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

Contents **vii**

List of Figures **xi**

List of Tables **xv**

Abbreviations **xviii**

1 Introduction **1**

1.1 Malaria 1

1.1.1 Overview 1

1.1.2 The Parasite: *P. falciparum* 2

1.2 *Plasmodium falciparum* erythrocyte membrane protein 1 (*Pf*EMP1) 6

1.2.1 Cytoadherence 6

1.2.2 Antigenic variation 8

1.2.3 Regulation of gene expression 8

1.2.4 Organization, chromosomal location and grouping of *var* genes 9

1.2.5 Polymorphism, sequence diversity and mechanisms of generating diversity in *var* genes 16

1.3 New frontiers and existing challenges 20

1.4 Next generation sequencing technologies and short read assembly 20

1.4.1 Second generation sequencing technologies 20

1.4.2 Short read assembly 22

1.5 Overview of thesis 27

2 Evaluating existing short read assembly methods 29

2.1 Introduction 29

2.2 Methods 30

2.2.1 Library preparation and sequencing 30

2.2.2 Choice of short read assemblers 31

2.2.3 Assembly benchmarking using error free in silico reads . 33

2.2.4 Evaluating the velvet assembler on real and simulated reads 35

2.2.5 Evaluating mapping based assembly approaches 36

2.2.6 Sequencing errors 37

2.3 Results 38

2.3.1 Comparing *de novo* assembly tools 38

2.3.2 Velvet assembly of whole genome data 41

2.3.3 Velvet assembly on Chromosome 1 of 3D7 41

2.3.4 Velvet assembly on *var* genes 42

2.3.5 Reference guided assembly 46

2.3.6 Understanding sequencing errors in *P. falciparum* 50

2.4 Discussion 55

3 New approaches to assemble *var* genes from short reads 60

3.1 Introduction 60

3.2 Methods: Proposed assembly approach 62

3.2.1 Preprocessing sequence data 62

3.2.2 Generating seed contigs 66

3.2.3 Iterative scaffolding and extension 68

3.2.4 Evaluating the assembly approach 69

3.3 Results 73

3.3.1 Defining regions of interest 73

3.3.2 Evaluating the new assembly approach using culture-adapted samples 74

3.3.3 Evaluating new assembly approach using clinical samples 79

3.3.4 Additional evaluations 83

3.4 Discussion 89

4 Understanding mechanisms of *var* gene diversity using next generation sequencing 92
4.1 Introduction 92
4.2 Methods 94
4.2.1 Sample preparation and sequencing 94
4.2.2 Reference genomes 95
4.2.3 Alignment and processing of sequence data 96
4.2.4 Genome-wide scan for recombination breakpoints 97
4.3 Results 104
4.3.1 Sequence data and mapping to reference genomes 104
4.3.2 Detecting genome wide crossover events 105
4.3.3 Signature of recombination in *var* genes 111
4.4 Discussion 124

5 Assembly of *var* genes from clinical samples 128
5.1 Introduction 128
5.2 Methods 131
5.2.1 Sequence data 131
5.2.2 Initial Motifs and iterative assembly of clinical samples . 131
5.2.3 QC and filtering 132
5.2.4 Similarity between *var*-contigs 134
5.2.5 Network analysis and clustering 135
5.3 Results 137
5.3.1 Samples and sequence data 137
5.3.2 Initial Motifs 137
5.3.3 Iterations and additional motifs 138
5.3.4 Results of the initial assembly 139
5.3.5 Initial quality control steps 142
5.3.6 A first look at the motif sharing *var*-contigs 147
5.3.7 Using full length sequences 147
5.3.8 Clustering *var*-contigs and detecting population structure 159
5.4 Discussion 165

6 Final discussion, conclusions and future directions 167

6.1 Assembly of *var* genes 168

6.2 *Var* gene diversity via ectopic gene conversion 171

6.3 Global diversity of *var* genes 171

6.3.1 The first full-length assembly of *var* genes from clinical samples 171

6.3.2 Discovery of unexpected and highly conserved *var*-contigs172

6.4 Future directions 177

6.5 Publications and press releases 177

Appendices 179

References 185

List of Figures

1.1 The spatial distribution of *P. falciparum* malarial endemicity in 2010 2

1.2 The life cycle of the human malaria parasite *P. falciparum* within two hosts 5

1.3 *Pf*EMP1 mediate adhesion of the infected RBC 7

1.4 Organization of multi-gene families in *P. falciparum* 12

1.5 Pathways of DNA double-strand break repair by homologous recombination 19

1.6 Example of graph-based assembly approaches 25

2.1 Comparing SOAPdenovo and Velvet on synthetic reads simulated from *var* genes of the 3D7 genome 39

2.2 The effect of sequencing errors and uneven coverage on assembly of *var* genes 43

2.3 A histogram of shared sequences in 3D7 *var* genes identified by a pairwise blast alignment 44

2.4 Visualising the assembly graph of *var* genes from a velvet assembly (real reads, k=61) 45

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 47

2.6 Read mapping coverage plots over *var* genes on the left subtelomere of Chromosome 8 48

2.7 Number of reads mapped per kb over *var* genes of the 3D7 genome 49

2.8 Overall error rates at low (*Q*5), medium (*Q*15) and high (*Q*25) quality bins for forward and reverse reads 51

2.9 Error rates per-cycle for Illumina's GA2 and HiSeq platforms at a quality cutoff of 5 (Q5) 52
2.10 Error rates per-cycle of Illumina's GA2 and HiSeq platforms at a quality cutoff of 25 (Q25) 53
2.11 Overall substitution patterns of Illumina reads from six libraries 55

3.1 A work flow diagram of the iterative assembly approach developed to assemble *var* genes 63
3.2 A working definition of subtelomeric regions 73
3.3 Size distribution of subtelomeric regions in the reference genome 3D7 74
3.4 The cumulative number of shared motifs for 10 iterations of the four culture-adapted samples 3D7, IT, HB3 and DD2 76
3.5 Assembly statistics of the four lab adapted samples after 10 iterations 77
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 80
3.7 Assembly statistics of 50 clinical samples 81
3.8 A histogram of the number of contigs with DBLα(*var*-contigs) per sample for the 50 clinical isolates 82
3.9 Comparing ORFs obtained from individual and mixed assembly by varying the proportion of ORFs covered. 87
3.10 An ACT view of read coverage over *var*-contigs generated from individual and mixed assemblies 88

4.1 Breakpoint detection using SNPs and paired end coverage . . . 99
4.2 Patterns of Paired end coverage (PEC) over a 3D7 *var* gene with/without recombination 103
4.3 Visualising SNPs of the five progeny using the 3D7 parent as a reference 105
4.4 Comparative SNP-map of Chromosome 3 107
4.5 Genome wide breakpoints per window size 109
4.6 Breakpoints per chromosome at a window size of 20 kb (sliding by 10 kb) for the five progeny X1 to X5 110

4.7 Per-chromosome inheritance patterns for each progeny and average inheritance with /out progeny 2 112
4.8 Paired and Orphaned read coverage over 3D7 *var* genes 114
4.9 Chromosome view of *var* genes for progeny 4 and 5 115
4.10 Subtelomeric *var* genes involved in ectopic recombination in progeny 4 and 5 117
4.11 Evidence of gene conversion in progeny 4 (X4) from paired end coverage analysis 118
4.12 Evidence of gene conversion in progeny 5 (X5) from paired end coverage analysis 119
4.13 *De novo* assembly of reads that map to PFA0765c and PFB0010w in progeny 4 (X4) 121
4.14 *De novo* assembly of reads that map to PFA0005w and PFB1055c in progeny 5 (X5) 122

5.1 Iterative assembly work flow for *var* genes in clinical samples . 133
5.2 A global map of clinical samples used in this chapter 137
5.3 Assembly statistics of the initial 743 samples 139
5.4 Scatter plots of non-core read counts with the four assembly statistics 141
5.5 Box plots showing assembly stats by geographical region 143
5.6 Number of *var*-contigs by country of origin 144
5.7 Density plots of ORF sizes for clinical samples, 3D7 and IT . . . 146
5.8 Grouping $\sim 50,000 var - contigs$ 148
5.9 A histogram of samples and populations that share 10 aa long motifs 149
5.10 Frequency of shared motifs 150
5.11 Motif sharing *var*-contigs from a pmatch analysis of all-vs-all *var*-contigs 151
5.12 Artemis view of non-core reads from four samples (PA0036, PA0020, PH0136 and Control sample) aligned to *var*-contigs of PA0036 153
5.13 A screen shot of PF3D7_0632500 from PlasmoDB 153

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 . . 154
5.15 A scatter plot of match length and percent identity for the *var*-contig PA0036_VAR_1 156
5.16 Match length for different cutoff values of percent identity using all *var*-contigs 157
5.17 Amino acid identity network of *var*-contigs from 424 samples . 161
5.18 A network of *var*-contigs from Cambodian samples 163
5.19 Scatter plots of cluster size vs the number of unique samples represented in each cluster 164

List of Tables

2.1 Samples used to evaluate existing assembly approaches and determine error rates of the Illumina sequencing technology in *P. falciparum* . . . 32

2.2 Assembly statistics of Velvet and SOAPdenovo at different *k-mer* sizes . . . 40

2.3 Assembly statistics of contigs with the DBLα*domain* . . . 40

2.4 Whole genome Velvet assembly of real data from two PCR-free libraries of 3D7 . . . 41

2.5 Assembly results of simulated and real reads on chromosome 1 of 3D7 . . . 42

3.1 A summary of the four culture-adapted samples and non-core reads used to evaluate the iterative assembly approach . . . 75

3.2 Iterative assembly results for *var* genes of the 3D7 genome . . . 78

3.3 Coverage of *var* genes in the 3D7 genome . . . 78

3.4 Clinical samples used to test the iterative assembly appoach . . . 83

3.5 Comparing iterative assembly with *de novo* assembly using 5 clinical samples . . . 84

3.6 Assembly statistics of the five clinical samples using the iterative approach . . . 84

3.7 Comparing individual assembly with mixed assembly of four clinical samples: Assembly statistics . . . 85

3.8 Comparing ORFs of individually assembled contigs with ORFs of the mixed assembly . . . 86

4.1 Raw reads and mapping statistics to 3D7, HB3 and combined reference genomes . . . 104
4.2 Number of cross-over events per chromosome . . . 106
4.3 Assigning *var* genes of the progeny to either the 3D7 or HB3 parent based on a visual inspection of paired read coverage . . . 116
4.4 Genes bridged by mate-pairs: Number of reads where mates map to different genes (mismatch ≤ 2) . . . 123

5.1 Samples used for initial assembly of *var* genes in clinical samples 138
5.2 Summary of assembly results for the initial 743 samples . . . 142
5.3 A list of samples used for initial assembly, pmatch /BLAST analysis and social network analysis . . . 159

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