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).

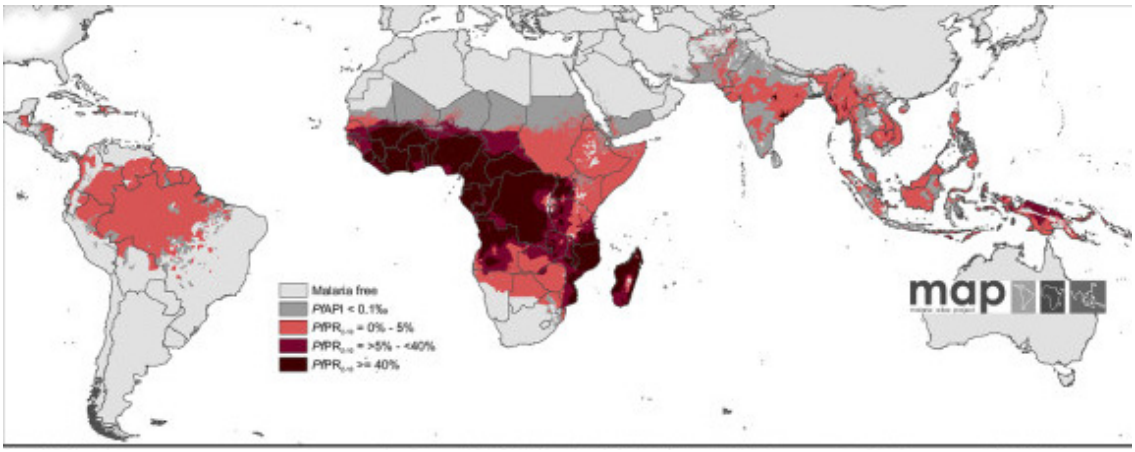

Figure 1.1: The spatial distribution of *P. falciparum* malarial endemicity in 2010. The map shows endemicity predictions based on *P. falciparum* Parasite Rates (*Pf*PR). Predictions were categorized as low risk *Pf*PR_2-10 $\leq$ 5% light red; intermediate risk *Pf*PR_2-10>5% to <40%, medium red; and 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.

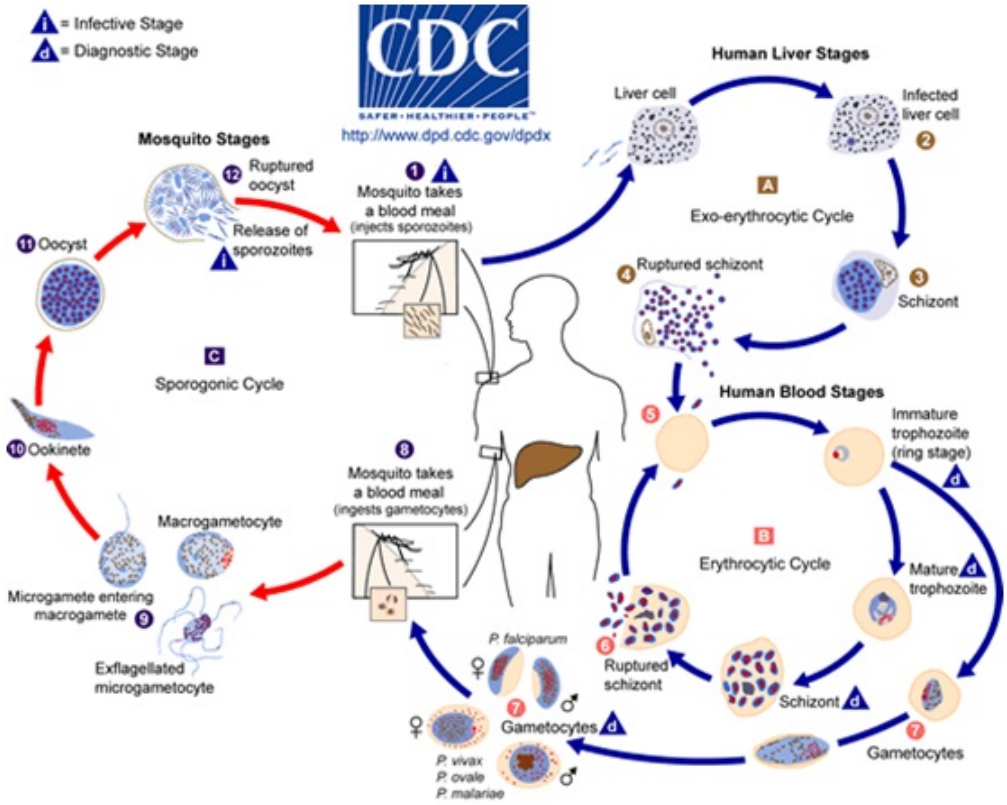

Figure 1.2: The life cycle of the human malaria parasite *P. falciparum* within two hosts (Taken from CDC: `http://www.cdc.gov/malaria/about/biology`)

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

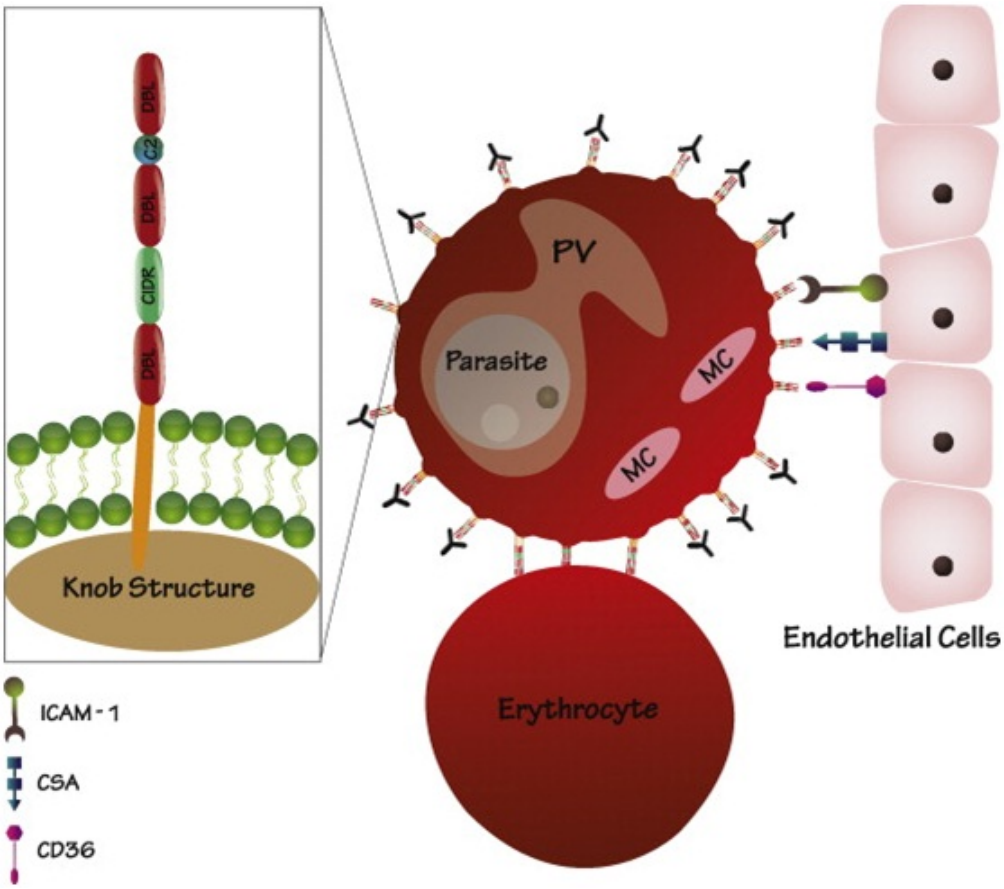

Figure 1.3: *Pf*EMP1 mediates adhesion of the infected RBC. *Pf*EMP1 is expressed by the malaria parasite *P. falciparum* on the knobs formed on the surface of infected erythrocytes. The variable extracellular regions DBLs and CIDR mediate adhesion through binding to several endothelial receptors such 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

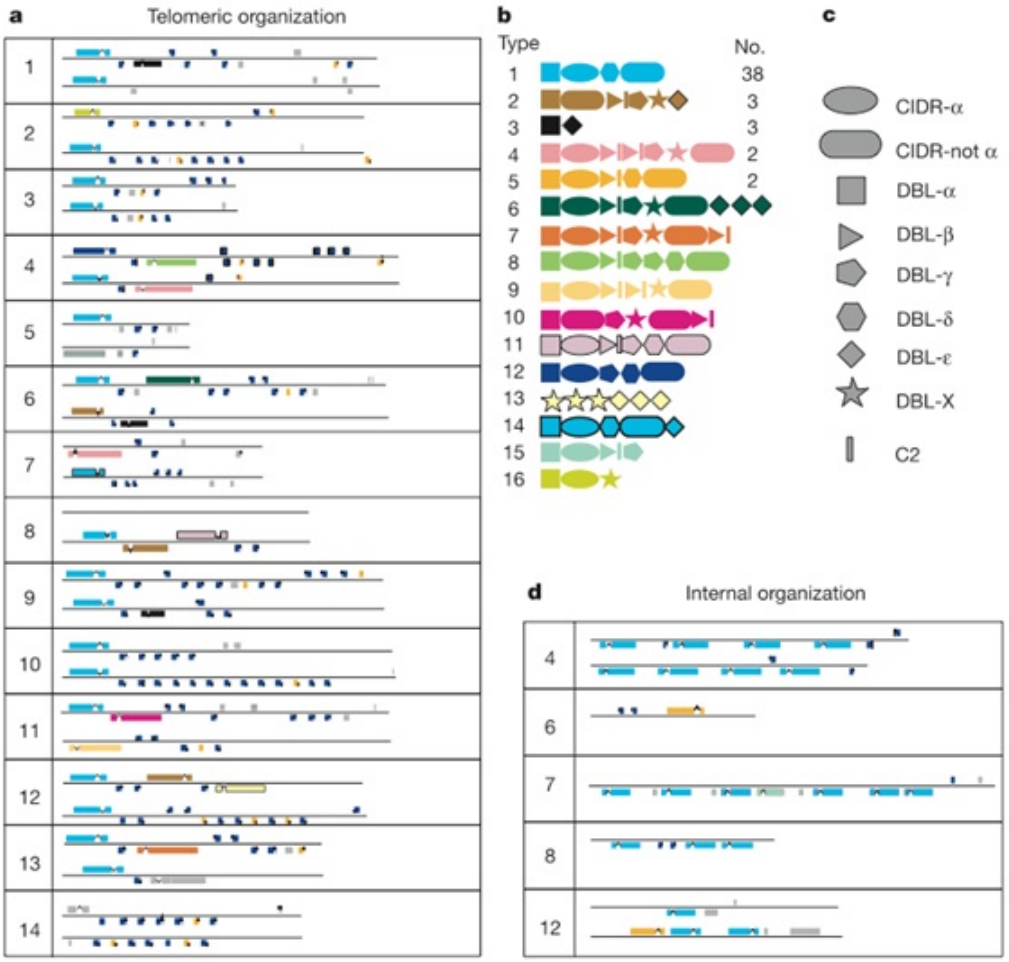

Figure 1.4: Organization of multi-gene families in *P. falciparum*. a, Telomeric regions of all chromosomes showing the relative positions of members of the multi-gene families: *rif* (blue) *stevor* (yellow) and *var* (colour coded as indicated; see b and c). Grey boxes represent pseudogenes or gene 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.

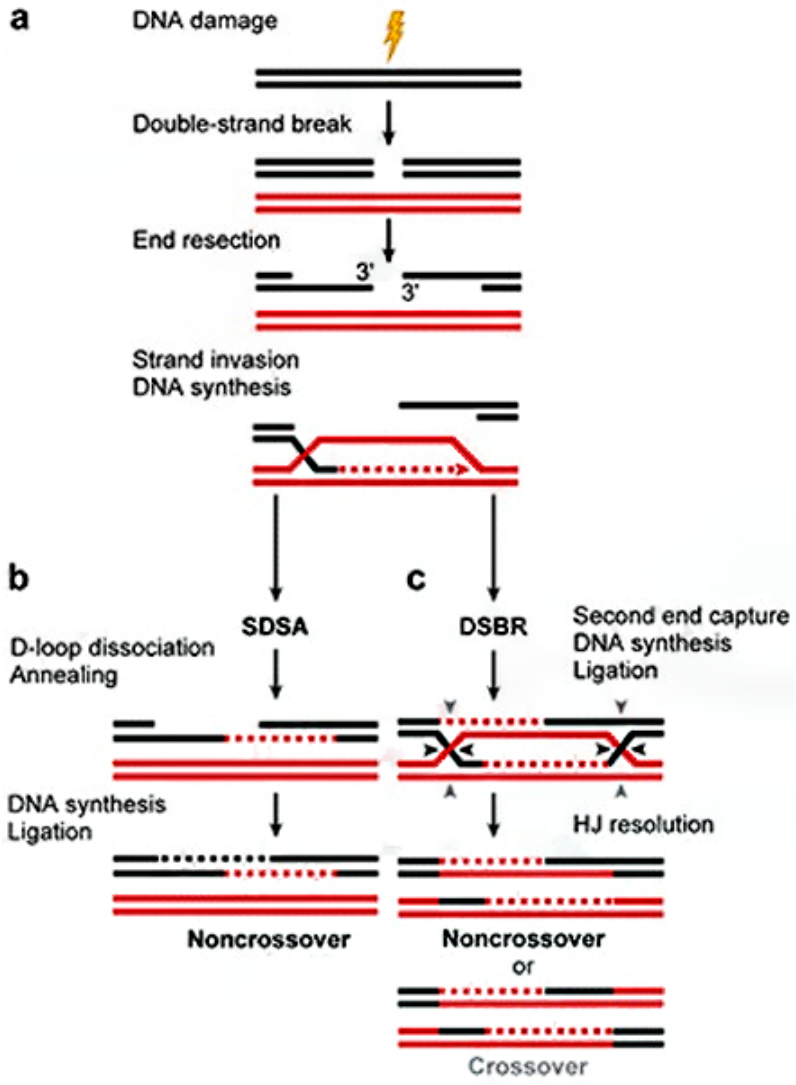

Figure 1.5: Pathways of DNA double-strand break repair by homologous recombination. Double-strand breaks (DSBs) can be repaired by distinctive homologous recombination (HR) pathways, such as synthesis-dependent strand annealing (SDSA) and double-strand break repair (DSBR). (a) After DSB formation, the 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-

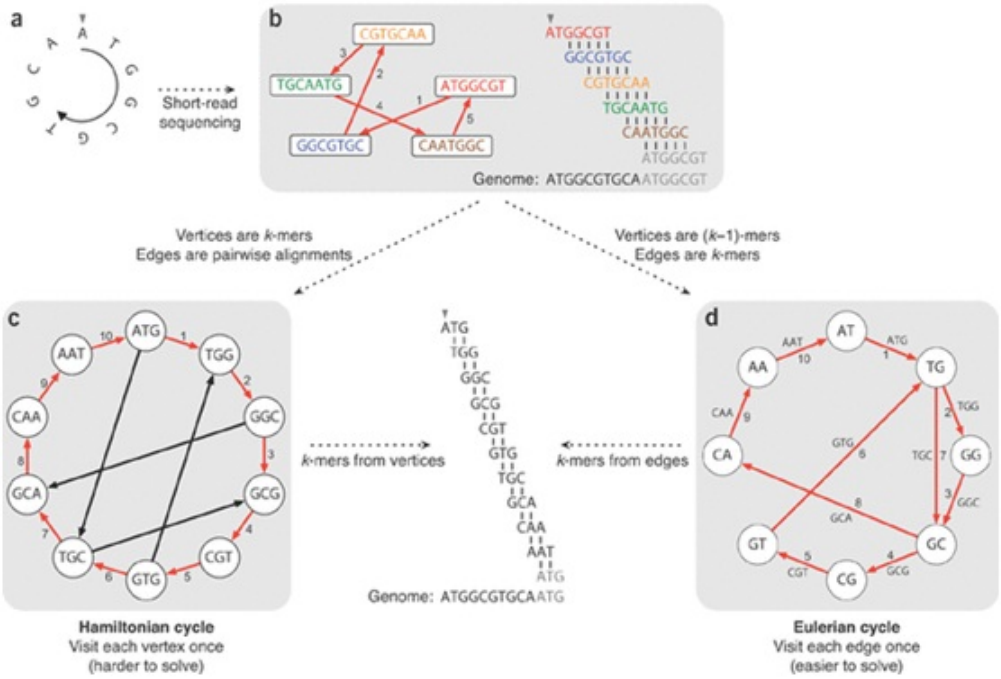

Figure 1.6: Example of graph-based assembly approaches (a) An example of a small circular genome. (b) Most assemblers developed to assemble reads from traditional Sanger sequencing platforms represent reads as nodes in a graph, and edges represented alignments between reads. Walking along a Hamiltonian 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.