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`)