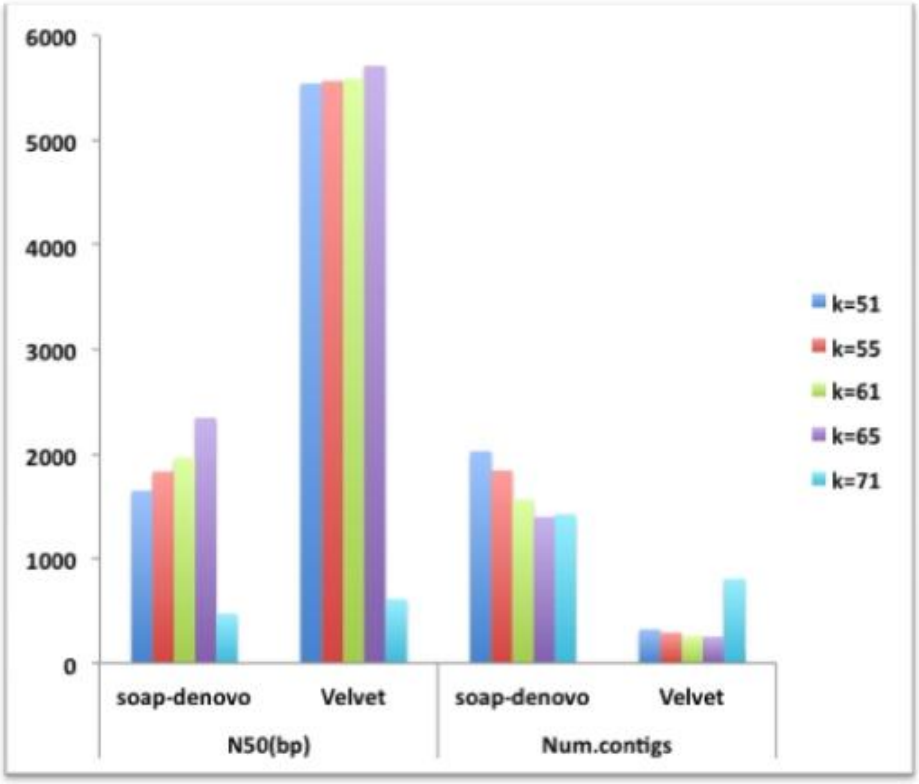


Figure 2.1: Comparing SOAPdenovo and Velvet on synthetic reads simulated from *var* genes of the 3D7 genome. Optimal assembly results were obtained at a *k-mer* size of 65 with Velvet generating a better assembly with the highest N50 and least number of contigs.

k	SOAPdenovo				Velvet			
	Sum (kb)	N50 (bp)	Num.contigs	Largest (bp)	Sum (kb)	N50 (bp)	Num.contigs	Largest (bp)
51	514	1648	2027	9517	21	5543	324	10418
55	514	1831	1841	9525	425	5567	291	10422
61	512	1963	1560	10489	430	5586	258	10489
65	**509**	**2346**	**1398**	**10493**	**433**	**5713**	**252**	**10493**
71	464	471	1422	3842	420	612	804	4521

Table 2.2: Assembly statistics of Velvet and SOAPdenovo at different *k-mer* sizes. Optimal assembly results were found at *k-mer* size of 65.

	Sum (kb)	N50 (bp)	#contigs with DBLα
SOAPdenovo	192	5277	47
Velvet	293	6738	50

Table 2.3: Assembly statistics of contigs with the DBLα domain. Velvet generated the highest number of contigs with the DBLα domain suggesting a better assembly with better representation of the *var* repertoire.

misassembled contigs were detected in both assemblies. A total of 50 contigs contained a complete DBLα domain in the Velvet assembly. The sum of these contigs accounted for $\sim$68% of the expected sum of *var* sequences with DBLα in the 3D7 genome as determined by our method ($\sim$428 kb, 54 genes). Conversely, SOAPdenovo assembled 47 contigs with the DBLα domain that only contained 45% of the expected total sequence (192 kb compared to Velvet's 293 kb).

Velvet was therefore chosen as a better tool to further optimise assembly of *var* genes in *P. falciparum*. The main objective of these experiments was to establish a robust method that could be used in the context of clinical samples. However, the initial comparisons were done using simulated data and real reads from the reference genome as separating out the complete *var*-specific reads would not be possible for a real clinical sample due to high polymorphism. Having established Velvet as a better choice, the following sections present further tests on real and simulated data from the 3D7 genome. Assembly results were examined in a decreasing order of complexity from the whole genome to a single chromosome and finally the *var* gene family.

2.3.2 Velvet assembly of whole genome data

A whole genome *de novo* assembly of 26.6 million reads (paired 76 bp, two lanes) from a PCR-free library (Table 2.1, NP_i200.1 and NP_i200.2) of the reference clone 3D7 was extremely fragmented (Table 2.4). Assembly results from the two lanes were comparable in the number and size distribution of the contigs generated. The number of contigs was ∼20,000 with an average N50 size of 1,370 bp. The total number of assembled bases (i.e. sum of contigs) was however closer to the expected genome size of ∼23 Mb (∼18 Mb to 19 Mb).

A filtered set of reads was obtained by excluding reads that aligned to the reference genome with more than two mismatches. Reads that passed the mapping-based filtering accounted for 73% of the total and covered 97% of the genome with at least one read (94% covered with at least 5 reads). The assembly of filtered reads was slightly better after removing 27% of the reads that were either contaminants or had a lower quality (Table 2.4). However, the results show that whole genome *de novo* assembly of short reads is still impractical for *P. falciparum*.

	NP_i200.1		**NP_i200.2**	
	All reads	**Filtered reads**	**All reads**	**Filtered reads**
Total bases (Mb)	18.87	18.25	19.59	19.2
Num. contigs	19676	18276	19899	18508
N50 (bp)	1361	1442	1387	1491
Average (bp)	958.9	998.8	984.7	1037.6
Largest (bp)	20771	16989	19471	23005
Unused reads (%)	31.22	18.1	41.3	16.87

Table 2.4: Whole genome Velvet assembly of real data from two PCR-free libraries of 3D7. The second data set of 'filtered reads' was obtained by removing reads that aligned to the genome with more than two mismatches. Assembly results were highly fragmented for both libraries.

2.3.3 Velvet assembly on Chromosome 1 of 3D7

Reads that aligned to chromosome 1 of the reference genome were obtained from the 'filtered set' described in the previous section. A mapping coverage

of 94% was observed on comosome 1. The average fragment size (166 bp) was shorter than the expected standard library fragment size of 200 bp. Assembly of these reads was highly fragmented (N50=1 kb, sum of contigs=402 kb) compared to assembly of error free simulated data with a similar number of reads and insert size distribution (N50=15 kb, sum of contigs=614 kb). However, introducing errors and uneven coverage to simulated reads had a significant effect on assembly quality dropping the N50 contig size to 1.3 kb (Table 2.5). These results suggest that sequencing errors and uneven coverage may explain the low quality assemblies.

	Real reads	**Simulated with error**	**Error free simulation**
N50	1076	1255	15278
Average	849	1019	3593
Larges	4979	5457	40793
Total bases	401616	637948	614369

Table 2.5: Assembly results of simulated and real reads on chromosome 1 of 3D7. Real reads and uneven-simulated reads with errors (column 3) had comparable results. Reads simulated without errors and with even-coverage (column 4) resulted in better assembly.

2.3.4 Velvet assembly on *var* genes

Initially, 1.9 million read-pairs where one or both of the reads aligned to *var* genes were assembled generating the best assembly at a k-mer size of 65 with N50 contig size of ∼1.7 kb (sum of contigs=457 kb, Number of contigs=603, Largest contig=9.85 kb). The number of contigs that contained the DBLα domain was 40 (∼74% of the expected) generating a total of 165 kb bases (∼37% of the expected ∼450 kb) lower than reported for simulated reads (Num. contigs with DBLα=50, sum of contigs=293 kb). A closer look at errors and read coverage shows that regions with highest errors correspond with contig breakpoints (Figure 2.2). In addition, low and uneven coverage also affected assembly quality. Next, the effect of repeated sequences on *var* assembly was investigated. Although the most commonly shared sequences were smaller than 100 bp, stretches of above 1 kb were also identified from the pairwise blast search of 3D7 *var* genes (Figure 2.3). A visual inspection of the underlying de Bruijn graph

revealed extremely dense nodes which represented highly similar sequences within members of the family (Figure 2.4A). The effect of repetitive sequences on the *var* assembly graph was examined by removing reads that aligned to genes that contained shared sequences above the fragment size of 200 bp. The graph was significantly simplified although still far from ideal (Figure 2.4B). However, such simplifications are likely to affect quality by generating gaps in the final assembly.

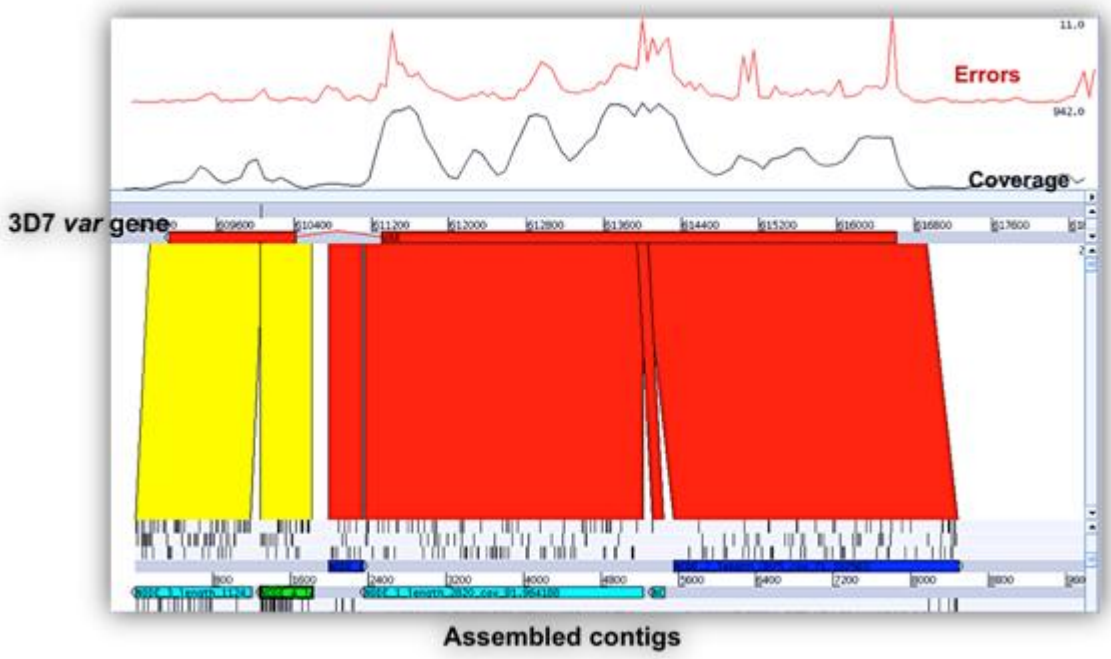


Figure 2.2: The effect of sequencing errors and uneven coverage on assembly of *var* genes. A comparative view of assembled contigs with *var* genes of the reference genome is shown. The top panel shows mismatch (error) count plot (red) and read mapping coverage plots (black). The bottom panel shows assembled contigs (bottom) that were ordered and orientated against a 3D7 *var* gene (top). The red and yellow blocks represent synteny matches. The color of contigs indicates whether they align in the forward (green) or reverse (blue) strands and if there is an overlap between neighboring contigs (cyan). Black bars represent stop codons.

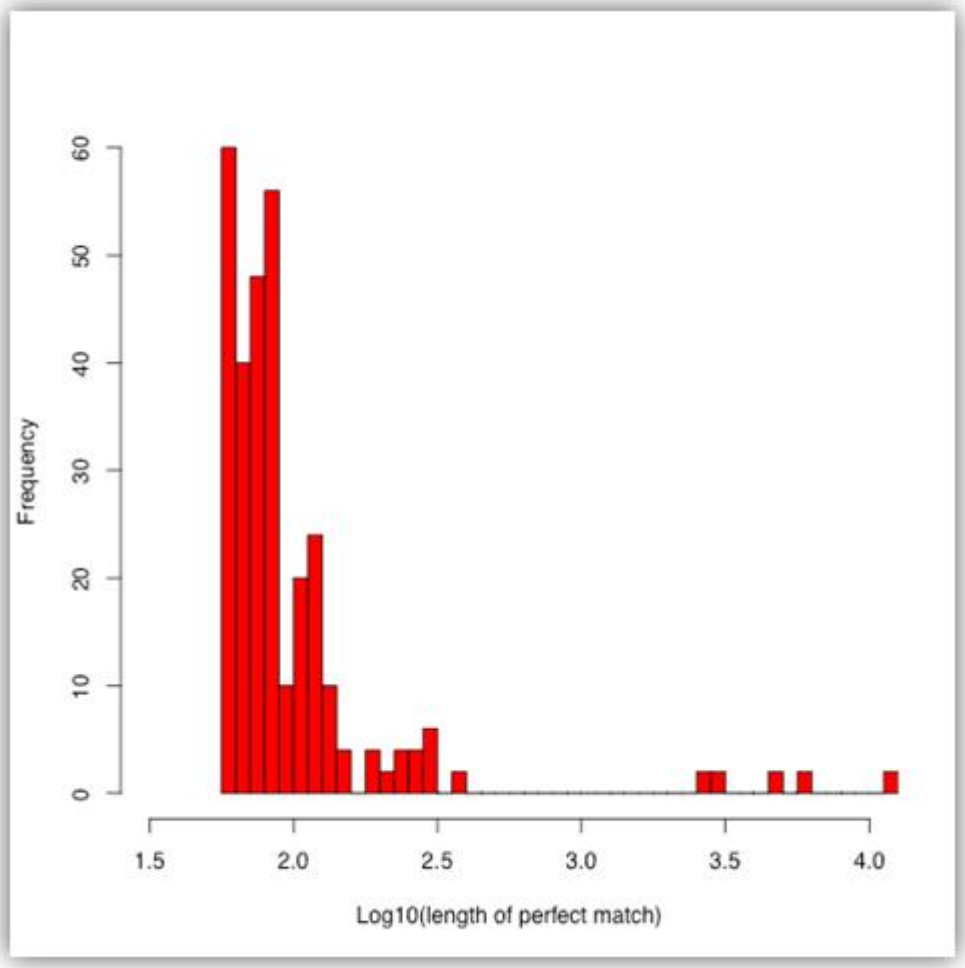


Figure 2.3: A histogram of shared sequences in 3D7 *var* genes identified by a pairwise blast alignment. Perfectly matching sequence blocks of length up to 4 kb were found between *var* genes of the 3D7 genome.

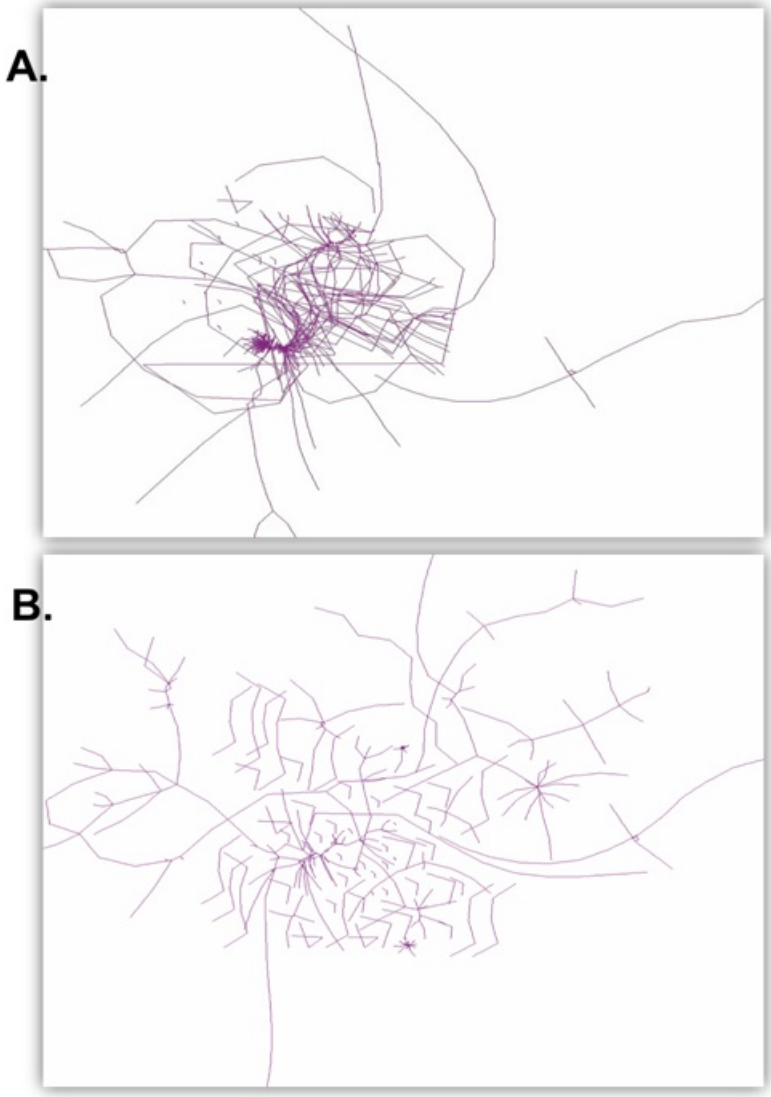


Figure 2.4: Visualizing the assembly graph of all *var* genes from a velvet assembly (real reads, k=61). In these plots, node sequences were represented as lines and curves that are joined with other nodes at their tips. The dense regions (nodes) on the graph represent repetitive sequences that have multiple connections with other nodes A) The assembly graph of real reads that align to *var* genes of the the 3D7 genome. B) A simplified assembly graph after removing reads that align to shared sequences between *var* genes with a match length of above 200 bp (as shown in Figure 2.3).

2.3.5 Reference guided assembly

In addition to *de novo* assembly, a possibility of using mapping based assembly approaches was investigated. The challenge to genome assembly posed by the inherent features of the genome (particularly the high A+T bias) is illustrated using a *k-mer*-based uniqueness plot computed by counting the frequency of sequences of length 30 bp. The uniqueness plot indicates how well short reads could be uniquely mapped to genomic regions. The correlation between sequence complexity, read coverage and G+C content is shown for chromosome 1 (Figure 2.5) and the left subtelomeric region of chromosome 8 (Figure 2.6). The increase in G+C indicates higher information content and therefore mappabiliy. Although both alignment and assembly benefit from paired-end information, current assemblers construct graphs using *k-mer*s from all reads independently. Read pair information is used at later stages of the assembly to simplify the graph and resolve repeats. A *k-mer*-based uniqueness plot therefore shows regions of the genome where the assembly would terminate contigs due to ambiguity. The subtelomeres of *P. falciparum* contain repeat blocks which have a lower uniqueness compared to core regions of the genome. The overall spikiness of uniqueness and G+C content affects mapping coverage across the genome more specifically in subtelomeric regions (Figures 2.5 and 2.6).

Despite the problem of uniqueness in the sub-telomeric regions, at first sight it appears that the relatively high G+C content and uniqueness of the *var* genes should aid their assembly by mapping. However the extreme polymorphism of these genes presents a much greater additional problem when sequence reads from different genotypes are used. Figure 2.7 demonstrates this point by comparing the homologous mapping coverage of the reference genotype 3D7 to the coverage of the reference from three field isolates of *P. falciparum* and to its closest known relative, the chimpanzee parasite *P. reichenowi*.

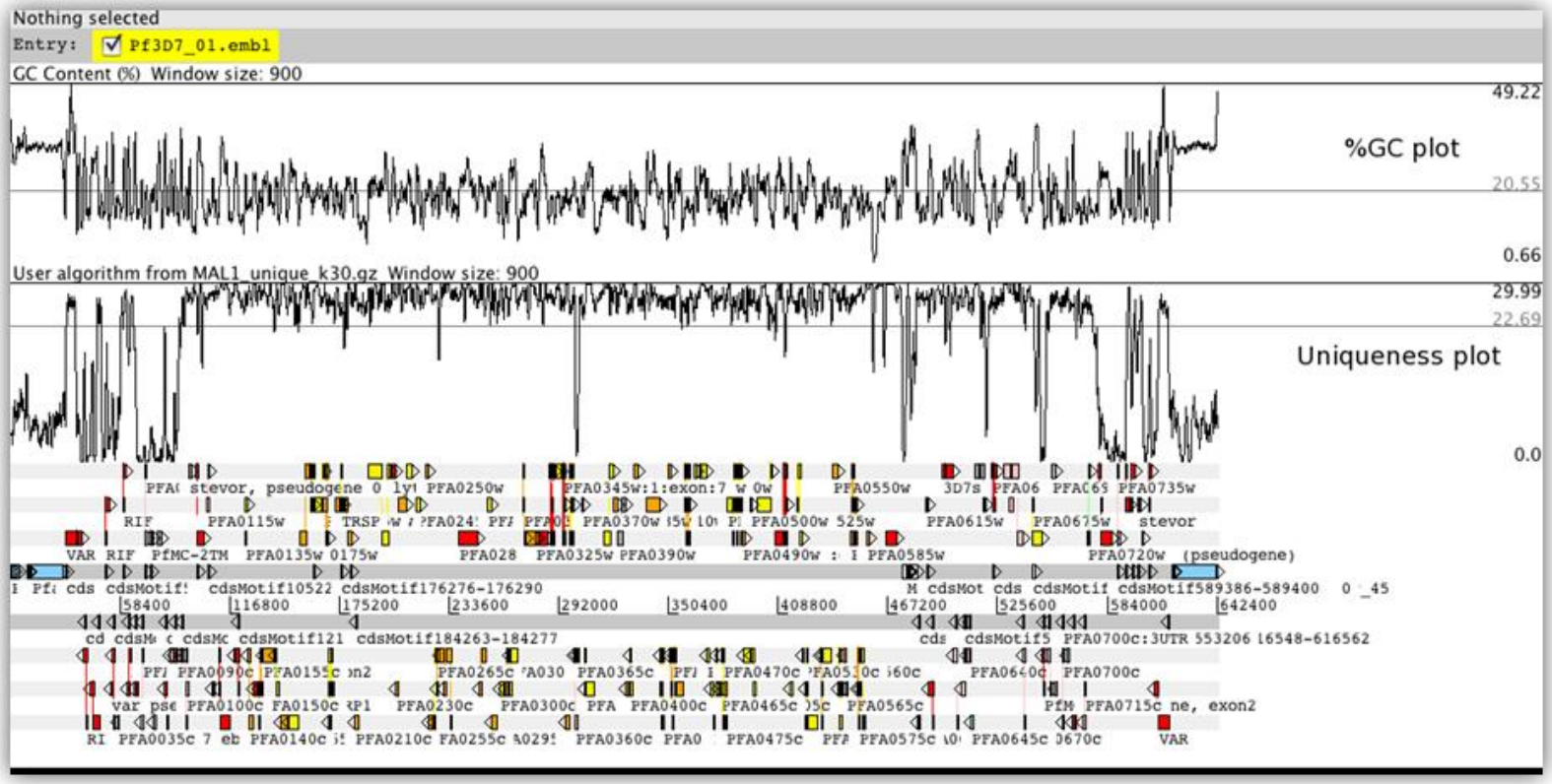


Figure 2.5: A plot of G+C content and uniqueness (based on a *k-mer* size of 30 bp) on Chromosome 1 of the 3D7 genome. An Artemis view of G+C content (top panel) and a uniqueness plot (middle panel). The bottom panel shows annotation information with the different blocks representing protein coding genes, pseudogenes and repeats in all the six reading frames.

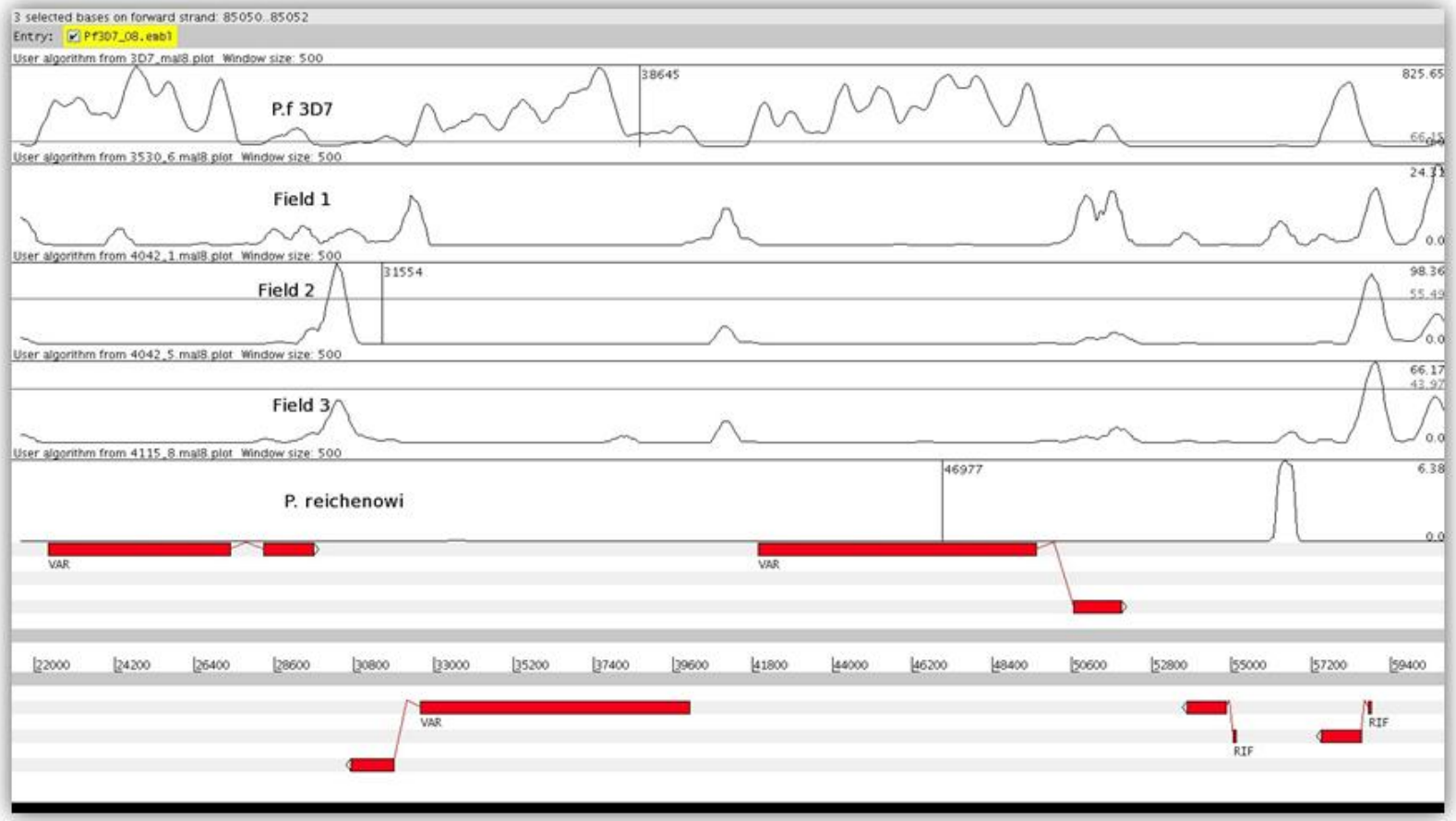


Figure 2.6: Read mapping coverage plots over *var* genes on the left subtelomere of Chromosome 8. Illumina reads from 3D7, three field samples and *P. reichenowi* were aligned to the 3D7 genome.This figure shows the difficulty of reliably aligning reads obtained from clinical samples to *var* genes (shown in red over the forward and reverse strands) of the reference genome 3D7.

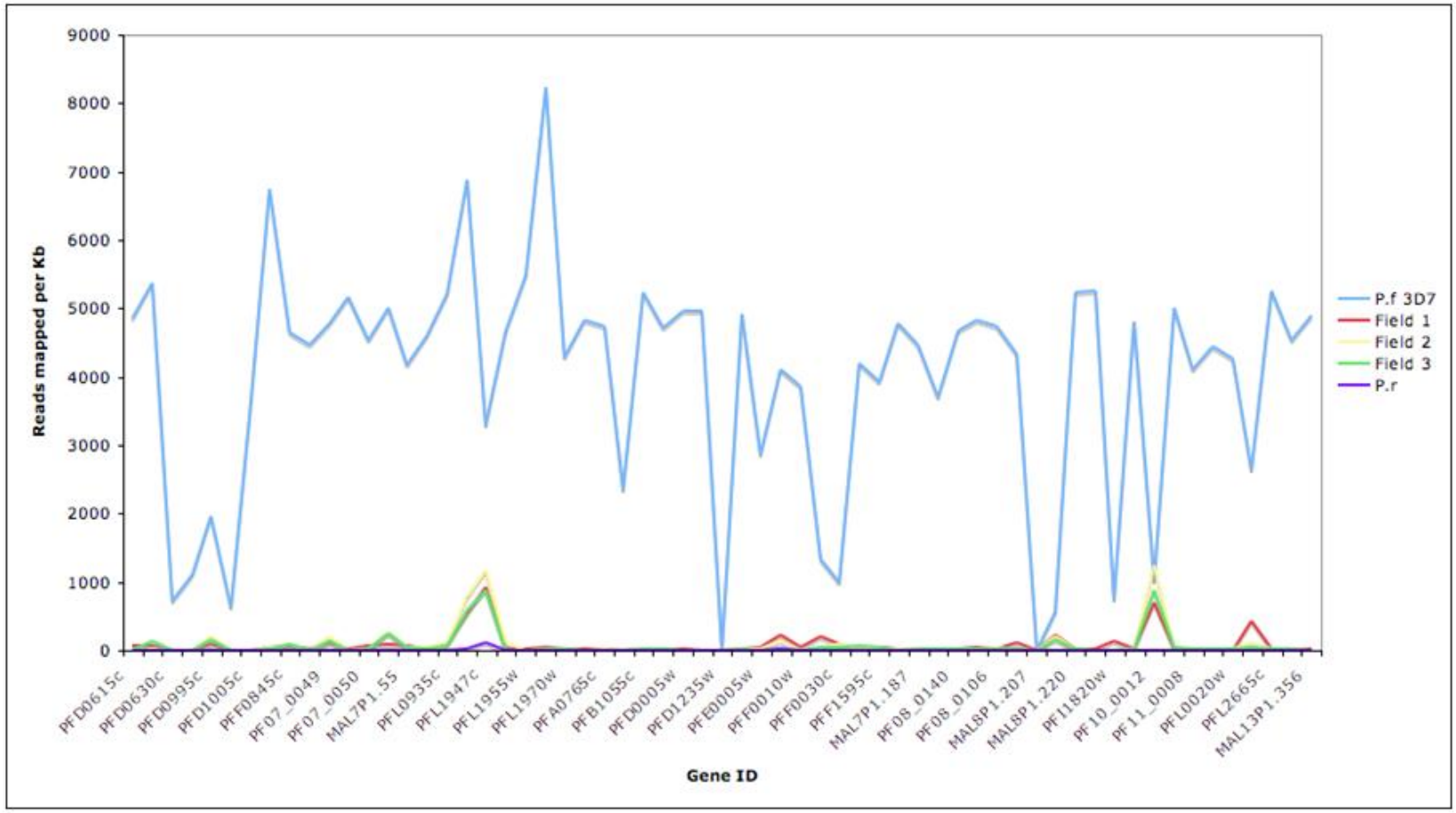


Figure 2.7: Number of reads mapped per kb over *var* genes of the 3D7 genome. Reads from 3D7, three field samaples and *P. reichenowi (P.r)* were uniquely aligned to version 2.1.4 of the reference genome 3D7. A count of reads mapped per kb is shown for the five genomes. The effect of repetitive sequences (shared matches between *var* genes) is shown by the lack of coverage over some *var* genes. Read mapping was very poor for field samples due to high polymorphism in *var* genes.

2.3.6 Understanding sequencing errors in *P. falciparum*

Following observations that suggest a role for sequencing errors in assembly quality, one aim of this section was to look at the extent and distribution of substitution errors in *P. falciparum* sequencing.

2.3.6.1 Overall sequencing errors

Initially, sequencing errors were independently computed for the forward and reverse reads at low ($Q5$), medium ($Q15$) and high ($Q25$) quality bins. Error rates were variable between runs and lanes (Figure 2.8). Low quality bins had higher error rates in all libraries. The variation in errors on the forward and reverse reads was not consistent between instruments. For examples, the Genome Analyser showed higher error rates on the second read at low ($Q5$) quality bins with the exception of NP_i500. On the other hand, error rates were comparable between the forward and reverse reads in HiSeq and MiSeq at all quality bins.

Libraries from the HiSeq 2000 instrument had the lowest error rates ($\sim$0.7%) compared to the Genome Analyser and MiSeq. The four MiSeq libraries (MS_i3K.1-4) had the highest error rates ($\sim$1 to 1.3%) in both the forward and reverse reads across all quality bins. However, these were part of a research and development experiment on long insert protocols and had a lower yield (Table 2.1). They were therefore excluded from further comparisons, as they may not reflect the quality of standard production libraries. The remaining six libraries were used for the analyses described in the following sections.

2.3.6.2 Per-cycle error rates

In order to identify positions that are particularly prone to errors in *P. falciparum*, error rates were computed for each position on both the forward and reverse reads (Figures 2.9 and 2.10).

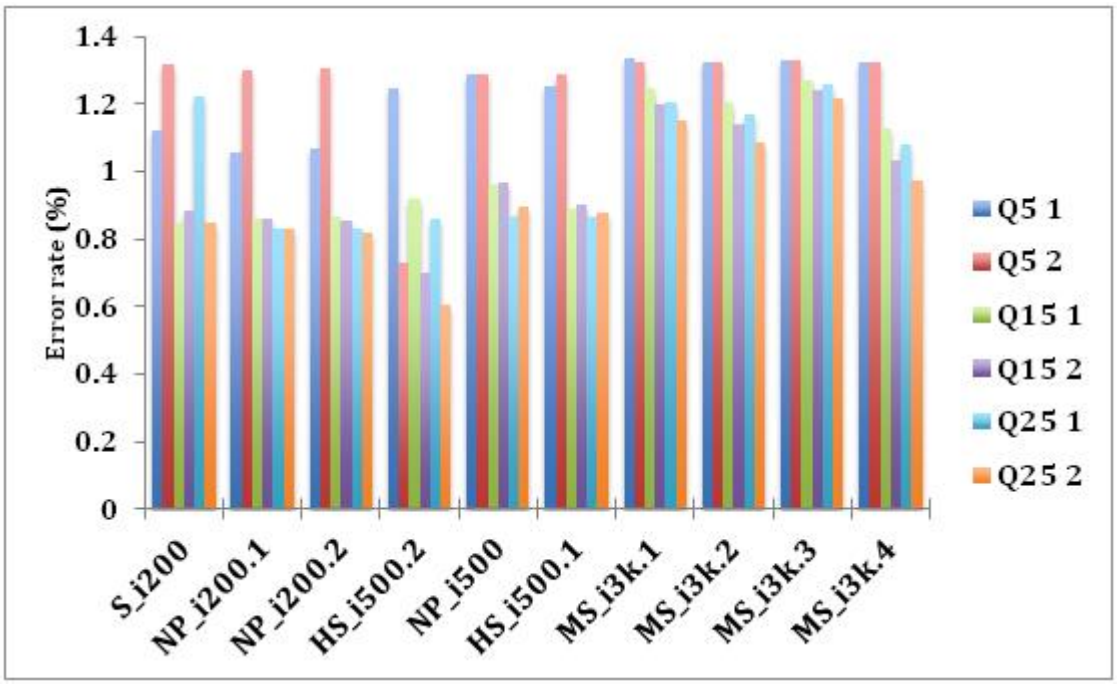


Figure 2.8: Overall error rates at low ($Q5$), medium ($Q15$) and high ($Q25$) quality bins for forward (eg. $Q5$ 1) and reverse (eg/ $Q5$ 2) reads. Error rates were computed for 10 libraries sequenced on GAII, HiSeq and MiSeq platforms representing standard and PCR-free library preparation protocols.

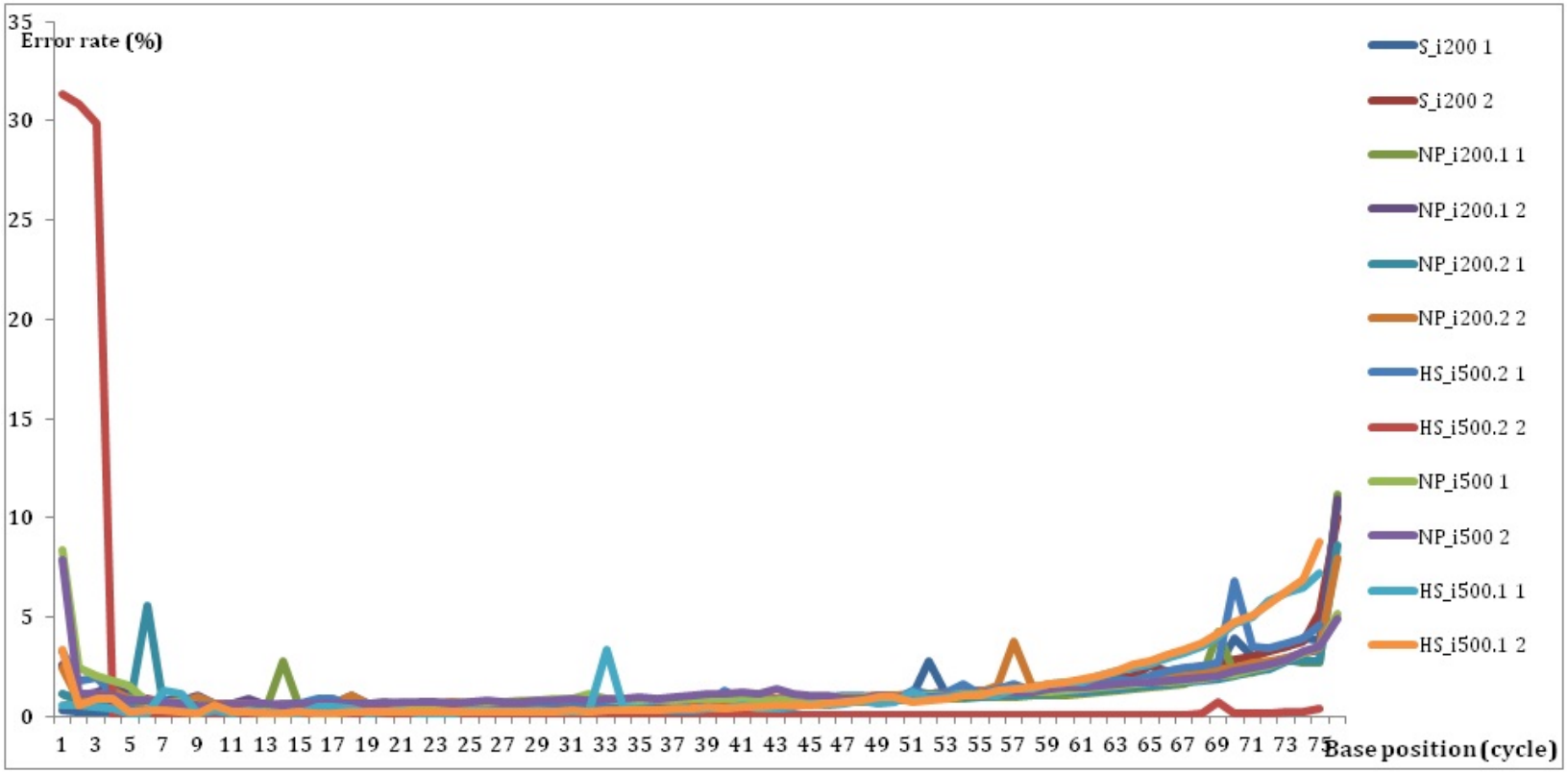


Figure 2.9: Error rates per-cycle for Illumina's GA2 and HiSeq platforms at a quality cutoff of 5 (Q5). Here, error rates (y-axis) were computed for each cycle (x-axis) or base position of the forward and reverse reads. Forward reads are represented by the suffix "1" (eg S_i200 1 represents the forward reads of the library S_i200).

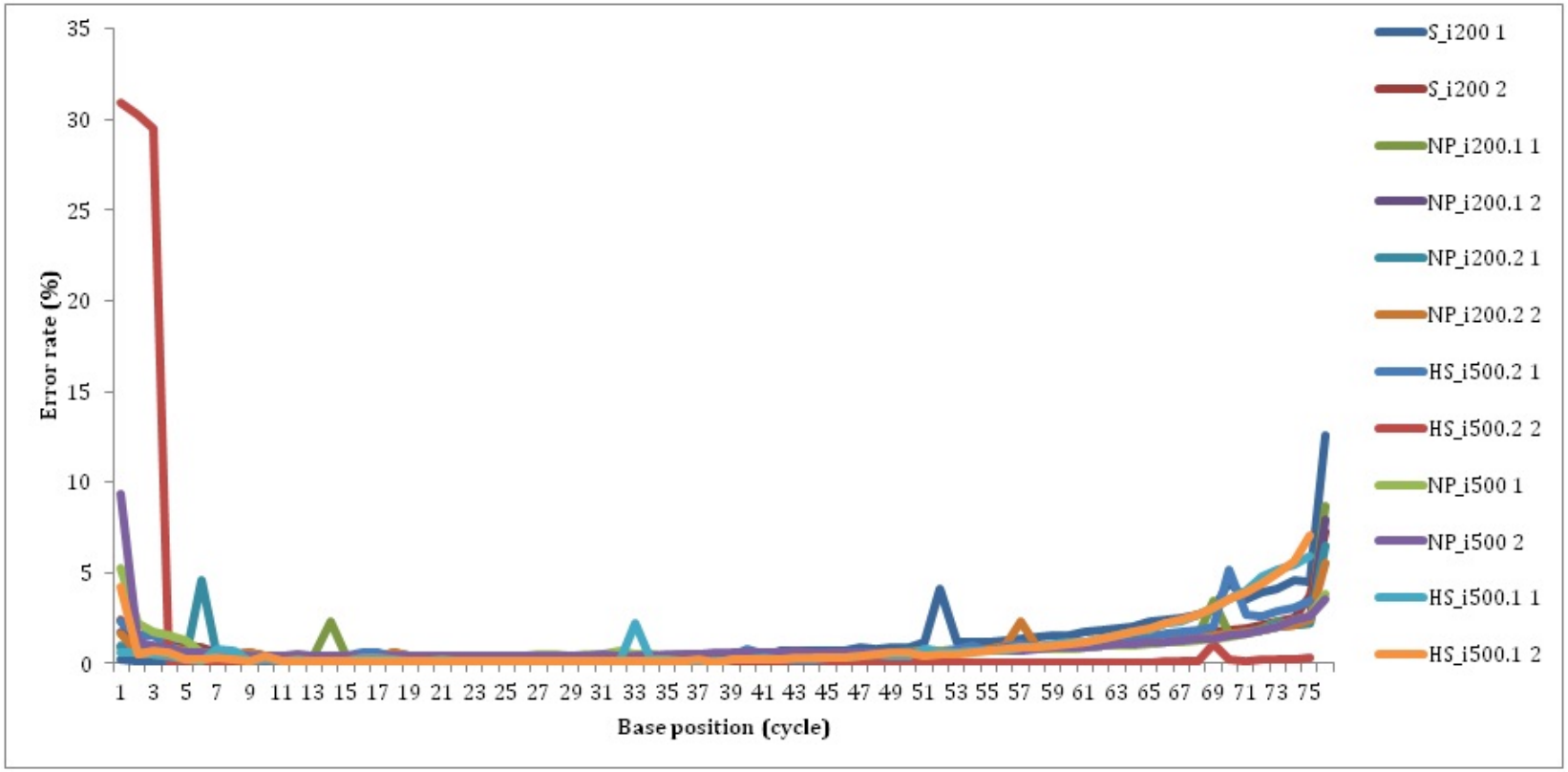


Figure 2.10: Error rates per-cycle of Illumina's GA2 and HiSeqplatforms at a quality cutoff of 25 (Q25). See Figure 2.9 for details.

The highest proportion of errors was found on the final five to seven positions. Surprisingly, a similar trend of increased errors were also observed at the beginning of reads in all libraries affecting both low ($Q5$) and high ($Q25$) quality bins. In five of the six libraries, errors in the first five bases accounted for $\sim$1 to $\sim$17% of the total while the last seven bases contributed a higher 24 to 43% of the total. The second read of the HiSeq library HS_i500.2 had an exceptionally high error rate in the first three bases causing $\sim$92% of all errors. Although error rate in the rest of the positions was very low, high error rates concentrated around a few bases will still affect the overall quality of the read. The occasional spikes in error rates such as those found on base 33 of the first read in library NP_i500.1 were potential indicators of random errors due to a number of reasons including tile-specific problems and issues associated with imaging. Despite the decrease in values as quality increases, the trends of error rates per cycle were consistent at low and high quality bins.

In addition to quantifying the extent of errors for each cycle of the sequencing process, patterns of substitution were investigated in order to understand systematic sources of bias. Proportion of errors due to the 12 potential substitutions (A$\rightarrow C, A \rightarrow G, A \rightarrow T \ldots T \rightarrow A, T \rightarrow C, T \rightarrow G$) were computed for each read of the six libraries (Figure 2.11A and 2.11B). Substitutions A-T, T-G and A-G were over-represented at low quality bins while T-G, A-T, A-G and T-C dominated high quality substitutions.

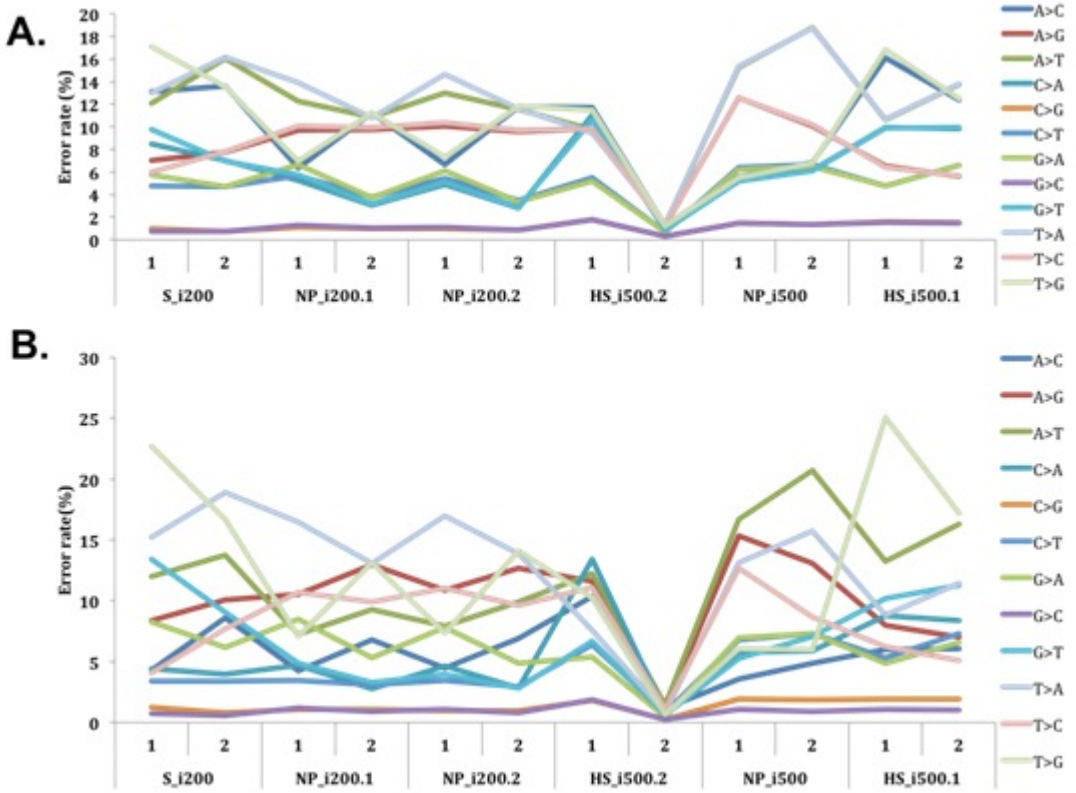


Figure 2.11: Overall substitution patterns of Illumina reads from six libraries. Substitution profiles of the forward and reverse reads of five libraries (as shown on the x-axis) and the rate of such substitutions were plotted for lower base quality of 5 (A) and a higher base quality cutoff of 25 (B).

2.4 Discussion

This chapter was aimed at evaluating the feasibility of using existing short read assembly methods to reconstruct the *var* gene family in *P. falciparum*. A method that works on *var* genes is expected to be effective on other gene families of *P. falciparum* such as the *rif* and *stevor* gene families.

Conclusion 1: *A comparison of short read assembly programs suggested that Velvet is a better choice than SOAPdenovo. However, assembly results from Velvet were not satisfactory.*

A larger proportion of the sequencing and analysis for this chapter was done in 2009 where development of short read assembly tools was in its early days. Assemblers that employ a de Bruijn graph structure were preferred over those with the traditional overlap-layout-consensus approach due to their potential to deal with the increase in sequence yield from next generation sequencing technologies. Although the underlying algorithm used to represent sequences is similar, de Bruijn graph-based assemblers have subtle differences in their pre and post graph processing of reads. For instance, SOAPdenovo applies a read filtering and error correction step based on a predefined *k-mer* frequency cutoff prior to building the graph. On the other hand, Velvet uses a similar approach to remove erroneous reads without correcting base calling errors. In addition, Velvet puts an extra effort in simplifying the assembly graph during both the construction stage and the assembly process by removing singletons. Further reduction in graph complexity and error correction is achieved by removing tips (a node or chain of nodes that have one loose end), bursting bubbles (i.e. merging paths based on sequence similarity and minimum number of reads represented by each path), removing paths that have fewer than the minimum cutoff and using read pair information. The extremely high A+T content of *P. falciparum* poses unique challenges in short read assembly and requires advanced heuristics of resolving ambiguous paths at various stages of the process.

Velvet's superior assembly results compared to SOAPdenovo were therefore attributed to its ability to simplify the assembly graph and remove erroneous paths. However, further evaluation of Velvet on real reads from whole genome, chromosome 1 and *var* genes of the 3D7 genome revealed that the results are not satisfactory as they are highly fragmented. The *k-mer* size was found to be a very important parameter that needs optimisation for a good quality assembly. Smaller *k-mer* values cause potential false overlaps that lead to ambiguities and assembly breaks. Conversely, larger *k-mer* sizes could generate assemblies with a higher contiguity, they require long overlaps and a higher read coverage.

Filtering reads with mismatches did not significantly improve the whole genome *de novo* assembly. This is due to a number of potential problems associated with read length, fragment size, low complexity, uneven coverage and

ambiguous overlaps from repeats that could not be resolved by the assembler. In addition, the decrease in coverage due to the filtering may also affect assembly quality. The overall results are consistent with previous whole genome *de novo* assembly attempts using a PCR-free sample preparation (Kozarewa, et al., 2009). A whole genome *de novo* assembly from a PCR-free library of *P. chabaudi* - a *Plasmodium* species with a higher G+C content - was substantially better (Thomas Otto, personal communication). Therefore low G+C content is one of the primary limiting factors in short read assemblies of *P. falciparum*.

Conclusion 2: *A reference-guided assembly was also not feasible due to a higher level of polymorphism than is acceptable by methods that employ a comparative assembly approach.*

In order for a reference-guided (or mapping based) assembly approach to work, a nucleotide similarity of above ∼90% is required (Pop, 2009). Subtelomeric regions of and *var* genes of *P. falciparum* are very divergent and not positionally conserved. The concept of using a fixed linear reference to guide the assembly of new sequences is thus not relevant. One consequence of the lack of mapping of reads from different genotypes to the telomeres is that all the reads that do not map to the reference will include those that cover the most polymorphic regions. It is therefore clear that assembly by mapping will be of little value when dealing with such highly polymorphic regions of the genome.

Conclusion 3: *Poor quality assembly of short reads was caused by a combination of technical/systematic reasons from the sequencing process and inherent features of the genome. Technical reasons include sequencing errors and uneven coverage due to an enzyme's limited capability to amplify certain regions of the genome. On the other hand, inherent genome features include biased A+T content and repeated sequences.*

Standard quality control pipelines of the Illumina platform often report error rates computed from a fraction of sequenced reads. Such reports provide a quick overview of quality in order to decide whether the data could be used for downstream analyses. However, the information is not enough to understand where the errors occur and whether some of the data could be

recovered. For instance, a closer look at a lane labeled as 'failed' may identify a subset of reads or cycles that contribute to the increased error rate which could then be excluded from further analysis. It is therefore important to further investigate error profiles for each position on the reads at low and high quality cutoff. Although sequencing errors are known to accumulate towards read-ends (Abnizova et al., 2012), it was surprising to see increased error rates at the beginning of reads.

In addition, unexpected high quality substitution errors specific to *P. falciparum* sequences were observed. The Illumina platform uses two lasers to initiate emission of fluorescence from four channels (A, G, C, T) where A,C and G,T pairs share each laser. Although Illumina's base calling algorithm, Bustard, uses a 16-parameter correction matrix (the cross talk matrix) to account for responses from sources other than incorporation of the intended base, cross talk effects are still visible at lower quality errors particularly for transversions (A-C and G-T). The observed high frequency of substitutions A $\leftrightarrow$ *G*, C $\leftrightarrow$ *T* and A $\leftrightarrow$ *T* is therefore less likely to be due to the crosstalk effect. This is potentially due to the extreme base composition and requires a special attention in applications such as variant calling. However, it is difficult to measure the effect of high/low quality substitution errors in assembly as short read assemblers have yet to take advantage of base quality information. Currently, such errors could be accounted for during the assembly process by trimming-off the first and last error prone bases. However, assembly tests performed by trimming read-ends did not improve the quality of contigs for two potential reasons. Firstly, the Velvet assembler has an efficient algorithm of removing erroneous nodes created due to sequencing errors that accumulate towards read-ends. Trimming of reads may thus have very little improvement over the initial assembly. Secondly, as trimming shortens the effective read length, the size of *k-mer*s that could be used for assembly also becomes smaller. As described previously, shorter *k-mer* sizes are likely to generate false overlaps and poor quality assembly.

Conclusion 4: *A slightly different approach to reference guided and whole genome or whole chromosome de novo assembly was required to reconstruct var genes and subtelomeric regions from the current sequence data.*

In summary, even when the best available methods were used, *var* genes could not be reliably assembled. Optimizing for assembling gene-families in particular and removing technical errors improved the results. Even so, technical and inherent bias meant that the assembly remains challenging, and we conclude that none of the current methods can effectively assemble highly polymorphic gene-families.

Chapter 3

New approaches to assemble *var* genes from short reads

3.1 Introduction

As described in Chapter 2, *de novo* assembly of *var* genes is challenging mainly due to high A+T content, shared sequence blocks and uneven read coverage. The polymorphic and mosaic nature of *var* genes adds further complexity to the problem of short read assembly. Despite the high sequence polymorphism in *var* genes, the presence of highly conserved short motifs was reported (Bull et al., 2007; Rask et al., 2010). For example the motif LARSFADIG is located on the DBLα region and found in nearly all *var* genes. Although the frequent recombination events that shuffle sequence blocks play an important role in the evolution of *var* genes (Frank et al., 2008; Kraemer et al., 2007; Rask et al., 2010; Taylor et al., 2000b), they also make use of existing short read assembly approaches inadequate. This chapter focuses on two major challenges in assembling *var* genes: identifying reads that belong to *var* genes and performing a targeted assembly on the collated reads.

Currently, there are no established methods for a targeted assembly of a gene, a group of genes or gene families. One approach could be to run a whole genome assembly followed by identifying contigs that resemble the genes of interest. However, in chapter 2, I demonstrated that this approach had

serious limitations; whole genome *de novo* assemblies of *P. falciparum* are highly fragmented, with the potential to collapse highly similar regions. Due to the high A+T content and low complexity, the uniqueness of most regions is also extremely low, resulting in false joins during the assembly or scaffolding stages. The problem becomes complicated while dealing with sequences from clinical isolates due to the presence of multiple infections, poor data quality and uneven read coverage. Although the blood-stage parasites are haploid, presence of multiple infections results in multiple haplotypes in individual patient samples. As a result, the complexity of the assembly graph increases for example due to bubbles and chimeric connections leading to a highly fragmented assembly. A new approach is thus required to rapidly identify short reads that come from the *var* gene family in order to perform a targeted assembly of short reads.

The problem of genome assembly could be illustrated with the analogy of solving a jigsaw puzzle where short reads are the pieces of the puzzle that need to be organized in order to reconstruct a complete genome or region of the genome. Assembling the *var* gene family could therefore be seen as solving $\sim$60 related puzzles from a mixture of millions of pieces including pieces from non-*var* puzzles (*var* genes are $\sim$0.2% of the genome). Thus a plausible approach would be to first identify the pieces that belong to the $\sim$60 puzzles as a whole and then to find a way to solve each puzzle from the mix.

This chapter aims to develop an alternative assembly approach for *var* genes that:

- addresses limitations of existing assembly approaches, specifically the two challenges described in the previous section:
  - rapidly identifying reads that belong to the *var* gene family, and
  - reconstructing members of the family
- is scalable to thousands of parasite isolates
- builds on existing methods already developed in our laboratory when applicable.

3.2 Methods: Proposed assembly approach

The assembly approach proposed in this chapter has two components. Firstly, identical regions of the *var* mosaic blocks were identified as shared motifs and used to assist with identification of reads that belong to the *var* genes. Secondly, an iterative assembly approach, that takes advantage of de Bruijn graph-based and overlap-layout-consensus assembly approaches, was introduced. The three main stages: pre-processing, generating seed contigs and iterative scaffolding/extension (Figure 3.1A-C) and six processes (1-6) comprise the new approach developed to address the assembly problem of *var* genes.

3.2.1 Preprocessing sequence data

3.2.1.1 BAM to preFasta

The purpose of this stage (Figure 3.1A. 1) is to reduce the dataset by excluding sequence reads that do not come from defined "regions of interest". In addition to minimizing the physical storage requirements (i.e. disk space), this step will eventually improve assembly quality by reducing data complexity as a result of the removal of reads from unwanted regions.

Input file 1: BAM files
Initially, raw FASTQ files of the samples will be stored in the BAM (Li et al., 2009a) file format. BAM files could be obtained as a result of an alignment process to a reference genome or alternatively, in the absence of a reference genome, BAM files will only store raw reads. Although the methods developed in this chapter are applicable in the absence of a reference genome, here, availability of a reference is assumed.

Input file 2: A file with regions of interest
A tab delimited file representing regions of interest is required to identify regions of the genome that will be included in the raw data. The format is shown below:

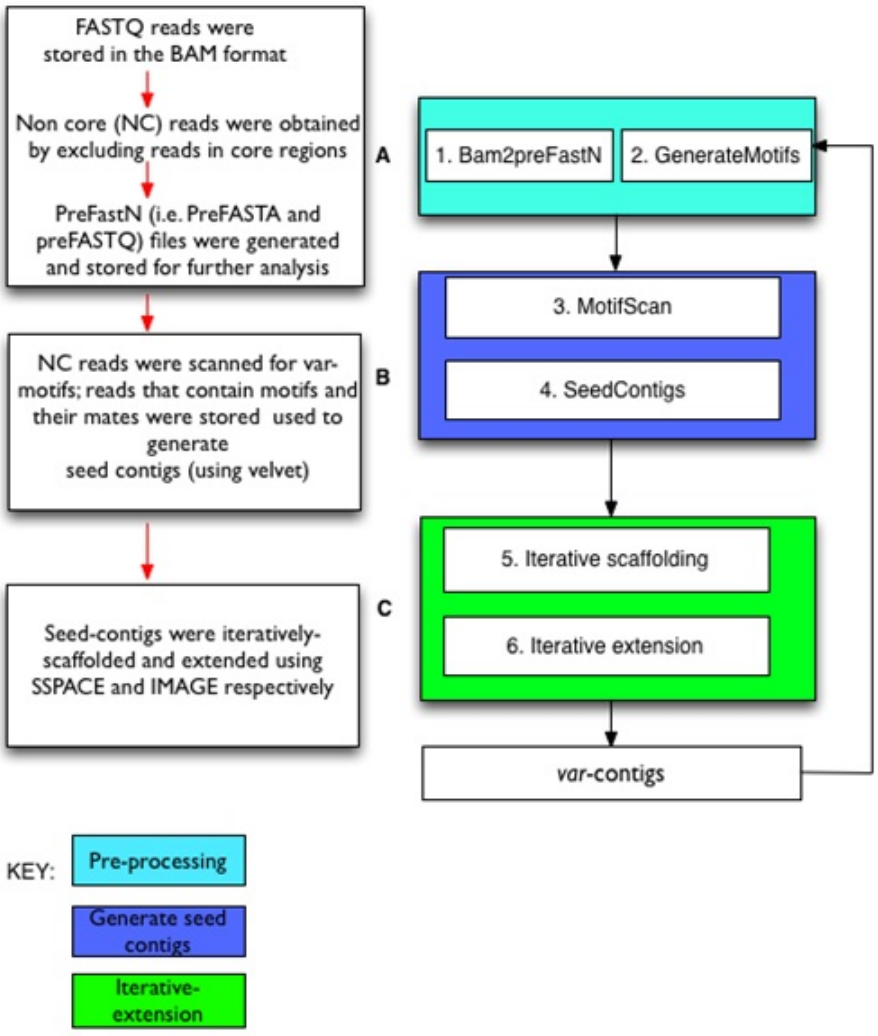


Figure 3.1: A work flow diagram of the iterative assembly approach developed to assemble *var* genes. The three stages and six processes of the iterative assembly approach developed to assemble *var* genes are shown. Conserved regions of *var* genes that were found in a minimum of two culture adapted samples were identified as initial motifs. Similarly, motifs for consecutive iterations were generated by finding conserved regions that were common in *var*-contigs of two or more samples. Core-regions were defined as central regions of the genome except the central *var*-clusters. Non-core reads were thus defined as reads that align to non-core regions of the genome and reads that did not align to the reference genome. See text for further details.

Ch1_Telo.01	MAL1	1	91653
Ch1_Telo.02	MAL1	565425	643292
Ch2_Telo.01	MAL2	1	67545
Ch2_Telo.02	MAL2	860464	947102

where columns 1 to 4 represent regionID, Chromosome, Start and End positions, respectively.

Defining regions of interest

First, a working definition of Subtelomeric regions was obtained by taking regions of the genome from chromosome-ends to the last subtelomeric copy of a *var*, *rif*, or *stevor* gene family. The genome was then divided into subtelomeric regions, central *var* gene clusters and the remaining core regions of the genome.

Regions of interest for assembly of the *var* gene family were defined by combining subtelomeric regions and central *var* clusters resulting in ∼2 Mb (10% of the genome) of non-core regions. In addition, reads that did not align to the genome were also included as they contain potentially novel sequences and highly polymorphic regions. The reads from non-core regions of the genome (i.e. subtelomeric reads, reads in central *var* regions and unmapped reads) are required during the iterative assembly process. It was therefore necessary to define a new file format that allowed efficient data storage.

Defining the preFasta and preFastq file formats

The FASTQ file format used to represent short read sequences from the Illumina platforms is shown below:

```
@Root_ID/1
Forward_read
+
Forward_quality
@Root_ID/2
Reverse_read
+
```

```
Reverse_quality
```

The FASTA file format has no quality values and is defined as:

```
>Root_ID/1
Forward_read
>Root_ID/2
Reverse_read
```

A file format suitable for storing large number samples was adopted in this thesis to store reads from non-core regions of the genome. In order to increase efficiency in data storage, *preFasta* and *preFastq* files contain the minimal sequence information required for subsequent storage and iterative assembly stages. The quality information in FASTQ files is still not used by assembly tools and could be discarded in the preFasta files. In addition, the standard FASTQ and FASTA indicators such as "@", ">", "/1", "/2" could also be discarded during storage and generated in real time while processing the data.

The preFasta file was therefore defined with three columns containing the minimum required information for a FASTA file:

```
Root_ID Forward_read Reverse_read
```

Similarily, a preFastq fill will contain additional two columns to include quality values for the forward and reverse reads:

```
Root_ID Forward_read Reverse_read Forward_Q Reverse_Q
```

This format ensured a significant saving in storage compared to the original FASTA and FASTQ files, which required four and eight lines respectively compared to the single line representation of preFasta and preFastq files. Finally, preFasta/preFastq files were compressed using the Unix command "gzip -9" to ensure additional savings in storage space.

3.2.1.2 Generate motifs

The next step in the pre-processing stage of the assembly pipeline was to generate a list of conserved or shared sequence elements, also known as motifs (Figure 3.1A.2). Initially, a BLAST search was done on a union file of *var* genes from three laboratory-adapted samples (i.e. all against all). However, this approach was not scalable with respect to the number of genes and size of input files.

A *k-mer* based hashing approach was developed to count the frequency of exact matches of sequences with a length of 'k' (k=10 for amino acids and k=30 for nucleotides). Motifs were then grouped according to the number of samples (or genomes) they came from. A 'shared motif' was defined as a motif found in a minimum of two samples. In order to avoid motifs from low-complexity regions that could potentially cause spurious matches, shared motifs that were also present in core regions of the genome were discarded. The information on motifs and motif sharing was represented in the following format:

```
ID  #Populations  #Samples  #Genes  GeneInfo SampleInfo PopInfo
```

GeneInfo, SampleInfo and PopInfo contain comma-separated lines listing the names and frequency of genes, samples and populations that share the motif.

The process of identifying motifs was repeated after each iteration of assembly and extension of contigs (described in sections 3.2.2 and 3.2.3). Addition of new motifs was expected to increase the motif database, enhancing the potential of capturing new sequences and novel variants of the *var* gene family.

3.2.2 Generating seed contigs

3.2.2.1 Scanning raw reads for motifs

Once a set of shared motifs (nucleotide or amino acid) were identified from members of the *var* gene family, raw reads were examined for the presence of the motifs. An exact match of a motif to either the forward or reverse read of a read pair was required prior to storing both reads for the initial stages of generating seed contigs (see next section).

Initially, identifying reads that contained a motif was done via a BLAST search using motifs and raw reads as database and query respectively. However, it became obvious that a BLAST-based scanning approach was too slow and not scalable as the number of motifs increased from a few thousands to millions.

A motifs scanning tool, *MotifScan*, that rapidly identified exact matches was developed to address this limitation. *MotifScan* was written in C++ and could be used to quickly scan for the presence of motifs in a large number of FASTA or FASTQ reads. Furthermore, motifs could be in amino acid or nucleotide formats. For amino acid motifs, reads were translated in all the six reading frames during the scanning process. *MotifScan* slides by one position over the length of each read or its amino acid translation until a match to the motif database is found. It is important to note that scanning until the sixth frame is the worst case scenario as a match to the motif database could be identified earlier. The result of this step was a FASTA/FASTQ formatted list of reads and their mates where the forward, reverse or both reads contained a motif. Including a read where its mate has a motif is an important aspect of the motif-scanning process as it provided additional information that could be used to extend and join seed regions in subsequent stages of the assembly process.

3.2.2.2 Generating seed contigs

Using the reads obtained from the motif scanning process, seed contigs were produced that could then be extended in the next steps. Velvet (Zerbino and Birney, 2008) was used to generate initial contigs as it was shown to have a better assembly quality when compared with other short read assemblers (Chapter 2). The scaffolding option was not used to minimise the risk of potential false joins. As described in chapter 2, choice of a *k-mer* size is an important parameter that needs optimisation for each iteration of assembly. Although Velvet was used in this thesis, the assembly approach described here is able to use a different assembler to generate seed contigs.

3.2.3 Iterative scaffolding and extension

The final stages of the pipeline (Figure 3.1C) involved sub-iterations of joining and extension of seed contigs.

3.2.3.1 Scaffolding

Seed contigs generated in the previous step were expected to be highly fragmented as the initial set of reads represented a very small fraction of the total. The scaffolding step was therefore important to join contigs that had strong read-pair support. At the beginning of this project, there was no stand-alone scaffolder that could be used independently. Short read assembly tools have built-in scaffolding modules that have a very limited flexibility. In order to address these limitations, a scaffolding tool was developed that took advantage of read pair information from standard sequencing libraries. The distribution of fragment sizes was used to estimate gap sizes and join contigs into scaffolds where there was sufficient and unambiguous evidence. SSPACE (Boetzer et al., 2011), a scaffolding software with similar principles became available during the course of development. After testing the performance, I decided to optimize SSPACE instead of continuing working on our scaffolder.

3.2.3.2 Iterative extension

Contigs and scaffolds generated in the previous steps account for less than the total nucleotide content of the gene family. In addition, scaffolds are expected to have gaps with unknown bases (Ns) that need to be filled during an iterative extension step described here. This step is intended to close gaps in scaffolds and extend contig-ends primarily using read pair information (Figure 3.1C.6).

Reads obtained from non-core regions of the genome were aligned to seed contigs generated in the previous steps and used as raw reads to initiate the extension process. By aligning reads to the seed contigs, it was possible to iteratively walk out of seed regions. The process involved a number of sub-iterations of mapping reads to seed contigs, identifying reads that align to contig-ends and performing a local assembly using Velvet. These principles

were implemented in IMAGE (Tsai et al., 2010), a tool developed in our laboratory by Jason Tsai for the purpose of closing gaps in draft assemblies. It was decided to optimize IMAGE for the iterative extension step of the assembly process. IMAGE begins by aligning short reads to contigs given a list of contigs, scaffolding information and raw FASTQ reads. After a number of optimisation steps on IMAGE's various options, the best extension and gap-closure results were obtained from choosing optimal parameters for the number of iterations, *k-mer* values used for local assembly and the minimum alignment score of reads when mapped to seed-contigs. A *k-mer* value of 41 was used over 5 to 7 iterations to obtain good quality gap closure and contig extension. Decisions to stop at 5 iterations or continue to 7 were made based on the number of gaps closed and the improvement in N50 contig size. A minimum alignment score of 70 for 76 bp reads (i.e. use reads that aligned with a score of above 70) was found to be critical as it minimised the effect of erroneous joins between contigs due to poor quality matches.

3.2.4 Evaluating the assembly approach

3.2.4.1 Testing on culture-adapted samples

Sequence data

In order to evaluate the performance and accuracy of the new assembly approach, a total of four laboratory adapted samples were used. DNA for sequencing of the reference isolate 3D7 and the IT sample was obtained from Prof. Chris Newbold's laboratory in Oxford (Chapter 2, section 2.2). Sequences for DD2 and HB3 were obtained from Prof. Dominic Kwiatkowski's laboratory at the Sanger Institute. The PCR-free library preparation protocol described in chapter 2 was used. Raw reads of the four samples were aligned to the 3D7 reference genome version 2.1.4 using SMALT `http://www.sanger.ac.uk/resources/software/smalt/` (A python script written by Martin Hunt in our laboratory was used to align on the following SMALT parameters: *-i 500 -r 10 -x -k 13 -s 6*). Reads are referred to be mapped in proper-pairs if both the forward and reverse reads align to the reference genome facing each other within the expected fragment size range (200-300 bp).

Var *genes*

Var genes for the 3D7 and IT genomes were obtained from GeneDB (`http://www.genedb.org/`). As described in the previous chapter, 3D7 is the reference isolate with a complete genome whereas the other genomes are still highly fragmented with a limited coverage of the *var* repertoire. Supercontigs of HB3 and DD2 genomes were obtained from the Broad Institute (`http://www.broadinstitute.org/`). Sequences annotated as *VAR*/PfEMP1 were identified resulting in 47 and 25 genes for HB3 and DD2 respectively.

Initial Motifs

Initially, *var* genes from 3D7, IT and HB3 genomes were used to generate motifs using Pmatch. However, due to the minimum length requirement of 14 aa, we decided to develop a *k-mer*-basd hashing approach to generate motifs. For the assembly of culture-adapted samples, initial motifs were generated from *var* genes of HB3 and DD2 genomes for two reasons. First, these samples have incomplete *var* repertoire and motifs generated from them would represent a minimal set of starting motifs. It is thus a good indicator of the approach's success in clinical samples. Second, initiating the process on motifs obtained from the other genomes will provide a means to evaluate the completeness and accuracy of the *var* repertoire produced for the 3D7 genome.

Iterations

The process was run for a total of 10 iterations. New motifs were generated at the end of each iteration and used as input for the next iteration.

Evaluating assembly

Assembly quality was evaluated by four commonly used measures: N50 contig size, sum of contigs and largest contig sizes. The completeness of the *var* repertoire was estimated by counting the number of contigs with the DBLα domain. In order to test the accuracy of the contigs generated by the process, the 3D7 genome was used as a reference. All contigs from the 3D7 assembly were aligned against *var* genes of the 3D7 genome.

3.2.4.2 Testing on clinical samples

Sequence data

A total of 50 samples were randomly selected from the Plasmodium Genome Variation (PGV) project at the Sanger Institute. Clinical samples were initially collected and sequenced by Prof. Dominic Kwiatkowski's lab at the Sanger Institute. The samples were randomly selected from 10 different countries representing West Africa, East Africa and South East Asia.

Initial motifs and iterations

Initial motifs for the assembly of 50 clinical samples were generated from *var* genes of 3D7, HB3 and IT genomes. *Var* genes for the three samples were obtained as described in the previous section. Amino acid motifs of length 10 aa were generated and checked for uniqueness. Motifs shared by a minimum of two samples and that were unique to non-core regions and the flanking upstream and downstream regions of 2 kb of the 3D7 genome were selected to initiate the process. The performance of the iterative assembly was enhanced by generating motifs from a six frame translation of contigs that contain the DBLα tag. In order to determine the number of iterations required to gather an optimal number of motifs (i.e. the iteration where the number of shared motifs reaches a saturation), the assembly was run for a total of 20 iterations. Shared motifs obtained at the end of each iteration were checked for quality and used as inputs for the next iteration.

Assembly statistics and quality check

Contigs that contain the DBLα domain were obtained from the 20^{th} iteration and assessed on how close they are from the expected assembly statistics (based on *var* genes of the 3D7 genome). In addition, the size distribution of open reading frames was examined to evaluate the accuracy of the assembly.

3.2.4.3 Additional evaluation using samples from the Illumina HiSeq platform

In order to further evaluate the assembly process, five clinical samples from the latest runs of the Illumina HiSeq platform were selected. These runs have a higher yield and longer read lengths (100 bp, paired end reads) compared to the previous samples used to test the assembly process.

Comparing with* de novo *assembly

Initially, the five samples were assembled using motifs generated at the end of the 20^{th} iteration. After running the iterative assembly process for three iterations, assembly quality of contigs that contained the DBLα tag were examined and compared with scaffolds of a *de novo* assembly (made by Velvet and obtained from Thomas Otto in our laboratory).

Mixed assembly

In order to test the performance of the new assembly approach in parasite samples that have multiple infections, raw reads from four of the five samples were selected. These samples were shown to have a single infection based on the number of contigs that contain the DBLα domain (i.e. expecting 60 genes per genome). Raw reads from non-core regions of the genome were first individually assembled for each sample (*k=71, cov-cutoff=auto*). The reads were then mixed and assembled using identical assembly parameters as the individual assembly. Open reading frames (ORFs) of contigs that contain the DBLα were obtained by translating to all six frames and choosing the frame with DBLα. The contigs from the two sets of assemblies were compared at the protein level using BLAST (*blastp - F F -m 8*). This provided a better assessment than a nucleotide based comparison as regions of extremely low G+C content such as introns and intragenic regions were excluded.

3.3 Results

3.3.1 Defining regions of interest

Regions of interest were defined using the reference genome 3D7 as described in section 3.2.1.1. A simple definition of subtelomeric regions was sought for the purpose of this thesis resulting in a total of ∼2Mb (10% of the genome) from the 28 subtelomeres. An example of the working definition of subtelomeric regions on chromosome 1 is shown in Figure 3.2. Analysis of the size distribution of subtelomeric regions in the *P. falciparum* genome using this working definition revealed that chromosomes 4 and 7 had the longest subtelomeric regions (Figure 3.3).

Although the working definition of subtelomeric regions adopted for this thesis may be different from that of the original genome annotation (defined using synteny with closely related species), it was possible to capture highly polymorphic regions that are currently being excluded in studies that rely on a unique alignment of short reads to call single nucleotide polymorphisms and copy number variations.

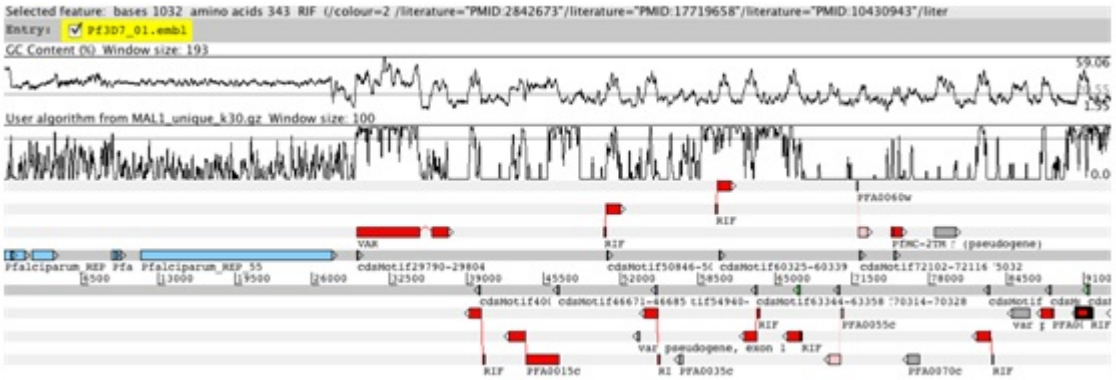


Figure 3.2: A working definition of subtelomeric regions adopted for this thesis: Regions of the genome from the end of each chromosome to the last subtelomeric copy of a *var*, *rif* or *stevor* family were identified as subtelomeric. The red blocks represent coding genes, grey boxes represent pseudo-genes, cyan blocks at the left-end represent repeats. The two plots on the top panel show the G+C content and *k-mer*-based uniqueness plots.

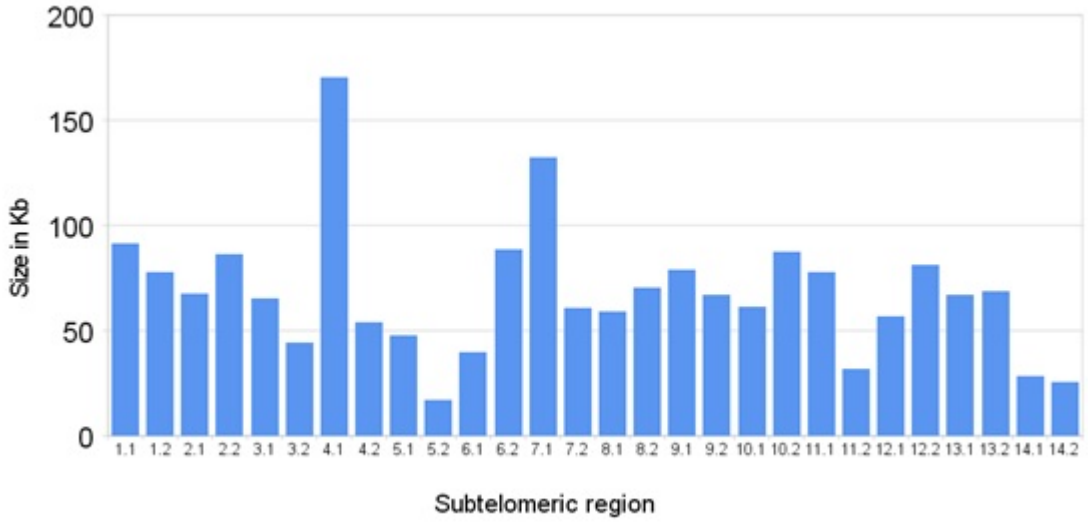


Figure 3.3: Size distribution of subtelomeric regions in the reference genome 3D7 by Chromosome. Sizes range from 17.5 to 170 kb covering ∼10% of the genome.

3.3.2 Evaluating the new assembly approach using culture-adapted samples

3.3.2.1 Sequence data

A summary of raw reads for the four culture-adapted samples are shown in Table 3.1. The number and proportion of non-core reads varied in the four samples from ∼9 to 34% depending on the number of reads aligned to the genome and reads that did not align (eg. due to poor read quality).

The HB3 genome had a significantly higher number of non-core reads due to the higher number of reads that did not align to the 3D7 genome (∼33%). This could be due to a decrease in base quality of the second read after the $\sim 50^{th}$ cycle (Appendix B, Figure B-2).

	Total read pairs	%All reads aligned to 3D7 (Aligned in proper pair)	Non-core read pairs (%total)
3D7	12,488,019	96(93)	3,453,588(28%)
DD2	20,650,235	91(84)	2,327,890 (11%)
HB3	18,501,446	67(73)	6,369,447(34%)
IT	27,013,569	97(86)	2,477,392 (9%)

Table 3.1: A summary of the four culture-adapted samples and non-core reads used to evaluate the iterative assembly approach. Reads that aligned in the correct orientation (facing each other) and within the expected insert size range are defined as 'aligned in proper pair'

3.3.2.2 Motif generation and iterative assembly

A total of 17,719 motifs (10 aa overlapping *k-mer*s) were found to be shared between *var* genes of HB3 and DD2 genomes. After excluding motifs that were also found in the core regions of the 3D7 genome, a total of 7,353 motifs remained to initiate the iterative assembly process. As the number of iterations increased, the number of motifs shared by a minimum of two of the four samples also increased (Figure 3.4). However, the rate of increase in new motifs declined after the $\sim 4^{th}$ iteration due to a potential saturation in the motif space. The highest improvement was found in the first two iterations where the number of motifs increased from the initial ∼7,000 to ∼180,000.

Assembly quality improved with more iterations (Figure 3.5). The most notable improvement was on the 2^{nd} iteration of the process. Overall, while the sum of contigs as well as N50 and largest contigs sizes increased with subsequent iterations, the number of contigs decreased indicating an improvement in assembly quality.

Although 3D7 had the best N50 contig size (∼5 kb), it also had the least number of contigs with DBLα and least value in sum of contigs compared to the other three samples. The N50 contig size is affected by the number of contigs available and thus may give a wrong impression of quality if not taken together with other measures as described here. The sum of contigs ranged from under 300 kb for 3D7 to ∼500 kb in the IT assembly. The number of contigs was comparable between samples during the first iteration (∼150 contigs per sample). However, in subsequent iterations the contig count for 3D7 stayed

below 100 while the other samples generated ∼300 contigs. This variation in the number of contigs is reflected in the N50 contig sizes as the highest N50 values for 3D7 correspond with the fewest contigs. The efficiency of the iterative extension process was also shown by the sizes of the largest contigs which increased from ∼5 kb to ∼12 kb during the course of the iterations. The number of contigs that contained the DBLα tag also showed improvement from the second iteration and converged to ∼40 to 50.

In summary, the iterative assembly generated a substantially higher number of *var*-contigs (i.e. contigs with the DBLα tag) compared to the original number of genes found in the HB3 and DD2 genomes used to initiate the motif generation process. It was possible to recover up to ∼80% of the expected *var* repertoire in the test samples by starting from a very limited set of initial motifs.

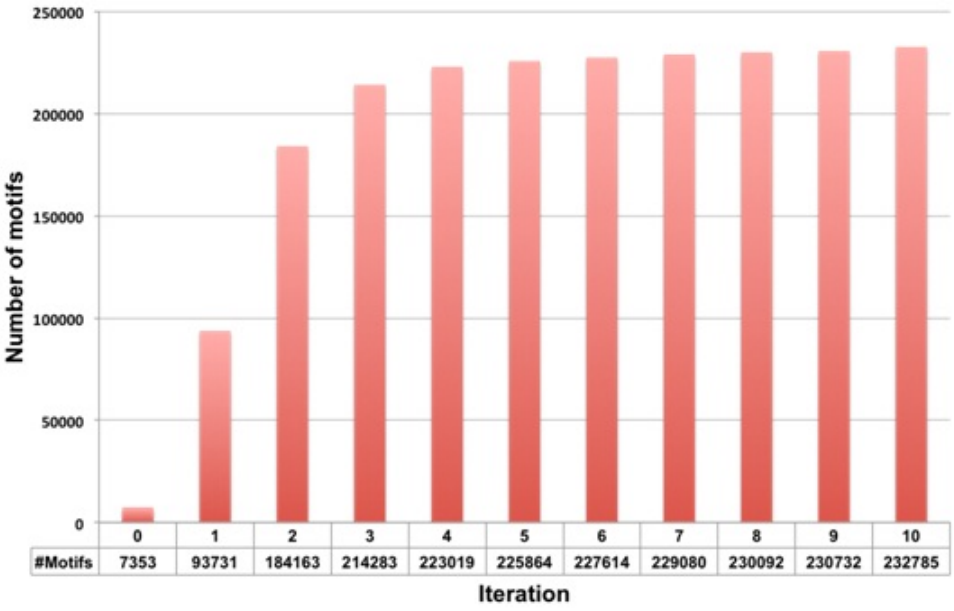


Figure 3.4: The cumulative number of shared motifs for 10 iterations of the four lab adapted samples 3D7, IT, HB3 and DD2. Initial motifs were obtained from HB3 and DD2 in order to test the assembly approach with a limited number of starting motifs.

Evaluating var contigs of the 3D7 genome

Optimal assembly results for 3D7 were obtained at the 5^{th} iteration (Table 3.2).

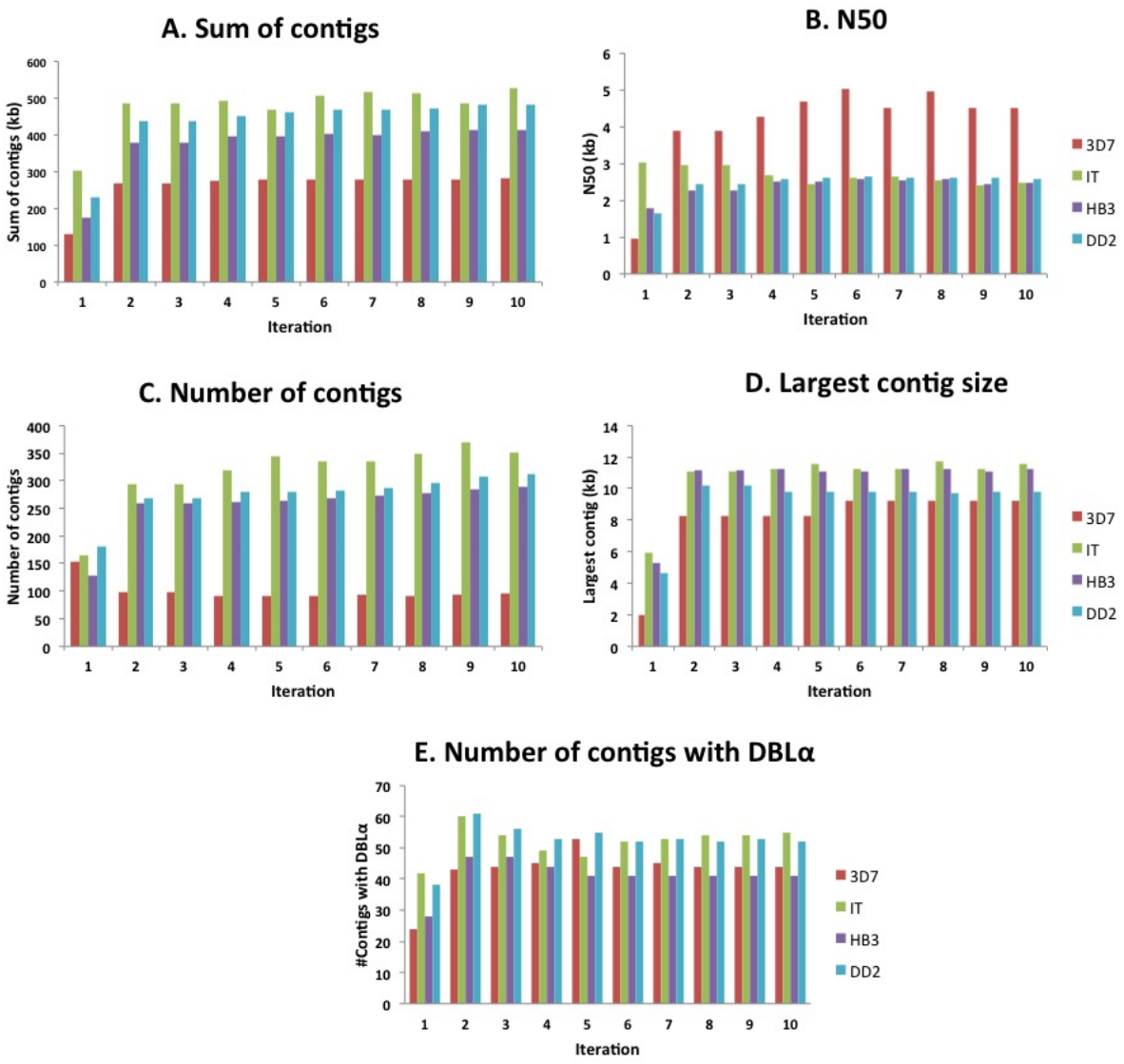


Figure 3.5: Assembly statistics of the four lab adapted samples for 10 iterations. Five assembly measures Sum of contigs, N50, Number of all contigs, largest contig and number of *var*-contigs were used to asses the assembly quality with an increase in iteration. A). Sum of contigs was used to measure the coverage of the *var* repertoire. The N50 contig size (B), the number of contigs (C) and the largest contig size (D) measure assembly contiguity. Completeness of the *var* repertoire is measured by counting contigs that contain the DBLα tag (E). The 3D7 assembly had the highest N50 and fewer contigs, but also had the least contig coverage as shown in A.

Comparing contigs generated from the iterative assembly to the 3D7 genome provided a measure of completeness and accuracy. Of the 46 contigs that contained the DBLα domain, only one misassembly was detected due to a chimeric connection between two genes of the central *var* cluster. Although it does not contain the DBLα domain, var2CSA was assembled in one contig producing a full length (intact) gene. The completeness of the repertoire in the 3D7 genome was therefore estimated to be ∼75% (i.e. expecting 61 *var* genes in the genome). Coverage of the *var* repertoire in 3D7 was evaluated by aligning all 93 contigs to the genome. Inspection of comparisons performed using BLAST and Abacas revealed that 46 of the 61 *var* genes in 3D7 were covered by contigs (Table 3.3). A total of 21 genes were partially covered (minimum coverage of 50% at 99% identity) and the remaining 25 genes were fully covered by one or more contigs. The majority of the genes were fully or partially covered by single contigs (intact).

	All contigs	**Contigs with DBLα**
Sum	277761	196368
N50	5039	5540
Num.	93	45(1*)
Largest	9223	9223

*mis-assembeled contigs

Table 3.2: Iterative assembly results for *var* genes of the 3D7 genome.

	Fully covered	**Partially covered**
Exon1 only	1	7
Exon1(+ Ups)	2	3
Exon1(+ Intron)	7	6
Exon1(+Ups +Intron)	15	5
Total	**25**	**21**
Intact	21	16
Fragmented	4	5

Table 3.3: Coverage of *var* genes in the 3D7 genome. A total of 46 (of the expected 61) genes were covered by one or more contigs.

The iterative assembly approach was able to recover ∼75% of *var* genes in

the 3D7 genome. As described earlier, the initial motifs were obtained from HB3 and DD2. The results were thus very promising as they illustrate the potential of this approach in reconstructing a large proportion of the repertoire in unrelated clinical samples. The missing genes are expected to be due to insufficient seed motifs as the assembly was performed using motifs generated from HB3 and DD2. The single misassembled contig was a result of chimeric join between two genes of the central cluster. These genes share identical regions of larger than 1 kb and could not be resolved using a standard library (200-300 bp). In addition, the smaller fragment sizes in the 3D7 library (quartiles: 143,163,188) may also contribute to poor quality assembly.

3.3.3 Evaluating new assembly approach using clinical samples

The iterative assembly approach was further evaluated using 50 clinical samples from 10 countries (Table 3.4). Samples were chosen from standard PCR-free libraries (insert size 200-300 bp) with a read length of 76 bp.

The three laboratory clones 3D7, HB3 and IT were used to generate initial motifs. A total of 8,766 motifs were shared by at least two of the three samples and also passed quality control steps. The number of motifs increased with each iteration as observed in the lab-adapted samples. However, the rate of increase in motif acquisition was slower after the 10^{th} iteration (Figure 3.6). Each iteration involved sub-iterations of scaffolding and extension that helped improve the quality of assembly. Assembly results of the 50 samples are summarised in Figure 3.7. At the end of the 20^{th} iteration, the average number of contigs that contain the DBLα (n=2,793) was close to the expected value of $\sim$3,000 (i.e. expecting $\sim$60 per genome). In addition, the N50 contig length (6.4 kb) and the largest contig ($\sim$14 kb) sizes were also within the expected range of values for *var* genes with the DBLα tag in the 3D7 genome (sum=428 kb N50=7.7 kb; Number of contigs =54; Largest contig=12.5 kb). Box plots showing the distribution of 2,769 *var*-contigs within the 6 groups (Figure 3.7) show a similar distribution with the 3D7 and other clinical samples studies by Bull and colleagues (Bull et al., 2007).

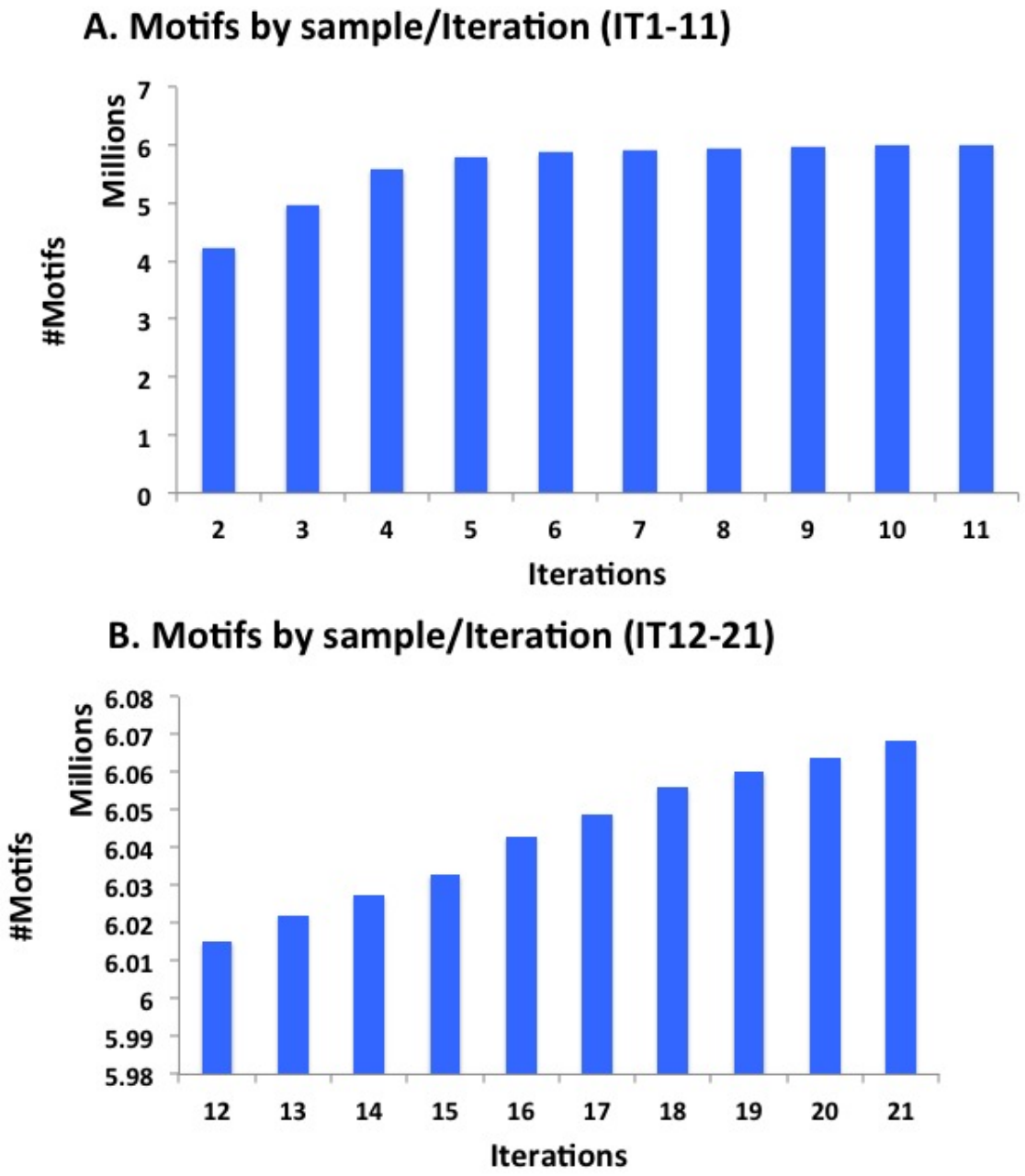


Figure 3.6: Number of motifs shared by at least two samples (i.e. *var*-contigs from two different samples) per iteration for 50 clinical samples. **A)**. The increase in number of shared motifs for the first 11 iterations is shown separately in A. The rate of increase in shared motifs was higher for the first ∼5 iterations. **B)**. The number of shared motifs continued to increase at a slower rate after the 12^{th} iteration as shown by the scale of the y-axis.

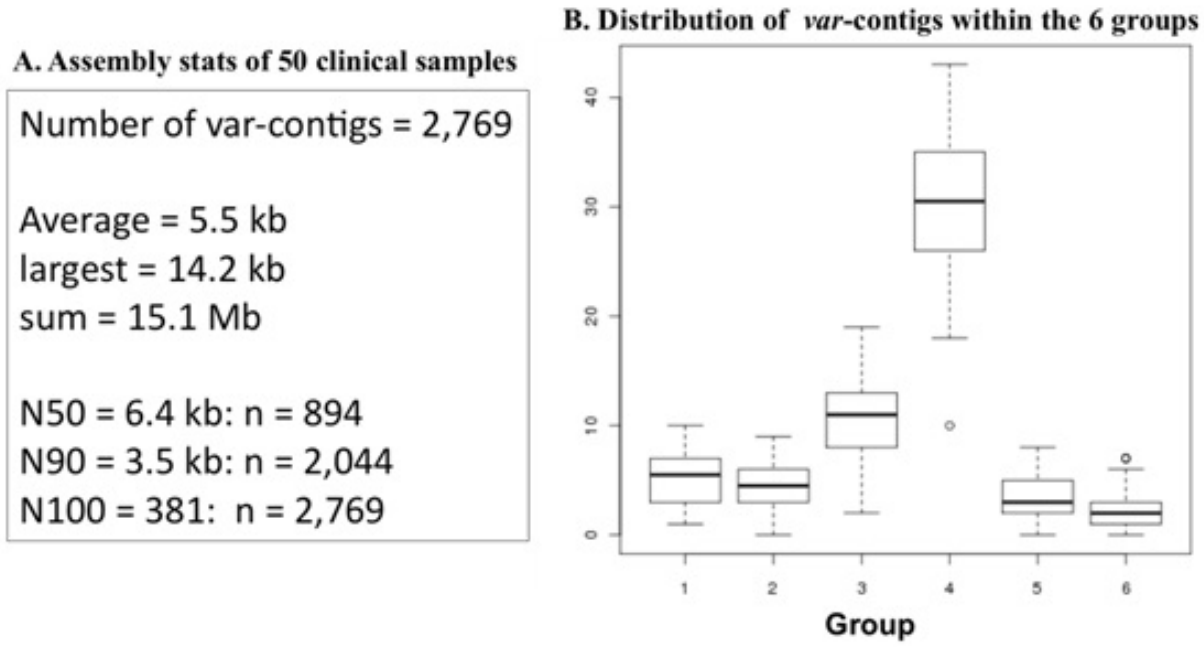


Figure 3.7: Assembly statistics of 50 clinical samples. **A)**. A total of 2,769 contigs had the DBLα tag representing $\sim$92% of the expected *var* repertoire. The number of contigs represented by the N assembly measures were also shown. For example N50 of 6.4 kb; n=894 indicates that a total of 894 contigs are above 6.4 kb in size. The sum of these contigs is equivalent to $\sim$7.5 Mb (i.e. half of the total sum of contigs). **B)**. *Var*-contigs were grouped into one of the six groups using the 'Cys-POLV' grouping method of Bull and colleagues (2007) (Bull et al., 2007).

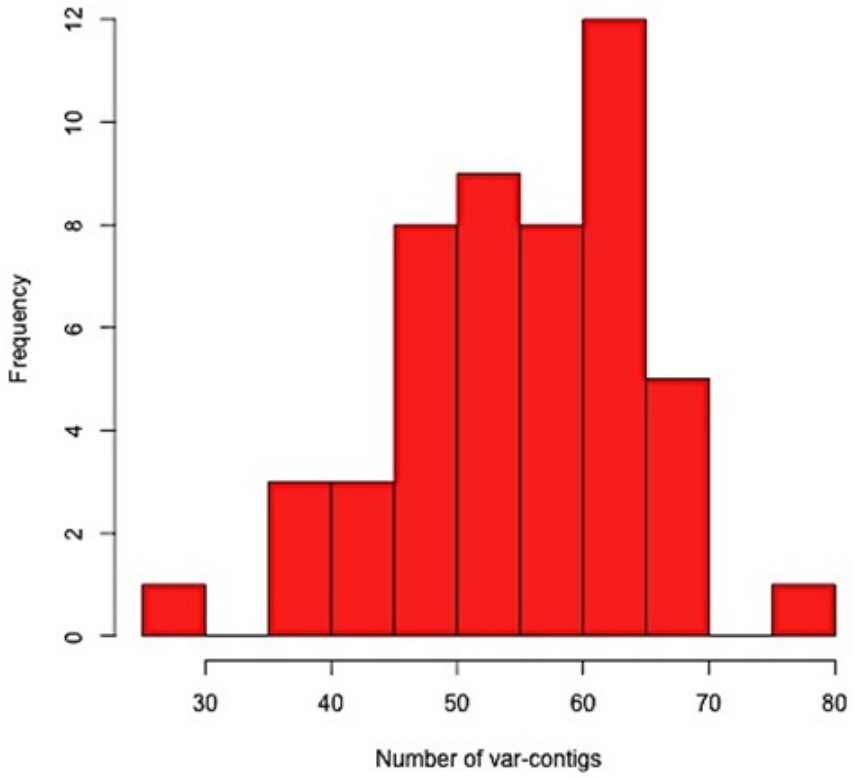


Figure 3.8: A histogram of the number of contigs with DBLα (*var*-contigs) per sample for the 50 clinical isolates. Two samples (PP0011 from Peru and PM0096 from Mali) were found on either end of the distribution with 26 and 78 *var*-contigs respectively.

Sample Origin	Sample IDs				
Gambia	PA0012	PA0032	PA0066	PA0081	PA0091
Kenya	PC0057	PC0070	PC0075	PC0080	PC0083
Thailand	PD0126	PD0127	PD0133	PD0134	PD0138
Ghana	PF0211	PF0231	PF0263	PF0288	PF0290
Cambodia	PH0142	PH0145	PH0380	PH0479	PH0483
Mali	PM0048	PM0090	PM0096	PM0132	PM0162
PNG	PN0027	PN0054	PN0056	PN0057	PN0059
Peru	PP0005	PP0006	PP0010	PP0011	PP0012
Bangladesh	PR0001	PR0002	PR0005	PR0006	PR0008
Uganda	PW0003	PW0009	PW0010	PW0013	PW0016

Table 3.4: Clinical samples used to test the iterative assembly approach. A total of 50 samples were chosen from 10 countries representing Africa, South East Asia and South America.

The average number of contigs with DBLα was $\sim$56 (Standard Deviation/SD=10) indicating that most samples contain the expected number of var genes. Isolates PM0096 and PP0011 were at the two extreme ends of the normal distribution with 78 and 26 *var*-contigs respectively (Figure 3.8). The sample with least number of *var*-contigs (PP0011) had poor data quality that affected the assembly (Appendix B). On the other hand, an increase in the number of *var*-contigs beyond the expected assuming a normal distribution ($\sim$76; considering 56+2SD) may indicate the presence of multiple infections. Additional challenges are envisaged with poor quality sequence data and mixed infections. Further evaluations taken to specifically look at these issues are described in the following sections.

3.3.4 Additional evaluations

3.3.4.1 Comparisons with *de novo* assembly

The sequencing technology and assembly tools have improved since the beginning of this thesis. It was thus important to evaluate if a *de novo* assembly of field isolates is practical due to significant improvements in yield and sequence quality. Although the assembly results were better than found in 2009, the iterative assembly approach described in this chapter generated the greatest

number of *var*-contigs (Table 3.5).

Clinical Sample	DBLα count	
	3 iterations	*De novo*
1	**57**	**48**
2	**56**	**45**
3	**54**	**39**
4	**61**	**48**
5	**97**	**86**

Table 3.5: Comparing iterative assembly with *de novo* assembly using 5 clinical samples (read length =100 bp; Velvet used for *de novo* assembly). The iterative assembly approach generated more *var*-contigs than a simple *de novo* assembly.

The iterative assembly approach was able to generate N50 contig sizes of up to 7.9 kb (Table 3.6) for Sample 1 (57 *var*-contigs) indicating both the efficiency of the method and benefits of long reads in assembling *var* genes.

Sample	Sum(bp)	N50(bp)	*var*-contigs	Largest(bp)
1	423,521	7,902	57	13,813
2	418,299	7,478	56	13,740
3	314,747	6,236	54	10,229
4	401,616	7,139	61	12,715
5	610,588	6,708	97	12,457

Table 3.6: Assembly statistics of the five clinical samples using the iterative assembly approach.

Sample 5 had the highest number of *var*-contigs suggesting presence of multiple infections. It was thus excluded from subsequent comparisons.

3.3.4.2 Mixed assembly

Reads from the first four of the five clinical samples (previous section) were first assembled individually for one iteration resulting in 58 to 67 *var*-contigs (Table 3.7). The assembly results were different from those shown in the previous section (comparisons with *de novo* assembly), as expected from the iterative process. The completeness of the *var* repertoire (i.e. number of contigs) and coverage (the sum of contigs) varies with each iteration due to the increase

in reads that are available to perform the iterative assembly. Concatenating contigs from the four samples resulted in a total of 249 *var*-contigs. Conversely, assembly of mixed reads generated in a total of 230 contigs with DBLα.

As these samples were obtained directly from patients, evaluating the quality of the assembly by comparing to the reference genome was not informative. We therefore compared Open Reading Frames (ORFs) from the two sets of assemblies using BLAST. The results revealed that ORFs from the mixed assembly overlapped with 33 to 84% of ORFs in the individual assembly (99%, match length of 300 aa) (Table 3.8).

Sample	Sum	N50	Num. scaffolds	Largest
1	316,696	5,984	67	10,745
2	323,601	6,263	65	10,990
3	134,458	4,011	58	8,420
4	317,902	6,915	59	11,271
Total	**1,092,657**	**6,163**	**249**	**11,271**
Mixed assembly	**813,249**	**5,348**	**230**	**11,114**

Table 3.7: Comparing individual assembly with mixed assembly of four clinical samples: Assembly statistics. Sample 3 had a relatively poor quality assembly compared to the other three samples.

As these samples were obtained directly from patients, evaluating the quality of the assembly by comparing to the reference genome was not informative. We therefore compared Open Reading Frames (ORFs) from the two sets of assemblies using BLAST. The results revealed that ORFs from the mixed assembly overlapped with 33 to 84% of ORFs in the individual assembly (99%, match length of 300 aa) (Table 3.8).

ORFs with the DBLα domain were first extracted from contigs of the two sets of assemblies (i.e. mixed and individual assemblies). ORFs of the mixed assembly were then compared with ORFs from the four samples resulting in an overlap of 33 to 84% of the total in each sample (99% identity and 300 aa). The fewer ORFs in Sample 3 are indicative of a poor quality assembly. Although 300 aa was a reasonable size to compare the two sets of ORFs, the results presented in Table 3.8 do not reveal the extent of overlap between the ORFs (i.e. the proportion of each ORF aligned at 99% identity, also called the

	Sample1	Sample2	Sample3	Sample4
Found in mixed assembly (total)	48(62)	52(68)	14(42)	48(57)
%Total	77	76	33	84

Table 3.8: Comparing ORFs of individually assembled contigs with ORFs of the mixed assembly. ORFs from the mixed assembly were compared with the ORFs obtained from the four samples. This table shows the count of ORFs of the individual assemblies that overlapped with ORFs of the mixed assembly with a minimum match length of 300 aa and a minimum identity of 99%.

coverage of ORFs). Therefore, additional comparisons were made based on the coverage of ORFs by varying the threshold from 5 to 100% (Figure 3.9). Up to 30% of contigs in the individual assembly were matched to mixed assembly over the full length of their ORFs. The remaining contigs were only partially covered with break points potentially caused by repetitive or shared sequence blocks. Aligning raw reads back to the contigs allowing multiple mapping for non unique reads revealed regions of excess coverage that correlated with contig-ends (i.e. break-points) in the mixed assembly (Figure 3.10).

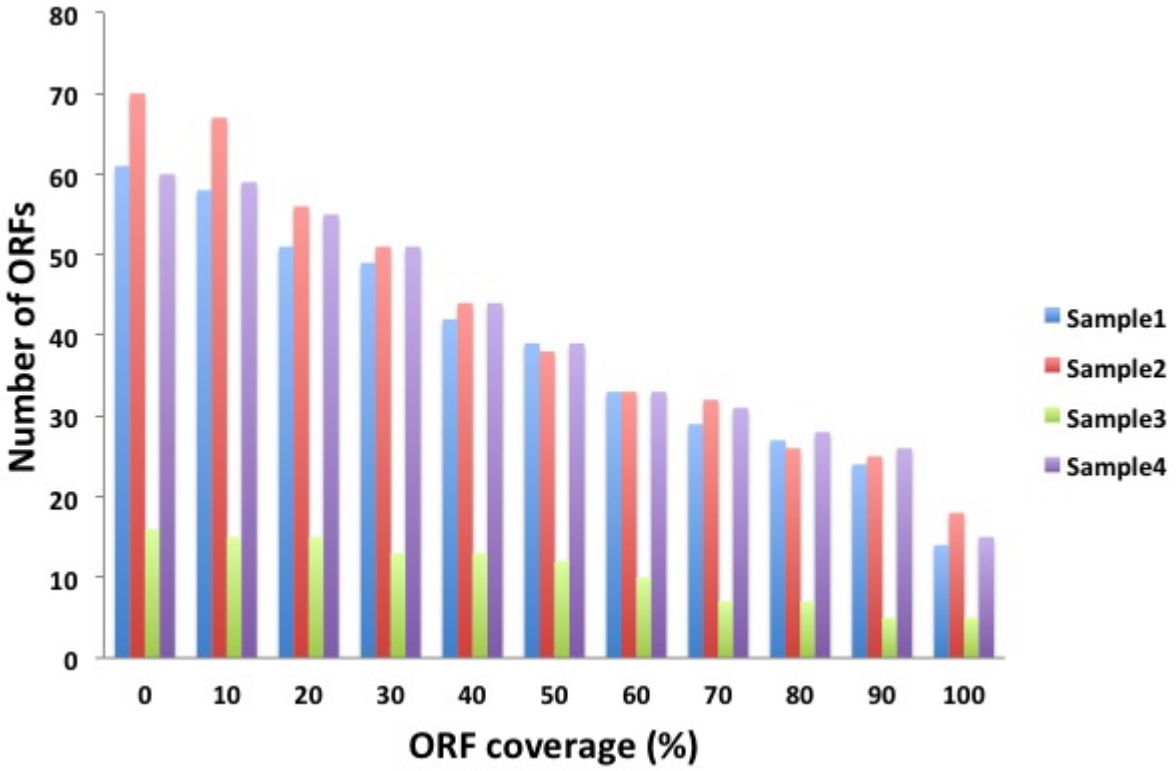


Figure 3.9: Comparing ORFs obtained from individual and mixed assembly by varying the proportion of ORFs covered. ORFs from the mixed assembly had a higher overlap with ORFs from individual assemblies at lower coverage thresholds. As the coverage requirement increased, the number of ORFs (from each individual assembly) that overlapped with ORFs from the mixed assembly decreased.

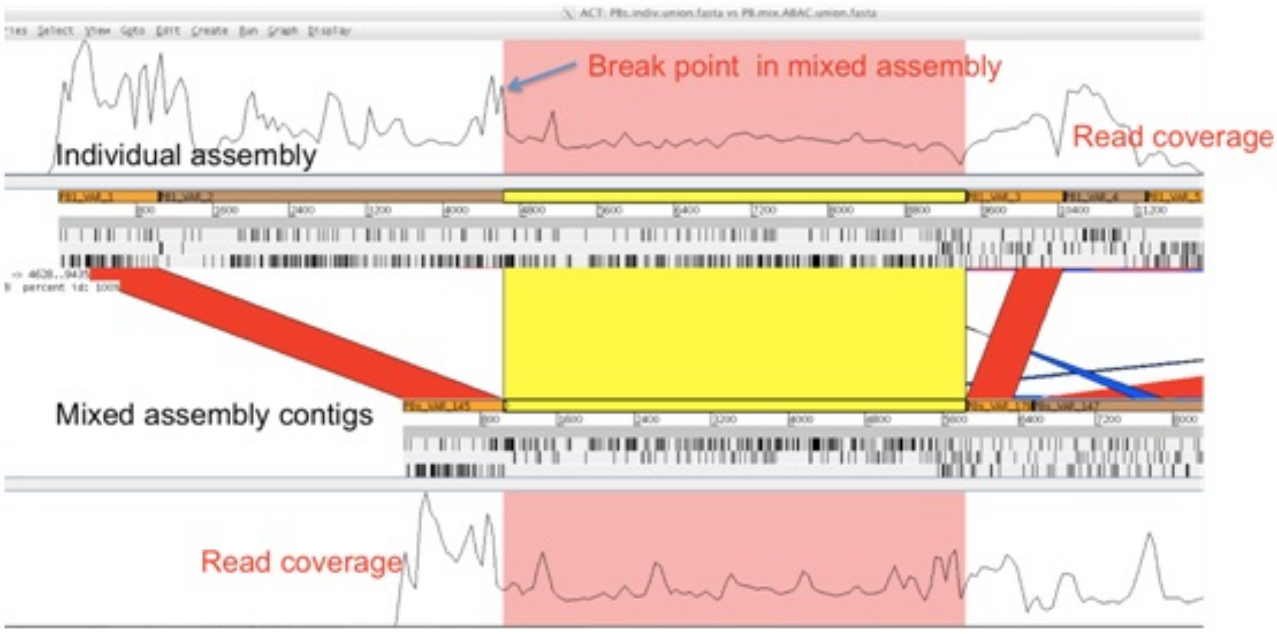


Figure 3.10: An ACT view of read coverage over *var*-contigs generated from individual and mixed assemblies. This example shows a BLAST comparison of *var*-contigs from the individual assemblies (top panel) against the mixed assembly (middle panel). Contigs are shown in alternating orange and brown blocks. The yellow, red and blue blocks (vertical) show synteny matches. Black bars in the middle panel represent stop codons. The contig on the top panel is partially covered (shown by the yellow match) by a contig from the mixed assembly with the breakpoint corresponding to a repetitive region (shown by the increased read coverage).

3.4 Discussion

The aim of this chapter was to develop an alternative approach that will address the limitations of existing assembly approaches in polymorphic gene families. This section presents a discussion of the findings in the following four conclusions.

Conclusion 1: *An iterative assembly approach based on conserved motifs provides a better way of reconstructing the var gene family in P. falciparum*

Gene families have blocks of highly similar and polymorphic regions and continue to evolve by accumulating additional polymorphism as well by recombination within these families. This ongoing microevolution leads to the maintenance of an extremely diverse gene repertoire. The *var* gene family is especially known to contain mosaic blocks that are highly recombinogenic and polymorphic (Bull et al., 2008; Frank et al., 2008; Kraemer et al., 2007). These regions pose significant challenges to standard short read assembly approaches. We have developed an assembly approach that takes advantage of the mosaic nature of *var* genes such that short conserved are used to initiate an iterative assembly. The new approach produced *var*-contigs (contigs that contain the universal DBLα tag) that were accurate and had high repertoire coverage. The efficiency, coverage and accuracy of the approach were demonstrated using sequences from four culture-adapted and 50 clinical samples. The potential of our approach to accurately assemble clinical samples was demonstrated first by assembling *var* genes of the 3D7 genome where initial motifs were generated from unrelated samples with incomplete *var* repertoire. The single misassembly detected in the 3D7 *var* assembly was due to a merge in two highly identical regions of *var* genes on chromosome 12 (central cluster). Such wrong joins are expected as identical segments between *var* genes of the 3D7 genome could be as long as $\sim$5 kb (Chapter 2). Resolving ambiguities in the assembly graph of such long shared segments is not possible with the standard library size of 200 - 300 bp. A fragment size longer than the shared unit is required to accurately assemble the full length of *var* genes that share long sequence stretches. Such cases of misassembly result in frame shifts during aminoacid translations and

were detected as part of the quality control process. Additional quality checks that take advantage of read-mapping coverage of raw reads are also being incorporated in the assembly pipeline. Examination of the read coverage patterns will reveal drops in paired-end coverage as potential signs of false joins.

Conclusion 2: *Motifs are an important part of this approach*

Motifs provided an important part of the assembly process. Identifying conserved (shared) elements across the length of *var* genes and including the mates of reads that contained motifs made it possible to generate sequence islands at the initial stages of the assembly process. These islands were considered as points of initiation for further iterative extension. A combination of iterative scaffolding and extension was used to close gaps between seed regions by walking-in and out of the sequence islands. Assembly quality was affected by data quality, yield and the coverage of motifs used to initiate the process. The increase in motif space at the beginning is characteristic of the initial stages of the assembly process where new motifs are being added to the collection. On the other hand, at later iterations, the majority of the motif would already be in the collection resulting in a decrease in the rate of accumulation.

Conclusion 3: *Iterations provide the mechanism for a controlled extension of the var gene repertoire*

Extremely low and extremely high read coverage are potential reasons for poor quality assembly. An iterative extension approach provided a means for identifying reads that could be used for a gradual extension of seed contigs. As the number of iterations increases the number of motifs identified also increased. A limited number of iterations was required to attain motif saturation ($\sim$5 for culture-samples and $\sim$7-10 for 50 clinical samples). This observation has implications on the number of iterations required to gather motifs in order to assemble a given set of samples. Motifs generated from the 50 clinical samples are representative of different geographical regions. Conditions for terminating iterations are determined based on the assembly quality (i.e. repertoire completeness as measured by the count of *var*-contigs, contiguity of assembled

contigs as measured by N50 and largest contig sizes, and repertoire coverage as measured by sum of *var*-contigs). Once an optimal number of iterations is achieved, assembly quality will stop improving or begin to deteriorate signaling an exit condition from the iterations.

Conclusion 4: *This approach provides a way of reconstructing var genes from clinical samples and get a complete view of the var repertoire for the first time*

Assembly results of 50 clinical samples resulted in the largest collection of *var* genes so far. The total number of contigs (n=2,769) found with the DBLα tag (N50=5.5 kb; Largest=14 kb; sum of contigs = $\sim$15 Mb) represented over 92% of the expected $\sim$3,000 contigs (expecting 60 *var* genes per genome). Samples with below 30 *var*-contigs were associated with poor quality in the raw data. Conversely, samples with over 70 *var*-contigs were shown to have multiple infections by a visual inspection of MSP1 genes. Assembly test of mixed samples resulted in shorter contigs than in the case of individual assemblies. However, within the majority of cases that were visually inspected, the breakpoints corresponded with the regions of high read coverage as expected. These regions are known to cause breaks in assembly due to ambiguities that could not be resolved using standard sized libraries (as discussed in the previous chapter). Samples used for mixed assembly tests were not normalised for coverage in order to represent over-representation of some genotypes in natural populations.

In summary, this chapter presented an alternative assembly approach to effectively reconstruct the *var* gene family from short reads of second-generation sequencing platforms. Applications of the method to perform a targeted assembly of the family were demonstrated using culture-adapted and clinical samples of *P. falciparum*.

Chapter 4

Understanding mechanisms of *var* gene diversity using next generation sequencing

4.1 Introduction

The constant exposure of surface antigens to the host immune system requires parasites to actively generate and display new variants of *Pf*EMP1 in order to effectively evade host defense mechanisms. In addition, the repertoire of PFEMP1 sequences within the parasite population as a whole needs to be sufficiently diverse such that infection with one genotype does not induce an effective response to other genotypes that may later infect the same patient. Understanding the mechanisms that generate the high level of diversity in *var* genes is therefore a critical step in elucidating the forces that drive their evolution (Frank et al., 2008).

One of the mechanisms suggested as a potential source of *var* diversity is ectopic gene conversion, a non-reciprocal transfer of genetic material between two non-allelic regions of high sequence similarity (Frank et al., 2008; Freitas-Junior et al., 2000). Gene conversion is usually initiated by a double strand break, where homologous recombination is used as a repair mechanism (Chen et al., 2007; Holliday, 1964) (See Chapter 1 for details). Our understanding of

gene conversion in eukaryotes is primarily derived from studies in yeast which show the presence of such events during both the meiotic and mitotic stages of cell division (Gao et al., 2005; Hicks, 2010; Symington et al., 1991). Despite the potentially deleterious effects (Chen et al., 2001, 2007), gene conversion could be advantageous for evolution of gene families and potentially their functional diversification (Maizels, 2005; Nielsen, 2003; Santoyo and Romero, 2005). It is also used by a number of pathogens to generate antigenic variation (Al-Khedery and Allred, 2005; Brayton et al., 2002; Santoyo and Romero, 2005) and also eliminate diversity (Jackson et al., 2012). In *P. falciparum*, double strand break and repair may take place at various points of cell division during both the sexual and asexual stages, providing a platform for meiotic and mitotic recombination events. The haploid nature of the parasite for the most part of its life cycle makes it relatively easy to study rates and mechanisms of recombination.

One method to study how new *var* genes are created would be to monitor the changes that take place on *var* genes as the disease progresses in human patients. However, this is not practical due to the short time scale. A number of genetic cross experiments between a series of parasite clones have been conducted in Chimpanzies. These studies have provided mechanistic and functional insights into the biology of *P. falciparum* parasites (Freitas-Junior et al., 2000; Jiang et al., 2011; Samarakoon et al., 2011; Walliker et al., 1987). Nonetheless, *var* genes pose a great challenge as over 60% are located in subtelomeric regions, where the application of existing laboratory-based and informatics methods is limited.

Previous attempts to understand mechanisms of *var* diversity within a genetic cross were focused on the DBLα region (Taylor et al., 2000a). Presence or absence of hybridisation bands on a Southern blot probed with parental DBLα amplicons to parental and five progeny clones was used to identify recombinant *var* genes. The results revealed 24 non-parental bands in the five progeny, $\sim$10 times more than the expected 2-3 events. However, most array hybridisation based studies exclude markers in subtelomeric regions and hyper variable multigene families to avoid potentially spurious results.

Recently, the 454 sequencing platform was used to study two progeny from a genetic cross between HB3 and DD2 (Samarakoon et al., 2011). The study

used a sliding window approach where a window of 15 SNPs was considered to compute allele frequencies and assign genotypes to either parent. A total of 24 and 27 crossover events were reported for the two progeny. Although this showed an increase from a previously reported number of 20 and 23 events respectively, it was based on ∼30% of the genome as it was not possible to reliably call SNPs form the rest of the genome. Conversely, a total of 25 and 22 non-crossover events were reported for the two progeny. The results demonstrate the value of a sequencing based approach to obtain higher resolution recombination maps. However, *var* genes and other hyper-variable regions were excluded.

In this chapter, five progeny from the original cross between 3D7 and HB3 parental clones (Walliker et al., 1987) were sequenced using Illumina's GAII platform to evaluate the rate of genome wide recombination events and distinguish large scale crossovers from non-cross over gene conversion events. Here, I will explore applications of short read sequencing and the Illumina technology in understanding mechanisms of *var* gene diversity.

4.2 Methods

4.2.1 Sample preparation and sequencing

Five clonal samples of the first genetic cross in *P. falciparum* between 3D7 and HB3 isolates (Walliker et al., 1987) were obtained from Prof. Chris Newbold's laboratory in Oxford. According to Walliker and colleagues, a mixture of infective gametocytes from the two parental clones were fed to ∼1500 mosquitoes which were then used to infect a splenectomized chimpanzee through 500 mosquito bites and intravenous injection. Recombinant parasites were then isolated from the infected chimpanzee and either cloned or passed through Pyrimethamine treatment followed by cloning followed by genotyping to identify recombinant progeny. Five progeny of these cloned progeny and two parental samples were prepared for sequencing on the Illumina GAII platform according to the PCR-free paired-end protocol (Kozarewa et al., 2009) as described in Chapter 2.

4.2.2 Reference genomes

4.2.2.1 Whole genome reference genomes

Reference genomes for the two parental clones 3D7 (version 3) and HB3 were obtained from geneDB (`http://www.genedb.org`) and the Broad Institute (`http://www.broadinstitute.org`) respectively. The HB3 genome is still in thousands of contigs that are joined to create larger sequence fragments (also called Supercontigs). Supercontigs are used to represent the state of a genome assembly that is highly fragmented and potentially incomplete. In order to obtain a better representation for HB3, supercontigs were ordered against the 3D7 genome using ABACAS (Assefa et al., 2009) as described in Chapter 2.

Draft genomes are particularly prone to inaccuracies that could be introduced at any point during the sequencing, assembly or scaffolding stages of the genome finishing process. Iterative Correction Of Reference Nucleotides (ICORN) (Otto et al., 2010a) was developed by Thomas Otto in our group to make use of the high sequence coverage of second generation sequencing platforms in order to detect and correct errors in a reference genome. The process involves iteratively aligning short reads to the reference genome and inspecting aligned reads with mismatches to detect high quality discrepancies. If the majority of the reads supports such discrepancies, a correction step is applied to replace the reference nucleotide by the base called from the reads. Subsequent mapping steps are used for further quality control based on the coverage of mapped reads. ICORN was used to generate a new reference genome for HB3 by iteratively correcting errors using Illumina reads of the HB3 isolate. Illumina reads (76 bp; paired-end reads; coverage $\sim$100X) were obtained from Prof. Dominic Kwiatkowski's group at the Sanger Institute. A combined reference genome of the two parents was generated by concatenating sequences from the 3D7 genome and the new HB3 reference.

4.2.2.2 *Var* gene sequences of 3D7 and HB3 genomes

Var genes of 3D7 were obtained from the latest annotation of version 3 of the genome (`http://www.genedb.org`). A total of 93 genes were obtained

including pseudo-genes and truncated Exon 2 genes. In addition to the *var* genes found from the annotation, HB3 genes were scanned for known conserved *var* motifs that were obtained from three lab adapted *P. falciparum* samples 3D7, HB3 and IT as described in Chapter 3. A multi-FASTA file of a combined *var* reference was generated by concatenating *var* genes from the two parents (93 from 3D7 and 97 from HB3).

4.2.3 Alignment and processing of sequence data

Raw reads were aligned to individual parental genomes and combined references (combined reference genome and combined *var* reference) using SMALT (*-i 500 -r 10 -x -k 13 -s 6*). SMALT is a fast and memory-efficient alignment tool that uses a traditional *k-mer*-based hashing method to index the reference genome. Reads that have a match to indices will then be aligned using a Smith-Waterman algorithm to ensure a highly sensitive output. In our experience, SMALT produces better alignment for *P. falciparum* genomes where the biased A+T content is a major challenge for short read aligners. Although the underlying principle is similar to its predecessor SSAHA (Ning et al., 2001), SMALT is more accurate, faster and user friendly. Special emphasis was given to the parameter *x* in order to map individual reads of a read pair to their best positions irrespective of the insert size between them. This property is unique to SMALT and important in identifying recombination breakpoints. The default option and other aligners will force read pairs to be aligned within the read-pair constraint leading to incorect alignment especially around breakpoints. Raw alignment results were stored in the Sequence Alignment/Map (SAM/BAM) format and processed using Samtools (version 0.1.18) (Li et al., 2009a).

Read pairs that did not align to a single position were excluded due to the difficulty of reliably determining where they should be placed. The Picard Suite from the Broad Institute (`http://www.broadinstitute.org`) was used to mark duplicate reads in the BAM file that were subsequently excluded from further analysis. Samtool's *mpileup* command and BCF tools (`http://www.vcftools.sourceforge.net`) were used for SNP calling (*samtools mpileup*). SNPs were then filtered based on quality (Q$\geq$ 30), number

of reads calling the SNP on each strand (read-depth$\geq$ 2) and position on the genome (eg. SNPs in low complexity regions are less reliable).

4.2.4 Genome-wide scan for recombination breakpoints

4.2.4.1 Detecting homologous recombination using comparative SNP maps

In order to identity regions of the genome that are involved in a reciprocal exchange (crossing-over), reads from the five progeny and the two parental samples were independently aligned to 3D7 and HB3 genomes using SMALT as described in the previous section. High quality SNPs (Q $\geq$ 30) generated from uniquely aligned reads were used to construct comparative SNP maps for each chromosome.

First, parental chromosomes were aligned to each other using BLAST (*blastn -F F -m 8*). The alignment output was then visualized in the Artemis Comparision Tool (ACT) (Carver et al., 2005). SNP files of the progeny and parental genomes were then uploaded to ACT using the Bamview utility (Carver et al., 2010). Recombination breakpoints and regions of homologous exchange were identified by visually inspecting these comparative maps for each chromosome. Regions of high and low SNP density were used to assign segments of chromosomes to either the 3D7 or HB3 genotypes. Accumulation of SNPs on one parent was expected to result in lack of SNPs on the other. Although this approach provides a simple and effective way of detecting regions of interest, it has major limitations in highly polymorphic regions, due to the difficulty of reliably calling SNPs. Differentiating real signals from background noise is therefore a major challenge in relying on visual inspection.

4.2.4.2 Using sliding windows to detect crossover and non-crossover recombination

In order to systematically identify breakpoints at a higher resolution and account for low complexity regions, a sliding window approach was developed. In addition to SNPs used in the previous section, supplementary evidence drawn from paired-end coverage of uniquely mapping reads was considered

prior to assigning genotypes.

Sliding windows of various sizes were analysed to assign regions of the genome to either 3D7 or HB3 based on the number of SNPs and proportion of each window covered by read-pairs. The following sections outline the approaches used to improve the signal to noise ratio while detecting recombination breakpoints.

Overlapping vs non-overlapping windows

Initially, non-overlapping sliding windows were considered to count SNPs and compute average paired-end coverage values. However, non-overlapping windows were found to underestimate breakpoints as described below (Figure 4.1). Assuming a minimum number of three SNPs (shown as red arrows), windows **a** and **b** (Figure 4.1A) will not be considered as they have below the expected number (two and one respectively). Conversely, window **e** of the overlapping panel is able to capture all three SNPs (Figure 4.1B).

Improving signal-to-noise ratio for smaller window sizes

In order to improve the signal-to-noise ratio for small windows, it was necessary to use a number of consecutive windows (n) instead of considering single windows. The minimum number of such windows was determined by the fragment size of the sequencing library (F) and the chosen window size (w). In order to capture small gene conversion events using paired reads, the length of the region covered by n windows (L) should not be greater than the fragment size, F. Assuming the length of the slide (s) to be half of the window size, the minimum number of windows could be obtained as follows:

$$s = \frac{w}{2} \tag{4.1}$$

$$\frac{w}{2}(n+1) = L \leq F \tag{4.2}$$

$$n \leq (\frac{2F}{w} - 1) \tag{4.3}$$

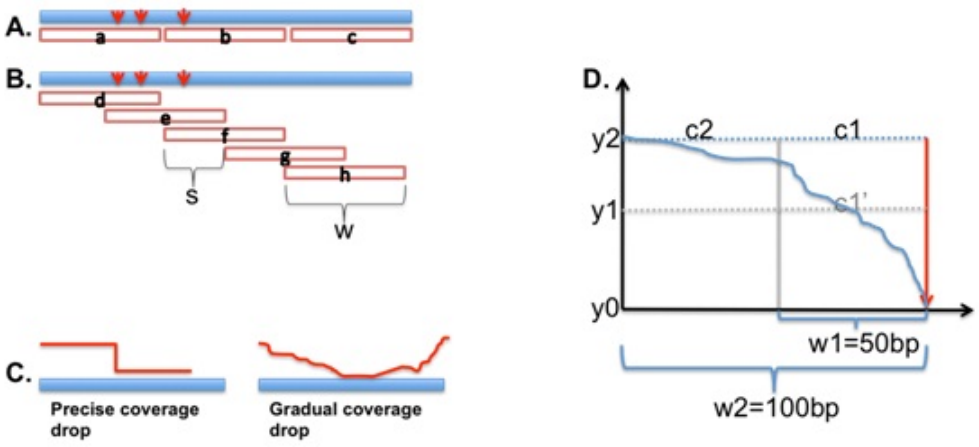


Figure 4.1: Breakpoint detection using SNPs and paired end coverage. **A)**. If a minimum of three SNPs were required to determine breakpoints, non-overlapping windows **a**, **b** and **c** are likely to miss genuine regions of recombination and gene conversion. **B)**. A sliding window approach overcomes limitations of non-overlapping windows. The window e is able to capture a region with three SNPs (red arrows) that would otherwise be missed by windows a and b in the non-overlapping approach. **w** represents the size of the window and **s** represent the sliding length. **C)**. Genome-wide scans for coverage drops are used to detect signal of recombination. However, it was necessary to distinguish between genuine coverage drops (precise coverage drops) and those with a gradual decrease in coverage. **D)**. The ratio of slopes over two regions length 50 bp and 100 bp was used to differentiate between precise and gradual coverage drops. (See below for details).

Gradient filter on coverage drops

While a decrease in coverage below the threshold could be an indication of a break-point, it could also be a result of poor-mapping due to the low complexity of the region in consideration. Distinguishing between precise and gradual coverage drops (Figure 4.1C) in a systematic way was thus adopted to minimise false positive calls. A filter based on differences in gradient (slope of the coverage plot) was developed as described below.

Two windows of lengths w_1 (50 bp) and w_2 (100 bp) were arbitrarily defined to the left of a given drop-point (Shown by the red arrows, Figure 4.1D). The x-axis represents position on the genome, while the y-axis represents paired-read coverage. The average mate pair coverage for windows w_1 and w_2 (i.e. c_1 and c_2); are represented by y_1 and y_2 on the y-axis of Figure 4.1D.

The slopes over windows w_1 and w_2 are defined as:

$$slope_1 = \frac{y_0 - y_1}{w_1} \tag{4.4}$$

$$slope_2 = \frac{y_0 - y_2}{w_2} \tag{4.5}$$

The ratio of the two slopes, r_S is therefore:

$$\begin{aligned} r_s &= \frac{slope_1}{slope_2} \\ &= \frac{w_2}{w_1} * \frac{y_1}{y_2} \\ &= 2(\frac{y_1}{y_2}) \end{aligned} \tag{4.6}$$

To determine whether a drop in coverage is gradual or precise, three possible values are considered for y_1 and y_2

1. $y_1 = y_2$
    - This represents the ideal case and shows no gradual drop in coverage (Figure 4.1C).

- Ratio of slopes (r_S) = 2

2. $y_1 > y_2$
   - The second scenario represents cases where there is an increase in coverage before the breakpoint.
   - Such cases are also acceptable.
   - Ratio of slopes (r_S) > 2

3. $y_1 < y_2$
   - There is a gradual drop in coverage.
   - Ratio of slopes (r_S) < 2
   - These events will be ignored.

Defining ambiguous regions and further quality filters

Ambiguous regions were defined as windows where both SNP count and coverage are below the minimum cutoff of three and five respectively. Such regions were grouped as non-unique (NU) or ambiguous and excluded from further analyses. In addition, polymorphic and low quality SNPs (Q<30), SNPs that have a strand bias (i.e. SNPs found on only one strand) and those common to all five progeny were excluded.

Final choice of window size

After evaluating the number of breakpoints at various window sizes (the results are shown in Figure 4.4), a window size of 20 kb was chosen to capture both homologous and non-homologous recombination breakpoints. Average Paired End Coverage (PEC), uniqueness (determined using a frequency count of *k-mer*s, k=30), and number of SNPs were computed over 20 kb windows (sliding by 10 kb). Windows that have over three SNPs were assigned to the HB3 parent and the boundaries were identified as break-points. Average PEC of above 5

was used to decide whether windows with less than three SNPs belong to the 3D7 parent or grouped as ambiguous (non-unique).

4.2.4.3 Detection and filtering of breakpoints in *var* genes

Initially, the iterative *de novo* assembly approach developed in Chapter 3 was applied with the aim of obtaining full length *var* genes for the five progeny. However, it was not possible to generate valid contigs due to the poor quality of raw data and the short read length (54 bp). The following sections describe alternative methods used to detect breakpoints in *var* genes.

Visual inspection

BAMview (Carver et al., 2010) was used to visualise coverage of uniquely and perfectly mapping read-pairs over parental *var* genes. Observed coverage plots were compared to expected patterns of recombination (Figure 4.2).

Mapping based detection of breakpoints

Raw reads were aligned to the combined *var* reference using SMALT and BWA to obtain a mapping output that contains uniquely and non-uniquely aligning reads respectively. The BAM files from both aligners were further processed to look for read pairs that map to different *var* genes on the same or different parents. However, the BWA output was used for the final analysis as the strict alignment criteria used in the SMALT mapping was found to exclude reads that align to homologous regions. Although the BWA mapping ensured inclusion of reads that are genuinely located on the *var* genes, the short read length (54 bp) makes it difficult to avoid spurious alignment of reads. Genes were therefore identified as potentially recombinant if they were bridged by a minimum of 50 read pairs. Initially, this cutoff may seem too high, however, it was chosen to account for the effect of spurious alignments. *De novo* assembly of reads that aligned to the genes bridged by paired-reads was then used to reconstruct *var* genes of the progeny and validate events (velvet v.1.2.03, (Zerbino and Birney, 2008)). Finally, the new contigs were ordered using ABACAS (Assefa et al., 2009) against parental genes and manually checked for formation of a valid *var*

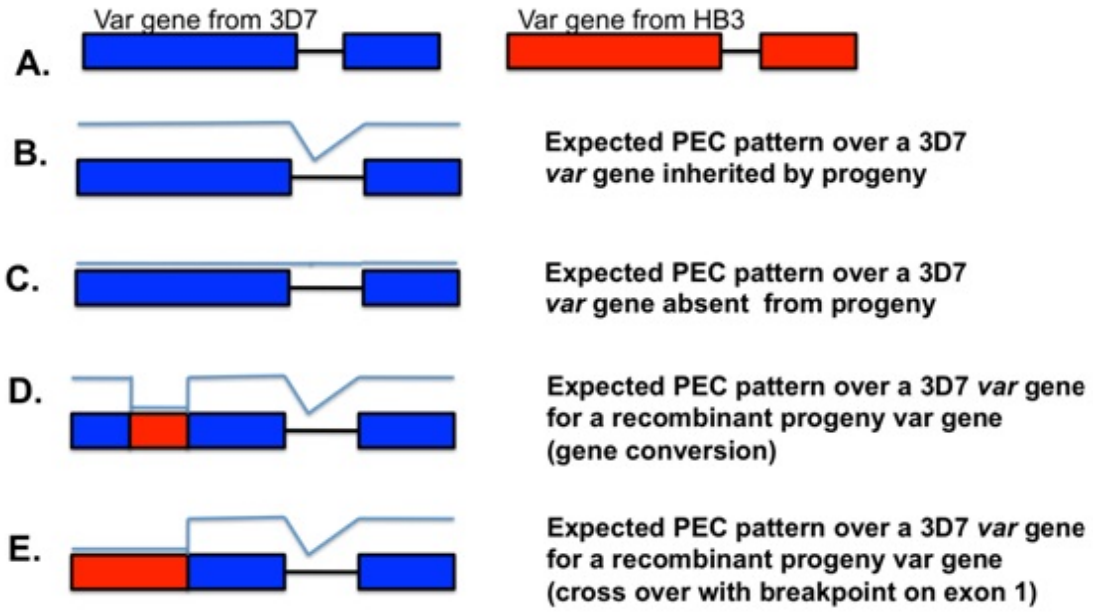


Figure 4.2: Patterns of Paired end coverage (PEC) over a 3D7 *var* gene with/without recombination. **A).** Parental *var* genes from the 3D7 and HB3 are represented in blue and red respectively. **B).** PEC over a 3D7 *var* gene suggests that the progeny had inherited the gene from the 3D7 parent. **C).** Conversely, lack of coverage may suggest that the 3D7 *var* gene was not inherited by the progeny. **D** and **E** show potential signals of recombination in the form of gene conversion and crossing over respectively. The drop in coverage on the first exon is expected to extend to the left for kb or Mb of sequence to establish a crossing-over event.

molecule.

4.3 Results

4.3.1 Sequence data and mapping to reference genomes

Five progeny obtained from the first genetic cross in *P. falciparum* between 3D7 and HB3 parental clones were sequenced using the Illumina GAII platform to a depth of 60- to 80-fold (paired reads of length of 54 bp; expected insert size of 200 bp). Raw reads were aligned to the two parental genomes and a combined reference genome. The 3D7 genome is the reference isolate (Gardner et al., 2002) with near-perfect base accuracy while HB3 is still highly fragmented and incomplete. A total of 93 and 47 genes were annotated as *var* (including pseudogenes and truncated exons) in 3D7 and HB3 respectively. Additional 50 genes were identified for HB3 using conserved *var* motifs resulting in 97 *var* genes for HB3.

During the initial whole genome analysis, reads that aligned to multiple positions were excluded, resulting in 34 to 60% of paired-reads mapping to the 3D7 parent (Table 4.1). The number of reads aligned to the HB3 reference was lower (26 to 52%) compared to 3D7 reflecting the poorer state of the HB3 reference rather than potential allelic bias. On the other hand, the reduction in mapping against the combined reference is primarily attributed to the high sequence similarity in the central regions of the parental genomes.

	X1	**X2**	**X3**	**X4**	**X5**
Raw reads ($x10^6$)	33.4	34.7	34.1	26.3	29.2
% Mapped	[52,46,26]*	39,29,22	47,38,21	68,58,29	61,52,30
% Mapped in pairs	44,39,23	34,26,20	42,34,19	60,52,27	52,46,27
Read length	54	54	54	54	54
Original ID	X33	XP8	X10	X4	XP2

*[3D7,HB3,Combined reference]

Table 4.1: Raw reads and mapping statistics to 3D7, HB3 and combined reference genomes. X1, X2, X3, X4 and X5 represent the five progeny.

4.3.2 Detecting genome wide crossover events

4.3.2.1 Visual identification

Initially, high quality SNPs (excluding SNPs with a quality score of below 30 and SNPs that were identified as heterozygous) generated by aligning progeny reads to the 3D7 parent were visualised using bamview. Figure 4.3 shows SNP views of the five progeny for chromosomes 1-5.

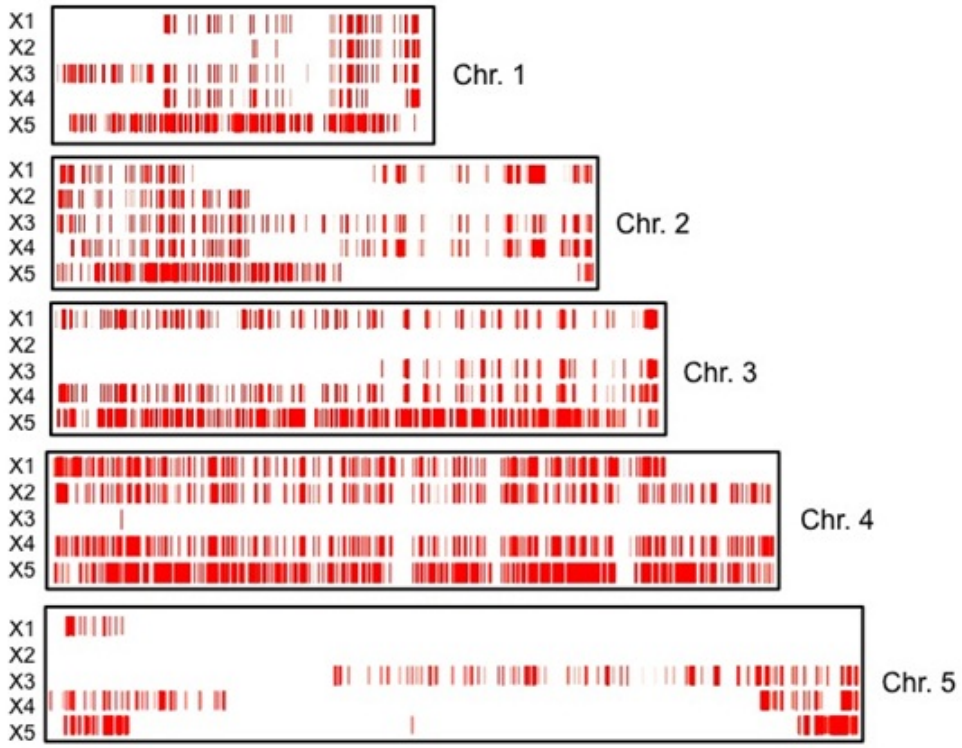


Figure 4.3: Visualising SNPs of the five progeny using the 3D7 parent as a reference. Here, SNPs are shown for the chromosomes 1 to 5. The quality of each SNP is reflected by the intensity of the red bars. The white horizontal blocks represent regions where no SNPs are called.

Although the red bars represented regions of chromosomes that were inherited from the HB3 parent, the white regions could either be inherited from the 3D7 parent or were ambiguous (i.e. non unique regions where reads could not be reliably aligned). In order to address this limitation, additional informa-

tion obtained from the reciprocal alignment (i.e. aligning reads from the five progeny to the HB3 reference) was used. In addition, Illumina reads of the 3D7 and HB3 clones were aligned to the 3D7 and HB3 reference genomes and used as control samples.

Comparative chromosome maps were thus constructed using the five progeny and both parental genomes (see Figure 4.4 for a map of Chromosome 3). A visual analysis of such maps at a chromosome level revealed regions of high (red bars) and low (white regions) SNP density. In most cases, regions of high SNP density on the 3D7 parent were found to have a lower density against the HB3 parent due to the reciprocal nature of homologous recombination during a crossover. SNP dense areas of a progeny against the 3D7 parent could be easily assigned to HB3 and vice versa. A total of 15 to 21 large-scale crossing over events were detected in each progeny with an average of $\sim$1 event per chromosome (Table 4.2).

Chr	**X1**	**X2**	**X3**	**X4**	**X5**
1	1	1	0	1	0
2	2	1	0	2	2
3	0	0	1	0	0
4	1	0	0	0	0
5	1	0	1	2	2
6	3	0	1	1	0
7	1	2	1	1	0
8	2	0	1	2	0
9	0	0	1	1	3
10	3	3	1	2	0
11	2	2	2	3	3
12	2	2	0	0	1
13	2	0	3	1	2
14	1	4	3	2	5
Total	**21**	**15**	**15**	**18**	**18**

Table 4.2: Number of cross-over events per chromosome. Large scale cross over events were detected using a visual inspection of comparative chromosome maps as shown in Figure 4.4.

Although initially the comparative chromosome maps as shown in Figure 4.4 may appear to clearly identify breakpoints, the white regions are still highly

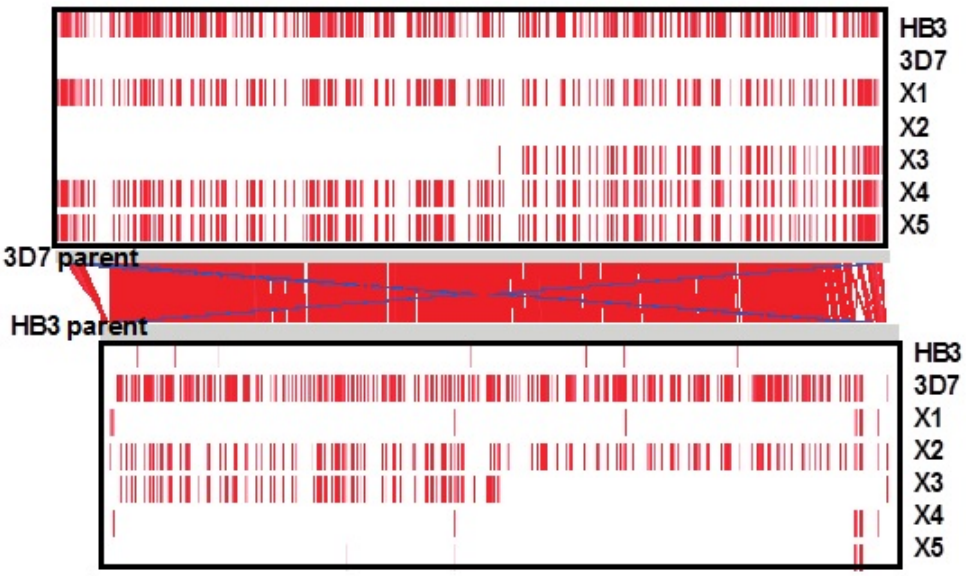


Figure 4.4: Comparative SNP-map of Chromosome 3: The top and bottom panels show SNPs called from progeny reads against Chromosome 3 of 3D7 and HB3 genomes respectively. The middle panel shows an Artemis Comparison Tool (ACT) view of synteny blocks (red lines) between the two chromosomes (generated by aligning the two chromosomes using *blastn -F F -m 8*). Copies of chromosome 3 in progeny X1, X4 and X5 are inherited from HB3; in progeny X2 it is inherited from 3D7; and in progeny X3, chromosome 3 is a result of a cross-over event. Although this figure looks cleaner compared to Figure 4.3 and the informatics approach described in the next section (Figure 4.6), assigning genotypes to the 'white' regions remains to be a challenge as they could either be from the 3D7 parent or non-unique regions.

ambiguous as they represent either the 3D7 genotype or non-unique regions. An informatic approach was therefore developed to merge the evidence from SNPs, paired-read coverage and reciprocal chromosome maps in order to detect large-scale crossover events as well as smaller gene conversion events.

4.3.2.2 Informatic identification

Genome-wide breakpoints were identified using a combination of high quality SNPs and paired-end coverage over overlapping sliding windows on the 3D7 genome. One aim of this experiment was to see if the shape of the distribution could be used to differentiate false positives from real events. The number of breakpoints detected was a function of window size and dropped from $\sim$500 to $\sim$14 per progeny with an increase in window size from 10 kb to 300 kb (Figure 4.5). The right tailed distribution was indicative of small scale gene conversion events (left tail) and the large scale crossover events (right tail). Although smaller window sizes ($<$ 20 kb) may provide a better resolution on the number and position of breakpoints, it was not possible to reliably assign genotypes mainly due to the poor data quality.

A window size of 20 kb was thus chosen to visualise patterns of recombination and gene conversion events across the 14 chromosomes (Figure 4.6)

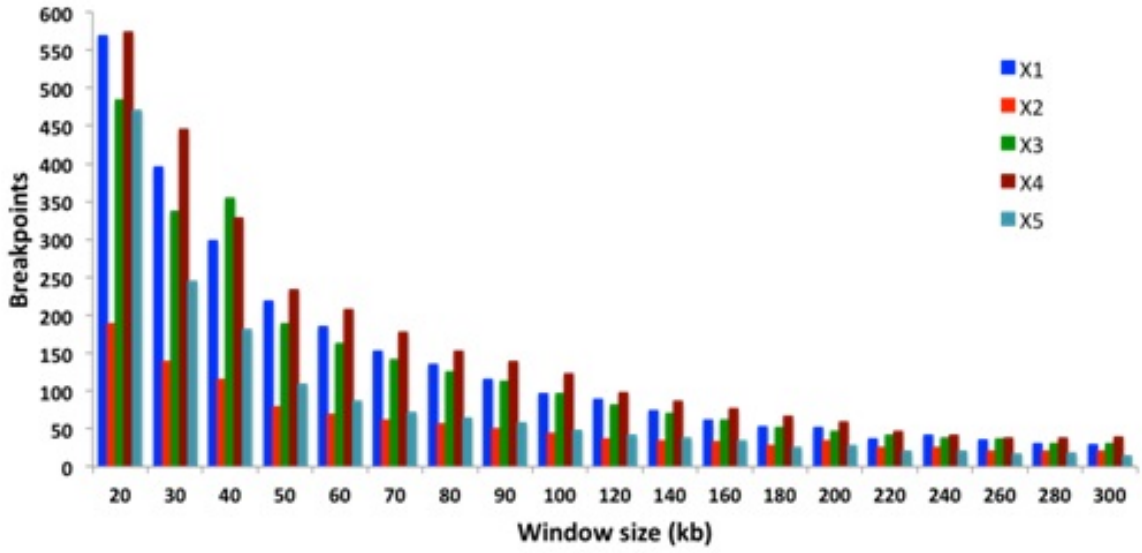


Figure 4.5: Genome wide breakpoints per window size. Breakpoints detected at larger window sizes were comparable with results of the visual identification.

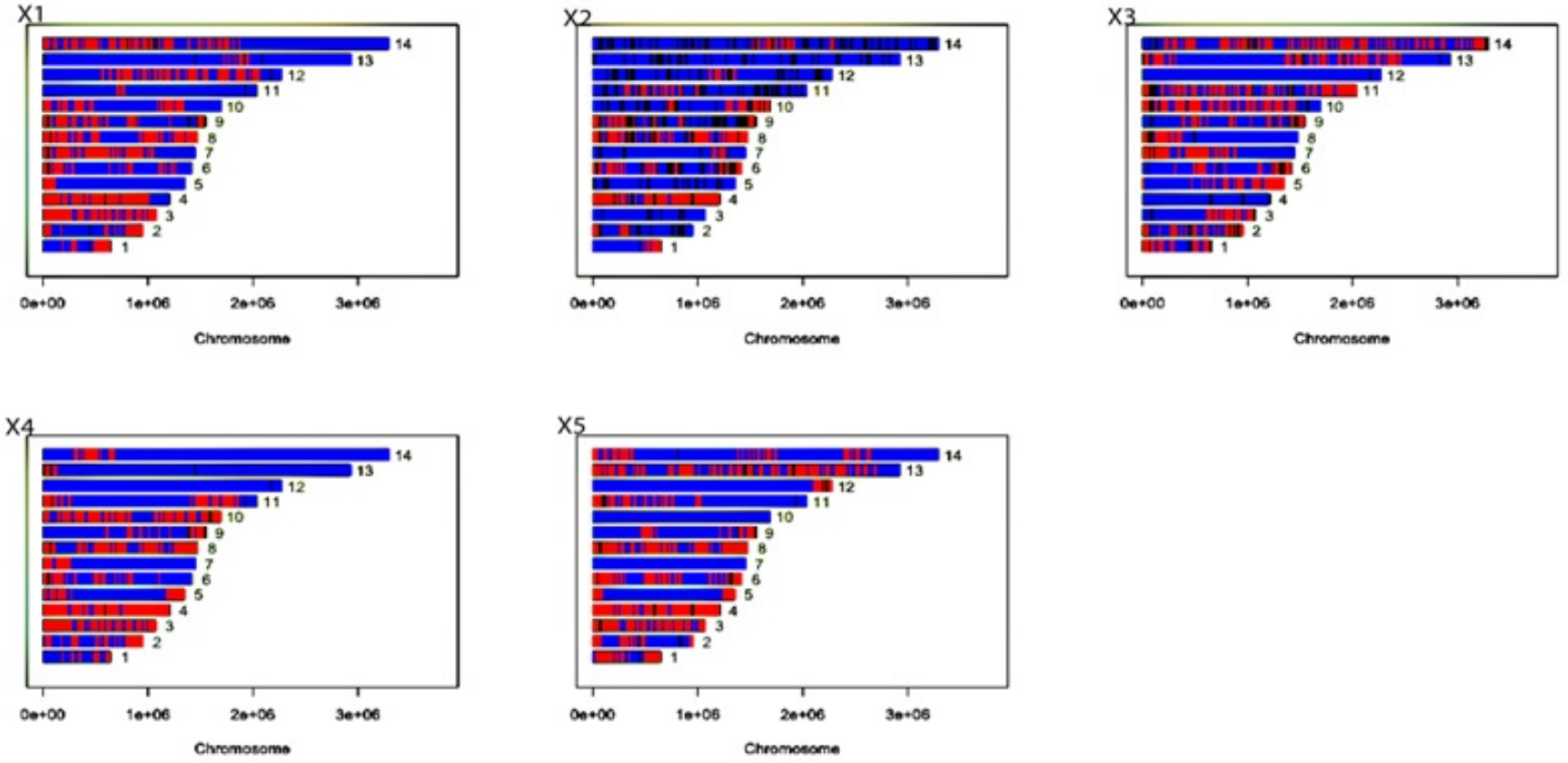


Figure 4.6: Breakpoints per chromosome at a window size of 20 kb (sliding by 10 kb) for the five progeny X1 to X5: Blue blocks show regions of chromosomes that were inherited from the 3D7 parent; red blocks show regions that were inherited from the HB3 parent; black regions represent non-unique (ambiguous) regions. Smaller window sizes provided a better resolution to identify large scale recombination events as well as potential non-crossover events.

Breakpoints detected at higher window sizes of ∼300 kb (Figure 4.6) were comparable to crossover events identified using a visual inspection of chromosome maps (Table 4.2). However, the numbers were not identical as the informatics identification included additional filtering and was done using sliding windows. Non-unique regions were also separately identified making the informatics approach more reliable instead of immediately assigning regions of low SNP density to the 3D7 parent. A smaller window size of 20 kb (sliding by 10 kb) was used to gain a better resolution of both crossover and putative gene conversion events (Figure 4.6).

Fewer breakpoints were observed in $X2(183)$ compared to other progeny $(X1 = 563, X3 = 478, X4 = 568, X5 = 465)$ due to lower read mapping (Table 4.1), which resulted in a larger proportion of non-unique regions. Further analysis of breakpoints at a window size of 20 kb revealed that most of the breakpoints (∼51%) were unique to one progeny compared to breakpoints shared by two (∼28%), three (∼18%) and four (∼3%) progeny. After excluding 7 breakpoints that were common to all, a total of 1,304 breakpoints were identified of which 270 were found in three or four of the five progeny.

In order to detect biases in inheritance, the proportion of the progeny genome assigned to each parent was analysed (Figure 4.7). The average inheritance over the progeny (Figure 4.7F and G) shows that chromosomes 5, 7 and 12 were predominantly inherited from 3D7 whereas the majority of chromosomes 2, 6, 8 and 9 came from HB3.

4.3.3 Signature of recombination in *var* genes

This section focuses on identifying crossovers and gene conversion events in *var* genes. As described previously, approaches based on read coverage and SNPs have a limited application in highly polymorphic regions such as *var* genes. Here, both a visual inspection and informatics approaches of detecting breakpoints are explored.

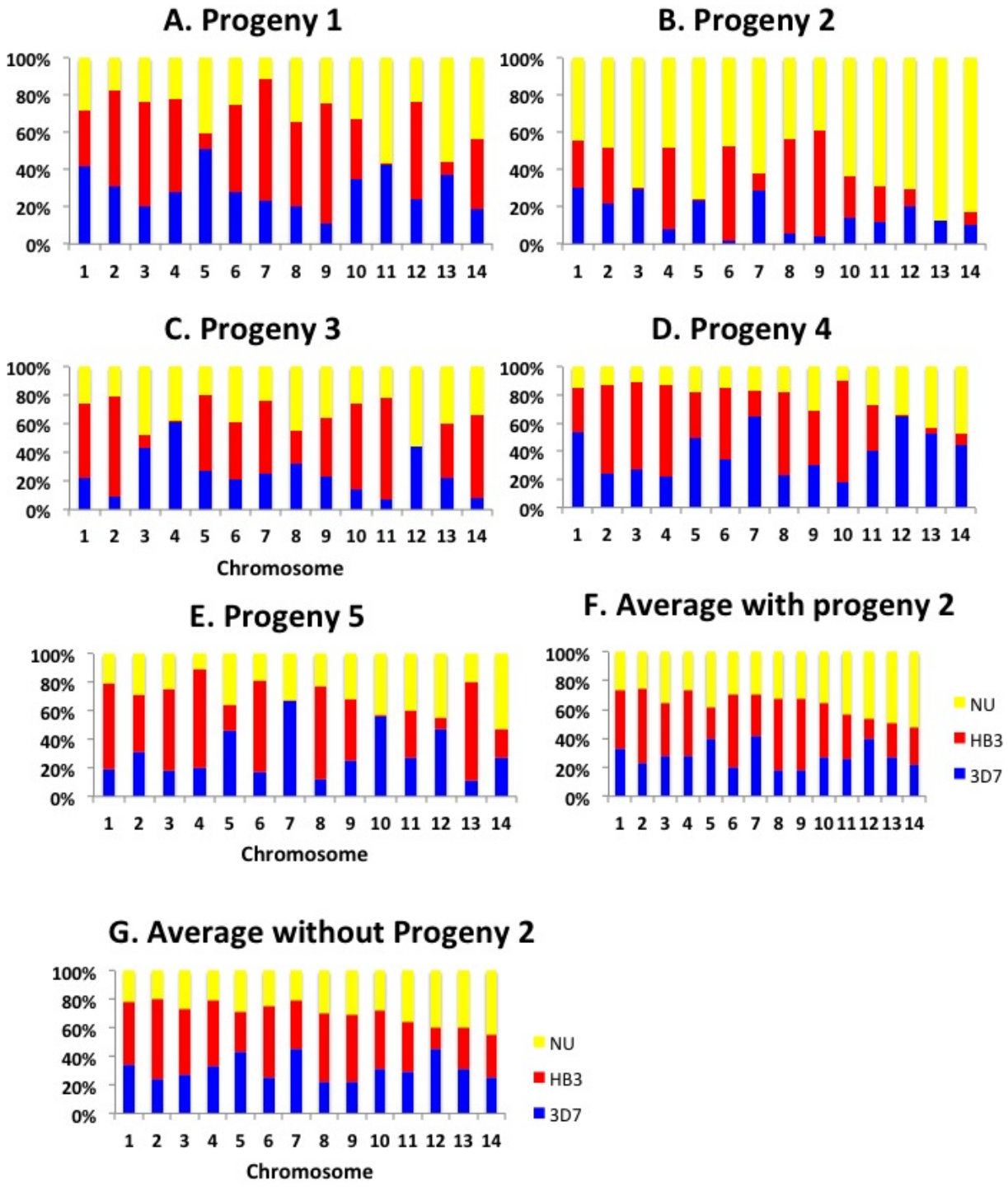


Figure 4.7: Per-chromosome inheritance patterns for each progeny and average inheritance profiles with / without progeny 2 to see the effect of poor quality data on the results. Blue and red blocks represent the proportion of chromosomes inherited from the 3D7 and HB3 parents respectively. Yellow bars represent regions of chromosomes that could not be reliably assigned to either parent.

4.3.3.1 Visual inspection of paired-read coverage

Firstly, BAM files of the five progeny were visualised and ambiguous patterns such as those characterised by lack of uniqueness and loss of coverage in all five progeny were excluded. Paired end coverage (PEC) plots for each *var* gene of the 3D7 genome were compared to expected patterns of recombination and gene conversion as shown in Figure 4.2. Regions of a gene that exhibited a clear loss of PEC were identified as breakpoints. Ideally, loss of PEC around a breakpoint is accompanied by an increased proportion of orphaned reads i.e. reads whose mates map to a different gene. Mates of orphaned reads were then investigated to find other genes that were involved in forming the recombinant in the progeny (Figure 4.8).

Although detecting regions of higher orphaned read coverage could be used to identify potential signature of recombination (Figure 4.8A), flanking regions of breakpoints often have higher levels of homology with donor and acceptor regions. Such homology, however essential to initiate recombination, may result in reads that align to multiple locations. It was therefore not possible to use this test in a genome-wide context. A closer look at read coverage over *var* genes in the 3D7 parent revealed that 24-50% of the genes were inherited from the 3D7 parent (Figure 4.9, Table 4.3).

A total of four genes in two progeny fulfilled both criterion i.e. loss of PEC and presence of unique orphaned reads that span two different genes (Table 4.3). The four genes with non-parental patterns of coverage were located on opposite ends of chromosomes one (PFA0005w, PFA0765c) and two (PFB1055c, PFB0010w) (Figure 4.10) in close proximity to telomeric-associated repetitive elements (TAREs). All four genes were of the group Type A, as they are transcribed away from the telomeres.

A detailed investigation of reads and their mates that mapped to these genes revealed that sequence blocks from two parental *var* genes were ordered to generate new genes in the progeny (Figures 4.11 and 4.12).

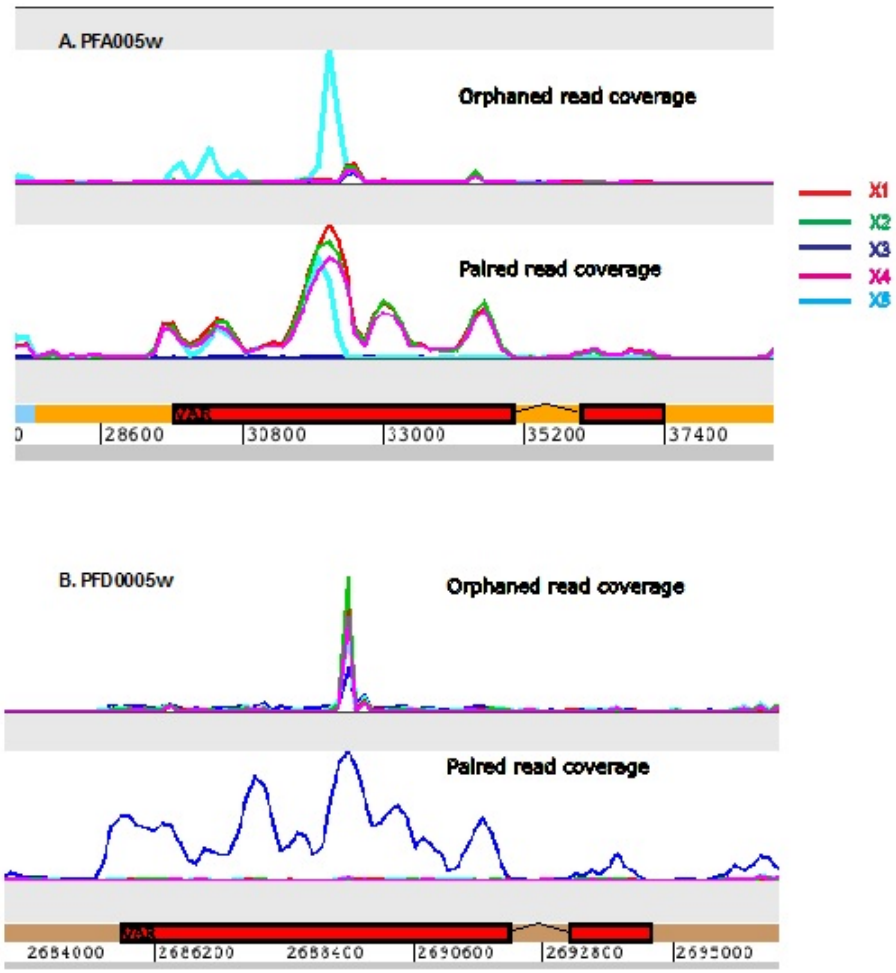


Figure 4.8: Paired and Orphaned read coverage over 3D7 *var* genes. **A)** Coverage pattern on progeny 5 shows a drop in PEC accompanied by an increase in orphaned read coverage indicating signatures of recombination for a gene located on Chromosome 1 of 3D7 (PFA005w). **B).** Coverage over a chromosome 4 *var* gene, PFD0005. The paired end coverage plot for progeny 3 (X3) indicated that PFD0005w was inherited from the 3D7 parent. Lack of coverage from other progeny may suggest that the gene was inherited from HB3 for progeny X1, X2, X4 and X5. No evidence of recombination is shown from the coverage plots.

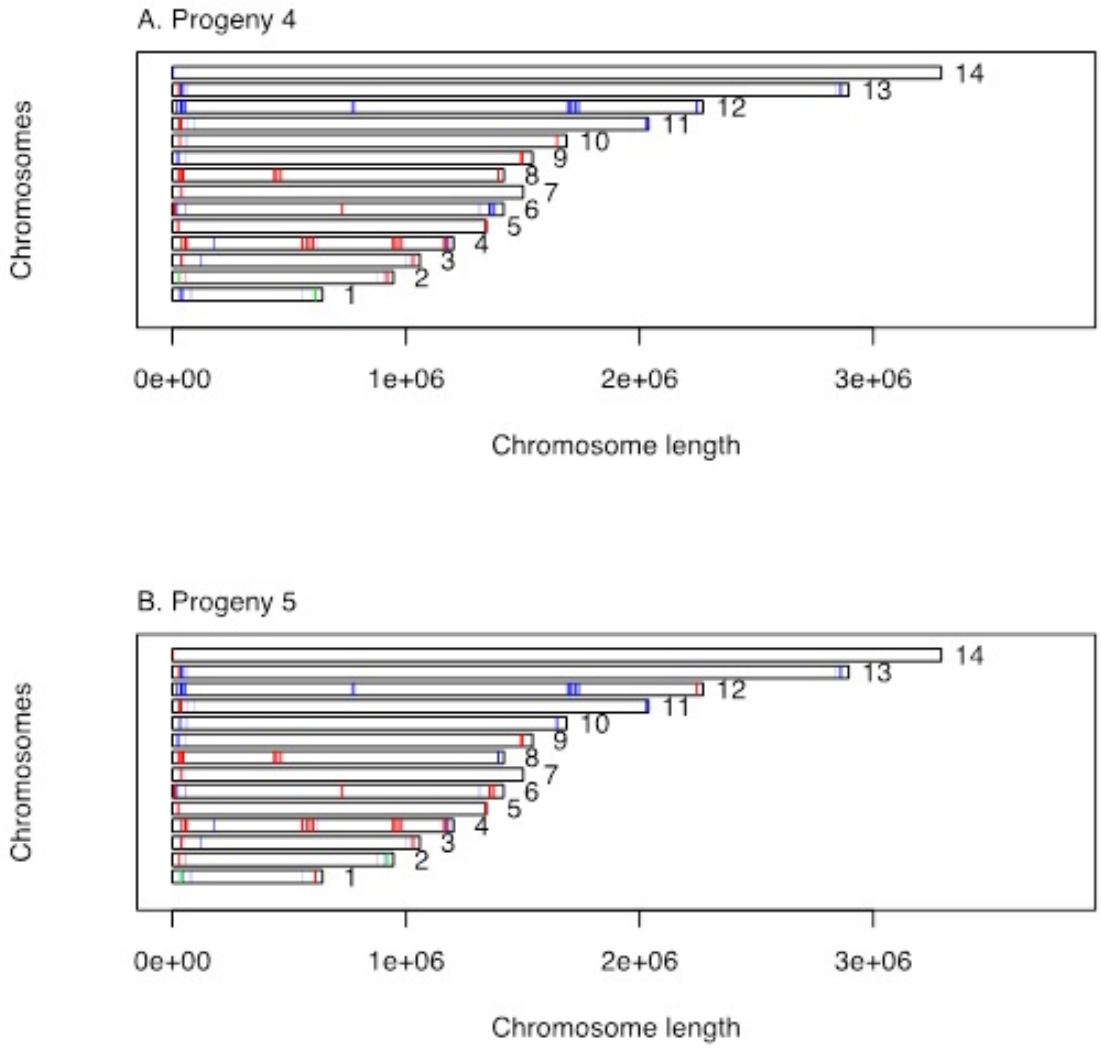


Figure 4.9: Chromosome view of *var* genes for progeny 4 and 5. The majority of *var* genes were inherited from either the 3D7 (shown in blue) or HB3 (shown in red) parents. The *var* genes on chromosomes 1 and 2 were involved in a non-homologous recombination (gene conversion) as indicated by the green bars. The results are consistent with the crossover data shown in Figure 4.6.

Var gene ID	chr.	X1	X2	X3	X4	X5
PFA0005w	1	3D7	3D7	HB3	3D7	**GC**
PFA0765c	1	HB3	HB3	HB3	**GC**	HB3
PFB0010w	2	HB3	HB3	HB3	**GC**	HB3
PFB0020c	2	HB3	HB3	HB3	HB3	HB3
PFB0045c	2	HB3	HB3	HB3	HB3	HB3
PFB0974c	2	HB3	3D7	HB3	HB3	3D7
PFB0975c	2	HB3	HB3	HB3	HB3	HB3
PFB1025c	2	HB3	HB3	HB3	HB3	HB3
PFB1045w	2	HB3	3D7	HB3	HB3	3D7
PFB1055c	2	HB3	3D7	HB3	HB3	**GC**
PFC0005w	3	HB3	3D7	HB3	HB3	HB3
PFC1120c	3	HB3	HB3	HB3	HB3	HB3
PFD0005w	4	HB3	HB3	3D7	HB3	HB3
PFD0020c	4	HB3	HB3	3D7	HB3	HB3
PFD0615c	4	HB3	HB3	3D7	HB3	HB3
PFD0625c	4	HB3	HB3	3D7	HB3	HB3
PFD0630c	4	HB3	HB3	3D7	HB3	HB3
PFD0635c	4	HB3	HB3	3D7	HB3	HB3
PFD0995c	4	HB3	HB3	3D7	HB3	HB3
PFD1000c	4	HB3	HB3	HB3	HB3	HB3
PFD1005c	4	HB3	HB3	3D7	HB3	HB3
PFD1015c	4	HB3	HB3	3D7	HB3	HB3
PFD1235w	4	HB3	HB3	HB3	HB3	HB3
PFD1254c	4	3D7	HB3	3D7	HB3	HB3
PFE0005w	5	HB3	3D7	3D7	HB3	HB3
PFE1640w	5	3D7	3D7	HB3	HB3	HB3
PFF0010w	6	HB3	HB3	3D7	HB3	HB3
PFF0030c	6	HB3	HB3	3D7	HB3	HB3
PFF0845c	6	HB3	HB3	HB3	HB3	HB3
PFF1580c	6	3D7	HB3	HB3	3D7	HB3
PFF1595c	6	3D7	HB3	HB3	3D7	HB3
MAL7P1.212	7	HB3	HB3	HB3	HB3	HB3
PF07_0048	7	HB3	3D7	HB3	3D7	3D7
PF07_0049	7	HB3	3D7	HB3	3D7	3D7
MAL7P1.50	7	HB3	3D7	HB3	3D7	3D7
PF07_0050	7	HB3	3D7	HB3	3D7	3D7
PF07_0051	7	HB3	3D7	HB3	3D7	3D7
MAL7P1.55	7	HB3	3D7	HB3	3D7	3D7
MAL7P1.56	7	HB3	3D7	HB3	3D7	3D7
MAL7P1.187	7	3D7	3D7	3D7	3D7	3D7
PF08_0142	8	HB3	HB3	3D7	HB3	HB3
PF08_0141	8	HB3	HB3	3D7	HB3	HB3
PF08_0140	8	HB3	HB3	3D7	HB3	HB3
PF08_0107	8	HB3	HB3	3D7	HB3	HB3
PF08_0106	8	HB3	HB3	3D7	HB3	HB3
PF08_0103	8	HB3	HB3	3D7	HB3	HB3
MAL8P1.207	8	HB3	HB3	HB3	HB3	HB3
MAL8P1.220	8	HB3	3D7	HB3	HB3	3D7
PFI0005w	9	HB3	HB3	3D7	3D7	3D7
PFI1820w	9	HB3	HB3	HB3	HB3	HB3
PFI1830c	9	HB3	HB3	HB3	HB3	HB3
PF10_0001	10	HB3	3D7	HB3	HB3	3D7
PF10_0406	10	3D7	HB3	3D7	HB3	3D7
PF11_0007	11	3D7	3D7	HB3	HB3	HB3
PF11_0008	11	3D7	3D7	HB3	HB3	HB3
PF11_0521	11	3D7	3D7	HB3	3D7	3D7
PFL0005w	12	3D7	3D7	3D7	3D7	3D7
PFL0020w	12	3D7	3D7	3D7	3D7	3D7
PFL0030c	12	3D7	3D7	3D7	3D7	3D7
PFL0935c	12	HB3	3D7	3D7	3D7	3D7
PFL0940c	12	HB3	3D7	3D7	3D7	3D7
PFL0947c	12	HB3	3D7	3D7	3D7	3D7
PFL1950w	12	HB3	3D7	3D7	3D7	3D7
PFL1955w	12	HB3	3D7	3D7	3D7	3D7
PFL1960w	12	HB3	3D7	3D7	3D7	3D7
PFL1970w	12	HB3	3D7	3D7	3D7	3D7
PFL2665c	12	3D7	3D7	3D7	3D7	HB3
MAL13P1.1	13	3D7	3D7	HB3	HB3	HB3
MAL13_0003	13	3D7	3D7	HB3	HB3	HB3
MAL13P1.356	13	3D7	3D7	3D7	3D7	3D7
PF14_0001	14	HB3	3D7	3D7	3D7	HB3

Table 4.3: Assigning *var* genes of the progeny to either the 3D7 or HB3 parent based on a visual inspection of paired read coverage. Four genes that were involved in a non-reciprocal recombination event were shown as GC.

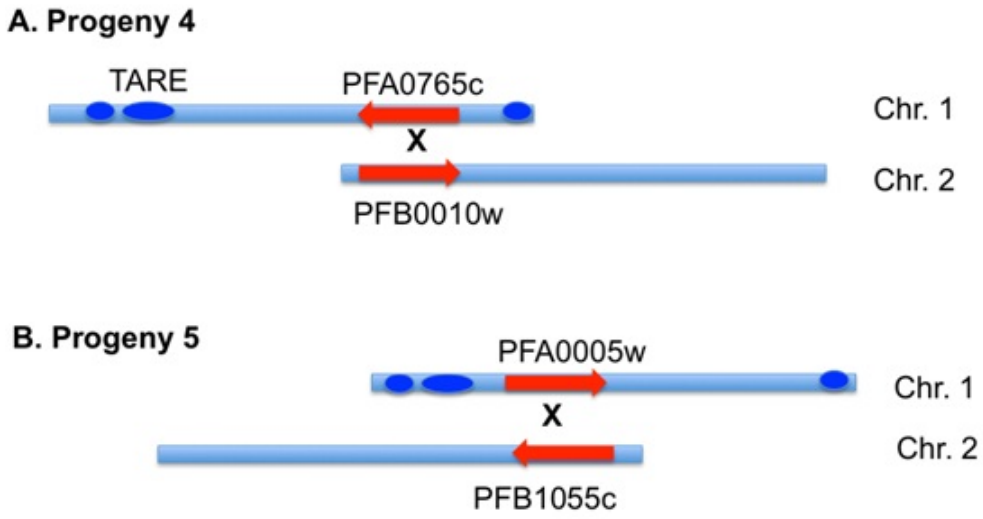


Figure 4.10: Subtelomeric *var* genes involved in ectopic recombination in progeny 4 and 5. *Var* genes on chromosomes 1 and 2 of the 3D7 parent showed evidence of a non-homologous recombination. The four genes were found to be of the group Type A.

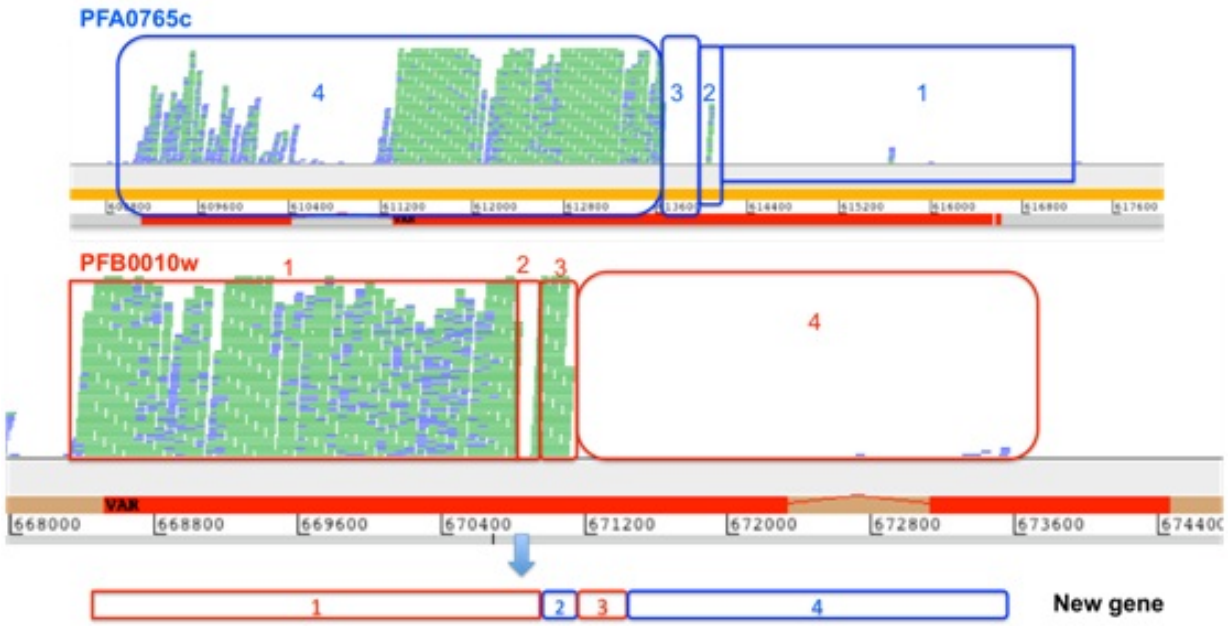


Figure 4.11: Evidence of gene conversion in progeny 4 (X4) from paired end coverage analysis. Coverage of paired reads is represented by a pileup of reads (blue and green stacks). A new *var* gene was formed by acquiring small blocks (1 to 4) from two genes on chromosomes 1 (PFA0765c) and 2 (PFB0019w). The upstream region and first half of Exon 1 was inherited from PFB0019w as shown by the increase in coverage over block 1 (red) on chromosome 2. Similarly, alternating blocks inherited from the two genes constituted the remaining regions of the new gene.

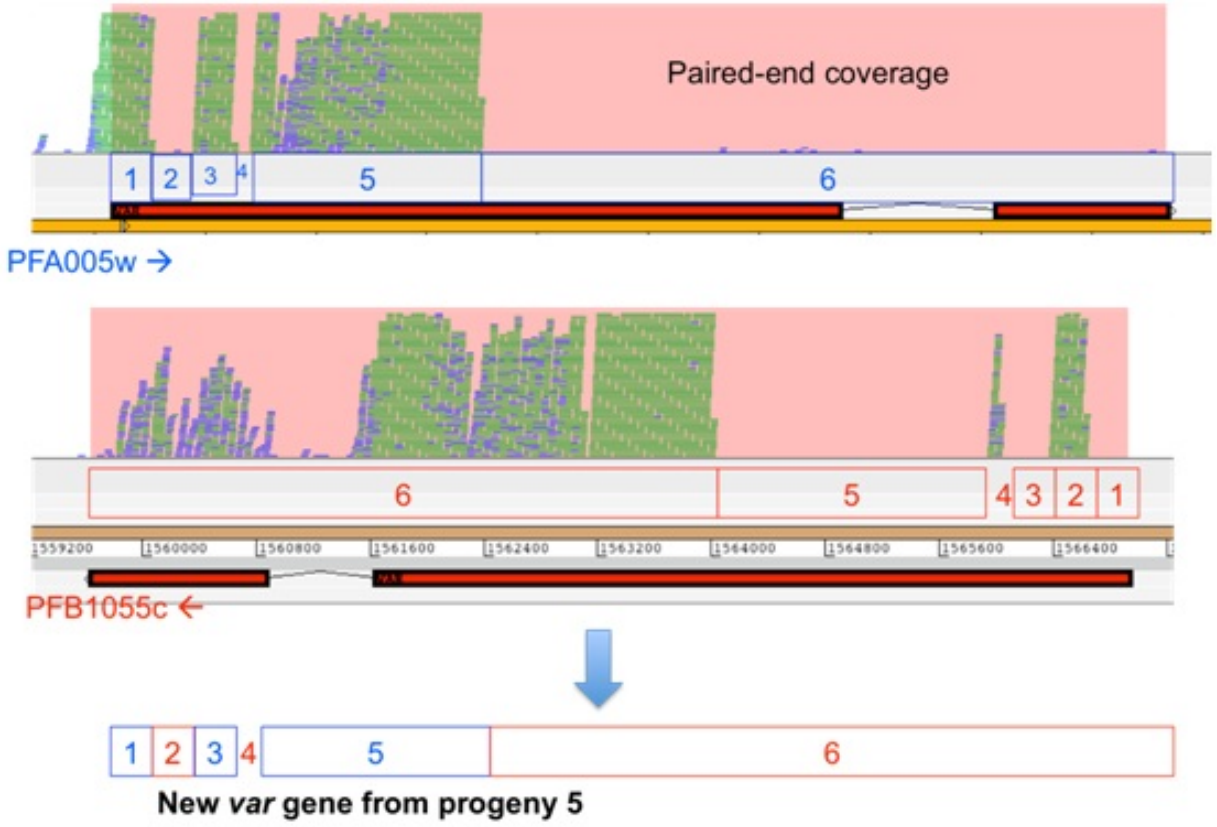


Figure 4.12: Evidence of gene conversion in progeny 5 (X5) from paired end coverage analysis. Description as above (Figure 4.11). Here, six alternating blocks from two *var* genes on chromosomes 1 and 2 of the 3D7 parent were used to generate a new *var* gene.

4.3.3.2 Reads mapping and *de novo* assembly to detect recombination in *var* genes

In order to systematically identify recombination breakpoints in *var* genes, progeny reads were aligned to a combined *var* reference (i.e. a multi-FASTA file containing *var* genes from both parents) allowing for a random placement of non-unique reads to one of the matching positions. Reads from the 3D7 parent were also aligned to the combined *var* reference and used as controls to detect false positive results. In addition to the previously identified two events: *PFA0005w X PFB1055c* and *PFA0765c X PFB0010w* (Table 4.3), a number of putative recombination events were detected (Table 4.4) by looking for read pairs that bridge two *var* genes of the same or different parent. Central *var* gene clusters on chromosomes 4 and 12 were also found in the putative list.

Events found in all or most of the progeny as well as those found in the 3D7 control sample were excluded, as they are likely to be caused by regions of high sequence similarity between *var* genes. The remaining events were further evaluated using *de novo* assembly of raw reads that align to both genes. The two events previously identified in progeny X4 and X5 were confirmed by *de novo* assembly of short reads (Figures 4.13 and 4.14). Lack of evidence from the control samples and formation of contigs with long open reading frames prove that the events are genuine.

The two events previously identified in progeny X4 and X5 (Figures 4.11 and 4.12) were confirmed by *de novo* assembly of short reads as shown in Figures 4.13 and 4.14. Lack of evidence from the control samples and formation of contigs with long open reading frames prove that the events are genuine. Both cases represent non-allelic gene conversions between telomeric *var* genes of chromosomes one and two of the 3D7 parent. Recombinant *var* genes were made of four to six alternating sequence blocks obtained from the parental genes. Block sizes as short as $\sim$200 bp were detected in both events. It was however not possible to determine whether these events were mieotic or mitotic as the parental clones underwent both sexual and asexual development stages. Gene conversion events could thus be a result of mieotic or mitotic non-homologous recombination exchanges.

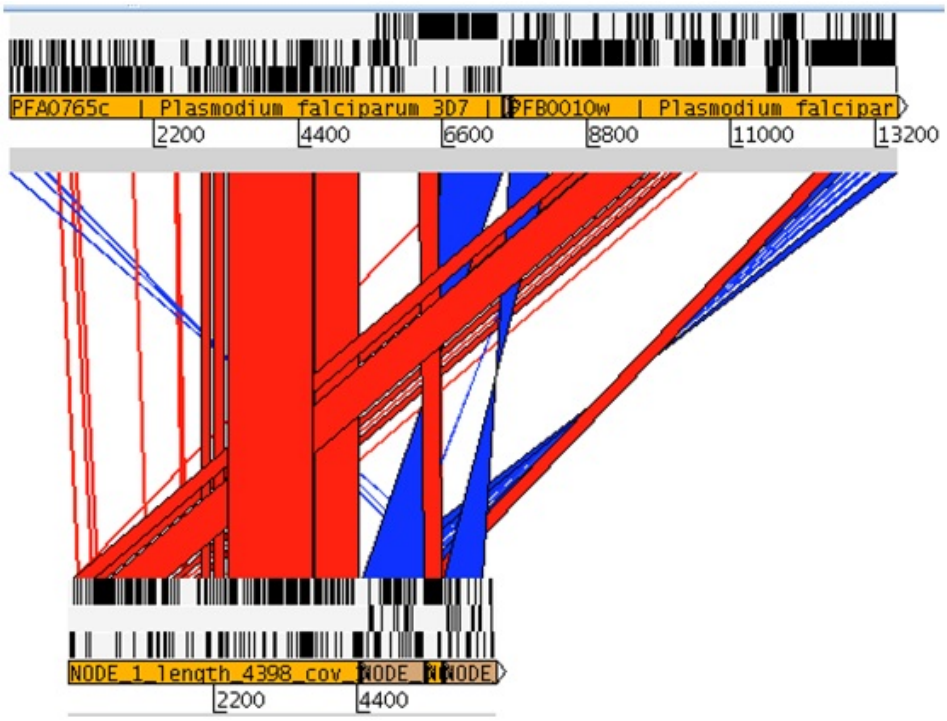


Figure 4.13: *De novo* assembly of reads that map to PFA0765c and PFB0010w in progeny 4 (X4). The top panel shows sequences of the two genes PFA0765c and PFB0010w in the forward strand. The three frames of the forward strand are shown where black lines indicate stop codons and long white regions indicate open reading frames. The bottom panel shows four contigs that were generated by the *de novo* assembly of reads that mapped to PFA0765c and PFB0010w. The red and blue bars indicate matches between sequences from the top and bottom panels (blue bars for reverse complement matches).

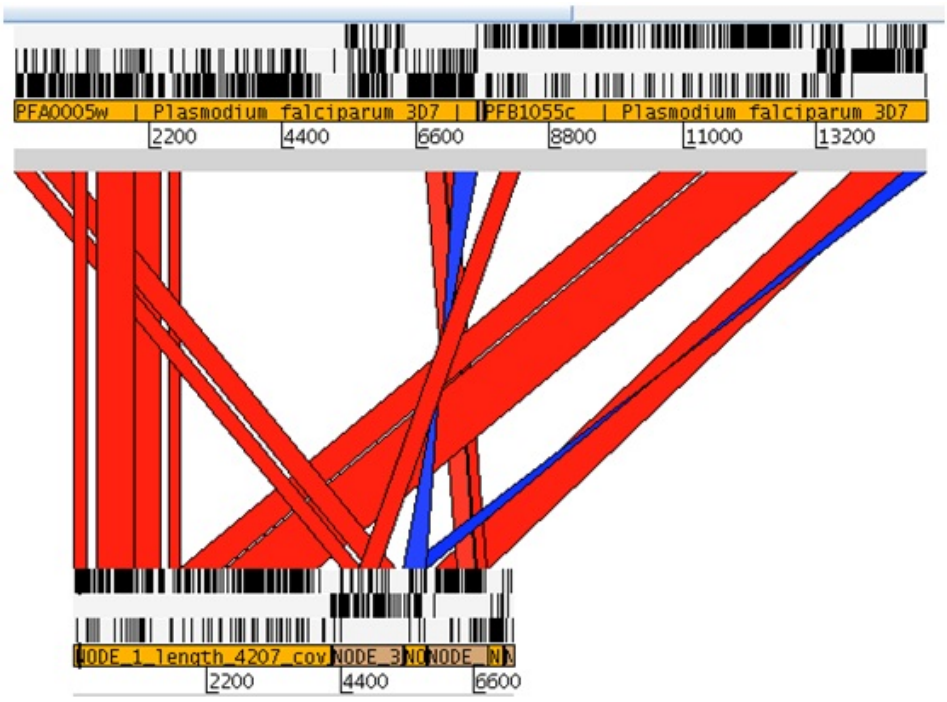


Figure 4.14: *De novo* assembly of reads that map to PFA0005w and PFB1055c in progeny 5 (X5). See description for Figure 4.13

Gene 1	Gene2	X1	X2	X3	X4	X5	3D7
PFA0005w	**PFB1055c**	-	-	-	-	**3295**	-
PFA0765c	**PFB0010w**	-	-	-	**1071**	-	-
PFC0115c	PFD1025w	87	97	-	92	84	-
PFD0140w	PFD0615c	-	-	**216**	-	-	-
PFD0615c	PFD0625c	-	-	**217**	-	-	-
PFD0625c	PFHG_02419	-	-	**302**	-	-	-
PFD0630c	PFHG_02419	-	-	**214**	-	-	-
PFD0635c	PFHG_02419	-	-	**93**	-	-	-
PFD0655w	PFD0995c	-	-	**69**	-	-	-
PFD0995c	PFD1015c	-	-	**129**	-	-	-
PFD1000c	PFD1015c	-	-	**122**	-	-	-
PFD1005c	PFD1015c	-	-	**493**	-	-	-
PFD1235w	PFL1955w	-	**62**	-	-	-	-
PFF1580c	PFF1595c	-	-	-	171	-	-
PF08_0107	PFHG_02419	-	-	77	-	-	-
PFL0005w	PFHG_04770	71	69	-	-	58	-
PFL0005w	PFL0020w	240	-	-	-	64	-
PFL1955w	PFL1960w	-	483	724	379	404	395
PFL1960w	PFHG_02419	-	321	480	228	325	-
PFL1960w	PFL1970w	-	540	690	306	265	136
MAL13P1.1	PF13_0003	83	70	-	-	-	-
PF13_0003	PFHG_03234	**56**	-	-	-	-	-
PFHG_02272	PFHG_04910	-	**68**	-	-	-	-
PFHG_02429	PFHG_04081	94	80	-	-	-	-
PFHG_03234	PFHG_05483	-	**125**	-	-	-	-
PFHG_03671	PFHG_05483	-	406	-	215	-	-
PFHG_03839	PFHG_03999	244	138	-	163	253	-
PFHG_03839	PFHG_04859	242	128	-	206	269	-
PFHG_03999	PFHG_04859	236	204	-	300	260	-
PFHG_04081	PFHG_04368	160	180	-	109	-	-
PFHG_04081	PFHG_04928	105	130	-	101	82	-
PFHG_05483	PFHG_05502	-	**86**	-	-	-	-

Table 4.4: Genes bridged by mate-pairs: Number of reads where mates map to different genes (mismatch ≤ 2). Reads from the 3D7 parent were also aligned to *var* parental genes and used as a control. Gene-pairs that were bridged by mate-pairs in all progeny or in the 3D7 were thus excluded.Gene pairs shown in boldface were recombinant genes that were confirmed by *de novo* assembly as shown in Figure 4.13. Values shown in boldface represent gene pairs that were bridged by a minimum of 50 read-pairs and not common to all progeny.

The remaining potential recombination events shown in Table 4.4 could not be confirmed using *de novo* assembly due to the formation of two or more separate contigs representing individual genes instead of a recombinant gene.

4.4 Discussion

This chapter explored applications of the Illumina sequencing technology to investigate the mechanisms used by parasites to generate diversity in *var* genes. The following conclusions summarise the results of sequence analysis on five progeny from a genetic cross experiment between the 3D7 and HB3 parental clones.

Conclusion 1: *Despite challenges of using very short (54 bp) reads, it was possible to obtain a global view of recombination, even in subtelomeric regions.*

The main challenge in using short reads for progeny sequence analysis was mappability. On average, the number of reads mapped to individual parental genomes is expected to be ∼50% of the total. However, variation from the expected value was observed due to recombination events that favor over-representation of one parent. The lowest percentage of read mapping to the combined reference genome was primarily due to the high sequence identity between core regions of parental chromosomes. A lower proportion of reads aligned to the HB3 reference (26 to 52%) compared to 3D7 potentially due to the incomplete nature of the genome.

Similar results were obtained from the two complementary methods of detecting genome wide crossover events (15 to 21 events). These observations are in agreement with the expected number of meiotic crossover events and previous reports (Jiang et al., 2011). An approach based on detecting SNP dense areas of a progeny may provide a quick overview of large-scale breakpoints. However, shorter fragments of low SNP density may also be caused by lack of coverage as a result of low complexity and the difficulty of reliably calling variants. It is therefore crucial to understand the genomic context of the region and decide whether a region is ambiguous prior to assigning genotypes. While some of such ambiguous regions may be common to all progeny due to inherent sequence features of the genomes, others are specific to a given sample as a result of the issues with the library preparation and the sequencing.

Despite the advantages of using short reads to obtain a nucleotide level resolution of breakpoints, notable limitations were observed in detecting subtelomeric breakpoints with a read length of 54 bp. The synteny between core

regions of 3D7 and HB3 allows reads to be aligned to either parent with mismatches (SNPs) that could be analysed to identify regions inherited from each parent. This approach was found to be adequate for detection of larger crossing over events. However, due to high polymorphism in *var* genes and subtelomeres, aligning short reads within tolerable allowance of mismatches was not possible. A different approach was thus required to detect smaller gene conversion events. For conversion tracts larger than the fragment size, following mates of reads that map near breakpoints was a better way of identifying donor and acceptor regions. Longer reads and larger insert libraries will be needed to reliably detect breakpoints as well as indels in subtelomeric regions where recombination rate is ∼10X higher than core regions of the genome (Taylor et al., 2000c).

Conclusion 2: *This chapter demonstrates the use of short-read sequencing to obtain a detailed resolution of ectopic gene conversion and shuffling of sequence blocks employed by parasites to generate new* var *genes (Figures 4.11 to 4.14)*

In addition to chromosome level breakpoints, short reads were used to identify segments of *var* genes that were inherited by each progeny. Regions of high read coverage when aligned to the 3D7 genome represented segments of the genome that were transferred to the progeny. Low (or zero) coverage on the other hand may imply a genuine absence of those regions in the progeny, implying they were inherited from the HB3 parent. However, lack of coverage could also be a result of low-complexity and repeats that prevent short reads from uniquely aligning in the subtelomeric regions where most *var* genes reside (Gardner et al., 2002). Signatures of recombination within and between parental *var* genes were thus detected using a combination of paired-end coverage inspection and targeted mapping of reads.

Although ectopic gene conversion was reported as a potential mechanism of generating new variants (Frank et al., 2008; Freitas-Junior et al., 2000), previous approaches based on hybridisation of the DBLα domain offered a limited insight on the extent of shuffling over the full length of genes.

Var genes that were involved in recombination via shuffling of sequence blocks were identified in two of the five progeny on chromosomes 1 and 2 of the

3D7 parent. The genes on Chromosome 1 were located adjacent to telomeric-associated repetitive elements (TAREs), which may be involved in initiating homologous recombination. This finding is consistent with previous studies that associated repeats and low complexity regions with hot-spots and elevated rates of recombination (Jiang et al., 2011). Random clustering of telomeres, during both the sexual and asexual stages of the parasite, is believed to provide a means for enhanced shuffling of sequence blocks in *var* genes (Freitas-Junior et al., 2000).

The genes that were identified to be recombining were members of the group Type A in both progeny. In addition, the genes were located on the 3D7 parent and between chromosomes 1 and 2. Recombination in *var* genes is believed to occur within defined hierarchies, such that members of Type A genes recombine more often with other type A than non-Type A genes (Bull et al., 2008; Kraemer et al., 2007). Our findings are also consistent with this observation, and also shed a new light on the extent of recombination via shuffling of blocks. However, understanding mechanisms of such frequent events (up to six recombining blocks were detected between two *var* genes in progeny 5) at the molecular level needs further investigation.

Central *var* clusters are known to be involved in spontaneous recombination (Deitsch et al., 1999), and may be ideal regions for gene conversion due to their organisation and high sequence similarity. We thus expected to see central *var* gene clusters in our putative list of recombinant genes (Table 4.4). However, it was not possible to reliably confirm these recombination events due to limitations imposed by the short read length. In addition, the fragment size (200-300 bp) was shorter than identical sequence fragments (repeated sequences) that are characteristic of these genes as described in Chapter 2.

In terms of absolute numbers, fewer than expected non-parental *var* genes were identified to be involved in recombination (only two of the progeny *var* genes were confirmed to have signatures of recombination) compared to an earlier study by Taylor and colleagues (Taylor et al., 2000a) that identified 24 events. The short read length (54 bp) and small fragment size of the library (200 bp) were major limiting factors. It was also not possible to *de novo* assemble *var* genes using approaches described in Chapter 3. Additional potential reasons

include low complexity, high percent identity between parental genomes and poor sequence quality resulting in unreliable read mapping.

Chapter 5

Assembly of *var* genes from clinical samples

5.1 Introduction

The study of *var* genes especially in clinical samples taken directly from patients will significantly improve our understanding of their diversity and evolutionary history. Previous studies that involve sequence analysis have mainly looked at diversity and expression of *var* genes on a limited number of clinical isolates. A number of studies have also shown association of sequence features with specific disease phenotypes (Ariey et al., 2001; Bull et al., 2005; Cham et al., 2010; Falk et al., 2009; Jensen, 2004; Kaestli et al., 2004, 2006; Kalmbach et al., 2010; Kirchgatter and Portilo, 2002; Kyriacou et al., 2006; Lavstsen et al., 2005; Montgomery et al., 2007; Nielsen et al., 2002; Normark et al., 2007; Rottmann et al., 2006).

However, all except two of the previous studies that used a sequence analysis approach of *var* genes have focused on the DBLα region (Kraemer et al., 2007; Rask et al., 2010). Although it was possible to accurately classify *var* genes into existing groups and make associations with disease severity using sequences taken from the DBLα domain, a large proportion of the *var* repertoire is still excluded. In the first comparative study to use full length genes, Kraemer and colleagues (Kraemer et al., 2007) analysed a near complete *var* repertoire of

three culture-adapted samples 3D7, IT and HB3. While 3D7 has a complete set of genes (i.e. the expected 60 protein coding genes), the other two had an incomplete repertoire. The results confirmed the presence of extreme diversity in *var* genes with only three genes (var1CSA, var2CSA and Type 3 *var*) showing a higher degree of conservation in all the three genomes. The second study by Rask and colleagues (Rask et al., 2010) used an additional four genomes (DD2, PFCLIN, RAJ116 and IGH) to provide a new definition of domain boundaries and identify recombination hotspots. A combination of phylogenetic and iterative homology block detection methods was used to define 628 homology blocks that could represent *var* genes with a better resolution than existing domain boundaries. However, due to the limits on the number and diversity of sequences used (311), these blocks may not accurately represent *var* genes in natural populations.

Understanding the order of sequence blocks and mosaic domains is of great importance. Recent studies have associated specific domain cassettes (as defined by Rask and colleagues (Rask et al., 2010)) with disease severity and a rosetting phenotype (Avril et al., 2012; Claessens et al., 2012; Lavstsen et al., 2012). Such association studies may facilitate the discovery of important antigens that could be used as potential vaccine targets for severe malaria. Obtaining full-length sequence information on *var* genes may thus be a step forward in such attempts. Moreover, lack of full-length information continues to be a major roadblock in understanding *var* gene diversity.

The focus of this thesis was to develop a new approach for the assembly of *var* genes from short reads of second generation sequencing platforms. As described in previous chapters, assembly of *var* genes using existing tools was not practical due to high polymorphism and the mosaic nature of *var* genes (Chapter 2). An iterative assembly approach that takes advantage of the inherent mosaicism in *var* genes was thus developed (Chapter 3) and evaluated on culture-adapted and a small number of clinical samples (50 samples). Here, the new approach is applied on a larger number of clinical samples.

Assembly of clinical samples adds another layer of complexity due to a number of difficulties associated with the quality of the input DNA and raw sequence data. Contamination with host DNA could result in a lower amount

of starting material, and therefore low yield of sequence data. In addition, systematic errors and bias towards certain sequence features due to the sequencing chemistry may affect data quality and result in reads with errors. Multiple genotypes circulating in a single individual contain highly similar as well as polymorphic haplotypes that affect the structure of the de Bruijn graph and quality of the resulting assembly. Although some of the challenges are being addressed by improvements in library production protocols used for sample preparation and sequencing (Oyola et al., 2013), the effect of poor quality data and uneven coverage still poses a unique challenge in assembly of *var* genes.

In this chapter, the iterative assembly approach described in Chapter 3 was applied to a larger collection of clinical isolates consisting of 743 samples taken from Africa, South East Asia and South America.

5.2 Methods

5.2.1 Sequence data

Clinical samples of *P. falciparum* were obtained from the Plasmodium Genome Variation (PGV) project at the Sanger Institutes malaria programme (`www.sanger.ac.uk/research/areas/malariaprogramme/`). Methods of sample preparation and the sequencing technology have seen a significant improvement over the last few years. Samples sequenced during the early days of the project were especially of low yield and poor quality with shorter read lengths of 37 and 54 bp. It was therefore decided to exclude samples that had a read length of below 76 bp. Samples that were not prepared using the PCR-free protocol (Chapter 2) were also excluded.

5.2.2 Initial Motifs and iterative assembly of clinical samples

The assembly work flow for a large number of clinical samples is illustrated in Figure 5.1. As described in Chapter 3, a total of 50 clinical samples were assembled for 20 iterations to evaluate the iterative assembly approach developed in this thesis (Chapter 3).

To obtain the maximum number of seed motifs for the clinical sample assembly, the assembly of the 50 samples was repeated for another iteration. Initial motifs to assemble 743 samples were thus obtained from the 21st iteration by translating contigs with the DBLα tag (*var*-contigs). Although the open reading frame (ORF) that contains the DBLα tag could be used to identify the correct reading frame, presence of frame-shifts meant some of the long DBLα may be excluded. *Var*-contigs were thus translated in all the six frames to generate the initial set of shared motifs. Once started with these motifs, the assembly of the clinical samples was repeated for three iterations, with each iteration involving sub-iterations of scaffolding and extension. Seed contigs were generated by optimising *k-mer* sizes for the assembly in two categories. For reads with a length of 76 bp, *k-mer* sizes of 51, 65 and 71 were used. For reads of length 100 bp and above, an additional *k-mer* size of 81 was used. Assembly results with different *k-mer* values were compared based on the number and

N50 sizes of *var*-contigs. For each sample the *k-mer* value that resulted in the highest number of contigs with DBLα and the highest N50 value was chosen to generate seed contigs. A final list of motifs (length=10 aa) was generated from *var*-contigs of the third iteration by considering a single frame with the longest Open Reading Frame (ORF) in either the forward or reverse strand. If the longest frame does not contain the DBLα tag, ORFs longer than 300 aa from the three frames on the strand of choice were chosen.

5.2.3 QC and filtering

Assembly quality was measured using the number of *var*-contigs, sum of *var*-contigs, N50 and largest contig sizes. The number of *var*-contigs was used as a measure of repertoire completeness for each assembled sample. Initially, samples that had below 30 *var*-contigs were excluded as they were found to be a result of low yield or poor quality sequence data. Samples with more than 70 *var*-contigs were defined as having multiple infections. Initially, these cutoff values were determined based on the expected number of *var* genes ($\sim$60) from previous studies on laboratory clones and clinical isolates. Additional quality checks include comparing assembly statistics of *var*-contigs with expected values from *var* genes of the reference genome 3D7. Using the method proposed by Bull et al. (2007), *var*-contigs were grouped into one of the six groups. The count of contigs in each group was compared with that of assembly results from the 50 samples (Chapter 3) and the reference genome 3D7. The most reliable method of checking assembly quality of *var*-contigs would be to compare with *var* genes of the reference genome. However, as described in the previous chapters, such approaches are not practical for the highly polymorphic *var* gene family. One approach adopted in my thesis to overcome this limitation was to look at ORFs instead of nucleotide sequences of contigs. In addition to providing a better measure of contiguity, using ORFs could minimize the effect of low complexity regions in introns and upstream and downstream regions of *var*-contigs. ORFs with a minimum length of 300 aa were obtained from each *var*-contig and stored as separate entries. For example, two ORFs of a *var*-contig (VAR1) from Sample1 were represented as follows:

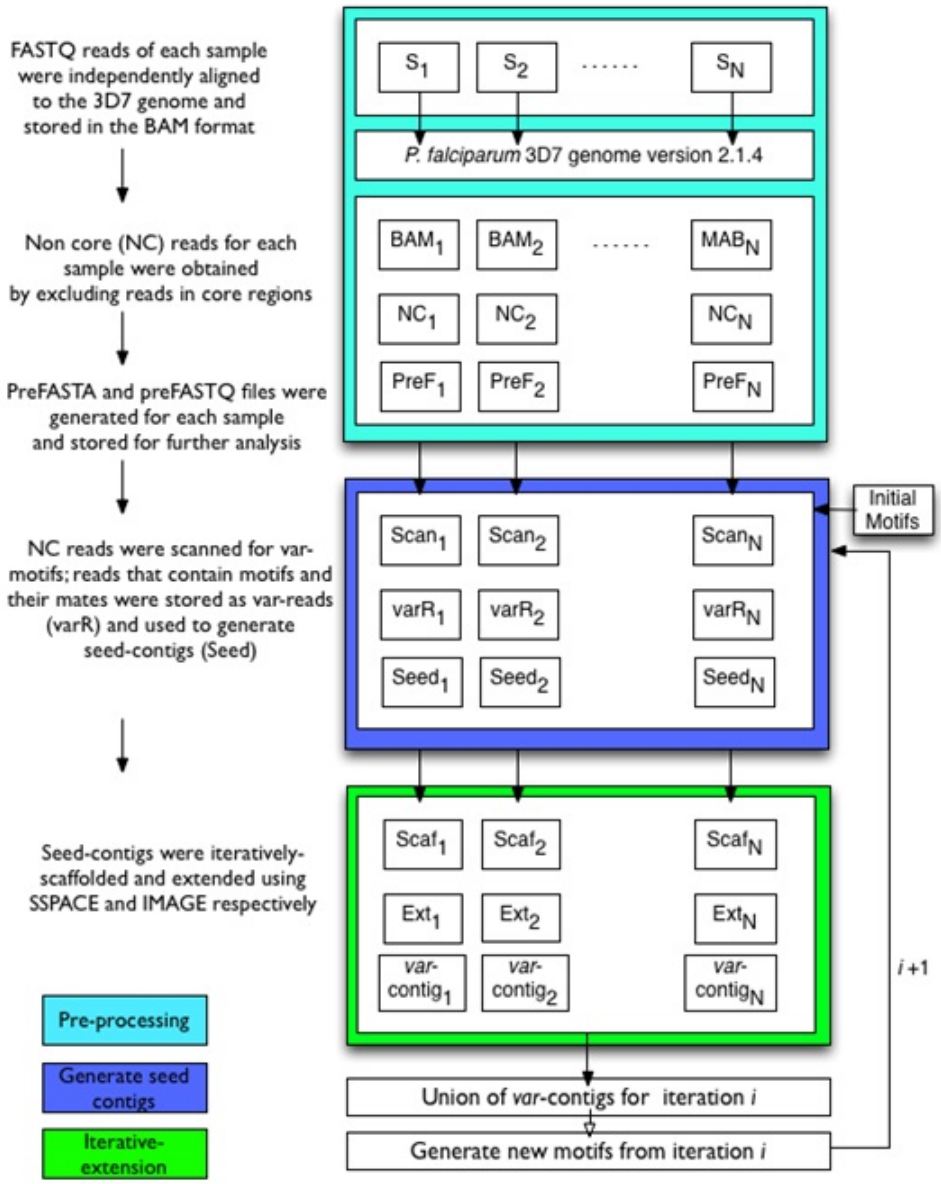


Figure 5.1: Iterative assembly work flow for *var* genes in clinical samples. Processes of the three stages of the *var* assembly are shown in boxes with cyan, blue and green backgrounds. Decisions on further iterations are made based on the quality of *var*-contigs from the current iteration. Assembly results of N=743 clinical samples are presented in this chapter (i=3).

Sample1_VAR1.ORF1
Sample1_VAR1.ORF2

The ratio of ORFs to *var*-contigs was used as a further quality control measure. Ideally, a full length *var*-contig should result in two ORFs representing each exon. It is however expected to find one ORF as most *var*-contigs may only capture the first exon. Introns and second exons are likely to cause ambiguities due to the high A+T content, and therefore generate smaller contigs that do not contain the DBLα domain.

5.2.4 Similarity between *var*-contigs

Similarity and relatedness of *var*-contig were analysed on three levels. Initially, we intended to use short-motifs generated from *var*-contigs during the assembly process. However, as the quality of contigs improved, it was possible to use longer matches using Pmatch (for perfect matches) and BLAST (allowing mismatches) as described below.

5.2.4.1 Pmatch analysis

Perfectly matching sequences were detected between any two *var*-contigs using Pmatch (minimum length=14 aa). Pmatch is written in C (Richard Durbin, Sanger Institute; unpublished) and rapidly identifies pairwise identical matches given two multi-fasta files of amino acid sequences. Amino acid translations of *var*-contigs were used as both query and subject for the pmatch analysis (i.e. all against all matching).

Var-contigs were translated by choosing the longest ORF in the strand where the DBLα tag was found. As mentioned in the previous section, if the longest ORF did not contain the DBLα tag, ORFs above 300 amino acids on the three frames of the chosen strand were concatenated. This method of detecting ORFs by jumping across the three frames of the main strand (i.e. the strand with DBLα) minimised the risk of missing ORFs due to frame shifts caused by misassemblies. Although this approach may also introduce the possibility of chimeric ORFs, the minimum length requirement of 300 amino acids was used

to account for such effect. The output of Pmatch required up to ∼250 Gb of storage disc space for a single analysis. It was thus necessary to convert the results to a simple motif-sharing format as defined in Chapter 3.

5.2.4.2 BLAST

Initially, nucleotide and amino acid BLAST (Altschul et al., 1990) databases of all *var*-contigs were generated using *formatdb*. In order to speed up the matching process, *var*-contigs of each sample were separately stored in a file and used as query during the BLAST search (*blastall -p blastp/blastn -e 0.001 -F F -m 8*). The output was compressed (*gzip -9*) prior to storage for further analysis.

5.2.4.3 Defining similarity between *var*-contigs and repertoires

Mosaic blocks of *var* genes result in a fragmented alignment profile between *var*-contigs. Similarity between two *var*-contigs was thus defined as a function of the total number of positions matched (i.e. the proportion of identical positions between the two *var*-contigs to the total length aligned). Similarity between *var* repertories was then computed as the average of all pairwise similarity values.

Similarity between two *var*-contigs V_1 and V_2 with n different blocks of matching sequences $m_1 \ldots m_n$ is defined as:

$$S_{v_1,v_2} = \frac{2 * \sum m_i}{L_1 + L_2}; \tag{5.1}$$

where m_i is length of match i, $i \in$ [1,n];
and L_1 L_2 are the full lengths of the two *var* contigs

5.2.5 Network analysis and clustering

Analysis of social networks was first used in *var* gene studies by Bull and colleagues (Bull et al., 2008). It was shown to be a better approach to study population structures and recombination hierarchies in the DBlα region than the phylogenetic tree based approaches which were shown to be impractical due to higher rates of recombination (Barry et al., 2007). The results of BLAST matching between *var*-contigs were processed to generate a graph of connected

var-contigs. The first set of samples that have a single infection (i.e. samples that contain below 70 *var*-contigs) were analysed. A graph was constructed by considering *var*-contigs as nodes. An edge between two nodes was added to the graph if two contigs have a match that fulfills the minimum identity and length requirements (eg. 99% and 1000 aa respectively for amino acid networks shown in the results section). A pairwise similarity index for two *var*-contigs was computed as described in the previous section. Each edge was thus updated according to weights obtained from the pairwise similarity values. These values range from 0 (no match) to 1 (two contigs are identical). A customised script (*blast*2*Gexf.pl*) was written to convert BLAST output files to the Graph Exchange XML Format (GEXF) (`http://gexf.net/format/`). First developed at the Gephi project in 2007, the GEXF format is widely used in representing a complex graph structure in terms of nodes and edges of a graph. In addition, a number of attributes such as weight and colour of nodes could be included in the graph file. Node colours were defined according to country of origin. In addition to visualising clusters, the Markov Clustering Algorithm (`http://micans.org/mcl/`) was used to generate clusters of *var*-contigs that share identical sequence blocks (Inflation parameter was tested at I=0.2, 1.2, 2, 4 and 6; final choice=1.2).

5.3 Results

5.3.1 Samples and sequence data

A total of 725 samples passed the selection criteria (i.e. samples prepared using PCR-free protocol and with a minimum read length 76 bp) at the beginning of this analysis (Figure 5.2, Table 5.1). These samples represented 13 countries from West Africa, East Africa, South East Asia and South America. The majority of samples came from The Gambia, Ghana and Cambodia. An overview of samples used in this chapter and their geographical origins are shown in Figure 5.2 and Table 5.1.

Figure 5.2: A global map of clinical samples used in this chapter. 725 samples were obtained from 13 countries representing West Africa (WA), East Africa (EA), South East Asia (SEA) and South America (SA). In addition to the 725 samples with known countries of origin, 18 samples from various countries that became available during the course of the study were also included.

5.3.2 Initial Motifs

Assembly of the 50 clinical samples of Chapter 3 plus an additional iteration (10 countries; 21 iterations) resulted in a total of 10.7×10^6 motifs from *var*-contigs.

These motifs were used to initiate the assembly process on the 743 samples (Table 5.1).

5.3.3 Iterations and additional motifs

The initial set of motifs were generated using amino acid translations in all the six frames. Although it was important to have a large number of motifs during assembly in order to increase the efficiency of the iterative process, such a high number is not required for the final analysis of shared motifs, as it will lead to unnecessary redundancy and an inflated count of overlapping genes.

At the end of the third iteration, a final list of 3.5 x10^6 motifs (10 aa long sequences as described in Chapter 3) were obtained from *var*-contigs using the longest ORF with the DBLα tag. Compared to generating motifs from amino acid translations of all the six frames, this approach reduced the number of motifs by ∼70%.

ID	Country	Region	#samples
PA	Gambia	WA	168
PF	Ghana	WA	122
PM	Mali	WA	32
PK	Burkina Faso	WA	3
PT	Malawi	EA	55
PC	Kenya	EA	25
PE	Tanzania	EA	15
PR	Bangladesh	SEA	3
PD	Thailand	SEA	82
PH	Cambodia	SEA	191
PN	Papua New Guinea	SEA	7
PV	Vietnam	SEA	11
PP	Peru	SA	11
Others			18
Total			**743**

Table 5.1: Samples used for initial assembly of *var* genes in clinical samples. A total of 743 samples were obtained from 13 countries as shown in Figure 5.1 (725 samples). Additional 18 samples from various countries became available during the course of the project and were also included.

5.3.4 Results of the initial assembly

Assembly results for the 743 samples were summarised by four commonly used assembly metrics: sum of contigs, N50 size of contigs, number of contigs and largest contig size (Figure 5.3). The variation observed in the quality of each assembly is shown by the distribution of values for these four measures. The sum and N50 of *var*-scaffolds show a wider distribution range reflecting the poor quality assembly on the extreme left side of the distribution as well as a mixture of genotypes (multiple infections) on the far right end of the distribution. Conversely, the number of *var*-scaffolds and the largest scaffold size were narrowly distributed with median values of ∼60 and ∼10 kb respectively, reflecting a high quality assembly in terms of repertoire completeness.

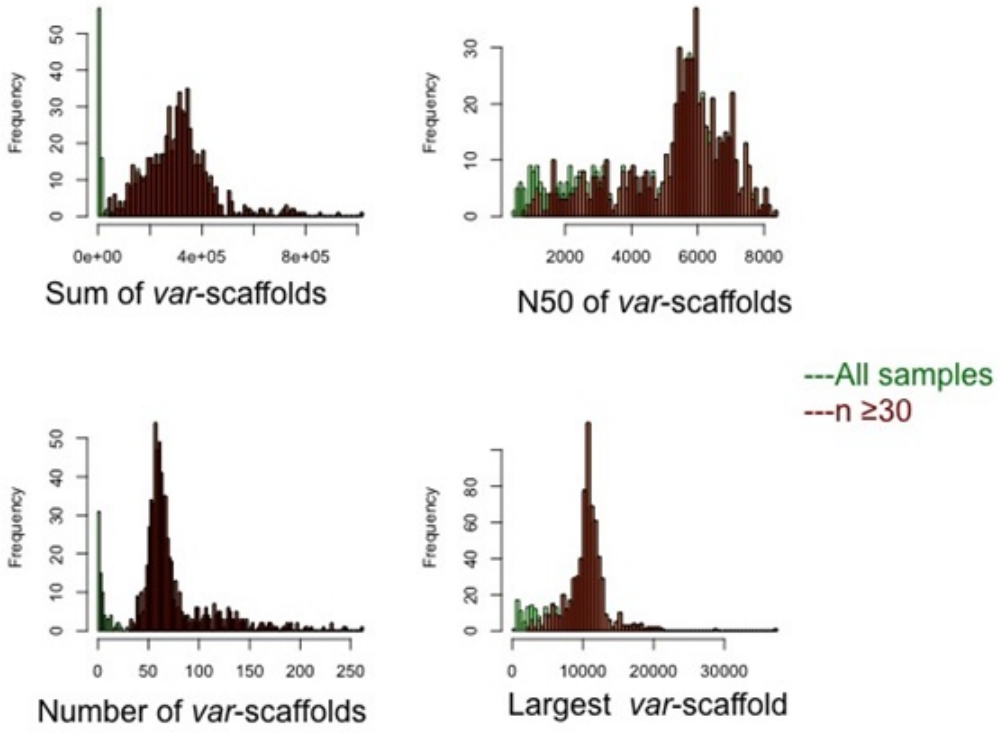


Figure 5.3: Assembly stats of the initial 743 samples. Green shades represent all samples while dark red shades represent samples with above 30 *var*-contigs. Sum of scaffolds, N50 and largest scaffold sizes were measured in base pairs (bp). An additional summary of the four measures is shown in Table 5.2.

In addition, summary of assembly results were shown by the range, mean and median values of five assembly statistics (Table 5.2). Sum of *var*-contigs was used as a measure of overall coverage of the *var* repertoire. The N50 and number of *var*-contigs measure the contiguity and repertoire completeness of *var*-contigs respectively. Mean values for the sum of *var*-contigs, N50 size and number of *var*-contigs for the 743 samples (∼351 kb, ∼5 kb and 68 respectively) revealed an overall highly representative assembly. The initial histogram plots of all 743 samples were shown in green bars in Figure 5.3. A total of 647 samples had above 30 *var*-contigs (i.e. 647 of the 725 samples of interest). The remaining 78 samples had a poor quality assembly with as few as one contig containing the DBLα tag. The count of non-core reads was investigated to find a reason for such fewer number of *var*-contigs in the 78 samples. Overall, there was a positive correlation between non-core read count and the four assembly measures (R^2 = ∼0.5; p<0.0001). The number of non-core reads was also noticeably low for the 78 samples (Figure 5.4 E, F). Samples with the least non-core read count came either from very recent multiplexed libraries (eg. PH0553-C, PH0581-C, PF0539-C had less than 300,000 read-pairs) or libraries that were sequenced at the beginning of the project (eg. PP0011-C, PF0007-C, PK0032-C and PC0034-C had over 1 Million read-pairs but with poor read quality). Multiplexed libraries were observed to generate inconsistent yield within the different samples that are sequenced in one lane (Magnus Manske, personal communication and preliminary assessment of recent multiplexed libraries). Conversely, sample PC0034-C had the fourth largest number of non-core reads (∼75% of total reads) suggesting issues with data quality instead of yield. Further investigation revealed unusually long insert sizes and a large number of duplicates (∼8% of total reads) and reads where the mate aligns to a different chromosome (∼10% of total reads).

A closer look at the assembly results was obtained by breaking the analysis down to regions (Figure 5.5) and countries (Figure 5.6). The total number of bases in each assembly provided a measure of how well the *var* repertoire is covered. Sum of *var*-contigs for each region revealed that samples from West Africa and East Africa had the largest range (733 bp to 835 kb and 1.1 kb to 640

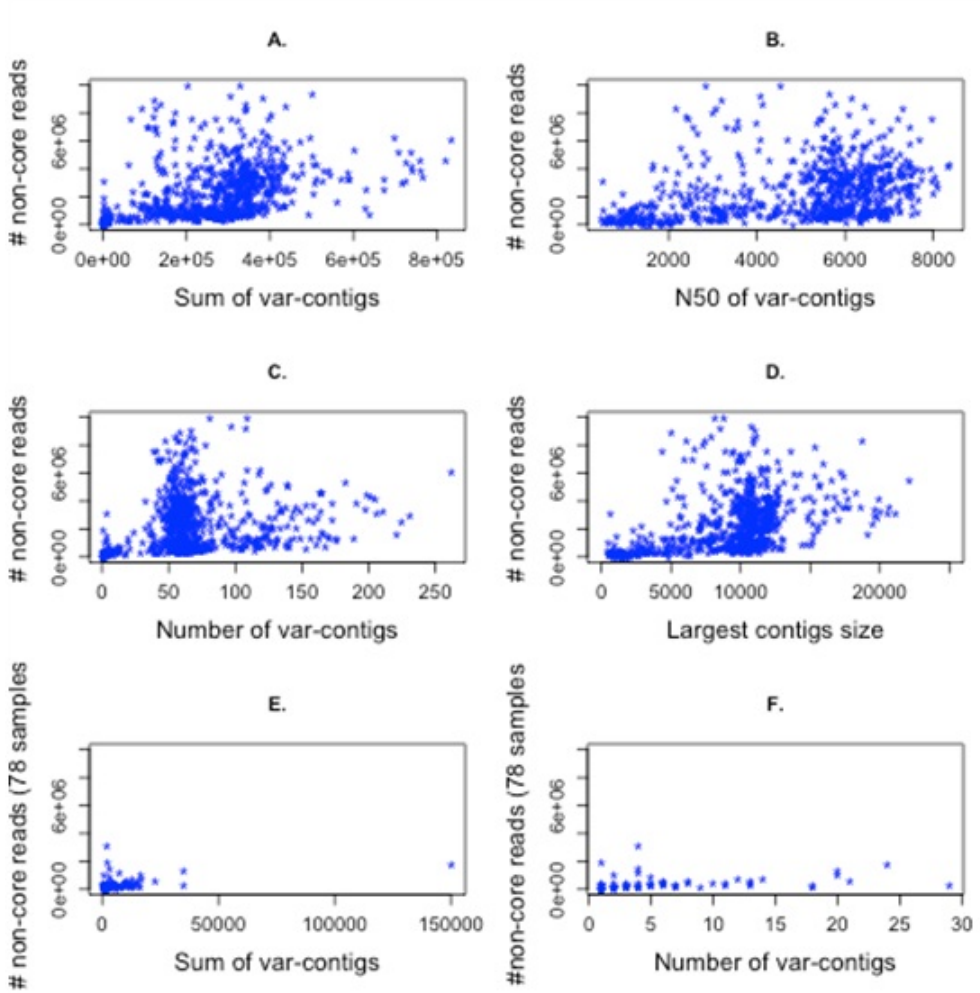


Figure 5.4: Scatter plots of non-core read counts with the four assembly statistics. **A-D)**. Sum of *var*-contigs, N50 contig size, number of *var*-contigs and Largest contig size for all samples. **E-F)**. Sum and number of *var*-contigs are separately shown for samples with less than 30 *var*-contigs.

	Min	Median	Mean	Max
Sum(bp)	477	288,500	350,800	835,100
N50(bp)	477	5,599	4,893	8,362
Num. contigs	1	61	68	262
Largest(bp)	477	10,420	9,705	37,040
N-count(bp)	0	2	403	15,010

Table 5.2: Summary of assembly results for the initial 743 samples. A graphical representation of the Sum, N50, Number of *var*-contigs and Largest contig size is shown in the four histograms of Figure 5.3.

kb respectively) compared to samples from South East Asia and South America. At first sight this may appear to be due to the large number of samples from West Africa (n=325) and East Africa (n=95). However, the narrow distribution in South East Asia can not be accounted for as they have a comparable number of samples (n=294). It is therefore likely that the distribution of sum of contigs as well as *var*-scaffolds is indicative of multiplicity of infection (MOI). West and East African populations were found to display higher values of multiplicity of infection, with up to a five-fold increase in the number of *var*-scaffolds (Figure 5.6). In addition, a visual inspection of aligned reads over the MSP1 gene for samples with the highest number of *var*-contigs confirmed more than one haplotype (Appendix B, Figure B-4). Samples that contain less than 30 *var*-contigs were excluded from further analysis in this chapter. However, they will be included in the future when improving the assembly by, for example, using additional iterative steps.

5.3.5 Initial quality control steps

Quality of assembled contigs was initially assessed using three approaches: count of ambiguous (unknown) bases, size of ORFs, and distribution of *var*-contigs in to the six groups (Bull et al., 2007).

Number of 'N's

Firstly, the total count of 'N's in each sample (i.e. number of gaps in the sum of *var*-scaffolds) was considered and found to be extremely low (median=2,

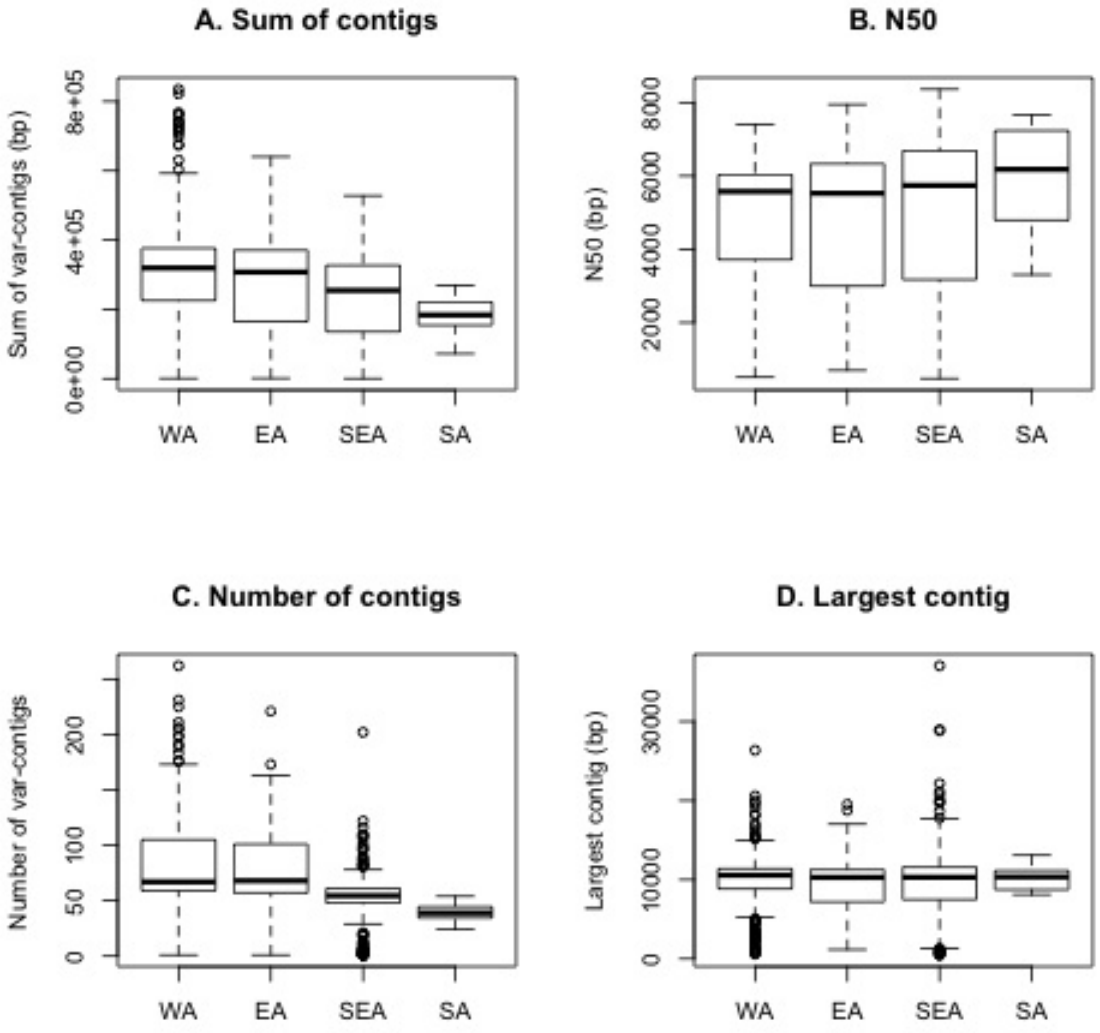


Figure 5.5: Box plots showing assembly statistics by geographical region. The four statistics were separately shown for West Africa, East Africa, South East Asia and South America. Box limits represent median, first and third quantiles; whiskers represent the upper and lower bounds while outliers are shown by the dots. A). Sum of contigs represents the total number of bases in *var*-contigs for the four regions B). N50 contig size distribution of *var*-contigs was $\sim$5 kb on average and consistent across the four regions. C). The number of *var*-contigs showed a similar pattern of variation with West African samples displaying a higher degree of variability compared to South East Asia and South American samples.

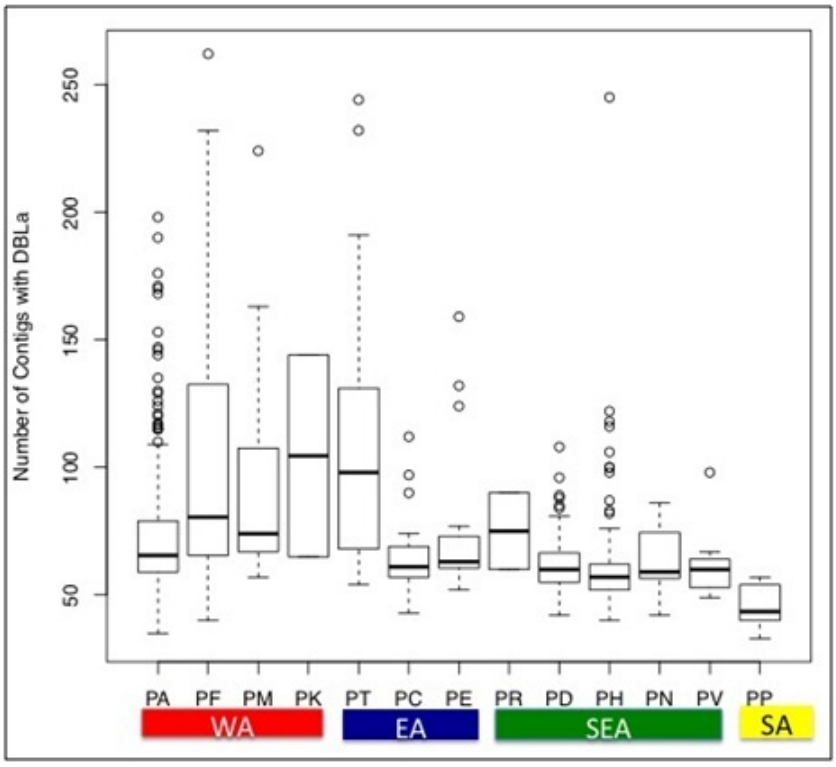


Figure 5.6: Number of *var*-contigs by country of origin. A better resolution on the distribution of the number of *var*-contigs is shown using box plots on a country level. Box limits represent median, first and third quantiles; whiskers represent the upper and lower bounds while outliers are shown by the dots. West African (WA) and East African (EA) samples had higher variability in the number of *var*-contigs than samples from South East Asia (SEA) and South America (SA).

mean=∼200). Together with the high N50 values (mean=∼5 kb), the fewer gaps observed in the assembly are indicative of a higher level of contiguity. The top ten samples with the highest number of Ns were found in The Gambia. However, these samples also had above 120 *var*-scaffolds and a sum of *var*-scaffolds between ∼500 kb and ∼840 kb suggesting the presence of multiple genotypes in the samples. The size of the largest *var*-scaffold was between ∼9 kb and ∼13 kb. In order to make sure the largest contigs are segregated by individual genotypes instead of creating false joins, a further quality cheek was conducted by investigating ORFs.

Size of ORFs

Secondly, in order to evaluate the effects of misassembly on assembly contiguity, the number and size of ORFs was examined. The number of ORFs for each sample was expected to be higher than the number of *var*-contigs (or *var*-scaffolds; but the term *var*-contigs is used here after in this section for simplicity) as *var*-contigs may have multiple ORF entries. If the ORF of a given *var*-contig is not interrupted by stop codons due to false joins that result in frame shifts, the upper limit for the number of ORFs is expected to be twice the number of contigs. A total of 51,140 ORFs were obtained with a minimum length of 300 amino acids. The ratio of ORFs to *var*-contigs was expected to be between one and two for samples with a good quality assembly. A ratio lower than one indicates that most contigs are shorter than ∼900bp. Conversely, a ratio of above two is a sign of frame-shifts as a result of potential mis-assembly. The overall ratio of ORFs (n=51,140) to *var*-contigs (n=50,131) was nearly one as the assembly process mainly captures exon 1 of the *var* repertoire. The sum of ORFs was equivalent to ∼91% of the sum of *var*-contigs (∼205 Mb). The remaining 9% of sequence is due to UTRs, introns and exon 2 sequences. N50 size of ORFs (1,761 aa) was also comparable with the N50 size of *var*-contigs (5,705 bp). The size distribution of ORFs from the assembled clinical sample was also comparable with ORFs from 3D7 and IT genomes (Figure 5.7).

In addition to the overall ORF distribution for all *var*-contigs, a closer look at the ratio of ORFs to *var*-contigs for each sample revealed that two samples PA0106 and PA0107 had the highest number of ORFs (170 and 156 respectively),

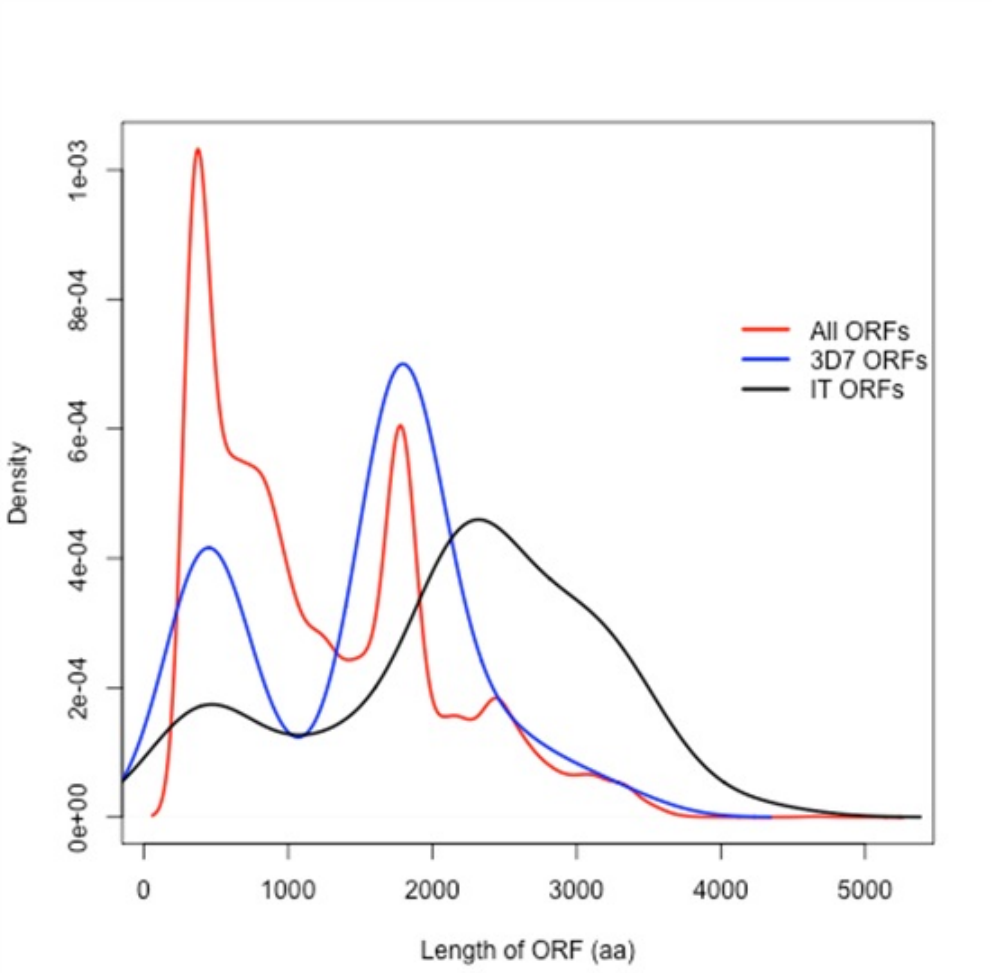


Figure 5.7: Density plots of ORF sizes for clinical samples, 3D7 and IT. The size distribution of ∼50,000 ORFs (red) is shown together with two culture-adapted samples 3D7 and IT with complete repertoires containing 83 and 74 ORFs respectively. The mean ORF sizes were 1217, 1484 and 2168 for all, 3D7 and IT respectively.

although the number of *var*-contigs was 59 and 56 respectively. The remaining samples showed no evidence of excessive frame shifts as the ratio of ORFs to *var*-contigs was within the expected range of one and two.

Grouping var-contigs

Finally, the number of *var*-contigs that are represented by the six groups (as defined by Bull et al. (2007)) followed a similar distribution as that of the 50 samples (Chapter 3) and the three culture-adapted samples 3D7, IT and HB3 where the majority of the genes fall into group 4 (Figure 5.8).

Taken together, these results confirm that the contigs and scaffolds generated from clinical samples were of a high quality.

5.3.6 A first look at the motif sharing *var*-contigs

Motifs generated from *var*-contigs of the third assembly iteration revealed that ∼40% of the total were unique to single samples. The remaining motifs were shared by a minimum of two samples with a heavily right-tailed distribution (Figure 5.9A).

5.3.7 Using full length sequences

5.3.7.1 Pmatch

The pmatch output file was converted to a shared-motif format, revealing perfect amino acid matches of length 14 to 3,404 aa. A large proportion (∼98%) of shared motifs were between 14 and 100 aa (Figure 5.10), and shared by the majority of *var*-contigs (Figure 5.11).

However, unexpected long perfect matches of length above 1,000 aa were also observed (∼600 motifs). Interestingly, these long motifs were shared between *var*-contigs of the same population as well as different populations. The longest motif shared by samples from different countries was 3,404 aa long and found in three samples (PA0036, PA0020 and PH0136), two from The Gambia and one from Cambodia.

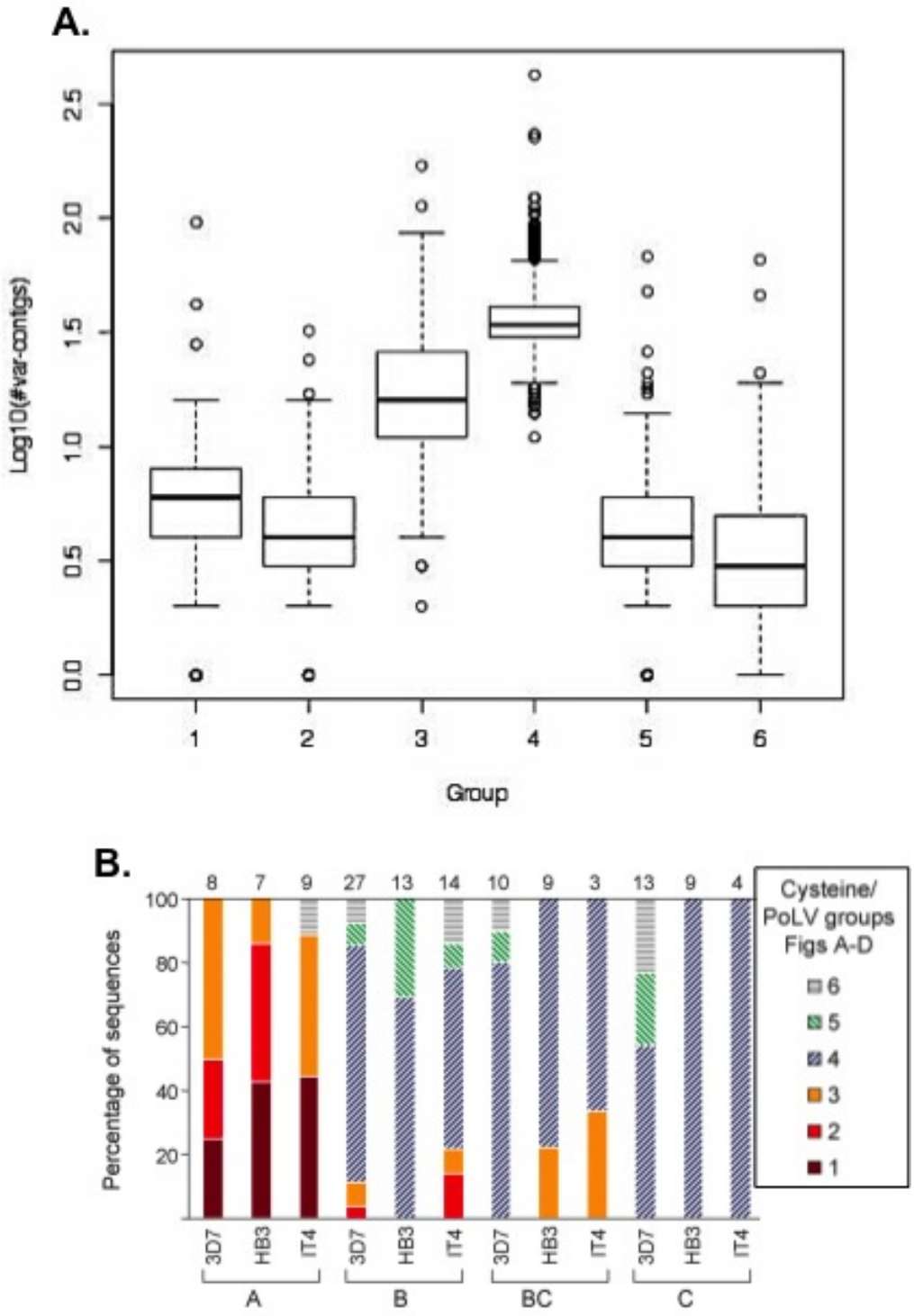


Figure 5.8: Grouping ∼50,000 *var*-contigs using the method proposed by Bull et al. (2007). A) Box plots show the distribution of *var*-contigs in the six groups. B) Correlation of the six groups with existing classification (A, B, C, BC) based on three culture-adapted samples (Adapted from Bull et al. (2007)).

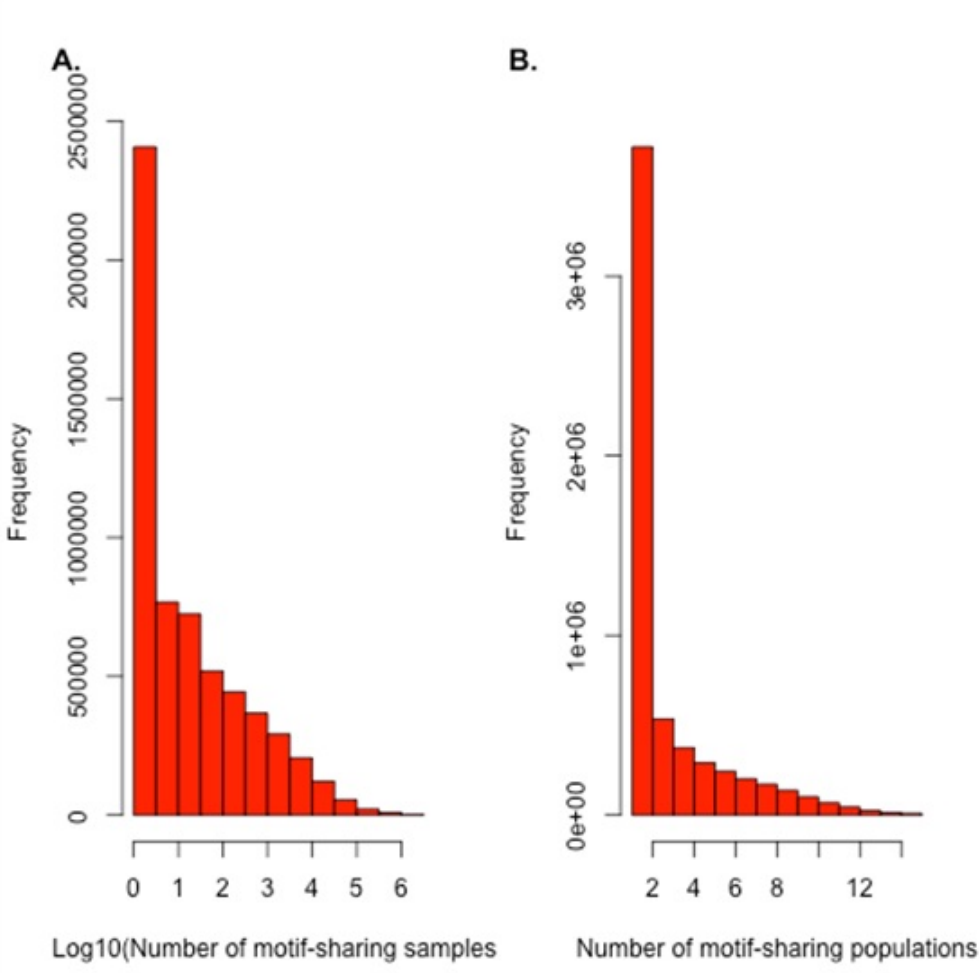


Figure 5.9: A histogram of samples **(A)** and populations **(B)** that share 10 aa long motifs. The majority of motifs were shared by one or two samples and populations with a heavy right-tailed distribution.

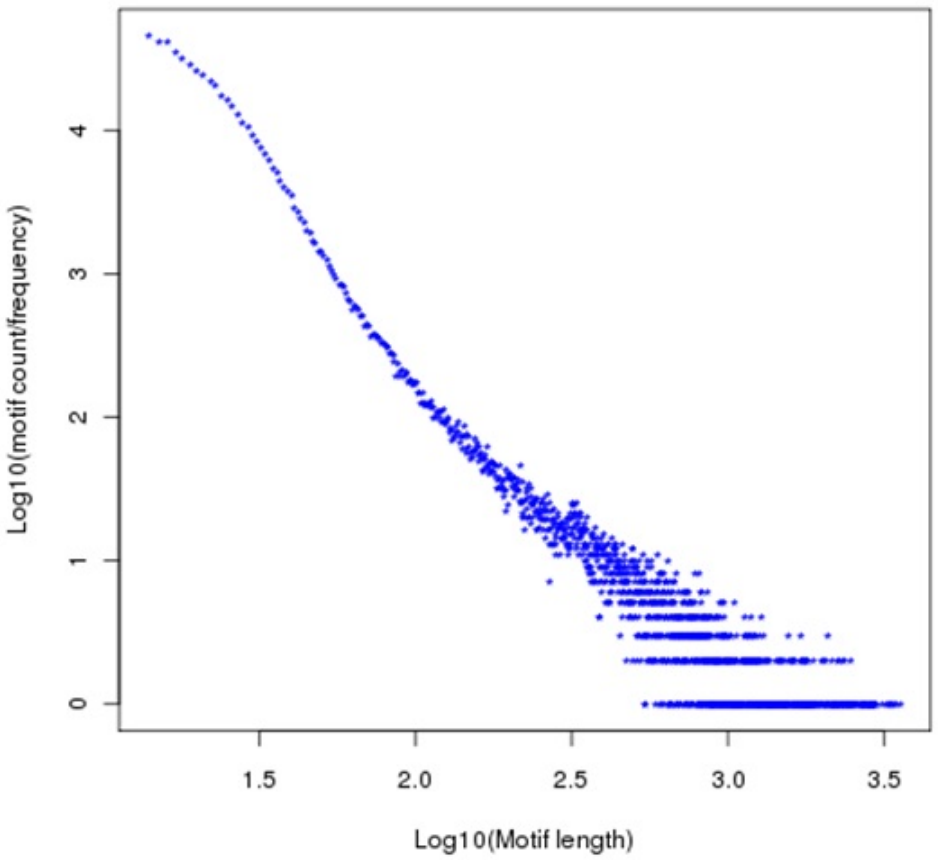


Figure 5.10: Frequency of shared motifs. This plot shows the number of shared motifs as a function of motif-length (x-axis). A large proportion of shared motifs were below 100 aa.

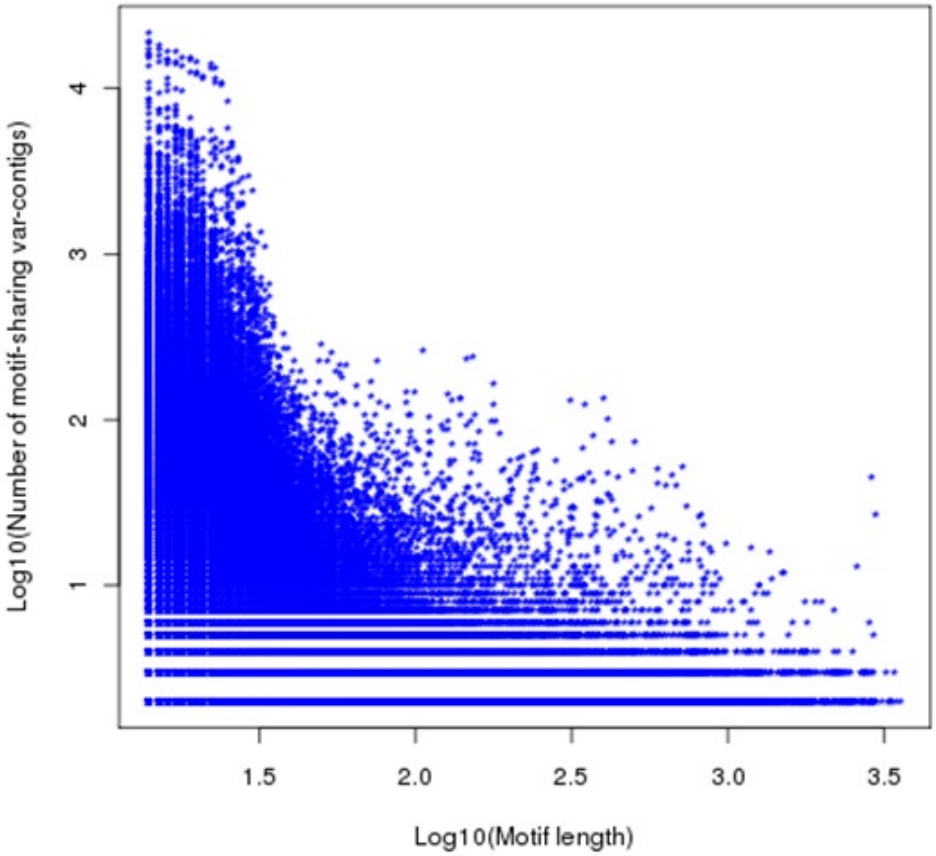


Figure 5.11: Motif sharing *var*-contigs from a pmatch analysis of all-vs-all *var*-contigs. The scatter plot shows a negative correlation (R^2=-0.2; p-value $< 2.2x10^{-16}$) between motif length and the number of *var*-contigs that share a motif.

Initially, we hypothesised that the three identical *var*-contigs were a result of contamination during sample handling. In order to verify this, raw reads from the three samples (PA0036, PA0020 and PH0136) and an additional control sample (from Thailand) were aligned to all *var*-contigs of the sample PA0036 (Figure 5.12). If there was contamination, non-core reads of the other samples would align to *var*-contigs other than PA0036_VAR1. However, the alignment results showed a distinct full coverage signal over PA0036_VAR1 from all the three samples. Other *var*-contigs of PA0036 were only covered by non-core reads of PA0036 (i.e. mapping to itself as expected). It was reassuring to confirm that non-core reads from a control sample (from Thailand) did not align to any of the *var*-contigs. The long perfect matches between single genes therefore appear to represent genuine biological events. It is expected to see a higher degree of conservation between *var* genes of the central clusters. It is thus intuitive to assume *var*-contigs with long perfect matches come from a central region of chromosomes. However, aligning the 10 largest motifs to the *P. falciparum* genome (via BLAST (Altschul et al., 1990)) revealed top matches to subtelomeric *var* genes such as PF3D7_0632500 on chromosome 6 (Figure 5.13). It is important to note that this may not be the best way of identifying central *var* clusters as the target (i.e. 3D7) is only one genome. A flanking sequence of *var* genes could provide a better marker to identify central *var* genes based on similarities of Ups sequences (see Chapter 1 for details).

To investigate whether the long motifs are associated with specific *var* groups (1 to 6), we looked at the number of distinct *var*-groups that are represented by a motif. The results show that the majority of *var*-contigs that share longer motifs were represented by fewer groups (1 to 2) than shorter motifs which can contain up to all six groups (Figure 5.14).

Next, analysis of the six groups represented by *var*-contigs that share a motif revealed that the majority of *var*-contigs that shared long motifs were of groups 1, 2 and 3 ($\sim$60% of *var*-contigs for a motif length of above 500 aa and $\sim$75% for motifs above 1,000 aa). These three groups are shown to contain a DBLα with two cysteine residues and correspond with Type A *var* genes (Bull et al., 2005, 2007, 2008).

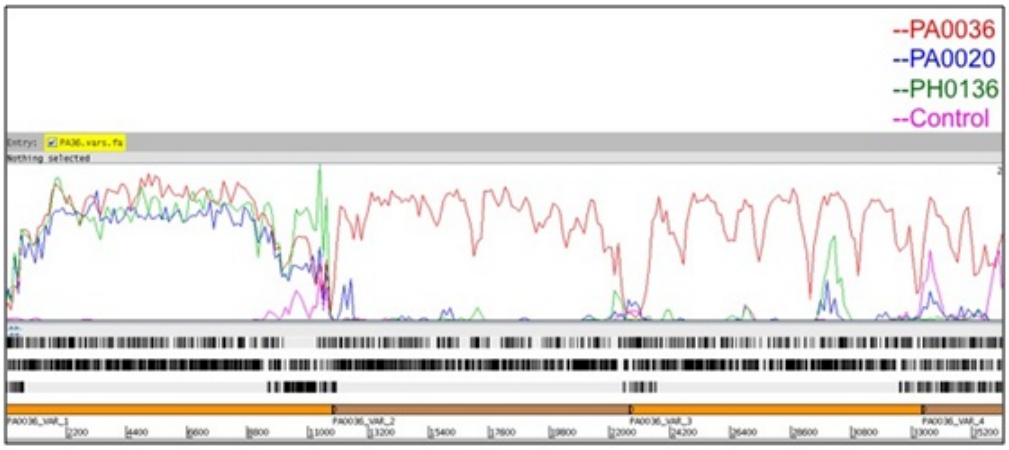


Figure 5.12: Artemis view of non-core reads from four samples (PA0036, PA0020,PH0136 and Control sample) aligned to *var*-contigs of PA0036. The top panel shows paired-read coverage plots for PA0036 (red), PA0020 (blue), PH0136 (green) and the Control sample. The middle panel shows the three reading frames of the forward strand for *var*-contigs of PA0036. Black bars represent stop codons, ORFs are represented by long open white blocks. The bottom panel shows *var*-contigs of PA0036 starting at PA0036_VAR1. Read coverage from samples PA0020 (blue), PH0136(green) was visible over PA0036_VAR1 while the remaining *var*-contigs of PA0036 remain uncovered.

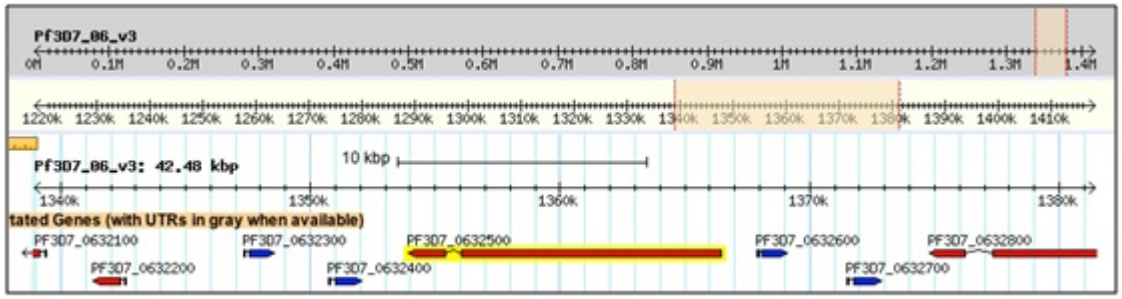


Figure 5.13: A screen shot of PF3D7_0632500 from PlasmoDB: The subtelomeric gene PF3D7_0632500 was the closest match (∼47% identity over the full length) to the long motif shared by three samples from different geographical regions.

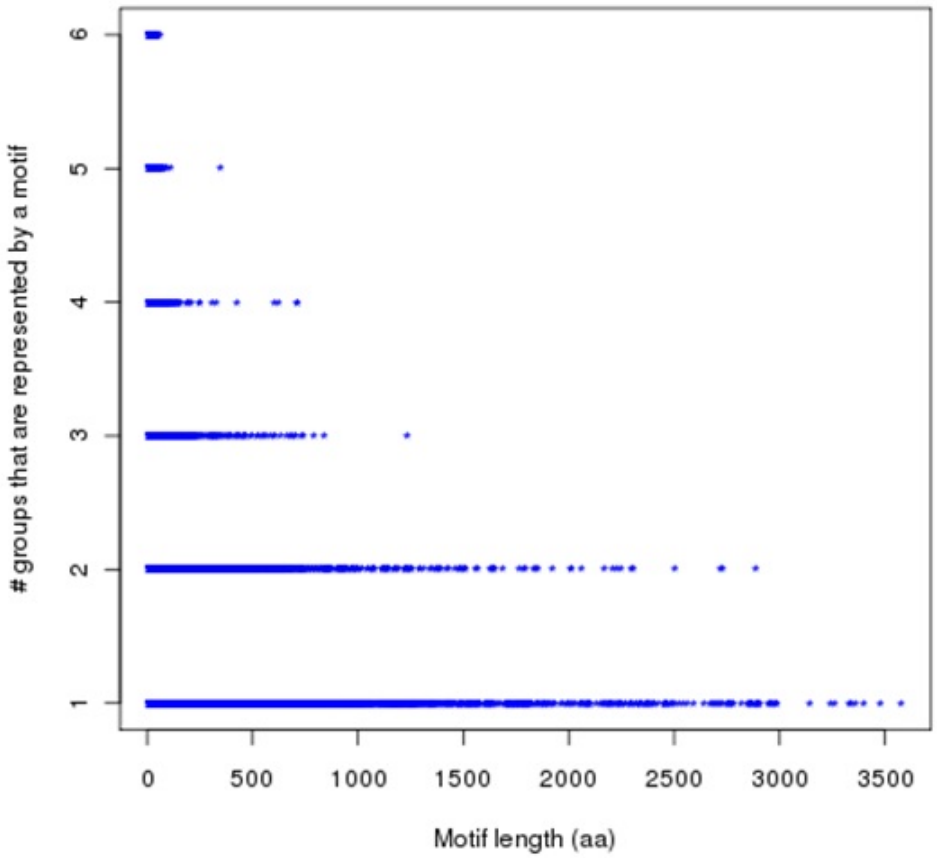


Figure 5.14: Number of groups represented by *var*-contigs that share a pmatch motif for 426 samples that had a minimum of 30 *var*-contigs. Long motifs were shared by *var*-contigs that belong to one or two distinct groups. Conversely, the short motifs are shared by *var*-contigs from all groups.

5.3.7.2 Amino acid and nucleotide BLAST matches

The pmatch analysis was only able to show perfect matches between *var*-contigs or conserved blocks within *var*-contigs. In order to investigate similarity between *var*-contigs while allowing mismatches, a BLAST search of all-against-all *var*-contigs was conducted. The results revealed long identical matches over the full length of *var*-contigs, confirming observations of the pmatch analysis. A closer look at nucleotide alignments of *var*-contigs also showed long identical matches that span over exons, introns and upstream regions. The observed sequence similarity between *var*-contigs of the same and different geographical origins(match lengh >5 kb; identity >99%) was higher than expected from previous studies. The results of the BLAST search provided a better way of processing the output and quickly identify matches of a given *var*-contig. For example, *var*-contig PA0036_VAR1 had a match with total of 59 *var*-contigs from 57 samples in 6 countries at a minimum identity of 99% and match length of 5 kb. The 59 *var*-contigs represented The Gambia (n=21), Ghana (n=6), Mali (n=4), Burkina Faso (n=1), Thailand (n=2) and Cambodia (n=25). 54 of the 59 *var*-contigs were of group 3 *var* genes, while the remaining five were of group 1. As mentioned previously, these two groups belong to Type A *var* genes.

Match length vs percent identity of a match

A scatter plot of nucleotide matches of PA0036_VAR1 to other *var*-contigs revealed a positive correlation ($R^2 = 0.3; p - value < 2.2x10^{-16}$) between match length and the percent identity of a match (Figure 5.15). As the match length decreased, the identity of a match also decreased. The observed relationship between match length and percentage identity of a match was further investigated by looking at all *var*-contigs at various identity cutoff values. Longer matches were predominantly found at higher percent identity thresholds (Figure 5.16). These results are interesting as they have implications on the time-scale of events that contributed to maintaining diversity in *var* genes. For example, the long perfect matches may be a result of recent population expansion events where there was not enough time for recombination to break these long haplotype blocks. Conversely, shorter matches indicate a longer time scale since the

event allowing more SNPs to accumulate.

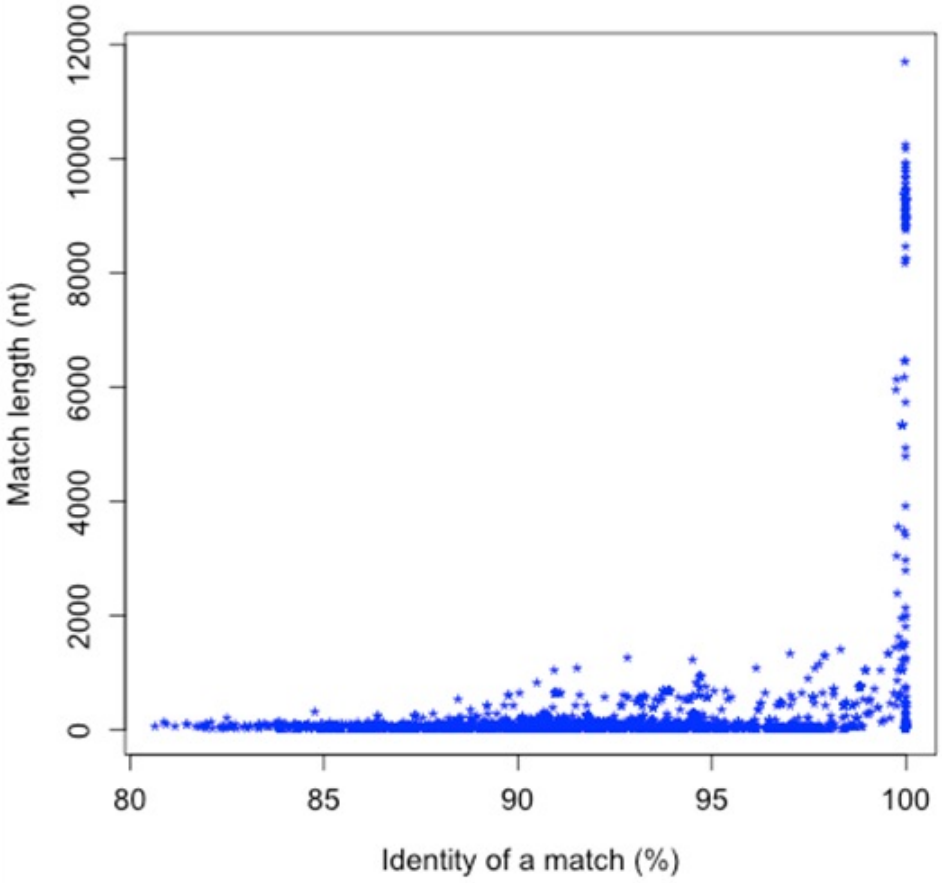


Figure 5.15: A scatter plot of match length and percent identity for the *var*-contig PA0036_VAR_1. Longer matches had the highest percent-identity (R^2=0.3, p-value $< 2.2x10^{-16}$)

In both cases (Figure 5.15 and Figure 5.16), the highest match length values were observed at high identity thresholds.

BLAST matches of ORFs

To minimize the effect of low complexity regions, such as introns, amino acid translations of *var* contigs (using the longest ORF with DBLα) were used to

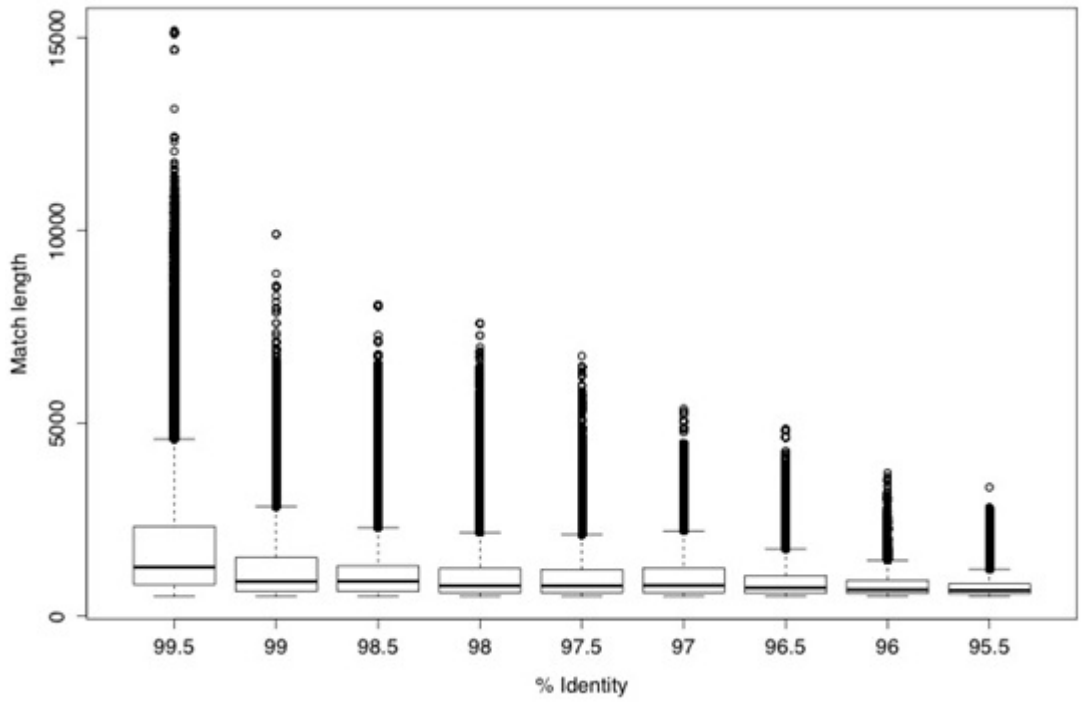


Figure 5.16: Box plots of match length (bp) for different cutoff values of percent identity using all *var*-contigs. Box limits represent median, first and third quantiles; whiskers represent the upper and lower bounds while outliers are shown by the dots.

generate a list of potential matches between *var*-contigs. Long and perfect amino acid matches, similar to those observed from Pmatch, were also identified using the BLAST search. Previous studies have identified three strain-transcending *var* genes: var1CSA, var2CSA and Type 3 *var* genes, that were found in clinical and culture-adapted isolates studied so far including 3D7 and IT (Kraemer et al., 2007). In the 3D7 genome, three *var* genes were identified as Type 3 *var* genes: PFA0015c (PF3D7_0100300), PFF0020c (PF3D7_0600400) and PFI1820w (PF3D7_0937600).

In order to test if the long-perfect matches were homologues of the known strain-transcending *var* genes, ORFs from *var* genes of the 3D7 genome were aligned to ORFs of the ∼50,000 *var*-contigs. As var2CSA (PFL0030c/PF3D7_1200600) does not have the DBLα tag, it was not included in the current list of *var*-contigs. We expected the Type 3 *var* genes and var1CSA (PFE1640w-ps/PF3D7_0533100) to have sequence homology with the highly conserved *var*-contigs. However, no match was observed at higher identity and length thresholds of 99% and 1,000 aa respectively. Lowering the match length cutoff to 50 aa identified matches between 13 *var*-contigs and two of the three Type 3 *var* genes (PFF0020c had a match with two *var*-contigs and PFI1820w to 11 *var*-contigs). The 13 *var*-contigs represented samples from The Gambia, Ghana, Mali, Kenya, Thailand and Cambodia, but they were not part of the long perfect matching *var*-contigs described earlier in this section.

The longest identical ORF matches of length 4,668 aa were observed between *var*-contigs of Thailand and Cambodia. Similarly, *var*-contigs from other countries including Kenya and The Gambia were also found to have perfect matches of up to ∼4,000 aa. The Markov Clustering Algorithm was able to detect distinct clusters of *var*-contigs based on a pairwise similarity measure derived from the BLAST matches of ORFs. A social network analysis was then used to visualise the population level structure of the global collection of *var*-contigs as described in the following section.

5.3.8 Clustering *var*-contigs and detecting population structure

There is no established method for the analysis of diversity and population structure in the *var* gene family. This is mainly due to high levels of recombination that prevent orthology /ancestry from being established. A social network analysis was used a better approach to deal with the complexity resulting from a highly polymorphic nature of *var* genes (Bull et al., 2008). Here results of a preliminary analysis of amino acid similarity networks were presented as an exemplar method of establishing structures in populations of *var* genes.

To simplify the analysis, samples with over 70 *var*-contigs were excluded, as they are likely to have multiple infections (Table 5.3). Populations from Burkina Faso and Bangladesh were also excluded, as they were represented by single samples (Table 5.3). In addition, two of the 102 Gambian samples (PA0106 and PA0107) were excluded, as they had a high ratio of ORFs to *var*-contigs due to a highly fragmented assembly. A total of 424 samples from 11 countries remained for subsequent analysis.

ID	Country	Region	#samples	#PasseedQC1 (DBL $\geq$ 30)	#PassedQC2 ($\leq$ 70)
PA	Gambia	WA	168	162	102
PF	Ghana	WA	122	108	43
PM	Mali	WA	32	32	13
PK	Burkina faso	WA	3	2	1
PT	Malawi	EA	55	43	13
PC	Kenya	EA	25	25	20
PE	Tanzania	EA	15	15	11
PR	Bangladesh	SEA	3	2	1
PD	Thailand	SEA	82	79	63
PH	Cambodia	SEA	191	153	138
PN	Papua New Guinea	SEA	7	7	5
PV	Vietnam	SEA	11	9	8
PP	Peru	SA	11	10	10
			725	**647**	**428**

Table 5.3: A list of samples used for initial assembly (Column 4), pmatch /BLAST analysis (Column 5) and social network analysis (Column 6). Burkina faso and Bangladesh were excluded from the final list as they were represented by a single sample. Two of the 102 Gambian samples were also excluded due to a poor quality assembly.

A total of 26,832 ORFs were obtained from *var*-contigs of the 424 samples. Using a minimum match length of 1,000 aa (identity cutoff of 99%), a total of 1,553 clusters were identified. The majority of clusters (∼68%) only contained 2 or 3 *var*-contigs. They were thus excluded from the network diagram in Figure 5.17. Conversely, the largest cluster (Figure 5.17. Cluster 1) had 141 *var*-contigs representing 127 samples and 10 populations (The Gambia=26, Ghana=14, Mali=4l, Vietnam=2, Thailand=30, Kenya=4, Malawi=2, Peru=4, Thailand=53 and PNG=2). Other clusters contained *var*-contigs that represented variable numbers of samples and populations.

The network diagram (Figure 5.17) represented a total of 6,985 nodes (i.e. ∼26% of the total ORFs in 424 samples) that overlap with other ORFs at a minimum percent identity of 99% and a match length of at least 1,000 aa. In addition to results from the clustering algorithm, the visual inspection revealed interesting identity matches as shown by the five clusters (Figure 5.17, 1-5). The second cluster has a single *var*-contig from Kenya (PC0016_VAR27) clustered with *var*-contigs from Thailand and Cambodia. Similarly, cluster 4 has single *var*-contigs from The Gambia, Malawi and Vietnam mixed with contigs from Thailand and Cambodia. Conversely, most of the *var*-contigs in cluster 3 were from The Gambia and Cambodia, while mixed with single *var*-contigs from Thailand, Ghana and Mali. Finally, in cluster 5, single *var*-contigs from Tanzania, Mali and Thailand were clustered with samples from Peru, Kenya and The Gambia. These observations highlight the widespread distribution of highly conserved *var*-contigs.

A higher degree of overlap between *var*-contigs is observed with the majority of clusters representing samples primarily from South East Asia. These samples also form some of the largest cluster sizes (connected components) compared to African samples. A larger proportion of *var*-contigs from African samples formed smaller clusters and were excluded during the filtering based on degree of a node (i.e. *number of connections*<4). This is an interesting observation as samples from Africa have much older *var* repertoires that have been exposed to natural forces such as mutation and recombination. As a result, a lower degree of similarity is expected in African samples compared to Thailand and Cambodia.

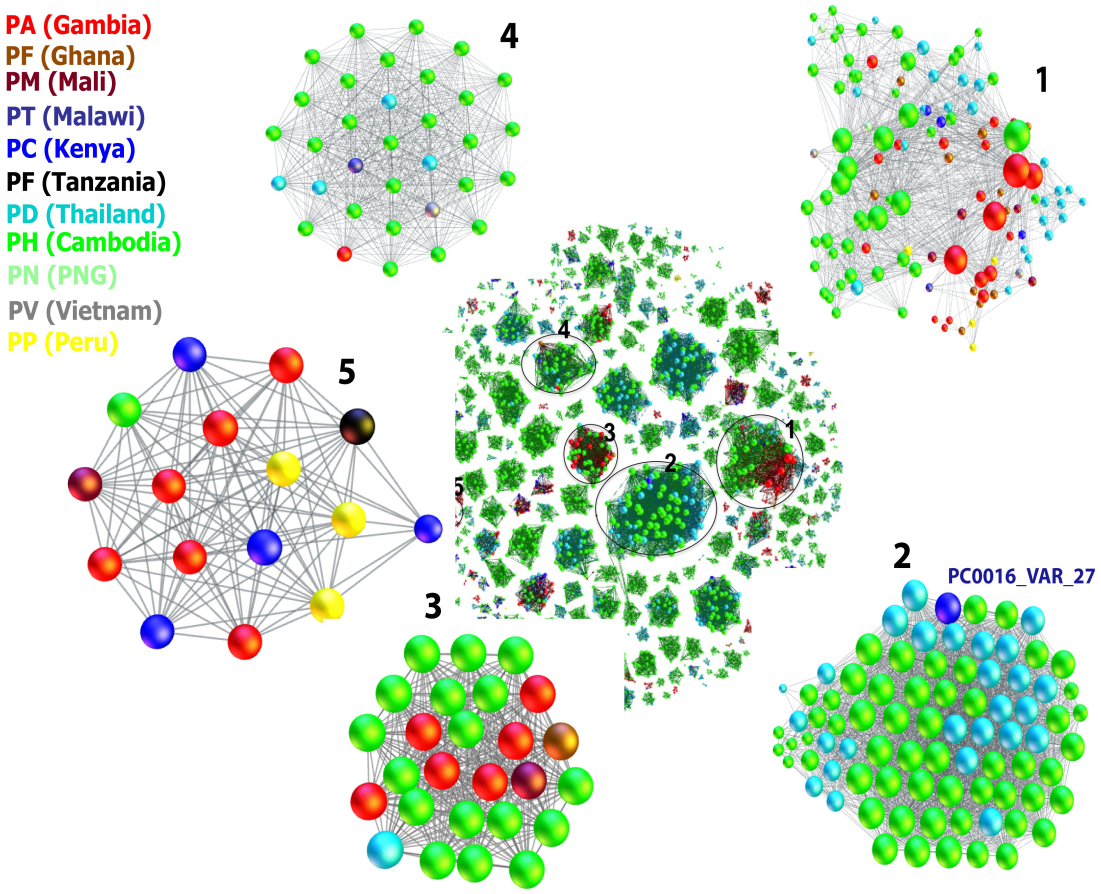


Figure 5.17: Amino acid identity network of *var*-contigs from 424 samples. The main figure shows *var*-contigs that share highly similar sequence blocks (above 1,000aa and a minimum identity of 99%). *Var*-contigs that have fewer than 4 connections with other contigs were excluded to simplify the graph. The number of connections of a node (*var*-contig) is represented by the size of each circle. A closer view of five representative clusters is also shown (1-5). Cluster 1 contained 141 ORFs*var*-contigs representing 127 samples and 10 populations (The Gambia=26, Ghana=14, Mali=4l, Vietnam=2, Thailand=30,Kenya=4, Malawi=2, Peru=4, Thailand=53, and PNG=2). Other clusters contained *var*-contigs that represented variable numbers of samples and populations (See text for details).

Country-specific clustering analysis was done on 138 Cambodian samples revealing sub-populations of *var*-contigs as shown in Figure 5.18. 42% of the total ORFs (n=8,065) were grouped into 629 clusters containing 2 to 55 *var*-contigs. Initially, such clusters may appear to be a result of clonal expansion events. However, there was a very strong positive correlation ($R^2 > 0.99$) between number of *var*-contigs and the number of unique samples in each cluster (Figure 5.19) suggesting a wider distribution of highly conserved *var*-contigs.

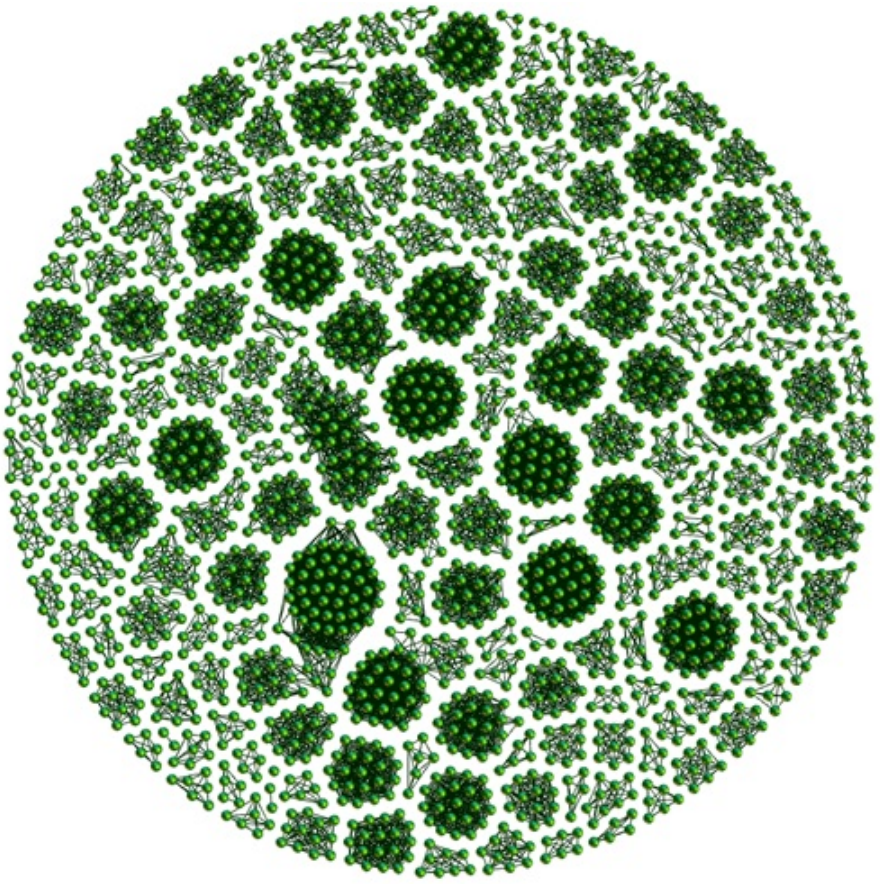


Figure 5.18: A network of *var*-contigs from Cambodian samples ($degree > 3$). A total of 3,380 *var*-contigs (nodes) were represented in this network (number of edges=15,598). Initially, the distinct sub-groups of *var*-contigs may seem due to a clonal expansion event. However, as shown in Figure 5.19, each cluster contains different samples suggesting the presence of long and unexpected conserved sequences in a large number of samples.

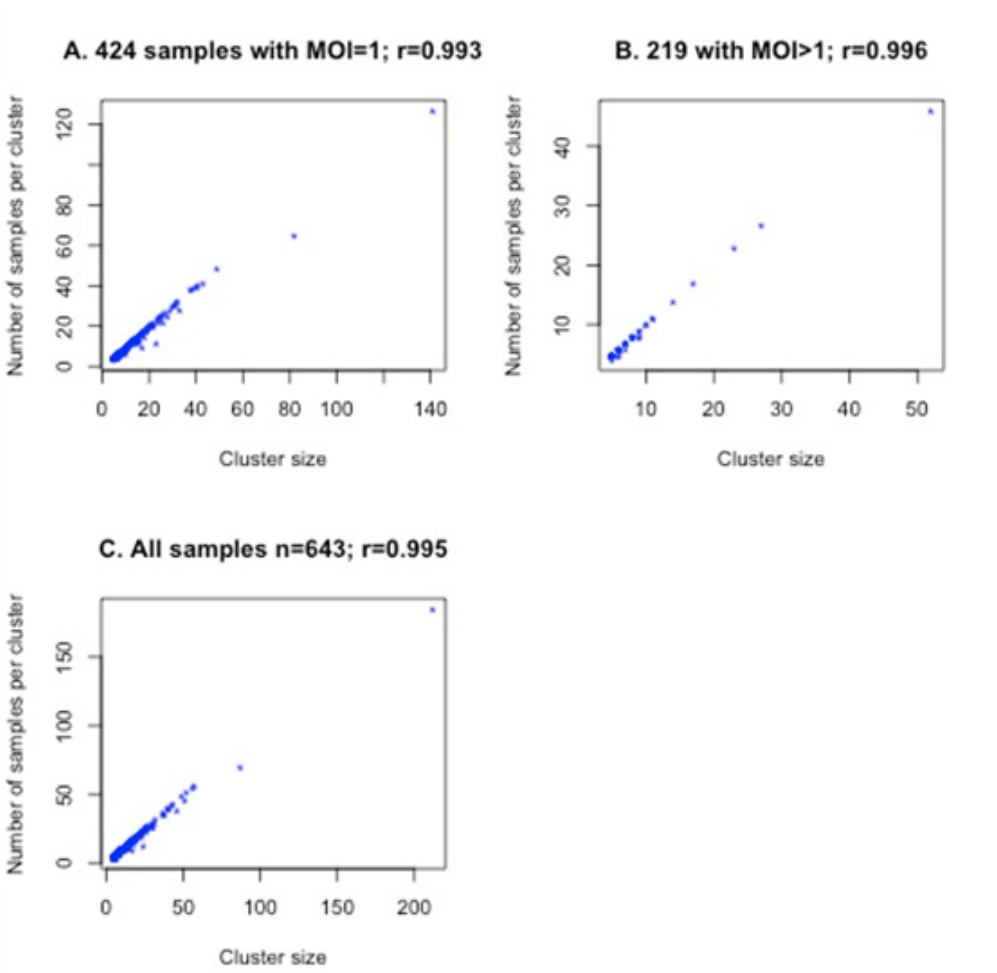


Figure 5.19: Scatter plots of cluster size vs the number of unique samples represented in each cluster. A very strong correlation is observed (as shown by R^2 values) between cluster size and number of distinct samples. **A)**. Scatter plot for samples with a minimum of 30 *var*-contigs and a maximum count of 70 *var*-contigs. These samples are considered to have a single infection with a multiplicity of infection (MOI) of 1. **B)** Scatter plot for samples that have above 70 *var*-contigs (MOI>1). **C)**. Scatter plot for all samples.

5.4 Discussion

This chapter presented assembly results of *var* genes from clinical samples. The following two conclusions summarise the results obtained and their implications.

Conclusion 1: *The results show the first full length assembly of var genes and the largest collection of var genes from clinical samples.*

It was possible to generate the largest collection of full length *var*-contigs using an iterative assembly approach developed to specifically assemble *var* genes. Although the assembly was initiated with a small number of motifs from three lab-adapted samples, it was able to generate high quality assembly of *var* genes on over 700 samples. The increase in the number of samples used to iteratively generate seed motifs is one reason for high quality assembly with as few as three iterations. The results show the first report of a targeted assembly of highly polymorphic gene families in general and of the *var* gene family in particular.

As expected, sequence quality and yield affect quality of the final assembly. This effect is likely to be pronounced in the iterative assembly approach compared to a simple *de novo* assembly of all reads due to the need to have enough reads that contain seed motifs to initiate the process. However, it could also be seen as an early quality check as samples with poor quality and low yield will not have enough reads to proceed with the assembly. Despite the continual improvement in sequencing yield and protocols developed to remove contaminants, variability in quality and yield are characteristics of sequences obtained from natural populations.

The expected number of *var* genes with the presence of a single genotype provided a simple measure of assembly quality during the initial stages of the assembly. Excluding samples with assembled *var*-contigs of below 30 was further justified by looking at the number of non-core reads and quality of the raw data.

As described in Chapter 2, the Illumina platform is prone to substitution errors at the beginning and ends of reads. Initially, we intended to incorporate

trimming of reads in the assembly workflow. However, assembly attempts by trimming reads at error-prone ends did not improve the assembly of *var* genes. Although Velvet was chosen to generate seed contigs, the assembly approach is modular such that a different assembly tool could easily be used if future tests justify the choice. Additional iterations and quality control steps could be included by simply restarting the assembly process from where it stopped (i.e. start from iteration 4).

In summary, the iterative assembly approach together with the quality control steps applied in this chapter were shown to be effective in producing high quality full length draft *var* genes.

Conclusion 2: *Unexpected cross-continental var-contigs were identified between samples of unrelated countries.*

Preliminarily analysis of similarity between *var*-contigs obtained from the *de novo* assembly of clinical samples revealed unexpected and long sequences of high similarity. Nucleotide and amino acid alignments as well as perfect-match searches confirmed the presence of continent-transcending *var*-contigs (CTVs).

We have established that these similarities were not because of DNA contamination and informatics issues (eg. misassemblies). Potential explanations for the existence of continent-transcending *var*-contigs in such diverse parasite populations and their implications are presented in the next chapter.

Chapter 6

Final discussion, conclusions and future directions

The aim of this thesis was to explore applications of second-generation sequencing to address the following three questions:

1. Can we assemble (reconstruct) *var* genes from short reads of the Illumina sequencing platforms?
2. What are the mechanisms used by human malaria parasites to generate extreme diversity in *var* genes?
3. What is the global diversity of *var* genes?

In this chapter I aim to explore to what extent the above three objectives have been achieved and the contributions made by this thesis.

6.1 Assembly of *var* genes

As I described in Chapters 1 and 5, previous studies have suggested the presence of extreme sequence diversity in *var* genes. However, these studies used a very small proportion of the *var* repertoire as they relied on the DBLα region, a short conserved domain that accounted for $\sim$5% of the *var* repertoire per isolate. This thesis proposes that a better understanding of the global diversity of *var* genes and the mechanisms used by parasites to generate diversity could be achieved by studying full-length *var* sequences of natural populations.

In 2007/8, the malaria programme at the Wellcome Trust Sanger Institute initiated the Plasmodium Genome Variation project with the aim of sequencing thousands of clinical samples taken directly from patients in endemic areas. As described in Chapter 2, despite developments in algorithms for alignment and *de novo* assembly of short reads, subtelomeric regions and highly polymorphic gene families such as *var* genes were excluded from analyses, as they were intractable. The first report of whole-genome re-sequencing of hundreds of parasite isolates (Manske et al., 2012) focused solely on the core genome, calling the final genotypes from $\sim$40% the genome.

The increase in the number of sequenced samples and limitations of currently available *de novo* and reference-guided assembly approaches, thus certainly meant that further research was necessary and thus facilitated the research detailed in this thesis. As shown in Chapter 2, initial attempts to use existing *de novo* and reference-guided assembly tools were not successful due to a number of factors associated with inherent sequence features (such as high A+T content, high polymorphism and repeated sequences) as well as technical artefacts in the sequencing process (such as sequencing errors, poor quality of sequence data and uneven read coverage). An iterative assembly approach was therefore developed in Chapter 3 to assemble *var* genes from short reads of clinical samples.

An initial concern at the onset of the project was that it might not be possible to assemble *var* genes due to limitations of read-lengths and insert sizes. It was also suggested that the problem of assembling *var* genes could be solved in time with improvements in sequencing technologies. At the beginning of my project,

Illumina's GA2 platforms were able to routinely generate 54 bp reads with test runs of 76 cycles (76 bp paired-reads). Over the three-year period, read lengths have increased to ∼100 bp, although fragment sizes remained at 200-300 bp on standard production-level sequencing libraries.

The iterative assembly approach developed in this thesis was thus a viable alternative to address limitations of existing short read sequencing platforms and assembly methods. Thousands of clinical samples are already sequenced using Illuman's platforms with a read length of below 100 bp. Although it is likely that future improvements in sequencing technologies may improve read length and fragment sizes, re-sequencing of these clinical samples would not be possible due to the limitations in the initial starting DNA. A primary reason for this limitation being that there is no leftover DNA for most samples. This project includes description of new approaches developed to address assembly of highly polymorphic gene families during the last 3-4 years.

As future developments in sequencing platforms that promise longer reads (eg. oxford nanopore: `http://www.nanoporetech.com` and PacBio: `http://www.pacificbiosciences.com`) become viable, assembly tools developed for long capillary reads may serve as alternatives to the iterative assembly approaches described in this thesis. However, the time-scale of their commercial availability is not yet known at the time of writing.

Contributions

The two main contributions of the assembly approach described in this thesis are:

1. The introduction of a targeted-assembly method for gene families using conserved amino acid motifs and

2. The use of an iterative-assembly strategy that combines de Bruijn graph and overlap-layout-consensus (OLC)-based assembly approaches.

The first part takes advantage of the higher degree of conservation in amino acid sequences of polymorphic gene families compared to their DNA sequences. One previous approach that used amino acid sequences (Salzberg et al., 2008) relied on the presence of full-length gene models that were used to guide

the assembly process. However, this method was only tested on bacterial genomes. In addition, its application to highly polymorphic *var* genes was not practical due to the difficulty of obtaining closely related genomes that could be used to assemble *var* genes from clinical samples. The assembly approach described here is different as it only uses conserved regions instead of full-length sequences and these regions are only used to collect reads rather than as a template for the assembly process.

The second part combined advantages of the de Bruijn graph and OLC-based approaches to assemble *var* genes from short reads. Until recently, de Bruijn graph-based assemblers were the only tools available to efficiently handle the large sequence data generated by high-throughput second generation sequencing platforms. The Velvet assembler was thus used to generate seed contigs. However, the iterative extension steps use the OLC principle as contig-ends were gradually extended based on the overlap between the contig-ends and newly generated contigs. The process first identifies read-pairs where one or both reads align to contig-ends. These reads were then assembled to generate new contigs that could extend contig-ends or close a gap between two contigs as determined by the overlap step.

Additional applications of the assembly approach developed in this thesis include assembly of other multigene families in *P. falciparum* (eg. *rif* and *strvor* genes) and other pathogens (eg. VSG genes in Trypanosomes). Future work includes evaluating assembly and scaffolding methods that have become available recently (eg. SGA (Simpson and Durbin, 2012) and Cortex (Iqbal et al., 2012)).

6.2 *Var* gene diversity via ectopic gene conversion

Results from sequence analysis of a genetic cross (Chapter 2) confirmed ectopic gene conversion as a mechanism for *var* gene diversity. Recombinant genes were identified in two of the five progeny. The genes involved were located on Chromosomes 1 and 2 of the 3D7 parent and were of the group Type A. In addition, analysis of *var*-contigs using nucleotide and amino acid identity matches revealed that highly conserved *var*-contigs tend to be within the same group. These two observations provide additional evidence for the presence of recombination hierarchies in *var* genes. The quality of the raw data used (as described in Chapter 4) was not optimal. This presented a limitation as the read length and insert size were too short for analysis of recombination and gene conversion in *var* genes based on *de novo* assembly.

6.3 Global diversity of *var* genes

Two major contributions have been made towards understanding the global diversity of *var* genes:

1. The first assembly of full-length *var*-contigs.
2. Discovery of unexpected and long identical *var*-contigs.

6.3.1 The first full-length assembly of *var* genes from clinical samples

The first full-length assembly of *var* genes from clinical isolates was presented in this thesis (Chapter 5). Assembly results from a large number of clinical samples ($\sim$800 isolates) were shown to have a higher repertoire-completeness (i.e. the number of contigs identified as *var* genes was close to the expected number of *var* genes), and contiguity (i.e. contig N50 size, largest contig size and ORF sizes were comparable with the expected values from previously completed genomes such as 3D7). Such availability of full-length *var* genes is a major progress towards understanding the population structure and diversity of *var*

genes in natural populations. One aim of sequencing large numbers of parasites from clinical samples is to determine changes in population structure and detect regions of the genome under selection as a result of external pressures such as drugs and vaccines. Current methods of analysis that use SNPs obtained from less polymorphic coding regions of the genome (Manske et al., 2012) do not take the rapidly changing subtelomeres and *var* genes into account. Investigating the diversity and population structure of *var* genes is therefore an immediate next step to my thesis project.

One challenge associated with assembly of *var* genes in clinical samples was ensuring that assembled contigs had an acceptable quality. Measuring assembly quality is not straightforward due to the various parameters that need to be considered. For example, recent efforts of benchmarking assembly tools such as Asemblathon 1 (Earl et al., 2011) have illustrated the need to consider multiple metrics in measuring assembly quality. In this thesis, in addition to the four commonly used metrics (i.e. number of contigs or scaffolds, sum of contigs, N50 and largest contig sizes), the size of ORFs was used to check the quality of *var*-contigs. Frame-shifts are often signs of a misassembled contig while dealing with protein coding genes. The lengths of ORFs from *var*-contigs in clinical samples were therefore compared with ORF sizes of *var*-genes from the 3D7 and IT genomes.

Although the new approach generated high quality contigs, genes that contain duplicated regions larger than the fragment size of the library will remain difficult to assemble in one piece. Developments towards large insert libraries and longer read lengths will thus improve assembly of such *var* genes.

6.3.2 Discovery of unexpected and highly conserved *var*-contigs

Preliminarily analysis of *var*-contigs based on nucleotide and amino acid similarities (Chapter 5) revealed distinct clusters of highly conserved *var*-contigs within and between populations. The validity of these contigs was confirmed by looking at the sizes of ORFs (Figure 5.7) and aligning short reads back to *var*-contigs (Figure 5.12). Clusters that contain *var*-contigs from as many as 6 countries were identified. These observations were surprising as the *var*-contigs

had percent-identities of up to 100% over a match length of ∼5-10 kb. Previous studies may have missed such highly conserved long sequences due to the focus on the DBLα region. Here, two potential reasons are explored as explanations for the presence of highly conserved *var*-contigs within and between populations.

Recent transmission via a traveller

Initially, we hypothesized that continent-transcending *var*-contigs (CTVs) may have been a result of recent transmission events, perhaps carried from one location to another by a traveller. It is important to note that the conservation in these *var*-contigs does not extend to the rest of the *var* repertoire between any two samples. In addition, a single CTV could be found in as many as 57 samples from six populations. Parasites undergo a number of cell division steps within both the human and mosquito hosts that may lead to mutations, crossing over and gene conversion events. It is thus likely for a mutation to accumulate over time unless the transmission was a very recent event. Therefore, for a CTV to have been recently transferred to a new location via a traveller and to be found in the isolates sampled by our team, we assume that parasites that carry a CTV are circulating at a reasonably high frequency in the population. Recently transmitted or acquired alleles are expected to have high frequencies when they are associated with advantageous mutations such as drug resistant alleles. As parasites that carry the advantageous mutation spread across populations, neutral regions in the vicinity of the beneficial allele also rise to prominence. This phenomenon, also known as hitchhiking, is reported in a number of pathogens including *P. falciparum* (Walliker, 2005).

A closer look at known drug resistance genes (and regions) in *P. falciparum* revealed two genes that are in close proximity to *var* genes of the central cluster on chromosomes 4 and 7. The genes *pfcrt* (*P. falciparum* chloroquine resistance reporter gene) and *dhfr* (dihydrofolate reductase) were identified to confer resistance to quinolone-based (eg. Chloroquine) and antifolate antimalarial drugs respectively. The distance between *var* genes and the two drug resistance genes were ∼105 kb and ∼134 kb on Chromosomes 7 and 4 respectively. It was thus likely that central *var* genes on chromosomes 4 and 7 may have been

linked with drug resistance genes and spread around the world. Fowler and colleagues (Fowler et al., 2006) have reported a similar observation suggesting a potential linkage between central *var* clusters and drug resistance genes (*pfcrt* and *dhfr*). However, other drug resistance genes such as dhps (On chromosome 8) and mdr1 (On Chromosome 5) were ∼400 kb to ∼900 kb away from *var* genes. A recently identified drug resistance region on chromosome 13 (Cheeseman et al., 2012) was also ∼1 Mb away from *var* genes on either subtelomere. It was therefore less likely for subtelomeric *var* genes and central *var* clusters on chromosomes 6, 8 and 12 to have a strong linkage with drug resistance genes.

In summary, a higher degree of conservation between central *var* genes of chromosomes 4 and 7 may be expected as they are in close proximity to drug resistance genes. Ideally, genes that have a higher degree of linkage could be detected by scanning genomic regions for reduced levels of heterozygosity or formation of highly segregating haplotypes across multiple populations. However, the data presented in chapter 5 do not extend beyond the *var* gene regions. Further confirmation of this hypothesis thus requires analysing additional information on the core regions of the genomes (eg. read coverage over central *var* genes and look for read-pairs that connect to the rest of the chromosome on the edge of the *var* arrays) and the use of meta-data such as the dates of isolation.

Functional importance

The second explanation assumes a functional relevance for highly conserved *var*-contigs. In a recent study using full length genes from the seven genomes (Rask et al., 2010), Buckee and Recker (Buckee and Recker, 2012) noted that rare *Pf*EMP1 domains, primarily of the group Type A, were highly conserved and longer than genes from other groups. The high sequence conservation in type A genes is believed to be due to binding related functional constraints. The most recent reports associating type A genes with severe and cerebral malaria have shown evidence of their involvement in binding to brain endothelium cells (Avril et al., 2012; Claessens et al., 2012; Lavstsen et al., 2012).

The results presented in chapter 5 showed that the majority of highly conserved continent-transcending *var*-contigs were of the groups 1 to 3 according to the grouping by Bull and colleagues (Bull et al., 2007). As groups 1 to 3

correspond with type A *var* genes, it is likely that the high similarity between these genes is maintained due to their functional importance, specifically their role in parasite virulence. It is however important to note that the definition and measurement of virulence is not straightforward as it refers to the result of a complex interactions between parasites and their hosts. In this context, we define virulence as the parasites ability to cause damage with the aim of reducing the fitness of the host (Schmid-Hempel, 2011).

Understanding the underlying drive towards an increased level of virulence is also challenging as a highly virulent parasite may kill the host resulting in no adaptive advantage for the parasite. It was therefore suggested that virulence could be a simple side effect of host-parasite interactions whereby there was not enough time for adaptive evolution to take place (Schmid-Hempel, 2011). Alternatively, virulence may also emerge as a result of short-term adaptations (also known as short-sighted evolution) where parasites gain advantage in the short term, for example by colonizing specific tissues as in the case of cerebral malaria. Such increased virulence leads to severe forms of the disease leading to the death of the host and may not be advantageous in the long term. As the majority of CTVs are associated with severe malaria, their presence may be as a result of short-sighted evolution. However, the detailed mechanisms of their emergence, the speed of their spread and maintenance of such genes across populations requires further investigation.

If these were bacteria, the phenomena would be explainable by lateral gene transfer as bacteria have specialized mechanisms to facilitate capture of foreign DNA. However, as the events reported here must involve recombination, it is harder to come up with a plausible explanation for the spread across a global population. One potential explanation for how virulent genes may be maintained in a given population may be via recurrent mutations or unusual gene conversions. To explore this further, it is important to note that any positive selective pressure on parasites should be associated with an advantageous trait. Initially, it may appear that highly virulent parasites are at a disadvantage as they are likely to kill their host. Although this may certainly be the case, it is however likely that such virulent parasites may also multiply rapidly, hence balancing the negative effects of virulence. It is worth noting that reproduction

and transmission are the two major objectives of infective parasites. In *P. falciparum*, gametocytes are produced in the asexual stages of the life cycle where they remain in the blood stream to be taken by mosquitoes for the sexual stages of the parasite's development. Despite the potential damage to the host, higher degree of virulence may have a direct impact on gametogenesis as increased virulence (and parasitaemia) could lead to higher rates of production in male and female gametocytes. Highly virulent parasites may thus have an increased transmission ability.

In summary, highly similar continent-transcending *var*-contigs of the central cluster on Chromosomes 4 and 7 may have been transmitted via a traveller and raised to high frequency due to drug resistance genes. However, this did not account for the higher degree of conservation in other *var* genes. The strong association of most conserved *var*-contigs with Type A *var* genes suggests that their cross-continental conservation of telomeric *var* genes may be due to the role they play in causing severe malaria infections which may have a fitness advantage by increasing transmission of parasites via enhancing gametogenesis.

Because members of the *var* gene family are prone to higher rates of recombination, selective sweeps and population structure could be visible by comparing the full length *var* repertoire within and between samples of populations. The results presented in Chapter 5 suggest the potential of studying full-length *var* genes as potential markers for more recent evolutionary changes. It is also raises the question whether there really is 'extreme diversity' in *var* genes globally. Ongoing work will focus on determining how diverse or 'extremely diverse' *var* genes are by continuing the comparative analysis on full-length *var*-contigs generated in Chapter 5.

Our interest in understanding global *var* repertories from field-isolated parasite samples is motivated by the potential of associating expression profiles of *Pf*EMP1 variants with disease phenotypes. The unexpected and long continent-transcending *var*-contigs were found to correspond with Type A *var* genes. These results are very interesting as type A *var* genes are associated with severe malaria infections. Future analyses could thus focus on assembling full-length *var* transcripts from patients with different levels of disease severity and looking at the number of gametocytes.

6.4 Future directions

On going and future work towards addressing the three main objectives are outlined below.

- We are currently working on further quality control measures on var gene assemblies include PCR-confirmation of *var*-contigs. In addition, investigating the diversity and population structure of *var* genes is also on going as an immediate next step to my thesis project.

- I will explore additional applications of the assembly approach developed in this thesis including assembly of other multigene families in *P. falciparum* (eg. *rif* and *strvor* genes) and other pathogens (eg. VSG genes in Trypanosomes).

- Our interest in understanding global var repertories from field-isolated parasite samples is motivated by the potential of associating expression profiles of *Pf*EMP1 variants with disease phenotypes. The unexpected and long continent-transcending *var*-contigs were found to correspond with Type A *var* genes. These results are very interesting as type A *var* genes are associated with severe malaria infections. Future analyses will focus on assembling full-length *var* transcripts from patients with different levels of disease severity and looking at additional phenotypic information.

6.5 Publications and press releases

- Some of the tools used in the assembly and quality control of contigs were described in two software packages: ABACAS (Assefa et al., 2009, Bioinformatics) and PAGIT (Swain et al., 2012, Nature Protocols).

- The method described in Chapters 3 and the results from Chapter 5 were presented at the 8^{th} international BioMalPar Conference (14-16^{th} May 2012, Heidelberg, Germany) where I was awarded the 'best oral presentation' prize.

- The work detailed in Chapters 4 and 5 are currently under preparation for separate publications.
- My PhD project was also featured by the Wellcome Trust Sanger Institute's press office (`http://www.sanger.ac.uk/about/press/features/assefa.html`)

Appendices

Appendix A:
Storage space estimates for an iterative assembly of *var* genes in *P. falciparum*

ID (from Figure A-1)	Process	Disc Space per process (Gb)	Temporary storage (Gb)	Sum (Gb)
1	Get non-core reads	1.5	0	1.5
2	Get var-reads	0.4	0.4	
3	Assembly	15	15	
4	Scaffolding	15	15	
5	IMAGE	12	12	
6	Scaffolding	15	15	
7	IMAGE	12	12	
	Output files (Permanent storage)			0.5 2

Table A-1: Storage (Disc) space estimates for one lane of sequence data per iteration. Disc space required for each step of the seven processes are shown for one iteration. Although temporary files were deleted after each process (as shown in column 4), processing ∼800 samples (Chapter 5) required up to ∼12 Tb of temporary storage space at a given time.

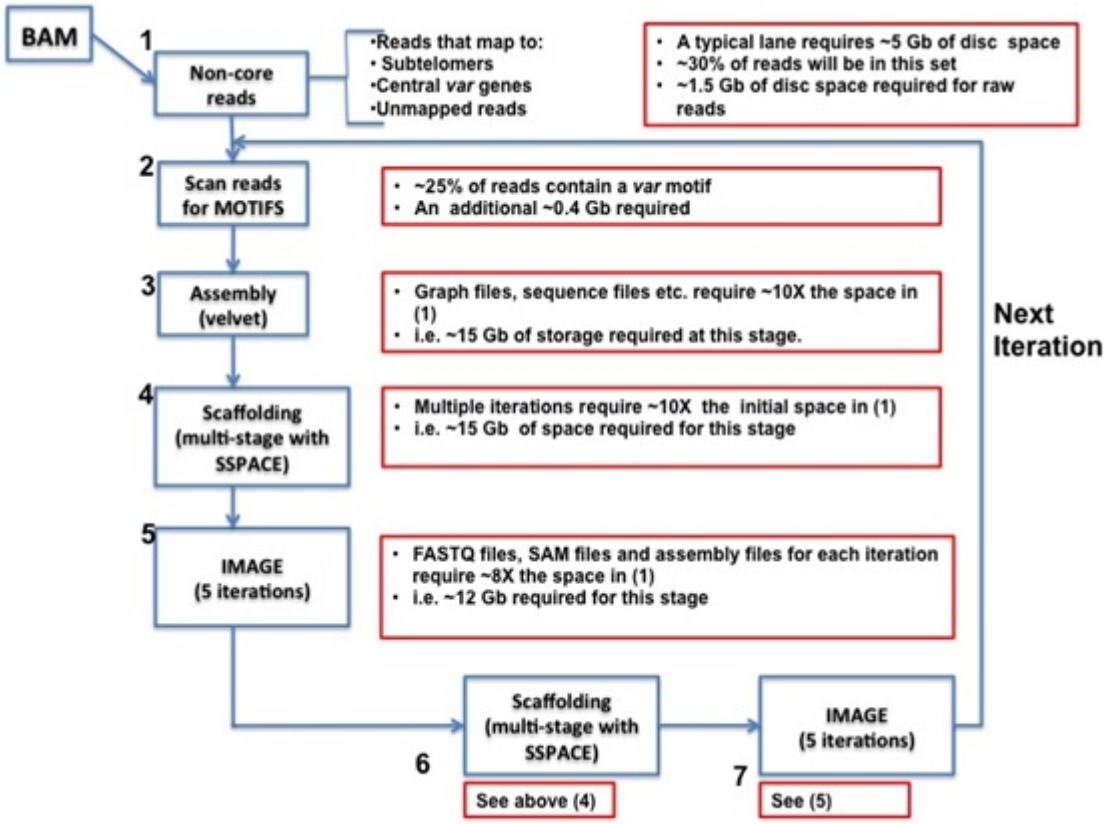


Figure A-1: Storage (Disc space) estimates for one iteration of *var* assembly per sample. Temporaty and permanent storage required for a single iteration of assembly is shown (red boxes) for each process. Additional information on the temporary and permanent storage estimates is shown in Table A-1.

Appendix B:
Quality of raw data and its effect on assembly

Quality and insert size plots were obtained from the Sanger Institute's raw data quality control archives.

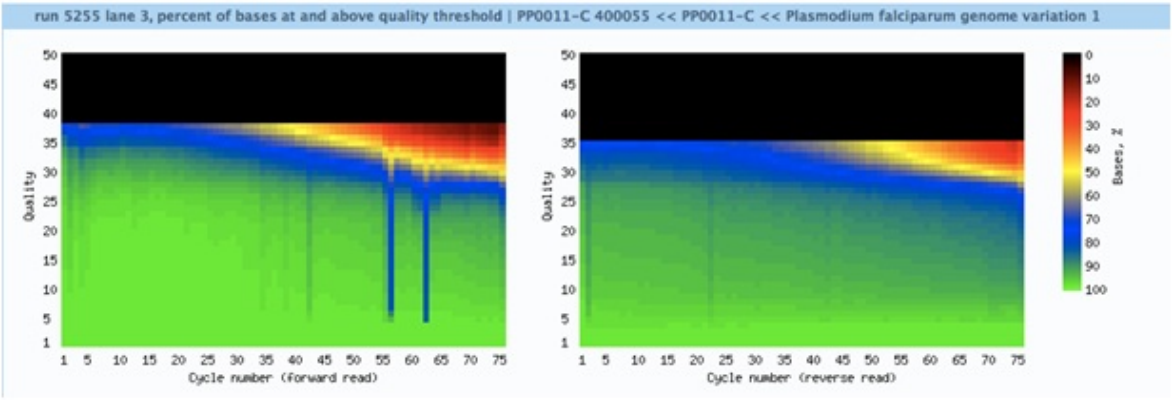


Figure B-1: Base quality plot for the forward (left panel) and reverse (right panel) reads of sample PP0011. Assembly results for PP0011 (Chapter 3, 50 clinical samples) were affected by the drop in quality on cycles 55 and 73. The number of *var*-contigs was 26 for PP0011 (Chapter 3, section 3.3.3). The quality of each base (y-axis) is shown for the 76 cycles (x-axis). The colors represent the percentage of bases that have a quality value shown on the y-axis.

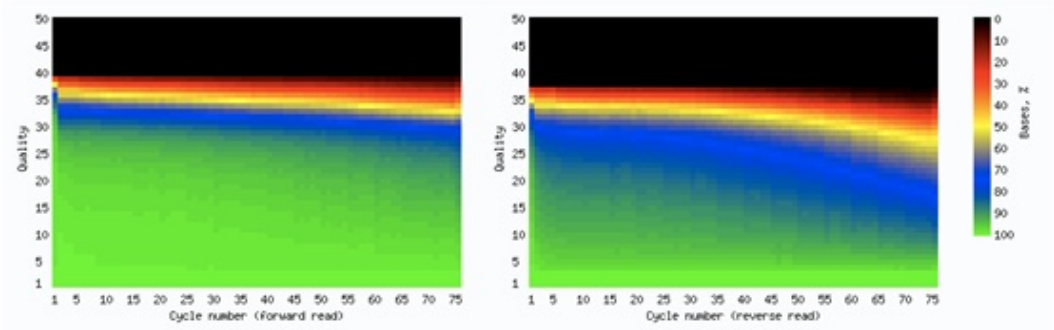


Figure B-2: Base quality plot for the HB3 genome used in Chapter 3. The second read (right panel) shows a decay in quality from cycle ∼45 onwards. See Figure B-1 for description. The HB3 genome had a significantly higher number of non-core reads due to the larger proportion of reads that did not align to the 3D7 genome (∼33%). The decrease in quality on the second read has affected the number of reads that aligned to the reference genome (Chapter 3, section 3.3.2.1).

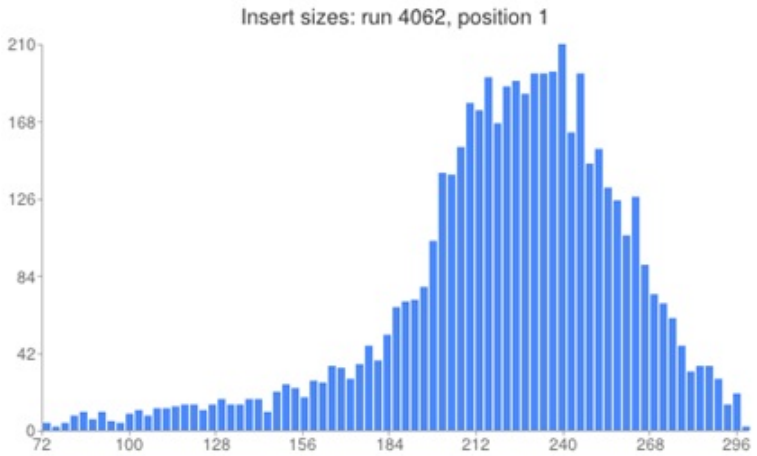


Figure B-3: Hb3 insert sizes (normal sizes)

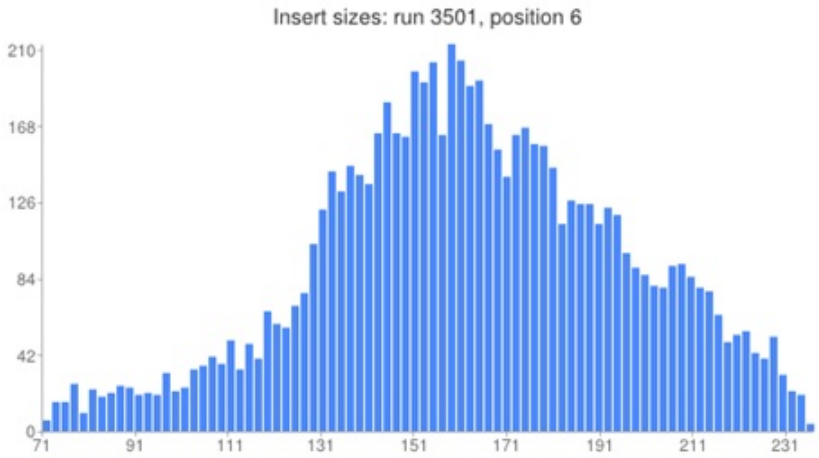


Figure B-4: A distribution of insert sizes for the 3D7 genome used in Chapters 2 and 3. The poor quality assembly for *var* genes of the 3D7 genome (Chapter 3, testing the iterative assembly approach on culture-adapted samples) was partly due to the shorter than the expected insert sizes. As shown here, the actual insert sizes were ∼160 bp, while the expected sizes were 200-300 bp.

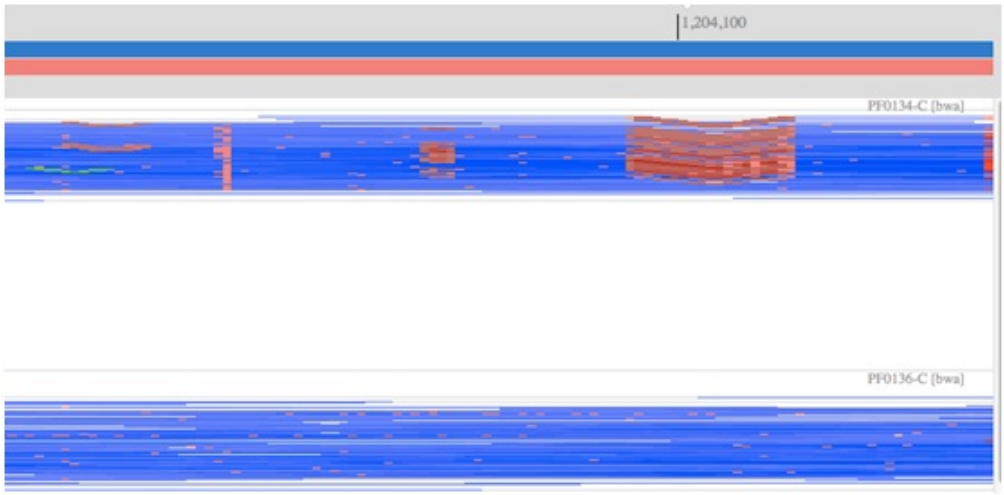


Figure B-5: Number of *var*-contigs is indicative of multiplicity of infection: Samples PF0134-C and PF0136-C had 262 and 65 *var*-contigs respectively. This figure shows a LookSeq view of SNPs (shown in red) for the two samples (Top panel for PF0134-C and bottom panel for PF0136-C). Read coverage and SNPs over the MSP1 gene show segregating haplotypes for sample PC0134-C suggesting multiple infections.

Appendix C:
Availability of Software developed to assemble *var* genes

Software developed in this thesis, additional scripts and documentation are available from:

`https://sourceforge.net/projects/varassembly/`

References

Abnizova, Irina et al. (Apr. 2012). "Analysis of Context-Dependent Errors for Illumina Sequencing". In: *Genome Science and Technology* 10.02, p. 1241005 (cit. on pp. 26, 58).

Al-Khedery, Basima and David R Allred (Dec. 2005). "Antigenic variation in Babesia bovis occurs through segmental gene conversion of the ves multigene family, within a bidirectional locus of active transcription". In: *Molecular microbiology* 59.2, pp. 402–414 (cit. on p. 93).

Altschul, Stephen F et al. (Oct. 1990). "Basic local alignment search tool". In: *J. Mol. Biol* 215.3, pp. 403–410 (cit. on pp. 135, 152).

Anderson, Tim (Apr. 2009). "Mapping the spread of malaria drug resistance". In: *PLoS medicine* 6.4, e1000054 (cit. on p. 2).

Andrej Trampuz, Matjaz Jereb Igor Muzlovic Rajesh M Prabhu (2003). "Clinical review: Severe malaria". In: *Critical Care* 7.4, p. 315 (cit. on p. 4).

Ariey, Frédéric et al. (July 2001). "Association of Severe Malaria with a Specific Plasmodium falciparumGenotype in French Guiana". In: *The Journal of infectious diseases* 184.2, pp. 237–241 (cit. on pp. 15, 128).

Assefa, Samuel et al. (2009). "ABACAS: algorithm-based automatic contiguation of assembled sequences." In: *Bioinformatics (Oxford, England)* 25, pp. 1968–1969 (cit. on pp. 33, 95, 102, 177).

Avril, Marion et al. (May 2012). "A restricted subset of var genes mediates adherence of Plasmodium falciparum-infected erythrocytes to brain endothelial cells". In: *PNAS* 109.26, E1782–E1790 (cit. on pp. 15, 129, 174).

Barry, AE et al (2007). "Population genomics of the immune evasion (var) genes of Plasmodium falciparum," in: *PLoS Pathogens* 3.3, e34 (cit. on pp. 11, 16, 17).

Barry, Alyssa E et al. (Mar. 2007). "Population Genomics of the Immune Evasion (var) Genes of Plasmodium falciparum". In: *PLoS Pathogens* 3.3, e34 (cit. on p. 135).

Baruch, D I et al. (July 1995). "Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes". In: *Cell* 82.1, pp. 77–87 (cit. on p. 6).

Batzoglou, Serafim et al. (Jan. 2002). "ARACHNE: A Whole-Genome Shotgun Assembler". In: *Genome Research* 12.1, pp. 177–189 (cit. on p. 23).

Baumeister, Stefan et al. (Nov. 2009). "The malaria parasite Plasmodium falciparum: cell biological peculiarities and nutritional consequences". In: *Protoplasma* 240.1-4, pp. 3–12 (cit. on p. 3).

Bentley, David R (Dec. 2006). "Whole-genome re-sequencing". In: *Current Opinion in Genetics & Development* 16.6, pp. 545–552 (cit. on p. 20).

Bentley, David R et al. (Nov. 2008). "Accurate whole human genome sequencing using reversible terminator chemistry". In: *Nature* 456.7218, pp. 53–59 (cit. on p. 20).

Biggs, B A (July 1990). "Knob-independent cytoadherence of Plasmodium falciparum to the leukocyte differentiation antigen CD36". In: *Journal of Experimental Medicine* 171.6, pp. 1883–1892 (cit. on p. 6).

Boetzer, M et al. (Feb. 2011). "Scaffolding pre-assembled contigs using SSPACE". In: *Bioinformatics (Oxford, England)* 27.4, pp. 578–579 (cit. on pp. 23, 27, 68).

Brayton, Kelly A et al. (2002). "Antigenic variation of Anaplasma marginale msp2 occurs by combinatorial gene conversion". In: *Molecular microbiology* 43, pp. 1151–1159 (cit. on p. 93).

Buckee, Caroline O and Mario Recker (Apr. 2012). "Evolution of the Multi-Domain Structures of Virulence Genes in the Human Malaria Parasite, Plasmodium falciparum". In: *PLoS Computational Biology* 8.4, e1002451 (cit. on p. 174).

Bull, P C et al. (Mar. 1998). "Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria". In: *Nature medicine* 4.3, pp. 358–360 (cit. on p. 15).

Bull, P C et al. (July 2000). "Plasmodium falciparum–Infected Erythrocytes: Agglutination by Diverse Kenyan Plasma Is Associated with Severe Disease

and Young Host Age". In: *Journal of Infectious Diseases* 182.1, pp. 252–259 (cit. on p. 15).

Bull, Peter C et al. (Nov. 2005). "Plasmodium falciparum variant surface antigen expression patterns during malaria". In: *PLoS Pathogens* 1.3, e26 (cit. on pp. 11, 15, 128, 152).

Bull, Peter C et al. (July 2007). "An approach to classifying sequence tags sampled from Plasmodium falciparum var genes". In: *The International Journal of Biochemistry & Cell Biology* 154.1, pp. 98–102 (cit. on pp. 14, 60, 79, 81, 132, 142, 147, 148, 152, 174).

Bull, Peter C et al. (June 2008). "Plasmodium falciparum antigenic variation. Mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks". In: *Molecular microbiology* 68.6, pp. 1519–1534 (cit. on pp. 16, 17, 89, 126, 135, 152, 159).

Butler, J et al (May 2008). "ALLPATHS: de novo assembly of whole-genome shotgun microreads". In: *Genome Research* 18.5, pp. 810–820 (cit. on p. 26).

Calderwood, M S et al. (June 2003). "Plasmodium falciparum var Genes Are Regulated by Two Regions with Separate Promoters, One Upstream of the Coding Region and a Second within the Intron". In: *Journal of Biological Chemistry* 278.36, pp. 34125–34132 (cit. on p. 9).

Carver, T J et al. (Aug. 2005). "ACT: the Artemis comparison tool". In: *Bioinformatics (Oxford, England)* 21.16, pp. 3422–3423 (cit. on pp. 34, 97).

Carver, Tim et al. (2010). "BamView: viewing mapped read alignment data in the context of the reference sequence." In: *Bioinformatics* 26, pp. 676–677 (cit. on pp. 36, 97, 102).

Chaisson MJ, Pevzner PA (2008). "Short read fragment assembly of bacterial genomes." In: *Genome Res.* 18.2, p. 324 (cit. on pp. 24, 26).

Cham, G K K et al. (Oct. 2010). "Hierarchical, Domain Type-Specific Acquisition of Antibodies to Plasmodium falciparum Erythrocyte Membrane Protein 1 in Tanzanian Children". In: *Infection and immunity* 78.11, pp. 4653–4659 (cit. on pp. 15, 128).

Cheeseman, Ian H et al. (Apr. 2012). "A Major Genome Region Underlying Artemisinin Resistance in Malaria". In: *Science* 336.6077, pp. 79–82 (cit. on pp. 2, 174).

Chen, Donald S et al. (Feb. 2011). "A Molecular Epidemiological Study of var Gene Diversity to Characterize the Reservoir of Plasmodium falciparum in Humans in Africa". In: *PLoS ONE* 6.2, e16629 (cit. on pp. 11, 16).

Chen, Jian-Min et al. (2001). "Gene Conversion in Evolution and Disease". In: *Life Sciences* (cit. on p. 93).

Chen, Jian-Min et al. (Sept. 2007). "Gene conversion: mechanisms, evolution and human disease". In: *Nature Reviews Genetics* 8.10, pp. 762–775 (cit. on pp. 92, 93).

Chen, Qijun (2007). "The naturally acquired immunity in severe malaria and its implication for a PfEMP-1 based vaccine." In: *Microbes and infection / Institut Pasteur* 9, pp. 777–783 (cit. on pp. 15, 16).

Cheng, Qin et al. (Nov. 1998). "stevor and rif are Plasmodium falciparum multicopy gene families which potentially encode variant antigens". In: *Molecular and Biochemical Parasitology* 97.1-2, pp. 161–176 (cit. on p. 3).

Chookajorn, T et al. (Jan. 2007). "Epigenetic memory at malaria virulence genes". In: *PNAS* 104.3, pp. 899–902 (cit. on p. 9).

Chookajorn, Thanat et al. (Oct. 2008). "Mutually exclusive var gene expression in the malaria parasite: multiple layers of regulation". In: *Trends in parasitology* 24.10, pp. 455–461 (cit. on p. 8).

Claessens, Antoine et al. (May 2012). "A subset of group A-like var genes encodes the malaria parasite ligands for binding to human brain endothelial cells". In: *PNAS* 109.26, E1772–E1781 (cit. on pp. 15, 129, 174).

Compeau, Phillip E C et al. (Nov. 2011). "How to apply de Bruijn graphs to genome assembly". In: *Nature* 29.11, pp. 987–991 (cit. on p. 25).

Conway, D J et al (1999). "High recombination rate in natural populations of ..." In: *Proc Natl Acad Sci U S A.* 96.8, pp. 4506–4511 (cit. on p. 16).

Dayarian, Adel et al. (2010). "SOPRA: Scaffolding algorithm for paired reads via statistical optimization". In: *BMC Bioinformatics* 11.1, p. 345 (cit. on p. 23).

Deitsch, Kirk W et al. (1999). "Intra-cluster recombination and var transcription switches in the antigenic variation of Plasmodium falciparum". In: *Molecular and Biochemical Parasitology* 101, pp. 107–116 (cit. on p. 126).

Dohm, Juliane C et al. (Oct. 2007). "SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing". In: *Genome Research* 17.11, pp. 1697–6435207 (cit. on p. 26).

Dzikowski, Ron et al. (Aug. 2007). "Mechanisms underlying mutually exclusive expression of virulence genes by malaria parasites". In: *Nature* 8.10, pp. 959–965 (cit. on p. 9).

Earl, Dent et al. (Feb. 2011). "Assemblathon 1: A competitive assessment of de novo short read assembly methods". In: *Genome Research* 21, pp. 2224–2241 (cit. on p. 172).

Enderes, C et al. (Oct. 2011). "Var Gene Promoter Activation in Clonal Plasmodium falciparum Isolates Follows a Hierarchy and Suggests a Conserved Switching Program that Is Independent of Genetic Background". In: *Journal of Infectious Diseases* 204.10, pp. 1620–1631 (cit. on p. 9).

Erlich, Yaniv et al. (July 2008). "Alta-Cyclic: a self-optimizing base caller for next-generation sequencing". In: *Nature methods* 5.8, pp. 679–682 (cit. on p. 21).

Fairhurst, Rick M et al. (Aug. 2012). "Abnormal PfEMP1/knob display on Plasmodium falciparum-infected erythrocytes containing hemoglobin variants: fresh insights into malaria pathogenesis and protection". In: *Microbes and Infection* 14.10, pp. 851–862 (cit. on p. 6).

Falk, Nicole et al. (Aug. 2009). "Analysis of Plasmodium falciparum varGenes Expressed in Children from Papua New Guinea". In: *Journal of Infectious Diseases* 200.3, pp. 347–356 (cit. on pp. 15, 128).

Flick, Kirsten and Qijun Chen (Apr. 2004). "var genes, PfEMP1 and the human host". In: *Molecular and Biochemical Parasitology* 134.1, pp. 3–9 (cit. on p. 6).

Fowler, Elizabeth V et al. (Mar. 2002). "Genetic diversity of the DBL region in Plasmodium falciparum var genes among Asia-Pacific isolates". In: *Molecular and Biochemical Parasitology* 120.1, pp. 117–126 (cit. on p. 11).

Fowler, Elizabeth V et al. (Oct. 2006). "Physical Linkage to Drug Resistance Genes Results in Conservation of varGenes among West Pacific Plasmodium falciparumIsolates". In: *Journal of Infectious Diseases* 194.7, pp. 939–948 (cit. on p. 174).

Frank, Matthias and Kirk Deitsch (Aug. 2006). "Activation, silencing and mutually exclusive expression within the var gene family of Plasmodium falciparum". In: *International Journal for Parasitology* 36.9, pp. 975–985 (cit. on p. 8).

Frank, Matthias et al. (Aug. 2008). "Frequent recombination events generate diversity within the multi-copy variant antigen gene families of Plasmodium falciparum". In: *International Journal for Parasitology* 38.10, pp. 1099–1109 (cit. on pp. 18, 60, 89, 92, 125).

Freitas-Junior, L H et al. (2000). "Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum." In: *Nature* 407, pp. 1018–1022 (cit. on pp. 9, 18, 92, 93, 125, 126).

Fried, M and P E Duffy (June 1996). "Adherence of Plasmodium falciparum to Chondroitin Sulfate A in the Human Placenta". In: *Science* 272.5267, pp. 1502–1504 (cit. on p. 14).

Frith, M C et al. (Apr. 2010). "Incorporating sequence quality data into alignment improves DNA read mapping". In: *Nucleic acids research* 38.7, e100–e100 (cit. on p. 26).

Gaida, Annette et al. (2011). "Cloning of the repertoire of individual Plasmodium falciparum var genes using transformation associated recombination (TAR)." In: *PLoS ONE* 6, e17782 (cit. on p. 16).

Gao, Ling et al. (2005). "Meiotic recombination hotspots in eukaryotes". In: *Yi chuan Hereditas Zhongguo yi chuan xue hui bian ji* 27, pp. 641–650 (cit. on p. 93).

Gardner, M J et al. (Oct. 2002). "Genome sequence of the human malaria parasite Plasmodium falciparum". In: *Nature* 419.6906, pp. 498–511 (cit. on pp. 3, 9, 11–13, 16, 29, 104, 125).

Gething, Peter W et al. (Dec. 2011). "A new world malaria map: Plasmodium falciparum endemicity in 2010". In: *Malaria Journal* 10.1, p. 378 (cit. on p. 2).

Giladi, Eldar et al. (Oct. 2010). "Error Tolerant Indexing and Alignment of Short Reads with Covering Template Families". In: *Journal of Computational Biology* 17.10, pp. 1397–1411 (cit. on p. 26).

Greenwood, B M et al (2005). "Malaria". In: *Lancet* 365.9469, pp. 1487–1498 (cit. on p. 1).

Gupta, Sunetra et al. (Mar. 1999). "Immunity to non-cerebral severe malaria is acquired after one or two infections". In: *Nature medicine* 5.3, pp. 340–343 (cit. on p. 15).

Haldar, Kasturi and Narla Mohandas (May 2007). "Erythrocyte remodeling by malaria parasites". In: *Current Opinion in Hematology* 14.3, pp. 203–209 (cit. on p. 3).

Havlak, P et al. (Apr. 2004). "The Atlas Genome Assembly System". In: *Genome Research* 14.4, p. 721 (cit. on p. 23).

Hayton, Karen et al. (July 2008). "Erythrocyte Binding Protein PfRH5 Polymorphisms Determine Species-Specific Pathways of Plasmodium falciparum Invasion". In: *Cell Host and Microbe* 4.1, pp. 40–51 (cit. on p. 18).

Hernandez, D et al (May 2008). "De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer". In: *Genome Research* 18.5, pp. 802–809 (cit. on pp. 23, 26).

Hernandez-Rivas, Rosaura et al. (2010). "Telomeric Heterochromatin in Plasmodium falciparum". In: *Journal of Biomedicine and Biotechnology* 2010, pp. 1–12 (cit. on p. 9).

Hicks, Wade M (2010). "Mitotic Gene Conversion". In: *Science* 82 (cit. on p. 93).

Holliday, Robin (1964). "A mechanism for gene conversion in fungi". In: *Genetics Research* 5, pp. 282–304 (cit. on p. 92).

Huang, X (Sept. 2003). "PCAP: A Whole-Genome Assembly Program". In: *Genome Research* 13.9, pp. 2164–2170 (cit. on p. 23).

Huang, X and A Madan (Sept. 1999). "CAP3: A DNA sequence assembly program". In: *Genome Research* 9.9, pp. 868–877 (cit. on p. 23).

Hughes, Katie R et al. (Feb. 2009). "Continued cytoadherence of Plasmodium falciparum infected red blood cells after antimalarial treatment". In: *Molecular and Biochemical Parasitology* 169.2, pp. 71–78 (cit. on p. 6).

Huson, Daniel H et al. (Sept. 2002). "The greedy path-merging algorithm for contig scaffolding". In: *Journal of the ACM* 49.5, pp. 603–615 (cit. on p. 27).

Hviid, Lars (Oct. 2011). "The case for PfEMP1-based vaccines to protect pregnant women against Plasmodium falciparummalaria". In: *Cellular microbiology* 10.10, pp. 1405–1414 (cit. on p. 16).

Imelfort, Michael and David Edwards (Nov. 2009). "De novo sequencing of plant genomes using second-generation technologies". In: *Briefings in Bioinformatics* 10.6, pp. 609–618 (cit. on p. 26).

Iqbal, Zamin et al. (Jan. 2012). "De novo assembly and genotyping of variants using colored de Bruijn graphs". In: *Nature Genetics* 44.2, pp. 226–232 (cit. on p. 170).

Jackson, Andrew P et al. (Feb. 2012). "Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species". In: *PNAS* 109.9, pp. 3416–3421 (cit. on p. 93).

Jeck, WR et al (2007). "Extending assembly of short DNA sequences to handle error". In: *Bioinformatics* 23, pp. 2942–2944 (cit. on p. 26).

Jemmely, Noelle Yvonne et al. (May 2010). "Small variant surface antigens and Plasmodiumevasion of immunity". In: *Future Microbiology* 5.4, pp. 663–682 (cit. on p. 13).

Jensen, A T R (Apr. 2004). "Plasmodium falciparum Associated with Severe Childhood Malaria Preferentially Expresses PfEMP1 Encoded by Group A var Genes". In: *Journal of Experimental Medicine* 199.9, pp. 1179–1190 (cit. on pp. 15, 128).

Jiang, Hongying et al. (2011). "High recombination rates and hotspots in a Plasmodium falciparum genetic cross." In: *Genome Biology* 12, R33 (cit. on pp. 13, 93, 124, 126).

Kaestli, Mirjam et al. (May 2004). "Longitudinal Assessment of Plasmodium falciparum varGene Transcription in Naturally Infected Asymptomatic Children in Papua New Guinea". In: *The Journal of infectious diseases* 189.10, pp. 1942–1951 (cit. on pp. 15, 128).

Kaestli, Mirjam et al. (June 2006). "Virulence of Malaria Is Associated with Differential Expression of Plasmodium falciparum varGene Subgroups in a CaseControl Study". In: *The Journal of infectious diseases* 193.11, pp. 1567–1574 (cit. on pp. 15, 128).

Kalmbach, Yvonne et al. (July 2010). "Differential varGene Expression in Children with Malaria and Antidromic Effects on Host Gene Expression". In: *The Journal of infectious diseases* 202.2, pp. 313–317 (cit. on pp. 15, 128).

Kantibhattacharyya, M et al. (June 2004). "Molecular players of homologous recombination in protozoan parasites: implications for generating antigenic variation". In: *Infection, Genetics and Evolution* 4.2, pp. 91–98 (cit. on p. 18).

Kim, Pan-Gyu et al. (2008). "A Scaffold Analysis Tool Using Mate-Pair Information in Genome Sequencing". In: *Journal of Biomedicine and Biotechnology* 2008, pp. 1–8 (cit. on p. 27).

Kingsford, Carl et al. (2010). "Assembly complexity of prokaryotic genomes using short reads". In: *BMC Bioinformatics* 11.1, p. 21 (cit. on p. 24).

Kirchgatter, Karin and Hernando A del Portilo (Jan. 2002). "Association of severe noncerebral Plasmodium falciparum malaria in Brazil with expressed PfEMP1 DBL1 alpha sequences lacking cysteine residues." In: *Molecular Medicine* 8.1, pp. 16–23 (cit. on pp. 15, 128).

Kozarewa, Iwanka et al. (Apr. 2009). "Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes". In: *Nat. Methods* 6.4, pp. 291–295 (cit. on pp. 30, 31, 94).

Kraemer, Susan M and Joseph D Smith (Nov. 2003). "Evidence for the importance of genetic structuring to the structural and functional specialization of the Plasmodium falciparum var gene family". In: *Molecular microbiology* 50.5, pp. 1527–1538 (cit. on p. 17).

— (Aug. 2006). "A family affair: var genes, PfEMP1 binding, and malaria disease". In: *Current opinion in microbiology* 9.4, pp. 374–380 (cit. on pp. 8, 11, 14).

Kraemer, Susan M et al. (2007). "Patterns of gene recombination shape var gene repertoires in Plasmodium falciparum: comparisons of geographically diverse isolates." In: *BMC genomics* 8, p. 45 (cit. on pp. 9–11, 13, 16, 17, 34, 60, 89, 126, 128, 158).

Kyes, S A (Aug. 1999). "Rifins: A second family of clonally variant proteins expressed on the surface of red cells infected with Plasmodium falciparum". In: *Proceedings of the National Academy of Sciences* 96.16, pp. 9333–9338 (cit. on p. 3).

Kyes, S A et al. (Sept. 2007). "Antigenic Variation in Plasmodium falciparum: Gene Organization and Regulation of the var Multigene Family". In: *Eukaryotic Cell* 6.9, pp. 1511–1520 (cit. on pp. 9, 10).

Kyriacou, Helen M et al. (Dec. 2006). "Differential var gene transcription in Plasmodium falciparum isolates from patients with cerebral malaria compared to hyperparasitaemia". In: *Molecular and Biochemical Parasitology* 150.2, pp. 211–218 (cit. on pp. 15, 128).

Langmead, Ben et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In: *Genome Biology* 10.3, R25 (cit. on p. 35).

Lavstsen, Thomas et al. (2003). "Sub-grouping of Plasmodium falciparum 3D7 var genes based on sequence analysis of coding and non-coding regions". In: *Malaria Journal* 2.1, p. 27 (cit. on p. 13).

Lavstsen, Thomas et al. (2005). "Expression of Plasmodium falciparum erythrocyte membrane protein 1 in experimentally infected humans". In: *Malaria Journal* 4.21 (cit. on pp. 15, 128).

Lavstsen, Thomas et al. (May 2012). "Plasmodium falciparum erythrocyte membrane protein 1 domain cassettes 8 and 13 are associated with severe malaria in children". In: *PNAS* 109.26, E1791–E1800 (cit. on pp. 15, 129, 174).

Le Roch, K G et al. (Feb. 2012). "Genomics and integrated systems biology in Plasmodium falciparum: a path to malaria control and eradication". In: *Parasite Immunology* 34.2-3, pp. 50–60 (cit. on p. 20).

Leech, J H et al. (June 1984). "Identification of a strain-specific malarial antigen exposed on the surface of Plasmodium falciparum-infected erythrocytes". In: *Journal of Experimental Medicine* 159.6, pp. 1567–1575 (cit. on p. 6).

Li, H and R Durbin (July 2009). "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics (Oxford, England)* 25.14, pp. 1754–1760 (cit. on p. 35).

Li, Heng et al. (Aug. 2009a). "The Sequence Alignment/Map format and SAMtools." In: *Bioinformatics* 25.16, pp. 2078–2079 (cit. on pp. 36, 62, 96).

Li, Ruiqiang et al. (Dec. 2009b). "The sequence and de novo assembly of the giant panda genome". In: *Nature* 463.7279, pp. 311–317 (cit. on pp. 23, 24, 26).

Li, Ruiqiang et al. (Feb. 2010). "De novo assembly of human genomes with massively parallel short read sequencing". In: *Genome Research* 20.2, pp. 265–272 (cit. on pp. 23, 32).

Lieber, Michael R (2010). "The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End Joining Pathway". In: *Annual review of biochemistry* 79, p. 181 (cit. on p. 18).

MacPherson, G G et al. (June 1985). "Human cerebral malaria. A quantitative ultrastructural analysis of parasitized erythrocyte sequestration." In: *The American Journal of Pathology* 119.3, p. 385 (cit. on p. 4).

Maier, Alexander G et al. (June 2009). "Malaria parasite proteins that remodel the host erythrocyte". In: *PLoS ONE* 7.5, pp. 341–354 (cit. on p. 3).

Maizels, Nancy (2005). "Immunoglobulin gene diversification." In: *Annual review of genetics* 39, pp. 23–46 (cit. on p. 93).

Manske, Magnus et al. (July 2012). "Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing." In: *Nature* (cit. on pp. 17, 168, 172).

Mardis, Elaine R (Mar. 2008). "The impact of next-generation sequencing technology on genetics". In: *Trends in Genetics* 24.3, pp. 133–141 (cit. on p. 20).

Miller, Louis et al. (Feb. 2002). "The pathogenic basis of malaria". In: *Nature* 415.6872, pp. 673–679 (cit. on pp. 4, 8).

Milner Jr, Danny A et al. (Dec. 2008). "Severe malaria in children and pregnancy: an update and perspective". In: *PLoS ONE* 24.12, pp. 590–595 (cit. on p. 4).

Montgomery, Jacqui et al. (Aug. 2007). "Differential var gene expression in the organs of patients dying of falciparum malaria". In: *Molecular microbiology* 65.4, pp. 959–967 (cit. on pp. 8, 15, 128).

Mugasa, Joseph et al. (2012). "Genetic diversity of expressed Plasmodium falciparum var genes from Tanzanian children with severe malaria". In: *Malaria Journal* 11.1, p. 230 (cit. on p. 16).

Mullikin, J C and Z Ning (Dec. 2002). "The Phusion Assembler". In: *Genome Research* 13.1, pp. 81–90 (cit. on p. 23).

Murray, Christopher Jl et al. (2012). "Global malaria mortality between 1980 and 2010: a systematic analysis". In: *The Lancet* 379, pp. 413–431 (cit. on p. 1).

Myers, E W (Jan. 1995). "Toward simplifying and accurately formulating fragment assembly". In: *Journal of computational biology : a journal of computational molecular cell biology* 2.2, pp. 275–290 (cit. on p. 22).

Myers, Eugene W et al. (Mar. 2000). "A Whole-Genome Assembly of Drosophila". In: *Science* 287.5461, pp. 2196–2204 (cit. on p. 23).

Narzisi, Giuseppe and Bud Mishra (2011). "Comparing de novo genome assembly: the long and short of it." In: *PLoS ONE* 6, e19175 (cit. on p. 23).

Neghina, Raul et al. (Dec. 2010). "Malaria, a journey in time: in search of the lost myths and forgotten stories." In: *The American journal of the medical sciences* 340.6, pp. 492–498 (cit. on p. 1).

Newbold, C I (1999). "Antigenic variation in Plasmodium falciparum: mechanisms and consequences." In: *Current opinion in microbiology* 2, pp. 420–425 (cit. on p. 6).

Newbold, Chris et al. (June 1999). "Cytoadherence, pathogenesis and the infected red cell surface in Plasmodium falciparum". In: *International Journal for Parasitology* 29.6, pp. 927–937 (cit. on p. 6).

Nielsen, K M (May 2003). "Gene Conversion as a Source of Nucleotide Diversity in Plasmodium falciparum". In: *Molecular biology and evolution* 20.5, pp. 726–734 (cit. on p. 93).

Nielsen, Morten A et al. (Dec. 2002). "Plasmodium falciparum Variant Surface Antigen Expression Varies Between Isolates Causing Severe and Nonsevere Malaria and Is Modified by Acquired Immunity". In: *Plasmodium falciparum Variant Surface Antigen Expression Varies Between Isolates Causing Severe and Nonsevere Malaria and Is Modified by Acquired Immunity* 168, pp. 3444–3450 (cit. on pp. 15, 128).

Ning, Z et al. (Oct. 2001). "SSAHA: a fast search method for large DNA databases." In: *Genome Research* 11.10, pp. 1725–1729 (cit. on p. 96).

Normark, Johan et al. (Oct. 2007). "PfEMP1-DBL1Œ± amino acid motifs in severe disease states of Plasmodium falciparum malaria". In: *PNAS* 104.40, pp. 15835–15840 (cit. on pp. 15, 128).

Otto, Thomas D et al. (July 2010a). "Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology." In: *Bioinformatics* 26.14, pp. 1704–1707 (cit. on pp. 22, 26, 95).

Otto, Thomas D et al. (2010b). "New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq." In: *Molecular microbiology* 76.1, pp. 12–24 (cit. on p. 10).

Oyola, Samuel O et al. (Jan. 2013). "Efficient Depletion of Host DNA Contamination in Malaria Clinical Sequencing". In: *Journal of Clinical Microbiology* 51.3, pp. 745–751 (cit. on p. 130).

Ozarkar, Aarti et al. (2009). "Analysis of PfEMP1—var Gene Sequences in Different Plasmodium falciparum Malarial Parasites". In: *Scholarly Research Exchange* 2009, pp. 1–10 (cit. on p. 16).

Pain, A et al. (Oct. 2008). "The genome of the simian and human malaria parasite Plasmodium knowlesi." In: *Nature* 455.7214, pp. 799–803 (cit. on p. 2).

Pasternak, Noa D and Ron Dzikowski (July 2009). "PfEMP1: An antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite Plasmodium falciparum". In: *The International Journal of Biochemistry & Cell Biology* 41.7, pp. 1463–1466 (cit. on pp. 3, 6, 7).

Pop, M (Dec. 2003). "Hierarchical Scaffolding With Bambus". In: *Genome Research* 14.1, pp. 149–159 (cit. on pp. 23, 27).

Pop, Mihai (July 2009). "Genome assembly reborn: recent computational challenges". In: *Briefings in Bioinformatics* 10.4, pp. 354–366 (cit. on pp. 24, 26, 57).

Pop, Mihai and Steven L Salzberg (Mar. 2008). "Bioinformatics challenges of new sequencing technology". In: *Trends in Genetics* 24.3, pp. 142–149 (cit. on p. 24).

Prabakaran, Ponraj et al. (2011). "454 antibody sequencing - error characterization and correction". In: *BMC research notes* 4, p. 404 (cit. on p. 26).

Ranson, Hilary et al. (2009). "Insecticide resistance in Anopheles gambiae: data from the first year of a multi-country study highlight the extent of the problem". In: *Malaria Journal* 8.1, p. 299 (cit. on p. 2).

Rask, Thomas S et al. (Sept. 2010). "Plasmodium falciparum Erythrocyte Membrane Protein 1 Diversity in Seven Genomes – Divide and Conquer". In: *PLoS computational biology* 6.9, e1000933 (cit. on pp. 11, 15, 16, 34, 60, 128, 129, 174).

RBM, WHO (2010). "Roll Back Malaria Partnership: Malaria in Africa". In: (cit. on p. 1).

Recker, Mario et al. (2011). "Antigenic variation in Plasmodium falciparum malaria involves a highly structured switching pattern." In: *PLoS Pathogens* 7, e1001306 (cit. on p. 8).

Roberts, D J et al. (June 1992). "Rapid switching to multiple antigenic and adhesive phenotypes in malaria". In: *Nature* 357.6380, pp. 689–692 (cit. on p. 8).

Rogerson, Stephen J et al. (Dec. 2007). "Malaria in Pregnancy: Linking Immunity and Pathogenesis to Prevention". In: *Am J. Tropy. Meed. Hyg* 77.6, pp. 14–22 (cit. on p. 15).

Rottmann, Matthias et al. (Dec. 2006). "Differential Expression of var Gene Groups Is Associated with Morbidity Caused by Plasmodium falciparum Infection in Tanzanian Children". In: *Infection and immunity* 74.7, pp. 3904–3911 (cit. on pp. 15, 128).

Rowe, J Alexandra et al. (May 2009). "Adhesion of Plasmodium falciparum-infected erythrocytes to human cells: molecular mechanisms and therapeutic implications". In: *Expert Reviews in Molecular Medicine* 11, e16 (cit. on p. 8).

Ruiqiang Li, Hongmei Zhu Jue Ruan Wubin Qian Xiaodong Fang Zhongbin Shi Yingrui Li Shengting Li Gao Shan Karsten Kristiansen Songgang Li Huanming Yang Jian Wang Jun Wang (Feb. 2010). "De novo assembly of human genomes with massively parallel short read sequencing". In: *Genome Research* 20.2, pp. 265–272 (cit. on pp. 24, 26, 32).

Salanti, A (Nov. 2004). "Evidence for the Involvement of VAR2CSA in Pregnancy-associated Malaria". In: *Journal of Experimental Medicine* 200.9, pp. 1197–1203 (cit. on p. 14).

Salzberg, Steven L et al. (Sept. 2008). "Gene-Boosted Assembly of a Novel Bacterial Genome from Very Short Reads". In: *PLoS Computational Biology* 4.9, e1000186 (cit. on p. 169).

Samarakoon, Upeka et al. (2011). "High-throughput 454 resequencing for allele discovery and recombination mapping in Plasmodium falciparum". In: *BMC genomics* 12, p. 116 (cit. on p. 93).

San Filippo, Joseph et al. (June 2008). "Mechanism of Eukaryotic Homologous Recombination". In: *Annual review of biochemistry* 77.1, pp. 229–257 (cit. on pp. 18, 19).

Santoyo, Gustavo and David Romero (2005). "Gene conversion and concerted evolution in bacterial genomes." In: *FEMS microbiology reviews* 29, pp. 169–183 (cit. on p. 93).

Scherf, A et al. (1998). "Antigenic variation in malaria : in situ switching , relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum". In: *The EMBO Journal* 17, pp. 5418–5426 (cit. on pp. 6, 8).

Scherf, Artur et al. (2008). "Antigenic Variation in Plasmodium falciparum". In: *Annual review of microbiology* 62.1, pp. 445–470 (cit. on pp. 8, 9).

Schmid-Hempel, Paul (Feb. 2011). *Evolutionary Parasitology*. The Integrated Study of Infections, Immunology, Ecology, and Genetics. Oxford University Press (cit. on p. 175).

Sharma, Yagya D (Jan. 1991). "Knobs, knob proteins and cytoadherence in falciparum malaria". In: *Current Opinion in Microbiology* 23.9, pp. 775–789 (cit. on p. 6).

Shendure, Jay and Hanlee Ji (Oct. 2008). "Next-generation DNA sequencing". In: *Nature* 26.10, pp. 1135–1145 (cit. on p. 22).

Simpson, Jared T and Richard Durbin (Apr. 2012). "Efficient de novo assembly of large genomes using compressed data structures." In: *Genome Research* 22.3, pp. 549–556 (cit. on pp. 23, 170).

Simpson, JT et al (June 2009). "ABySS: a parallel assembler for short read sequence data". In: *Genome Research* 19.6, pp. 1117–1123 (cit. on pp. 23, 24, 26).

Smith, J D et al. (July 1995). "Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes". In: *Cell* 82.1, pp. 101–110 (cit. on p. 6).

Su, X Z et al. (July 1995). "The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes". In: *Cell* 82.1, pp. 89–100 (cit. on p. 6).

Sue Kyes et al. (Dec. 2001). "Antigenic variation at the infected red cell surface in malaria". In: *Annual review of microbiology* 55, pp. 657–707 (cit. on pp. 6, 8).

Sutton, Granger G et al. (Jan. 1995). "TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects". In: *Genome Science and Technology* 1.1, pp. 9–19 (cit. on p. 23).

Swain, Martin T et al. (June 2012). "A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs". In: *Nature Protocols* 7.7, pp. 1260–1284 (cit. on p. 177).

Swamy, L et al. (Apr. 2011). "Plasmodium falciparum var Gene Silencing Is Determined by cis DNA Elements That Form Stable and Heritable Interactions". In: *Eukaryotic Cell* 10.4, pp. 530–539 (cit. on p. 9).

Symington, L S et al. (1991). "Genetic Analysis of a Meiotic Recombination Hotspot on Chromosome III of Saccharomyces Cerevisiae". In: *Genetics* 128, pp. 717–727 (cit. on p. 93).

Szostak, Jack W et al. (May 1983). "The double-strand-break repair model for recombination". In: *Cell* 33.1, pp. 25–35 (cit. on p. 18).

Taylor, H M et al. (Jan. 2000a). "A study of var gene transcription in vitro using universal var gene primers". In: *Molecular and Biochemical Parasitology* 105.1, pp. 13–23 (cit. on pp. 10, 16, 93, 126).

Taylor, H M et al. (Oct. 2000b). "Var gene diversity in Plasmodium falciparum is generated by frequent recombination events". In: *Molecular and Biochemical Parasitology* 110.2, pp. 391–397 (cit. on pp. 10, 11, 18, 60).

Taylor, Helen M et al. (Jan. 2000c). "A study of var gene transcription in vitro using universal var gene primers". In: *Molecular and Biochemical Parasitology* 105.1, pp. 13–23 (cit. on p. 125).

Tilley, Leann et al. (June 2011). "The Plasmodium falciparum-infected red blood cell". In: *The International Journal of Biochemistry & Cell Biology* 43.6, pp. 839–842 (cit. on p. 3).

Treangen, Todd J and Steven L Salzberg (Nov. 2011). "Repetitive DNA and next-generation sequencing: computational challenges and solutions". In: *Nature Reviews Genetics* 13.1, pp. 36–46 (cit. on p. 24).

Trimnell, Adama R et al. (Aug. 2006). "Global genetic diversity and evolution of var genes associated with placental and severe childhood malaria". In: *The International Journal of Biochemistry & Cell Biology* 148.2, pp. 169–180 (cit. on p. 11).

Tsai, Isheng J et al. (2010). "Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps". In: *Genome Biology* 11.4, R41 (cit. on p. 69).

Voss, Till S et al. (Feb. 2006). "A var gene promoter controls allelic exclusion of virulence genes in Plasmodium falciparum malaria". In: *Nature* 439, pp. 1004–1008 (cit. on p. 9).

Walliker, D (Nov. 2005). "The hitchhiker's guide to malaria parasite genes". In: *Trends in Parasitology* 21.11, pp. 489–493 (cit. on p. 173).

Walliker, David et al. (July 1987). "Genetic Analysis of the Human Malaria Parasite Plasmodiumfalciparum". In: *Science* 236.4809, pp. 1661–1666 (cit. on pp. 18, 93, 94).

Warimwe, GM et al (2009). "Plasmodium falciparum var gene expression is modified by host immunity". In: *Proc Natl Acad Sci U S A.* 106.51, pp. 21801–21806 (cit. on p. 15).

Warren, Ren L et al. (Dec. 2006). "Assembling millions of short DNA sequences using SSAKE". In: *Bioinformatics (Oxford, England)* 23.4, pp. 500–501 (cit. on p. 26).

Webb, James L A (Dec. 2008). *Humanity's Burden*. A Global History of Malaria. Cambridge University Press (cit. on p. 1).

Wei-Chun Kao, Kristian Stevens Yun S Song (Oct. 2009). "BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing". In: *Genome Research* 19.10, p. 1884 (cit. on p. 21).

wellems, T E et al. (May 1990). *Chloroquine resistance not linked to mdr-like genes in a Plasmodium falciparum cross*. URL: `http://www.nature.com/nature/journal/v345/n6272/pdf/345253a0.pdf` (cit. on p. 18).

WHO (2011). "World Malaria Report 2011". In: (cit. on pp. 1, 2).

Zerbino, Daniel and Ewan Birney (Mar. 2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs". In: *Genome Research* 18.5, gr.074492.107–829 (cit. on pp. 23, 24, 26, 31, 67, 102).

Zhao, Xiaohong et al. (Nov. 2010). "EDAR: An Efficient Error Detection and Removal Algorithm for Next Generation Sequencing Data". In: *Journal of Computational Biology* 17.11, pp. 1549–1560 (cit. on p. 26).